Bayesian Networks for spatio-temporal integrated catchment assessment

by

Chiedza Dondo

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Civil Engineering

FACULTY OF ENGINEERING AND THE BUILT ENVIRONMENT

UNIVERSITY OF CAPE TOWN

September 2010

ACKNOWLEDGEMENTS

I would like to thank the supervisors for all their hard work and input, for being eternally positive and always encouraging me when the going got tough and the results I got were not favourable. Thanks to all of them for holding my hand to the finishing line and for not letting me quit when I lost faith in my abilities.

Thanks to the staff and students from the Catchment Research and Management Team at the Australian National University: Prof A. Jakeman, Dr. J. Ticehurst, Dr. C. Pollino, Dr W. Merritt, Dr. B. Croke and Prof. J. Norton, for their assistance and input during my visit and treating me like part of their team.

I would like to acknowledge the Water Research Commission and the Council for Geoscience for funding this research and taking a chance on a new idea. To the staff members who contributed, either academically or logistically, I say thank you.

To my parents, thank you for making me the hardworking, persistent person that I am and allowing me to pursue the career of my choice. Thank you for letting me grow, for standing by and supporting the decisions I have made over the years, some of which were not commendable

To my husband, thanks for being patient with an often emotional and stressed student and providing valuable input and support when needed. Lastly, thank you to my sisters, my brother, family, friends, mentors and strangers that have encouraged and inspired me throughout this long and arduous journey.

ABSTRACT

Author: Chiedza Dondo[1]

Title: Bayesian Networks for spatio-temporal integrated catchment assessment

Supervisors: Dr. U.K. Rivett[1]

Dr. L.P. Chevallier[2]

Dr. A. Potgieter[3]

Departments: [1]Civil Engineering

[3]Computer Science

Organisation: [2]Council for Geoscience

Date: September 2010

In this thesis, a methodology for integrated catchment water resources assessment using Bayesian Networks was developed. A custom made software application that combines Bayesian Networks with GIS was used to facilitate data pre-processing and spatial modelling. Dynamic Bayesian Networks were implemented in the software for time-series modelling.

The structures of three Bayesian Network models were created automatically using a Hybrid Genetic Algorithm (HGA) which was implemented in a custom developed software product. The creation of the networks was done in a one step process with the discretisation of the continuous datasets. The discretisation was done using an equal binning method and the three networks resulted from variations in the number of intervals defined for the bins. The three networks were scored using the error rate, the logarithmic metric, the Brier score and the spherical score. From this evaluation, the states of the continuous variables were finalised and the optimum Bayesian Network model (the one with the most favourable scores) emerged. The model was then populated with the data collated for the Great Kei catchment in the Eastern Cape Province in South Africa.

The results were used to explore scenarios on the likely impacts of variations of some query variables over other variables in the network. This was performed through sensitivity analyses, scenario analyses and "what if" analyses. The findings from the model conform to existing knowledge on the study area which illustrated that Bayesian Networks can be successfully applied in integrated catchment assessment. The use of Bayesian Networks for spatial prediction was successfully proven with an example on the effects of surface water EC in one catchment on other neighbouring catchments. This information can be used in assessing the likely impacts of changes in surface water quality on connected catchments.

Lastly, the capability of Dynamic Bayesian Networks for temporal prediction was demonstrated. Dynamic Bayesian Networks were tested for predicting monthly rainfall and temperature and the results compared to that obtained from the static Bayesian Network. The results showed that Dynamic Bayesian Networks provided better predictions mainly because of the ability to incorporate evidence from the preceding months.

The major finding is that there is need for adequate data at the required scale. This was evident from the fact that some well-known relationships from theory could not be established using the automatic structure mining method used. The importance of selecting the appropriate discretisation technique was also highlighted by the different patterns obtained with variations in the discretisation levels. In the absence of the required data, expert knowledge should be collected and used to inform the modification of the relationships obtained using automatic methods and for the infilling of gaps in data.

DECLARATION

I, the undersigned hereby declare that this thesis is my own unaided and original work, apart from the normal guidance from my supervisors and where acknowledged. This work has not been submitted for a degree at any other tertiary institution.

SIGNED: _____

DATE: _____

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

"Change will not come if we wait for some other person or some other time. We are the ones we've been waiting for. We are the change that we seek."

US President, Barack Obama, 2008

"If you're walking down the right path and you're willing to keep walking, eventually you'll make progress."

US President, Barack Obama, 2008

CHAPTER ONE: BACKGROUND AND INTRODUCTION

1. Introduction

Historically, water resources were viewed as limitless and therefore planning for services was focused on the economics of demand and supply. As the demand increased, water resources got depleted and the cost and challenges associated with providing services grew, as did the environmental problems (Hermanowicz, 2005; van Wyk *et al.,* 2001). The paradigm has since shifted and the focus is now on conservation and the application of sustainability[1] principles in water resource assessment, planning and management. Sustainability focuses on maintaining ecosystems and all their components and processes in a condition such that they continue to provide the goods and ecological services (Andreasen *et al.,* 2001).

For sustainability, the widely accepted aspects of it, that is, the social, physical/ecological and economic, have to be integrated in an assessment and monitoring framework (Aspinall and Pearson, 2000; Macleod *et al.*, 2007; Quinlan and Scogings, 2004; Said *et al.*, 2006; Everard, 2004). The most common and suitable framework through which water resource assessment has been applied is Integrated Water Resources Management (IWRM).

IWRM can be defined as:

> *"the process which promotes the co-ordinated development and management of water, land and related resources, in order to maximise the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems" (GWP,* 2001)

[1] In 1987 the Brundtland Report, also known as "Our Common Future", defined sustainability as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs."

IWRM and can occur at two levels, at a broad level, for example at national level, or at a detailed community level (Ness *et al.*, 2007). At detailed level, the catchment scale, as specified at the 1992 Dublin Conference on Water and the Environment, is the appropriate unit for analyses (Everard, 2004; Walker *et al.*, 2006).

A catchment can be defined as:

> *"a basin shaped area of land, bounded by natural features such as hills or mountains from which all runoff water flows to low points which include any body of water such as a creek, river, lake, estuary, wetland, sea or ocean (Pine Rivers Catchment Association, 2010).".*

The selection of the catchment as a suitable unit for analyses is based on the understanding that land, water and ecosystem matters are associated with the hydrologic cycle and ecological processes that occur within the catchment (Argent *et al.*, 1999; Aspinall and Pearson, 2000; Amakali and Shixwameni, 2003). The use of the catchment also allows decision-makers to create a holistic view of interrelated components within an area (Pollard, 2002).

Water resources assessment is a key element of IWRM and involves the study of the current and future status of water resources and supply services in a catchment. It focuses on the availability, accessibility and demand. It involves a holistic evaluation of the relevant parameters on the quality and quantity of the surface and groundwater resources in a region or catchment. The analyses of uses, needs, demands and social or economic impacts under different scenarios or alternatives can also be evaluated (GWP, 2001; Batchelor, *et al*., 2005).

1.1 Problem statement

The majority of water resources assessment models used in South African rural catchments are generally conceptual, empirical or physical-based. Physical-based models are quantitative and founded on mathematical equations and are the most commonly used. These are either in the form of partial differential equations, based on mass-balance concepts or empirical regression equations. These equations are then resolved using numerical methods to output the values for the variables of interest. Jewitt *et al.,* 2000 provides an in-depth discussion of these models.

Physical models are complex, but provide a more detailed representation of processes than empirical or conceptual models and can evaluate numerous parameters. They require good quality input and output data at the relevant spatial and temporal scales, mostly for calibration, which are often not available for most catchments in South Africa (Silberstein, 2006; Hughes, 2004). These models are good at simulating daily river flow and transport dynamics (Hansen *et al.,* 2007). According to Parkin *et al.,* 1996, one of the major challenges arises when they are used in predicting hydrological response in ungauged catchments. Examples of problems include the inability of equations to represent actual field processes and over-parameterisation of the models.

Analysing parameters in great detail can suggest that these models may be more "accurate", but this might also not be true (Ochieng, 2007). Various studies highlighted in Ochieng (2007) illustrate how similar results can also be obtained from simple models which are highly useful when there is inadequate data or when the quality of the data is not acceptable.

In physical-based models, the uncertainty associated with model data inputs, parameter estimation and outputs are generally not presented. This is vital in IWRM where processes and data are often complex and highly uncertain in time and space. Uncertainties are inherent and cannot be eliminated although in some cases they can be reduced. It is vital in modelling that the uncertainty be quantified or at the very least be documented, as decision-makers need to know the uncertainty around an outcome. This enables them to judge whether or not they have enough information to make a particular decision.

The quantification or documentation of uncertainty requires a basic understanding of what uncertainty is and the basic sources of uncertainty (Krzysztofowicz, 2001). Analyses that involve the evaluation of uncertainty are especially relevant in the South African context where there have been minimal or no noticeable contributions to uncertainty analysis in water resources assessment models (Hughes, 2004; Institute of Water Research, 2008a).

Another challenge is that in most catchments, there is lack of clarity as to the sustainability of their water resources in future if people continue using resources the way they are doing now, if certain resources run out or if different management strategies are set up (Chakrabarti *et al.*, 2001; Loucks, 2000). Models that support scenario analyses and prediction under uncertainty and that can assist in assessing the effects of future drivers like climate change, population growth and land use change on resources management and planning in the catchment are required.

Catchment processes occur over varying spatial and temporal scales. The assessment of water resources needs to encompass the analyses of spatial and temporal changes. Spatial changes are differences across space or throughout the catchment at a given time and temporal variations are changes occurring at different periods in time. The next section outline the purpose of this research, which is addressing the issues discussed above.

1.2 Research purpose

This research proposes a Bayesian Network based spatial and temporal modelling methodology for catchment water resources assessment. This methodology can use uncertain quantitative and qualitative data of varying spatial and temporal scales and levels of uncertainty. It provides capabilities for the analyses and documentation of model uncertainties and their implication at detailed catchment scale. It also allows for the prediction and simulation of scenarios and the monitoring of the change in catchment sustainability over specified periods of time.

Bayesian networks are graphical models that allow for the representation and reasoning of any uncertain domain (Pearl, 1988; Korb and Nicholson, 2004). They show the relationships between datasets in the specific domain, for example water resources assessment and represent the strength of these relationships as probabilities. They enable the documentation of uncertainty by representing beliefs about values as probability distributions; the higher the uncertainty the wider the probability distribution.

Bayesian Networks have the ability to incorporate multiple qualitative and quantitative data available at different spatial and temporal scales (Ticehurst *et al.*, 2007). There is no need to convert the data to common units. Expert knowledge from various sources and with different uncertainties can also be incorporated (Uusitalo, 2007). This is important in catchment modelling, where they are incomplete records of some datasets and social variables (which are difficult to quantify) need to be evaluated (Sadoddin *et al.*, 2005; Jakeman *et al.*, 2003).

Bayesian Networks can reason with missing data. They are suited to incrementally mine new probabilistic patterns from data. When new data are input into the Bayesian Network model, new probabilities are calculated. Because they are solved analytically, they offer rapid response during query analyses when the model is updated. Bayesian Networks techniques have also been proven in some studies to provide predictive properties (van Wyk *et al.*, 2001).

Standard Bayesian Networks model static situations with a fixed set of variables, that is, they reflect the condition at one point in time. Most catchment processes evolve over time so time-series modelling is essential. Dynamic Bayesian Networks (DBNs) were developed to model processes that evolve over time and are suitable for time-series modelling (Russell *et al.*, 2004 cited by Amir, 2004; Pearl, 1988). Bayesian Networks and Dynamic Networks are discussed in more detail in Chapter 3.

Specific software is required for Bayesian Network modelling. Although it is acknowledged that there are software products available commercially or as open-source packages, this research required a custom developed software application which was created because of the reasons listed in the following paragraph.

The major motivation arose from the fact that this research intends to explore the use automatic methods for Bayesian Network structure learning. The aim was to use a novel structure mining technique, the Hybrid Genetic Algorithm (HGA) and this necessitated the development of custom software. This is discussed in more detail in Section 5.1. There was also a need to implement Dynamic Bayesian Networks in a simple and user-friendly framework. The other requirement was the need to integrate GIS display capabilities for the easy presentation of result when performing scenario analyses.

1.3 Main objective

The main objective of this thesis is to develop a Bayesian Network-based methodology for spatial and temporal catchment water resources assessment in South Africa.

1.4 Specific objectives

The specific objectives of this thesis are to:

i) develop a Bayesian Network model for water resources assessment using automatic techniques;
ii) collate data for the selected study area, the Great Kei catchment in the Eastern Cape Province in South Africa;
iii) use the collected data in a custom developed software, based on a Hydrid Genetic Algorithm, to automatically create the network;
iv) apply the developed Bayesian Network model to evaluate spatial variations in water resources parameters in the study area;
v) evaluate and validate the Bayesian Network model mainly using sensitivity and scenario analyses; and
vi) illustrate the use Dynamic Bayesian Networks for temporal prediction of some aspects of the catchment over specified times.

1.5 Research contribution

The contribution of this research effort is threefold, namely:

A new methodology for integrated catchment assessment

A novel approach in South Africa for assessing and monitoring catchment water resources by considering relationships between multiple variables is proposed. Approaches that provide answers on the extent to which different factors affect water resources sustainability, through scenario analyses, are not fully developed. The methodology developed in this research highlights the assessment process, from data collection, data processing, the modelling required and the presentation of the results to the stakeholders.

Integration of data and knowledge under uncertainty

The methodology contributes towards a new way of integrating data and knowledge about any uncertain domain in modelling. Subjective estimates from experts can be used where measured data are not available and this allows incomplete datasets to be successfully used in modelling. The effects of the uncertainties in input parameters are presented and assessed through sensitivity and scenario analyses. The developed model is easily updateable when accurate data becomes available. This makes it useful for adaptive management and research activities.

Time-series monitoring and prediction

A new approach in South Africa for time-series monitoring and prediction using Dynamic Bayesian Networks is developed. Traditional approaches to time-series analyses are based on regression techniques performed using two variables, and suffer from the assumptions of stationarity and linearity (Jain *et al*., 2007). These predictions are based on a finite period, and it is difficult to incorporate prior knowledge and deal with multi-dimensional inputs. The approach developed in this research is based on the analysis of the temporal relationships between multiple variables.

1.6 Thesis outline

Chapter 2 discusses IWRM and water resources assessment, the major factors that need to be integrated in modelling and some of the major issues that makes the modelling process complex. The methods and tools available for water resources assessment in South Africa are presented. Following a discussion of these methods, a case is made for the selection of Bayesian Networks as a suitable technique to facilitate and handle the complex issues required in integrated assessment.

A brief theory of Bayesian Networks is provided in Chapter 3. The various steps involved in creating Bayesian Networks and the theory on the development and use of Dynamic Bayesian Networks for time-series modelling are presented. The integration of Bayesian Networks and GIS is discussed with some examples from literature.

Chapter 4 presents some examples from literature on the application of the different types of Bayesian Networks in water resources assessment. A discussion on the advantages and some of the limitations of Bayesian Networks is presented at the end of the chapter.

The research methodology followed in this research is discussed in Chapter 5. In this chapter, the software development process and the functionalities of the custom developed application are examined. There is also discussion on the data available for the study area and how it was used in the modelling process. The process of designing the Bayesian Network model using the custom developed application and the results obtained are presented.

Chapter 6 presents the results of the modelling process and the different sensitivity, scenario and "what if" analysis performed. Results on the use of Bayesian Network for spatial prediction and Dynamic Bayesian Networks for time-series modelling are discussed. The implications of these results on the management of the catchment water resources are presented.

Chapter 7 presents conclusions drawn from this research and provides recommendations for future research work.

CHAPTER TWO: CATCHMENT WATER RESOURCES ASSESSMENT

2 Introduction

Integrated water resource management (IWRM) was identified in Chapter 1 as currently the best approach to ensure sustainable water resources use and management for catchments globally and in South Africa. Some of the issues that make integrated assessment complex are being discussed in this chapter. This is followed by a presentation of examples of water resources assessment methodologies/tools commonly used in South Africa. This leads to a discussion on the justification for the use of Bayesian Networks in assessment.

2.1 Integrated water resources management assessment and modelling

IWRM is the systematic use of technical and non-technical measures and activities to ensure the effective and efficient management of water resources. The primary goal of IWRM is to optimize the relationship between the capacity of the available resources to provide sustainable services, such as water of a given quantity and quality, and utilization of the resource (Ashton, 1996).

The key concepts of IWRM are equity, efficiency and sustainability. IWRM aims:

i) to promote more equitable access to water resources and the benefits that are derived from water to tackle poverty;
ii) to ensure the efficient use of scarce water for the benefit of the greatest number of people; and
iii) to achieve more sustainable utilisation of water for a better environment.

Water resources issues are scale dependent and the spatial scales range from local to international. International scale results in a broad analysis whereas the catchment scale is the most suitable for a detailed analysis. The suitability of the catchment scale as the smallest and most detailed unit for analysis was emphasised at the 1992 Dublin Conference on Water and the Environment (Everard, 2004; Walker *et al.*, 2006).

At a catchment scale, IWRM can feed into a broader Integrated Catchment Management (ICM) Framework. Ashton, 1996 presents a diagram that illustrates the relationship between IWRM and ICM (Figure 2- 1). Traditionally, water resources managers assumed that for sustainable development, it was sufficient to control water use and protect the integrity of the water resources hence the focus on water quantity and quality management. Increasing, there was the realisation that the complex issues of land use patterns and stakeholder involvement could no longer be ignored and had to be considered in the management of catchments.

Water quantity and quality management requires the least level of integration and involves the lowest management of complexity. In water quality and quantity management, there are fewer parameters to evaluate and the minimum number of stakeholders to consider. IWRM offers more integration and management of the complex catchment issues. Other environment and social factors besides just water quality and quantity have to be evaluated. This also increases the number of stakeholders who participate in the process. ICM is a broader context as it recognises all environmental resources in a catchment (Pollard, 2002). The most "ideal" concept is Integrated Development Management, which encompasses the development of all resources and covers a larger geographical area and has the highest value to society.

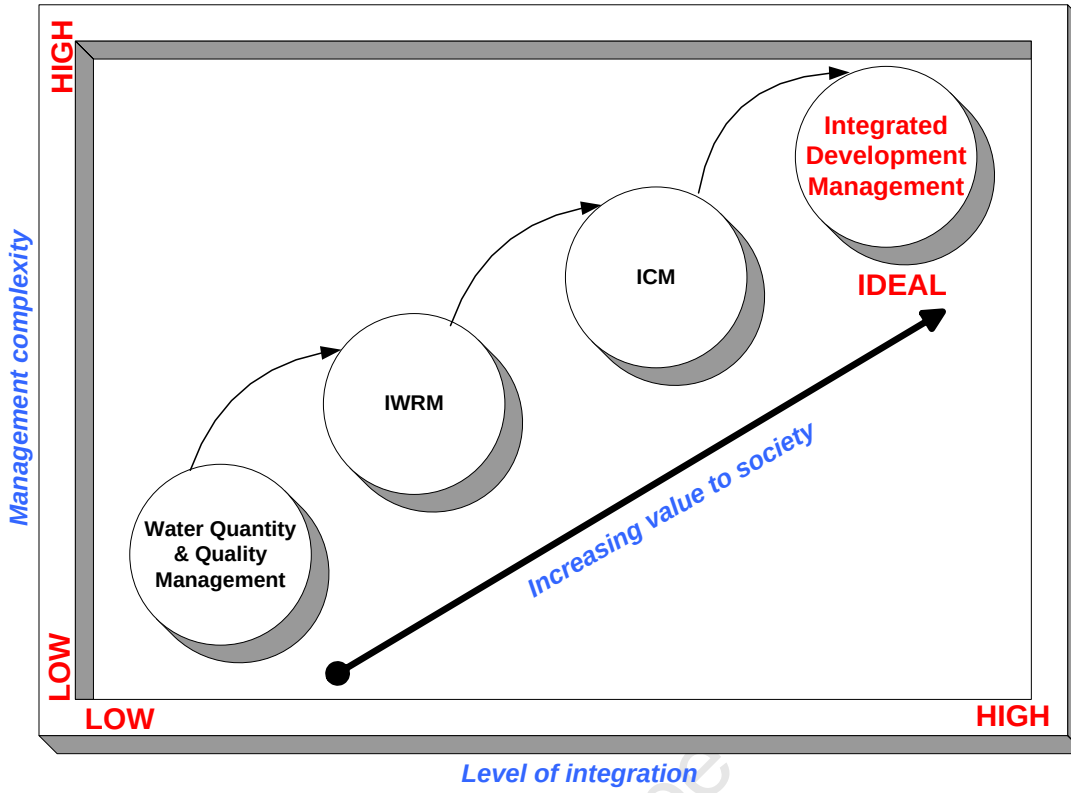


Figure 2- 1: Relationship between IWRM and ICM in terms of the level of integration required and the complexity of the management processes (adapted from Ashton, 1996).

A key component of IWRM is water resources assessment, which involves the analysis of the following components (Batchelor *et al.*, 2005):

i) demand assessment- which provides an overview of the current and future demands for water and the hindrances to providing for these demands;

ii) environmental assessment- which considers current and potential societal impacts on the environment and the functioning and protection of the ecosystems;

iii) social assessment- assessing how social and institutional structures affect water availability, demand, access and social structures; and

iv) risk or vulnerability assessment- investigating the likelihoods of extreme events like flood, droughts and the impacts on society and also the impacts of other factors like climate change, political change or epidemics.

Batchelor *et al.*, 2005 also stated that according to their experience, water resources assessment needs to be carried out in several steps of increasing complexity ranging from light assessment, problem-focused assessment and to a comprehensive assessment (Table 2- 1).

Table 2- 1: The three levels of water resources assessment presented (after Batchelor *et al.*, 2005)

Light (or rapid)	**Problem-focused**	**Comprehensive**
-Initial identification of priority problems.	-Follows a rapid assessment and focuses on an individual problem/group of problems.	-Developing a comprehensive information base of water related issues in the area of interest.
-Assessment of easily accessible quality controlled secondary data on physical aspects of resource availability and service provision. -Primary data collection done to fill gaps.	-Detailed assessment of quality controlled secondary data with additional primary data collected if necessary.	-Consolidation, quality control and assessment of secondary data on physical aspects of resource availability and service provision. -Primary data to fill gaps, sometimes as part of a long-term management programme.
-Secondary data, and rapid techniques for collecting societal information.	-Rapid and participatory appraisal techniques for collecting information specific to problems.	-Participatory evaluation techniques for collecting societal information and detailed measurement of physical information.
-Initial assessment of the causes of problems.	-Detailed assessment of causes of individual problems.	Detailed assessment of root causes of problems, linkages between problems and externalities that influence water availability and use.

Water resources assessment requires more integrated approaches to modelling (Jakeman *et al.*, 2005). As highlighted in Table 2- 1, it must incorporate various primary and secondary data on water resources and other related environment and social issues. It must integrate societal and stakeholder issues in evaluation and must examine the linkages between the different data and issues.

Integrated assessment modelling is complicated by the fact that catchments are complex systems characterised by self-organization, adaptation, and heterogeneity across time and space scales (Poch *et al.*, 2004; Pahl-Wostl, 2007). Due to that inherent complexity it is difficult to distinguish cause from effect (Costanza *et al.*, 1993; Jewitt *et al.*, 2000; Blöschl, 2006; Argent *et al.*, 2000).

Paul-Wostl, 2007; van der Sluijs, 1996; Costanza *et al.*, 1993 describe the complexity of catchments as follows:

i) they exhibit discontinuous and chaotic behaviour (when very small changes in the system parameters can have disproportionally large impacts on the system behaviour);
ii) the state of a system and the effect of interventions at a certain moment in time depends on history and context;
iii) systems are hierarchical and not all system properties can be observed and due to non-linear, evolutionary processes and new system states may be observed in the future without any similarity to historical occurrence;
iv) they escape attempts at external control by adaptation and human beings; they may behave differently than anticipated by evolving and learning; and
v) for some extreme states it may be impossible to quantify probabilistic judgements due to non-linear developments.

Section 2.3 provides a discussion of the major issues to be integrated in water resources assessment that arise from the complexities of catchment systems highlighted above.

2.2 Aspects in integrated water resources assessment

Figure 2- 2 illustrates the main aspects to be integrated in the assessment of water resources.

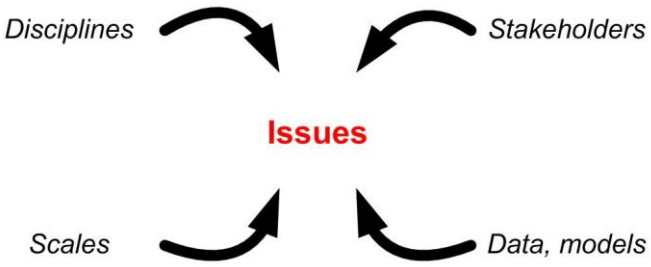


Figure 2- 2: The issues commonly "integrated" in integrated modelling and assessment (adapted from Parker *et al.*, 2002).

The subjects to be integrated in assessment include different disciplines and stakeholders involved in catchment management, different data and models and the varying scales of catchment system behaviour (Letcher *et al.*, 2005 cited in Ekasingh and Letcher, 2005). These aspects are discussed in more detail in the following sections.

2.2.1 Stakeholder involvement in assessment

A stakeholder in IWRM is defined as someone who uses water and is concerned about the protection of the resources (DWAF, 2004b). This ranges from officials from national and regional levels of governments to the community and local water users (Quinlan and Scogings, 2004). Their interests and responsibilities may be corresponding, overlapping or conflicting. These stakeholders need to be engaged and involved from the start of the assessment process so that they get a common understanding of the issues, the processes and models applicable to the catchment. They can learn from each other's experiences and this assists in facilitating the implementation of any management initiatives (Greiner, 2004). Cain *et al.*, 1999 state that the implementation of many integrated resource management plans fail due to the lack of participation of local institutions during the development process.

The stakeholders can be included at the following stages (Jessel and Jacobs, 2005; Letcher and Jakeman, 2002):

i) in the initial stages of the project, when the problems are identified and the different options are discussed;
ii) during model development to gather input on the data, methodology and assumptions to be made during processing; and
iii) lastly for presentation of the results and model testing and validation.

This stakeholder involvement is often iterative throughout the assessment process.

In the initial stages of the project, the first step is the identification of the stakeholders and the creation of an awareness of the participatory process. It is more important to include the community and all water users and environment groups because they have a better understanding of their needs. They can articulate any problems they might be experiencing with the supply, use and management of the resources.

In the initial stage, the modelling objectives can be specified and priorities set on the problems to solve and some potential actions required in solving these problems can be listed. There are still challenges for community involvement; especially in developing countries like South Africa. Because most people in these catchments are poor and uneducated, it is difficult for them to communicate their needs and challenges and also for them to understand the options and scenarios provided to them by the scientists, service providers and managers. In most cases, larger water users often dominate the needs of the poor communities (DWAF, 2004b).

During model development, input on the data to use, the methodology and assumptions to be made can also be gathered from stakeholders and experts in water-related disciplines. The model and its components must initially be presented in a conceptual approach that is simple and easy to interpret especially for the non-expert stakeholders. Using these conceptual diagrams, the stakeholders can provide information on the perceived interactions between the different aspects being modelled (Cain *et al.*, 1999). Group surveys can also be used to gather more qualitative data on the catchment systems and the results can be included in modelling.

The information gathered through stakeholder participation can then be used in assessment to created the model and produce results. The resulting model and the outputs from the modelling exercise must also be presented in a graphical that is simply enough for the stakeholders to understand. The stakeholders can then review the results and revise the model until a consensus is reached and the model and results are acceptable (Burgman, 2005)

Stakeholder participation is a more complex process than outlined in this section. The aim was to provide a brief discussion on its importance and some of the complications. More information on stakeholder participation can be obtained from Lupini, 2004; Janssen *et al.*, 1996; Burns *et al.*, 2006; Quinlan and Scogings, 2004; Pollard *et al.*, 2001; Jessel and Jacobs, 2005; Letcher and Jakeman, 2002.

2.2.2 Multiple data, models and databases

Researchers worldwide have acknowledged that catchment assessment requires the integration of data, models and databases. Data characterising catchments often originate from heterogeneous sources. Some datasets are qualitative; others are quantitative and most are complex. The data are either collected over long periods of time, but with gaps, or over short field measurements. The periods of time when the measurements are done do not often coincide between the different databases. Other problems include measurement errors and the difficulty of getting a dataset that covers the entire geographical area (Hughes, 2001). Sometimes the data are collected for other purposes outside modelling and this leads to the unavailability of the relevant data at the required scale (Cherkassy *et al.*, 2006).

If many data values are missing, then the quality of models and information decreases and the uncertainty associated with the data increases. Data are needed in models to constrain the discussion and improve the understanding and limitations of the output results. The demand for data to calibrate and validate models rises with increase in model complexities (Silberstein, 2006). This is a challenge for water resources assessment especially in South Africa, where a decline in the infrastructure available for hydrological data gathering and monitoring has been experienced over the years (Hughes, 2004).

Different approaches are available for dealing with missing data. The general question is whether one must interpolate, extrapolate or produce other estimate values. The traditional, simplistic approach is to delete cases with missing data. The disadvantage of this is that it introduces bias in the results if the sample is not representative of the whole population. The bias might be improved by reweighing the remaining cases using response probabilities which are estimated from the data (Schafer *et al.,* 2002). Other commonly used methods available for estimating missing values are mostly based on statistical analysis, Feelders *et al.,* 2000 provide some examples. The choice of the suitable method is usually informed by the type of model used. The method used in this research is discussed in more detail in Section 3.1.4.

2.2.3 Multiple scales of systems behaviour

The issue of scale is important in integrated assessment modelling (Jewitt, 1998; Lovell *et al.,* 2002). Scale refers to:

> "*the spatial, temporal, quantitative or analytical dimensions used by scientists to measure and study objects and processes" (Gibson et al., 2000).*

The three different types of scale for catchment assessment, in both the spatial and temporal dimensions, identified by Jewitt, 1998; Loucks *et al.,* 2005 are:

i) process scale: the temporal and spatial scales that natural phenomena exhibit and are outside human control;
ii) sampling scale: the scale at which humans elect to collect samples of observations or study phenomena; the density of the measuring network and the sampling frequency determine the sampling spatial and temporal scales; and
iii) information scale: the spatial and temporal scale of the information required in decision-making; information at scales smaller than needed is considered noise and information at larger scales than required is considered irrelevant.

When modelling, the purpose is to select a scale that provides information at the required level of detail, taking into consideration the available processes and their spatial and temporal scales. Changes in scale affect the importance or relevance of variables (Meentemeyer, 1989; Parker *et al.,* 2002; Costanza *et al.,* 1993) and subsequently the output. "Real world" processes operate at different scales and these need to be considered (Allen *et al.,* 1992; Ehleringer *et al.,* 1993 cited in Agarwal *et al.,* 2000).

The question of choosing the appropriate modelling and simulating scale is challenging and the complexities vary from one study area to another. Wiens, 1989 states that the choice should be informed by the question that is being asked about the system. According to Jakeman *et al.,* 2003, the scale should be fine enough to capture the needed level of detail variability but not finer than that allowed by data availability and quality. As a result, it might be more appropriate to select the final scale after an initial analysis of the data.

When spatial and temporal data are available at a different scale to that required in modelling, they have to be aggregated or disaggregated. Aggregation is taking data from site specific observation to a coarse[2] scale of study. Disaggregation is taking the output of large scale observations and deducing changes that occur at finer resolution. Disaggregation affects the predictability of the variables. Predictability measures the reduction in uncertainty of one variable given the knowledge of others based on categorical data. Predictability is high when the spatial and temporal scales are similar (see Figure 2-3) (Costanza *et al.*, 1993; Meentemeyer, 1989; Wiens 1989).

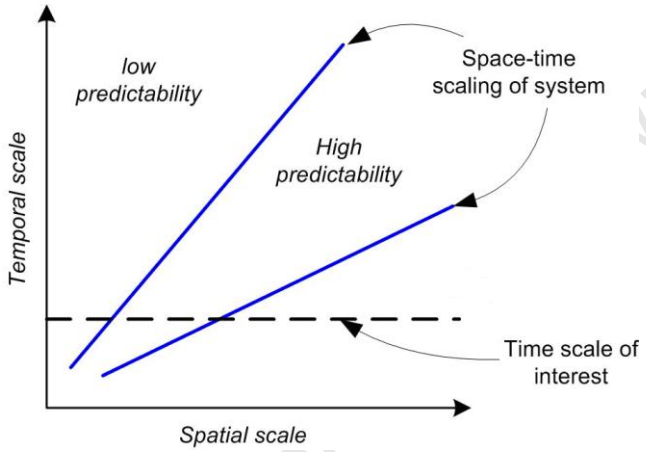


Figure 2- 3: Relationship between spatial and temporal scales and predictability (adapted from Wiens, 1989).

At fine or detailed spatial scales there is low predictive capacity. This is because of the reduction in spatial autocorrelation as there is more variability in the attributes (Wiens, 1989). This variability at detailed scale leads to an increase in uncertainty. Uncertainty is reduced at broad spatial and temporal scales as the local heterogeneity of features or processes is averaged out at broad scales. At broader scales, there is an increase in autocorrelation and patterns are more predictable (Wiens, 1989).

[2] Scale can be characterised by the resolution. Resolution is the level of disaggregation of the study and is measured by the space between units forming the grain of the study. Resolution may be characterized as fine or broad scale. Fine-scale models depict geographically small units of analysis, while broad-scale models usually have larger spatial units of analysis (Agarwal *et al.*, 2000).

These are some of the issues considered in choosing the relevant spatial and temporal scales for modelling for this research. The scales chosen for this research are discussed in Chapter 5.

2.2.4 Uncertainty

Uncertainties[3] are inherent and cannot be eliminated although in some cases they can be assessed. It is vital in modelling that the uncertainty be quantified or at the very least be documented and this requires a basic understanding of what uncertainty is and its sources (Krzysztofowicz, 2001). Burgman, 2005 divides uncertainty into two major groups, linguistic uncertainty and epistemic uncertainty. Linguistic uncertainty arises due to the fact that language is not exact (Burgman, 2005). It is most prevalent when stakeholders' are involved in the modelling process and as such is not really relevant to this research.

Epistemic uncertainty reflects incomplete knowledge and includes (Burgman, 2005):

i) measurement error;
ii) systematic error;
iii) natural variation;
iv) subjective judgement; and
v) model uncertainty.

Measurement errors are apparently random and arise from the fact that the equipment and people used in measuring data are imperfect. This type of uncertainty can be dealt with by applying statistical techniques to numerous measurements and reporting aspects such as confidence intervals when data and results are presented. These confidence intervals are however not sufficient to capture all the uncertainty so measured data often reflect changes in measuring methods, changes in the instruments used and the people gathering the data.

Systematic errors occur when measurements are biased. It is the difference between the true value of a parameter and the value to which the mean of the measurements converges to with increases in sample size.

[3] a person is uncertain if s/he lacks confidence about the specific outcomes of an event. Reasons for this lack of confidence might include a judgement of the information as incomplete, blurred, inaccurate, unreliable, inconclusive, or potentially false.

Systematic errors can result from a deliberate judgement of a scientist to incorrectly include/exclude some data from assessment. They can also result from a consistent and unintended error in calibrating equipment or recording measurements. These errors once recognised can be removed by applying a correction when the size and direction of the bias is known. In order to avoid these errors, independent studies can be carried out comparing estimates with scientific theory and careful attention to detail can be employed.

Natural variation is environmental change in time and space and is difficult to predict. This results from lack of adequate knowledge about the dynamic processes and initial conditions of a system. The true values of the aspect under consideration changes with changes in the driving variables such that the values are difficult to measure across a full range of temporal and spatial values. Probability distributions and intervals can be used to address this uncertainty.

Subjective judgement uncertainty results where they are inadequate measurements. Experts' judgements can instead be used and these are usually based on observations and experiences and contain uncertainty. Subjective judgement can be handled by assigning a degree of belief (subjective probability) which should coincide with the results of the data, if they were available.

Model uncertainties can be subdivided into different types. Firstly, there are those due to the model structure, which are attributable to the incomplete understanding and the simplified description of modelled processes as compared to the real situation. Uncertainties can be caused by the variability of observed input and output values over regions smaller than the spatial and /or temporal scale of the model. Errors can also result from linking models of different spatial and temporal scales. Model technical uncertainty can arise from the computer implementation of the model; this can be due to approximations, resolutions and software bugs.

Silberstein, 2006 discusses the relationship between parameter and model structure uncertainty (Figure 2- 1). Figure 2- 1A shows a structurally sound model that fits reality. Cumulative errors grow as the number of parameters grows. But as the model gets complex, the structure errors become less (Loucks *et al.*, 2005).

If the model is not correct as in Figure 2- 1B, increasing parameters does not reduce the errors. The model only reaches structural and not parameter accuracy. In B, there is opportunity to improve the model and reduce the errors using real data.

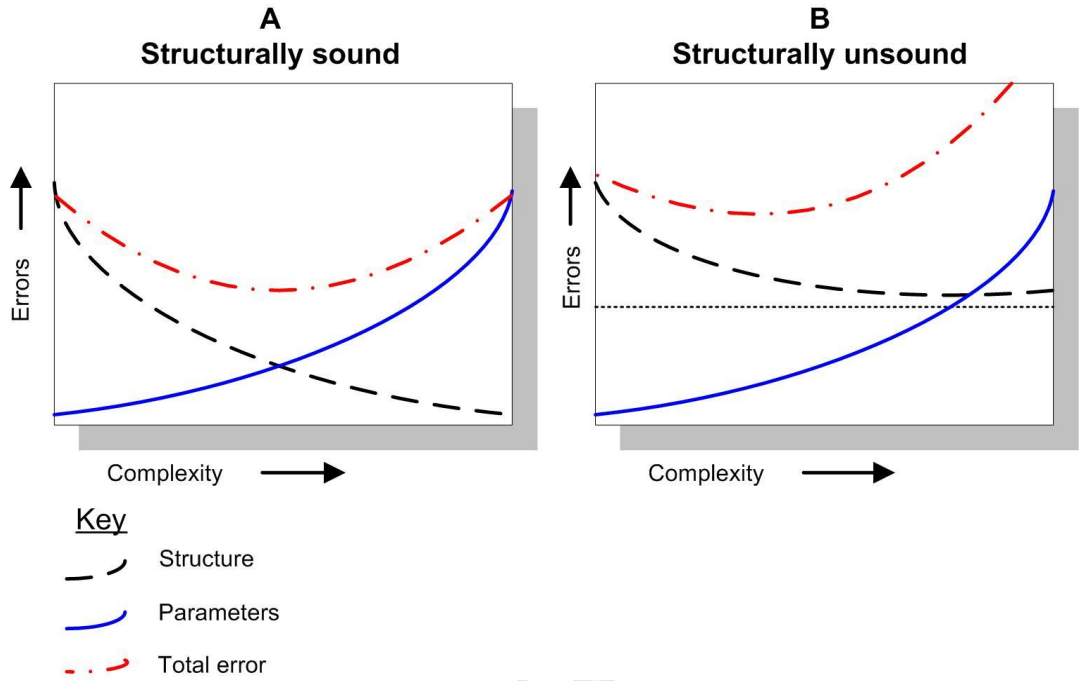


Figure 2- 4: The relationship between parameter and model structural error for a correct model (A) and a flawed model (B) (adapted from Silberstein, 2006).

It is important to communicate the uncertainty in data and models to the users of the modelling results and decision-makers. If the value of, for example, rainfall varies and this variation cannot be predicted in time with certainty, it is called a random variable. It cannot be determined with certainty what the value of a random variable is, instead, only the probability or likelihood that it will be within some specific range of values can be provided. This implies that the probabilities of observing particular ranges of such a variable are defined by a probability distribution (Loucks *et al.*, 2005).

An uncertainty analysis needs to be performed and this attempts to describe the entire set of possible outcomes and their associated probabilities of occurrence (Loucks *et al.*, 2005; Umakhanthan, 2002). Although an extensive sensitivity assessment is recommended, often the results can be difficult to interpret due to the large number and complexity of relationships tested.

In most cases, priorities need to set on the variables to assess and this can be done on a trial and error basis or using prior or expert knowledge (Jakeman *et al.,* 2006). The modelling of uncertainty due to model structure is complex and is often omitted from most studies.

2.3 Modelling for water resources assessment

Models are instrumental in water resources assessment for decision support. Models are simplifications of real-world systems and are designed to enable process understanding and exploration of system behaviours. They provide a mechanism for data testing; provide responses to question on the future states of systems and enable the investigation of scenario options (Silberstein, 2006; GWP, 2001; Batchelor, *et al*., 2005; Argent, 2004). Models are used as they are cheaper and faster than carrying out real experiments and enable the prediction of aspects that people are unable to do in reality (Silberstein, 2006). Barnes, 1995 states that models should be generalisable, adaptable and have a wide scope.

There are various types of models designed for different purposes and at the bare minimum; they must fulfil the three criteria provided below. Although in most cases all three will not be fulfilled to the same full extent, tradeoffs have to be made depending on the modelling objectives (Costanza *et al.*, 1993):

a) realism; the ability to simulate system behaviour in a qualitative, realistic way;
b) precision; simulating system behaviour in a quantitatively precise way; and
c) generality; which is the ability of the model to represent a wide range of systems' behaviour.

When models are developed, only certain aspects of the catchment system are abstracted in order for their behaviour to be understood and predicted. The decision on which aspect to extract is usually based on the understanding the modeller has of the processes in the catchment (Wagener, 2003; Loucks *et al.,* 2005). The abstraction can only be done with limited accuracy and therefore some caution has to be taken before applying models since their lack of knowledge of the system cannot be compensated. Ultimately, not every problem requires a model (Dent, 2001 cited in Dube, 2006).

Modelling guidelines are therefore required to safeguard against or assist in management of the common errors resulting from the incorrect application of models. An example of these guidelines is provided by Jakeman *et al.,* 2006 and illustrated in Figure 2- 5 and discussed in the following paragraphs.

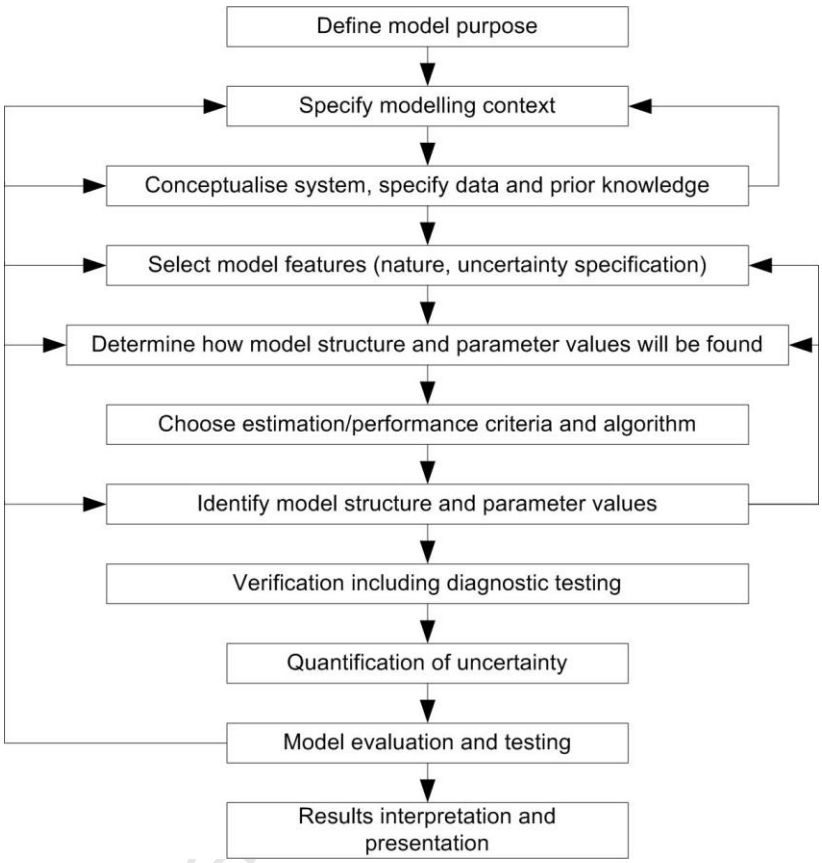


Figure 2- 5: Iterative steps in developing a catchment model (adapted from Jakeman *et al.*, 2006).

Defining the purpose of the modelling exercise

In the first instance, the purpose of the modelling exercise need to be clearly outlined as it affects the selection of the model to use (Barnes, 1995; Loucks *et al.*, 2005). Examples of some of the purposes which are pertinent to this research include (Jakeman *et al.,* 2006):

i) Acquisition of a better qualitative understanding of the system;
ii) Data assessment for discovering the limitations, inconsistencies and gaps in data;
iii) Interpolation; estimation of variables that cannot be measured or filling in missing data;
iv) Summarising of data; and
v) Prediction or short-term forecasting (Dawes *et al.*, 2001).

Specifying the modelling context

This second step involves defining the scope and objectives of modelling. If this is not done adequately when the modelling process commences, the following might occur (Jakeman *et al.*, 2006):

a) the scope might be extended beyond what is required to respond to the pertinent question;
b) the underestimation or disregard of the difficulties and limitations of the data and techniques;
c) oversimplification or over-elaboration of the system components;
d) dependence on existing familiar models that are not ideal for the problem under study; and
e) overlooking existing knowledge and previous experiences.

In this step, the issues to be modelled and the interest groups are identified. The time required in completing the model, the resources available for modelling and the outputs are specified. It is also vital to specify the temporal and spatial scope, scale and resolution of the modelling context. The scope is the boundary of the system being modelled and defines what should be included in the model and what should be left out (Loucks *et al.*, 2005). The choice of the boundary is related to the selection of the scale of study. The delineation of the boundary and the scale are iterative processes that are discovered through trial and error (Jakeman *et al.*, 2006).

Selection of model features/types

The features of the model must be selected to conform to the data specifications and the system conceptualised in the previous step. Examples of features include the type of variables covered, and how they are processed (that is whether or not the model is lumped or distributed or if the system is a black box and so on). These features can assist in improving the conceptualisation and determining the calibration techniques available. The choice of model features depends on the modelling purpose, the objective, and the quality and quantity of prior knowledge (Jakeman *et al.*, 2006; Barnes, 1995).

Caminiti, 2004 discusses a procedural approach to selecting a model with the involvement of the relevant stakeholders. Ideally, comparisons between different model types must be done before the selection of a suitable one. But in practice this is seldom done due to limited resources and personal preferences.

At this stage, an initial assessment of the expected uncertainty is made and a plan for testing and evaluating this uncertainty has to be outlined (Loucks *et al.*, 2005). This will help in setting the predictive power and the extent to which observed past behaviour can be adopted into changed or future scenarios. Although it is best to revise/change the selected model type once an evaluation exercise has been carried out, this is often not possible due to financial and human resources constraints (Jakeman *et al.*, 2006).

Determine how model structure and parameter values will be obtained

In this step, the method to be used in finding the model structure is defined. When there is not enough data for modelling, science-based theoretical knowledge that informs of the relations between data can be used. Sometimes the structure may be discovered using trial and error but whatever the method; the rule is always "avoid more complication than is necessary to fulfil the objectives" (Jakeman *et al.*, 2006; Petersen *et al.*, 2003).

After deciding how the model's structure will be determined, the next step is to make a choice on how the parameters are to be estimated. The parameters can be estimated from observed data or they can be inferred from secondary data. Another option is to calibrate the parameters by comparing model outputs to observed data (Jakeman *et al.*, 2006). The next step is choosing the estimation criteria and algorithm.

Choice of estimation or performance criteria and algorithm

At this stage, an algorithm or technique for parameter estimation has to be selected. This should be computationally simple, must be able to quantify uncertainty in the results, should be as statistically efficient as feasible and must have the capability to test for over-parameterisation. Over-parameterisation is when a model has more parameters than necessary to describe the overall system behaviour (Dawes *et al.,* 2001). Over-parameterisation often leads to the model being fitted to inconsistent or irrelevant noise in data; it weakens the predicting power of models and leads to misinterpretation of results (Jakeman *et al.,* 2006). Barnes, 1995 states that the number of parameters in a model should be constrained by the data available at the appropriate scale.

Identify model structure and parameter values

The suitable model structure and parameter values are discovered iteratively. Generally, this is analysing whether or not some parameters can be dropped or if the addition of other parameters is necessary. Formal statistical techniques can also be applied to differentiate among different model structures. They provide a way of assessing the model structure based on performance on other various data sets, parameter estimates and prior knowledge (Jakeman *et al.,* 2006; Refsgaard *et al.,* 2005).

Verification

Donigian, 2002 states that since models are approximations of reality they cannot precisely represent natural systems. The implication is that model outputs need to be interpreted and should not always be perceived to be true (Snowling and Kramer, 2001). The consensus amongst the research community is that assessing the quality of a model is crucial and this can be done through verification (Refsgaard *et al.,* 2005; Wagener, 2003).

Although the terms ‘verification’ and ‘validation’ are often used interchangeably in practice; Oreskes *et al.,* 1994 cited in Snowling and Kramer, 2001 differentiates the two and provides the following definition for verification:

"*verification is testing to show that the model's numerical solution is close to the analytical solution*"

Model verification should test the robustness of the model to practical insignificant changes in data and the deviations of data and the system from the assumptions made during model development. A qualitative approach to model verification can be taken and this involves the use of expert knowledge. If the model does not perform as well as expected, the assumptions, the structure and the data included would have to be revised.

Quantitative verification can involve consideration of some of the following (Jakeman *et al.*, 2006; Wagener, 2003):

a) goodness of fit which is achieved by comparing the means and variances of observed data against modelled outputs;
b) consistency of the model in cross-validation against different sections of input-output records of data;
c) absence of correlation between models errors and observed inputs (this indicates unmodelled input-output behaviour);
d) speed of computation;
e) the certainty with which parameter values converge when more observations are processed using the model.

The methods available for verification differ with the type of model selected. In this research Bayesian Networks will be used and this implies that the methods for verification applicable to Bayesian Networks are the only relevant ones for presentation. These are discussed in detail in Chapter 4.

Quantification of uncertainty

The uncertainties in data and models have been discussed in Section 2.2.4. It is vital in modelling that the uncertainty be quantified and this requires a basic understanding of what uncertainty is and of the basic sources of uncertainty (Krzysztofowicz, 2001). While they are ways of assessing, quantifying and reducing uncertainties, they cannot be totally removed.

Some models are able to provide estimates of uncertainty from data, measurements or conditions. These estimates are provided in probabilistic terms. This requires the testing of the model over the range of every uncertain input and parameter. This task can become complex with integrated models and it is rarely performed. A common simpler type of approach is to carry out sensitive analysis to monitor the sensitivity of model outputs to changes in parameters. The results can be difficult to interpret with increases in the number of parameters and the relations being tested. Trial and error may be used to prioritise the aspects of the variables to be tested. Sensitive analysis can be performed with Bayesian Networks and this is explained in more detail in Section 3.1.6.

Model evaluation and testing

This is the evaluation of the model according to its objectives (Jakeman *et al.,* 2006). Some of the questions to ask include the following (Loucks *et al.,* 2005):

a) Did the model fulfil its objectives?
b) Are the results valid and were the quality requirements realised?
c) Was the choice of spatial and temporal discretisation adequate?
d) Was the choice of model restrictions correct?
e) Was the correct model selected?
f) Was the numerical approach correct?
g) Was the implementation done correctly?
h) Are the sensitive parameters clearly defined?
i) Was uncertainty analysis performed?

There exists no single accepted statistic or test that determines how a model is evaluated and both graphical comparisons and statistical tests are required in the process of evaluation. Another critical issue is how the performance criteria for model uncertainty can be defined. A likely solution would be to set a threshold of uncertainty, for example 65% and if the model does not achieve this then it is rejected. This will not work in all the cases as sometimes less accurate models may be more useful in providing information on the dynamics of the system being modelled. Refsgaard *et al.,* 2004 argue against using this approach and advocate for the consideration of the context of the model when evaluating and testing it. The performance criteria therefore vary from case to case and are context specific.

Sometimes the creation of a flawed model might be the best way of understanding the domain. This is according to Silberstein, 2006 who provides the following quote by Dooge, 1988:

> *"Perhaps the most dangerous thing in hydrology may be a model that fits with expectations."*

This implies that if humans accept that their limited understanding is adequate, then there will be no progress in modelling. Some of the methods that can be used in model evaluation include expert elicitation and extended peer review.

a) Expert elicitation

This is a structured way of acquiring subjective judgements from experts in the domain of study. The experts are selected and the nature of the problem and the elicitation procedure are explained to them. The quantities to be assessed are defined and a familiar scale is selected. A subjective probability density function for describing the uncertainty is then created and verified with the experts.

b) Extended peer review

Stakeholders are involved in quality assurance. They can improve the quality of the problem formulation, contribute to knowledge on local conditions and provide personal observations, which can give rise to new research. The only obstacle to the process is making the stakeholders understand the complex processes. Other methods used for uncertainty assessment, which are mainly statistical, which are specific to Bayesian Networks, are described in Chapter 4.

Model results interpretation and presentation

Interpreting results from model computation and prediction is a very crucial step. The result should be analysed according to the modelling objectives. Any unanticipated results should be presented and explained. Any summaries should state the uncertainties in the result, the results usability and restrictions. The results must be scientifically correct and complete and be presented in a easy to understand format with the use of visual tools like graphs, time-series plots, and GIS (Loucks *et al.*, 2005; Caminiti, 2004).

2.4 Water resources assessment approaches in South Africa

Table 2- 2 lists some of the commonly available approaches for water resources assessment in South Africa. The discussion focuses on tools and methodologies developed by the Department of Water Affairs and the Water Research Commission (WRC), whose involvement in the water industry are mandated by the government.

The Department of Water Affairs is the custodian of all South Africa's water resources and is responsible for policy formulation in the sector. It also has responsibility for water services provided by the local government. The WRC was set-up by government and is mandated under the Water Research Act No. 34 of 1971 to support water research and build research capacity in South Africa (WRC, 2007).

Table 2-2: Modelling approaches for water resources

Methodology	Focus	Discussion	Reference
MIKE BASIN	Catchment sustainability	-sustainability defined by comparing demand with water available - monthly values of water quantity or annual runoff are estimated using a CARMA[4] time series model, rivers are represented as networks with branches and nodes and based on input data, the software calculates the amount of water available in each branch and node -other elements of sustainability like water quality and resource are not included	Kjeldesen *et al.*, 2001
ACRU model	Surface water	-a physical model integrating water budgets and runoff and produces components of the hydrological system with risk analysis -applied in hydrology, crop yield modelling, irrigation supply and demand and water resources assessment	Schulze and Smithers, 2003,
Water Resources Yield model (WRYM)	Surface water	-a network based model to analyse complex water systems on a monthly time-step -used for catchment analysis to assess long and short-term yield capabilities under various operating and growth scenarios	de Jager and van Rooyen, 2008
Water Resources Planning Model (WRPM)	Surface Water	-more detailed than the WRYM and used for detailed operations runs -models varying demands over time and under new resource supply schemes	DWAF, 2009

[4] Continuous-time autoregressive moving average time series.

Methodology	Focus	Discussion	Reference
Water Resources Simulation Model	Surface water	-a rainfall-runoff model -includes groundwater, surface water interaction and stream flow reductions -used to produces maps and data on the status of water resources of South Africa at catchment scale	DWAF, 2009
WQ2000	Water quality	-a network model -provides an initial assessment of the impacts of proposed development options on water quality -provide a regional overview of salinity on a monthly time-step	Herold and le Roux, 2004
WQS	Water quality	-a deterministic monthly time-step hydro-salinity simulation network model -	DWAF, 1988
Aquifer management system (AMS)	Catchment sustainability	-a monitoring database with a GIS front-end and is connected to a water balance model -rainfall, groundwater levels and quality can be monitored over time	DWAF, 2009
South African Groundwater Decision Tool (SAGDT)	Groundwater quality	-a fuzzy logic based risk assessment tool for groundwater resource contamination and sustainability -also assesses human health risks associated with contaminated groundwater -assesses impacts of groundwater levels on ecosystems	Dennis and van Tonder, 2004
Decision support system (DSS) for ICM for the Kruger National Park Rivers Research Programme	Surface water and catchment assessment	-numerical simulation of hydrological data -has models for simulating the effects of stream flow on fish population dynamics, geomorphology and vegetation -the predictive capacity and scenario assessment tools are based on knowledge-based and rule-based systems	Jewitt, 1998; Jewitt *et al.*, 2000.
Catchment sustainability indicators	Catchment sustainability	-twenty indicators prioritised to test the sustainability of catchments -selection and prioritising of indicators involved relevant stakeholders in catchment management	Walmsley *et al.*, 2004

The majority of water resources assessment models used in South African rural catchments are generally conceptual, empirical or physical-based (for example the WRYM model and ACRU). Physical-based models are quantitative and founded on some mathematical equation. These are either in the form of partial differential equations, based on mass-balance concepts or empirical regression equations. These equations are then resolved using numerical methods to output the values for the variables of interest Jewitt *et al.*, 2000 provides an in-depth discussion of these models.

Physical models are complex, provide a more detailed representation of processes than empirical or conceptual models and can evaluate numerous parameters. This requires good quality input and output data at the relevant spatial and temporal scales, mostly for calibration, which are often not available in most catchments (Silberstein, 2006; Hughes, 2004). The models are good at simulating daily river flow and transport dynamics (Hansen *et al.*, 2007). According to Parkin *et al.*, 1996, one of the major challenges arises when they are used in predicting hydrological response in ungauged catchments. Examples of problems include the inability of equations to represent actual field processes and over-parameterisation of the models.

Analysing parameters in great detail can suggest that these models may be more "accurate", but this might also not be true (Ochieng, 2007). Various studies highlighted in Ochieng, 2007 illustrate how similar results can also be obtained from simple models which are highly useful when there is inadequate data or when the quality of the data is not acceptable. In physical-based models, the uncertainty associated with model data inputs, parameter estimation and outputs are generally not presented. This is vital in IWRM where processes and data are often complex and highly uncertain in time and space.

The ultimate objective of modelling is to produce a representation of a system that is capable of simulating all the important dynamics of the processes. It is unlikely that traditional physical modelling techniques can solely provide the flexibility required to analyse physical and socio-economic factors in an integrated approach (Dent, 2000 cited in Dube, 2006). Instead, a framework is required which enables the interaction of the output data and results from these models with other economic evaluation techniques, decision making methodologies and statistical analyses tool. Such a framework must enable multi-scale analysis to handle the qualitative data adequately and allow uncertainties in data and model predictions to be quantified and incorporated in subsequent analysis. Ideally it should facilitate the involvement of the stakeholders affected by management decisions resulting from the models output. Such a framework can be provided by Bayesian Networks (Batchelor *et al.*, 2005).

The main emphasis of IWRM is on the inclusion of stakeholders in water in planning and modelling for integrated assessment. The stakeholders provide information at different stages of assessment from problem formulation, through to model development and result presentation. A structured formal framework is required in order to capture their knowledge of the system.

Clearly, some of the model uncertainties discussed in Section 2.2.4 can be reduced by the use of expert knowledge. Model structure uncertainties arising from lack of adequate knowledge of the system can be reduced by including the conceptual models elicited from the stakeholders/experts. Model technical uncertainties due to approximations and computer bugs can be checked and identified after the presentation of results to the stakeholders.

Bayesian Networks provide the best graphical framework for this because of their simplicity which allows for the inclusion of participants in modelling. Also because they are based on simple probability and not a black-box system, the process of production of the results can be easily tracked and any mistakes rectified.

Although indicator-based approaches like the one presented by Walmsley *et al.*, 2004, provide an integrated way of accessing catchment sustainability, they still lack the framework for documenting or accounting for the uncertainty in data. They offer limited predictive capabilities that allow the indication of the likely impact of future scenarios.

The issues of identifying and quantifying sources of modelling uncertainties in water resources estimation models in South Africa were investigated by Sawunyama, 2008. The research focuses on understanding, characterising and quantifying the uncertainty in water resources estimation models. It centres on hydrology models currently used in South Africa like the ACRU, and the Pitman model. Work done so far has analysed uncertainties associated with parameter estimation of input climate data. Variance based measures and sensitivity analyses were used to assess the combined contribution of uncertainty sources to model outputs. The results also showed the uncertainty sources which were the dominant contributors to model output uncertainty and a regional comparison was done for selected catchments in South Africa.

There is still a gap because the study by Sawunyama, 2008 focuses on existing hydrological models and does not explore novel methods that incorporate uncertainties and enable integration of hydrological models with other socio-economic and environmental model outputs in an assessment study. Bayesian Networks are a suitable method for modelling under these conditions.

2.5 Conclusions

This chapter commenced with a summarised discussion of IWRM and its relation to ICM in the South African context in Section 2.1. In this section, water resources assessment was identified as a key aspect of IWRM. The different levels of water resources assessment were proposed and the necessity of using more integrated approaches to assessment was highlighted by considering the complexity of catchment systems.

The integrated modelling approach to water resources assessment has to address issues such as multiple spatial and temporal scales; and the integration of different types and sources of data and databases and uncertainty. Definitions of the different types of uncertainties available in data and models and how they can be reduced, eliminated or evaluated are provided in Section 2.2. Following this, some guidelines and steps to be followed in modelling are provided in Section 2.3 and these will inform the methodology for this research.

Section 2.4 provides example of the commonly used approaches for water resources assessment. An in-depth discussion on these approaches and some of their limitations highlight the appropriateness of Bayesian Networks for addressing some of the issues. The next chapter, Chapter 3, examines the theory, creation, use and evaluation/verification of Bayesian Networks and Dynamic Bayesian Networks.

CHAPTER THREE: THEORY ON BAYESIAN NETWORKS

3 Introduction

This chapter commences with a brief introduction to Bayesian Networks. Following this is a discussion on the process of creating Bayesian Networks. The steps outlined start with the definition of the modelling purpose and the selection of variables to include in modelling. The creation of the network and population of the probabilities of the variables in the network follows. The use of Bayesian Inference, for querying the network, is briefly discussed. The created network needs to be evaluated and verified and the recommended techniques for performing this are presented. After these steps, the advantages of combining GIS techniques and Bayesian Networks are highlighted. Lastly, Dynamic Bayesian Networks are examined as tools for use in time-series modelling for water resources assessment.

3.1 Bayesian networks

Bayesian networks are graphical models that allow for the representation and reasoning of an uncertain domain (Pearl, 1988; Korb and Nicholson, 2004). A Bayesian network is a directed acyclic graph (DAG) made up of a set of random variables from the problem domain, which are represented as nodes (Mihajlovic and Petkovic, 2001). A graph is made up of nodes (or vertices) and edges (or arcs). A cycle is a path that starts and ends at the same node. A DAG is a directed graph with no cycles.

An arc in a DAG points from one node called the parent node, (for example node *A* in Figure 3- 1) to a child node, (that is node *B* in Figure 3- 1). *A* is an ancestor of *C* and *C* is a descendant of *A* (Mihajlovic and Petkovic, 2001).

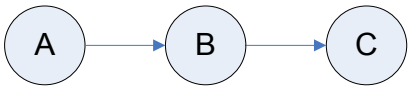


Figure 3- 1: Example of a simple 'serial' connection of three variables.

In a network, variables like *A*, *B* or *C* represents mutually exclusive propositions and can be in any number of states. For example:

Variable: river = *states: {perennial, non-perennial}*
Variable: temperature = *states: {hot, cold}*
Variable: rainfall = *states: {0, 40}*

A variable can be observable or latent (hidden). A latent variable is one in which the states are inferred but not observed directly. To correctly represent the dependence and independence relationships amongst the variables, the direction of the arcs must connect *cause* to e*ffect*. For example, in the network in Figure 3- 1, the direction of the arc from *A* to *B* as shown by the arrow means that *A* is a cause of *B* (Pearl, 1988; Kjaerulff *et al.*, 2008).

The arcs in the Bayesian Network represent the interactions/relationships between the variables (Murphy, 1998). These relationships are expressed as probabilistic dependencies which are calculated through a set of *conditional probability matrices*. For example a variable *A* has states $\{a_1, a_2, \ldots, a_m\}$ and B has states, $\{b_1, b_2, \ldots, b_n\}$.The following is the resulting conditional probability matrix.

$$P(B|A) = \begin{bmatrix} P(b_1|a_1) & P(b_2|a_1) & . & P(b_n|a_1) \\ P(b_1|a_2) & P(b_2|a_2) & . & P(b_n|a_2) \\ . & . & . & . \\ . & . & . & . \\ P(b_1|a_m) & P(b_2|a_m) & . & P(b_n|a_m) \end{bmatrix}$$

If the variables being considered are discrete, the conditional probability matrices are represented as conditional probability tables (CPTs). The CPTs define the probability or likelihood of a variable being in a particular state given the state(s) of its parents (Bromley *et al.*, 2005). The CPT describes the effect the parent variable has on a child variable. If the data is continuous, conditional probability distributions are used. The term CPT is used in this research because it only discusses discrete variables.

Nodes at the top of the network (parent nodes) contain marginal probabilities. Marginal probabilities represent the likelihood of the variable existing in any of the predefined states. A change in the likelihood of the state of a variable is spread through the network using its CPTs. This means that for each parent and each possible state of that parent, there is a row in the CPT describing the likelihood of the child node being in some state (for example Figure 3- 2). A typical CPT for the network is shown in Table 3- 1.

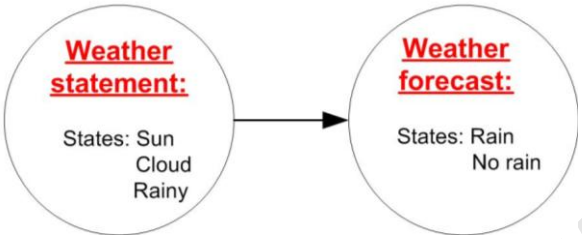


Figure 3- 2: Network showing two variables, weather statement and weather forecast. Weather statement is the parent node and has three states and weather forecast, the child node, has two states.

Table 3- 1: An example of a CPT associated with the weather forecast example in Figure 3-2

	Weather forecast	
Weather statement	*Rain (%)*	*No rain (%)*
Sunny	4	40
Cloud	16	20
Rainy	80	40

The first entry in the first row of probabilities can be interpreted as "the probability that it will rain given that the weather is sunny = 4%". The sum of the probabilities of the different states of the node (the columns) must equal to 100%.

3.1.1 Creating Bayesian Networks

The process followed in developing the Bayesian Network is summarised in Figure 3- 3.

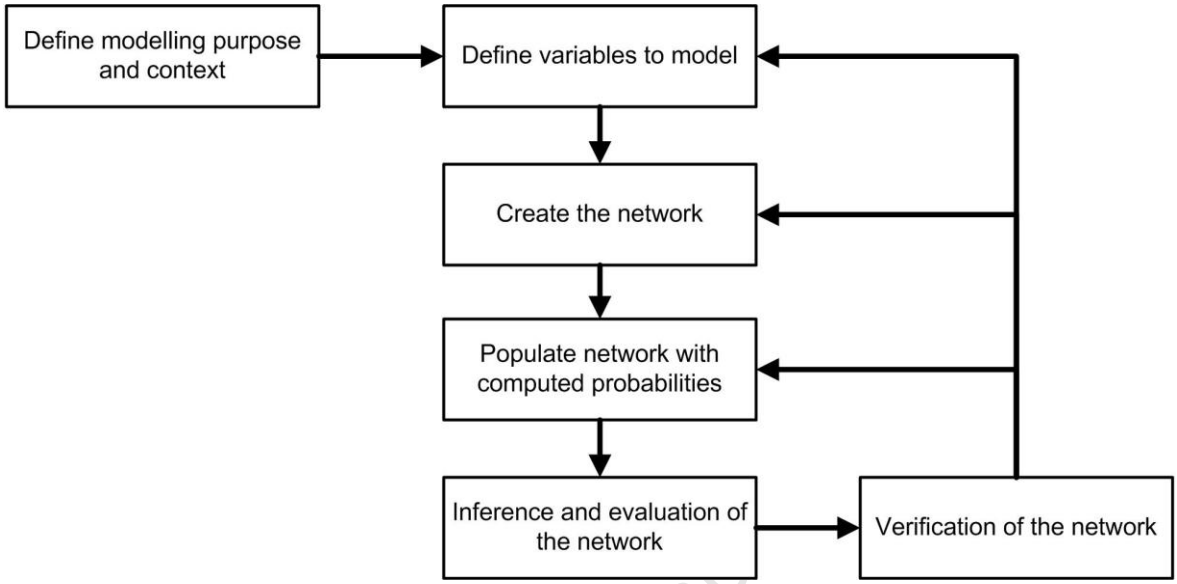


Figure 3- 3: Steps in creating a Bayesian Network.

In the first instance, the purpose of the modelling exercise and the context needs to be clearly outlined as it affects the selection of the variables to use (Barnes, 1995; Loucks *et al.*, 2005). This step is discussed in detail in Chapter 5, when the research methodology is presented. The variables to be modelled and their states are then identified. The network structure is created, either manually or automatically, showing the "cause and effect" relationships between the variables. Using independence and dependence assumptions, the conditional probabilities are computed from the input data.

Following this is the verification and evaluation of the network and its use for scenario analysis and testing. From the results of scenario analysis and model verification and evaluation, conclusions might be drawn about the irrelevance of some variables and the need to modify the network might arise. The parameters of the network can also be continuously updated when new information is obtained. The process of creating a network is therefore iterative (de Santa Olalla *et al.*, 2005; Kjaerulff *et al.*, 2008). The steps introduced here are discussed in detail in the following sections.

3.1.2 Definition of variables to use in Bayesian Network modelling

Ideally, the definition of variables that are relevant to a study and their states should be carried out in a formal, well-structured process that involves relevant stakeholders or experts in the problem domain (Baran and Jantunen, 2004; Jakeman *et al.,* 2006). In this research, although stakeholder input and knowledge was solicited, this was not carried out in a formalised process due to time and financial constraints. The main challenge to this process is finding suitable and knowledgeable experts who have the time to assist. Conflicting or subjective information often arises from stakeholder participation. In order to manage this, guidelines have been created from various studies and these outline the most appropriate ways of eliciting expert knowledge (Korb and Nicholson, 2004; Uusitalo, 2007).

The choice of the variables can also be informed by policy, or published literature in the application domain. The availability of data and the modelling scenarios of interest also affect the selection of variables. The variables can either be discrete or continuous. A discrete variable has a finite or countable number of possible values and a continuous variable can take on any value in some interval. Discrete variables are more commonly used than continuous variables as these are easy to interpret and are more readily defined and better handled by the Bayesian Network technology (Korb and Nicholson, 2004; Ismail, 2003). The variables selected for this research are discussed in Chapter 5.

In defining the states of the variables, either a supervised or an unsupervised approach is used. If the data is continuous, this process is called discretisation. In a supervised approach, the states are set by the user using exploratory data analysis, or information from experts, policy or literature. In an unsupervised method, a suitable algorithm is used to automatically portion the data; the suitable intervals for the thresholds being selected based on statistical analysis. There are however, some challenges with discretisation and these can be demonstrated with an example. In Table 3- 2 in the second column, there is a list of rainfall readings that were taken over the seven days of the week.

Table 3- 2: Discretisation of continuous data

Day	Rainfall (R) (mm)	Discretised Intervals = 4	Discretised Intervals = 6
Monday	2.5	x1	y1
Tuesday	4.1	x1	y2
Wednesday	7.5	x2	y3
Thursday	11.1	x3	y4
Friday	13.3	x3	y5
Saturday	15.0	x3	y6
Sunday	18.6	x4	y6

The third column in Table 3- 2 shows the result of discretising the data using the intervals defined in Table 3- 3. The fourth column in the Table 3- 2 shows the result of discretising the data using the intervals defined in Table 3- 4.

Table 3- 3: The discretisation of data using four intervals

State	Interval
x1	0< R <=5
x2	5< R<=10
x3	10 <R <=15
x4	>15

Table 3- 4: The states defined from data after discretisation using six intervals

State	Intervals
y1	0<R<=3
y2	3<R<=6
y3	6<R<=9
y4	9<R<=12
y5	12<R<=14
y6	>14

More dependencies and relationships are likely to be found when the data are discretised using fewer intervals. On the other hand, having more intervals enables the mining of more complex relationships (for example in Table 3- 3, if four intervals are used the difference between the rainfall readings taken on Thursday, Friday and Saturday is lost in the analysis whereas with six intervals, it is captured). There is a need to strike a balance and the suitable interval can sometimes only be selected after a sensitivity analysis/scoring assessment is performed. The problem is that in catchment modelling, there is hardly enough data to enable partitioning into many intervals. Conditional distributions become weakly defined if there are a few data points per distribution.

The process of finding the best technique for discretisation or the most optimal intervals is crucial and is still a subject for research (Uusitalo, 2007). In most studies, an iterative approach is adopted where a scoring measure or sensitivity analysis might lead to conclusions that the partition of data into many intervals does not bring considerable changes and will therefore not be warranted.

3.1.3 Building the Bayesian network

The building of the network is defined as the process of first creating a graphical network and then calculating the conditional probability tables (CPTs) associated with the variables. This process is called Bayesian Network learning (Nilsson, 1998) and can be done either manually with the assistance of domain experts or automatically, with the software "learning" the structure from the input data (de Santana *et al.*, 2007).

The process of structural learning through eliciting expert knowledge becomes challenging in complex situations where the relationships between variables are not well known. Experts might fail to agree on different aspects of the structure and also some hidden patterns in data might potentially be left unexplored (Uusitalo, 2007).This research utilises an automatic learning method.

The algorithms mostly used for automatic learning can be divided into the following broad categories (Cheng *et al*., 1997):

a) search and scoring algorithms; and
b) dependency relationship analysis algorithms.

The search and score algorithm starts with a graph that has no arcs. It searches through the space of possible existing structures, adding new arcs each time. After this, it calculates a score to assess if the new structure is better than the previous one (Hand *et al*., 2001). The process continues iteratively until no new arc can be added and no new structure is better than the previous one in terms of the score calculated.

Different scoring metrics can be used, for example, Bayesian scoring, entropy based measures and minimum description length measures (Cheng *et al.,* 1997; Cooper and Herskovits, 1992). The discussion of these is outside the scope of this research. Figure 3- 4 illustrates the concept of the 'search and score' algorithms as illustrated by Meek, 2003.

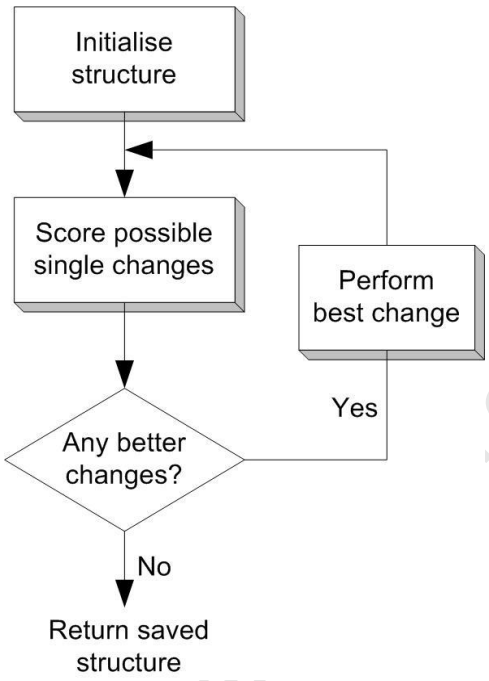


Figure 3- 4: Search and score Bayesian Network structure learning algorithm (adapted from Meek, 2003).

Most of these algorithms apply heuristic search methods[5] and the nodes must be ordered to reduce the search space (de Santana *et al.,* 2007; Cheng *et al.,* 1997). Examples of these algorithms are described in more detail in Cheng *et al.,* 1997. Some researchers have argued against producing a single best solution/network after computation and have developed algorithms that produce several networks and these are then "averaged" to generate one "optimal" network (Chickering *et al.,* 2000). This is elaborated in Section 3.1.7.

[5] Heuristics are techniques that experimentally provide good computer algorithm performance but do not guarantee a solution. They use knowledge about a domain and this helps improve the efficiency of a search, minimise search space and guide reasoning in the domain. (Carbonell, 2003; Donald, 1998).

The dependency relationship analysis algorithm works on the premise that a structure encodes the dependencies among the variables in the underlying network to be created. Algorithms in this category attempt to discover these dependencies in data and use them to deduce the Bayesian Network Structure (Cheng *et al.,* 1997; Margaritis, 2003). Once conditional dependency is identified in the data between two variables, an arc is placed between them in the network. After the locations of arcs are identified, their directions are assigned so as to correctly represent the conditional independencies (Bouckaert, 2007).

Some of the problems common to both types of algorithms include the following (Cheng *et al.*, 1997; Wong and Leung, 2004):

a) they require that nodes should be ordered, which cannot be achieved in most cases;
b) for the search and score algorithms, the search spaces are usually large; and
c) an exponential number of dependency tests have to be done in the dependency analysis algorithms.

Due to the shortfalls of these types of algorithms, research has been focused on creating hybrid models that involve a combination of the two (Wong *et al.*, 2004). This research utilises a novel approach, an unsupervised Hybrid Genetic Algorithm (HGA) (Osunmakinde *et al.*, 2007), which is described in more detail in Section 5.1. In the Bayesian Network learning process, once the structure of the network has been created, the next step is the learning of parameters for the conditional probability tables (CPTs) from the data.

3.1.4 Computing conditional probability tables

The process of computing conditional probabilities tables for each variable/node is called parameter learning (Xenos, 2004; Kiiveri, *et al.,* 2001). The calculations of CPTs can be done either through domain expert knowledge elicitation or from known or existing data (Wang, 2004). The use of domain experts leads to the realisation of subjective estimates which are usually error-prone.

The process of eliciting probabilities from domain experts can also become unfeasible when the amount of variables to consider increases. The bias in estimating probabilities can be in the form of overconfidence, experts can attribute higher than justifiable values or in some cases they can make the variable more probable because it is more significant to their analysis. In an attempt to reduce some of the bias, tools and guidelines have been established to guide the process (Uusitalo, 2007).

During automatic learning of the CPTs from known data the following situations can be prevalent:

i) all the variables are observable;
ii) some variables cannot be measured (in this case they are called *hidden/latent variables*) and as shown in Figure 3- 5, these represent variables whose values are not given in the data and are introduced to simplify the network and enable the definition of more states for the given variables (this allows for creation of more complex models) (Cowell *et al.,* 1999); or
iii) the variables were not measured due to some reason like instrument failure (*missing variables*).

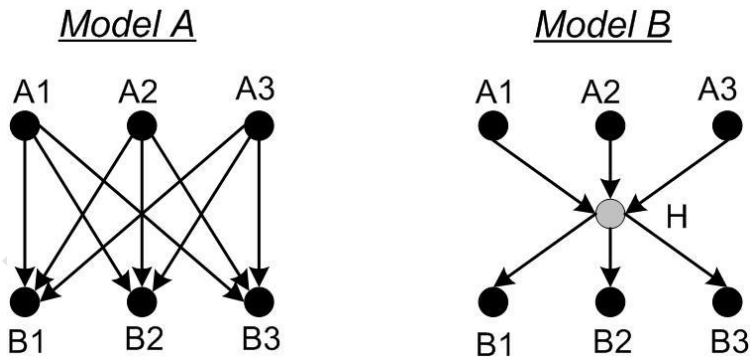


Figure 3- 5: The introduction of a hidden variable in a model. Model A is the initial model and model B is the result of adding a hidden variable (node) H in the model. Model B is simplified and more computationally efficient (Nilsson, 1998).

If all the variables are observable, then ordinary statistics can be used to calculate probabilities for the CPTs. This is illustrated in Table 3- 5 with three variables, rainfall, temperature and recharge.

Table 3- 5: Training set data used to populate the CPT (adapted from Nilsson, 1998)

Rainfall	Temperature	Recharge	Number of instances
high	low	high	$Count_1$
high	low	low	$Count_2$
high	high	low	$Count_3$
high	high	high	$Count_4$
low	low	low	$Count_5$
low	low	high	$Count_6$
low	high	low	$Count_7$
low	high	low	$Count_8$
Total			$\sum_{i=1}^{8} count_i$

Assume Table 3- 5 is provided as a training set to be used in CPT learning. A count can be done of the number of times the variables exist in the following states:

{rainfall= high, temperature=low, recharge =high}

The conditional probabilities are then estimated by the ratio of the corresponding counts (Cowell *et al.,* 1999).

If there are hidden or missing variables, algorithms like the *Expectation Maximization algorithm (EM)* can be used to estimate the values of the hidden data (Cowell *et al.,* 1999). The unknown variables are divided into parameters and hidden variables. *EM* uses point estimates to represent parameters (Θ) and distributions (S) are used in the place of the hidden variables. The process occurs in two steps, the Expectation (E) step and the Maximization (M) step. In the E step, given certain values, the posterior distribution of hidden variables S is solved. In the M step, the parameters (Θ) are updated to maximise the likelihood given a fixed distribution over hidden variables S (Cowell *et al.*, 1999; Pearl, 1988). This algorithm is commonly used in Bayesian Networks to fill in the missing values for variables.

3.1.5 Bayesian Inference

Bayesian inference is the calculation of the posterior probability for a set of query nodes based on given values for some evidence nodes (Pearl, 1988; Korb and Nicholson, 2004). The power of Bayesian networks is in their capability to follow how the changes in certainty of one variable affect the certainty of other variables in the network, that is, when evidence is available. Evidence is defined as:

> "*a definite finding that a node A has a particular value i.e. A = a".*

Bayesian inference can be represented by the following equation:

$$P(H|e) = \frac{P(e|H) \times P(H)}{P(e)} \text{ (from Bayes' Rule)}$$

Where:

H = the hypothesis involving query variables
e = a set of evidence variables

Bayesian inference can be diagnostic (bottom-up) or predictive (top-down) or a mixture of both as illustrated in the example in Figure 3- 6.

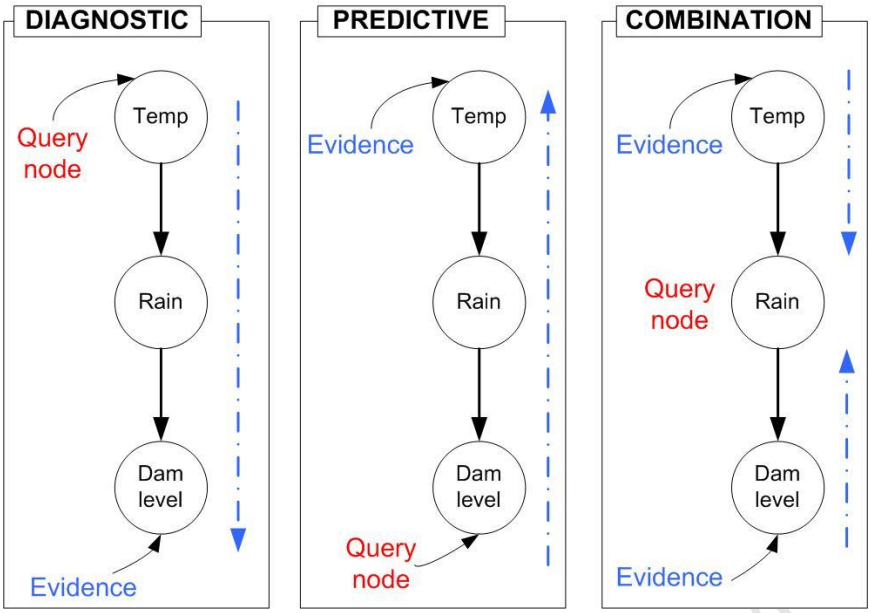


Figure 3- 6: Diagnostic inference, predictive inference and a combination of both.

Diagnostic inference is determining the probabilities of causes from given effects. Predictive inference is calculating the probabilities of effects from specified causes. A combination of both types of inference is when the probability of a query node being in any state is established by providing evidence of its effects and causes. Figure 3- 6 gives examples of the three methods.

Bayesian inference can be exact or approximate. In exact inference, the probabilities are calculated from measured data and the results obtained are assumed to be correct. This can become computationally unfeasible for some networks and approximate inference is used instead (Pearl, 1988; Korb and Nicholson, 2004).

In approximate inference the probabilities are estimated using randomised sampling algorithms for example Monte Carlo and Gibbs sampling. Approximate answers are given to queries and the accuracy of results depends on the number of samples generated from data, the more the data, the more accurate the results. A discussion on approximate sampling is outside the scope of this research as this was not used. More information is provided by Pearl, 1988.

3.1.6 Evaluation of the network

After the network has been developed and populated, it needs to be evaluated. The first type of evaluation should involve domain experts to assess the following (Korb and Nicholson, 2004):

i) if all the variables used are relevant and the range of values is exhaustive;
ii) if the variables are named and defined appropriately;
iii) whether or not the ranges used for discretisation are appropriate; and
iv) if there is consistency in the states of the different variables.

This should be followed by a review of the probabilities.

The second type of evaluation is sensitivity analysis. Sensitivity analysis assesses how sensitive the network or the probabilities of the query nodes[6] are to changes in the inputs/evidence values (evidence sensitivity) or changes in parameters (parameter sensitivity) (Korb and Nicholson, 2004; Jensen, 2001; Bednarski *et al.*, 2004).

Evidence sensitivity analysis is the testing of how responsive the results of a propagation of evidence are to variations in the set of evidence. Some of the questions that evidence sensitivity analysis provide answers to are (Kjaerulff *et al.*, 2008):

i) the minimum or maximum probabilities obtained from observing a variable;
ii) which evidence supports or goes against a postulated scenario[7];
iii) the evidence that can be used to discriminate different scenarios; and
iv) what if a selected observed variable had been observed to a different value than the current one?

The answers to these questions can provide an understanding of the conclusions reached by the model (Kjaerulff *et al.*, 2008).

Parameter sensitivity analysis is the analysis of how sensitive the results of evidence propagation are to variations in the value of a parameter in the model. Through sensitivity analysis, parameters of the network that have a large or small impact on the probability of a hypothesis given evidence can be identified.

[6] The node (or variable) that is being interrogated during modelling.
[7] A scenario is defined in this section as an outline of an hypothesized chain of events.

When there is not enough data or knowledge, parameter sensitivity analysis can be used to concentrate on the most influential parameters and the result can be used to focus the model development/revision process (Kjaerulff *et al.*, 2008).

Apart from sensitivity analysis, another useful analysis is the value of information. In Bayesian Network modelling, there is the option to gather additional information to improve the solution. Before this can be done, it is convenient to assess the usefulness/value of making additional observations and adding this information before it can be sourced. In the value of information analysis the task is to establish the value of information gathered from different sources and its impact on the predicted results or sensitivity of the network (Kjaerulff *et al.*, 2008). This assists in reducing the cost of further data collection exercises.

Sensitivity analysis is performed on the resulting Bayesian Network model developed in this research. The concepts of sensitivity to evidence and parameter sensitivity analysis introduced here are elaborated in Chapter 6 with examples showing how the measures are calculated in this research.

3.1.7 Verification of the network

The last stage in modelling is the verification of the network. Verification is the process of ensuring that the network is an accurate representation of the domain being modelled (Korb and Nicholson, 2004). Various methods are available for this and the easiest method involves the splitting of the nodes or variables into two subsets:

i) the query nodes; and
ii) the evidence nodes.

Sample data are used to assess how well the network predicts the values of the *query nodes* when the *evidence nodes* take the values observed in the sample set. The most popular method for verification is the *predictive accuracy method*. The initial sample data is divided into two sets, the training and the test set. The training set is used to learn the structure and the parameters of the Bayesian Network to be created (Korb and Nicholson, 2004).

The test set is then used to assess the accuracy of prediction of the network through the following procedure:

i) a target variable or node to be tested (query node) is selected;
ii) any available evidence is added for all the other nodes;
iii) taking the predicted value for the target variable as that which has the highest probability; and
iv) comparing the predicted value with the actual observed value in the sample test data.

The predictive accuracy method can be used to verify the network and make recommendations for revision or for the selection of the "optimum" Bayesian Network model to use from several options arising from the modelling process. Different models can result based on the differences in expert perceptions (if their participation is elicited) or the variations in the method used for either discretising the continuous data or the use of different algorithms for automatically mining the network structure. The process of selecting the "optimum" network is called Bayesian Network selection.

Researchers like Chickering and Heckerman, 2000; Hoeting *et al,* 1999 have done studies that show that the best approach is Bayesian Network averaging as opposed to selection. Bayesian Network averaging is the summing of all possible networks and producing an "optimum" network that would be an average of possible solutions. However, Chickering and Heckerman (2000) also state that in real world applications, the sum of all possible models is complex and the averaged model is difficult to interpret. In most situations, Bayesian Network selection is adequate.

The choice of an "optimum" model can be made after scoring the available models and making a comparison of the results. Several scoring measure are available and these ranges from more complex functions to simple statistical measures. More complex scoring functions includes those defined by Chickering and Heckerman, 2000; Steck and Jaakkola, 2003. The development of the complex scoring functions is a subject of ongoing research and most are not implemented in commercial or available open-source software.

In their paper of forecasting categorical data, Zhang and Casey, 1999 state that there is no advantage in using elaborate scoring functions if no more useful information is conveyed to the user by representing the forecast information in multiple categories. Easily scored and easily interpreted probability rules can be used.

The derivation of scoring measures is generally based on the following premise. Suppose a forecaster makes some predictions and then gives the client/user probabilities $p_i, \ldots p_n$ for all the events. The client takes a decision based on these probabilities and if the *i*th event occurs, then the client pays the forecaster $R_i(p_i, \ldots p_n)$ abbreviated $R_i(p)$. $R_i(p)$ is called the pay-off or probability scoring metric. It is a function of the difference between the actual events and the assessed probabilities (Morgan and Henrion, 1990)

The commonly used simple universal probability score metrics are (Stanski *et al.*, 1989; Pearl, 1978):

i) the Brier score (or quadratic score) (Cofiño *et al.*, 2002);

ii) the logarithmic score; and

iii)spherical metric

The Brier score is a quadratic function for the differences between the assessed probabilities and the fraction that happen. It has been widely used in studies on rainfall probability forecasting (Morgan and Henrion, 1990). The score is the mean square probability error and is expressed as follows:

$$1 - 2 \times P_C + \sum_{i-1}^{n} (P_i)^2$$

The logarithmic metric is given by the following equation:

$$-\log(P_C)$$

The spherical score is given by the following equation:

$$\frac{P_C}{\sqrt{\sum_{i=1}^{n} P_j^{\,2}}}$$

Where: n = number of states of a variable

P_c = the probability predicted for the correct state

P_i = the probability predicted for state i

The score is then averaged over all the cases and the mean value is produced. The quadratic score ranges from 0 to 2 with 0 being the best performance and 2 being the worst. The logarithmic score values are calculated using the natural log and are between 0 and infinity, with zero indicating the best performance. The spherical payoff is between 0 and 1, with 1 being best.

The predictive accuracy and scoring metrics discussed above are used for the validation process performed in this research and more information on the results obtained is provided in Chapter 5.

The discussion thus far has followed through the process of creating and evaluating Bayesian Networks. One aspect that has not been discussed yet is the presentation of results. Most software packages output the results as tables and graphs showing the probabilities. In catchment modelling, where most of the datasets are spatial, it is not the most adequate way to represent the results. The next section discussed other more suitable presentation tools.

3.2 Combining GIS and Bayesian Networks

Geographic Information System (GIS) is an appropriate tool for the presentation of modelling results. GIS provides capabilities for data pre-processing, spatial analysis, presentation and post-processing (McKinney *et al.*, 2002; Koutsoyiannis *et al.*, 2003). Combining Bayesian Network with GIS functionality is imperative in catchment modelling because most of the data are spatial and the end users or decision-makers are accustomed to using maps in planning and management. The solution in most case studies has been the combination of Bayesian Network modelling software with an existing GIS software product or the development of a custom product (McKinney *et al.*, 2002).

There are different approaches to linking GIS with Bayesian Networks and these are (McKinney *et al.,* 2002):

i) loose coupling, in which the two are separate systems and data are transferred between the model and the GIS system;
ii) a tight coupling where the data management for the GIS and the model are integrated and they share a database; and
iii) an embedded system in which the BN modelling environment and GIS are in a single manipulation framework which includes the storage and mining of GIS data and presentation of inference results in the same application.

The loose coupling approach is the most prevalent in literature. In the simplest application, GIS is used to pre-process the data. An example is when the study area is subdivided into a grid made up of cells (pixels) of similar sizes. Each pixel would then contain value(s) for the variable(s) used in modelling. The final probabilities can be obtained by estimating the proportion of pixels that fall in each "state" of each variable (Stassopoulou *et al.,* 1998; Sadoddin *et al.,* 2005).

After pre-processing in the GIS software, the results are used to populate CPTs in the Bayesian Network model. The work done by Stassopoulou *et al.,* 1998 documents one of the earliest studies in combining GIS and Bayesian Networks. Pourret *et al.,* 2008 acknowledge the progress to date by highlighting researchers that have made substantial contributions. Examples cited included work done in France by Cavarroc and Jeansoulin, applications being developed at Pennsylvania State University, land analysis studies by Scarlatti and Rabino and models developed at the University of Nice.

In some variations of the loose coupling approach, after modelling, the results are exported to GIS format. They are linked to points or polygons indicating areas/regions of the area for spatial display and presentation. This approach was used by Smith *et al.,* 2007 in their study of using Bayesian Belief Networks to predict habitat suitability maps. GIS data was used to populate some of the variables. Maps representing all GIS variables were intersected using ArcGIS version 9.1 to form a single layer. Statistical analyses were used to find correlation relationships between the variables and these were used to define the structure of parts of the network.

The resultant attribute tables (with about 25 000 records) were used to create a case file used in calculating the CPTs in the final model. The results of inference and scenario analysis were then exported and joined to the GIS layer to produce the final habitat suitability map.

One of the major drawbacks of the loose coupling approach is that each time inference is performed; the results have to be exported from the Bayesian Network software and re-imported into the GIS package before being finally displayed. This process can be time-consuming especially in real-time decision-making.

Ames (2005) presents a tight coupled approach in which the MapWindow® GIS software is combined with the GeNIe® Bayesian Network modelling software. This approach involves the depiction of the catchment on a map as a river network made up of nodes and arcs. Each node/point in the river network is modelled as a variable and the relationships between the different locations in the catchment are represented by arcs. When inference is performed the results are displayed in real time on the map. A major shortfall of this method is that it does not provide the capability of analysing the relationships between multiple variables at each node/point.

The most ideal approach is an embedded system and the study by Grêt Regamy *et al.,* 2006 is an example. It is a custom-developed embedded system which combines Hugin® with ArcGIS 8.3® for avalanche risk assessment. The Hugin® API library was linked to the Visual Basic programming language. ESRI MapObjects 2.2 was used to call the objects in the library from the Visual Basic environment in ArcGIS. The aim was to estimate the risk on a cell by cell basis. The study area was divided into a 5 metre x 5 metre grid with each cell having values representing the multiple variables that are input into nodes in the Bayesian Network. These values are then used in an avalanche model to produce a probability model that can be used to populate the CPTs. The Bayesian Network is then evaluated and the output results, which are the annual risks for each cell expressed in monetary terms, are linked and displayed on the map.

The software development for the research documented in this thesis combines a loose-coupling approach and an embedded system. The specifics are discussed in more detail in Section 5.1.

3.3 Dynamic Bayesian Networks for time-series analysis

Standard Bayesian Networks model static situations with a fixed set of variables, that is, they reflect the condition at one point in time. Most catchment processes evolve over time so time-series modelling is essential. Different approaches have been documented for time-series modelling the simplest one being the inclusion of time as a random variable (or node) in the Bayesian Network. This complicates the network as it increases the number of nodes. Also, in some instances, the temporal variable cannot be defined as a random variable (Mihajlovic and Petkovic, 2001)

Dynamic Bayesian Networks (DBNs) were specifically developed to model processes that evolve over time (time-series modelling) (Russell *et al.,* 2004 cited by Amir, 2004; Pearl, 1988). In catchment modelling, there are documented examples of their use in modelling and predicting river and lake water pollution (Lamon and Stow, 2004 cited in Shihab, 2008) and in groundwater quality modelling (Shihab, 2008).

In a DBN, a time slice *t* is used to represent a snapshot of the evolving process. The DBN comprises a sequence of sub-models each representing the state of the system at a particular point in time. The changes in the sub-models of the DBN over time can be of two types: the changes of the CPTs over time and the change of the structure (that is the introduction and removal of arcs and/or nodes from the DBN structure over time). The introducing and removing of arcs and nodes of the network structure over time is a complex task and only a few techniques have been developed to deal with this issue. The change of CPTs in the sub-models over time is easier to implement and is commonly used.

Each variable *X* in a DBN is associated with a time slice *t* and denoted X^t .The number of time slices required to model a problem is called a *time span T*. The number of variables associated with each time slice is called the *slice size n*. The relationships between variables in the same time slice are depicted by *intra-slice* arcs, which are represented by CPTs (Jensen, 2001) (see Figure 3- 7).

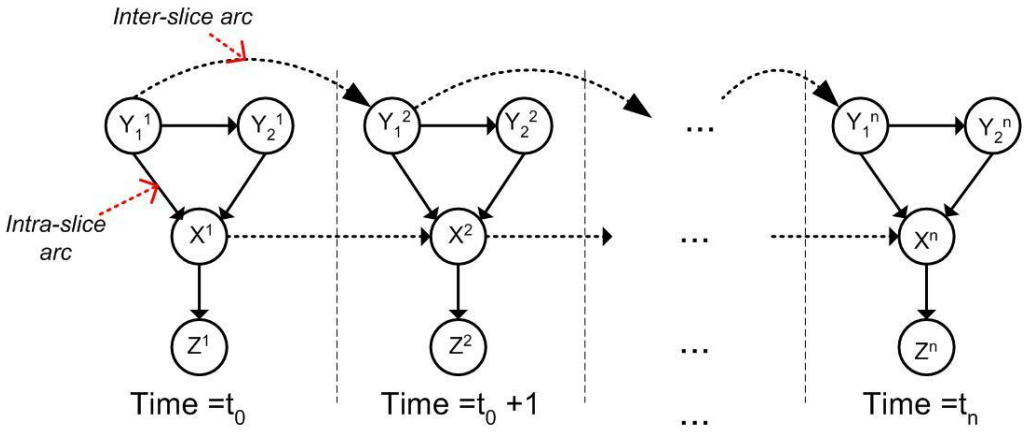


Figure 3- 7: Illustration of a Dynamic Bayesian Network showing intra-slice arcs as solid arrows and inter-slice arcs as dotted arrows.

The relationships between variables at successive time steps are indicated by *inter-slice (or temporal)* arcs which integrate conditional probabilities between variables from the different time slices (Korb and Nicholson, 2004; Mihajlovic and Petkovic, 2001).

These relationships include:

i) the relationship between the same variable over time (this link is always present because the value of a variable at one time-step affects its value at the next one); and
ii) the relationships between different variables over time.

The interface of a time slice is the set of variables with parents in the previous time-slice. For example in the network in Figure 3- 7 at time t_0+1, the interface is the set $\{Y_1^2, X^2\}$. If the structure of the model is the same for all time slices, then the model is called a *repetitive temporal model* (Darwiche, 2001). This is the model used in this research.

In order to specify a DBN the following three sets of parameters are required (Korb and Nicholson, 2004):

i) state transition probability distribution functions (pdfs) that specify time dependencies between states;
ii) observation pdfs that specify the dependencies of observation nodes to other nodes at time slice *t*; and
iii) the initial CPTs for the beginning of the process which is the first time slice t_0.

A DBN is *stationary* if the transition probability distribution functions do not change between the time steps. A DBN is *stationary* and *first-order Markovian* if it satisfies the following conditions (Darwiche, 2001; Kjaerulff *et al.*, 2008; Mihajlovic and Petkovic, 2001):

i) it has the same structure at any time slice *t*;
ii) the state of the system at time *t+1* depends only on its immediate past, that is the state at *t*;
iii) the only cross-arcs allowed are those signifying change from time slice *t* to *t*+1 (also called temporal arcs).

As Dynamic Bayesian Networks are a subclass of Bayesian Networks, most structure-based algorithms developed for Bayesian Networks are directly applicable (Darwiche, 2001). The following tasks are also applicable to DBN:

i) inference (see Section 3.1.5);
ii) decoding: the process of finding the most likely sequence of hidden variables given the observations (Murphy, 2002);
iii) learning: when given the number of sequences of observations, parameters of a DBN that best suit the observed data are estimated and the best model for the system is created (the same methods used in static Bayesian Networks are applicable); and
iv) pruning: this is the process of determining which models are semantically important for inference in the structure and removing the insignificant ones from the network.

Decoding and pruning are not relevant to this research so they are not discussed in detail, more information can be obtained from Murphy, 2002b; Mihajlovic and Petkovic, 2001. CPTs parameter learning is done using the same methods for Bayesian Networks stated in Section 3.1.4.

In DBNs, the states of the model do not have to be directly observable and only a subset of states can be observed at each time slice. Inference is the procedure used to calculate all the unknown states in the network and is shown by the example in Figure 3- 8.

The shaded circles (nodes) represent the states that need to be estimated and the unshaded circles (nodes) are observations. Given the values of observation nodes in every time slice, y_i, the aim of inference is to estimate the values of hidden nodes x_i where node i receives values from 0 to T-1 (Mihajlovic and Petkovic, 2001).

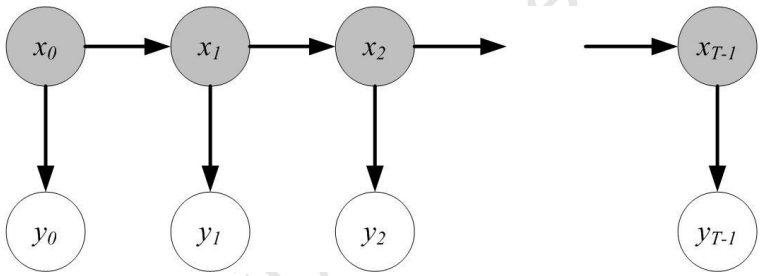


Figure 3- 8: Illustration of the concept of inference on DBN, the unshaded circles are observation nodes and the shaded circles are to be estimated during inference.

As illustrated in Figure 3- 9, inference can be divided into the following main categories (Murphy, 2002a):

a) filtering;
b) prediction; and
c) smoothing.

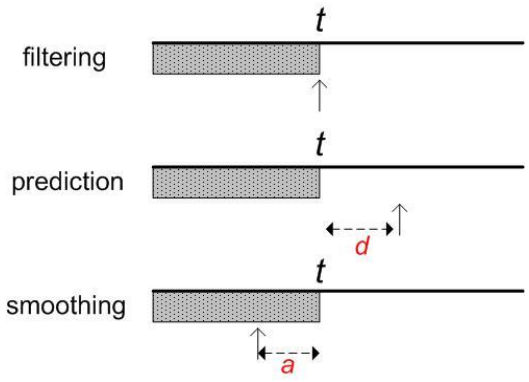


Figure 3- 9: Inference in Dynamic Bayesian Networks.

In Figure 3- 9, the present time/state is *t.* Filtering involves monitoring the probability of some present states given evidence about the past and the present. This is useful for monitoring and tracking. Prediction deals with the forecasting of future observations (for example at time $t + d$) or the estimation of hidden states based on past observation data.
This is the most relevant type of inference in catchment modelling. Smoothing is the estimation of the state of the past (for example at time $t - a$), given all evidence up to the current time.

This research restricts to the inference capabilities of Dynamic Bayesian Networks. They are used to predict some variables of the catchment and more information on their application is provided in Chapter 5.

3.4 Conclusions

The chapter provided a discussion of Bayesian Networks and the steps involved in creating them. Firstly, the modelling purpose and the variables to be included have to be outlined. The variables can be continuous or discrete but due to the limitations of most of the commonly Bayesian Network modelling software products, only discrete data can be used. Continuous data therefore has to be discretised before use and this process can be informed by expert knowledge, standards, policy or by using existing data. The challenges of the different approaches are discussed in Section 3.1.2.

For creating the structure of the network, two options are presented in Section 3.1.3; manual construction using expert knowledge and automatic mining from existing data. The advantages and disadvantages of using either method are presented in Section 3.1.3. Following the creation of the structure is the population of CPTs and this can also be performed using either expert knowledge or data as outlined in Section 3.1.4. Bayesian Inference techniques for querying and using the network are examined in Section 3.1.5. Section 3.1.6 discusses the approaches used in model evaluation and sensitivity analysis is highlighted as one commonly used method. In terms of the verification of the Bayesian Network, different scoring metrics that are used are outlined in Section 3.1.7.

The chapter discusses the options and advantages of combining GIS with Bayesian Networks using examples from literature in Section 3.2. The chapter concludes with a discussion on Dynamic Bayesian Networks, which are an extension of Bayesian Networks but are more suitable for modelling time-series process as is presented in Section 3.3

The next chapter reviews case studies from literature on the application of Bayesian Networks in catchment studies. International case studies and some examples from South Africa are provided. Some of the challenges and advantages of Bayesian Network modelling introduced in this chapter are discussed in more detail.

CHAPTER FOUR: BAYESIAN NETWORK APPLICATIONS IN WATER RESOURCES MANAGEMENT

4 Introduction

Chapter Two recommended Bayesian Networks as an appropriate modelling technique for water resources assessment. A discussion of how they are created, evaluated and presented was provided in Chapter Three. This chapter reviews some of the attributes of Bayesian Networks that make them a favourable approach and their limitations by critiquing examples of applications from literature. Although it is preferable and attempts were made to provide an exhaustive discussion of examples, in reality due to time limitations this is a challenge. The cases discussed in this section are therefore those mostly focused on addressing water-related issues.

The purpose of this literature review is to assess how some of the complex issues[8] of water resources assessment discussed in Chapter Two were addressed in the different applications. The discussion highlights the following issues:

i) the modelling scope and objectives and the software used;
ii) the process used to decide the variables to include in the network and the states of the variables;
iii) how missing data are handled or the methods used to estimate it or fill in the gaps;
iv) how the structure of the network was created, whether this was learnt automatically using a software or manually, using expert knowledge;
v) the method used during parameter learning to populate the CPTs;
vi) whether or not there was time-series analysis involved;
vii) if GIS was used in the application, this could be data pre-processing, management or results presentation; and
viii) the methods used to validate and verify the network.

The cases are summarised in Table 4- 1 and discussed in the following section.

[8] The issues discussed include the need to integrate multiple data, models and databases and the multiple scales of system behaviour.

4.1 International case studies

Table 4- 1: International case studies of Bayesian Networks (BN) applications in catchment modelling summarised

Case study	Scope	Summary	Software	Variables and states	Structure and parameter learning	Verification and validation
Said *et al.*, 2006 (Idaho, USA)	Water sustainability	-assessed impacts of management decisions on water sustainability at catchment scale -considered water supply, demand, quality, rainfall, slope, land cover, population, soil	Hugin	-experts -data	-experts -data	-sensitivity analysis
Ticehurst *et al.*, 2007 (Australia)	Water quality	-evaluated the sustainability of coastal lake catchments and impacts of management actions -included water quality, socio-economic factors, wetlands, fish population, flood risk and threatened flora and fauna	Interactive Component Modelling System (ICMS)	-experts -data -literature	-experts -data, models -literature	-experts -sensitivity analysis
de Santa Olalla *et al.*, 2005 (Spain)	Irrigation water use and demand	-assessed the impacts of different management options on water resources availability -included rainfall, water demand and use, available water income from agriculture, socio-economic data	Hugin	-experts -data -literature	-experts -data	-experts -sensitivity analysis
Ames, 2002 (Utah, USA)	Water quality	-evaluated the risks/benefits of legal requirements in phosphorus management -included phosphorus concentration, stream flow	Netica	-experts -data	-experts -data, models	-experts -sensitivity analysis
Bromley *et al.*, 2005 (UK)	Water demand management	-assessed the impact of pricing policy on water demand -included housing numbers and types, water use, climate, leakage data, greywater reuse and community awareness of policy	Hugin	-experts -data	-experts -data, models	-sensitivity analysis
Arancibia and Moriarty, 2003 (Peru-Sierra)	Watershed management	-evaluated the impacts of different land use scenarios -included management capacity, rainfall, available water autonomy of the community, income from agriculture and project sustainability	Netica	-data -literature	-data -literature	-scenario analysis

Case study	Scope	Summary	Software	Variables and states	Structure and parameter learning	Verification and validation
Soncini-Sessa *et al.*, 2007 (Italy)	Irrigation water use and demand	-evaluation of policy alternatives for irrigation -included irrigated area, irrigation system, biomass, soil, evapotranspiration, plant growth, air temperature, solar radiation, canopy cover - dynamic modelling for time-series analysis of soil state and biomass	TwoLe	-experts -literature -models	-experts -models -literature	-predictive accuracy
Chee *et al.*, 2005 (Australia)	Water quality	-modelled effects of environmental flow management on freshwater catfish -included daily flow, river bottom water quality and salinity, viability of catfish, spawning area, breeding population	Netica	-experts -literature -models	-experts -literature -models	-sensitivity analysis -scenario analysis
Koivulaso *et al.*, 2005 (Ireland)	Water quality	-assessed impacts of climate change on dissolved organic carbon (DOC) in catchment runoff -included soil moisture, runoff, air temperature, DOC, peat decomposition	Bayesian Network Tools in Java (BNJ)	-experts - simulation models	-models -experts	-scenario analysis
Bacon *et al.*, 2002 (Wales)	Agriculture	-cost/benefit analysis of land management policies -includes land use option and their costs/benefits, social/environmental benefit, farmers attitude and perceptions	Netica	-experts -models	-models -experts	-sensitivity analysis -scenario analysis
Pollino *et al.*, 2007 (Australia)	Water quality	-risk assessment of native fish communities -data included river flow, water quality, temperature, community change, biological potential, diversity, future abundance	Netica	-experts -literature	-experts -data -models	-sensitivity analysis -predictive accuracy (error rates) -information reward -expected value
Smith *et al.*, 2007 (Australia)	Habitat suitability	-mapped mammal suitability -data included soil types, distance to water, land tenure, rainfall, grazing pressure, ground cover, acacia density -GIS intersection analysis used to produce some relationships and populate CPTs -results produced as a habitat suitability map	Netica, ArcGIS 9.1	-experts -literature	-GIS model -experts -data	-predictive accuracy -sensitivity analysis -error matrix method -kappa statistic

Case study	Scope	Summary	Software	Variables and states	Structure and parameter learning	Verification and validation
Borsuk *et al.*, 2004 (North Carolina, USA)	Water quality	-predicted ecosystem response to alternative management strategies -data included nitrogen concentration, river flow, water temperature, oxygen concentration, fish health, carbon production, wind frequency, algal density, shellfish survival	Analytica	-experts -literature	-experts -models -literature	-predictive accuracy -sensitivity analysis
Sadoddin *et al.*, 2005 (Australia)	Water quality	-assessed ecological impacts of salinity management - included temperature, precipitation, evapotranspiration, vegetation management, community attitude, recharge, groundwater level and flow, stream water quality, landscape and habitat condition -GIS used to generate potential land cover under different management scenarios	ICMS	-experts -literature	-models -literature	-scenario analysis -sensitivity analysis
Tattari *et al.*, 2002 (Finland)	Water management	-assessed impacts of buffer zones on water protection and biodiversity -data included soil type, vegetation cover, bird, insect and plant diversity, erosion, water quality, plant height and coverage, slope, fertilisers and pesticides	Fully Connected Belief Networks (FC BeNe)	-experts	-experts	-sensitivity analysis
Kipkemboi *et al.*, 2007 (Kenya)	Environmental impact assessment	-assessed the effects, on ecosystem integrity, of using natural wetlands as seasonal fishponds -data included land use, biodiversity, human diseases, resource us conflicts, ecological and social impact, eutrophication, ponds, species change, hydrology change	Netica	-data -literature	-data -literature	-scenario analysis
Dawsey *et al.*, 2007	Water distribution systems	-dynamic modelling for real time assessment of water drinking systems -data included water pump status, water flow measurements, biological contaminant and industrial demand	No software, conceptual model presented	-data	-simulation models -data	-scenario analysis
Shihab, 2008 (Oman)	Groundwater quality	-dynamic modelling of groundwater pollutants -data included total dissolved solids, electrical conductivity, chemical oxygen demand (COD), nitrate concentration and pH	Hugin	-data -literature	-data -experts	-sensitivity analysis

4.2 Case studies from South Africa

The application of Bayesian Network is still a novel research area in South Africa, or so it seems from the literature search. It could be that there are more applications that have not been published. Due to the scarcity of literature documenting the application of Bayesian Networks in water resources modelling, the discussion is generalised to include environmental modelling. In terms of catchment modelling, the Council for Scientific and Industrial Research (CSIR) is one of the leading organisations actively researching the application of Bayesian Networks. Some of the examples from literature of work done by the CSIR are discussed in Section 4.2.3 (Musango and Peter, 2007).

4.2.1 Marine applications of Bayesian Networks

Curtis *et al.,* 2005 highlight a marine application of Bayesian Networks in predicting sub-surface chlorophyll from satellite data and environmental factors like the depth of ocean floor, the season and region. The software used is BayesiaLab®. The missing/hidden variables are estimated using the EM algorithm. The variables, the states and the structure of the network are automatically learnt from the data. The graphic display capabilities of GIS are used to show the prediction results. The modelling results are verified using the predictive accuracy method.

Grandin *et al.,* 2006 discuss the use of Bayesian Networks and Dynamic Bayesian Networks in predicting sub-surface chlorophyll profiles in a system called the Plankton Prediction System. Their work furthers that done by Curtis *et al.,* 2005. The learning of the structure of the network is done automatically using a hill climb search method and the approximate conditional likelihood scoring algorithm. These are both examples of search and score methods for structure learning as discussed in Section 3.1.3.

Parameter learning is carried out automatically using the maximum likelihood estimation and hidden/missing variables are approximated using the EM algorithm. CPTs are populated using data from satellite images and the learning is done on a pixel by pixel basis (this is explained in detail in Section 3.2). The results of learning and inference are displayed graphically using GIS capabilities. It was stated in the thesis that due to time limitations, the accuracy of the predictions could not be assessed.

4.2.2 Dynamic Bayesian Networks in weather forecasting

McIntosh, 2008 proposes the use of Dynamic Bayesian Networks for combining weather data from multiple Global Circulation Models (GCMs) to produce higher resolution more accurate seasonal forecasts for Southern Africa. There is no documented information on the software used or the modelling methodology. The results of modelling have not yet been published.

de Kock, 2008 documents the use of Dynamic Bayesian Networks in weather forecasting in Southern Africa. The software used was the open source software Genie®. The structure of the DBN was mined from data using simulated annealing. The development of the structure was done iteratively, during the process, different candidate structures was produced. Data were input into the networks and predictions were produced. A scoring function was then used to select the optimum structure. A simple model was developed with three variables, maximum temperature, minimum temperature and rainfall. The missing data were filled in using the EM algorithm. The resulting network was evaluated using the predictive accuracy method.

4.2.3 Bayesian Networks for assessing the impacts of climate change on agriculture

The first example is the one highlighted by Musango and Peter, 2007. It considers the application of Bayesian Networks in assessing the impacts of climate change on agriculture. Hugin® is used to assess the mostly quantitative data. Experts or existing literature sources/models are used to derive the variables and their states. The model results are validated using sensitivity analysis.

In the second example, Peter *et al.,* 2007 assess the impacts of climate change on biofuels production using Hugin®. The study has since advanced and the amount of variables and data has increased considerably. As a result, it was discovered that Hugin® became unstable and the organisation has since developed custom software to counter some of these limitations.

4.3 Discussion

This discussion elaborates on the points summarised in the applications listed thus far. Although the aim was to get examples from different countries in the world, from the international case studies reviewed, it seems that most documented applications for catchment modelling are from Australia. This is a relevant point due to the fact that the climate of Australia is similar to that of South Africa and both countries experience comparable issues in catchment management, for example water quality problems and water scarcity issues. Therefore, some of the modelling procedures from Australian case studies may be applicable in South Africa with minimal adjustment (Dye and Croke, 2003).

The rest of the discussion critiques other aspects including: the software used, the data types, scale issues, handling of missing data, variable selection and discretisation, structure and parameter learning and the use of GIS.

<u>The software</u>

According to the evaluation above, Hugin® and Netica® are the most commonly used software products. Hugin® is one of the most advanced software packages and provides numerous capabilities. It allows for the automatic learning of the structure from data. Dynamic Bayesian Networks can also be created using an object-oriented approach. The major deterrent of Hugin® is its cost. Although it can handle missing data, if there is a significant amount of data missing problems might be experienced (Uusitalo, 2007).

Netica® is also commonly used and is comparatively more affordable than Hugin®. Netica® has most of the capabilities that Hugin® has except its lack of structural learning capabilities. Other software products include Analytica® (a commercial software) and Bayesian Network Tools in Java (BNJ). BNJ® is available free of charge but does not support automatic structure or CPT learning, these all have to be manually specified.

Some of the applications use custom software/approaches for example TwoLe® (Soncini-Sessa *et al.,* 2007), Interactive Component Modelling System (ICMS)® (Ticehurst *et al.,* 2007; Sadoddin *et al.,* 2005) and Fully Connected Belief Networks (FC BeNe)® (Tattari *et al.*, 2002). Murphy, 1998; and Uusitalo, 2007 provide a list of available software packages for Bayesian Networks and their capabilities.

Data and scale issues

Most of the examples involve the integrating of qualitative and quantitative data of different uncertainties in modelling. One of the major challenge mentioned in Chapter 2 is the discrepancy between the scale at which the data are collected and the modelling scale. The data available might not be at the required scale. In other studies, if the data are not available at the right scale and the variable is not important, it is eliminated from the study. If the variable is necessary, expert knowledge or functional equations are used to fill the gaps (Ticehurst *et al.,* 2007). In some cases, based on stated or published assumptions and the modelling purpose, it might be justifiable to interpolate or aggregate the data to the required scale (Borsuk *et al.,* 2004).

Dynamic Bayesian Networks for time-series modelling

Several examples have been highlighted that investigate time-series modelling. The most simplistic approach is the study done by Pollino *et al.*, 2007. They used time as a variable in a static BN variable that is linked to future abundance and diversity of the native fish communities. The future diversity and abundance are predicted over a one year time frame. Therefore, although some temporal aspects of Bayesian Networks were explored, DBN were not fully implemented.

The studies in weather forecasting by McIntosh, 2008 and de Kock, 2008 illustrate the successful use of Dynamic Bayesian Networks for predicting climatic conditions like rainfall and temperature. The only drawback is that these examples show the use of a few parameters, for example the three modelled in de Kock, 2008; therefore there is no indication of the performance of DBN when applied to a multitude of variables as commonly encountered in catchment systems.

Other examples provided include Dawsey *et al.,* 2007, who use Dynamic Bayesian Networks in modelling water distribution systems, although the paper only documents the conceptual model and no discussion is provided on the implementation of the model. The study by Soncini-Sessa *et al.,* 2007 on the evaluation of irrigation policy alternatives in Italy also addresses some temporal aspects of Bayesian Networks. It considers the states of the soil and the biomass at different time-steps. Shihab, 2008 discusses the successful use of DBN in groundwater quality modelling.

Missing data

There are various approaches to handling missing data. If the variable is considered unimportant by the end-user or modellers (this decision can arise after an intensive sensitivity analysis), it can be removed from modelling. If for a variable there are missing records/measurements, results from functional relationships and simulation models or expert knowledge can be used to fill the gaps. Koivusalo *et al.,* 2005 discuss the use of outputs from different simulation models to populate the network. The only foreseeable problem with this approach is the propagation of errors from the simulation model to the Bayesian Network in cases when the simulation model is not properly calibrated (Ames *et al.,* 2003).

In some studies where there is missing data, that is the data available are not adequate to cover all the sates defined for a certain variables, the dataset can be input into the software with the gaps. Most standard software products, for example Hugin®, Netica® and Genie® can handle missing data as they use algorithms like the EM algorithm to estimate CPTs with missing data (Uusitalo, 2007; Pollino *et al.,* 2007).

Variables, structure and parameter learning

In all the studies, a combination of literature sources, models or expert knowledge is used to define the variables and for discretisation of the continuous data. In some studies, this is combined with automatic statistical discretisation methods that are available in most Bayesian Network modelling software packages.

In creating the structure, most studies use expert knowledge. The use of experts is common because methods for automatically creating "optimal" models are still in the development phases. One of the main reasons for this is that these procedures are computationally hard (Uusitalo, 2007). A variation of this is when the modeller first develops the network automatically using existing data or manually from literature sources. The result is presented to domain experts, in an iterative process, to get their input and make modifications before an agreed-upon structure is produced (Pollino *et al.,* 2007).

Similarly for parameter learning, the use of experts is common especially when there is lack of data to represent a particular variable (Baran and Jantunen, 2004). Eliciting probabilities from experts limits the number of variables to consider because as they increase, the probability tables become unmanageable. Baran and Jantunen, 2004 suggest four as the maximum number of variables to use as parents/children in a relationship with one variable.

The use of GIS in Bayesian Networks

Pullar and Springer, 2000 discuss the importance of combining GIS and catchment models. GIS can be used to pre-process data and provides an interactive platform for decision-makers to quickly modify parameters and visualise the results of simulation within a spatial context. The commercial Bayesian Network software packages discussed in this section have no GIS capabilities.

Examples of the use of GIS are provided by Koivulaso *et al.,* 2005; Stassopoulou et *al.,* 1998; Sadoddin *et al.,* 2005. Stassopoulou *et al.,* 1998 and Sadoddin *et al.,* 2005 use GIS for pre-processing data and to populate the CPTs. Both studies use a cell-based approach in which the CPTs of some attributes are estimated from the proportion of cells with values that fall into each variable state. Smith *et al.,* 2007 expands this model further by using GIS intersection techniques and statistics to define the relationships between GIS variables. The results of scenario analysis from the study are presented in GIS format as a habitat suitability map. The GIS software used is ArcGIS version 9.1.

Ames, 2002 documents the development of a system for displaying the results of Bayesian Network modelling and inference in a GIS by combining GeNIe® with the Mapwindow® GIS software. The variables or "nodes" in the Bayesian Network represent different points along the river network.

Model evaluation and verification

In almost all studies, there is some evaluation and verification technique. It is either the use of expert judgement (Ticehurst *et al.,* 2007; Chee *et al.,* 2005), sensitivity analysis or scenario analysis. These are the bare minimum techniques for any model results that have been recommended in literature (Ames, 2002). Sensitivity analysis can be used to estimate the effects of uncertainty in inputs on the model outputs.

The most comprehensive procedures for evaluation and verification are documented by Smith *et al.,* 2007. Smith *et al.,* 2007 discuss the tabulation of an error matrix after comparing the model prediction results to field data. Overall accuracy and kappa statistics are calculated and used to assess the model accuracy.

Some of the main challenges of systematically verifying the BN models are presented by Ticehurst *et al.,* 2007. They stated that in most cases, there is an unavailability of time-series data to verify BN models. Also, due to the fact that the outputs of the model are future impacts of management decisions, data are often not available until the decisions have been implemented. The same issues are discussed by Ames, 2002.

4.4 Advantages and limitations of Bayesian Networks

Bayesian Networks represent beliefs about values as probability distributions; the higher the uncertainty the wider the probability distribution. This is one of the main advantages because the evaluation of uncertainty prevents overconfidence in the outcome. Another advantage is the ability to exhibit good predictive accuracy with little data. With the EM algorithm, the probabilities can be estimated even when there are missing values. Because Bayesian Networks are easily updateable, new information can be included in the model as it is acquired, leading to a reduction in the uncertainty of the results (Uusitalo, 2007).

Bayesian Networks, as highlighted in the different examples provided, have the ability to incorporate multiple qualitative and quantitative data available at different spatial and temporal scales (Ticehurst *et al.*, 2007). There is no need to convert the data to common units of measure. Expert knowledge from various sources and with different uncertainties can also be incorporated into the model (Uusitalo, 2007).

Due to the fact that they are solved analytically, Bayesian Networks provide rapid response during query analysis when the model is updated. This is vital especially when performing scenario analysis and presenting the outcome of these for decision-making. Bayesian Networks do not only go from cause to effect, but analysis can be done from effect to cause in order to perform diagnosis, which assesses the different causes of given scenarios or effects (Uusitalo, 2007).

In terms of limitations, one of the shortfalls of Bayesian Networks is that they are best-suited to dealing with discrete data. This means that continuous data, for example rainfall, that is common in catchments have to be discretised (see Section 3.1.2) before being modelled. When using fewer intervals for discretisation, more dependencies are likely to be discovered whereas more ranges bring the ability to model complex relationships, but this requires more data. A balance can be reached on a trial and error basis but, in reality, the data poses restrictions on the ability to capture complex dynamics (Uusitalo, 2007).

The other challenge involves the elicitation of structure and probabilities from experts. For creating the structure of the network, experts might be unsure of the processes or might fail to agree on a common structure. Although having more variables in the network can improve the accuracy of prediction by ensuring the capture of complex relationships, their inclusion is not useful if the experts have no knowledge on their interactions (Borsuk *et al.*, 2004).

When eliciting probabilities from experts; the challenges include the following: the experts not understanding the concepts and objectives, and the inability to represent their knowledge as probability distributions (Barton *et al.*, 2008; Henriksen *et al.*, 2003). Anderson, 1998 cited in Baran and Jantunen, 2004 provides other pitfalls of the process of eliciting probabilities from experts. The combination of different expert opinions to form probability estimates is also a challenge. These issues of expert elicitation are discussed in literature and there are set guidelines for informing the process (Uusitalo, 2007).

Another limitation of Bayesian Networks is their inability to model feedbacks or loops in the network, which could be beneficial in catchment modelling. Borsuk *et al.*, 2004 suggest the use of Dynamic Bayesian Networks to tackle this shortfall.

As can be inferred from the discussion, the limitations of Bayesian Networks by far outweigh the advantages. Due to continuous research in the area, ways of countering some of these challenges are being developed.

4.5 Conclusions

In this Chapter, case studies on the application of Bayesian Network for modelling were presented. From international case studies, there are numerous examples applied at catchment scale. In South Africa, the applications are generalised to include environmental applications due to the scarcity of published case studies that focus on catchment scale modelling. The only available examples are those based on the work by the CSIR and these were discussed in Section 4.2.3.

The different cases were critiqued in Section 4.3, followed by a discussion on the advantages and limitations of Bayesian Network in Section 4.4. From this discussion, the following issues are drawn:

i) Bayesian Networks have not been applied to catchment water resources assessment in South Africa;
ii) when there is scarcity of adequate data, for example in South Africa, experts have to be relied on for variable selection, discretisation and population of CPTs and this implies that the number of variables and their states have to be simplified to avoid complications;
iii) the application of Dynamic Bayesian Networks to time-series modelling of catchment systems is not fully developed;
iv) the use of GIS in Bayesian Network or Dynamic Bayesian Network modelling has mostly been limited to data pre-processing and display of results. The full integration of mining, inference and display capabilities is not adequately addressed;
v) in areas when there is lack of data, it is advisable to use experts for structure and CPT learning, although there are still problems with their engagement and interpretation of their input; and
vi) the validation and verification of the model is vital and at the minimum, a sensitivity analysis must be performed.

These conclusions drawn from the literature review highlight the issues and relevant techniques to be used for the modelling methodology adopted in this thesis and presented in Chapter 5.

CHAPTER FIVE: RESEARCH METHODOLOGY

5 Introduction

The aim of this research is to develop a Bayesian Network model for catchment water resources assessments. The development of the network structure can be either through manual methods, expert elicitation or automatic methods. The advantages of the different methods were discussed in Chapters Three and Four and an automatic method was selected for the development of the model documented in this research.

This chapter also presents the custom software developed for automatic data discretisation and Bayesian Network structure learning and inference results visualisation. The variables used for water resources assessment and their significance in modelling are discussed. The process of developing the Bayesian Network is outlined as well as the assumptions made during the modelling process.

5.1 Software development

Although it is acknowledged that there are software products available commercially or as open-source packages for Bayesian Network modelling; this research used a custom developed software application for the following reasons:

i)The requirements of this research which intended to explore the use of novel automatic methods for Bayesian Network structure learning using the Hybrid Genetic Algorithm (HGA); which is currently not implemented in the available software packages;

ii)The need to implement Dynamic Bayesian Networks in a simple and user-friendly framework; and

iii)The need to integrate GIS display capabilities for easy result presentation when performing Bayesian Inference.

5.1.1. Development principles

The software was developed by a South Africa company, Complex Adaptive Systems. For this project, the author had a conceptual model of the software, some specifications and a list of the capabilities that were required from the software. Within the company that developed the software where postgraduate students carrying out investigations in the following issues which were pertinent to this research:

i)The creation of novel and more efficient algorithms for automatic structure learning of Bayesian Networks from data (Osunmakinde, 2009); and

ii)The development and implementation of Dynamic Bayesian Network.

The custom software for this project was created to encompass the aforementioned aspects using the standard software development cycle outlined in Figure 5-1.

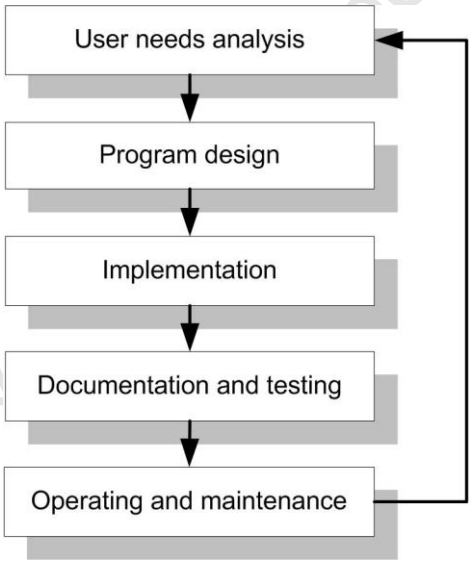


Figure 5- 1: Software design procedures.

Initially, a user needs analysis was carried out by the developers. The problem was defined and an understanding of the tasks the software should perform was gained. The features and the expected functions of the software were outlined. A logical flow process diagram and a user case diagram were created to show the functionalities of the software and how it would be used. Only after an agreement was reached did the processes move to the nest step, which was program design.

The program design and implementation phases involved mainly the developers. In the design stage the developers defined the logical steps (or tasks) required to fulfil the software objectives using flowcharts and pseudocode. The user interface was designed and was presented to the authors for approval. In the implementation phase the tasks were then coded using C Sharp and Microsoft Visual Studio.NET. The software used to store the database was Microsoft SQL.

In the custom developed software, the following are the steps the user takes in modelling:

i. Firstly, import the data from a comma delimited text file into the database;
ii. If there are missing records in the data, use the Expectation Maximisation (EM) algorithm to fill in the missing data in the imported text file;
iii. Create a project and specify the location of the data file and the map (shapefile) of the study area;
iv. Discretise any continuous data in the input file and at the same time automatically mine the structure of the static or the dynamic network;
v. The resulting Bayesian Network is presented in the work area and the results can be exported to image format (see Figure 5- 2);
vi. Inference is performed in the map area and the results are imported onto a map in the query area (see Figure 5- 3); and
vii. The map results can be exported to an image file and the probability results/values (from inference) can be saved to a text file.

5.1.2. Data discretisation in the software

The discretisation algorithm implemented in the software is an equal-width binning method. It is an unsupervised discretisation approach which divides the attribute values into *k* equal sizes. The user specifies *k*, which is the number of intervals or classes required and the equal-width function searches for the maximum and minimum attribute values and these are used to determine data intervals (see Section 3.1.2). As was discussed in Section 3.1.2, discretisation affects the relationships that are mined from data.

In this research, three different discretisation levels are defined for the data and the three resulting Bayesian Networks are created (the algorithm used to create the networks is described in Section 5.2.3). The resulting networks are used to predict certain selected measurable variables in the network. The predicted and measured values are compared and scores are calculated to decide on the goodness of fit or predictive precision of the three networks. The discretisation intervals of the highest scoring network are used in the final modelling and the highest scoring network is selected as the optimal network. This is discussed in more detail in Section 5.2.5.

5.1.3. Automatic structure learning

The learning algorithm implemented in the software was developed by Osunmakinde, 2009 as part of his PhD thesis. The algorithm is a Hybrid Genetic Algorithm (HGA). A genetic algorithm (GA) is a model which generates solutions to optimization problems using the techniques inspired by genetics and evolution theories from biology. The basic elements include the selection of solutions based on their goodness of fit, the reproduction of populations for the crossover of genes and the mutation for the random change of genes. A GA eventually finds better solutions to a problem in a similar manner to the way species evolve to adapt to their environments.

For the HGA, with an input training dataset the learning begins with a population of parents, for example {[x_1], [x_2], [x_3], [x_4]} and an independent network: (x_1), (x_2), (x_3), (x_4) is formed. Every parent in the population is evaluated with a minimum description length (MDL) scoring measure. The aim of MDL is to score the networks and it favours the ones that minimise the sum of the encoding length of the model and the length of the encoding of the data given the model. This means that a network with the most compact description of the data and includes the description of the model itself is favourable (Korb and Nicholson, 2004).

Each parent, for example [x_1] produces offspring during several crossover processes with the rest of the other parents, for example [x_2], [x_3], [x_4]. The process ensures that there are no cycles; a network checker is present to confirm this. If there are 4 parents, for example, there is a 25% probability of producing offspring from each pair of parents.

To avoid backtracking during an evaluation of the structure of an offspring, an inner loop that uses mutual information (MI) checks if an offspring is probabilistically fit as a sub-candidate network. MI is a score that measures the sharing of information between two random variables.

For each survived new offspring, for example $[x_1, x_2]$, the prior probability that $[x_1]$ or $[x_2]$ is a parent of each other is 50% so it is difficult to choose which of the two should be the parent. A new offspring $[x_2, x_1]$, is produced by mutating $[x_1, x_2]$ using the Extended Dependency Analysis (EDA). EDA is a heuristic search technique for finding significant relationships between variables in large datasets.

It is also complex to decide which of the mutated offspring is a parent of the other since there is a 50% probability that either is a parent. The two subcomponents will produce two candidate sub-networks which compete with each other until a winning sub-network emerges. The more fitting candidate sub-network is guided by Shannon's information content in which the EDA selects a candidate sub-network that minimises the score of the whole Bayesian Network. The optimisation process is repeated until the solution space is exhausted and the best Bayesian Network model emerges. The resulting network showing cause and effect relationships mined from the software is displayed in the work area (see Figure 5- 2).

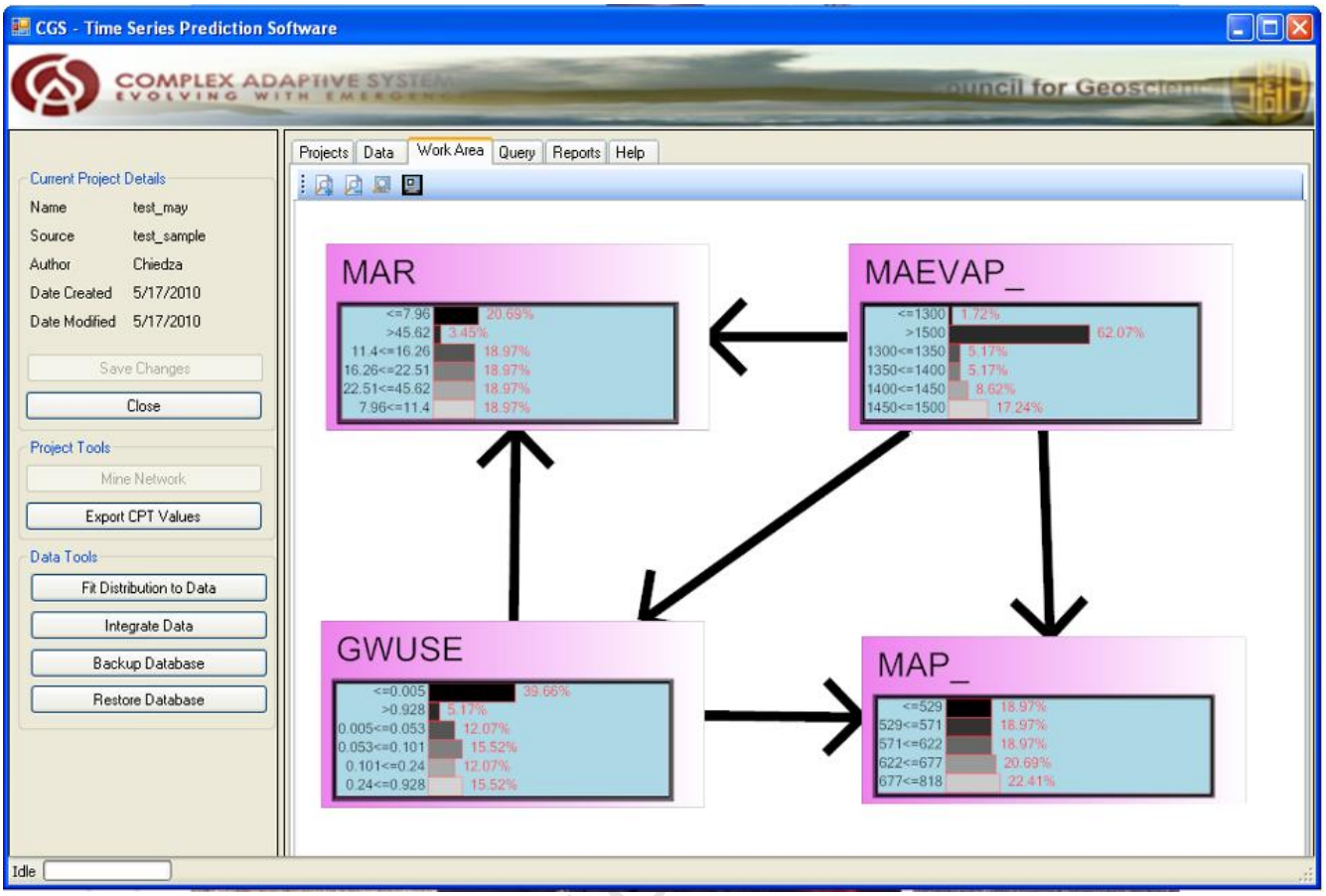


Figure 5- 2: The work area showing a sample structure of a Bayesian Network model.

5.1.4. Inference and presentation of the results

After the model has been created and the Bayesian Network is populated, the software offers inference capabilities. These are used in combination with the map of the study area to display the results and most likely states of the query variables per region. This is done in the "query "part of the software interface. This is useful as an initial analysis of parameter sensitivity analysis.

The user sets the query variable and inputs evidence for the evidence nodes. Inference is performed and the probabilities of the different states of the other variables are provided for each region on the study area map (see Figure 5- 3).

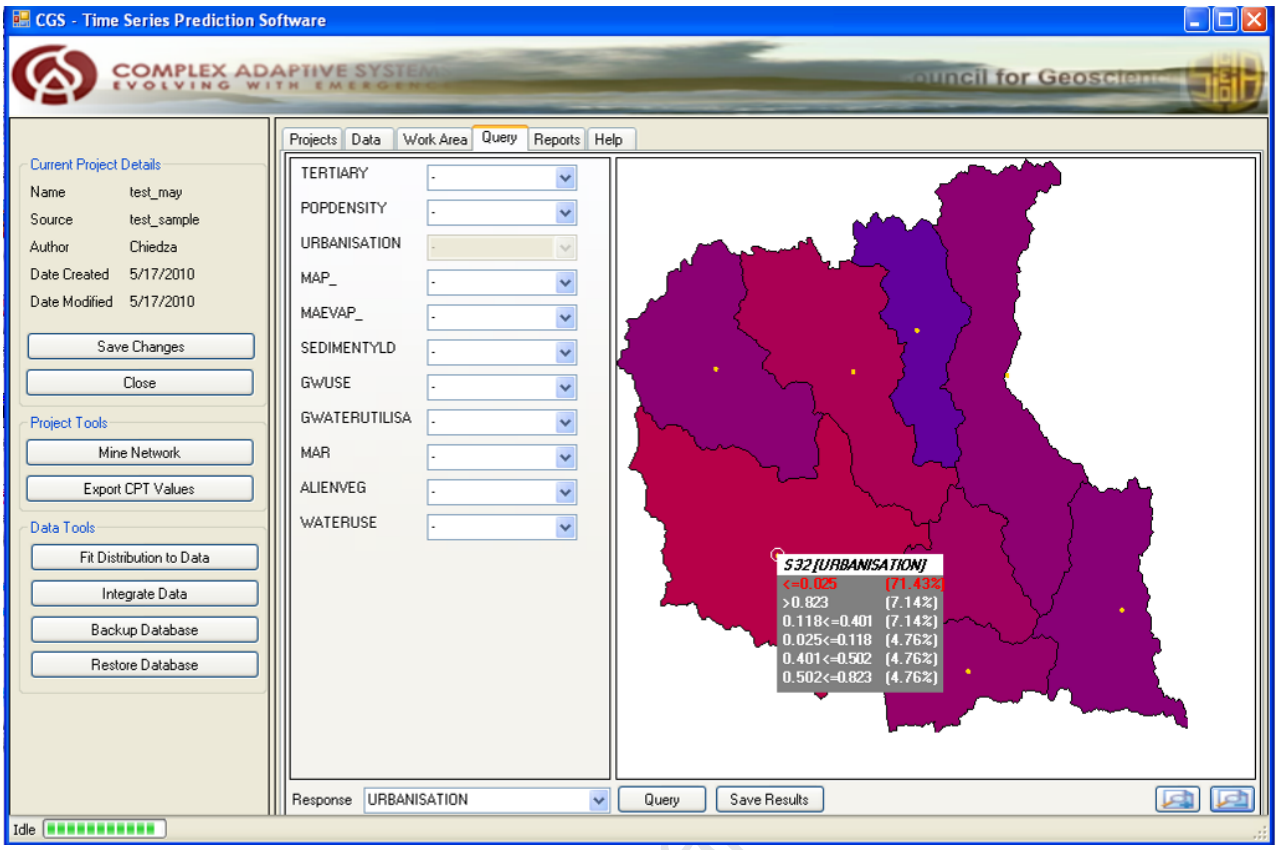


Figure 5- 3: Inference results displayed graphically. The colours range from blue to red, with blue indicating low probability and red high probability. This example shows the Great Kei catchment, the study area described in Section 5.2.1.2).

One of the limitations of the software is the lack of sensitivity analyses capabilities. For this research, the open source Genie® software was used as a supplement mainly for its sensitivity analysis functions. Before a discussion on how implementation was done in the software products presented, it is vital to outline the modelling procedure utilised. This is discussed in the following section.

5.2 Modelling procedure

The modelling procedure used is adapted from the one documented by Jakeman *et al.*, 2006 and shown in Figure 5- 4. In the first step, the purpose of the modelling exercise needs to be clearly outlined.

Examples of some of the modelling purposes pertinent to this research include (Jakeman *et al.*, 2006):

i) acquisition of a better qualitative understanding of the system;

ii) data assessment, which involves discovering the limitations, inconsistencies and gaps in data;

iii) summarising of the data; and

iv) prediction or short-term forecasting of catchment phenomenon (Dawes *et al.*, 2001).

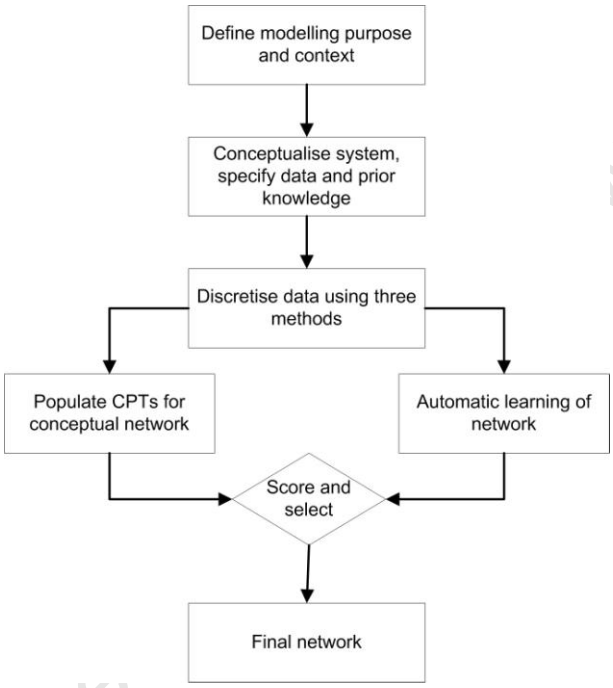


Figure 5- 4: Research modelling procedure.

The modelling context has to be defined at the outset. This involves identifying the issues to be modelled or omitted and the scope of modelling. The spatial and temporal scales of study also have to be selected (Loucks *et al.*, 2005). If the modelling context is not adequately outlined, the following might occur (Jakeman *et al.*, 2006):

i) the scope might be extended beyond the requirements to respond to the pertinent question;

ii) an underestimation of or disregard for the difficulties and limitations of the data and techniques;

iii) oversimplification or over-elaboration of the system components; and

iv) overlooking of existing knowledge and previous experiences.

The second step involves the conceptualisation of the system. The data, prior knowledge and assumptions about the processes are defined in this step. The procedure initially starts qualitatively, with an assessment of the information available for the area, the monitoring systems in place and whether or not there is compatibility with the defined modelling purposes.

Prior knowledge can be acquired from experts, literature review or analysis of the data. The result is a conceptual model of the study area indicating the relationships between the different variables. The conceptual model in this research was created from literature review and work done by Walmsley *et al.*, 2004

In the following step, in this research, the continuous data in are discretised using three methods (this is discussed in more detail in Section 5.5) and three Bayesian Networks are automatically mined from data. The three networks are then scored using statistical metrics and the model with the best score is selected as the model to use and the discretisation classes in the winning model are used. The conditional probability tables (CPTs) of the variables are then populated from the available data.

5.2.1 Modelling scope

The modelling scope pertains to the boundaries of modelling, that is, the spatial and temporal scales used for data analysis. As was discussed in Section 2.4, these affect the predictability and outcome of the results and it is therefore vital to state them explicitly. Their discussion is beneficial especially so that potential users of the results can be aware of their limitations.

5.2.1.1 Modelling scale

The modelling scale includes the temporal and spatial scales. The temporal scale indicates the time when the data used were gathered and the spatial scale indicates the geographical area that was covered in analysis.

Temporal scale

The rainfall, temperature and runoff data used in the model were collected from 1950 to 1999 and modelled at monthly time-steps. For the other datasets; data was only available for 1995, when a water resources assessment was done for the study area at catchment scale. This is the only detail data available to date and the estimates were correlated with the other datasets. The assumption being made is that this represents the status quo as at 1995. The specific aspects of the variables, for example the extent of missing records and the temporal availability are provided in Section 5.3.

Spatial scale

The question of choosing the appropriate modelling and simulating scale presents a challenge. Wiens, 1989 states that this should be informed by the questions one wants to ask about the system. According to Jakeman *et al.,* 2003, the operational scale should be fine enough to capture the needed level of detail variability but not finer than that allowed by the data available and the quality thereof.

Schulze, 2000 concludes that the choice of the appropriate scale to use varies from case to case. Meentemeyer, 1989 discusses the determinants and constraints to the selection of the appropriate scale. The following section presents an overview of the study area and discusses the spatial scales of modelling and analysis. The spatial scale of study used in this research is catchment scale. In terms of ICM and IWRM, in the South African context, the catchment scale is deemed appropriate. According to Water Law, most ICM and IWRM decision are and will in the future be considered at this scale. Appendix A provides a detailed discussion on the subdivision of South Africa into the various catchment levels.

In this research, the data (or cases) used for modelling are available at quaternary scale and at monitoring stations and points located in the quaternary catchments. The results of analyses are then aggregated and presented per tertiary catchment (see Figure 5- 5).

5.2.1.2 Overview of the study area

The study area is catchment S, the Great Kei River primary catchment in the Eastern Cape Province in South Africa. Figure 5- 5 shows the quaternary and tertiary catchments in the area. The quaternary catchments are all not from the original demarcation, they had to be modified to suit the purposes of the research. An equal number of quaternary catchments were required per tertiary catchment. The Great Kei primary catchment covers an area of about 20 485 km^2 and is made up of three sub-areas, the Upper Kei, the Middle Kei and the Lower Kei which are subdivided based on the river systems (Figure 5- 6).

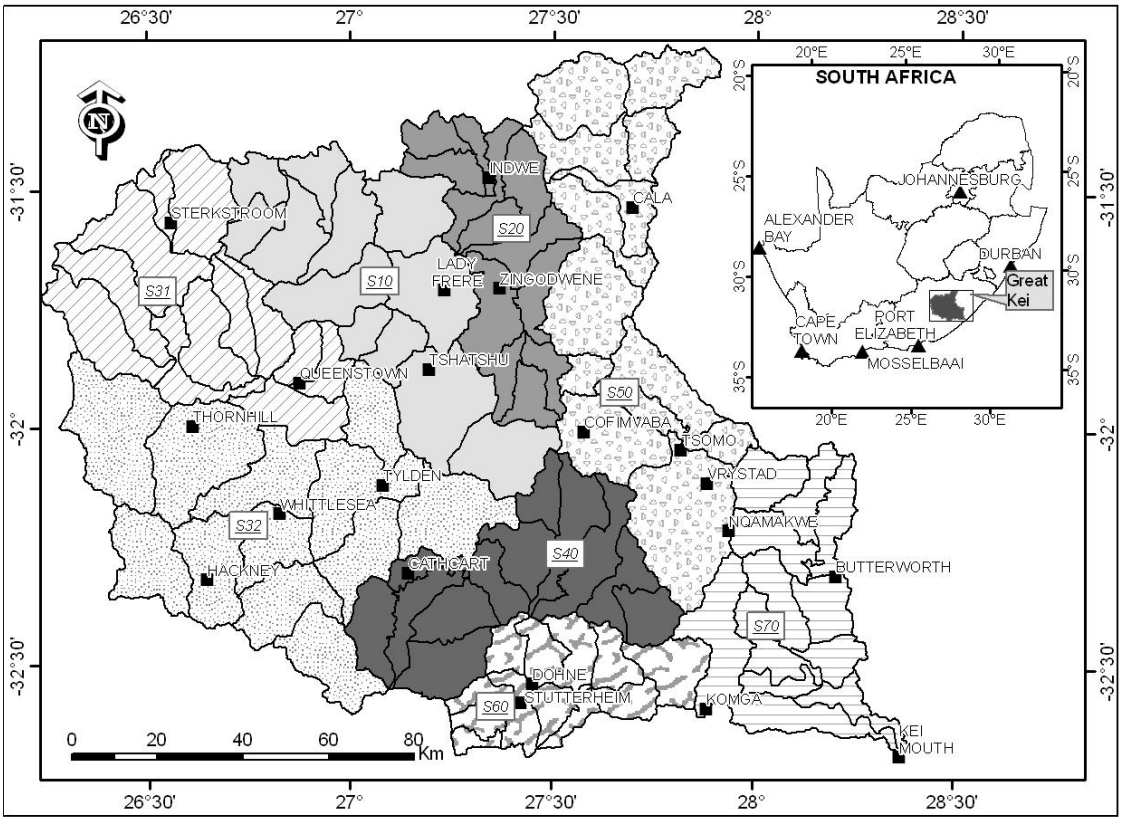


Figure 5- 5: The tertiary catchments of the Great Kei river catchment S and the modified quaternary catchments.

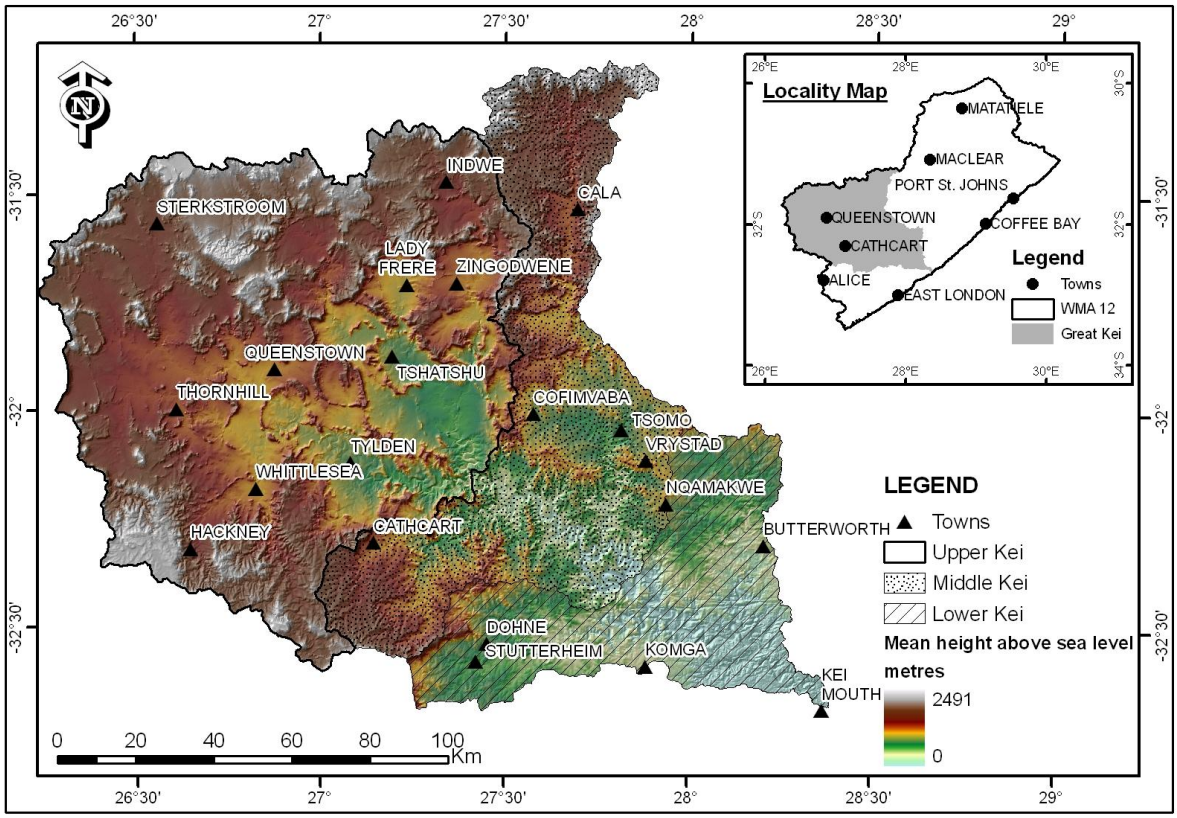


Figure 5- 6: Map of the Great Kei river catchment S and the three sub-areas.

The Upper Kei sub-area is about 11 500 km^2 in extent. The climate in the catchment is harsh, varying from hot and dry in summer (October to April) and very cold in winter (May to September). Maximum temperatures can get to as low as 14°C in winter and as high as 40°C in summer. Frost and snowfalls are experienced in the high lying ground in winter. Rainfall is mainly experienced in the summer months (Figure 5- 7). The mean annual precipitation (MAP) ranges from around 400 mm in the west to 900 mm in the east and in the mountains in the south of the catchment it can reach 1 000 mm (Schulze, 2004). The seasonal variation of rainfall is high with frequent dry spells and droughts (DWAF, 2004).

The Middle Kei sub-area experiences hot and dry summer months to very cold winters. Temperatures in summer can go over 40°C and in winter they can drop to as low as -10°C with snow and frost on high lying ground. The rainfall varies over the catchment with the MAP from 500 mm in the valley to a maximum of 1 200 mm in the mountains. Rain falls predominantly in summer with high intensity thunderstorms and hail (Figure 5- 7).

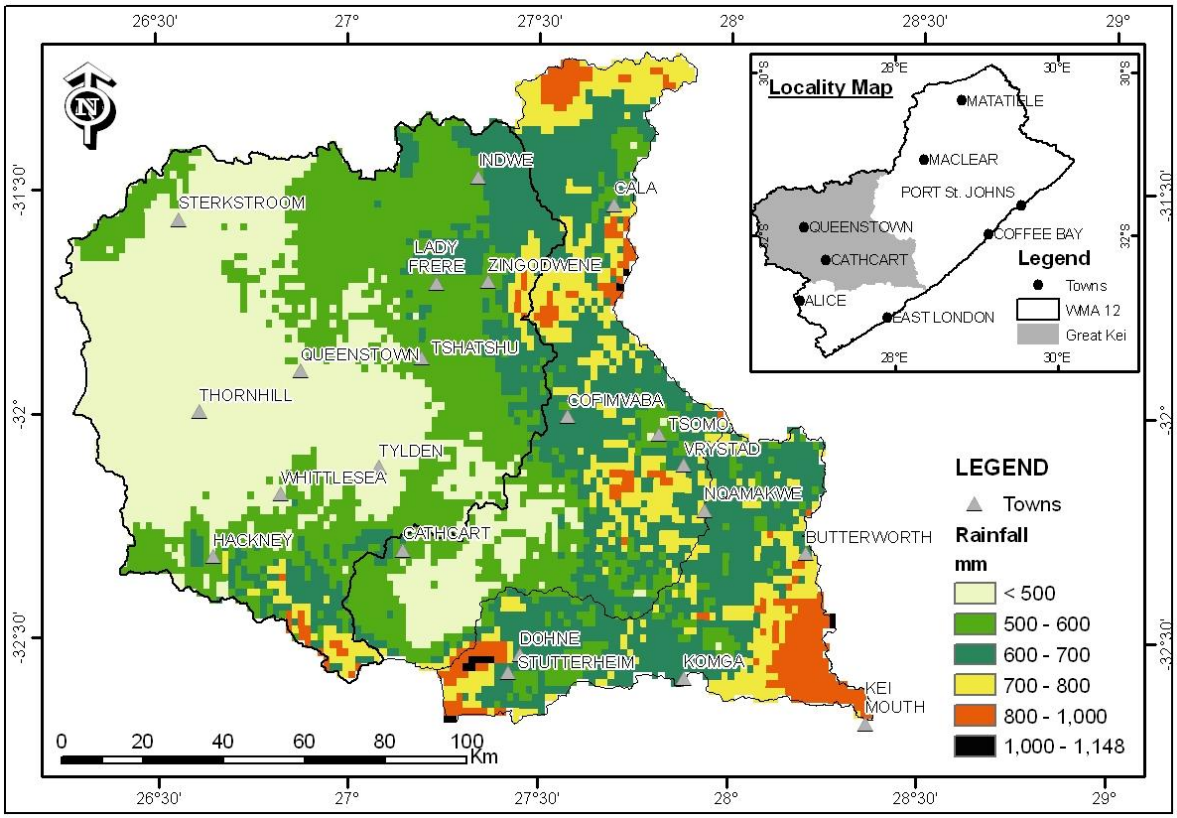


Figure 5- 7: Map showing the mean annual rainfall in the catchment.

The Lower Kei sub-area is a summer rainfall region with June and July being the driest. The temperature ranges are similar to the Middle Kei catchment. The MAP varies from 1 000 mm along the coast to 700 mm inland around Butterworth and 1 200 mm in the Amatola mountains close to the town of Dohne (Figure 5- 7).

The soils and vegetation

The vegetation of the Upper Kei sub-area is predominantly grassveld and savanna in the plateaus with varying levels of invasion by Acacia Karoo (thornveld). In the lower areas of the Black Kei River valley there is valley thicket. There are no significant indigenous forests and there are some invasive alien plants (black and silver wattle) around Queenstown, which do not yet cause a threat. The soils in the area are poorly developed, being shallow and rocky and mostly not suitable for crop production (DWAF, 2004). Figure 5- 8 shows the vegetation classification of the area from Acocks, 1988.

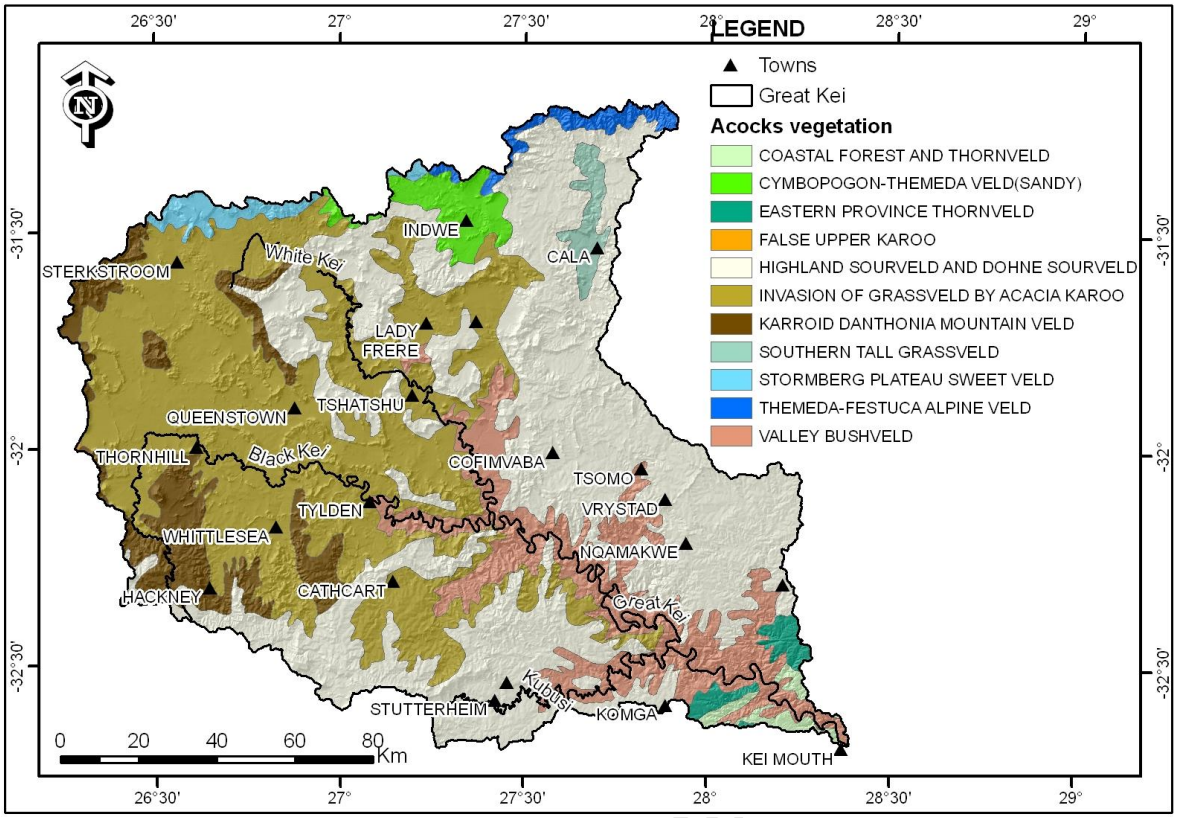


Figure 5- 8: The vegetation of the study area (Acocks, 1988).

The natural vegetation of the Middle Kei sub-area consists mainly of grassland with a mix of savanna in some areas. Valley thicket is most prevalent in the Great Kei Valley and there are some parts that have been invaded by the alien black and silver wattle. The soils are mostly moderate to deep clayey loams with very shallow and rocky soils. The soils are not suitable for crop production. Alluvial soils are in the river valleys (DWAF, 2004) (Figure 5-8).

The plateau area in the Lower Kei sub-area is covered by grassland with large areas of valley thicket in the Great Kei valley. There is considerable commercial forestry near the Kubusi River. Alien black and silver wattle trees invade the sub-area, largely around the Nqamakwe area (Figure 5- 8). The soils are generally shallow and low in fertility.

Land use and settlements

Land use patterns are influenced by the fact that the area was, until 1994, divided into the former Ciskei, Transkei and Republic of South Africa (RSA) (Figure 5- 9). The Ciskei and Transkei areas were for black people under the apartheid system and the RSA was for whiter people.

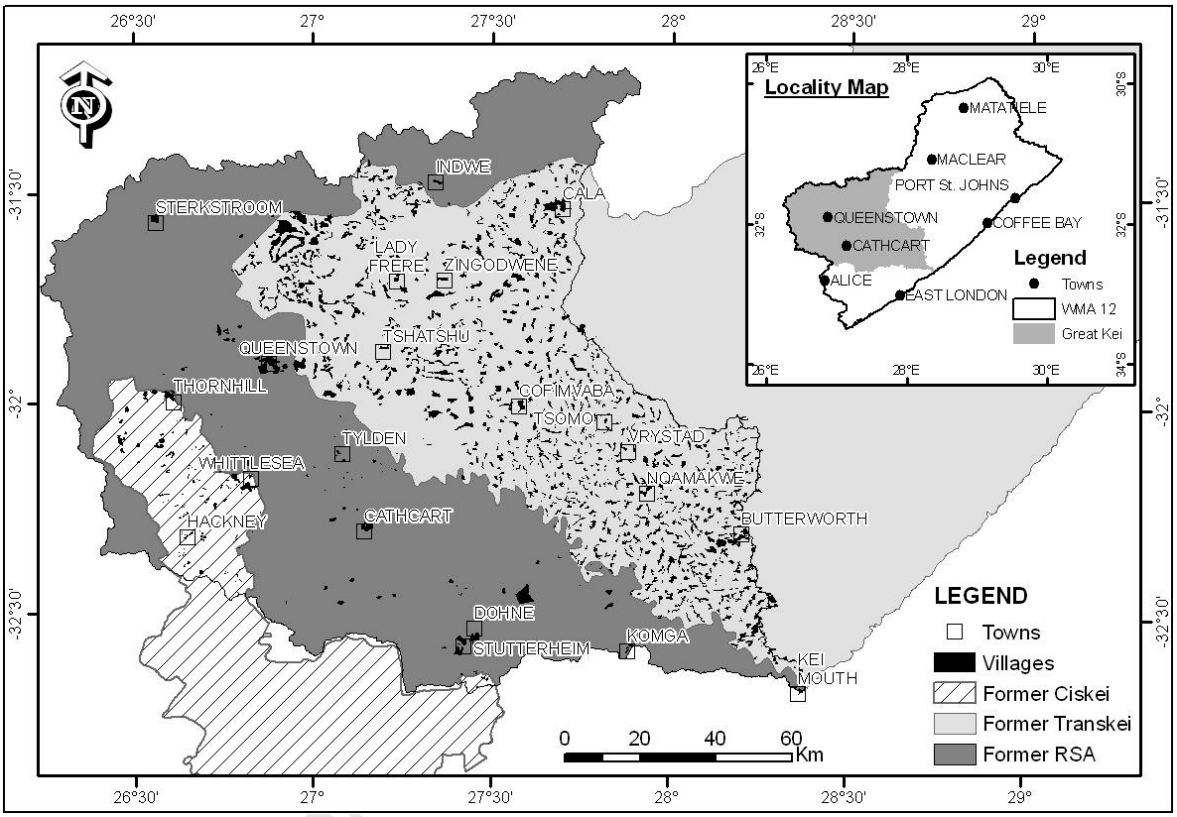


Figure 5- 9: The division of the study area into Transkei, Ciskei and former South Africa, before 1994.

The Upper, Middle and Lower Kei sub-areas are characterised by dispersed rural village settlements and communal subsistence farming and grazing in the former Ciskei and Transkei areas (DWAF, 2004). Parts that formerly belonged to the white areas comprise mostly privately owned commercial stock and game farms. In the communal lands, severe degradation and erosion has occurred and this can be attributed to poor land practices.

Parts mainly affected by degradation include the upper areas of the White Kei River, the Upper Black Kei River and the Tsomo River (Figure 5- 9). This has resulted in the increased turbidity of rivers and sedimentation of dams especially in the Upper Kei area and the Tsomo River.

Demography

In the Upper Kei sub-area, the former Transkei and Ciskei areas are largely rural (about 70% of the population) with villages. These villages have population densities of 55 persons per km^2 as opposed to 4 persons per km^2 for the area with largely privately owned farms (Figure 5- 10). Queenstown is the major town where the services and development are concentrated. The second largest urban area is Whittlesea and the other urban areas are Sterkstroom, Indwe and Lady Frere. Because of the lack of employment opportunities, outward migration is expected from the Upper Keu area and this means that the population will no grow substantially in future.

The Middle Kei sub-area is mostly rural with many small villages and three small towns, Cala, Cofimvaba and Tsomo. The area with largely privately owned farms is sparsely populated. The population is not expected to grow significantly in the future due to the scarcity of employment opportunities (employments rates are around 70%).

The Lower Kei sub-area is mostly rural with two main towns, Butterworth and Nqamakwe. The majority of the people in Butterworth reside in informal settlements around the town. This has implications on water resources use, demand and management. The Stutterheim and Komga towns are centres servicing the privately owned farms in the area (Figure 5-10).

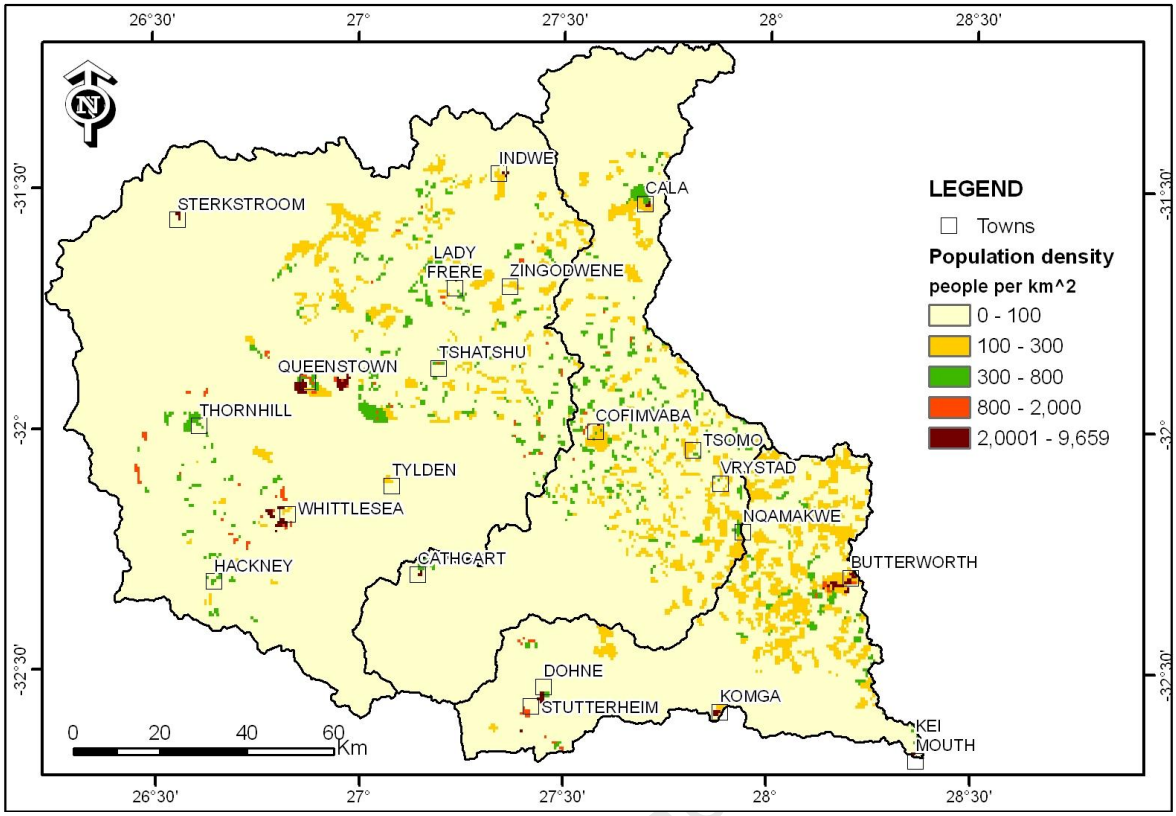


Figure 5- 10: Map showing the population density for the study area, data were obtained from the WR2005 study (Middleton and Bailey 2008).

Surface and groundwater systems

The main rivers of the Upper Kei sub-area are the White Kei River, the Indwe River, the Klaas Smits and Heuningklip Rivers and the Black Kei and Klipplaat Rivers. The main dams are Xonxa, Macubeni, Doring River, Lubisisi, Bongolo, Glenbrock, Mitford, Tentergate, Limietskloof, Thrift, Waterdown, Bushmanskrantz, Oxkraal and Shiloh (Figure 5- 11). There is a considerable amount of bulk raw water supply to cater for the urban domestic and irrigation requirements. Of the estimated 17 000 ha of land irrigated, 8 600 ha have assured water supply from dams in formal irrigation schemes. The rest of the irrigated areas (Figure 5- 11) rely on run-of-river irrigation in the Klipplaat, Heuningklip and Klass Smits or groundwater extracted from well points. The following are the irrigation schemes in the area (DWAF, 2004):

i) Klipplaat government from Waterdown Dam (1 905 ha);

ii) Oxkraal from Oxkraal Dam (541 ha); and

iii)Nthabethemba for Xonxa, Qamata and Lubisi Dams (5 893 ha).

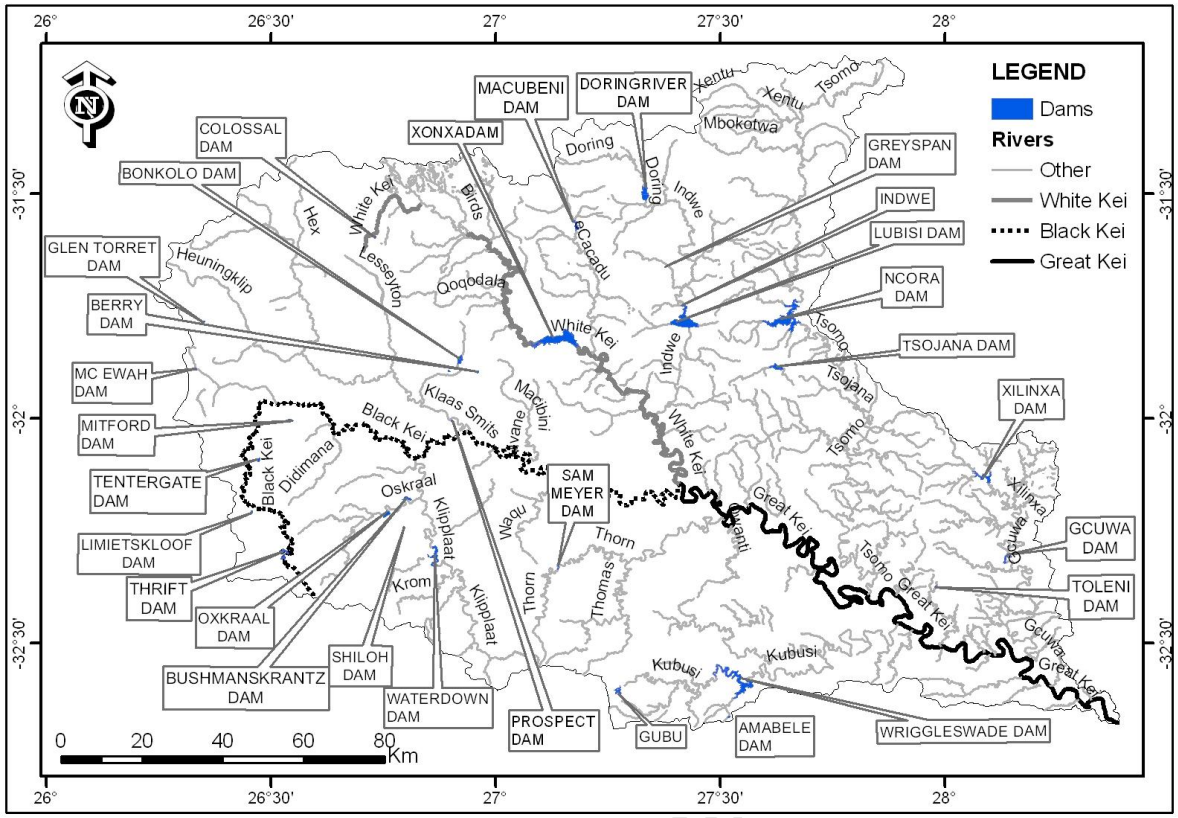


Figure 5- 11: Rivers and dams in the study area.

In the Middle Kei sub-area, there are three major dams, the Ncora, Tsojana and Sam Meyer Dams. The Ncora Dam was constructed mainly for hydropower generation and to transfer water (estimated at about 105 million m^3/annum) to an adjacent catchment to the east also for power generation. The Ncora irrigation scheme also runs off this dam (20 million m^3/annum) (DWAF, 2004). The Tsojana Dam is for domestic supply and the Sam Meyer Dam is for both domestic and recreational use. Irrigation from Tsojana Dam is based on run-off river flows.

The Lower Kei sub-area comprises the lower reaches of the Great Kei River and its two main tributaries, the Gcuwa River and the Kubusi River tributaries (Figure 5- 11). The main dams are the Gubu and Wriggleswade dams on the Kubusi River and the Xilinxa and Gcuwa dams on the Gcuwa River. These dams are owned by DWA and are for domestic water supply and interbasin transfer to the catchment area south of the study area. The Xilinxa Dam also caters for industrial requirements (DWAF, 2004). Figure 5- 12 shows the irrigated areas and the crops grown in each area with their proportions.

Table 5- 1: Water resources supply for the different sub-areas (DWAF, 2004)

Sub-area	Town	Surface water supply	Groundwater supply	Comments
Upper Kei	Queenstown, Whittlesea	Bongolo dam; pipeline from Waterdown dam	—	irrigation around Klaas Smits and Heuningklip rivers
	Lady Frere	Macubeni dam	—	—
	Indwe	Doring River dam	—	—
	Sterkstroom, Ilinge and Hewu	—	groundwater scheme	—
Middle Kei	Cofimvaba	pipeline from Tsojana	—	dam also supplies surrounding rural areas
	Cathcart	Sam Meyer dam	—	—
	Cala	weir on Tsomo River	Zindhlwane stream, fed by springs	—
Lower Kei	Komga	pump station on Kei River	boreholes; springs	groundwater sources are high in nitrates and are mixed with treated surface water before use
	Butterworth	Gcuwa weir (from the Xilinxa Dam)	—	Xilinxa Dam also supplies surrounding rural areas
	Nqamakwe	Toleni dam	boreholes	—
	Kai Mouth	coastal streams	—	—

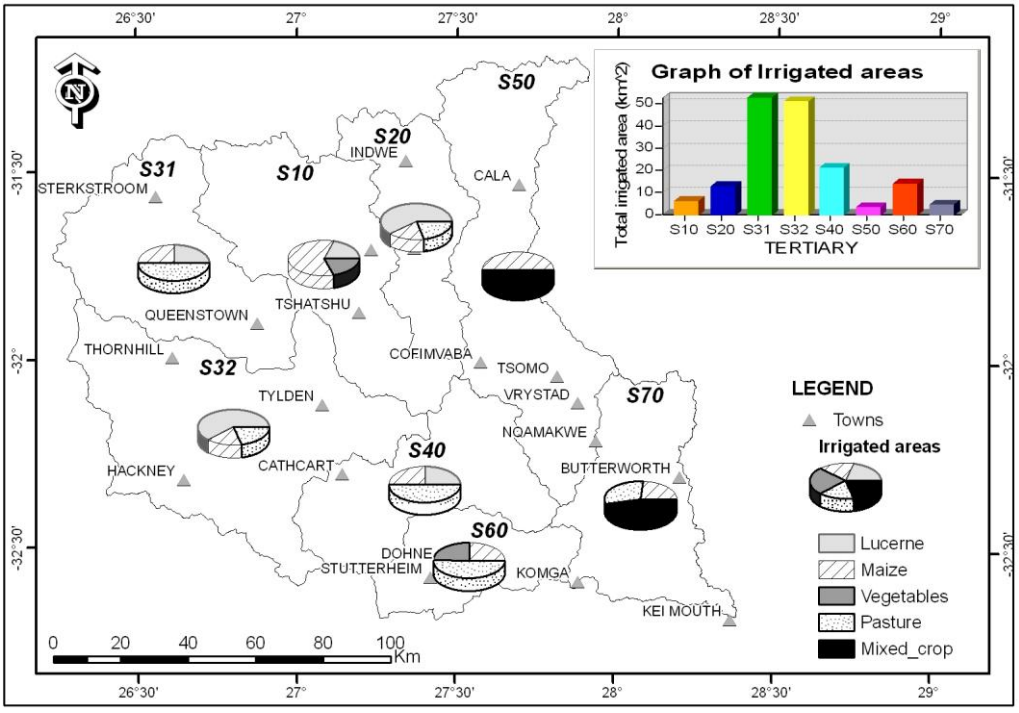
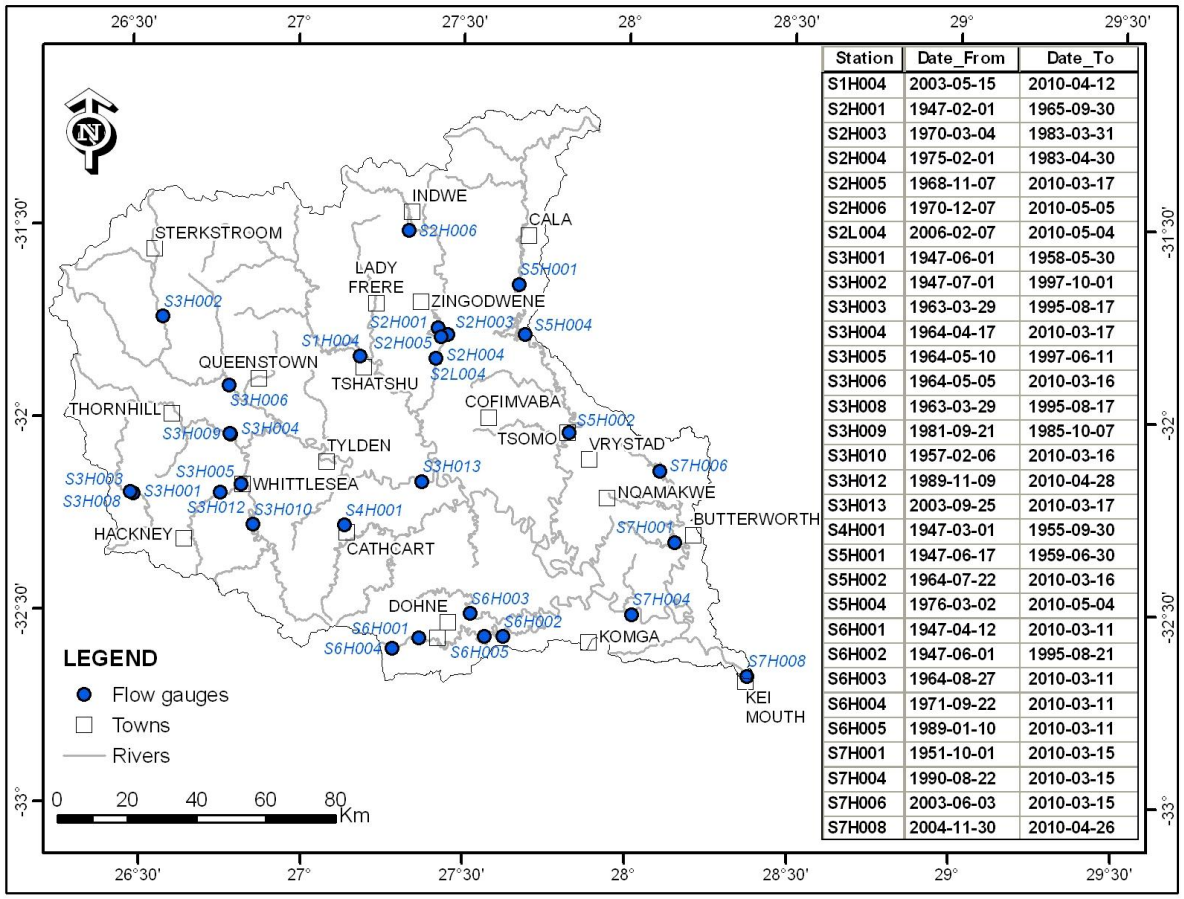


Figure 5- 12: Irrigated areas in the study area and the crops cultivated.

There are high sedimentation rates in the Upper Kei catchment and Ncora dam and these reduce the yield of dams. The determination of accurate values of dam yield is challenging due to the unavailability of accurate flow gauging data. This reduces the degree of confidence in the hydrological calculations. The situation is more likely to change in the future is some regions when new flow gauges are installed. Figure 5- 13 shows the status of flow gauges in the study area.



Station	Date_From	Date_To
S1H004	2003-05-15	2010-04-12
S2H001	1947-02-01	1965-09-30
S2H003	1970-03-04	1983-03-31
S2H004	1975-02-01	1983-04-30
S2H005	1968-11-07	2010-03-17
S2H006	1970-12-07	2010-05-05
S2L004	2006-02-07	2010-05-04
S3H001	1947-06-01	1958-05-30
S3H002	1947-07-01	1997-10-01
S3H003	1963-03-29	1995-08-17
S3H004	1964-04-17	2010-03-17
S3H005	1964-05-10	1997-06-11
S3H006	1964-05-05	2010-03-16
S3H008	1963-03-29	1995-08-17
S3H009	1981-09-21	1985-10-07
S3H010	1957-02-06	2010-03-16
S3H012	1989-11-09	2010-04-28
S3H013	2003-09-25	2010-03-17
S4H001	1947-03-01	1955-09-30
S5H001	1947-06-17	1959-06-30
S5H002	1964-07-22	2010-03-16
S5H004	1976-03-02	2010-05-04
S6H001	1947-04-12	2010-03-11
S6H002	1947-06-01	1995-08-21
S6H003	1964-08-27	2010-03-11
S6H004	1971-09-22	2010-03-11
S6H005	1989-01-10	2010-03-11
S7H001	1951-10-01	2010-03-15
S7H004	1990-08-22	2010-03-15
S7H006	2003-06-03	2010-03-15
S7H008	2004-11-30	2010-04-26

Figure 5- 13: The flow gauges in the study area with a table insert to show the duration of monitoring for each station. 'Date_From' is the beginning of the monitoring at that point and 'Date_To' indicates when monitoring was stopped.

Flow gauging weirs are used to measure runoff data and monitoring is done at flow-gauging stations and on major reservoirs. The data are provided by DWA on their website as daily average rates in units of cm^3/second or monthly values in million m^3 (DWA, 2010). Quality codes are assigned to the data to indicate the confidence levels in the values and these show whether the data are good, observed, edited, above rating, estimated, missing or unreliable.

The flow gauge table shown in Figure 5- 13 shows that approximately 61% of the gauges are not currently being monitored. The earliest record of flow was in 1947 and the station S6H001 around Dohne has the longest record of monitoring data, about 63 years.

Groundwater has been widely used in the rural areas due to the low cost as compared to the cost associated with the supply of surface water. Groundwater supplies are of low quantity due to poor siting techniques. Due to the lack of proper operational and management skills, groundwater supplies has been perceived as being unreliable and has therefore not been developed to full capacity. The wide occurrence of perennial surface streams also contributes towards less exploration of groundwater. Overall, groundwater resources are underutilised for most of the study area except for the Klaas Smits area, to the north of Thornhill town where its use is moderate (DWAF, 2004).

Alluvial groundwater extraction is done mainly for irrigation along the Klaas Smits and Heuningklip Rivers. The potential for large scale groundwater use is unknown due to unavailability of adequate knowledge on the groundwater quantity and quality. This lack of knowledge can be attributed to the lack of proper monitoring.

Groundwater occurrence and quality in the area is influenced by the geology (Figure 5- 14). The geology of the Great Kei is characterised by horizontal to very gently dipping rocks of the Karoo Supergroup (shale, mudstones and sandstones) with dolerite intrusions in the form of dykes, sills and ring structures. The sandstone–mudstone ratio and the frequency of alternation is a determining factor for the matrix transmissivity and regional to local water flow. The ring structures are conducive to the development of fractured aquifers that created secondary porosity and store large amounts of water. The dolerite rings have controlled settlement patterns as people reside close to the base of the rings to have access to water from the springs (Chevallier *et al.*, 2001).

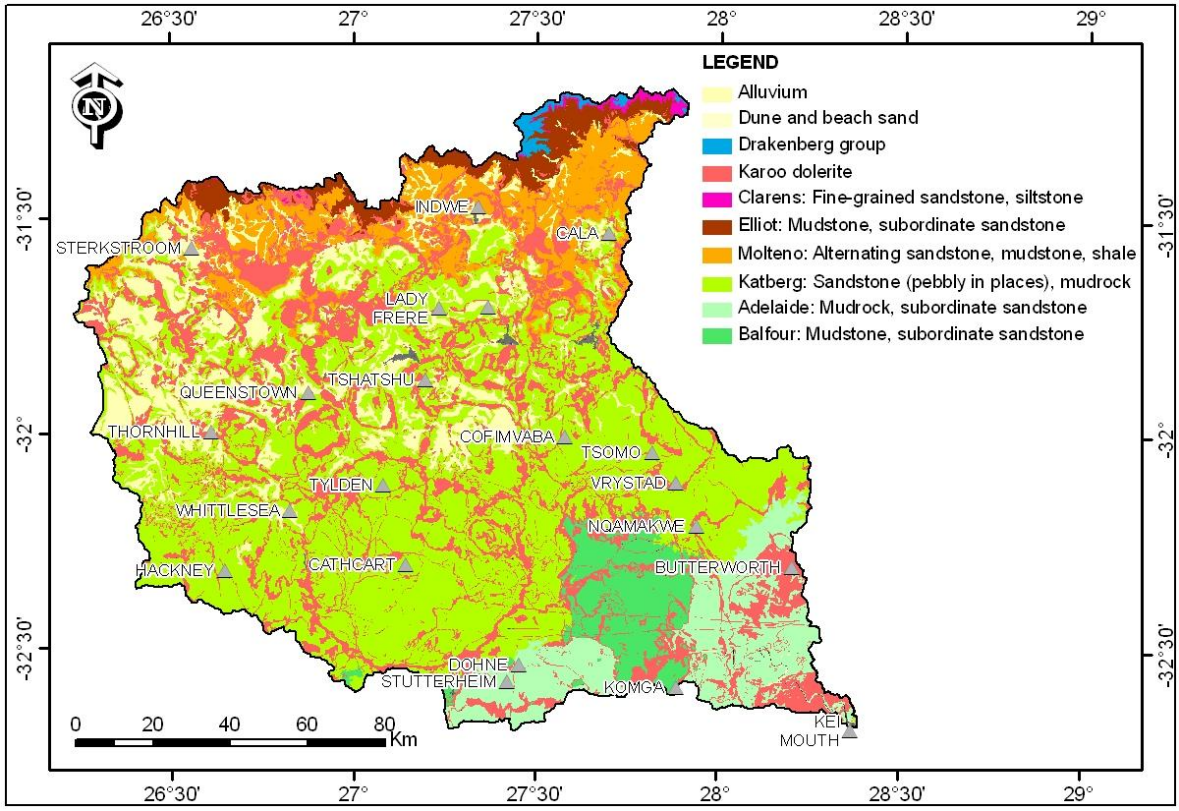


Figure 5- 14: The geology of the study area.

The past section outlined the spatial and temporal modelling scopes of the study area. The following step in the modelling procedure is the conceptualisation of the system and this is discussed in the following section.

5.2.2 Conceptualisation of the system

A generalised conceptual model of the developed Bayesian Network is shown in Figure 5-15. The development of the model is based on the following methods:

i) Expert knowledge (Walmsley *et al.*, 2004);

ii) Data analysis; and

iii) Information from literature review.

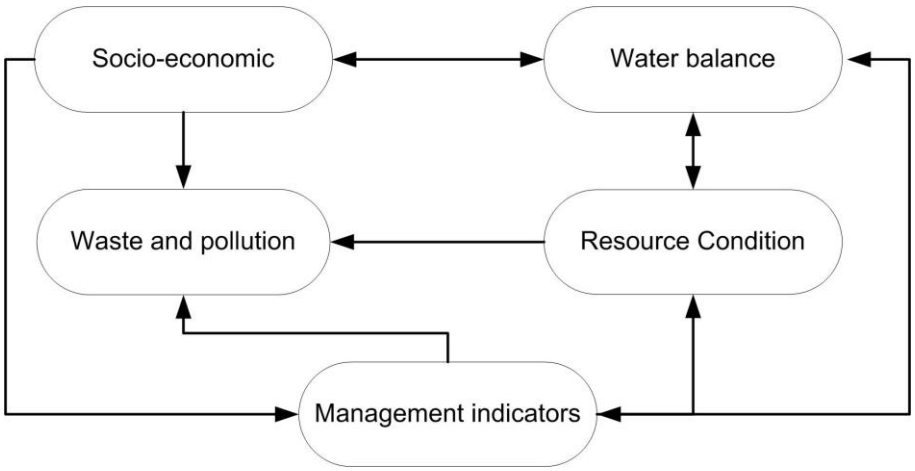


Figure 5- 15: A conceptual diagram showing the main components of the catchment water resources assessment model.

The conceptual model highlights the main aspects in water resources assessment. Walmsley *et al.,* 2004 carried out a study which listed the most important indicators for catchment sustainability analysis. The list of indicators (from which the conceptual model was developed) resulted from extensive participation and consultation with the relevant stakeholders in catchment management in South Africa. More details on the process is provided in Section 5.2.3.

Figure 5- 15 shows the possible linkages between the different components. Socio-economic factors like population and water use affect the availability of water (water balance) and the quality of surface and groundwater (waste and pollution). The pollution of water systems affects the condition of the ecosystem which can also in turn affect the water balance. Management indicators like the technical capacity and the institutional and policy arrangements in place in the catchment affect the condition of the ecosystem and the quality and quantity of the water. The next section discusses the variables considered under each component.

5.2.3 Variable definition

The variables were defined based mainly on the research done by Walmsley *et al.,* 2004. The variables are grouped according to the classes provided in the conceptual model discussed in Section 5.1. The indicators presented by Walmsley *et al.,* 2004 are similar to those used in other organisations and countries as shown in Table 5- 2.

Table 5- 2: Examples from literature of the indicators considered under the different aspects of water resources assessment

Reference	Waste and pollution	Water balance	Socio-economic	Resource condition	Management indicators
Africa state of environment reporting (UNECA, 2001)	-solid and hazardous waste -water quality -atmosphere and climate	-water resources -water demand and use	-population -urbanisation -poverty, human health	-resource degradation -natural disasters -biodiversity	-ongoing sustainable development initiatives in Africa
South Africa state of environment reporting (DEAT)	-air quality and population -surface and groundwater quality -energy use	-water stress -available water -dams' capacity and levels	-population -human health -human settlements	-ecosystem stress -land degradation -invasion of alien species	-budget for the environment - environmental governance
Catchment condition assessment (Australia) Walker *et al.,* 2004	-dryland salinity risk -suspended sediment load	-available water -water use/demand		-soil degradation -hillside erosion ratio -native vegetation and protected areas	-level of institutional capacity
Canada water sustainability index (Government of Canada, 2010)	-level of waste treatment -water quality index	-water resources' availability -water demand and use	-population -human health -access to water services	-ecosystem stress -aquatic life protection -fish species monitoring	-financial capacity of community -training for water and waste treatment operators
Catchment sustainability assessment indicators South Africa (Walmsley *et al.*, 2004)	-solid waste generated -surface and groundwater quality	-water available -water demand and use	-population density urbanisation -land use -statistics on access to water and sanitation -	-natural vs. alien vegetation -% of wetland area -status of riparian vegetation -habitat and fish integrity -river health	-water use efficiency -state of community satisfaction -level of institutional development in the area

Reference	Waste and pollution	Water balance	Socio-economic	Resource condition	Management indicators
Water Poverty Index (Dlamini, 2005)	-surface and groundwater quality	-water available -water use	-% of poor households with -access to water services	-% degraded land	-human and financial capacity of community

All the indicators listed in Table 5-2 were selected after consultation with the relevant stakeholders and intensive literature review. The indicators and variables outlined in Walmsley *et al.*, 2004 were developed by Walmsley, 2003 as part of a PhD research.

In order to prioritise these variables, the relevant stakeholders in the water sector and involved in catchment studies were consulted. The groups of officials included in the prioritisation and testing of the indicators are shown in Figure 5- 16.

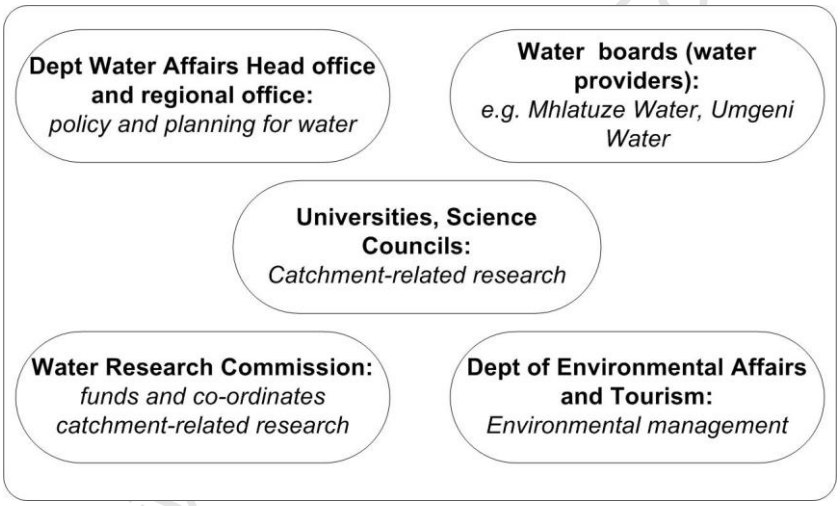


Figure 5- 16 : Stakeholders who participated in the selection of catchment sustainability indicators (Walmsley *et al.*, 2004).

The following sections discuss the variables and the data required to represent the variables by considering these issues:

i) justification for inclusion of the variable, that is, its importance in terms of water resources assessment;
ii) availability of data, the source, the spatial and temporal data coverage and scale;
iii) frequency of data collection and update;
iv) accuracy of data and sources of uncertainty; and
v) the definition of the different states used to represent the variable.

The variables are grouped and discussed according to the following categories; water balance, waste and pollution, socio-economic condition, resource condition and management indicators.

5.2.3.1 Water balance indicators

Water balance indicators refer to the groundwater and surface water availability, use and demand. The variables considered include total available water resources, rainfall, temperature, runoff, groundwater recharge, groundwater exploitation potential, groundwater use and water demand. These variables are discussed further below.

Total available water resources

Total available water per capita is the amount of available surface and groundwater resources per annum. In most catchments, water is supplemented by imports from other catchments or countries via pipeline and these amounts have to be incorporated. The data used were the result of the 1995 study by DWA. The results were produced by considering the mean annual rainfall, the natural runoff, evaporation losses, looses to the sea, return flows, surface water-groundwater interactions and water abstraction (DWAF, 2002b).

The outputs of the 1995 study were yearly figure for each quaternary catchment in units of million m^3 per annum. This amount was considered constant throughout the modelling period for each quaternary catchment.

Rainfall

Rainfall is the natural supply of water to the catchment and is stored in lakes, dams, reservoirs, wetlands, estuaries and groundwater. The main sources of daily rainfall readings in South Africa are non-recording rain gauges, which provide point rainfall data and are cheap and reliable (Seed *et al.,* 1990 cited in Lynch, 2004). Major sources of rainfall data in South Africa include the South African Weather Service (SAWS), the Agricultural Research Council (ARC) and the Department of Water Affairs (DWA) (Lynch, 2004). Figure 5- 17 shows the rainfall stations used in this study.

It is not possible to obtain accurate rainfall readings at a station due to systematic and random errors. Deficiencies in values are estimated at ±10- 20% (Kroese *et al.,* 2006). Other sources of uncertainty in rainfall data include (Schulze, 1994; Barca *et al.,* 2005):

i) insufficient number of gauges and non-representative site locations;
ii) instrument inaccuracy caused by the gauge type, height, windshield, exposure and wind speed;
iii) rain drop diameter, intensity and duration;
iv) instrument failure, which means the absence of readings for that time or the incorrect reading of the time and date at which the recording is made.

Some quality control measures have to be performed to reduce these errors. This means that the collected rainfall data that is obtainable from the different recording stations cannot be used directly without corrections. Lynch, 2004's study provides a complete corrected quality controlled set of rainfall data from the years 1880 to 2000 for the rainfall stations in Southern Africa. This data can be obtained as daily, monthly or yearly values in units of mm. The monthly time-series data for stations in the study area from the year 1950 to 1999 were used in this research (see Figure 5- 17). For quaternary catchments with no rainfall station, the values from the nearest station in the neighbouring catchment were assigned.

The data were discretised into the states shown in Table 5- 3. The classes were based on Mzezewa *et al.,* 2010 and according to the characterisation, high rainfall monthly values are greater than >100 mm. The rest of the classes were based on percentiles from data were the 20^{th}, 40^{th}, 60^{th} and 80^{th} percentile are 8.5; 25.4; 48.2 and 83.4 mm respectively.

Table 5- 3: The discretisation states for rainfall

Parameter	States	Description
Rainfall (mm)	Class 1	<10
	Class 2	10-25
	Class 3	25-50
	Class 4	50-100
	Class 5	100-150
	Class 6	>150

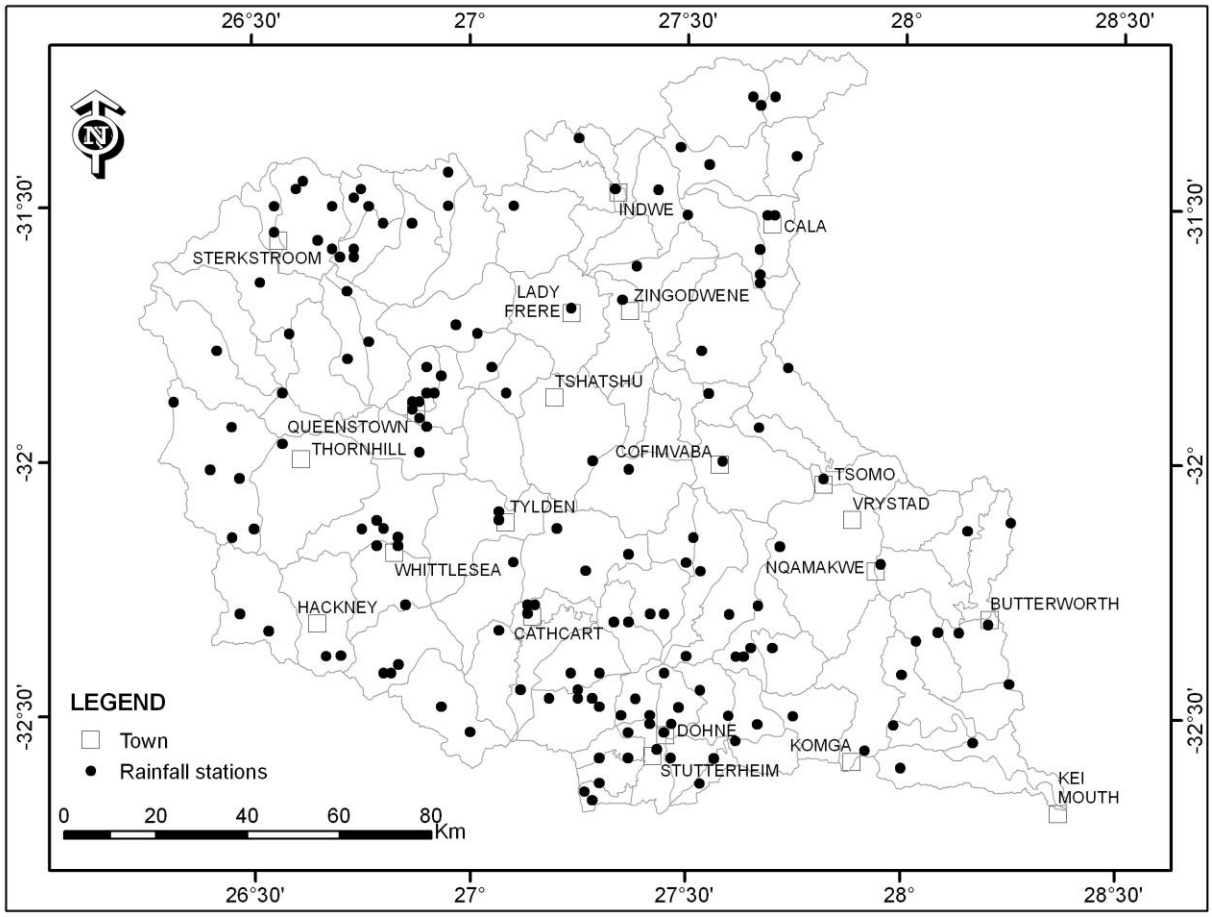


Figure 5- 17 : Rainfall stations used in the study.

Temperature

Temperature refers to air temperature and daily values are measured at various meteorological stations in South Africa in units of degrees Celsius (Schulze *et al.*, 2003). The main problems with the temperature data are that in some instances the records are missing due to instrumental failure and in other cases erroneous reading are obtained. This research uses the quality controlled dataset provided by Schulze *et al.*, 2004. This dataset used recorded temperature values at monitoring stations and filled in the gaps and checked for errors by considering the effects of altitude and also through logical checks The maximum temperature is used in this research as it provides the highest response when used in modelling and for predicting runoff. The temperature stations used are shown in Figure 5- 18.

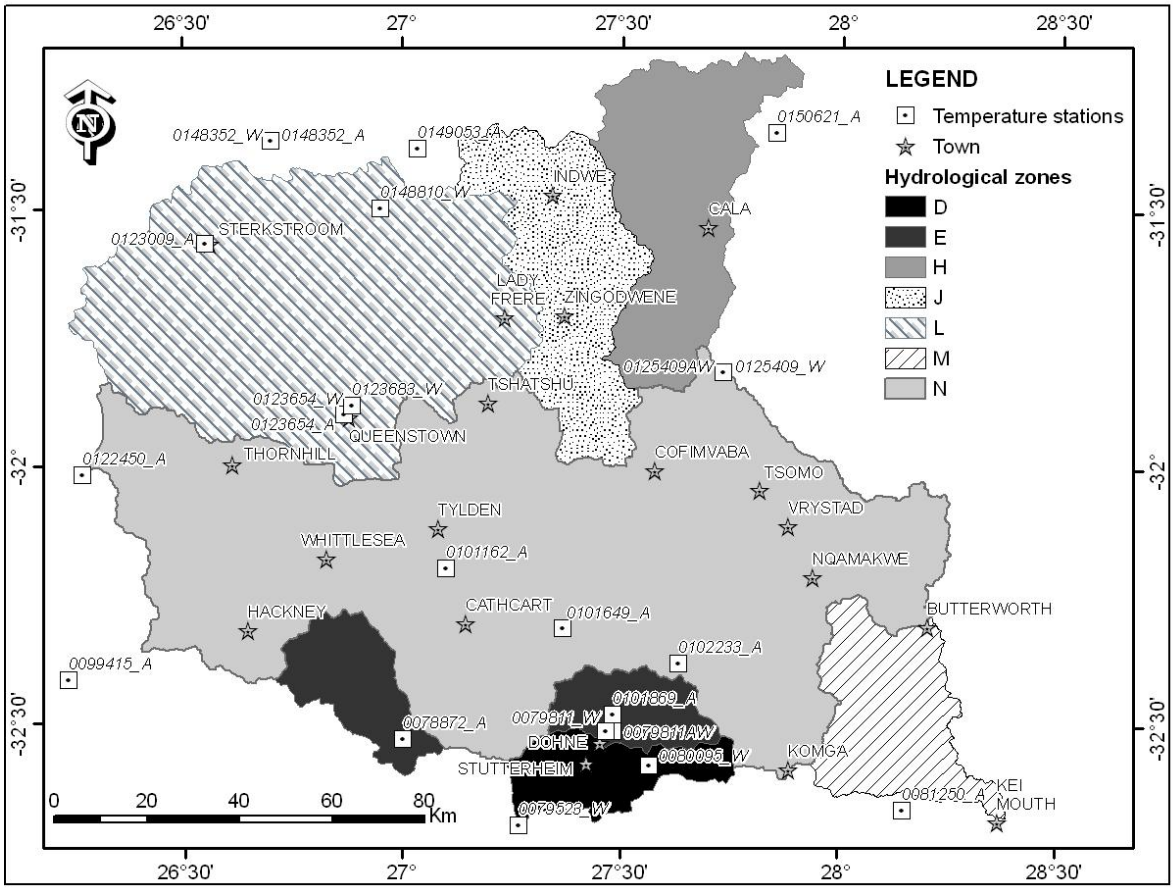


Figure 5- 18 : Temperature stations used in the study.

As shown in Figure 5- 18, the study area is subdivided into seven hydrological zones. According to a study by Midgley *et al.*, 1990, areas in the same hydrological zone have similar hydrological responses. This means that areas falling in a hydrological zone, for example the regions covered by Zone M can be assigned the same rainfall and temperature readings from the station 0081250_A. This was the approach used for assigning temperature values to the quaternary catchments. Average monthly time-series temperature data from January 1950 to December 1999 were used in the study.

Runoff

Runoff is the amount of water yield through a catchment and consists of stormflow9, baseflow10 as well as seepage[11], normal flow[12] and overflow from any reservoirs within the catchment (Tarboton *et al.*, 1992). It is a reflection of the quantity of water available in the catchment. Although daily flow data are measured at gauged stations in the catchments, they are unsuitable for direct use in modelling because they contain gaps due to missing records. The data are sometimes available for a short observation period or they may represent different sequences of dry and wet years (Smakhtin, 2000).

A time-series dataset of quality controlled runoff values was used in this research and this can be obtained from the Water Resources Simulation Model (WRSM) 2000 software that was produced from a WRC project completed in 2008 (Middleton and Bailey, 2008).

The study produced predicted monthly time-series values per quaternary catchment based on the Pitman hydrology model (Pitman *et al.*, 2007). The simulated monthly values from year 1950 to year 1999 for each quaternary catchment were used. The discretisation of runoff was done using the classes defined by Middleton and Bailey, 2008 but these were refined to capture the variations in the data. The results are shown in Table 5- 4.

Table 5- 4: The discretisation states for runoff

Parameter	States	Description
Runoff (million m^3)	Class 1	<0.2
	Class 2	0.2-0.4
	Class 3	0.4-2
	Class 4	2-5
	Class 5	5-10
	Class 6	>10

[9] Stormflow is water that flows over or near the surface of a catchment during and after a rainfall event and contributes to flows in the river within the catchment.
[10] Baseflow consists of water from previous rainfall events that has percolated through the soil horizons into the intermediate and groundwater zones and then contributes as a delayed flow to the streams in the catchment.
[11] Seepage is the water that seeps through the base and the walls of the reservoir.
[12] Normal flow is the water that should be released legally from a reservoir to supply downstream users.

Groundwater recharge

Groundwater recharge is the amount of water that enters the aquifer[13] per unit time. It results from the percolation and infiltration of water through the upper surface and the accumulation of water to the upper surface of the saturated zone (Parsons, 2004).

The amount of recharge available per quaternary catchment for the whole country is estimated as a proportion of rainfall. The percentage values used were generated as part of a Groundwater Resources Assessment (GRAII) study carried out by the Department of Water Affairs. The study considered the effects of land cover, slope, soil type, geology and rainfall on recharge for the whole South Africa (DWAF, 2005).

In this research, the percentages were multiplied with the monthly rainfall data to get estimate recharge values for the month for each quaternary catchment from the year 1950 to the year 1999. The recharge percentages are assumed constant throughout the modelling timescale. This is justifiable as they resulted from a study that considered years of monitoring data. The accuracy of the recharge values is influenced by the accuracy of the rainfall data as they are directly related. The discretisation of recharge was performed based on the levels defined for rainfall and the results are shown in Table 5- 5.

Table 5- 5: The discretisation states for recharge

Parameter	States	Description
Recharge (mm)	Class 1	<1.5
	Class 2	1.5-5
	Class 3	5-10
	Class 4	10-20
	Class 5	20-50
	Class 6	>50

[13] A water bearing geological formation capable of yielding groundwater in useful amounts.

Groundwater exploitation potential

The groundwater exploitation potential map used was estimated from the groundwater harvest potential. Harvest potential is the maximum volume of groundwater that is available for abstraction without depleting the aquifer systems and takes into account recharge, storage and drought periods (DWAF, 2005). The groundwater exploitation potential was then calculated by multiplying the Harvest Potential with an exploitation factor determined from borehole yield data. The exploitation potential indicates the portion of the Harvest potential that can practically be exploited. The volume units of the potential are million m^3 per annum (DWAF, 2002b).

Groundwater exploitation potential is affected by the geology of the area. Geology is important as the make-up of rocks verify the degree to which they have been weathered. The aspect of geology considered is this research is the sandstone ratio (Figure 5- 19).

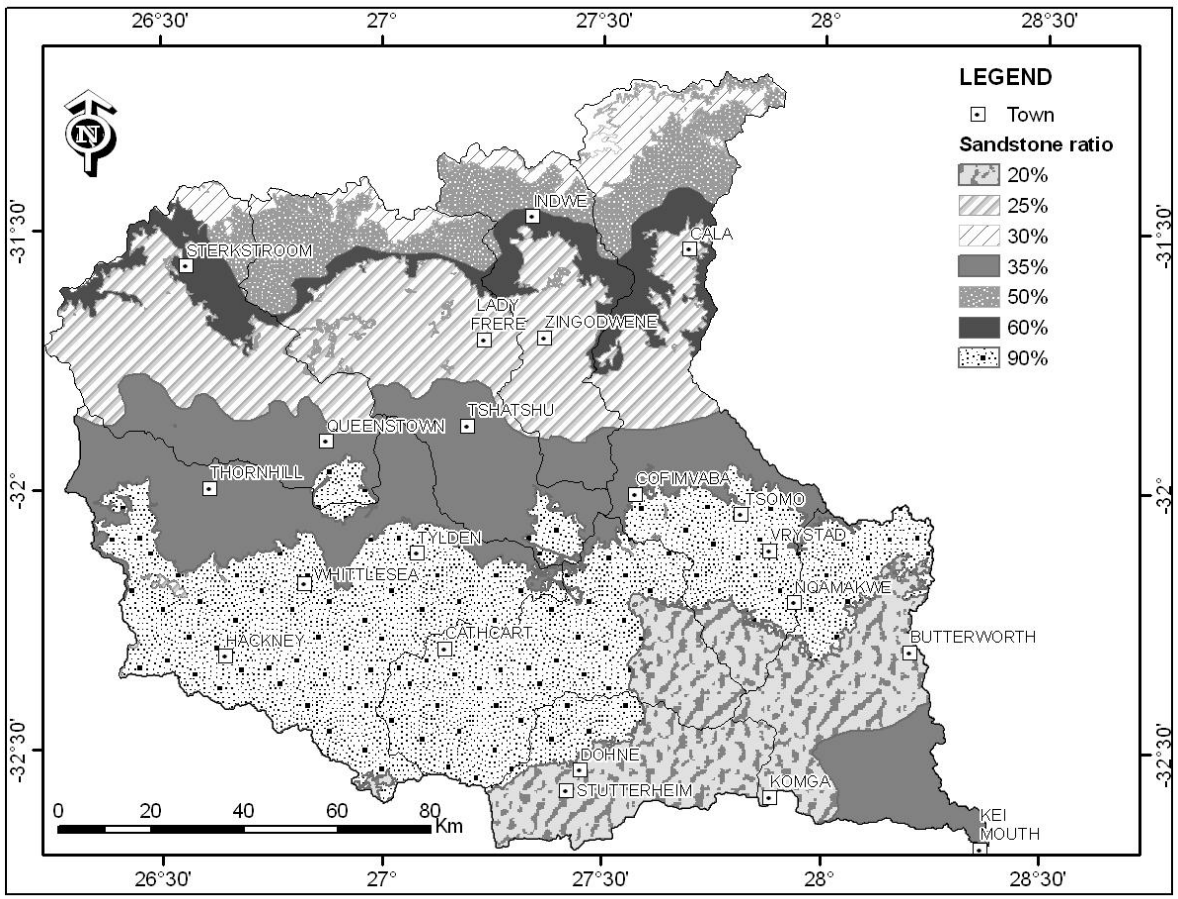


Figure 5- 19 : Sandstone ratios for the geology of the study area (after Dondo *et al.*, 2010).

The sandstone ratio influences the definition of the groundwater aquifer type. A high sandstone ratio shows good porosity and permeability and areas were the geology has high sandstone ratio have high groundwater potential (Chevallier *et al.*, 2001). Four variables for geology were set-up representing the proportion of the area with the following sandstone ratios:

i) 0-35% sandstone ratio;
ii) 35-50% sandstone ratio;
iii) 50-75% sandstone ratio and
iv) 75-100% sandstone ratio.

The variables were discretised using the classes shown in Table 5- 6 and the values were considered constant throughout the modelling period.

Table 5- 6: The discretisation states for sandstone variables

Parameter	States	Description
0-35%; 35-50%; 75-100% sandstone ratio	VeryLow	<25
	Low	25-50
	Medium	50-75
	High	>75
50-75% sandstone ratio	VeryLow	<10
	Low	10-25
	Medium	25-50
	High	>50

Water demand

Water demand is the volume of water used by all water-use sectors including domestic, mining, agriculture, commercial and industrial. This demand must be below the water supply in order for the catchment water resources to be sustainable. If demand is above supply, it could indicate a stressed catchment (Walmsley *et al.*, 2004). This variable can be related to land use change, population and human settlement, water demand use and availability and biodiversity change.

The data used for this variable was the result of the 1995 water resources assessment study done by DWA. The water requirements were quantified using monthly time series of natural flows from a rainfall-runoff simulation model. A yearly figure, in units of million m^3 per annum, was provided for each quaternary catchment. For the purposes of this study, these water demand figures were considered constant throughout the modelling period. The data used include urban, rural and irrigation water requirements (DWAF, 2002b).

Volume of groundwater utilised

The volume of groundwater utilised is estimated by considering groundwater use by the different economic sectors in each quaternary catchment. There is a scarcity of data for this variable throughout most quaternary catchments in South Africa. This is largely due to the fact that pre-1994; there was no legal requirement for owners of private boreholes like farmers and private companies to register their water use. Post-1994, a water policy is now in place that makes it mandatory for the registration and licensing of boreholes so with time more reliable estimates will be available. This variable can affect land use, water use demand and availability, human settlement and population change and habitat condition (Walmsley *et al.*, 2004).

The data used was the result of the 1995 water resources assessment study by DWA (DWAF, 2002b). The assessment produces a yearly figure for each quaternary catchment. For the purposes of this study, this groundwater use amount; provided in units of million m^3 per annum was considered constant throughout the modelling period for each quaternary catchment.

5.2.3.2 Waste and pollution

Waster and pollution is a threat to water resources in the catchment. Because waste estimates are difficult to obtain especially in rural catchments, only pollution variables are considered. Pollutants originate mainly from industrial, mining, and agricultural sources.

Organic and inorganic chemicals, plant nutrients, oxygen-demanding wastes, radioactive elements, sediments and microbiological components are the major sources of pollution in catchment ground and surface water (DEAT, cited in Walmsley, 2002). The variables considered include surface water quality and groundwater quality.

Surface water quality

The surface water quality components considered are electrical conductivity (EC), Phosphorus (P) and Nitrogen (N) (Kelbe *et al.*, 1999). According to Walmsley *et al.*, 2004, EC is the most important variables in assessing catchment sustainability as it measures salinity. EC is recorded in milliSiemens per metre (mS/m). It is a gauge of the ability of water to conduct an electric current. This ability results from the presence of ions such as carbonate, bicarbonate, sulphates, sodium, potassium, calcium, nitrate, chloride and magnesium (DWAF, 1996). The rocks and soils have an effect on EC which can range from 40 mS/m for rainfall with little infiltration to 600mS/m for groundwater in contact with saline or mudrock. Domestic and industrial effluent discharges and runoff from urban, industrial and cultivated areas can also affect EC values. EC can therefore be used to indicate likely pollution sources in the catchment.

Nitrogen (N) and Phosphorus (P) are the key water quality variables linked to non-point source or diffuse pollution. Agricultural activities have been acknowledged worldwide as the major source of this type of pollution. High nitrate levels in drinking water can be hazardous to children and livestock and high P and N levels in surface waters can cause eutrophication which results in nuisance plant growth, decline in some types of fish and excessive loss of water through evapotranspiration (van Ginkel, 2002 cited in Annandale *et al.*, 2004).

According to Cullis *et al.*, 2004, there is generally a relationship between water quality and the land use in a catchment. Fertiliser loadings from agriculture, increased erosion and agricultural practises like ploughing can also impact on the quality of the water.

The poor quality surface water may negatively affect the growth of aquatic plants or the fitness of the water for use, be it for agricultural, domestic or industrial (Cullis *et al.,* 2004; Kelbe *et al.,* 1999; & Parsons, 2004). It is therefore important to investigate the relationship between surface water quality and land use.

Phosphorus can occur in numerous organic and inorganic forms and may be present as dissolved and particulate species. Inorganic soluble phosphorus, in the form orthophosphates (HPO_4^{2-} and $H_2PO_4^-$) is the phosphorus immediately available to aquatic biota and therefore needs to be properly managed for the prevention of algal bloom. Phosphorus concentrations in unimpacted rivers are between 0.01-0.05 milligrams per litre (mg/ℓ). Nitrogen also occurs in many forms and ammonia (NH_3) and ammonium (NH_4^-) are the reduced forms of inorganic nitrogen. Their proportions in water are influenced by water temperature and pH (Cullis *et al.,* 2004). Nitrogen concentrations in unimpacted rivers are less than 0.5 mg/ℓ. Figure 5- 20 shows the surface water quality sites used in this study.

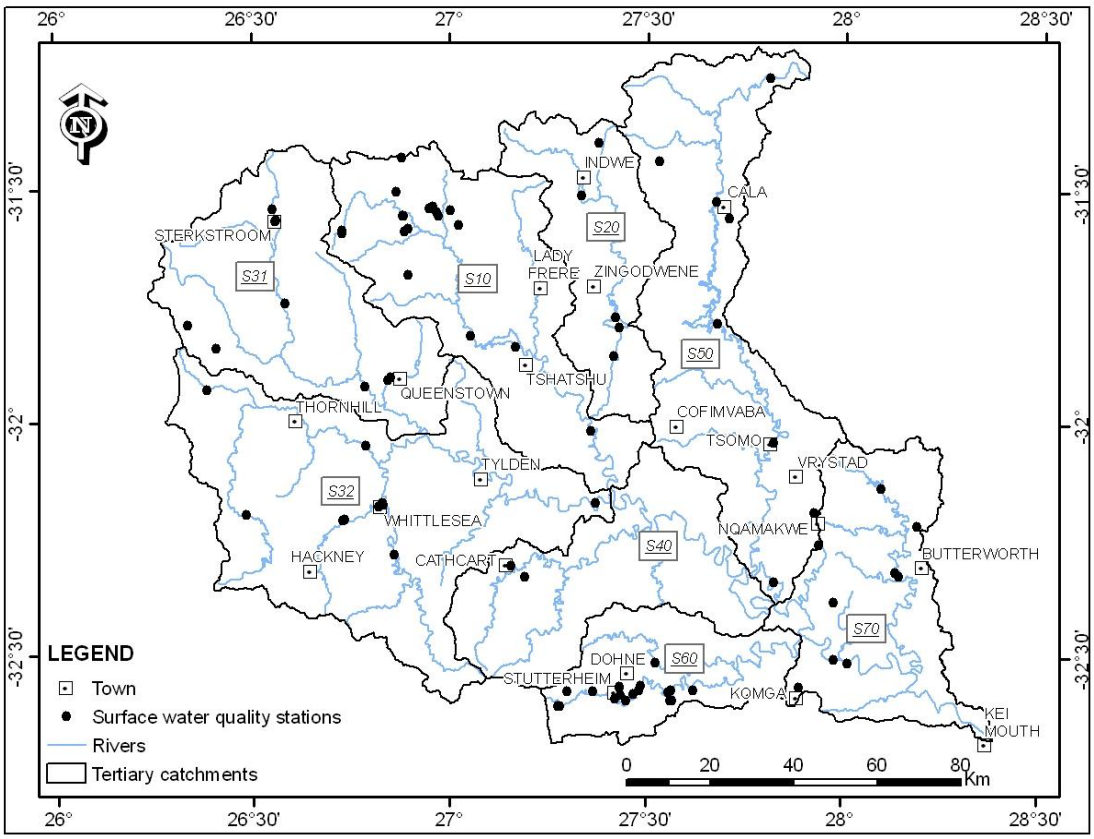


Figure 5- 20 : Surface water quality stations used in the study.

Surface water quality readings are available at monitoring points set up in catchments throughout South Africa. Figure 5- 20 shows the stations in the study area that have adequate data for modelling.

The parameters measured at these stations are PO_4^{-3}, which is used to represent inorganic phosphorus and NH_4^- and NO_3 (the sum of these two parameters represents inorganic nitrogen) (Cullis *et al.*, 2004). The units of measure for both parameters are mg/ℓ. Due to the unavailability of time-series data, average values were calculated for the catchments and these were considered constant for the entire modelling period. This is justifiable as most of the readings are within acceptable limits (see Appendix C for a list of the data).

The surface water quality data have some of the following sources of uncertainty (Rode *et al.*, 2005):

i) instrument errors and calibration errors;
ii) selection of unrepresentative sampling sites; and
iii) inadequate sampling frequency.

Groundwater quality

Groundwater quality is an indicator of whether or not the resource is under stress or if there is any threat to its sustainability (Cape Water Programme, 1999). The groundwater quality variables included in this research are electrical conductivity (EC) and sodium adsorption ratio (SAR) (Lesch and Suarez, 2009). The boreholes with quality data are shown in Figure 5- 21. EC, which is measured in milliSiemens per metre (mS/m), was converted to TDS by multiplying the values with a conversion factor of 6.5 (DWAF, 1999 cited in Cullis *et al.*, 2004). The standards for domestic and irrigation quality are shown in Table 5- 7.

SAR is used for evaluating the sodium hazard associated with an irrigation water supply. Irrigation waters with high SAR levels can lead to high sodium levels over time which affects soil infiltration and percolation rates. Excessive SAR levels lead to soil crusting and poor aeration (Lesch and Suarez, 2009).

SAR is calculated using the following equation (with the cation measurements in milliequivalents per litre (meq/ℓ) :

$$SAR = \frac{Na}{\sqrt{(Ca + Mg)/2}}$$

Because the sodium, calcium and magnesium values in boreholes are measured in mg/ℓ, these were first converted to meq/ℓ using the following equations (Lesch and Suarez, 2009):

i) Sodium (Na in meq/ℓ) = $\frac{1}{22.99}$ x Na (Na in mg/ℓ)

ii) Calcium (Ca in meq/ℓ) = $\frac{2}{40.08}$ x Na (Na in mg/ℓ)

iii) Magnesium (Na in meq/ℓ) = $\frac{2}{24.31}$ x Na (Na in mg/ℓ)

Table 5- 7: The classes of TDS for domestic and irrigation water use based on the South Africa water quality guidelines

Parameter	Classes	Description
TDS	<585	Excellent water quality
	585-1 755	Good water quality
	1 755-3 510	Poor water quality
	>3 510	Unacceptable
SAR	<6	Good water quality
	6-12	Fair water quality
	12-20	Poor water quality
	>20	Unacceptable

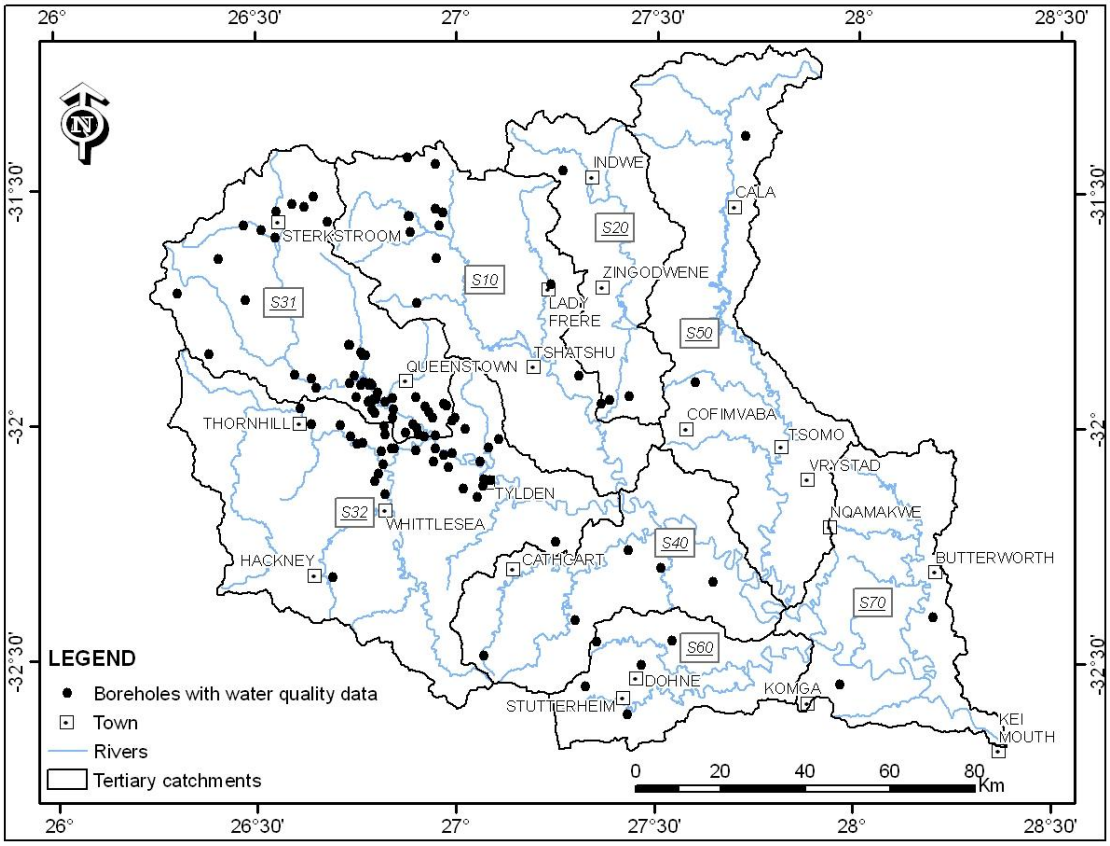


Figure 5- 21 : Boreholes with water quality data in the study area.

Due to the unavailability of time-series data, average values were calculated for the quaternary catchments and these were considered constant for the modelling period. This is justifiable as most of the readings are within acceptable limits (see Appendix C). The following are some of the sources of uncertainties in groundwater quality data (Nilsson *et al.*, 2006):

i) the choice of the location of the monitoring sites might not be optimal;

ii) the position of some of the sites is not accurately measured;

iii) some of the values are recorded erroneously;

iv) the frequency and time of sampling might not be adequate; and

v) the data might have been observed from different seasons which will not be corresponding for all stations in the quaternary catchment

5.2.3.3 Socio-economic condition

Socio-economic indicators include social statistics like the population density and the amount of people without adequate access to water and sanitation services. This affects the water use, future demand and pollution of water resources in the catchment. It also affects the ecosystem health.

Population density

The population density was obtained by dividing the population of each quaternary catchment with its surface area (in km^2). The population estimates used were calculated per quaternary catchment and were based on the Census carried out in 1996. Population density is vital because population growth is stated as the primary cause of future water requirements in catchments in South Africa. It also impacts on the pollution of water resources in the area and the ecosystem health. For water resource assessment it is vital to model the interactions between people and water resources quantity and quality (DWAF, 2004).

The population values were considered constant throughout the modelling period. Ideally, time-series data should have been used but these were not available.

Urbanisation

Urbanisation is a ratio of the population of the catchment that is in urban areas over the total population of the catchment. The number of people living in urban areas affects the water use and infrastructure requirements and also has impacts on the pollution management and ecosystem health. The data used was from the 1996 Census and is provided per quaternary catchment. It was discretised into the states shown in Table 5- 8.

Table 5- 8: The classes used for discretisation of the urbanisation variables based on data analysis

Parameter	States	Description
Urbanisation	Class1	<0.1
	Class2	0.1-0.25
	Class3	0.25-0.5
	Class4	0.5-0.75
	Class5	>0.75

Percentage of households with no access to adequate water

The information on access to water and sanitation was gathered during the 1996 census and is available at municipal ward level. Adequate access to water was defined as access to piped water in and the rest of the households without this level of services were deemed unserved.

Adequate sanitation was defined as either flush or chemical toilets. The data used in this research is the number of households with no access to a flush or chemical toilet (STATSSA, 1996). The data is provided per municipal ward in the different regions and the data were allocated to the quaternary catchments by averaging for all wards lying in the catchment. The values were discretised using the states shown in Table 5- 9.

Table 5- 9: The defined states for water and sanitation access

Parameter	States	Description
Water access	Class1	<10
	Class2	10-20
	Class3	20-40
	Class4	40-50
	Class5	>50
Sanitation access	Class1	<10
	Class2	10-25
	Class3	25-50
	Class4	50-65
	Class5	>65

Uncertainties in socio-economic data like population and access to water and sanitation arise from the following (Kaluer *et al.,* 2005):

i) infrequency in data collection as censuses are performed every 5 years so for years in between the figures have to be projected;

ii) lack of consistency in estimates of population between different demographers and study areas and studies in other catchments in South Africa suggest that the undercount in censuses could be up to 19% (DWAF, 2007); and

iii) mismatch in spatial units due the fact that population data are collected and enumerated at statistical or managerial units (for example wards, districts) whereas in catchment studies they are required at catchment level. This uncertainty can be divided into that due to boundary imprecision between catchments and administrative units and the scaling effect of assigning values to catchments (see Figure 5- 22).

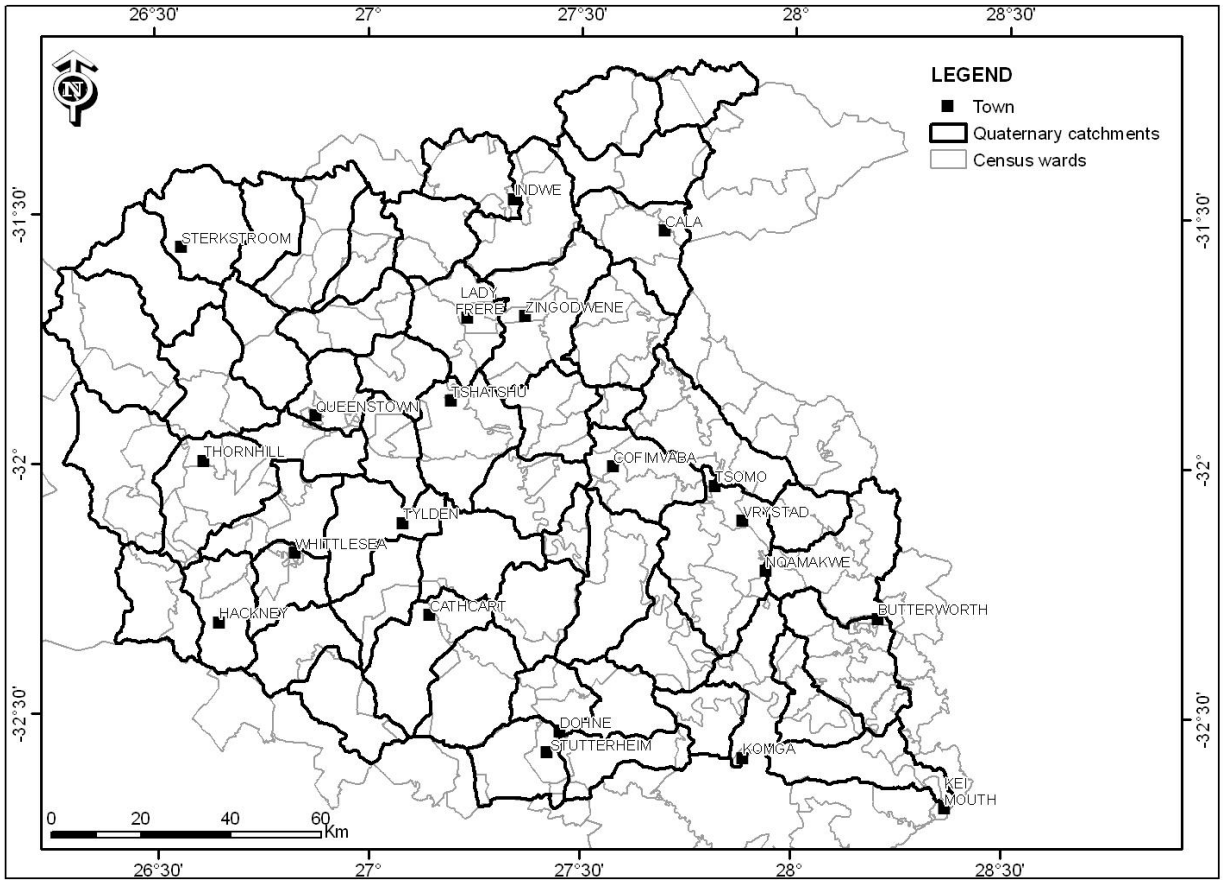


Figure 5- 22 : The spatial unit mismatch between wards and quaternary catchments.

5.2.3.4 Resource condition

Resource condition indicators highlight the health of the ecosystem and the river health. Walmsley *et al.,* 2004 proposes variables like the South African Scoring System (SASS), the Fish Assemblage Integrity Index, the Index of Habitat Integrity, and the Riparian Vegetation Index. These variables are ideal for monitoring the river health as part of the River Health Monitoring plan established by DWA. Although the plan is to monitor these variables for all rivers, currently no rivers in the study area have been fully assessed. The other stated variable with data for the study area is the proportion of the catchment covered by alien vegetation. The two other variables included in this research are the irrigated area and the degraded land.

Irrigated area

The irrigated area represents the regions in a quaternary catchment where crops are grown under irrigation (the units are km^2). The areas are the maximum irrigated regions when adequate water is available in the dams and rivers. The data were derived from the land cover data acquired from the CSIR (DWAF, 2004). The datasets available are for the year 1994 and 2001. This data resulted from the classification of Landsat 7® satellite imagery. The 1994 data is used and it is assumed to be representative of the conditions for the complete modelling period.

Degraded land

This variable represents the proportion of the catchment that is degraded. Land degradation leads to increases in soil erosion and changes the flow patterns in rivers and affects the water quality. Erosion results in an increase in the sediment loads of rivers and dams thereby affecting the quality and quantity of water resources and may affect aquatic life (Le Maitre *et al.,* 2007). Degraded land includes degraded forest, grassland, thicket, woodland, dongas and sheet erosion scars. The data were obtained from the land cover data from CSIR (DWAF, 2004). The data was discretised according to the classes shown in Table 5- 10.

Table 5- 10: The classes for the proportion of the area of degraded land in a quaternary catchment

Parameter	States	Description
Degraded land	VeryLow	<0.1
	Low	0.1-0.25
	Medium	0.25-0.5
	High	0.5-0.75
	VeryHigh	>0.75

The sources of uncertainty in land cover data include the following (Castilla *et al.*, 2005):

i) positional errors which depend on the quality of the orthorectification performed on the image before classification;
ii) the classification and generalisation errors when the data were created;
iii) errors due the fact that although the land cover gives a snapshot of the data at the time of data acquisition, the data is used to represent the state at other times, which is erroneous; and
iv) positional and aggregation errors that arise when assigning values to cells in a grid.

These uncertainties can be reduced by using data from more detailed surveys but at this broad catchment scale, this is often not practical.

Area covered by alien vegetation

This is the area of the catchment covered by alien vegetation. Alien vegetation threatens the ecosystem as they use water and this affects the availability of water resources by reducing runoff (DWA, 2004). The alien vegetation manifestations for South Africa were mapped by the CSIR using expert knowledge elicited from workshops carried out in different regions. The results were supplemented by existing detailed localised maps and GIS datasets (DWA, 2004).

The major sources of uncertainties in the data arise from the following (Görgens, 1998 cited in DWA, 2004):

i) the quality of data depends on the level of expert knowledge available, the nature of the terrain and the extent and complexity of the invasion;
ii) the mapping of alien vegetation ending unnaturally along administrative boundaries;
iii) mapping of riparian infestations along rivers at the coarse scale could have led to significant under-estimates of river lengths and, therefore, of infested riparian areas.
iv) riparian infestation identification in a particular catchment with the simple statement: "all rivers are invaded" led to all the river lengths appearing in the area being assigned a uniform infested "buffer" strip of a specific width; and
v) small rivers not reflected at the mapping scale were not accounted for and infestation along these rivers was not quantified.

The continuous data were discretised according to the classes illustrated in Table 5- 11.

Table 5- 11: Shows the discretisation classes for the alien vegetation variable

Parameter	States	Description
Alien vegetation	Class 1	<1
	Class 2	1-5
	Class 3	5-10
	Class 4	10-20
	Class 5	>20

5.2.3.5 Management indicators

Management indicators represent the institutional arrangements and capacity of the catchment for IWRM. The management indicators are not included in the network but are used to assist in the discussion of the results obtained for each primary catchment. These include the percentage of unaccounted for water in the catchment and the presence of reliable hydrological and water quality monitoring in the catchment.

Percentage of unaccounted for water in the catchment

Unaccounted for water is water lost during distribution from the source to the end user. This amount includes water from reticulation system leaks, unauthorised water connections, faulty water meters and domestic plumbing leaks. These affect the sustainability of water services (DWA, 2004). Water losses result in the increased abstraction of water and this affects the quantity of available water resources and increases the cost of water services provision.

The main challenge is that the data are available from municipalities at the municipal administrative scale, which do not coincide with the catchments boundaries. The data used for catchments are likely to be extrapolated from the municipal scale. The data used were a product of the 1995 assessment conducted by DWA (DWAF, 2002b). In order to derive the estimates assumptions had to be made about the type of raw water supply, the distance of travel of the water and the nature of the distribution systems. The data are provided in million m^3 per annum per tertiary catchment and are shown in Figure 5- 23. Catchments S31 and S32 have the most losses due to the fact that they are largely agricultural irrigated areas.

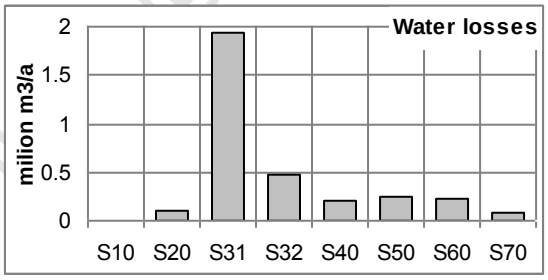


Figure 5- 23 : The water losses in the tertiry catchments.

The presence of reliable hydrological and water quality monitoring

This is the number of reliable hydrological and water quality points per 100 km^2 of the catchment. The two variables are important because continued monitoring of the catchment water availability and ongoing evaluation of the pollution levels in the catchment are important for sustainability (Walmsley *et al.*, 2004).

Active and reliable monitoring provides historical information on the water resources in terms of flow patterns. This is vital for monitoring the sustainability of the water resource. Also, with more data the uncertainties in the prediction of future scenarios of water use, quality and demand can be reduced. Figure 5- 24 shows the density of rainfall, runoff and surface water quality stations per unit area of the catchments. The distribution of groundwater chemistry monitoring holes is also shown on the map

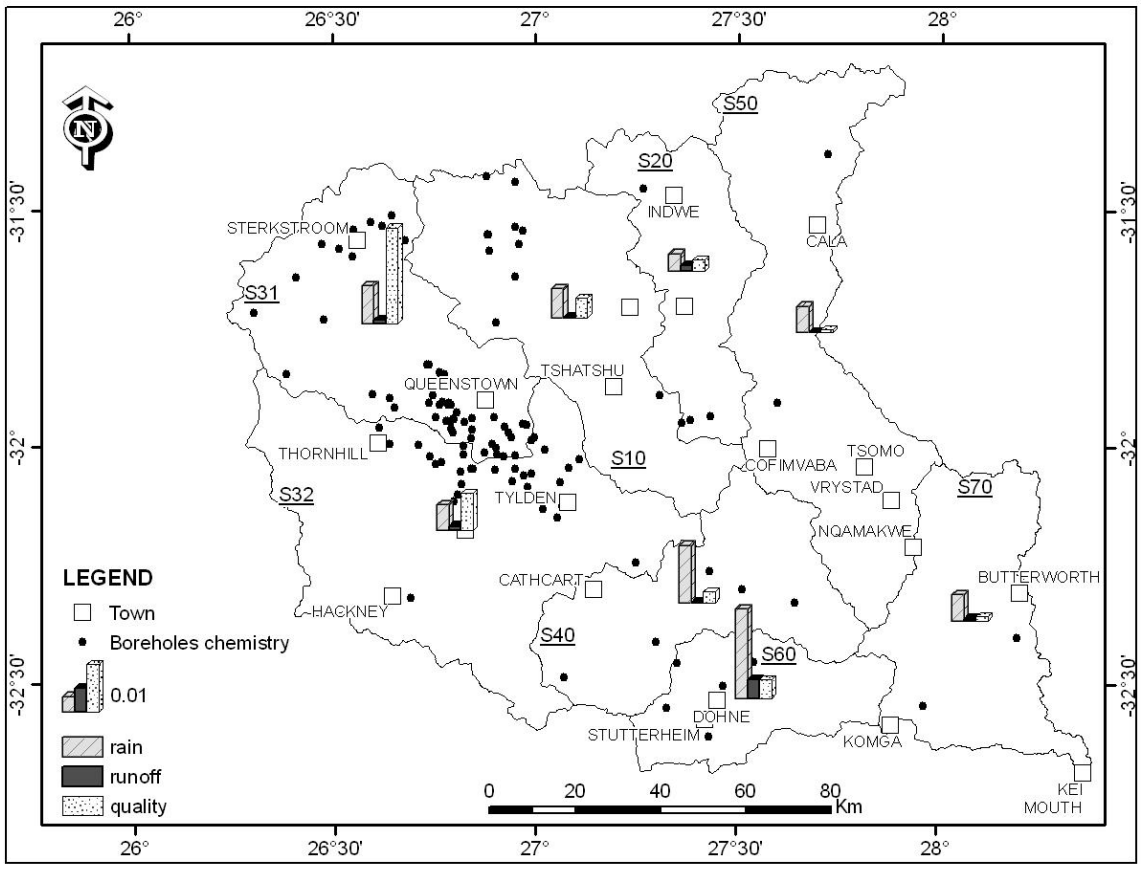


Figure 5- 24 : The density of stations in the quaternary catchments. The density is the ratio of monitoring station per 100 km^2.

Figure 5– 24 shows that most of the groundwater and surface water quality stations are in catchments S31 and S32 and the rest of the catchments are not being properly monitored. S31 and S32 are largely farming areas so there was and is still an ongoing need to monitor water quality at stations along the river. The density of rainfall points is highest in catchments S60 and S40, with S60 having the highest density of river flow gauges. S40 and S60 are also irrigated but not to the extent of tertiary catchments S31 and S32.

If water resources have to be monitored and assessed for sustainability; adequate monitoring stations have to be set up in the rest of the catchments.

The previous section provided information on the variables being used in modelling. After the data was combined for each quaternary catchment, Pearson correlation coefficients were calculated to measure dependences between the variables. This is discussed in the next section.

5.2.4 Pearson correlation matrix of all the variables

The Pearson correlation coefficient is a measure of the linear dependence between two variables. The results are values ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation). A value of 1 means that all data points are located on a straight line for which Y increases when X increases. A value 0 implies the lack of a linear correlation between the variables. A value of -1 means that all data lies on a straight line of on which Y decreases when X increases.

A Pearson correlation matrix was created for all variables and the results were analysed using the thresholds outlined in Table 5- 12. The matrix is shown in Table 5- 13.

Table 5- 12: Thresholds for interpreting the results of Pearson correlation analysis

Correlation	Negative	Positive
None	-0.09 to 0.0	0.0-0.09
Small	-0.3 to -0.1	0.1 to 0.3
Medium	-0.5 to -0.3	0.3 to 0.5
Large	-1.0 to -0.5	0.5 to 1.0

In Table 5- 13, the shaded cells highlight correlated variables indicting the relationships of interest to the study. The variables with medium to large correlations are indicated in bold italics. The variables are described in Table 5- 14.

Table 5- 13: The Pearson correlation matrix showing the correlation between all the variables

	Rain	Temp	Rech	Runoff	Pop_Den	Urbanisation	San Access	AlienVeg	Gwateruse	Wat Access	Irr_area	Wat_dem	100% sand	1/4 sand	1/2 sand	3/4 sand	Sur_EC	Surf_P	Surf_N	Degraded	GW_SAR	GW_TDS	GWPot	WaterAvail
Rain	1																							
Temp	***0.30***	1																						
Rech	***0.81***	0.19	1																					
Runoff	-0.01	-0.06	0.05	1																				
Pop_Den	0.06	0.00	0.16	0.00	1																			
Urbanisation	0.03	0.02	0.08	-0.04	***0.47***	1																		
San Access	-0.08	0.14	-0.16	-0.07	0.04	***0.33***	1																	
AlienVeg	-0.04	0.08	-0.03	-0.01	-0.16	-0.06	0.21	1																
Gwateruse	-0.07	0.08	-0.14	-0.07	-0.26	-0.10	0.13	-0.07	1															
Wat Access	-0.09	0.11	-0.20	-0.10	0.07	0.35	***0.95***	0.11	0.22	1														
Irr_area	-0.08	0.06	-0.20	-0.06	-0.10	0.06	***0.45***	0.15	***0.34***	***0.53***	1													
Wat_dem	0.00	0.05	0.05	-0.02	***0.75***	***0.56***	0.28	-0.08	-0.11	***0.33***	0.22	1												
100% sand	0.01	0.20	0.03	0.04	-0.12	0.03	0.23	***0.36***	-0.18	0.14	0.16	0.04	1											
1/4 sand	0.02	-0.03	0.04	-0.02	0.23	-0.12	-0.22	-0.26	0.16	-0.15	-0.03	0.01	-0.63	1										
1/2 sand	-0.03	-0.21	-0.09	-0.03	-0.06	0.10	-0.01	-0.18	-0.01	0.04	-0.19	-0.03	-0.56	-0.22	1									
3/4 sand	-0.02	-0.17	-0.04	-0.03	-0.11	0.05	-0.12	-0.15	0.17	-0.09	-0.10	-0.10	-0.49	-0.21	***0.57***	1								
Sur_EC	-0.09	0.13	-0.21	-0.08	-0.25	-0.08	0.22	***0.31***	0.20	0.20	0.19	-0.04	0.23	-0.20	-0.03	-0.12	1							
Surf_P	-0.03	0.05	-0.09	-0.05	0.16	0.10	0.26	-0.05	***0.66***	0.28	***0.30***	0.18	-0.15	0.17	-0.02	0.05	0.17	1						
Surf_N	-0.07	0.12	-0.17	-0.05	-0.11	0.10	***0.30***	***0.48***	0.30	0.24	0.29	-0.02	0.25	-0.12	-0.18	-0.13	***0.54***	**0.48**	1					
Degraded	0.01	0.01	-0.02	0.08	0.18	-0.15	-0.27	-0.10	-0.11	-0.25	-0.11	0.10	0.04	0.05	-0.10	-0.08	-0.04	-0.08	-0.11	1				
GW_SAR	0.03	-0.01	0.05	0.06	0.02	-0.01	-0.05	-0.03	-0.10	-0.11	-0.12	0.00	-0.03	-0.07	0.14	0.02	-0.07	-0.03	-0.13	-0.16	1			
GW_TDS	-0.07	0.13	-0.14	-0.04	-0.11	-0.06	0.28	0.22	0.11	0.27	0.31	-0.02	***0.35***	-0.07	***-0.40***	-0.22	0.22	0.13	0.25	-0.14	0.09	1		
GWPot	-0.05	0.07	-0.12	0.11	-0.13	-0.25	-0.02	-0.01	0.20	0.04	0.23	-0.05	0.12	0.02	-0.19	-0.10	0.09	0.05	0.05	***0.57***	-0.14	0.26	1	
WaterAvail	-0.02	-0.11	-0.20	0.06	-0.11	-0.23	-0.13	-0.27	0.00	-0.06	0.18	-0.10	-0.26	0.12	0.19	0.15	-0.19	-0.04	-0.27	0.27	-0.03	0.13	0.43	1

Table 5- 14: The description of the variables used in the correlation matrix in Table 5- 13

Variable	Description		Variable	Description
Rain	Rainfall values (mm)		100% sand	Ratio of area with geology that has 100% sandstone ratio
Temp	Maximum temperature (°C)		1/4 sand	Ratio of area with geology that has 25% sandstone ratio
Rech	Groundwater recharge (mm)		1/2 sand	Ratio of area with geology that has 50% sandstone ratio
Runoff	Runoff (million cm^3)		3/4 sand	Ratio of area with geology that has 75% sandstone ratio
Pop_Den	Population density		Sur_EC	Surface water EC values
Urbanisation	Urbanisation		Surf_P	Surface water phosphorus values
San Access	% of population with adequate sanitation		Surf_N	Surface water nitrogen values
AlienVeg	Ratio of area with alien vegetation		Degraded	Proportion of area that is degraded
Gwateruse	Volume of groundwater used		GW_SAR	Groundwater SAR values
Wat Access	% of population with adequate water		GW_TDS	Groundwater TDS values
Irr_area	Ratio of area that is irrigated		GWPot	Groundwater exploitation potential
Wat_dem	Water demand		WaterAvail	Total water resources available

The results of Pearson correlation analysis show large; positive correlations between the following variables:

- rainfall – recharge;
- water demand – urbanisation;
- water demand – population density;
- water access – sanitation access;
- surface water phosphorus – groundwater use;
- surface water nitrogen – surface water EC; and
- groundwater potential – degraded land.

These correlations show the link between water demand and population and the effects of water use on water quality.

Medium correlations exist between the following variables:

- rainfall – temperature
- population density – urbanisation
- water/sanitation access – urbanisation
- sandstone – alien vegetation
- surface water nitrogen – alien vegetation
- surface water EC – alien vegetation
- irrigated area – groundwater use
- surface water nitrogen – groundwater use
- water demand – water access
- surface water phosphorus – irrigated area;
- groundwater TDS – irrigated area
- groundwater TDS – sandstone
- surface water phosphorus – surface water nitrogen
- water available – groundwater potential

These correlations show the link between groundwater quality, land use and geology.

The Pearson correlation results are used later in Section 5.2.6 when the final Bayesian Network model is created.

5.2.5 Data discretisation and network structure learning

The continuous variables used for modelling in this research were discretised automatically using equal-width binning. The main challenge of automatic discretisation (as discussed in Section 3.1.2) is the choice of the appropriate levels of discretisation that adequately capture the relationships in the data and provide good prediction capabilities. In equal-width binning, the user specifies *k*, which is the number of intervals or classes required and the equal-width function searches for the maximum and minimum attribute values and these are used to determine data intervals.

In this research, three different discretisation levels are defined for the data by varying the number of intervals specified for the equal-width binning.

According to Uusitalo, 2007, most of the studies illustrate the use of between 2-10 intervals for variables used in modelling in ecological studies. Based on this information, the three methods used for discretising the continuous variables in this study are:

i) Number of intervals automatically defined by the software during the structure learning process;

ii) Number of intervals (k_1) specified by the user =2; and

iii) Number of intervals (k_2) specified by the user =4.

Table 5- 15 provides a description of the variables used in modelling.

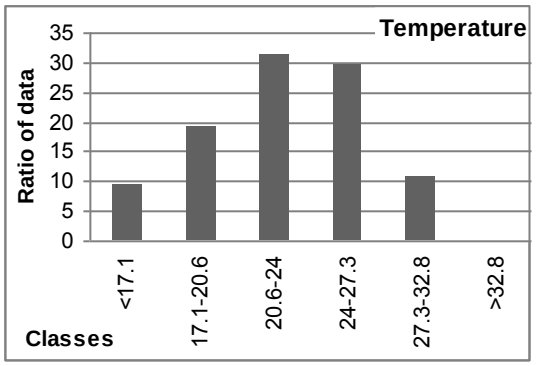
Table 5- 15: The description of the variables used in modelling

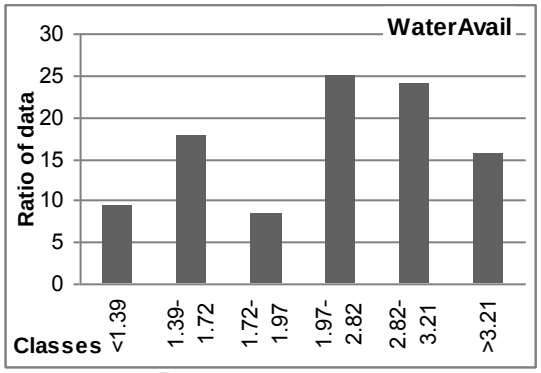
Variable	Description	Variable	Description
Rainfall	Rainfall values (mm)	Sandstone	Ratio of area with geology that has 100% sandstone ratio
Temperature	Maximum temperature (°C)	Quart sand	Ratio of area with geology that has 25% sandstone ratio
Recharge	Recharge (mm)	Half sand	Ratio of area with geology that has 50% sandstone ratio
Runoff	Runoff (million cm^3)	Threquasand	Ratio of area with geology that has 75% sandstone ratio
PopulationDensity	Population density	SurEC	Surface water EC values
Urbanisation	Urbanisation	Surf_P	Surface water phosphorus values
SaniAccess	% of population with adequate sanitation	Surf_N	Surface water nitrogen values
AlienVeg	Ratio of area with alien vegetation	Degraded	Proportion of area that is degraded
Gwateruse	Volume of groundwater used	GW_SAR	Groundwater SAR values
Water_ Access	% of population with adequate water	GW_TDS	Groundwater TDS values
Irri_area	Ratio of area that is irrigated	GWPotential	Groundwater exploitation potential
Wat_demand	Water demand	WaterAvail	Total water resources available

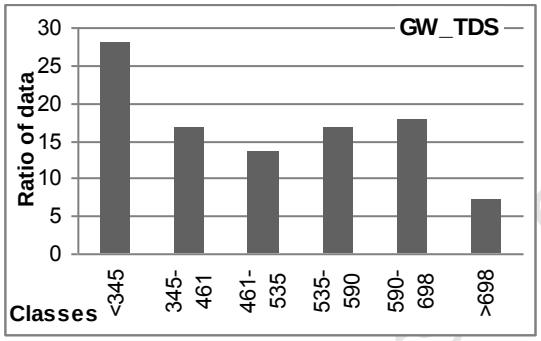
In a one-step automatic procedure, the software discretised the continuous variables and created the Bayesian Network structure. Three Bayesian Networks resulted from the three discretisation levels; result A from the automatic setting of the number of intervals by the software, result B for number of intervals $k_1 = 2$ and result C for number of intervals $k_2 = 4$. The results are presented and discussed in the following sections.

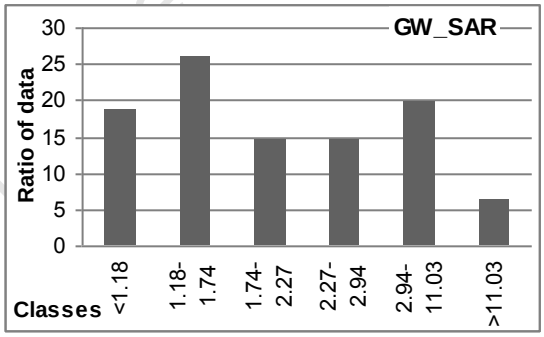
5.2.5.1 Results A with unsupervised selection of discretisation intervals

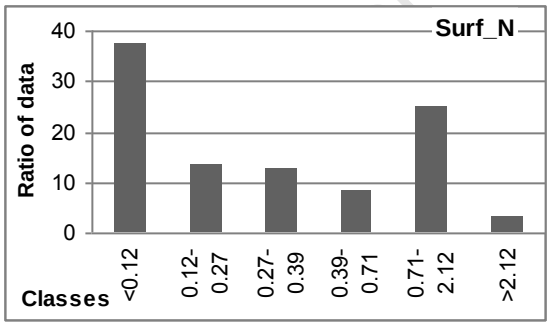
The variables were discretised using the equal binning method and allowing the custom-developed software to define the appropriate number of intervals automatically and concurrently producing the Bayesian Network structure. The discretisation results are shown in the following graphs.

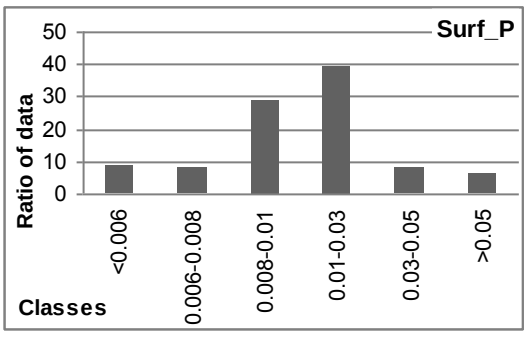


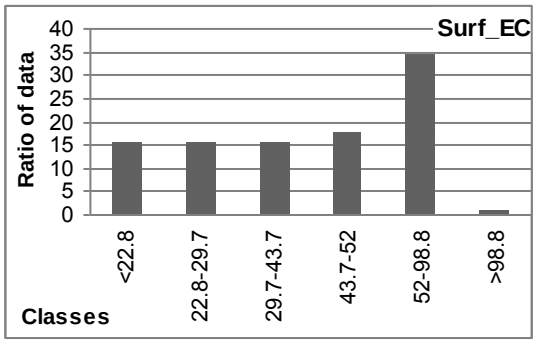


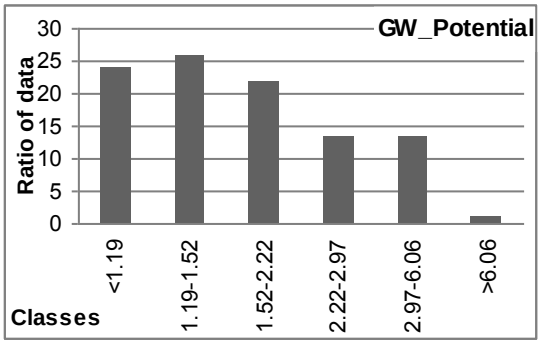


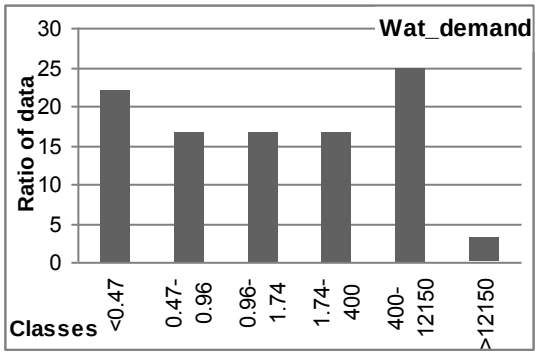


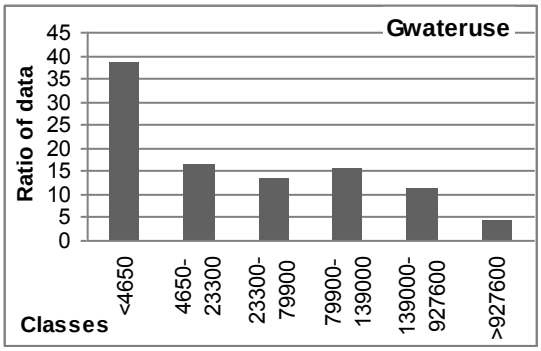


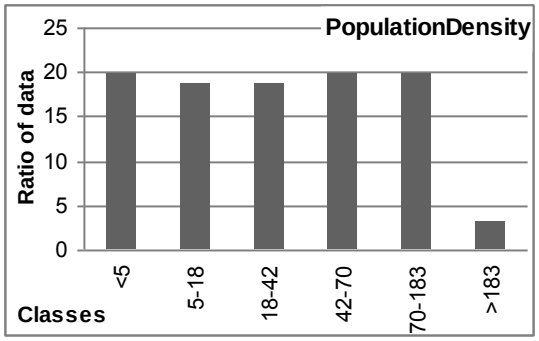


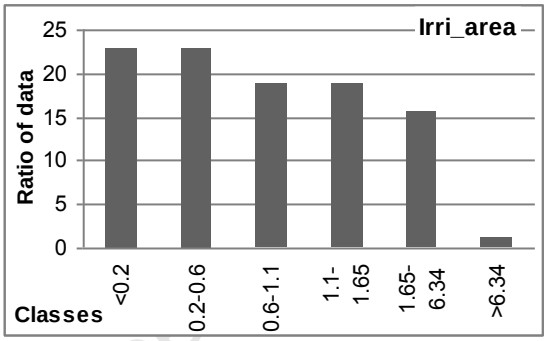












The results show that all the variables were discretised using six intervals. The resulting network is shown in Figure 5- 25.

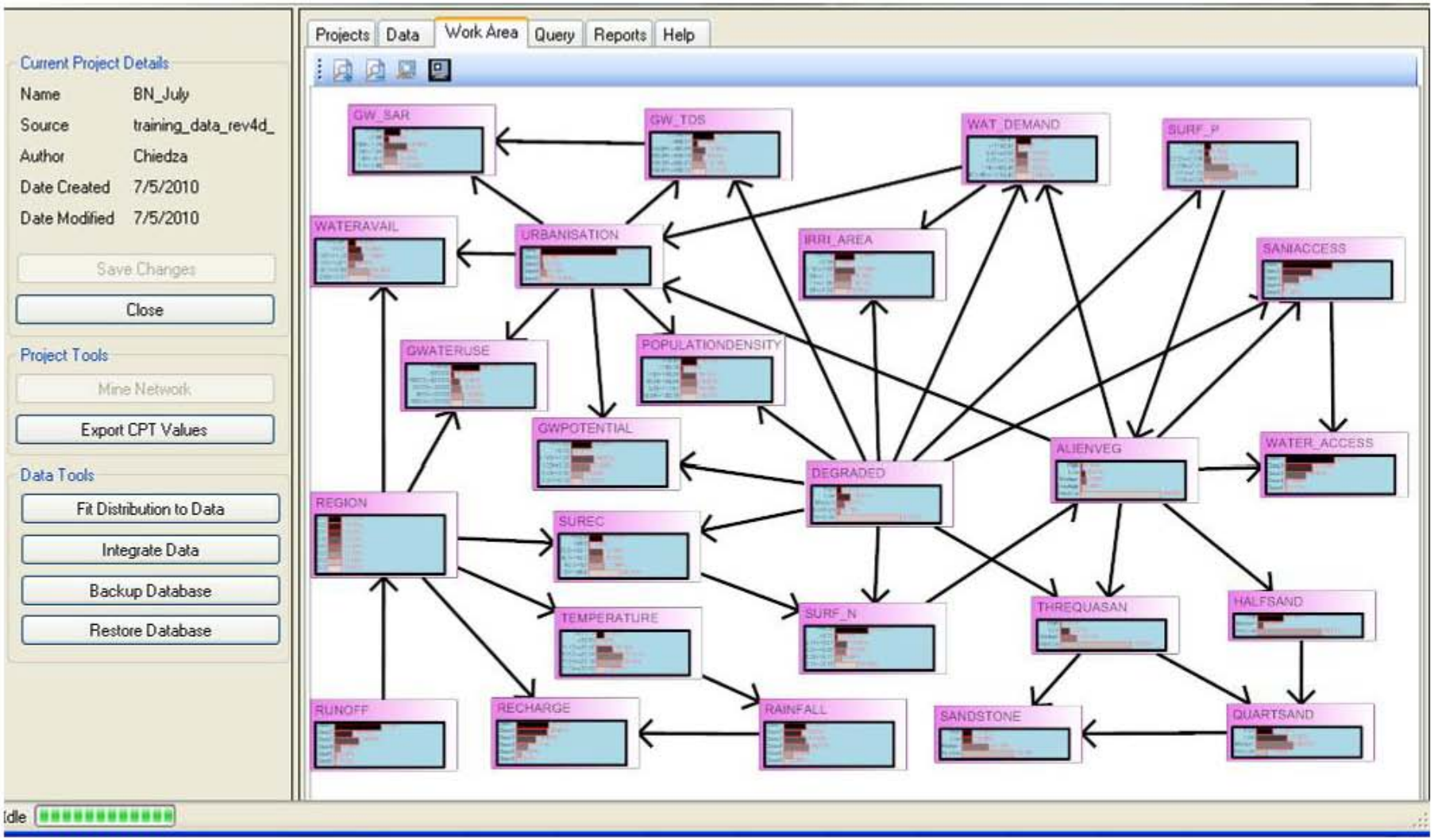
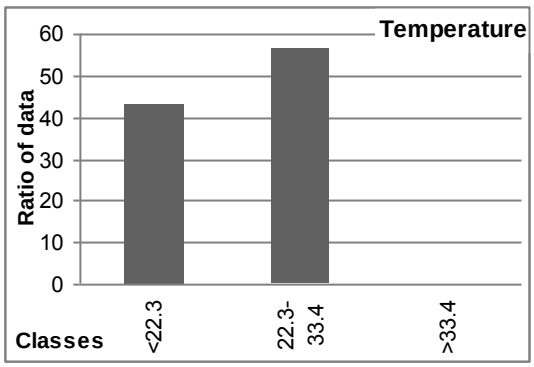
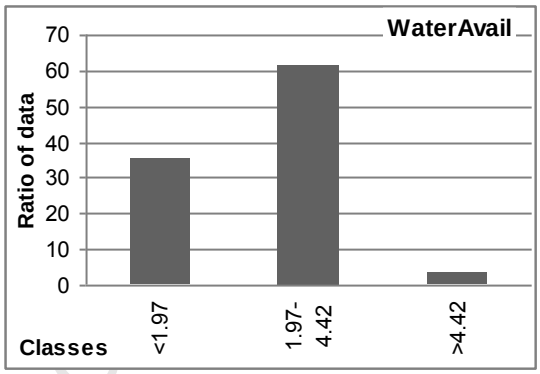


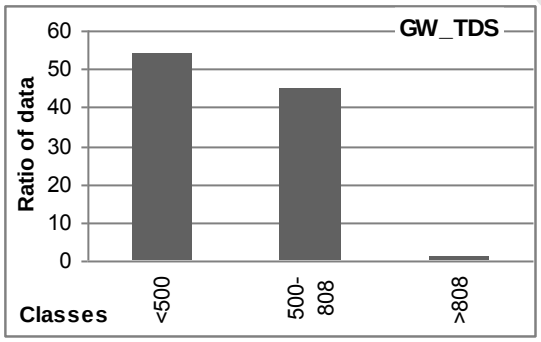
Figure 5- 25 : Bayesian Network - Result A, created and displayed in the custom developed software.

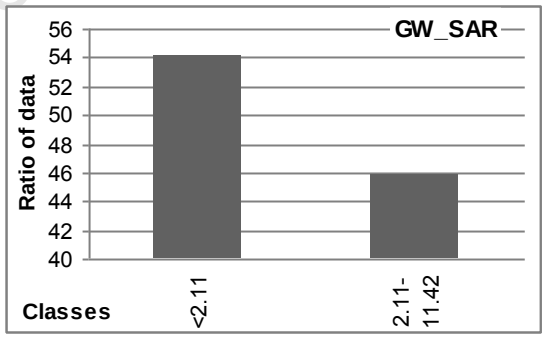
5.2.5.2 Result B with two discretisation intervals

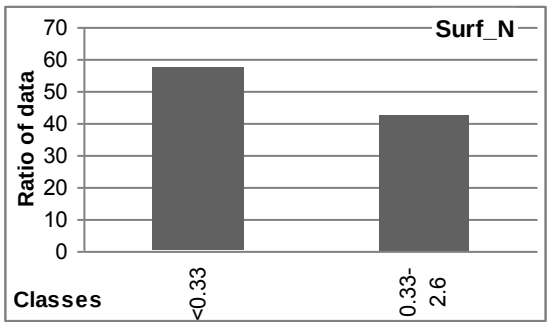
The variables were discretised using the equal binning method and setting the number of intervals $k_1 = 2$ and concurrently producing the network structure. The discretisation results are shown in the following graphs. The algorithm could not discretise some of the variables into two classes but instead three. Due to the fact that the algorithm simultaneously mines the structure as it discretises the data, this means that for these variables in order to get optimum patterns, three classes where more ideal than two.

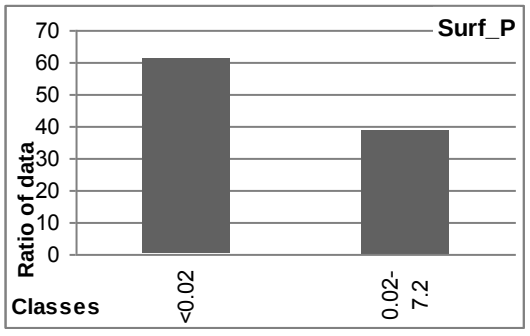


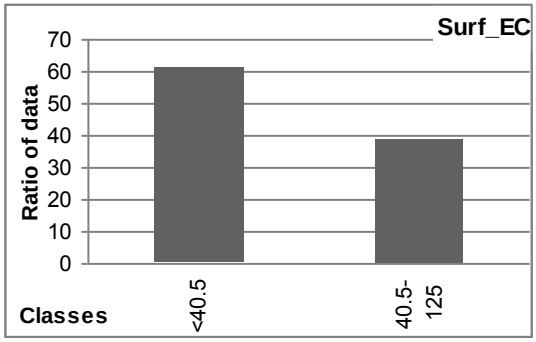


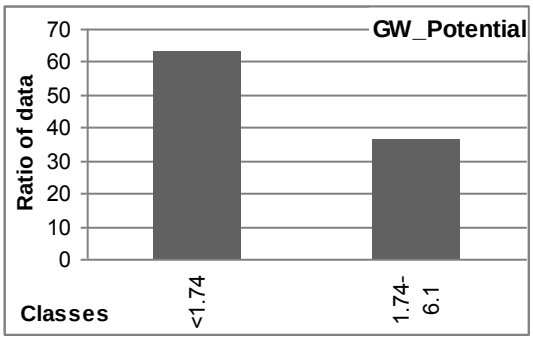


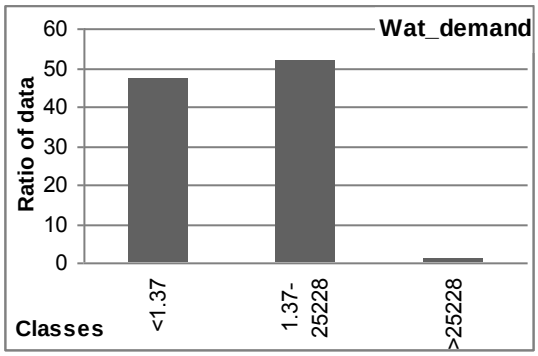


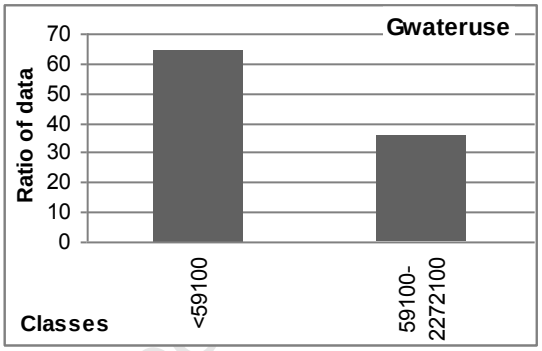


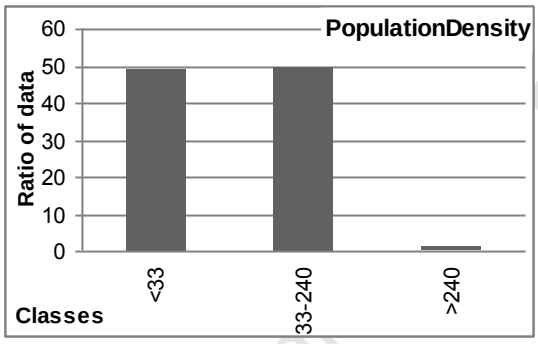


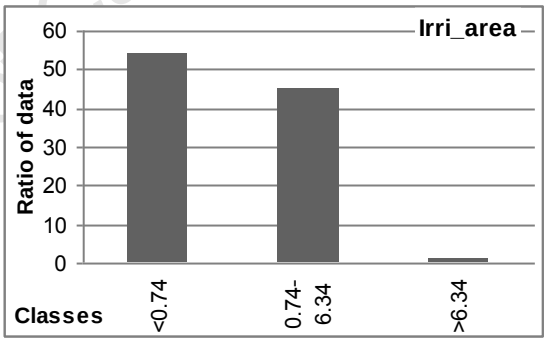












The results show that most of the variables were discretised using three variables. The only variables discretised using two intervals were groundwater SAR, surface water nitrogen, EC and phosphorus and groundwater potential. The resulting network is shown in Figure 5-26.

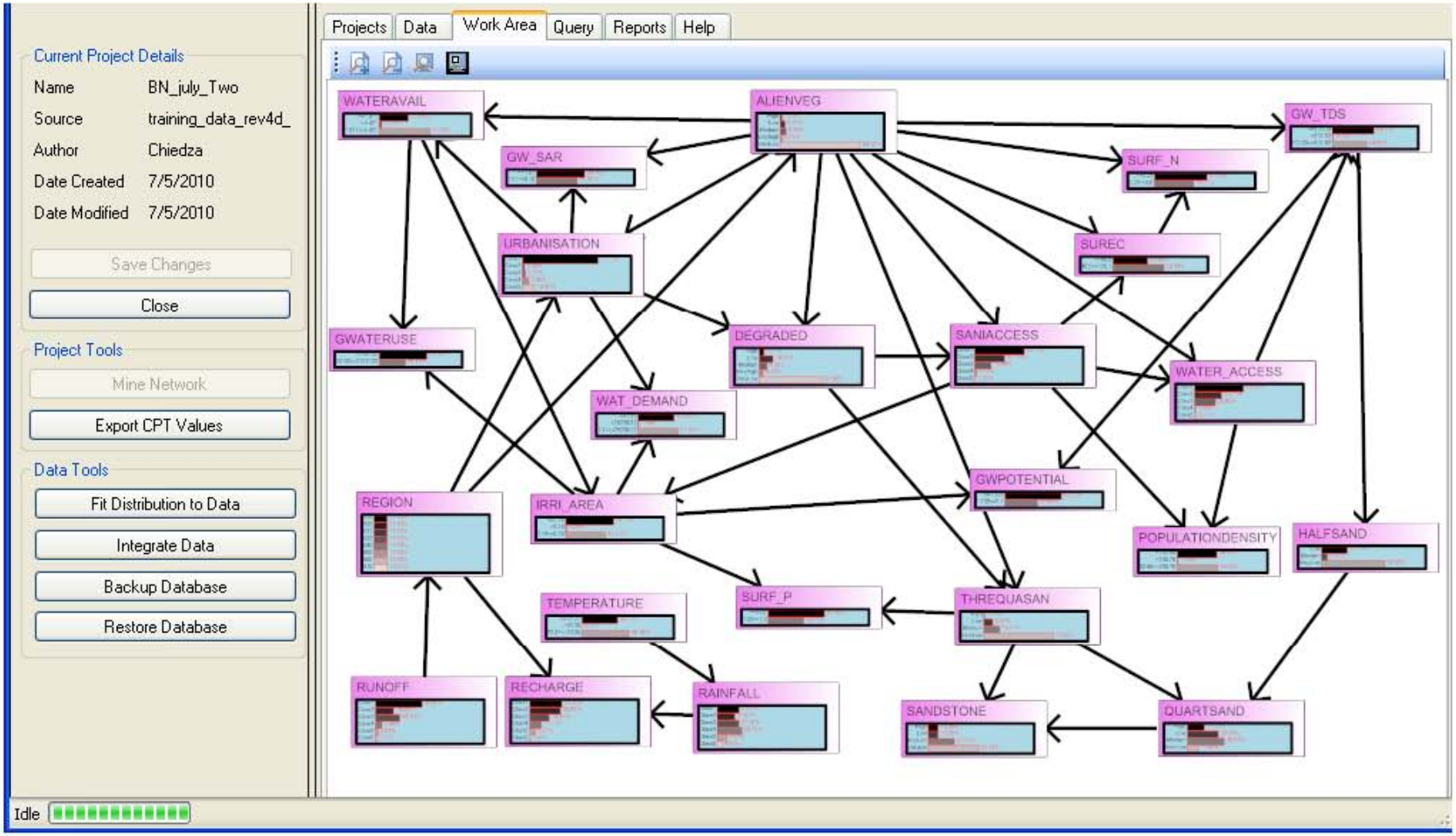
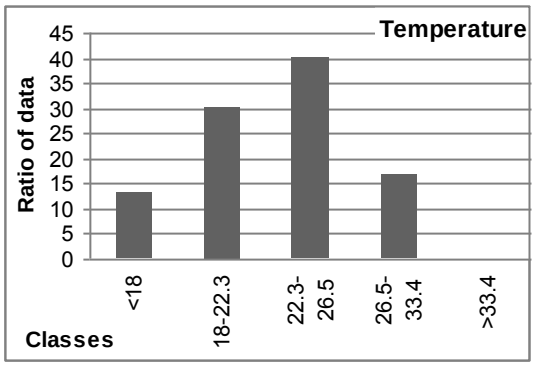
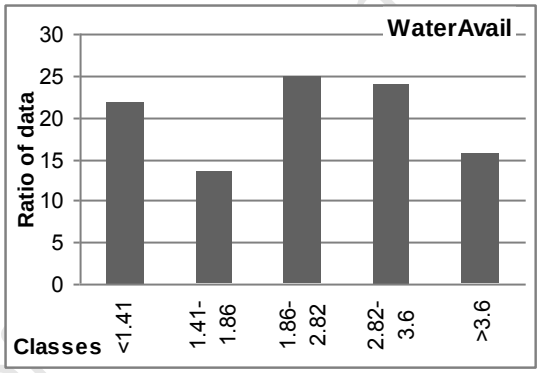


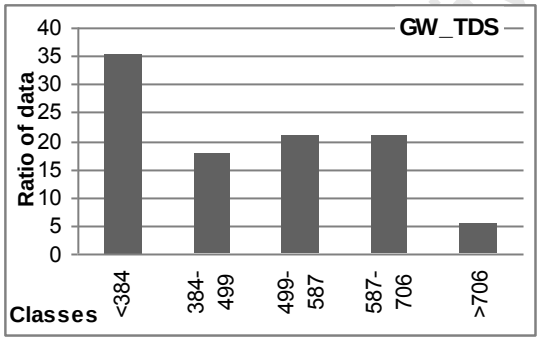
Figure 5- 26 : Bayesian Network Result B.

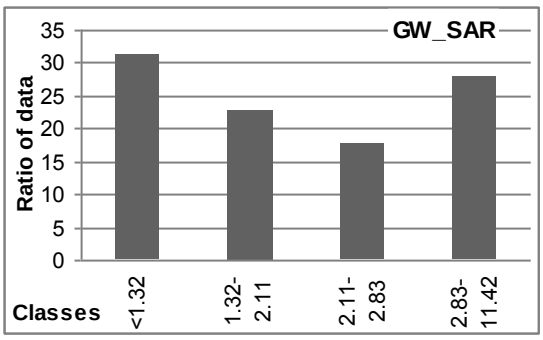
5.2.5.3 Result C with four discretisation intervals

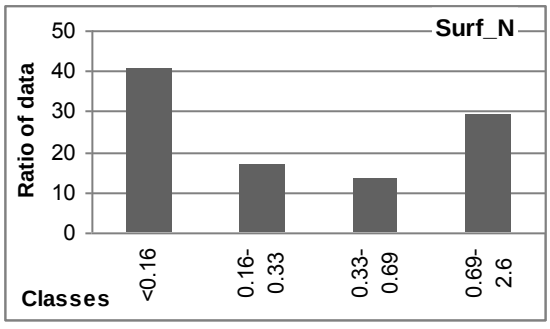
The variables were discretised using the equal binning method and setting the number of intervals $k_1 = 4$ and concurrently producing the network structure. The discretisation results are shown in the following graphs. The algorithm could not discretise some of the variables into four classes but instead five. Due to the fact that the algorithm simultaneously mines the structure as it discretises the data, this means that for these variable in order to get optimum patterns, five classes where the most ideal.

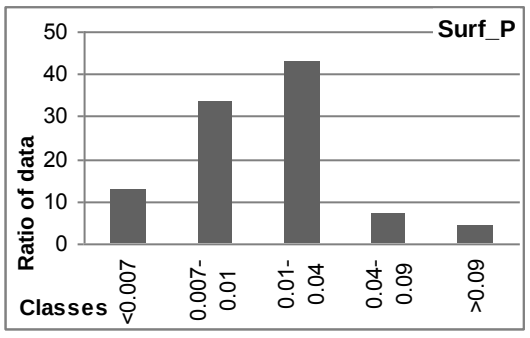


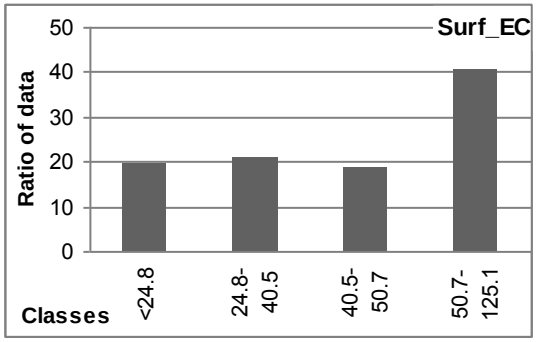


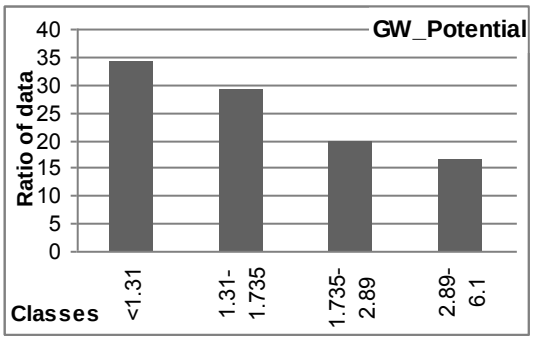


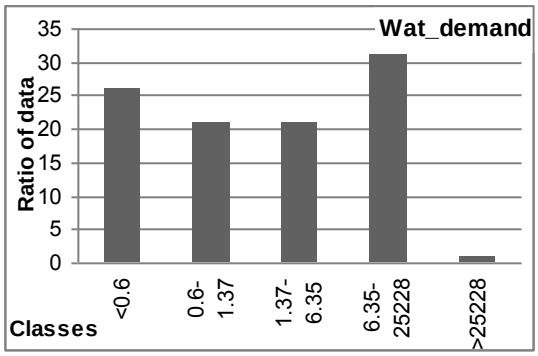


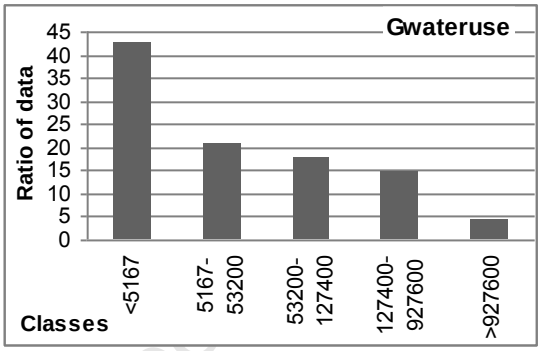


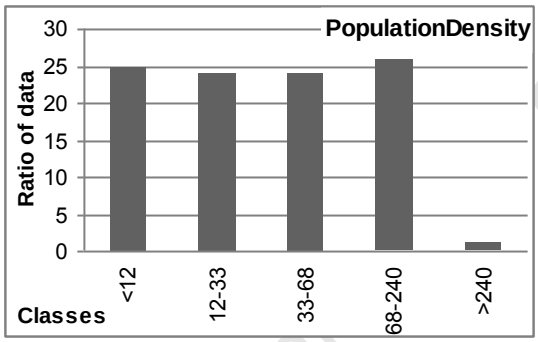


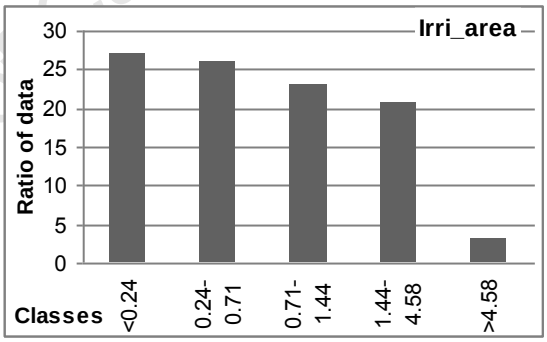












The results show that most of the variables were discretised using five variables. The only variables discretised using four intervals were groundwater SAR, surface water nitrogen and EC. The resulting network is shown in Figure 5- 27.

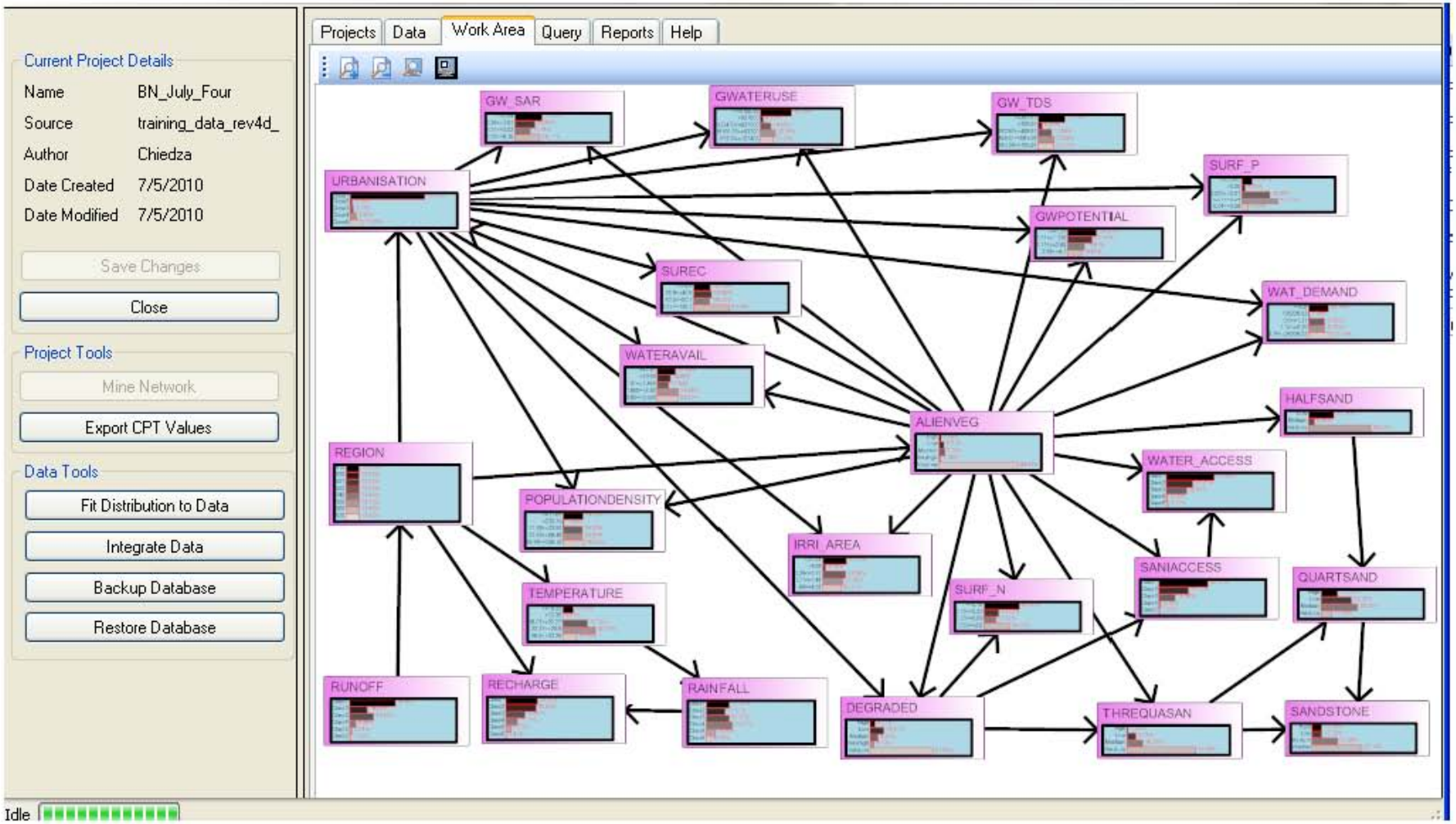


Figure 5- 27 : Bayesian Network-Result C.

5.2.4.4. Scoring the three networks

In order to select the appropriate discretisation levels and for the selection of the optimal network, the variables in the three networks were scored using the following measures:

i)The error rate;

ii)The logarithmic metric;

iii)The Brier score (quadratic loss); and

iv)The spherical score.

The error rate gives a percentage of cases when the values of the variables being tested were incorrectly predicted by the network. This means that the better the prediction, the lower this value would be. The logarithmic score values are calculated using the natural log and are between 0 and infinity, with zero indicating the best performance. The quadratic score ranges from 0 to 2 with 0 being indicative of the best performance and 2 being the worst. The spherical payoff is between 0 and 1, with 1 being best. The results obtained are shown in Figures 5- 28, 5- 29 5- 30 and 5 -31.

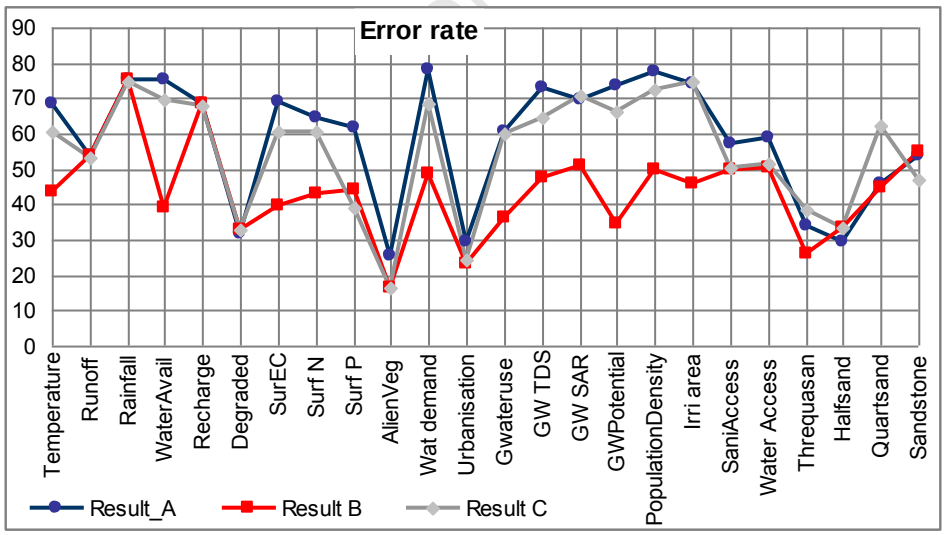


Figure 5- 28 : The calculated error rates for variables of the three networks.

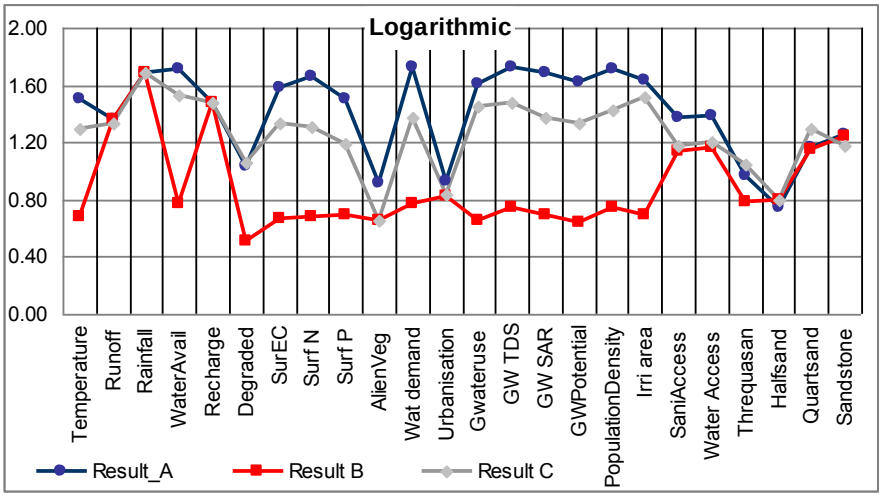


Figure 5- 29 : The logarithmic score calculated for variables of the three networks.

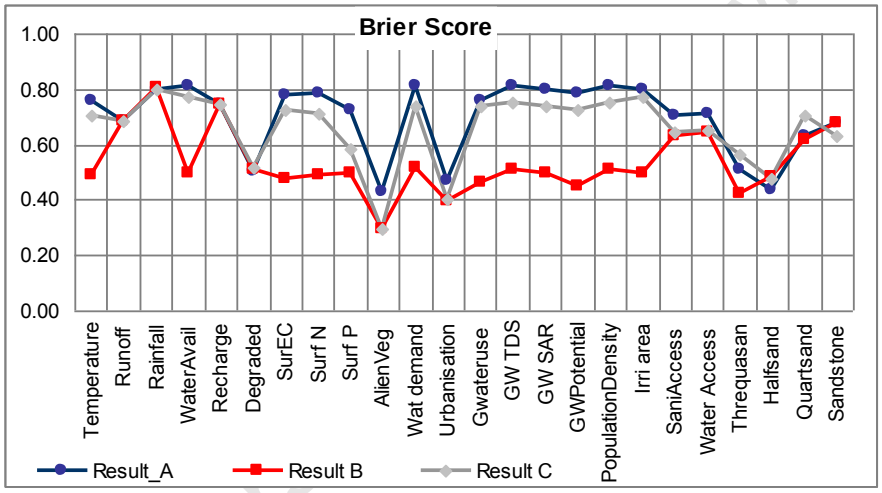


Figure 5- 30 : The Brier scores for variable in the three networks.

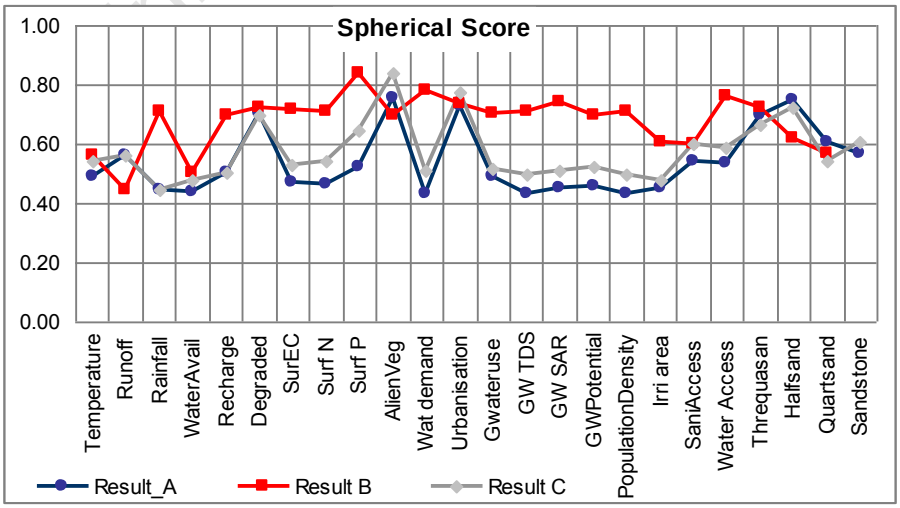


Figure 5- 31 : The spherical scores calculated for variables in the three networks.

The results of the error rate and the three scores show that network result B, with $k_1 = 2$ has the best performance of the three networks. It shows that using a small number of intervals in discretising the data improves predictability. The major disadvantage is that more complex relationships are not captured. The following section discusses the selection of the final discretisation levels and network.

5.2.6 The final variable states and Bayesian Network model

The scoring results showed that network result B had the best performance in terms of predictability. This means that the network was adapted and used as the "most optimum" network and the discretisation levels used were considered the most appropriate or relevant. The discretisation levels had to be modified slightly to accommodate knowledge from literature on the significance of the breakpoints (Uusitalo, 2007). The network also had to be modified to capture well-known relationships and relationships of interest to the study.

The variables listed in Table 5- 16 were discretised using the thresholds discussed in Section 5.2. These are rainfall, runoff, groundwater recharge, sandstone ratios, urbanisation, water and sanitation access, degraded land and alien vegetation.

Table 5- 16: The classes used to discretise some of the variables

Parameter	Range	Description
Temperature (°C)	<18	VeryLow
	18-22.3	Low
	22.3-26.5	Medium
	>26.5	High
Water available	<1.41	VeryLow
	1.41-1.97	Low
	1.97-3.6	Medium
	>3.6	High
Groundwater TDS	<380	VeryLow
	380-500	Low
	500-706	Medium
	>706	High

Parameter	Range	Description
Groundwater SAR	<0.94	VeryLow
	0.94-1.51	Low
	1.51-2.02	Medium
	>2.02	High
Surface water nitrogen (mg/ℓ)	<0.33	Low
	0.33-0.69	Medium
	>0.69	High
Surface water phosphorus (mg/ℓ)	<0.01	Low
	0.01-0.04	Medium
	>0.04	High
Surface water EC	<40.5	Low
	40.5-50.7	Medium
	>50.7	High
Groundwater exploitation potential	<1.74	Low
	1.74-2.89	Medium
	>2.89	High
Water demand	<1.37	VeryLow
	1.37-6.35	Low
	6.35-400	Medium
	>400	High
Groundwater use	<5 167	VeryLow
	5 167-53 200	Low
	53 200-127 400	Medium
	>127 400	High
Population density	<33	VeryLow
	33-68	Low
	68-240	Medium
	>240	High
Irrigated area	<0.74	VeryLow
	0.74-1.44	Low
	1.44-1.58	Medium
	>4.58	High

The final network for water resources assessment is shown in Figure 5- 32. The discussion of the results is provided in Chapter 6. The arrows show the relationships between the variables and the percentages next to each state of a variable are the probabilities or likelihoods of each state. The probabilities were produced from the input data by counting the number of instances the variable existed in each state (see Section 3.1.4).

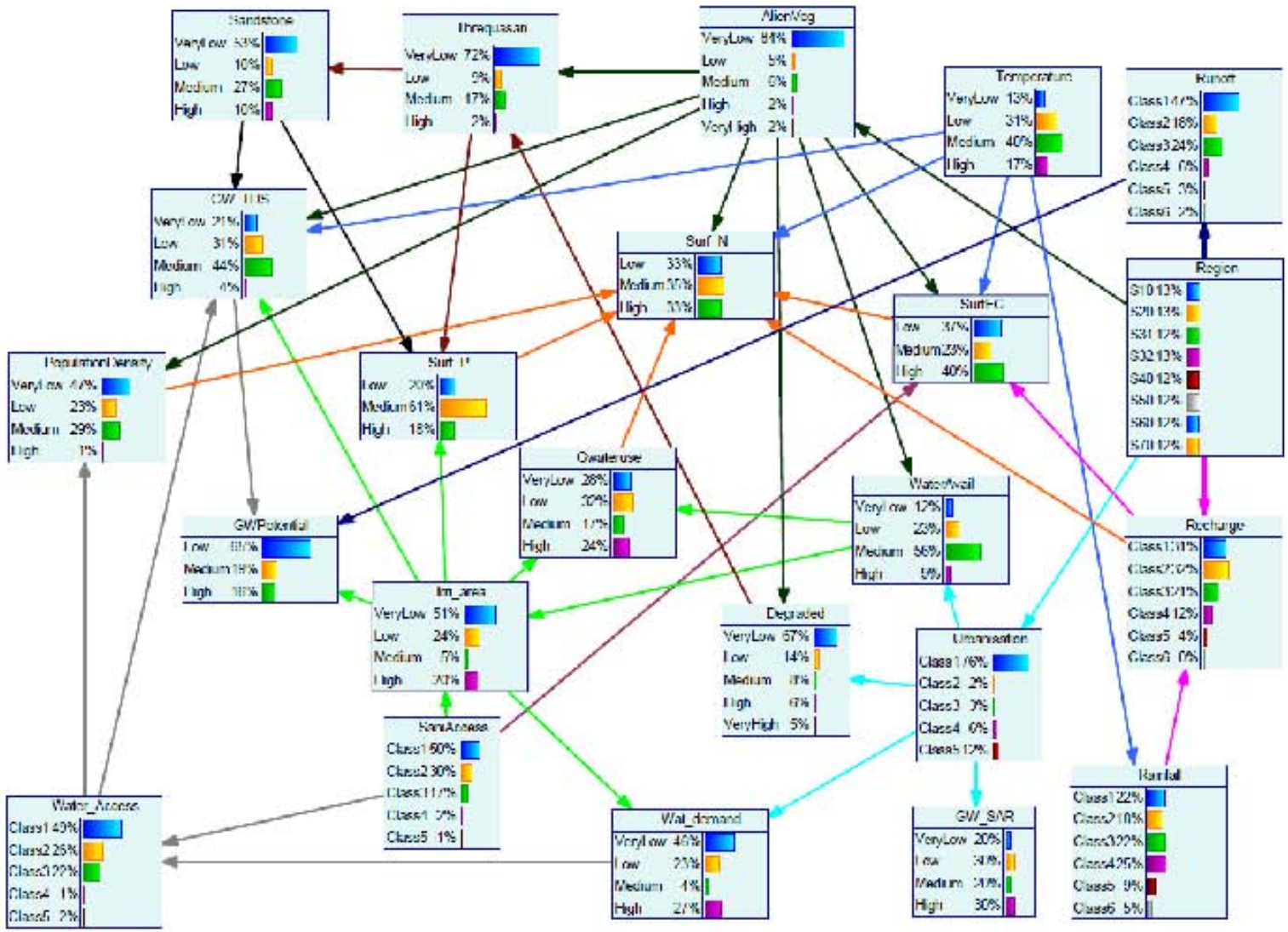


Figure 5- 32 : The final network for water resources assessment.

5.2.7 The Bayesian Network model for spatial variation in water quality

In catchment modelling, it is vital to model the spatial variation in some parameters for example water quality. Besides being able to model the effects of changes in the different selected variables in the entire catchment, there is also the need for more detailed analysis to evaluate the effects at the sub-catchment level. This involves assessing the effects of changes in the status of one sub-catchment on the neighbouring ones or modelling the results of changes in areas upstream to areas downstream of the catchment.

In this case, surface water EC values were assessed as an example. The approach used was proposed by Cofiño *et al.*, 2002 who utilised a spatial network of rainfall stations to predict the rainfall in neighbouring regions. In this research, monthly EC averages from the year 1980-2007 were used. In months were there were missing records, the yearly average was used. Figure 5- 34 shows the relationships between EC in the different regions of the study area as shown in Figure 5- 33. The relationships were automatically created using the custom developed Bayesian Network software.

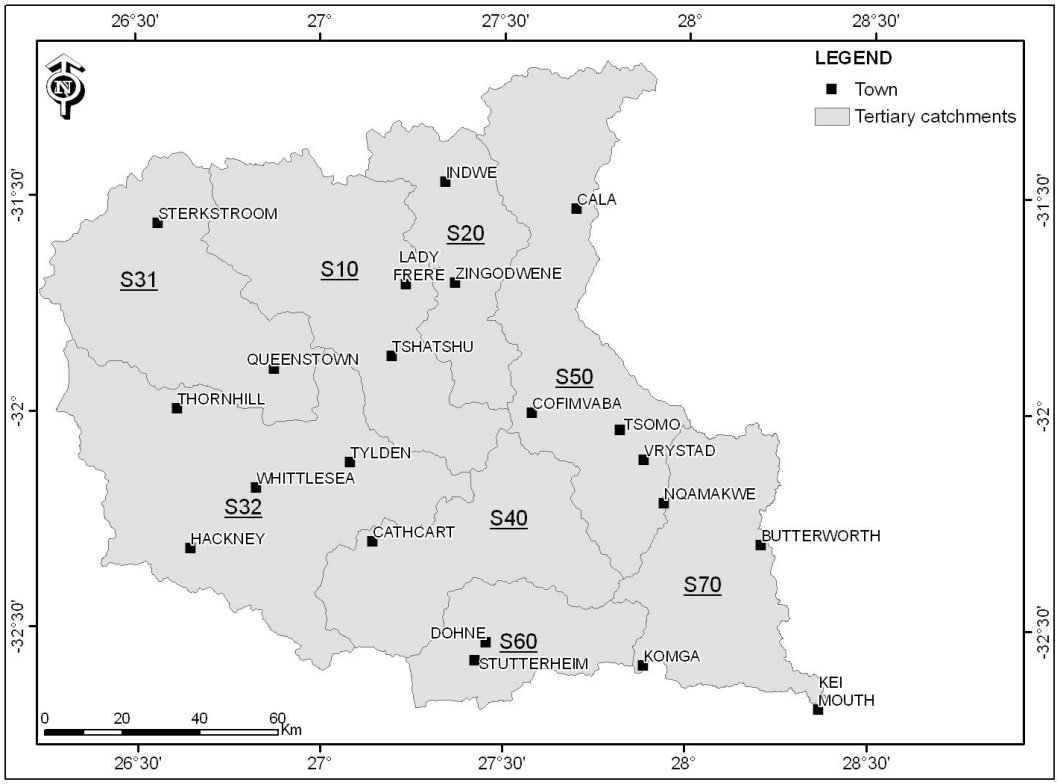


Figure 5- 33 : The modelling regions (tertiary catchments) in the study area.

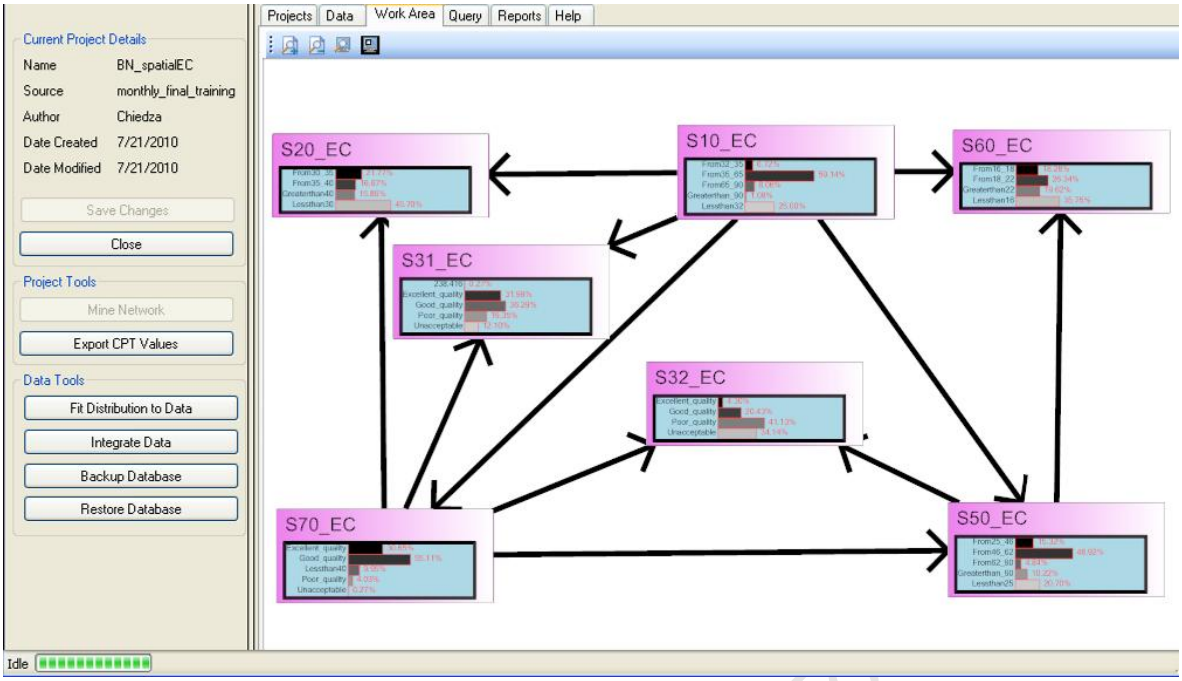


Figure 5- 34 : The network for analysing the spatial analysis of EC.

Pearson correlation values were calculated for the surface water EC from the different catchments and the results are shown in Table 5- 17.

Table 5- 17: The Pearson correlation matrix showing the correlation between EC values in the different regions

	S10_EC	S20_EC	S31_EC	S32_EC	S50_EC	S60_EC	S70_EC
S10_EC	1						
S20_EC	***0.36***	1					
S31_EC	0.15	0.06	1				
S32_EC	-0.10	-0.20	***0.43***	1			
S50_EC	0.08	0.00	-0.24	-0.08	1		
S60_EC	0.19	-0.03	0.11	0.14	0.18	1	
S70_EC	-0.03	-0.10	-0.18	0.03	***0.31***	0.12	1

The results show the effects of change in water quality in one catchment on the neighbouring catchments. It shows how the water quality upstream affects the quality of water in rivers downstream. For example, the most downstream catchment S70 is affected by catchments S10, S32, S40 and S50 which are upstream.

Pearson correlation statistics show strong correlation between S10 and S20; S31 and S32; S50 and S70. The network shown in Figure 5- 34 can be used in scenario testing or sensitivity analysis to assess the effect of changes in quality or pollution levels on neighbouring catchments. Examples of these analyses are provided in Chapter 6.

5.2.8 Temporal Bayesian Network modelling

The aim of this analysis is to illustrate the applicability of Dynamic Bayesian Networks in time-series modelling. With the availability of sufficient high quality data, some variables in the catchment can be modelled using time-series analysis. This can assist in the prediction of the future values of those variables based on past knowledge.

The Dynamic Bayesian Network time-series modelling is done using the relationship between rainfall, temperature, recharge, surface water EC and surface water nitrogen. The parameters (which are a subset of the network) were selected to illustrate the approach. The aim is to predict rainfall and temperature for a month in a year based on knowledge about the conditions of the same month in the previous years. The data used are monthly readings for the time period, 1950 to 1999.

The monthly data was input into the custom developed software and the relationships between the variables were mined automatically. Figure 5- 35 shows the structure of the model used and it represents three time steps (that is time = 0, which is January, time = 1 for February and time = 3 for March and so on). This means that for each time, rainfall or temperature will be predicted by varying the input of the other related variables in the preceding and current months. A more detailed discussion of the types of analysis performed and the results are provided in Section 6.6.

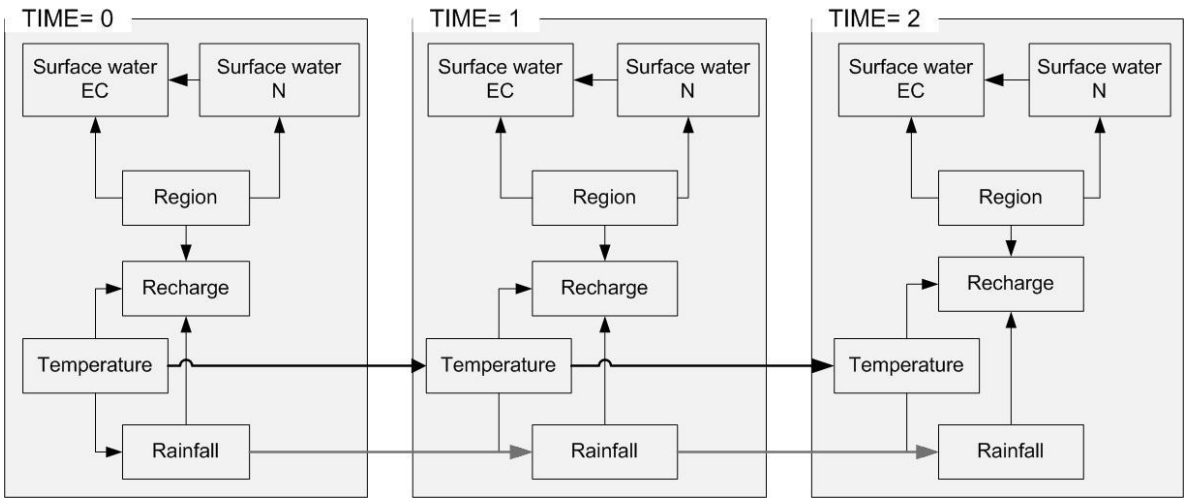


Figure 5- 35: Dynamic Bayesian Network for temporal modelling.

A more detailed discussion of the results obtained is provided in Section 6.6.

5.3 Conclusions

The chapter commenced with a section outlining the development and functions of the custom developed Bayesian Network software used in this research. The modelling methodology ensued with a discussion of the steps followed during the process. The study area was presented as well as the variables to be included in modelling. Using the custom developed software and the available data, three Bayesian Networks were created automatically by using three different methods for automatic discretisation of the continuous data. These three networks were evaluated using the error rate, the Brier score, the spherical score and the logarithmic loss and the most favourable network was selected as the final network.

Two other types of Bayesian Networks were presented, a spatial network for the prediction of surface water EC and a Dynamic Bayesian Network for the temporal prediction of rainfall and temperature. The following chapter discusses the results obtained using the different networks by applying the models to data from the selected study area.

CHAPTER SIX: PRESENTATION AND DISCUSSION OF RESULTS

6 Introduction

This chapter presents the results obtained from modelling in the study area. The outcomes of evaluating the Bayesian Network using sensitivity and scenario analyses are discussed. The results obtained from spatial prediction and time-series modelling are examined. The application of these analyses and results on water resources management are discussed.

6.2 Presentation of the results

The final network and the results obtained are illustrated in Figure 6- 1. The resulting network created by the custom developed software was duplicated in the Genie® software. Genie® was used because it has the capabilities for performing sensitivity analyses which the custom developed software lacks. The results are discussed by comparing them to the conceptual diagram shown in Figure 5- 15 and also by making references to literature sources. The relationship between the socio-economic indicators and the pollution variables shown in the conceptual diagram is also reflected in the results, which show a link between population density and surface water nitrogen concentrations. The relationship between the number of people with access to water (water access variable) and groundwater TDS is also another example of the link between socio-economic indicators and pollution variables.

The relationship between water balance and socio-economic variables is reflected in the results by the relationships between urbanisation and the volume of water available and the demand for water resources. Pollution and resource condition are related as is indicated by the link between alien vegetation and groundwater TDS, surface water nitrogen concentration and surface water phosphorus concentrations.

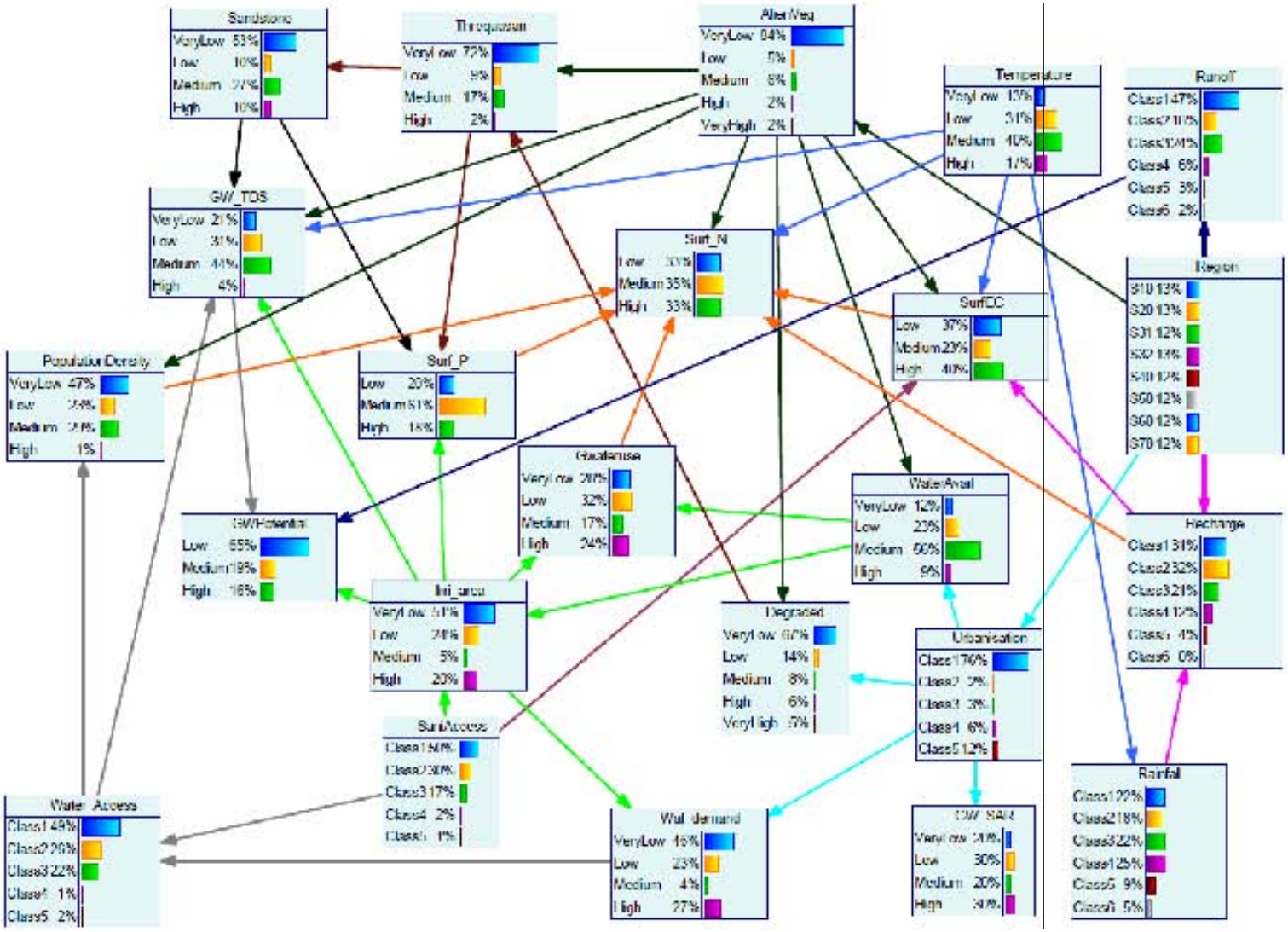


Figure 6- 1: Final Bayesian Network model and results.

The linkage between resource condition and water balance is reflected in the results which show relationships between the irrigate area in each region to groundwater TDS and surface water nitrogen concentration. There are also relationships between alien vegetation and surface water nitrogen and phosphorus concentrations.

The relations between alien vegetation and the related variables are not useful due to the unavailability of enough variation in the data used for the alien vegetation variable. All tertiary catchments, except S40 have very low occurrences of alien vegetation. The alien vegetation variable was included in the network because its relationship with the other variables might be applicable and tested in another catchment with more data or when more detailed analyses are performed at a finer scale.

Runoff is not directly related to rainfall as expected, they are related spatially through the region (which represents the tertiary catchment). Perhaps if the discretisation levels or the spatial and temporal scale change, a different pattern will emerge. Most of the catchments have low runoff except catchments S50, S60 and S70 where the highest probable class of runoff is Class 3. The reason being that these areas receive proportionally higher rainfall than the other catchments (see Figure 5 -8).

The rest of the relations are examined in detail in the following sections using the evaluation techniques outlined in Section 6.2.

6.3 Evaluation of the network

As was discussed in Chapter 3, after the network is developed and the probabilities are populated, it has to be evaluated. The first type of evaluation should involve domain experts to assess the following (Korb and Nicholson, 2004):

i) if all the variables used are relevant and if the ranges are exhaustive;
ii) if the variables are named and defined appropriately;
iii) whether or not the ranges used for discretisation are appropriate; and
iv) if there is consistency in the states of the different variables; and
v) if the probabilities are consistent with the data.

The second type of assessment is scenario testing. Scenario testing is important as it enables the investigation of the behaviour of the model for different expert scenarios. It assesses whether the model behaves as expected in light of past experiences and also whether it performs according to current research (Bednarski, *et al.,* 2004). Scenario testing is discussed in Section 6.2.1.

Another form of evaluation is sensitivity analysis. The sensitivity of the network is usually evaluated based on the scenarios being modelled. Two types of sensitivity analysis are used in this research. The first is a general analysis of how sensitive the network or the probabilities of the query nodes[14] are to changes in the inputs/evidence values (evidence sensitivity). The second type assesses how sensitive the probabilities of the query nodes are to changes in the parameters (parameter sensitivity) (Korb and Nicholson, 2004; Jensen, 2001; Bednarski *et al.*, 2004).

Scenario testing and sensitivity analyses are investigated in the following sections relative to the following scenarios:

i) variations in urbanisation;
ii) variations in surface water EC and groundwater TDS
iii) changes in irrigated areas; and
iv) changes in water demand.

These scenarios were selected because; according to studies done in the area, these are some of the main issues currently being faced and are also future challenges. The changes in probabilities for the related variables are explored under each case. The convention to be used in this discussion is that change is considered to be "significant" is the probabilities change by at least 15% otherwise it is negligible.

[14] The node (or variable) that is being interrogated during modelling.

6.3.1 The effects of variations in urbanisation

The aim is to assess the effects of changes in urbanisation on the other variables in the catchment. The anticipated trend in the study area is the increase in rural to urban migration due to economic reasons (DWA, 2004). An analysis of variations in urbanisation is vital in determining the impacts of this anticipated increase on water use, quality and availability. The results can be then used in decision-making for future resource management planning.

The effects are explored by considering the following

i) the base case, which is represented by the probabilities obtained from data analysis, that is the initial probabilities which are averaged over the entire study area;
ii) probabilities when urbanisation is very low or low; and
iii) probabilities when urbanisation ranges from medium to high.

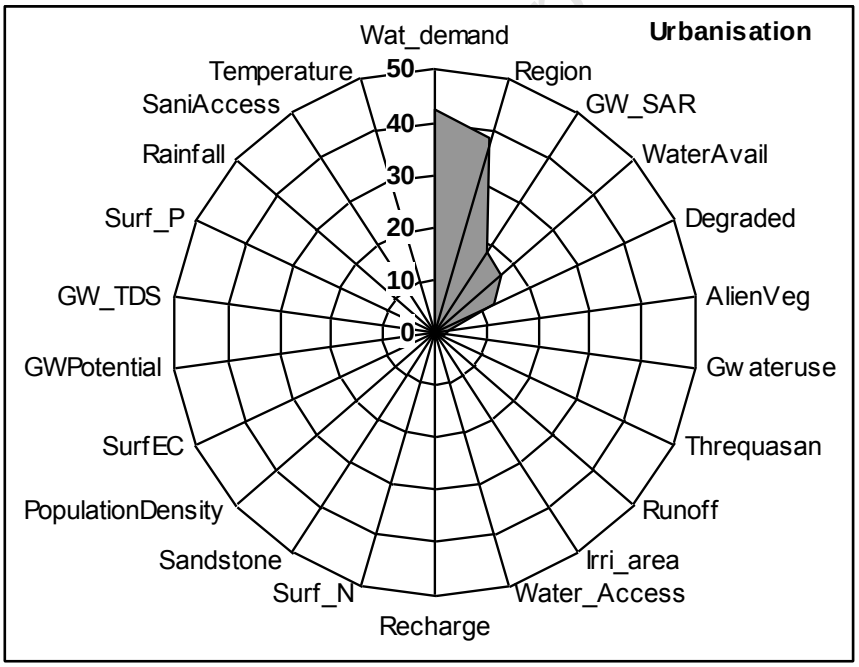


Figure 6- 2: Results of sensitivity analysis for urbanisation.

The results of sensitivity analysis for the urbanisation variable show that it mainly affects water demand, region, groundwater SAR, water availability and degradation (see Figure 6-2). The variable affected the most is water demand. The strong effect of region reflects the regional variations of urbanisation in the study area.

The regional variations of the different levels of urbanisation in the catchments are shown on the map in Figure 6- 3. The highly urbanised catchments (those with greater that 50% of the population in urban areas) are S20, S31, S40 and S60. S10 and S70 have the least levels of urbanisation as they show high prevalence of Class 1 urbanisation. Tertiary catchments S10, S50 and S70 are in the former Transkei homelands and there are mostly dispersed rural village settlements were urbanisation is low (see Figure 5- 9).

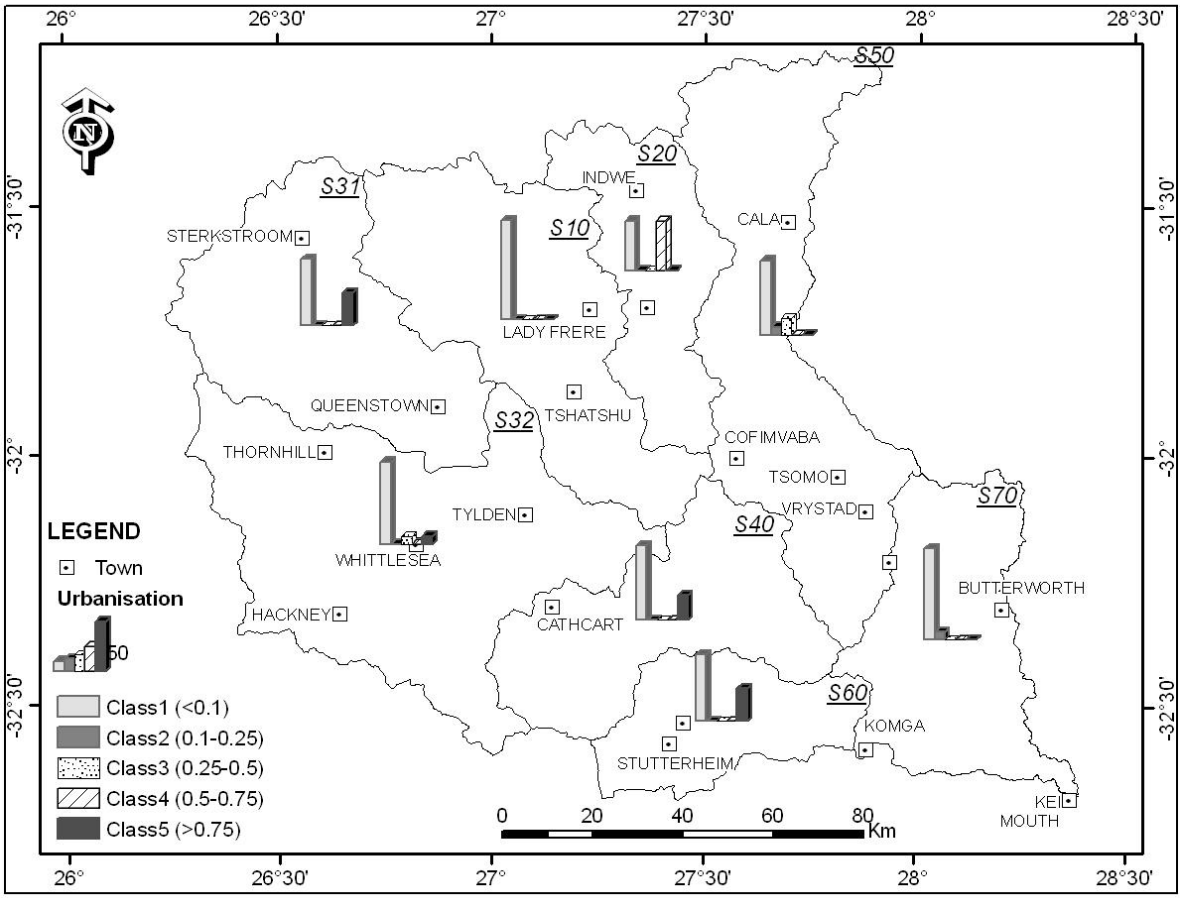


Figure 6- 3: Regional variations of urbanisation classes.

The effects of variations in urbanisation on water demand and water availability are shown in the graphs in Figure 6- 4.

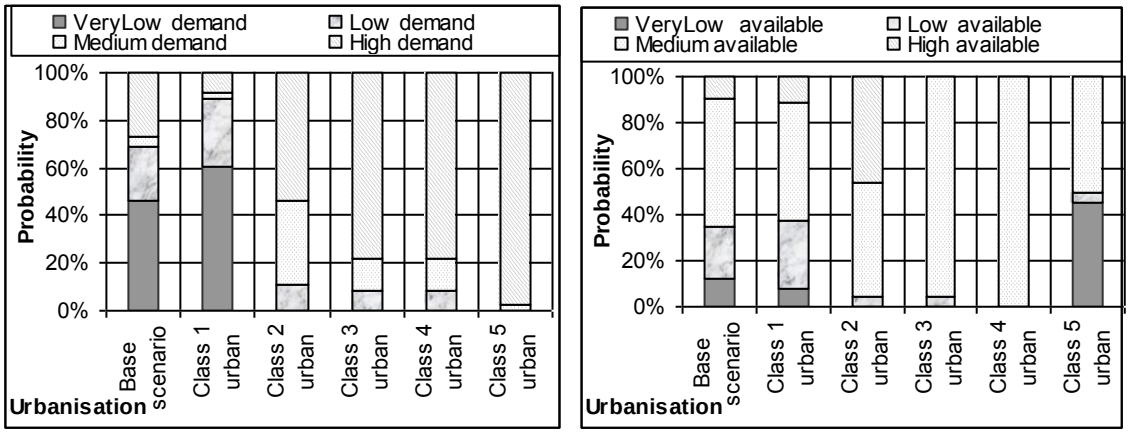


Figure 6- 4: Effects of changes in urbanisation on water demand (graph on the left) and water available (graph on the right).

In terms of water demand, the results show that the more urbanised the area becomes, the greater the demand for water resources. The graph on the left in Figure 6- 4 shows an increase in the probability of high demand; with increases in urbanisation levels. Class 1 urbanisation (the lowest) shows "very low" demand for water resources. This has implications for future water demand when the levels of urbanisation increase due to rural–urban migration.

The water available is the amount of available surface and groundwater resources per annum. The general trend seems to be that the higher the urbanisation level, the more the amount of water available/allocated. The outlier is the Class 5 urbanisation state, which shows approximately equal proportions of verylow/low and medium states of water availability. Class 5 urbanisation is most prevalent in S31, S40 and S60. This means that in these areas, they might be problems in terms of water availability if urbanisation increases. S31 will be the most affected as it is an irrigation area and is in a low rainfall region (see Figure 5- 8).

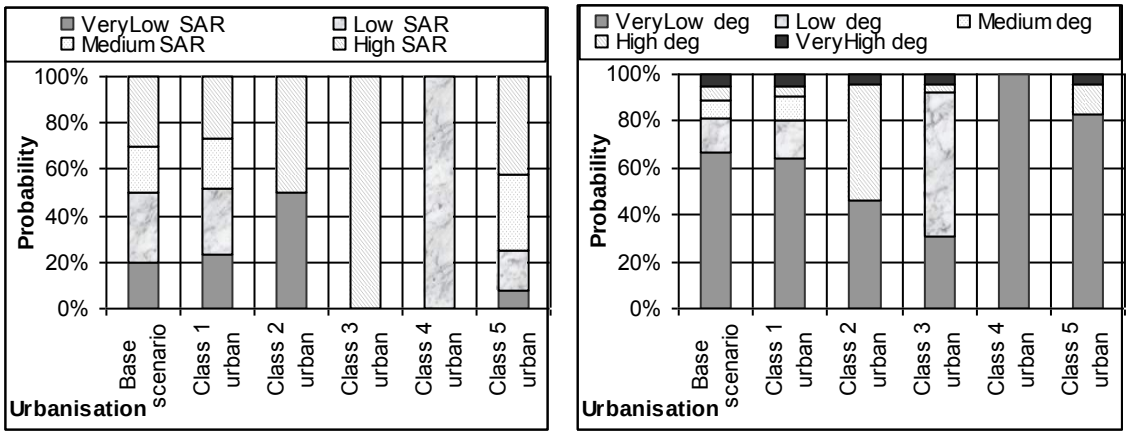


Figure 6- 5: Effects of changes in urbanisation on groundwater SAR (graph on the left) and land degradation (graph on the right).

According to the graphs in Figure 6 -5, Class 1 (the lowest) urbanisation is prevalent in all catchments hence there are approximately equal probabilities for all states of groundwater sodium adsorption ratio (SAR). Class 2 urbanisation in most likely in catchment S50 and they are equal proportions of very low and high SAR values. The high SAR values can be attributed to the nature of the settlements in the area. S50, the mostly informal rural area has the second lowest proportion of people with access to adequate sanitation, with the first being catchment S20.

This relationship between lack of sanitation and poor water quality is proposed in literature. The hypothesis is that runoff from informal settlements with poor access to sanitation can impact on groundwater quality. The results obtained in this research support this hypothesis although more data are required to substantiate them and the effects of other potential causes like geology need to be isolated.

Considering the graph on the right in Figure 6– 5, it can be concluded that there is no real trend between degradation and urbanisation except that it reflects the regional variations in degradation which are similar to those in urbanisation. Land degradation is highest in catchments S50 and S70 (where Class 2 or low urbanisation is most prevalent). These are areas characterised by dispersed rural village settlements and communal subsistence farming and grazing in the former Transkei. These poor land practices over the years have lead to severe land degradation.

6.3.2 Changes in groundwater TDS

The results of sensitivity analysis for the groundwater TDS variable are shown in Figure 6-6.

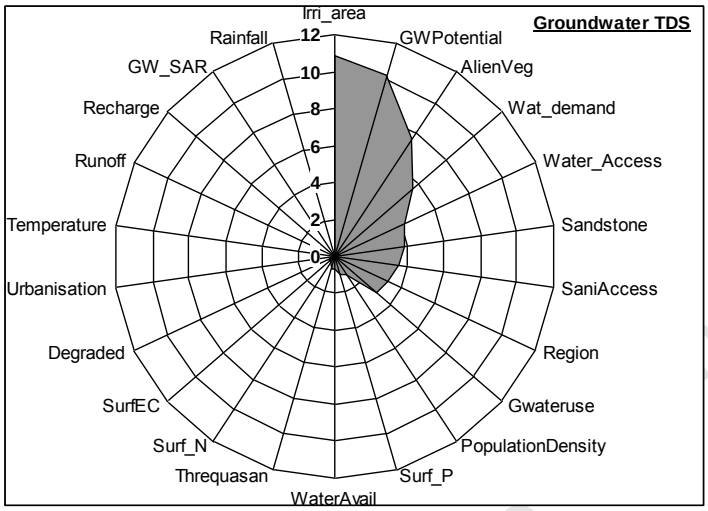


Figure 6- 6: The results of sensitivity analysis of groundwater TDS.

The results show that groundwater TDS is mainly affected by the irrigated area, groundwater potential, alien vegetation, water demand, water access, sandstone and sanitation access (see Figure 6- 6). The effects of changes in irrigated areas on groundwater TDS are discussed in Section 6.2.4. The outcomes of the variations in groundwater potential on groundwater TDS are shown in the graphs in Figure 6- 7.

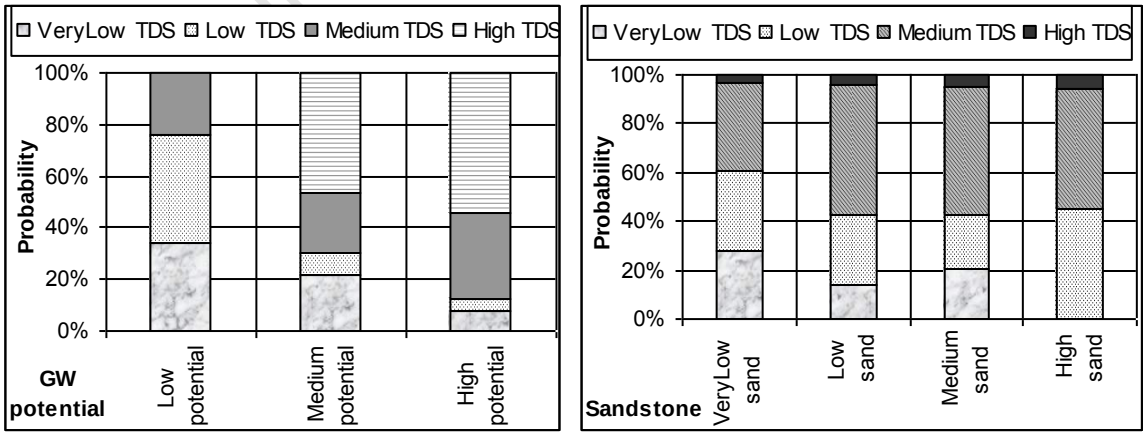


Figure 6- 7: The results of variations of groundwater TDS with groundwater potential (graph on the left) and the relationship between groundwater TDS and sandstone ratio (graph on the right).

Figure 6- 7 shows an increase in groundwater TDS values with increases in groundwater potential. The groundwater potential takes into account recharge which is mainly dependent on the geology. From literature, the geology of the area, made up of shale, mudstone and sandstones has been known to be a cause of the high salinity values (DWA, 2004). This is reflected on the graph on the right of Figure 6- 7 which shows a reduction in the probability of "very low" TDS with increase in the sandstone ratio of the geology. Such findings, although they do not bring any knew information serve as a test for the conformance of the model with existing knowledge.

Analyses of the results between alien vegetation and groundwater TDS shows no apparent pattern. The problem is the distribution of alien vegetation which is mostly very low in all the catchments. There is not enough variation in the data to establish a trend. Most probably by changing the spatial scale of analysis and the discretisation levels, more information on the relationship can be obtained.

The outcomes of variations in water demand and water access on groundwater TDS are shown in the graphs in Figure 6- 8.

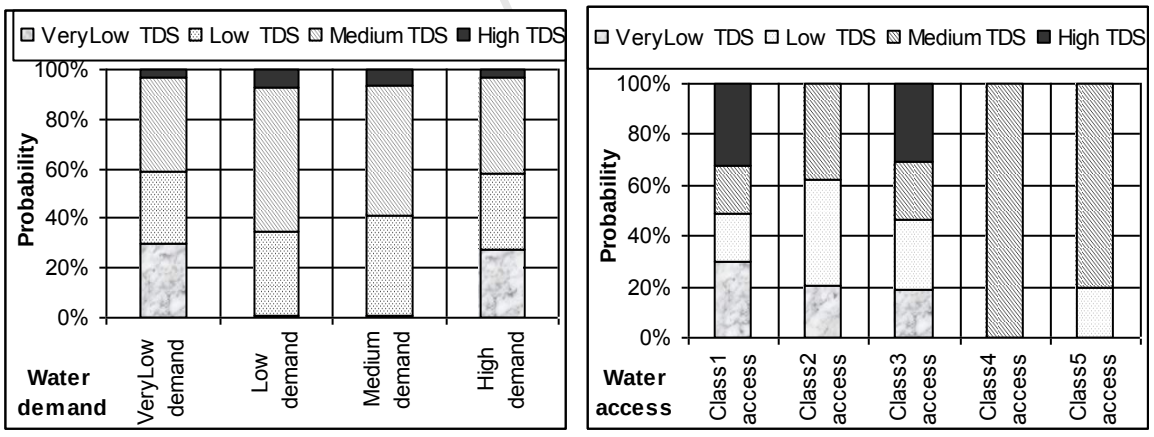


Figure 6- 8: Effects of changes in groundwater TDS on water demand and water access.

Water demand is the volume of water used and includes domestic, agricultural, commercial and industrial values. High demand areas like urban areas are expected to have high TDS levels due to the runoffs from sewage treatment plants, unlicensed solid waste sites, urban runoff and poorly designed and maintained sewerage systems.

Likewise, rural areas were the demand for water is low can also have elevated TDS levels due to runoff from informal settlements with inadequate sanitation. From the graph to the left in Figure 6- 8 this trend is reflected as it shows equal likelihoods of the different classes of TDS for "very low water demand" and "low water demand".

In terms of water access, the general trend is that the greater the proportions of households with access to adequate water, the higher the groundwater TDS levels (see the graph on the right in Figure 6- 8). This supports the hypothesis that urban areas, where most of the people have adequate water resources, are affected by pollution from runoff and sewerage. This is shown by a drop in the likelihood of "very low" groundwater TDS by as much as 50% when moving from Class 1 (<10% of the households) to Class 3 (25-50% of the households) of adequate water access. This type of analysis is useful in assessing the impacts of provision of water resources on the water quality in a catchment.

6.3.3 Variations in surface water EC

The results from analysis show that the surface water EC concentrations are mostly affected by sanitation access, alien vegetation, region, water access and recharge (Figure 6- 9).

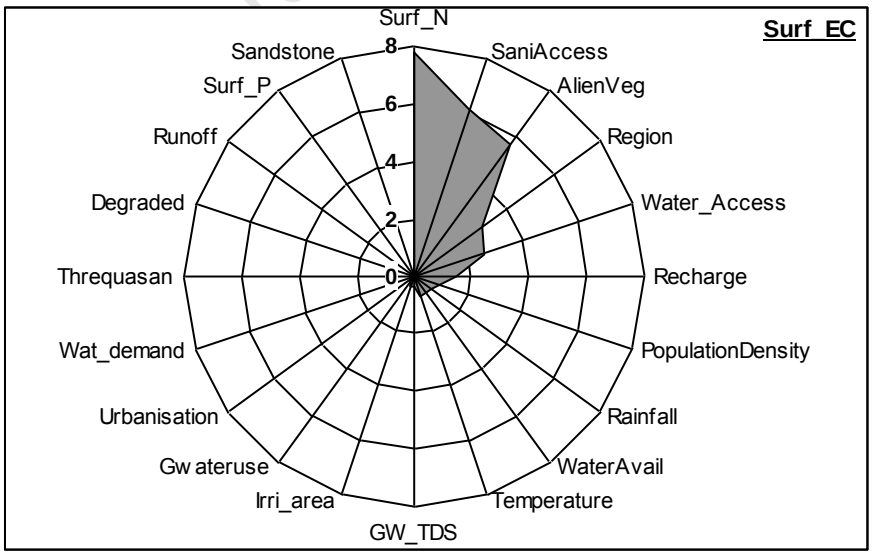


Figure 6- 9: The results of sensitivity analysis on surface water EC.

The variations of surface water EC with surface water nitrogen levels and sanitation access are shown in the graphs in Figure 6- 10.

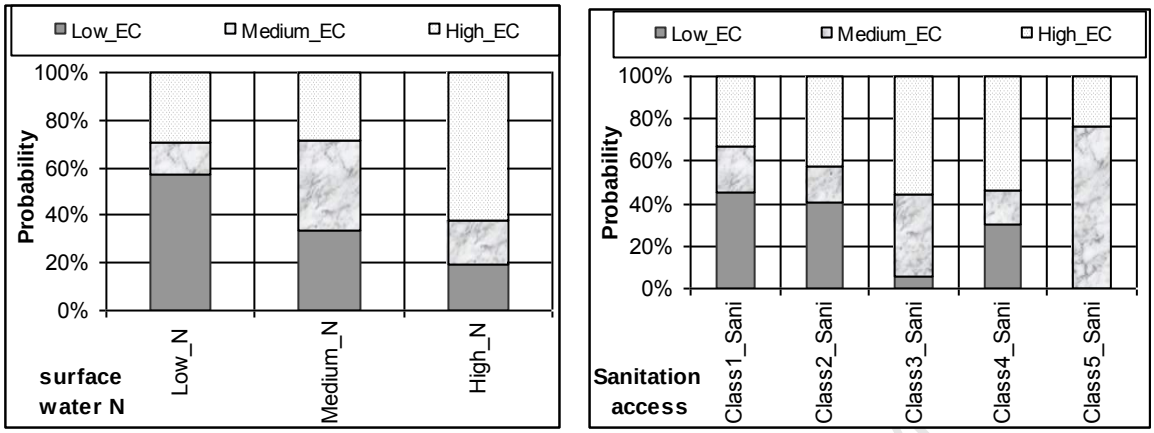


Figure 6- 10: The effects of variations in surface water nitrogen values on surface water EC (graph on the left) and the effect of changes in sanitation access on surface water EC (graph on the right).

Areas in the catchment with low nitrogen levels also have low EC values as illustrated on the graph on the left in Figure 6– 10. Increases in nitrogen levels result in increases in the probability of occurrence of high EC values. Significant decreases of the probability of low EC with increases in nitrogen increases (about 20% going from the low to high states) suffice. This relationship is as expected from theory and can be used to assess the effects of non-point source pollution, mainly from agricultural activities on EC. Where nitrogen levels are high and remediation/management options are planned to reduce the values, the effects of such options on surface water EC can be evaluated and quantified.

The relationship between sanitation access and surface water EC shows that the higher the proportion of households provided with adequate water services, the poor the surface water quality (Figure 6- 10). This is illustrated by the drop in the probability of low EC values by about 37% when sanitation access increases from Class 1 (10% of households served) to Class 5 (>65% of the households served). Similar results are obtained when assessing the relationship between water access and surface water EC. These findings substantiate the results discussed in Section 6.2.2 which show high groundwater TDS with increases in water access. The explanation of this trend is the same as that provided for groundwater TDS, that is; households with high sanitation and water access are mostly in the urban areas where pollution of water from runoff, stormwater and sewerage works is rife.

Analyses of the results between alien vegetation and surface water EC show no pattern. The problem is the distribution of alien vegetation which is mostly very low in all the catchments. There is not enough variation in the data to establish a trend.

The sensitivity of surface water EC to the region variable represents the spatial variation of EC (see map in Figure 6- 11). The general trend is that the catchments which are largely irrigated, for example S31, S32 and S40 have high EC values, due to non-point source pollution from agricultural activities. The high EC values in catchment S10 are most likely caused by runoff from informal settlements with inadequate sanitation. These findings have to be verified with more accurate detailed data before any firm conclusions can be drawn.

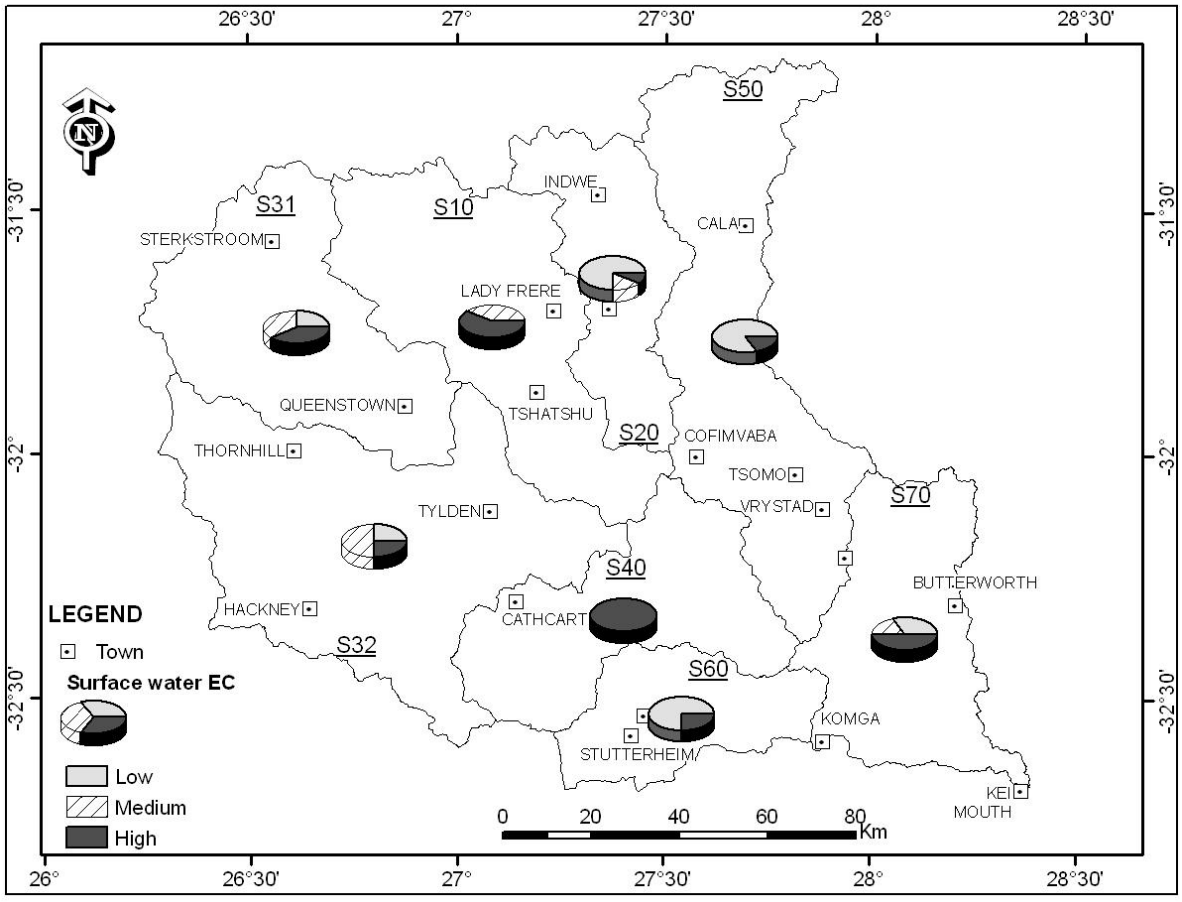


Figure 6- 11: The regional spatial variations of surface water EC.

In terms of recharge (Figure 6– 12), the general trend is that the higher the recharge (which is directly related to rainfall), the better the water quality.

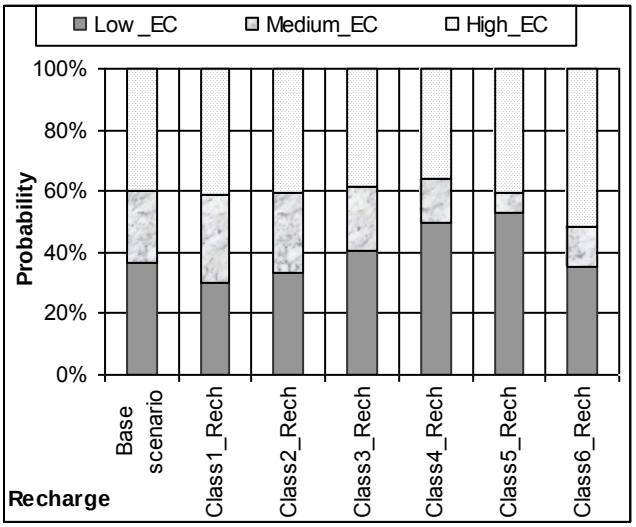


Figure 6- 12: Effects of changes in recharge on surface water EC.

This trend is illustrated by the increase in the probability of low EC with increases in recharge, significant changes by 23% from Class1 (low) to Class 5 (high) recharge are realised. The only finding that opposes this hypothesis is the one for Class 6 recharge. The likely explanation is that this class is mostly prevalent in tertiary catchments S60 and S70. In these areas other factors contribute to the elevated levels, in S60 the most likely cause is pollution from agriculture and in S70, runoff from informal settlements with inadequate sanitation might be affecting EC levels.

6.3.4 Changes in irrigated areas

The results from analysis show that the irrigated areas affect groundwater use, groundwater TDS, water availability and surface water P levels. Irrigated areas are also related to sanitation and water access and groundwater potential. Measures of the sensitivity of "irrigated area" to changes in these variables are shown in Figure 6- 13.

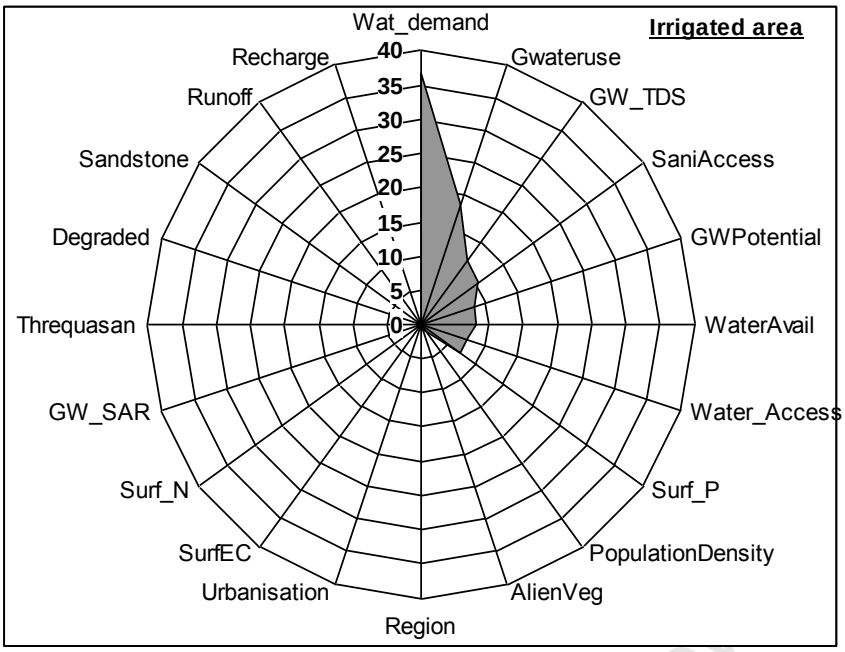


Figure 6- 13: Results of sensitivity analysis on the "irrigated area" variable.

The relationship between water demand and irrigated area is discussed in Section 6.2.5. The relationships with sanitation and water access reveal the spatial variation of irrigation. Most irrigated areas are not urbanised therefore the levels of access are low. The scenarios of the effects of changes in irrigated areas over the groundwater use and TDS are illustrated in the graphs in Figure 6- 14.

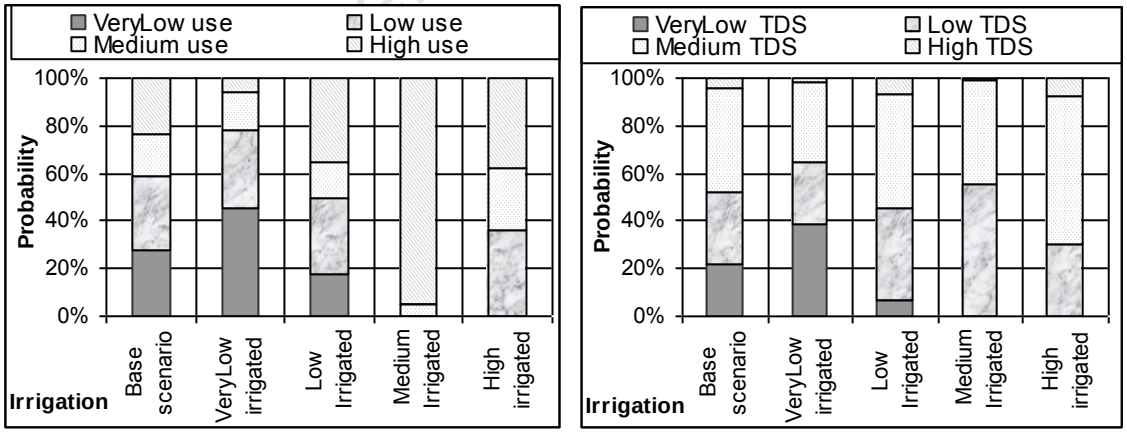


Figure 6- 14: The relationship between irrigated areas and groundwater use (graph on the left) and the link between irrigated areas and groundwater TDS (graph on the right).

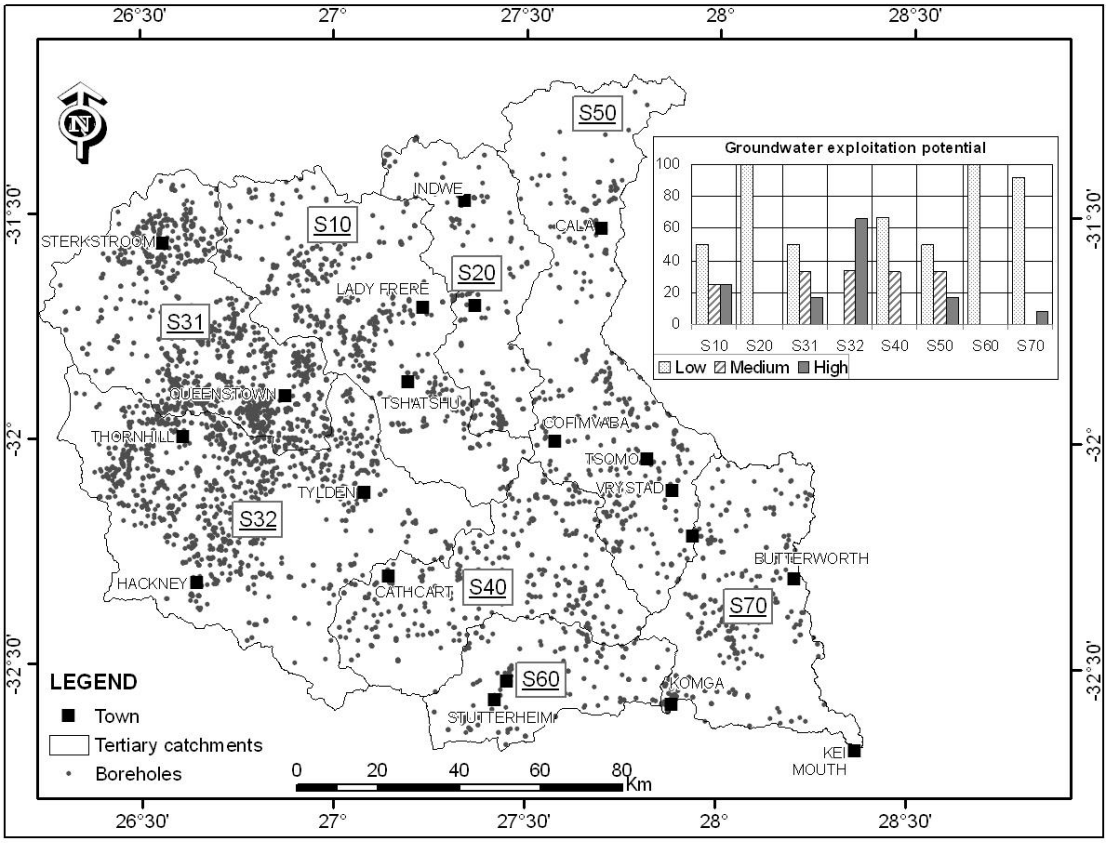


Figure 6- 15: Boreholes drilled in the study area and a graph showing the distribution of the groundwater exploitation potential.

Largely irrigated areas have high groundwater exploitation potential (see Figure 6- 15). This can be due to the fact that the potential is calculated using borehole data. In most irrigated areas, monitoring and groundwater use has been ongoing for a number of years so there is a wealth of knowledge on groundwater characteristics. This brings a bias as these areas are assigned high values when the exploitation potential is calculated. This is illustrated by Figure 6- 15 which shows the distribution of boreholes in relation to the groundwater exploitation potential in the study area.

The scenarios of the effects of changes in irrigated areas on the surface water P are illustrated in the graphs in Figure 6- 16.

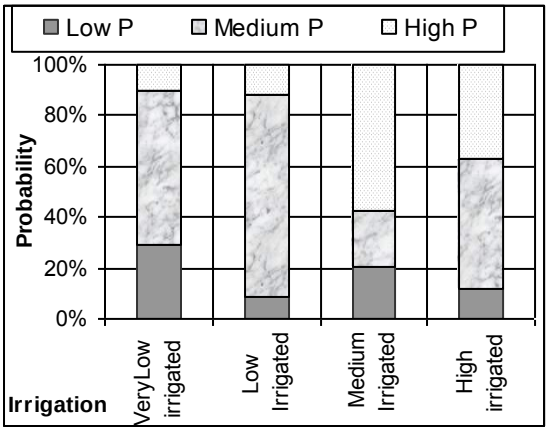


Figure 6- 16: The effects of irrigation on surface water P concentration.

Highly irrigated areas have high surface water P levels. This conforms with theory from literature that agricultural sources contribute to phosphorus in water through their use of fertilisers. Such a relationship can be used to assess and quantify the effects of changes in irrigation activities on the phosphorus levels in surface water.

6.3.5 Variations in water demand

The sensitivities of the water demand variable to changes in the other variables in the network are shown in Figure 6- 17. It shows that water demand is affected by irrigation, urbanisation and groundwater use.

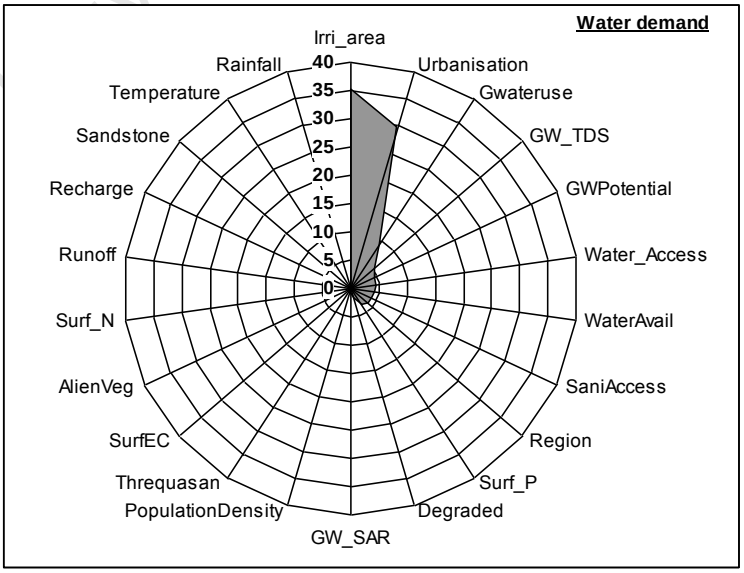


Figure 6- 17: Results of sensitivity analysis of water demand.

The scenarios investigated are to assess the combined effects of irrigation and urbanisation on water demand. The example illustrates the power of Bayesian Networks in their ability to assess the effect of a multitude of parameters on a query variable simultaneously. The two scenarios are:

i) Very low irrigation and Classes 1 and 5 urbanisation on water demand; and
ii) High irrigation and Classes 1 and 5 urbanisation on water demand.

The results are shown in the graphs in Figure 6- 18.

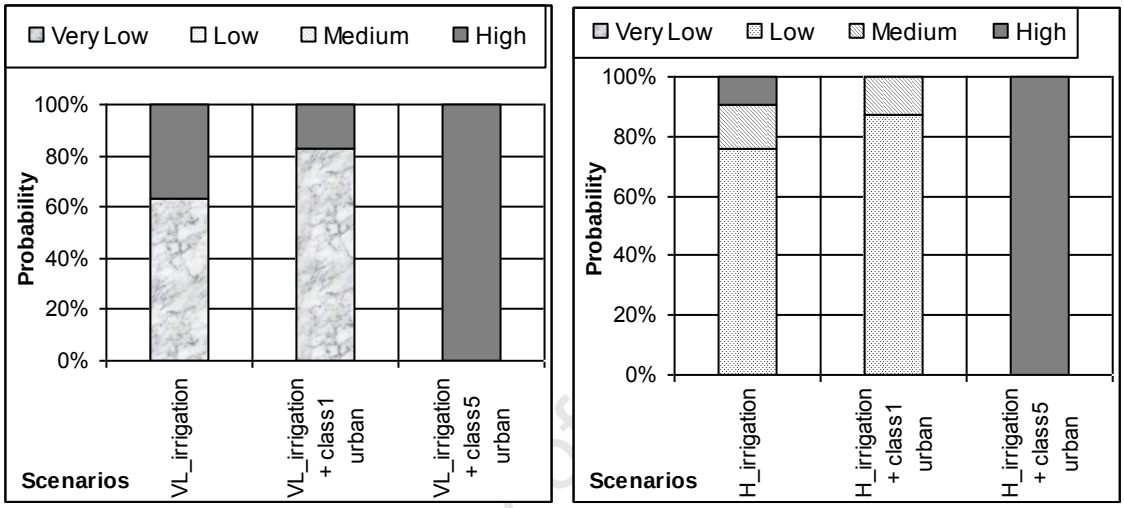


Figure 6- 18: Combined effects of urbanisation and irrigation on water demand.

The graph on the right of Figure 6- 18 shows that when low irrigation is considered separately, there is a 65% probability for "very low" water demand and 25% for "high" water demand. These probabilities change considerable when urbanisation is also taken into account. With the combined effect of very low irrigation and urbanisation, the likehood for "low" water demand increases by 17% to 82%. With "very low" irrigation and high urbanisation, "high" water demand is the most likely scenario.

The graph on the left of Figure 6- 18 shows that with high irrigation only, there is a 70% probability for "low" water demand. These probabilities change considerable when urbanisation is also taken into account. With the combined effect of very low irrigation and urbanisation, the likehood for "low" water demand increases by 17% to 82%. With "high" irrigation and high urbanisation, "high" water demand is the most likely scenario

The results show that when the combined effects of irrigation and urbanisation on water demand are assessed, urbanisation has a greater effect than irrigation. Since water demand is the volume of water used by all water-use sectors including domestic, mining, agriculture, commercial and industrial, such an analysis can be used to monitor the changes in water demand when the water use pattern by all water users change.

6.3.6 "What if" scenario analysis

"What if" scenarios are used to assess the effects to all variables in the network when values of some variables change. In the example being provided in this section, the following scenario is investigated:

"*what happens to all variables in catchment/region S32 if the population density and urbanisation become high and low rainfall is prevalent?"*

The effects are shown in Figures 6- 19 and 6 -20. Figure 6- 19 is the original network created from data and Figure 6- 20 is the result after the "what if" scenario is performed.

The results show that with increases in population density and urbanisation and a decrease in rainfall in catchment S32, the following are likely effects:

i) an increase in water demand; under the "what if" query it is 100% likely to be high;
ii) a deterioration in surface water quality as indicated by the increase in the likelihoods of higher values of surface water phosphorus and nitrogen;
iii) the groundwater sodium adsorption ratio and TDS are likely to increase showing a decline in the groundwater quality; and
iv) a decrease in the amount of water available as shown by the 20% increase in the likelihood of "verylow" for the "water available" variable from 12% in the base scenario to 32% under the "what if" scenario.

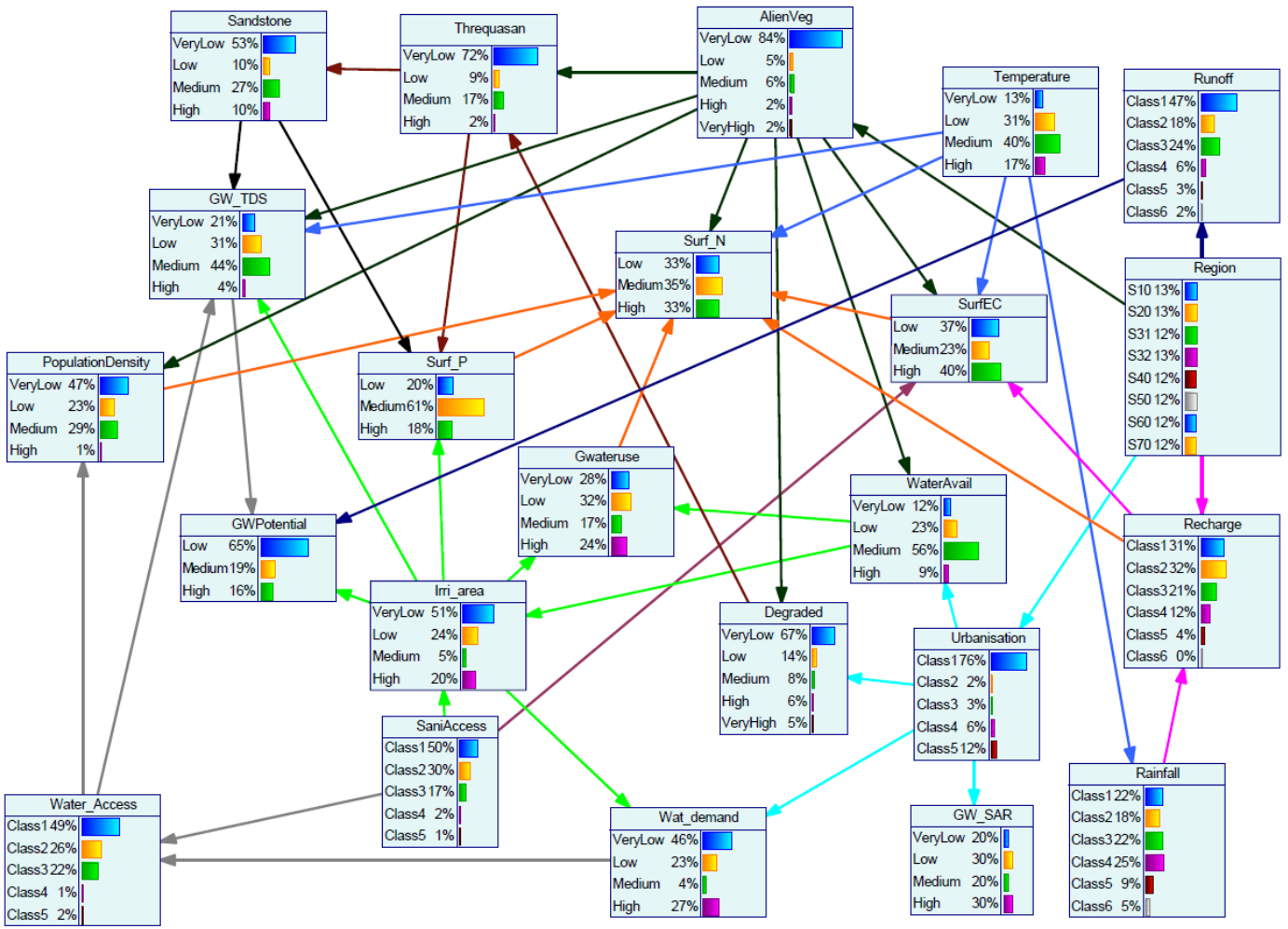


Figure 6- 19: The original Bayesian Network for data analysis.

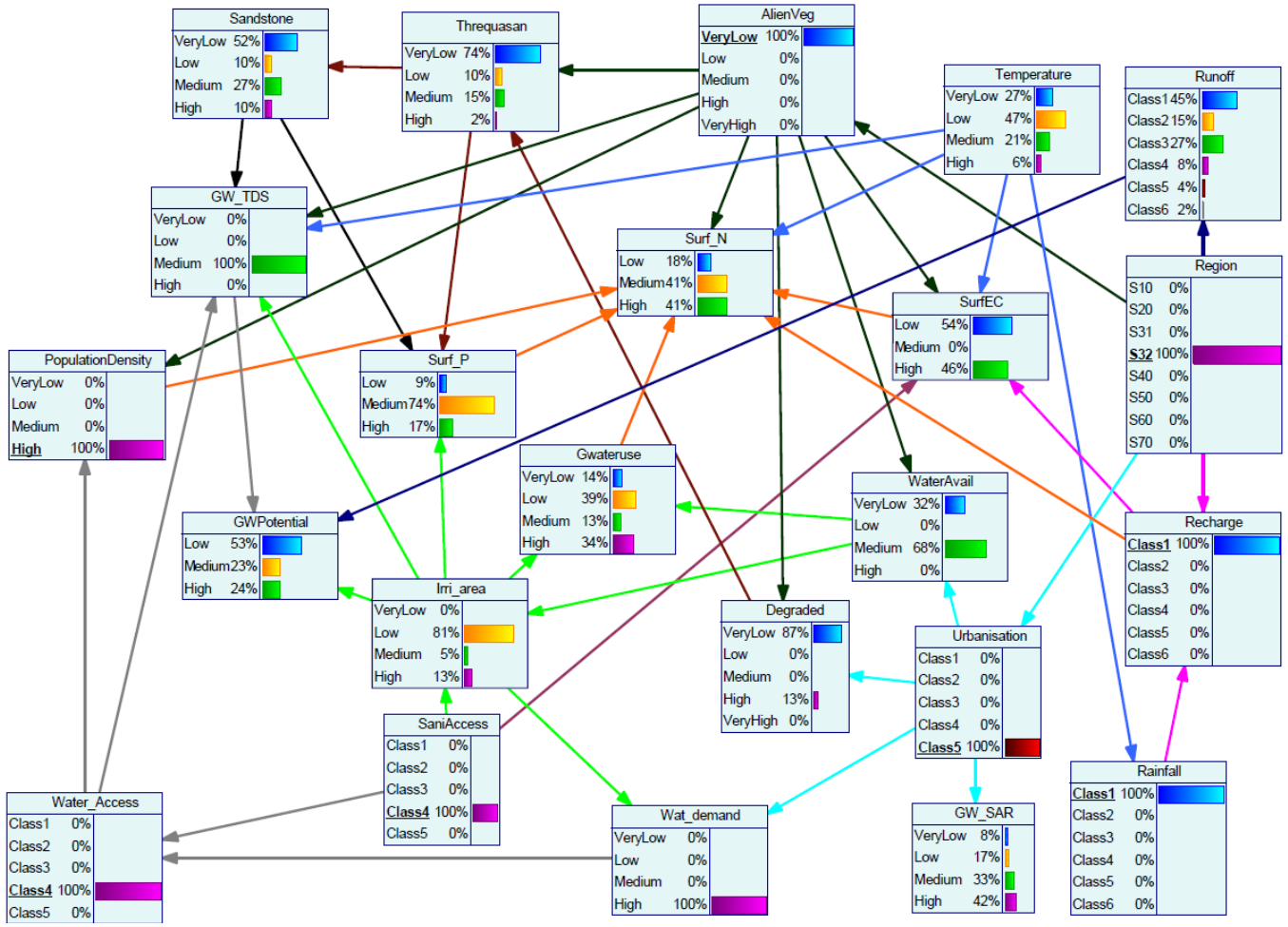


Figure 6- 20: The resulting Bayesian Network after performance of the "what if" scenario analysis.

6.4 Hypothesis testing: spatial prediction of EC

The network shown in Figure 6- 21 (see Section 5.2.7 on how it was created) can be used for predicting surface water EC values for a catchment using the "known/observed" values of its neighbouring catchments. The schematic diagram of the network was drawn using the GeNie® software.

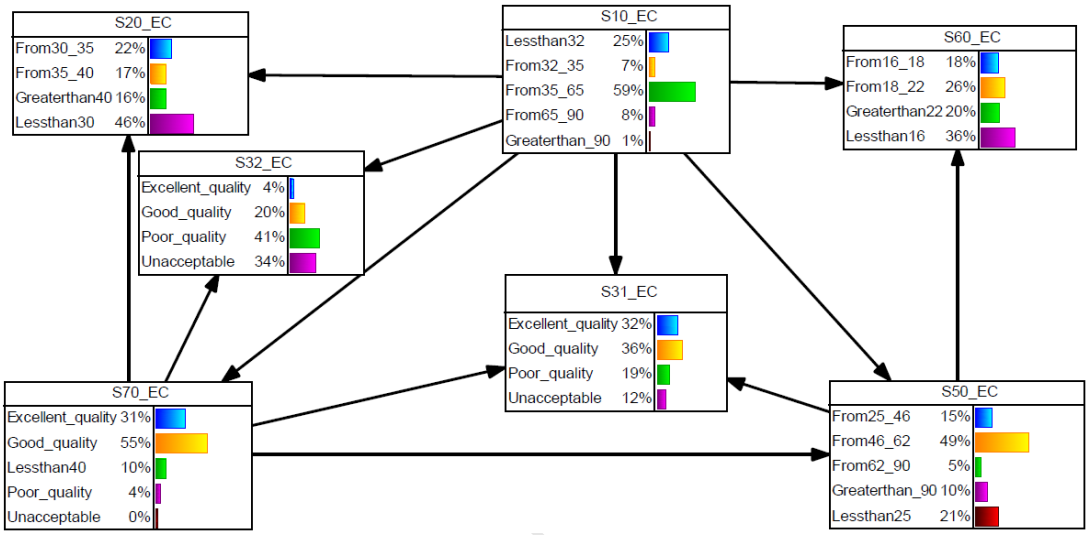


Figure 6- 21: The resulting network for spatial EC prediction.

Table 6- 1: Shows the discretisation of the EC values for the different tertiary catchments

Region	Range	Classes
S10	<32	Lessthan32
	32-35	From32_35
	35-65	From35_65
	65-90	From65_90
	>90	Greaterthan_90
S20	<30	Lessthan30
	30-35	From30_35
	35-40	From35_40
	>40	Greaterthan40
S31	>90	Excellent_quality
	90-270	Good_quality
	270-540	Poor_quality
	>540	Unacceptable
S32	>90	Excellent_quality
	90-270	Good_quality
	270-540	Poor_quality
	>540	Unacceptable

Region	Range	Classes
S50	<25	Lessthan25
	25-46	From25_46
	46-62	From46_62
	62-90	From62_90
	>90	Greaterthan_90
S60	<16	Lessthan16
	16-18	From16_18
	18-22	From18_22
	>22	Greaterthan22
S70	<40	Lessthan40
	40-90	Excellent_quality
	90-270	Good_quality
	270-540	Poor_quality
	>540	Unacceptable

Tertiary catchment S40 was not included because there was not enough data (see Appendix C which shows the EC data used in analysis). S10 was used, although it also has sparse data. S31 and S32 have the worst water quality compared to the other catchments, with some values in the poor and unacceptable ranges. This can be attributed to the fact that these are the two catchments which contain the highest areas of irrigated lands. Catchments S10, S20, S50 and S60 have lower EC values than rest of the catchments.

The aim of the study was to assess the effects of changes in the EC values of catchments upstream on the most downstream catchment, S70. A sensitivity analysis on S70 shows that the EC values are sensitive to changes in EC values in catchments S50, S32, S10, S31 and S20 in descending order of influence (Figure 6- 22).

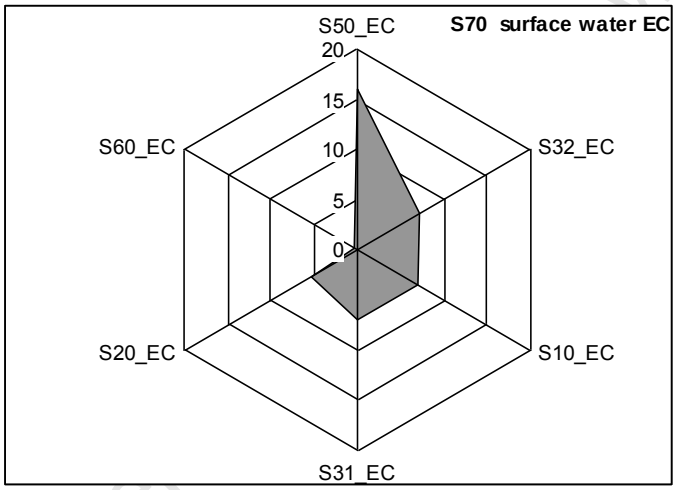


Figure 6- 22: Sensitivity analysis on catchment S70.

The hypothesis being tested is the following:

"increases in EC values in upstream catchments also cause an increase in values for downstream catchments"

As was highlighted in Section 6.2, evidence sensitivity analysis can be used to support a postulated hypothesis. It will be used in this case to test the hypothesis postulated above. Tertiary catchment S70 is the "query variable" and catchments S50, S32, S10, S31 and S20 are the catchments whose evidence will be varied to assess whether or not they support the hypothesis.

This example also illustrates the diagnostic capabilities of Bayesian Networks. Diagnostic inference is determining the probabilities of causes from given effects. The effects in this scenario are low and high values of surface water EC in catchment S70. The results of inference are shown in the Figures below.

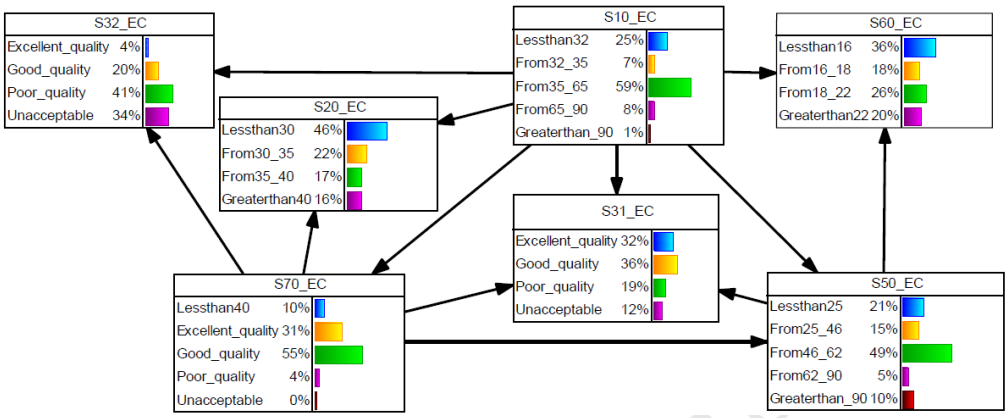


Figure 6- 23: The base scenario which represents good water quality in S70.

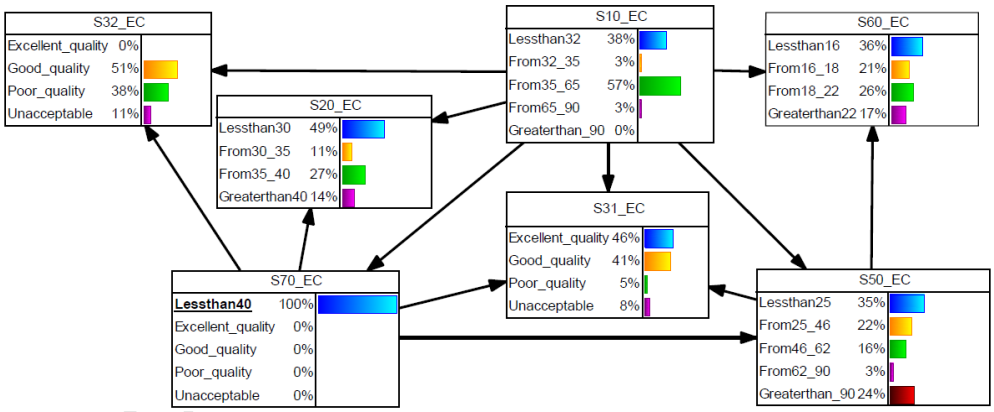


Figure 6- 24: The causes of very low EC values (less than 40 mS/m) at S70.

By comparing the scenario shown in Figure 6- 24 and the base scenario, it is evident that EC values less than 40 mS/m in S70, correspond to low values in upstream catchments S10, S20, S50, S31 and S32. This is reflected in the increase in the probabilities of occurrence of the low states of EC in these upstream catchments. This scenario supports the postulated hypothesis and indicates that water quality evidence at the upstream catchments can be used to make inferences about water quality at downstream catchments. This, of course can only be applicable after other factors that influence water quality from within the query catchment are eliminated.

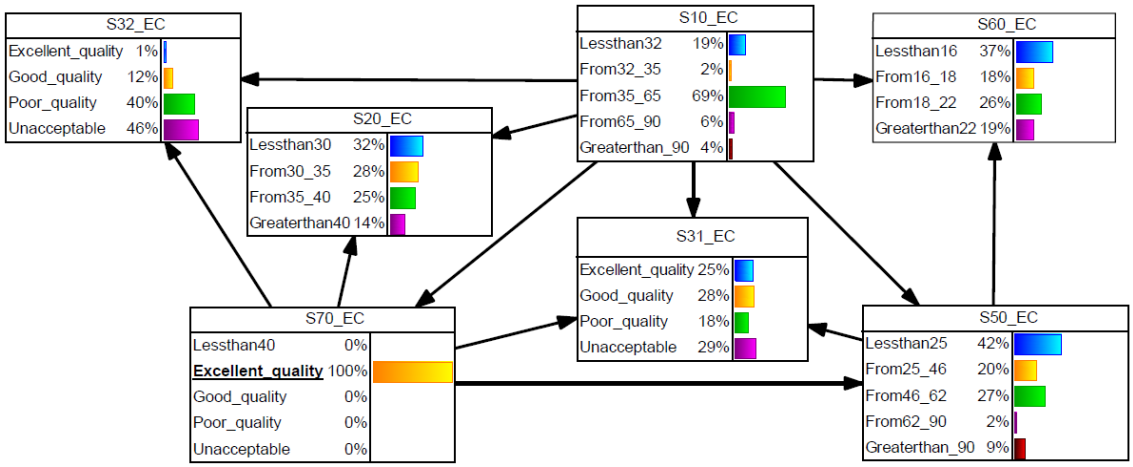


Figure 6- 25: Analysis for excellent water quality in catchment S70.

Comparing the scenario shown (Figure 6- 25) and the base scenario (Figure 6- 23) shows that "excellent water quality" is related to low values at points in rivers in catchments S10, S20, S50 and S31 to some extent. Findings in S32 do not support the hypothesis as they relate "excellent quality" in S70 to "unacceptable" water quality in S32.

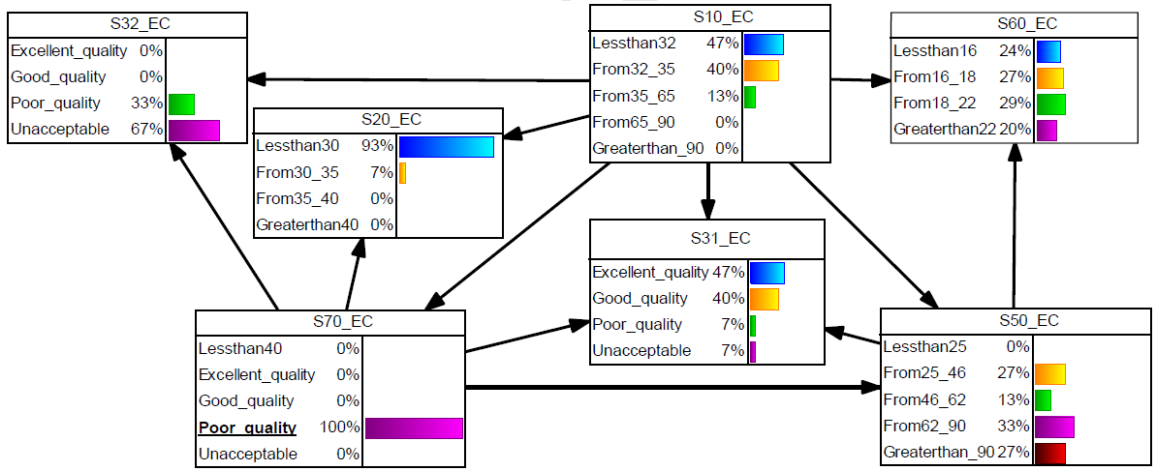


Figure 6- 26: Analysis for poor water quality in catchment S70.

High EC levels in catchments S32 and S50 correlate with "poor water quality" in catchment S70 (see Figure 6- 26). This support the hypothesis and indicates that if EC values increase in catchments S32 and S50, there is a high likelihood that this will affect the quality of surface water in catchment S70.

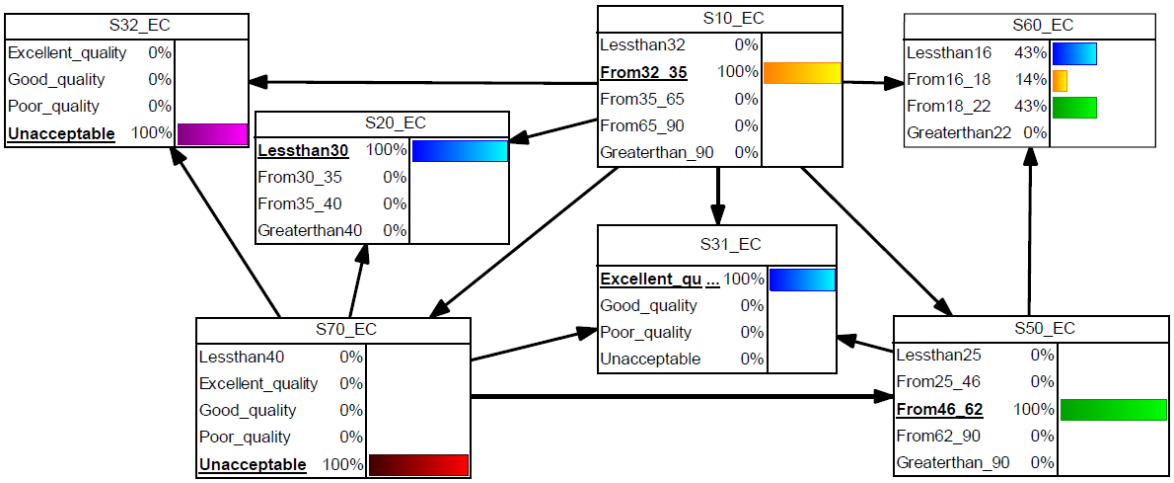


Figure 6- 27: Analysis for unacceptable water quality in catchment S70.

The results for unacceptable water quality in S70 show that only catchment S32 supports the hypothesis, unacceptable levels in S32 are related to "unacceptable levels in S70 (Figure 6- 27). These results can be used to make inferences about the likely impact of further deterioration in surface water quality in S32 on quality in S70. The results can also be used to assess the effectiveness of different water quality management strategies set up for catchment S32 on other catchments, for example S70.

This is a first step in getting a general idea of the spatial interaction of monitoring points along the river. The results thus produced have to be verified using more data and at more detailed spatial scales, for example evaluation of a spatial network of points along the river in the same catchment. Such a network can thereafter be used to monitor the movement of pollutants from source to sink.

The results can also be verified by using more detailed temporal scales of analysis, for example using daily or hourly readings were available. This is vital because the initial data used in this research were averaged monthly readings. The assumption made was that the values were constant throughout the day which is not strictly true. The use of daily readings can reveal more patterns and verify the results obtained in this study.

6.5 Dynamic Bayesian modelling results

Initially, sensitivity analyses were performed on the static Bayesian Network to assess the temperature, rainfall, recharge, surface water N and surface water EC. The results show that the relationships between surface water EC and N and temperature are weakly defined. The recharge, surface water EC and N variables were therefore omitted from the temporal analysis. As a result, temporal predictions were carried out only for rainfall and temperature. This also seemed appropriate considering the fact that these two variables had the most accurate, and variable amount of data. In the presence of adequate data, a larger number of variables in a more complex network can be used for temporal predictions.

In terms of the weather, the study area has two seasons, summer and winter. The winter season is from May to September and the summer season is from October to April. Rainfall and temperature values for the different seasons are shown in Figures 6- 28 and 6- 29.

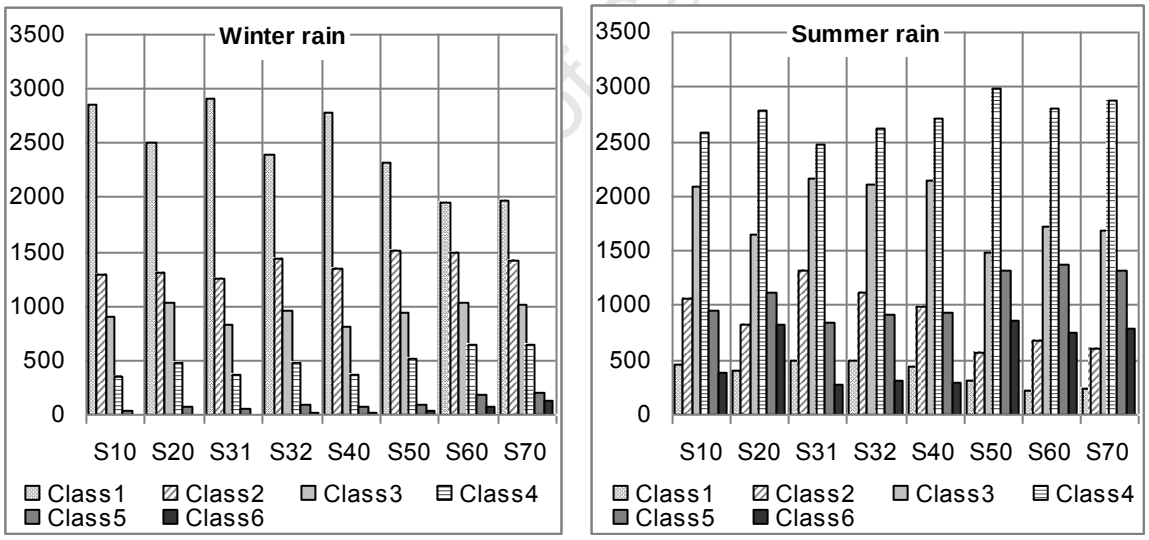


Figure 6- 28: The frequencies of the different classes of rainfall for the winter and summer seasons.

The rainfall graphs show that the winter season receives generally lower rainfall than the summer season. In terms of winter, Figure 6- 29 shows that winter temperatures are lower than summer temperatures.

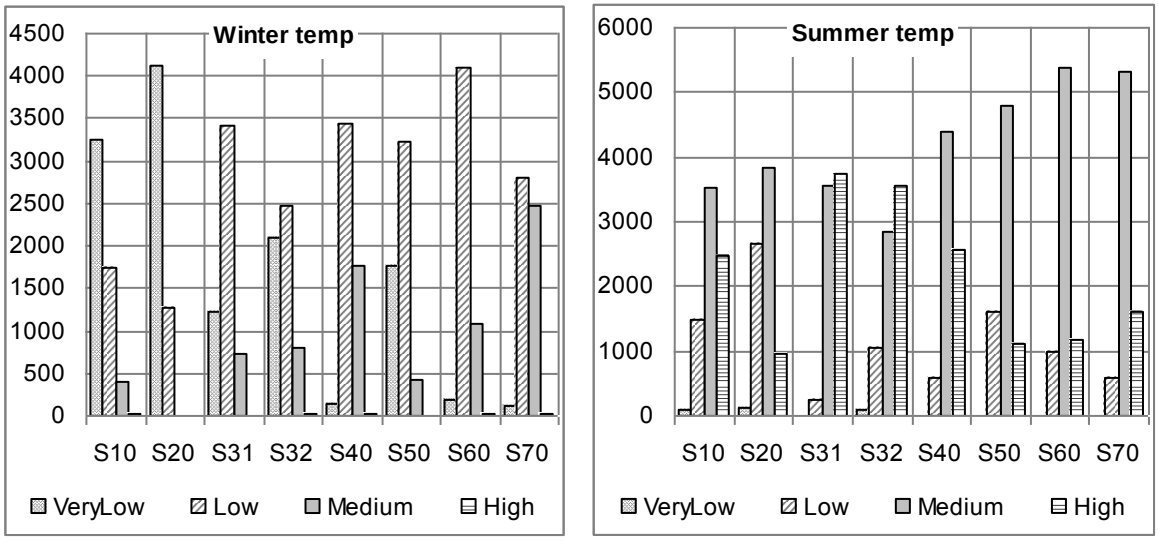


Figure 6- 29: The frequencies of the different classes of temperature for the winter and summer seasons.

In terms of the relationship between rainfall and temperature, the general trend is that the lower the temperature, the lower the rainfall is in a specific catchment and the higher the temperature, the more the monthly rainfall received (Figure 6- 30).

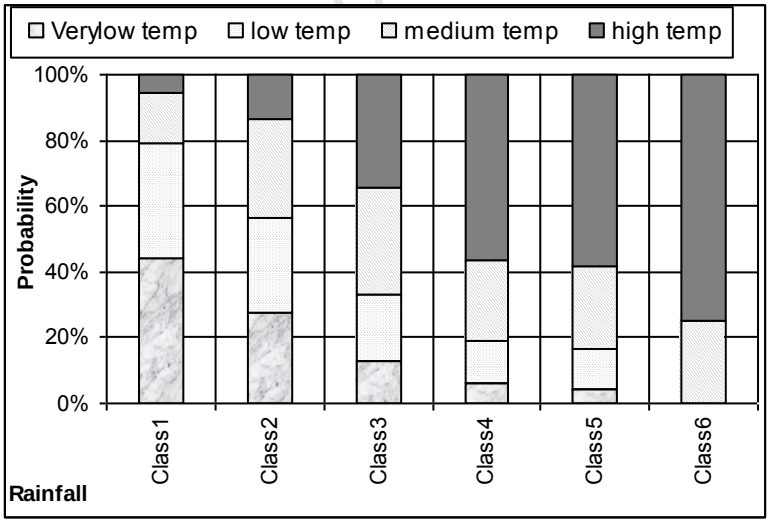


Figure 6- 30: Relationship between temperature and rainfall.

Scenario analysis was performed in the custom developed software and the evidence for temperature and rainfall for selected months of the year were entered as input.

Inference is carried out to predict the rainfall for a certain month based on previous readings and the resulting probabilities per tertiary catchment are saved as text files. An example of the results from inference is shown in Figure 6- 31.

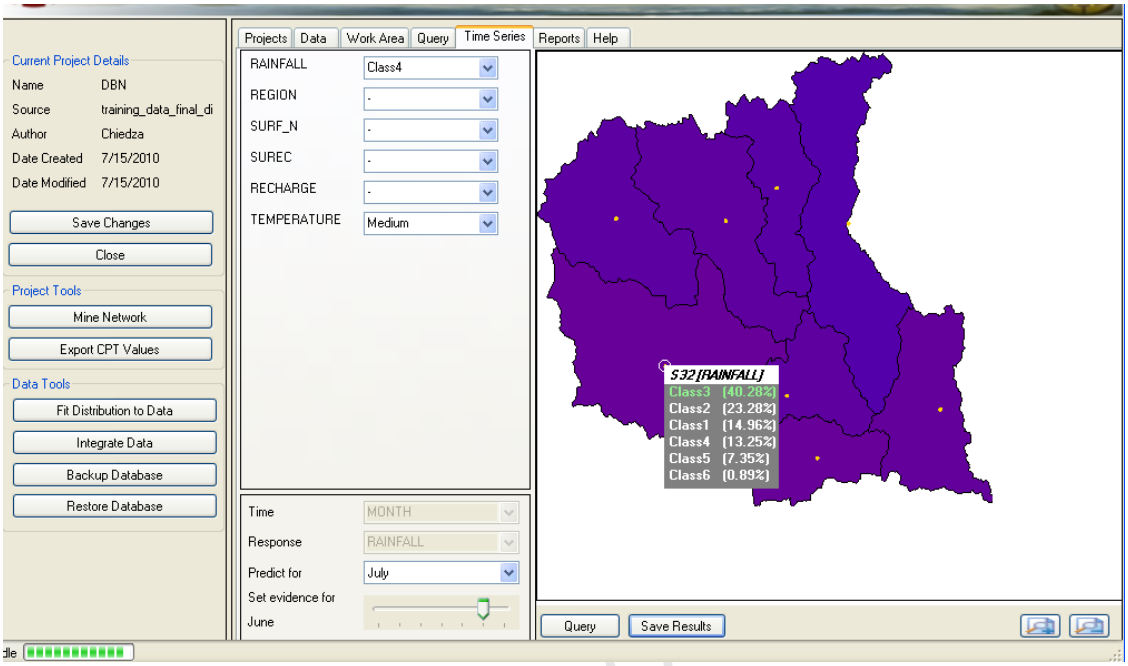


Figure 6- 31: Screenshot showing the inference capabilities of the custom developed software.

The predictions for rainfall and temperature are performed for tertiary catchments S10 and S60 for the following months:

i) July, the middle of winter and

ii) April, during the summer season.

S10 and S60 were selected for analysis because they have the highest densities of rainfall stations and have more reliable data than the other regions. The two catchments also have varying patterns in terms of rainfall and temperature. In both winter and summer, catchment S10 receives lower rainfall that S60. Regarding the winter season, temperatures are higher in S60 than in S10.

The Dynamic Bayesian Network was used to make temporal predictions for rainfall in July. Different scenarios were evaluated by varying the evidence from preceding months to July. The evidence and prediction results are shown in Table 6- 2.

Table 6- 2: The evidence and predictions for rainfall for July in tertiary catchments S10 and S60

	Evidence													Rainfall Prediction	
Result A	January		February		March		April		May		June		July	*S10*	*S60*
	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall		*Class 1*	*Class 3*
	Medium		Medium		Medium		VeryLow		Low		VeryLow			*42%*	*38%*
Result B	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall		*Class 1*	*Class 4*
	High	Class 1	Medium	Class 5	Medium	Class 3	High	Class 1	VeryLow	Class 5	Low	Class 4		*30%*	*24%*
Result C	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	*Class 1*	*Class 1*
	Medium	Class 3	Medium	Class 4	Medium	Class 2	Low	Class 3	Low	Class 6	Low	Class 1	VeryLow	*55%*	*39%*
Result D	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall		*Class 3*	*Class 3*
	High`	Class 3	High	Class 4	Medium	Class 2	VeryLow	Class 6	Medium	Class 1	Low	Class5		*48%*	*42%*
Result E	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	*Class 1*	*Class 3*
	Medium	Class 6	High	Class 6	Low	Class 1	High	Class 4	Low	Class 3	VeryLow	Class5	Low	*38%*	*30%*
Result F	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall		*Class 3*	*Class 3*
	Medium	Class 4	Medium	Class 6	Low	Class 5	VeryLow	Class 1	Low	Class 6	Low	Class3		*29%*	*33%*
Result G	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall		*Class 3*	*Class 3*
	Medium	Class 4	Medium	Class 6	Low	Class 1	VeryLow	Class 1	Low	Class 6	VeryLow	Class5		*41%*	*45%*
Result H	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall		*Class 3*	*Class 3*
	Medium	Class 2	Medium	Class 6	Low	Class 1	VeryLow	Class 3	Medium	Class 2	VeryLow	Class2		*48%*	*44%*
Result I	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	*Class 3*	*Class 3*
	High	Class 5	High	Class 6	Medium	Class 4	Medium	Class 6	VeryLow	Class 4	Low	Class6	Low	*39%*	*44%*
Result J	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall		*Class 3*	*Class 3*
	Medium	Class 4	Medium	Class 6	High	Class 3	High	Class 5	VeryLow	Class 5	Low	Class3		*36%*	*39%*

Using the static Bayesian Network, the expected predictions for July rainfall for catchments S10 and S60 are shown in Figure 6- 32.

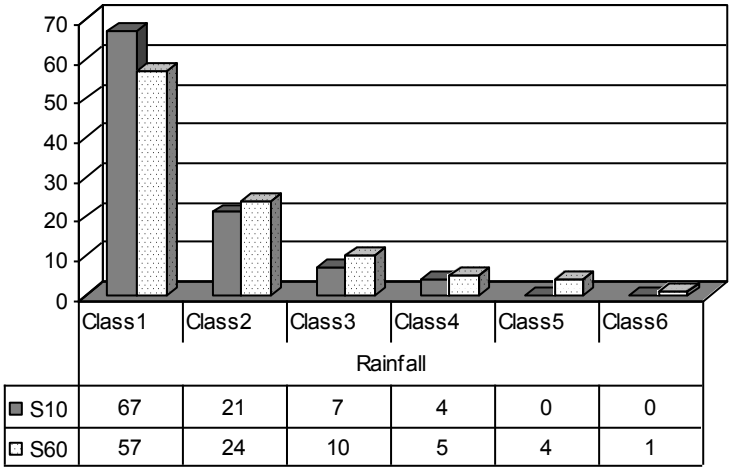


Figure 6- 32: The most probable states for rainfall for July.

The results in Table 6- 2 show that Class 3 was calculated as the most likely state for both catchments S10 and S60. This does not reflect the result in Figure 6- 32 which was obtained from analysing the data using the static Bayesian Network. Only in four of the cases of prediction for S10 were the predictions in accordance with the result from the static network. For catchment S60, only one case managed to predict the Class 1 value.

The results show that having evidence of the preceding months improves the prediction of rainfall for a specific month. This is illustrated by the changes in the probabilities of the states of rainfall with variations in the evidence for the preceding months. The discrepancy between the results of the static network and the dynamic network can be explained by considering the seasonal variations in rainfall for the two regions as shown in the graphs in Figure 6 -28. In terms of winter rain, for the month of July, in catchment S60 there is only a slight difference between the likelihoods of Class 1 rain and that of Class 3 rain as compared to region S10 where Class 1 rainfall is dominant. With the Dynamic Bayesian Networks, seasonal variations in temperature prediction were possible whereas the static network only considered the averages of the whole dataset.

Table 6- 3: The evidence and predictions for temperature for July in tertiary catchments S10 and S60

	Evidence													Temp Prediction	
Result A	January		February		March		April		May		June		July	*S10*	*S60*
	Temp	Rainfall	Temp											*Medium*	*Medium*
	Medium		Medium											*33%*	*52%*
Result B	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall								*Medium*	*Medium*
	Medium	Class 2	High	Class 4	Medium	Class 6								*58%*	*79%*
Result C	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall						*Medium*	*Medium*
	High	Class 1	Medium	Class 5	Medium	Class 5	Medium	Class 3						*50%*	*75%*
Result D	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall				*Medium*	*Medium*
	Medium	Class 2	Medium	Class 1	Low	Class 1	VeryLow	Class 4	VeryLow	Class 5				*44%*	*68%*
Result E	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall		*Medium*	*Medium*
	High	Class 3	High	Class 2	Medium	Class 3	Low	Class 2	Low	Class 3	VeryLow	Class6		*59%*	*75%*
Result F	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Rainfall	*High*	*Medium*
	Medium	Class 4	Medium	Class 6	Low	Class 4	High	Class 1	VeryLow	Class 2	VeryLow	Class 4	Class 3	*45%*	*77%*
Result G	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Rainfall	*Medium*	*Low*
	High	Class 5	High	Class 3	Medium	Class 3	Medium	Class 6	Low	Class 1	Low	Class 3	Class 6	*51%*	*53%*
Result H	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Temp	Rainfall	Rainfall	*Low*	*Low*
	Medium	Class 6	Medium	Class 4	Low	Class 2	Medium	Class 5	VeryLow	Class 4	Low	Class2	Class 6	*50%*	*75%*

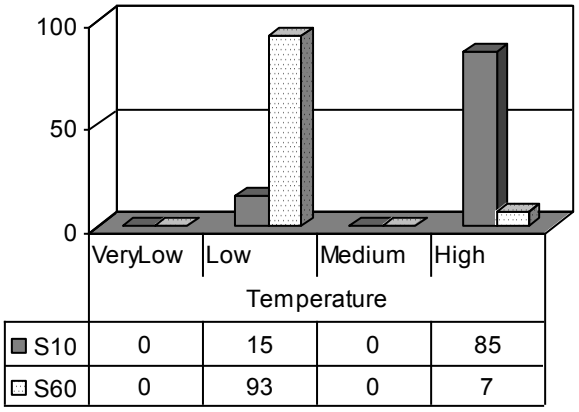


Figure 6- 33: The most probable states of temperature for July.

From Figure 6 -33, the results of using a static network, it is evident that the most probable state for temperature in catchment S10 is "high" and in catchment S60 it is "low". The results from the Dynamic Bayesian Network are different; most cases predict "medium" temperature for both catchments. Only three cases produced similar results to the static network although the second most likely state for most cases for both catchments was similar, that is "low". This shows the importance of having evidence from past months; it improves the accuracy of the prediction.

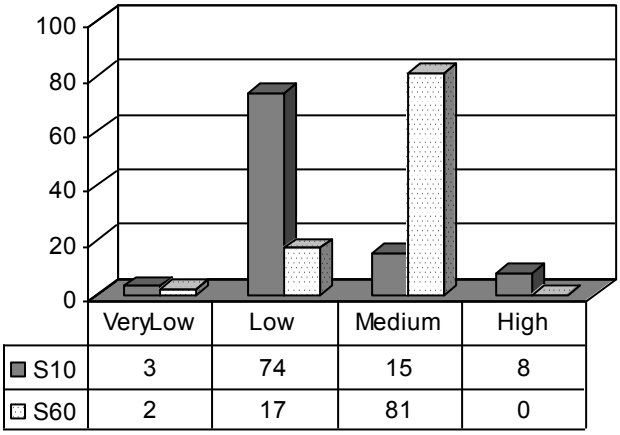


Figure 6- 34: The most probable states of temperature for April.

The most probable states for temperature for April from the static Bayesian Network are as shown in Figure 6– 34.

Table 6- 4: The evidence and predictions for temperature for April in tertiary catchments S10 and S60

	Evidence							Temp Prediction			
Result A	January		February		March		April	*S10*		*S60*	
	Temperature	Rainfall						*VeryLow*	*Medium*	*Low*	*Medium*
		Class 1						*32%*	*30%*	*58%*	*38%*
Result B	Temperature	Rainfall						*Low*	*Medium*	*Low*	*Medium*
		Class 2						*32%*	*30%*	*57%*	*38%*
Result C	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall		*Low*	*Medium*	*Low*	*Medium*
	High	Class 3	Medium	Class 5	Medium	Class 3		*40%*	*32%*	*68%*	*29%*
Result D	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall		*Low*	*Medium*	*Low*	*Medium*
	Medium	Class 2	High	Class 4	Medium	Class 4		*37%*	*31%*	*64%*	*32%*
Result E	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Rainfall	*Low*	*Medium*	*Low*	*Medium*
	High	Class 6	Medium	Class 2	Low	Class 5	Class 6	*40%*	*26%*	*67%*	*30%*
Result F	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall		*Low*	*Medium*	*Low*	*Medium*
	Medium	Class 2	High	Class 6	Low	Class 1		*39%*	*31%*	*62%*	*34%*
Result G	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Rainfall	*VeryLow*	*Low*	*Low*	*Medium*
	High	Class 6	Medium	Class 3	High	Class 3	Class 1	*51%*	*30%*	*72%*	*24%*
Result H	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Rainfall	*Medium*	*Low*	*Medium*	*Low*
	Medium	Class 5	High	Class 1	Medium	Class 6	Class 6	*31%*	*28%*	*49%*	*44%*
Result I	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Rainfall	*VeryLow*	*Low*	*Low*	*Medium*
	High	Class 3	Medium	Class 2	Medium	Class 5	Class 3	*35%*	*33%*	*67 %*	*30%*

The results for temperature prediction using the Dynamic Bayesian Network for April and the evidence used are shown in Table 6- 4. Figure 6- 34 shows that the most likely state for temperature in S10 is "low" and in S60 it is "medium". For S10, more than 50% of the cases predicted "low" temperature, which is similar to the results from the static network. But according to the graph showing the seasonal variation in summer temperature, "medium" is the most likely state for temperature in S10. The prediction from the dynamic network is more accurate and incorporates the seasonal variations in rainfall and temperature.

The results are less accurate for catchment S60, where the majority of cases predicted "low" temperature. "Medium" is predicted as the second most likely state after "low" temperature. The results are not similar to the ones from the static network due to the input of evidence of rainfall from the preceding months.

6.6 Conclusions

This chapter presented the modelling results and examined the different scenarios that are relevant to the research. Scenarios investigated include variations of urbanisation, groundwater and surface water quality, changes in irrigated areas and variations in water demand. Evaluation of the scenarios was done using sensitivity analyses which provided information on the most influential variables in the prediction of the query variable. An example of a "what if" scenario was presented and this showed the full power of Bayesian Networks in modelling. The spatial prediction of EC values in a catchment based on evidence on other neighbouring catchments or more upstream catchments was explored. Lastly, time series analysis was performed for rainfall and temperature using Dynamic Bayesian Network.

Based on the modelling results obtained in this chapter, conclusions can be drawn about the capabilities and shortcomings of the techniques presented. These are discussed in the following chapter, which concludes the research. The conclusion includes a summary of the research, highlights major findings and outlines further research work that is required to address some of the shortcomings encountered.

CHAPTER SEVEN: CONCLUSIONS AND RECOMMENDATIONS

7 Introduction

This chapter concludes the research. It examines the extent to which the objectives stated in Chapter 1 were fulfilled and the implications of the results obtained. Any outstanding issues, challenges and recommendations for future research are discussed.

7.2 Summary of research

The main objective of this thesis was to develop a Bayesian Network-based methodology for spatial and temporal assessment of water resources in the Great Kei catchment in the Eastern Cape Province, South Africa. This has been successfully fulfilled as is discussed in the methodology presented in Chapters 5 and the results discussed in Chapter 6. A model has been developed and was applied to the Great Kei catchment in South Africa.

The specific objectives of this research were to:

i) develop a Bayesian Network model for water resources assessment using automatic techniques;
ii) collate data for the selected study area, the Great Kei catchment in the Eastern Cape Province in South Africa;
iii) use the collected data in a custom developed software, based on a Hydrid Genetic Algorithm, to automatically create the network;
iv) apply the developed Bayesian Network model to evaluate spatial variations in water resources parameters in the study area;
v) evaluate and validate the Bayesian Network model mainly using sensitivity and scenario analyses; and
vi) illustrate the use Dynamic Bayesian Networks for temporal prediction of some aspects of the catchment over specified times.

The tasks and methodologies used to address each specific objective are discussed in the following paragraphs.

Literature Review

The research commenced with literature reviews which are presented in Chapters 2, 3 and 4. The aim of the literature review in Chapter 2 was to provide information on water resources assessment and discuss the issues pertinent to assessment at catchment scale. These included the need to integrate multiple datasets at different spatial and temporal scale, the importance of stakeholder participation in the process and issues of uncertainty in the input data and the model output. The available models for water resources assessment in South Africa were presented and critiqued and this led to the justification for the use of Bayesian Networks to assess some of the complexities identified for water resources assessment. This Chapter was used to make the case for the use of Bayesian Networks.

Chapter 3 presented literature on the theory of Bayesian Networks and Dynamic Bayesian Networks. The steps used to create Bayesian Networks were outlined and these ranged from the creating of the structure, the use of the data to calculate probabilities and how inference is performed using the results. The use of sensitivity and scenario analyses to evaluate the network and gets results to address specified queries is discussed. The different scoring measures that are commonly used to verify Bayesian Networks were discussed. The benefits of combining GIS and Bayesian Networks, especially for catchment modelling were presented using examples from literature. Lastly, the use of Dynamic Bayesian Networks for temporal modelling was examined. The information provided in this chapter assisted in the creation of the methodology for the research.

The last literature review chapter, Chapter 4 examined in detail examples from literature on the use of Bayesian Networks and Dynamic Bayesian Networks in water resources management. International case studies and applications from South Africa were discussed. The issues considered included the purpose of the modelling procedure, the software used, the methods for pre-processing the data, the evaluation and verification techniques and the presentation of the data. The chapter concluded with a discussion on the advantages and limitations of Bayesian Networks. The examples and the information presented in this chapter also informed the development of the methodology applied in this research.

Bayesian Network conceptual model development

Chapter 5 discussed the conceptual model which was developed for water resources assessment. The model was based on a study that was carried out in South Africa in 2003 to create a list of indicators suitable for catchment sustainability assessment. The list of indicators emerging from this research was based on review of existing literature, expert knowledge and extensive stakeholder participation. These indicators where selected for use as variables in the modelling in this research. The conceptual model was developed based on the results of this study.

Data collection and automatic mining of relationships

Chapter 5 commenced with a discussion of the custom developed software for Bayesian Network and Dynamic Bayesian Network modelling. This software was developed based on an unsupervised Hybrid Genetic Algorithm (HGA). The software was used to discretise continuous data, to mine the structure of the network, compute CPTs. The results from inference/queries performed in the software are displayed on a map and the resulting probabilities are exported to text files

The input data into the software is a single table/spreadsheet with a list of cases for all variables. Information on the data gathered for the study area, the scale and frequency of collection, the completeness of the records and sources of uncertainty in the data are listed in Chapter 5. Most of the data used, except rainfall, temperature, runoff and water quality was based on the once-off water assessment study carried out in the study area in 1995.

The data were aggregated to quaternary catchment scale and monthly modelling time steps for the years from 1950-1999 were used. The only variables that varied monthly were rainfall, runoff and temperature. For the rest of the data, a constant average value was used to represent the variable. This is discussed in detail in Chapter 5.

The automatic learning involved two steps, firstly the automatic discretisation of the continuous data and then learning the structure and CPTs from the data. In cognisance of the fact that the levels selected to discretise the data affect the relationships and ultimately the structure of the network, the approach adopted was to create three networks using three discretisation levels. The resulting networks were scored in terms of predictive accuracy and then the network with the best scores was selected as the most "optimum to use". The discretisation levels from the winning network where selected for discretising the continuous data. The process is presented in Chapter 5.

Application of model to the study area

The developed Bayesian Network model was applied to the Great Kei catchment in the Eastern Cape Province in South Africa. The features and capabilities of Bayesian Networks are demonstrated in Chapter 6 using examples. The combination of Bayesian Networks and GIS enabled the rapid visualisation of inference results. "What if" queries and scenario analyses were facilitated by visual interpretations for the different regions in the Great Kei catchment.

Evaluation and verification of the network

An initial process for the verification of the network was performed and this was discussed in Chapter 5. The aim of the process was to assist in the selection of the "optimum" discretisation levels for the data and the network to adopt. This involved the use of scoring measures to inform the selection process.

Chapter 6 examined the use of sensitivity and scenario analyses, both evaluation techniques, to test the network. The analyses were used to successfully investigate the effects of urbanisation, variations in water quality and changes in water demand on various aspects of the catchment. The evaluation results highlighted some previously unknown patterns and confirmed some expected system behaviours. This assists in improving the knowledge of the study area

The results of the study on the effects of variations in urbanisation showed that urbanisation affects water demand and water availability. More urbanised areas have a greater demand for water resources when compared with rural areas. This has implications for future water supply as the study area is expected to experience an increase in rural to urban migration in future. People are expected to migrate to urban areas for employment and more economic opportunities. Decision-makers have to cater for this increase in their planning.

In terms of groundwater TDS and surface water EC, high water demand areas, for example urban areas had high levels and this was attributed to runoffs from sewerage treatment plants, unlicensed solid wastes and poorly designed sewage systems. These results conform to relationships documented in literature. For integrated catchment management, this type of analysis can be used to provide quantitative measures of the impacts of the implementation of proposed water management systems on groundwater pollution in the catchment.

Poor water quality was also evident in rural areas were informal settlements are rife. Due to the lack of adequate sanitation in these areas, the pollution was attributed to the runoff from these areas. Highly irrigated areas, which are a source of non-point pollution, also had increased likelihoods for elevated EC and nitrogen values. In order to verify these results, the effects of geology on quality have to be removed isolated.

The power of Bayesian Networks lies in its ability to assess the effects of changes in multiple variables on one query aspect. This provides the capability of discovering "the most likely cause" of a scenario and also the combined effects of the different variables. The combined effects of irrigation and urbanisation on water demand were used as an example. The results showed that when the two factors are combined, urbanisation has a greater effect than irrigation. This highlighted the importance of including urbanisation statistics in planning for future water resources supply and management. Such a multivariable simultaneous analysis can be used to monitor changes in water demand with changes in the use pattern of all water users.

Spatial prediction

A Bayesian Network was created to illustrate the use of the technique for spatial prediction. The variable tested was surface water EC. Such a network is useful in catchment management in accessing the effects of increased pollution in areas upstream on downstream sub-catchments. The hypothesis was that an increase in EC values upstream leads to an increase in values further downstream and vice versa. Results from some catchments supported this hypothesis. More accurate monitoring data are required to verify these results.

Temporal analysis

The application of Dynamic Bayesian Networks in time-series modelling is illustrated by examples in Chapter 6. The main advantage of Dynamic Bayesian Networks, which must be pursued for future research, is its use in predicting the values of variables in the future, where no comparative data are available. Temporal analyses were performed for rainfall and temperature because these two variables had the most accurate data. Rainfall predictions were made for the month of July. The results of the temporal analysis were compared with those from the static Bayesian Network. Prediction using the Dynamic Network managed to reflect the seasonal variations and this was not possible with the static network. Having knowledge about the past improves the prediction of the future values of a variable.

7.3 Recommendations and future work

Bayesian Networks have been shown to improve the understanding and prediction of factors of the catchment. As was presented in Chapter 6, the extent to which Bayesian Networks can fulfil the objectives depends on the availability of accurate and complete datasets. One of the advantages of using Bayesian Networks is that the results of modelling are presented with a statement of confidence that illustrates the belief in the results as opposed to a one-number answer.

The next step in this research after developing the model should be the presentation of the results to the relevant stakeholders to get feedback. The stakeholder feedback could be used to evaluate the variables included, their states, the CPTs, and the outcome of modelling. This input and the other outcomes of the research may further refine the model, perhaps through changes in the variables included, the scale of application and the relationships defined. The ease with which Bayesian Networks can be updated with new information lends them suitable for this kind of adaptive and iterative process.

This research proposed a simplified representation of what constitutes catchment water resources assessment. This was mainly due to time, data and financial resources. It is likely that in reality, the concept is more complex with more datasets and relationships than those defined. Future work could involve investigating the more comprehensive and broader integrated catchment management.

The major limitation to the model results and usage is the availability and accuracy of the data used. As presented in Chapter 5, there are problems with the spatial and temporal coverage of most datasets. Of particular importance is the groundwater and surface water values were yearly (or in some cases as much as 5 years) averages were correlated with monthly temperature and rainfall data collected from 1950-1999. Considering the relationships obtained between water quality and some socio-economic indicators, it is imperative to have better temporal and spatial coverage of water quality data.

It is recommended that more monitoring stations be established and monitoring must ideally take place on a daily basis. It is also highly recommended that the model be tested with more accurate site specific data preferably obtained from monitoring on daily intervals to assess if similar patterns are discovered. An important point to note is that the network structure influences the sensitivity analysis as nodes close to the one being tested have the most influence. When assessing the impacts of changes on a "query" variable, the variables directly related to the query variable will have the most effect as opposed to the ones further from it, which might be an artificial trend. This implies that expert knowledge needs to be incorporated into the creation of the structure of the network. This will enable the network to be a true representative of the problem domain and make it provide the answers required by the decision-maker.

As illustrated in this thesis, one of the major strengths of Bayesian Networks is their ability to provide useful models and results with limited or missing data. An alternative approach is to use outputs from simulation models or statistical estimation or interpolation algorithms as input data to populate the network. The choice of the suitable estimation or interpolation methods is based on literature or outcomes from other studies in similar catchments in South Africa. No assessment was done to test their suitability to the study area or the datasets used in this thesis as this was not within the scope of this thesis. Future research work should involve a rigorous evaluation of the available, relevant techniques before applying the data in modelling.

Most of the datasets used in this research and generally in catchment modelling are continuous. The commonly available software for modelling cannot handle this type of data and it has to be discretised. As this thesis highlighted, the definition of the different ranges for discretisation affect the relationships between variables and their sensitivities. Further work can be on evaluating and producing better methods for discretising data or investigating algorithms that can handle continuous data.

The use of expert knowledge has been shown to be vital in various aspects of modelling, firstly in developing the structure of the network and also in estimating probabilities when data is incomplete or some immeasurable variable needs to be included. Although automatic structure mining techniques are useful for defining relationships, these techniques cannot be used solely, without expert knowledge, especially with limited data. The automatic mining of structure in Bayesian Networks is still a subject of active research and future work could focus on the development of improved algorithms to assist.

The spatial and temporal scales of the data affect predictions. The selection of the scales in this research was informed by data availability and policy. It could be that the selected scales are not representative of the catchment processes. More work should involve a detailed investigation of the different scales and lead to recommendations on the optimal scales to use in decision-making and planning activities in the catchment.

This research documented the types of uncertainties inherent in the datasets used and their effects on modelling. Some aspects of parameter uncertainty were analysed but these were based on sensitive analyses. The investigation of uncertainties in the modelling results which arise from the model structure, which is an important aspect of any modelling exercise, is not addressed in this research. Future work could involve the development of a procedure for model uncertainty analysis. A suggestion is the combination of Bayesian Networks with Monte Carlo simulation. The suitability of tools like the Data Uncertainty Engine (DUE) (Brown and Heuvelink, 2005) can also be explored.

The use of Dynamic Bayesian Networks in time-series analysis is briefly explored in this research. The aim was to test their applicability in catchment process modelling. Initial results showed that they can be successfully applied. Future work could involve a rigorous analysis of their applicability and the inclusion of more variables and data in analysis.

The investigation of the effects of climate change on catchment processes is a topical research issue. The main tools that have been used to assess these effects in South Africa are physical-based hydrology models. The capabilities of Bayesian Networks highlighted throughout this thesis can facilitate such types of analyses. Bayesian Networks can enable stakeholder participation in the catchment management process.

This research indicated the use of Bayesian Networks in enhancing the understanding of catchment dynamics, the performance of basic what if queries and the prediction of catchment factors. For decision-making support, the application of Bayesian Decision Networks (BDN) should be explored. BDN have the same principles as Bayesian Networks, the difference being that they can be used to evaluate the likely benefits/costs of different interventions to known catchment problems. They can also be used in assessing the impacts of climate and political change on catchment resources. Future work should explore their application in integrated catchment management.

REFERENCES

Acocks, J.P.H. 1988. Veld Types of South Africa 3rd edition, Memoirs of the Botanical Survey of South Africa 57: 1-46.

Agarwal, C., Green, G.L., Grove, M., Evans, T., & Schweik, C. 2000. A Review and Assessment of Land-Use Change Models Dynamics of Space, Time, and Human Choice. [Online]. Available: http://hero.geog.psu.edu/archives/AgarwalEtALInPress.pdf [Accessed: 10 January 2007]

Aghazadeh, N., & Mogaddam, A.A. 2010. Assessment of Groundwater Quality and its Suitabilityfor Drinking and Agricultural Uses in the Oshnavieh Area, Northwest of Iran. *Journal of Environmental Protection*, (1):30-40.

Amakali, M., & Shixwameni, L., 2003. River basin management in Namibia. *Physics and Chemistry of the Earth,* 28(20):1055-1062.

Ames, D.P., Neilson, B.T., Stevens, D.L., & Lall, U. 2005. Using Bayesian networks to model watershed management decisions: an East Canyon Creek case study. *Journal of Hydroinformatics*, 7(2005):267-282.

Ames, D.P.2002. *Bayesian Decision Networks for watershed management.* Logan, Utah. Utah State University. (PhD-thesis).

Andreasen, J.K., O'Neill, R., Noss, R., & Slosser, N.C. 2001. Considerations for the development of a terrestrial index of ecological integrity. *Ecological Indicators*, 1(1):21-35.

Arancibia, A., & Moriarty, P.2003. Towards the use of Bayesian Networks for decision support in watershed management: A case study from the Peru Sierra. [Online]. Available: http://www.bvsde.paho.org/bvsacd/agua2003/aran.pdf [Accessed: 10 January 2007]

Argent, R. M. 2004. An overview of model integration for environmental applications-components, frameworks and semantics. *Environmental Modelling & Software*, 19(3):219-234.

Argent, R. M., Vertessy, R. A., & Watson, F. G. R. 2000. A framework for catchment prediction modelling in Hydro 2000. *Proceedings of 26th National and 3rd International Hydrology and Water Resources Symposium, Vol. 2.The Institution of Engineers*, Australia, Perth, pp. 706-711.

Argent, R.M., Grayson, R.B., & Ewing, S.A. 1999. Integrated Models for Environmental Management: Issues of Process and Design. *Environmental International*, 25(6):693-699.

Ashton, P. 1996. *Integrated Catchment Management: Balancing resource utilization and conservation.* [Online]. Available: http://www.awiru.co.za/pdf/astonpeter.pdf [Accessed: 11 January 2008]

Aspinall, R., & Pearson, D. 2000. Integrated geographical assessment of environmental condition in water catchments: Linking landscape ecology, environmental modelling and GIS. *Journal of Environmental Management*, 59(4):299-319.

Bacon, P. J., Cain, J. D., & Howard, D. C. 2002. Belief network models of land manager decisions and land use change. *Journal of Environmental Management*, 65(1):1-23.

Baran, E., & Jantunen, T. 2004. Stakeholder consultation for Bayesian decision support systems in environmental management. *Proceedings of the Regional Conference on Ecological and Environmental Modelling (ECOMOD 2004)*, Universiti Sains Malaysia, 15-16 September 2004, Penang, Malaysia.

Barnes, C.J. 1995. The art of catchment modeling: What is a good model? *Environment International*, 21(5):747-751.

Barton, D.N., Saloranta, T., Moe, S.J., Eggestad, H.O., & Kuikka, S. 2008. Bayesian belief networks as a meta-modelling tool in integrated river basin management- Pros and cons in evaluating nutrient abatement decision under uncertainty in a Norwegian river basin. *Ecological Economics*, 66(1):91-104.

Basson, M.S., van Niekerk, P.H., & van Rooyen, J.A. 1997. *Overview of water resources availability and utilization in South Africa*. Pretoria: Department of Water Affairs & Forestry Report RSA/00/0197.

Batchelor, C., Moriarty, P., & Laban, P. 2005. *Using Water Resources Assessments within the EMPOWERS IWRM planning cycle*. [Online]. Available: http://www.empowers.info/content/download/1224/10181/file/EMPOWERS%20WP%20 [Accessed: 10 May 2007]

Bednarski, M., Cholewa, W., & Frid, W. 2004. Identification of sensitivities in Bayesian networks. *Engineering Applications of Artificial Intelligence*, 17(4):327-335.

Bless, C., & Kathuria, R. 1998. *Fundamentals of Social Statistics: An African Perspective*. 2nd rev. ed. South Africa: Juta & Co. Ltd.

Blöschl, G. 2006. Hydrologic synthesis: Across processes, places, and scales. *Water Resources Research*, 42, W03S02, doi: 10.1029/2005WR004319.

Borsuk, M.E., Stow, C.A., & Reckhow, K.H. 2004. A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. *Ecological Modelling*, 173(2):219-239.

Bouckaert, R. 2007. *Bayesian Network Classifiers in WEKA for Version 3-5-7*. [Online]. Available: http://www.cs.waikato.ac.nz/~remco/weka_bn/node1.html [Accessed: 10 September 2007]

Bromley, J., Jackson, N.A., Clymer, O.J., Giacomello, A.M., & Jensen, F.V. 2005. The use of Hugin® to develop Bayesian networks as an aid to integrated water resource planning. *Environmental Modelling and Software*, 20(2):231-242.

Brown, J.D., & Heuvelink, G.B.M. 2005. *Data Uncertainty Engine (DUE) User's Manual*. [Online]. Available: www.harmonirib.com [Accessed 10 January 2008].

Burgman, M. 2005. *Risks and Decisions for Conservation and Environmental Management*. United Kingdom: Cambridge University Press.

Burns, M., Audouin, M., & Weaver, A. 2006. Advancing sustainability science in South Africa. *South African Journal of Science*, 102.

Butterworth, J., & Soussan, J. 2001. Water Supply and Sanitation & Integrated Water Resources Management: why seek better integration? WHIRL Project Working Paper 1. *The WHIRL project workshop on Water Supply & Sanitation and Watershed Development: positive and negative interaction*, Andhra Pradesh, India, 5-14 May 2001.

Cain, J., Batchelor, C., & Waughray, D. 1999. Belief Networks: A framework for the participatory development of natural resources management strategies. *Environment, Development and Sustainability*, 1(2): 123-133.

Caminiti, J.E. 2004. Catchment modelling – a resource manager's perspective. *Environmental Modelling and Software*, 19(11):991–997.

Chakrabarti, M. 2001. Towards an operational definition of sustainability. *Paper presented at the Seminar on Poverty and Sustainable Development, organized by UNESCO*, University of Montesquieu-Bordeax IV, Paris, 22-23 November 2001. [Online]. Available: http://www.cbnrm.net/library/documents/index.html [Accessed 29 October 2004]

Chee, Y.E., Burgman, M., & Carey, J.2005. *Use of a Bayesian Network Decision Tool to Manage Environmental Flows in the Wimmera River, Victoria*. [Online]. Available:

www.waterscience.com.au/pdf/Risk_based_Approaches_Report_4.pdf [Accessed: 10 September 2007]

Chen, S.H., Jakeman, A.J., & Norton, J.P. 2008. Artificial Intelligence techniques: An introduction to their use for modelling environmental systems. *Mathematics and Computers in Simulation*, 78(2):379-400.

Cheng. J., Bell, D., & Liu, W. 1997. *Learning Bayesian Networks from Data: An Efficient Approach Based on Information Theory*. [Online]. Available: www.cs.ualberta.ca/~jcheng/Doc/report98.pdf [Accessed: 10 March 2008].

Chevallier L, Goedhart M and Woodford AC. 2001. *The influence of dolerite sill and ring complexes on the occurrence of groundwater in Karoo fractured aquifers: a morpho-tectonic approach*. WRC No. 937/1/01, Pretoria: Water Research Commission.

Chickering, D.M., & Heckerman, D. 2000. A comparison of scientific and engineering criteria for Bayesian model selection. *Statistics and Computing*, 10(1):55-62.

Cooper, G.F., Herskovits, E. 1992. A Bayesian Method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309-347.

Costanza, R., & Ruth, M. 1998. Using Dynamic Modeling to Scope Environmental Problems and Build Consensus. *Environmental Management*, 22(2):183-195.

Costanza, R., Wainger, L., Folke, C., & Mäler, K. 1993. Modeling Complex Ecological Economic Systems. *BioScience*, 43(8):545-555.

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., & Spiegelhalter, D.J. 1999. *Probabilistic Networks and Expert Systems*. New York: Springer-Verlag.

Cullis, J., Görgens, A.H.M., & Rossouw, J.N. 2004. *First Order estimate of the contribution of agriculture to non-point source pollution in three South African catchments: Salinity, Nitrogen and Phosphorus*. Pretoria: Water Research Commission Report No 1467/2/04.

Curtis, R., Fenn, R., & Oberholster, D. 2005. *Predicting Plankton from Satellite Data.* University of Cape Town, Cape Town: Department of Computer Science Technical Report No. CS05-19-00. [Online]. Available: http://pubs.cs.uct.ac.za/archive/00000261/01/Technical_Report_final.pdf [Accessed: 5 April 2008]

Darwiche, A. 2001. Constant-Space reasoning in dynamic Bayesian Networks. *International Journal of Approximate Reasoning*, 26(3):161-178.

Dawes, W.R., Walker, G.R., & Stauffacher, M. 2001. Practical modelling for management in data-limited catchments. *Mathematical and Computer Modelling*, 33(6):625-633.

Dawsey, W.J., Minsker, B.S., & Amir, E. 2007. *Real time assessment of drinking water systems using a Dynamic Bayesian Network*. [Online]. Available: http://reason.cs.uiuc.edu/eyal/papers/bayesian-water-wewrc07.pdf [Accessed: 10 October 2008].

de Jager, F.G.B., & van Rooyen, P.G. 2008. *Water Resources Yield Model (WRYM) User Guide – Release 7.5.6.3.* [Online]. Available: http://www.usersupport.co.za/general.php?model=WRYM [Accessed: 10 January 2009]

de Kock, M.2008. *Weather forecasting using Dynamic Bayesian Networks.* Cape Town, South Africa. University of Cape Town. (BSc-thesis).

de Santa Olalla, F.J.M., Domínguez, A., Artigao, A., Fabeiro, C., & Ortega, J.F. 2005. Integrated water resources management of the Hydrogeological Unit ''Eastern Mancha'' using Bayesian Belief Networks. *Agricultural Water Management*, 77(1):21–36.

de Santana, A.L., Francês, C.R., Rocha, C.A., Carvalho, V., Vijaykumar, N.L., Rego, L.P., & Costa, J.C. 2007. Strategies for improving the modeling and interpretability of Bayesian networks. *Data & Knowledge Engineering*, 63(1):91-107.

Dennis, I., & van Tonder, G.J. 2004. The architecture and application of the South African Groundwater Decision Tool. (*In*: Stephenson, D., Shemang, E.M., & Chaoka, T.R. (eds.), *Water Resources of Arid Areas*. London: Taylor & Francis Group.

Dondo, C., Chevallier, L., Woodford, A.C., Murray, R., Nhleko, L.O., Nomnganga, A. & Gqiba. D. 2010. *Flow Conceptualisation, Recharge and Storativity Determination in Karoo Aquifers, with special emphasis on Mzimvubu - Keiskamma and Mvoti - Umzimkulu Water Management Areas in the Eastern Cape and KwaZulu-Natal Provinces of South Africa*. WRC Report (in press), Pretoria: Water Research Commission.

Donigian, A.S. 2002. *Watershed Model calibration and validation: The HSPF experience*. [Online]. Available: http://hspf.com/TMDL.Nov02.Donigian.Paper.doc [Accessed: 10 May 2008]

Driscoll, F.G. 1986. *Groundwater and Wells*. 2nd ed. St. Paul, Minnesota: Johnson Division.

Dube, R.A. 2006. *Appropriate positioning of modeling as a decision support tool for surface water resources planning in South Africa*. Pretoria. University of Pretoria. (PhD-thesis).

Dye, P.J., & Croke, B.F.W. 2003. Evaluation of streamflow predictions by the IHACRES rainfall-runoff model in two South African catchments. *Environmental Modelling & Software*, 18 (8):705-712.

Ekasingh, B., & Letcher, R.A. 2005. Successes and failures of attempts to embed socioeconomic dimensions in modeling for integrated natural resource Management: Lessons from Thailand. (*In* Zerger, A. & Argent, R.M. (eds.), *MODSIM 2005 International Congress on Modelling and Simulation.* Australia: Modelling and Simulation Society of Australia and New Zealand. p. 170-176.) [Online]. Available: http://www.mssanz.org.au/modsim05/papers/ekasingh.pdf [Accessed: 10 August 2007]

Eng, J. 2006. *ROC analysis: web-based calculator for ROC curves*. Baltimore: Johns Hopkins University [Online]. Available: http://www.jrocfit.org [Accessed: 15 August 2008].

Everard, M. 2004. Investing in sustainable catchments. *The Science of the Total Environment*, 324(1):1-24.

Gabaix, X., Laibson, D. 2008. *The Seven Properties of Good Models.* [Online]. Available: http://www.economics.harvard.edu/faculty/laibson/files/NYU%20Methodology%20may%2020.pdf [Accessed: 15 August 2009]

Ghahramani, Z. 1997. Learning Dynamic Bayesian Networks. (*In* Giles, C.H. & Gori, M (eds.), *Adaptive Processing of Sequences and Data Structures.* Berlin: Springer-Verlag. p. 168-197.) [Online]. Available: http://www.gatsby.ucl.ac.uk/~zoubin/papers/vietri.pdf [Accessed: 10 June 2007]

Gibson, C.C., Ostrom, E., & Ahn, T.K. 2000. The concept of scale and the human dimensions of global change: a survey. *Ecological Economics*, 32(2):217-239.

Global Water Partnership (GWP). 2001. *Integrated Water Resources Management.* Online: Available: http://www.waterland.net/gfx/content/Policy_Choices_and_Challenges.doc [Accessed: 13 August 2007]

Grandin, R., Potgieter, A., & Zield, J. 2006. *Predicting Sub-surface Plankton Profiles using Dynamic Bayesian Networks*. [Online]. Available: http://stefanor.uctleg.net/past-honours projects/grandin_mcintosh_symington/files/report_prediction.pdf [Accessed: 10 June 2008]

Greiner, R. 2004. Systems framework for regional-scale modelling and assessment. *Mathematics and Computers in Simulation*, 64(1):41-51.

Grêt Regamey, A., & Straub, D. 2006. Spatially explicit avalanche risk assessment linking Bayesian networks to a GIS. *Natural Hazards and Earth System Sciences*, 6:911-926.

Hand, D., Mannila, H, & Smyth, P. 2001. *Principles of Data Mining*. Cambridge: The MIT Press.

Hansen, J.R., Refsgaard, J.C., Hansen, S., & Ernstsen, V. 2007. Problems with heterogeneity in physically based agricultural catchment models. *Journal of Hydrology*, 342(1-2):1-16.

Hare, M., & Deadman, P. 2004. Further towards a taxonomy of agents-based simulation models in environmental management. *Mathematics and Computers in Simulation*, 64(1):25-40.

He, C., Malcolm, S.B., Dahlberg, K.A., & Fu, B. 2000. A conceptual framework for integrating hydrological and biological indicators into watershed management. *Landscape and Urban Planning*, 49(1):25-34.

Heckerman, D., Geiger, D., & Chickering, D.M. 1995. Learning Bayesian Networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197-243.

Hemson, D., Meyer, M., & Maphunye, K. 2004. *Rural Development: The provision of basic infrastructure services*. [Online]. Available: http://www.sarpn.org.za/documents/d0000683/Rural_development_012004.pdf [Accessed: 10 January 2006]

Henriksen, H.J., & Rasmussen, P., Brandt, G., von Bülow, D., & Jensen, F.V. 2003. Public participation modelling using Bayesian networks in management of groundwater contamination. *Environmental Modelling & Software*, 22(8):1101-1113.

Hermanowicz, S.W. 2005. *Sustainability in Water Resources Management: Changes in Meaning and Perception*. [Online]. Available: http://repositories.cdlib.org/cgi/viewcontent.cgi?article=1049&context=wrca [Accessed: 12 August 2006]

Herold, C.E., & le Roux, P.J. 2004. WQ2000: *Development of an interactive surface water quality information system for South Africa*. Pretoria: Water Research Commission Report 950/1/04.

Hoeting, J.A., Madigan, D., & Raftery, A.E. 1999. *Bayesian Model Averaging: A Tutorial.* Technical Report 9814. Department of Statistics, Colorado State University. [Online]. Available: http://www.stat.colostate.edu./research/1999.html [Accessed: 10 May 2010]

Hope, L.R., & Korb, K.B. 2004. *A Bayesian Metric for Evaluating Machine Learning Algorithms*. [Online]. Available: http://www.csse.monash.edu/~korb/pubs/inforeward.pdf [Accessed: 10 May 2008]

Huang, J., & Ling, C.X. 2005. Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299-310.

Hughes, D.A. 2001. Providing hydrological information and data analysis tools for the determination of ecological instream flow requirements for South African rivers. *Journal of Hydrology,* 241(1):140-151.

Hughes, D.A. 2004. Three decades of hydrological modelling research in South Africa. *South African Journal of Science,* 100(11):638-642.

Hughes, D.A., & Parsons, R. 2004. *Ground water Pitman model Version 3-Model Description.* [Online]. Available: http://www.ru.ac.za/institutes/iwr/software/reserve/Pitman/GW-SW-V3.htm [Accessed: 30 May 2008]

Institute of Water Research (IWR), 2008b. *Identification, estimation, quantification and incorporation of risk and uncertainty in Water Resources Management tools in South Africa*. [Online]. Available: http://www.ru.ac.za/static/institutes/iwr/uncertainty/DEL2_Literature_Review.pdf [Accessed: 7 January 2009]

Institute of Water Research (IWR), 2008a. The Pitman Monthly Model. [Online]. Available: http://www.ru.ac.za/institutes/iwr/software/reserve/helpdss/model11.htm [Accessed: 30 May 2008]

Ioris, A.A.R., Hunter, C., & Walker, S. 2008. The development and application of water management sustainability indicators in Brazil and Scotland. *Journal of Environmental Management*, 88(4):1190-1201.

Ismail, M.K. 2003. *An empirical investigation of the impact of discretization on common data distributions*. Australia: RMIT University (MTech-thesis). [Online]. Available: http://goanna.cs.rmit.edu.au/~vc/papers/ismail.pdf [Accessed: 10 April 2008]

Jain, A., & Kumar, A.M. 2007. Hybrid neural network models for hydrologic time series forecasting. *Applied Soft Computing,* 7(2):585-592.

Jakeman, A.J., & Letcher, R.A. 2003. Integrated assessment and modelling: features, principles and examples for catchment management. *Environmental Modelling & Software,* 18(6), 491-501.

Jakeman, A.J., Letcher, R.A. & Norton, J.P. 2006. Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software,* 21(5):602-614.

Jakeman, A.J., Letcher, R.A., Newham, L.T.H., & Norton, J.P. 2005. Integrated catchment modelling: issues and opportunities to support improved sustainability outcomes.

Proceedings of the 29th Hydrology and Water Resources Symposium, Canberra, 21-23 February, pp. 1-15 [Online]. Available: http://icam.anu.edu.au/Jakeman_2005_Water%20Capital.pdf [Accessed: 10 April 2007]

Janssen, W., & Goldsworthy, P. 1996. Multidisciplinary Research for Natural Resource Management: Conceptual and Practical Implications. *Agricultural Systems*, 51(3):259-279.

Jensen, F. 2001. *Belief Networks and Decisions Graphs*. New York: Springer 2001.

Jessel, B., & Jacobs, J. 2005. Land use scenario development and stakeholder involvement as tools for watershed management within the Havel River Basin. *Limnologica – Ecology and Management of Inland Waters*, 35(3):220-233.

Jewitt, G.P.W. 1998. *Resolution of scale issues in an Integrated Catchment Information System for the rivers of the Kruger National Park*. Cape Town: University of Stellenbosh. (PhD-thesis).

Jewitt, G.P.W., & Görgens, A.H.M. 2000. *Scale and model interfaces in the context of integrated water resources management for the rivers of the Kruger National Park*. Pretoria: Water Research Commission Report 627/1/00.

Jewitt, G.P.W., & Schulze, R.E. 1999. Verification of the *ACRU* model for forest hydrology applications. *Water SA*, 25(4):483-489.

Johnson, A.K.L., Shrubsole, D., & Merrin, M. 1996. Integrated Catchment Management in northern Australia: From concept to implementation. *Land Use Policy* 13(4):303-316.

Kaluer, B., Mysiak, J., & Sigel, K. 2005. Socio-economic data for river basin management. (*In* Van Loon, E., & Refsgaard, J.C. (eds.), *Guidelines for assessing data uncertainty in river basin management studies*. Geological Survey of Denmark and Greenland, Copenhagen. P. 182.) [Online]. Available: http://harmonirib.geus.info/ [Accessed: 10 December 2008]

Karodia, H. 1998. Implications of the new Water Policy: Problems and Solutions. *Proceedings of the 1998 Royal Society of South Africa Working Conference*, Cape Town, 11-13 November 1998. [Online]. Available: www.rssa.uct.ac.za/conferen/pre14.htm [Accessed: 15 October 2007]

Kelbe, B., Germishuyse, T. 1999. *A Study of the Relationship between Hydrological Processes and Water Quality characteristics in the Zululand Coastal Region*. Pretoria: Water Research Commission Report No. 346/1/99.

Khu, S., Savic, D., Liu, Y., & Madsen, H. 2004. A fast Evolutionary-based Meta-Modelling Approach for the Calibration of a Rainfall-Runoff Model. *Proceedings of the 2004 International Environmental Modelling and Software Society iEMSs 2004*, University of Osnabrück, Germany, 14-17 June 2004. [Online]. Available: http://www.iemss.org/iemss2004/pdf/evocomp/khuafas.pdf [Accessed: 10 December 2008]

Kiiveri, H.T., Cacceta, P., & Evans, F. 2001. Use of conditional probability networks for environmental monitoring. *International Journal of Remote Sensing*, 22(7):1173-1190.

Kipkemboi, J., van Dam, A. A., & Denny, P. 2007. Environmental impact of seasonal integrated aquaculture ponds ('fingerponds') in the wetlands of Lake Victoria, Kenya: an assessment, with the aid of Bayesian Networks. *African Journal of Aquatic Science 2007*, 32(3):219–234.

Kjaerulff, U.B., & Madsen, A.L. 2008. *Bayesian Networks and Influence Graphs: A Guide to Construction and Analysis*. New York: Springer Science and Business Media, LLC.

Kjeldsen T.R., & Rosbjerg, D. 2001. A framework for assessing the sustainability of a water resources system. (*In* Schumann, A.H., Acremann, M.C., Davis, R., Marino, M.A., Rosbjerg, D., & Xia Jun. (eds.), *Regional Management of Water Resources*. IAHS Publ. no 268, p. 107-113.)

Koivusalo, H., Kokkonen, T., Laine, H., Jolma, A., & Varis, O. 2005. Exploiting simulation model results in parameterising a Bayesian network – A case study of dissolved organic carbon in catchment runoff. (*In* Zerger, A., & Argent, R.M. (eds.) *MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand*, p 170-176.) [Online]. Available: http://www.mssanz.org.au/modsim05/papers/koivusalo.pdf [Accessed: 10 December 2008]

Korb, K.B., & Nicholson, A.E. 2004. *Bayesian artificial intelligence*. Chapman and Hall/CRC Press.

Koutsoyiannis, D., Karavokiros, G., Efstratiadis, A., Mamassis, N., Koukouvinos, A, & Christofides, A. 2003. A decision support system for the management of the water resource system of Athens. *Physics and Chemistry of the Earth*, 28(14-15):599-609.

Kroese, N., Visser, P., Nhlapo, A., Terblanche, D., & Banitz L. 2006. *Daily Rainfall mapping over South Africa (DARAM): Infrastructure and capacity building*. Pretoria: Water Research Commission Report No. 1426/1/06.

Krzysztofowicz, R. 2001. The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, 249(1):2-9.

Lanini, S. 2006. Water management impact assessment using a Bayesian Network model. *Proceedings of the 7th International Conference in Hydroinformatics*, Nice, France, 4-7 September 2006. Chennai, India: Research Publishing Services.

Le Maitre, D.C., Milton, S.J., Jarmain, C., Colvin, C.A., Saayman, I & Vlok, JHJ. 2007. Linking ecosystem services and water resources: landscape-scale hydrology of the Little Karoo. *Frontiers in Ecology and the Environment*; 5(5):261–270

Letcher, R.A., & Jakeman, A.J. 2002. Experiences in an Integrated Assessment of Water Allocation Issues in the Namoi River Catchment, Australia. *Proceedings of the 2002 Integrated Environmental Modelling and Software Society*, Lugano, Switzerland, 24-27 June 2002. [Online]. Available: http://www.iemss.org/iemss2002/proceedings/pdf/volume%20tre/177_letcher.pdf [Accessed: 10 November 2007]

Loucks, D.P. 2000. Sustainable water resources management. *Water International*, 25(1):3-10.

Loucks, D.P., van Beek, E., Stedinger J.R., Dijkman, J.P.M., & Villars, M.T. 2005. *Water Resources Systems Planning and Management: An Introduction to Methods, Models and Applications*. France: UNESCO Publishing.

Lovell, C., Mandondo, A. & Moriarty, P. 2002. The question of scale in integrated natural resource management. *Conservation Ecology*, 5(2):25. [Online]. Available: http://www.consecol.org/vol5/iss2/art25/ [Accessed: 10 November 2007]

Lupini, L. 2004. "Policies that work" in a context of integrated catchment management for sustainable use of water to enhance rural development. *Proceedings of the 87th EAAE-Seminar: Assessing rural development policies of the CAP*, Vienna/Wien, Austria, 21-23 April, 2004.

Lynch, S.D. 2004. *Development of a raster database of annual, monthly and daily rainfall for Southern Africa*. Pretoria: Water Research Commission Report 1156/1/03.

Macleod. C.J.A., Scholefield, D., & Haygarth, P.M. 2007. Integration for sustainable catchment management. *Science of the Total Environment*, 373(2-3):591-602.

Macleod. C.J.A., Scholefield, D., & Haygarth, P.M. 2007. Integration for sustainable catchment management. *The Science of the Total Environment*, 373 (1):591-602, January 2007.

Margaritis, D. 2003. *Learning Bayesian Network Model Structure from data*. Pittsburgh: Carnegie Mellon University. (PhD-thesis). [Online]. Available: reports-archive.adm.cs.cmu.edu/anon/2003/CMU-CS-03-153.ps [Accessed: 10 April 2008]

McDonald, D.A., & Pape, J. 2002. *Cost recovery and the crisis of service delivery in South Africa.* South Africa: HSRC Publishers.

McIntosh, H. 2008. *Dynamic Bayesian Networks for Ensemble Seasonal Forecasting.* [Online]. Available: http://www.csag.uct.ac.za/files/bosberaad%20Hayley.ppt#257,1,Slide 1 [Accessed: 10 April 2008]

Meek, C. 2003. *An Overview of Learning Bayes Nets from Data* [Online]. Available: http://www.jsmf.org/meetings/2003/meek.pdf [Accessed: 5 March 2008]

Meentemeyer, V. 1989. Geographical perspectives of space, time, and scale. *Landscape Ecology*, 3(3):163-173.

Mehrjardi, R.T., Jahromi, M.Z., Mahmodi, Sh., & Heidari, A. 2008. Spatial Distribution of Groundwater Quality with Geostatistics (Case Study: Yazd-Ardakan Plain). *World Applied Sciences Journal*, 4(1):09-17.

Midgley, D.C., Middleton, B.J., & Pitman, W.V. 1990. *The Surface Water Resources of South Africa 1990*. Pretoria: Water Research Commission.

Middleton, B.J., & Bailey, A.K. 2008. Water Resources of South Africa, 2005 Study (WR2005): Book of maps: Version 1. Pretoria: Water Research Commission. Report No. TT 382/08.

Mihajlovic, V., & Petkovic, M. 2001. *Dynamic Bayesian Networks: A State of the Art.* CTIT technical reports series, TR-CTIT-34. ISSN 13813625. [Online]. Available: http://purl.org/utwente/36632 [Accessed: 18 April 2008]

Morgan, M. G., & Henrion, M., 1990). Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis, New York: Cambridge University Press.

Mukheibir, P., & Sparks, D. 2003. *Water resource management and climate change in South Africa: Visions, driving factors and sustainable development indicators.* [Online]. Available: http://www.erc.uct.ac.za/publications/Water%20resource%20management%20and%20climate%20change%20in%20SA%20-%202003.pdf [Accessed: 25 January 2006]

Murphy, K. 1998. *A Brief Introduction to Graphical Models and Bayesian Networks*. [Online]. Available: http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html [Accessed: 8 May 2008]

Murphy, K. P. 2002a. *Dynamic Bayesian Networks*. [Online]. Available: http://www.cs.ubc.ca/~murphyk/Papers/dbnchapter.pdf [Accessed: 5 May 2008]

Murphy, K. P. 2002b. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Berkeley, University of California. (PhD-thesis.)

Musango, J., & Peter, C. 2007. A Bayesian approach towards facilitating climate change adaptation research on the South African agricultural sector. *Agrekon*, 46(2):245-259.

Mwelwa, E.M. 2004. *The Application of the monthly time step Pitman Rainfall-Runoff Model to the Kafue River Basin of Zambia*. Grahamstown, South Africa: Rhodes University (MSc-thesis.)

Mzezewa, J., Misi, T., & van Rensburg, L.D. 2010. Characterisation of rainfall at a semi-arid ecotope in the Limpopo Province (South Africa) and its implications for sustainable crop production. *Water SA*, 36 (1):19-26.

Nelken, R., & Shieber, S.M. 2006. *Computing the Kullback-Leibler Divergence Between Probabilistic Automata Using Rational Kernels*. Harvard University School of Engineering and Applied Sciences (SEAS), Technical Report TR-07-06. [Online]. Available: ftp://ftp.deas.harvard.edu/techreports/tr-07-06.pdf [Accessed: 10 April 2008]

Ness, B., Urbel-Piirsalu, E., Anderberg, S., & Olsson, L. 2007. Categorising tools for sustainability assessment. *Ecological Economics*, 60(3):498-508.

Nevill, J. 2000. *Measuring and Assessing Sustainability: water resources and ecosystems*. [Online]. Available: http://www.netspace.net.au/~jnevill/freshwater.htm [Accessed: 10 November 2008]

Nicholson, A., Watson, S., & Twardy, C. 2003. *Using Bayesian Networks for Water Quality Prediction in Sydney Harbour*. [Online]. Available: www.csse.monash.edu.au/bai/talks/NSWDEC.ppt [Accessed: 10 December 2008]

Niedermayer, D.1998. *An Introduction to Bayesian Networks and their Contemporary Applications*. [Online]. Available: http://www.niedermayer.ca/papers/bayesian/bayes.html [Accessed: 10 April 2008].

Nilsson, B., Højberg, A.L., Refsgaard, J.C., & Troldborg, L. 2006. Uncertainty in geological and hydrogeological data. *Hydrology and Earth System Sciences. Discussions*, 3:2675-2706.

O' Regan, B., & Moles, R. 2006. Using system dynamics to model the interaction between environmental and economic factors in the mining industry. *Journal of Cleaner Production*, 14(8):689-707.

Ochieng, G.M. 2007. *Hydrological and water-quality modelling of the Upper Vaal Water Management Area using a stochatic mechanistic approach*. Tswane, Tswane University of Technology. (PhD-thesis.)

Pahl-Wostl, C. 2007. The implications of complexity for integrated resources management. *Environmental Modelling & Software*, 22(5):561-569.

Parker, P., Letcher, R., Jakeman, A., Beck, M.B., Harris, G., Argent, R.M., Hare, M., Pahl-Wostl, C., Voinov, A., Janssen, M., Sullivan, P., Scoccimarro, M., Friend, A., Sonnenshein, M., Barker, D., Matejicek, L., Odulaja, D., Deadman, P., Lim, K., Larocque, G., Tarikhi, P., Fletcher, C., Put, A., Maxwell, T., Charles, A., Breeze, H., Nakatani, N., Mudgal, S., Naito, W., Osidele, O., Eriksson, I., Kautsky, U., Kautsky, E., Naeslund, B., Kumblad, L., Park, R., Maltagliati, S., Girardin, P., Rizzoli, A., Mauriello, D., Hoch, R., Pelletier, D., Reilly, J. Olafsdottir, R., & Bin, S. 2002. Progress in integrated assessment and modelling. *Environmental Modelling and Software*, 17(3):209-217.

Parkin, G., O'Donnell, G., Ewen, J., Bathurst, J.C., O'Connell, P.E., & Lavabre, J. 1996. Validation of catchment models for predicting land-use and climate change impacts: 2. Case study for a Mediterranean catchment. *Journal of Hydrology*, 175(1-4):595-613.

Parsons, R. 2004. *Surface Water: Groundwater interaction in a South African Context-A Geohydrological Perspective*. Pretoria: Water Research Commission Report No. TT 218/03.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 2nd rev. ed. San Francisco: Morgan Kaufmann.

Pearl, J. 1978. An economic basis for the certain methods of evaluating probabilistic forecasts. *International Journal of Man-Machine Studies*, 10:175-183

Peter, C., Musango, J.K., & De Lange, W.2007. Using Bayesian networks to model the impact of climate change scenarios on biofuels production from irrigated agriculture – analyzing water, energy and food sector interdependencies. *Presented at IPCC TGICA Expert Meeting,* Nadi, Fiji, 20-22 June 2007.

Petersen, B., Gernaey, K., Devisscher, M., Dochain, D., Vanrollegem, P.A., 2003. A simplified method to assess structurally identifiable parameters in Monod-based activated sludge models. *Water Research*, 38:2893-2904

Pine Rivers Catchment Association, 2010. What is a Catchment? [Online]. Available: http://www.prca.org.au/index.php?option=com_content&view=article&id=16&Itemid=16 [Accessed: 10 June 2010]

Pitman, W.V., Kakebeeke, J.P., & Bailey, A.K. 2007. *WRSM2000. Water Resources Simulation Model for Windows:Users Guide July 2008*. 6th ed. Pretoria: Department of Water Affairs and Forestry.

Poch, M., Comas, J., Rodríguez-Roda, I., Sànchez-Marrè, M., & Cortés, U. 2004. Designing and building real environmental decision support systems. *Environmental Modelling & Software*, 19(9):857-873.

Pollard, P., Devlin, M., & Holloway, D. 2001. Managing a complex river catchment: a case study on the River Almond. *Science of the Total Environment*, 265 (1):343-357, January 2001.

Pollard, S. 2002. Operationalising the new Water Act: contributions from the Save the Sand Project – an integrated catchment management initiative. *Physics and Chemistry of the Earth*, 27(11):941-948.

Pollard, S., & Walker, P. 2000. *Catchment management and water supply and sanitation in the Sand River Catchment, South Africa: description and issues*. [Online]. Available: http://www.nri.org/projects/WSS-IWRM/Reports/Working_papers/WHIRL%20working%20paper%201_final.pdf [Accessed: 5 September 2008]

Pollard, S.R., Kotze, D., Ellery, W., Cousins, T., & Jewitt, G., 2006. *Towards wetland and livelihood improvements: An integrated socio-ecological approach to the rehabilitation of a communal wetland in north-eastern region of South Africa*. [Online]. Available: http://www.award.org.za/File_uploads/File/Towards%20wetland%20Pollard%20et%20al%202006%20WI%20final.pdf [Accessed: 10 June 2008]

Pollino, C.A., Woodberry, O., Nicholson, A., Korb, K., & Hart, B.T. 2007. Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment. *Environmental Modelling and Software*, 22(8):1140-1152.

Potts, J.M., Folland, C.K., Jolliffe, U.T., & Sexton, D. 1996. Revised "LEPS" Scores for Assessing Climate Model Simulations and Long-Range Forecasts. *Journal of Climate*, 9:34-53.

Pourret, O., Naim, P., & Marcot, B. 2008. *Bayesian Networks: A Practical Guide to Applications*. England: John Wiley & Sons Ltd.

Pullar, D., & Springer, D. 2000. Towards integrating GIS and catchment models. *Environmental Modelling and Software*, 15(5):451-459.

Quinlan, T., & Scogings, P. 2004. Why bio-physical and social scientists can speak the same language when addressing Sustainable Development. *Environmental Science & Policy*, 7(6):537-545

Refsgaard, J.C., & Henriksen, H.J. 2002. Modelling guidelines- a theoretical framework. (*In* Refsgaard, J.C. (eds.), *State-of-the-Art Report on Quality Assurance in modelling related to river basin management*. HarmoniQuA-report, D-WP1-1.) [Online]. Available: http://harmoniqua.wau.nl/public/Reports/SOA%20chapters/SOA.pdf [Accessed: 10 July 2008]

Refsgaard, J.C., & Henriksen, H.J. 2004. Modelling guidelines-terminology and guiding principles. *Advances in Water Resources*, 27(1):71-82.

Refsgaard, J.C., Henriksen, H.J, Harrar, W.G., Scholten, H., & Kassahun, A. 2005. Quality assurance in model based water management - review of existing practice and outline of new approaches. *Environmental Modelling & Software*, 20 (10):1201-1215.

Refsgaard, J.C., van der Sluijs, J.P., Brown, J., & van der Keur, P. 2006**.** A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources*, 29(11):1586-1597.

Refsgaard, J.C., van der Sluijs, J.P., Højberg, A.L., & Vanrolleghem, P.A. 2007**.** Uncertainty in the environmental modelling process – A framework and guidance. *Environmental Modelling & Software*, 22(11):1543-1556.

Riddell, E.S., Lorentz, S.A., Ellery, W.N., Kotze, D., Pretorius, J.J., & Nketar, S. N. 2007. *Water Table Dynamics of a Severely Eroded Wetland System, Prior to Rehabilitation, Sand River Catchment, South Africa***.** [Online]. Available: http://www.award.org.za/File_uploads/File/Riddell%20%20IAH%20Sept%202007.pdf [Accessed: 10 May 2008]

Robertson, D., Wang, Q.J., Haines, C. 2004. *Bayesian Networks to Assist Land and Water Management*. [Online]. Available: www.csse.monash.edu.au/~ctwardy/Bnmelb/robertson.pdf [Accessed: 10 May 2008]

Rode, M, & Suhr, U.2005. Surface water quality data. (*In* Van Loon, E., & Refsgaard, J.C. (eds.), *Guidelines for assessing data uncertainty in river basin management studies*. Geological Survey of Denmark and Greenland, Copenhagen. P. 182.) [Online]. Available: http://harmonirib.geus.info/ [Accessed: 10 December 2008]

Ryan, J., McAlpine, C., & Ludwig, J. 2007. GLAMS: A Graphical Method for Capturing Land and Water Management Practices in Agroecosystems. *Ecosystems*, 10(3):432-447.

Sadoddin, A., Letcher, R.A., Jakeman, A.J., & Newham, L.T.H 2005. A Bayesian decision network approach for assessing the ecological impacts of salinity management. *Mathematics and Computers in Simulation*, 69(1-2):162-176.

Said, A., Sehlke, G., Stevens, D.K., Glover, T., Sorensen, D., Walker, W., & Hardy, T. 2006. Exploring an innovative watershed management approach: From feasibility to sustainability. *Energy*, 31(13):2037-2050.

Sawunyama, T. 2008. *Evaluating uncertainty in water resources estimation in Southern Africa: A case study of South Africa*. Grahamstown: Rhodes University (PhD-thesis.)

Sayer, J.A. & Campbell, B.M. 2004. *The Science of Sustainable Development: Local Livelihoods and the Global Environment*. United Kingdom: Cambridge University Press.

Schafer, J.L., & Graham, J.W. 2002. Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2):144-177. [Online]. Available: http://folk.ntnu.no/eiriksko/Shafer.pdf [Accessed: 1 June 2008].

Shihab, K. 2008. Dynamic Modeling of groundwater pollutants with Bayesian Networks. *Applied Artificial Intelligence*, 22:352-376.

Schultz, C.B., Watson, M.D., & Taviv, I. 2000. An inter regional water resources planning model. *Proceedings 1st WARFSA/WaterNet Symposium: Sustainable Use of Water Resources*, Maputo, 1-2 November 2000. Maputo: WARFSA WaterNet.

Schulze, R. 2000. Transcending scales of space and time in impact studies of climate and climate change on agrohydrological responses. *Agriculture, Ecosystems and Environment*, 82(1):185-212.

Schulze, R.E. 1994. *Hydrology and agrohydrology. A text to accompany the ACRU 3.00 Agrohydrological modelling system*. Pretoria: Water Research Commission Report No. TT69/95.

Schulze, R.E. 2003. On Models and Modelling for Integrated Water Resources Management: Some Thoughts, Background Concepts, Basic Premises and Model Requirements. (*In*: Schulze, R.E. (eds.), *Modelling as a Tool in Integrated Water Resources Management: Conceptual Issues and Case Study Applications*. Pretoria: Water Research Commission, Report No. 749/1/02. p. 20-46.)

Schulze, R.E., & Maharaj, M. 2003. *Development of a Database of Gridded Daily Temperatures for Southern Africa*. Pretoria: Water Research Commission Report No. 1156/2/04.

Schulze, R.E., & Smithers, J.C. 2003. The *ACRU* Modelling System as of 2002: Background, Concepts, Structure, Output, Typical Applications and Operations. (*In*: Schulze, R.E. (eds.), *Modelling as a Tool in Integrated Water Resources Management: Conceptual Issues and Case Study Applications*. Pretoria: Water Research Commission, Report No. 749/1/02, p 47-83.)

Schulze, R.E., 2005. Some foci of integrated water resources management in the "South" which are oft-forgotten by the "North": A perspective from southern Africa. *Water Resources Management*, 2(1): 269-294.

Sheppard, E., & McMaster, R.B. 2004. *Scale and Geographic Inquiry: Nature, Society, and Method*. USA: Blackwell Publishing Ltd.

Silberstein, R.P. 2006. Hydrological models are so good, do we still need data? *Environmental Modelling and Software*, 21(9):1340-1352.

Smakhtin, V.U. 2000. Estimating daily flow duration curves from monthly streamflow data. *Water SA*, 26(1):13-18.

Smith, L., Mottiar, S., & Whiter, F. 2003. *Testing the limits of market-based solutions to the delivery of essential services: Nelspruit Water Concession*. [Online]: Available: http://www.cps.org.za/cps%20pdf/RR99.pdf [Accessed: 10 May 2008]

Smits, S., Pollard, S., du Toit, D., Butterworth, J., & Moriarty, P. 2005 *Modelling scenarios for water resources management in the Sand River Catchment, South Africa*. [Online]: Available: ftp://ftp.fao.org/agl/emailconf/wfe2005/WHiRL.pdf [Accessed: 10 May 2008]

Smith, C. S., Howes, A.L., Price, B., & McAlpine, C.A. 2007. Using a Bayesian Belief network to predict suitable habitat of an endangered mammal- The Julia Creek dunnart (Sminthopsis douglasi). *Biological Conservation*, (2007):333-347.

Snowling, S.D., & Kramer, J.R. 2001. Evaluating modelling uncertainty for model selection. *Ecological Modelling*, 138(1-3):17-30.

Soncini-Sessa, R., Castelletti, A., & Weber, E. 2003.A DSS for planning and managing water reservoir systems. *Environmental Modelling and Software*, 18(5):395-404.

South Africa. Council for Scientific and Industrial Research (CSIR), 2004. *2004 Eastern Cape State of the Environment Report.* [Online]. Available: http://www.environment.gov.za/soer/ecape/download_report.htm [Accessed: 10 December 2008]

South Africa. Department of Environment Affairs and Tourism (DEAT), 2008. *State of the Environment Report*. [Online]. Available: http://www.environment.gov.za/soer/reports/index.htm [Accessed: 10 November 2008]

South Africa. Department of Water Affairs and Forestry (DWAF), 2009. User Support Website: http://www.usersupport.co.za/index.php [Accessed: 10 May 2008]

South Africa. Department of Water Affairs and Forestry (DWAF), 1996.*South African Water Quality Guidelines: Volume 1: Domestic Use*. 2nd ed. Pretoria: Department of Water Affairs and Forestry.

South Africa. Department of Water Affairs and Forestry (DWAF), 1988 .*Water Quality Modeling, Volume A: Water Quality Calibration Model.* Pretoria: Department of Water Affairs and Forestry

South Africa. Department of Water Affairs and Forestry (DWAF), 2005 .*Groundwater Resource Assessment II: 3bE – Groundwater-Surface water interactions*. Pretoria: Department of Water Affairs and Forestry

South Africa. Department of Water Affairs and Forestry (DWAF), 2002a. *National Water Resource Quality Status Report: Inorganic Chemical water quality of surface water resources in SA- the big picture*. [Online]. Available: http://www.dwaf.gov.za/iwqs/water_quality/NCMP/RepNatA.htm [Accessed 10 July 2008]

South Africa. Department of Water Affairs and Forestry. (DWAF), 2002b. *Mzimvubu to Keiskamma Water Management Area: Water Resources Situation Assessment: Main Report: Volume 1*. Pretoria: Department of Water Affairs and Forestry.

South Africa. Department of Water Affairs and Forestry. (DWAF), 2004. *Development of an Internal Strategic Perspective for the Amatole – Kei Area of the Mzimvubu to*

Keiskamma Water Management Area (Wma No. 12). [Online]. Available: http://www.dwaf.gov.za/Documents/Other/WMA/12/AmatoleKeiISPAug04Intro.pdf [Accessed: 10 August 2007]

South Africa. Department of Water Affairs and Forestry (DWAF), 2008 .*WARMS- Water Use Registering and Licensing.* [Online]. Available: http://www.dwaf.gov.za/Projects/WARMS/default.asp [Accessed: 10 December 2008]

South Africa. Department of Water Affairs (DWA), 2010 .*Hydrological Services-Surface Water (Data, Dams and Flow Information)* [Online]. Available: http://www.dwa.gov.za/Hydrology/ [Accessed: 10 December 2008]

South Africa. Department of Environmental Affairs. (DEAT), 2005. State of the environment. [Online]. Available: http://soer.deat.gov.za/State_of_the_Environment.html [Accessed: 10 January 2009]

Stanski, H.R., Wilson, L.J., & Burrows W.R. 1989. *Survey of common verification methods in meteorology*. [Online]. Available: http://www.cawcr.gov.au/projects/verification/Stanski_et_al/Stanski_et_al.html [Accessed: 1 December 2008]

Stassopoulou, A., Petrou, M., & Kittler, J. 1998. Application of a Bayesian Network in a GIS Based Decision Making System. *International Journal of Geographical Information Science*, 12 (1):23-46.

Stave, K.A. 2003. A system dynamics model to facilitate public understanding of water management options in Las Vegas, Nevada. *Journal of Environmental Management*, 67(4):303-313.

Steck, H., & Jaakkola, T.S. 2003. *Predictive Discretization during Model Selection.* [Online]. Available: http://dspace.mit.edu/handle/1721.1/6709 [Accessed: 10 May 2010]

Tarboton, K.C., & Schulze, R.E. 1992. *Distributed hydrological modelling system for the Mgeni catchment*. Pretoria: Water Research Commission Report No. 234/1/92.

Tattari, S., Schultz, T., & Kuussaari, M. 2003. Use of belief network modelling to assess the impact of buffer zones on water protection and biodiversity. *Agriculture, Ecosystems and Environment*, 96(1):119-132.

Thomas, E.P., Seager, J.R., & Mathee, A. 2002. Environmental health challenges in South Africa: policy lessons from case studies. *Health & Place,* 8(4):251-261.

Ticehurst, J.L., Newham, L.T.H., Rissik, D., Letcher, R.A., & Jakeman, A.J. 2007. A Bayesian network approach for assessing the sustainability of coastal lakes in New South Wales, Australia. *Environmental Modelling and Software*, 22(8):1129-1139.

Umakhanthan, U. 2002. *Estimation of the spatio-temporal heterogeneity of rainfall and its importance towards robust catchment simulation, within a hydroinformation environment*. Sydney, Australia. University of New South Wales. (PhD-thesis.)

United Nations Economic Commission for Africa (UNECA). 2001. *State of the Environment in Africa.* [Online]. Available: www.uneca.org/water/State_Environ_Afri.pdf [Accessed: 10 July 2008].

United States Environmental Protection Agency (EPA). 2006. *Glossar*y. [Online]. Available: http://www.epa.gov/reva/glossary.htm [Accessed: 10 June 2008]

Uusitalo, L. 2007. Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, 203(3-4):312-318.

van Asselt B.A.M., & Rotmans, J. 2002. Uncertainty in integrated assessment modelling. *Climate Change*, 54 (1-2):75-105, 2002.

van der Sluijs, J. 1996. Integrated Assessment Models and the Management of Uncertainties. [Online]. Available: http://www.iiasa.ac.at/Publications/Documents/WP-96-119.pdf [Accessed: 10 June 2008]

van Kerkhoff, L. 2005. Integrated research: concepts of connection in environmental science and policy. *Environmental Science & Policy*, 8(5):452-436.

Van Wyk, E., van Wilgen, B.W., & Roux, D.J. 2001. How well has biophysical research served the needs of water resource management? Lessons from the Sabie-Sand catchment. *South African Journal of Science*, 97:349-356.

Vanclooster, M., Romanowicz, A.A., Paniconi, C., Troch, P., Holman, I., & and Marechal, D. 2002. *Meta-hydrological Modelling Approaches Supporting the Implementation of Water Management Decision Support Systems*. [Online]. Available: http://siti.feem.it/mulino/dissem/intcon/vancloo.pdf [Accessed: 10 June 2008]

Varis, O., & Kuikka, S. 1999. Learning Bayesian decision analysis by doing: lessons from environmental and natural resources management. *Ecological Modelling*, 119 (2):177-195.

Vegter, J.R. 1995. *An explanation of a set of National Groundwater maps*. Pretoria: Water Research Commission Report No. TT74/95.

Wagener, T. 2003. Evaluation of catchment models. *Hydrological Processes*, 17(16):3375-3378.

Walker, J., Dowling, T. & Veitch, S. 2006. Assessment of catchment condition in Australia. *Ecological Indicators*, 6(1):205-214.

Walmsley, J., Carden, M., Revenga, C., Sagona, F., & Smith, M. 2001. Indicators of sustainable development for catchment management in South Africa - Review of indicators from around the world. *Water SA*, 27(4):539-550.

Walmsley, J.J. 2002. Framework for measuring sustainable development in catchment systems. *Environmental Management, 29*(2):195-206.

Walmsley, R.D., Walmsley, J.J.I., & Walmsley, C. 2004. *Testing and development of catchment sustainability indicators*. Pretoria: Water Research Commission. Report No. KV 156/04.

Wang, H. 2004. *Building Bayesian Networks: Elicitation, Evaluation and Learning*. Pittsburgh, University of Pittsburgh (PhD-thesis.)

Wiens, J.A. 1989. Spatial Scaling in ecology. *Functional Ecology*, 3(4):385-397.

Wong, M.L., & Leung, K.S. 2004. An efficient data mining method for learning Bayesian Networks using an evolutionary algorithm-based hybrid approach. *IEEE Transactions on Evolutionary Computation*, 8(4):378-404.

Wong, T, & Wong, H. 1996. *Genetic Algorithms*. [Online]. Available: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/hmw/article1.html [Accessed: 30 April 2008]

World Wildlife Fund (WWF). 2007. *Integrated River Basin Management*. [Online]. Available: http://www.panda.org/about_wwf/what_we_do/freshwater/our_solutions/rivers/irbm/index.cfm [Accessed: 20 November 2007]

Woyessa, Y.E., Hensley, M., & van Rensburg, L.D. 2006. *Catchment management in semi-arid area of central South Africa: Strategy for improving water productivity*. [Online]. Available: http://www.wrc.org.za/downloads/watersa/2006/WISA%20special%20ed/5.pdf [Accessed: 10 December 2008]

Woyessa, Y.E., Pretorius, E., & van Heerden, P. 2004. *The application of the SAPWAT model in irrigation water management planning for the Sand-Vet irrigation scheme:*

contribution towards an integrated catchment management system. [Online]. Available: http://www.ewisa.co.za/literature/files/258.pdf [Accessed: 10 December 2008]

Xenos, M. 2004. Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers & Education*, 43(4):345-359.

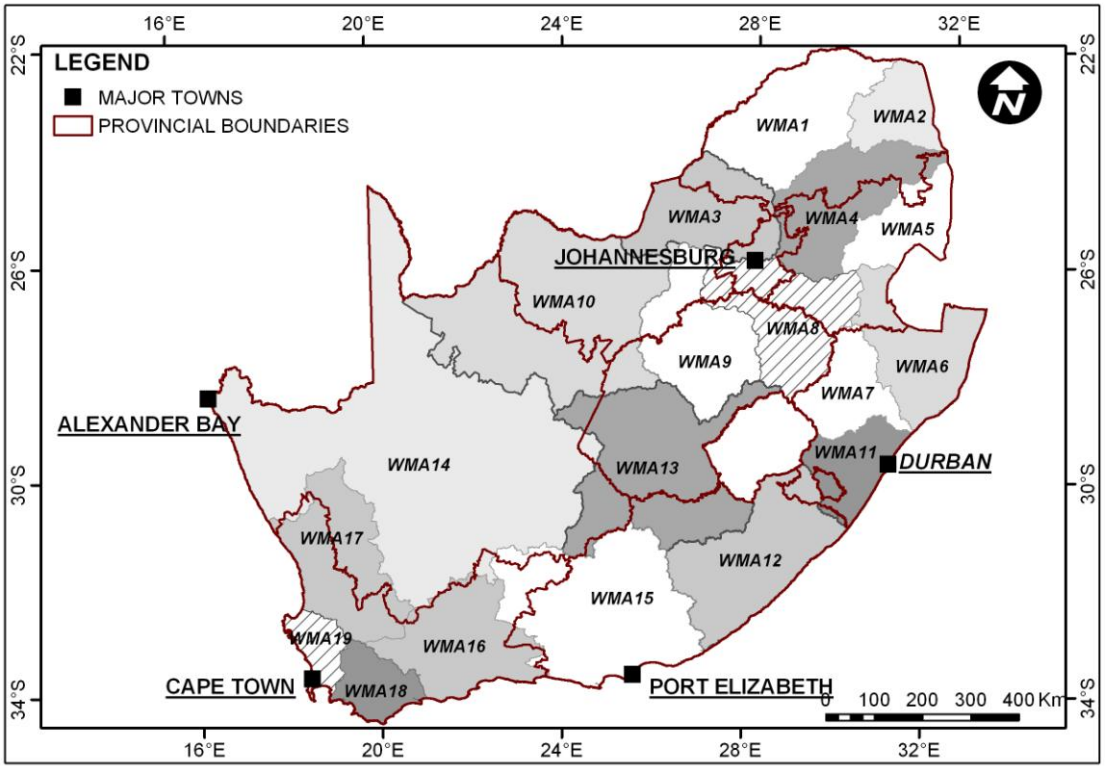
Zhang, H., & Casey, T. 2000. Verification of Categorical Probability Forecasts. *Weather and Forecasting*, 15(1): 80-89.

APPENDIX A: CATCHMENTS IN SOUTH AFRICA

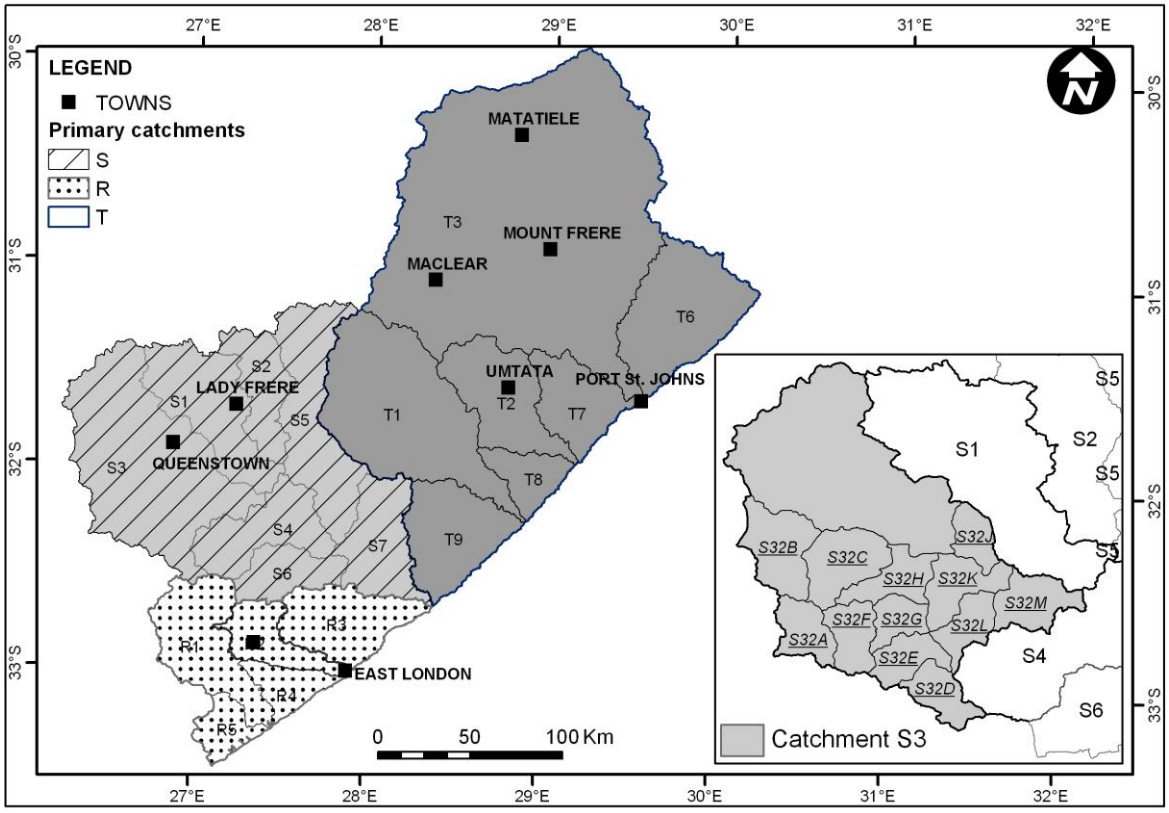
A water management area is defined in the National Water Act No. 36 of 1998 as:

> *"an area established as a management unit in the national water resource strategy within which a catchment management agency will conduct the protection, use, development, conservation, management and control of water resources," National Water Act 36 of 1998.*

The Minister of Water Affairs and Forestry has the mandate to establish these water management areas and has identified and set up nineteen to cover South Africa (Department of Water Affairs and Forestry 1999). The demarcation of the WMAs was based on catchment boundaries, socio-economic development patterns, operational efficiency and the interests of the public and communities in that area. Meetings and discussions amongst various stakeholders preceded the finalisation of the WMAs.



The grouping of the water management areas is based on primary catchments. The primary catchments are further divided into secondary catchments, which are partitioned into tertiary catchment and these into quaternary catchments.



The quaternary catchments have been selected to have similar runoffs: the greater the runoff volume, the smaller the catchment area and vice versa. The quaternary catchments are numbered alpha-numerically in downstream order (Department of Water Affairs and Forestry, 2004b). A quaternary catchment number, for example S32D, may be interpreted as follows:

> *"The letter S denotes Drainage Region S (referred to as a primary catchment). The number 3 denotes secondary catchment 3 of Drainage Region S. The number 2 shows that the secondary catchment has been sub-divided into tertiary catchments and this one is number 2. The letter D shows that the quaternary catchment is the fourth in sequence downstream from the head of secondary catchment S3"*

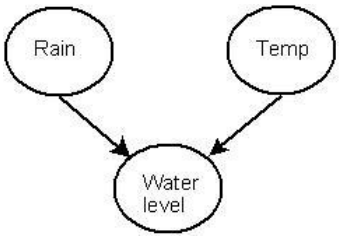
APPENDIX B: BAYESIAN NETWORK NUMERICAL EXAMPLES

This example was adapted from Cowell *et al.,* 1999 is provided to demonstrate the following concepts of reasoning in Bayesian Networks:

a) conditional probability;
b) Bayes' Rule;
c) the chain rule of joint probabilities; and
d) assumed independence amongst variables.

Given the following information

The Bayesian Network illustrating the relationship between rainfall (rain), temperature (temp) and groundwater level (water level):



Let us represent the variables with the following letters:

Rain = *R*, Temp = *T*, Water Level =*W*

Given the following marginal probabilities:

P (R =high) = 0.1 i.e. the probability that rainfall is high

P (T = high) = 0.2 i.e. the probability that temperature is high

You also have the following conditional probabilities [*P (W=high | R,T)*]:

P (W = high | R=low, T=low) = 0

P (W = high | R=low, T=high) = 0.5

P (W = high | R=high, T=low) = 1

P (W = high | R=high, T=high) = 1

After measuring the groundwater level of a specific borehole and finding that it is high, i.e. *W = high*, it is required to find the conditional probabilities for rainfall and temperature.

Calculations

a) Using probability axioms

P (R =low) =1- 0.1 = 0.9

P (T = low) =1- 0.2 = 0.8

b) Using Bayes' Theorem, the required conditional probabilities are represented by the following formula:

$$P(R, T \mid W=high) = \frac{P(W = high \mid R,T) \times P(R,T)}{P(W = high)}$$

c) Using conditional probabilities and independence between variables:

P (R, T) = P (R) × P (T | R) (using formula given in Equation 4-7)

But using Equation 4-6 and conditional independence of rain (R) and temperature (T) as defined in Section 4.3:

P (R, T) = P (R) × P (T), from this, it follows that:

P (R=low, T=low) = P (R=low) × P (T=low) = 0.9 × 0.8 = 0.72

P (R=low, T=high) = P (R=low) × P (T=high) = 0.9 × 0.2 = 0.18

P (R=high, T=low) = P (R=high) × P (T=low) = 0.1 × 0.8 = 0.08

P (R=high, T=high) = P (R=high) × P (T=high) = 0.1 × 0.2 = 0.02

d) Using Equation 4-5 which illustrates the concept of marginal probability and using conditional independence between variables, *P (W=high)* can be calculated as follows:

P (W=high, R,T) = P (W=high | R,T) × P (R | T) × P (T)

= P (W=high | R,T) × P (R) × P (T)

i.e. because R is independent from T.

P (W=high | R,T) has already been provided. As an example:

For R = low and T =high

P (W=high, R, T) = 0.5×0.9×0.2 =0.09

Using marginalisation, P (W=high) can be obtaining by summing all occurrences of W= high to give 0.19. The rest of the results are presented in Table 4-6.

Results of probability calculations.

R[P(R)]	low [0.9]		high [0.1]		Total	Comments
T[P(T)]	low[0.8]	high [0.2]	low[0.8]	high[0.2]		
P(R,T)	0.72	0.18	0.08	0.02	1	Calculated
P (W=high | R,,T)	0	0.5	1	1		Provided
P (W=high, R,T)	0	0.09	0.08	0.02	0.19	Calculated
P (R, T | W=high)	0	0.47	0.42	0.11	1	Final answer

This means that if high groundwater level has been measured, the following can be deduced from the calculations:

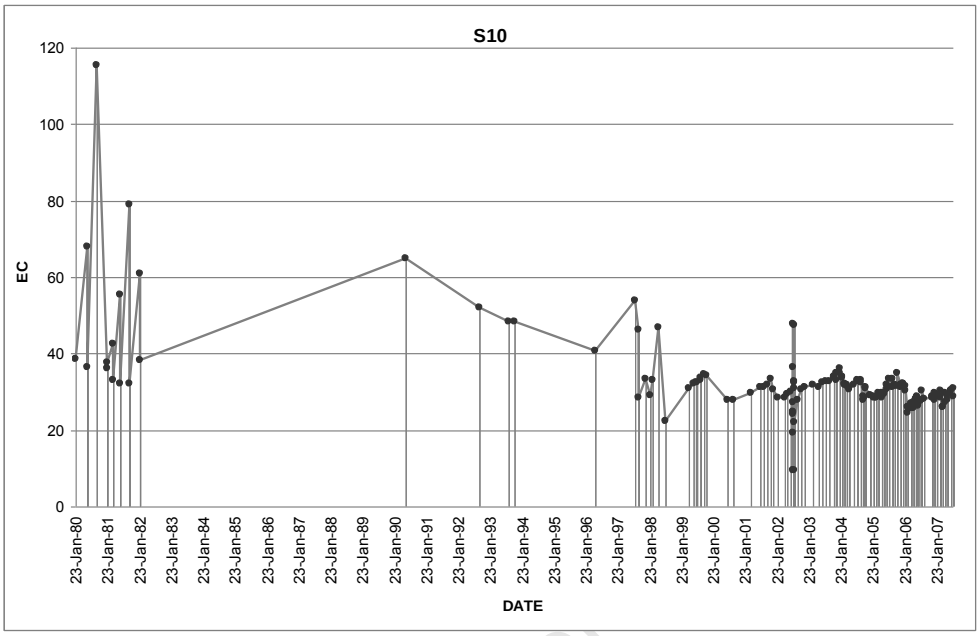
The probability that rainfall is high = 0.42+0.11=0.53

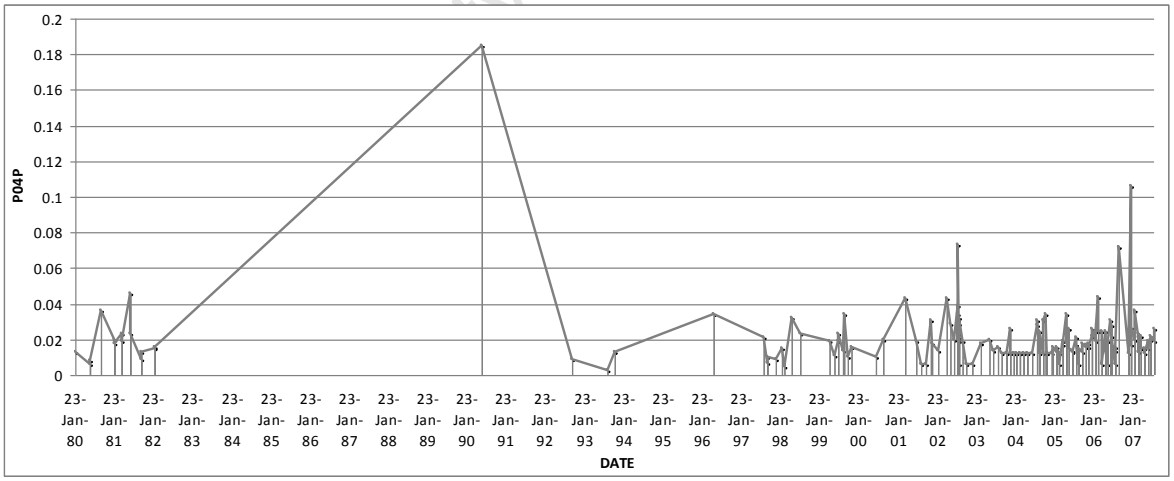
The probability that temperature is high = 0.47+0.11=0.58

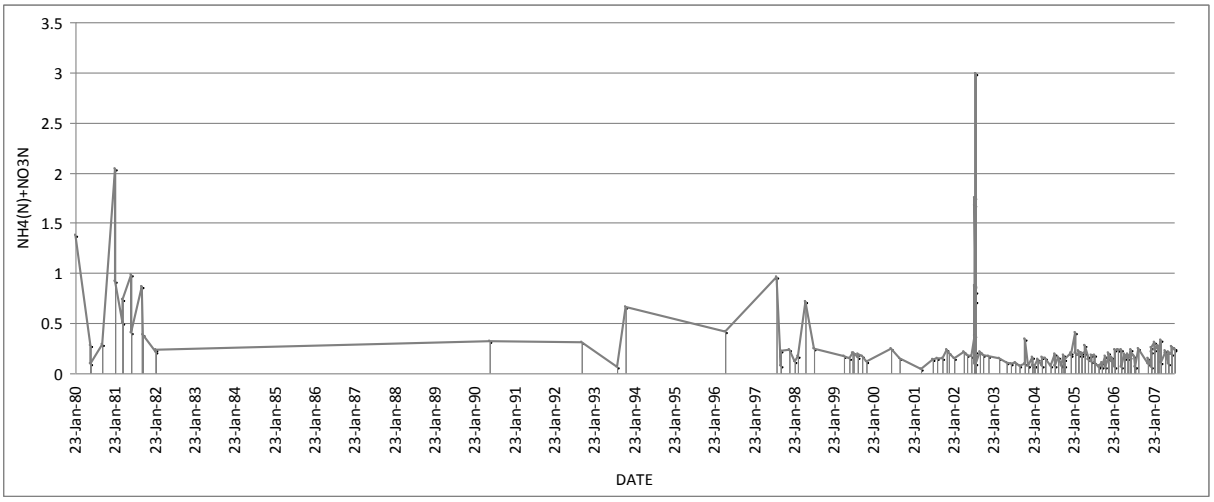
APPENDIX C: SURFACE AND GROUNDWATER QUALITY DATA

SURFACE WATER QUALITY GRAPHS

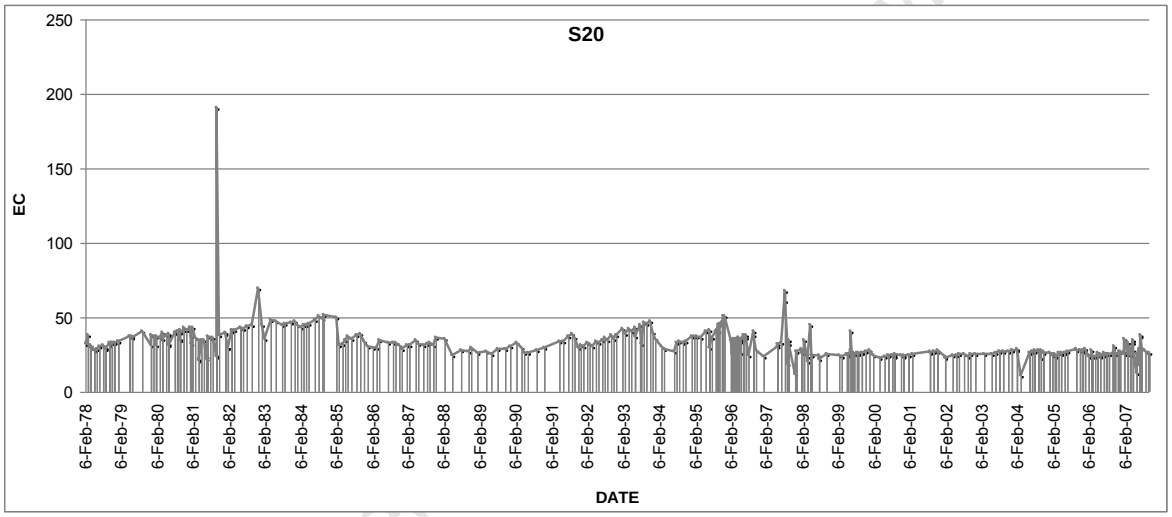
REGION S10

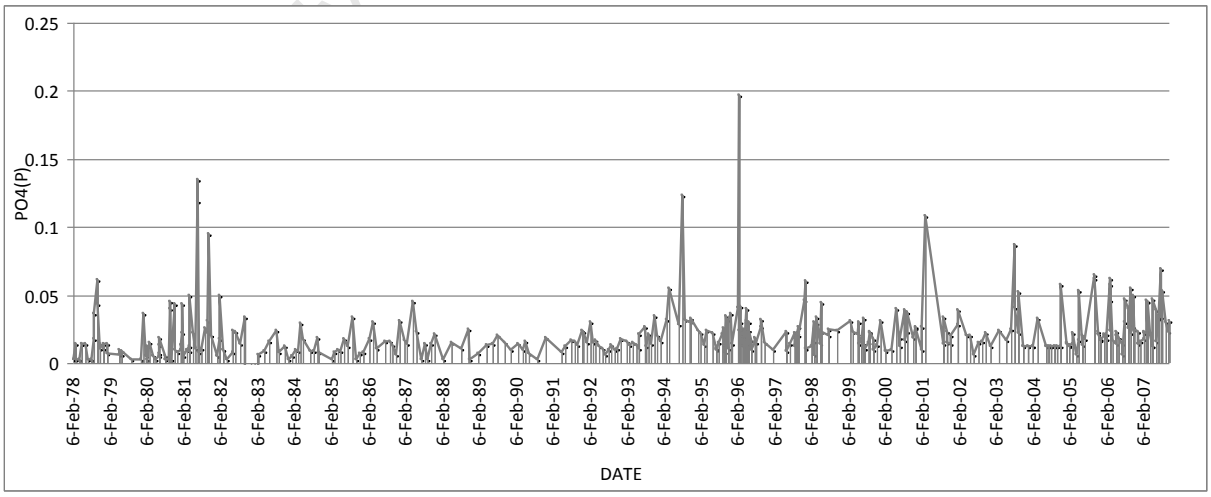


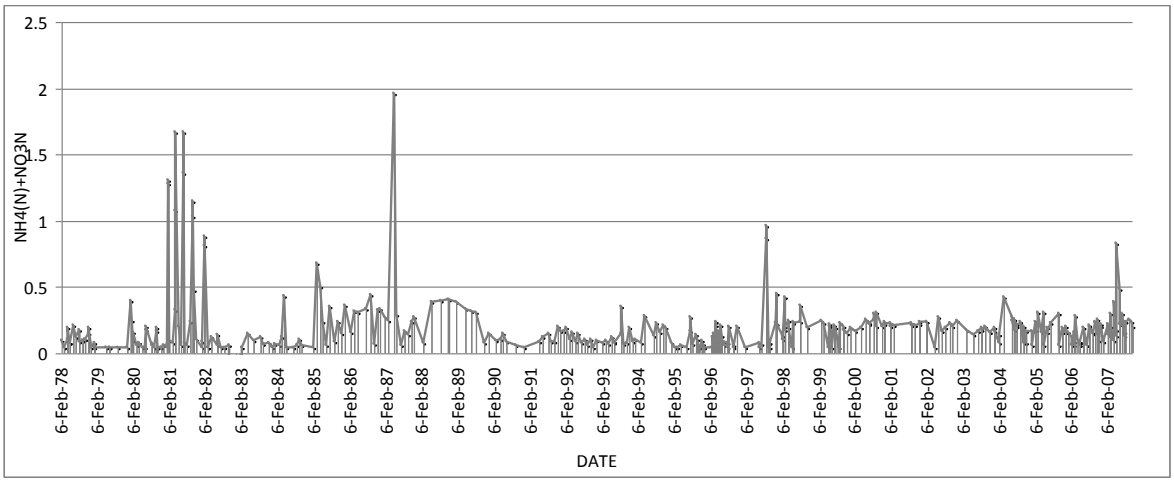




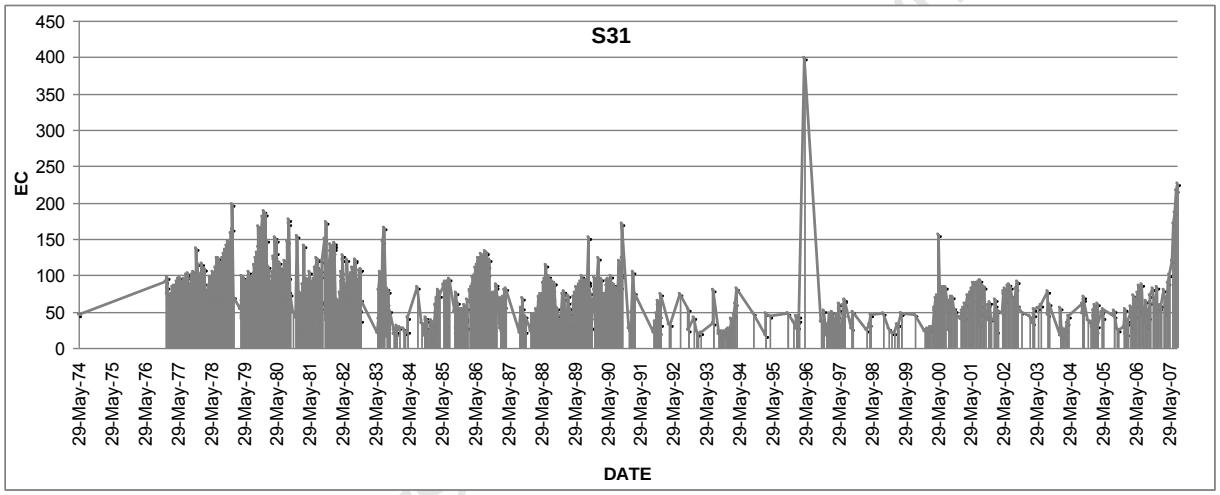
REGION S20

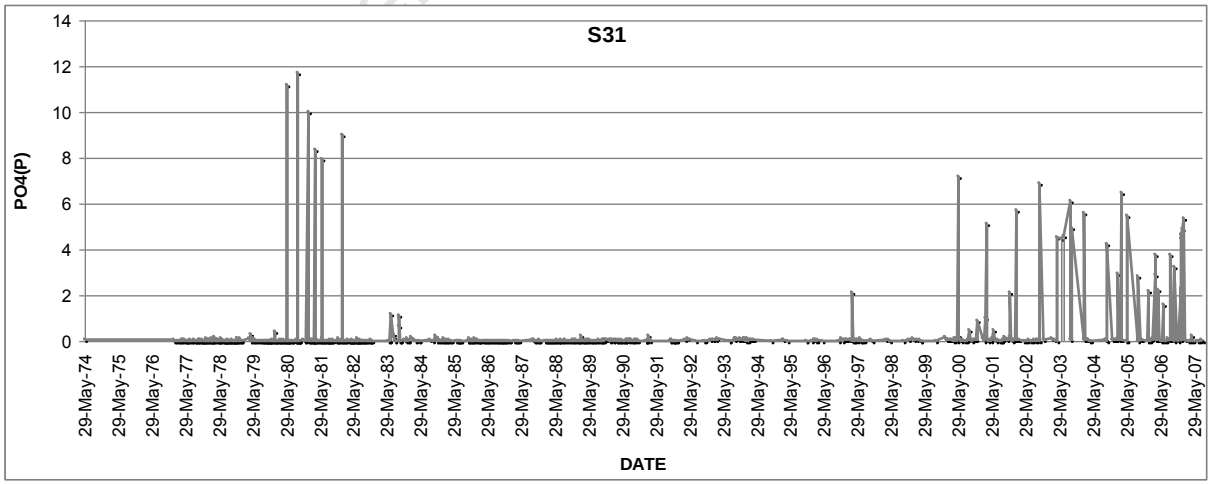


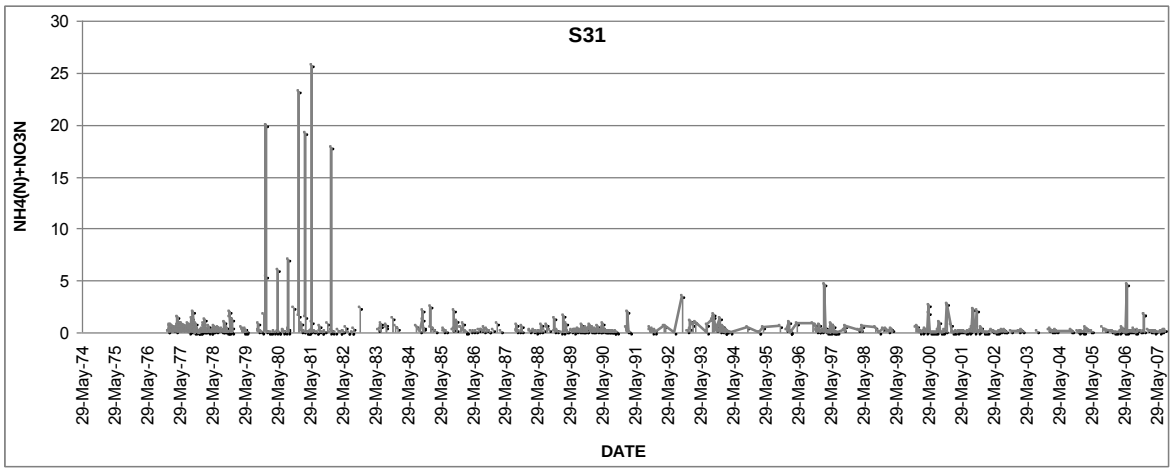




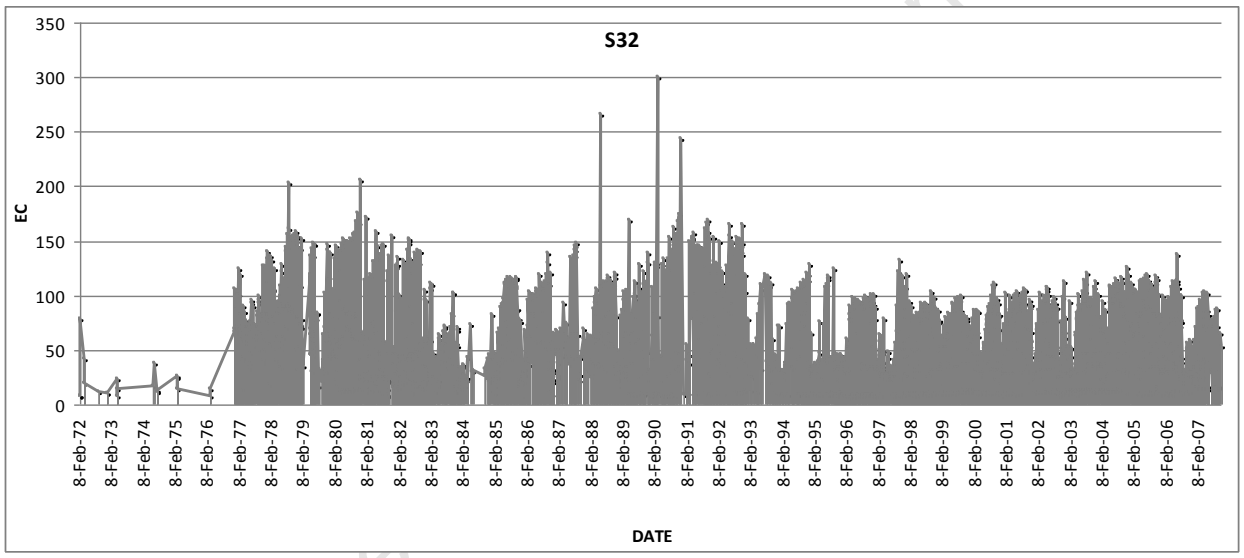
REGION S31

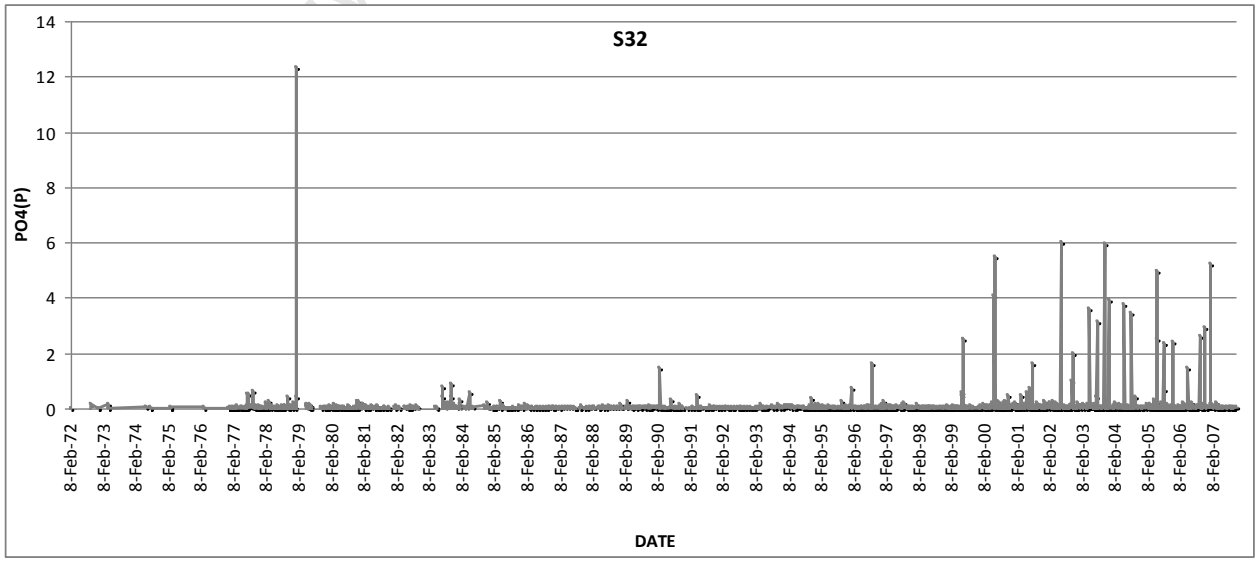


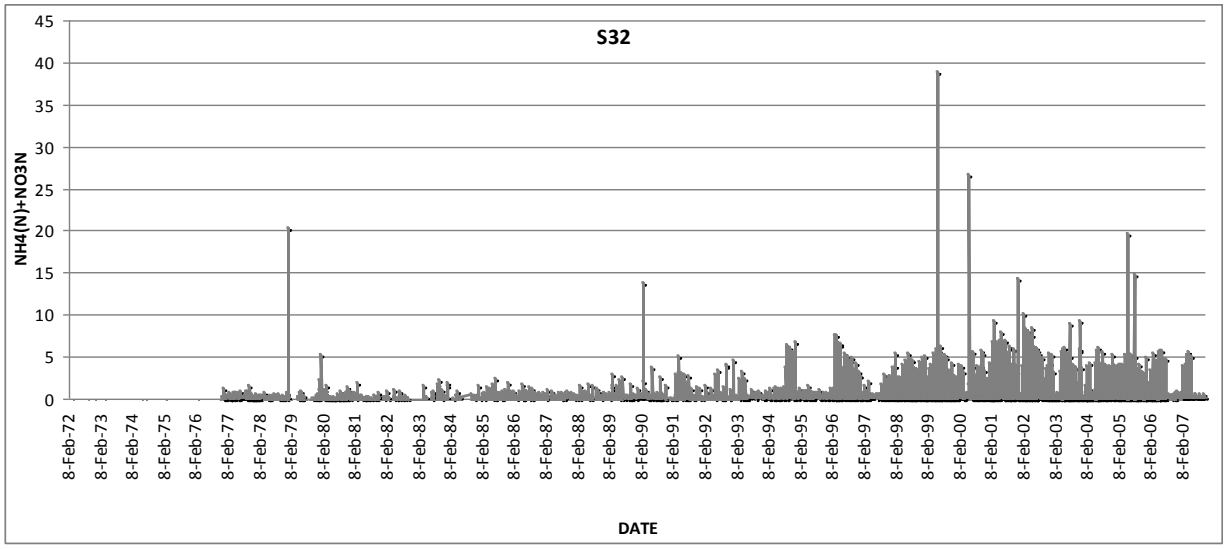




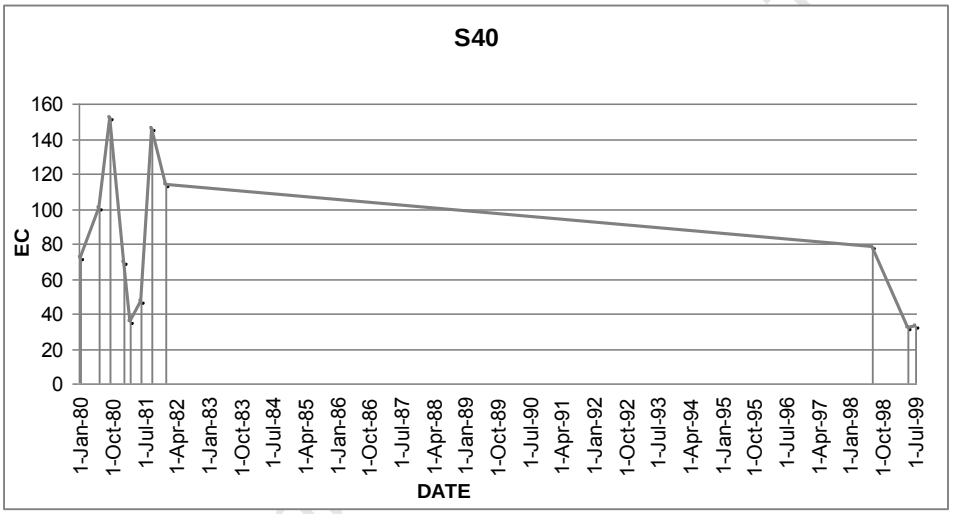
REGION S32

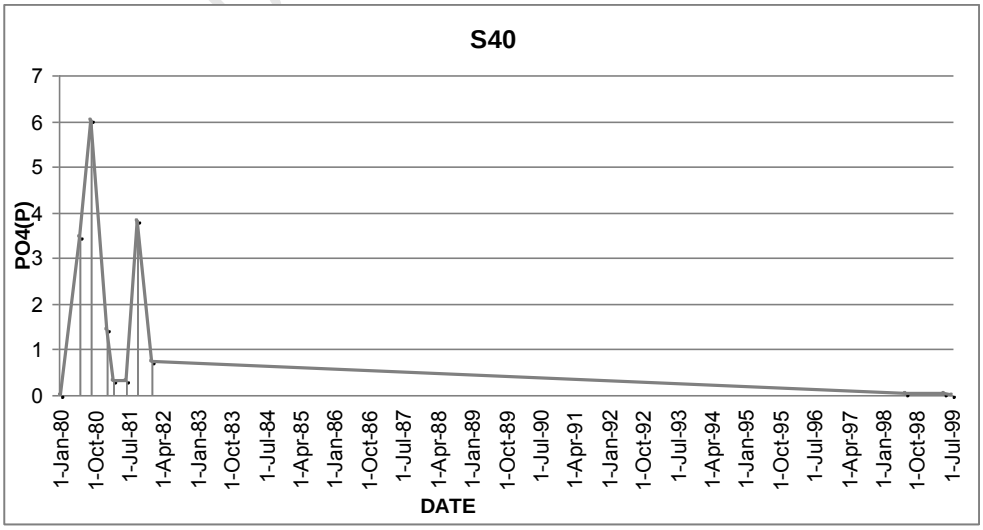


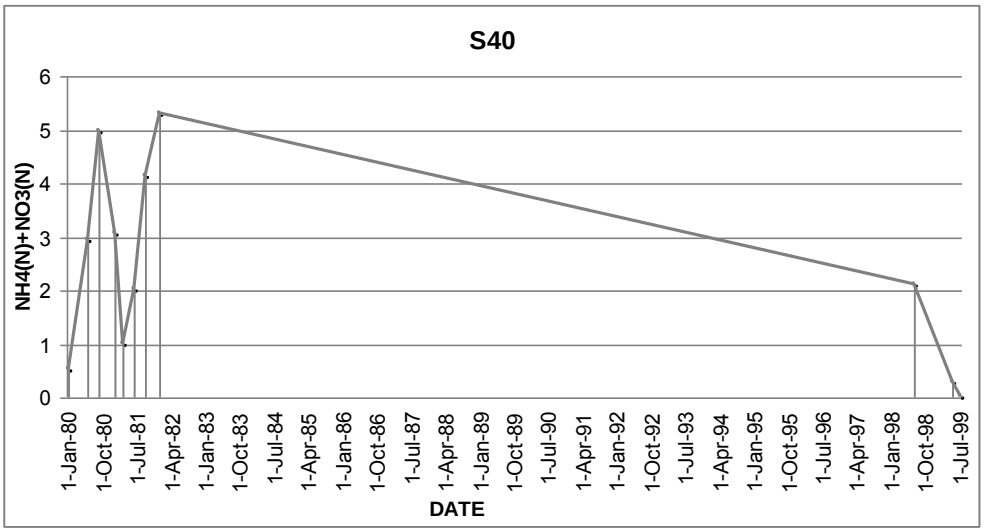




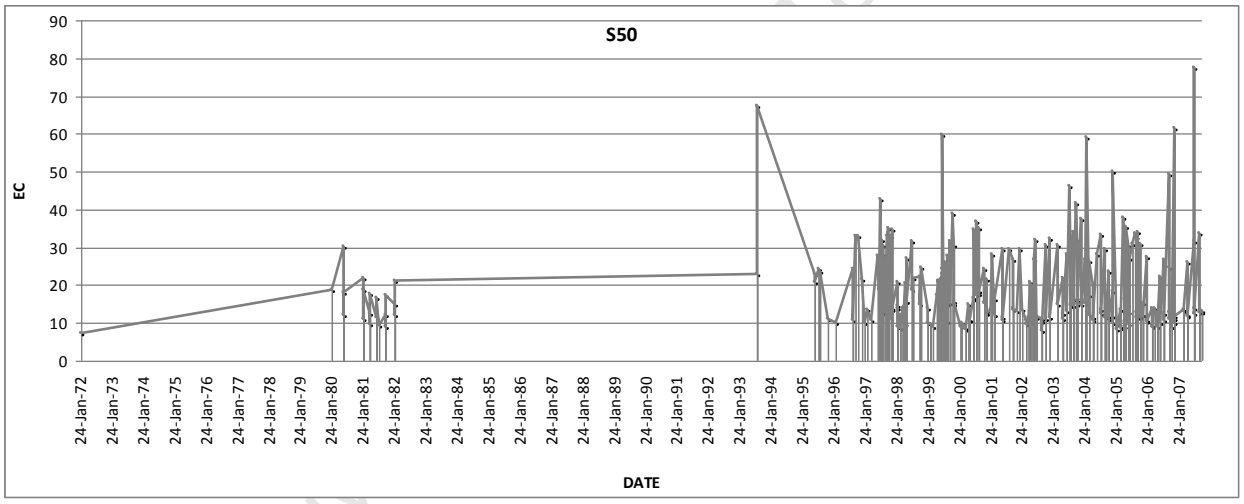
REGION S40

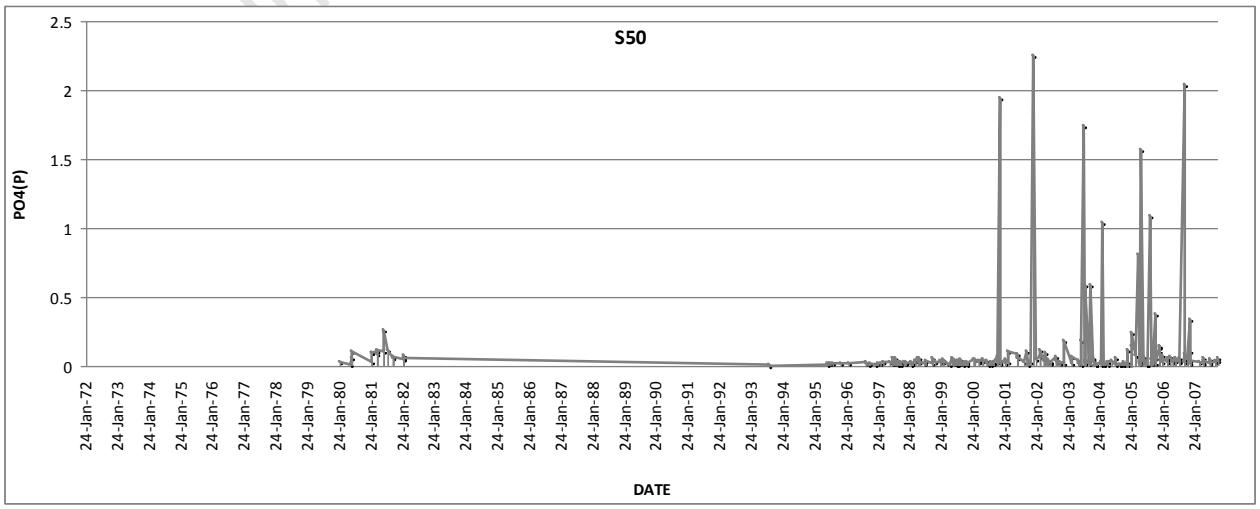


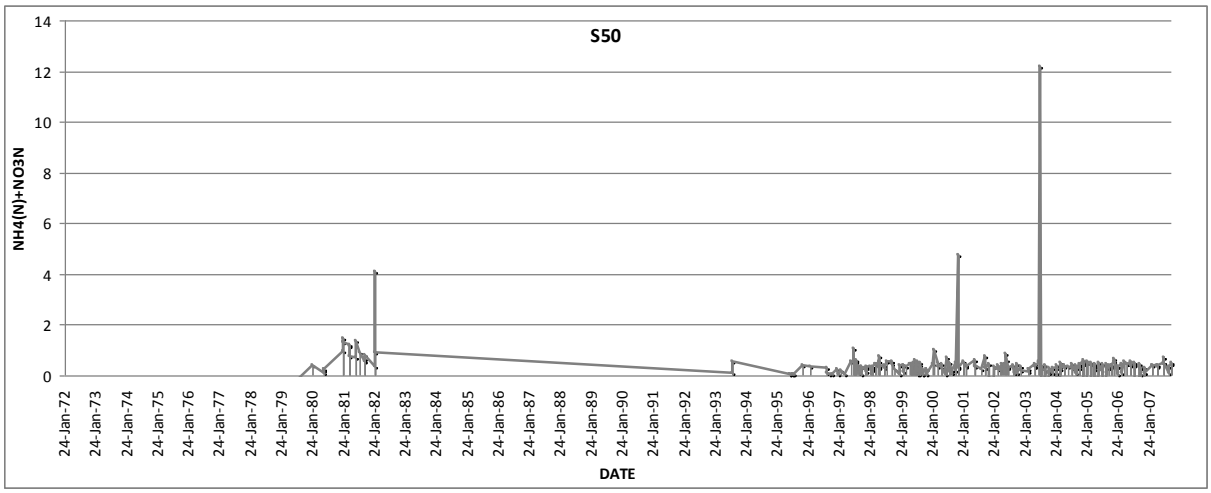




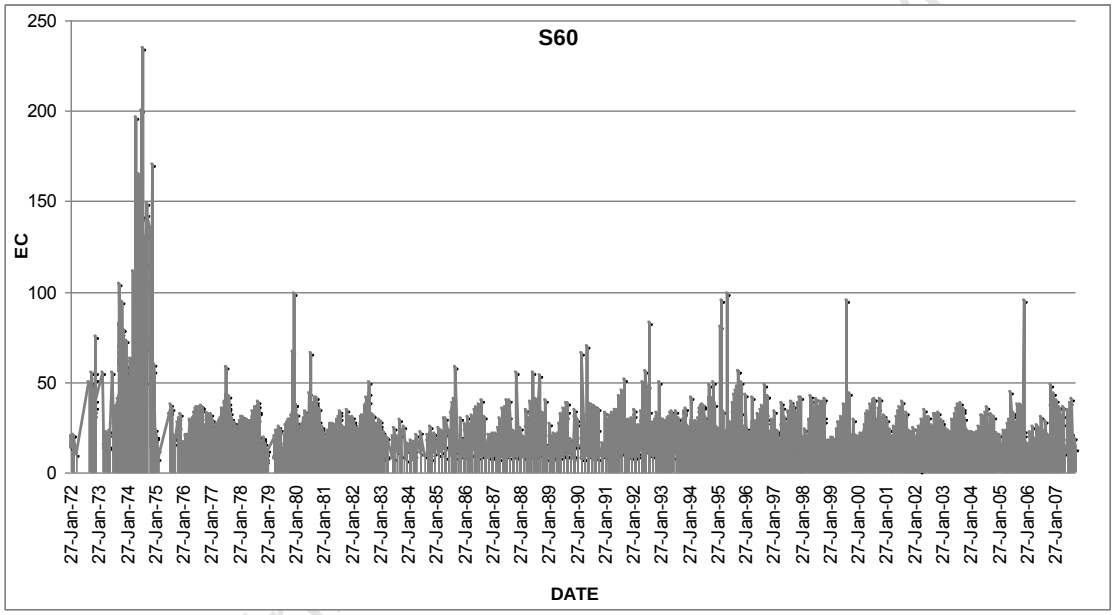
REGION S50

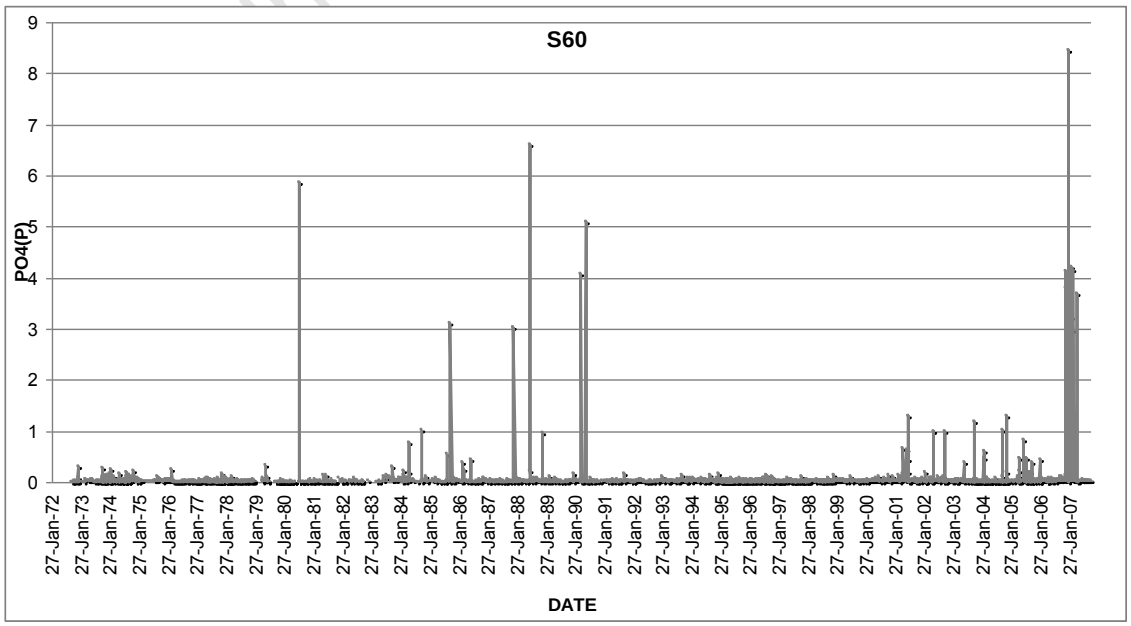


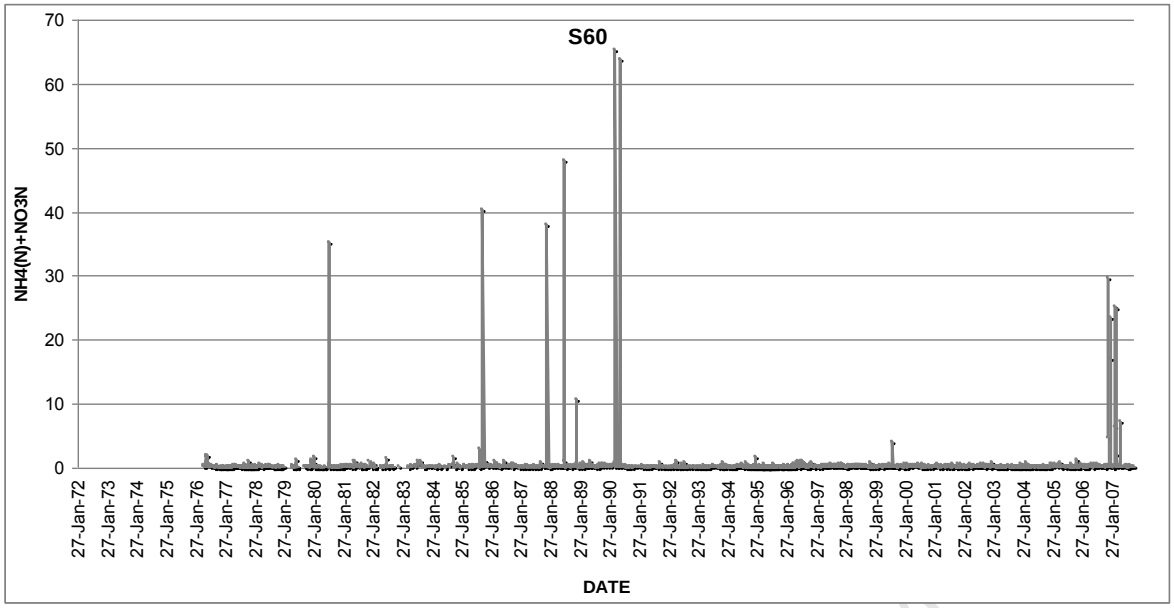




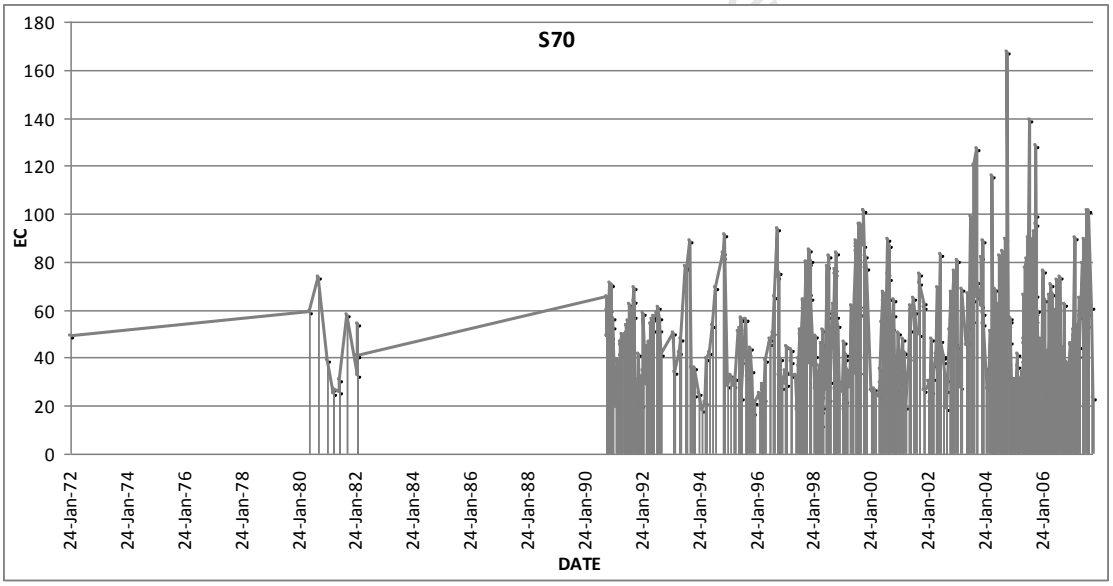
REGION S60

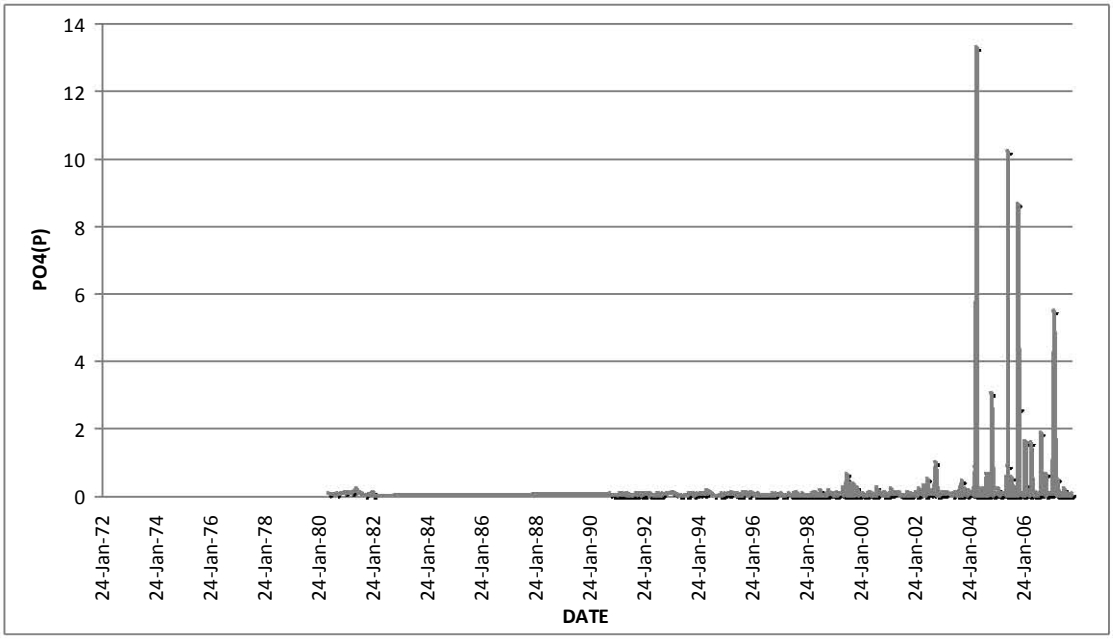


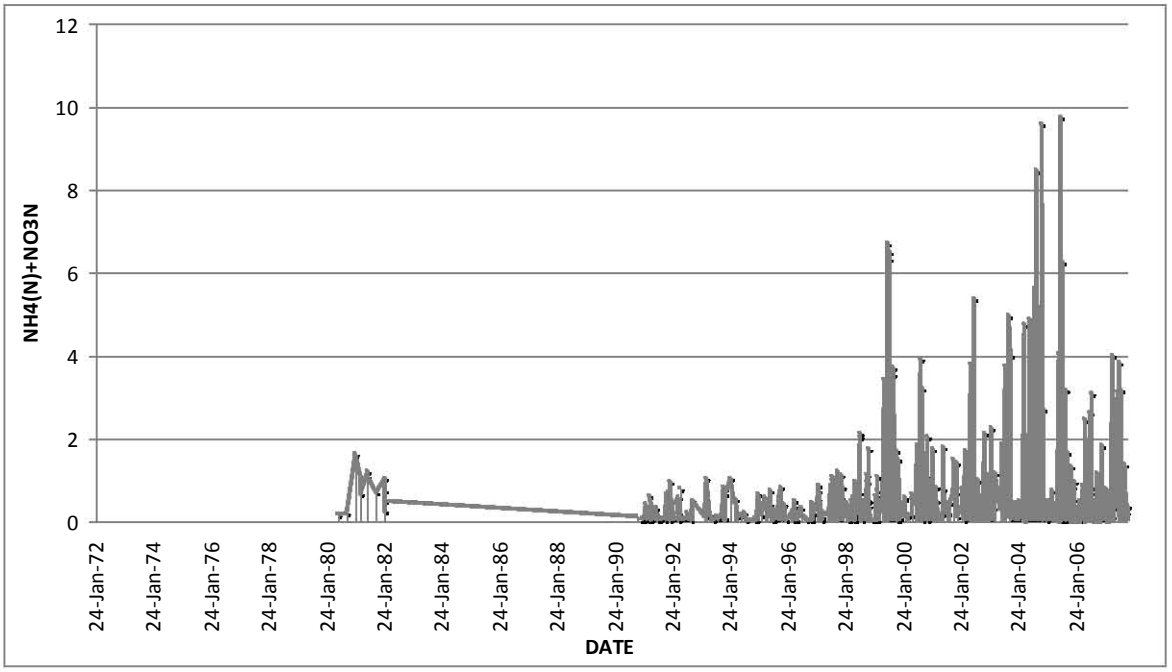




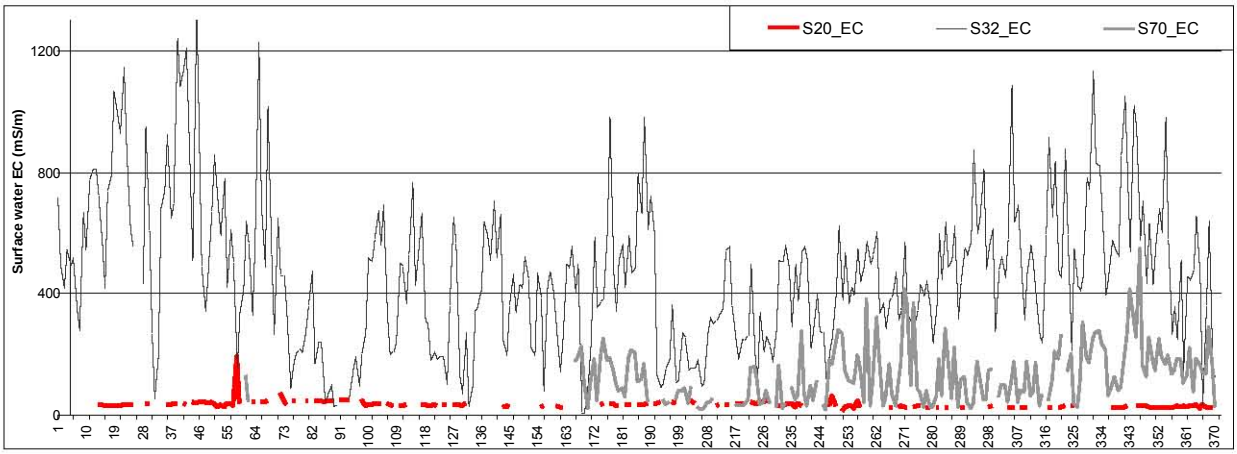
REGION S70

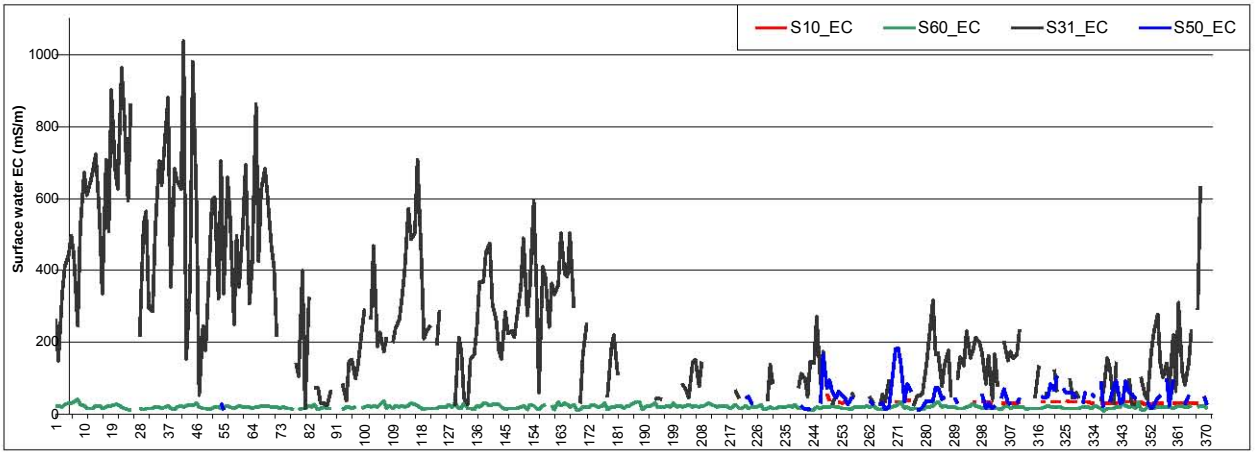



PO4(P)
14
12
10
8
6
4
2
0
24-Jan-72
24-Jan-74
24-Jan-76
24-Jan-78
24-Jan-80
24-Jan-82
24-Jan-84
24-Jan-86
24-Jan-88
24-Jan-90
24-Jan-92
24-Jan-94
24-Jan-96
24-Jan-98
24-Jan-00
24-Jan-02
24-Jan-04
24-Jan-06
DATE


NH4(N)+NO3N
12
10
8
6
4
2
0
24-Jan-72
24-Jan-74
24-Jan-76
24-Jan-78
24-Jan-80
24-Jan-82
24-Jan-84
24-Jan-86
24-Jan-88
24-Jan-90
24-Jan-92
24-Jan-94
24-Jan-96
24-Jan-98
24-Jan-00
24-Jan-02
24-Jan-04
24-Jan-06
DATE

DATA USED FOR SPATIAL PREDICTION OF EC





APPENDIX D: STATES AND CPTs OF THE VARIABLES

Sandstone:

Threquasan	VeryLow	Low	Medium	High
VeryLow	0.34	1	1	1
Low	0.14	0	0	0
Medium	0.37	0	0	0
High	0.14	0	0	0

Alien vegetation:

Region	S10	S20	S31	S32	S40	S50	S60	S70
VeryLow	1	1	1	0.92	0.083	0.92	0.92	0.92
Low	0	0	0	0	0.42	0	0	0
Medium	0	0	0	0.083	0.17	0.083	0.084	0.084
High	0	0	0	0	0.17	0	0	0
VeryHigh	0	0	0	0	0.17	0	0	0

Temperature:

VeryLow	0.13
Low	0.31
Medium	0.4
High	0.17

Runoff:

Region	S10	S20	S31	S32	S40	S50	S60	S70
Class1	0.61	0.56	0.72	0.45	0.6	0.25	0.35	0.22
Class2	0.13	0.19	0.11	0.15	0.13	0.24	0.21	0.28
Class3	0.18	0.2	0.14	0.27	0.19	0.32	0.31	0.36
Class4	0.049	0.042	0.026	0.081	0.054	0.094	0.096	0.07
Class5	0.017	0.0076	0.0052	0.035	0.021	0.054	0.028	0.043
Class6	0.01	0.0025	0.0012	0.017	0.0092	0.045	0.0065	0.032

Groundwater use:

WaterAvail	VeryLow				Low				Medium				High			
Irri_area	VeryLow	Low	Medium	High	VeryLow	Low	Medium	High	VeryLow	Low	Medium	High	VeryLow	Low	Medium	High
VeryLow	0.6	0	0	0	0.15	0	0	0	0.52	0.21	0	0	0.67	0	0	0
Low	0.4	0	0	1	0.31	0	0	0.33	0.3	0.37	0	0.38	0.33	0	0	0
Medium	0	0.5	0.5	0	0.39	1	0.62	0.33	0.13	0.053	0	0.13	0	0	0	0.67
High	0	0.5	0.5	0	0.15	0	0.38	0.34	0.042	0.37	1	0.5	0	1	1	0.33

Water available:

AlienVeg	VeryLow					Low					Medium					High					VeryHigh				
Urbanisation	Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5
VeryLow	0.096	0	0	0	0.37	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Low	0.21	0	0	0	0	1	0.5	0.5	0.5	0	0.4	0.5	0.5	0.5	0	1	0.5	0.5	0.5	0.5	1	0.5	0.5	0.5	0.5
Medium	0.57	0.5	1	1	0.63	0	0.5	0.5	0.5	0	0.6	0.5	0.5	0.5	0	0	0.5	0.5	0.5	0.5	0	0.5	0.5	0.5	0.5
High	0.13	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Rainfall:

Temperature	VeryLow	Low	Medium	High
Class1	0.45	0.33	0.12	0.072
Class2	0.24	0.22	0.14	0.14
Class3	0.18	0.2	0.22	0.26
Class4	0.1	0.16	0.32	0.36
Class5	0.017	0.048	0.13	0.14
Class6	0.0054	0.037	0.073	0.033

GW_SAR:

Urbanisation	Class1	Class2	Class3	Class4	Class5
VeryLow	0.23	0.5	0	0	0.083
Low	0.29	0	0	1	0.17
Medium	0.21	0	0	0	0.33
High	0.27	0.5	1	0	0.42

Urbanisation:

Region	S10	S20	S31	S32	S40	S50	S60	S70
Class1	1	0.5	0.67	0.83	0.75	0.75	0.67	0.92
Class2	0	0	0	0	0	0.083	0	0.084
Class3	0	0	0	0.084	0	0.17	0	0
Class4	0	0.5	0	0	0	0	0	0
Class5	0	0	0.33	0.083	0.25	0	0.33	0

Irrigated_area:

WaterAvail	VeryLow					Low					Medium					High				
SaniAccess	Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5
VeryLow	1	1	0.25	0	0	0.62	0.7	0.33	0	0	0.66	0.13	0.29	0	0	0.83	0	0.5	0	0
Low	0	0	0	0.33	0.33	0.13	0.1	0	0.33	0.33	0.28	0.33	0.57	1	1	0	0.33	0	0	0.33
Medium	0	0	0	0.33	0.33	0	0	0	0.33	0.33	0.034	0.13	0	0	0	0	0.33	0	0	0.33
High	0	0	0.75	0.33	0.33	0.25	0.2	0.67	0.33	0.33	0.034	0.4	0.14	0	0	0.17	0.33	0.5	1	0.33

Degraded land:

AlienVeg	VeryLow	Low	Medium	High	VeryHigh	VeryLow	Low	Medium	High	VeryHigh	VeryLow	Low	Medium	High	VeryHigh	VeryLow	Low	Medium	High	VeryHigh	VeryLow	Low	Medium	High	VeryHigh
VeryLow	0.61	1	0.8	0	1	0.5	0	0	0	0	0.34	0	0	0	0	1	0	0	0	0	0.87	1	1	0	0
Low	0.16	0	0	1	0	0	0	0	0	0	0.66	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Medium	0.11	0	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
High	0.048	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0	0.5	0.5	0.5	0.5	0	0.5	0.5	0.5	0.5	0.13	0	0	0.5	0.5
VeryHigh	0.065	0	0	0	0	0	0.5	0.5	0.5	0.5	0	0.5	0.5	0.5	0.5	0	0.5	0.5	0.5	0.5	0	0	0	0.5	0.5

Water demand:

Irri_area	VeryLow					Low					Medium					High				
Urbanisation	Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5	Class1	Class2	Class3	Class4	Class5
VeryLow	0.83	0	0	0	0	0.77	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Low	0	0	0	0	0	0.23	0	0	0	0	1	0	0	0	0	0.87	0.5	0.5	0.5	0
Medium	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0.62	0.13	0.5	0.5	0.5	0
High	0.17	1	1	1	1	0	0	1	1	1	0	1	0	0	0.38	0	0	0	0	1

Water access:

SaniAccess	Class1				Class2				Class3				Class4				Class5			
Wat_demand	VeryLow	Low	Medium	High	VeryLow	Low	Medium	High	VeryLow	Low	Medium	High	VeryLow	Low	Medium	High	VeryLow	Low	Medium	High
Class1	0.93	1	0.88	0.81	0.18	0.18	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Class2	0.071	0	0.12	0.19	0.64	0.55	0	0.83	0.16	0.25	0.099	0.17	0	0	0	0	0	0	0	0
Class3	0	0	0	0	0.18	0.27	1	0.17	0.84	0.75	0.9	0.83	0	0	0	0	0	0	0	0
Class4	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0	1	0	0	0	0
Class5	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	1	0	1	1	1	1

Runoff:

Region	S10	S20	S31	S32	S40	S50	S60	S70
Class1	0.61	0.56	0.72	0.45	0.6	0.25	0.35	0.22
Class2	0.13	0.19	0.11	0.15	0.13	0.24	0.21	0.28
Class3	0.18	0.2	0.14	0.27	0.19	0.32	0.31	0.36
Class4	0.049	0.042	0.026	0.081	0.054	0.094	0.096	0.07
Class5	0.017	0.0076	0.0052	0.035	0.021	0.054	0.028	0.043
Class6	0.01	0.0025	0.0012	0.017	0.0092	0.045	0.0065	0.032

Population density:

Water_Access	Class1					Class2					Class3					Class4					Class5				
AlienVeg	VeryLow	Low	Medium	High	VeryHigh	VeryLow	Low	Medium	High	VeryHigh	VeryLow	Low	Medium	High	VeryHigh	VeryLow	Low	Medium	High	VeryHigh	VeryLow	Low	Medium	High	VeryHigh
VeryLow	0.27	1	0	1	0	0.52	1	1	0	1	0.75	0.5	1	0	1	0	0	0	0	0	0.5	0	0	0	0
Low	0.34	0	0	0	0.5	0.29	0	0	0.5	0	0.062	0.5	0	0.5	0	0	0.5	0.5	0.5	0.5	0	0.5	0.5	0.5	0.5
Medium	0.39	0	1	0	0.5	0.19	0	0	0.5	0	0.19	0	0	0.5	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
High	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0