A quantitative evaluation of the effectiveness of groundwater management plans

Emma Kathryn White B.A., B.Sc.

ORCID ID 0000-0003-0907-6135

PhD Thesis

The University of Melbourne

Department of Infrastructure Engineering

December 2019
But how will you look for something when you don’t in the least know what it is? How on earth are you going to set up something you don’t know as the object of your search?

Plato, The Meno

Let us settle ourselves, and work and wedge our feet downwards through the mud and slush of opinion, and prejudice, and tradition, and delusion, and appearance, that alluvion which covers the globe...till we come to a hard bottom and rocks in place, which we can call reality, and say, This is, and no mistake; and then begin...

Henry David Thoreau, Walden
Declaration of Published Work

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis. This thesis solely contains my original work towards the degree of Doctor of Philosophy. Any additional material utilised has been acknowledged in the text. This thesis is fewer than 100,000 words in length, excepting figures, tables, bibliographies and references, and includes the following multi-author papers published or submitted to peer reviewed journals.

- Chapter 3 consists of a paper published by *Water Resources Research* on 6/6/2016
- Chapter 4 consists of a paper published by *Hydrogeology Journal* on 4/7/2019
- Chapter 6 consists of a paper in preparation for *Groundwater* as of 15/12/2019

The core theme of this thesis is the evaluation of the effectiveness of groundwater management plans. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, working within the Department of Infrastructure Engineering under supervision by Prof Andrew Western, Dr Tim Peterson, Dr Elisabetta Carrara and Dr Justin Costelloe.

Emma Kathryn White, December 2019
Abstract

Groundwater is the world’s largest freshwater resource, constituting 96% of available freshwater. However, overexploitation due to poor or absent management in many regions of the globe has resulted in adverse environmental and socio-economic impacts. Groundwater management seeks to balance and mitigate the detrimental impacts of development, and management plans are commonly used to outline strategies to share and distribute water. But, plans are seldom systematically and quantitatively assessed for effectiveness and many plans are not conducive to a quantitative evaluation. A comprehensive framework for evaluating plan effectiveness is lacking in hydrogeology and currently, it is unknow how effective many plans are at achieving their stated objectives. Because plans are the primary means of managing groundwater, it is crucial they are efficient and effective. The objective of this thesis was to develop a methodology to quantitatively evaluate the effectiveness of groundwater management plans and was conducted in three parts that each addressed one of the following research questions:

1. What components do plans require to be quantitatively assessed for effectiveness?
2. How can the effectiveness of groundwater management plans towards achieving objectives be evaluated?
3. How much calibration data is required to constrain model predictions so they are useful and informative?

In the chapter 3, groundwater management was structured as a system control problem to capture the feedback between an aquifer system and management action that occurs in reality, where management action is dictated by aquifer state. Within the control framework, a novel testability assessment rubric that determined if plans met the requirements of a control loop,
and subsequently, whether they could be quantitatively tested, was developed. Seven components of a management plan equivalent to basic components of a control loop were determined, and the requirements necessary to enable testability were defined. Each component was weighted based upon proposed relative importance, then segmented into rated categories depending on how well the requirements were met. Component importance varied but, a defined objective or acceptable impact was necessary for plans to be testable. Plans lacking an objective or acceptable level of developmental impact could not be tested for effectiveness. The rubric was developed within the context of the Australian groundwater management industry, and while use of the rubric is not limited to Australia, it was applied to 15 diverse Australian groundwater management plans and approximately 47% were found to be testable. These results are significant because testability is an important, but often overlooked, prerequisite of the evaluation process and only by quantitatively assessing the effectiveness of groundwater management plans, can we systematically learn how to develop better management plans.

Once it’s been established that a plan can be tested, the challenge of how to test it remains. This is the focus of chapter 4, where a plan testing methodology is developed using a numerical groundwater model. Historically, management modelling has focused on optimisation studies which quantify decision variables in order to find the “best” management strategies to create plans. In contrast, this study does not search for optimal strategies but instead, aims to simulate the sequential decision-making process implicit in environmental management, so that the effectiveness of management scenarios, when implemented as intended, can be evaluated. The purpose of this section was to develop and demonstrate a methodology to quantitatively evaluate the effectiveness of groundwater management plans by simulating sequential management decisions that evolve based on aquifer/management feedback. Groundwater management was structured as a system control loop to capture the
aquifer/management feedback and management decisions were based on realistically sparse observation times and locations. The method provides an indication of how a plan may proceed in reality under alternate timings and frequencies of management decisions and in systems with differing response times. A synthetic example quantified the impact of a generic plan, specifying environmental objectives, extraction restrictions and entitlement limits, relative to no-management by combining a numerical model of reality with management rules under a stochastic climate. A simple synthetic model was used to demonstrate the methodology to test the effectiveness of sequential decision making in groundwater management. This allowed the effectiveness of a management plan to be evaluated within a system control framework.

Model simplicity was chosen over site specific complexity to keep the results generalisable and avoid concentrating on case-specific idiosyncrasies. However, the method is easily extended to complex groundwater systems and management actions, provided they can be modelled adequately. The management decision-making frequency was varied from daily to decadal. Generally, effectiveness decreased as the interval between management interventions increased and intervals greater than annual showed minimal improvement compared to entitlement only. The timing of management decisions relative to the irrigation season also impacted plan effectiveness, and when decisions were made prior to the irrigation season, quarterly management was less effective than annual and biannual management.

It was determined that the prediction of effectiveness was sensitive to model parameters (hydraulic conductivity). Considering that aquifers are vastly complex, heterogeneous, anisotropic systems this sensitivity to parameterisation cast doubt upon the utility of the methodology when applied to complex realistic scenarios. To investigate the feasibility of using this method considering the high degree of uncertainty in groundwater systems, chapter 5 comprised of a calibration constrained, predictive uncertainty analysis.
Given the high degree of parameter uncertainty in groundwater models, section 3 (chapter 5) aimed to determine how much calibration data was necessary to quantitatively evaluate management plan effectiveness. A synthetic study was used to evaluate the uncertainty around predictions generated from four different groups of model realisations that were created based on increasing amounts of observation data, i.e. one prior model and three different posterior models. A numerical model of a synthetic, unconfined aquifer system (reality) with domestic, monitoring and extractions wells was used to generate three calibration datasets (groundwater levels and extraction rates), that varied in length of time and monitoring network extents. The aquifer was managed by a management plan with the objective of preventing dewatering of a domestic well by implementing extraction restrictions in pumping wells when threshold groundwater level triggers were reached. The plan was considered to have failed if the well became dry. Management was simulated in the reality system for fifty years and a “true” plan effectiveness (number of failures) was produced and used as a baseline to compare prior and posterior models. Four simple models of the reality system were built based on alternatively, prior knowledge, and a calibrated solution to each of the three different observation datasets. Each model was used to predict the effectiveness of management decision-making on a monthly basis and if the aquifer was managed under a maximum entitlement volume, and finally, unmanaged.

The model predictive uncertainty for the four simplified models of the synthetic reality system was quantified, through a calibration-constrained uncertainty analysis using Null-Space Monte Carlo methods. The objective was to evaluate if more extensive observation datasets can increase system understanding through calibration to such as degree that the plan effectiveness predictions approach the true reality effectiveness. Because models cannot simulate the true complexity of natural systems, simplifications are required. In this study, time-varying recharge that occurred in reality was simplified in the prior and posterior
models as time-invariant, in line with standard industry practices. Due to the simplification of recharge, the hydraulic conductivity parameter assumed inappropriately high values during the calibration to compensate. This resulted in calibration-induced bias and caused the model to make erroneous predictions. Calibration-induced bias is difficult to identify because often an acceptable fit to data is achieved and does not indicate a problem. However, the calibrated model predicted that the plan would not fail, when in fact, the plan did fail, which has important repercussions for environmental management. Even with use of rigorous uncertainty analysis methods, the effectiveness of management could not be determined due to the limitations of numerical models which raised serious question over our ability to model management.
Acknowledgements

This project was co-funded by Australian Research Council (LP130100958), The Bureau of Meteorology; Victorian Department of the Environment, Land, Water and Planning; Victorian Department of Economic Development, Jobs, Training and Resources, and Northern Territory Water and Power and their financial backing made my research possible. I would like to thank the following people who helped and guided me and were instrumental in the successful completion of this thesis.

To my principal supervisor Prof. Andrew Western, who taught me so much about environmental modelling, groundwater management and the PhD process. Thank you for your humour, kindness and ability to distil disparate threads of thought into a cohesive direction. I very much appreciated your guidance and faith in me and your ability to explain complex concepts in a simple and accessible way. Most of all thank you for your wisdom, guidance and talking me out of withdrawing during a crisis.

To Dr Tim Peterson, who was invaluable with technical knowledge, extensive programming/computing experience and provided directional concepts grounded in lateral thought. Thank you for your knowledge of system dynamics and resilience and clarity of thought, which were instrumental in defining the broad issues relevant to management that steered my research direction. I very much appreciate your expertise, enthusiasm and dedication to my project and for teaching me the meaning of “CPU” with infinite patience.

To Dr Elisabetta Carrara (Bureau of Meteorology), who kept things real and was able to spot the tree within the forest. Your extensive understanding of management, legislation and the practicalities of governing groundwater was invaluable and, your propensity to ask the questions that needed to be voiced was extremely important to me. Thank you, Betty for
your unique perspective, wealth of experience and knowledge and most of all, your laughter, never-wavering support, and soul-affirming hugs.

To Dr Justin Costello, who provided me perspective, geological knowledge and grounded my research in reality. Justin generously gave his time to unsnarl MODFLOW knots, talk about rocks, bioregional assessments, and reaffirm why my project mattered. Thank you for teaching me humility and that ultimately, I am only accountable to myself.

To Dr Murray Peel who was my committee chair and was always fair and organised and looked out for my best interests. Thank you for your stories, laughter and for bringing chocolates to committee meetings.

Thanks to the many people around the world who answered my random emails and calls and provided me with invaluable assistance. In particular, I would like to thank John Doherty of Watermark Computing who replied to my innumerable questions about PEST and was always keen to talk about calibration or uncertainty analysis. Thanks to Daniel, Lev and the SPARTAN team for all of their help getting the PEST suite running on SPARTAN and tirelessly debugging the PEST dependencies.

Thanks to my family and friends for helping me through this process, it has not been an easy time for me. I appreciate all the distractions and especially the espresso martinis during the PhD planning sessions. I would like to thank my parents for instilling in me a love of science and more recently, for all their childcare duties. And of course, my children, Kai and Naomi for their love, understanding and wonderful drawings saying Mama’s work is “grate”. Thank you for being so completely perfect.

Most of all, I want to thank my partner Kevin Hayley, who has been my greatest supporter and champion during this challenging time, and in my life in general. Thank you
for sharing your knowledge and for countless dinnertime discussions about groundwater modelling and inversion. Thank you for your strength and your smile and for spending our 9th wedding anniversary helping me comb through lines of MATLAB code searching for an errant comma - I could not have done this without you.
Table of Contents

Declaration of Published Work................................................................................................................. i
Abstract .................................................................................................................................................... ii
Acknowledgements ............................................................................................................................... vii
Table of Contents ................................................................................................................................ x
List of Figures ....................................................................................................................................... xii
List of Tables ......................................................................................................................................... xiv

Chapter 1: Introduction ............................................................................................................................. 1

Chapter 2: Groundwater management effectiveness............................................................................. 12
  2.1 Research Motivation ......................................................................................................................... 12
  2.2 Literature Review .............................................................................................................................. 13
  2.3 Research Objectives ......................................................................................................................... 18
  2.4 Structure of Thesis ........................................................................................................................... 19

Chapter 3: Can we manage groundwater? .......................................................................................... 21

Chapter 4: Do groundwater management plans work? ..................................................................... 42

Chapter 5: Calibration and Uncertainty Analysis Background Information....................................... 67
  5.1 Background Information ................................................................................................................. 68
    5.1.1 Uncertainty in Natural Systems ................................................................................................ 68
    5.1.2 Uncertainty in Groundwater Modelling ................................................................................ 69
  5.2 The Calibration Process .................................................................................................................... 71
    5.2.1 Solving Systems of Linear Equations .................................................................................... 72
    5.2.2 The Jacobian Matrix ............................................................................................................... 73
    5.2.3 Solving ill-posed inversions with Regularisation ................................................................. 75
  5.3 Parameter/Predictive Uncertainty Analysis .................................................................................. 76
    5.3.1 Null Space Monte Carlo Techniques ................................................................................... 77
  5.4 Concluding Remarks ....................................................................................................................... 80

Chapter 6: Can we model management? ............................................................................................. 82
  6.1 Introduction ..................................................................................................................................... 85
    6.1.1 Managing Groundwater under Uncertainty .......................................................................... 85
    6.1.2 Purpose of Study ....................................................................................................................... 87
  6.2 Methodology .................................................................................................................................. 88
6.2.1 Generation of Calibration Datasets Using Synthetic Reality System .......... 90
6.2.2 Model Parameterisations and Prior Distribution ......................................... 93
6.2.3 PEST Calibrations, Parameter and Predictive Uncertainty ......................... 96
6.3 Results ........................................................................................................ 98
6.3.1 PEST Calibrations and NSMC Analysis.................................................. 98
6.3.2 Management Plan Effectiveness............................................................... 103
6.4 Discussion and Conclusions ....................................................................... 109
6.4.1 Model Simplifications and Predictive Bias ............................................. 110
6.4.2 Implications and Conclusions ................................................................. 115
6.4.3 Major Findings and Conclusions ............................................................ 118
6.5 Appendix One ............................................................................................ 120
6.6 Appendix Two ............................................................................................ 121
6.7 References ................................................................................................. 123

Chapter 7: Major Findings and Conclusions .................................................... 126
7.1 Main Findings and Conclusions ................................................................. 127
7.2 Challenges and Limitations ....................................................................... 134
7.3 Future Research ........................................................................................ 136
7.4 Scientific Contribution .............................................................................. 137

Appendix A ...................................................................................................... 140
Appendix B ...................................................................................................... 142
References ...................................................................................................... 158
List of Figures

Chapter 3

Figure 1: Classical Control Theory Loop.
Figure 2: Groundwater Management Control Theory Loop.

Chapter 4

Figure 1: The Act of Management Modelling Loop
Figure 2: Model Domain.
Figure 3: Groundwater management plan
Figure 4: Average daily water balance in the unmanaged scenario
Figure 5: Selected results showing daily, four monthly and unmanaged scenarios that illustrate the range of hydraulic conductivities (x-axis) within which management of this aquifer geometry has an impact upon the failure rate
Figure 6: Percentage of Days Plan Objectives failed
Figure 7: Temporal (top) and volumetric (bottom) reliability of irrigation supply wells during the irrigation season in the medium aquifer – Mid-season Timing
Figure 8: Temporal (top) and volumetric (bottom) reliability of irrigation supply wells during the irrigation season in the medium aquifer – Early-season Timing

Chapter 5

Figure 1: Singular Value Decomposition of the Jacobian Matrix
Figure 2: Null-Space Monte Carlo Analysis
Chapter 6

Figure 1: Reality hydraulic conductivity field used to generate the three calibration datasets.

Figure 2: Pilot point spacing for each adjustable parameter.

Figure 3: Hydraulic conductivity fields for one prior realisations and the three calibrated models.

Figure 4: Drawdown at monitoring well one (top) and monitoring well two (bottom).

Figure 5: Pre and Post NSMC analysis hydraulic conductivity fields for short-sparse (a & b); Long-sparse (c & d) and Long-extensive (e & f) datasets.

Figure 6: Predictions of plan effectiveness of the prior distribution for monthly (top), entitlement-only (middle) and unmanaged (bottom) management scenarios.

Figure 7: Predictions of plan effectiveness made after NSMC methods were applied to the model calibrated with short-sparse (three years sparse). Monthly (top), entitlement-only (middle) and unmanaged (bottom) management scenarios.

Figure 8: Predictions of plan effectiveness made after NSMC methods were applied to the model calibrated with long-sparse (twenty-five years sparse) for monthly (top), entitlement-only (middle) and unmanaged (bottom) management scenarios.

Figure 9: Predictions of plan effectiveness made after NSMC methods were applied to the model calibrated with long-extensive (twenty-five years extensive) for monthly (top), entitlement-only (middle) and unmanaged (bottom) management scenarios.

Figure 10: Predictions of plan effectiveness simulated with NSMC long-extensive realisations under (a) static simplified climate and (b) the variable reality system climate for monthly management, entitlement-only and unmanaged scenarios.
List of Tables

Chapter 3

Table 1: The primary components of an engineering control theory loop with control system example and aquifer system equivalents.

Table 2: Plan Testability Assessment Components.

Table 3: Plan Testability Assessment Rubric.

Table 4: Questionnaire to determine Management or Monitoring Assessment Type.

Table 5: Australian Groundwater Management Plans selected for Testability Assessment using developed Rubric

Table 6: Australian Management Plan Testability Results.

Chapter 6

Table 1: Calibration Specifications for three BeoPEST calibrations.

Table 2: Hydraulic conductivity percentile values for reality, prior and the three calibrated models.

Table 3: Percentage of days domestic well one dewatered at the plan failed for reality, prior, base calibrated models and each NSMC parameter realisation
... small drops form in the region above the earth, and these again join others, until rain water falls in some quantity; similarly, inside the earth quantities of water, quite small at first, collect together and gush out of the earth...

Aristotle Meteorologica I XIII

Chapter 1: Introduction

In Greek mythology the river Styx marked the boundary of Tartarus – and its chilly waters were considered so sacred, so precious, that the gods of Mt Olympus swore binding oaths upon them [Graves, 2017]. Vivid imagery in epic poems of great subterranean streams surging from below – the primeval waters of the Styx which never fails, but leaps out of the rocks [Hesoid, c 725 B.C] – ignited a belief in boundless groundwater and the myth of the underground river: a belief which persists in public perception to this day. The relative invisibility and highly uncertain nature of groundwater compared to surface water, contributes to a paucity of understanding by the general public and hence; a devaluing of groundwater in the public eye. The purpose of this section is to provide a summary of the evolution and importance of groundwater management.

Development of groundwater evolved from utilisation of springs during the Paleolithic period (3.3 Ma – 9,650 BC) through to basic hand dug wells in Neolithic times (~7,000 BC), to deeper more complex wells during the Bronze Age (3,000 – 600 BC), with wells up to 20 m deep excavated in Greece, Syria and India [Angelakis et al., 2016]. Wells were prevalent in Classical Greece and during Roman times with complex underground distribution systems constructed, regulations mandating that wells be maintained in good order (in case of warfare) and,
groundwater contributed an important component of urban water supply [Angelakis et al., 2016]. Qanats, a type of underground canal, designed to direct water through a subterranean conduit for surface storage and distribution, have been utilised in Syria for over 2,000 years [Lightfoot, 1996]. And approximately 2,500 years ago, at the dawn of the Persian Empire, groundwater was managed in Iran using qanats that are still utilised today and supply 15% of Iran’s water demands [Ahmadi et al., 2010]. In Australia, Indigenous Australians long utilised groundwater by excavating waterholes and tunnels and managed groundwater resources by reducing evaporation and protecting water sources from contamination [Moggridge, 2005]. However, it was the advent of centrifugal pumps in the 19th century and advances in drilling methods, that enabled deep groundwater to be pumped at relatively low cost, and dramatically increased global groundwater development with little thought given to management or regulation.

Groundwater is the world's largest freshwater resource and provides drinking water for almost 2 billion people and half the irrigation water for global food production [Aeschbach-Hertig and Gleeson, 2012; Famiglietti, 2014; Gleeson et al., 2010]. Bourgeoning populations have driven water demand, and groundwater levels have declined in many regions of the world including Australia, the Middle East, China, USA, India, Northern Africa, and Southern Europe due to excessive extractions [White et al., 2016]. Overexploitation of groundwater has resulted in adverse environmental and socio-economic impacts such as, dwindling well yields and growing pumping costs [Konikow and Kendy, 2005], water quality deterioration [Fogg and LaBolle, 2006; Tushaar Shah et al., 2001], stream and wetland desiccation [Wada et al., 2010], land subsidence [Giordano, 2009], saline water intrusion [Reichard and Johnson, 2005; Werner et al., 2011; Werner et al., 2013], among other undesirable occurrences [Gleeson et al., 2010].

Groundwater is a shared resource and frequently spans geographical and political boundaries
[Blomquist and Ingram, 2003] and with increasing populations under a changing climate, water shortages are becoming more frequent [Green et al., 2011] underscoring the urgency of effective management.

While all groundwater usage has an impact [Alley and Leake, 2004], groundwater management seeks to mitigate the undesirable impacts of groundwater usage to within acceptable levels. Worldwide, groundwater management plans are increasingly being used as the primary way to manage groundwater and ensure water security and environmental protection. In many regions, groundwater is a finite resource with intense usage competition and therefore; statutory groundwater management (GWM) plans are essential to outline volume and extraction regulations to prevent overexploitation, inequitable distribution or well interference. Generally, the main management actions employed by GWM plans include well licensing, extraction capping, water sharing via entitlement/allocation systems, water extraction restrictions, groundwater trading, managed aquifer recharge, entitlement carryover, buybacks and emerging methods such as treatment of wastewater for drinking purposes [Pilz, 2010; Radcliffe, 2010; The Murray Darling basin authority, 2012], and novel localized methods of control such as restricting electricity use to reduce pumping in groundwater depleted areas of India [Shah et al., 2008]. Legislation mandating use of management plans to regulate groundwater in Australia has led to a proliferation of plans of varying structure, developed by a litany of management organisations that vary between, and within, states and territories. Currently, there are 228 plans in Australia [Bureau of Meteorology, 2018], developed under numerous legislative frameworks and of contrasting scope, rigour and employing different management actions to achieve particular objectives. However, the impact upon the aquifer system of many plans, distinct from the impact of other hydrological drivers, is currently unknown. Many drivers, such as climate
and previous management action influence aquifer state and the implicit assumption that a beneficial aquifer state results from a plan despite the lack of established causality is pervasive in hydrogeology. Quantitatively evaluating the impact of the plan upon aquifer state is the only way to establish causality. However, not only are plans rarely quantitatively tested, many plans are not conducive to quantitative analysis. The overarching theme of this thesis is the quantitative analysis of groundwater management plans.

Groundwater management, in contrast to surface water management, is relatively new and faces unique challenges due to the relative invisibility of groundwater and changing climatic and anthropological impacts. In Australia, the El Niño-Southern Oscillation exerts a profound influence, with El Niño conditions bringing droughts and devastating bushfires, and La Niña accompanied by violent rainfall, floods and cyclones [BoM, 2019]. It is a land of extremes and in many regions, groundwater is the sole source of water and provides one fifth to one-third of total consumptive usage [Bureau of Meteorology, 2019; Harrington and Cook, 2014]. Water availability varies greatly across the country, both in space and time [Letcher and Powell, 2011], making water resource management very difficult and motivating development of management frameworks. It is useful to briefly describe the history of water resource management in Australia to place current GWM plans in context.

European settlers adopted a water management style more suited to the verdant fields of Europe than the arid plains of Australia. Initially, surface water was governed by the riparian doctrine (water can be used only by landholders occupying the land it flows through), which imposed a user limit by requiring upstream users not to adversely impact downstream users. Groundwater, however, did not have a user limit, and landholders could use as much as they pleased even if flagrant usage impacted others. By the late nineteenth century, the common law riparian doctrine
had been abolished and replaced by licensing and allocation due to irrigation development. Various acts nationalised the beds and banks of waterways and the rights to water and allowed formation of irrigation trusts to license and distribute water [Government of New South Wales, 1896; 1912; Government of Victoria, 1886; 1905].

During Federation, water management was vested in control of the states under section 100 of the Australian constitution and consequently, achieving unanimous agreement on trans-boundary policy direction was challenging [Government of Australia, 1900]. Water management in Australia followed a predictable trajectory: drought and times of crisis prompted policy and legislative action. This cycle is demonstrated by over a century of water policies initiated during times of drought when political pressure reached a crescendo [Geoscience Australia, 2014], for example, the River Murray Waters Agreement between Victoria, New South Wales and South Australia, could not be agreed upon in time of Federation but, was finally reached in 1914 following several years of severe drought [Australian Commonwealth, 1915; Musgrave, 2011].

In the early twentieth century, with a very European conception of water availability and variability, development of water resources in Australia was vigorously pursued, and marked by government sponsored irrigation schemes and construction of large storage dams [Musgrave, 2011]. Development surged forward with scant mind to regulation and sustainability through most of the twentieth century, with objectives based solely upon engineering factors and legislated obligations to consider external consequences lacking [Crase, 2010]. In the late 1960s, the failure of a supply bore in the western Victorian town of Nhill, ultimately led to the development of the [1969] Victorian Groundwater Act [Clark and Myers, 1969], which aimed to control groundwater usage and required a permit to drill bores [Victorian Government, 2019]. Then extensive droughts in the late 1970s and early 1980s resulted in graphic images of the
Murray River, stagnant far from the mouth [Australian Government, 2015], and skeletal Red Gums bordering dusty rivers, which, combined with the Darling’s striking algal bloom in the early 1990s [Australian Government, 2015]; catapulted the importance of effective water management into the public consciousness and, shifted long entrenched perceptions of irrigation and water management [Musgrave, 2011]. In response, a Water Reform Framework agreement was reached in 1994 by the Council of Australian Governments with the intention of achieving an efficient and sustainable water industry with greater private sector and community engagement. Mandating that future water development be based upon environmentally sustainable development [Brutland, 1987], the COAG reforms meant that, for the first time, management of water sought a balance between economic, social and environmental needs [McKay, 2011]. The main elements of the framework were the development of water markets and trading (including allocation systems and separating land titles from water licenses); recognition of environmental water requirements and compelling states to specify allocation volumes for the environment; shifting regulation and policy away from water providers; and pricing reforms aiming to reflect the true cost of water [Gardner et al., 2009].

The Murray-Darling Basin (MDB) Agreement was reached shortly after, which permanently capped surface water extractions from the basin in 1997 [McKay, 2011]. However, groundwater extractions remained uncapped and consequently, increased significantly [National Water Commission, 2005]. Due to decades of overuse, with thousands of wells that flowed unimpeded, groundwater levels in the Great Artesian Basin (GAB) had declined. A rehabilitation program, funded by the Commonwealth Government, began in 1989 with the purpose of capping bores and piping water instead of allowing it to flow through open drains [McKay, 2011]. Yet, in the late 1990s more than 1,500 artesian bores flowed unabated into over 34,000 km of open bore
drains due to poor technology. The GAB Strategic Management Plan led to an initiative that, over 15 years, resulted in upgrades to 750 bores and installation of more than 31,000 km of distribution and drainage pipes [Department of Agriculture, 2019; Great Artesian Basin Consultative Council, 2000]. The GAB Strategic Management Plan was reviewed in 2015 and a public consultation on a new plan closed in 2018.

Because aquifer systems are relatively invisible, with intense development pressures and long lag times; the regulation of groundwater has always trailed that of surface water [Nelson, 2012]. Groundwater management has gained traction in Australia over the past thirty years and is still developing. In other parts of the world, similar trends towards a greater focus in managing groundwater resources is seen, albeit with differing approaches and at differing stages. For example, the Canadian province of British Columbia passes the Water Sustainability Act [2016] which requires users, for the first time, to obtain an authorisation to use groundwater. California passed the Sustainable Groundwater Management Act [2014], legislation that allows local agencies to customise groundwater sustainability plans and provides them with certain powers to maintain aquifers in an acceptable state. Prolonged dry conditions across southern Australia from 1996 to mid-2010 – termed the Millennium Drought [BoM, 2015] – resulted in unprecedented reforms to Australian water management through the National Water Initiative (NWI) [Council of Australian Governments, 2004] and the National Plan for Water Security (NPWS). The primary aims of the NWI was to return overallocated systems to a sustainable level of development; create water management plans; facilitate a nationally compatible water market where shares could be traded; recognise surface-groundwater connectivity and manage connected systems as a single resource and; provide statutory provision for environmental water [Council of Australian Governments, 2004; Gardner et al., 2009]. The states and territories were
required to implement the NWI in their jurisdictions and to change their existing policy and legislation to fit in with the NWI [Council of Australian Governments, 2004].

In 2007, the National Plan for Water Security (NPWS) was announced, accompanied by $10 billion dollars of commonwealth funding to modernise irrigation infrastructure; address over-allocation and redefine governance in the Murray Darling Basin (MDB); continue the restoration of the Great Artesian Basin (GAB); investigate feasibility of water development projects in northern Australia and, improve water information, necessary for informed decision-making, by expanding the role of the Bureau of Meteorology to amass and distribute water data in a cohesive manner [Gardner et al., 2009; Howard, 2007]. Under the NPWS, the Commonwealth was tasked with developing a basin wide plan for the MDB that set a sustainable cap at a catchment level and sought to restore ecological health throughout the basin. Subsequently, the Federal Water Act 2007 [Government of Australia, 2007] and Water Amendment Act [Government of Australia, 2008], gave the Commonwealth power to manage the basin as a whole and enabled creation of an administrative authority with the role of developing a basin wide management plan (the Murray Darling Basin Authority MDBA). The Basin Plan was completed in 2012 with the objectives of achieving a sustainable and long-term adaptive management framework of surface and groundwater, optimising social, environmental and economic outcome and, improving water security in the basin [Murray Darling Basin Authority, 2012].

New management zones (water resource areas) with new extraction limits for surface and groundwater (sustainable diversion limits) were created and required to be managed under water resource plans developed by the each of the basin states and accredited by the MDBA by June 2019. The plans outline strategies to either manage surface-groundwater conjunctively, or detail how surface water or groundwater resources are to be managed in each water resource area. As
of August 2019, only two plans have completed the accreditation phase and are considered complete: the South Australian Murray Region Plan [Department for Environment and Water, 2018], and the Warrego-Paroo-Nebine plan [Department of Natural Resources and Mines, 2016], both of which outline strategies to manage groundwater and surface water. There are four more plans in the accreditation phase (two surface-groundwater plans, one groundwater plan and one surface water plan) and seven more are in the assessment phase [MDBA, 2019]. The individual plans are intended to reflect the goals of the Basin Plan as a whole.

Until completion of the MDBA accredited plans, GWM plans created by the states under their water legislation remain in place. These GWM plans are quite diverse and have various techniques and management approaches that define such things as the objectives of the plan, and the maximum volumes that can be extracted from a particular area. The name for the maximum volume varies depending on location, i.e. termed the permissible consumptive volume (PCV) in Victoria and an allowable annual volume (AAV) in South Australia. Many regions are already allocated up to their PCV limits and the only way to acquire new water is to trade with existing groundwater license holders. Plan objectives vary but include groundwater levels, prevention of an adverse impact such as saline intrusion [Department of Water (DOW), 2013], minimum river flow volumes [DNREAS, 2009; Upper Ovens, 2011], groundwater quality objectives [DOW, 2009; Murray Darling Basin Authority, 2012], maximum rate of groundwater extraction [Department of Primary Industries Parks Water and Environment, 2012] and rate of groundwater level decline [GWM, 2001; 2017], pressure head decrease [SAAL NRM, 2009], or storage volume decline [DLRM, 2016].

Generally, in Australia, water is shared using the concept of a consumptive pool (the volume of water that is available each year). The consumptive pool is shared amongst users via an
entitlement and allocation system – where an entitlement is the maximum volume a user can access per year, and an allocation is the volume the user can access in a particular year (usually a percentage of the entitlement volume), depending on the total volume of the consumptive pool. The manner in which allocations are determined varies across the country. For example, some Victorian plans contain prescriptions mandating entitlement restrictions based on threshold groundwater levels, such as the Katunga, Mid-Loddon and Lower Campaspe Valley Water Supply Protection Area (WSPA) plans [GMW, 2009; 2017; Goulburn Murray Water, 2012]. The Katunga WSPA is managed to the maximum annual groundwater recovery levels from the previous five irrigation seasons at several monitoring bores, and percentages of allocations are provided depending on levels. Previously, allocations in the Katunga WSPA were determined based on the rolling average volume of groundwater usage during the previous five years. However, the volume based approach was amended to a groundwater level management style during the 2017 review [GMW, 2017]. The Mid-Loddon plan cuts licensed volumes by 30% if the maximum three year rolling average recovery levels in a monitoring well fall below a threshold [GMW, 2009]. The Lower Campaspe management plan determines allocations based on the three-year rolling average of maximum groundwater level recovery in a monitoring well. Depending on the level, the plan stipulates step-wise restrictions to licensed volumes from 100% to 40% [Goulburn Murray Water, 2012]. The Upper-Ovens plan introduces a roster system to share water when streamflow volumes at a monitoring location in the Ovens River decline to threshold volumes and extractions are banned if volumes continue to decline [Upper Ovens, 2011]. Allocations in the Deutgam groundwater management zone are primarily determined based on groundwater level triggers and trigger levels are defined for 50, 25 and 0% allocations [SRW, 2015]. Furthermore, water managers have the authority to ban groundwater extractions in
extreme situations under Section 33 of the Victorian Water Act [Government of Victoria, 1989].

Clearly, there is a multitude of plans with different mechanisms to address various groundwater problems. A legally binding nationally consistent management plan development guide similar to the National Water Initiative (2010) non-statutory guidelines would help remedy this discord between states and territories. Notwithstanding the diversity of approaches, success is dependent on effective plans that are well implemented. The focus of this thesis is on developing methods for assessing the effectiveness of GWM plans in particular.
Chapter 2: Groundwater management effectiveness

2.1 Research Motivation

The central contention of this thesis is that the effectiveness of most groundwater management plans is unknown. Considering the importance of groundwater for global resource security and the predominance of GWM plans as a means to manage groundwater, this lack of evidence of plan effectiveness is a large research gap. An inability to quantify the effectiveness of existing plans, potentially inefficient plans, hinders the creation of more effective plans. The research motivation was the development of a comprehensive framework to quantitatively evaluate the effectiveness of groundwater management plans and this thesis is divided into three main sections with the following objectives and research questions:

1. The objective of the first section is to assess the testability of management plans by determining the requirements plans must have in order to be quantitatively testable. The research question this section seeks to answer is: What components do plans require to be quantitatively assessed for effectiveness?

2. The objective of the second section is determining how to test management plans and a plan evaluation method is presented. The research question this section seeks to answer
is: *How can the effectiveness of groundwater management plans towards achieving stated objectives be quantified?*

3. The objective of the third section is to expand upon the outcomes of the second section and explore the feasibility of using the developed method in a more realistic hydrogeological setting that is subject to spatial variability and model uncertainty. The research question this section seeks to answer is: *Given model parameter uncertainty, how feasible is evaluating management plan effectiveness using numerical groundwater models?*

### 2.2 Literature Review

Water management and resource sustainability is an amalgamation of political, socioeconomic and climatic factors and effective management requires identifying when, and if, a particular management plan is capable of achieving a defined goal. Unfortunately, in most cases the degree to which a groundwater management plan impacts an aquifer is unknown, and it is assumed that a plan works as long as the aquifer remains in an ‘acceptable’ state. But, the state of an aquifer is a result of a combination of external and management related factors including precipitation recharge, discharge to streams and groundwater dependent ecosystems, evapotranspiration, and pumping extractions and understanding the influence of each factor upon the groundwater system is crucial for effective management. However, as stated by Walters and Holling [1990], the efficacy of measures regulating environmental systems cannot be claimed with certainty as management policies can be described as uncertain perturbation experiments [*Walters and Holling, 1990*]. Proclaiming a plan as effective presents a challenging proposition as absolute truth is non-existent in science and furthermore, most management studies focus on optimisations instead of separating the impact of management from other hydrological drivers.
Yet annual reports detailing progress of management plans frequently state the ‘successful implementation’ of groundwater management plans \[GMW, 2014\], implying plans work and are responsible for realisation of management objectives despite a lack of definitive evidence showing plans are responsible for meeting aquifer state objectives when it is unknown if management is causing groundwater levels to fluctuate or whether it is due to climate.

Furthermore, ascertaining that a plan works cannot be done except in a closed system \[Oreskes et al., 1994\] and decomposing groundwater dynamics into components caused by management and climate is exceedingly challenging \[Shapoori et al., 2015a\]. But, aquifers are open, heterogenous systems, and are often represented by simple conceptual models, with hydrogeological components that are not accurately or precisely known and complicated by scale issues, data scarcity \[Oreskes et al., 1994; Poeter, 2007\] and un-foreseen circumstances. So, a quandary is reached; we need to know how to manage groundwater effectively but even when management appears to be effective by meeting objectives, we can’t tell whether objectives are met due to management or happenstance. Currently, the best we can do is amass evidence indicating management is likely working.

Entire libraries have been written on groundwater management; how it should and should not be managed, methods of calculating sustainable pumping rates and reviews of aquifer modelling studies, yet precious little has been published examining quantitative analysis of the effectiveness of groundwater management plans and elucidating causes of groundwater management success or failure. Even qualitative studies evaluating management effectiveness are scarce; rare examples include Jacobs et al [2004] who reviewed knowledge gained from groundwater management in Arizona where Colorado River water and effluent replaced
groundwater, and Sophocleous [2010] who examined experiences from groundwater management in the High Plains Aquifer in the US.

Currently, quantitative techniques for assessing the effectiveness of groundwater management or determining the probability management plans and techniques will succeed are few, despite being required to establish whether groundwater management works. Gallagher [2015] developed the Groundwater Operational Management Package (GWOMP), an integrated MODFLOW-MT3DMS-SEAWAT modelling package that simulates an aquifer under various management regimes to allow the testing of allocation volumes, carryover and groundwater trading with a range of triggers and operational rules. The package was successfully implemented on the Pioneer Basin in Queensland and aquifer response under management statistically evaluated [Government of Queensland, 2002].

Numerical groundwater models have long been used to predict effects of management decisions on aquifer yields and evaluate water management strategies [Sophocleous et al., 1999], impact of groundwater extraction [Ebraheem et al., 2002; Gorelick, 1983] upon aquifers, optimise extraction rates [Casola et al., 1986; Makinde-Odusola and Mariño, 1989; McPhee and Yeh, 2004; Ajay Singh, 2012; A. Singh, 2014; Tankersley and Graham, 1994; Wagner, 1995], manage seawater intrusion [Reichard and Johnson, 2005; Rejani et al., 2008], and investigate implications of economic considerations [Booker et al., 2012; Bredehoeft and Young, 1970; Bromley, 1991; Gisser and Sánchez, 1980; Koundouri, 2004a; b] amongst others. The Western Australian Department of Water (2009) used MODFLOW [Harbaugh, 2005] to construct a 13 layer model simulating 15 different management scenarios upon confined aquifers and Mulligan [2014] used an agent based coupled economic-groundwater model to evaluate the impact of taxes and quotas on aquifer head levels.
Structuring groundwater management as a systems control problem forms the basis of much of the work described in this thesis where the effectiveness of a management plan based on the concept of sustainable yield is quantitatively evaluated by modelling the feedback driven, adaptive, sequential decision-making process of management. Managing in this way is a form of adaptive management, because the aquifer state dictates the management action.

During the Millennium Drought, rigid groundwater management rules resulted in suspension of NSW Water Sharing Plans [Hamstead et al., 2008] and showed the need for flexible and adaptive plans and that extremes of climate be considered during planning [Hamstead et al., 2008]. However, irrigators and other users require a degree of certainty in order prevent severe financial consequences so a clearly defined adaptive management policy could outline potential management interventions and clarify anticipated impacts upon the community. Inflexible control methods that suppress natural disturbance regimes can decrease system resilience, the manner in which the system responds to and ability to absorb stress, and cause disastrous ecological collapse [Folke, 2002]. Therefore, allowing the management to change depending on the aquifer state allows for more flexibility than blanket management rules.

A feedback driven approach to management is passive adaptive. Adaptive management responds to changing knowledge and conditions by reviewing past management and aims to continually improve of management of a system through learning from success or failure of previous management attempts. Adaptive Management can be divided into three types of learning; trial and error, passive and adaptive. Trial and error is learning from random experience [Allan and Curtis, 2005]. Passive adaptive management selects the single best action based on past events and information whereas active adaptive management uses available data to construct several policy experiments to determine the policy that enhances learning and delivers acceptable
management performance \cite{Hollings1978, Walters1990}. Active Management,
that involved policy experiments to test uncertain aspects of policy, has been extensively applied

The feedback mechanism allows management to adapt and improve by learning from experience
and improved knowledge of the system and the system’s response to change. The response of the
system to change is dictated by system connectivity, uniformity, variability and critical shift
threshold, and provides crucial knowledge for informed management policy \cite{Hollings1978} and
system resilience \cite{Peterson2012}. Because aquifer response to management will likely
differ to anticipated response; adaptive control allows adjustment of control based upon actual
system response, such as Jones \cite{1992} updating parameter estimates with hydraulic head values.

Allan and Curtis \cite{2005} claim adaptive management conflicts with the culture of natural resource
management where emphasis is upon activity, control, comfort and clarity instead of reflection,
learning and embracing complexity and variability and they highlight the scarcity of adaptive
management use despite it being a political catch phrase. Inadequate funding, monitoring, project
reviews and evaluations, and the goal-oriented nature of management were cited as factors
contributing to dearth of use. Lee \cite{1999} appraised adaptive management using a policy design
framework and found the objectives of scientists and managers differ, the former recognises
surprise and pursues implications whereas the latter pursues objectives. Experiment based
management can be slow and expensive but is reasonable in complex over-exploited natural
systems because complexity implies management outcomes may be surprising \cite{Lee1999}. An
adaptive control technique is particularly applicable to hydrogeology because of high aquifer
uncertainty and complicated feedback mechanisms between aquifer management and system state.

2.3 Research Objectives

The literature review provided in section 2.2, detailed various established approaches to groundwater management; however, an exhaustive search of groundwater management planning and modelling literature uncovered no quantitative assessments of groundwater management plan effectiveness. As a result, an evaluation of the quantitative effectiveness of plans was prioritised over more qualitative assessment methods such as, for example, methods based upon surveys of water users and water managers that have been used previously to investigate the effectiveness of local groundwater management districts [Kromm and White, 1995] and agricultural policy adoption [Bosch et al., 1995]. The initial objective was to quantitatively evaluate the effectiveness of Australian groundwater management plans; however, during the data collation stage, it was discovered that many plans were not conducive to analysis. The high variability in plan structure and contact made an evaluation difficult because many plans were not written in a manner conducive to analysis. Consequently, before plans could be tested, it had to be determined if they could be tested. Therefore, a method to evaluate whether plans could be assessed for effectiveness was developed so that plans lacking required components could be identified. The testability of plans is an often overlooked, but critical prerequisite of the quantitative evaluation process. Plans cannot or are not being evaluated for effectiveness and a method to assess whether plans contain the required components to be testable is lacking in the hydrogeology field. This lack of plan evaluation is due to several factors:
• Management plans are assumed to be effective if an aquifer remains in an agreeable state despite an absence of established causality.

• Management modelling has focused on optimisation studies to create the “optimal” management strategies, instead of evaluating the effectiveness of established plans (if implemented as intended).

• A comprehensive framework to evaluate plan effectiveness is lacking in hydrogeology.

Notwithstanding plan reviews, which as long as the aquifer is in an acceptable state are generally fairly routine, plans are often implemented in a “set and forget” fashion, where once the plan is in place and providing no adverse impacts appear, plan effectiveness is assumed. This method of managing is inadequate because plans need to be robust and capable of achieving objectives under changing climate conditions and in light of the large uncertainties inherent in management approaches based upon recharge and storativity estimates. For example, a plan may seem successful when implemented under a favourable climate and then fail when drought conditions prevail, or the system may not have been stressed such that the mechanisms of the plan are enacted. Effective management is capable of coping with climatic stress.

2.4 Structure of Thesis

As this thesis is by publication, further literature reviews were conducted in each of the three stand-alone papers and were tailored to support the research question posed in that specific paper. This thesis is structured into seven chapters.
Chapter 1 describes the importance of groundwater and the evolution of groundwater management, focusing on an Australian perspective. Additionally, the potential consequences of groundwater mismanagement are detailed.

Chapter 2 provides the broad background and motivation for the research and the thesis objectives.

Chapters 3, 4 and 6 are structured as stand-alone research papers, each addressing one of three research questions. Chapter 5 provides necessary background information for the study described in chapter 6.

Finally, chapter 7 describes the major findings and conclusions of the entire PhD and details the challenges, limitations and the scientific contribution of the work and future research directions are proposed.
Chapter 3: Can we manage groundwater?

A method to determine the quantitative testability of groundwater management plans

This chapter was published as the following journal article:

Can we manage groundwater? A method to determine the quantitative testability of groundwater management plans

E. K. White¹, T. J. Peterson¹, J. Costelloe¹, A. W. Western¹, and E. Carrara²

¹Department of Infrastructure Engineering, University of Melbourne, Victoria, Australia, ²Bureau of Meteorology, Docklands, Victoria, Australia

Abstract Groundwater is the world’s largest freshwater resource and due to overextraction, levels have declined in many regions causing extensive social and environmental impacts. Groundwater management seeks to balance and mitigate the detrimental impacts of development, with plans commonly used to outline management pathways. Thus, plan efficiency is crucial, but seldom are plans systematically and quantitatively assessed for effectiveness. This study frames groundwater management as a system control problem in order to develop a novel testability assessment rubric to determine if plans meet the requirements of a control loop, and subsequently, whether they can be quantitatively tested. Seven components of a management plan equivalent to basic components of a control loop were determined, and requirements of each component necessary to enable testability were defined. Each component was weighted based upon proposed relative importance, then segmented into rated categories depending on the degree the requirements were met. Component importance varied but, a defined objective or acceptable impact was necessary for plans to be testable. The rubric was developed within the context of the Australian groundwater management industry, and while use of the rubric is not limited to Australia, it was applied to 15 Australian groundwater management plans and approximately 47% were found to be testable. Considering the importance of effective groundwater management, and the central role of plans, our lack of ability to test many plans is concerning.

1. Introduction

Groundwater provides the main source of drinking water for almost two billion people, half the irrigation water used for global food production and represents the world’s largest freshwater resource [Aeschbach-Hertig and Gleeson, 2012; Famiglietti, 2014; Gleeson et al., 2010]. With increasing demand and excessive extraction, water levels have declined in many regions of the world, including Australia, the Middle East, China, USA, India, Northern Africa, and Southern Europe. These declines have resulted in decreased well yields and increased pumping costs [Konikow and Kendy, 2005], water quality deterioration [Fogg and LaBolle, 2006; Shah et al., 2001], stream and wetland desiccation [Wada et al., 2010], land subsidence [Girdano, 2009], and other adverse environmental and social impacts [Gleeson et al., 2010; Wada et al., 2010]. Therefore, with most of the easily accessible groundwater resources developed or overdeveloped, and the era of groundwater exploration mostly over; detailed evaluation and careful management of known aquifers has become of the utmost importance [Freeze and Cherry, 1979; Gleeson et al., 2012; Pigram, 2006]. The extent and severity of these impacts suggests effective groundwater management remains elusive, and considering plans are a primary method of managing groundwater, ensuring they are robust and effective is vital. Logically, the only way to tell if plans are robust and effective is by testing them. And it follows that if plans cannot be tested, effectiveness will remain unknown. We contend there are two critical aspects of determining the effectiveness of a management plan:

1. Plans are constructed in a way to make them testable.
2. An evaluation is undertaken to assess effectiveness. There are two components to this evaluation process, (a) the aquifer system has been stressed such that plan management actions are activated and (b) causality of any observed change in aquifer state has been attributed to management interventions.

This research concerns the first part; that is, the testability of plans. The purpose of this paper is to present a method to assess if management plans can be tested for effectiveness. The evaluation of how effective plans are is a topic of current research but is beyond the scope of this paper. It is not contended that
testability equates to effectiveness but simply, that testability is a crucial, yet oft disregarded, prerequisite of evaluating the effectiveness of management plans.

1.1. An Acceptable Aquifer State Does Not Equate to an Acceptable Management Plan

Measures used to describe the condition of an aquifer system include hydraulic head levels, well yields, minimum stream or base flow volumes, fluxes, and status of groundwater-dependent ecosystems (GDEs), which combined, comprise the state of the aquifer system. Many external and management-related drivers impact this aquifer state, including precipitation, recharge, discharge, evapotranspiration, pumping extractions, previous management actions, and geological environment [Konikow and Kendy, 2005]. Isolating the individual impact of management actions toward achieving objectives, or maintaining head at acceptable levels, from that of external drivers is very difficult.

Consequently, an acceptable aquifer state can often be misinterpreted as evidence of plan success. But the state of an aquifer is not equal to plan efficacy, for example, consider a scenario where a management plan has governed a semiconfined aquifer for 10 years in a region utilizing both surface and groundwater. Throughout the decade, the region experienced a severe drought and surface water availability declined, prompting increased groundwater demand. Head levels dropped until groundwater restrictions were triggered, and, despite sequential implementation of increasingly severe restrictions, levels continued to fall. When the drought broke, recharge increased, extractions reduced and heads recovered to predrought levels. Conversely, consider the same aquifer during a decade in which conditions were particularly wet, with plentiful recharge, minimal pumping, and management intervention was not required. Consequently, the plan mechanisms designed to manage the system under stress were not enacted. The same management plan is perceived differently based on climatic conditions, because clearly, both climate and management influence head levels and the respective impacts influencing an aquifer are difficult to disentangle [Shapoori et al., 2015].

As a result of interwoven impacts and the realities of groundwater management, assumptions are made. We believe the assumption that a groundwater management plan is the causative factor if an aquifer remains in an acceptable state, is pervasive in groundwater management. Yet evidence of the validity of this assumption is difficult to amass. It cannot be stated that a plan is effective unless the system has been stressed such that management actions are activated, and causality of any observed system change established. Proclaiming plans as effective without verification is ill-advised and borrowing the words of Sir Arthur Conan Doyle, “the temptation to form premature theories upon insufficient data is the bane of our profession.” [Doyle, 1915]. Efficacy must be demonstrated not assumed. Despite routine measurement of aquifer states, plan efficacy is rarely analyzed and in many cases, cannot be analyzed. Nonetheless, plan review reports frequently state “successful implementation” of plans e.g., [Golburn Murray Water (GMW), 2014a, 2014b, 2014c, 2014d, 2014e], as opposed to quantitatively analyzing the specific role of the plan in successful actualization of objectives. Furthermore, stable water levels are often used as a benchmark to determine whether management is appropriate, for example, several Australian plan review reports state “monitoring and metering indicate no significant changes in the condition of the resource or water usage patterns… Therefore, it is considered that the groundwater resources are being managed sustainably” [Southern Rural Water, 2014b, 2014c, 2014d]. And regularly the success of groundwater management plans is measured against potentiometric surface or broad, qualitative objectives defined in legislation such as “equitable management” [Government of Victoria, 1989] or the “protection, maintenance and enhancement of ecosystem value” [Department of Water (DOW), 2006], despite significant causative uncertainty between management actions and aquifer dynamics. Due to this uncertainty, much ambiguity surrounds the effectiveness of plans. The first step toward resolving this ambiguity is determining if a plan can be tested. That is, does the plan contain the required components that would allow a quantitative analysis to separate and determine the individual effect upon the aquifer of climate and management, and provide quantitative, defensible, proof of plan legitimacy? Such individual analysis seldom occurs, despite its potential to minimize erroneous management conclusions and to lead to more effective plans. Consequently, the efficiency of numerous plans is unknown.

1.2. The Challenge of Environmental Management

Implicit in the management of environmental systems is the assumption that sequential decision making improves outcomes. Considering that aquifers, like many environmental systems, are complex,
heterogeneous, and open systems where hypotheses cannot be proven [Oreskes et al., 1994], and sequential management decisions are complicated by extensive lag times, scale issues, and data scarcity and uncertainty, this assumption warrants examination. Extractions from aquifers, as with most managed natural resources, are governed by a process in which managers have incomplete system knowledge and receive incomplete data which they use to consider resource demands and other values to decide upon an acceptable usage. Clearly the effectiveness of such decision making is dependent upon the reliability of system knowledge, perceived vs actual risks, and time lags between management actions and the observed response of the system. Quantitatively exploring the impact of sequential environmental management and the coupled social-biophysical dynamics is challenging, and to date the focus has been on simple theoretical numerical modelling studies. Two notable early investigations of potential effects of sequential environmental management are Janssen et al. [1999] and Carpenter et al. [1999], who explored the management of phosphorous loading to a lake simulated with a positive feedback. Both studies found that various management approaches caused the lake dynamics to oscillate between the desirable oligotrophic, and undesirable, eutrophic states. More recently, Lade et al. [2013] explored harvesting of a nonspecific common-pool resource where overexploitation was regulated by social ostracism of noncomplying harvesters. Importantly, and unlike Janssen et al. [1999] and Carpenter et al. [1999], the natural resource itself did not have a positive feedback but its interaction with society resulted in a positive feedback and consequent overexploitation of the resource. And through modelling alternate management control methods, Anderies et al. [2007] found panaceas can be more detrimental than inaction in the case of fisheries management. The authors stated that inappropriate use of feedback in natural resource management can destabilize or destroy ecological systems.

These management investigative studies show that the feedback cycles present in many socioenvironmental systems may, through interactions with management under inadequate system knowledge, be stimulated in unanticipated ways leading to system failure. This concept of stress testing management was explored in a hydrogeological context by Guillaume and El Sawah [2014], who used simple one and two cell groundwater models that captured aquifer system dynamics and satisfied predetermined assumptions, to allow stakeholders to explore conditions under which management plans succeeded and failed to achieve objective conditions. Because each policy failed under particular circumstances, insight into system behavior under management was gleaned, plan limitations identified, and the fact that inappropriate preconceptions can result in management failure highlighted. This is important because in groundwater management, reality will always differ from preconceptions due to uncertainty, and because real aquifers and societal responses are orders of magnitude more complex than the aforementioned modelled systems. Thus, it is reasonable to conclude from these studies that the management of a system under uncertainty, such as an aquifer, can result in unforeseen impacts due to unexpected system dynamics. In view of that, our groundwater management plans may not in fact be doing what we think they are.

1.3. How is Groundwater Managed?

All groundwater development has an impact [Alley and Leake, 2004]. Acceptable levels of impacts vary depending on the value of the resource, socioeconomic factors, and the importance of the impacted environment. In some regions development pressure is high and aquifer systems are stressed, while in others, development is low, environmental, and social impacts are mostly absent and the aquifers are not under pressure. It follows that the level of acceptable impact is correlated to developmental stress and balanced against system value [Richardson et al., 2011]. As a consequence, priorities of management plans and acceptable extraction rates vary: in systems with little development stress where interventionist management is unnecessary, management plans designed to monitor the aquifers for impacts without control mechanism are appropriate. In contrast, regions with intense usage competition, historical, or anticipated water shortage stresses, statutory management plans detailing control measures are required to prevent and manage overexploitation. Broadly, groundwater extractions can be considered as either sustainable or nonsustainable depending upon the hydrogeological environment and management time frame. Sustainable groundwater development is defined as the “use of groundwater resources in a manner that can be maintained indefinitely without causing unacceptable environmental, social or economic consequences” [Alley and Leake, 2004]. Unsustainable development is defined as consciously finite and consuming. However, as articulated stated in American Society of Civil Engineers [1987] “it is not a sin to mine groundwater.” The idea that all aquifers can be developed sustainably is erroneous [Kolf and Woolley, 2004], and as long as
more water is extracted from the aquifer than is replenished through recharge, the aquifer is mined [Custodio, 2002]. Social, economic, or political reasons may necessitate groundwater mining to sustain communities or economic productivity in arid regions with little recharge or alternate water sources, and in this way groundwater use is determined through a (often implicit) cost benefit analysis. As long as the benefits afforded by groundwater use exceed the social, economic and environmental costs, usage will proceed until exhaustion, or until costs rise to outweigh benefits or the price of alternate sources [Koundouri, 2004; Tietenberg and Lewis, 2000]. Clearly, water has economic value in all its competing uses and its management as a commodity is crucial for fair, sustainable, and efficient usage [ICWE, 1992], and yet, tariffs are generally below full cost of supply [Rogers et al., 2002]. Pricing water at the true cost of the resource prompts rationing and would result in reduction of demand, supply supplementation, and efficient transfer of water to the most valuable use [Rogers et al., 2002]. Presently, in most cases, water pricing does not reflect true costs and alternate rationing and management methods are employed, including extraction limits, water licensing, water sharing [Council of Australian Governments (COAG), 2004], buybacks [Commonwealth of Australia, 2014], water trading [National Water Commission, 2016b], managed aquifer recharge [Dillon et al., 2009], and carryover. [Department of Sustainability and Environment, 2012; National Water Commission, 2016a] among others. These techniques are increasingly enacted through management plans and leave water managers with the problematic task of deciding where to invest limited resources based upon fragmentary knowledge and often, conflicting factors. Writing a comprehensive, multidisciplinary management plan under uncertainty is a herculean task, with many aspects to consider, stakeholders to appease, and political whims and pressures to navigate through. As a consequence, testability is rarely in the forefront of groundwater management planning.

1.4. How Are Management Plans Tested?

Since the latter part of the 20th century, there has been an increasing focus on groundwater policy and many regions now mandate groundwater management plan development.

Qualitative guides on groundwater management, policy guidelines, [Department of Environment and Primary Industries, 2014; Department of Environment Land and Water Planning (DELWP), 2015; Murray-Darling Basin Authority, 2013; National Water Commission, 2010; United Nations Educational Scientific and Cultural Organization et al., 2003], sustainability strategies [Aeschbach-Hertig and Gieson, 2012; Garduno and Foster, 2010; Gieson et al., 2010; Hamstead, 2009], and nuanced technical investigations abound. To date, the focus has been on writing plans to satisfy legislative requirements and community concerns with testability not prioritized. To our knowledge, very few studies have examined the effectiveness of groundwater management plans distinct from aquifer state, elucidated causes of groundwater management success or failure. Considerable effort has been put into using numerical models to evaluate technical aspects of groundwater management. For example, numerical models have been used to predict impacts of groundwater extraction upon aquifers [Ebraheem et al., 2002; Gorelick, 1983], optimize extraction rates [Bear and Levin, 1967; Casola et al., 1986; Makinde-Oduosa and Marino, 1989; McPhee and Yeh, 2004; Singh, 2012, 2014; Tankersley and Graham, 1994; Wagner, 1995], adjust control based on actual system response [Jones, 1992], manage seawater intrusion [Reichard and Johnson, 2005; Rejani et al., 2008], and investigate implications of economic considerations [Booker et al., 2012; Bredehoft and Young, 1970; Bromley, 1991; Giwer and Sanchez, 1980; Koundouri, 2004; Mulligan et al., 2014], among others. These studies do not, however, investigate if sequential management decisions result in successful resource management or assess the effectiveness of the plan separate from the state of the aquifer.

Many numerical modelling studies have also investigated the impact of various aquifer management decisions [Government of Western Australia, 2009; Guillaume and El Sawah, 2014]. Gallagher [2015b] developed the Groundwater Operational Management Package (GWOMP), an integrated MODFLOW-MT3DMS-SEAWAT modeling package that simulates an aquifer under various management regimes to allow for testing of allocation volumes, carryover, and groundwater trading with a range of triggers and operational rules. The package was implemented on the Pioneer Basin in Queensland and the Campaspe Catchment, Victoria. The aquifer response under various management scenarios was simulated [Gallagher, 2015a; Queensland Government, 2002]. While these exciting studies hold great potential to investigate the intricacies of management, they did not separate the effectiveness of the plan from the state of the aquifer and because they were single site studies; nor did they elicit general attributes of a groundwater management plan that lead it to be testable.
This paper takes the first step toward the quantification of management plan effectiveness distinct from aquifer state, by determining the components of a plan that permit it to be quantitatively evaluated, and is ordered as follows. First, engineering control theory is presented as a framework to structure the process of management within an aquifer system and subsequently; to create a robust, novel, and holistic rubric for assessing plan testability. The rubric, comprising of seven components upon which management plans are evaluated, divides testable plans from nontestable plans. Following the description of the rubric developmental process, a series of 15 Australian groundwater management plans covering a representative range of water uses and challenges, are assessed for testability. Results lead us to contend that many groundwater management plans cannot be quantitatively tested for effectiveness. And that; is a problem.

2. Framework Development

2.1. Control Theory and Application to Hydrogeology

Control is a discipline of engineering concerned with controlling the behavior of dynamic systems modified by feedback, where maintaining system state at a particular value or within certain constraints is the primary objective [Astrom and Murray, 2008]. As Control Theory is a highly precise form of system management, loop effectiveness can be tested and quantified using various performance assessment methods and statistics [Huang and Shah, 2012; Harris et al., 1999], and for that reason it was selected to structure groundwater management plan testability. Control Theory is applied in a diverse range of fields, and the main iterative actions of the control loop are sensing, computation, and actuation, which together form a five component feedback loop (comparator, controller, actuator, plant, and sensor) shown in Figure 1 and detailed in Table 1 [Astrom and Murray, 2008; Owen, 2015].

In a control system, the desired system value is input as a reference, compared to the actual value by the comparator, and depending upon the error, corrective action is computed by the controller and effected upon the system by the actuator [Astrom and Murray, 2008; Owen, 2015]. This results in maintenance of a system close to the desired conditions, and depending upon the manner in which the control mechanism is engaged (with a particular gain or on/off), the system state may oscillate around the desired state. Uncertain system dynamics, and noise in sensing and actuation systems, can introduce uncertainty into the control loop [Astrom and Murray, 2008] and external disturbances, both predictable and unpredictable, act
Table 1. The Primary Components of an Engineering Control Theory Loop

<table>
<thead>
<tr>
<th>Loop Components</th>
<th>Classical Control Theory Definition</th>
<th>What it does</th>
<th>Automotive Cruise Control</th>
<th>Groundwater Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>The objective state that the system strives to maintain and constantly compares to actual system state</td>
<td>Defines the zero error state of a control system</td>
<td>Speed</td>
<td>A measurable management objective such as hydraulic head, well yield, etc.</td>
</tr>
<tr>
<td>Comparator</td>
<td>Where actual value is compared to desired value</td>
<td>Calculates error between actual and desired value and inputs it into controller</td>
<td>Onboard computer</td>
<td>Water Managers</td>
</tr>
<tr>
<td>Controller</td>
<td>The mechanisms that exerts some control measure on the system to maintain the reference</td>
<td>If error is not equal to zero, controller decides appropriate action</td>
<td>Onboard computer</td>
<td>A groundwater management plan dictating control measures and trigger thresholds</td>
</tr>
<tr>
<td>Actuator</td>
<td>The means by which the controller influences the plant</td>
<td>Exerts a force upon plant to actuate desired response</td>
<td>Throttle</td>
<td>Extraction rate</td>
</tr>
<tr>
<td>Plant</td>
<td>The system acted upon by the controller</td>
<td>Fluctuates between system states depending on disturbances and control measures</td>
<td>Car</td>
<td>Aquifer system</td>
</tr>
<tr>
<td>Sensor</td>
<td>Device or observation that monitors actual system state</td>
<td>Measures actual value and inputs it into comparator</td>
<td>Speedometer</td>
<td>Aquifer state is monitored with transducers, water level measurements, and water user’s observations</td>
</tr>
</tbody>
</table>

*The example of an automotive cruise control system is contrasted with aquifer system equivalents shown in the groundwater management control system.*

upon the system and may cause control failure if the controller is unable to adapt to changing conditions [Koenig, 2009]. As there are certain external conditions under which the system cannot maintain the reference, selection of an appropriate and achievable objective is required for successful system control.

The concept of control theory is closely comparable to the objective of natural resource management, which generally aims to maintain developmental impacts within acceptable levels and studies utilizing varying types of control appear in the literature. Anderies et al. [2007] used robust control to manage the uncertainty in social-ecological dynamics typical of fisheries management. They found a trade-off between the robustness and vulnerability of various parameters, where increasing the robustness of certain parameters, was at the expense of others. Roseta-Palma and Xepapadeas [2004] considered the use of robust control “natural” in their assessment of surface water management decisions given the high level of uncertainty associated with climatic variables. Foo et al. [2014] designed a control system to improve management of a river system and balance irrigation requirements within environmental constraints, and several additional studies investigated water management under control, through irrigation channel operation [Choy and Weyer, 2008; Ooi and Weyer, 2008].

However, to date, studies considering groundwater management as a control problem are few. Exceptions include, Ahr [2000] who designed a feedforward/feedback control system to manage groundwater during times of drought by forecasting groundwater levels based on current measurements and calculated extraction cuts when forecasted levels dropped below targets. Brown and Rogers [2006] presented an adaptive groundwater management model with an embedded feedforward control system that used forecasted seasonal rainfall and groundwater elevations to calculate a price for water that maximized social benefits. To our knowledge, groundwater management plans have not been placed within a conceptual framework encompassing the entire managed system and doing so, provides structure and a foundation to assess their testability.

Control theory applied to a managed aquifer system results in the control loop shown in Figure 2.

A reference aquifer state (i.e., hydraulic head levels, well yield, minimum stream, or base flow volumes) is defined as the desired system state and actual aquifer state is monitored, compared to the reference, and if a discrepancy is detected, corrective action is dictated by the groundwater management plan. However, the groundwater management plan does not act upon the aquifer itself, but upon water users and consequently, within the controller is nested control subsystem involving management rules, users, and metering, required to ensure users compliance with management rules (Figure 2). Within this subsystem loop, rules
are dictated to users, translated to extraction rates, and compliance is monitored through metering. A sub-system is not required in many control systems where control laws are implemented by embedded microprocessors, but is necessary in natural systems due to the fact that policy is implemented by people and institutions, which adds additional error and uncertainty [Anderies et al., 2007]. Just because the plan stipulates restrictions does not necessarily mean users will comply. This is due to the fact that aquifers are common pool resources and any water not pumped today may be pumped by others tomorrow [Gisser and Sanchez, 1980; Negri, 1989]. Measuring extractions through metering and other methods confronts the issue of noncompliance.

Fluctuations in extraction rates and external disturbances, including agricultural market variations and climate perturbations, act upon the aquifer system and result in changes to user behavior (extraction rates) and the aquifer state (Figure 2). The dashed lines in Figure 2 represent the path of a monitoring plan which as described in section 1.3, is appropriate for unstressed systems. Monitoring plans are concerned with observing the system state to ensure it remains within acceptable levels and generally contain the proviso that more interventionist management will proceed if monitoring indicates the occurrence of unacceptable impacts. The solid line control loop represents the path of a management plan where control measures are periodically enacted.

Groundwater management is, however, more complex and nuanced than is captured by classical control theory. The frequency of the control loop cycle is low, generally yearly, and complex time lags exist between action and response. In addition, due to three-dimensional aquifer heterogeneity, considerable uncertainty surrounds the plant and the monitoring network. More advanced approaches exist (i.e., adaptive [Åström and Wittenmark, 2013], optimal [Kirk, 2012], stochastic [Åström, 2012], and robust control [Dullerud and Paganini, 2000]) that may allow a greater exploration of the intricacies of groundwater management but are less conducive to a broad assessment of multiple plans and are beyond the scope of this paper.

2.2. Methodology
Control systems are structured to form feedback loops that almost always contain the five components described in section 2.1. Therefore, in order to set up a groundwater management control loop,
hydrogeological elements equivalent to the five components, the reference, and external disturbances were determined (Table 1) through an iterative process of control loop appraisal and management plan review. Components of a groundwater management control loop are as follows:

1. Reference: Management Objective
2. Comparator: Data Review and Analysis
4. Actuator: Method of Control
5. Plant: Aquifer characterization
6. Sensor: Monitoring
7. Disturbances: Driver Monitoring

The role of each component in a groundwater control system is listed in Table 2, and together the components form the basis of the testability rubric. In general, the components forming the rubric were identified in cited groundwater management guidelines, and plan guidelines recommended definition of acceptable level of stress, aquifer characterization, control measures, and the description of a measurable management objective for groundwater dependent values [DELP, 2015; NWC, 2010]. Commonly, the guidelines focused predominantly on risk assessment and stakeholder engagement which befitted their purpose of effective management rather than testability. The purpose of the rubric is to determine whether a management plan contains the required components of a control loop because, only plans forming complete control loops are testable. Certain factors relevant to comprehensive groundwater management plans such as market/trading rules; environmental water entitlements, and connectivity of ground and surface water while not mentioned explicitly in the framework, may be incorporated into management plans through the definition of the objective. The objective does not necessarily just pertain to groundwater head, but to any measurable aspect of the aquifer system.

To obtain a relative and systematic indication of the testability of various plans, an index method was adopted to allow calculation of a numerical testability value. The method adopted was similar to the DRASTIC methodology developed by the United State Environmental Protection Agency (USEPA). DRASTIC allows a numerical value reflecting the relative susceptibility of groundwater to pollution, to be systematically calculated, in any hydrogeological setting [U.S. Environmental Protection Agency (USEPA), 1987]. Greater index values indicate higher vulnerability to pollution. DRASTIC uses seven physical factors relevant to contaminant transport to construct a numerical ranking system composed of weights, categories, and ratings. The relative importance of each factor is evaluated in respect to all other factors and assigned a weight between 1 and 5, with most important factors receiving a weight of five and least important weights of one. Then each factor is divided into categories with differing impacts on pollution potential and assigned a rating from 1 to 10, again based on significance. Factors strongly influencing pollution potential, such as shallow water table or highly conductive sediments, are given high ratings and less important factors (deep water tables, confining layers) are rated lower [USEPA, 1987].

A similar numerical ranking system was applied to the control loop components. Relative weights were determined based upon the user estimated importance of the particular component to a functioning control loop. Weights ranged from 1 to 5, with 5 being most important for testability (Table 3). A SMART management objective, defined as Specific, Measureable, Assignable, Realistic, and Timely [Doran, 1981], and an adequate monitoring network, were determined to be the most significant components to a functioning control loop and rated at five (Table 3). High weights were assigned to objective and monitoring because it is impossible to test plans without an objective, acceptable impact, or if monitoring is unable to determine system relative to the objective. Conversely, driver monitoring was weighted at one because while unpredictable external disturbances can influence system state, driver monitoring is not crucial for control loop function. It is important to make the distinction that an appropriate method of control is important for effective function of a control loop, but is less important for a testable loop and was weighted at three.

For each component, certain requirements must be met in order to fully adhere to the control theory framework. The components were divided into categories based on the level to which specified requirements were met—from completely to incompletely—and were evaluated against each other to determine relative importance (Table 3). Ratings were assigned to categories and varied between 1–10, with 10 being greatest adherence to the control framework and 1 being the least (Table 3). For example, to fully meet criteria 1
Table 2. Plan Testability Assessment Components*  

<table>
<thead>
<tr>
<th>Component</th>
<th>Relevance of Component</th>
<th>Control Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMART Objective or Acceptable Impact (O)</td>
<td>An objective serves as a reference for the control loop and is required for comparison between current and desired state. An objective dictates management actions and informs the control loop. Appropriate objectives include aquifer systems states such as waterlevels, stream baseflows, or well yields. The SMART acronym [Goran, 1981] is used in management to define meaningful objectives and stands for: Specific: Measurable: Achievable/Assignable: Realistic/Relevant: Timely: OR, Quantitative level of acceptable aquifer impact.</td>
<td>Reference</td>
</tr>
<tr>
<td>Aquifer and Use Characterization (A)</td>
<td>Reliable hydrogeological data, specifically the amount, location, availability, and demand upon groundwater resources is crucial to make informed management decisions, accurately evaluate potential impacts, determine acceptable levels of impact, anticipated use and establish a baseline state. (Aquifer type, yield, waterbalance, head, GDEs, pumping rate). If the potential response of a system to stimuli is unknown, management by classical control theory is challenging.</td>
<td>Plant</td>
</tr>
<tr>
<td>Method of Control (C)</td>
<td>The control is the actuation mechanism engaged by controller (the groundwater management plan) to effect system change and minimize the error between objective and actual system state. Control mechanisms utilized by management plans include hydrogeological investigation-based license entitlement capping, water restrictions, restricted groundwater trading, carryover, buyback, and managed aquifer recharge and have been demonstrated (through experience or numerical modeling simulations) to influence the aquifer state.</td>
<td>Actuator</td>
</tr>
<tr>
<td>Aquifer System Monitoring (M)</td>
<td>Monitoring, used to evaluate system state and inform management decisions, must be sufficient to determine system state to allow comparison with objective so that dependent on the aquifer state; the controller can dictate necessary action.</td>
<td>Sensor</td>
</tr>
<tr>
<td>Data Review and Analysis (D)</td>
<td>At appropriate frequency monitoring, data should be analyzed and aquifer state computed and compared to the objective state. The discrepancy between objective and actual state determines control actions. For example, if a threshold aquifer level is reached, the groundwater management plan may mandate restrictions and water entitlements.</td>
<td>Comparator</td>
</tr>
<tr>
<td>Driver Monitoring (E)</td>
<td>Drivers include external disturbances and groundwater extractions. External disturbances can perturb control loops. Management-related system impacts must be separated from impacts due to external drivers to determine plan efficiency and driver (e.g., weather, extractions) monitoring data are required for this purpose.</td>
<td>Disturbance</td>
</tr>
<tr>
<td>Frequency and Review (F)</td>
<td>The frequency of the control loop cycle determines how often the actual state is compared to desired state. This influences speed at which issues are addressed. Review timeframe of a plan indicates if a plan is static or dynamic.</td>
<td>Controller</td>
</tr>
</tbody>
</table>

*The following components are required for a management plan to form a control loop and thus manage an aquifer using systems control.

...and receive a rating of 10, the plan must define a SMART objective, and quantitatively define an acceptable level of system impact. Plans that define neither are assigned a rating value of 1 (Table 3).

An additive model allows for calculation of a relative testability index, T:

\[ T = 5O_1A_1C_1E_1F_1 + 3O_2A_2C_2E_2F_2 + 10O_3A_3C_3E_3F_3 \]  

(1)

Where the right hand variables are as defined in Table 3, the subscript r denotes the rating for the criteria and the subscript w denotes the weight of the criteria. The greater the testability index, the more closely the plan adheres to the control framework and the greater the likelihood the plan is conducive to quantitative assessment. To reflect the diversity of plans, assessments were divided into two types: monitoring and management, depending on whether regional aquifer stress necessitates observation or management of...
Table 3. Plan Testability Assessment Rubric*

<table>
<thead>
<tr>
<th>Component Category</th>
<th>Rating (R)</th>
<th>Weight (W)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMART Objective and Acceptable Impact (O)</td>
<td>10</td>
<td>5</td>
<td>Plan defines SMART objective. Plan quantitatively defines an acceptable level of impact.</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Plan defines SMART objective. Plan subjectively defines an acceptable level of impact.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Plan subjectively defines objective. Plan subjectively defines an acceptable level of impact.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Plan subjectively defines objective. Plan does not define acceptable level of impact.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Plan does not define objective or acceptable level of impact.</td>
<td></td>
</tr>
<tr>
<td>Aquifer Use and Characterization (A)</td>
<td>10</td>
<td>1</td>
<td>Plan describes baseline system state, hydrogeological parameters and environment.</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Plan describes baseline system state, but not hydrogeological parameters and environment.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Plan describes hydrogeological parameters and environment but not baseline system state.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Plan does not describe hydrogeological parameters, environment or baseline system state.</td>
<td></td>
</tr>
<tr>
<td>Method of Control (C)</td>
<td>10</td>
<td>2</td>
<td>Plan has method of control demonstrated through simulation or practice to influence aquifer state sufficiently to meet objective; or method of control is based on hydrogeological analysis.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Plan has unproven method of control</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Plan has no method of control</td>
<td></td>
</tr>
<tr>
<td>Aquifer System Monitoring (M)</td>
<td>A</td>
<td>10</td>
<td>Monitoring design as described in plan is adequate to determine aquifer state relative to objective or acceptable impacts.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Monitoring design as described in plan is inadequate to determine aquifer state relative to subjective objective or acceptable impacts.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Monitoring network is not described in plan and adequacy is unknown OR objective or acceptable impacts are unknown.</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>4</td>
<td>Monitoring design can establish if aquifer is responding to development in a manner consistent with understanding of aquifer characteristics and if not; a review is triggered.</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Monitoring design can establish aquifer response and behaviour is not consistent with current aquifer understanding, but no review is triggered.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Monitoring design is unable to establish aquifer response, or monitoring design is unknown.</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>1</td>
<td>Plan contains a means to determine adequacy of monitoring and method to rectify if inadequate.</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>No means to determine or rectify monitoring inadequacies or monitoring design is unknown.</td>
<td></td>
</tr>
<tr>
<td>Data Review and Analysis (D)</td>
<td>A</td>
<td>10</td>
<td>Sufficiently reliable monitoring data were periodically analyzed and compared to objective.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Variable quality data are analyzed and compared to objective.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Analysis frequency detailed but not analytical process.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Data are unreliable and not analyzed or compared to objective, or analytical process is not detailed.</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>1</td>
<td>Review frequency consistent with decision making frequency and aquifer response time.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Review frequency specifically described.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Review frequency variable.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Review frequency inconsistent with decision making frequency and aquifer response time, or no decisions are made (no adaptability).</td>
<td></td>
</tr>
<tr>
<td>Driver Monitoring (E)</td>
<td>R</td>
<td>10</td>
<td>External disturbances monitored and accounted for in plan. Groundwater extractions are monitored and recorded.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Certain external disturbances monitored and accounted for.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>External disturbances not monitored or accounted for in plan. Bores not metered.</td>
<td></td>
</tr>
<tr>
<td>Frequency and Review (F)</td>
<td>A</td>
<td>10</td>
<td>Plan contains clear quantitative triggers for review and switch to different management action or plan type.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Plan contains qualitative trigger for review OR switch of management actions.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Plan does not contain trigger for review or switch of management actions.</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>2</td>
<td>Contingency plan for unexpected response.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Inadequate contingency plan.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>No contingency plan.</td>
<td></td>
</tr>
</tbody>
</table>

*Plan Testability Assessment Rubric. Each component is divided into several rated categories dependent upon degree of fit to the control theory framework, and higher-rated categories have a greater fit to the framework. Each component is weighted dependent upon perceived importance to control loop, with 5 being greatest and 1, least. Notes: R denoted Rating and W denotes Weighting.

**Not required for Monitoring Assessment.

ex extractions. Table 4 outlines a series of questions to determine if a management or monitoring assessment is appropriate. Calculation of testability indexes for monitoring plans does not include the control criteria (C 50).

2.3. Data

In Australia, groundwater plan development is mandated under the Federal Government’s National Water Initiative [COAG, 2004], and the nation is divided into 361 groundwater management areas with 256 commenced plans. While policy guidelines exist, such as the National Water Initiative Water Planning Guidelines [National Water Commission, 2010], a statutory, nationally consistent groundwater management plan
development methodology is lacking in Australia, with States granted autonomy to develop plans. As a consequence plans vary broadly in methodology. Due to the volume of management plans within Australia, it was necessary to select a representative sample of plans as case studies to evaluate the testability assessment method. The rationale behind selection of each plan is detailed in Appendix Table A1. Fifteen plans were assessed from across Australia with a broad spread of jurisdictions, geological environments, aquifer stresses, and primary uses of the resource, including: urban water supply, irrigation, oil/gas development. Objectives, impacts, and issues represented the breadth of challenges faced in Australia and included water supply, salinity control, seawater intrusion, mining impacts, and reliance upon nonrenewable water sources. Selected plans are shown in Table 5, and for ease of reading individual plans are referred to by corresponding Plan Number.

2.4. Application of Testability Assessment Method

The methodology used to apply the rubric (Table 3) to the plans was as follows:

1. Plans and supporting documentation were reviewed for control components defined in rubric.
2. Plans were assigned a score out of 10 for each criteria and total score was summed to yield a T index value.

In this assessment, it was assumed that groundwater monitoring occurred as specified in the plan, and only data specified in the management plan and one supporting document were considered.

3. Results

This section presents the overall requirements considered necessary for a plan to be testable, followed by a detailed presentation of the seven components of the rubric, and the degree each plan met them (Table 6).

3.1. Plan Testability Assessment Overview

Assuming causality has been established, and from a purely physical perspective, whether a plan improves aquifer state, avoids unacceptable impacts or achieves objectives are the only standards upon which effectiveness can be measured. Therefore, a plan must describe a quantitative objective or level of impact to be

Table 4. Questionnaire to Determine Management or Monitoring Assessment Type

<table>
<thead>
<tr>
<th>Question</th>
<th>Monitoring Plan</th>
<th>Management Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the degree of aquifer stress? Is the aquifer overallocated?</td>
<td>Low-No</td>
<td>High-No</td>
</tr>
<tr>
<td>What is the projected increase in pumping demand?</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Is aquifer system and external monitoring conducted?</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Is management intervention required?</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 5. Australian Groundwater Management Plans Selected for Testability Assessment Using Testability Rubric

<table>
<thead>
<tr>
<th>Management Plan</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Hopkins Corangamite Groundwater Catchment Statement</td>
<td>SRW [2014a]</td>
</tr>
<tr>
<td>2 Warrion WSPA GWMP</td>
<td>SRW [2010]</td>
</tr>
<tr>
<td>3 Upper Oven River WSPA Water Management Plan</td>
<td>GMW [2011]</td>
</tr>
<tr>
<td>4 Shepparton Irrigation District LMP</td>
<td>GMW [2015a, 2015b]</td>
</tr>
<tr>
<td>5 Lower Campaspe WSPA GWMP</td>
<td>GMW [2012a, 2012b]</td>
</tr>
<tr>
<td>6 Water Sharing Plan for Upper and Lower Namoi Groundwater</td>
<td>DOW [2006]</td>
</tr>
<tr>
<td>7 Water Sharing Plan for the NSW Great Artesian Basin</td>
<td>DOW [2011]</td>
</tr>
<tr>
<td>9 Alice Springs Water Resource Strategy</td>
<td>Department Of Natural Resources Environment The Arts And Sport [2009]</td>
</tr>
<tr>
<td>10 Tindall Limestone Aquifer Water Allocation Plan</td>
<td>Department Of Natural Resources Environment</td>
</tr>
<tr>
<td>11 Gnaangara Groundwater Areas Allocation Plan</td>
<td>DOW [2009]</td>
</tr>
<tr>
<td>12 Pilbara Groundwater Allocation Plan</td>
<td>DOW [2013]</td>
</tr>
<tr>
<td>13 Water Allocation Plan Far North Prescribed Wells Area</td>
<td>South Australian Arid Lands Natural Resource Management Board [2009]</td>
</tr>
<tr>
<td>14 Sassafras Wesley Vale Water Management Plan</td>
<td>Department of Primary Industries Parks Water and Environment [2012]</td>
</tr>
<tr>
<td>15 Murray Darling Basin Plan</td>
<td>Murray Darling Basin Authority [2012]</td>
</tr>
</tbody>
</table>
considered testable regardless of the overall T Index. Several plans considered untestable achieved medium T Index values showing the importance of accompanying use of the rubric with a detailed discussion of each component score. For this purpose, an interpretation of T Index values for each component is provided in Table 7. Use of Table 7 in conjunction with the rubric illuminates positive and negative aspects of plans, explains why particular scores may have been obtained, and suggests mechanisms for increasing plan T Index scores.

Results of the Management Plan Testability Assessment are shown in Table 6. Seven of the 15 (46%) assessed plans were determined to be testable and normalized values of T Index calculated using the proposed weighting system ranged from 0.23 to 0.85 (with one being full marks). Foresee of discussion, ratings were divided into high (10–7), medium (6–4), and low (3–1) divisions. It is reiterated here that T Indexes indicate testability and are not reflective of the overall effectiveness of each plan, which is currently unknown. Plans receiving high T Index values are not necessarily effective plans and vice versa.

### 3.1.1. Objective or Level of Acceptable Impact

The requirements for component one were the definition of a SMART objective or defined level of acceptable impact (Table 2) and seven plans [3, 9, 10, 11, 12, 14, 15] received a high rating for component one. Objectives receiving a high score included minimum river flow volume or level [3, 10], rate of aquifer depletion [9], maximum extraction volumes [14], groundwater levels, and quality [11, 12, and 15]. The remaining plans [1, 2, 4, 5, 6, 7, 8, 13] only qualitatively stated objectives, which included mitigation of particular impacts, for instance, salinity [4] or more generally that resources be managed in an equitable and sustainable manner and adverse impacts avoided [1, 2, 5]. The common limitation of these objectives was the omission of defined measurable acceptable impacts.

### 3.1.2. Aquifer Characterization

A description of the baseline system state, hydrogeological parameters, and environment was required for a high rank of aquifer characterization and this was achieved by 73% of plans. All plans named target aquifers and the level of hydrogeological detail varied from extensive descriptions, maps, or figures of geological environment, aquifer characteristics, and baseline state [3, 4, 9, 10, 11, 13, 15], to a brief description of main target aquifers, system state, and hydrogeological parameters [2, 5, 6, 14]. Plans receiving a lowscore...
lacked detail on baseline system state [1, 7, 8, 12]. Generally, the degree of aquifer characterization increased with intensity of aquifer use.

### 3.1.3. Method of Control

Two monitoring plans [1, 2] were not assessed against the control criteria due to low aquifer stress and projected pumping demand (Table 4). Of the remaining plans all except one, which contained a control measure where allocation volume was not restricted [8], contained control measures based on hydrogeological analysis or methods demonstrated to influence aquifer head levels (Table 2), and subsequently, received a high ranking. All plans required, or contained provisions detailing the planned introduction of [14], the licensing of commercial extractions. All except one [4], where pumping was encouraged for salinity mitigation, contained extraction volume caps, frequently manifested as current licensed volumes. Extraction limits were commonly managed through various annual or seasonal entitlement allocations with volume determinations based on estimated available water [6], rolling average recovery levels [5], model predictions [10], and recharge estimations [11]. Most plans detailed a restriction strategy with three plans outlining staged restrictions, based on river flow rates [3, 10], or head levels in trigger wells [5]. Other plans reporting restriction triggers were at the discretion of management authorities under various Water Acts [6, 7, 9], and included a subjective definition of adverse impacts. Additional methods of control included trading (most plans), carryover [6, 7, 15], or potential for implementing carryover [5], entitlement buy-back [15], [Commonwealth of Australia, 2014] and managed aquifer recharge [11, 9].

### 3.1.4. Aquifer Monitoring

The monitoring criteria were divided into three subcriteria (a) whether monitoring can determine system state relative to objective or acceptable impacts, (b) aquifer response, and (c) adequacy of monitoring. In order for the first part to be assessed, plans were required to describe the monitoring network. Plans omitting

<table>
<thead>
<tr>
<th>Component</th>
<th>Interpretation of T - Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMART Objective and Impact (O)</td>
<td>Plans scoring medium to low for this component require work explicitly defining what the plan aims to achieve, or avoid, in a manner that can be measured. Examples include: indicators of aquifer state such as groundwater levels, quality, or rate of decline, maximum extraction volumes, and minimum streamflow volumes. Plans scoring medium to low are considered untestable and require prompt review.</td>
</tr>
<tr>
<td>Aquifer Use and Characterization (A)</td>
<td>A medium to low score indicates the need for further hydrogeological investigation. Such investigation would require greater detail on aquifer characteristics, baseline system state, hydrogeological parameters, main uses, geologic maps, environment, groundwater-dependent ecosystems (GDEs).</td>
</tr>
<tr>
<td>Method of Control (C)</td>
<td>Management plans scoring medium to low contain an inadequate method of control. Methods of control demonstrated to influence aquifer state include: extraction caps or restrictions with specified hydrogeological trigger levels; entitlement, and allocation systems; fees and tariffs; water trading, etc.</td>
</tr>
<tr>
<td>Aquifer System Monitoring (M)</td>
<td>Plans will score low if monitoring network is not described (number and location of monitoring bores, monitoring purpose and frequency, historical trends); an objective is undefined; if it cannot be determined if the monitoring network is capable of measuring or capturing changes in objective parameters; or if unexpected system responses (i.e., rapid propagation of drawdown cones) cannot be detected by current network. In order to ensure the network monitors everything necessary, the objectives and requirements of the monitoring network must be clearly stated. Plans lacking this detail will score low because it cannot be determined if monitoring is adequate.</td>
</tr>
<tr>
<td>Data Review and Analysis (D)</td>
<td>Plans scoring low require either greater investment in data collection to ensure more reliable data or more frequent assessment of the data because low scores indicate data are generally unreliable or are not used to their full potential.</td>
</tr>
<tr>
<td>Driver Monitoring (E)</td>
<td>Plans scoring medium to low indicate that expansion of external driver monitoring is required. External drivers include climate data such as temperature, precipitation, humidity, and evapotranspiration. Additionally, monitoring of volume extractions through metering of wells, field log books, or a review of extractions rates reported to regulatory agencies is required for a high score in this component.</td>
</tr>
<tr>
<td>Frequency and Review (F)</td>
<td>Plans scoring low lack alternate management triggers such as minimum flow rates, groundwater levels in wells and GDEs or water quality, maximum extraction volumes. A lack of plan review triggers, such as minimum groundwater levels or change in availability of water, could also result in a low score for this component.</td>
</tr>
</tbody>
</table>

*This table is designed to be used in conjunction with Table 3 to allow examination of individual component scores.

*Not required for Monitoring Assessment.
monitoring descriptions received a low rating [1, 7, 8]. The majority of plans described monitoring bores, gauging stations, frequency, type and purpose of monitoring, historical trends, thresholds and triggers adequately, and hence, from a qualitative assessment, appear capable of detecting unexpected aquifer responses. Whether the monitoring network was capable of determining aquifer state compared to an acceptable impact could only be determined for plans with defined acceptable impacts (i.e., scoring highly for Criteria 1) and plans [2, 5, 6] with subjective objectives received a low score. Despite the extensive monitoring network described in [4], the lack of a defined impact precluded this plan from receiving a high score.

Review triggers were present in two plans relating to groundwater level in wells [5] and change in water availability, land use, or aquifer understanding [11]. The stated monitoring purpose of most plans was evaluation of aquifer head levels but monitoring was also conducted for water quality purposes [4] and extraction rate [9]. Assessing the adequacy of a monitoring network is difficult and resource intensive and none of the plans contained specific provisions to determine if monitoring was adequate. Where specific objectives or aquifer impacts were specified, monitoring design was appropriate to capture changes in the aquifer system. Several plans [13, 14] reported a high degree of uncertainty surrounding the hydrogeological environment and recommended expansion of the monitoring network. Consequently, they were assigned medium scores due to questionable data reliability. The sparse nature of data collection introduces additional uncertainty, but provided data are adequate, infrequently monitored plans, and plans covering regions with high hydrogeological uncertainty can be testable [14].

3.1.5. Data Review and Analysis
Most plans described the data collection and analysis processes. Plan review frequency and data reliability were predominantly high. Generally, aquifer state was compared to the objective state annually when triggers defined in the plan were used to determine allocation volumes for the following year. This was consistent with the data review and reporting period. However, two plans dictated immediate introduction of restrictions when threshold river flow rates [5] or groundwater heads in wells [14] [Chris Cleary, 2016] were reached, making response and review time considerable faster. Aquifer response time was mentioned in the background information of the NSW Great Artesian Basin Plan [7] but remaining plans did not cover aquifer lag times, which varied from rapid in the unconsolidated sedimentary aquifers of the Upper Ovens System [3], to very slow in fossil aquifers where extractions have greatly exceeded recharge replenishment, such as the Amadeus Aquifers of Alice Springs [9] and the Great Artesian Basin [7, 8].

3.1.6. Driver Monitoring
Driver monitoring includes climate and extractions volumes. Considering that climate data, including temperature, humidity, and rainfall, are recorded by the Bureau of Meteorology at many sites across Australia (http://www.bom.gov.au/climate/data/), to receive a high rating plans must specify monitoring of extractions and degree of system stress. Generally, in unmetered wells, extraction rates were recorded by water users in log books and reported to the appropriate regulatory body, and in metered wells, meters were read directly or measured by telemetric systems. Data are held by managing authorities in each state or territory; for example, Victorian extraction information is stored in the Victorian Water Registry [Government of Victoria, 2014]. Metering requirements varied from nonexistent [4], to mandatory [5, 9, 10], to volume-dependent requirements such as all licensed wells producing volumes greater than 10 mL/yr [1, 3], or wells producing greater than 0.5 mL/yr [11] requiring meters. In the NSW plans, whether bores require meters or log books was dependent on the water source [6, 7]. All plans except two, where metering was not yet conducted [14], or planned to be discontinued [4], achieved a high score for driver (climate and extraction) monitoring. However, it is acknowledged that extractions from many wells are not metered or monitored and extraction volumes are crude estimates.

3.1.7. Feedback and Review
Clear quantitative alternate management triggers were present in 11 plans and included; river flow rates [3], model predicted base flow volumes [10], groundwater levels in wells [5, 9, 14], or GDEs [12, 15], and water quality [15]. The response to triggers was generally a decrease in water available to users, or increased monitoring and altered extraction regimes [12]. Other plans had extraction volume triggers, for example, reaching, or exceeding allocation volumes that corresponded to actions such as license and allocation capping or reduction, introduction of efficiency measures, and plan review and update [6, 11]. The supporting documentation of one plan indicated extraction volumes could trigger a review of extractions and environmental requirements [7]. None of the plans detailed contingency plans for unexpected aquifer responses, however, two plans contained plan review triggers; specific groundwater level [5] and a change in the availability of
4. Discussion

Uncertainty is a fundamental reality of groundwater management that at times can lead to a management style that is more reactive than proactive, even somewhat, impromptu. It is recognized that despite the best intentions, circumstances such as an extended droughts, climate change impacts, or approval of water intensive resource developments, may conspire to prevent the implementation of management plans precisely as written. It is beyond the scope of this paper to evaluate the meticulousness of plan implementation, but it is recognized that this is an important aspect of water governance.

4.1. Testability Rubric Performance

The testability rubric gave an indication of plan testability by determining if required criteria were present and isolated structural shortcomings of the plans. Overall, the rubric was robust in that it was capable of evaluating plans spanning a range of aquifer types and extraction drivers. Fundamentally, the rubric highlighted the critical importance of clearly defined acceptable levels of developmental impacts and SMART objectives. Plans lacking defined levels of impact or specific objectives; could not be quantitatively assessed for effectiveness.

4.2. Testability Rubric Critique

4.2.1. Subjectivity of Ranking System

The testability rubric is a subjective rating method where numerical scores were assigned to various attributes to produce categories of potential testability. The point rating system was developed using professional judgment, and is somewhat subjective, and arguments could be made for altering values, and subsequently, testability indexes. This subjectivity is unavoidable because criteria were rated and weighted to reflect user—decided importance. Criteria with the greatest weighting are keystone criteria and subsequent criteria are dependent upon them. For example, if an objective is not defined it makes it difficult to assess if a method of control is adequate to meet objectives, or whether monitoring can assess the objective, or if the analysis method is capable of comparing the reference to the system state. It was determined that an objective or acceptable impacts is fundamental to management and analysis, but depending upon management objectives; other criteria are less critical. For example, if the management objective is maintenance of water levels, it could likely be achieved through careful monitoring of water levels and detailed aquifer characterization would not be required. Moreover, a change in one criterion impacts other criteria, i.e., if the objective is changed from monitoring water levels to maintaining a wetland, a modification of the monitoring network is required.

The capacity of individual components to skew T Index score was examined by comparing plan scores obtained under the proposed weighting system with those obtained using an equal weighting system, where each component was assigned a weighting of one. Testability indexes for each plan were then normalized and the percentage change due to the altered weighting system was calculated for each plan. Using an equal weighting system resulted in up to a 33% percentage increase in T Indexes for plans [1, 6] that scored low in highly weighted categories such as Objective and Monitoring. The T Index of plans with high overall scores generally decreased on the order of 5–10% (Table 6). Clearly, T Index scores are sensitive to highly weighted components. This is an intentional artifact of the method and reflects component importance. The degree of sensitivity of the scoring system to the weighting of individual components was
explored by alternatively scoring a plan highly in all components except for a highly weighted, and subsequently, low weighted component. The percentage increase from a low to a highly weighted component score was 16% (Table 6). As shown on Table 6, several plans attained T Index values of around 0.65 [5, 6] using the proposed rubric weighting, yet were considered untestable due to a lack of SMART objective. When equal weighting was used, the T Indexes increased by approximately 7%, and illustrated the necessity of evaluating the T index value in conjunction with an examination of individual component scores.

The rubric was developed to be applicable to a range of plans and is not site specific. Hence, it is inevitable that some plan deficiencies and nuances may be overlooked. Aquifers are complex, dynamic systems with multiple scales; micro and macro-scale phenomena may have multiple layers of feedback loops. Many real-world complexities are not accounted for in the framework, including the extensive variation in groundwater flow rates and considerable lag times between pumping and observed aquifer response, which may cause amplification or dampening. For example, a water level drop at a sensitive DGE may prompt an allocation cut and reduced pumping, but due to the delayed response time, by the time pumping reduces, levels may have recovered or alternately, dropped perilously low [Bredehoeft and Young, 1970]. Very few plans mentioned aquifer response time even though consideration of this parameter is crucial to set appropriate management objectives and realistic expectations [Meals et al., 2010].

4.2.2. Application Challenges

Challenges encountered applying the rubric included difficulties reviewing a large diversity of plans using a single framework because aspects present in some plans were absent in others, and variations in plan format made pinpointing the presence of criteria difficult. Plans generally fell into two format types: legislative and report style. Legislative plans [6, 7, 8, 15] were scribed in characteristic legislative language; “archaic, foreign and with uncommon words, long complex sentences, repetition and a total absence of color and humanity” [Maley, 1987]. The legalistic language made comprehension challenging and Maley [1987] reports even lawyers and judges find legislative language can present at times “an impenetrable barrier to understanding,” which “for ordinary people; might as well be in a foreign language.” The comprehensibility and accessibility of such plans was an issue, with the loss of assessment criteria within legislative jargon or burial beneath irrelevant information and elegant, yet circumspect, language. Plans can be theoretically sound yet the manner in which they are written introduces unnecessary complication and confusion. Previous research indicates that when users understand the rationale behind management actions and the likely effect upon the future availability of water, they are more likely to comply with management actions [Nelson and Casey, 2013]. Management plans are a primary way to convey management actions to users, and while not necessary for testability; accessibility and comprehensibility are important for plan effectiveness.

In practice, management of groundwater is very difficult due to great uncertainty, limited resources, imperfect data to base decision making upon and various other pressures. Consequently, evaluation of plans is expensive, time-consuming and not currently possible in some regions. To make the evaluation process easier, the testability assessment rubric aims to prove a quick and inexpensive method to highlight areas of plans requiring improvement and, where possible, may also be used to quantitatively assess plans as part of a review framework.

4.2.3. Are the Plans Working?

The period since implementation of the seven testable plans ranged from 3 to 9 years and provided empirical evidence of aquifer state under plan management. Whether stated objectives were reached during this period and the associated degree of management intervention was examined, to illustrate the fact that often there is a lack of causality between achieving objectives and management intervention. For example, a specific objective of one plan [3] was prevention of extraction-related cease to flow events in the Upper Ovens River and tributaries. During the four years the plan was in place, a roster system for groundwater extraction was enacted three times ranging in duration from 3 to 8 weeks at a time. Twice allocations were cut to 75% of entitlements for periods of 1 month and approximately 6 weeks, respectively [GMW, 2012a, 2012b, 2014a, 2015a, 2015b]. Ultimately there were no recorded cease to flow events [Pethybridge, 2016], yet without further stress to the plan and an analysis of the individual impact of the roster system and allocation cuts upon aquifer base flow to streams, reaching objectives may have been simply happenstance of a few relatively wet years.

Another plan stated an objective as maintenance of groundwater levels within historical norms [14], and set allocation volumes equal to historical use. There were provisions for extraction restrictions within the
plan, however, levels remained above thresholds during the 3.5 years of plan management and no restrictions were imposed. Thus the primary control mechanisms of the plan were not activated. In both examples, plan objectives were reached, but we contend the effectiveness (or ineffectiveness) of the plans remains unknown for two reasons:

1. Causality for achieving objective aquifer states must be attributed to management interventions, but this was not clearly demonstrated [3].
2. The control measures of the plan must be activated due to systems stress [14].

These two points comprise the evaluation of plan effectiveness detailed in the introduction. Evaluation of effectiveness is the second of the two components required to determine the effectiveness of management plans, the first being that the plan is testable. Determining the effectiveness of plans is a current topic of research and a numerical method to evaluate the effectiveness of testable plans is presently under development.

5. Conclusion

Groundwater management plans lacking a clearly defined objective or acceptable level of developmental impact cannot be tested for effectiveness and approximately 47% of analyzed plans were not conducive to quantitative assessment. These results are significant because through testing, plans improve, and effective groundwater management is crucial to ensure water resource security. Control theory presented a useful framework to structure groundwater management upon in order to develop a testability rubric capable of assessing a diverse range of groundwater management plans. The importance of groundwater as a resource demands effective management, and that plans are constructed in a manner to make them testable, is a fundamental prerequisite of an assessment of the effectiveness. The presented testability assessment rubric provides a pathway for the development of rigorous management plans that can be quantitatively analyzed. Only by quantitatively assessing the effectiveness of groundwater management plans, can we systematically learn how to develop better management plans.

Appendix A

Table A1 outlines the rationale behind the management plan selection process.

<table>
<thead>
<tr>
<th>Table A1. Plan Selection Rationale*</th>
<th>Management Plan</th>
<th>Main WaterUse</th>
<th>Issues</th>
<th>Selection Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Hopkin Corangamite Groundwater</td>
<td>Urban</td>
<td>Paucity of data in some areas.</td>
<td>Hopkins Corangamite Local Management Plan is an example of a non-statutory Local Management Plan. Geelong sources most of its water from the West Barwon Reservoir in the Otways covered by the Gerangame plan. Victorian Plan</td>
<td></td>
</tr>
<tr>
<td>Catchment Statement and Gerangame</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mete Local Management Plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Warrion WSPA Groundwater Manage-</td>
<td>Dairy, Irrigation, D&amp;S</td>
<td>Minimal aquifer stress, aquifer uncertainty</td>
<td>The Warrion Plan is a statutory Monitoring Plan with no explicit control measures. Victorian Plan</td>
<td></td>
</tr>
<tr>
<td>ment Plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Upper Ovens River Water Supply</td>
<td>Irrigation, Urban</td>
<td>Fast aquifer response, high surface and water connectivity</td>
<td>The Ovens plan conjunctively manages surface and groundwater, has a very short system lag time and immediate restrictions when thresholds are reached. Victorian Plan</td>
<td></td>
</tr>
<tr>
<td>Protection Area Water Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Shepparton Irrigation District</td>
<td>Salinity control, Irrigation</td>
<td>Salinity</td>
<td>Groundwater pumping is encouraged under the Local Management Plan to mitigate impacts of salinity, alternate objective. Victorian Plan</td>
<td></td>
</tr>
<tr>
<td>Local Management Plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Lower Campaspe WSPA Groundwater</td>
<td>Irrigation, D&amp;S, Urban</td>
<td>Salinity in some areas, water shortages</td>
<td>High seasonal drawdowns, sequential restrictions form a part of the plan, good example of triggers and corresponding actions. Victorian Plan</td>
<td></td>
</tr>
<tr>
<td>Management Plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Water Sharing Plan for the Upper</td>
<td>Irrigation, Farming</td>
<td>Water shortages, high demand, over allocation</td>
<td>Example of a highly legislative water sharing plan. High value resource. New South Wales Plan</td>
<td></td>
</tr>
<tr>
<td>Lower Namoi Groundwater Sources</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### References


American Society of Civil Engineers (1987), *Ground Water Management Manual of Practice*, No. 40, 281 pp., N. Y.


### Acknowledgments

The authors acknowledge Australian Research Council Linkage Project LP130100958 and funding partners, Bureau of Meteorology (BoM) and the Department of Environment, Land, Water and Planning (DELWP) for valuable contributions. The authors are indebted to Luk Peeters, Tony Jakeman, and an associate editor for the constructive and insightful comments during the review process that greatly improved this paper. All data used during this analysis are detailed in the paper and listed in the references. For further information on groundwater management policy in Australia, please contact the corresponding author.
Water Resources Research


Department of Water (DOW) (2006), Water Sharing Plan for the Upper and Lower Namoi Groundwater Sources, Gov. of N. S. W., Australia.

Department of Water (DOW) (2011), Water Sharing Plan for the NSW Great Artesian Basin Shallow Groundwater Sources, Gov. of N. S. W., Australia.


Gallagher, M. (2015a), Description of the Campaspe Groundwater Operational Management Modelling, edited by Emma White, Department of Natural Resources and Mines, Queensland.


Murray-Darling Basin Authority (2012), Basin Plan, Commonw. of Aust., Canberra.


Ooi, S. K., and E. Weyer (2008), Control design for an irrigation channel from physical data, *Contral Eng. Pract.*, 16(9), 1132–1150.


And it never failed that during the dry years the people forgot about the rich years, and during the wet years they lost all memory of the dry years. It was always that way.

John Steinbeck, East of Eden

Chapter 4: Do groundwater management plans work?

Modelling the effectiveness of groundwater management scenarios.

This chapter was published as the following journal article:

Do groundwater management plans work? Modelling the effectiveness of groundwater management scenarios

E. K. White 1, J. Costelloe 1 & T. J. Peterson 1 & A. W. Western 1 & E. Carrara 2

Received: 31 October 2018 / Accepted: 15 June 2019
Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract
In contrast to management optimisation methods, which quantify decision variables to create plans, this study does not seek the “best” strategy. Instead, it simulates the sequential decision-making process implicit in environmental management, so that the effectiveness of management scenarios, when implemented as intended, can be evaluated. The purpose was to develop a methodology to quantitatively evaluate the effectiveness of groundwater management plans by simulating sequential management decisions that evolve based on aquifer/management feedback. A groundwater management scheme was structured as a system control loop to capture the aquifer/management feedback, and management decisions were based on realistically sparse observation times and locations. The method indicates how a plan may proceed in reality under alternate timings and frequencies of management decisions and in systems with differing response times. A synthetic example quantified the impact of a generic plan, specifying environmental objectives, extraction restrictions and entitlement limits (maximum volume/year that users are permitted), relative to no-management by combining a numerical model of “reality” with management rules under a stochastic climate. The management decision-making frequency varied from daily to decadal. Generally, effectiveness decreased as the interval between management interventions increased and intervals greater than annual showed minimal improvement compared to entitlement only. The timing of management decisions relative to the irrigation season also impacted plan effectiveness, and when decisions were made prior to the irrigation season, quarterly management was less effective than annual and biannual management. By testing the capacity of plans to achieve objectives, groundwater management can be systematically and objectively improved.

Keywords Groundwater management · Numerical modelling · Water resources conservation · Groundwater protection

Introduction
On the eve of his death, Socrates described the flow of “monstrous, unceasing, subterranean rivers”, which, he believed, originated from the centre of the earth (Plato 2003). Despite the presence of underground water being understood in antiquity, the nature and providence of it was highly uncertain. And today, despite all our scientific advances and technological innovations, there is still great uncertainty in the location, volume and yield of aquifer systems, which regularly confounds groundwater managers. In an 1861 Ohio (USA) court case, management of groundwater was described as “hopelessly uncertain and therefore; practically impossible” (Frazier vs Brown 1861). More than 150 years later, groundwater management is still described as “flying blind” (Currell et al. 2016), “a free for all” (Famiglietti 2014); management plans are described as “neglected” (Foster et al. 2015), and groundwater legislation as the “Cinderella of water laws” (McKay 2006). It is precisely that uncertainty identified as far back as the nineteenth century that is to blame, because it is hard to manage something that you cannot see.

The World Economic Forum has consistently rated water crises in the top five global risks for the past 7 years (World Economic Forum 2018) and in many regions, the value of groundwater as a water resource cannot be overstated.
Therefore, effective management is crucial. Yet, it is only relatively recently that the frightening prevalence of over-extraction and groundwater level declines across the globe (Famiglietti 2014; Giordano 2009; Konikow and Kendy 2005; Wada et al. 2010), are being acknowledged and reflected in legislation, with groundwater management plans becoming increasingly common (California Legislature 2014; Government of Australia 2007). Despite this progress, one of the important questions that is being neglected is: How well do groundwater management plans work?

Plans are rarely quantitatively assessed for effectiveness and, in many cases, are not conducive to quantitative analysis (White et al. 2016). In the past, aquifer state— e.g. hydraulic head levels, well yields, minimum stream or base flow volumes, fluxes, and status of groundwater-dependent ecosystems (GDEs) — has been taken as a proxy of plan effectiveness, despite the potential for plans to have unexpected consequences and the frequent absence of established causality between management action and aquifer state. Stemming from an implicit belief that sequential decision-making management improves environmental outcomes, it is often assumed that management is responsible when an aquifer remains in an acceptable state (White et al. 2016). However, it cannot be concluded that management is effective until the system has been stressed such that plan mechanisms are enacted, and causality of any system change attributed to management action. This research presents a numerical modelling method that evaluates plan effectiveness. To the authors’ knowledge, there are no existing approaches that evaluate plan effectiveness because management modelling tends to search for the “best” management strategy and neglects the effectiveness of the management process itself.

Historically, groundwater management modelling has focused on simulation-optimisation studies (Gorelick and Zheng 2015), and reviews include (Singh 2014b, 2012, 2015; Gorelick 1983; Wagner 1995; Yeh 1992; Loucks and Van Beek 2005; Ahlfeld and Mulligan 2000; Ahlfeld and Heidari 1994). Management optimisations are generally solved for the decision variables (i.e. extraction/injection rates or management strategy) of “least cost” (Wagner 1995) or “maximum benefit” within specified constraints. The optimal solution is found by minimizing (or maximising) a particular objective function, for example MODFLOW GWM (Ahlfeld et al. 2005; Banta and Ahlfeld 2013) has been developed to determine optimal solutions to various groundwater management problems. MODFLOW GWM has been applied to numerous management challenges, including potential baseflow decline due to groundwater extraction (Fienen et al. 2018), optimising recharge and injection rates for managed aquifer recharge (Ebrahim et al. 2015), evaluating model solver error in optimal management solutions (Ahlfeld and Hoque 2008), and maintaining aquifer productivity by optimizing extractions (Banta and Ahlfeld 2013). However, in these studies, the “optimal” management solution was sought instead of the effectiveness of management under a stochastic climate.

A vast range of water resource problems have been informed by optimisation studies, including optimal allocation of water resources (Habibi Davijani et al. 2016, Singh 2014c, 2014a, Bear and Levin 1967), multi-objective water resource management optimisation (Reed et al. 2013), monitoring network design (Reed and Minsker 2004), saline intrusion management (Emch and Yeh 1998; Rejani et al. 2008; Reichard and Johnson 2005; Park and Aral 2004; Werner et al. 2013; Sreekanta and Datta 2011), groundwater policy and management evaluation (Mulligan et al. 2014; Esteban and Dinar 2012; Brown et al. 2015; Bredehoeft et al. 1995), hydro-economic modelling (Peña-Haro et al. 2011), and optimizing management uncertainty (White et al. 2018). Balancing the conflicting uses of groundwater by evaluating trade-offs and navigating political currents are herculean tasks and optimisation modelling has proved an invaluable tool. However, management of water resources is more complex than what models can produce due to the interplay between the physical, social, political, ecological and biophysical systems that govern aquifer systems (Loucks 1992). As a result, certain management problems are not readily optimisable because the optimisation process is likely too restrictive to handle policy evaluation and resource allocation problems involving rules, compromises and hierarchical decision-making (Gorelick and Zheng 2015).

Groundwater management is dynamic, influenced by feedback and sequential decision-making, and is often determined by aquifer state, all of which make optimisation difficult. Many management plans implement extraction restrictions or other management actions when certain groundwater trigger levels are reached (GMW 2011, 2006, 2012). In this way, the aquifer state dictates the management action, yet this natural system/human feedback is rarely captured by traditional optimisation approaches, which aim to identify an “optima” and do not modify pumping rates based on heads modelled during simulations. Furthermore, the outcomes of coupled natural and human systems are unpredictable due to the potential for nonlinear feedback and irregular systems dynamics (Sivapalan et al. 2012; Gordon et al. 2008; Elsawah and Guillaume 2016; Peterson et al. 2012). Sequential decision-making has been shown, due to feedback and unpredictable system dynamics, to produce unexpected outcomes in other environmental fields including, contaminant management of surface waters (Janssen and Carpenter 1999; Carpenter et al. 1999) fisheries management (Anderies et al. 2007), and exploitation of an unspecified shared resource (Lade et al. 2013). Due to potential unforeseen aquifer dynamics, groundwater management plans may not actually be working the way they are thought to be (White et al. 2016). Gorelick and Zheng (2015), call for new types of quantitative policy and planning models that combine simulation methods, decision-making
processes and objectives, and address the complex interactions between human behaviour and aquifer systems.

Agent-based modelling has been applied to groundwater management problems to highlight water-usage trade-offs, feedbacks, and decision-consequences, and to evaluate the dynamics between agent behaviour and aquifer systems (Castilla-Rho 2017; Castilla-Rho et al. 2015). Mulligan et al. (2014) combined groundwater and economic models to evaluate management policy while considering human behaviour in the form of agents (farmers). Guillaume and El Sawah (2014) state that modelling does not necessarily translate into how a plan will operate in reality, and used iterative closed-question modelling to stress-test groundwater management planning. Plan limitations and how preconceptions influence management success were demonstrated (Guillaume and El Sawah 2014). Understanding how beliefs can compromise management success is important, and may help frame realistic expectations, but the study was not concerned with quantifying plan effectiveness.

In contrast to optimisation studies, which quantify decision variables to create plans (Brown et al. 2015), this study does not seek the “best” strategy but instead, aims to test the sequential decision-making management process, the act of management itself. Modelling the effectiveness of the sequential decision-making process of management has, to our knowledge, not been conducted in a groundwater context. A simplified system analysis approach was used to frame groundwater management as a system control problem, which allowed management to change during the simulation in response to water level fluctuations. System control has previously been used in groundwater studies (Jones et al. 1987; Bauser et al. 2010; Ghorbanidehno et al. 2017; Ahn 2000; Tankersley and Graham 1994; Andricevic 1990) but the authors are not aware of any studies that both utilise system control to evaluate sequential decision-making and consider climatic uncertainty with stochastic forward forcing data.

The purpose of this study was to develop and demonstrate a method to quantitatively evaluate the effectiveness of a groundwater management plan by simulating the sequential decision-making process of management with a numerical groundwater model. Additionally, the impact upon plan effectiveness of making management decisions on various timescales and within aquifers with varying response times was assessed. Synthetic management scenarios were used that encapsulated elements from multiple groundwater management plans and simulations were designed to answer the following questions:

1. How effective is the sequential decision-making process that implements water restrictions and entitlement volumes (here defined as maximum volume a user can take per year), at maintaining groundwater levels at a wetland and two domestic wells compared to no management and entitlement only?
2. To what degree does altering the management decision-making period (how often management decisions are made, i.e. the frequency with which groundwater levels are compared to triggers) improve the effectiveness of management?
3. What is the impact upon irrigation supply reliability required to achieve environmental and domestic objectives?

The methodology presented in this paper provides a statistical evaluation of how well a plan achieves objectives when management decisions are made frequently, compared to infrequently and at different times of the year. Comparing scenario effectiveness allows a qualitative assessment of the trade-offs between achieving the plan’s environmental objectives, maintaining irrigation supply and balancing management budgetary constraints under a variable climate. By modelling the act of management, the trade-offs between differing objectives and management actions can be considered in conjunction with stakeholders to set the most appropriate management period, and determine the most important uses and acceptable levels of impacts. For example, achieving 100% of an environmental goal may result in too many days of supply deficit or cause an unacceptable reduction in supply reliability. Manipulating the levels of “unsatisfactory” demand can indicate how often supply will not be met in order to achieve an environmental goal. Stakeholders can not only determine an acceptable proportion of days with restricted supply, but also, the degree of supply restrictions necessary to achieve environmental objectives and explore the consequences of management decisions. Stakeholder involvement in groundwater management promotes cooperation, accountability and a sense of resource ownership that can lead to more equitable and sustainable water management (Barthel et al. 2017), and the importance of stakeholders is being recognised by the policy system (Head 2010). This study constitutes the first step towards evaluating the effectiveness of real plans.

Methodology

A simple example is presented to demonstrate the method by evaluating the capacity of a synthetic management plan to achieve a measurable objective. The act of management was simulated by creating a numerical model of a managed aquifer that represented "reality", and in which, the effect of management decisions could be evaluated. However, management decisions were only based upon groundwater levels at two monitoring wells and not informed by the entire simulation suite. A management plan was formulated for the aquifer and
the impact of plan implementation at various frequencies—daily, monthly, four-monthly, yearly, bi-yearly, five-yearly, decadal—of management decision-making was assessed compared to an unmanaged baseline.

**Modelling the act of management**

Structuring groundwater management as a system control problem, where dynamic systems are modified by feedback in order to maintain particular system states (Astrom and Murray 2008; Åström and Wittenmark 2008) was introduced in White et al. (2016) and allows for aquifer/management feedback and adaptive control of aquifer systems. A control loop was created by programming management rules (plan) in Python (Python Software Foundation 2019) and combining the plan with a MODFLOW-NWT (Niswonger et al. 2011) numerical groundwater model.

Simulation of the act of management requires adaptive change to management action in response to groundwater level fluctuations that occur during a modelling simulation (Fig. 1). This functionality is unavailable with standard MODFLOW without altering the source code; therefore, the process was facilitated using FloPy, a python package that reads and writes MODFLOW files (Bakker et al. 2016). FloPy allowed the comparison of modelled heads to groundwater trigger levels and well depths that were defined in Python, a process that is analogous to groundwater monitoring and management supervision. To prompt immediate management action when triggers were reached, a separate MODFLOW model was run for each daily time step in the climate record. In this way the entire climate record was simulated, day by day, as today’s aquifer state informed tomorrow’s management action. Starting hydraulic heads, pumping rates, recharge and evapotranspiration values were updated daily by replacement of MODFLOW input files. At certain decision-making intervals, when trigger levels were reached, extraction rates were amended during the model simulation. The dewatering of both domestic and irrigation wells was detected by comparing simulated heads to predefined well depths in the Python code. If dewatering threshold depths were reached, the pumping rate for that well was assigned to zero for the subsequent model run. Unlike traditional management models where stresses are predetermined prior to the simulation, this methodology enabled management to change in response to simulated levels in the aquifer.

Management evolving in response to water level fluctuations minimises potential mismatch between management action and aquifer response that may occur with static management. Naturally, if drivers are stationary, an aquifer system reaches a dynamic equilibrium. The time required to reach a new equilibrium after application of a hydraulic stress is termed the aquifer response time (Rousseau-Gueutin et al. 2013; Walton 2011). Consequently, there is a delay between management action and the observed result, which is termed the lag time (Meals et al. 2010). When management timescales are discordant with aquifer response times, unsustainable development and over-exploitation can occur (Gleeson et al. 2012). Robust plans that perform well under a diverse range of climatic conditions are required (Jakeman et al. 2016), so that managers can account for climatic extremes and unpredictability, and understand the impact of climate upon the efficiency of management plans (Gorelick and Zheng 2015; Alley 2016; Alley et al. 2002). In the following example, implementation of a plan in different responding systems at various decision-making frequencies and under various potential future climates is demonstrated.

**Synthetic example**

**Numerical groundwater model**

A synthetic, homogeneous model of an unconfined aquifer was created using MODFLOW-NWT (Niswonger et al. 2011) and designated as the “reality” in which management scenarios were explored. A simple groundwater conceptualisation sufficed because the purpose was not to model a real system under management, but to demonstrate how sequential decision making could be incorporated into a groundwater model. While simple, this system captured required dynamics and if management was shown to be ineffective in simple scenarios, then additional scenario complexity would be superfluous. The example system was designed to mimic a simplified crop irrigation operation situated on the floodplain of an alluvial upland valley. The farm was subject to extraction restrictions to maintain groundwater levels at a wetland and two domestic wells. The dimensions of the aquifer were 2.5 km × 2.5 km, and discretised by a 352 × 229 rectangular finite difference grid with background spacing of 15 m. The grid was incrementally tightened to 2.5-m intervals around each pumping well (Fig. 2) to capture groundwater level fluctuations due to pumping that caused well dewatering. The vertical thickness of the aquifer was 92 m in the north east and reduced to a thickness of 70 m in the south west corner. A no-flow boundary was applied to the horizontal base to represent impermeable bedrock. Natural groundwater flow direction was from uplands in the north east to the wetland region in the south west adjacent to a river boundary comprising the western margin of the model domain. The eastern boundary was defined as general head, and the northern and southern as no flow boundaries.

Groundwater pumping volumes were estimated based on simulated soil moisture content from a vertically integrated one-dimensional (1D) soil moisture model. The pumping occurred in three irrigation zones with dimensions of 350 m × 352 m × 229 m.
Fig. 1 The act of management modelling loop (adapted from White et al. 2016). The inner black loop cycled daily and updated starting heads with heads simulated the previous day. Pumping rates were updated if groundwater levels declined below the depth of the well. When the management cycle occurred (shown in red) groundwater levels in monitoring wells were compared to triggers in the plan and restrictions were implemented for the subsequent management period. In the example, the frequency the management cycle iterated was varied from daily to decadal.

450 m (each containing four pumping wells) in the eastern region. While, realistically, an irrigation operation of this scale is likely to have only one or two wells, it was necessary to spread extraction volumes across multiple wells for each irrigation zone to maintain model stability and focus on the effect of the management plan rather than having extractions limited by aquifer flow capacity. The volume extracted from irrigation wells was reported as a total combined irrigation volume from all wells. Two domestic wells each extracted 2 ML/year in the north-western quadrant, and two monitoring wells, with four specified groundwater level triggers, were located adjacent to the wetland and the domestic wells respectively.

Stochastic climate replicates were used to reflect climatic variability and explore management under various climatic conditions. Rainfall and evapotranspiration data recorded from 1990 to 2016 at the Nhill weather station in Western Victoria were extracted from the Australian Water Availability Project (AWAP) database (Raupach et al. 2008; CSIRO 2016). The stochastic Climate Library (SCL) (Srikanthan et al. 2006) was used to generate 20 stochastic replicates of daily historical precipitation (climate replicates). The climate replicates shared the statistical characteristics of the historical AWAP data and accounted for climatic variability and uncertainty. For this study, the climate is assumed to be statistically stationary, although the method could be extended to include nonstationary climate.

Soil moisture model Groundwater usage varies with climate and irrigation demand fluctuates seasonally. To capture this variability, a simple soil moisture model was developed to determine demand, recharge, and actual groundwater evapotranspiration values based on the climate replicates. The soil moisture model used was a vertically integrated 1D model adapted from Peterson and Western (2014) to include irrigation demand:

$$\frac{dS}{dt} = P - \text{PET} - K_S \frac{S}{S_{\text{cap}}} \left[ D - \left( S - S_{\text{cap}} \right) \frac{S_{\text{cap}}}{S_{\text{cap}}} \right]$$

where $S$ [L] is the soil moisture storage, $t$ [T] is time, $P$ [L/T] is the daily precipitation, PET [L/T] is the daily areal potential evapotranspiration, $S_{\text{cap}}$ [L] is the soil moisture storage capacity (set as 100 mm), $K_S$ [L/T] is the soil vertical saturated hydraulic conductivity, $\alpha$ [\text{--}] is a power term (set to 2) controlling the threshold response of free-drainage to the soil moisture and $D$ [L/T] is the irrigation depth within the irrigation regions and is defined as follows:

$$\delta \left( \frac{S_{\text{cap}} - S - P}{S - S_{\text{cap}}} \right) \leq \left[ 1 - \frac{S}{S_{\text{cap}}} \right]$$

where $t_{\text{start}}$ [T] is the start of the irrigation season (1 November), $t_{\text{end}}$ [T] is the end of the irrigation season (31 March), $\delta$ [\text{--}] is a soil moisture threshold (set at 0.8) defining the average relative soil moisture at which irrigation occurs, $N(0, \sigma)$ is a random normal function with a mean of zero and standard deviation of $\sigma$ [\text{--}] (set to 0.25). The parameters $\delta$ and $\sigma$ produced randomness in the irrigation demand and timing so that the random changes that occur in real-world irrigation operations could be captured.

47
Fig. 2 Synthetic example model domain of 2.5 km × 2.5 km showing three irrigation areas (patched regions), wetland (yellow circle) and grid discretisation. There were four pumping wells (black circles) per irrigation area, two domestic wells (green triangles) and two monitoring wells (red circles), which were used to make management decisions for the entire aquifer. The western boundary was a river (thick blue line on LHS of figure) and the eastern a constant head boundary (thick blue line on RHS of figure). North and south boundaries were no flow. Groundwater flow was from the north east to the south west (thin blue lines).

Irrigation water was only applied to the crop when the soil reached the soil moisture threshold of $\sigma$, which occurred approximately every 10 days, and for the remaining days of the month, irrigation demand was zero. However, due to the daily time-step of the numerical model, daily irrigation demand volumes were required; therefore, the daily pumping demand for each irrigation zone was derived by summing the monthly demand and averaging it over each day in the month. Daily demand volumes were then multiplied by the area of each zone to get a daily zone demand.

To account for energy already consumed by soil water evapotranspiration (see Eq. 1), the potential evapotranspiration from groundwater was simulated as:

$$\text{PE}_{\text{groundwater}} = \frac{1}{4} \times \text{PET} \times \left(1 - \frac{S}{S_{\text{cap}}}ight)$$

Potential evapotranspiration, drainage and demand generated by the soil moisture model provided the “reality” groundwater model with input values for evapotranspiration, recharge and well extraction volumes.

Groundwater management plan

The objective of the management plan, shown on Fig. 3, is maintenance of groundwater levels at the wetland and two domestic wells. Protection of environmental and domestic uses was selected as the objective due to its prevalence in Australian groundwater management planning (GWM 2001, 2009; SRW 2010; NREATS 2009; DOW 2009; SAAL NRM 2009) and the increasing consideration of groundwater dependent ecosystems (GDEs) in management plans (DELWP 2015b, a; DLRM 2016). The plan consisted of entitlement volume limits, groundwater level triggers and extraction restrictions upon three irrigators within the model domain. While more complex management scenarios involving water trading, entitlement carryover and non-compliance could be simulated, they are beyond the scope of this study.

Modelled scenarios

Scenarios included various combinations of management decision-making period, timing and aquifer response times.
Management period The management period varied from daily to decadal to explore the benefit provided by more frequent, yet more resource intensive management, compared to less resource costly management. In addition to the seven management period scenarios, two baseline scenarios of the aquifer were simulated: entitlement-only and unmanaged. Entitlement only was designed to evaluate the impact on effectiveness of setting an entitlement volume compared to no management. In this scenario, extractions from the aquifer were subject to entitlement volumes but were not modified by restrictions and, pumping only ceased when the wells became dry. Due to the daily time-step of the model, daily maximum extraction volumes were required as model input; therefore, entitlement volumes were set at typical daily usage volumes and imposed as daily maximums. In the entitlement-only scenario and all the managed scenarios, demand from pumping wells was compared to the maximum daily entitlement volume, which if exceeded, resulted in amendment of pumping rate down to entitlement volume. The unmanaged scenario had neither entitlement volume nor extraction restrictions and demand was only constrained by the area under crop cultivation. In this way, the aquifer was a common pool resource as defined by Hardin (1968), and supply was prioritised above all else. In the managed scenarios, groundwater levels were periodically compared to trigger levels defined in the plan at the two monitoring wells. If trigger levels were reached in either monitoring well, restrictions were applied to the pumping wells for the subsequent management period. For example, if the management period was monthly, entitlement volumes were reduced for the entire subsequent month. Domestic wells were not subject to water restrictions. Sustainable groundwater management is infinitely more complex than the simple scenario modelled here and depends upon legal, social, environmental, political and economic factors. It

---

<table>
<thead>
<tr>
<th>Zone</th>
<th>Irrigation Area 1</th>
<th>Irrigation Area 3</th>
<th>Irrigation Area 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entitlement (m^3/day)</td>
<td>896</td>
<td>918</td>
<td>924</td>
</tr>
</tbody>
</table>

There are two domestic wells utilising the aquifer (DS1, DS2) with an estimated usage of 5.5 m^3/day.

**Monitoring Wells**

There are two monitoring wells in the aquifer. Monitoring well one was situated to monitor groundwater levels near the domestic wells to ensure levels do not decline below dry depths. Monitoring well two was situated to monitor groundwater levels near sensitive GDE wetland in the south west of the management area. The purpose of monitoring well two was to ensure that groundwater levels do not decline below the critical threshold at the GDE.

**Method of Control – Extraction Restrictions**

Threshold groundwater levels in the two monitoring wells were established as trigger levels for restrictions upon pumping well extractions. If trigger levels at monitoring well one or monitoring well two are reached, restrictions will be implemented.

<table>
<thead>
<tr>
<th>Trigger</th>
<th>Pumping Rate Cut %</th>
<th>Head (m above AHD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Monitoring well one</td>
</tr>
<tr>
<td>Trigger 1</td>
<td>25</td>
<td>72.20</td>
</tr>
<tr>
<td>Trigger 2</td>
<td>50</td>
<td>71.50</td>
</tr>
<tr>
<td>Trigger 3</td>
<td>75</td>
<td>70.70</td>
</tr>
<tr>
<td>Trigger 4</td>
<td>100</td>
<td>70.00</td>
</tr>
</tbody>
</table>

**Data Review and Analysis**

Groundwater level measurements at the monitoring wells will be reviewed and compared to trigger levels on a frequency varying from daily to decadal (management period).

**Driver Monitoring**

Groundwater extractions are monitored at each pumping well. Climate is monitored at a nearby weather station.

**Success Measures**

The success measure was the maintenance of groundwater levels at the wetland and domestic wells

- Domestic wells
  - If groundwater levels in the two domestic wells decline below the depth of the well (each well is 7.5 m deep), water users are unable to access water and the plan is considered to have failed.
- Wetland
  - If groundwater levels at the centre of the wetland decline below dry depth, the plan is considered to have failed.
was assumed irrigators comply with restrictions for the purposes of this study, so that the impact of plans, if implemented as intended, could be evaluated; however, the authors acknowledge non-compliance is a great challenge for managers. Restrictions were chosen because during the Australian Millennium Drought (1999-2009), restrictions were imposed in certain management areas and several plans outlined sequential restrictions based on usage volumes or groundwater recovery levels. Evaluating the impact sequential decision had upon objective failure rate was a study objective.

Management timing The effect of making management decisions at different times of the year was explored by replicating all simulations at two different decision-making timings during the (southern hemisphere) irrigation season: mid-season (January) and early season (November). For the mid-season timing, annual, bi-annual, five- and ten-yearly management decisions were made on January 1st of each year, and four-monthly (120 day) management decisions were made on January 1st, May 1st and September 1st of each year. For the early-season (November) timing, annual and greater decisions were made on November 1st, and four-monthly decisions were made on November 1st, March 1st, and July 1st of each year. For both decision timings, daily management decisions were made each day, and monthly decisions were made on the first of each month.

Aquifer response time To determine the impact of aquifer response time upon plan effectiveness, hydraulic conductivity \( K_s \) was varied from 0.026 to 1.43 m/day in six steps, resulting in six aquifers that responded to hydraulic disturbance on different timescales. Specific yield remained constant at 0.2 for each aquifer hydraulic conductivity/response time. All scenarios were replicated in the six aquifers and the plan effectiveness determined. To assess the interaction between aquifer response time, management period, and plan effectiveness, all scenario combinations were repeated for each of the 20 stochastic climate replicates to evaluate management under climatic variability.

Evaluation framework

In this example, plan effectiveness was defined by the maintenance of minimum groundwater levels at a wetland and two domestic wells and was directly evaluated by calculating the objective failure rate—how often the wetland levels declined below minimum depths and domestic wells went dry—during the simulation period. Additionally, the plan was indirectly assessed by evaluating the impact of restrictions upon irrigation supply reliability to determine any trade-offs required to achieve the environmental and domestic objective.

Objective failure rate (frequency with which wetland and domestic wells go dry) When levels declined below the threshold dry depths at the wetland and domestic wells, the plan failed. Each day groundwater levels at the wetland and domestic wells were compared to objectives and the number of failures across all climate replicates was calculated for each scenario. The objective failure frequency \( F_{freq} \) of each replicate was calculated using Eq. (4):

\[
F_{freq,i} = \frac{O_{fails,i}}{N_{days}} \delta \frac{\%}{\text{Replicate}}
\]

where \( F_{freq,i} \) is the plan objective failure frequency for the \( i \)th replicate; \( O_{fails,i} \) is total number of days in which objective was not met for the \( i \)th replicate; and \( N_{days} \) is total number of time intervals in simulation.

Equation (4) yielded 20 failure frequency values for each management period scenario, which were averaged to produce a mean failure rate for each management period (Eq. 5). This yielded one value of objective failure frequency at each location for each management period, which allowed comparison between management periods.

\[
F_{freq} = \frac{\sum F_{freq,i}}{N_{replicates}} \delta \frac{\%}{\text{Replicate}}
\]

where \( F_{freq} \) is the average failure frequency for that management period and \( N_{replicates} \) is the number of replicates (20).

Irrigation supply reliability Reliability is the probability that demand can be met by the aquifer system and is defined in terms of either time (probability that demand will be met within a particular time period) or volume—total supplied volume divided by the total demanded volume (McMahon and Adeloye 2005). While irrigation supply was not an explicit objective of the plan, temporal and volumetric reliability were determined to illustrate the potential irrigation supply trade-off required to achieve the environmental/domestic objective. Reliability of irrigation wells was calculated during the irrigation season when pumping occurred.

Temporal reliability was calculated with Eq. 6 at a daily scale using a total extraction volume from all irrigation wells.

\[
R_t = \left( \frac{N_i}{N} \right) \delta \frac{\%}{\text{Replicate}}
\]

where \( R_t \) is the temporal reliability; \( N_i \) is total number of time intervals in which demand was met; and \( N \) is total number of time intervals in the simulation (McMahon and Adeloye 2005). Temporal reliability does not account for the length or severity of the shortage and treats a few long severe shortages the same as many short mild ones (McMahon and Adeloye 2005). Considering that a severe shortfall can result in crop failure and wetland desiccation, shortage severity is an
important consideration for managers. Volumetric reliability 
(Rv) provides an indication of the magnitude of the shortfall 
and was calculated on a daily time-step using Eq. (7): 
\[
R_v = 1 - \frac{\sum (D_j - D'_j)}{\sum D_j}
\]

Where \( D'_j \) is the supply from the aquifer during the \( j^{th} \) 
failure period; \( D_j \) is the demand during the \( j^{th} \) period; \( N \) is 
the number of periods in the simulation.

Results

While scenarios were modelled in aquifers with six different 
response times (\( K_s \) values), three of the six aquifers did not 
show a difference between objective failure rate at daily man-
agement and the entitlement-only scenario, i.e. management 
was indistinguishable from unmanaged. Daily management 
(extraction restrictions + entitlements) and the entitlement only 
(no restrictions) failure rates area shown in Fig. 4 and man-
age only differs to the entitlement only scenario within
the grey shaded region. Figure 4 shows the failure rate for each $K_s$ value at the wetland (Fig. 4a,b); domestic wells (Fig. 4c,d); and the temporal irrigation supply reliability (Fig. 4e,f) for both daily and entitlement-only management. Management had the greatest impact on aquifers with $K_s$ values between 0.143 and 0.56 m/day (grey-shaded region) for each figure pair. In aquifers with a $K_s$ value outside of this range, the impact of management on failure rate and reliability is negligible, because fast responding systems accessed water from greater distances and were better able to meet demand than slower responding systems that primarily met demand from storage. In these systems, daily management provided no greater benefit over simply setting an entitlement volume.

This is further illustrated with average daily water balances for the unmanaged scenarios of the slow, medium and fast aquifers (with $K_s$ values of 0.026, 0.26 and 1.43 m/day respectively; Fig. 5) that illustrate the discrepancy in water availability between aquifer response times. Figure 5 shows that systems responding on different scales have very different water availability and consequently, exhibit very different outcomes due to the plan.

The decrease in irrigation supply reliability due to daily management compared to setting an entitlement volume can be seen by comparing Fig. 4e,f, and illustrates the economic toll required to achieve environmental objectives. Management only influenced failure rate in aquifers where demand and aquifer capacity to supply were relatively balanced, which is, realistically, the type of aquifer in which a small-scale irrigation operation of this type would be installed and managed under a plan of this nature. Irrigation at an inappropriate scale would not be pursued in an aquifer that consistently failed to meet demand, nor would extraction restrictions be required when demanded volume was easily supplied. Consequently, exploring management impacts for a small irrigation operation upon nearby receptors is most appropriate in the medium aquifer ($K_s = 0.26$ m/day), where not only is management required, but has a measurable impact upon the domestic and environmental outcomes. The observed sensitivity of plan effectiveness to response time is expected and reinforces that management plans must be tailored to specific systems and that blanket application of management is unwise and most likely ineffective or unnecessary. The remainder of the results section will focus on the impact of various management periods upon the plan effectiveness and well reliability in the medium aquifer. Results of the seven management decision-making periods are presented for decision timings synchronised to January (mid-irrigation season) and November (early-irrigation season) timings.

![Fig. 5 Selected average daily water balance in the entitlement-only baseline aquifer scenario for $K_s = 0.026$, 0.26 and 1.43 m/day. While recharge remains constant across aquifer types, the slow aquifer has the lowest inflow from constant head (red). This is because the drawdown cone caused by pumping extractions (wells_out) in the slow aquifer does not propagate as far to access distant water. As a result, there is no additional source of water available to the slow aquifer and any extractions (red) must be predominantly supplied by storage (black). Consequently, the slow aquifer has the lowest well extraction volume (purple) and the aquifer is unable to meet the demand. The medium and fast aquifers have higher transmissivity and can access water from the constant head boundary (red). As a result, the volume of outflow due to pumping wells (purple) is greater, reflecting the ability to meet supply](image-url)
Plan objective failure rate

The failure rate at the wetland and domestic wells for mid-season management decision timing is shown on Fig. 6a. The number of days in which supplied volume was less than demanded volume (days of limited supply) is shown in green on the secondary axis and indicates periods when restrictions are in place, entitlement volumes are in place or the well is dry. Generally, the failure rate increased as the period between management decisions increased and the unmanaged scenario had the greatest occurrence of plan failures (Fig. 6a). The fewest instances of objective failure at the wetland (5.1%), domestic well one (3.6%) and domestic well two (0%) occurred under daily management (Fig. 6a). Increasing the

Fig. 6 Percentage of days when plan objectives failed. Wetland (blue), Domestic well one (red), Domestic well two (yellow). The number of days when supply was less than demand is shown in green on the secondary axis. a January decision (mid-irrigation season): Daily management had the lowest plan failure rate and entitlement-only (in grey-shaded region) had the second highest failure rate but the fewest days when supply was less than demand. Unmanaged had the highest failure rate and was variable. Generally, failure frequency increased as the length of management period increased, except for 120-day management, which showed slightly more plan failures than annual management due to an interaction between timing of the irrigation season and the management decision. b December decision (prior to irrigation season): Shifting the timing of the decision from mid-season (January) to early-season (November) decreased the failure rate of the 120-day management period and resulted in a trend of increasing failure rate. The management period trend plateaued at annual management period. All management periods of annual or greater had the same failure rate as the entitlement-only scenario. Unmanaged was highly variable and reported the greatest failure rate. Wetland (blue); domestic well one (red); domestic well two (yellow)
management period from daily to monthly resulted in less than a 0.5% increase in failure rate at each monitoring location, and a 2.5% decrease in the number of days of limited supply.

An increase in the management period from daily to annual, increased the wetland and domestic well one failure rate by 2.8 and 3.6% respectively. The average number of days with restricted supply also increased by 2.2% when managed annually compared to daily (Fig. 6a). The exception to the increasing failure with increasing decision-making period trend was the 120-day management period, which reported a small increase of failure rate compared with annual management. This was due to an interaction between the management decision-making timing and the irrigation season and is further discussed in section ‘Management timing’. A Welch t-test found a statistical difference between the failure rates of each management period at the three locations. Statistical significance analyses are provided for the wetland, domestic well one and domestic well two in Appendix 1, Appendix 2 and Appendix 3 respectively. While there are differences between management periods, the failure rate results could be divided into four groups—daily/monthly, management periods of 120 days and greater, entitlement-only, and unmanaged—between which the statistical difference is large (Fig. 6a; Appendix 1, Appendix 2, and Appendix 3). This means that the biggest differences in failure rate occur between these four groups. The trade-off between environmental, domestic and commercial uses that is required when managing groundwater is clearly illustrated on Fig. 6a. For management periods of 120 days and greater, imposing restrictions to attain plan objectives increases the days when supply is less than the demanded volume. The managed scenarios all show a greater proportion of days where demand was not met compared to the entitlement-only scenarios due to imposition of restrictions (Fig. 6a).

The failure frequency at the wetland and domestic wells for early-season timing is shown on Fig. 6b. Daily and monthly management show the lowest failure rate. Managed scenarios of annual or greater, had the same failure rate as the entitlement-only scenario, showing that in this case, management provided no benefit for the early-season timing. A t-test found no statistical difference between objective failure rate or irrigation supply reliability between entitlement-only and management periods of annual and greater (Appendix 1, Appendix 2, Appendix 3 and Appendix 4) indicating that restrictions were not an effective management action for this combination of management timing and period. Regardless of management timing, entitlement volumes improved environmental and domestic outcomes compared to the unmanaged scenario.
Irrigation supply reliability

Temporal and volumetric reliabilities for mid-season (January) timing are shown on Fig. 7. More frequent management increased reliability compared to less frequent management because extractions were restricted for a shorter period. The entitlement-only scenario provided the highest temporal reliability with a median value of 0.57. Temporal reliability in the unmanaged scenario was highly variable and significantly lower than the entitlement-only scenario (Fig. 7). Interestingly, a t-test showed that irrigation supply reliability of all management frequencies was significantly lower than the entitlement-only scenario, and that there were statistical differences between the management periods (p values are shown in Appendix 4). There was a statistically significant difference between the reliabilities of daily, monthly and 120-day periods and the management periods of annual and greater (Appendix 4). Additionally, there was a large difference between entitlement-only and all managed scenarios. Mean reliability values for daily, monthly and 120-day management frequencies ranged from 0.38 to 0.41 and were statistically indistinguishable between monthly and 120-day management. The longer management frequencies, annual to decadal, all had similar reliability values of 0.36 to 0.34 that were lower than the shorter management period values (Fig. 7a). Selection of an appropriate management period is therefore an important management consideration, as discussed in section ‘Management period’.

The average number of extreme shortfalls—arbitrarily defined as days where supply is less than 20% of demand and shown as a blue line on reliability plots (Figs. 7 and 8)—for each scenario was determined by summing the number of days of extreme shortfalls for each of the twenty replicates per period. Extreme shortfalls occurred on less than 0.6% of days across all scenarios and were lowest for daily and monthly management frequencies. This indicates that while there were more days where full demand was not met (due to implementation of restrictions); the occurrence of extreme shortfalls was curtailed under daily and monthly management as compared to the unmanaged scenario. The largest number of extreme shortfalls occurred under decadal management and the 120-day management period reported a similar number of extreme shortfalls to the entitlement-only scenario (Fig. 7a). Results were similar for volumetric reliability, which provide an indication of

Fig. 8 a Temporal and b volumetric reliability of irrigation supply wells early-season in the medium aquifer - November decision. The blue line shows the mean percentage of days of extreme shortfalls (supply less than 20% of demand) which is lowest for a daily management period. Supply reliability for annual and greater management is the same as the entitlement-only scenario (grey-shaded region). Reliability is lowest daily to 120-day management when restrictions are enacted. Reliability under the unmanaged scenario is highly variable.
the severity of the shortfall. Volumetric reliability was greatest for entitlement only and unmanaged, though the unmanaged scenario showed a high degree variability with wells supplying all or nothing. Of the managed scenarios, daily management had the highest volumetric reliability followed by monthly (Fig. 7b). Longer management scena-
narios (5 years and decadal) showed higher variability in volumetric than temporal reliability. Temporal and vol-
umetric reliability for early-season (November) timing is shown on Fig. 8 and management at greater frequency than four monthly (120 days) has no benefit compared to enti-
tlement only.

Discussion

A simple synthetic model was used to develop and demon-
strate a methodology to test the effectiveness of sequential
decision making in groundwater management. This allowed
for an assessment of the effectiveness of a management plan
within a system control framework. Model simplicity was
chosen over site-specific complexity to keep the results
generalisable and avoid concentrating on case-specific idio-
syncrasies. However, the method is easily extended to com-
plex groundwater systems and management actions, provided
they can be modelled adequately. The example demonstrates a
methodology for using numerical groundwater models to
evaluate groundwater management plans. In a realistic con-
text, this method could be used to test current or proposed
plans in management areas where numerical models exist.
The method could be part of a stakeholder engagement pro-
cess where the simulation results are shared with the commu-
nity to explain the impact and consequences of various man-
agement actions. The open source platform of both
MODFLOW and Python makes this method freely available
and customizable.

Aquifer response time

As expected, plan effectiveness was highly sensitive to
hydraulic conductivity and in this example, management
was only effective within aquifers with \( K_s \) values of be-
tween 0.143-0.56 m/day. This band of \( K_s \) values in which
a management signal was discernible corresponded to a
situation where demand was relatively similar to the aqui-
fer’s ability to provide water. While this might seem like a
restricted set of conditions, it is where groundwater re-
source exploitation is maximised within sustainability con-
straints. The management plan provided no benefit com-
pared to entitlement-only when outside of this range be-
cause water availability was either physically constrained
(low \( K_s \)) or plentiful (high \( K_s \)), resulting in uniform failure
or uniform success respectively (Fig. 4). This simple
example illustrates the importance of incorporating aquifer
response time into management design and demonstrates
the folly of a “one size fits all” management approach.
Figures 4 and 5 show the same plan had very different
outcomes in systems responding on differing scales. The
slow responding aquifer predominantly accessed water
from storage and was unable to meet extraction demand,
resulting in overexploitation and almost uniform plan fail-
ure (Figs. 4 and 5). If management action occurs too rap-
idly in a slowly responding aquifer, then the impact of the
action may have long-reaching implications that will not
become apparent for some time. Walton (2011), states that
when management planning horizons are shorter than aqui-er response times, the impact of extractions may be
underestimated. Long-term climatic and geological chang-
es must be accounted for in systems responding on long
timeframes (Alley et al. 2002), which considering the rate
and potential impacts of climate change, further compli-
cates management.

In contrast to the slow system, the drawdown cone prop-
agated rapidly through the fast aquifer and accessed water
from the constant head boundary condition, easily main-
taining extraction demand from wells and negating the
need for a plan (Figs. 4 and 5). In this simple scenario,
the plan was successful because water from further away
could be accessed and the localised receptors (wetland and
domestic wells) that were plan objectives were not adver-
sely impacted. However, in more complex scenarios, if
management action occurs too slowly then it may fail
to prevent adverse impacts, for example, drawdown cones
due to extractions propagating through the aquifer and des-
iccating a wetland. Furthermore, receptors to extractions
change with the aquifer response time. It can be seen in
Fig. 5 that the constant head boundary provides most of the
infloows for the fast aquifer, which is balanced by an in-
crease in river and well extractions, compared to the slow
aquifer which pulls from storage. The drawdown cone of a
fast aquifer will be of lower amplitude but further reaching
and so the potential receptors will be different because they
will be further away. The difference between receptors de-
pending upon response times should also be considered in
management design.

Management period

More frequent management resulted in fewer plan failures
and increased temporal reliability compared to less fre-
quent management. Management periods greater than an-
ual showed minimal improvement compared to entitle-
ment only scenarios, while reducing temporal and volu-
metric reliability. The frequency that management deci-
sions were made directly impacted the effectiveness of
the plan. This is an important point because the cost of

56
Management action must be balanced by the benefit provided. Daily management is expensive and thus, very uncommon. This study showed that increasing the investment in management, in this scenario, decreased the plan failure rate; however, the increase in plan failure rate between daily and monthly management periods was relatively small and may be acceptable to managers due to the lower resource expenditure required for monthly compared to daily management. Additionally, as demonstrated by this study, the objective failure rate increases with increasing length of management period and the best environmental/domestic outcomes are often achieved at the detriment of supply. These are considerations and trade-offs that managers face and that can be informed by modelling the act of management.

Management timing

The time of year management decisions were made directly impacted plan effectiveness. For example, when management decisions were made in January (mid-way through the irrigation season), the 120-day management period was the exception to the increasing failure with length of management period trend and was consistently less effective than annual and bi-annual management. This was due to an interaction between the timing of the management decisions (January, May and September) and the irrigation season (November-March). The 120-day management cycle compared groundwater levels to triggers in January, May and September and, because May and September were during the nonirrigation season when levels were high, restrictions were not enacted. By the time the next decision was made in January, 2 months after irrigation began, water levels had declined and, restrictions were implemented for the remainder of the irrigation season. However, for the first part of the season (November and December), pumping was unrestricted, and the aquifer was essentially unmanaged; whereas, for annual and bi-annual management frequencies, decisions were also made in January, but the restrictions persisted for 1 or 2 years until the next decision was made.

When levels in monitoring wells were compared to triggers in November, heads were above triggers and no restrictions were implemented. If management decisions were made often (daily or monthly), this was not a concern because the water level declines during the irrigation season were detected and pumping rates were reduced. However, if the management period was annual or greater, decisions were only ever made during the nonirrigation season, and the allocations that were provided when levels were high remained in place until the following November. This shows that when decisions are made prior to the irrigation season, annual, and all subsequent, management periods exhibited the same plan failure rate as the entitlement-only scenario (Fig. 7). Often, allocation volumes are announced prior to the commencement of irrigation seasons to provide irrigations with certainty so they can decide upon cropping types and schedules. However, this analysis suggests that when decisions are made prior to the irrigation season, and the management period is annual or greater, extraction restrictions provide no benefit for the system considered. Nonetheless, this is highly dependent upon aquifer response time and had the systems responded on a decadal or greater timeframe, the impact of management would be quite different. The difference between the entitlement-only scenario and the unmanaged scenarios shows that, in this case, entitlement volume limits increased the effectiveness of the plan. This is important because applying an entitlement volume in a management region is relatively inexpensive compared to other management intervention and these results indicate entitlement volumes improve management outcomes compared to no management. The results of the study illustrate the unpredictability of aquifer system management and underscores the need for cautious management.

Usage of methodology

This method is intended as a management tool that can be used in various ways:

1. To evaluate how effective a given plan is at achieving stated objectives. Scenario modelling can assist in guiding the development of plans and help inform managers of appropriate objectives, triggers and management techniques. Additionally, the method can identify the most effective management period, timing, inappropriate or unachievable objectives, pumping demands or unacceptable impacts and highlight the consequences of management decisions. For example, if the method shows objectives cannot be achieved under any decision-making period, that indicates either the objectives or the mechanisms of the plan may be unsuitable for that particular aquifer resulting in plan redevelopment. The high failure rate in slower responding systems (Figs. 4 and 5) indicates that plans based on inflated entitlement volumes, where the aquifers’ ability to supply is unbalanced with the (potential) demanded volume, may not work due to this disparity, depending on how entitlements are utilised during periods of stress. If an aquifer does not have the capacity to meet unrealistic entitlement volumes, then the plan will be ineffectual regardless of management decision frequency.

2. Trigger levels can be adjusted and simulated to determine the potential impact of making decisions under various degrees of risk. For example, if stakeholders thought
occasional wetland desiccation was acceptable, the trigger levels may be set lower. Quantification of the impact on effectiveness helps with risk-based management.

3. As part of a cost-benefit analysis where managers can test various decision-making frequencies to decide whether any reduction in failure rate is worth the increased cost of more frequent management. The cost benefit analysis can also be used to balance environmental protection with supply reliability—for example, managers and stakeholders can evaluate if a particular proportion of days with reduced supply is worth it to achieve the environmental objective.

4. Stakeholders are more likely to accept management actions when they are part of the process and understand the consequences of management action and are equipped to evaluate trade-offs. Effective communication between lawmakers, managers and stakeholders is very important (Nelson 2013) and this method could facilitate understanding and collaboration.

Alternate management scenarios

While this study demonstrated the methodology with a simple system, and predictions of effectiveness depended on model geometry and hydrogeological settings, the method could and should be applied with a realistic site-specific model to inform specific problems. For example, alternative scenarios that can be used with this method include various well numbers, locations and extraction volumes. In the demonstration case, results are sensitive to hydraulic conductivity and model geometry influences the prediction of effectiveness. Changes in aquifer dimensions, transmissivity, boundary types and locations alter the aquifer response time (Walton 2011), upon which effectiveness depends. Increased extractions are likely to increase plan failures due to larger extraction related drawdown cones. Moving wells closer to a boundary such as the river could increase reliability because recharge from the river would be accessed. While this scenario would likely not increase plan failures, other adverse impacts that are not plan objectives may occur, such as a reduced stream flow.

Various combinations of well numbers and location and extraction volume could be explored with the method to aid management design and explore potential impacts to effectiveness and reliability. The method is not limited to water level objectives, various SMART—specific, measurable, achievable, realistic, timely—objectives can be evaluated with the method including water quality thresholds, and extraction volume requirements. Any measurable objective that can be adequately modelled can be assessed—for example, the management decision-making period timing was selected to be uniform in order to compare the effectiveness of various periods. However, realistically, decision periods often fluctuate and inserting randomness into the decision-making period would allow an evaluation of the impact of varying decision timing. Water trading and Entitlement carry overs could easily be incorporated into the methodology and various percentages of entitlements and lengths of carryovers could be explored.

Limitations

Groundwater models are uncertain, labour-intensive, and are often prohibitively expensive for groundwater managers who must operate on finite budgets, which limits the applicability of this method and others. Also, the variation in failure rates between aquifer response time (Fig. 4) indicates the method is sensitive to hydraulic conductivity. Therefore, if the model of the groundwater system is highly uncertain, the capacity to evaluate plan effectiveness will be compromised. Management may then be deemed either unnecessary or ineffectual due to erroneous assumptions about the groundwater response during modelling. Models with sparse calibration datasets and hence a wide range of plausible parameters may be of little value for testing management effectiveness. In the example, a numerical model that was perfectly known sufficed as reality and was used to evaluate a plan. However, application of the methodology to a real system subject to variable aquifer properties faces additional challenges due to parameter/predictive uncertainty. Consideration of both conceptual model and model parameter uncertainty that accounts for incomplete knowledge of the subsurface (Bredehoeft 2005) would be required.

Uncertainty in groundwater management modelling is a well-established challenge (Guillaume et al. 2016). The reality of a natural system is always unknown and the feasibility of assessing management with a model that is incapable of perfectly replicating the system is yet to be determined. The level of model fidelity required to adequately represent system reality so that a plan can be assessed is difficult to ascertain and is the subject of ongoing research. In the example, two types of uncertainty were considered, firstly climatic uncertainty was assessed using stochastic climate replicates and secondly, parameter uncertainty was considered in a simplified and heuristic manner with the six different $K_s$ values to demonstrate the impact upon plan effectiveness of parameter variations. It is recognised this is a very basic representation of parameter uncertainty. While rigorous uncertainty studies have been conducted (Sreeranth et al. 2016; Doherty et al. 2010; Gallagher and Doherty 2007), the purpose here was simply to highlight the impact of parameter change upon plan effectiveness, not to realistically represent
uncertainty in a synthetic example. An evaluation of the impact of parameter uncertainty on the ability to evaluate management plans is a future research direction currently being pursued.

This simple assessment demonstrates the importance of monitoring data in the groundwater management planning process because the effectiveness of the plan varied widely depending upon hydraulic conductivity values used. These observations suggest that a high degree of system understanding is required to evaluate the effectiveness of management plans. Lacking such data, experimentation upon plan effectiveness may not be possible and resources could be better utilised increasing system understanding.

**Conclusion**

Currently many groundwater management modelling studies focus on finding the optimal solution of management problems and neglect the management process itself. Groundwater management plans, the primary means of managing groundwater, are not systematically and quantitatively evaluated for effectiveness and currently it is unknown how well they work. This paper develops and demonstrates a method to quantitatively assess the effectiveness of sequential decision-making inherent to groundwater management. Groundwater management was structured as an engineering control loop to capture the aquifer/management feedback and determine if the successive decision-making process improves outcomes compared to no management. The methodology was demonstrated using a simple numerical model constituting “reality” where the impact of implementation of a management plan (consisting of environmental objective, extraction restrictions and entitlement limits) could be assessed at seven different management decision-making frequencies (daily, monthly, four-monthly, yearly, bi-yearly, five-yearly, decadal).

In the synthetic case study, management effectiveness was found to be highly sensitive to aquifer response time and a small change in hydraulic conductivity had a large impact on plan success rate. Of the six response times simulated, management was effective in only three (within the range of $K$, values 0.143–0.56 m/day), where demand and capacity were relatively balanced. Outside of this range, management had a negligible impact compared to no management. The observed sensitivity to response time is expected and reinforces that management plans must be tailored to specific systems and that blanket application of management is unwise and most likely ineffective or unnecessary.

In the example case, management improved domestic and environmental outcomes at all decision-making periods compared to the unmanaged scenario. More frequent management resulted in fewer plan failures and increased reliability compared to less frequent management. Management periods greater than annual showed minimal improvement compared to entitlement only, while reducing reliability. Generally, as the length of management period increased, the plan effectiveness decreased. The best environmental and domestic outcomes occurred at daily and monthly management periods when the decision-making period was shorter but resulted in a decrease in temporal irrigation supply reliability. The need for trade-offs was demonstrated because as environmental outcomes increased, supply reliability decreased. If management decisions were made prior to the irrigation season, management was less effective than if decisions were made in the middle of the irrigation season. Annual or greater management period was no more effective than entitlement only when decisions were made prior to the irrigation season. However, the results of the example depend upon system response time and may vary for a large system with a long response time.

The sensitivity of the method to aquifer response time suggest that high degree of system understanding is required to evaluate the effectiveness of management plans.

The flexible methodology allows for more complex management involving trading and economic implications or analysis of multiple conceptual models. An understanding of the compromises required to achieve objectives, can assist in the evaluation of priorities and inform the community consultation process of plan development. The systematic evaluation of management plan effectiveness can lead to improvements in the planning process. The main conclusions from this study are:

- Plan effectiveness can be evaluated by simulating the sequential decision-making process of management and the timing and frequency decisions are made impacts plan effectiveness.
- Implementation of restrictions at an inappropriate timescale provides no greater improvement to plan effectiveness than setting an entitlement limit.
- When management decisions are made on annual or longer periods, as frequently occurs in practice, plan effectiveness can be highly uncertain. Additional factors such as noncompliance, which was not considered in this study, could further decrease effectiveness in the considered scenarios.

This study constitutes a step towards evaluating the effectiveness of established groundwater management plans.

**Acknowledgements** Climate data used in this study can be found on the AWAP (Australian Water Availability Project) database.

**Funding information** The authors acknowledge Australian Research Council Linkage Project LP130100958 and funding partners, Bureau of Meteorology (BoM) and the Department of Environment, Land, Water and Planning (DELWP) for valuable contributions.
### Appendix 1

Table 1  Welch’s statistical significance t-test for January decision timing: wetland fail data at daily (1), monthly (30), four-monthly (120), yearly (365), two-yearly (730), five-yearly (1,825) and ten-yearly (3,650), entitlement-only and unmanaged

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>30</th>
<th>120</th>
<th>365</th>
<th>730</th>
<th>1,825</th>
<th>3,650</th>
<th>Entitlement Only</th>
<th>Unmanaged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
</tr>
<tr>
<td>1</td>
<td>0.0000</td>
<td>1.0000</td>
<td>1.9675</td>
<td>0.0576</td>
<td>11.1706</td>
<td>0.0000</td>
<td>9.1878</td>
<td>0.0000</td>
<td>10.0371</td>
</tr>
<tr>
<td>30</td>
<td>-1.9675</td>
<td>0.0576</td>
<td>0.0000</td>
<td>1.0000</td>
<td>12.0179</td>
<td>0.0000</td>
<td>9.2665</td>
<td>0.0000</td>
<td>10.3732</td>
</tr>
<tr>
<td>120</td>
<td>-11.1706</td>
<td>0.0000</td>
<td>-12.0179</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>-2.0628</td>
<td>0.0461</td>
<td>-1.0103</td>
</tr>
<tr>
<td>365</td>
<td>-9.1878</td>
<td>0.0000</td>
<td>-9.2665</td>
<td>0.0000</td>
<td>2.0628</td>
<td>0.0461</td>
<td>0.0000</td>
<td>1.0000</td>
<td>1.0048</td>
</tr>
<tr>
<td>730</td>
<td>-10.0371</td>
<td>0.0000</td>
<td>-10.3732</td>
<td>0.0000</td>
<td>1.0103</td>
<td>0.3188</td>
<td>-1.0048</td>
<td>0.3213</td>
<td>0.0000</td>
</tr>
<tr>
<td>1,825</td>
<td>-9.7986</td>
<td>0.0000</td>
<td>-9.6697</td>
<td>0.0000</td>
<td>-0.4596</td>
<td>0.6489</td>
<td>-2.0407</td>
<td>0.0490</td>
<td>-1.2239</td>
</tr>
<tr>
<td>3,650</td>
<td>-9.6755</td>
<td>0.0000</td>
<td>-9.3874</td>
<td>0.0000</td>
<td>-1.1577</td>
<td>0.2564</td>
<td>-2.5586</td>
<td>0.0156</td>
<td>-1.8246</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-10.3732</td>
<td>0.0000</td>
<td>1.0103</td>
<td>0.3188</td>
<td>-1.0048</td>
<td>0.3213</td>
<td>0.0000</td>
</tr>
<tr>
<td>Entitlement Only</td>
<td>-22.5117</td>
<td>0.0000</td>
<td>-26.6924</td>
<td>0.0000</td>
<td>-15.6019</td>
<td>0.0000</td>
<td>-16.5850</td>
<td>0.0000</td>
<td>-15.6811</td>
</tr>
<tr>
<td>Unmanaged</td>
<td>-6.8994</td>
<td>0.7873</td>
<td>-6.4570</td>
<td>0.0000</td>
<td>-3.9730</td>
<td>0.0008</td>
<td>-4.3996</td>
<td>0.0003</td>
<td>-4.1756</td>
</tr>
</tbody>
</table>

\( t \) t-score; \( p \) p-value

Table 2  Welch’s statistical significance t-test for November decision timing: wetland fail data at daily (1), monthly (30), four-monthly (120), yearly (365), two-yearly (730), five-yearly (1,825) and ten-yearly (3,650), entitlement-only and unmanaged

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>30</th>
<th>120</th>
<th>365</th>
<th>730</th>
<th>1,825</th>
<th>3,650</th>
<th>Entitlement Only</th>
<th>Unmanaged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
</tr>
<tr>
<td>1</td>
<td>0.0000</td>
<td>1.0000</td>
<td>1.5225</td>
<td>0.1365</td>
<td>7.6762</td>
<td>0.0000</td>
<td>13.1746</td>
<td>0.0000</td>
<td>13.1746</td>
</tr>
<tr>
<td>30</td>
<td>-1.5225</td>
<td>0.1365</td>
<td>0.0000</td>
<td>1.0000</td>
<td>6.8929</td>
<td>0.0000</td>
<td>12.9785</td>
<td>0.0000</td>
<td>12.9785</td>
</tr>
<tr>
<td>120</td>
<td>-7.6762</td>
<td>0.0000</td>
<td>-6.8929</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>6.8407</td>
<td>0.0000</td>
<td>6.8407</td>
</tr>
<tr>
<td>365</td>
<td>-13.1746</td>
<td>0.0000</td>
<td>-12.9785</td>
<td>0.0000</td>
<td>-6.8407</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>730</td>
<td>-13.1746</td>
<td>0.0000</td>
<td>-12.9785</td>
<td>0.0000</td>
<td>-6.8407</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1,825</td>
<td>-13.1746</td>
<td>0.0000</td>
<td>-12.9785</td>
<td>0.0000</td>
<td>-6.8407</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>3,650</td>
<td>-13.1746</td>
<td>0.0000</td>
<td>-12.9785</td>
<td>0.0000</td>
<td>-6.8407</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-6.8407</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Entitlement Only</td>
<td>-13.1746</td>
<td>0.0000</td>
<td>-12.9785</td>
<td>0.0000</td>
<td>-6.8407</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Unmanaged</td>
<td>-8.1176</td>
<td>0.0000</td>
<td>-7.5954</td>
<td>0.0000</td>
<td>-4.9125</td>
<td>0.0001</td>
<td>-2.0103</td>
<td>0.0562</td>
<td>-2.0103</td>
</tr>
</tbody>
</table>

\( t \) t-score; \( p \) p-value
### Appendix 2

#### Table 3  Welch’s statistical significance t-test for January decision timing: domestic well one fail data at daily (1), monthly (30), four-monthly (120), yearly (365), two-yearly (730), five-yearly (1,825) and ten-yearly (3,650), entitlement-only and unmanaged

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>30</th>
<th>120</th>
<th>365</th>
<th>730</th>
<th>1,825</th>
<th>3,650</th>
<th>Entitlement Only</th>
<th>Unmanaged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0000</td>
<td>1.0000</td>
<td>2.3992</td>
<td>0.0224</td>
<td>19.9388</td>
<td>0.0000</td>
<td>19.4861</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>-2.3992</td>
<td>0.0224</td>
<td>0.0000</td>
<td>1.0000</td>
<td>21.4270</td>
<td>0.0000</td>
<td>20.9522</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>-19.9388</td>
<td>0.0000</td>
<td>-21.4270</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>-0.6537</td>
<td>0.5173</td>
<td></td>
</tr>
<tr>
<td>365</td>
<td>-19.4861</td>
<td>0.0000</td>
<td>-20.9522</td>
<td>0.0000</td>
<td>0.6537</td>
<td>0.5173</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>730</td>
<td>-21.6845</td>
<td>0.0000</td>
<td>-23.5561</td>
<td>0.0000</td>
<td>-1.6404</td>
<td>0.1092</td>
<td>-2.3144</td>
<td>0.0261</td>
<td></td>
</tr>
<tr>
<td>1,825</td>
<td>-15.0438</td>
<td>0.0000</td>
<td>-14.6688</td>
<td>0.0000</td>
<td>-2.6539</td>
<td>0.0131</td>
<td>-3.0673</td>
<td>0.0048</td>
<td></td>
</tr>
<tr>
<td>3,650</td>
<td>-17.1724</td>
<td>0.0000</td>
<td>-17.0122</td>
<td>0.0000</td>
<td>-3.8005</td>
<td>0.0007</td>
<td>-4.2531</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>Entitlement Only</td>
<td>-37.0851</td>
<td>0.0000</td>
<td>-40.7383</td>
<td>0.0000</td>
<td>-18.8642</td>
<td>0.0000</td>
<td>-19.6200</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>Unmanaged</td>
<td>-7.9792</td>
<td>0.0000</td>
<td>-7.7023</td>
<td>0.0000</td>
<td>-4.9775</td>
<td>0.0001</td>
<td>-5.0758</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-4.7330</td>
<td>0.0001</td>
<td>-4.2591</td>
<td>0.0036</td>
<td>-4.0694</td>
<td>0.0006</td>
<td>-1.8775</td>
<td>0.0754</td>
<td></td>
</tr>
</tbody>
</table>

**t** t-score; **p** p-value

#### Table 4  Welch’s statistical significance t-test for November decision timing: domestic well one fail data at daily (1), monthly (30), four-monthly (120), yearly (365), two-yearly (730), five-yearly (1,825) and ten-yearly (3,650), entitlement-only and unmanaged

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>30</th>
<th>120</th>
<th>365</th>
<th>730</th>
<th>1,825</th>
<th>3,650</th>
<th>Entitlement Only</th>
<th>Unmanaged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0000</td>
<td>1.0000</td>
<td>2.5881</td>
<td>0.0147</td>
<td>21.2813</td>
<td>0.0000</td>
<td>36.0620</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>-2.5881</td>
<td>0.0147</td>
<td>0.0000</td>
<td>1.0000</td>
<td>23.9658</td>
<td>0.0000</td>
<td>40.1143</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>-21.2813</td>
<td>0.0000</td>
<td>-23.9658</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>18.1318</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>365</td>
<td>-36.0620</td>
<td>0.0000</td>
<td>-40.1143</td>
<td>0.0000</td>
<td>-18.1318</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>730</td>
<td>-36.0620</td>
<td>0.0000</td>
<td>-40.1143</td>
<td>0.0000</td>
<td>-18.1318</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>1,825</td>
<td>-36.0620</td>
<td>0.0000</td>
<td>-40.1143</td>
<td>0.0000</td>
<td>-18.1318</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>3,650</td>
<td>-36.0620</td>
<td>0.0000</td>
<td>-40.1143</td>
<td>0.0000</td>
<td>-18.1318</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Entitlement Only</td>
<td>-36.0620</td>
<td>0.0000</td>
<td>-40.1143</td>
<td>0.0000</td>
<td>-18.1318</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Unmanaged</td>
<td>-7.9747</td>
<td>0.0000</td>
<td>-7.6843</td>
<td>0.0000</td>
<td>-4.8507</td>
<td>0.0001</td>
<td>-1.8574</td>
<td>0.0783</td>
<td></td>
</tr>
</tbody>
</table>

|       | 7.9747| 0.0000| 7.9747| 0.0000| 7.9747| 0.0000| 7.9747| 0.0000         |           |
### Appendix 3

Table 5  
Welsh’s statistical significance t-test for January decision timing: domestic well/ two fail data at daily (1), monthly (30), four-monthly (120), yearly (365), two-yearly (730), five-yearly (1,825) and ten-yearly (3,650), entitlement-only and unmanaged

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>30</th>
<th>120</th>
<th>365</th>
<th>730</th>
<th>1,825</th>
<th>3,650</th>
<th>Entitlement Only</th>
<th>Unmanaged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t(0.0000)</td>
<td>p(1.0000)</td>
<td>t(15.3104)</td>
<td>p(0.0000)</td>
<td>t(-23.9253)</td>
<td>p(0.0000)</td>
<td>t(-22.7615)</td>
<td>p(0.0000)</td>
<td>t(-15.1419)</td>
</tr>
<tr>
<td></td>
<td>1.825</td>
<td>3,650</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Entitlement Only</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
</tr>
<tr>
<td></td>
<td>Unmanaged</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
</tr>
</tbody>
</table>

\( t \) t-score; \( p \) p-value

Table 6  
Welsh’s statistical significance t-test for November decision timing: domestic well/ two fail data at daily (1), monthly (30), four-monthly (120), yearly (365), two-yearly (730), five-yearly (1,825) and ten-yearly (3,650), entitlement-only and unmanaged

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>30</th>
<th>120</th>
<th>365</th>
<th>730</th>
<th>1,825</th>
<th>3,650</th>
<th>Entitlement Only</th>
<th>Unmanaged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t(0.0000)</td>
<td>p(1.0000)</td>
<td>t(14.3057)</td>
<td>p(0.0000)</td>
<td>t(33.3087)</td>
<td>p(0.0000)</td>
<td>t(33.3087)</td>
<td>p(0.0000)</td>
<td>t(33.3087)</td>
</tr>
<tr>
<td></td>
<td>1.825</td>
<td>3,650</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Entitlement Only</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
</tr>
<tr>
<td></td>
<td>Unmanaged</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
<td>-33.3087</td>
</tr>
</tbody>
</table>

\( t \) t-score; \( p \) p-value
### Appendix 4

#### Table 7  Welch’s statistical significance t-test for January decision timing: irrigation supply reliability data at daily (1), monthly (30), four-monthly (120), yearly (365), two-yearly (730), five-yearly (1,825) and ten-yearly (3,650), entitlement-only and unmanaged

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>30</th>
<th>120</th>
<th>365</th>
<th>730</th>
<th>1,825</th>
<th>3,650</th>
<th>Entitlement Only</th>
<th>Unmanaged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
</tr>
<tr>
<td>1</td>
<td>0.0000</td>
<td>1.0000</td>
<td>3.3577</td>
<td>0.0018</td>
<td>3.7136</td>
<td>0.0007</td>
<td>-2.7490</td>
<td>0.0091</td>
<td>-3.2960</td>
</tr>
<tr>
<td>30</td>
<td>-3.3577</td>
<td>0.0018</td>
<td>0.0000</td>
<td>1.0000</td>
<td>1.0351</td>
<td>0.3080</td>
<td>-6.1725</td>
<td>3.5956</td>
<td>-6.3222</td>
</tr>
<tr>
<td>120</td>
<td>-3.7136</td>
<td>0.0007</td>
<td>-1.0351</td>
<td>0.3080</td>
<td>0.0000</td>
<td>1.0000</td>
<td>-6.0140</td>
<td>0.0000</td>
<td>-6.2582</td>
</tr>
<tr>
<td>365</td>
<td>2.7490</td>
<td>0.0091</td>
<td>6.1725</td>
<td>0.0000</td>
<td>6.0140</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>-0.8333</td>
</tr>
<tr>
<td>730</td>
<td>3.2960</td>
<td>0.0022</td>
<td>6.3223</td>
<td>0.0000</td>
<td>6.2582</td>
<td>0.0000</td>
<td>0.8333</td>
<td>0.4101</td>
<td>0.0000</td>
</tr>
<tr>
<td>1,825</td>
<td>4.4866</td>
<td>0.0001</td>
<td>7.2571</td>
<td>0.0000</td>
<td>7.1483</td>
<td>0.0000</td>
<td>2.2228</td>
<td>0.0329</td>
<td>1.3520</td>
</tr>
<tr>
<td>3,650</td>
<td>3.1968</td>
<td>0.0031</td>
<td>5.7251</td>
<td>0.0000</td>
<td>5.8552</td>
<td>0.0000</td>
<td>1.0990</td>
<td>0.2799</td>
<td>0.3611</td>
</tr>
<tr>
<td>Entitlement Only</td>
<td>-25.8166</td>
<td>0.0000</td>
<td>-23.8010</td>
<td>0.0000</td>
<td>-17.9143</td>
<td>0.0000</td>
<td>-28.1187</td>
<td>0.0000</td>
<td>-25.7217</td>
</tr>
<tr>
<td>Unmanaged</td>
<td>1.9090</td>
<td>0.0713</td>
<td>2.7342</td>
<td>0.0132</td>
<td>2.9884</td>
<td>0.0072</td>
<td>1.1893</td>
<td>0.2488</td>
<td>0.9333</td>
</tr>
</tbody>
</table>

$t$ t-score; $p$ p-value

#### Table 8  Welch’s statistical significance t-test for November decision timing: irrigation supply reliability data at daily (1), monthly (30), four-monthly (120), yearly (365), two-yearly (730), five-yearly (1,825) and ten-yearly (3,650), entitlement-only and unmanaged

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>30</th>
<th>120</th>
<th>365</th>
<th>730</th>
<th>1,825</th>
<th>3,650</th>
<th>Entitlement Only</th>
<th>Unmanaged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
<td>p</td>
<td>t</td>
</tr>
<tr>
<td>1</td>
<td>0.0000</td>
<td>1.0000</td>
<td>3.0889</td>
<td>0.0037</td>
<td>3.5959</td>
<td>0.0009</td>
<td>25.9106</td>
<td>0.0000</td>
<td>25.9106</td>
</tr>
<tr>
<td>30</td>
<td>-3.0889</td>
<td>0.0037</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.7153</td>
<td>0.4789</td>
<td>21.9878</td>
<td>0.0000</td>
<td>21.9878</td>
</tr>
<tr>
<td>120</td>
<td>-3.5959</td>
<td>0.0009</td>
<td>-0.7153</td>
<td>0.4789</td>
<td>0.0000</td>
<td>1.0000</td>
<td>19.4823</td>
<td>0.0000</td>
<td>19.4823</td>
</tr>
<tr>
<td>365</td>
<td>-25.9106</td>
<td>0.0000</td>
<td>-21.9878</td>
<td>0.0000</td>
<td>-19.4823</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>730</td>
<td>-25.9106</td>
<td>0.0000</td>
<td>-21.9878</td>
<td>0.0000</td>
<td>-19.4823</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1,825</td>
<td>-25.9106</td>
<td>0.0000</td>
<td>-21.9878</td>
<td>0.0000</td>
<td>-19.4823</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>3,650</td>
<td>-25.9106</td>
<td>0.0000</td>
<td>-21.9878</td>
<td>0.0000</td>
<td>-19.4823</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Entitlement Only</td>
<td>-25.9106</td>
<td>0.0000</td>
<td>-21.9878</td>
<td>0.0000</td>
<td>-19.4823</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Unmanaged</td>
<td>1.8835</td>
<td>0.0748</td>
<td>2.6978</td>
<td>0.0141</td>
<td>2.8894</td>
<td>0.0091</td>
<td>8.2505</td>
<td>0.0000</td>
<td>8.2505</td>
</tr>
</tbody>
</table>

63


Bredelhoft JD, Reichard EG, Gorelick SM (1995) If it works, don’t fix it: benefits from regional ground-water management. In: Groundwater models for resources analysis and management. CRC, Boca Raton, FL, pp 101-121


DELWP (2015a) Ministerial guidelines for groundwater licensing and the protection of high value groundwater dependent ecosystems. Government of Victoria, Victoria, Australia


DOW (2009) Gnarangara groundwater areas allocation plan. Dept of Water, Government of Western Australia, Perth, Australia


Frazier vs Brown (1861) Frazier vs Brown. Ohio Supreme Court, Columbus, OH


NREATS (2009) Water allocation plan for the Tindall Limestone Aquifer Katherine 2009-2019. Dept. of Natural Resources Environment, the Arts and Sport, Northern Territory Government, Katherine, NT, Australia


Srikanthan S, Chiew F, and Frost, A (2006), Stochastic Climate Library Version 2.2. eWater, Canberra

SRW (2010) Groundwater management plan Koo Wee Rup water supply protection area. Southern Rural Water, Maffra, Australia


The deep ocean stream from whom all rivers take their waters, and all branching seas, all springs and all deep sunk wells.

Homer, The Illiad

Chapter 5: Calibration and Uncertainty Analysis Background Information.
5.1 Background Information

5.1.1 Uncertainty in Natural Systems

Natural systems exhibit a large degree of heterogeneity that makes groundwater management modelling challenging [Gorelick and Zheng, 2015], because models can never capture the complexity of natural systems. Simplification of real-world complexity and incomplete knowledge of the subsurface leads to uncertainty [Bredehoeft, 2005]. Compounding this intrinsic uncertainty is groundwater management uncertainty, which arises due to ambiguity in system characterisation, potential future climate and stress forecasts, varying management objectives, political whims, and unpredictable stakeholder behaviour [Guillaume et al., 2016]. This uncertainty not only needs to be accounted for, and data acquisition improved, so that management can be better modelled [Ojha et al., 2015], but also, understood and incorporated into management practices and decision-making [Barnett et al., 2012; Doherty, 2015]. Yet, even simply conceptualising the magnitude of uncertainty is a challenge: the “known unknowns” are considerable, let alone all the nebulous “unknown unknowns” [Hunt and Welter, 2010]. Uncertainty proliferates due to scale issues and data scarcity, with estimates of hydraulic parameters often varying across several orders of magnitude, not only across, but also within hydro-stratigraphic layers [Fetter, 2001; Freeze and Cherry, 1979].

Sources of uncertainty that impact predictions can be grouped into four major types: measurement uncertainty, parameter uncertainty, structural uncertainty and future scenario uncertainty (such as future climate input) [Anderson et al., 2015; Middlemis et al., 2019]. Measurement uncertainty stems from inaccuracies incurred when measurements were taken, for example, errors in measurement techniques or instrument accuracy [Anderson et al., 2015]. Parameter uncertainty is due to the fact that model parameters are non-unique, i.e. more than one
set of parameters will fit the observation data [Doherty, 2015]. Structural uncertainty arises due to the inability to represent the complexity of a natural system with a model and is due to necessary simplifications to ensure a tractable model [Hunt and Welter, 2010]. Finally, the uncertainty in future stresses and system dynamics is termed scenario uncertainty [Middlemis et al., 2019] and results because modellers are not fortune tellers. All four sources contribute to model predictive uncertainty [Doherty, 2010].

5.1.2 Uncertainty in Groundwater Modelling

A concise evolution of groundwater uncertainty analysis is provided by Anderson et al [2015], and Pappenberger and Beven [2006], persuasively refute arguments opposing uncertainty estimation in hydrological models. There are volumes of studies concerning the impact on predictive uncertainty of each type of uncertainty including dominant sources such as structural [Doherty and Welter, 2010; Ferré, 2017; Gupta et al., 2012; Højberg and Refsgaard, 2005; Refsgaard et al., 2006; Rojas et al., 2008], and parameter [Beven and Binley, 1992; Gallagher and Doherty, 2007; Moore and Doherty, 2005; Shapoori et al., 2015b; Tonkin and Doherty, 2009; Tonkin et al., 2007]. Current practices indicate parameter uncertainty is a major source of uncertainty in groundwater models [Middlemis et al., 2019; Moore and Doherty, 2005], and the contribution of parameter uncertainty to total predictive uncertainty is the focus of this study. Parameter error arises because modellers can’t provide parameter detail on a commensurate scale of variation that occurs in real natural systems and influences the response of that system to stress. As a result, small scale parameter variability is not represented in models [Doherty, 2015]. During model construction, modellers are faced with the “too simplistic” or “too complex” quandary. Use of too simplistic a parameterisation scheme omits the fine scale parameter detail that may be important to predictions. And models that are too complex become unwieldy to run,
difficult to calibrate and are often highly unstable and still fail to uniquely identify parameters. The use of many parameters in a groundwater model (termed highly parameterised model) involves assigning more parameters than can be uniquely identified by a calibration dataset and results in a model that lacks a unique solution and is called an *ill-posed* inverse problem [Doherty and Hunt, 2010; Doherty et al., 2010].

Highly parameterised models can be constructed by assigning parameter values at certain points termed pilot points and interpolating the intervening values using geostatistical methods. The use of highly parameterised inversion (HPI) is advantageous because it enables the extraction of maximum information from the calibration dataset and provides a null-space (a region of parameter space where parameter values are insensitive to the data) that allows exploration of uncertainty [Doherty et al., 2010]. Compared to simple zone-based parameterisations, HPI reduces structural noise so that areas of high data worth (i.e., *where the parameters are sensitive to the calibration dataset*) [Fienen et al., 2010], can be determined. Parameter values that are sensitive to the calibration dataset are altered to fit calibration data and insensitive parameter values are not constrained and can be used to explore parameter uncertainty. Hunt et al [2019], recently applied advances in parameter estimation methods to the often used and thirty year old Freyberg model [Freyberg, 1988]. They found highly parameterised models using pilot points constrained by geological knowledge provided the best model predictions. The usage of a large numbers of parameters combined with regularised inversion (*process that provides a method to solve ill-posed inverse problems by attaining parameter uniqueness*), extracts more information from an observation dataset than usage of simple zone-based parameterisation [Moore and Doherty, 2006].
In the study described in chapter 6, a synthetic groundwater model of an unconfined aquifer was designated as reality and a highly parameterised model of the reality system was constructed using pilot points [e.g. Tonkin and Doherty, 2009]. The purpose of the study was to explore the impact of model parameter uncertainty on the uncertainty around predictions of plan effectiveness. Because the model used was highly parameterised, the inversion was ill-posed. Ill-posed environmental modelling inverse problems are widely solved using PEST, a model independent, open source software package that enables parameter estimation and parameter/predictive uncertainty analysis [Doherty, 2018a; b; Doherty and Hunt, 2010]. A central step in the least-squares calibration approach used by PEST involves solving linear equations and matrix manipulations. The point of a least-squares calibration is to improve the model by minimising the sum of squared differences between observations and simulated model outputs by altering the parameter values to better reproduce historical observation data. Some background information on linear algebra techniques and how ill-posed inverse problems are solved are given below to explain why certain approaches were used.

5.2 The Calibration Process

The purpose of calibration is to alter model parameter values until the model acceptably reproduces historical observation data. During calibration, observation data, parameter values and the action of the model are represented by a system of linear equations which are expressed in matrices and vectors of the form

\[ y = Ax \quad \text{Equation 5.1} \]

where \( A \) is a \( n \times m \) matrix that represents the action of a model on its parameters; model parameters are represented by the vector \( m \times 1 \) \( x \); and \( y \) is a \( n \times 1 \) vector that represents model
predictions corresponding to the observations [Doherty, 2010; Doherty, 2015], where \( m \) is the number of parameters and \( n \) is the number of observations. The matrix \( A \) changes as the calibration process advances (iterations) and is further discussed in section 5.2.2. A local calibration method was used in the study described in chapter 6 and while global calibration methods exists, discussion of such methods are beyond the scope of this background section that provides supporting information for chapter 6.

5.2.1 Solving Systems of Linear Equations

In order to solve the system of linear equations, linear algebra techniques and matrix manipulations are used. A system of linear equations has a solution if at least one set of numbers exists that can replace the variables in the equations and make each equation read as true [Strang et al., 1993]. In the calibration context, the variables are parameter values that when acted upon by the model, produce the model estimates of the observation data. Model parameters will never be definitively known due to uncertainty. However, based on expert judgment, values that are considered likely and unlikely can be defined [Doherty, 2015] and form a parameter probability distribution. A parameter distribution based upon expert judgement and a conceptual model alone is termed the prior distribution. During calibration the values of parameters are informed and shaped by the observation dataset and form a parameter probability distribution termed the posterior. In the study described in chapter 6, a model based on the prior parameter distribution and three posterior models are used in a synthetic study to predict plan effectiveness which is then compared to a known reality plan effectiveness value. Multiple parameters can be informed by the same observation data-point (i.e. a change in different parameters has the same impact on an observation data-point). Consequently, a particular data-point, informs a linear combination (sum of scalar multiples) of parameters rather than a single parameter. When multiple parameters
overlap and are informed by the same observation data-point the values of those parameters can’t be uniquely identified, because the linear equations are not independent. This is important because it means that the equations, and hence problem, can’t be solved uniquely.

In contrast, a system with a unique solution consists of linear equations with the same number of independent equations and variables [Doherty, 2010]. Further, a system of linear equations is only independent if no equations are multiples, or combinations of multiples, of any of the other equations in the system [Norman, 1995; Strang et al., 1993]. Systems of this nature (with unique solutions) are termed well-posed or overdetermined inverse problems. Often in hydrogeology, there are more parameters that can be uniquely identified with the dataset and this type of calibration is termed an ill-posed or underdetermined inverse problem and means that certain parameters are not informed or only partially informed by the calibration dataset. When the system of equations is dependent, (i.e. not independent) there are an infinite number of solutions [Strang et al., 1993]. The system of equations used in chapter 6 is an ill-posed inverse problem, that frequently occur in environmental systems [Doherty, 2015].

5.2.2 The Jacobian Matrix

Generally, a calibration dataset consists of a set of historical observations (such as groundwater levels) at certain locations or points in time. The calibration process starts by simulating each of the observations once to obtain an initial mismatch between historical and modelled observations. Then, the value of one model parameter is changed a little and the observations are re-simulated to see what impact the parameter change had on each simulated observation point. The small alteration in parameter value and subsequent re-simulation of every observation point is done, one at a time, for every parameter in the model. The result is an $n \times m$ matrix called the Jacobian ($J$) (which if the model was linear then $J$ would be the matrix $A$ in equation 5.1), where
$n$ is the number of observations (rows) and $m$ is the number of parameters (columns). The equation described in equation 5.1 describes a linear model and an assumption of model linearity forms the basis for much of the methods described in this chapter. However, most numerical groundwater models are non-linear, including the model used in this thesis. Because the linear algebra rules relate to linear models, linear approximations of the actions of a non-linear model on its parameters are made at points in parameter space and that is what the Jacobian contains. Each entry in the Jacobian is the derivative of a difference between the initial simulated observation and the simulated observation after a small change in parameter value, divided by the change in the parameter value. The result is a matrix of the change in the model misfit due to a small parameter perturbation (i.e. the change in residual and there is a value for every point in the calibration dataset (i.e. every time in the time-series). Part of the process used by PEST to solve the ill-posed inverse problem typical of groundwater models, for the parameters of minimum error variance involves inversion of a matrix constructed using the Jacobian matrix and a weight matrix ($Q$) that is proportional to the inverse of measurement noise [Doherty, 2015].

$$J^TQJ \quad \text{Equation 5.2}$$

Where $J^T$ is the inverse of the Jacobian matrix; $J$ is the Jacobian matrix and $Q$ is the weight matrix. Inversion of this weighted matrix can only occur under certain conditions (i.e. when the system of equations is independent) [Doherty, 2015]. Linear algebra principles dictate that if there is a mismatch between the number of observation data values and the amount of unique information provided by the parameters then the weighted matrix can’t be inverted because information conflicts could exist (i.e. different parameters might impact the same observation point in differing ways) [Doherty, 2015]. Matrix inversion is used by PEST [Doherty, 2018a; b]
to solve the systems of linear equations in Eq 5.1, so matrix manipulation must be conducted to make the system of equations independent which will then allow the inversion of the weighted matrix. While this process requires an assumption of linearity that is rarely met, the model is linearized by substitution of the Jacobian matrix, for the model $A$ in equation 5.1 at a point in parameter space. The inverse problem is solved by minimising the squares sum of differences between observed and simulated values using a regularised Gauss-Newton Levenberg Marquardt approach [Doherty, 2015] and additional detail on the Gauss-Newton Levenberg Marquardt can be found in Marquardt [1963].

The process of calculating a Jacobian requires that the model is run at least $m$ (number of parameters) +1 times, which for models with many parameters can be highly computationally intensive. However, the Jacobian provides much useful information, such as an indication of how sensitive the modelled output is to a change in each parameter and is used during the solution of the inverse problem to determine an updated parameter set.

### 5.2.3 Solving ill-posed inversions with Regularisation

In the study detailed in chapter 6, a solution to the ill-posed inverse-problem is achieved using regularisation. Regularisation provides a way to solve ill-posed inverse problems by identifying the parameter field of lowest error variance and ensuring a prediction is approximately in the centre of the posterior parameter distribution [Doherty, 2018a]. To ensure that the matrices are the right shape and contain the required information content to uniquely identify the parameters, data is added or removed. As a result, the inverse problem is solved by a combination of data addition (Tikhonov Regularisation) and data removal (Truncated Singular Value Decomposition (SVD)) [Doherty, 2015]. Tikhonov regularisation adds information to the problem in the form of prior information by providing “likely” parameter values, and truncated SVD breaks the
Jacobian down into eigenvalues and eigenvectors to determine which eigenvectors are sensitive to the data.

Matrix decomposition makes the matrix manipulations easier and yields a unique solution [Doherty, 2015] because insensitive eigenvectors are removed or “truncated” [Doherty, 2010; Doherty and Hunt, 2010; Doherty et al., 2010; Tonkin and Doherty, 2009]. Once the inverse problem has been solved, a parameter update vector is calculated, and the model and the residuals are re-calculated. If the reduction in objective function remains larger than a user-input convergence threshold, then the process is repeated (this is called an iteration). Each iteration of the calibration process produces a Jacobian matrix and a parameter set and when the residual falls below the convergence criteria, the calibration is complete.

Use of this method to solve the inverse problem assumes model linearity whereas as detailed above, in reality, groundwater models are rarely linear. Therefore, the process needs to be done over iterations, each of which, refines the parameter estimates (i.e. multiple parameter update vectors need to be calculated based on linearizing a model at various points in parameter space). For more detailed explanation on this process see Doherty [2010; 2015].

5.3 Parameter/Predictive Uncertainty Analysis

Non-uniqueness is a well-established concept in groundwater modelling [McLaughlin and Townley, 1996], and only a single solution is provided from the calibration methods described above. However, there are multiple, alternate parameter sets that would also fit the observation data. All these potential parameter sets form a distribution that, as described in Section 5.2.3, is assumed to be Gaussian. The point in a distribution of minimum error variance is the centre. So, the solution found during calibration is the parameter set at the centre of the parameter
distribution. But, the actual, true, parameter values could lie anywhere in the distribution. Monte Carlo (MC) uncertainty analysis methods involve randomly sampling from a specified parameter distribution with the premise that when enough samples are taken, then the sampled distribution will approach the true distribution. However, MC methods, such as GLUE [Beven and Binley, 1992], that assigns a likelihood value to each parameter set based on the misfit between modelled and observed values, require many model runs, which is not practical for models with long run times and many parameters (such as the model used in chapter 6 that ran in 55 minutes). Markov Chain Monte Carlo (MCMC) methods take a sample of parameter realisations that form a Markov chain instead of sampling from the prior and are more efficient than likelihood rejection sampling methods but, are problematic with long running models and many parameters [Doherty, 2015]. Therefore, Null Space Monte Carlo (NSMC) methods [Tonkin and Doherty, 2009] were chosen for the study detailed in chapter 6.

5.3.1 Null Space Monte Carlo Techniques

A calibration-constrained subspace technique termed Null-Space Monte Carlo [Tonkin and Doherty, 2009] was used in chapter 6, that explored uncertainty with considerably fewer model runs than MC methods because the Jacobian matrix was used in NSMC analysis instead of model parameters and simulated observations [Doherty, 2015]. NSMC methods [Tonkin and Doherty, 2009] allowed exploration of uncertainty by projecting random realisations of prior parameter values on to the parameter “null space” where parameters were insensitive to the data and maintaining parameter values at close to calibrated values in the parameter “solution space” where parameter sensitivity was high. Tonkin and Doherty [2009] define the following steps to NSMC methods:

1. Calibrate model using regularisation.
2. Separate parameter space into two subspaces (solution space and null space) using Singular Value Decomposition (SVD) (Figure 1).

3. Generate stochastic realisations of model parameters based on prior distribution

4. Calculate difference between each stochastic realisation and calibrated model parameters

5. Project differences onto the parameter null space and add difference to calibration solution space (Figure 2).

6. Recalibrate model if parameter differences have caused the model to no longer be calibrated.

7. Run model with each parameter realisation to assess uncertainty.

A null space exits if there are parameters or linear combinations of parameters that are not informed by the observation data [Doherty, 2015]. Consequently, parameters and combinations of parameters in the null space could not be estimated during the calibration and those in the solution space could.
Figure 1: Singular value decomposition of the Jacobian matrix (USVᵀ) into three component matrices (U, S, V₁ and V₂). For additional detail, see Doherty (2015).

Figure 2: Null-Space Monte Carlo Analysis where prior parameter realisations are projected onto the parameter null-space and solution space parameters remain close to calibrated values. For additional detail see Tonkin and Doherty 2009.
Singular Value Decomposition was used during NSMC analysis to decompose the Jacobian matrix (around the calibrated models) down into three component matrices. The process provided a new orthogonal (linearly independent) coordinate system for the parameter range space, subdivided the parameter space into two orthogonal parameter sub-spaces (null-space and solution space) and gave an indication of the information content of the calibration dataset [Doherty, 2015]. The new coordinate system was described by \( m \) (number of parameters) \textit{eigenvectors} (special vectors that when multiplied by a matrix, result in multiples of themselves) that were each associated with a number called an \textit{eigenvalue} [Doherty, 2015; Strang et al., 1993]. Each eigenvector in a particular coordinate space is linearly independent to all other eigenvectors (i.e. points in a unique or orthogonal direction in parameter space). Eigenvectors contain linear combinations of parameter values that are weighted by singular values (square root of eigenvalues). Eigenvalues weight each eigenvector based on how much information about the linear combinations of parameters, that eigenvector contains, is in the calibration dataset. The entries in each eigenvector (which sum to one) are the weights or proportions of the original parameters that are associated with that eigenvector (linear combination of parameters). Large singular values and associated eigenvectors are assigned to the solution space and small singular values and associated eigenvectors are assigned to the null-space. This division of parameter space allows an exploration of the impact of parameter uncertainty on predictive uncertainty and is the purpose of chapter 6.

\textbf{5.4 Concluding Remarks}

Chapter 6 describes a calibration-constrained predictive uncertainty analysis that was conducted using PEST and employed NSMC methods to investigate whether a reduction in model
parameter uncertainty resulted in a commensurate reduction in predictive uncertainty of predictions of management plan effectiveness. A synthetic study was designed where a numerical groundwater model constituted reality and several simpler models, that were calibrated to observation datasets of differing length and extents, were created to approximate the synthetic reality. It was theorised that decreasing parameter uncertainty through calibration to increasingly extensive datasets would decrease the predictive uncertainty of plan effectiveness. A highly parameterised model was required because parameters had to be defined on a scale that, while not capturing the true complexity of reality, attempted to capture the pertinent process that controlled groundwater flow. As a result, the inverse problem was ill-posed and PEST was used to solve ill-posed inverse problems.
Chapter 6: Can we model management?
Quantifying data requirements for management plan evaluation.

This chapter is in preparation as the following article:

---

*It is flowing, ever flowing, in a free unstinted measure
from the silent hidden places where the old earth hides her treasure*

A.B. Banjo Patterson, Song of the Artesian Water
Can management be modelled?

E. K. White. Corresponding author
Department of Infrastructure Engineering, University of Melbourne, Victoria, Australia, 3010, whitee1@student.unimelb.edu.au

A. W. Western.
Department of Infrastructure Engineering, University of Melbourne, Victoria, Australia, 3010, a.western@unimelb.edu.au

T. J. Peterson.
Department of Infrastructure Engineering, University of Melbourne, Victoria, Australia, 3010, timjp@unimelb.edu.au

E. Carrara.
Bureau of Meteorology, Docklands, Victoria, Australia, 3001, elisabetta.carrara@bom.gov.au

K.H. Hayley.
Groundwater Solutions, Melbourne, Australia, 3051, kevin.h.hayley@gmail.com

J. Schumacher
Groundwater Solutions, Melbourne, Australia, 3051, jschumacher@fluid-domains.com

Conflict of interest: None

Key words: Groundwater management, Numerical modelling, Uncertainty analysis, Null Space Monte Carlo, Water resources conservation, Groundwater protection.

Article Impact Statement: Management plan effectiveness cannot be accurately predicted using a numerical groundwater model due to uncertainty and calibration-induced model bias.
Abstract

Given the high degree of parameter uncertainty typical in groundwater models, this study aimed to determine how much calibration data was necessary to quantitatively evaluate management plan effectiveness. A synthetic study was used to evaluate the uncertainty around predictions generated from four different groups of model realisations, created based on increasing amounts of observation data generated by a numerical groundwater model designated as reality. Four simple models of the reality system were built based on prior knowledge, and a calibrated solution to each of the three different observation datasets (three posterior distributions). Each model was used to predict the effectiveness of management decision-making on a monthly basis and both entitlement-only and unmanaged scenarios. The predictive uncertainty was quantified for prior and posterior models, through a calibration-constrained uncertainty analysis using Null-Space Monte Carlo methods. The objective was to evaluate if more extensive observation data can sufficiently increase system understanding through calibration so that the plan effectiveness predictions approach the reality effectiveness. Because models cannot simulate the complexity of natural systems, simplifications are required; for example, time-varying recharge was set as static. Due to this simplification, other parameters such as hydraulic conductivity, assumed inappropriately high values to compensate resulting in calibration-induced model bias that caused the models to make erroneous predictions of plan effectiveness. Even with use of current best practice uncertainty analysis methods, the effectiveness of management could not be determined due to the limitations of the numerical models utilised, which raised serious question over our current ability to model management.
6.1 Introduction

6.1.1 Managing Groundwater under Uncertainty

When Emerson penned the phrase “what we know is a point to what we do not know”, he could well have been describing groundwater modelling (Emerson 1836). Point measurements from wells are used to extrapolate hydro-geological conditions, from which numerical models of groundwater flow are constructed. But these conceptualisations are based upon limited data and the true, underground conditions are anyone’s guess. Uncertainty is an intrinsic and unavoidable aspect of modelling and can only be reduced through data collection and ground truthing (Anderson, Woessner, and Hunt 2015). And sometimes, depending on the prediction and the data, even data collection can do little to reduce uncertainty (Doherty 2015, Zech 2016). Furthermore, extensive data collection is usually prohibitively expensive, particularly for groundwater managers, who usually must operate under finite budgets and make aquifer wide decisions based on a handful of wells (Australian Government 2017, Jakeman et al. 2016). Consequently, modelling is often used to inform groundwater management decisions, most commonly through optimisations of various extraction regimes or management strategies (Gorelick and Zheng 2015, Singh 2012, 2014, Gorelick 1983, Wagner 1995, Loucks and Van Beek 2005).

But groundwater management often involves a system control approach where extractions are modified in response to groundwater level variations, forming a feedback loop (White et al. 2016). Model can also be used in this context to predict the effectiveness of management plans (White et al. 2019), because the performance of many plans, distinct from climate or other external forcing and previous management intervention, is often unknown (White et al. 2016). Furthermore, because it is difficult to distinguish between changes in plan effectiveness due to the actions of the plan and, changes due to alternate model representation; the accuracy of predictions of plan effectiveness is currently unknown (White...
et al. 2019). However, often groundwater models are used to support groundwater management decisions. In this paper we investigate if models can be used to reliably estimate plan effectiveness and how estimations of effectiveness change with increasing amounts of observation data. A synthetic study was used to determine how well the model parameters must be known in order to predict plan effectiveness within reasonably narrow uncertainty bounds using a numerical groundwater model.

Models are imperfect simulators of natural systems and may fail to include vital system characteristics (Jakeman et al. 2016). Consequently, models can’t predict what is going to happen (Oreskes et al., 1994); but, within a reasonable level of confidence, may predict what won’t happen (Doherty 2018). Therefore, management plans should be created with SMART (specific, measurable, achievable, realistic, timely) objectives (Doran 1981) that ask not if something will happen but instead, if it won’t happen. Furthermore, plans must be testable (White et al. 2016), with objectives that have relatively low predictive uncertainty (Guillaume et al. 2016), so whether or not the objectives are likely to be met can be evaluated. Evaluation of predictive uncertainty is important for robust management modelling because ensuring adverse impacts are statistically unlikely to occur under particular management scenarios is critical for an informed risk assessment (Doherty 2010; Walker 2017).

Modelling supporting important decisions should be guaranteed to exaggerate uncertainty around predictions of adverse impacts (Doherty and Simmons 2013) and a comprehensive analysis of the uncertainty of parameters, and the dependent model predictions, is crucial when using models to support decision-making (Doherty, Hunt, and Tonkin 2010). Decision-makers, however, often prefer a single value to base actions upon compared to a range of potential values that are generated during uncertainty analysis. So, a crucial groundwater modelling challenge is incorporating various types of uncertainty into user friendly decision-support systems with a quantifiable risk that can be used by water managers and policy
makers (Gorelick and Zheng 2015). Clear communication of the likelihood of adverse impact occurrence is relevant in a decision-making context and is the reason predictive uncertainty of plan objectives should small enough to instil confidence that the adverse impacts are unlikely to occur. However, in practice, the acceptable level of uncertainty will be a pragmatic judgment call that will depend upon the severity of risk and consequences.

6.1.2 Purpose of Study

Plans often involve a series of rules and management decisions based on groundwater levels that over time, form a sequential decision-making process. In contrast to optimisations which seek the “optimal” management strategy, testing the sequential decision-making process in a modelling framework may provide an indication of the efficacy of the plan. White et al., 2019, presented a methodology to simulate the sequential decision-making process of groundwater management as a feedback loop using a simple synthetic aquifer and, determined the plan effectiveness for various combinations of decision-making frequency and timings. It was found that management plan effectiveness was highly sensitive to aquifer response time, which was dominantly controlled by hydraulic conductivity and model geometry. Motivations for this study included quantifying the uncertainty around plan effectiveness and determining whether reducing parameter uncertainty through calibration, decreases predictive uncertainty.

In this study, following the methodology outlined in White et al (2019), groundwater management was framed as a systems control problem to capture the aquifer/management feedback, where groundwater levels dictate management actions that are predefined in a management plan. A synthetic groundwater model was designated as reality and used for the dual purpose of generating observation data and providing a reality effectiveness for comparison purposes. The uncertainty around predictions of management plan effectiveness predicted using four simplified groups of models of the reality groundwater model was
quantified. Each simple model was informed by a different observation dataset, generated by
the *reality* system, that varied in length of time and number of monitoring wells. The
objective was to evaluate if more extensive observation datasets could increase system
understanding through calibration to such a degree that the plan effectiveness predictions
approached the *reality*. The hypothesis was that increasing system understanding through
calibration to a more extensive dataset would increase the accuracy of the prediction and that
the model predicted failure rate would approach the *reality* failure rate. The main question
this study sought to answer was "How much observation data do you need to manage
groundwater?" (i.e. how much calibration data is required to constrain model predictions, so
they are useful and informative?). Given model parameter uncertainty, the purpose was to
determine the feasibility of using numerical models, informed by different observation
datasets, to accurately and precisely predict management plan effectiveness by simulating
management as a feedback-driven sequential decision-making process.

6.2 Methodology

During construction of a groundwater model, geological interpretation, previous
investigations and professional judgement is utilised to develop a conceptual model of the
system that includes information such as potential parameter values. The potential parameter
values are usually represented as a probability distribution, within which, the true parameter
values likely lie. A parameter distribution based upon the conceptual model alone is termed
the *prior* distribution (*based on expert judgment*). In this paper, *prior* models were populated
with parameter values that were randomly generated from a log-uniform distribution and
based on limited information. Bayesian methods state that as more information is gained,
estimations of parameter values can be revised since the observation data informs and
constrains parameter values (Doherty 2015) (Equation 1).

\[ P(k|h) \propto P(h|k)P(k) \]  \hspace{1cm} \text{Equation 1}
Where $P(k|h)$ is a likelihood function calculated from the model fit to observation data, $P(k)$ is the prior probability and $P(k|h)$ is the posterior probability, and $k$ is a parameter and $h$ represents historical measurements (Doherty 2010). Parameter values that are constrained by calibration datasets are termed posterior parameter distributions (informed by calibration dataset), and generally, have smaller uncertainty bounds than the prior.

In this study, three calibration datasets of groundwater levels and extractions rates were generated using a numerical groundwater model representing a synthetic reality system. The reality system was subject to an entitlement volume and calibration level and extraction data were produced under an “entitlement only” scenario, where extractions volumes are limited to 80% of maximum demand. Subsequently, two groups of models, prior models and posterior models, were created to approximate the reality system based on increasing degrees of system information. The three posterior distributions were created through the processes of calibration and Null-Space Monte Carlo (NSMC) (Tonkin and Doherty 2009) analysis. The PEST suite of software tools allows model-independent uncertainty analysis (Doherty, Hunt, and Tonkin 2010, Doherty and Hunt 2010) and were used in this study. Fifty realisations of each of the four groups of models were generated and predictions of plan effectiveness were compared to determine the amount of observation data required to sufficiently constrain parameter uncertainty.

The prior and posterior model ensembles were run at a daily time-step under three management scenarios: (i) monthly management decision-making (review of water levels and implementation of any required restrictions as specified in the plan), (ii) entitlement only, and (iii) an unmanaged scenario. The plan used by White et al (2019) and in this study is provided in Appendix 1 and additional details on methodology are provided in White et al (2019).
The plan effectiveness based on the percentage of fail days at domestic well one i.e. number of days the well becomes dry was calculated for each scenario (section 6.3.2). The distributions of predictions of plan effectiveness were assessed and compared to the true reality effectiveness. Calibration and uncertainty analysis were facilitated with the PEST suite of software (Doherty 2015). FloPy (Bakker et al. 2016) and PyEMU, a collection of python modules that allows linear and non-linear uncertainty analysis and geostatistical interpolation (White, Fienen, and Doherty 2016), were used to interact with MODFLOW and PEST. The predicted plan effectiveness for both prior and posterior models and was conducted in the following sequence:

1. Generation of three calibration datasets using a synthetic reality system.
2. Generation of prior parameter distribution.
3. For each of the three calibration datasets: a model was calibrated with BeoPEST and a calibration constrained parameter uncertainty analysis using NSMC methods was conducted with PEST.
4. Predictive uncertainty analysis of management plan effectiveness using four different parameter ensembles associated with the prior, and three posterior distributions.

6.2.1 Generation of Calibration Datasets Using Synthetic Reality System

A numerical groundwater model of a synthetic, unconfined aquifer governed by a management plan was generated using MODFLOW-NWT and FloPy (Bakker et al. 2016, Niswonger, Panday, and Ibaraki 2011). A synthetic study was conducted because the purpose was to determine if the model could accurately predict management plan effectiveness and consequently, the true effectiveness needed to be known for comparison purposes. As detailed in section 6.1.2, the synthetic model was designated as reality and had two purposes (1) to generate calibration data (e.g. groundwater hydrographs representing synthetic observation bores) and (2) to provide a true reference or basis to compare the modelled plan
effectiveness. For the purposes of this study, the synthetic observation data was assumed not to contain measurement error and the conceptual model of the system (i.e. model domain, aquifer dimensions, well configuration, model boundary conditions, etc) did not change between reality, prior and posterior models.

The model shared the conceptualisation used by White et al, 2019 and was constructed to mimic a small-scale irrigation operation in an alluvial upland valley. A generic groundwater management plan that governed the aquifer was defined in Python programming language had an objective of preventing dewatering of a domestic well (DS1) by maintaining groundwater levels in the vicinity of the well above the dry depth (Figure 1) (White et al. 2019, Python Software Foundation, 2019). The irrigation operation was allocated a particular volume of water each year (an entitlement) and was subject to extraction restrictions in the supply wells if predefined trigger levels in two monitoring wells were reached. The impact of the management plan was evaluated in a simple synthetic aquifer system by comparing the number of days the domestic well went dry under monthly management compared to two baseline scenarios (entitlement volume only and unmanaged).

The aquifer dimensions were 2.5 km by 2.5 km and discretised by a 352 x 229 rectangular grid with a background spacing of 15 m. Grid spacing contracted around pumping and monitoring wells to capture head fluctuations due to extractions (White et al, 2019). Due to sloping topography, the aquifer thickness decreased from 92 m in the northeast to 70 m in the south west. West and east domain margins were represented by a river and a constant head boundary respectively. No flow boundaries comprised the base of the model and the north and south boundaries of the model domain. Groundwater was extracted from twelve pumping wells (black circles on Figure 1) in three irrigation zones and time-varying demand, recharge and evapotranspiration were estimated on a model grid scale using a vertically integrated 1-D soil moisture model and based on climate data recorded at Nhill in western Victoria and
obtained from the AWAP dataset (White et al, 2019). Hydraulic conductivity was represented on a model grid scale by a spatially correlated field with a mean value of 0.26 m/day and was stochastically generated using a geostatistical structure described by two variograms (exponential and spherical) with anisotropy bearings of 72 ° that aligned with the principle direction of groundwater flow (Figure 1). The reality system was used to generate three calibration datasets that varied from a sparse monitoring network measured for a short timeframe to an extensive monitoring network monitored for a long time:

1. **Short-sparse dataset**: Three years of water level measurements at two monitoring wells and extraction rates at twelve supply wells.
2. **Long-sparse dataset**: Twenty-five years of water levels at two monitoring wells and extraction rates at twelve supply wells.
3. **Long-extensive dataset**: Twenty-five years of water levels at twenty monitoring wells and extraction rates at twelve supply wells.
6.2.2 Model Parameterisations and Prior Distribution

Simplifications were required so the reality system could be approximated with a finite number of parameters and allowed a calibrated solution to be achieved without an excessive number of modelling simulations. Therefore, the domains of the prior and posterior models were parameterised using pilot points (Tonkin and Doherty 2009, Christensen and Doherty 2008) with four adjustable parameters: 1. hydraulic conductivity ($k_s$), 2. recharge ($R$), 3.
Evapotranspiration (ET) and specific yield ($S_y$). Pilot point locations for hydraulic conductivity, recharge, evapotranspiration and specific yield were spaced at 100, 500, 500, and 1,250 m intervals across the model domain and are shown on Figure 2. Use of pilot points resulted in a smoother hydraulic conductivity field that lacked the small-scale heterogeneity of the reality and a total of 609 parameters and avoided implementation of subjective parameter zoning.

Use of Highly Parameterised Inversion (HPI) methods was facilitated with the use of the University of Melbourne high performance computing cluster (Lafayette et al. 2016) but, can also be employed using cloud computing resources (Hunt et al. 2010, Hayley 2017). Highly parameterised models combined with subspace methods are widely considered best practice when exploring uncertainty in groundwater models. The maximum amount of information is obtained from the calibration dataset with use of HPI and exploration of uncertainty is enabled by the presence of a null-space (Doherty, Hunt, and Tonkin 2010).

Considering that a daily modelling time-step was required by the management evaluation method, and that MODFLOW input files require arrays for stresses such as recharge at each time step, the computational burden of a time-varying recharge was prohibitive. Therefore recharge, evapotranspiration and specific yield varied across the model domain but did not change over time. Applying static recharge to a groundwater model is standard industry practice and time-varying parameterisation is rarely employed (Knowling, White, and Moore 2019). The demanded groundwater extraction volumes used in the prior and posterior models were the extraction rates generated by the reality model. These extraction rates were modified by both management action (in monthly scenario) and well dewatering events.
Figure 2: Pilot point spacing for each adjustable parameter. (a) pilot points for hydraulic conductivity were spaced at 100 m intervals; (b) pilot points for recharge were spaced at 500 m intervals; (c) pilot points for evapotranspiration were spaced at 500 m intervals and (d) one pilot point was used to define specific yield at the center of the model domain.

The prior parameter distribution consisted of fifty realisations of each model parameter generated at pilot point locations. Realisations of the prior hydraulic conductivity \(0.08 \leq k_z \leq 0.8\), mean \(\mu = 0.26\) m/day), recharge \(6 \times 10^{-7} \leq R \leq 0.08\), \(\mu = 5 \times 10^{-5}\) m/day \((5 \times 10^{-2}\) mm/day), and evapotranspiration \(1 \times 10^{-7} \leq ET \leq 0.01\), \(\mu = 6 \times 10^{-5}\) m/day \((6 \times 10^{-2}\) mm/day) distributions were generated using PEST utility RANDPAR (Doherty 2018b) from log uniform distributions with the specified means and upper and lower bounds defined above. The parameter realisations were generated at pilot point locations across the model.
domain. Specific yield \((0.2 \leq S_y \leq 0.5, \mu = 0.01)\) was generated from 1,250 m spaced pilot points. The bounds of the prior parameter distributions represented the inferred, probable bounds based on the conceptual hydro-geological model. Ordinary kriging was used to interpolate parameter values from the pilot point locations to the MODFLOW-NWT model grid locations (Doherty 2016). This resulted in fifty prior models with different, but equally plausible, parameter values. Each of the fifty prior models was run at a daily time-step under a monthly management decision-making scenario, an entitlement only scenario and an unmanaged scenario. A selected prior model is shown on Figure 3a, as it can be seen, the interpolated prior hydraulic conductivity field is smoother than the reality field due to simplification and interpolation.

6.2.3 PEST Calibrations, Parameter and Predictive Uncertainty

Three calibration datasets (see Table 1) were used to calibrate a model of the reality system using the simplified parameterisation detailed in Section 6.2.3. Initial conditions for each model consisted of the mean values from the prior parameter distribution. Three non-linear calibrations were performed with BeoPEST (Schreuder 2009), a variant of PEST for use on high performance computing (HPC) systems. PEST employs the Gauss Newton local calibration method to achieve a solution, and while global calibrations methods exist, they are not practical with models with long run times and many parameters. As with all local calibration methods, the potential exists for the calibrated solution to be a local optima instead of the global optima. However, this was partially mitigated with use of NSMC methods and null-space projections and allows for a highly parameterised exploration of predictive uncertainty in computationally tractable manner.

Parallel BeoPEST runs were employed on a SLURM high performance cluster (Lafayette et al. 2016). A forward transient model was constructed as the calibration model and used to determine the parameters with lowest error variance for each calibration dataset. An initial
steady state calibration was conducted and specifications for the subsequent transient
calibrations are shown in Table 1. Initial groundwater level measurements and drawdowns
(or differences from initial) were used as targets in the calibration. For each calibration
dataset, a unique parameter realisation was obtained that had the lowest minimum predictive
error variance (Doherty 2015) and three models were successfully calibrated.

<table>
<thead>
<tr>
<th>Calibration Dataset</th>
<th>Short-sparse</th>
<th>Long-sparse</th>
<th>Long-extensive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of calibration dataset</td>
<td>3 years</td>
<td>25 years</td>
<td>25 years</td>
</tr>
<tr>
<td>Monitoring wells</td>
<td>2</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Daily MODFLOW stress periods</td>
<td>1,096</td>
<td>9,125</td>
<td>9,125</td>
</tr>
<tr>
<td>Adjustable parameters</td>
<td>609</td>
<td>609</td>
<td>609</td>
</tr>
<tr>
<td>Calibration targets</td>
<td>2,192</td>
<td>18,264</td>
<td>182,640</td>
</tr>
</tbody>
</table>

Table 1: Calibration specifications for the three BeoPEST calibrations. Calibration targets were a combination of groundwater level measurements and drawdown measurements.

NSMC analysis was conducted on each of the three base calibrated models. The number of
singular values used in the NSMC analysis was 7, 10 and 46 for the short-sparse, long-sparse
and long-extensive datasets respectively. Considering that the number of singular values
provides an indication of the amount of information in the calibration dataset, it was expected
to increase with increasing length and extent of calibration dataset. The calibration and
NSMC analysis on the three datasets provided three alternative ensembles of fifty calibration-
constrained parameter realisations (Tonkin and Doherty 2009) that were all used in predictive
scenarios. One iteration of recalibration was conducted on each set of the posterior
realisations after null space projection to ensure parameter variability and reasonable data

misfit (Doherty, Hunt, and Tonkin 2010). The following parameter realisations were used to run the management model under monthly, entitlement-only and unmanaged management frequency for a period of 50 years to determine the effectiveness of the plan under each parameterisation:

- Reality
- 50 Random Prior Realisations
- 50 Posterior Realisations (Base model calibrated to short-sparse)
- 50 Posterior Realisations (Base model calibrated to long-sparse)
- 50 Posterior Realisations (Base model calibrated to long-extensive)

6.3 Results

6.3.1 PEST Calibrations and NSMC Analysis

The calibrated hydraulic conductivity parameter fields for one prior realisation and the three calibration datasets are shown on Figure 3 and detailed in Table 2. Modelled versus simulated heads for each well used in the calibration and declines in the objective function over each calibration iteration are provided in Appendix B. The short-sparse, long-sparse and long-extensive datasets resulted in decreases from the initial objective function of two orders of magnitude: 97 %, 96% and 88 % respectively.

Observed and modelled groundwater levels produced by each calibrated model at monitoring well two for the period 1962 to 1964 are shown on Figure 4 for one prior realisation and each calibrated model. This period constitutes the timeframe of the short calibration dataset and the first three years of the long calibration datasets.
Figure 3: Hydraulic conductivity fields (m/day) for one prior realisation (a) randomly generated from a uniform distribution with $\mu = 0.26$ and $0.08 < x < 0.8$. The base calibrated model hydraulic conductivity fields are shown for the short-sparse dataset (b); long-sparse (c) and long-extensive (d). Model structural error can be seen where the hydraulic conductivity is inappropriately high in the posterior fields (b, c, d) because it is compensating for a static recharge and evapotranspiration. See Fig 1 for the reality field. Domestic wells (orange triangles), sparse monitoring wells (yellow circles), wetland (blue circle) and pumping wells (black circles).
Figure 4: Drawdown at monitoring well one (top) and monitoring well two (bottom). Reality model observation data is shown in blue. The best-fit calibrated model results using short-sparse dataset (orange); long-sparse dataset (red); long-extensive dataset (green); and one prior realization (purple). The modelled results from the long-sparse dataset has the best fit to the observation data and the modelled results using prior knowledge has the poorest fit to data.

The long-sparse calibration shown in red in Figure 4, has the best fit to the observed data at both monitoring wells. Due to the simplification of recharge as static in time, the models do not capture the time varying recharge events that occur in the reality model (blue line). These recharge events can be seen in the stepwise increases, where drawdown becomes negative due to recharge influx. The calibrated models captured the seasonal declines due to the time-varying extractions over the summer months but missed the recharge events that occurred during the winter months (smaller peaks).
The calibrated (a, c, f) and post NSMC analysis (b, d, f) hydraulic conductivity fields for each of the three datasets are shown on Figure 5. Projection of the random prior field onto the null-space of the calibrated models (base calibrated models) can be observed by comparing Figure 5 b, d and f with a, c and e respectively. This shows random parameter realisations from the prior distribution in regions of the parameter space that are insensitive to the observation data (the null space).
Figure 5: Pre and Post NSMC analysis hydraulic conductivity fields for short-sparse (a & b); Long-sparse (c & d) and Long-extensive (e & f) datasets. Prior projection on the null space can be seen on b, d and f, where random realizations of the prior have been projected onto the insensitive parameter space and allows exploration of parameter uncertainty compared to the base calibrated models. Domestic wells (orange triangles), sparse monitoring wells (yellow circles), wetland (blue circle) and pumping wells (black circles).
### 6.3.2 Management Plan Effectiveness

Similar to White et al. (2019), the management plan was considered to have failed if groundwater levels at domestic well one decreased below the base of the well. Unlike White et al. (2019), where three locations (two domestic wells and one wetland) were used as plan objectives, only one location was used in this study for simplicity and clarity. Water levels were compared to the threshold levels and the number of days of plan failure for each realisation was determined over the simulation period.

In the reality system the plan failed 2.5% of days under monthly management. The failure rate increased under entitlement only and unmanaged scenarios to 7.5% and 8.1% of days respectively (Table 3). The reality plan effectiveness was compared to the distributions of plan effectiveness predicted by the prior and posterior models. The frequency with which groundwater level at domestic well one declined below dry depths and the plan failed for the reality, prior, base calibrated models and each of the base models post NSMC analysis is presented in Table 3, with distributions shown on Figures 6 – 9.
<table>
<thead>
<tr>
<th>Management Frequency</th>
<th>Predicted Days of Plan Failure (% of days in simulated period that the plan failed)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reality</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Monthly</td>
<td>2.5</td>
</tr>
<tr>
<td>Monthly (mean)</td>
<td>2.5</td>
</tr>
<tr>
<td>Entitlement Only</td>
<td>7.5</td>
</tr>
<tr>
<td>Entitlement Only (mean)</td>
<td>7.5</td>
</tr>
<tr>
<td>Unmanaged</td>
<td>8.1</td>
</tr>
<tr>
<td>Unmanaged (mean)</td>
<td>8.1</td>
</tr>
</tbody>
</table>

Table 3: Percentage of days domestic well one dewatered at the plan failed for reality, prior, base calibrated models and each NSMC parameter realisation.
When the plan failure rate was predicted using 50 parameter realisations generated with prior information, the plan failure rate varied from 0 - 22% for monthly management and 0 - 40% under entitlement only and unmanaged scenarios. A histogram of the prior predictive uncertainty is shown on Figure 6 and the reality percentage of failure days lay within the distribution of failure predicted by the prior models, although always in the upper tercile of the distribution. The mode of the distribution indicated a zero failure rate. The monthly management distribution was narrower than the entitlement-only and unmanaged scenarios which were highly variable. However, the predictive variability precludes a clear determination of whether the plan could prevent domestic well one from dewatering under monthly management.

![Figure 6](image)

Figure 6: Predictions of plan effectiveness made with fifty realisations of the prior distribution for monthly (top), entitlement-only (middle) and unmanaged (bottom) management scenarios are shown in blue. The reality failure rate is shown under each management scenario as a red dotted line. The range of predictions for each management scenario under prior information is quite variable and the mean is higher than the reality failure rate.

The posterior distribution was defined under three calibration datasets and each is discussed separately below. Short-sparse realisations showed occasional plan failures. The predictive
uncertainty was quite wide for each management frequency, varying from 0 – 15 % for monthly and 0 – 40 % for entitlement-only and unmanaged scenarios (Table 3). Reality plan effectiveness was within the bounds of the effectiveness predicted using the short-sparse posterior distribution. Compared to the prior realisations, calibration to three years of data from two monitoring wells reduced the predictive uncertainty bounds for monthly management by approximately 30 % (from 0 – 22% to 0 - 15%) (Figure 7). Like in the prior case, the mode of these distributions is also at zero failure.

The posterior realisations produced by calibration and NSMC analysis to the long sparse dataset from two monitoring wells resulted in a plan failure rate of 0 for all management scenarios and realisations (Figure 8), meaning after calibration, all 50 realisations reported a failure rate of zero. In reality, the plan failed 2.5 % of days under monthly management and so the calibrated model did not capture these days of failure (Figure 8). This was because the particular combination of parameters attained during the calibration resulted in simulation of hydraulic heads that were slightly higher than the reality heads and extraction related drawdowns did not cause heads to decline below threshold levels. The increase in available head was due to a combination of greater recharge values, lower evaporation values and a greater connection to boundary conditions where water could be accessed. Figures 5 c and d show the expansion of a low conductivity zone between the extraction wells and domestic well 1 that occurred during the NSMC analysis and effectively insulated the domestic well from the impact of extractions.

The long-sparse posterior models under-predicted the drawdown, and consequently, under-predicted the plan failure rate. The prediction is consistent (high precision), but it is inaccurate. Figure 8 shows that the reality plan effectiveness does not lie within the uncertainty bounds generated by using long-sparse realisations because the distribution is degenerate (only one possible value for a variable to assume) (Strang et al. 1993).
Predictions of monthly failure rate produced by calibration and NSMC analysis to the long extensive dataset from twenty monitoring wells varied from 6 – 20% of days under monthly management and 21-40% of days in the entitlement-only and unmanaged (figure 9). The predictive uncertainty generated using realisations from the long-extensive posterior distribution narrowed but the prediction was inaccurate. None of the true reality failure rates of 2.5; 7.5 and 8.1% of days in the simulated period were within the predicted bounds (Figure 9). Calibration to twenty-five years of data from two monitoring wells increased the precision of predictions of plan effectiveness but decreased the accuracy in comparison to the prior and short-sparse dataset. The long-extensive posterior models over-predicted the reality failure rate.

**Figure 7:** Predictions of plan effectiveness made after NSMC methods were applied to the model calibrated with short-sparse (three years sparse). Monthly (top), entitlement-only (middle) and unmanaged (bottom) management scenarios are shown in blue. The reality failure rate is shown under each scenario as a red dotted line. The base calibrated model failure rate predicted with the pre-NSMC calibrated model is shown as a green line. The range of predictions for each management scenario was less variable than the prior, but still quite variable. The true reality effectiveness lay within the distribution of predicted effectiveness and the calibrated model prediction was significantly higher than the reality and lay outside of the predictive uncertainty bounds of the post NSMC model realisations.
Figure 8: Predictions of plan effectiveness made after NSMC methods were applied to the model calibrated with long-sparse (twenty-five years sparse) for monthly (top), entitlement-only (middle) and unmanaged (bottom) scenarios are shown in blue. The reality failure rate is shown under each management scenario as a red dotted line. The base calibrated model failure rate is shown as a green line. The base calibrated model over-predicted the failure rate and each NSMC management scenario under-predicted the plan failure rate and there were zero plan failures out of the fifty realisations in this scenario. The true reality effectiveness does not lie within the predicted distribution of effectiveness.

Figure 9: Predictions of plan effectiveness made after NSMC methods were applied to the model calibrated with long-extensive (twenty-five years extensive) for monthly (top), entitlement-only (middle) and unmanaged (bottom) management scenarios are shown in blue. The reality failure rate is shown under each scenario as a red dotted line. The base calibrated model failure rate is shown as a green line. The base calibrated model over-predicted the failure rate and was further from the true value than the NSMC realisations. Each NSMC management scenario over-predicts the plan failure rate. The true reality effectiveness does not lie within the predicted distribution of effectiveness.
6.4 Discussion and Conclusions

The data requirements for successfully evaluating management plan effectiveness with a numerical model were examined by conducting a calibration constrained predictive uncertainty analysis using NSMC methods that are considered best practice methods (Middlemis et al. 2019, Barnett et al. 2012). This process involved making predictions using groups of model realisations conditioned to four different, increasing levels of system observation data (prior-based on expert knowledge and three different calibration datasets – short-sparse, long-sparse and long-extensive). Increasing the level of system observation was unable to improve the prediction of the plan failure rate. In fact, in some cases, the calibration process made model predictions worse at observation wells and less able to accurately predict failure rate than the prior models.

Model parameter simplifications, where time-variant recharge was represented as constant and parameters were defined at a coarser spatial resolution than in the reality system, were required during the calibration process and caused the parameter values to assume inappropriately high hydraulic conductivity values (approximately 0.8 m/day in the base calibrated models compared to the reality model that had a mean hydraulic conductivity value of 0.26 m/day) to compensate for the simplifications (Table 2). As a result, parameter values took on the maximum possible value (within the bounds defined by the prior parameter distribution) for hydraulic conductivity across large regions of the model domain and shifted away from prior, expert, knowledge despite the imposition of PEST regularisation constraints resulting in erroneous predictions. By contrast, calibration resulted in more precise predictions; however, calibration-induced bias caused calibrated models to make inaccurate predictions. This study shows that currently, despite uncertainty analysis with NSMC methods, management plan effectiveness is unlikely to be accurately determined with a numerical groundwater model with the data availability assumed here, which we feel covers a
range from poorly monitored to very well monitored in terms of current monitoring networks. We acknowledge that our results pertain to a very simple situation and more complex systems would likely compound modelling challenges.

6.4.1 Model Simplifications and Predictive Bias

In the *reality* model, groundwater levels were dominated by time-varying recharge and parameters were defined on the scale of the model grid (80,379 grid cells). But in the *prior* and *posterior* models, simplifications were required to approximate the complexity of the *reality* system with a smaller number of parameters (609) to make parameter estimation tractable. As such, recharge and evapotranspiration were assigned as time-invariant at pilot point locations, and the *reality* grid scale, hydraulic-conductivity heterogeneity was approximated by sparser pilot points with smooth interpolation between points. Treating recharge and evapotranspiration as time invariant is a common model simplification and time-invariant parameterisation is rare (Knowling, et al. 2019). Due to these simplifications, the *prior* and *posterior* models lacked periodic recharge events and ultimately received less recharge and evapotranspiration than the *reality* system. Consequently, during the calibration, the value of recharge could not be adjusted to enable a fit to observed data. Therefore, other parameters were likely adjusted instead during calibration, which in this case, was the hydraulic conductivity (Table 2). By forcing the hydraulic conductivity to assume unrealistically high values to offset the impact of static recharge, bias was introduced into the model. The error due to simplification resulted in “compensatory parameters” where parameters assume incorrect values to fit the data but cause erroneous predictions (Doherty and Christensen 2011). This phenomena is termed “null-space entrainment”, where parameters that belong in the null-space, (i.e. are insensitive to the data), are included in the solution space and altered to take on incorrect, and inappropriate surrogate values that ultimately cause the model to make incorrect predictions (Doherty and Christensen 2011,
Knowling, White, and Moore 2019, White, Doherty, and Hughes 2014). And predictions made with “calibrated” but “wrong” models can be worse than predictions made simply using expert, *prior* knowledge (White, Doherty, and Hughes 2014), which is the case in this study. The “cost” of calibration is potentially, a degradation in predictive accuracy (Moore and Doherty 2006). The contraction in the width of the histograms post calibration-constrained uncertainty analysis (Figures 7, 8 & 9) show that the calibration nominally reduced the parameter uncertainty; however, the shift away from the *reality* value (i.e. increase in mean failure rate from 3.4% using *prior* realisations to 17 % using the NSMC long-extensive realisations) illustrates how calibration made the model “more wrong”. While the *prior* and short-sparse predictive uncertainty bounds were wide (0-22% and 0-15 %), the true *reality* value lay within those bounds. With the addition of calibration data and the erroneous conditioning of parameters it entailed, the true failure rate was not captured by the predictive uncertainty bounds (Figures 7, 8 & 9).

The long-sparse model under-predicted the failure rate of 2.5 % and conveyed a degree of certainty that the plan would not fail under monthly management, that was entirely false and conversely, the long-extensive model over-predicted the failure rate. These results have important implications for management-decision making because plans may be considered to be working when they are not (long-sparse) or vice versa (long-extensive). Considering that the mode of each of the management scenarios for the prior, short-sparse and long-sparse distributions was zero, an examination of the confidence intervals of predictions is vital for informed management as use of the mode, in this case, implies a misplaced confidence that failure will not occur. As stated in section 6.1.1, modelling to support management decisions must overstate the potential for adverse impacts to occur (Doherty and Simmons, 2013), and this requires a shift towards the management viewpoint of attaining confidence that bad things *won’t* happen (Doherty, 2015).
Figures 3b, c and d show high values for hydraulic conductivity along the eastern and western boundaries of the model domain which represent a constant head and river boundary respectively. The artificially high parameter values are due to the model attempting to access water from these boundaries to compensate for the simplified climate forcing (static recharge and evapotranspiration). This bias or model error is not obvious because the fits to observed data are quite good (Figure 4), in particular for the long-sparse dataset, and does not imply a fundamental problem with the model. However, simplification of complex systems results in bias that can be compounded by the calibration process (White, Doherty, and Hughes 2014). The degree of calibration induced bias is prediction unique and depends on the number and configuration of parameters, observation network configuration, data processing methods and the fit to data obtained during calibration (White, Doherty, and Hughes 2014). The NSMC method introduced variability to the parameter fields and figure 5 shows that NSMC posterior realisations are less biased than the base calibrated models due to the impact of the prior field projection onto the null-space. The prior hydraulic conductivity fields most resemble the reality model (Figures 1 & 3) and the calibrated model fields do not look similar to the reality field. Despite having initial parameter values set at the prior (using a level of prior knowledge that would be unlikely in all but a synthetic study), the calibration was unsuccessful at producing a reliable model.

In the reality system, domestic well one was situated in a low hydraulic conductivity zone (Figure 1) that somewhat insulated it from the impact of extractions by providing a buffer between the extraction wells. However, in the calibrated fields, the increased values of hydraulic conductivity between the domestic well and the extraction wells caused simulated extraction-induced drawdown cones to propagate faster and further (Figure 3). This resulted in more frequent predictions of failure in the calibrated scenarios. Additionally, parameterisation is coarser in the prior and posterior models than the true reality model.
because attempting to capture small scale parameter variations would make the model too complex (Figures 1 & 3) resulting in too many adjustable parameters for the calibration/NSMC analysis to be computationally feasible. Yet, if the system detail that influences uncertainty is omitted from models it can lead to an underestimation of uncertainty around the predictions of management interest (Doherty and Christensen 2011). Nonetheless, some degree of parameter simplification is always required.

Such simplifications are often achieved in practices using a “zone-based parameterisation” to represent the model domain, where heterogeneity is simplified by limiting the values parameters can assume to a finite number of values to represent “zones”. Zone based parameterisation smooths out the heterogeneity and coarsens the resolution that parameters are defined upon. When the parameter resolution is too coarse to fully represent the system, the model will likely make incorrect predictions (Moore and Doherty 2006) even when it is considered to be “well calibrated”. Christensen and Doherty (2008) state that even pilot point parameterisation can cause significant structural error depending on the prediction of interest. The predictions made using the base calibration models overpredicted the reality failure rate of 2.5, 7.5 and 8.1 and with the exception of entitlement-only and unmanaged scenarios in the long-extensive dataset, base calibrated model predictions did not fall within the posterior uncertainty bounds (Figures 7, 8 & 9).

Calibration-induced predictive bias has been observed and reported in numerous other groundwater modelling studies (Knowling, White, and Moore 2019, Moore and Doherty 2005, 2006, Doherty and Christensen 2011, Doherty and Welter 2010, White, Doherty, and Hughes 2014). Knowling et al (2019), used a Bayesian approach with a synthetic model to investigate the impact of parameterisation in risk-based decision making. In their study, the impact of groundwater extractions on streamflow at three different parameter resolutions, fine, intermediate and coarse was simulated. Outputs simulated using the coarse parameter set
under-predicted the true number of days predicted stream-flows fell below a specified threshold mirroring the results observed in this study where the number of failure days was underpredicted by the long-sparse calibrated model (Table 3).

White and Doherty (2014) describe a 1-D MODFLOW (Harbaugh 2005) model with 10 model cells arranged in a single row that was designated as reality and used to generate observation data. A structural error was introduced into the reality model and the resulting “defective” model was calibrated to the reality observation data and resulted in incorrect values for hydraulic conductivity. The predictions made using the defective model were more incorrect than predictions made using prior knowledge (White, Doherty, and Hughes 2014). In another example, Moore and Doherty (2005), calibrated a simple contaminant transport model and found that the predictions of travel time were compromised due to parameter simplifications required during calibration. The degree of calibration-induced bias is prediction specific (White, Doherty, and Hughes 2014) and indicates that careful consideration of required predictions and simplifications is critical during the model calibration process.

Models are often applied for unknown future climates and just because the model matches the observation that does not guarantee that it can simulate future, unknown climates (Oreskes, Shrader-Frechette, and Belitz 1994). To examine the impact of climate simplifications on the plan effectiveness and assess the degree of calibration-induced bias, the long-sparse posterior distribution was run under the true, reality system climate. The frequency of plan failures under both the static climate scenario and the variable reality climate scenario are shown on Figure 10 and it can be seen that the predictive uncertainty shifted closer to the reality system failure rate under the true recharge and evapotranspiration conditions, but, the prediction was still wrong and did not include the actual reality failure rate within bounds. This clearly shows the presence of calibration-induced model bias and even when the reality system
climate is perfectly known, the model is unable to accurately predict the failure rate because it has been conditioned to be wrong. When the prior realisations were simulated under a variable reality climate and compared to the long-sparse posterior, the effectiveness was slightly further from the reality value than the effectiveness determined with the long-extensive posterior, but the difference was not significant.

Figure 10: Predictions of plan effectiveness simulated with NSMC long-extensive realisations under the variable reality system climate for monthly management, entitlement-only and unmanaged scenarios. The reality system failure rate is shown under each management scenario as a red dotted line, while the base calibrated model failure rate is shown as a green dotted line. When the reality climate is used, the predictive uncertainty of plan effectiveness shifts closer to the reality value but is still inaccurate due to parameter simplifications between the reality model and the posterior distribution. This is an example of model parameter and structural error resulting in erroneous model predictions.

6.4.2 Implications and Conclusions

Recognition and acknowledgement that modelled outputs will contain bias (Walker 2017) is required when testing management objectives. Therefore, use of models to predict management outcomes should be conducted with careful consideration of the management
prediction and the potential sources of uncertainty. Predictions that are informed by historical data tend to be similar in type and location to the historical data. But, if the model prediction occurs under different conditions or is related to a differing aspect of system behaviour to the historical dataset, then there may be high uncertainty in those predictions (Doherty 2015). Hunt and Werner (2010), state that the presence of “unknown unknowns” or structural uncertainty precludes the neat demarcation of error bands around model predictions but, that such predictions still provide value and can be used to evaluate differences in predictive uncertainty between various scenarios. A benefit of linear analysis is that once the uncertainty around a prediction has been calculated, the change (or differences) in predictive uncertainty due to increased data can also be determined (Doherty 2015). An evaluation of differences between modelled outputs, such as potential management scenarios, has less systematic bias than predictions that depend upon small-scale system detail (Hunt 2017). Consequently, comparing the reduction in predictive uncertainty between particular management actions is an important usage of models (Hunt and Welter 2010). The results of uncertainty analysis comparing differing management scenarios are more representative that a calculation of the outcomes of individual scenarios (White, Doherty, and Hughes 2014).

Sepulveda and Doherty (2015), showed that a model’s capacity to predict changes in groundwater levels due to extractions is greater than the ability to predict absolute values. While the absolute effectiveness of plans cannot be determined, evaluation of system behaviour with and without the plan could provide an indication of whether a plan is likely to have a measurable impact on the aquifer. For example, it can be seen on Figure 10 that the predicted days of plan failure reduces between unmanaged and monthly management frequency, ergo, management is doing something. Indeed, the estimation of predictive uncertainty has been used to design monitoring well networks (Fienen et al. 2010) and could also be used to guide management plan development. Because calibration induced bias is
very prediction specific (White, Doherty, and Hughes 2014), careful consideration of plan
objective is necessary when evaluating plans.

The use of observation data in the history matching process should be approached with
cautions to minimise the potential for bias to be introduced (Knowling and Werner 2016).
Parameter simplification must be conducted with consideration of the prediction and sensitive
parameters (Knowling, White, and Moore 2019). Recently developed Iterative Ensemble
Smother Methods (White 2018) that allow for extremely high numbers of parameters may
be a way to mitigate the effect of calibration-induced bias and allow for the methods
developed in this study to be more successfully applied in future studies. Statistical methods
and various types of observation data (not just groundwater levels) should be used during
calibration to help constrain models (Voss 2011). However, there are calls that a greater
emphasis should be put on prior expert knowledge to minimise the problem of calibration-
induced bias (Voss 2011).

This synthetic example is limited by the usage of only one climate realisation, and multiple
climate scenarios were not explored. Future studies could expand this work to a series of
stochastic climate realisations to explore climate variability. Three calibration datasets were
used and additional datasets or alternate monitoring and extraction well configurations may
produce different results and could be investigated. Fifty NSMC realisations were generated
for each of the three base calibrated models. These posterior distributions could be better
defined with additional realisations; however, the number of realisations was sufficient for
proof of concept. While beyond the scope of this study, further studies could utilise greater
numbers of NSMC realisations.
6.4.3 Major Findings and Conclusions

This study quantified the uncertainty around predictions of management plan effectiveness and evaluated the data requirements of assessing plan effectiveness with a numerical groundwater model. The purpose was to determine whether management plan effectiveness could be predicted precisely and accurately by reducing model parameter uncertainty using recommended best practice techniques such as NSMC methods. The prior predictive uncertainty bounds around plan effectiveness were wide, however, the plan effectiveness predicted by the reality system was contained within the bounds. However, after calibration to three synthetic datasets of increasing length and extent generated by the reality system, parameter uncertainty decreased and the posterior distributions were narrower than the prior. But, despite obtaining a good fit to calibration data, calibration-induced bias caused by model simplifications resulted in the predictive uncertainty to shift away from the reality and become “more wrong”. Predictions made with the posterior distributions obtained using the two long calibration datasets resulted in, more precise, but more inaccurate predictions than those made using the prior. The reality failure rate did not fall within the bound of these posterior distributions. The bias in the posterior models was due to the fact that complex systems must be simplified to enable them to be modelled and systemic model issues remained after the rigorous calibration and uncertainty analysis. One widely used simplification adopted here was that time-varying climate forcing (recharge) was simplified as constant in the model which is a common simplification in many groundwater models. This resulted in other parameters taking on a compensatory role and assuming inappropriate values in order to achieve a fit to the observation data during calibration. In this study, hydraulic conductivity assumed erroneously high values of 0.8 m/day when the mean of the prior was set to 0.26 m/day and caused the model to make biased predictions.
Even using industry standard groundwater modelling and best practice uncertainty techniques (NSMC methods), the predictive uncertainty of plan effectiveness is too great to calculate plan failure rate using this method with a numerical groundwater model and calibration-induced predictive bias caused predictions to be misleading. Calibration-induced model bias is an important issue because it imparts undue confidence in predictions (i.e. the plan will not fail, when it does indeed fail), and leaves decision-makers blind to the true risk. Calibration should be conducted with the upmost consideration of potential bias and the importance of particular inputs and system dynamics.

This study suggests that the effectiveness of plans cannot be determined accurately at this point even by modelling management as a systems control loop and that the use of numerical models to manage groundwater has numerous challenges. However, the quantification of the predictive uncertainty of differences between the effectiveness of various plans or management action can provide useful insights such as informing monitoring network design and determining if system behaviour is sensitive to proposed management. Currently, inadequate observation datasets prevent the model under study from being informed enough to accurately predict management plan effectiveness. Evaluating the change in predictive uncertainty of plan effectiveness could be a way forward for management modelling. Because despite the challenges of making decisions when uncertainty bounds are wider than is perhaps ideal, as eloquently stated by Andre Journel “… it is preferable to have a model of uncertainty than an illusion of reality” (Yarus and Chambers 1994). This sentiment is particularly pertinent in a management context where decisions can have significant and long reaching consequences.
Appendix One: Groundwater Management Plan outlining objective, extraction restrictions and triggers (White et al. 2019).

Objective
The objective of the plan is to maintain the following groundwater levels:
- Groundwater levels must remain 70.15 m above Australian height datum (AHD) at the wetland centre.
- Groundwater levels must remain 73.76 m above AHD at domestic well one and 71.56 m above AHD at domestic well two.

Aquifer Use and Characterization
The aerial extent of the aquifer is 2.5 by 2.5 km and vertical depth is 80 m. Management area comprises the entire model domain (Figure 1). There are three licensed water users extracting water from the aquifer for irrigation using twelve wells. Entitlement volumes for each water user are shown below.

<table>
<thead>
<tr>
<th>Zone</th>
<th>Irrigation Area 1</th>
<th>Irrigation Area 3</th>
<th>Irrigation Area 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entitlement (m$^3$/day)</td>
<td>896</td>
<td>918</td>
<td>924</td>
</tr>
</tbody>
</table>

There are two domestic wells utilising the aquifer (DS1, DS2) with an estimated usage of 5.5 m$^3$/day.

Monitoring Wells
There are two monitoring wells in the aquifer. Monitoring well one was situated to monitor groundwater levels near the domestic wells to ensure levels do not decline below dry depths. Monitoring well two was situated to monitor groundwater levels near sensitive GDE wetland in the south west of the management area. The purpose of monitoring well two was to ensure that groundwater levels do not decline below the critical threshold at the GDE.

Method of Control – Extraction Restrictions
Threshold groundwater levels in the two monitoring wells were established as trigger levels for restrictions upon pumping well extractions. If trigger levels at monitoring well one or monitoring well two are reached, restrictions will be implemented.

<table>
<thead>
<tr>
<th>Trigger</th>
<th>Pumping Rate Cut %</th>
<th>Head (m above AHD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monitoring well one</td>
<td>Monitoring well two</td>
</tr>
<tr>
<td>Trigger 1</td>
<td>25</td>
<td>72.20</td>
</tr>
<tr>
<td>Trigger 2</td>
<td>50</td>
<td>71.50</td>
</tr>
<tr>
<td>Trigger 3</td>
<td>75</td>
<td>70.70</td>
</tr>
<tr>
<td>Trigger 4</td>
<td>100</td>
<td>70.00</td>
</tr>
</tbody>
</table>

Data Review and Analysis
Groundwater level measurements at the monitoring wells will be reviewed and compared to trigger levels on a frequency varying from daily to decadal (management period).

Driver Monitoring
Groundwater extractions are monitored at each pumping well. Climate is monitored at a nearby weather station

Success Measures
The success measure was the maintenance of groundwater levels at the wetland and domestic wells
- Domestic wells: If levels in the two domestic wells decline below the depth of the well (each well is 7.5 m deep), water users are unable to access water and the plan is considered to have failed.
- Wetland: If groundwater levels at the centre of the wetland decline below dry depth, the plan is considered to have failed.
6.6 Appendix Two
Appendix 2: The Act of Management Modelling Loop (from White et al. 2019). The inner black loop cycled daily updating starting heads and pumping rates if well dry depths were reached. When the management cycle occurred (shown in red) groundwater levels in monitoring wells were compared to triggers in plan and restrictions implemented for the subsequent management period. In the synthetic scenario, when management was implemented, the cycle iterated monthly.
6.7 References


Chapter 7: Major Findings and Conclusions

An exhaustive literature review highlighted that despite the importance of effective groundwater management, studies quantitatively testing plans for effectiveness are virtually non-existent and, subsequent analysis showed that in many cases; plans are not conducive to quantitative evaluation. That management improves environmental outcomes is a pervasive assumption in hydrogeology that is frequently made without being quantitatively substantiated. A review of groundwater management literature (chapters 1, 2, 3) identified a lack of studies either questioning whether plans could be assessed or conducting an evaluation of the effectiveness of groundwater management plans. Management modelling investigations have generally evaluated the impact of a particular pumping regime or focused on optimising management strategies. The absence of a comprehensive framework to evaluate groundwater management plan effectiveness was identified as a key research gap and the purpose of this thesis was to develop such a framework.

After establishing the requirements necessary for plans to be amendable to quantitative analysis and applying this to fifteen exiting management plans of varying breadth and scope in chapter 3, a modelling method that determined how well a plan achieves stated objectives was developed and demonstrated in chapter 4. Finally, the feasibility of applying the evaluation methodology to
complex, realistic sites using a numerical groundwater model, given model parameter uncertainty was explored in chapter 6, using methods outlined in chapter 5. By contrast, the focus of chapter 7 is to detail the main findings, conclusions and to reflect upon the rigour with which the research questions were addressed in the course of this research. Additionally, the limitations and challenges encountered are described, and future research directions expanding upon this work are recommended.

7.1 Main Findings and Conclusions

As detailed in chapter 2, this thesis was structured in three parts in order to address the following research questions:

1. What components do plans require to be quantitatively assessed for effectiveness?
2. How can the effectiveness of groundwater management plans towards achieving stated objectives be quantified?
3. Given model parameter uncertainty, how feasible is evaluating management plan effectiveness using numerical groundwater models?

The first research question (What components do plans require to be quantitatively assessed for effectiveness?) was addressed in chapter 3 where groundwater management was structured as a system control problem so that a novel testability assessment rubric could be developed. The rubric was developed within the context of the Australian groundwater management industry, and while use of the rubric is not limited to Australia, it was applied to 15 Australian groundwater management plans and approximately 47% were found to be testable. Testability was found to be an important aspect of groundwater management planning but does not equate to effectiveness i.e. just because a plan is testable does not necessarily mean it is effective.
Considering that testability is a critical prerequisite for quantitative analysis, plan development must be driven so that plans are conducive to future methods of assessment. It is through the coupled assessment and development process that groundwater management can be systematically improved. The main findings that emerged from researching the first research question were:

- Management plans are very diverse and there is no consistent developmental guide for effectiveness evaluations.
- Plans that contain all the components of a functioning control loop are guaranteed to be testable.
- A SMART (specific, measurable, achievable, realistic, timely) objective or acceptable level of impact is required for plans to be testable.
- Plans should be developed to be testable so that evaluations of effectiveness can be conducted.

The components required to ensure plan testability were identified by structuring groundwater management as a system control problem and developing the plan testability assessment rubric. Consequently, research question one was fully addressed.

To address the second research question (How can the effectiveness of groundwater management plans towards achieving stated objectives be quantified?), several sub-questions were considered in chapter 4:

- How effective is the sequential decision-making process of management at maintaining groundwater levels at domestic wells and a wetland compared to no management?
• To what degree does altering the decision-making period improve the effectiveness of management?

• What is the decline in irrigation supply reliability required to achieve environmental and domestic supply goals?

Toward answering the four sub-questions, a methodology to quantitatively evaluate the effectiveness of groundwater management plans by simulating the sequential decision-making process implicit in environmental management was developed and described in chapter 4. The sequential decision-making process of management, informed by aquifer/management feedback, was simulated at varying degrees of decision-making frequency, and it was determined that the effectiveness depended on when (timing during the year) and how often management decisions were made. When management decisions were made in November (prior to the irrigation season), as opposed to January (during the irrigation season) the management plan provided no additional benefit than simply setting an entitlement limit. These results have important implications for management, where decisions are often governed by budgetary constraints. The presented methodology could be used to guide management plan development by indicating potential outcomes and evaluating the trade-offs between environmental and water supply goals.

The modelling method where the sequential decision-making process of management was modelled provides an indication of how a plan may proceed under alternate timings and frequencies of management decisions and in systems with differing response times. A synthetic example quantified the impact of a generic plan, specifying environmental objectives, extraction restrictions and entitlement limits, relative to no-management by combining a numerical model of reality with management rules under stochastic climate replicates. A simple synthetic model was used to develop and demonstrate that the effectiveness of groundwater management plans
could be quantitatively evaluated by simulating sequential management decisions that evolve based on aquifer/management feedback. The management decision-making frequency varied from daily to decadal. This modelling approach allowed the effectiveness of a management plan to be evaluated within a system control framework. In the presented example, more frequent management improved environmental outcomes and indicated that implementation of restrictions at the wrong time of year provided no greater improvement to plan effectiveness than setting an entitlement limit. When management decisions were made prior to the irrigation season, quarterly management was less effective than annual and biannual management due to an interaction between the decision-making timing and the timing of the irrigation season.

Management decisions concerning allocation volumes are frequently made based on the groundwater levels in the winter during the period of recovery of groundwater levels and the synthetic example implied that management decisions may be more effective when made during the irrigation period so that groundwater level decline may be controlled with restrictions when required. When management decisions were made on annual or longer periods in the example, as frequently occurs in practice, plan effectiveness was shown to be highly uncertain. Additional factors not considered in this study, such as non-compliance, could further decrease effectiveness in the considered scenarios. Model simplicity was chosen over site specific complexity to keep the results generalisable and avoid concentrating on case-specific idiosyncrasies. Management effectiveness was found to be highly dependent upon hydraulic conductivity and system detail such as well configuration and model geometry.

While a range of hydraulic conductivity values (0.026 m/d – 1.43 m/d) were used to approximate different system response times and resulted in six aquifer replicates, management intervention for this particular model geometry only had a measurable impact on objective failure rate and
reliability in three of the aquifer system replicates (i.e. those with hydraulic conductivity values between 0.143 m/d and 0.56 m/d). Management had the greatest impact between hydraulic conductivity values of 0.143 and 0.56 m/s and outside of this “Goldilocks” range, the impact of management was negligible. Aquifers responding on slow timeframes (low hydraulic conductivity) showed low reliability and could not meet irrigation demand despite management intervention. Whereas, fast responding aquifers displayed high reliability and easily meet irrigation demand, making management intervention redundant. Management had an impact when irrigation demand was generally balanced with the aquifer’s capacity to provide, which is, realistically, the type of aquifer that a small-scale irrigation operation would be installed within and managed under a plan of this nature. Irrigation at an inappropriate scale would not be pursued in an aquifer that consistently failed to meet demand, nor would extraction restrictions be required when demanded volume was easily supplied.

Because the study was a synthetic study, a higher level of system detail was known, whereas in reality, the uncertainty would be much higher and could further cloud the impact, or lack thereof, of management action upon the system state. However, the method is easily extended to complex groundwater systems and management actions, provided they can be modelled adequately. The plan testing methodology can be tailored to answer site-specific questions with more complex models. However, the method was sensitive to model parameterization (hydraulic conductivity) which raised the question of how well the aquifer properties need to be known and led directly to the calibration-constrained predictive uncertainty analysis. The method enables quantitative evaluation of the effectiveness of alternative management scenarios and decision-making frequencies and, allows a qualitative assessment of trade-offs between conflicting uses. By
developing and demonstrating a potential method to quantify the effectiveness of groundwater managements plans, research question two was fully addressed.

To address the third research question (*Given model parameter uncertainty, how feasible is evaluating management plan effectiveness using numerical groundwater models?*) in chapter 6, a calibration-constrained predictive uncertainty analysis was conducted to assess the feasibility of using the methodology developed in chapter 4 to evaluate plan effectiveness in a complex, heterogenous and highly uncertain system. The purpose was to investigate whether decreasing parameter uncertainty through calibration to increasingly extensive datasets, would result in a commensurate decrease in the uncertainty of predictions of plan effectiveness. A key question was determining how much calibration data was required to constrain model predictions of plan effectiveness so that predictions were accurate and precise. However, despite obtaining a good fit to calibration data, calibration-induced bias caused the predictive uncertainty to shift away from the *reality* and become “more wrong”. Predictions made with the *posterior* distributions obtained using the two long calibration datasets resulted in, more precise, but more inaccurate predictions than those made using the *prior*. Due to simplifications of *reality* in the simple model, the calibration process caused the model to become biased and make incorrect predictions, either over-predicting or under-predicting the plan failure rate. This has important implications and can cause misplaced confidence that a plan will not fail when it does in fact periodically fail, and can leave decision-makers blind to the true risk. Alternatively, an over-prediction of plan failure rate could result in over-restrictive extraction limits that are not in fact required. Potential biases and the importance of system inputs and dynamics should be considered during the calibration process. While the effectiveness of plans cannot be determined accurately at this point, quantification of the predictive uncertainty of differences between the effectiveness of various
plans or management action can provide useful insights and potentially guide management plan development. The implications of calibration-induced model bias are wide and highlight the challenges and difficulties of modelling management. Calibration-induced bias is prediction specific so the desired predictions and available datasets must be well thought out when assessing plans. Even with use of rigorous uncertainty analysis methods, the effectiveness of management could not be determined due to the limitations of numerical models which raised serious question over our ability to model management. This study investigated the feasibility of evaluating management plan effectiveness using numerical groundwater models and found that with current best practice uncertainty analysis, groundwater management plan effectiveness cannot be accurately predicted at this time and further studies are required. Following on from all this, research question three was fully addressed.

Currently, it is unclear when management improves environmental outcomes and evaluation of management effectiveness is challenging and rarely conducted. Management plans must be written to be conducive to quantitative analysis by containing SMART objectives or acceptable levels of impact. Furthermore, plans should not be assumed to be effective in the absence of established causality. While it is not yet possible to determine plan effectiveness accurately and precisely using numerical groundwater models, evaluation of the relative impact of plans or management scenarios is important. Potential applications of this research could enable an initial estimation of plan effectiveness without the need to install extensive and costly monitoring wells that may provide similar indications. Importantly, the way management plan objectives are worded needs to shift from aiming to achieve a beneficial outcome to preventing a bad outcome from occurring. While this distinction may seem nuanced, it is important to think of the problem in this manner because models, after all, are imperfect simulators of natural systems and users
must be mindful of their limitations. And when information is limited, as occurs in all environmental modelling scenarios to varying degrees, the precautionary principle (assumption of environmental harm in the absence of scientific information [Fullem, 1995]), should be utilised to shift the focus to the preventing bad things from happening. Models are unable to predict what will happen, but, are able to predict what likely will not happen [Doherty, 2015]. Consequently, management plans should be created with SMART objectives that aim to prevent the occurrence of adverse impacts instead of achieving beneficial outcomes. Furthermore, plans must be testable (White et al. 2016), with objectives that have relatively low predictive uncertainty (Guillaume et al. 2016), so whether or not the objectives are likely to be met can be evaluated. Evaluation of predictive uncertainty is important for robust management modelling because ensuring adverse impacts are statistically unlikely to occur under particular management scenarios is critical for an informed risk assessment (Doherty 2010; Walker 2017). This thesis provides a pathway for quantitative, evidence-based improvement of management planning.

7.2 Challenges and Limitations

The following limitations were encountered:

- While numerical groundwater models are powerful tools, they have many limitations including an inability to represent the true complexity of natural systems, cost and availability to managers. The considerable cost of building numerical groundwater models and computational burden of the testing methodology limits its applicability. Access to high performance computing is required which, with several exceptions, is unusual in management organisations. Various forms of uncertainty limit the applicability of the methodology presented in chapter 4 because it was sensitive to aquifer properties, which in reality, are highly uncertain. The model used in the example
was very simplistic and likely had much model structural error which was beyond the scope of this investigation. There would have been error in the recharge and evapotranspiration and measurement error. Additionally, the compliance of groundwater users to a management plan is a further unknown and unpredictable human behaviour could further complicate plan efficiency.

- A synthetic study was used for the modelling studies in this thesis. The synthetic system used in the study described in chapter 4 was simplistic in the assumption of a homogenous aquifer and results were dependent upon model geometry. Only one model geometry was investigated, and results may differ if alternate well configurations were used. Groundwater extraction were represented with multiple extraction bores to prevent extraction bores from dewatering by spreading the pumping across several wells. However, the scenario of multiple bores in a small area is unlikely and would not occur in reality. Results may differ if one bore was used instead.

- Real management plans or catchments were not studied in this research. The management plan used in this study was an amalgamation of several management plans. Generalisations of management are inferred using synthetic systems and assessment of real plans in real systems would likely result in different outcomes.

- Discerning the impact of management upon aquifer systems was very difficult in a synthetic study and difficulties are compounded in real systems that are highly uncertain.

- The limitations of the predictive uncertainty analysis include the fact that only one climate replicate from one location in western Victoria was used. Additionally, the calibration data sets were somewhat arbitrary and the locations of the two monitoring wells were selected based on the same configuration used in chapter 4 so many of the
same limitations would apply (uncertainty, simple model etc). The irrigation scenario simulated was simplistic and crop type was not accounted for a simple 1-D soil moisture model was used. The configuration of the model domain could impact plan effectiveness and results may differ if different configurations were used.

7.3 Future Research

All the plans that were assessed with the rubric were in Australia and future research could expand the assessment to international plans. This may result in amendments and additions to the rubric as new plan components are encountered and may help identify more effective plan structures. Evaluation of a real plan using a numerical groundwater model of a management area is a logical extension of the study addressing research question two. A study of this nature could provide insights into the outcomes of various management scenarios and assist in guiding the management process. The evidence suggests that use of modelling to manage groundwater is hard and data limitations present barriers to well-informed models. Therefore, improving data collecting and preferentially amassing data that will inform numerical models (with predictions of objectives that are of a similar type to datasets) needs to be a priority. Different data collection techniques could be explored for example, incorporation of remote sensing data into the evaluation process may improve predictions of effectiveness and should be an area of future research.

Further research and groundwater management modelling efforts should focus on quantifying and where possible, reducing predictive uncertainty and using reductions in predictive uncertainty to guide management investment, such as improving monitoring networks. Additional modelling studies that evaluate systems with different model configurations (i.e. number and location of wells, extraction regimes, climate replicates) as well as models used to
inform existing groundwater management plans could yield information on the type and amount of data that best informs the calibration process and ultimately predictions of plan effectiveness. A very pragmatic initiative that could improve computational ease and efficiency would be to develop modules within MODFLOW and/or other models that can directly simulate implementation of management rules. Management modelling has many challenges and limitations, and the identification of such limitations can make sure models are fit for purpose and yielding the best results possible. Further research into methods to manage and limit calibration-induced model bias is warranted. Simulations of plan effectiveness with alternate methods that are yet to be developed is an area of future research.

7.4 Scientific Contribution

The new understanding based on my research includes; plans are important; they need to be tested and testable so that it can be confirmed that they are working as intended; and while, the effectiveness of groundwater management plans can be successfully evaluated by simulating the sequential decision making process of management, model shortcomings such as calibration-induced bias cast doubt over the reliability of predictions made with numerical groundwater models. This study highlighted the importance of rigorous uncertainty analysis when using numerical groundwater models, particularly in a management context. This research contributes to the body of scientific knowledge in the following ways:

- Identified the need to develop plans to be testable and developed a methodology to assess the testability of management plans. The method allows a rapid assessment of plans to isolate any structural shortcomings that would preclude testability. An interpretation of
the testability score enables practitioners to see where plans can be systematically improved.

- A methodology to simulate the sequential decision-making process of management in a groundwater context using a numerical model was developed and demonstrated. This is a novel way to model groundwater management and allows for representation of aquifer/management feedback where the simulated levels dictate management action, as occurs in reality. Modelling a feedback driven system is more realistic than simulation of a pre-determined pumping rate and can provide new insights into aquifer management effectiveness.

- This research demonstrated the impact of parameter uncertainty on the plan effectiveness testing methodology where management was modelled as a feedback-driven sequential decision-making process.

- Finally, this thesis represents the first comprehensive framework for the quantitative evaluation of management plan effectiveness.

The information outlined in this thesis constitutes a first step towards quantitatively evaluating plan effectiveness. This thesis clarifies not only plans that are not being tested but many plans can’t be tested because they are not written with SMART objectives or acceptable levels of impacts and it provides recommendations on how such plans can be improved, including recommendations on how to write plans to be testable.

A method to test plan effectiveness using a model is also presented and the limitations of the method are explored with an uncertainty analysis. The testing methodology can be customised to site-specific complexity and can be used by managers to simulate how well-established plans
work. It can provide insight into issues such as how frequently decisions on management interventions should be made.

Specifically, this thesis constitutes a novel contribution to groundwater management literature for the following reasons:

- This thesis presents the first groundwater management testability assessment rubric.
- This thesis presents the first quantitative evaluation of the sequential decision-making process of groundwater management.
- This thesis presents the first study investigating how much data is required to test management plan effectiveness.

***
Appendix A

Publications

Peer Reviewed Journal Articles


Conference Abstracts


Appendix B

Supplementary Information: Calibration
Results for Chapter 6
Short Sparse Dataset

Monitoring Well One Calibrated Fit: Observed (blue) and Simulated (orange) drawdown

Monitoring Well Two Calibrated Fit: Observed (blue) and Simulated (orange) drawdown
Change in Measurement Objective Function (SSE) Value During Calibration
**Long Sparse Dataset**

Monitoring Well One Calibrated Fit: Observed (blue) and Simulated (orange) drawdown

Monitoring Well Two Calibrated Fit: Observed (blue) and Simulated (orange) drawdown

Change in Measurement Objective Function (SSE) Value During Calibration
**Long Extensive Dataset**

Monitoring Well One Calibrated Fit: Observed (blue) and Simulated (orange) drawdown

Monitoring Well Two Calibrated Fit: Observed (blue) and Simulated (orange) drawdown
Monitoring Well Three Calibrated Fit: Observed (blue) and Simulated (orange) drawdown

Monitoring Well Four Calibrated Fit: Observed (blue) and Simulated (orange) drawdown
Monitoring Well Five Calibrated Fit: Observed (blue) and Simulated (orange) drawdown

Monitoring Well Six Calibrated Fit: Observed (blue) and Simulated (orange) drawdown
Monitoring Well Seven Calibrated Fit: Observed (blue) and Simulated (orange) drawdown

Monitoring Well Eight Calibrated Fit: Observed (blue) and Simulated (orange) drawdown
Monitoring Well Eleven Fit: Observed (blue) and Simulated (orange) drawdown

Monitoring Well Twelve Fit: Observed (blue) and Simulated (orange) drawdown
Monitoring Well Thirteen Fit: Observed (blue) and Simulated (orange) drawdown

Monitoring Well Fourteen Fit: Observed (blue) and Simulated (orange) drawdown
Change in Measurement Objective Function (SSE) Value During Calibration
References


Alley, W. M., and S. A. Leake (2004), The Journey from Safe Yield to Sustainability, Ground Water, 42(1), 12-16.


Australian Commonwealth (1915), River Murray Waters Act, in 46, edited, Australia.


BoM (2015), Recent rainfall, drought and southern Australia’s long term rainfall decline Bureau of Meteorology.

BoM (2019), What is la nina and how does it impact Australia?, edited, Bureau of Meteorology.


Council of Australian Governments (2004), Intergovernmental Agreement on a National Water Initiative between the Commonwealth of Australia and Governments of New South Waters, Victoria, Queensland, South Australia, the Australian Capital Territory and the Northern Territory., edited, Commonwealth of Australia, Canberra.


Department of Natural Resources and Mines (2016), Warrego-Paroo-Nebine Water Resource Plan.


Department of Water (DOW) (2013), Pilbara groundwater allocation plan, edited, Government of Western Australia.


DNREAS (2009), Water Allocation Plan Tindall Limestone Aquifer (Katherine), Department of Natural Resources Environment The Arts and Sport, Katherine.


DOW (2009), Gnangara groundwater areas allocation plan, edited by D. o. Water, Government of Western Australia.


Gardner, A., R. Bartlett, and J. Gray (2009), Water Resources Law, Lexis Nexis Butterworths Chatswood, NSW.


GMW (2009), Mid-Loddon Groundwater Management Area Local Management Rules, edited by Goulburn Murray Water.


Government of Australia (1900), Commonwealth of Australia Constitution Act edited.


Harrington, N., and P. Cook (2014), Groundwater in Australia, National Centre for Groundwater Research and Training, Australia.
Højberg, A. L., and J. C. Refsgaard (2005), Model uncertainty – parameter uncertainty versus conceptual models, Water Science and Technology, 52(6), 177-186.
McIntyre, N. W., C (2011), Developing adaptive groundwater management arrangements in the southeast of South Australia: a pilot study, Department for water, South Australia.


Moggridge, B. J. (2005), Aboriginal People and Groundwater, University of Technology, Sydney.


Murray Darling Basin Authority (2012), Basin plan, Commonwealth of Australia.


Poeter, A. (2007), All models are wrong, how do we know which are useful?, Groundwater, 45(4), 1.


Shapoori, V., T. Peterson, A. Western, and J. Costelloe (2015a), Decomposing groundwater head variations into meteorological and pumping components: a synthetic study, Hydrogeol J, 23(7), 1431-1448.


The Murray Darling basin authority (2012), Basin plan.


Walters, C. J., and C. S. Holling (1990), Large-scale management experiments and learning by doing, Ecology, 71(6), 2060-2068.


Wolf, L. M., C (2011), Management of groundwater in Australia: new models and approaches for adaptive management and large scale reuse, edited by CSIRO.
Now we came out, and once more saw the stars

Dante, Inferno