Rigorous Validation of Hydrologic Models in Support of Decision-Making

by

Robert Chlumsky

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Civil Engineering (Water)

Waterloo, Ontario, Canada, 2017

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Hydrologic models are often relied upon to inform decisions in hydrology and water resources applications. Typically, hydrologic models are validated (i.e., deemed fit-for-purpose) using the split-sample test introduced by Klemeš in 1986, where a model is shown to reproduce historical data that was not used in training the model. However, simple history matching is a necessary but insufficient condition to show reliability in decision-making, and there are many examples in the literature of models with excellent diagnostic metrics but insufficient skill for informing decisions. Furthermore, the current methods of model validation and uncertainty estimation are not easily understood by non-modellers, and decision-makers are often handed model outputs that are not in a readily usable form to inform decision-making. There exist many calls in literature for both the development of rigorous model validation methodologies, and techniques to address the gap between scientists and decision-makers.

As a response to the gap in literature for new model validation methodologies and ways of improving communication between scientists and decision-makers, this thesis introduces Decision Crash Testing (DCT), which is capable of directly evaluating a model's performance in a decision-making context. The DCT framework uses a series of synthetic reality experiments to recreate the model-building process and test whether the model correctly informs the decision in a set of hypothetical realities. These provide a baseline estimate of the difficulty of the decision and the probability that the model is capable of informing the correct decision, which is much more easily interpretable by decision-makers.

The DCT framework is demonstrated using two case studies derived from reservoir management applications in an Ontario watershed. The development of the hydrologic models that underlie both case studies is a deviation from the typical approach of using level-pool routed inflow estimates for model calibration; instead, the observed stage data is calibrated to directly, and the reservoirs are explicitly represented in the hydrologic model. This method avoids some of the pitfalls of calibrating to estimated inflows with known numerical artefacts. The two case studies illustrate the ability of the DCT framework to assess the decision-making ability of the evaluated model, to provide a framework within which to meaningfully assess improvements the evaluated model for specific decision-making applications, and to test the impacts of various decision formulations on the ability of the model to inform decision-making.

ACKNOWLEDGEMENTS

I would first and foremost like to thank my supervisor, Dr. James R. Craig, for his support and inspiration over the last two years. Your intellectual curiosity and dedication to your students are truly remarkable, and you have made my graduate experience thus far quite enjoyable.

I would also like to thank my supervisor’s primary colleague and partner in crime Dr. Bryan Tolson, with whom I have also worked closely with many times, and from whom I have also learned much.

I also wish to extend a thank you to some of the staff at Ontario Power Generation, namely Joan Frain, Mark Nussli, Michael McNiven, and Connor Werstuck, who helped to make this project possible.

My experience would also have been lacking without the continuous help and support of my many peers and colleagues at the University of Waterloo, including those in the Collaborative Water Program, who have broadened my horizons far beyond what I expected when I started the program. I would like to extend a particular thank you to Dr. Juliane Mai, my mentor and friend who helped me many times during my thesis work, as well as Hongli Liu, whom I had the pleasure of working with closely on a number of projects.

Finally, I would like to thank my friends and family, and especially my partner Lauren, for all of their continued love and support.

Table of Contents

List of Figures viii

List of Tables x

Chapter 1 Introduction 1

1.1 Goals and Objectives 2

1.2 Thesis Organization 2

Chapter 2 Background 4

2.1 Hydrologic Modelling 4

2.1.1 What is a hydrologic model? 4

2.1.2 Classifications of Hydrologic Models 4

2.1.3 Model Building Procedure 7

2.1.4 Model Calibration and Validation 8

2.2 Critical Evaluation of the Concept of Validation in Hydrologic Modelling 9

2.2.1 Philosophy of Validation 9

2.2.2 Current Validation Practices in Hydrology 10

2.2.3 Call for Improved Validation of Hydrologic Models 11

2.3 Decision-Making in Hydrology 12

2.3.1 Decision-Making Process 12

2.3.2 Decision-Making Example in Hydrology 14

2.3.3 Methods, Frameworks and Sensitivity Analysis in Decision-Making 15

2.3.4 Uncertainty in Hydrology and Decision-Making 18

2.3.5 Linking Hydrologic Models and Decision-Making 19

2.4 Linking Hydrologic Model Validation to Decision-Making 21

Chapter 3 Methods and Model Development 22

3.1 Decision Crash Testing 22

3.1.1 DCT Methodology 22

3.1.2 DCT Similarity and Decision Models 25

3.1.2.1 DCT Similarity and Decision Models in Non-Binary Decisions 30

3.1.3 DCT Considerations 32

3.1.3.1 Synthetic Reality Generation 32

3.1.3.2 Analysis and Pre-screening of Synthetic Realities 34

3.1.3.3 Convergence of DCT Metrics 35

3.1.4 Applications of DCT 35

3.1.4.1 Direct Evaluation of the Hydrologic Model for Decision-Making 36

3.1.4.2 Model Failure Analysis 37

3.1.4.3 Improvement of the Model-Building Procedure 38

3.1.4.4 Evaluation of the Decision Formulation 39

3.1.5 Analysis of DCT Method Choices 39

3.1.6 Limitations of DCT 40

3.1.7 Comparison of DCT to Other Methods in Literature 41

3.2 Case Study Background 42

3.2.1 Reservoir Management in Canada 42

3.2.2 Study Area and Case Study Introduction 42

3.2.3 Hydrologic Model Structure 44

3.2.4 Data Availability and Discretization 46

3.2.5 Reservoir Representation 50

3.2.6 Model Calibration Procedure 52

3.2.7 Model Results and Use in DCT 55

Chapter 4 Decision Crash Testing in a Reservoir Management Application 66

4.1 Case Study Setup 66

4.1.1 Reservoir Management Decision 66

4.1.2 Decision Quantity Definition 68

4.1.3 Mountain Chute Model Updates 70

4.1.4 Synthetic Reality Generation 72

4.1.5 Model Evaluation Setup 78

4.2 Results and Discussion 79

4.2.1 Follow-up DCT Experiment in Model Improvement 82

4.2.2 Adjusted Decision Formulations for Sharp Binary Decisions 87

4.3 Case Study Conclusions 92

Chapter 5 Decision Crash Testing in a Data Collection Application 94

5.1 Case Study Setup 94

5.1.1 Data Gauging Decision 94

5.1.2 Decision Quantity Definition 96

5.1.3 Synthetic Reality Generation ... 97

5.1.4 Model Evaluation Setup ... 101

5.2 Results and Discussion ... 102

5.3 Case Study Conclusions ... 106

Chapter 6 Conclusions ... 108

6.1 Contributions to Literature ... 109

6.2 Future Opportunities for DCT ... 110

References ... 112

Appendix A Reservoir Model Development - Supporting Materials ... 119

Appendix B DCT Case Study 1 - Supporting Materials ... 123

Appendix C DCT Case Study 2 - Supporting Materials ... 128

List of Figures

Figure 1. Three types of spatial discretization in hydrologic models 6

Figure 2. Model building steps adapted from literature 7

Figure 3. Decision quantity distribution from synthetic reality generations for binary decisions 26

Figure 4. Decision quantity pairs from synthetic reality generations and model outputs 26

Figure 5. An example discrepancy distribution which is just the sample distribution, or histogram, of model skill from a DCT experiment 28

Figure 6. Decision quantity distribution from synthetic reality generation for non-binary decisions . 31

Figure 7. Conceptual diagram of soil profiles in the model 45

Figure 8. Process diagram for the developed hydrologic model 46

Figure 9. Depth to bedrock map from borehole logs in the Madawaska watershed 48

Figure 10. Flow and level data locations in the Madawaska watershed 49

Figure 11. Modelled areas in the Madawaska discretization 50

Figure 12. Truncated cone representation of reservoir geometry 51

Figure 13. Calibrated model results for Bark Lake Dam inflow model 56

Figure 14. Inflow comparison from calibrated model and level pool routing for Bark Lake Dam model 57

Figure 15. Bark Lake model split-sample test results for reservoir stage 57

Figure 16. Bark Lake model split-sample test results in flow comparison 58

Figure 17. Uncorrected level pool routing-estimated inflow series at Bark Lake Dam 59

Figure 18. Comparison of stage-fitted and level pool routing-fitted inflows 60

Figure 19. Comparison of Bark Lake model stage and level pool routing simulated stage 61

Figure 20. Calibrated model results for Mountain Chute model 62

Figure 21. Calibrated model results for Mountain Chute model 62

Figure 22. Comparison of upstream reservoir releases to simulated reservoir inflows at Mountain Chute 63

Figure 23. Mountain Chute model split-sample test results for reservoir stage 64

Figure 24. Mountain Chute model split-sample test results in flow comparison 64

Figure 25. Location of key features in reservoir management DCT case study 67

Figure 26. Synthetic reality generated summer flows at subbasin 48 for base and adjusted reservoir operations 68

Figure 27. Comparison of simulated and observed flows using a seasonal rating curve representation at Kamaniskeg Dam 71
Figure 28. Kamaniskeg Dam stage with seasonal outflow representation 72
Figure 29. Comparison of synthetic reality simulated reservoir stages at Mountain Chute 74
Figure 30. June proportion of flows above threshold for synthetic realities 75
Figure 31. Synthetic reality decision quantity distributions of DQ_1 75
Figure 32. August proportion of flows above threshold for synthetic realities 76
Figure 33. Synthetic reality decision quantity distributions of DQ_2 76
Figure 34. Convergence plot of the AMAMS for DQ_1 78
Figure 35. Decision quantity 1 plot for reservoir management decision 80
Figure 36. Decision quantity 2 plot for reservoir management decision 80
Figure 37. Convergence plot of the similarity score for DQ_1 with 100 trials 82
Figure 38. Decision quantity 1 plot for reservoir management decision in follow-up experiment 84
Figure 39. Decision quantity 2 plot for reservoir management decision in follow-up experiment 84
Figure 40. Bootstrapped distributions of similarity score for each DCT experiment 85
Figure 41. Normality plots for bootstrapped distributions of similarity scores 86
Figure 42. Similarity score as a function of decision quantity threshold for June flows 88
Figure 43. Similarity score calculation for fuzzy boundary decision formulation 89
Figure 44. Decision quantity 1 plot for fuzzy decision formulation 90
Figure 45. Similarity score sensitivity for sharp and fuzzy boundary and fuzzy decision formulations 91
Figure 46. Location of potential flow gauges in data gauging DCT case study 95
Figure 47. Example of random noise added to Bark Lake Dam stage time series 99
Figure 48. Decision quantity distributions for Subcase 1 100
Figure 49. Decision quantity distributions for Subcases 2 and 3 100
Figure 50. Decision quantity 1 plot for Subcase 2 in data gauging decision 104
Figure 51. Decision quantity 2 plot for Subcase 2 in data gauging decision 105
Figure 52. Decision quantity 3 plot for Subcase 2 in data gauging decision 105
Figure B1. Linear regression fit for Winter outflows from Kamaniskeg Dam 124
Figure B2. Linear regression fit for Spring outflows from Kamaniskeg Dam 125
Figure B3. Linear regression fit for Summer outflows from Kamaniskeg Dam 125
Figure B4. Linear regression fit for Autumn outflows from Kamaniskeg Dam 126

List of Tables

Table 1. Contingency table example for binary decisions ... 27

Table 2. Meteorological Stations used in Madawaska models ... 47

Table 3. Key parameters used in synthetic reality generation for reservoir management case study .. 73

Table 4. Synthetic reality summary statistics for reservoir management decision ... 77

Table 5. Summary metrics for DCT results for reservoir management decision ... 79

Table 6. Comparison of DCT metrics for base and follow-up experiments ... 83

Table 7. Summary metrics for synthetic reality generations in data gauging case study ... 101

Table 8. Summary metrics for DCT results in data gauging case study ... 103

Table A1. Parameters calibrated in each step of calibration procedure ... 121

Table A2. Base reservoir parameters for calibrated models ... 122

Table B1. Reservoir parameters for Kamaniskeg Dam ... 123

Table B2. List of parameters used in calibration of evaluated model for reservoir management case study ... 123

Table B3. Fitted piecewise linear regression coefficients for seasonal outflows from Kamaniskeg Dam ... 124

Table B4. Seasonal rating curve for Kamaniskeg Dam ... 127

Table C1. List of parameters used in calibration of evaluated model for data gauging case study ... 128

Chapter 1
Introduction

Models are useful tools in many applications, and can be used to perform experiments and analyses that would not be possible or feasible in the real world. In hydrology, models are used for a number of applications, including flood forecasting, watershed planning, source water protection, and reservoir management. Thus, understanding how the performance of hydrologic models can be evaluated, and building a trust in the models that we use, is of fundamental importance.

The methods by which hydrologic models are evaluated as fit for use in these applications is an open question which has not seen much development, generally speaking, since the 1980s. The existing methods for validation of hydrologic models are long established but inconsistently applied, and even more rarely are the limitations and uncertainties in models effectively communicated to non-modellers. Despite the widespread use of models to inform decision-making, there exists a gap between the information presented by modellers and the information that is useful for decision-makers. There are many reasons for this, including the difficulty in communicating uncertainty, the complexity in understanding the large number of hydrologic models that exist, and the variety of applications in which the models are deployed. Furthermore, surprisingly few methods exist that validate models specifically for decision-making applications, which is a non-trivial matter but would facilitate improved understanding of model outputs between modellers and decision-makers.

This thesis introduces a method called Decision Crash Testing (DCT), which is a framework with which to validate models for use in specific decision-making applications. This method builds on various concepts that exist in the literature, and is inspired by the crash testing analogy of Andréassian et al (2009), which encourages modellers to stress test models to their breaking points to reveal their limitations. This novel method is discussed at length and is illustrated with two reservoir management case studies.

The development of the hydrologic models used to support the reservoir management case studies is also presented in this thesis. The development and calibration of hydrologic models involving reservoirs is a non-trivial task, particularly when there is a lack of data and/or a lack of knowledge of upstream reservoir operations. Typically, reservoir inflow forecasting models are calibrated using inflow estimates rather than measured data, which can cause models to match spurious peaks from the inflow estimates without distinction from real events, among other issues. A novel application of a new method is presented for the calibration of hydrologic reservoir models, which uses only the available data in calibration and avoids some of the issues with existing methods.

1.1 Goals and Objectives

This thesis has two main goals. The primary goal of this thesis is to demonstrate the gap that exists in the literature for hydrologic model validation methods that address the application of models in decision-making, then to present the DCT framework as a potential solution to this gap in literature, and illustrate its use with two detailed case studies. The secondary goal of this thesis is to present a novel method for calibrating hydrologic models involving reservoirs, which overcomes some of the data availability issues with conventional approaches. This thesis includes the first known application of this method to a North American watershed.

This thesis has three main objectives that follow from these goals.

1. To develop, evaluate, and apply a novel method for the development and calibration of a hydrologic model in support of reservoir management, which is deployed in building the case studies in this thesis.
2. To present a full description of the DCT framework, highlighting how it fills the gap that exists in literature, and how it can be improved in future work.
3. To illustrate the efficacy and utility of the DCT framework by applying it to two reservoir management case studies in Ontario.

1.2 Thesis Organization

This thesis is organized into 6 chapters, the first of which serves as this introduction to the thesis.

Chapter 2 provides relevant background information on hydrologic modelling, the current state of hydrologic model validation, and the application of environmental models to decision-making. The need for both improved validation methods and a bridging of the gap between modellers and decision-makers is illustrated, providing the motivation for the DCT framework.

Chapter 3 discusses the methodology relevant to this thesis. This chapter provides a full description of the DCT framework, discussing the methodology, the ways in which DCT can be deployed, the limitations of DCT, and a comparison of DCT to existing methods in literature. This chapter also presents the development of the hydrologic models supporting the case studies in this thesis, which includes a description of the novel method for calibrating reservoir models. The contribution to literature from this method are illustrated in this chapter.

Chapter 4 introduces the site-specific application of DCT to a decision regarding the adjustment of seasonal reservoir operations. The details of the DCT setup, with respect to the modes of deployment in Chapter 3, are discussed. The insights gained from decision crash testing are discussed, and a number of additional experiments in model improvement and alternative decision formulations are demonstrated.

Chapter 5 details the second application of DCT to a data gauging decision to support reservoir management. The results and implications from the DCT experiments are highlighted.

Chapter 6 summarizes the major contributions to literature from this thesis work, namely the novel method for reservoir model calibration, and the further development and application of the DCT framework. The main conclusions of the case studies and directions for future research, including potential future improvements to the DCT framework, are also presented.

Chapter 2
Background

The background chapter of this thesis will provide the reader with the necessary background to understand the significance of the research. A general overview of hydrologic models is discussed, the current state of hydrologic model validation is discussed, the types of methods available for using models in a decision-making context are reviewed, and the gap that exists in decision-making between modellers and decision-makers is illustrated.

2.1 Hydrologic Modelling

This section provides an introduction to hydrologic modelling, including the purposes and types of hydrologic models and the steps involved to build a hydrologic model.

2.1.1 What is a hydrologic model?

The study of hydrology refers to, in basic terms, the study of the movement of water in the environment. The definition from the United States Geological Survey defines hydrology as, “the science that encompasses the occurrence, distribution, movement and properties of the waters of the earth and their relationship with the environment within each phase of the hydrologic cycle” (USGS, 2016). A hydrologic model is a numerical model that approximates a real system with various schemes for representing hydrologic processes, such as infiltration, runoff, evapotranspiration, etc. This can also be extended to include man-made processes and structures, such as reservoir outflow, tile drains, water extractions, etc. Once built, models are used to predict or emulate the behaviour of watershed systems. Hydrologic models have a broad application in the literature and otherwise, including water balance models, reservoir management, watershed management, groundwater studies, etc. Hydrologic models can also be coupled with water quality models for applications such as nutrient management and contaminant transport studies.

2.1.2 Classifications of Hydrologic Models

There are a few different categories by which hydrologic models can be classified. The first distinction is made in the purpose of the model, which influences the level of complexity and amount of data required to build and run the model. The purpose of the model can be broadly classified as either forecasting, simulation, or scientific. A forecasting model is used to produce some hydrologic variable(s) (usually river flows) into the future using forecasted forcing data, while a simulation model produces some hydrologic variable(s) for the same period as the data supplied (Klemeš, 1986). A scientific model is used to advance the state of hydrologic science via hypothesis testing and confirmation of conceptual models, and may not be useful in forecasting or simulation. Klemeŝ (1986) notes that forecasting models can be validated continuously as new measured data is captured, while simulation models cannot be validated in the same way. The validation of scientific models has the goal of verifying that the model produces the correct output for the right reasons (Biondi et al., 2012). This is not a strict requirement of forecasting models, which can have adequate forecasting capabilities without being scientifically correct. Forecasting models can be relatively simple and still perform well for their intended purpose, and in fact, complex models have not been shown to provide a large improvement over simpler models in terms of forecasting ability (Pagano et al., 2014). The comparison to observed data cannot be made with simulation models, which may be asked to project into futures that will never happen. In any case, the purpose of the model is one of the most important distinctions to make prior to its validation and use.

Another distinction is made in the spatial representation of the study area, which generally categorizes models as lumped, semi-distributed, or distributed models, in increasing levels of finer spatial discretization. A lumped model treats the entire study area or basin as a single unit, in which all state variables are represented with an average value across the entire basin. A distributed model makes predictions that are distributed in space by dividing the basin into distinct units using a 2D or 3D grid (or mesh); the hydrologic processes are simulated separately within each grid cell (Beven, 2012), and water must be transported or routed between grid cells. Semi-distributed models divide the basin into distinct subbasin units, generally using elevation data that can be used to identify surface water drainage areas. The behavior within each subbasin unit can be further discretized using hydrologic response units (HRUs), which are defined as units with unique hydrologic responses to precipitation, radiation, and other forcings. HRUs are constructed using common groupings of spatial data within each subbasin, including landuse, soil types, vegetation cover, terrain type, and management policy (Flügel, 1997). The HRU concept is similar to definitions of other units in hydrology, such as the Representative Elementary Area (REA) (Wood et al., 1988), the Grouped Response Unit (GRU) (Kouwen et al., 1993), the Representative Elementary Watershed (REW) (Reggiani et al., 1998), and hydro-landscape units (Dehotin and Braud, 2008).

Semi-distributed models are an intermediate level of discretization between lumped models and distributed models, since they allow a finer level of discretization than a lumped model without creating as many units as a distributed model. The pictorial representation of each of the three levels of discretization is shown in Figure 1.

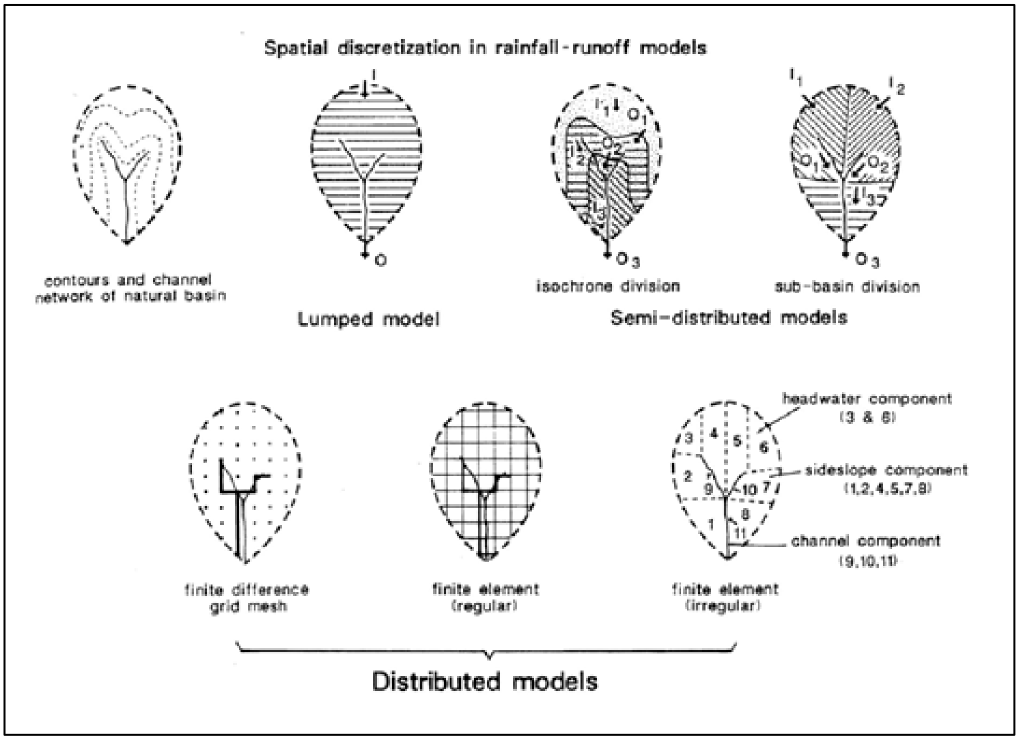
Spatial discretization in rainfall-runoff models
I
contours and channel network of natural basin
O
Lumped model
I1
O1
I2
O2
I3
O3
isochrone division
I1
I2
O1
O2
I3
O3
sub-basin division
Semi-distributed models
finite difference grid mesh
finite element (regular)
finite element (irregular)
headwater component (3 & 6)
sideslope component (1,2,4,5,7,8)
channel component (9,10,11)
Distributed models

Figure 1. Three types of spatial discretization in hydrologic models (Jones, 1997)

The decision in the level of discretization to use in a model is a tradeoff between the accuracy of the model in representing the heterogeneity of the basin, and the larger number of model parameters (and amount of computation) that are needed with increasing levels of discretization. There is a general notion that models with more parameters at finer discretizations are more physically realistic than models with fewer parameters, but due to the difficulty in properly building a model with many parameters, studies have found that distributed models can only marginally outperform or even underperform lumped models (Grayson et al., 1992, Reed et al., 2004).

Another distinction is made in deterministic models and stochastic models. Deterministic models always produce the same output for a given set of inputs. On the other hand, Stochastic or probabilistic models incorporate randomness or uncertainty in some component of the model, and can therefore produce different outputs from the same set of inputs. The majority of models used in hydrology are deterministic, although this distinction is not always clear (Beven, 2012). Deterministic models can be run using varied inputs or varied parameters to produce stochastic outputs. It is also possible to generate a distribution of model outputs using a set of different models with varied model codes, inputs, and parameterizations; this set of models is called an ensemble (Refsgaard et al., 2014a).

2.1.3 Model Building Procedure

The general model building procedure is presented in Figure 2, which is adapted from Refsgaard et al (2007) and Gupta et al (2008). This includes steps that are common in hydrologic modelling, although in practice the overall process is iterative, and some steps may be combined or otherwise vary from the sequence presented.

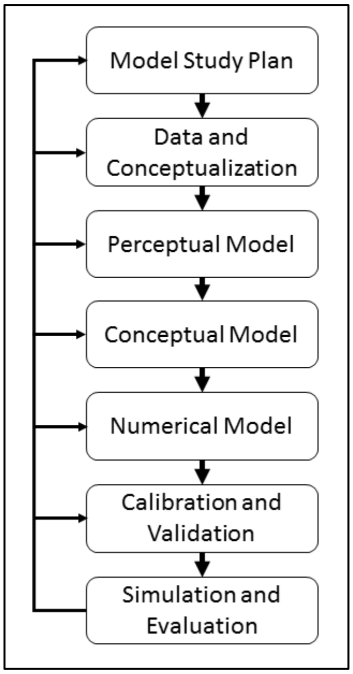
Model Study Plan
Data and Conceptualization
Perceptual Model
Conceptual Model
Numerical Model
Calibration and Validation
Simulation and Evaluation

Figure 2. Model building steps adapted from literature

The process begins with a Model Study Plan, in which the purpose of the model, the requirements of the model (in terms of output and level of accuracy), the resources available to build the model, and the parties involved in building the model are defined (Refsgaard et al., 2007). The initial definition of the purpose of the model is an important and often overlooked step, particularly if the model is to be used in a decision-making context (Liu et al., 2008), where there may be conflicting priorities in the desired use of the model.

The Data and Conceptualization step involves the collection of all relevant knowledge about the area to be modelled, as well as a development of a plan for modelling the area (Refsgaard et al., 2007). This includes specifying the resolution of data to be used, the sourcing of data, and an initial overview of the types of processes that will need to be included in the model.

The perceptual model is defined as the modeller's mental understanding of the key processes in the study area (Refsgaard and Henriksen, 2004), which is influenced by their knowledge and previous experience (Gupta et al., 2008). This is followed by the development of the conceptual model, defined as a model with a clear set of descriptions for model elements, such as model assumptions, boundary conditions, required inputs and outputs, process representations, etc. (Gupta et al., 2008, Liu et al., 2008, Refsgaard and Henriksen, 2004). In some literature, the perceptual model and the conceptual model are considered as one step (Liu et al., 2008, Refsgaard and Henriksen, 2004).

The next step is the numerical model, in which the site-specific model is implemented (using some kind of computer program) as a numerical approximation of the real system (Gupta et al., 2008, Refsgaard et al., 2007). A distinction can be made between the model code, a generic computer program that can be used to implement a conceptual model, and the site-specific model, which is built for the particular study area (Refsgaard and Henriksen, 2004).

Following the completion of a functional numerical model, the model undergoes calibration and validation. In these steps, the model's performance is optimized via parameter adjustments in calibration, and then is tested for accuracy and reliability using independent data via validation (Refsgaard et al., 2007). This step is the focus of the research in this thesis, and is thus discussed in greater detail in section 2.1.4 and section 2.2.

The final step is Simulation and Evaluation, in which the model is used for its intended purpose. This may be for simulation, forecasting, scientific investigations, and/or for use in decision-making. An uncertainty analysis of model outputs may also be included in this step to assist in the evaluation of model results (Refsgaard et al., 2007).

2.1.4 Model Calibration and Validation

Two key steps in the model-building process are model calibration and validation. Calibration refers to the process by which the model is trained or 'tuned' to be able to replicate a set of observed data. Since hydrologic model parameters are generally unknown (or imprecisely known), calibration is typically used to estimate parameter values that provide a good match of observed data. In other words, calibration is performed to optimize the model parameters for history matching (Refsgaard et al., 2014a). Calibration can be done manually, in which a modeller adjusts parameter values themselves, or automatically, which makes use of an algorithm to estimate parameter values. Often the modeller will use a combination of both manual and automatic techniques to calibrate a numerical model (Boyle et al., 2000). The topic of automatic calibration for hydrologic modelling is a large research area, and many different calibration algorithms have been deployed, e.g., see Ostrich Optimization Software Toolkit documentation (Matott, 2016).

Model validation in hydrology is generally considered the step in the model-building process where the model is analyzed and evaluated for performance (Biondi et al., 2012), generally using data which is independent from that used in calibration. The validation step is done as a check on the model to ensure it is fit for purpose, and does more than just match the historical data provided in calibration. A poor result in model validation generally means that the model will not perform well in future use, and would warrant some iteration of other steps in the model-building procedure, ranging from re-calibrating the model to reconsidering the conceptual model.

Model validation is less well-defined than model calibration, and both the definition and possibility of validation has been the subject of debate in many articles. This is discussed further in the following section.

2.2 Critical Evaluation of the Concept of Validation in Hydrologic Modelling

This section discusses the various philosophical approaches to validation from literature, the current state of validation methods in the literature, and the call for improved validation methods.

2.2.1 Philosophy of Validation

Model validation is a commonly performed task, however it is worthwhile to ask the question, ‘Can a hydrologic model be validated?’ This question has been discussed in literature at length. The first common reference in this discussion is to Popper, who promoted the view that models cannot be validated, only invalidated (Popper, 1959). Konikow and Bredehoft agreed that hydrologic models cannot be validated, and that ‘validation’ should be abandoned as term because it misleads the reader into believing that the model has more predictive power than it actually does (Konikow and Bredehoeft, 1992). Orsekes et al stated that models should be evaluated in relative terms, and not in an absolute sense (Oreskes et al., 1994). Klemeŝ argued that validation should refer to the testing of a model’s “operational adequacy” for applied purposes (such as forecasting), and that models used to advance hydrology as a science cannot be verified (Klemeš, 1986). This is echoed by Rykiel, who stated that validation is “certainly possible”, as long as (1) the purpose of the model, (2) the performance criteria required of the model for acceptable use, and (3) the context that the model will be used in, are specified (Rykiel, 1996). Rykiel also discussed ‘engineering validation’, and remarked that model validation cannot prove scientific theory. A more recent article by Refsgaard et al (2014) stated that, “it is not possible to carry out model verification or model validation if these terms are used universally without restriction to domains of applicability and levels of accuracy”. Other articles also distinguish between validation for the purpose of model performance and validation of scientific models (Biondi et al., 2012), and state that models must be evaluated in the context of their objectives (Jakeman et al., 2006).

The general conclusion drawn from the literature is that models can only be validated for acceptable performance in a limited application, and not in the more general sense, which is distinct from what happens in practice. This emphasizes the need to validate models for a specific purpose and levels of accuracy, and the inadequacy of applying validation concepts to models for general use.

2.2.2 Current Validation Practices in Hydrology

The classic paper by Klemeŝ (1986) defined a hierarchy of four main validation methodologies with application to hydrologic models, which remains the standard for validation procedures to this day. The methodologies are summarized below.

1. *Split-sample test*: The data are split into two segments, one of which is used for calibration and the other for validation. The model should only be acceptable if the model performance and error characteristics are similar in both calibration and validation. Klemeŝ recommended splitting the data 50/50 between calibration and validation if the data period was ‘sufficiently long’, else a 70/30 split between calibration and validation. Typically in practice, this test is the only one applied for validation, likely because it is the easiest to apply and only one dataset is required (Refsgaard et al., 2014b).

2. *Proxy-basin test*: A model is calibrated using data from one basin and validated using data from a second basin, and vice-versa. This is considered a test for the transposability of the model to other regions, particularly for ungauged basins (defined as basins with “inadequate records of hydrological observations to enable computation of hydrological variables” (Sivapalan et al., 2003)).

3. *Differential split-sample test*: The data is split into two sets based on a differential characteristic, for example, into wet years and dry years. The model is then calibrated to one data set and validated on the other. This test was requested by Klemeŝ to be required whenever a model is to be used under conditions different from those in the flow record, since this method tests the model's ability to perform under conditions that are not observed in calibration data.

4. *Proxy-basin differential split sample test*: This combines the proxy-basin and differential split sample test, and is considered the most rigorous of the four methodologies proposed by Klemeŝ. The data from two basins would be collected, say basins A and B, and in each basin the data would be split according to the differential split-sample test, for example as wet years denoted set 1 and dry years denoted set 2. Then the model would be calibrated/validated from data in different basins and data groupings, for example, wet years in basin A (A1) and validated on dry years in basin B (B2). The reverse of this would also be done, i.e., the model calibrated to dry years in basin A (A2) and validated using wet years in basin B (B1).

These methodologies proposed by Klemeŝ have been applied a few times in the literature, e.g., (Donnelly-Makowecki, 1999, Refsgaard and Knudsen, 1996, Seibert, 2003, Xu, 1999). Coron et al also presented a generalized split-sample test, in which overlapping periods are used for generating multiple calibrated models, and these models are then validated in every other available independent data period (Coron et al., 2012). Outside of these few examples from hydrology, the methods proposed by Klemeŝ have not found wide use in earth science disciplines (Refsgaard et al., 2014b); only split-sample tests are commonly applied, and the other three (more rigorous) levels of testing are rarely applied (Andréassian et al., 2009).

2.2.3 Call for Improved Validation of Hydrologic Models

The recent literature presents a renewed call for an improvement in the validation tests used for hydrologic models, and cites the lack of interest in developing new tests since Klemeŝ (1986) (Refsgaard et al., 2005). The split-sample test is considered a weak one and is inadequate for climate change studies in particular, while the differential split-sample test is more appropriate although not commonly used (Kirchner, 2006, Refsgaard et al., 2014a). Other authors argue for the development of improved evaluation methods altogether (Gupta et al., 2008), and state that further development of suitable testing schemes is a major challenge (Refsgaard and Henriksen, 2004). Andréassian et al call for the development of more rigorous and demanding testing of hydrologic models, analogous to crash tests in the automotive industry, in order to test the limits of models and to help improve the models (Andréassian et al., 2009). Other authors also call for more rigorous testing of hydrologic models (Biondi et al., 2012, Jakeman et al., 2006, Refsgaard et al., 2005).

One suggested reason that an agreed upon and rigorous validation procedure has not yet been proposed is due to the difficulty in developing a procedure that can be used generally for the large number of different types of hydrologic models that exist, and remain general across many different applications of hydrologic models (Biondi et al., 2012). Whatever the reason for the lack of uptake on rigorous validation methods, the call for the development and adoption of rigorous validation methods is clearly documented in the literature.

2.3 Decision-Making in Hydrology

This section presents background on the application of hydrologic models to decision-making. This section begins by giving a brief review of the decision-making process and the types of computationally-based decision-making methods that have been deployed in the hydrologic literature. Subsequently, this section discusses how uncertainty in modelling is incorporated into the decision-making process, illustrates the gap that exists in the literature, and ties the concepts of model validation and decision-making together to provide the motivation for the main work done in this thesis.

2.3.1 Decision-Making Process

The decision-making process is generally defined as the process through which a decision is scoped, evaluated, and executed in practice. The particular steps included in the decision-making process vary both in the literature and in practice, depending on the particular application. A generic decision-making approach in which a computational hydrologic model is used to support the decision is presented below (Liu et al., 2008, Refsgaard et al., 2007, Walker et al., 2003).

1) **Problem Formulation**: The initial problem and goals of a decision are explicitly stated, as well as the limitations of the potential decisions to make. A list of decisions to evaluate, or a single decision to test, is defined. The questions of ‘why is modelling required?’, ‘what level of accuracy is required?’, and ‘how will the model be used to support decision-making?’ should be answered in this step. This step will also identify the key parties involved in the

decision-making process (i.e., the modellers, the decision-maker(s), the stakeholders or members of the public, policy experts, collaborators, and perhaps external reviewers).

2) **Data Collection and Model Development**: Any data that is required to support model development and/or decision-making is identified and collected, and a computational model is built to replicate the real system. This step encompasses the development of the initial conceptual model of the complex system, the detailed model-building steps, the calibration and validation of the model for decision-making, and the linking of the hydrologic model to any other models that may be required (e.g., economic models, climate models, etc.).

3) **Model Simulation and Decision Identification**: In this step, the model is used to generate predictions of system behaviour that inform the selection of one or more preferred decisions or course of actions for the future. This step encompasses activities related to scenario development (i.e., formulation of plausible future states of the world for evaluating the performance of decisions), decision assessment, decision modification or adjustment, etc. This step is likely to be iterative and involve collaboration between modellers, policy experts, and stakeholders in determining the preferred decision. Ultimately, the decision-maker uses the insights from the computational model(s) and other parties to make a decision.

4) **Decision Implementation**: Once a decision is selected by the decision-maker, the decision may be implemented. This involves the communication of the decision to the relevant parties (including uncertainty analysis results, if performed), and a continuous collaboration between parties to ensure the decision is implemented correctly and efficiently.

5) **Monitoring, Evaluation, and Follow-up**: The impacts of the decision are monitored and evaluated to ensure success. If the decision is unsuccessful, or one or more key assumptions have been found to be untrue, then the process may be repeated or the decision may be adjusted.

While there is a wide variety in decision-making approaches in the literature, many steps are common between methods. There is also a general agreement in literature that the decision-making process is typically iterative. Each step in the decision-making process can also be iterative, and involve review with stakeholders, modellers and policy experts.

The case studies in this thesis will focus on the first three steps, in which the plausible decisions are identified, the model to support the decision is developed, and the model is used to inform the decision. The decision implementation and post-monitoring are outside the scope of this thesis, since the focus is on the evaluation of the model itself. The distinction between informing a decision and making a decision is highlighted; a model may *inform* a decision by producing predictions and/or simulations that illustrate the preferred decision, however the decision-maker must still *make* the decision, using the information from both the model and other sources. In relation to the process above, the model supports the decision by informing it in Step 3, and the decision is made in Step 4 by the decision-maker, which may or may not be the same as the decision informed from the model.

2.3.2 Decision-Making Example in Hydrology

To put the decision-making process in context for hydrologic applications, an example of a decision to build a reservoir is presented. This example is discussed for the first three steps of the decision-making process to be consistent with the focus of this thesis, i.e., the use of a model to inform the decision.

1. **Problem Formulation**: The decision is formulated as a binary decision in terms of whether to build a reservoir at a specific location to mitigate flooding, or to not build a reservoir. The decision is based on what the expected peak flow is over the next 10 years, and whether the cost of building the reservoir offsets the potential damage from a large flood. To simplify this tradeoff, the decision is set up such that an expected peak flow greater than X m^3/s suggests that the reservoir should be built, and vice-versa, where the threshold is calculated based on a cost-benefit analysis of flood damage versus cost of reservoir construction. It is decided that a model is required to estimate the peak flow over the next ten years.

2. **Data Collection and Model Development**: The relevant data is collected to build the hydrologic model, and the model itself is built to support the decision. To build a hydrologic model for reservoir management, the required data may include: soil maps, topography, vegetation types, upstream drainage area, etc. The historical streamflow data would also be collected for calibration and validation of the model. Once the data is collected, the model is built and tested in its ability to predict peak flows.

3. **Model Simulation and Decision Identification**: In this step, the model is actually run to predict peak flows in the next ten years, and the results are used to inform the decision. For example, if the model is run and the estimated peak flow does not exceed the determined

threshold of *X*, the decision to not build the reservoir is identified by the model as the preferred decision.

Following this step, decision to build or not build the reservoir would be made by the decision-maker, which would likely be the same as the model-informed decision. The peak flows would be monitored to ensure that the model was in fact capable of predicting peak flows accurately, and ensure that the selected course of action does not need to be altered.

In this example, the model may have informed the decision by proposing that the reservoir not be built, based on the provided decision formulation and threshold for peak flow. However, the decision-maker may still make the decision to build the reservoir. For example, if the peak flow predicted by the model is close to the threshold, the decision-maker may take the conservative decision to build the reservoir anyway, given the uncertainty present in the model. Other factors may also play a role; for example, the decision-maker may decide to build the reservoir due to high public pressure to do so, regardless of the modelling results.

As is typical with real-decision making, the process may be iterative. For example, if the decision is reached to not build a reservoir, the initial problem could be reformulated to consider building a hydroelectric generating station with the reservoir. The generating station may be expected to recover sufficient cost to offset the cost of construction, and thus the decision may be selected to build the reservoir. In terms of the decision setup, a recovered cost would reduce the threshold of expected peak flow required to justify building the reservoir, thus potentially changing the decision if the expected peak flow was within that range.

2.3.3 Methods, Frameworks and Sensitivity Analysis in Decision-Making

There exist a number of methods in the literature for using a computational model, or a set of computational models, for decision-making applications in the field of hydrology. The methods can be broadly classified based on their approach to defining the preferred decision from a set of alternatives. One class of methods focuses on an optimality approach, in which a metric is presented for evaluating the performance of a given decision, and the preferred decision is selected based on the best metric in a likely future state. For example, multi-criteria decision analysis (MCDA) uses multiple objectives to evaluate a set of decisions, and selects a preferred decision from a tradeoff curve, e.g., (Ahmadisharaf et al., 2016, Matrosov et al., 2015, Yu et al., 2016). Methods also exist that incorporate some concept of ‘robustness’ into their optimization for selection of the preferred decision. Robustness is defined here as satisfying some minimum performance criteria across a range of plausible future states rather than being optimal in the most likely future state. For example, in a water supply problem, overdesigning the water treatment and supply facilities for a larger-than-expected capacity is not optimal in the most likely future state (due to an increase in cost), but it is much more robust to a plausible high demand scenario, whereas the 'optimal' solution is more likely to be inadequate in the high demand scenarios. Examples of these types of methods includes Robust Optimization (Gorissen et al., 2015) and Info-Gap Decision Theory (Ben-Haim, 2010). Uncertainty can also be incorporated into decision-making using Bayes theorem, which is done in the literature with Bayesian Decision Analysis (Davis et al., 1972, Varouchakis et al., 2016) and Bayesian Networks (Carmona et al., 2013, Pang and Sun, 2014). In these cases, the probabilities of given events are described statistically and thus the expected value of a decision can be computed and used to select an optimal decision under risk. A specific application of Bayesian decision analysis is applied to linking climate models with risk assessment in water resources in 'decision scaling' (Brown et al., 2012).

A different philosophy in decision-making is to use models to explore the conditions under which a decision can fail and focus on selecting or designing a better decision in the face of uncertainty, rather than selecting the optimal best decision from a given set. These methods tend to be more iterative in their procedures, as they require rebuilding a decision and re-evaluating under what circumstances the current preferred decision may fail. In these methods the concepts of robustness and adaptivity, the latter defined as leaving options 'open' for future adjustments as more information is known, are commonly discussed. These methods tend to use elements of Exploratory Modelling Analysis (EMA) (Bankes, 1993), which uses models as experiments to explore various assumptions or scenarios about a system. Examples of these robustness- and adaptivity-seeking methods include Robust Decision Making (Groves and Lempert, 2007, Matrosov et al., 2013), Real-Options (Steinschneider and Brown, 2012), and Dynamic Adaptive Policy Pathways (Haasnoot et al., 2013, Kwakkel et al., 2015, Kwakkel et al., 2016).

Several tools for supporting decision-making also exist. One such example is Scenario Discovery (Bryant and Lempert, 2010), which is tool to analyze scenarios and identify key scenarios to inform decision-making. The method involves examining the performance of scenarios and classifying them based on failure criteria, which can be done using a number of classification algorithms, e.g., Self-

Organizing Maps, Patient Rule Induction Method, etc. (Bankes et al., 2013). Scenario Discovery is often used to support Robust Decision Making in its analysis of scenarios.

Aside from decision-making methods which use a computational model to inform a decision, there exists a much smaller body of literature on methods that evaluate a model in a decision-making context, or provide insight into the use of a model for decision-making applications. There are two sensitivity-based methods in this category, one of which is the Management Option Rank Equivalence (MORE) method (Ravalico et al., 2010). The MORE method explores the simulation model parameters in decision space to find what is called the rank-equivalence boundary, where a change in parameter value would change the decision selection outcome. Using a Pareto optimization approach to explore parameters in decision space, various solutions to the minimum required changes in parameter values to change the preferred decision (conditional on the values of other parameters) can be reported, and used to determine which parameters are most critical in the model for decision-making applications (Ravalico et al., 2009). This method is similar to previous studies, such as the local or one-at-a-time sensitivity analysis performed for decision-making under climate change (Dessai and Hulme, 2007).

The other sensitivity-based method is the *de Novo* planning framework (Kasprzyk et al., 2012), the key feature of which is a global sensitivity analysis performed on the decision variables with respect to performance criteria, which is used to iteratively test formulations of the decision rules. In this method several decision formulations can be tested, and the preferred decision formulation can be identified.

Another method for evaluating a model in decision-making, not based on sensitivity analysis, is the Iterative Closed Question Modelling (ICQM) approach (Guillaume et al., 2015). ICQM is a general methodology for making decisions under uncertainty, and uses the formulation of boundaries in decision space to draw conclusions about decision-making. The boundaries formed in the ICQM methodology include the epistemic knowledge of the plausible scenarios, defining assumptions of what is plausible in the physical system, and normative boundaries, defining the different types of plausible decisions to be made. The explicit mapping of plausible decisions and scenarios helps to define which decisions will meet the objectives, which plausible scenarios are of concern, etc. The method offers little guidance on making a decision explicitly; ideally, the iteration of incorporating more knowledge to adjust the epistemic boundary, and the inclusion or exclusion of decision possibilities to adjust the normative boundary, eventually concludes when only one feasible decision remains. This method is likely more useful as an exercise to learn more about the state of knowledge and plausible options available for decision-making than a decision-making method itself.

2.3.4 Uncertainty in Hydrology and Decision-Making

Since decision-making is an endeavor that generally involves individuals and organizations of varying backgrounds, areas of expertise, and technical language, a clarification of what is meant by ‘uncertainty’ is warranted. There is a divide in the treatment of uncertainty by natural scientists and decision-makers in terms of how uncertainty is discussed and incorporated into the decision-making process. For example, in the natural sciences the term ‘uncertainty’ often refers to statistical uncertainty, while the decision-makers’ interpretation tends to pertain to the balancing of conflicting objectives (Walker et al., 2003). Decision-makers will tend to take a risk-based approach to decision-making and focus on system vulnerabilities, while scientists tend to focus on reducing uncertainties (Höllermann and Evers, 2017). These differences are important to consider in bridging the gap between scientists and decision-makers.

Perhaps one of the most comprehensive classifications of model-based decision support is the article by Walker et al (2003), in which uncertainty is defined generally as, “any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system”. The framework for classifying uncertainty proposed by Walker et al has been revisited by several articles (Brugnach et al., 2008, Refsgaard et al., 2007), and was formally updated in 2010 (Kwakkel et al., 2010).

Uncertainty in these articles is described by three dimensions: location, level, and nature. The location of uncertainty refers to the component of decision support where the uncertainty lies, whether it be in the system description and problem formulation, in the model structure or implementation, and/or the inputs to the model. For example, the uncertainty in output from a hydrologic model typically comes from a combination of: uncertain measured inputs (e.g., precipitation, temperature), uncertain conceptualization of the model (bucket storage model vs. detailed Richards equation), unclear completeness of the process representations, and the parametric uncertainty from calibration to uncertain and limited observation data. The level of uncertainty refers to a spectrum of the magnitude of uncertainty, ranging from deterministic (known with certainty) to statistical, to total ignorance (we do not know what we do not know). The nature of the uncertainty refers to three main categories: epistemic uncertainty, variability uncertainty, and ‘ambiguity’. Epistemic uncertainty is due to imperfect knowledge, and can thus be reduced with more study. As an example, the spatial distribution of rainfall could be reduced with more gauges, and could thus be considered a mostly epistemic uncertainty. Variability uncertainty is inherent uncertainty due to random behaviour, and thus cannot be reduced further (this is also known as aleatory uncertainty in the natural sciences (Beven, 2008, Beven, 2013)). Finally, ambiguity refers to the interpretation of the nature of uncertainty by multiple heterogeneous stakeholders and decision-makers. In other words, the framing of the decision can be a source of uncertainty as well. For example, the framing of a groundwater extraction issue can be set as "excessive water consumption" by the ecologist, "illegal water extraction" by policymakers, and "insufficient water supply" by a farmer (Brugnach et al., 2008). This would lead to a different discussion of uncertainty depending on the framing, and various components of uncertainty (e.g., regional groundwater storage) may be highly relevant to one party and of no concern to another. An examination of decision-making from the practitioner perspective also adds the dimension of procedural uncertainty (Höllermann and Evers, 2017), which is related to the actual decision-making process and implementation of a decision. This can include uncertainties in steps required to make and implement a decision, such as the competency of parties involved in decision-making, and the political or social acceptance of a proposed decision.

There is also the concept of deep uncertainty, which is frequently used in the robust decision making and adaptive management literature (Bryant and Lempert, 2010, Groves and Lempert, 2007, Haasnoot et al., 2013). Deep uncertainty describes, in basic terms, a situation where the system model and its components cannot be adequately described statistically, and the uncertainty creates a condition where traditional optimality-based approaches are not sufficient in decision-making (Groves and Lempert, 2007).

The inclusion of the various kinds of uncertainty in a model-based decision-making framework is a difficult problem, and is further complicated by differences in perspective of modellers and decision-makers. The identification of how what kind of uncertainty to consider, and how important it will be in the model and decision-making process, is an important first step to set reasonable expectations of accuracy and performance (Jakeman et al., 2006). The inclusion of uncertainty in decision-making, and how it is treated by relevant parties in the decision-making process, is explored more in the following section.

2.3.5 Linking Hydrologic Models and Decision-Making

Environmental models are used extensively in decision-making (Bennett et al., 2013), and there is a general trend of increased use of models to support decision-making and management in the environmental sciences (Gupta et al., 2008, Jakeman et al., 2006, Liu et al., 2008, Refsgaard and

Henriksen, 2004). This emphasizes the need for ensuring that computational models can reliably provide value in decision-making contexts. The incorporation of uncertainty from hydrologic modelling into decision-making is a recognized difficulty. For example, uncertainty bounds on model outputs (e.g., the 5th and 95th percentiles of the output quantity of interest) are the most common means of describing model uncertainty, despite the difficulty that practitioners experience in incorporating the uncertainty bounds into a decision (Höllermann and Evers, 2017). It has been suggested that uncertainty in modelling exercises cannot be understood by decision-makers or by the public (Pappenberger and Beven, 2006), with some evidence that this is the opinion of scientists and water managers alike (Höllermann and Evers, 2017). In a more general sense, information that is critical for the decision-maker is not always provided by scientists in a form that is understandable to decision-makers (Isendahl et al., 2009, Liu et al., 2008), leaving non-modellers to interpret the quality of the modelling results without the proper experience to do so (Jakeman et al., 2006).

A few methods explicitly map the model results into a decision to handle uncertainty, e.g., (Brown et al., 2012, Guillaume et al., 2015), which would help to translate the uncertainty into terms that decision-makers can find useful. However, these methods are limited both in the literature and in their application, and these methods do not assess the ability of the model itself to support decision-making explicitly.

Suggestions to improve the understanding of models to inform decision-making have been made in literature. The concept of participatory modelling, in which models are developed in direct consultation with stakeholders, has been shown to improve the decision-making process, e.g., (Carmona et al., 2013, Xue et al., 2016). Various studies have agreed that cooperation with stakeholders and transparency in the decision-making process are important for improving the acceptance of decisions supported by models (Carmona et al., 2013, Höllermann and Evers, 2017, Jakeman et al., 2006). Scientists have a tendency to skip to the numerical modelling step with only a general idea of the future use of the model (Liu et al., 2008), which fails to establish the purpose of the model and can reduce the model's credibility in a decision-making context. Providing a clearly defined model purpose and an explicit conceptual model early in the process can help to make the model more transparent and bridge the gap between modellers, decision-makers and stakeholders (Liu et al., 2008). To this end, there is also a call in the literature to include criteria for model evaluation other than performance, such as fitness for purpose, flexibility to transient management needs, and transparency to stakeholders (Jakeman et al., 2006).

Other suggestions have included the reframing of uncertainty as risk to allow decision-makers to focus on vulnerabilities (Höllermann and Evers, 2017), which is much more readily understood in decision-making than uncertainty. A similar type of thinking has suggested that uncertainty can be incorporated into decision-making via scenario analysis, which is frequently used by decision-makers (Liu et al., 2008). It has also been suggested that binary decisions are preferred by decision-makers (Höllermann and Evers, 2017), thus framing decisions as binary where possible may improve the usefulness of models for decision-making.

The call for improved communication between scientists and decision-makers in environmental decision-making is ubiquitous and a subject of increasing interest (Keur et al., 2010, Liu et al., 2008). The linking of hydrologic model outputs to the decision-making process remains a challenge (Höllermann and Evers, 2017), and in general the integration of environmental science into decision-making is regarded as one of the most difficult challenges of environmental management (Liu et al 2008).

2.4 Linking Hydrologic Model Validation to Decision-Making

There are no existing methods known to the author that directly validate a hydrologic model in a simulation (versus forecasting context) explicitly for a decision-making purpose, only methods (such as the ones discussed in section 2.3) which inform the decision-making process generally, or inform ability of a model to perform in a decision-making application only indirectly. In addition, the current methods of evaluating model performance and uncertainty are framed in a manner that is not necessarily useful for decision-makers, as discussed in section 2.3.1. The call in literature for new developments in validation methodologies is also discussed in section 2.2.3.

This thesis addresses these gaps by presenting a method for rigorously validating hydrologic models in a manner that is directly useful for decision-makers, i.e., directly in a decision-making context. The thesis methodology is presented in the following chapter.

Chapter 3
Methods and Model Development

This chapter contains an explanation of the methodology behind decision crash testing, the focus of this thesis, which ties together the discussions of validation and decision-making in Chapter 2. The second half of this Chapter will describe the model development used to support the two case studies in this thesis. The method of model development for hydrologic models involving reservoirs discussed in this chapter is considered a contribution to the literature on its own merit, and is thus described in sufficient detail.

3.1 Decision Crash Testing

Decision Crash Testing (DCT) is a novel method for evaluating the ability of a model to inform decision-making. The DCT framework was developed by Dr. Bryan Tolson and Dr. James Craig at the University of Waterloo and first reported in Tolson and Craig (2016), is further developed and expanded upon in this thesis. This section will introduce the DCT methodology, discuss the advantages of DCT, illustrate how DCT fills some of the existing gaps in literature, and discuss how it is different from existing methods.

3.1.1 DCT Methodology

DCT was inspired by the crash test analogy of Andréassian et al (2009), which was in turn inspired by the classic paper on model validation methodologies of Klemeŝ (1986) (XXX Tolson and Craig 2016). In a situation where one wishes to use a model to inform a particular decision, DCT is intended to estimate the ability of that model to provide the correct decision by evaluating it in a set of hypothetical scenarios. This is an answer to the call in literature to both (a) provide more rigorous testing of hydrologic models in validation, and (b) bridge the gap between the uncertainty in hydrologic models and their interpretation in decision-making. An explicit outcome of this method is a measure of how likely the model is to inform the correct decision in a specific decision-making context, which is much more interpretable for decision-makers than other commonly used measures relaying model skill.

The following is a summary of the main steps in DCT evaluation.

1) A decision-making problem for the evaluated model to inform is formulated quantitatively. This includes the specification of a *decision quantity* (a value computed based on the

numeric output from the model on which the decision depends) and a set of rules for making a decision based on the decision quantity.

2) A *synthetic reality* or 'truth' model is generated, which is taken as a plausible representation of the natural system, and is somehow different from the base model being evaluated. The relevant state variable(s) and/or fluxes are extracted from the synthetic reality and supplied to the model undergoing evaluation for model building and/or calibration. In other words, the synthetic reality provides these 'observations' as a proxy for the real measurements required in building the model.

3) A correct decision is determined for the synthetic reality (i.e., if the synthetic reality was the real system, what would the correct decision be?) The decision must be consistent with the decision formulation rules in Step 1. This decision is assumed to be the correct decision only in this instance of the synthetic reality, and based on the decision rules defined in Step 1.

4) The model undergoing evaluation is calibrated to the 'observed' data generated from the synthetic reality in Step 2. Note that the 'model' here refers to both the numerical model and the methods involved in the model-building procedure; all of the choices involved in this process, including the calibration choices, may be evaluated in DCT. This step results in a complete model that can be used to select a decision.

5) The model built in Step 4 is used to solve the decision-making problem in Step 1. The corresponding decision is conditional on the model building procedure and supplied observed data.

6) The decision informed by our built model in Step 5 is compared to the 'correct' decision from the synthetic reality in Step 3 by means of the computed decision quantities. If these two decisions are the same, the decision informed by the model is recorded as a correct decision; otherwise, the decision is recorded as an incorrectly informed decision (since it is not consistent with the synthetic reality decision).

7) Steps 2 through 6 are repeated *N* times to create a statistical sample of size *N*. In each iteration, a new synthetic reality or 'truth' is assumed, synthetic observation data are generated, a new model is built, and the decision of the model is checked against the

synthetic reality decision. The result of each iteration is recorded and various post-processing diagnostics can be computed.

To provide a context for the above steps, the methodology is demonstrated with an example. Here, a binary decision on whether to build a reservoir as a function of expected peak flow is demonstrated. Let us assume that in this case, ten years of data from the real system are available to build a model, after which we must predict flows ten years into the future and decide whether to build a reservoir for flood mitigation if the expected future peak flow exceeds a given threshold. The DCT methodology would be setup as follows.

1) The decision is formulated as to build a reservoir if the peak flow in the ten years of the model prediction period exceeds some threshold, i.e., $max(Q) > Q_{thresh}$ (decision A). Otherwise, do not build the reservoir (decision B). In this example, the future peak flow in the next 10 years is the decision quantity, since this is the model output that dictates how the decision will be made.

2) A synthetic reality model is generated in which the flow values during the twenty-year period are considered as known (see section 3.1.3 on details for synthetic reality generation). The first ten years will be extracted as 'observed data' to build the model, and the next ten years of flows will be used to define a correct decision if all flow values were known.

3) The correct decision is based on the second ten-year set of flow values in our synthetic reality. Let us say that the maximum flow in the second ten-year set of synthetic flows was larger Q_{thresh}, meaning that the correct decision is to build the reservoir (decision A).

4) The model being evaluated is calibrated to the first ten-years of synthetic observation data. For simplicity let us assume that we cannot change our discretization and model structure (or have good reason not change them), and we would just like to test the calibration methodology of our modelling procedure. The calibration is performed on the first ten years of synthetic flows until the allocated calibration budget is consumed. By the basic procedure assumed here, the model is now considered 'ready' for use in decision-making.

5) The decision from our model built in Step 4 is evaluated. The model is run for the second ten-year set to predict peak flows in that period, and a decision based on the exceedance of

Q_{thresh} in that period is made. Let us say that the model does not predict any flow to exceed Q_{thresh}, thus the model-based decision is to not build a reservoir (decision B).

6) The correct decision, defined by the synthetic reality, and our model-based decision are compared. Since the synthetic reality decision (or correct decision) was to build a reservoir (decision A) and the model-based decision was to not build one (decision B), the decision is recorded as incorrect. The decision quantity (i.e., peak flow) from the synthetic reality and built model are also recorded.

7) Steps two through six are repeated *N* times, for example with 100 iterations.

From the iterative comparison of the model decision to the synthetic reality, inferences can be drawn on the ability of the model to inform the decision correctly. For example, the number of times the model proposed the correct decision out of 100 represents, in some sense, as an estimate of the likelihood that the model will inform the decision correctly in the real application. This estimate is actually representative of an upper bound on the decision-making skill of the model given the critical (but easily defendable) assumption is that informing the correct decision is easier for the model in this synthetic experiment than when the model is calibrated to real-world, non-synthetic data (i.e., the synthetic reality is easier to represent than actual reality). The decision formulations and details of analyses within the DCT framework are presented next.

3.1.2 DCT Similarity and Decision Models

In a mathematical formulation of the DCT framework, let us denote the decision quantity as ϕ. The correct (or true) decision quantity for synthetic reality iteration *n* will be denoted as ϕ_n, and the model decision quantity will be denoted $\hat{\phi}_n$, as an estimate of ϕ_n. Note that ϕ_n and $\hat{\phi}_n$ are paired, since DCT produces one modeled decision quantity per synthetic reality generation.

The sample of synthetic realities will produce a distribution of decision quantities, ϕ_n, with an associated probability density function, $f(\phi_n)$ (in the limit as $N \to \infty$, where *N* is the number of synthetic reality samples). In the case of a binary decision there exists a single *decision quantity threshold*, denoted ϕ^*, which divides the decision space into two plausible decisions (denoted decision A and decision B). These quantities are depicted in Figure 3.

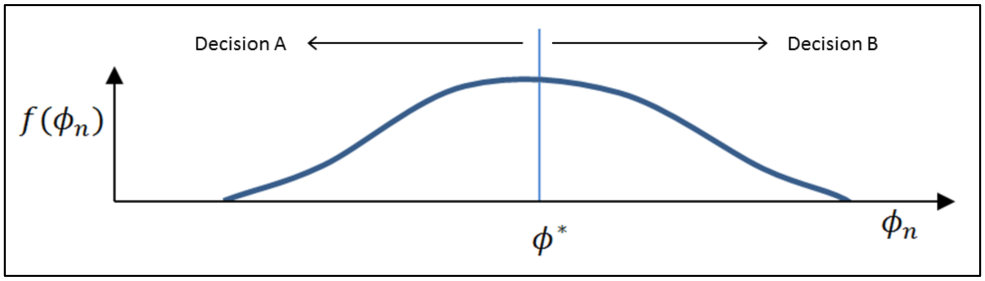

Figure 3. Decision quantity distribution from synthetic reality generations for binary decisions

In a simple example where we produce ten synthetic reality generations and thus ten pairs of ϕ_n and $\hat{\phi}_n$, we can plot the results of ϕ_n and $\hat{\phi}_n$ and visualize how the DCT diagnostics are computed. A sample plot for *N=10* points is shown in Figure 4.

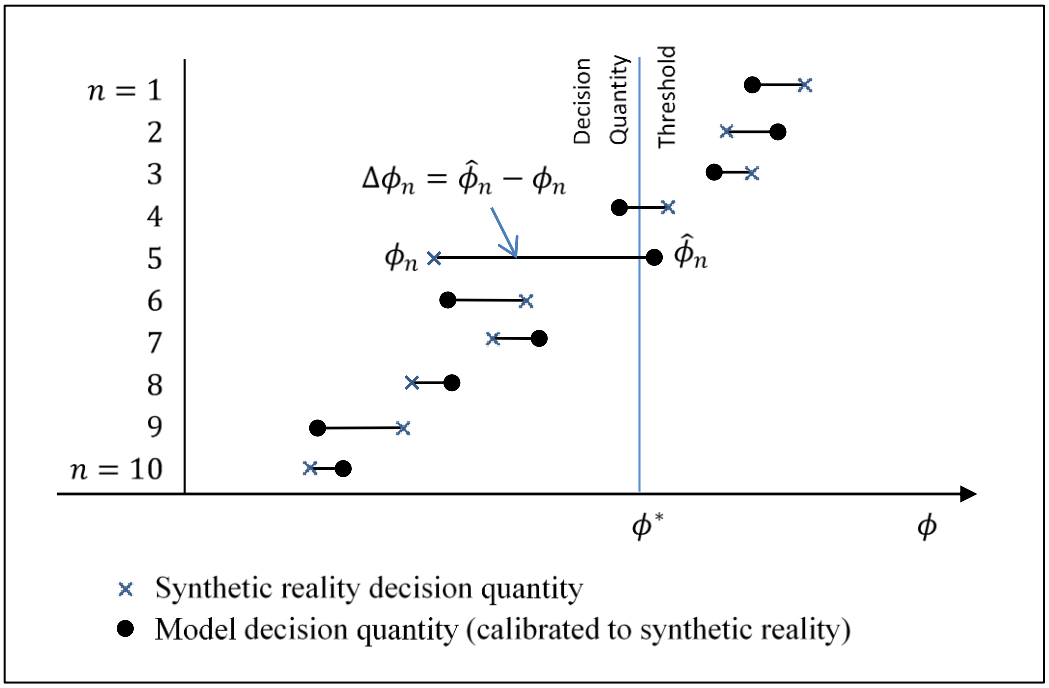

Figure 4. Decision quantity pairs from synthetic reality generations (ϕ_n) and model outputs ($\hat{\phi}_n$)

Figure 4 shows the ten pairings of ϕ_n and $\hat{\phi}_n$ in one-dimensional decision space, with the decision quantity threshold at ϕ^*, and a metric known as *model skill* denoted $\Delta\phi_n$. The model skill is the difference in decision quantity of the model and the synthetic reality, and is expressed mathematically as $\Delta\phi = \hat{\phi}_n - \phi_n$ for a given trial. In each of the ten instances, the model correctly informs the decision if ϕ_n and $\hat{\phi}_n$ are in the same region of decision space (i.e., on the same side of the decision quantity threshold ϕ^*). The model fails to inform the correct decision if ϕ_n and $\hat{\phi}_n$ are on opposite sides of the decision quantity threshold. From the example in Figure 4, the model fails to inform the correct decision for DCT iterate (or DCT observation number) *n=4* and *n=5* but informs the correct decision in all other trials, thus the percentage of correctly-informed decisions, called the similarity score, is 0.8 (i.e., model informed the correct decision 80% of the time). Generally, the similarity score is calculated as:

$$S_s = \frac{\#\{(\phi_n - \phi^*) \cdot (\ \hat{\phi}_n - \phi^*) > 0\}}{N} \tag{1}$$

where S_s is the similarity score, *N* is the number of trials, and the # is the number of elements in the set where the product of the two terms is greater than zero (i.e., the modelled and synthetic reality decision quantity are on the same side of the decision quantity threshold).

The similarity score can be compared to the probability of informing the correct decision purely by chance. In our example, a binary decision has a 50% chance of being made correctly at random; thus, a similarity score of 55% would mean that our model building procedure is about as good as a coin flip in making decisions, and 80% could be considered a fair bit better than a coin flip; how much better than random chance the model is required to be in order to be considered ‘good’ depends on the nature of the decision. Statistical tests on proportions can be applied to determine if a similarity score is significantly different than 0.5 in the binary case.

An additional analysis that can be applied to the DCT results is a contingency table, which has been used for analyzing probabilistic hydrologic forecasts in the literature (Biondi et al., 2012). A contingency table provides additional insight on whether the ability of the model to inform the correct decision is independent of the decision generated by the synthetic reality. For the simple binary decision from Figure 4, the contingency table is shown in Table 1.

Table 1. Contingency table example for binary decisions

		Synthetic Decision		
		A	B	Total
Model Decision	A	5	1	6
	B	1	3	4
	Total	6	4	10

The contingency table allows us to inspect the proportion of correct decisions made as a function of the synthetic reality generated. Ideally, the proportions of correct decisions informed by the model, separated by what the synthetic decision was, would be approximately equal, indicating that the model's ability to inform the decision is not a function of the synthetic reality generated. A large difference in these proportions would indicate a bias as a function of the synthetic decision, which is likely to negatively impact the ability of the model to inform the correct decision. In this example, the model correctly informs the decision 5/6 times for decision A and 3/4 times for decision B, which does not suggest any particular systematic bias from to the synthetic reality generated, although we would likely want more than 10 trials prior to drawing any conclusions here. Formal statistical tests to determine if the difference in proportions is significant are also available.

The individual model skill is the difference in decision quantity of the model and the synthetic reality, or expressed mathematically as $\Delta\phi = \hat{\phi}_n - \phi_n$ for a given trial. A smaller model skill means that the model decision quantity is closer to the synthetic decision quantity in a given trial, which can be discussed in average terms over the entire DCT experiment as well. Over a set of *N* trials, the histogram of model skill can be generated, as depicted in Figure 5. This distribution of model skill is referred to as the *discrepancy distribution*, since it characterizes the discrepancy between synthetic reality decision quantities and model-based decision quantities.

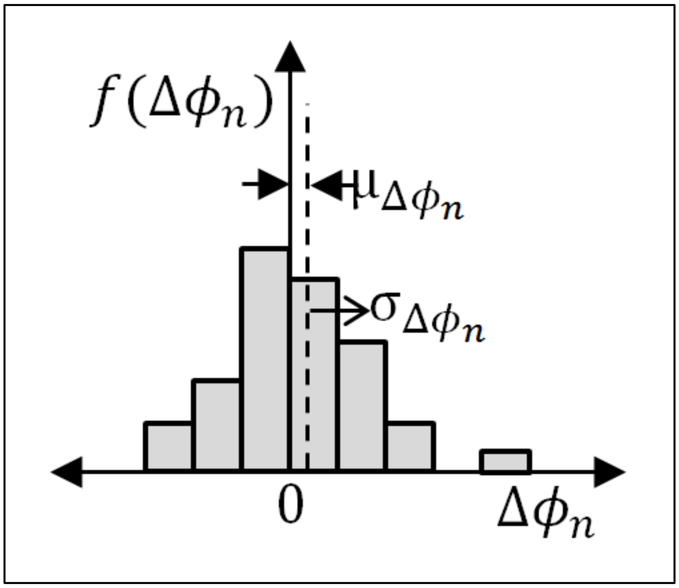

Figure 5. An example discrepancy distribution which is just the sample distribution, or histogram, of model skill from a DCT experiment

The discrepancy distribution is discussed here in terms of only mean and standard deviation for ease of explanation, although no assumption is made on the nature of this distribution. The discrepancy distribution is characterized by a mean value, $\mu_{\Delta\phi}$, and a deviation, $\sigma_{\Delta\phi}$, which are analogous to model bias and model precision in a decision-making context, respectively. Ideally, the discrepancy distribution will have (a) a mean value that is close to zero, representing a small bias in the decision quantity prediction of the model, and (b) a small deviation from the mean, representing a small variation between the synthetic decision quantity and the model-produced decision quantity in any given trial. In a decision-making context, a skilled model will be both unbiased and precise if both $\mu_{\Delta\phi}$ and $\sigma_{\Delta\phi}$ have near-zero values. A skilled model will also be represented by a near zero value in the mean absolute model skill, $\mu_{|\Delta\phi|}$.

Finally, a metric called *decision difficulty* is presented, which encapsulates the relative difficulty of the decision at hand. For a binary decision, the decision difficulty may be formulated as:

$$D_d = \frac{1 - |A - B|}{1 + |A - B|} \tag{2}$$

where D_d is the metric for the difficulty of the decision, A is the proportion of synthetic samples where decision A is the correct decision, and B is the proportion of synthetic samples where decision B is the correct decision (Craig, pers. comm., 2017). This is determined for the theoretical distribution as $N \rightarrow \infty$. In practice, A can be computed as:

$$A = \frac{\#\{\phi_n < \phi^*\}}{N} \tag{3}$$

where $\#\{\phi_n < \phi^*\}$ is the number of synthetic decision quantities less than the decision threshold (ϕ^*), and N is the number of synthetic reality samples. Similarly, B can be computed as:

$$B = \frac{\#\{\phi_n > \phi^*\}}{N} \tag{4}$$

The decision difficulty is bounded between 0 and 1, where 0 represents an easy decision in which decision A or B is the correct decision in each synthetic reality sample, and a value of 1 represents the most difficult decision, where the proportion of synthetic samples for decision A and decision B is equal. This is a useful metric specific to nature of the decision problem only (not the model being evaluated) to consider when discussing the required model skill. It is possible for the model skill to be poor but still achieve a high similarity score if the decision is an easy one, implying that the model is still adequate. The opposite is also possible, where the model skill is relatively good but the similarity score is poor, if the decision is a difficult one.

One limitation of the decision difficulty metric is that it does not take into account the distribution of the synthetic decision quantities with respect to the model skill required of the model. This means that the decision difficulty metric cannot distinguish the decision difficulty between, for example, a uniform distribution of decision quantities and a normal distribution, with a decision quantity threshold at the centre of each distribution. The uniform distribution would have a larger number of decision quantities farther from the decision quantity threshold than a normal distribution and thus represent an easier decision; however, this would not be reflected in the decision difficulty metric. This may be overcome by computing the *average maximum allowable model skill (AMAMS)*, as the average distance from the decision quantity to the decision quantity threshold. This may be calculated as:

$$AMAMS = \frac{1}{N}\sum_{n=1}^{N}\left|\phi_n - \phi^*\right| \tag{5}$$

where ϕ_n is the synthetic reality decision quantity, ϕ^* is the decision quantity threshold, and N is the number of trials. This metric can be interpreted as the maximum absolute model skill that the evaluated model must be capable of before it fails to inform the correct decision, on average. This is a useful metric to present alongside the decision difficulty in order to account for the distribution of the synthetic decision quantities, and for setting a reference model skill for later interpretation of DCT results. Since the AMAMS is computed using the absolute values of model skill. In the interpretation of the AMAMS it may be more useful to compare the mean absolute model skill, $\mu_{|\Delta\phi|}$, rather than the mean model skill, $\mu_{\Delta\phi}$.since the AMAMS is itself computed using the absolute values of model skill.

3.1.2.1 DCT Similarity and Decision Models in Non-Binary Decisions

The metrics presented in the previous section can be generalized for non-binary decisions. In the non-binary case for k plausible decisions, $k-1$ decision quantity thresholds can be defined to divide the decision space into k sections. This is shown in Figure 6 for the case of $k = 3$.

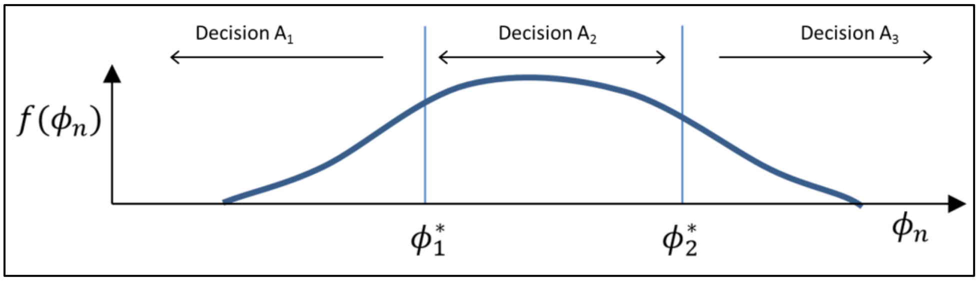

Figure 6. Decision quantity distribution from synthetic reality generation for non-binary decisions

The similarity score for the non-binary case is calculated as:

$$S_s = \frac{C}{N} \tag{6}$$

where *C* is the number of occurrences in *N* trials where ϕ_n and $\widehat{\phi}_n$ are in the same region of decision space, i.e., where they result in the same decision. In this more general case of *k* decisions, the similarity score could be compared to a value of *1/k*, i.e., the chance of informing the correct decision purely by chance in *k* decisions, if all decisions have equal probability of being selected. The threshold for considering a similarity score to be acceptable would depend on discussions between decision-makers, stakeholders, and modellers.

The more general formula for decision difficulty, in a case with *k* discrete decisions in the set of *A* (i.e., plausible decisions $A_1, A_2, \ldots, A_k$), is presented below.

$$D_d = \frac{(k-1) - \sum_i |A_i - A_j|}{(k-1) + \sum_i |A_i - A_j|} \quad i \neq j, j \in [1 \ldots k] \tag{7}$$

This formulation allows the decision difficulty to remain bounded between [0,1] for any case of *k* discrete decisions. The computation of the AMAMS is also adjusted from the non-binary case, and is calculated as:

$$AMAMS = \frac{1}{N}\sum_{n=1}^{N}|\phi_n - \phi_i^*|, \quad \min\{|\phi_n - \phi_i^*|\}, \quad i \in \{1 \ldots k\} \tag{8}$$

where ϕ_i^* is the nearest decision quantity threshold to ϕ_n in decision space. Note that the model skill definition does not change for non-binary decisions since the model skill is calculated independent of the number of decision quantity thresholds.

The next section provides a discussion of other considerations in the DCT framework, which are applicable to both binary and non-binary decisions.

3.1.3 DCT Considerations

The computational implementation of DCT warrants the consideration of specific choices that need to be made in deployment. These choices, each outline in the following sections, include the:

- method of generating synthetic realities,
- analysis of synthetic realities prior to use in the DCT experiment, and
- convergence of DCT metrics.

These considerations are discussed further in the following sections.

3.1.3.1 Synthetic Reality Generation

The generation of synthetic realities is one of the main steps in the DCT framework. The synthetic realities can be generated by any feasible means, such that each synthetic reality generation:

1. Includes sufficient information to determine the ‘correct’ decision, using the established decision formulation, and
2. Generates the synthetic observed data required for building the evaluated model.

There are a few methods for generating the synthetic realities, and more are likely to be developed in future work. A few examples of methods for generating synthetic realities include:

- the perturbation of historical observed data,
- the use of random parameters in an existing hydrologic model, or
- the use of perturbed forcings to an existing hydrologic model.

The perturbation of historical data can be done to create multiple stochastic generations of synthetic observation data, all of which would be relatively similar to the observed data. This technique would only be applicable if the decision formulation is based on the same observed data, such as streamflow measurements. Perturbation of the historical data can be done by (a) adding random error to the existing data directly, or (b) fitting the data with a time series or similar model for random generation. The disadvantages of the perturbation approach are that (a) only the time series of historical data can used in the DCT experiment and used to build the evaluated model, (b) there is no obvious method of determining whether the resulting perturbed synthetic reality is hydrologically plausible, and (c) the range of synthetic realities generated by the perturbation may be limited to those that are relatively similar to the supplied data. Some of these challenges are overcome by use of a perturbed model for generating synthetic realities, rather than perturbed data alone.

In using a model to generate synthetic realities, the use of random parameters and/or perturbed forcings can be used to generate model outputs, which serve as synthetic observed data for the evaluated model. For example, the potential evapotranspiration correction parameter could be randomly varied in the model to produce drier or wetter years of synthetic realities. Alternatively, the relative weighting of meteorological gauges could be randomly varied to produce stochastic forcings and model outputs, e.g., (Sgro, 2016). The advantage of a model-based approach is that (a) the outputs produced are more hydrologically consistent than simply perturbing a time series, and (b) the model can generate multiple outputs simultaneously for use in building the evaluated model (e.g., snow depth and streamflow). It is likely also easier to produce a wider range of plausible synthetic realities by perturbing the model rather than the time series directly since there are more controls over the output series. A check is recommended for each generation to ensure that the model outputs are plausible as a synthetic reality; for example, ensuring that the annual flow volume is within the 10th and 90th percentiles of the observed data, or ensuring that snow is present in the model for a given number of days during winter months for a model of central Ontario, may be reasonable checks on synthetic reality generations.

Since the synthetic reality is ultimately a substitute for reality, which is infinitely more complex than a model, it is more rigorous of a test to use a more complex model than the evaluated model for synthetic reality generations. The use of a more complex model also removes the possibility of matching the synthetic reality state with the model exactly. In the real application of the evaluated model, the data used to build the model will come from a much more complex system (reality) than the one represented in the model, thus the use of a more complex model to simulate reality is recommended. However, the use of a more complex model is not strictly required, since (a) in some cases, building a more complicated model than the evaluated model may not be possible or practical, (b) it may not possible to judge the relative complexity of two models if they are not nested in structure, and (c) the use of a simpler model for synthetic reality generations can still provide useful information via DCT, even if the test is somewhat less rigorous. However, the use of a more complex model than the evaluated model is recommended for synthetic reality generations where plausible and practically feasible.

3.1.3.2 Analysis and Pre-screening of Synthetic Realities

Once the synthetic realities have been generated, they can be analyzed prior to use in the DCT experiment. Typically, the generation of synthetic realities is much less computationally intensive than the model-building portion of the DCT experiment, thus the prior evaluation of synthetic realities is a logical step.

The entire set of synthetic realities generated may be examined and the distribution of decision quantities can be plotted prior to the model evaluation, from which the decision difficulty and AMAMS can be computed. These visualizations, along with the metrics, can provide a useful interpretation of the synthetic realities. Ideally, the decision difficulty is a non-zero value, with some synthetic decision quantities generated on either side of the decision quantity threshold. One way to adjust the synthetic realities is to resample them with a different distribution, which will result in a new decision difficulty. A simpler way to adjust the difficulty of the decision is to adjust the decision quantity threshold, although this may be limited by the physical interpretation of the decision quantity threshold and the method by which it was originally established.

While there are no strict requirements on the desired distribution of the synthetic realities, a uniform distribution is likely the best distribution to achieve to ensure a wide range of conditions are tested in the DCT evaluation. The synthetic realities may be resampled to any distribution, including a uniform one, if a sufficient number of excess realities are generated. A simple way to sample the distribution uniformly is by:

1. Fitting the distribution of synthetic realities with a known distribution, such as a gamma or normal distribution.
2. Obtaining the density function of the fitted distribution.
3. Resampling the synthetic realities, where each has a probability of being sampled equal to the inverse of its density.

Regardless of the resampling technique, the resampled distribution can be re-examined to ensure it is of an acceptable distribution prior to use in the DCT experiment. These manipulations of the synthetic realities may be iterated several times to achieve a desired distribution prior to the DCT experiment.

3.1.3.3 Convergence of DCT Metrics

Another consideration is the evaluation of convergence of the DCT criteria, such as the similarity score and decision difficulty. Ideally, the metrics produced by DCT have converged after a sufficient number of iterations such that the metrics would not change significantly with additional trials, i.e.:

$$S_s^N \approx \lim_{n \to \infty} S_s^n \tag{9}$$

where S_s^n is the similarity score produced after *n* trials. The same convergence can be considered for any of the DCT metrics with *N* trials. In many computational methods, this is determined with some convergence criteria based on a threshold for change in value from trial *N-1* to *N*. For example, if the computed similarity score changes by an assigned threshold value of 0.1% from trial *N-1* to *N*, the similarity score would be considered converged and DCT trials could stop. Convergence is a consideration in many Monte-Carlo based methods (such as Markov-Chain Monte Carlo), and various convergence assessment methods exist (Gentle, 2009). Since the sampling of synthetic realities can be done independently of the model evaluations, this could first be computed for the decision difficulty metric before evaluating the convergence of the similarity score.

3.1.4 Applications of DCT

Thus far, DCT has been presented as a framework for evaluating the ability of a model to inform decision-making. This is the main motivation and use of the DCT framework, as essentially a rigorous model validation method specifically designed for model evaluation in the context of decision-making. However, this section will provide a broader scope of the potential applications of DCT beyond this basic purpose. These potential applications include:

- **Direct evaluation**: is the model good enough to inform our decision?
- **Model failures**: under what conditions does the model tend to fail in informing the correct decision?
- **Model improvement**: how can the model be improved for informing decision-making? Which approach is better?

- **Decision formulation**: how should the decision be posed? What are the impacts of forming the decision one way versus another?

These applications can be used not only for testing individual models and decisions, but to generate rules of thumb for model deployment in general. Another direction of research has the potential to explore the implementation of the DCT method itself, including the impacts of choices made in the DCT evaluation and generating rules of thumb for deployment of DCT as a framework. This section will illustrate that DCT is a much more comprehensive framework with broader applications than model validation alone.

3.1.4.1 Direct Evaluation of the Hydrologic Model for Decision-Making

The direct evaluation of the model for decision-making is the primary application of DCT, which provides a baseline evaluation of the model. However, as discussed in section 3.1.3.2, the synthetic realities generated can be adjusted prior to use in the DCT experiment, with the potential of changing the decision difficulty substantially. Two main approaches are introduced for the direct evaluation of a model for decision-making with a single decision formulation:

1. Generate synthetic realities that are representative of random deviations from normal conditions, with a moderately difficult decision, or
2. Generate two sets of synthetic realities, one generated with a low decision difficulty and another with a maximally difficult decision (i.e., decision difficulty approximately equal to 1), with the purpose of using both sets and bounding the model performance.

The choice of which approach to utilize depends on the risk involved with the decision; for general model performance in decision-making the first approach is appropriate, while a decision with large consequences (such as flood prediction) would likely warrant a bounding on the model ability. In practice, both approaches can be used and the model can be evaluated for three different sets of synthetic realities with increasing levels of decision difficulty, which would provide a better bounding on the model skill in a decision-making context than a single experiment. It is important to note that any assessment of the model skill in a decision-making context is still likely an upper bound on the model’s ability, since the expectation is that the use of the model against reality is more difficult than against the most complex model output. However, this bounding would serve as a useful exercise in the direct evaluation of the model.

In the direct evaluation of a model, the interpretation of the similarity score is most informative when the similarity score is close to 1, indicating a well-performing model, or less than what would be expected with random chance (e.g., close to 0.50 or worse for a binary decision), indicating that little value is added by use of the model. Interpreting similarity scores between these two bounds is more subjective, since it depends on the specific application of the model and the similarity score desired by the decision-maker. However, any relative changes to the similarity score can safely be interpreted as improving or worsening the ability of the model to inform decision-making.

3.1.4.2 Model Failure Analysis

A more specific application of direct model evaluation is to use the DCT framework to investigate under what conditions the model fails to inform the decision correctly. This would be particularly important if the decision is one with a high consequence, such as one related to flood prediction or dam failure. While the model will, in any test, likely fail to inform the correct decision on some occasions due to random chance, the purpose of this application would be to investigate any systematic reasons that a model fails to inform the correct decision.

The general approach to determine the links between model failures and their causes (or at least correlations) to elements of the model-building procedure would be similar to that of Scenario Discovery (Bryant and Lempert, 2010), where a decision itself is iterated and modified until it becomes robust to failure, based on one or more performance criteria. Here, the same approach could be applied to the failures of a model to inform the correct decision.

During the DCT process, an additional step required for this type of application is the collection of metadata during the experiment. For example, the model parameters of both the synthetic reality models and/or the calibrated models undergoing evaluation would be collected, such that any link between poor estimates of model parameters and model failures could be identified. Any other elements of the model-building process that are determined during the DCT experiment should also be collected.

The Scenario Discovery approach could, for instance, make use of clustering algorithms to categorize the model results into failure and non-failure scenarios. A similar technique could be used on the model skill from a DCT experiment, since a large model skill is typically connected to a failure to inform the correct decision. Once the cluster of failure ‘scenarios’ is identified, the analysis would identify commonalities between these scenarios and determine how the decision can be adjusted to be more robust to those scenarios. In the DCT application, the cluster of evaluated models with large model skills would be analyzed for commonalities in order to determine how the model-building procedure could be adjusted to reduce the incidence of those poor model skills. For example, if this step identifies that the built models with poor model skills tend to underestimate a particular parameter, then the parameter range for that parameter could be adjusted in calibration to avoid a poor value. This process could be iterated until the model skill is deemed sufficient to the modeller, or it is not possible to identify common characteristics of models with poor model skills. Results such as these would likewise lead to useful rules of thumb for similar applications without having to go through the effort of applying DCT.

3.1.4.3 Improvement of the Model-Building Procedure

An important use of DCT is to directly test various model-building choices and their impact on the decision-making ability of the model, and thus determine how the model-building procedure can be improved to support decision-making, both for the specific model and for similar models or decisions. This could be done by simply running two or more DCT experiments with the same synthetic reality realizations but different model-building procedures, and comparing the results in terms of similarity score and other metrics. For example, to test the impact of the calibration algorithm on the ability of the model to inform decision-making, one experiment would be run with algorithm A and another with algorithm B; a difference in similarity score would indicate a difference in decision-making by the calibration algorithm, which could be evaluated using basic statistical tests. The same could be done for any model-building decision, such as the:

- Level of watershed discretization, e.g., lumped model vs 25 subbasins vs 250 subbasins;
- Use of various forcing data sources;
- Treatment of different hydrological process algorithms;
- Size of the calibration budget;
- Number or selection of parameters in calibration;
- Different objective functions in calibration.

As an additional example, a recent emergence in the calibration literature is the use of hydrologic signatures in calibration over the more common fit metrics, such as root mean square error and Nash Sutcliffe, e.g., (Hingray and S., 2010, Pokhrel et al., 2012, Shafii and Tolson, 2015). A set of DCT experiments in which one experiment uses hydrologic signatures in calibration, and a second uses root mean square error, could reveal whether the use of signatures has a significant impact on the ability of the model in a decision-making context. This demonstrates the potential for DCT to be used in providing evidence of meaningful improvements in the model-building procedure for decision-making applications.

3.1.4.4 Evaluation of the Decision Formulation

The definition of model-decision mapping, or in other words the decision rules that explicitly state how the decision will be based on model output, is a non-trivial task in the decision-making process. An important potential application of DCT is to test the impact of different decision formulation on the ability of the model to inform the correct decision. It may be possible that for a given decision set of *k* discrete decisions, a model will inform the correct decision more frequently if the decision quantity thresholds are adjusted, or the decision quantity is defined differently (i.e., based on a different model output(s)). DCT provides a framework to test various decision formulations directly.

The direct testing of the impact of a decision formulation is not commonly done in the literature, although a good example of this is found in the *de Novo* planning framework, where a global sensitivity analysis on the decision variables is used to iteratively test different decision formulations (Kasprzyk et al., 2012).

3.1.5 Analysis of DCT Method Choices

While the DCT framework is empirically sound, there are a number of things to be warned about regarding the robustness and reliability of the method itself. Research topics on DCT itself may include (Craig, pers. comm., 2017):

- Evaluating the relationship between the similarity score and its drivers, including model skill, model bias, the decision difficulty, and the synthetic reality sampling scheme;
- Testing various algorithms and strategies for sampling synthetic realities, including the impact of sampling an insufficient number of synthetic realities;
- Assessing the impact of measurement error on the ability of the model to inform decision-making, which can be done by adding an error filter function to the observed states generated by the synthetic reality for use in the model-building (similar to the filter applied in (Crow and Van Loon, 2006));

- Demonstrating under what conditions the similarity scores can be used to discriminate between the decision-informing skill of two competing models;
- Examining the conditions under which the model skill is dependent or conditional on the decision quantity.

These issues will not be addressed directly in this thesis but are, rather, open questions.

3.1.6 Limitations of DCT

The current DCT methodology, as presented, has a number of limitations. One of the main limitations is the subjectivity and lack of general guidance in generating synthetic realities for evaluating models, which has a large potential to influence overall DCT results. The discussions on this step in DCT thus far suggest that this limitation can be overcome by (a) the use of multiple synthetic reality sets in the DCT experiment, (b) the use of multiple decision formulations, and (c) by full reporting of the synthetic reality metrics, which provide a measure of the relative difficulty of the decision from the synthetic realities and decision formulation. Providing more guidance on the generation of synthetic realities is likely to be a research focus in future work.

Another major limitation lies in the types of model-building procedures that can be tested with DCT. In order for a model-building procedure to be tested with DCT, the model-building procedure must be setup programmatically such that it can be iterated many times. For example, testing the use of an automatic calibration procedure is relatively trivial, while testing the use of a manual calibration algorithm would be impossible.

Similarly, the decision formulation testable within DCT is limited to a quantitative formulation that can be evaluated automatically. A single decision quantity with a single decision quantity threshold is simple to evaluate programmatically for the purpose of DCT, but is not reflective of how a real decision-making process would proceed. In a real decision, there would be other factors and uncertainties to consider, as well as fuzzy decision areas where a judgement call by the decision-maker is required. These human factors that are clearly part of the decision-making process are difficult to capture in the DCT framework, thus simplified implementations of the process are used. The use of more sophisticated decision formulations that approach the complexity of the decision-making process are also potential areas for future research.

Finally, a practical limitation on the DCT framework is the computational cost of experiments. A typical DCT experiment may require 100 iterations of a model calibration with a budget of 2000 or more runs, which results in 200 000+ model runs without accounting for the runs required in the generation of synthetic realities. This cost increases proportionally for each synthetic reality set that is used. The computational cost of DCT illustrates the need for a fast model and/or sufficient computational resources, and makes the DCT method impractical for slower models.

3.1.7 Comparison of DCT to Other Methods in Literature

The various components of DCT can be found in the existing literature, and in particular there are a number of similarities between DCT and methods from the petroleum industry (Ballester and Carter, 2006, Carter et al., 2006). The problem is framed as an inverse problem, which is simply defined as the process of “inferring model parameters from the past system behavior (e.g., measurements)” (Carter et al., 2006). In these examples, the ability of a model-building process to generate reliable predictions is tested against a synthetic or truth model, although not in the context of informing a decision. The truth model is built using the same physics and resolution as the model undergoing testing, and is produced using a set of fixed, known parameters. The state variables produced by the truth model are presented to the tested model (as a proxy for observed data) for use in calibration. Carter et al (2006) also experimented with adding random error to this ‘observed’ data.

Other examples also exist where a ‘truth’ model is used to assess and improve the modelling process. In a separate area of literature, a ‘synthetic twin experiment’ is used to test the impact of incorrect error assumptions on data assimilation filters in a remote sensing application (Crow and Van Loon, 2006). A similar idea to a truth model also exists in reified analysis (Goldstein and Rougier, 2009), in which improvements to a model are guided by comparison to a hypothetical near-perfect model (which has been ‘reified’ out of abstraction).

In all of these instances, one extra step which is performed in DCT (and not in the methods mentioned here) is the mapping of the model outputs to decision space, such that the model quality is in specific reference to quality for informing decision-making, rather than a general ability to reproduce historical data. This is an important additional step in allowing decision-makers to understand the impacts of uncertainty for their practical purposes, which is not clear when the only discussion is on prediction accuracy. Any method that performs or considers decision-making must take the step to map model outputs to decision space. This can be as simple as prescribing a failure threshold for a single performance criterion, or something more complex. An example of more sophisticated mapping into decision space can be found in decision scaling (Brown et al., 2012) or Iterative Closed Question Modelling (Guillaume et al., 2015).

DCT is thus far unique in merging these two concepts together into one framework: the evaluation of models against synthetic realities (or truth models), and performing the comparison in a decision-making context by mapping the evaluations into decision space. Performing this evaluation in decision space is one of the key features of DCT that makes it more clear and accessible to decision-makers.

3.2 Case Study Background

This section will introduce the background for case studies used throughout this thesis, which are based on reservoir management problems in an Ontario watershed. An overview of the problem motivation, the characteristics of the local hydrology, and the development of the hydrologic models used in the case studies of this thesis are discussed in this section.

3.2.1 Reservoir Management in Canada

The proper management of hydroelectric and flood-control reservoirs is a large body of research, with many articles devoted to the optimization of reservoir releases (Ahmad et al., 2014, Choong and El-Shafie, 2015). Reservoirs are generally operated to consider several different objectives simultaneously, such as flood management, ecological low flow requirements, consumptive uses, and power generation (in the case of hydroelectric dams) (Chang et al., 2014, Hu et al., 2014, Kamodkar and Regulwar, 2014). Canada is home to many lakes and watercourses, and is the second largest producer of hydroelectricity in the world, demonstrated by the fact of 59.3% of the total electricity generated in Canada in 2014 coming from hydroelectric stations (Natural Resources Canada, 2017). As such, the proper management of hydroelectric reservoirs is of particular importance in Canada.

3.2.2 Study Area and Case Study Introduction

The case studies presented in this thesis are based on reservoir management problems in the Madawaska watershed, which is located in Ontario, southeast of Algonquin Provincial Park, and has a total drainage area of over 8500 km^2. The watershed is highly managed, home to a total of 41 reservoirs owned by a variety of government agencies and private companies, the two largest of which are Ontario Power Generation (OPG) and the Ministry of Natural Resources and Forestry (MNRF) (OPG, 2009).

The Madawaska watershed is located on the Canadian Shield, which is known for thin soils and bedrock outcrops, a product of glaciation. Many lakes, marshes and wetland features exist, partially due to glacial rebound and numerous depression storages (Fu et al., 2014). Overland flow seldom occurs with the exception of cases such as bedrock outcrops; otherwise, macropore infiltration through soils and interflow at the soil-bedrock interface are the dominant runoff processes (Buttle and D., 2004, Fu et al., 2014).

This research was initiated by a need on the part of OPG to reliably predict inflows into their reservoirs, particularly to Bark Lake Dam, which is the most upstream dam that is owned and operated by OPG in the Madawaska. One of the main complicating factors in predicting inflow is the presence of 12 upstream reservoirs that are owned and operated by MNRF for flood control, with reservoir operations and release rules that are unknown to OPG. Thus, the hydrology of the area is dominated by the operation of these reservoirs, and the traditional approach of using a standard hydrological model to predict inflows to the dam is not applicable. For prediction of inflows to other OPG-operated reservoirs in the watershed this is less of an issue, since there is co-ordination within OPG between the release from upstream reservoirs and downstream operations.

The other main complicating factor for the management of Bark Lake Dam is the lack of upstream flow data. Many data-driven methods exist in the literature and have been deployed for streamflow prediction, for example, artificial neural networks and time series models (Yaseen et al., 2015). However, these generally require extensive data sets to achieve good results in prediction, which are not available in this case. One possibility for obtaining historical inflow estimates (on which to build a data-driven model) is to use the level pool routing approximation (D'oria et al., 2012), where a simple mass balance on the reservoir is done to estimate inflow values. The daily inflow values can be calculated as:

$$Q_{in} = \frac{dV}{dt} + P - ET - Q_{out} \tag{10}$$

where Q_{in} is the inflow to the reservoir, dV/dt is the change in volume (or storage) of the reservoir determined from stage measurements, P is the precipitation intensity, ET is the evapotranspiration, and Q_{out} is the outflow from the reservoir (all values reported in mm/d). However, the level pool routing approach has been found to be sensitive to errors in the measured stage and outflow, and has the potential to generate spurious oscillations and even infeasible negative inflow values (D'oria et al., 2012). A look at potential causes of measurement errors, such as winter measurements of reservoir stage and issues in estimating outflow from tailwater stage-discharge rating curve, is likely enough to cause suspicion in the accuracy of the day-to-day measurements. The reliability of predictions from a black-box model built using these inflow estimates would be questionable.

The approach presented in this work overcomes these limitations by constructing hydrologic models which are calibrated to the available reservoir stage data, rather than flow data. The details of this are discussed more in the following sections, as well as details on the model structure and model building procedure. The methodology presented in this section is used to produce two models for supporting the DCT case studies, one model is of the Bark Lake Dam subcatchment and the other of the Mountain Chute subcatchment.

3.2.3 Hydrologic Model Structure

The hydrologic model structure was developed to capture the characteristics of the Canadian Shield, namely the importance of interflow, thin soils, depression storage, and fractured rock transport. The models are developed using the Raven hydrologic modelling framework (Raven Development Team, 2017), which is built for flexibility and thus ideal for both crafting a custom model of the Canadian Shield and tweaking the model for two different portions of the watershed without extensive effort.

The model structure was setup to have either bedrock outcrop (modelled with very thin soil) or deeper organic soil (with two soil profiles) in a given subbasin in order to accommodate the Canadian Shield landscape. Both soil types are assumed to be underlain by a fractured rock layer, which provides a transition to impermeable bedrock. The depth of the fractured bedrock layer also acts as an extra calibration parameter, where extra storage in the fractured bedrock can help account for extra storage present in the landscape contributing to flow but not accounted for in the model, such as depressions and wetlands. The conceptual soil profile is shown in Figure 7.

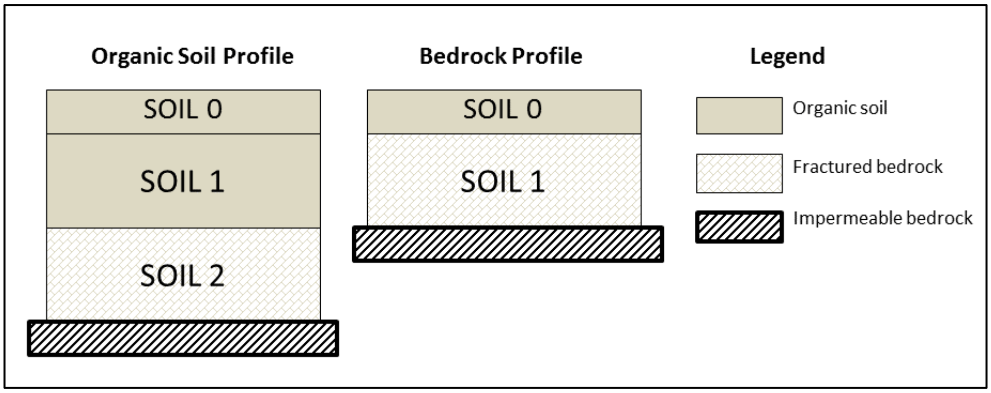

Figure 7. Conceptual diagram of soil profiles in the model

The top soil layer of each soil profile, SOIL 0, acts as a control on the infiltration rate for each soil profile. Infiltration moves water from the ponded water storage (water sitting on the top of the soil profile) into the first soil layer (SOIL 0). Soil evaporation also acts on the top soil layer, which moves water from the soil layer to the atmosphere. From the first soil layer, percolation moves water downwards through the soil profiles. Interflow is assumed to occur at the interface of the fractured bedrock interface and the soil layer above it, i.e., from SOIL 1 in the organic soil profile and SOIL 0 in the bedrock profile. Baseflow occurs from the fractured rock unit, and conceptually contributes a slower response than interflow.

Depression storage is conceptualized with a single storage unit per subbasin, where both the maximum depression storage and the percentage of ponded water that is routed to depression storage are calibratable parameters. This is the primary accounting for the numerous depressions and wetlands that exist in the watershed, although for the purpose of this model they are not connected to streams and other water features via groundwater.

Snow processes are controlled by a snow melt and refreezing algorithm. The snow melt is controlled by a simple snow melt algorithm, where potential melt is calculated using the degree day method and the actual snowmelt is controlled by calibratable maximum snowmelt rate. The snow refreeze is also calculated using a degree day method.

A few other more minor processes and corrections exist in the model, such as reservoir evaporation, canopy storage, and shortwave radiation corrections. The full process diagram for the developed model is shown in full in Figure 8.

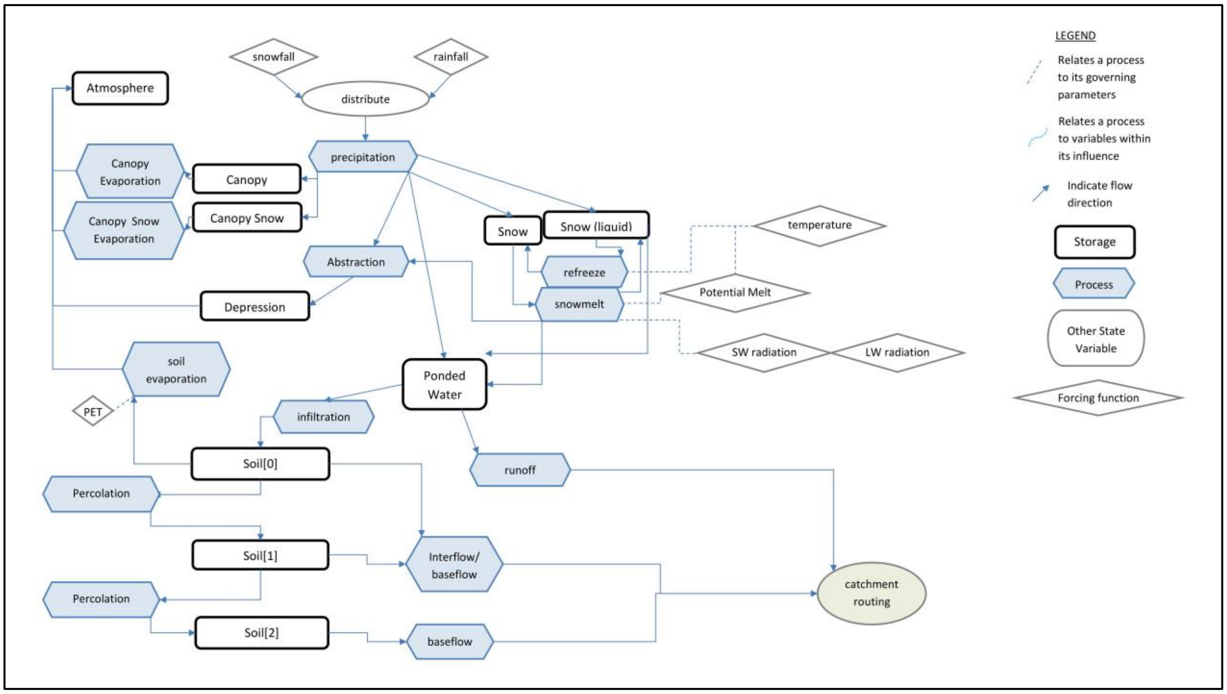

Figure 8. Process diagram for the developed hydrologic model

The model is relatively simple, which is a function of the limited data to support more complex algorithms; for example, detailed snowpack depth and density measurements could justify a more sophisticated set of snowmelt algorithms. This model structure is used in both of the case study models, with some additional configuration based on the presence of reservoirs in the system. The next section discusses the data available to support the model construction, and the watershed discretization.

3.2.4 Data Availability and Discretization

The development of the models for the two case studies made use of several data types and sources, including meteorological data, soil and landuse data, streamflow data, reservoir data, etc. This section will document the sources of the data used, the processing and quality control of that data, and their use in the model development. This is discussed as one for both case studies, since the two case studies both use subsets of the Madawaska watershed.

Meteorological data was obtained from Environment Canada Historical Climate Data, and seven of the closest stations to the Madawaska watershed were used. The stations were interpolated in the model using inverse distance weighting (IDW), such that the importance of each climate station in a given subbasin is a function of its proximity. The stations used in the developed models can be found in Table 2; the climate ID refers to the Environment Canada Climate Identifier key.

Table 2. Meteorological Stations used in Madawaska models

Station Name	Climate ID
Algonquin Park East Gate	6080192
Bancroft Auto	616I001
Chalk River AECL	6101335
Coe Hill	6161740
Haliburton 3	6163171
Ottawa CDA RCS	6105978
Petawawa Hoffman	610FC98

The model was driven using the total precipitation, minimum daily temperature, and maximum daily temperature forcings. These are the minimum forcings required to run a model in Raven; additional information is available from the Historical Climate Data, although only these forcings were used due to a large number of missing data points in the other forcing functions. Some data infilling was still required, since Raven does not allow for missing data points in forcing data. The missing data points for each time series were infilled using data points for the same day at each of the other stations where the data was available, where the data from other stations was weighted with IDW. Since the developed model also uses IDW to interpolate forcing functions, this approach in effect weights the other stations with data heavier on the days of missing data points. On occasion, the IDW infilling approach would produce a minimum temperature larger than the maximum (measured) temperature, or vice-versa. In such a case, the offending infilled value would be removed and instead interpolated using available data from that station only (i.e., interpolate a missing data value in time instead of from other stations).

The landuse data was obtained from the MNRF, and four primary land use types exist in the watershed (three forest types and a water land type). The soil data was obtained from Agriculture and Agri-Food Canada, which has soil profile data for the first 100 cm of soil from the surface. This information was used to generate the soil types in the model, including soil composition for the non-bedrock soil types. However, depth to bedrock data was required to build the model beyond a depth of 100 cm. This data was obtained from the Ontario Ministry of Northern Development and Mines, which provided a set of borehole logs that included depth to bedrock. The borehole logs showed three fairly distinct regions in the watershed where the depth to bedrock was distinctly different, thus the soil classes were additionally split by region in order to allow a better representation of soil depths (Figure 9).

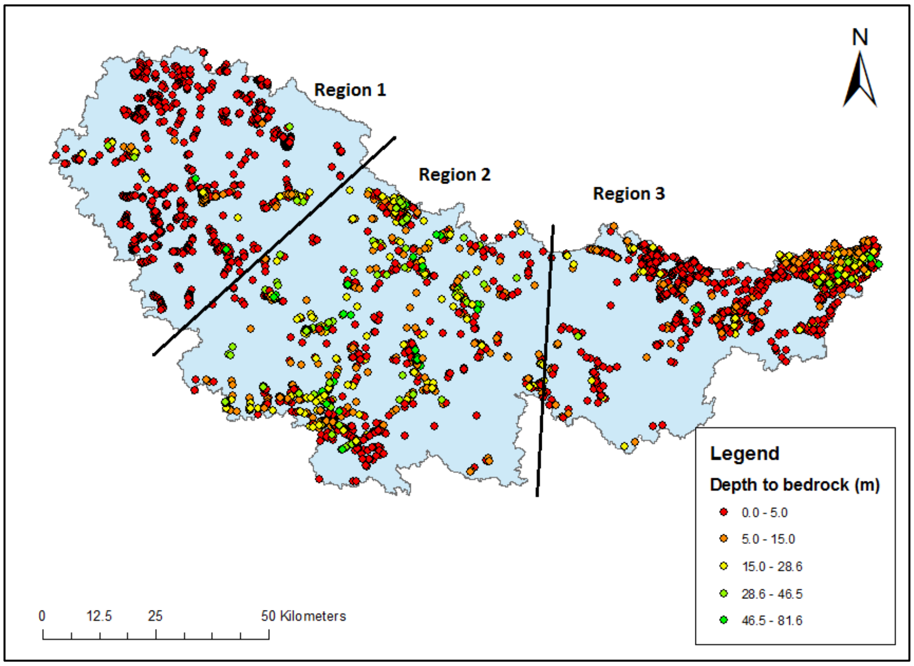

Figure 9. Depth to bedrock map from borehole logs in the Madawaska watershed

The depth to bedrock values were interpolated using IDW for each subbasin, and from there an average depth to bedrock was calculated for each of the soil classes.

Flow gauge data was obtained from both OPG and Water Survey of Canada (WSC). Missing values in flow gauge data were not infilled, since Raven can disregard missing data points in calculating diagnostics. The typical reservoir data available from OPG included reservoir stage and reservoir outflows, estimated from a tailwater rating curve (relating downstream water level to flow). Several rating curves and storage tables were provided by OPG for both OPG reservoirs and MNRF dams, however these were noted to be subject to error (particularly for MNRF dams). It is also noted that no flow gauge data was available upstream of Bark Lake Dam, since no flow or stage data associated with MNRF dams was available (shown in Figure 10).

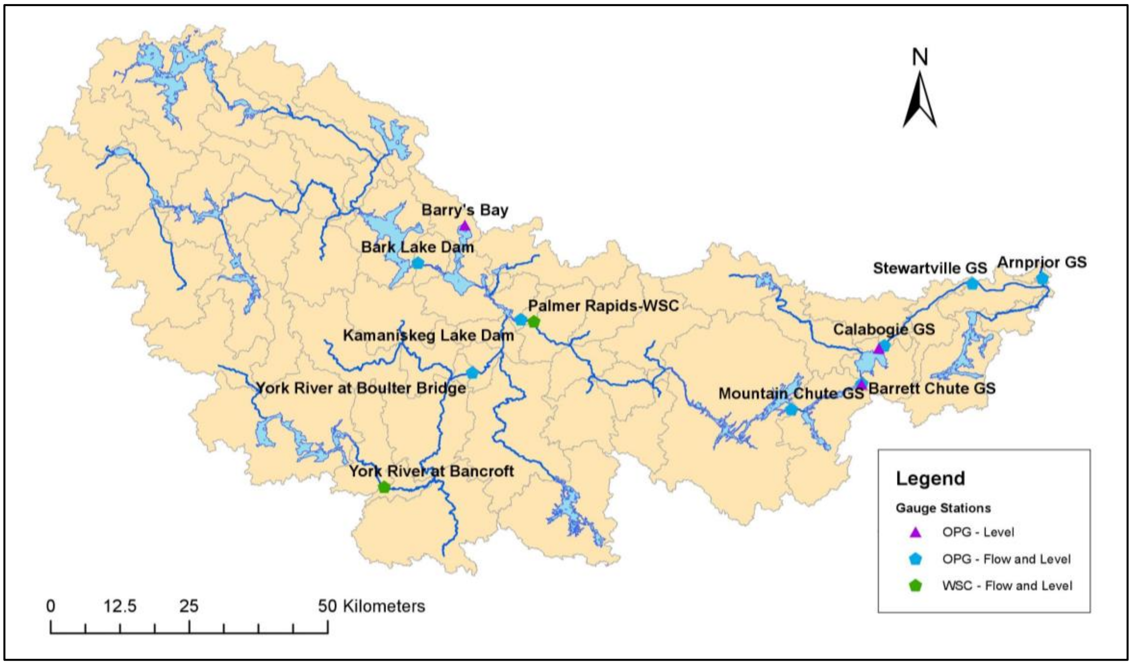

Figure 10. Flow and level data locations in the Madawaska watershed

The landuse, soil type, and elevation data were used to discretize the watershed into a total of 83 subbasins, including 18 reservoirs and 65 non-reservoir subbasins. For simplicity, a single dominant HRU is assumed for each subbasin. The two models for the case studies, the Bark Lake Dam model and the Mountain Chute model, can be shown in the discretized watershed in Figure 11.

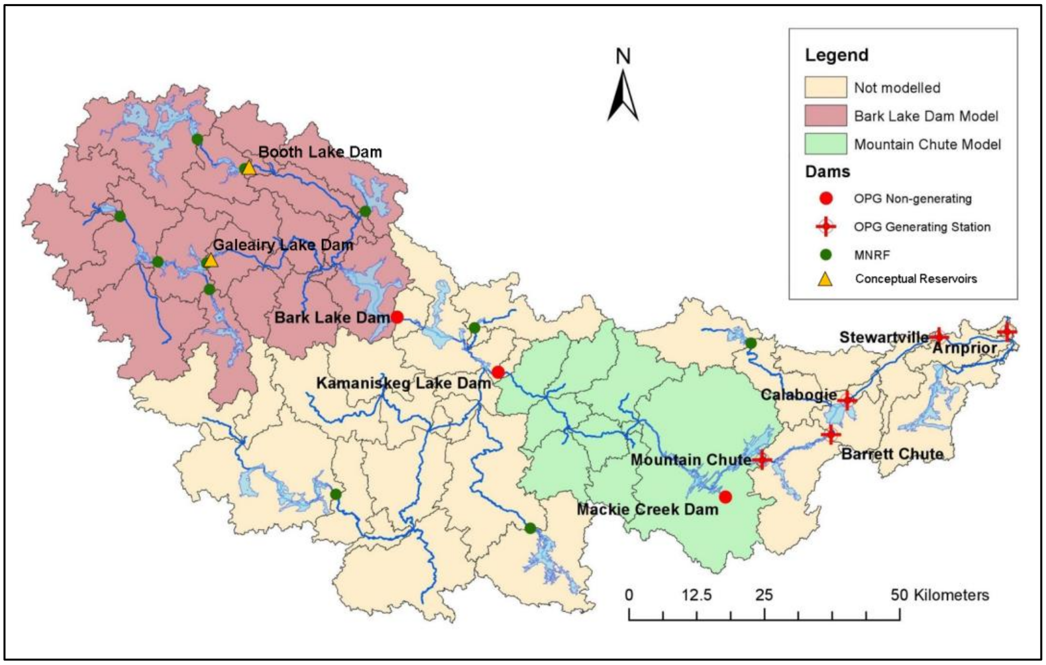

Figure 11. Modelled areas in the Madawaska discretization

One significant difference between the two models is the availability of data and the presence of MNRF dams. In the Bark Lake model, there is no streamflow data available upstream of Bark Lake Dam, and there are many dams controlled by MNRF, upstream of Bark Lake, which are not well characterized by the available information. In contrast, the Mountain Chute model is forced at the most upstream point by a WSC streamflow gauge at Palmer Rapids, and there are no MNRF dams in this area of the watershed. This makes the setup of the Mountain Chute model much more straight forward, since there are no dams to consider, save for Mountain Chute itself (a much smaller dam at Mackie Creek exists, but is ignored to its small drainage area and insignificant impact in comparison to other factors).

3.2.5 Reservoir Representation

An important component of reservoir models is the representation of reservoirs, which dictates the modelled reservoir relationships between stage, volume, area, and outflow. Typically, this relationship is captured using a table of values that explicitly defines the relationship, called a rating curve. The assumed shape of the reservoir can vary, and theoretically any shape or curve representation is acceptable as long as it can be demonstrated to be ‘operationally adequate’, to use the language of Klemeŝ (1986). In the models developed in this work, the reservoir geometry was simplified with a truncated cone representation, which has the advantage of accounting for the sloped sides of reservoirs (rather than a vertical sided representation) while maintaining enough simplicity in the shape to derive simple geometric equations. This geometric representation is shown in Figure 12.

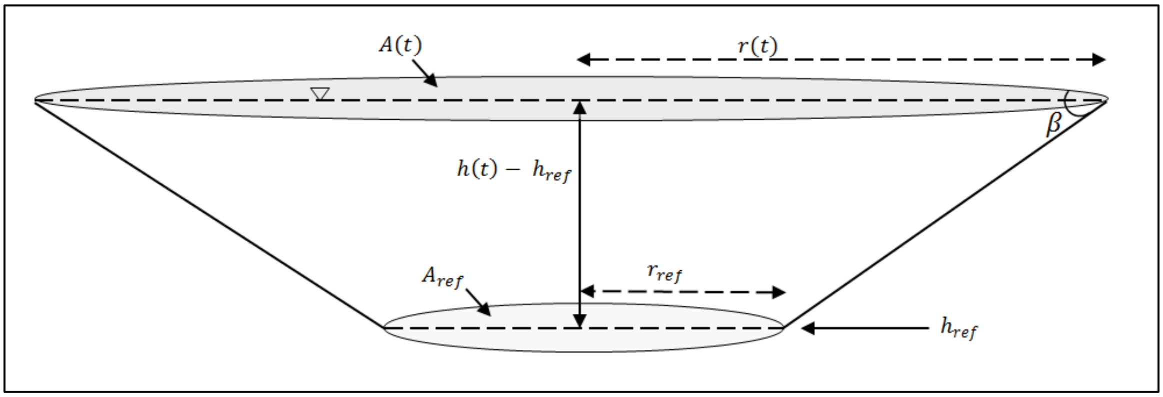

Figure 12. Truncated cone representation of reservoir geometry

Here, A_{ref} is the reference area, h_{ref} is the reference height, r_{ref} is the reference radius, $h(t)$ is the height at some time *t*, $A(t)$ is the reservoir area at time *t*, $r(t)$ is the radius at t, and β is the slope of the reservoir sides (in radians). There is also an associated reference volume of the reservoir, V_{ref}, which represents the volume of the reservoir at the reference height, and a volume of the reservoir at time *t*, denoted $V(t)$. This representation allows the calculation of the area and volume of the reservoir as a function of stage (Liu, pers. comm., 2017). First, the reference radius R_{ref} can be calculated using the reference area, which would be specified from available data and/or from calibration.

$$\mathrm{R_{ref}} = \sqrt{\mathrm{A_{ref}}/\pi} \tag{11}$$

Next, the $r(t)$ can be calculated as:

$$r(t) = R_{ref} + \left[h(t) - h_{ref}\right]/\tan(\beta) \tag{12}$$

From here, the $A(t)$ as a function of $r(t)$, simply as:

$$A(t) = \pi \cdot r(t)^2 \tag{13}$$

The volume at any given point in time, $V(t)$, can be calculated as:

$$V(t) = V_{ref} + \frac{1}{3} \cdot \pi(h(t) - \ h_{ref}) \left({R_{ref}}^2 + R_{ref} * r(t) + r(t)^2\right) \tag{14}$$

Together, these equations fully capture the stage-volume-area relationship. The outflow characteristic of a reservoir is calculated independently of the geometry, since the outflow is controlled. In this modelling setup, the reservoir outflow as a function of the stage at a given time, $h(t)$, is calculated as:

$$Q(h) = a(h - h_0)^b \tag{15}$$

where $Q(h)$ is the outflow as a function of reservoir stage (L^3/T), a is the power law coefficient (L^2/T), h is the reservoir stage (L), h_0 is a reference stage for flow (L), and b is the power law exponent (dimensionless).

Note that the reference stage for the flow rating curve is not necessarily the same as the reference stage in the volume rating curve equation. In the flow rating curve, the reference stage signifies the water level at which outflow begins, which is defined by reservoir operations and not geometry. The reference stage in the volume rating curve defines the 'bottom' of the reservoir as a reference point.

These equations together allow the representation of reservoir properties and behaviour in the hydrologic model. The geometric properties were informed by both the HydroLakes dataset (Messager et al., 2016) and through calibration. The HydroLakes dataset contains information on reservoirs around the globe, including reservoir water elevation, area, slope, volume, etc., and were a useful starting point in estimating the reservoir parameters prior to calibration of the reservoir area.

3.2.6 Model Calibration Procedure

The model calibration procedure deployed here makes use of the reservoir stage data rather than the typical estimated flow data. The only previous use of reservoir stage data directly in hydrologic model calibration found in the literature comes from the building of a rainfall-runoff model in Swaziland (Gijsbers, 2015), making this the first known and documented use of such a procedure in a North American watershed. Typically, inflows are estimated using level-pool routing and the model is calibrated to estimated inflows.

The reservoir outflow data is used to fix the daily outflow from the reservoir, thus the model should accurately reproduce the observed stage values, provided that (a) the model produces the correct inflow series, (b) the reservoir stage-volume relationship is well represented, and (c) the errors in measured data are unbiased. This approach is theoretically less sensitive to measurement errors than calibrating to inflows estimated from level pool routing, since the matching is done to the entire time series of data, and is thus much less influenced by spurious flow peaks caused by errors. It also calibrates directly to raw data (rather than estimated data), thus there is no presence of numerical artefacts beyond measurement error. Here the assumption is made that the data is unbiased, and thus the observed stage can be matched, provided the structure and calibration of the model is adequate.

A procedure was developed in order to calibrate these reservoir models using the available data. In this stepwise calibration, the model is calibrated sequentially to match the inflow volume, reservoir characteristics, snow parameters, and finally the stage data and daily volume changes overall. Each of these steps uses a different calibration metric, and the parameters calibrated in each step are selected as the parameters most influential to that particular metric. This allows the calibration to be done more efficiently with a relatively small number of parameters in each step. The calibration procedure is summarized in the steps below. The parameters calibrated in each step are discussed here briefly; the full list of calibrated parameters in each step can be found in Appendix A.

1. Calibrate the model to the match the inflow volume to the outflow volume in each year of simulation, understanding that there is consistently a near-zero net change in reservoir volume on October 1st of each year. This calibration is done first and is independent of the unknown reservoir relationships, since the incoming and outgoing water from the reservoir are not impacted by the modelled reservoir rating curve. The evaporation and precipitation on the reservoir surface are neglected here in comparison to the much larger flow volumes. The correction factors for potential evapotranspiration, the parameters for canopy interception and storage, and the parameters related to depression storage are calibrated in this step.

2. Adjust the reservoir rating curve parameters, namely the A_{ref} and β, to obtain an approximate solution for the rating curve parameters. This is done prior to the use of metrics that rely on fitting the historical stage data, which is impacted by the modelled rating curve. The starting point for the reservoir parameters comes from both the HydroLakes dataset and OPG data (parameter values provided in Appendix A). In practice, the reference area has much larger impact on the reservoir volume rating curve than the slope, thus β was not calibrated (the original values from HydroLakes were kept). A correction factor for potential evapotranspiration on lakes was also calibrated in this step.

3. The parameters related to snow and snowmelt are calibrated for the winter period, with a metric based on the daily change in volume in the reservoir during the winter months. The

metric is setup to penalize poor freshet response in the model, which is a key concern for reservoir operation. The calibrated parameters include the melt and refreezing factors, a correction factor for snowmelt in forests, and the temperatures for calculating the partition of rain and snow.

4. A more general set of parameters that control timing of flows, mainly soil parameters related to infiltration, interflow, etc. are calibrated. The metric used is based on the match to reservoir stage as well as the daily changes in reservoir volume.
5. An expanded version of the set of parameters in Step 4, which includes the soil routing parameters and a number of vegetation parameters, are calibrated as a final calibration run.

This procedure had to be iterated a few times in order to achieve a good fit to the data. It is noted that these steps are not fully independent; for example, the calculation of daily reservoir volume changes, used in Steps 3-5, requires a good estimation of the stage-volume rating curve in order to establish the stage-volume relationship. The details of the metrics used in each step are presented in Appendix A.

One issue that was identified in the calibration to reservoir stage is the long memory of reservoir stage; in other words, any errors in reservoir stage at one point in the simulation is propagated forward, which can create issues in matching stage at a later point in the simulation even if the inflow estimation was perfect. This is in contrast to the use of flow for calibration, since flow has a much shorter memory, and errors in flow estimation at one point in time do not necessarily impact the ability to match later flow events. The long memory issue can also impact the use of a warm-up period, since it becomes more difficult to ignore large errors in the stage simulation of the warm-up period, when results are typically ignored. There are a number of potential solutions to the long memory issue, such as (a) weighting later stage data heavier in the calculation of diagnostics, since the long memory implicitly makes the earlier time period more important in stage matching; (b) fixing the simulated stage to the measured stage periodically throughout the simulation; or (c) using metrics that use the stage data without fitting the stage explicitly, such as daily volume-change comparison metrics. In this work both (a) and (c) are used to reduce the impacts of the long memory in reservoir stage, however the second approach has been used in other studies, e.g., (Lin et al., 2015).

Additional effort was required in the Bark Lake Dam model to account for the presence of the upstream reservoirs. The problem in calibrating the upstream reservoirs is that there is little information on the reservoirs themselves and no information on the reservoir outflows. The large number of upstream reservoirs also creates a large number of parameters for each reservoir, creating a more difficult calibration problem. The large number of parameters were reduced by first consolidating the upstream reservoirs and simply representing all of the upstream reservoirs with only the Booth and Galeairy reservoirs, which are the two largest dams (in terms of drainage area) with the shortest upstream distance to Bark Lake. The next step was to use the HydroLakes data (and OPG rating curves) to fix the h_0 and h_{ref} parameters of these two upstream reservoirs, and finally to manually calibrate the other upstream reservoir parameters for volume and flow relationships. The upstream reservoir parameters were chosen to (a) reduce the peak flows from the reservoirs as much as possible, and (b) release approximately the same volume of water as enters the reservoir each year, meaning that stationarity in reservoir stage (and volume) from year-to-year is maintained. This calibration is based on the knowledge that the upstream MNRF dams are flood control dams (and thus presumably minimizing peak flows), and the assumption that the reservoirs have relatively consistent volume storage from year-to-year. The assumptions are considered the best possible information available regarding the upstream reservoirs to guide a manual calibration.

The Bark Lake and Mountain Chute models were both calibrated using the entire period of data (from 2007-10-01 to 2015-09-30), excluding a one-year warm-up period. In each calibration step, the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007) was deployed. The entire period was used in calibration in order to develop the best possible fit for use in DCT, although a set of split-sample tests was also performed as a check on model performance. This is discussed with results in the following section.

3.2.7 Model Results and Use in DCT

The calibrated model results for the two subcatchments are presented here, showing the simulated and observed stage for the entire simulation period. The results for the Bark Lake Dam model after undergoing calibration are shown in Figure 13.

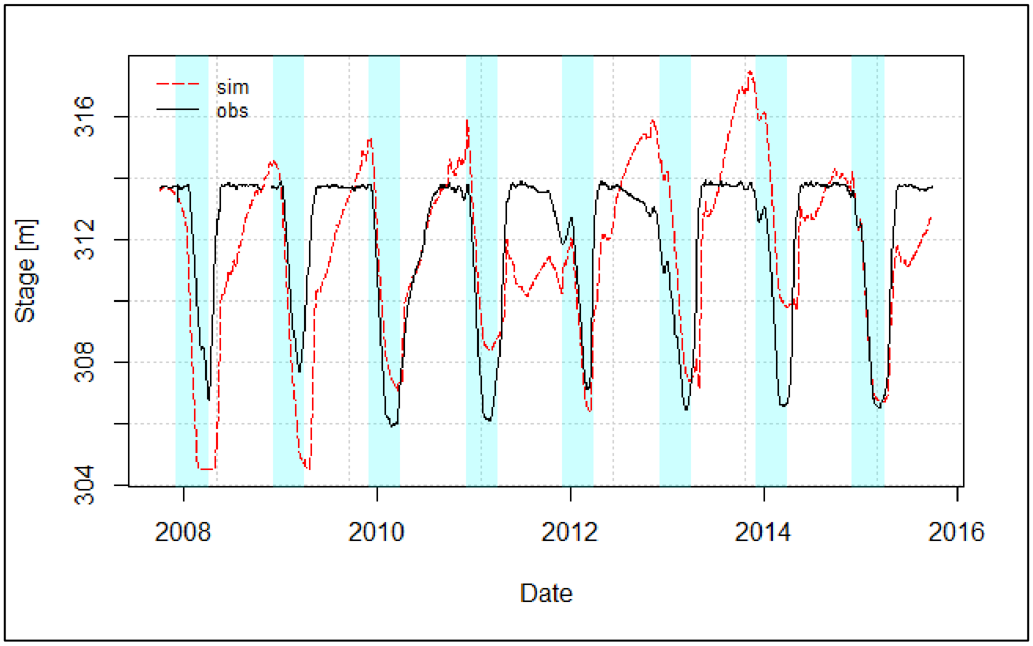

Figure 13. Calibrated model results for Bark Lake Dam inflow model (blue shading indicates the December 1st to March 31st winter period)

The results indicate in Figure 13 show a fair fit to reservoir stage with a Nash-Sutcliffe of 0.23 and an R^2 of 0.49, particularly considering that the simulation was continuous and errors in reservoir stage were propagated throughout. Also note that Nash-Sutcliffe metric for stage compares the simulated stage values against a mean stage value, which matches the stage well during many months of the year, resulting in a relatively low metric. As an additional check, the flow to Bark Lake Dam produced by the calibrated model and the flows estimated by OPG using level pool routing are compared; this is done as a general check on the model (Figure 14). The level pool routing inflow estimates are by no means ‘measured’ data, however, an agreement between the two flow estimates still serves as a good indication of model performance. The two flow estimates show a relatively favourable comparison in Figure 14, with an NSE of 0.41 and an R^2 of 0.46 (using the level pool routed flows as ‘observed’ data in the calculations). This suggests that the model is capable of reproducing the inflow series to some degree of agreement. However, the baseflow signature between the modelled and level pool routed flows are often in disagreement, indicating some issues here; this is likely attributed to the representation of upstream reservoirs, which causes the steady decline in the simulated flows at low flows.

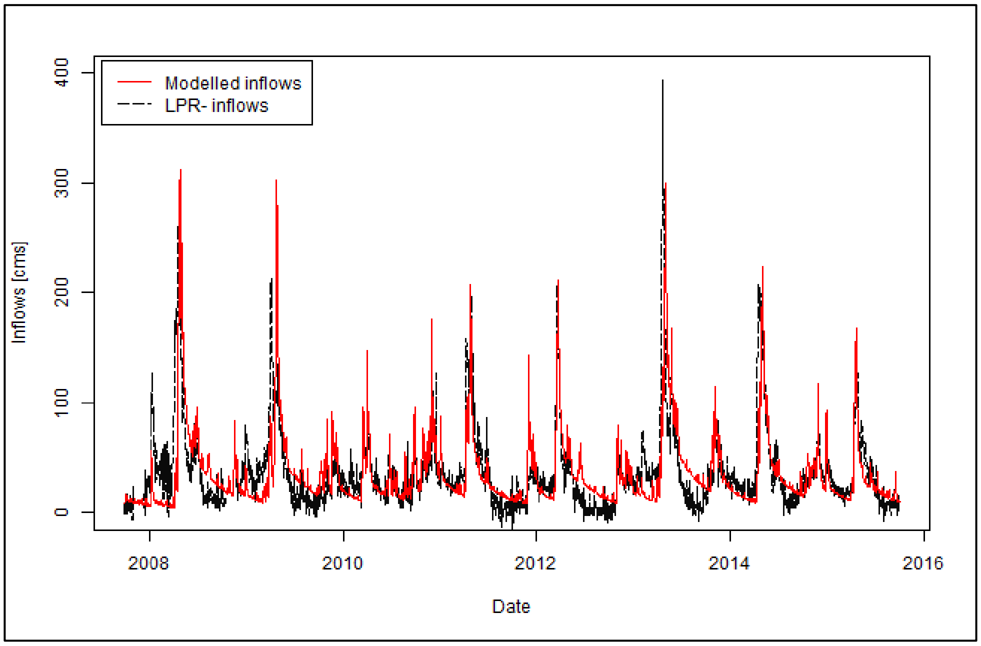

Figure 14. Inflow comparison from calibrated model and level pool routing for Bark Lake Dam model

As an additional validation test on the model, a split-sample test was applied, splitting the data from 2007-10-01 to 2011-09-30 and 2011-10-01 to 2015-09-30, and calibrating the model using the first data period only. The plot of the simulated stage for both periods is shown in Figure 15.

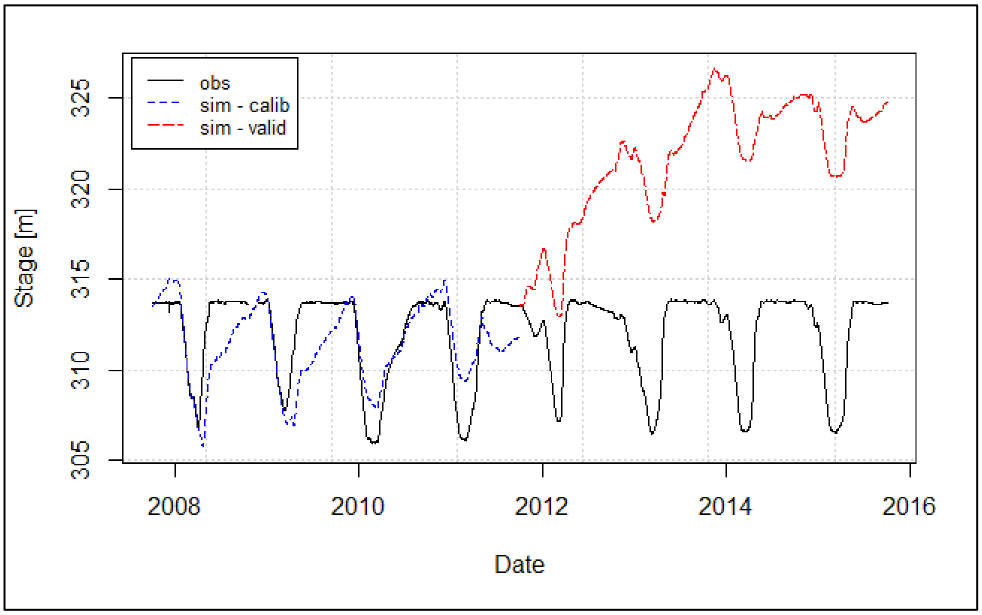

Figure 15. Bark Lake model split-sample test results for reservoir stage

The stage plot shows that the stage is relatively well-fitted in the calibration period but is significantly off in the validation period. Errors accumulate particularly in the 2012-2013 period and persist in subsequent years, due to the long memory of stage in a continuous simulation. However, the flows during the validation period are still fairly consistent with the level pool routing estimated flows, with a Nash-Sutcliffe of 0.53 and 0.59 in the calibration and validation periods, respectively. This agreement is shown in Figure 16.

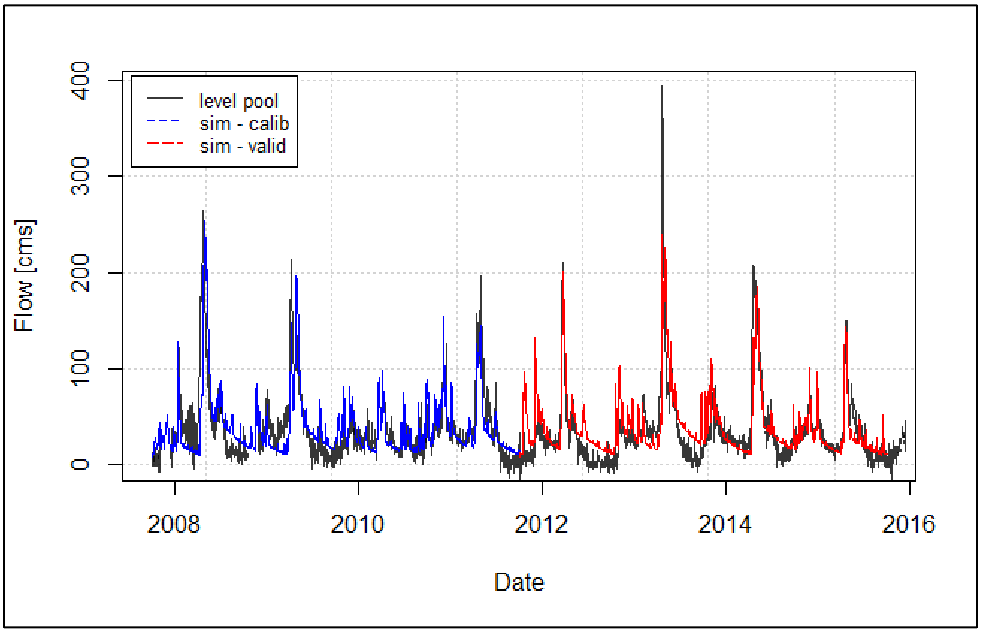

Figure 16. Bark Lake model split-sample test results in flow comparison

The simulated flows show a general overestimation of baseflow and underestimation of peak flows in comparison to the level pool routing flows. The relatively increased baseflows and decreased peak flows are likely due to both the explicit reservoir representation in the hydrologic model (which attenuate peak flow and increase low flows), and the artefacts of level pool routing, which tend to produce spurious peaks. Nonetheless, the model fit to flows is much better than would be expected after examining the reservoir stage fit alone. This is another example indicating that the long memory of stage can produce poor fits to reservoir stage without large detriment to the simulated flows.

The performance of the stage-fitted model can be compared to the results obtained by using the typical approach, i.e., estimating the inflow series from level pool routing and calibrating the hydrologic model to those flows. The reservoir volume change is calculated from the same stage-volume rating curve used in the Bark Lake model above. The uncorrected inflow series calculated using level pool routing on Bark Lake Dam (without inclusion of the precipitation and evapotranspiration terms) is shown in Figure 17.

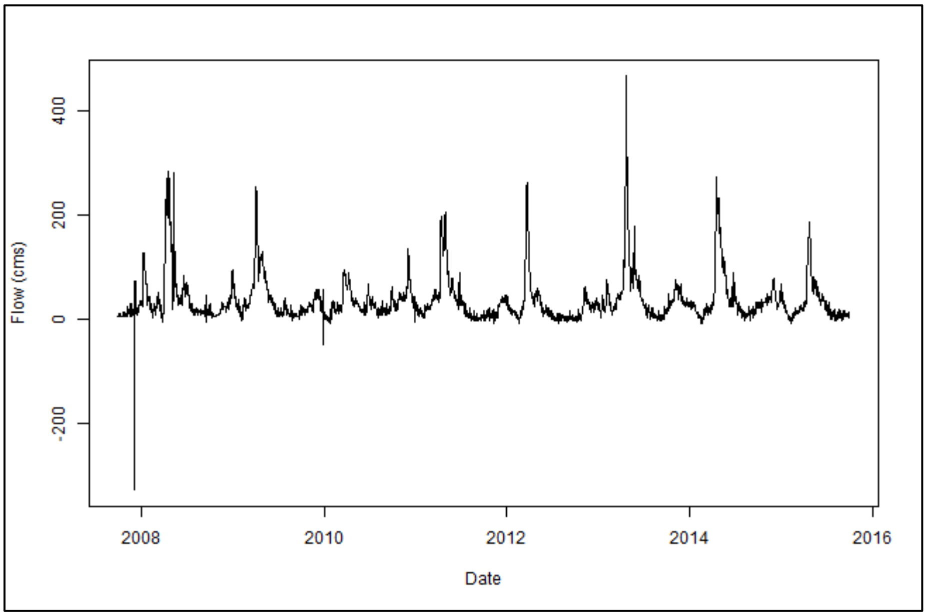

Figure 17. Uncorrected level pool routing-estimated inflow series at Bark Lake Dam

The level pool routed flows have a number of unrealistic characteristics, particularly the inflow of -327 m^3/s on December 6th, 2007, and the other smaller inflows in the series (comprising approximately 4% of the time series). There is no obvious method to determine whether the peaks are spurious or not, however spurious peaks are a known characteristic of the level pool routing method. This inflow series is corrected with the basic approach of setting all negative flows to a value of zero, and this corrected series is used to calibrate the hydrologic model. The model is calibrated with a budget of 10 000 runs using the DDS algorithm, and the calibration objective is set as the root-mean square error of the observed and simulated flows. The parameters included in the calibration are the same as those in Table C1 of Appendix C. The inflows generated from the stage-fitted model and the level pool routing-fitted inflows are shown in Figure 18.

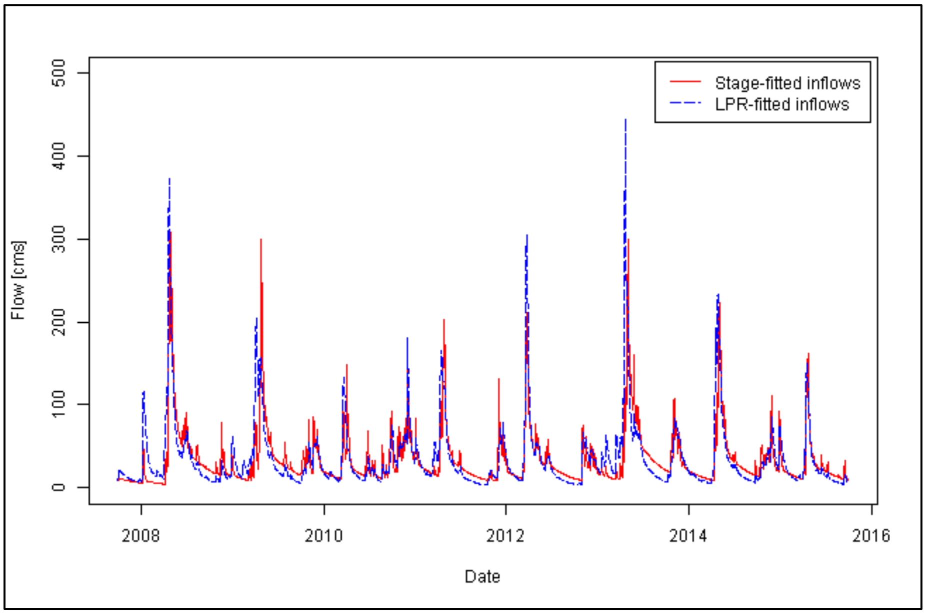

Figure 18. Comparison of stage-fitted and level pool routing-fitted inflows

The flows from both models are relatively similar with an R^2 of 0.63 and a volume difference of less than 2% (level pool routing inflow series has slightly larger volume overall). The base flows are generally larger in the stage-fitted model inflows, and the peaks are higher in the level pool routing model. Since there is no observed inflow data to compare to, no conclusion can be drawn from this plot regarding which provides the better fit, although it may be suggested that larger baseflows and smaller peaks are more likely in a reservoir-dominated watershed (since the reservoirs are operated for flood control), and the higher peaks in the level pool series may be spurious. The performance between these two inflow series is also examined using the simulated stage associated with each inflow series, which does have observed data for comparison. The difference in reservoir stage of these two inflow series is compared in Figure 19.

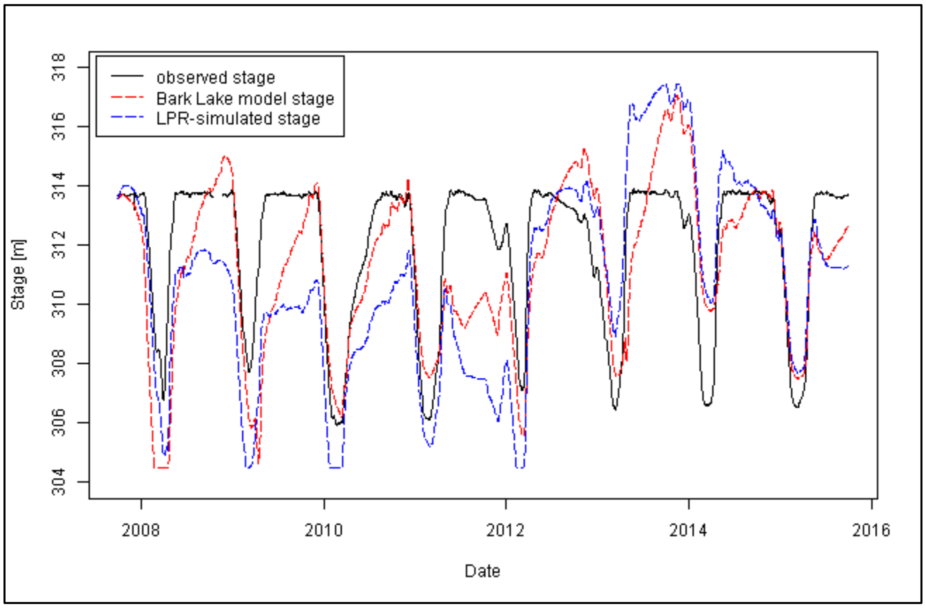

Figure 19. Comparison of Bark Lake model stage and level pool routing simulated stage

The stage-fitted flows provide an overall better fit to the observed stage than the level pool routed stage, with a Nash-Sutcliffe and R^2 of 0.23 and 0.49, respectively, while the level pool routed stage have a Nash-Sutcliffe and R^2 of -0.38 and 0.37, respectively. This demonstrates that the stage is somewhat better fitted when the stage data is used directly in calibration, which is not surprising.

The Mountain Chute model was also calibrated using the stage data. The calibrated model results of the Mountain Chute model are shown in Figure 20, and the comparison of simulated and level pool routing flows in Figure 21. The stage fit has a Nash-Sutcliffe of -0.01 and an R^2 of 0.43. The flow comparison for the Mountain Chute model is particularly favourable; the Nash-Sutcliffe and R^2 of the inflows are both calculated as 0.91.

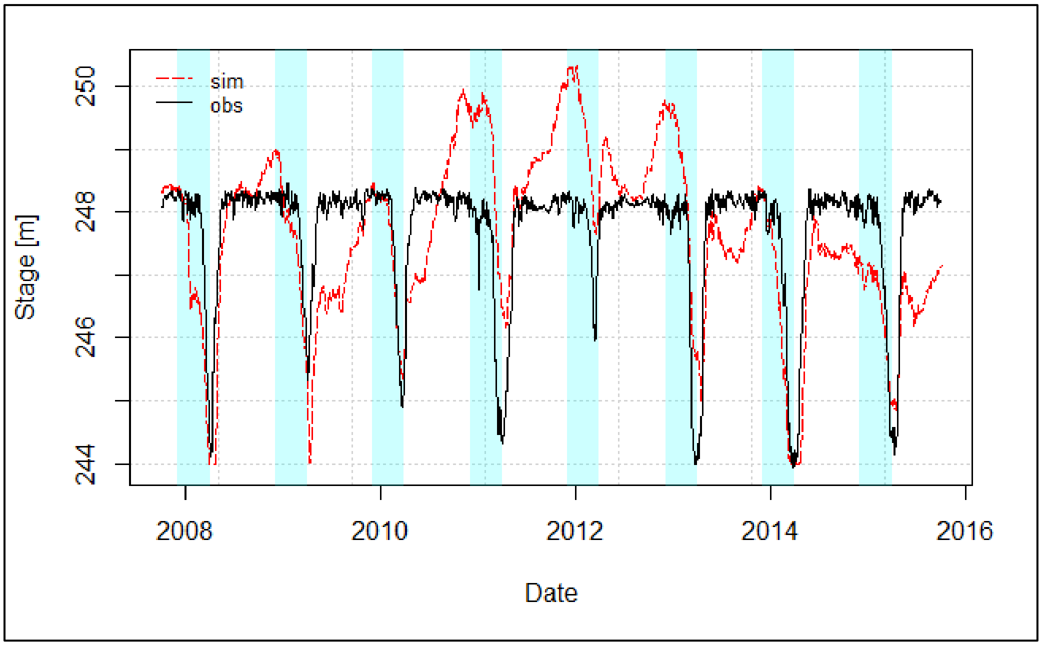

Figure 20. Calibrated model results for Mountain Chute model

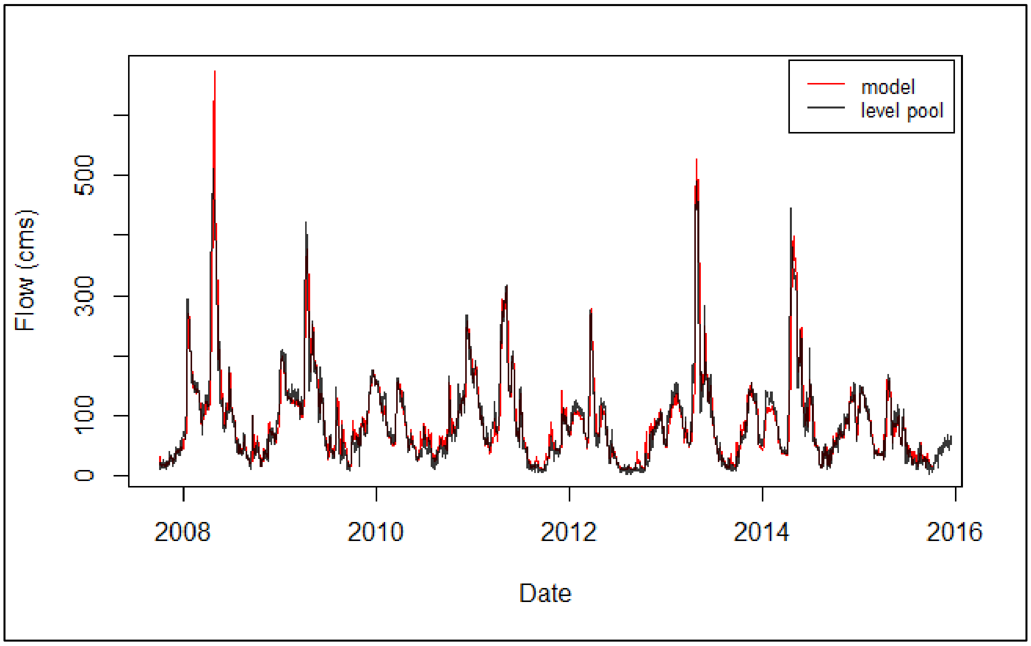

Figure 21. Calibrated model results for Mountain Chute model

The reason that the inflows are so close in the Mountain Chute model is due to the forced inflows from measured data in the most upstream point of the model, which is the dominant signal in the simulated inflow to the Mountain Chute reservoir (i.e., only a small amount of water is added to the simulated inflows by the landscape, compared to the magnitude of upstream reservoir releases). This makes the matching of inflows from level pool routing less impressive than the Bark Lake model, since the naïve model of assuming that forced subcatchment inflows are equal to inflows at Mountain Chute obtains a Nash-Sutcliffe of 0.86 and an R^2 of 0.91. The signal comparison from upstream released flows and model simulated inflows is shown in Figure 22.

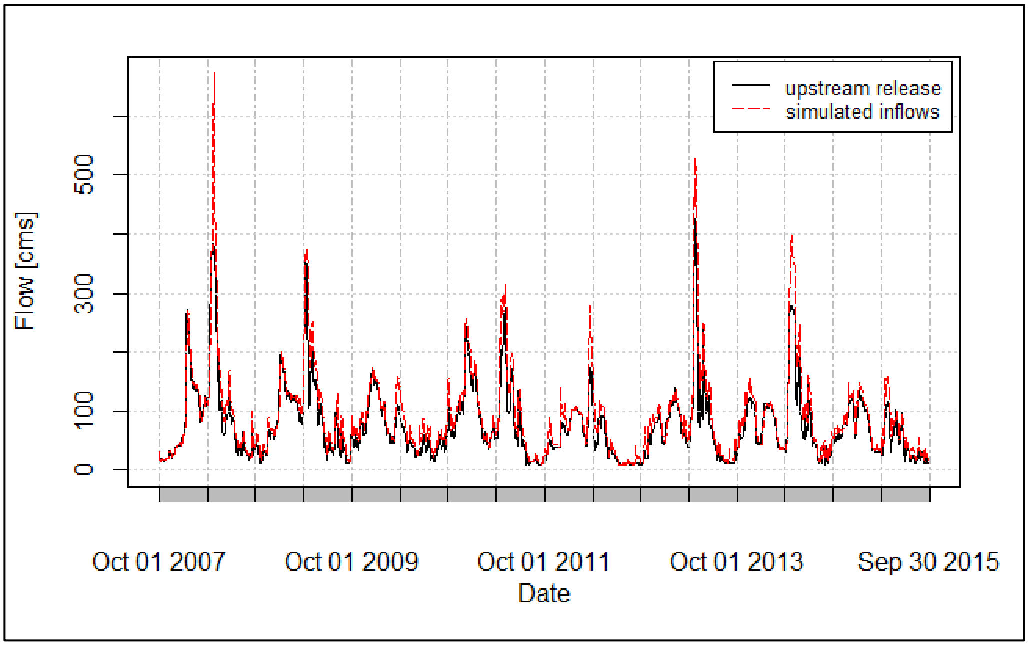

Figure 22. Comparison of upstream reservoir releases to simulated reservoir inflows at Mountain Chute

The split-sample test is again applied to the Mountain Chute model, for the same calibration and validation periods as the test applied on the Bark Lake model. The results for the reservoir stage and flows in calibration and validation show a similar pattern, where the validation stage is not well fitted but the flows simulated in validation still closely match the level pool routing estimates; the flows obtain a Nash-Sutcliffe and R^2 of 0.90 and 0.91 in the calibration period, respectively, and 0.91 for both metrics in the validation period.

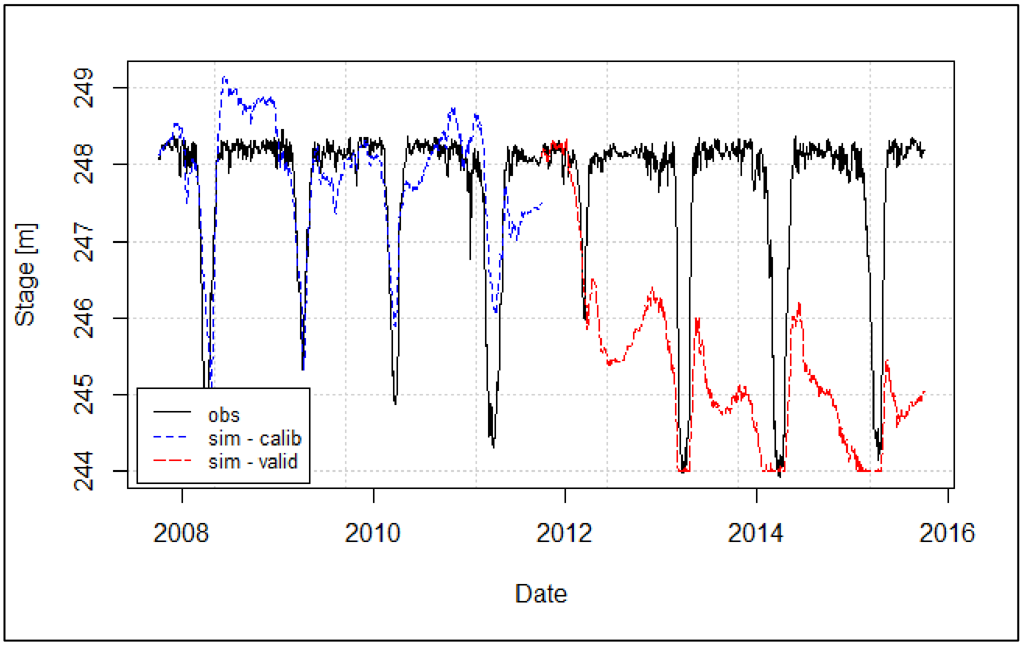

Figure 23. Mountain Chute model split-sample test results for reservoir stage

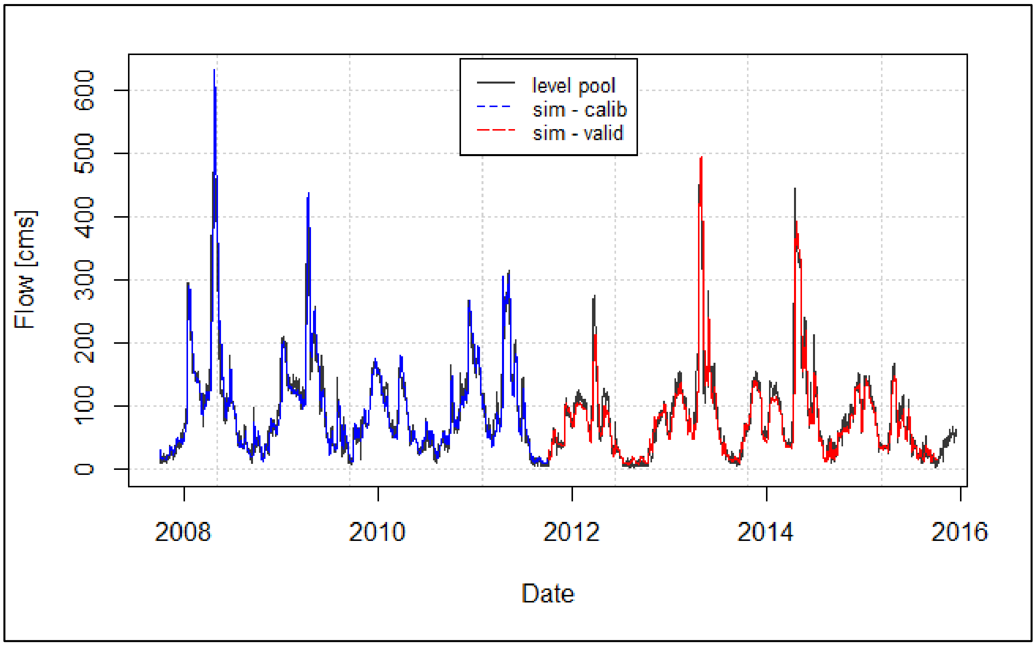

Figure 24. Mountain Chute model split-sample test results in flow comparison

These results for both models, particularly when models of both areas in the watershed show poor stage performance in validation but quite reasonable flow performance, raises the question of how well the stage needs to be matched in order to provide good flow estimates or insight into decision-making. This is explored further in subsequent thesis chapters, where the DCT framework is applied to these model structures to evaluate them in decision-making contexts.

The Bark Lake and Mountain Chute models calibrated to the full data period, where the stage and flows are both well-matched, are used in the synthetic reality generation step of the DCT framework. This provides a model with the best possible stage fit for use in the synthetic reality generation. It is stressed that for DCT, the synthetic reality model does not necessarily need to be fully validated for performance, it only requires the synthetic reality to produce plausible observations to test the evaluated model by. Therefore, the model calibrated to the full data period is sufficient for use in DCT as a synthetic reality generator.

The model structure presented in this chapter is also recycled in building the model undergoing evaluation. The use of these models in the DCT framework is explained in more detail in subsequent chapters. The next chapter presents a DCT application to a reservoir management decision in the Mountain Chute Dam subcatchment, which makes use of the Mountain Chute model.

Chapter 4
Decision Crash Testing in a Reservoir Management Application

This chapter discusses the setup and results of the first Decision Crash Testing (DCT) case study, applied to a reservoir management problem in the central portion of the Madawaska watershed. The details of the DCT setup are demonstrated in this chapter. This chapter also includes a demonstration of testing improvements to the model-building procedure within a DCT framework.

4.1 Case Study Setup

This section details the setup of the DCT framework in the context of a reservoir management decision in the Madawaska watershed. The context of the hypothetical reservoir management decision, the quantitative decision formulation, the additional model updates required for this case study, and the details of the DCT experimental setup are discussed in this section.

4.1.1 Reservoir Management Decision

In this case study, the hypothetical decision at hand is whether the outflows from Kamaniskeg Dam should be decreased in overall volume by 5% during the summer season in order to release more water in the autumn, and thus produce more profit from hydroelectric power generation. Downstream of the Kamaniskeg Dam along the Madawaska River exists the Lower Madawaska River Provincial Park, where many recreational activities including camping, canoeing, whitewater rafting, etc., are popular in the summer season. The location of the major features in this subcatchment are shown in Figure 25.

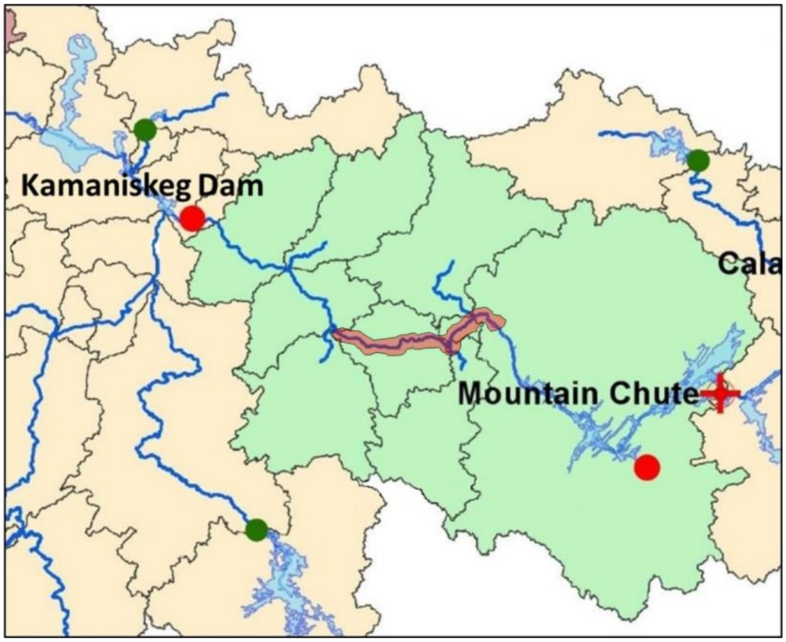

Figure 25. Location of key features in reservoir management DCT case study (Provincial Park location shown in transparent red)

In this hypothetical decision, the reservoir operator would yield greater profits by generating more electricity in the autumn season than in the winter season, and thus shifting some release from the summer season to the autumn season would be financially beneficial. However, releasing less water in the summer would impact the water flow and level in the Provincial Park during the summer season, to the detriment of the park patrons. Thus, the decision to adjust the flow releases at Kamaniskeg Dam is based on the impact of the adjusted reservoir operation on the flows in the Madawaska River. If the flows become reduced below a specified low flow threshold, the operational option is untenable.

The decision is setup as a binary decision, where either the current reservoir operation is kept, or the adjusted operation with decreased summer flows and increased autumn flows is implemented. In reality, there would be a spectrum of potential decisions, however it is kept to a binary decision for the purpose of this case study. The decision is also formulated to be based only on whether the impact of the decision on river flow is greater than a given threshold; the tradeoff between profits to the reservoir operator and value of higher flows for park patrons are greatly simplified by the use of a decision quantity threshold, which is assumed in this case study but would normally be determined through something more rigorous, such as a cost-benefit analysis.

The DCT implementation in this case study tests the ability of the model to inform the correct decision in a forecasting period, i.e., the model is supplied some calibration data for training, supplied future boundary conditions in the form of meteorological forcings and reservoir forcings, and used to evaluate the decision quantity for data unused in training. This is different from the use in the previous case study, where the model calibration period and decision evaluation period were identical.

4.1.2 Decision Quantity Definition

The decision quantity is meant to capture the impact of the reservoir operations on low flows in the Provincial Park, thus the outlet of subbasin 48 in the overall model discretization was selected as the point of evaluation, which is the subbasin along the main river branch just upstream of the Provincial Park. The summer flows from both possible reservoir operations (as simulated by a synthetic reality model with calibrated parameters) are shown in Figure 26.

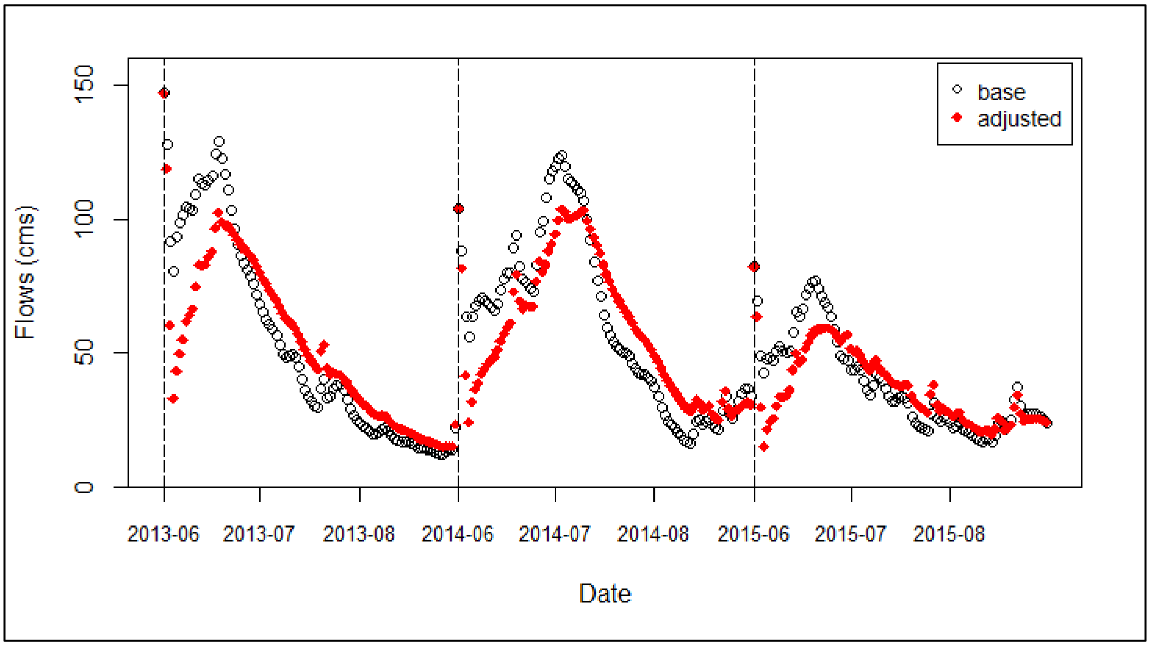

Figure 26. Synthetic reality generated summer flows at subbasin 48 for base and adjusted reservoir operations (vertical lines indicate breaks in time)

From the plot of summer flows, there is a consistent impact from the adjusted reservoir operation with a reduced summer rating curve. The June and early July flows are reduced by the adjusted reservoir operation, resulting in a slight increase in flows during the late July and August period. The increase in later summer flows occurs due to the relative retention of water in the reservoir earlier in the summer, leaving more water available to flow out during the later period. This shows that the period upon which the decision quantity is based is highly important, since the impact of the reservoir operations on the summer hydrograph are dependent on the time period examined.

In this case study, the decision quantities are based on the difference in proportion of low flow days during the respective evaluation periods, namely the months of June and August during the time period of 2012-10-10 to 2015-09-30. The threshold for each period, defining the low flow threshold, is set as 50 m^3/s and 20 m^3/s, respectively. The low flow thresholds are assumed for this case study, but reflect the hypothetical minimum streamflow values required for the reasonable enjoyment of the park for recreational purposes during the June and August months, respectively.

Mathematically, the proportion of flows during the simulated June months that are below the low flow threshold can be calculated as:

$$DV_{base} = \frac{1}{N_d} \sum_{i=1}^{N_d} S(Q_i > Q_t) \tag{1}$$

where DV_{base} is a decision variable corresponding to the base reservoir operation (dimensionless), N_d is the number of June days in the evaluation period, S is a function that evaluates the given statement and returns a value of 0 or 1 corresponding to whether it is true or false, Q_i is the daily average flow on the i^{th} June day (m^3/s), and Q_t is the minimum flow threshold required for ‘enjoyable’ use of the river (i.e., 50 m^3/s). The second decision variable, DV_{adj}, is defined in the same way but calculated for the adjusted reservoir operation instead of the base operation.

Using these definitions of decision variables, the decision quantity is calculated as:

$$DQ_1 = DV_{base} - DV_{adj} \tag{2}$$

where DQ_1 is the decision quantity calculated for June months (dimensionless). The second decision quantity, DQ_2, is calculated in the same way but for the August months with a low flow threshold of 20 m^3/s. In either case, the decision formulation is based on whether the decision quantity, representing a change in the proportion of low flow days for that period, is acceptable. The decision quantity threshold is set here as 0.2, which can be interpreted as, “the adjusted reservoir operation should be implemented if the percent reduction in low flow days is less than 20%”. This 20% decision quantity threshold is arbitrarily assigned for the purpose of this case study, but reflects the hypothetical tradeoff between the reduction in value of the Park’s enjoyment and the value of profits made by the adjusted reservoir management scheme. Note that in the case of the August period, the proportion of low flows days is expected to decrease under the adjusted reservoir operation, meaning that the use of DQ_2 as a decision quantity should lead to an easy decision for the model to inform. In terms of the model evaluation, the DCT framework is deployed to test whether the model (including provided data) is capable of estimating the change in proportion of summer flows with sufficient accuracy to inform the operational procedure of the reservoir.

The updates to the Mountain Chute hydrologic model that were required to implement this case study are discussed in the next section.

4.1.3 Mountain Chute Model Updates

A number of updates to the calibrated hydrologic model of Mountain Chute (discussed in section 3.2.7) were required for the application of DCT in this case study. The main adjustment was the added representation of the Kamaniskeg Dam to replace the forced inflows to the model, since the reservoir operations at Kamaniskeg needed to be represented. Here, inflows to the Kamaniskeg reservoir were computed using level pool routing and forced to the reservoir. The Kamaniskeg Dam reservoir parameters were estimated in a similar fashion described in section 3.2.5; reservoir parameter estimates were obtained from the HydroLakes and OPG data sources, and the reservoir area was calibrated to obtain a good model fit to the observed stage. The reservoir parameters for Kamaniskeg Dam are included in Appendix B.

An important update was the simulation of Kamaniskeg Dam outflows using a seasonal rating curve representation, rather than a fixed rating curve for the entire year. This was required in order to model the adjusted reservoir releases in different seasons. An analysis was performed to determine (a) which months are appropriate to cluster as seasonal outflow schemes, (b) the quantitative fitting of each season, and (c) the adjustments to the base seasonal outflow scheme required to decrease summer outflows by 5% (and increase autumn flows by 5%). The details of this analysis are included in Appendix B, however, the key results of this analysis include:

- The seasons are defined as expected, with the exception of March belonging to the winter months rather than the spring months. This was defined qualitatively by examining the historical stage-outflow scatterplots in each month, and the March pattern was much more consistent with winter months.

- The typical power law relationship was not appropriate for fitting flows in any season. A piecewise linear regression was used to fit the winter stage-outflow relationship, and a simple linear regression was used for the other three seasons. These linear regressions were used to create the stage-outflow relationships in the rating curve representation of the Kamaniskeg Dam, and the outflows produced by this seasonal relationship were verified against observed data.
- The summer flows in the adjusted reservoir management scheme were determined by manually reducing the summer flows in the base rating curve by a single factor, until such a point that the synthetic reality model with calibrated parameters had an average summer flow volume decrease of approximately 5%. This corresponded to an approximately 5% increase in autumn flows without adjusting the autumn stage-outflow relationship, since the excess volume held back in summer is naturally released in autumn under these circumstances.

The modelled outflows under the base seasonal rating curve were matched very closely by the observed flows downstream of Kamaniskeg Dam, with a Nash-Sutcliffe value of 0.96 and an R^2 value of 0.97. These outflows are compared in Figure 27.

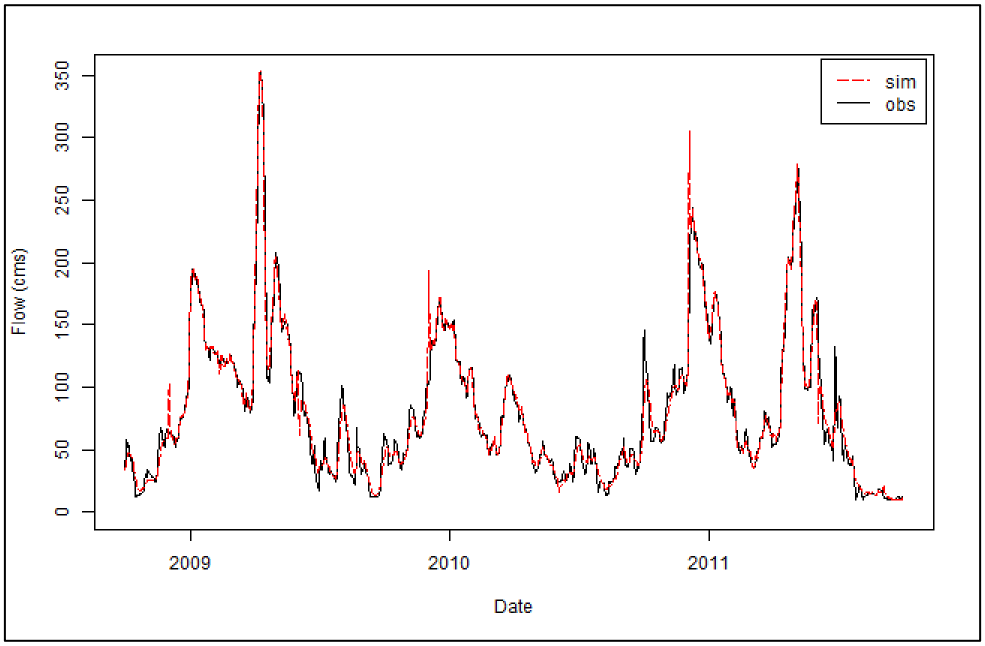

Figure 27. Comparison of simulated and observed flows using a seasonal rating curve representation at Kamaniskeg Dam

Although the flows were well-fitted by the representation of the Kamaniskeg Dam, the reservoir stage was shown to be more sensitive to the seasonal representation, as shown in Figure 28.

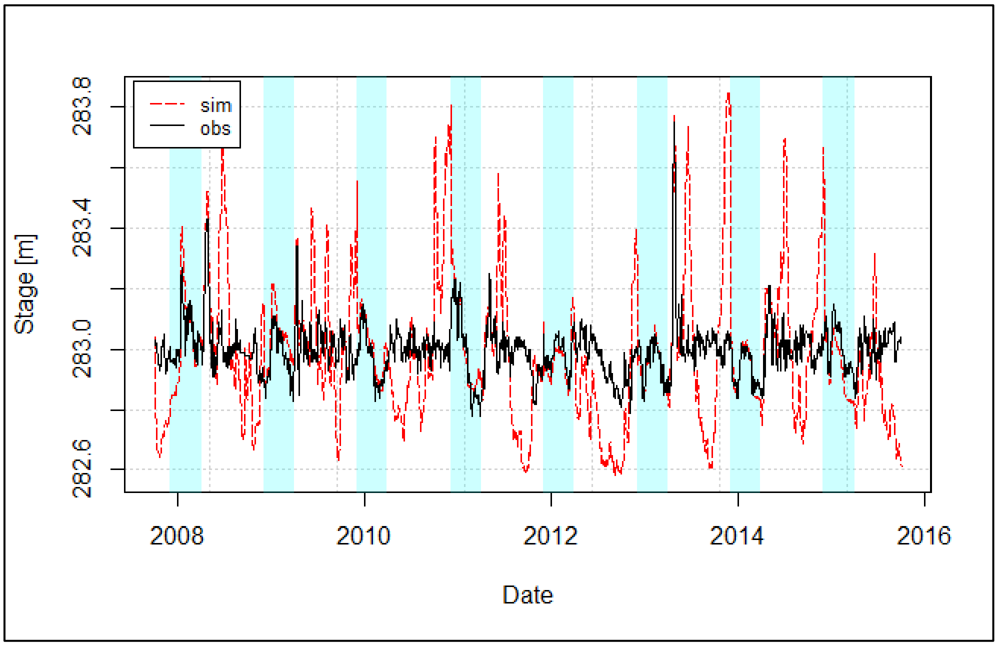

Figure 28. Kamaniskeg Dam stage with seasonal outflow representation

The sensitivity in reservoir stage is due to the linear fits to the seasonal stage-outflow relationships. In reality, the reservoir is operated to adjust flows to mitigate such changes in reservoir stage, which is not captured well by the linear fits to the stage-outflow relationship. However, the fit is deemed sufficient by the fit to the observed outflows.

One artefact of the seasonal rating curve is that there is spike in flows on the transition days between seasonal rating curves, as seen in Figure 28 for December 1st of each year, for example. Both of the reservoir management operations, the base scheme and the adjusted scheme with reduced summer flows, are described in the model with a table of seasonal rating curves. The rating curves provided to the model for both reservoir operations are shown in Appendix B.

4.1.4 Synthetic Reality Generation

In this case study, the synthetic reality model is used to generate synthetic reservoir stage data at Mountain Chute, to which the evaluated model is calibrated. The data is generated by the synthetic reality model for use in calibration for the 2007-10-01 to 2011-09-30 period, and the decision quantities are computed using the 2012-10-01 to 2015-09-30 period.

The synthetic realities were generated using random variations in key parameters to the updated Mountain Chute model. The parameters were varied using a uniform sampling of 11 key parameters within a specified range, shown in Table 3.

Table 3. Key parameters used in synthetic reality generation for reservoir management case study

Raven Parameter Name	Description	Base Value	Lower Bound	Upper Bound
HBV_BETA	Infiltration coefficient for rock layers [-]	10.98	5.00	20.00
HBV_BETA	Infiltration coefficient for soil layers [-]	13.75	5.00	20.00
MAX_PERC_RATE	Maximum percolation rate in fractured rock layers [mm/d]	57.26	10.00	100.00
MAX_PERC_RATE	Maximum percolation rate in fractured soil layers [mm/d]	196.00	10.00	200.00
MAX_INTERFLOW_RATE	Maximum interflow rate in fractured rock layers [mm/d]	10.73	1.00	50.00
MAX_INTERFLOW_RATE	Maximum interflow rate in fractured soil layers [mm/d]	34.55	1.00	50.00
PET_CORRECTION	PET Correction in rock layers [-]	0.71	0.40	1.30
PET_CORRECTION	PET Correction in soil layers [-]	0.66	0.40	1.30
SVF_EXTINCTION	Shortwave radiation extinction coefficient [-]	0.23	0.00	1.00
DEP_MAX	Maximum depression storage [mm]	229.00	10.00	500.00
LAKE_PET_CORR	PET correction for Lakes [-]	1.01	0.50	1.30
ABST_PERCENT	Percent runoff routed to depression storage [-]	0.30	0.05	0.50

The other parameters in the model, including the reservoir parameters at Kamaniskeg Dam and Mountain Chute, were not varied in the synthetic reality generation. The model boundary conditions were similarly not changed, which includes the inflows to Kamaniskeg Dam, the outflows from Mountain Chute, and the forcing functions (i.e., precipitation, temperature).

The synthetic reality model was run for the entire modelled period, from 2007-10-01 to 2015-09-30. The decision quantities were calculated to be consistent with the model evaluation period (2012-10-01 to 2015-09-30). The data passed to the evaluated model for calibration is the 2007-10-01 to 2011-09-30 period, thus the model evaluation period includes a one-year warm-up period prior to calculating decision quantities.

A check on the realism of the simulation was applied as a reservoir plausibility check on the Mountain Chute stage to ensure that the initial reservoir stage was within the 10% and 90% percentiles. An additional check was done to remove all synthetic realities that exceeded a stage of 256 m at Mountain Chute. Together, these quality checks removed the synthetic realities that produced an unsteady stage at Mountain Chute, which is shown in Figure 29.

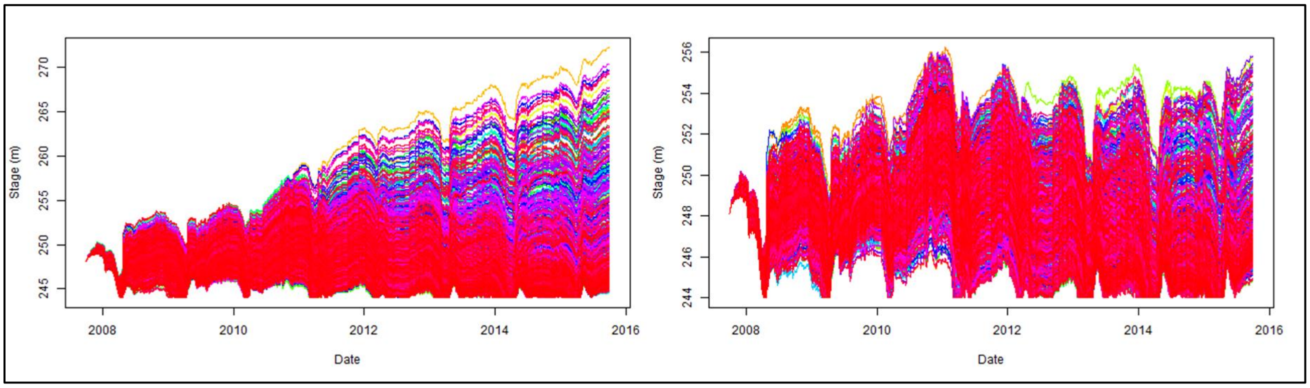

Figure 29. Comparison of synthetic reality simulated reservoir stages at Mountain Chute (pre-quality checks (left) and post-quality checks (right))

The removal of poor quality synthetic realities left 1325 generations remaining, from which 100 were randomly resampled with a uniform distribution (using the August decision quantity). For each of these 100 samples, the synthetic reality model was run with the given parameter set using the base reservoir operation, which generates the synthetic data series passed to the evaluated model. The model is then run again with the adjusted seasonal rating curve. These two runs are compared to calculate the decision variables and decision quantities. The distributions of the June decision variables and decision quantity is shown in Figure 30 and Figure 31, and the distributions are similarly shown for the August decision variables and decision quantity in Figure 32 and Figure 33; the distributions of 100 sampled synthetic realities, as well as the full distributions for 1325 synthetic realities, are both shown in these plots.

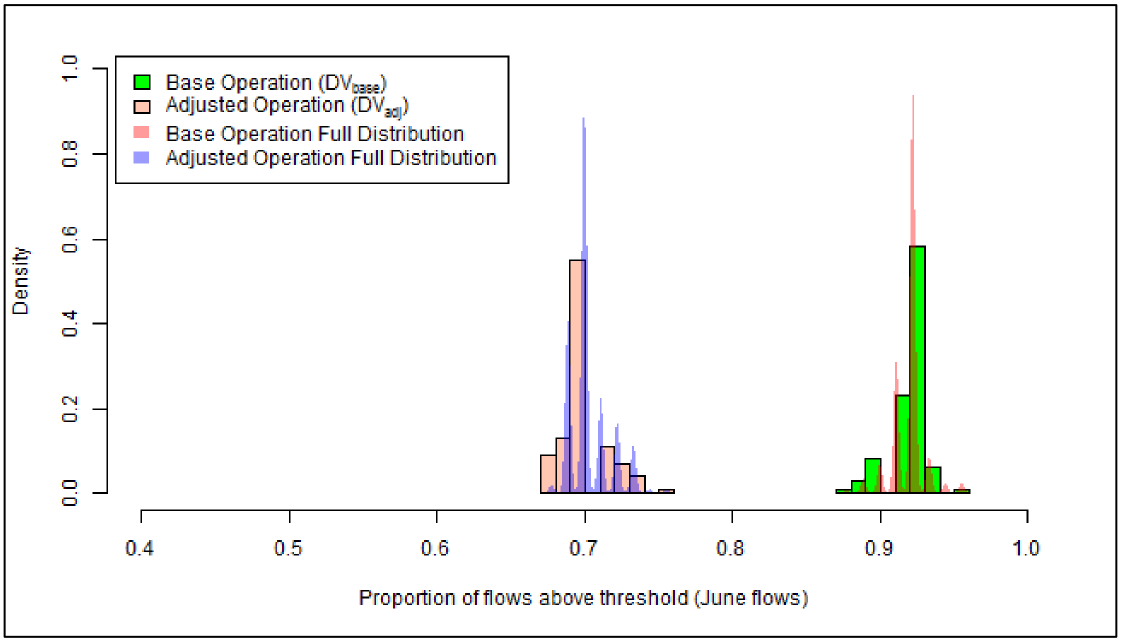

Figure 30. June proportion of flows above threshold for synthetic realities

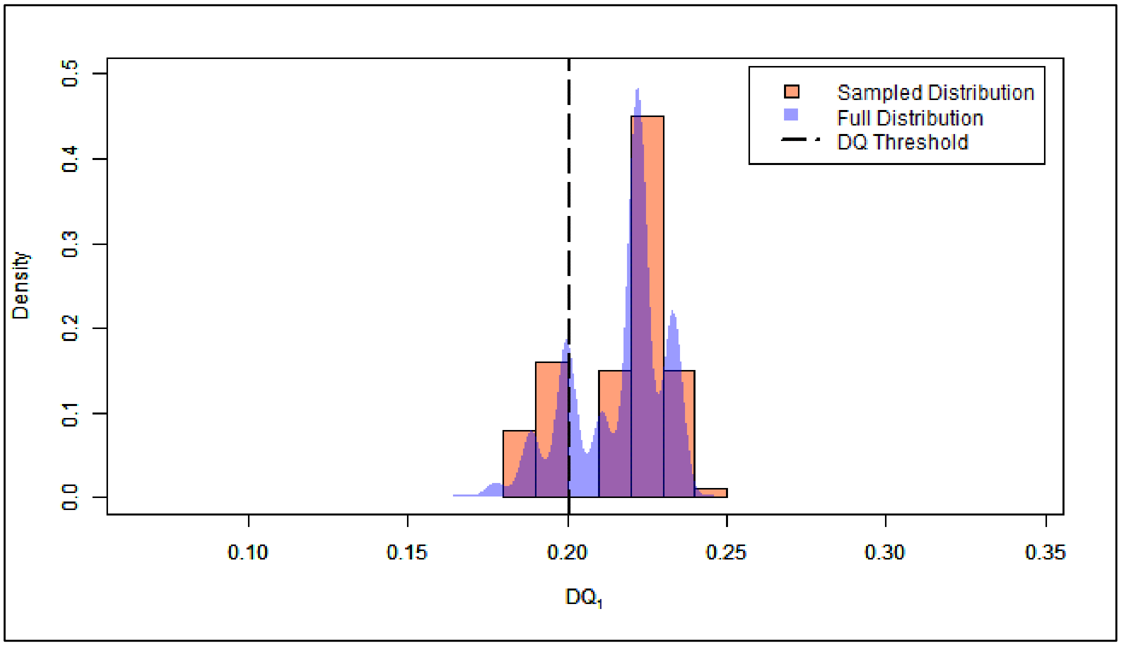

Figure 31. Synthetic reality decision quantity distributions of DQ_1

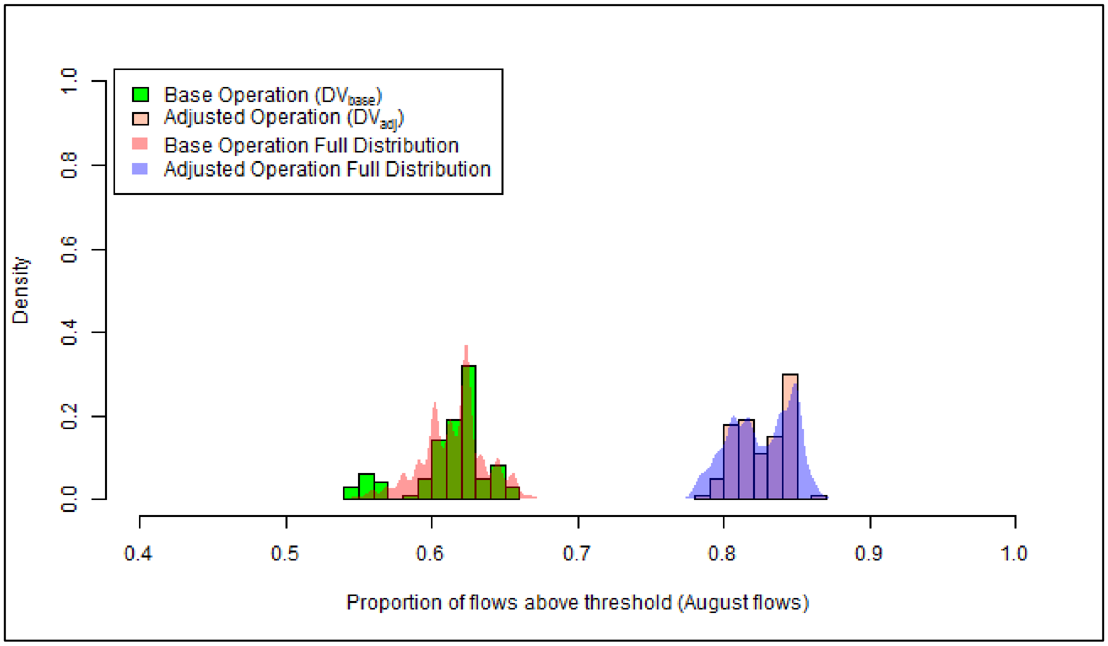

Figure 32. August proportion of flows above threshold for synthetic realities

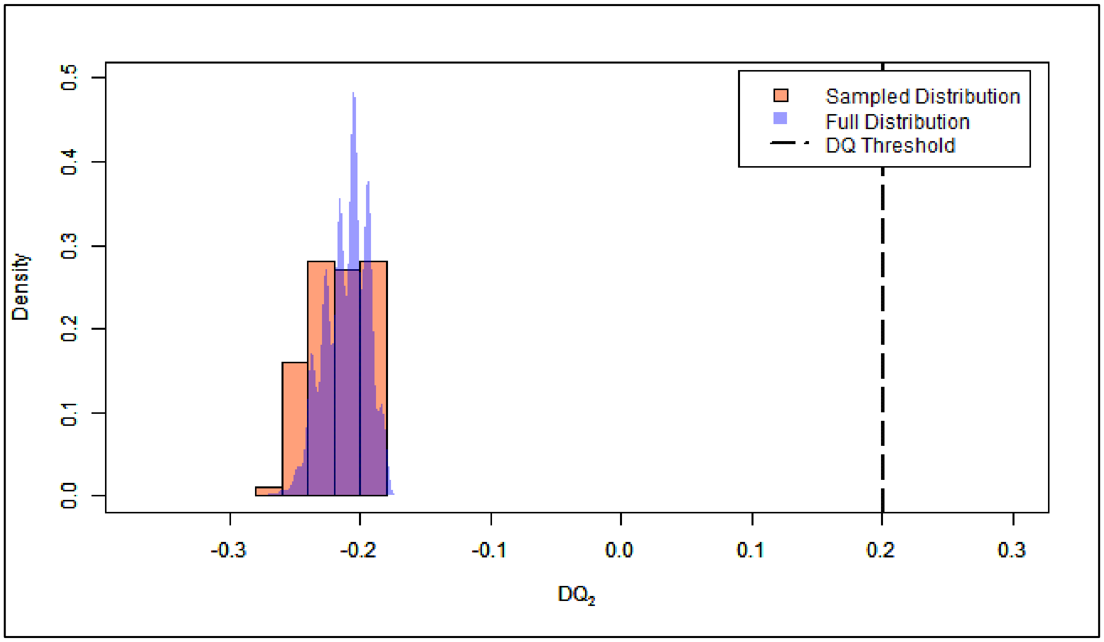

Figure 33. Synthetic reality decision quantity distributions of DQ_2

The impact of using a relatively small number of days for computing the decision quantities (i.e., <100 days) is shown in the sparseness of the synthetic decision quantity plots, particularly for June flows. For the sampled synthetic reality decision quantity distributions, the relevant summary metrics are presented in Table 4.

Table 4. Synthetic reality summary statistics for reservoir management decision

	Sampled Synthetic Realities (100)		All Synthetic Realities (1325)	
DQ	**Decision Difficulty**	**AMAMS**	**Decision Difficulty**	**AMAMS**
DQ1	0.32	0.02	0.35	0.00
DQ2	0.00	0.39	0.02	0.40

The summary statistics in Table 4 show that the decision is relatively difficult when the June decision quantity (DQ_1) is used, based on the larger decision difficulty and the smaller AMAMS. The August decision quantity (DQ_2), in contrast, has a decision difficulty of zero due to all of the synthetic decision quantities falling on one side of the decision quantity threshold (Figure 33), and has a much larger AMAMS. The summary statistics also show that the metrics are almost identical for the sampled synthetic realities as for the full set of synthetic realities.

These summary statistics are useful to compute prior to the DCT evaluation, since they provide an expectation of model performance and a benchmark for the required model skill. In order to provide evidence that the number of synthetic reality samples is enough to compute the decision difficulty and AMAMS, the convergence of these metrics with trial number can be plotted and evaluated qualitatively. This is shown for the AMAMS corresponding to DQ_1 in Figure 34 for both the sampled synthetic realities and the full set of 1325 synthetic realities.

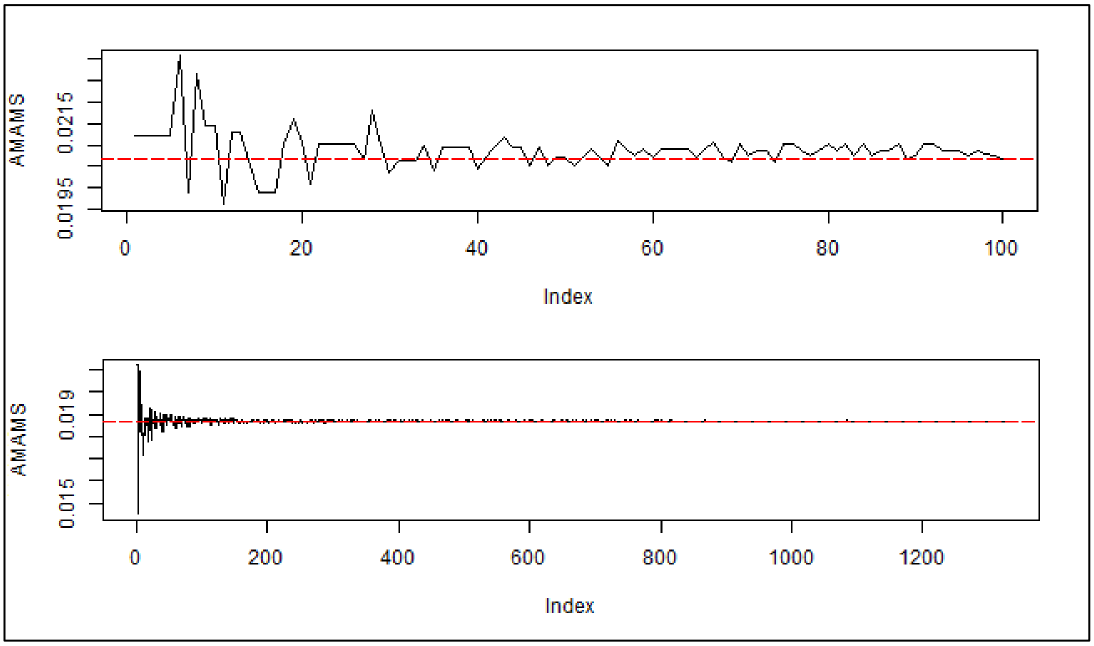

Figure 34. Convergence plot of the AMAMS for DQ$_1$

The convergence of the metric shown in Figure 34 indicates that the number of samples is likely enough to establish an estimate of the summary statistics for the synthetic reality generations, and is some evidence that the current sample size will also be sufficient for the remaining DCT metrics. This is certainly not conclusive prior to running the full DCT experiment, but is a good check prior to proceeding with the much more computationally expensive portion of the experiment.

4.1.5 Model Evaluation Setup

The model evaluation portion of this case study is similar to the previous one. The model structure is the same as the model used to generate synthetic realities, and an expanded set of the parameters varied in synthetic reality generation are included in the calibration parameter set (26 parameters in total are calibrated in each model evaluation step; the full list calibrated parameters is included in Appendix B). The list of parameters included in calibration is expanded to more closely reflect the fact that these parameters would not be known when the real model would be built and calibrated.

The model receives the 2007-10-01 to 2011-09-30 synthetic stage data at Mountain Chute Dam from the synthetic reality generation, and is calibrated to this stage series. The calibration procedure uses the DDS algorithm with a budget of 2000 runs, and uses the root-mean square error of the stage on Mountain Chute (inverse-weighted in time, as suggested in section 3.2.6) as a calibration objective. Once the model is calibrated, the model is run for the ‘future’ time period of 2011-10-01 to 2015-09-30, and the decision quantities are evaluated for the 2012-10-01 to 2015-09-30 period (again leaving one year as a warm-up period). This requires running the calibrated model in each iteration twice, once for each reservoir management scheme (similar to the synthetic reality generation). Once the model is run for each reservoir management scheme, the two decision quantities can be computed and the DCT metrics can be computed. These results are presented in the next section.

4.2 Results and Discussion

This section presents the DCT results for the reservoir management decision in the first case study. The summary of the DCT metrics can be found in Table 5.

Table 5. Summary metrics for DCT results for reservoir management decision

DQ	**Sim. Score**	**MS Mean*** $\mu_{\Delta\phi}$	**MS Abs. Mean*** $\mu_{\|\Delta\phi\|}$	**MS StDev*** $\sigma_{\Delta\phi}$	**R^2****	**Slope****
DQ$_1$	0.59	-0.002	0.016	0.021	0.002	0.000
DQ$_2$	1.00	-0.011	0.026	0.031	0.000	0.000

*MS = model skill

**from linear regression of MS onto calibrated RMSE

The summary metrics show that the model had a lower similarity score with DQ_1 than DQ_2, which is consistent with the lower decision difficulty computed for the synthetic reality generations with DQ_2. The mean model skill is close to zero in both cases, which suggests the model predicts the decision quantity in an unbiased way. An examination of the synthetic and model decision quantities reveals further information; these are shown in Figure 35 and Figure 36 for DQ_1 and DQ_2, respectively; the regions where the model fails to inform the correct decision are shown in Figure 35.

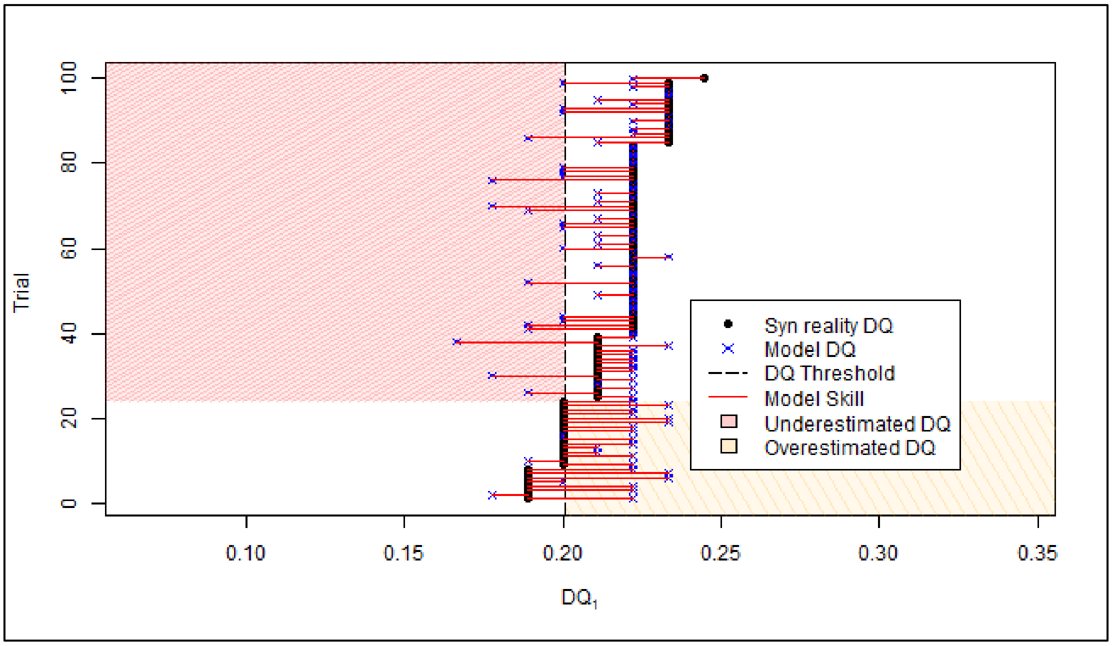

Figure 35. Decision quantity 1 plot for reservoir management decision

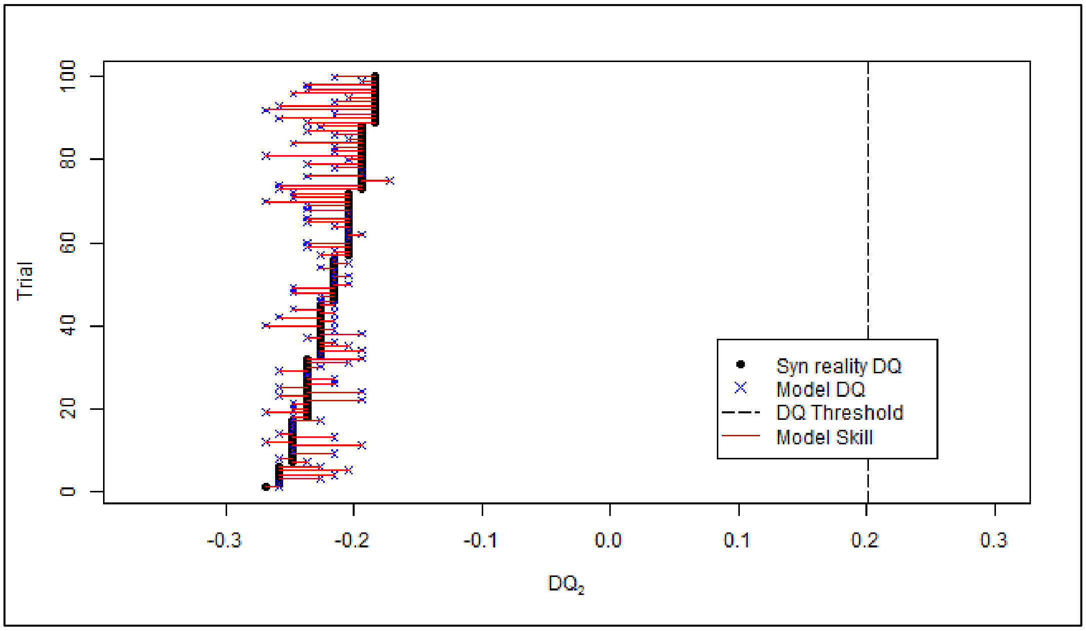

Figure 36. Decision quantity 2 plot for reservoir management decision

The decision quantity plots above show that the model has some skill in a decision-making context. It is not surprising that the model is capable of consistently informing the correct decision when the impact on August flows (DQ_2) is examined, since the decision difficulty is zero. However, examining the June flows (DQ_1) is a more honest test of the model's ability to inform decision-making for this particular decision quantity threshold. The model has a similarity score of 59%, which can be interpreted as an upper bound of 59% on the probability that the model would inform the correct decision in the real application. This is treated as an upper bound since the test here is much less rigorous than a hypothetical test against all instances of 'reality' would be (for instance, the real decision-making application would not provide all future forcings to be known with certainty). In this light, the 59% success rate is still approximately as good as a coin flip; this shows that even this relatively well-performing model is not much more useful than random chance for a sharp binary decision, and in order to build trust in the reliability of the model to inform a non-hypothetical binary decision successfully, the model building procedure should somehow be improved.

One possible detriment to the decision-making ability of the model is that the model skill appears to be a function of the synthetic decision quantity. Although the summary metrics indicate that the model skill is unbiased overall, there is a systemic trend evident in Figure 35 in the sign of the model skill as a function of synthetic decision quantity. In other words, the model tends to overestimate the decision quantity when the synthetic decision quantity is low, and underestimate the decision quantity when it is larger. A possible next step in the investigation of this model's performance would be to deploy a Scenario Discovery-based approach to determine under what conditions the model skill becomes biased, and use the DCT framework as a guide to rectify this systematic bias.

Finally, the synthetic reality analysis included a convergence check on the synthetic reality metrics, such as the AMAMS (see Figure 34). The same convergence check can be performed on any of the DCT metrics. The convergence of the similarity score for DQ_1 is shown in Figure 37, which shows that the similarity score does seem to converge on its estimated value.

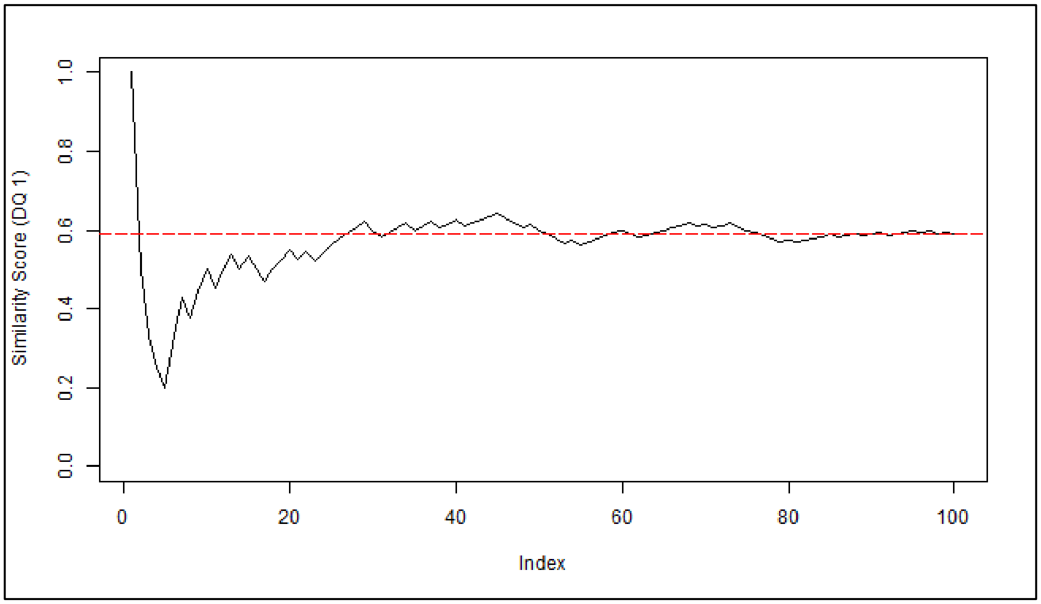

Figure 37. Convergence plot of the similarity score for DQ_1 with 100 trials

This qualitative convergence check provides reasonably sound evidence that including more runs will not significantly change the results presented here. Ideally, this would be further tested using a new set of 100 samples from the original 1325 synthetic realities generated.

4.2.1 Follow-up DCT Experiment in Model Improvement

The results of the previous section show that the model has a similarity score of 59% when the June flows are examined, which is likely not sufficient for reliability in the context of a sharp binary decision, with a decision quantity threshold in the toughest predictability region. In order to improve model reliability to inform the decision, the model-building procedure must somehow be adjusted. In this follow-up DCT experiment, a basic subsequent experiment is performed to demonstrate how DCT can provide the framework for guiding model improvements.

In this experiment, the calibration objective used in model calibration is changed. The original objective function in the base DCT experiment was the RMSE of the reservoir stage, inversely weighted in time. In this experiment, the calibration objective function is an average of the stage-based RMSE and the daily change in volume-based RMSE (this is the same objective as used in the building of the reservoir models, see the timing metric in Appendix A for details on the objective function). Otherwise, the same parameter set is calibrated, and the calibration budget of 2000 runs is also kept the same, thus any difference in the model performance is due to the adjusted objective function (and stochasticity in the calibration algorithm) only. In the follow-up experiment, 50 of the 100 synthetic realities used in the base experiment are sampled for use; this allows for direct comparison of model results between each experiment. The results of the base experiment are summarized in Table 6 (the base case results are recalculated for the 50 trials sampled).

Table 6. Comparison of DCT metrics for base and follow-up experiments

	Base Case		**Follow-up Experiment**	
Metric	**DQ_1**	**DQ_2**	**DQ_1**	**DQ_2**
Sim. Score	0.60	1.00	0.66	1.00
MS Mean	-0.003	-0.014	0.005	0.005
MS Abs. Mean	0.018	0.031	0.020	0.029
MS StDev	0.014	0.026	0.014	0.024
R^2	0.004	0.039	0.050	0.023
Slope	-0.001	0.003	-0.005	0.005

From the summary metrics, there is negligible difference in the mean model skill and the similarity score for DQ_2 is 1.0 in both experiments. However, the follow-up experiment shows a 6% increase in the similarity score for DQ_1, indicating that the adjustment in objective function had some improvement in the ability of the model to inform the decision. The decision quantity plots for each decision quantity are shown in Figure 38 and Figure 39.

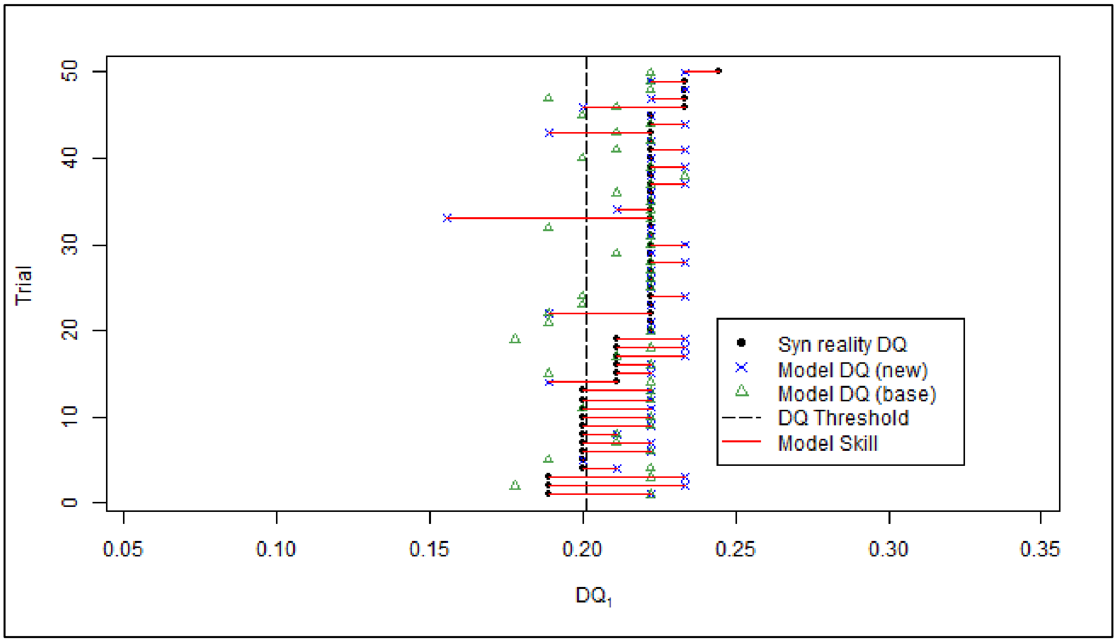

Figure 38. Decision quantity 1 plot for reservoir management decision in follow-up experiment

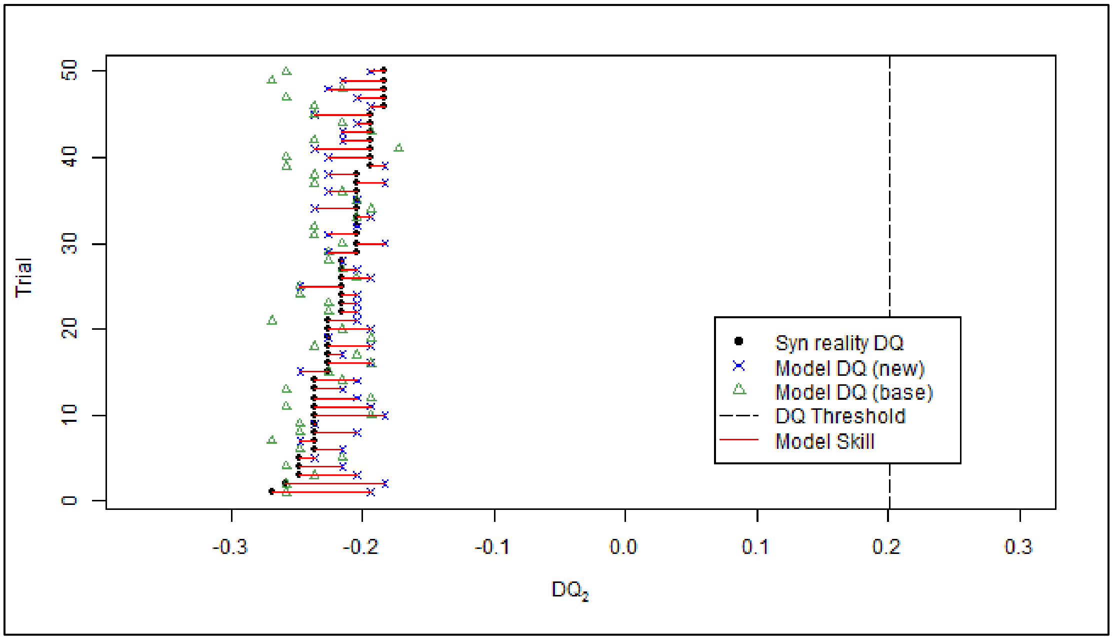

Figure 39. Decision quantity 2 plot for reservoir management decision in follow-up experiment

In order to provide more evidence of a genuine increase in the similarity score for DQ_1 due to this adjustment, a statistical test of the difference in similarity scores is applied. Since the similarity score is itself a single value in each experiment, the bootstrap method (Efron, 1979) is applied to obtain a distribution of the similarity score for each experiment. Note that in this case a hypothesis for a proportion could be applied without the need for bootstrapping. However, for more complex decision formulations the test for proportions may not be appropriate, and thus bootstrapping is demonstrated as a more general approach for hypothesis testing of the similarity score.

The bootstrapping is done by randomly sampling with replacement a total 50 samples from each set of experimental results, calculating a similarity score, and repeating this 50 times. This produces a distribution of similarity scores for each experiment, which is shown in Figure 40.

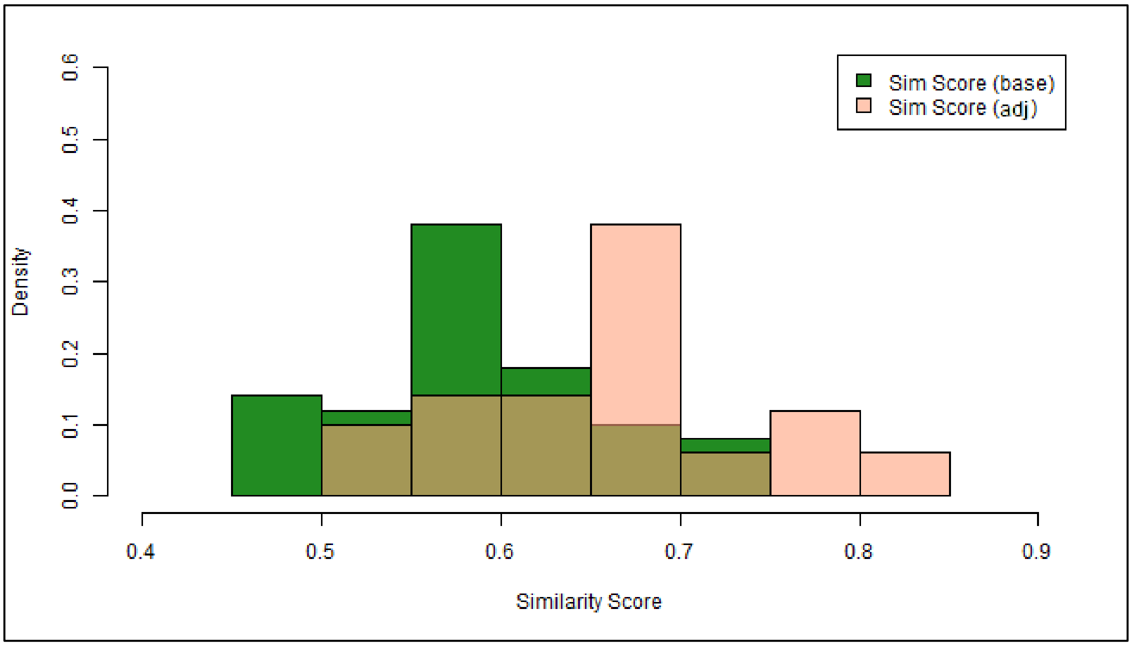

Figure 40. Bootstrapped distributions of similarity score for each DCT experiment

The plot of the distributions in Figure 40 suggests a difference in the mean of these two distributions; the sample means from bootstrapping are 0.59 and 0.66 for the base experiment and adjusted experiment, respectively. To test this difference statistically, the t-test is applied as a one-sided difference in the means of the sampled distributions. The null hypothesis, H_0, and the alternative hypothesis, H_A, are defined as:

$$H_0: \overline{S_{adj}} - \overline{S_{base}} = 0 \tag{3}$$

$$H_A: \overline{S_{adj}} - \overline{S_{base}} > 0 \tag{4}$$

where $\overline{S_{adj}}$ is the mean of the similarity score distribution in the adjusted experiment, and $\overline{S_{base}}$ is the mean of the similarity score distribution in the base experiment. The test is performed using unequal variances and unpaired results, which is referred to as Welch’s t-test (Welch, 1947). The normality of each distribution is an assumption of the test; the normality plots of each distribution show that this is not quite true, see seen in Figure 41.

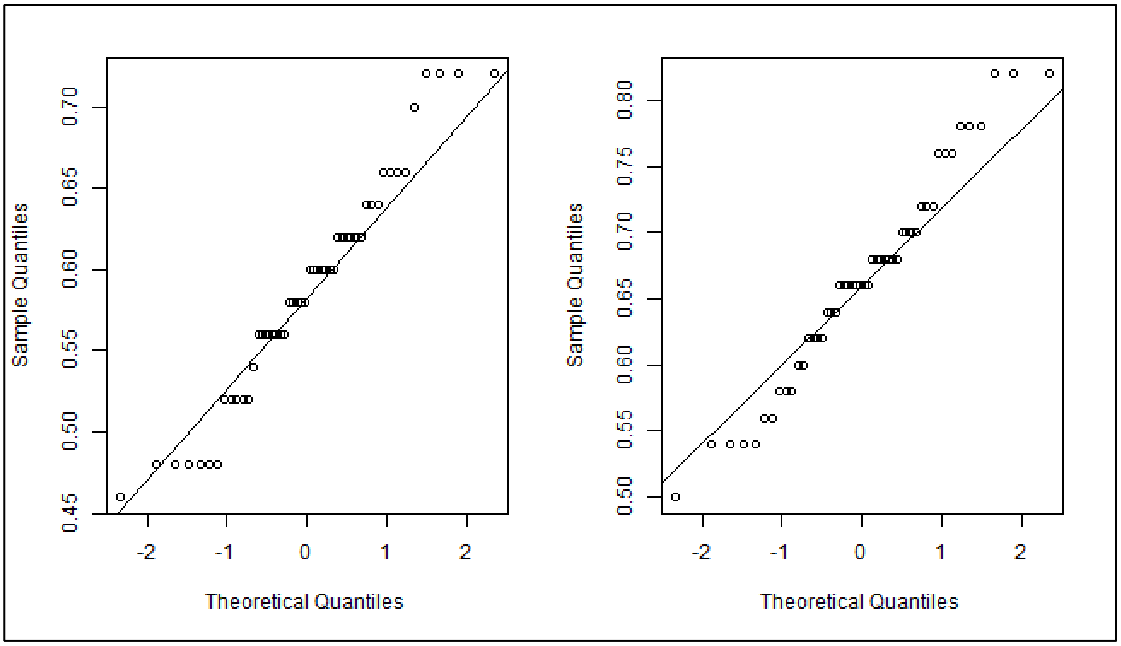

Figure 41. Normality plots for bootstrapped distributions of similarity scores (base similarity scores (left), adjusted similarity scores (right))

However, for the purpose of this t-test the strict normality assumption is relaxed, and the result of the test is treated as approximate. The t-test on difference in means the two similarity score distributions was performed using the ‘t.test’ function of the ‘stats’ library of the R Statistical Language (R Core Team, 2017). The result of the t-test shows that the null hypothesis is rejected and the true difference in means of the distributions is not equal to zero, with a p-value of approximately 6.5E-7. Thus, the conclusion can be made that there is a statistically significant difference in the mean similarity score of the base experiment and the adjusted experiment.

There are a number of different ways in which an improvement in the model-building procedure could be demonstrated in the DCT framework, however, this provides a simple example that demonstrates a statistically significant improvement in the ability of the model to inform the correct decision. This process would ideally be iterated until the similarity score is to the satisfaction of the decision-maker.

4.2.2 Adjusted Decision Formulations for Sharp Binary Decisions

In the main experiment, the similarity score with 100 trials was computed as 59% for June flows, which suggests that the evaluated model does not have much more decision-making ability in this binary decision than random chance. However, the model appears to estimate the change in proportion of low flow days with an average accuracy of approximately ±1.6%, which by most other standards would be considered excellent for decision-making. This warrants the question as to why the model appears to perform so poorly within the DCT framework, despite the perceived high level of accuracy in estimation of the decision quantity.

The answer may lie in the decision formulation, which is alluded by the reference to a ‘sharp’ binary decision. The sharp binary decision, where a decision is informed precisely by the decision quantity relative to the decision quantity threshold, is likely not consistent with the approach of a decision-maker. In a less theoretical decision-making process, a decision-maker would not consider a model-estimated decision quantity of 19% and 21% to be justification for making a different decision. This is the case for a number of reasons, primarily due to the uncertainty in model estimations. In the instance of this case study, a decision-maker would likely still choose to refrain from the adjusted reservoir operation if the model-estimated difference in low flow days is 19% in order to be conservative. Therefore, this section presents a new decision formulation that is more reflective of this fuzzier decision-making process, and re-evaluates the DCT results in this context. The sensitivity of the DCT results to the decision quantity threshold is also demonstrated.

The fuzzy decision formulations introduce a “fuzzy” region in decision space around the decision quantity threshold, where the decision would be made by the decision-maker and not automatically dictated by the model result. This is done to account for model uncertainty and other factors generally present in the decision-making process, such that a decision is clearly preferred by the model-estimated decision quantity if and only if the model informs the decision outside of this fuzzy boundary.

To demonstrate the sensitivity of the DCT results to the decision quantity threshold, the similarity score is plotted as a function of the decision quantity threshold used in the sharp binary decision, shown in Figure 42.

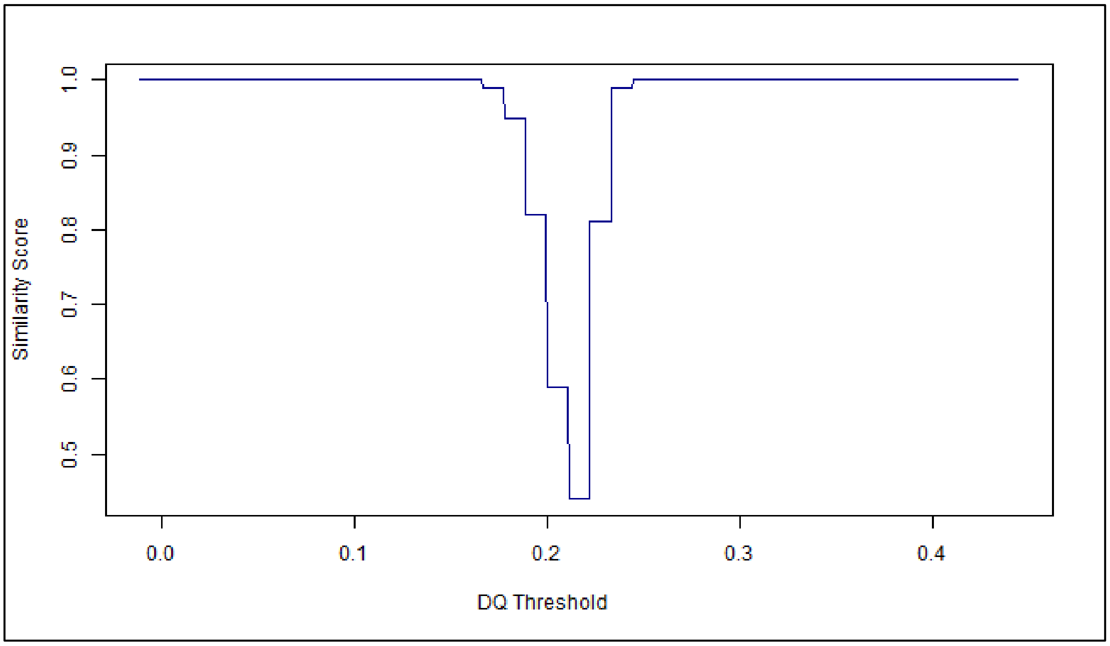

Figure 42. Similarity score as a function of decision quantity threshold for June flows

The plot in Figure 42 shows that the similarity score, a central metric in the DCT framework, is highly sensitive to the decision quantity threshold selected within the range of the original value of 20%. In this example, the similarity score decreases from a value of 1 to 0.44 with only a 0.045 change in the decision quantity. In this context, a small change in the decision quantity threshold can influence the DCT results drastically, which is not ideal.

The results of Figure 42 provide motivation to re-evaluate the DCT results with a less influential decision formulation, thus two fuzzy decision formulations are implemented in this section. In the first decision formulation, referred to as the 'fuzzy boundary' decision formulation, the model is deemed to have informed the decision correctly if:

1. The model and synthetic decision quantity are both within the fuzzy region of decision space, and
2. The model and synthetic decision quantity are both clearly in the same decision region of decision space (i.e., decision A or B) outside of the fuzzy region.

For the case where the model and synthetic decision quantity are on the same side of the sharp decision quantity threshold but one is in the fuzzy region and the other is not, the model is deemed to have partially informed the decision correctly, which is calculated as a function of the model skill.

This can be considered more reflective of the uncertain nature of less theoretical decision-making processes. For the purpose of this exercise, the fuzzy region is defined as the decision space between 17.5% and 22.5% for decision quantity 1 (the difference in proportion of June low flows). The fuzzy boundary decision formulation is shown graphically in Figure 43.

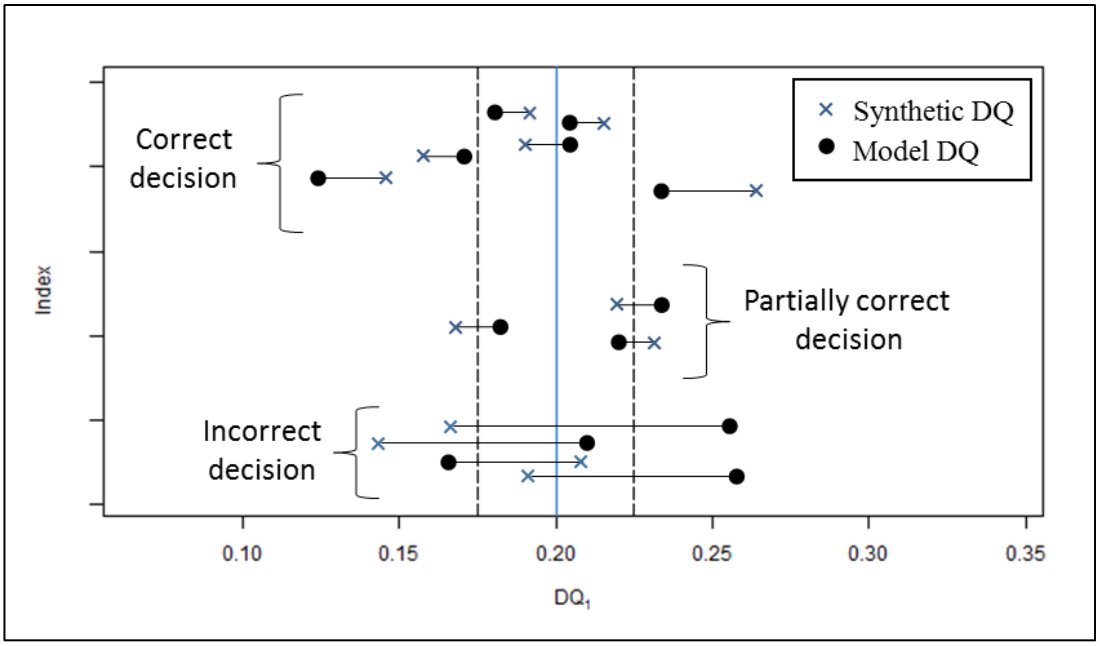

Figure 43. Similarity score calculation for fuzzy boundary decision formulation

In this fuzzy boundary decision formulation, the degree of partial correctness is calculated as:

$$max\left\{1 - {}^{|\Delta\phi_n|}/_{AMAMS}, 0\right\} \quad (5)$$

where $\Delta\phi_n$ is the model skill and $AMAMS$ is the average maximum allowable model skill. This allows the partial correctness to be bounded between 0 and 1 as a function of the model skill. In the computation of an adjusted similarity score, this value is used for the instances of partial correctness.

Evaluating the adjusted similarity score for the fuzzy boundary decision formulation outputs a value of 0.83, which is now significantly better than the previous similarity score of 0.59. This suggests that when a more realistic decision formulation is applied the model is interpreted as being much more likely (33% more) than random chance to inform the correct decision, whereas previously the model could only be interpreted as 9% more likely than random chance to inform the correct decision.

The results of the fuzzy boundary decision formulation are plotted in Figure 44; these are the same results as those in Figure 35, but replotted with the fuzzy boundary decision formulation.

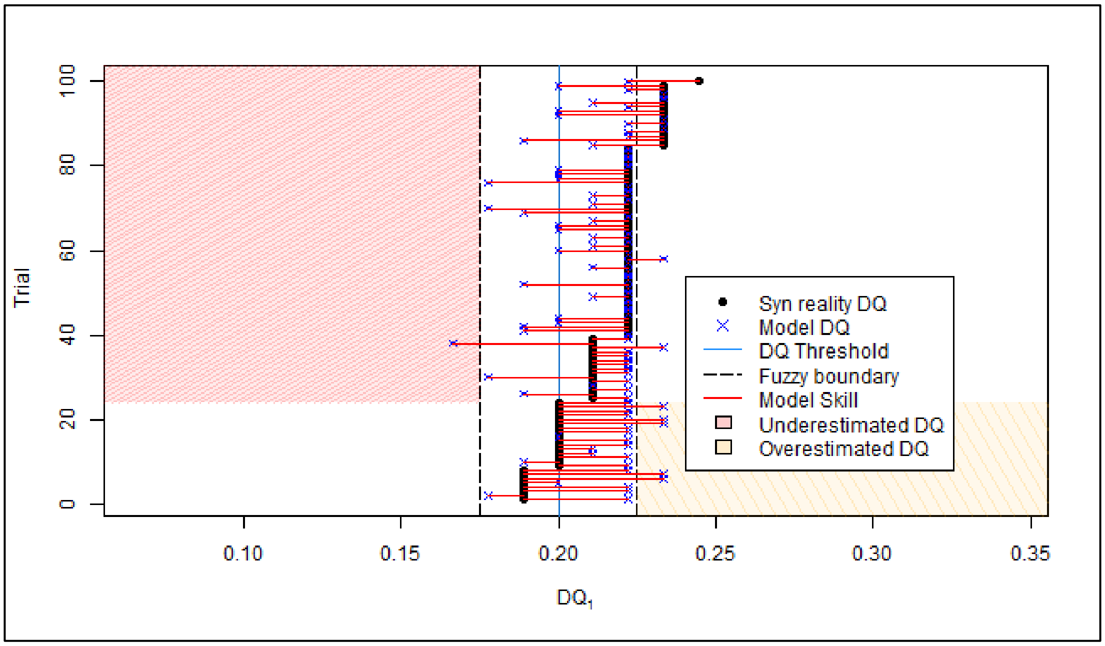

Figure 44. Decision quantity 1 plot for fuzzy decision formulation

The second adjusted decision formulation, referred to as simply the ‘fuzzy’ decision formulation, offers an alternative decision formulation with less complexity than the fuzzy boundary decision formulation. In this case, the decision uses the original decision quantity threshold of 20% to partition the decision space into two decisions, but evaluates instances where the model skill is small to be counted as partially correct (similar to the fuzzy boundary decision formulation). Here, any instances of partial correctness are evaluated on a scale of zero to one, calculated as:

$$max\left\{1 - {}^{|\Delta\phi_n|}/_{q}\, , 0\right\} \tag{6}$$

where q is a model skill threshold assumed as 0.025, and would normally would be prescribed by a decision-maker to convey the desired model accuracy. The q functions similarly to the *AMAMS* in equation (5) in providing a comparable value for interpreting the model skill. This formulation is another method for allowing some flexibility in interpreting the decision quantity pairings in a realistic way. With this fuzzy decision formulation, the computed similarity score is 0.62 instead of 0.59, which is a slight improvement.

The sensitivities of the similarity score for the fuzzy boundary decision formulation and the fuzzy decision formulation, in comparison to the sharp binary decision formulation, are shown in Figure 45. The fuzzy boundary decision thresholds are kept as ±0.025 from the decision quantity threshold for this calculation, and the *q* value is kept at 0.025 for the fuzzy decision sensitivity.

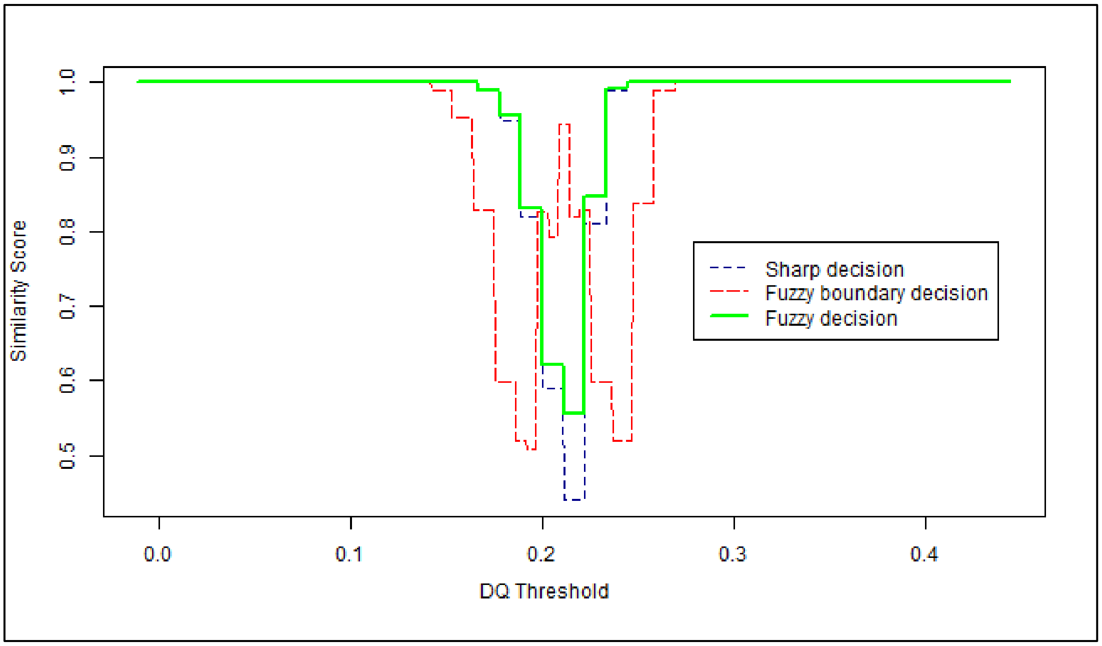

Figure 45. Similarity score sensitivity for sharp and fuzzy boundary and fuzzy decision formulations

The fuzzy boundary decision is shown to have a general improvement to the similarity score from the sharp decision formulation, although the complexity of the sensitivity to the decision quantity threshold in this case likely inhibits an understanding of the model's ability to inform decision-making as a function of the decision quantity threshold. In contrast, the fuzzy decision sensitivity shows an overall improvement to the similarity score with decision quantity threshold from the sharp decision formulation without a change to the shape of the curve. Thus, although the fuzzy boundary decision formulation yielded a larger improvement to the similarity score at a decision quantity threshold of 20%, the fuzzy decision formulation is less sensitive to the decision formulation, and is therefore easier to interpret.

In this exercise of creating a more realistic decision formulation, the importance of both implementing various decision formulations and assessing the formulations with a sensitivity analysis were demonstrated. The similarity score obtained from a single DCT experiment is likely to be a function of the decision quantity threshold and decision formulation implemented, and an understanding of the similarity score sensitivity to the decision quantity threshold is useful in understanding the DCT results. This highlights the importance of the decision formulation step in the DCT framework, and suggests that testing multiple decision formulations is a beneficial for obtaining more information about the effect of the selected decision formulation on the interpretation of the model's ability.

4.3 Case Study Conclusions

Subsequent to the analysis performed in this chapter, a likely set of conclusions from the DCT experiment would be:

1. If the June months are used as the basis for the decision, then the model needs some adjustments for use in decision-making, assuming that an upper bound of 59% likely to inform the right decision is not sufficient for reliability.
2. If the August months are used as the basis for the decision, then the model is deemed sufficient for informing the decision, and in fact the model is likely not required to inform the decision (beyond the analysis here).

This is useful information for the decision-maker to have, and is explained in a way that is interpretable for decision-making directly.

In the event that the decision should be based on June flows (DQ_1), DCT could be used as a framework for testing adjustments in the model-building procedure in a decision-making context. A likely first step to try to improve the model-building procedure would be to use a different calibration objective that captures more relevant information to the model's decision, which would be justified given that there was no correlation between the objective function and model skill.

A demonstration of this process was provided, where the objective function used in calibrating the evaluated model was adjusted, and the DCT experiment was re-run. The bootstrapping technique was used to obtain a distribution of the similarity score for each experiment, and a t-test was performed on the difference in mean similarity score of each distribution. The test concluded that the improvement in similarity score with the new objective function was statistically significant with a p-value of less than 1E-6, indicating that the adjustment of the objective function for calibration was successful in improving the ability of the model to inform decision-making. This provides an example of how one might test model improvements in a DCT framework, although other procedures are also possible.

The DCT framework could also be applied in a Scenario Discovery-inspired approach to investigate the cause of the systematic bias in model skill, or the cause of the particularly poor model trials. All of these potential follow-up experiments demonstrate the ability of DCT to provide not only a framework in which to evaluate the decision-making ability of a model, but a framework for guiding meaningful improvements in model-building for decision-making.

Finally, the importance of carefully selecting the decision formulation, the potential impact of the decision formulation on the interpreted ability of the model to inform decision-making, and the importance of testing the sensitivity of the model performance to the decision quantity threshold, was demonstrated. The use of multiple decision formulations is recommended for testing the impact of the decision formulation on the apparent ability of the model, particularly for binary decision formulations with a sharp decision quantity threshold.

The next chapter of the thesis examines the second case study of this thesis, which explores the ability of a hydrologic model to inform a data gauging decision for upstream reservoir releases in the absence of measured flow data.

Chapter 5
Decision Crash Testing in a Data Collection Application

This chapter details the setup and results of the second Decision Crash Testing (DCT) case study. This case study deploys the DCT framework to evaluate the ability of the model developed in the previous chapter to inform a hypothetical data gauging decision for the upper portion of the Madawaska watershed. This case study differs from the first in a number of key ways, including the evaluation of a model to inform decisions based on hindcasting rather than future predictions, and is complementary to the first case study in illustrating the potential application of DCT.

5.1 Case Study Setup

This section will discuss the setup of the DCT framework in the context of a data gauging decision in the Madawaska watershed. The context of the hypothetical data gauging decision, the quantitative decision formulation, and the details of the DCT experimental setup are discussed in this section.

5.1.1 Data Gauging Decision

The hypothetical decision used in this case study is based upon the question: "at which location should a new flow gauge be built?" In the Bark Lake Dam model described in section 3.2, there exist two upstream dams: Booth Lake Dam and Galeairy Lake Dam. Both of these are controlled dams with unknown stage-outflow rating curves, and the Bark Lake model would benefit from actual data about outflow from either dam. However, it is unclear which gauge site would be most beneficial. The potential flow gauge locations in this case study are immediately downstream of these two dams, with the goal of collecting the most possible information for inflow prediction at Bark Lake Dam. The location of these potential gauges in relation to the important dams in the watershed is shown in Figure 46.

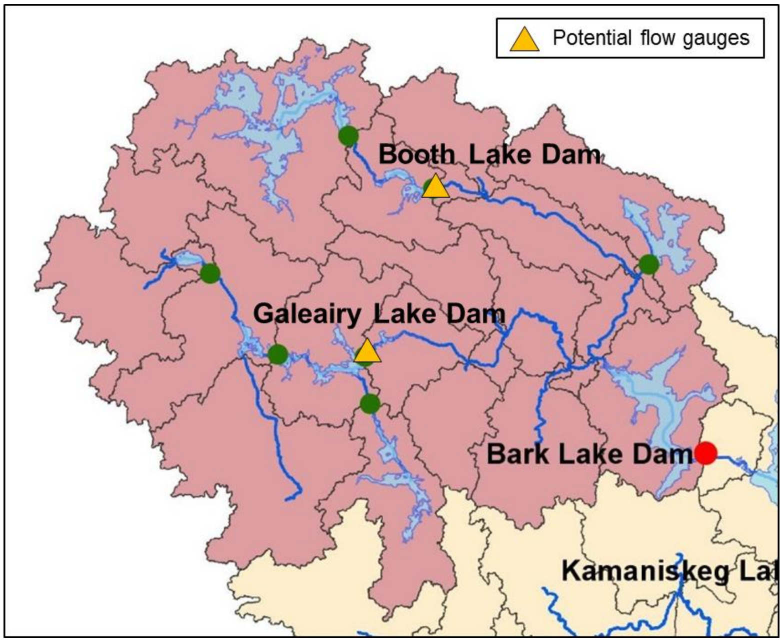

Figure 46. Location of potential flow gauges in data gauging DCT case study

Here it is assumed that both potential gauge locations are equal in terms of capital cost, maintenance cost, ease of access, etc., and the only deciding factor is the value of information obtained from a flow gauge at the given location. It is also assumed that building gauges at other locations, no gauges, or both gauges are not feasible options; for ease of the experimental setup, the decision is forced to be binary. In reality, this would not be the case. However, some simplification of the decision-making process is required for evaluation in the DCT framework.

The typical approach to answer this question would be to use the existing calibrated model from section 3.2.7 to inform a decision, using some decision formulation to link model outputs to a decision. For example, the decision could be based on a comparison of annual flow volumes. In this experiment, the ability of the model to provide the right decision to that question is tested using the DCT framework. In lieu of a model, it may be possible to answer this question by making some reasonable assumptions about which reservoir is likely to release more water. For example, selecting the reservoir with a larger upstream drainage area might be a reasonable decision; the drainage area of the area upstream to Galeairy Lake is approximately twice that of Booth Lake, thus by this criterion, the decision would be made to place the gauge downstream of Galeairy Lake Dam. However, since the goal is ultimately to build a model for inflow prediction at Bark Lake dam, the ability of the proposed model to inform this decision is tested in this chapter.

5.1.2 Decision Quantity Definition

The decision quantity is meant to quantitatively represent the amount of useful information for inflow prediction captured by a gauge at the given location. There are many ways that this could be defined. In this case study, three different decision quantities are used to represent the useful information captured at a given gauge location. These decision quantities are calculated separately, and used to evaluate the sensitivity of the DCT results to the choice of decision quantity used.

The first decision quantity captures the proportion of days where the daily average outflow from Booth Lake is larger than the daily average outflow from Galeairy Lake. This is defined as:

$$DQ_1 = \frac{2}{N}\sum_{i=1}^{N} S(Q_{67,i} > Q_{71,i}) \tag{1}$$

where DQ_1 is the first decision quantity (dimensionless), S is a function that evaluates the given statement and returns a value of 0 or 1 corresponding to whether it is true or false, $Q_{67,i}$ is the outflow from Booth Lake on the *i*th day of the evaluation period (m^3/s), and $Q_{71,i}$ is the outflow from Galeairy Lake on the *i*th day of the evaluation period (m^3/s). The evaluation period is from 2008-10-01 to 2011-09-30, thus *N* has a value of 1095 days. A factor of 2 is introduced a scaling factor to keep DQ_1 within the range [0, 2] and a decision quantity threshold of 1, such that if DQ_1 is greater than one, the decision is to place the gauge at Booth Lake, and if DQ_1 is less than one, the decision is to place the gauge at Galeairy Lake.

The second decision quantity relates the average annual flow volumes from both reservoirs, and is calculated as:

$$DQ_2 = \sum_{i=1}^{3} V_{67,i} \Bigg/ \sum_{i=1}^{3} V_{71,i} \tag{2}$$

where DQ_2 is the second decision quantity (dimensionless), $V_{67,i}$ is the cumulative outflow volume in the *i*th water year (defined from October 1st to September 30th) in m^3, and $V_{71,i}$ is the cumulative outflow volume in the *i*th water year in m^3. The same evaluation period as DQ_1 is used for all three decision quantities, thus there are three water years used in the summation. This decision quantity follows the same decision mapping rules as DQ_1 (i.e., a decision quantity threshold of 1).

The third decision quantity captures the proportion in average annual peak flows during the evaluation period, defined as:

$$DQ_3 = \sum_{i=1}^{3} Qp_{67,i} \Big/ \sum_{i=1}^{3} Qp_{71,i} \tag{3}$$

where DQ_3 is the third decision quantity (dimensionless), $Qp_{67,i}$ is the peak outflow from Booth Lake in the i^{th} day water year of the evaluation period (m^3/s), and $Qp_{71,i}$ is the peak outflow from Galeairy Lake in the i^{th} day water year of the evaluation period (m^3/s). This decision quantity also follows the same decision mapping rules as DQ_1.

In this case study, all three decision quantities are computed independently and used to gather inferences about the decision-informing ability of the model. This is one way to compare the impact of decision formulation on the model evaluation, and evaluate the sensitivity of the DCT results to the selection of the decision quantity used in the experiment.

5.1.3 Synthetic Reality Generation

The synthetic reality generation was performed using the calibrated Bark Lake Dam model, discussed in section 3.2.7. The purpose of each synthetic reality generation is to simulate a plausible reality in which to test the model undergoing evaluation. Each synthetic reality generation must:

1. Include sufficient information to determine the ‘correct’ decision, using a consistent decision formulation as the evaluated model, and
2. Generate the synthetic information required for building the evaluated model.

In this experiment, the hydrologic model undergoing evaluation will not be changed in terms of structure, but does require the synthetic stage time series at Bark Lake Dam to perform calibration. The synthetic reality must also produce the outflows from each of the two upstream reservoirs for evaluation of the ‘correct’ synthetic decision.

The synthetic realities were obtained using a random parameter sampling of the reservoir outflow power law parameters, *a* and *b*, for the Booth and Galeairy dams. All other parameters and boundary conditions in the model were unchanged. The parameters were randomly generated using a deviation of ±40% from the calibrated values for the set of *a* and *b* parameters, for a given reservoir at a time (i.e., in a given synthetic reality, the Booth Lake *a* and *b* parameters could both change by +30%, and the Galeairy Lake parameters could both change by -5%). Applying the same percentage increase or decrease to the set of *a* and *b* parameters was done to avoid the counteracting effects of one parameter increasing and the other decreasing, which provides a more uniform spread in the distribution of the decision quantities.

In order to ensure that the synthetic reality generations were plausible, a basic check on the stationarity of the reservoir stage in both upstream reservoirs was included. This metric, referred to hereafter as a ‘plausibility criterion’, checked the assumption that the overall volume (represented by stage) of the reservoirs is approximately stationary on an annual basis. This was quantified by checking whether the initial reservoir stage was within the 10% and 90% percentiles of the reservoir stage during the simulation period (2007-10-01 to 2011-09-30). This was done for both upstream reservoirs, and any randomly generated synthetic realities where either upstream reservoir failed this plausibility criterion were excluded from the DCT experiment.

The synthetic realities were generated prior to the DCT experiment, which allows for their examination and adjustment. A larger-than-required number of synthetic realities, 2000 in this case, were generated for analysis. The removal of synthetic realities with one or more nonstationary upstream reservoir stages left approximately 1600 synthetic realities for use in the DCT experiment. These were resampled and adjusted to create three forms of the case study (referred to hereafter as Subcases):

1. 100 synthetic reality generations randomly sampled, such that the decision quantities follow the natural distribution created by the synthetic reality generation.
2. 100 synthetic reality generations sampled such that DQ_3 follows an approximately uniform distribution.
3. The exact synthetic reality generations from #2 above, but with random noise added to the reservoir stage synthetic data provided to the evaluated model.

The third Subcase was setup to determine if errors in the synthetic data provided to the evaluated model would be a detriment to the ability of the model to inform the correct decision, relative to Subcase 2. The random noise was created by (a) adding a random, uniformly distributed 0.5% fluctuation to each point in the reservoir stage time series, and (b) applying an 8-point moving average smoothing to the perturbed time series. This ensures that the noisy stage series is still relatively smooth but still contains random deviations from the ‘true’ synthetic series. This procedure is conceptually similar to the procedure of Crow and Loon (2006), where a random noise filter was used to test the impact of errors in the observed data. An example of adding noise to a synthetic stage time series is shown in Figure 47.

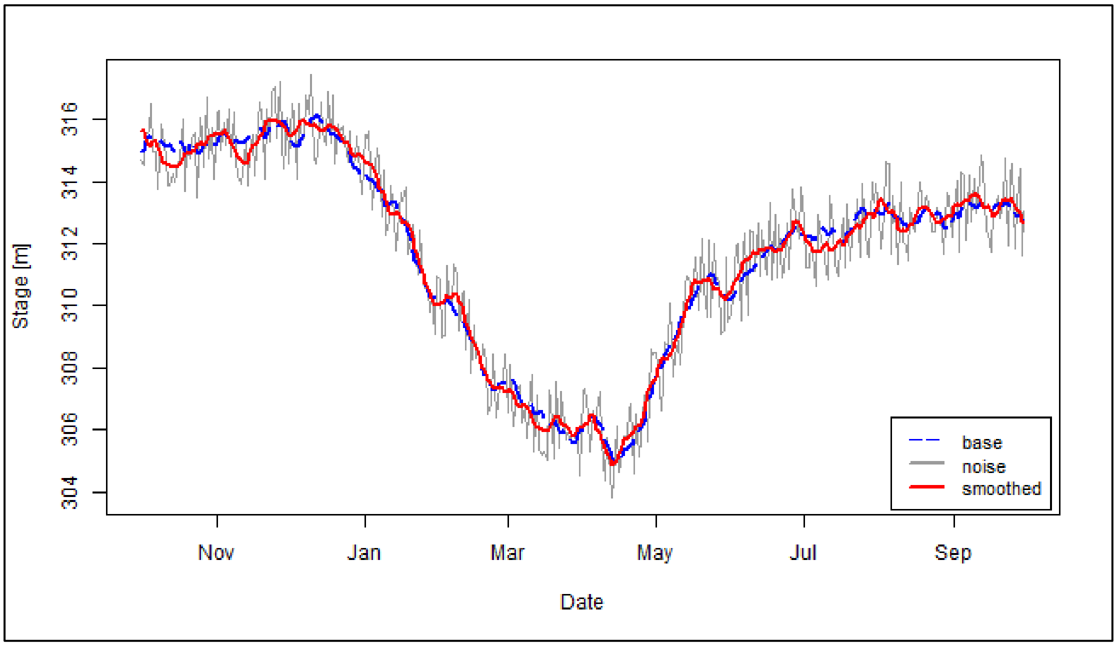

Figure 47. Example of random noise added to Bark Lake Dam stage time series

The synthetic reality generations produced a distribution of each decision quantity. The decision quantity distributions for Subcase 1 are shown in Figure 48, and in Figure 49 for Subcases 2 and 3.

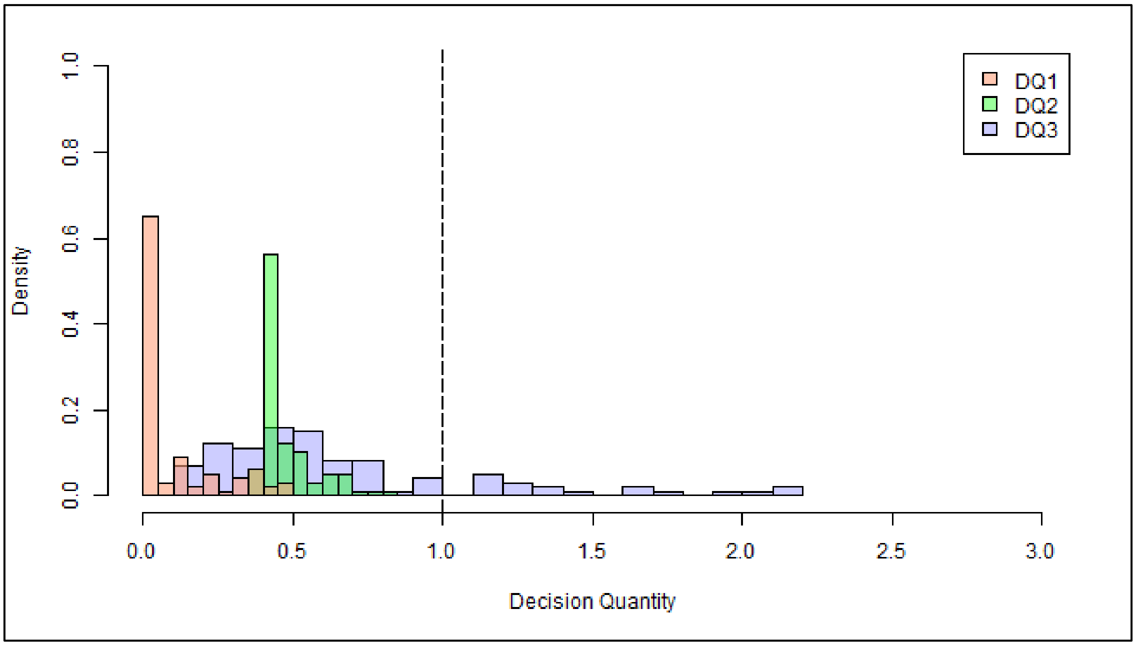

Figure 48. Decision quantity distributions for Subcase 1

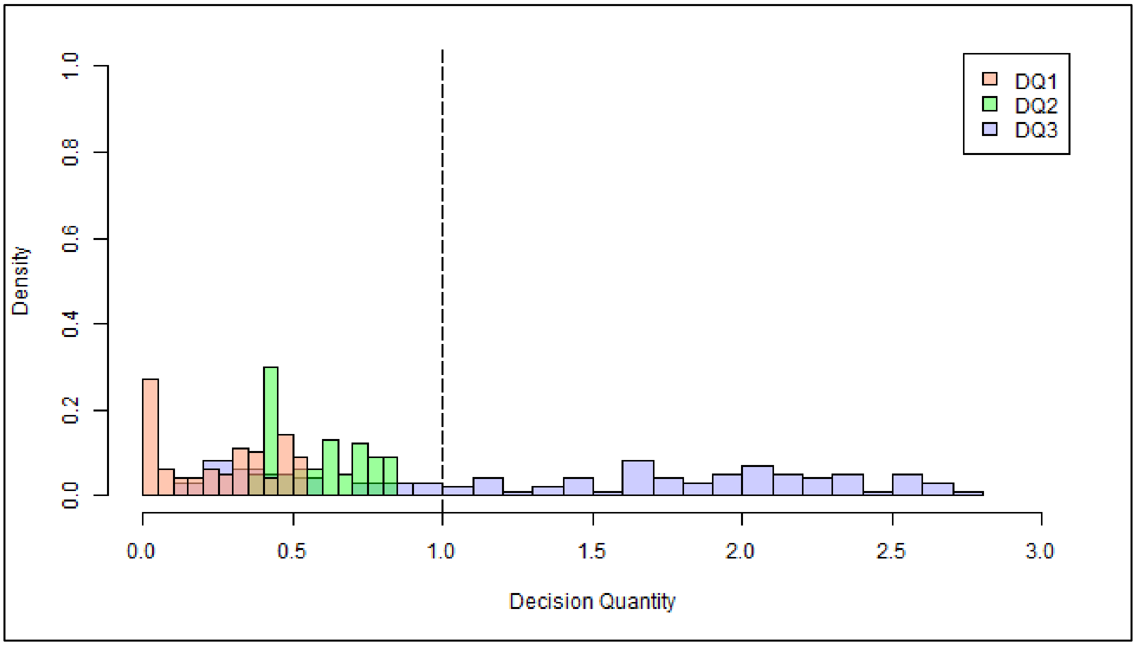

Figure 49. Decision quantity distributions for Subcases 2 and 3

These figures show that in each of the Subcases, the distributions for DQ_1 and DQ_2 fall entirely on the left side of the decision quantity threshold of 1.0, while DQ_3 is distributed over either side of the decision quantity threshold. Thus, the decision difficulty will be zero for DQ_1 and DQ_2 and non-zero for DQ_3, such that if the decision was based on DQ_3, the decision would be more difficult and the expectation would be a smaller similarity score. The comparison between Figure 48 and Figure 49 also shows the effect of the uniform sampling performed in creating a more uniform distribution of synthetic decision quantities.

Once the synthetic realities are deemed fit for use in the DCT experiment, a number of metrics can be computed to describe the synthetic reality generations. The decision difficulty and average maximum allowable model skill (AMAMS) were evaluated. The summary metrics describing the synthetic realities are presented in Table 7.

Table 7. Summary metrics for synthetic reality generations in data gauging case study

	Subcase 1		**Subcase 2 and 3**	
DQ	**Decision Difficulty**	**AMAMS**	**Decision Difficulty**	**AMAMS**
DQ_1	0.00	0.91	0.00	0.65
DQ_2	0.00	0.56	0.00	0.36
DQ_3	0.22	0.62	0.54	1.03

Table 7 shows that the uniform resampling of synthetic realities in Subcases 2 and 3 creates a generally more difficult decision for the model to answer than the sampling Subcase 1, shown by the generally larger decision difficulty and smaller AMAMS. The decision difficulty is also zero for DQ_1 and DQ_2 in all Subcases, given that all of the synthetic realities fall on the left side of the decision quantity threshold. This is useful information in and of itself, since it shows that if the decision formulation was based on either of these two decision quantities, the decision may lack the difficulty that would otherwise require the use of a computational model to inform the decision.

In each instance of a synthetic reality generation, the associated synthetic stage series at Bark Lake was provided to the evaluated model for use as data in calibration. The evaluation of the hydrologic model being tested is discussed in the next section.

5.1.4 Model Evaluation Setup

The evaluated model is of the same structural setup at the synthetic reality model, both of which are based on the Bark Lake Dam model discussed in section 3.2.7. The model receives the ‘observed’ synthetic reality stage data from the synthetic reality simulation and undergoes an automatic model calibration. The model is run from 2007-10-01 and is calibrated on the data from 2008-10-01 to 2011-09-30, which includes a one-year warm-up period. The model is calibrated to the RMSE of Bark Lake

Dam stage using the DDS algorithm, with a calibration budget of 2000 runs. The RMSE was inversely weighted in time, as suggested in section 3.2.6 as a solution to the long memory issue with stage calibration. The same reservoir plausibility criterion used in the synthetic reality generations is applied as a penalty function in the calibration for the evaluated model, allowing the automatic calibration to avoid solutions that would not have been sampled by the synthetic reality. This is something that makes the model’s ability to inform the correct decision easier; in reality, it would likely be assumed (but not known) that solutions which fail the plausibility criterion are not feasible. The decision quantities are evaluated for the same period as the calibration, thus the model is being tested in its ability to hindcast.

The parameters used in the calibration of the evaluated model include the two sets of *a* and *b* reservoir parameters for Booth Lake and Galeairy Lake, as well as 26 other hydrologic model parameters with defined ranges. Note that in the synthetic reality model only the four *a* and *b* parameters are varied, thus the 26 other parameters are included to provide a more realistic uncertainty in model parameters during calibration. The list of parameters included in calibration is included in Appendix C.

The model was calibrated for each Subcase with 100 synthetic reality generations, and relevant summary metrics were computed (presented in the next section). This required more than 600 000 model runs (2000 run calibration budget, 100 synthetic realities, 3 Subcases), illustrating the need for a very fast hydrologic model to make this type of study feasible with typically available computational resources.

5.2 Results and Discussion

This section presents the DCT results of the three Subcases in the data gauging decision. In this case study, the main pieces of the analysis are the DCT summary metrics (similarity score and model skill) and the associated plots, illustrating the distribution of each. Correlations between the calibrated objective function, RMSE, and model skill can also be inferred by examining a simple linear regression fit. The DCT summary metrics are shown in Table 8.

Table 8. Summary metrics for DCT results in data gauging case study

	Sim. Score	MS Mean* $\mu_{\Delta\phi}$	MS StDev* $\sigma_{\Delta\phi}$	R^2**	Slope**
Subcase 1					
DQ_1	1.00	-0.09	0.14	0.02	-0.11
DQ_2	1.00	-0.05	0.09	0.00	-0.03
DQ_3	0.82	-0.09	0.47	0.01	0.28
Subcase 2					
DQ_1	1.00	-0.25	0.19	0.00	-0.05
DQ_2	1.00	-0.16	0.15	0.00	0.02
DQ_3	0.35	-0.85	0.80	0.00	-0.26
Subcase 3					
DQ_1	1.00	-0.25	0.19	0.00	-0.06
DQ_2	1.00	-0.16	0.15	0.00	0.01
DQ_3	0.35	-0.85	0.80	0.00	-0.19

*MS = model skill

**from linear regression of MS onto calibrated RMSE

From the summary metrics in Table 8, a number of important generalizations can be drawn:

- The similarity score for DQ_1 and DQ_2 is 1 in all cases, indicating that the model informs the decision correctly in each trial. This is not too surprising, since the decision difficulty was 0 for those decision quantities (unlike DQ_3).
- The overall decision-making ability of the model with respect to DQ_3 is much better in Subcase 1, when the synthetic realities have not been resampled uniformly, and thus a greater number of synthetic realities remain on one side of the decision quantity threshold.
- There appears to be essentially no correlation between the calibrated model RMSE and the model skill, indicated by both the near-zero regression slopes and the near-zero R^2 values in all Subcases. This suggests that the calibration objective is not sufficient to describe the decision-making ability of the model.
- There was essentially no impact on the results by the added noise to the data, based on the comparison of results between Subcases 2 and 3. It is likely that the effects of introducing noisy synthetic data are dominated by the inability of the model to determine the decision quantity, and the noisy data may have a more dominant role in a better-performing model.

The summary metrics also showcase the importance of the selection of the synthetic realities used in the experiment. In the summary metrics for DQ_3, the similarity score (i.e., estimated probability of the model informing the correct decision) decreases from 82% to 35% when the synthetic reality is resampled uniformly. This is the difference between something that likely appears to be quite good to decision-makers versus something that is less likely to be correct than a coin flip (which has a 50% chance of informing the correct decision in this binary case).

The plots of synthetic and modelled decision quantities are also helpful in interpreting the DCT results. The decision quantity plots are presented for Subcase 2 only, but are similar to Subcase 1 and nearly identical to Subcase 3.

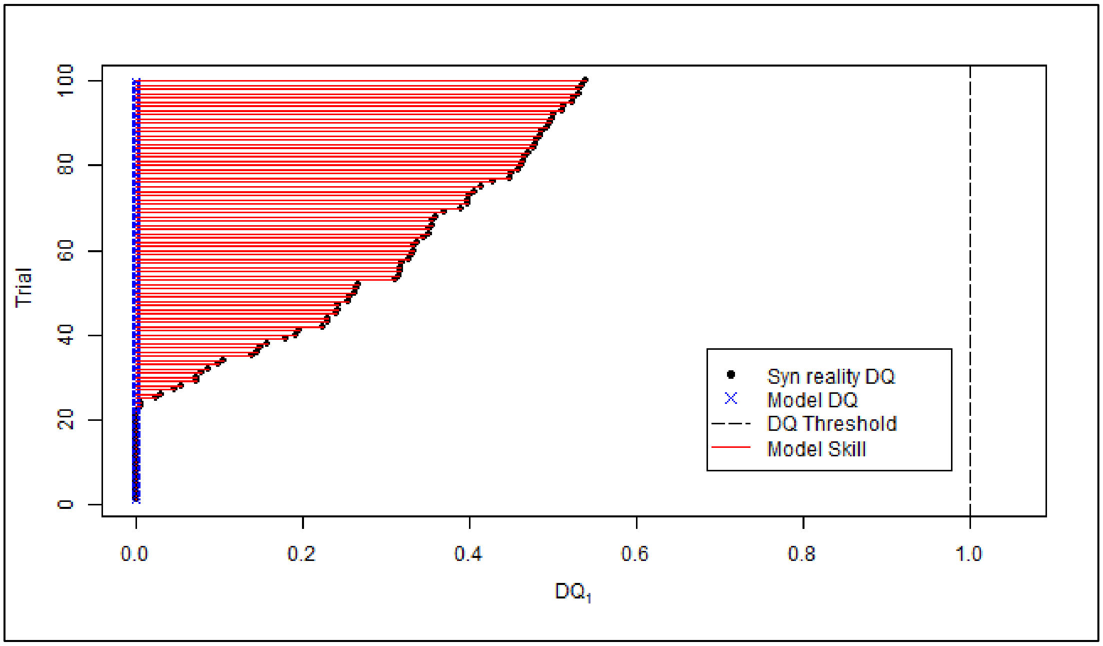

Figure 50. Decision quantity 1 plot for Subcase 2 in data gauging decision

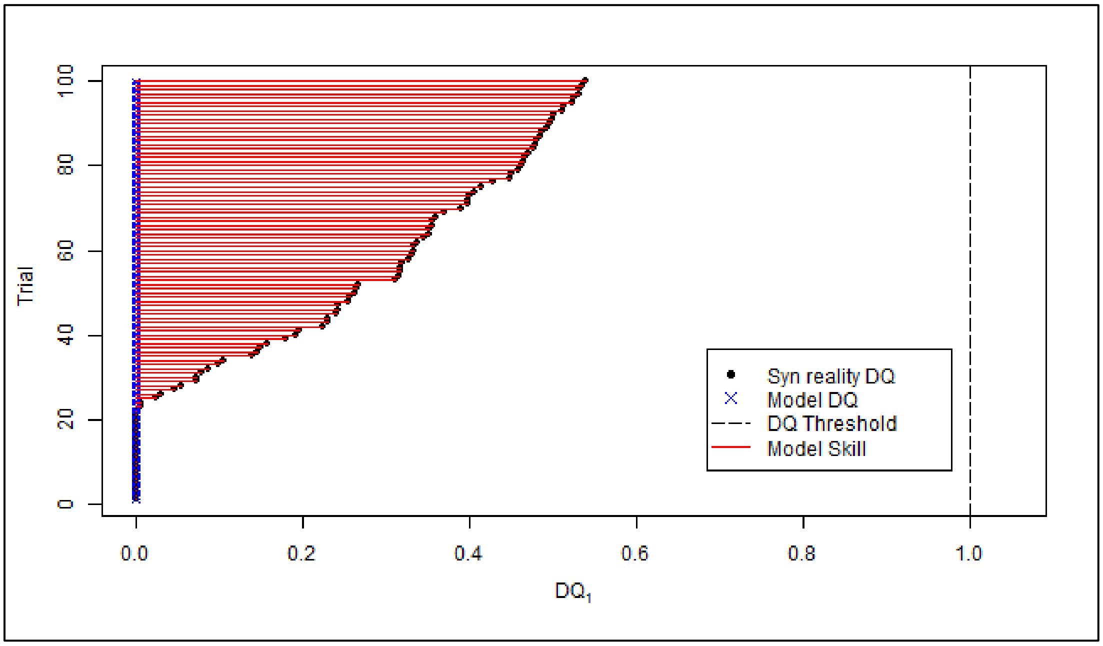

Figure 51. Decision quantity 2 plot for Subcase 2 in data gauging decision

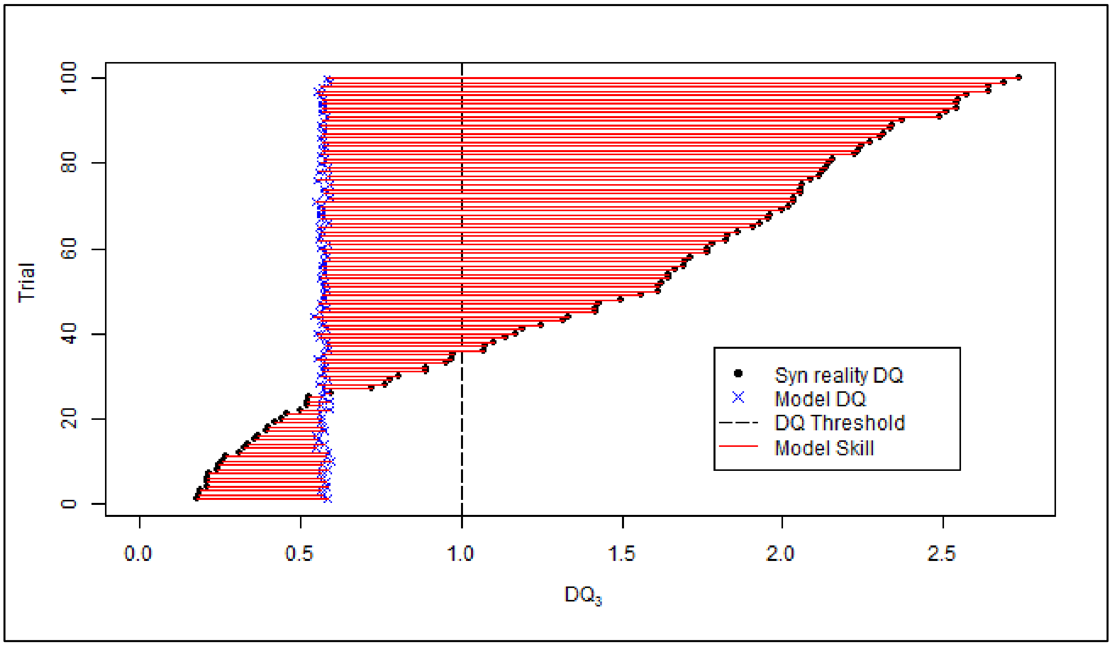

Figure 52. Decision quantity 3 plot for Subcase 2 in data gauging decision

The above figures show that the model decision quantity produced (for each of the three potential decision quantities) is almost identical, regardless of the synthetic reality decision quantity. This indicates that the model is essentially incapable of recreating the synthetic decision quantity using the chosen calibration approach, and the DCT metrics of model skill and similarity score are controlled by the synthetic reality distribution used. This again illustrates the importance of carefully analyzing and justifying the synthetic realities selected prior to use in DCT. However, the model still informs the decision correctly 100% of the time when decision quantities 1 and 2 are used, simply because the decision is quite easy in those cases.

The independence between the model-evaluated decision quantity and the synthetic decision quantity is also an indicator that either (a) the 'observed' stage data does not contain enough information for the model to discriminate between solutions during calibration and be useful in decision-making, or (b) the chosen calibration procedure was not sufficiently capable of training the model for decision-making applications. A combination of (a) and (b) is also possible, and it is difficult to determine which is the larger contributor without further tests. An interesting follow-up experiment would be to repeat the experiment with various calibration budgets and objective functions to determine which configuration, if any, allow the model to improve its decision-making ability. An improvement in the model's decision-making ability without change to the synthetic realities or information passed to the model from the synthetic reality would indicate that the calibration procedure of the model was inadequate, rather than the information.

5.3 Case Study Conclusions

This chapter explored the use of a hydrologic model to inform a decision regarding the placement of a flow gauge for improved data collection, with the purpose of building a forecasting model to support reservoir operations. In short, the DCT experiment showed that the evaluated model is incapable of reproducing the correct decision quantity, and tended towards the same decision quantity intendent of what the 'true' decision quantity was. The experiment also shows that the choice of decision formulation is highly important; under some decision formulations the decision was a very easy one, and likely would not even require a model. This was true when the decision was based on the proportion of days with higher reservoir outflows (DQ_1), and when the decision was based on cumulative outflow volumes (DQ_2). However, when the decision was based on average annual peak flow from the reservoirs (DQ_3), the decision became much more difficult, and the probability of the model informing the correct decision was worse than a coin flip, depending on the generation of the synthetic realities. This reveals the need in DCT for careful selection and reporting of the synthetic realities.

The DCT experiment performed here was quite revealing in the ability of the model to inform the correct decisions, as well as the impact of the decision formulation. If this decision was to be performed for a real project, the recommendation would likely be to determine which decision formulation is most applicable, and depending on the choice, either choosing to gauge at the Galeairy Lake Dam without the use of a model or investing more resources to improving the ability of the model to inform the decision. The DCT framework could be used to guide further improvements to the model if the latter pathway is selected, as demonstrated in the previous chapter.

Chapter 6
Conclusions

The goals of this thesis were to both (a) demonstrate a novel method of model calibration for hydrologic models with reservoirs, and (b) present the motivation for and application of Decision Crash Testing (DCT).

The first goal was fulfilled in Chapter 3, where the development and calibration of two hydrologic models was presented. The calibration procedure used the observed stage data directly in calibration, rather than inflow data estimated from level pool routing. The use of stage data for calibration overcame the issues of calibrating to estimated data with numerical artefacts, but also presented a number of new challenges; these were largely related to the long memory of reservoir stage, which meant that errors in stage accumulated and propagated throughout the simulation. However, it was also shown that the stage does not need to match perfectly in order to provide reasonable inflow estimates. A basic split-sample validation test on the two models showed that both had large deviations in reservoir stage during the validation period without significant detriment to the matching of inflows. The improvement in stage simulation from calibration directly to reservoir stage, rather than calibration to level-pool routing estimated inflows, was also demonstrated. The potential for this method to be deployed in future studies, for both research and operational applications, was illustrated.

The motivation for the DCT framework was presented in Chapter 2, where the call in literature for rigorous validation methodologies was discussed, and the lack of development in new validation methodologies was also highlighted. Chapter 2 also included the general conclusion from literature that validation of models is possible, but only in a specific context and scope of application. This fit well with the theme of using models for decision-making applications specifically, where the communication gap between scientists and decision-makers was identified. The ways in which model outputs are presented to decision-makers are currently not useful for decision-making. Thus, the motivation for DCT is twofold; DCT (a) answers the call in literature for new and rigorous validation methodologies, and (b) provides a framework for validating models in a specific application that is much more readily interpretable by decision-makers.

The DCT framework was demonstrated using two case studies in reservoir management. The first case study involved a decision based on the relative impacts of adjusted reservoir operation on downstream flows in the Mountain Chute subcatchment, and the second a data gauging decision for upstream flows in the Bark Lake subcatchment. In both cases, the DCT framework was demonstrably capable of providing useful information on both the difficulty of the decision and the ability of the model to inform the decision.

In the first case study, the model was found to be somewhat useful in informing the decision, but only slightly better than random chance, with a similarity score of 59%. However, the potential of the DCT framework for guiding meaningful model improvements in a decision-making context were demonstrated with the use of hypothesis testing. In addition, a re-formulation of the decision demonstrated how a more realistic representation of the decision-making process can influence the perceived ability of the model to inform the correct decision. The importance of testing the sensitivity of the similarity score to the decision quantity threshold in each formulation was also demonstrated.

In the second case study, the evaluated model was tested using three different decision quantities deployed simultaneously, which was done to explore the ability of the model using various plausible decision formulations. The model was found to be incapable of determining the correct decision quantity in all three cases, which were all essentially independent of the respective synthetic decision quantity. However, the decision was easy enough for two of the decision quantities that the model would likely not have been required to inform the decision, since the evaluated model was able to correctly inform the decision without being capable of estimating the decision quantity. This case study demonstrated the ability of DCT framework to gain information about the decision formulation itself, even in the case that the model is not deemed trustworthy.

In both case studies, the importance of carefully analyzing the synthetic realities that are generated, and testing multiple decision formulations simultaneously, was demonstrated. Both the method in which the synthetic realities are generated, and the decision quantity and decision quantity thresholds selected, can influence the DCT results to make a poor model appear highly skilled and vice-versa.

6.1 Contributions to Literature

This thesis presents two primary contributions to the literature. The first contribution is the novel calibration method for reservoir models. This has only been presented once prior by Gijsbers (2015), and this thesis provides the first known example of this method in a North American watershed.

The second and most vital contribution is the introduction and demonstration of DCT, which answers several challenges in the literature for novel validation methodologies and improved communication between scientists and decision-makers. The DCT framework has the potential for widespread use in model validation studies, both within and outside of water resources applications. The DCT framework was demonstrated with two case studies in reservoir management, which serve to both illustrate the potential for DCT to inform decision-making in a specific context, and provide a template for future studies in deploying the DCT framework.

6.2 Future Opportunities for DCT

This thesis presents several opportunities for future work and research involving the DCT framework, including both novel applications of the method and further development the DCT framework itself. DCT has a large potential for application to many fields of study outside of water resources due to its generic procedure. The main requirements for deployment of DCT are that:

1. A decision-maker needs to provide a decision,
2. A model is deployed to inform the decision,
3. The model-building procedure (or a portion of it) can be setup programmatically, and
4. The decision formulation is quantitatively explicit.

These requirements are relatively easy to satisfy for many decision-making exercises, and allow for DCT to be deployed in a wide range of applications, such as transportation engineering, structural analysis, machine learning, etc. It is hoped that future research will explore the application of DCT beyond the field of water resources.

The DCT framework itself also has opportunities for future development. As highlighted in this thesis, DCT can be applied for uses beyond direct model evaluation, such as testing for model improvements and evaluating the decision formulations themselves. This thesis also illustrated the use of future research in further understanding the DCT framework, including the generation of synthetic realities and the impact of the decision formulation on the model evaluation. Further to this, the DCT framework can also be expanded by:

- Expanding the decision formulation to use multiple decision quantities, and increasing the dimensionality of the decision space. This would require a multivariate generation of synthetic realities as well as a more complex model evaluation, although otherwise the same procedures could be followed.

- Implementing DCT for model ensembles instead of single models, where the decision is based on the output from a set of models rather than a single model.
- Incorporating uncertainty into the decision quantities, such that each decision quantity becomes a distribution or range of decision quantities. The decision formulation would need to be updated to handle uncertainty in the decision quantities from the evaluated model.

Many other expansions of the DCT framework presented here are possible, and many novel applications can be performed without major modification to the theory, simply by using different methods for generating synthetic realities or varied decision formulations. It is hoped that future research will see some exploration of these potential developments to the DCT framework.

References

Ahmad, A., El-Shafie, A., Razali, S., Mohamad, Z., 2014. Reservoir Optimization in Water Resources: a Review. Water Resour.Manage. 28, 3391-3405.

Ahmadisharaf, E., Kalyanapu, A.J., Chung, E., 2016. Spatial probabilistic multi- criteria decision making for assessment of flood management alternatives. Journal of Hydrology. 533, 365-378.

Andréassian, V., Perrin, C., Berthet,L., Le Moine, 2009. *HESS Opinions* " Crash tests for a standardized evaluation of hydrological models". Hydrol.Earth Syst.Sci.Discuss. 6, 3669-3685.

Ballester, P.J., Carter, J.N., 2006. Characterising the parameter space of a highly nonlinear inverse problem. Inverse Problems in Science and Engineering. 14, 171-191.

Bankes, S., 1993. Exploratory Modeling for Policy Analysis. Oper.Res. 41, 435-449.

Bankes, S., Walker, W.E., Kwakkel, J.H., 2013. Exploratory Modeling and Analysis. Encyclopedia of Operations Research and Management Science. 3, 532-537.

Ben-Haim, Y., 2010. Info-gap economics : an operational introduction, Houndmills, Basingstoke, Hampshire ; New York, Palgrave Macmillan.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., et al., 2013. Characterising performance of environmental models. Environmental Modelling and Software. 40, 1-20.

Beven, K.J., 2012. Rainfall-runoff modelling the primer, 2nd ed. ed. Chichester, West Sussex ; Hoboken, NJ; Hoboken, Wiley-Blackwell.

Beven, K., 2013. So how much of your error is epistemic? Lessons from Japan and Italy. Hydrol.Process. 27, 1677-1680.

Beven, K., 2008. On doing better hydrological science. Hydrol.Process. 22, 3549-3553.

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., Montanari, A., 2012. Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. Phys.Chem.Earth. 42-44, 70-76.

Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. Water Resour.Res. 36, 3663-3674.

Brown, C., Ghile, Y., Laverty, M., Li, K., 2012. Decision scaling: Linking bottom- up vulnerability analysis with climate projections in the water sector. Water Resour.Res. 48, n/a-n/a.

Brugnach, M., Dewulf, A., Pahl-Wostl, C., Taillieu, T., 2008. Toward a Relational Concept of Uncertainty: about Knowing Too Little, Knowing Too Differently, and Accepting Not to Know. Ecology and Society. 13, 30.

Bryant, B.P., Lempert, R.J., 2010. Thinking inside the box: A participatory, computer-assisted approach to scenario discovery. Technological Forecasting & Social Change. 77, 34-49.

Buttle, J.M., D., 2004. Hydrologic coupling of slopes, riparian zones and streams: an example from the Canadian Shield. Journal of Hydrology. 287, 161-177.

Carmona, G., Varela - Ortega, C., Bromley, J., 2013. Participatory modelling to support decision making in water management under uncertainty: Two comparative case studies in the Guadiana river basin, Spain. J.Environ.Manage. 128, 400.

Carter, J.N., Ballester, P.J., Tavassoli, Z., King, P.R., 2006. Our calibrated model has poor predictive value: An example from the petroleum industry. Reliability Engineering and System Safety. 91, 1373-1381.

Chang, J., Meng, X., Wang, Z., Wang, X., Huang, Q., 2014. Optimized cascade reservoir operation considering ice flood control and power generation. Journal of Hydrology. 519, 1042.

Choong, S., El-Shafie, A., 2015. State-of-the-Art for Modelling Reservoir Inflows and Management Optimization. Water Resour.Manage. 29, 1267-1282.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., et al., 2012. Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. Water Resour.Res. 48, n/a-n/a.

Craig, J.R., 2017. Personal Communication. Decision Crash Testing Theoretical Documentation.

Crow, W.T., Van Loon, E., 2006. Impact of Incorrect Model Error Assumptions on the Sequential Assimilation of Remotely Sensed Surface Soil Moisture. J.Hydrometeor. 7, 421-432.

D’oria, M., Mignosa, P., Tanda, M.G., 2012. Reverse level pool routing: Comparison between a deterministic and a stochastic approach. Journal of Hydrology. 470-471, 28-35.

Davis, D.R., Kisiel, C.C., Duckstein, L., 1972. Bayesian decision theory applied to design in hydrology. Water Resour.Res. 8, 33-41.

Dehotin, J., Braud, I., 2008. Which spatial discretization for distributed hydrological models? Proposition of a methodology and illustration for medium to large-scale catchments. Hydrology and Earth System Sciences. 12, 769-796.

Dessai, S., Hulme, M., 2007. Assessing the robustness of adaptation decisions to climate change uncertainties: A case study on water resources management in the East of England. Global Environ.Change. 17, 59-72.

Donnelly-Makowecki, L., 1999. Hierarchical testing of three rainfall– runoff models in small forested catchments. Journal of Hydrology. 219, 136-152.

Efron, B., 1979. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics. 7, 1-26.

Flügel, W., 1997. Combining GIS with regional hydrological modelling using hydrological response units (HRUs): An application from Germany. Math.Comput.Simul. 43, 297-304.

Fu, C., James, A.L., Yao, H., 2014. SWAT- CS: Revision and testing of SWAT for Canadian Shield catchments. Journal of Hydrology. 511, 719-735.

Gentle, J.E., 1943-, 2009. Methods of Computational Statistics,Computational statistics. Dordrecht ; New York, Springer, pp. 420--422.

Gijsbers, S.F.M., 2015. Thinking inside the box: Using reservoir levels to improve a conceptual rainfall-runoff model in the Umbeluzi River Basin.

Goldstein, M., Rougier, J., 2009. Reified Bayesian modelling and inference for physical systems. Journal of Statistical Planning and Inference. 139, 1221-1239.

Gorissen, B.L., Yanıkoğlu, İ, Den Hertog, D., 2015. A practical guide to robust optimization. Omega. 53, 124-137.

Grayson, R., Moore, I., Mcmahon, T., 1992. Physically based hydrologic modeling: 1. A terrain-based model for investigative purposes. Water Resour.Res. 28, 2639-2658.

Groves, D.G., Lempert, R.J., 2007. A new analytic method for finding policy-relevant scenarios. Global Environ.Change. 17, 73-85.

Guillaume, J.H.A., Kummu, M., Räsänen, T.A., Jakeman, A.J., 2015. Prediction under uncertainty as a boundary problem: A general formulation using Iterative Closed Question Modelling. Environmental Modelling and Software. 70, 97-112.

Gupta, H.V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol.Process. 22, 3802-3813.

Haasnoot, M., Kwakkel, J.H., Walker, W.E., ter Maat, J., 2013. Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. Global Environ.Change. 23, 485-498.

Hingray, B., S., 2010. Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments. Hydrol.Sci.J./J.Sci.Hydrol. 55, 1002-1016.

Höllermann, B., Evers, M., 2017. Perception and handling of uncertainties in water management—A study of practitioners' and scientists' perspectives on uncertainty in their daily decision-making. Environmental Science and Policy. 71, 9-18.

Hu, M., Huang, G.H., Sun, W., Li, Y., Ding, X., An, C., et al., 2014. Multi- objective ecological reservoir operation based on water quality response models and improved genetic algorithm: A case study in Three Gorges Reservoir, China. Eng Appl Artif Intell. 36, 332-346.

Isendahl, N., Dewulf, A., Brugnach, M., François, G., Möllenkamp, S., Pahl-Wostl, C., 2009. Assessing Framing of Uncertainties in Water Management Practice. Water Resour.Manage. 23, 3191-3205.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling and Software. 21, 602-614.

Jones, J.A.A., 1997. Global hydrology : processes, resources and environmental management, Harlow, Longman.

Kamodkar, R.U., Regulwar, D.G., 2014. Optimal multiobjective reservoir operation with fuzzy decision variables and resources: A compromise approach. Journal of Hydro-environment Research. 8, 428-440.

Kasprzyk, J.R., Reed, P.M., Characklis, G.W., Kirsch, B.R., 2012. Many-objective de Novo water supply portfolio planning under deep uncertainty. Environmental Modelling and Software. 34, 87-104.

Keur, P., Brugnach, M., Dewulf, A., Refsgaard, J., Zorilla, P., Poolman, M., et al., 2010. Identifying Uncertainty Guidelines for Supporting Policy Making in Water Management Illustrated for Upper Guadiana and Rhine Basins. Water Resour.Manage. 24, 3901-3938.

Kirchner, J.W., 2006. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water Resour.Res. 42, n/a-n/a.

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrological Sciences Journal. 31, 13-24.

Konikow, L.F., Bredehoeft, J.D., 1992. Ground-water models cannot be validated. Adv.Water Resour. 15, 75-83.

Kouwen, N., Soulis, E.D., Pietroniro, A., Donald, J., Harrington, R.A., 1993. Grouped response units for distributed hydrologic modeling. J.Water Resour.Plann.Manage. 119, 289.

Kwakkel, J.H., Walker, W.E., Marchau, V.A.W.J., 2010. Classifying and communicating uncertainties in model-based policy analysis. International Journal of Technology Policy and Management. 10.

Kwakkel, J.H., Haasnoot, M., Walker, W.E., 2016. Comparing Robust Decision-Making and Dynamic Adaptive Policy Pathways for model-based decision support under deep uncertainty. Environmental Modelling and Software. 86, 168-183.

Kwakkel, J., Haasnoot, M., Walker, W., 2015. Developing dynamic adaptive policy pathways: a computer-assisted approach for developing adaptive strategies for a deeply uncertain world. Climatic Change. 132, 373-386.

Lin, P., Yang, Z., Cai, X., David, C.H., 2015. Development and evaluation of a physically- based lake level model for water resource management: A case study for Lake Buchanan, Texas. Journal of Hydrology: Regional Studies. 4, 661-674.

Liu, Y., Gupta, H., Springer, E., Wagener, T., 2008. Linking science with environmental decision making: Experiences from an integrated modeling approach to supporting sustainable water resources management. Environmental Modelling and Software. 23, 846-858.

Liu, H., 2017. Personal Communication. Reservoir Rating Curve Documentation.

Matott, L.S., 2016. OSTRICH – An Optimization Software Toolkit for Research Involving Computational Heuristics; Documentation and User's Guide. , 8-9.

Matrosov, E.S., Huskova, I., Kasprzyk, J.R., Harou, J.J., Lambert, C., Reed, P.M., 2015. Many-objective optimization and visual analytics reveal key trade-offs for London's water supply. Journal of Hydrology. 531, 1040-1053.

Matrosov, E.S., Woods, A.M., Harou, J.J., 2013. Robust Decision Making and Info-Gap Decision Theory for water resource system planning. Journal of Hydrology. 494, 43-58.

Messager, M.L., Lehner, B., Grill, G., Nedeva, I., Schmitt, O., 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nature Communications: 13603.

Natural Resources Canada, 2017. About Renewable Energy. 2017.

OPG, 2009. Madawaska River Water Management Plan.

Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical mode. Science. 263, 641.

Pagano, T.C., Wood, A.W., Ramos, M.-., Cloke, H.L., Pappenberger, F., Clark, M.P., et al., 2014. Challenges of Operational River Forecasting. J.Hydrometeorol. 15, 1692-1707.

Pang, A.P., Sun, T., 2014. Bayesian networks for environmental flow decision- making and an application in the Yellow River estuary, China. Hydrology and Earth System Sciences. 18, 1641.

Pappenberger, F., Beven, K.J., 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water Resour.Res. 42, n/a-n/a.

Pokhrel, P., Yilmaz, K.K., Gupta, H.V., 2012. Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. Journal of Hydrology. 418 419, 49.

Popper, K., 1959. The logic of scientific discovery, London, Hutchinson of London.

R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ravalico, J.K., Dandy, G.C., Maier, H.R., 2010. Management Option Rank Equivalence (MORE) – A new method of sensitivity analysis for decision- making. Environmental Modelling and Software. 25, 171-181.

Ravalico, J.K., Maier, H.R., Dandy, G.C., 2009. Sensitivity analysis for decision- making using the MORE method—A Pareto approach. Reliability Engineering and System Safety. 94, 1229-1237.

Raven Development Team, 2017. Raven: User's and Developer's Manual v2.7.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-., et al., 2004. Overall distributed model intercomparison project results. Journal of Hydrology. 298, 27-60.

Refsgaard, J.C., Henriksen, H.J., 2004. Modelling guidelines—terminology and guiding principles. Adv.Water Resour. 27, 71-82.

Refsgaard, J.C., van Der Sluijs, J. P., Højberg, A.L., Vanrolleghem, P.A., 2007. Uncertainty in the environmental modelling process – A framework and guidance. Environmental Modelling and Software. 22, 1543-1556.

Refsgaard, J.C., Henriksen, H.J., Harrar, W.G., Scholten, H., Kassahun, A., 2005. Quality assurance in model based water management – review of existing practice and outline of new approaches. Environmental Modelling and Software. 20, 1201-1215.

Refsgaard, J., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T.A., Drews, M., et al., 2014a. A framework for testing the ability of models to project climate change and its impacts. Climatic Change. 122, 271-282.

Refsgaard, J.C., Knudsen, J., 1996. Operational Validation and Intercomparison of Different Types of Hydrological Models. Water Resour.Res. 32, 2189-2202.

Refsgaard, J.C., Madsen, H., Andréassian, V., 2014b. A framework for testing the ability of models to project climate change and its impacts. Climatic Change. 122, 271-282.

Reggiani, P., Sivapalan, M., Majid Hassanizadeh, S., 1998. A unifying framework for watershed thermodynamics: balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics. Adv.Water Resour. 22, 367-398.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecol.Model. 90, 229-244.

Seibert, J., 2003. Reliability of Model Predictions Outside Calibration Conditions. Hydrology Research. 34, 477--492.

Sgro, N.A., 2016. Formal Hypothesis Testing for Prospective Hydrological Model Improvements.

Shafii, M., Tolson, B.A., 2015. Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. Water Resour.Res. 51, 3796-3814.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., et al., 2003. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. Hydrol.Sci.J./J.Sci.Hydrol. 48, 857-880.

Steinschneider, S., Brown, C., 2012. Dynamic reservoir management with real-option risk hedging as a robust adaptation to nonstationary climate. Water Resour.Res. 48.

Tolson, B.A., Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resour.Res. 43, n/a-n/a.

Varouchakis, E.A., Palogos, I., Karatzas, G.P., 2016. Application of Bayesian and cost benefit risk analysis in water resources management. Journal of Hydrology. 534, 390-396.

Walker, W.E., Harremoes, P., Rotmans, J., van der Slujis, J.P., van Asselt, M.B.A., Janssen, P., et al., 2003. Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. Integrated Assessment. 4, 5-17.

Welch, B.L., 1947. The Generalization of ` Student's' Problem when Several Different Population Variances are Involved. Biometrika. 34, 28-35.

Wood, E.F., Sivapalan, M., Beven, K., Band, L., 1988. Effects of spatial variability and scale with implications to hydrologic modelling. J.HYDROLOGY. 102.

Xu, C., 1999. Operational testing of a water balance model for predicting climate change impacts. Agric.For.Meteorol. 98, 295-304.

Xue, J., Gui, D., Zhao, Y., Lei, J., Zeng, F., Feng, X., et al., 2016. A decision- making framework to model environmental flow requirements in oasis areas using Bayesian networks. Journal of Hydrology. 540, 1209-1222.

Yaseen, Z.M., El-Shafie, A., Jaafar, O., Afan, H.A., Sayl, K.N., 2015. Artificial intelligence based models for stream-flow forecasting: 2000–2015. Journal of Hydrology. 530, 829-844.

Yu, S., He, L., Lu, H., 2016. A tempo- spatial- distributed multi- objective decision- making model for ecological restoration management of water- deficient rivers. Journal of Hydrology. 542, 860-874.

Appendix A

Reservoir Model Development - Supporting Materials

Calibration Procedure:

See section 3.2.6 for a discussion of the calibration procedure steps. Here the metric used in each step are presented.

1. Total Volume – calculated as:

   $$M_V = Q(|(\overrightarrow{V_{in}} - \overrightarrow{V_{out}})|; p = 0.9) \qquad (1)$$

   where M_V is the volume metric (m^3), $\overrightarrow{V_{in}}$ is a vector of cumulative inflow volumes for each water year in the simulation (m^3), $\overrightarrow{V_{out}}$ is a vector of cumulative outflow volumes for each water year in the simulation (m^3), Q is the quantile function, and p is the percentage of time that values in an empirical cumulative distribution function should be less than the returned value. The quantile function is used to reduce the sensitivity of the metric to the largest values.

2. Reservoir Properties – calibrate the $\mathrm{A_{ref}}$ to root-mean square error (RMSE), computed as:

   $$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (2)$$

   where y_i is the *i*th observed daily average reservoir stage (m), $\hat{y}_i$ is the *i*th simulated daily average reservoir stage (m), and n is the number of days. The RMSE of the reservoir stage is linearly inverse weighted (earlier errors weighted less than later errors) to counteract the implicit heavier weighting of earlier errors.

3. Winter Metric – calibrate the snow parameters, metric computed as

   $$M_W = Q(|(\overrightarrow{dV_{sim}} - \overrightarrow{dV_{obs}})|; p = 0.8) \qquad (3)$$

   where M_W is the winter metric, $\overrightarrow{dV_{sim}}$ is the vector of simulated daily reservoir volume changes (m^3/d), $\overrightarrow{dV_{obs}}$ is the vector of observed daily reservoir volume changes (m^3/d), Q is the quantile function, and p is the percentage of time that values in an empirical cumulative distribution function should be less than the returned value, set to 90% here.

4. Timing – for Steps (4 and 5), the final calibration sets are both calibrated using an average of the RMSE calculated for the reservoir stage (same as the metric from (2) above) and the RMSE calculated on the daily change in reservoir volume. The RMSE formula is the same as equation 7 above, and the daily change in reservoir volume is calculated as:

$$dV_i = V_i - V_{i-1} \tag{4}$$

where dV_i is the daily change in volume on the *i*th day of the simulation, V_i is the volume of the reservoir on the *i*th day, and V_{i-1} is the volume of the reservoir on the previous day of the simulation. Reservoir volumes are linearly interpolated from the provided rating curve using the simulated or observed stage value.

Table A1. Parameters calibrated in each step of calibration procedure

Raven Parameter Name	Description	Total Volume	Reservoir Properties	Winter	Timing 1	Timing 2
AirSnowCoeff	Air snow heat transfer coefficient [1/d]					
IrreducibleSnowSaturation	Maximum liquid water content of snow [pct SWE]			✓		
RainSnowTransition[0]	RainSnow minimum temperature [°C]			✓		
<NA>	RainSnow difference min and max temperature [°C]			✓		
POROSITY	Porosity [-]				✓	✓
HBV_BETA	Infiltration parameter in HBV method [-]				✓	✓
MAX_PERC_RATE	Maximum percolation rate [mm/d]				✓	✓
MAX_INTERFLOW_RATE	Maximum interflow rate [mm/d]				✓	✓
BASEFLOW_COEFF	Baseflow coefficient for rock [1/d]				✓	✓
BASEFLOW_N	Baseflow parameter for rock [-]				✓	✓
PET_CORRECTION	PET correction factor [-]	✓				
<NA>	Thickness of fractured rock layer [m]				✓	✓
<NA>	Thickness of top soil layer [m]					
MAX_HT	Maximum vegetation height [m]					✓
MAX_LAI	Maximum leaf area index [m^2/m^2]					✓
MAX_LEAF_COND	Maximum leaf conductance [mm/s]					✓
SAI_HT_RATIO	Ratio of stem area index to height [m2/m2/m]					✓
MAX_CAPACITY	Maximum canopy storage for rain [mm]	✓				
MAX_CAPACITY_SNOW	Maximum canopy storage for snow [mm]	✓		✓		
RAIN_ICEPT_PCT	Canopy interception percentage for rain [-]	✓				
SNOW_ICEPT_PCT	Canopy interception percentage for snow [-]	✓		✓		
SVF_EXTINCTION	Shortwave radiation extinction coefficient [-]	✓				
IMPERMEABLE_FRAC	Impermeable fraction of land [-]	✓				
FOREST_COVERAGE	Percentage of land covered with vegetation [-]					
DEP_MAX	Maximum depression storage [mm]	✓				
OW_PET_CORR_FOREST	Open water PET correction for forests [-]	✓				
OW_PET_CORR_LAKE	Open water PET correction for lakes [-]	✓				
FOREST_SPARSENESS	Forest sparseness factor [-]					
MELT_FACTOR	Melt factor for degree day method [mm/d/°C]			✓		
MIN_MELT_FACTOR	Minimum melt factor [mm/d/°C]			✓		
HBV_MELT_FOR_CORR	Forest melt factor correction [-]			✓		
REFREEZE_FACTOR	Refreeze factor for degree day method [mm/d/°C]			✓		
HBV_MELT_ASP_CORR	Aspect melt factor correction [-]			✓		
LAKE_PET_CORR	PET correction for Lakes [-]	✓	✓			
ABST_PERCENT	Percent runoff routed to depression storage [-]	✓				
<NA>	Reference height for reservoir [m]					
<NA>	Reference volume for reservoir [m^3]					
<NA>	Reference area for reservoir m^2]*		✓			
<NA>	Reservoir side slope [-]					

*Bark Lake Dam area in the Bark Lake model, and Mountain Chute Dam area in Mountain Chute model, respectively

Reservoir Parameters

Table A2. Base reservoir parameters for calibrated models

Parameter	Booth Lake	Galeairy Lake	Bark Lake	Mountain Chute
h_{ref} [m]	2.82E+02	3.88E+02	3.05E+02	2.44E+02
V_{ref} [m^3]	0.00E+00	2.00E+08	0.00E+00	0.00E+00
A_{ref} [m^2]	2.78E+07	1.01E+07	4.79E+07	4.47E+07
β [-]	6.86E+00	6.51E+00	5.76E+00	5.36E+00
h_0 [m]	3.89E+02	3.89E+02	-	-
a [m^2/d]	1.00E+00	7.00E-01	-	-
b [-]	1.00E+00	1.10E+00	-	-

Appendix B

DCT Case Study 1 - Supporting Materials

Table B1. Reservoir parameters for Kamaniskeg Dam

Parameter	Value
h_{ref} [m]	2.82E+02
V_{ref} [m^3]	0.00E+00
A_{ref} [m^2]	2.78E+07
β [-]	6.86E+00

Table B2. List of parameters used in calibration of evaluated model for reservoir management case study

Raven Parameter Name	Description	Base Value	Lower Bound	Upper Bound
RainSnowTransition[0]	RainSnow minimum temperature [°C]	-2.52	-8.00	1.00
<NA>	RainSnow difference min and max temperature [°C]	1.92	0.50	6.00
HBV_BETA	Infiltration parameter for rock in HBV method [-]	10.98	0.10	20.00
HBV_BETA	Infiltration parameter for soil in HBV method [-]	13.75	0.10	20.00
MAX_PERC_RATE	Maximum percolation rate for rock [mm/d]	57.26	0.10	100.00
MAX_PERC_RATE	Maximum percolation rate for soil [mm/d]	196.00	0.10	200.00
MAX_INTERFLOW_RATE	Maximum interflow rate for rock [mm/d]	10.73	0.05	50.00
MAX_INTERFLOW_RATE	Maximum interflow rate for soil [mm/d]	34.55	0.05	50.00
BASEFLOW_COEFF	Baseflow coefficient for rock [1/d]	0.23	0.01	1.00
BASEFLOW_COEFF	Baseflow coefficient for soil [1/d]	0.00	0.00	0.00
BASEFLOW_N	Baseflow parameter for rock [-]	0.61	0.50	5.00
BASEFLOW_N	Baseflow parameter for soil [-]	0.00	0.00	0.00
PET_CORRECTION	PET correction factor for rock [-]	0.71	0.10	1.30
PET_CORRECTION	PET correction factor for soil [-]	0.66	0.10	1.30
<NA>	Thickness of fractured rock layer [m]	294.72	0.10	300.00
SVF_EXTINCTION	Shortwave radiation extinction coefficient [-]	0.23	0.00	1.00
IMPERMEABLE_FRAC	Impermeable fraction of land [-]	0.00	0.00	1.00
FOREST_COVERAGE	Percentage of land covered with vegetation [-]	0.90	0.50	1.00
DEP_MAX	Maximum depression storage [mm]	229.00	0.00	500.00
MELT_FACTOR	Melt factor for degree day method [mm/d/°C]	2.77	0.50	10.00
MIN_MELT_FACTOR	Minimum melt factor [mm/d/°C]	2.73	0.50	10.00
HBV_MELT_FOR_CORR	Forest melt factor correction [-]	1.15	0.50	1.50
REFREEZE_FACTOR	Refreeze factor for degree day method [mm/d/°C]	4.55	0.50	10.00
HBV_MELT_ASP_CORR	Aspect melt factor correction [-]	0.53	0.50	1.50
LAKE_PET_CORR	PET correction for Lakes [-]	1.01	0.10	1.30
ABST_PERCENT	Percent runoff routed to depression storage [-]	0.30	0.00	0.50

Kamaniskeg Seasonal Reservoir Fitting

The seasonal rating curve at Kamaniskeg Dam was fitted using a piecewise linear regression. Only spring flows are fitted with a set of 3 linear regressions (with breakpoints in regression sets at stage values of 283.1 m and 283.3 m), the other seasons were fitted with a single linear regression. These regression values are shown in Table B3.

Table B3. Fitted piecewise linear regression coefficients for seasonal outflows from Kamaniskeg Dam

Parameter	Winter	Spring	Summer	Autumn
Intercept 1	-119285.86	-77978.94	-28465.61	-24170.32
Slope 1	421.87	275.91	100.74	85.57
Intercept 1	NA	-306895.19	NA	NA
Slope 1	NA	1084.65	NA	NA
Intercept 1	NA	-44645.84	NA	NA
Slope 1	NA	158.82	NA	NA

The plots of piecewise linear regression by season are shown in the following four figures.

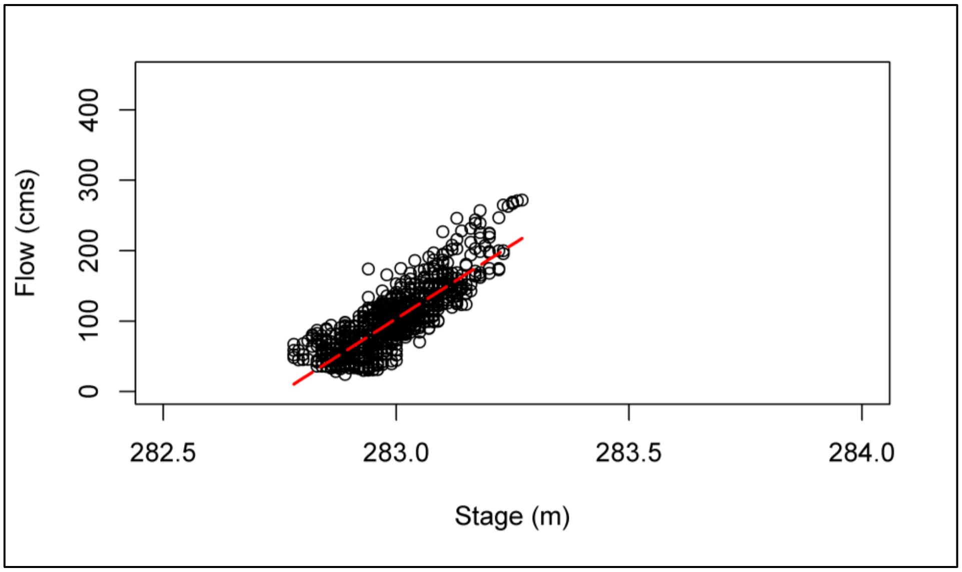

Figure B1. Linear regression fit for Winter outflows from Kamaniskeg Dam

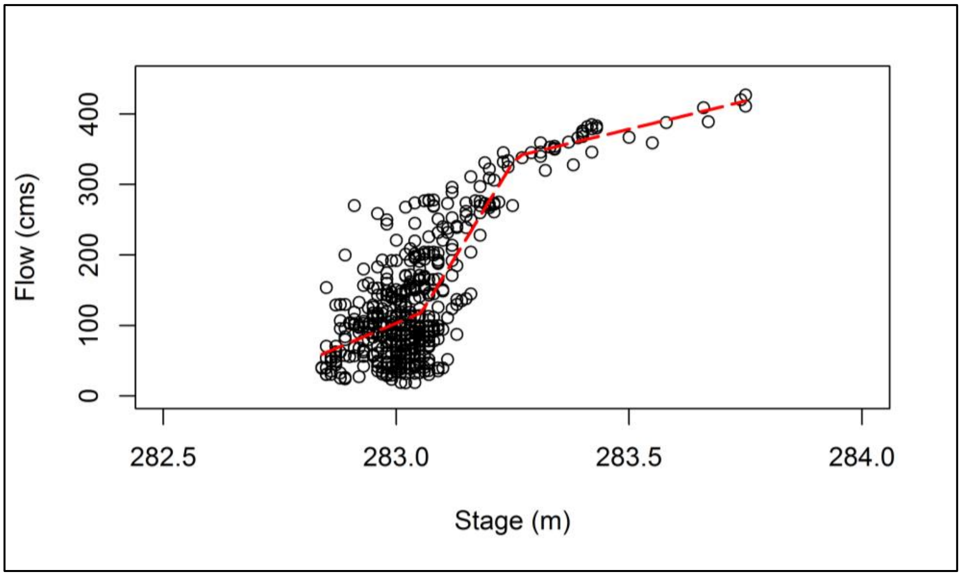

Figure B2. Linear regression fit for Spring outflows from Kamaniskeg Dam

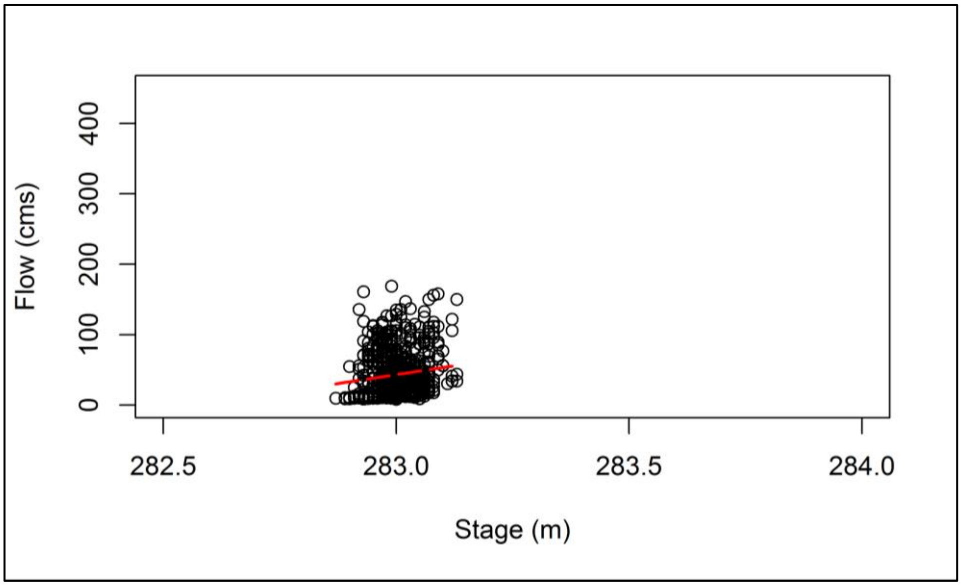

Figure B3. Linear regression fit for Summer outflows from Kamaniskeg Dam

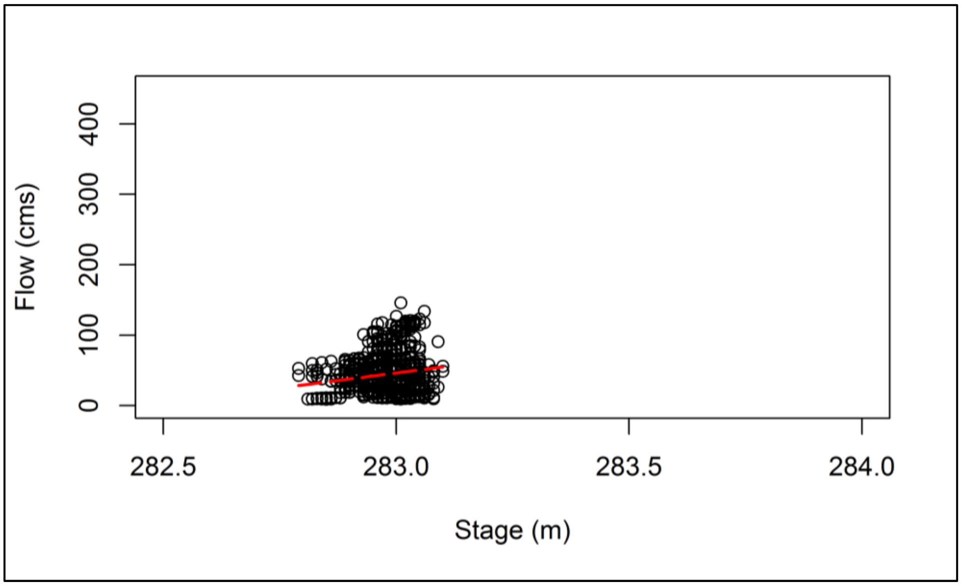

Figure B4. Linear regression fit for Autumn outflows from Kamaniskeg Dam

The seasonal rating curve, as provided to the Raven model for base operations, is shown in Table B4. The adjusted operation table simply divides the summer flows by a factor of 3.28.

Table B4. Seasonal rating curve for Kamaniskeg Dam

		Day of year	**274**	**91**	**152**	**244**	**335**
Stage [m]	**Volume [m3]**	**Area [m2]**	$Q_{default}$	Q_{spring}	Q_{summer}	Q_{autum}n	Q_{winter}
282.00	0	27800000	0.00	0.00	0.00	0.00	0.00
282.25	6954856	27838854	0.00	0.00	0.00	0.00	0.00
282.50	13919429	27877735	0.00	0.00	0.00	0.00	0.00
282.75	20893726	27916643	25.08	34.40	18.17	25.08	0.00
283.00	27877753	27955579	46.48	103.37	43.36	46.48	103.67
283.25	34871518	27994541	67.87	331.01	68.54	67.87	209.14
283.50	41875026	28033531	89.26	378.58	93.73	89.26	314.61
283.75	48888285	28072548	110.66	418.28	118.91	110.66	420.07
284.00	55911302	28111591	132.05	457.99	144.10	132.05	525.54
284.25	62944083	28150662	153.44	497.69	169.28	153.44	631.01
284.50	69986635	28189761	174.84	537.40	194.46	174.84	736.48
284.75	77038966	28228886	196.23	577.10	219.65	196.23	841.94
285.00	84101081	28268038	217.62	616.80	244.83	217.62	947.41
285.25	91172987	28307218	239.01	656.51	270.02	239.01	1052.88
285.50	98254692	28346425	260.41	696.21	295.20	260.41	1158.35
285.75	105346202	28385658	281.80	735.92	320.39	281.80	1263.82
286.00	112447523	28424919	303.19	775.62	345.57	303.19	1369.28

Appendix C

DCT Case Study 2 - Supporting Materials

Table C1. List of parameters used in calibration of evaluated model for data gauging case study

Raven Parameter Name	Description	Base Value	Lower Bound	Upper Bound
RainSnowTransition[0]	RainSnow minimum temperature [°C]	-7.41	-10.00	2.00
<NA>	RainSnow difference min and max temperature [°C]	4.79	0.50	10.00
POROSITY	Soil porosity [-]	0.68	0.10	0.80
HBV_BETA	Infiltration parameter in HBV method [-]	5.02	0.10	20.00
MAX_PERC_RATE	Maximum percolation rate [mm/d]	49.99	0.10	50.00
MAX_INTERFLOW_RATE	Maximum interflow rate [mm/d]	12.98	0.05	50.00
BASEFLOW_COEFF	Baseflow coefficient [1/d]	0.23	0.01	10.00
BASEFLOW_N	Baseflow parameter [-]	0.72	0.50	4.00
PET_CORRECTION	PET correction factor [-]	0.13	0.10	1.20
<NA>	Thickness of fractured rock layer [m]	106.44	0.10	200.00
MAX_CAPACITY	Maximum canopy storage capacity [mm]	18.90	0.50	30.00
MAX_SNOW_CAPACITY	Maximum canopy storage capacity for snow [mm SWE]	9.87	0.50	15.00
RAIN_ICEPT_FACT	Rain interception fraction [-]	0.06	0.01	0.10
SNOW_ICEPT_FACT	Snow interception fraction [-]	0.07	0.01	0.10
SVF_EXTINCTION	Shortwave radiation extinction coefficient [-]	0.64	0.00	1.00
IMPERMEABLE_FRAC	Impermeable fraction of land [-]	0.07	0.00	0.10
DEP_MAX	Maximum depression storage [mm]	127.00	1.00	300.00
OW_PET_CORR	Open water correction factor [-]	0.53	0.25	1.00
MELT_FACTOR	Melt factor for degree day method [mm/d/°C]	1.90	1.00	15.00
MIN_MELT_FACTOR	Minimum melt factor [mm/d/°C]	1.92	0.50	4.00
HBV_MELT_FOR_CORR	Forest melt factor correction [-]	0.96	0.50	1.50
REFREEZE_FACTOR	Refreeze factor for degree day method [mm/d/°C]	2.05	2.00	5.00
HBV_MELT_ASP_CORR	Aspect melt factor correction [-]	1.08	0.50	1.50
LAKE_PET_CORR	PET correction for Lakes [-]	0.50	0.50	1.30
ABST_PERCENT	Percent runoff routed to depression storage [-]	0.46	0.00	0.50
<NA>	Booth Lake outflow power law parameter a [m^2/d]	1.00	0.00	20.00
<NA>	Booth Lake outflow power law parameter b [-]	1.00	0.00	10.00
<NA>	Galeairy Lake outflow power law parameter a [m^2/d]	0.70	0.00	20.00
<NA>	Galeairy Lake outflow power law parameter b [-]	1.10	0.00	10.00