Formal Hypothesis Testing for Prospective Hydrological Model Improvements

by

Nicholas Sgro

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

Waterloo, Ontario, Canada, 2016

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

New algorithms for simulating hydrological processes are regularly proposed in the hydrological literature. However, the tests used to evaluate the effectiveness of these algorithms are typically no more than history matching – an improved model hydrograph is (often inappropriately) interpreted as an improved model. These tests ignore the considerable uncertainty inherent to hydrological models which may obscure the results of any comparisons.

In this work, a simple and more stringent method is proposed for comparing two model algorithms in terms of their ability to provide distinguishably different validation results under the impact of uncertainty in observation data and forcings. by generating distributions of performance indicators (e.g. Nash Sutcliffe) which can then be compared using basic statistical methods. The results show that at times modelling decisions are indistinguishable even when a single performance indicator shows improvement. As may be expected, our ability to identify the preferred hydrologic algorithm is significantly diminished when increased model/data uncertainty is incorporated into the evaluation process. This suggests that more robust testing is needed than what is typically reported in literature proposing model enhancements.

The thesis goes on to provide an example of how the new model comparison procedure can be include multiple performance measure and hydrological signatures to provide a diagnostic evaluation of modelling decisions. Diagnostics approaches to model evaluation can be found in the literature, but this work shows that data uncertainty affects each signature differently, and often introduces significant variability in model performance. Finally, it is demonstrated how the testing procedure can be used to examine the interactions between modelling decisions for better understanding of model deficiencies and the underlying hydrology.

Acknowledgements

I would like to thank my supervisor, Dr. James Craig for his help and guidance during my time at the University of Waterloo.

I am also grateful to BC Hydro for the funding, access to data, and providing me with a space to work during my summers in Vancouver. In particular, I would like to thank Dr. Georg Jost for his insights while I was at BC Hydro.

This research would not have been possible without funding from the Natural Sciences and Engineering Research Council (NSERC).

Finally, I would like to thank my family for their support and encouragement throughout my studies, and life in general.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Hydrological modelling is an important tool for scientists, engineers and decision makers in a variety of areas, including: hydroelectric power, water supply, ecosystem sustainability, agriculture, and flood protection. Hydrologists also rely on models to help enhance their understanding of hydrological systems, diagnose problems with our current understanding of the water cycle, and to direct future experimental work and data requirements. The variety and importance of the uses of hydrological models means that improving models should be of interest not only to hydrologists, but to the general public as well.

Despite the importance of hydrological modelling, the best way to represent the water cycle at the watershed scale is still an open question. This is due in large part to the complexity of the hydrological cycle, which is made up of many sub-processes (e.g., infiltration, evapotranspiration, snowmelt), each of which are difficult to represent at the catchment scale. Heterogeneity and unknown variables make each process nearly impossible to model without many simplifications. This has led to many different representations of each process as model developers choose different methods for simplifying each process. The ambiguity in how to best represent hydrological processes has resulted in a vast number of different models developed for different objectives and involving different conceptualisations, making it difficult to decide which model or process submodel to use when beginning a new project.

Developing a hydrological model can problematic due to the large number of decisions that must be made, such as, how to spatially and temporally discretize the watershed, which processes to include in the model, and which algorithms to use for the included processes. Making these decisions is typically done in an ad hoc fashion, relying on the judgment of the model developer. The impact of individual modelling decision is rarely examined critically. One reason for this lack of critical examination is that rigorously testing alternative model configurations is difficult, owing to the complexity of the models and the limited data available.

The issue of comparing modelling decisions is even more difficult when you consider the large amount of uncertainty involved in the observed data, the model parameterization, and the hydrological model structure. The uncertainty inherent in hydrological modelling means that the results of any model comparisons are also uncertain, meaning that the true results may be obscured when comparing modelling decisions. It has been recognized in the literature that uncertainty inherent to hydrological modelling results in many different models and parameter sets performing nearly equally well. This issue, termed 'equifinality' by (K. Beven, 1993), is a widely accepted problem in hydrology and other fields of environmental modelling. While the problem of equifinality cannot be completely overcome given our current data and understanding of hydrology, it is hoped that the work in this thesis will provide a more powerful method for distinguishing between model performances that appear to be equifinal using traditional methods, and reduce the number of models that are deemed equifinal. In order to improve water management decisions and reduce the risk of hydrological hazards, it is important that we begin making hydrological modelling decisions in a more systematic and scientifically sound manner.

1.1 Purpose

The purpose of this thesis is to develop, test and apply a methodology to compare individual modelling decisions using statistical hypothesis testing procedures. It is hoped that this contribution will help provide a deeper understanding of the cause of differences between different hydrological models, help to ascertain whether proposed improvements are true improvements, and help in the development of future models by focusing attention on the decisions with the greatest impact.

The evaluation of modelling decisions will be performed by isolating individual modelling decisions using a flexible modelling framework, and then generating distributions of fitness metrics for each option. These distributions will be generated by considering sources of uncertainty affecting the calibration of the model in order to generate a collection of equally plausible results. Each result will then be evaluated, and statistical testing used to check for significant differences between the distributions.

1.1.1 Objectives

1. Develop and apply a procedure for assessing proposed model changes which is more rigorous than the ad hoc testing used in the current scientific literature.
2. Examine the impact of considering increasing levels of uncertainty on the ability to effectively distinguish between modelling decisions.
3. Demonstrate how the new testing procedure can be used in conjunction with a diagnostic model evaluation for increased understanding of model differences.
4. Apply the newly developed procedure to test multiple modelling decisions simultaneously to show how modelling decisions interactions can be examined.

1.2 Thesis Organization

This thesis is organized into six chapters.

Chapter 2 provides background information and a literature review of the topics related to this thesis. An overview of hydrological model development process, a review of hypothesis testing in hydrological modelling, and a summary of the sources of uncertainty found in hydrological models are all provided.

Chapter 3 proves details of the methods used in the thesis, including a description of the basin used for the case study, details of some of the uncertainty found in the hydrological observations, a description of the base model and the potential model changes being tested, and an overview of the statistical methods used.

Chapter 4 develops the testing procedure used to rigorously test modelling decisions, shows how uncertainty can be applied in different ways with varying degrees of effectiveness and computational requirements, and examines how the increasing levels of uncertainty affect the ability to distinguish between modelling decisions.

Chapter 5 provides an example of how the new model comparison procedure can be used in conjunction with multiple performance measure to provide a diagnostic evaluation of modelling decisions. The chapter goes on to demonstrate how the diagnostic approach can be used to examine the interactions between modelling decisions for better understanding of model deficiencies and the underlying hydrology.

Chapter 6 summarizes the major conclusion drawn from this study, and offers recommendations of areas requiring future research.

Chapter 2
Background

2.1 Hydrological Modelling Procedure

The procedure for developing a hydrological model can be broken down into five steps, as shown in Figure 1.

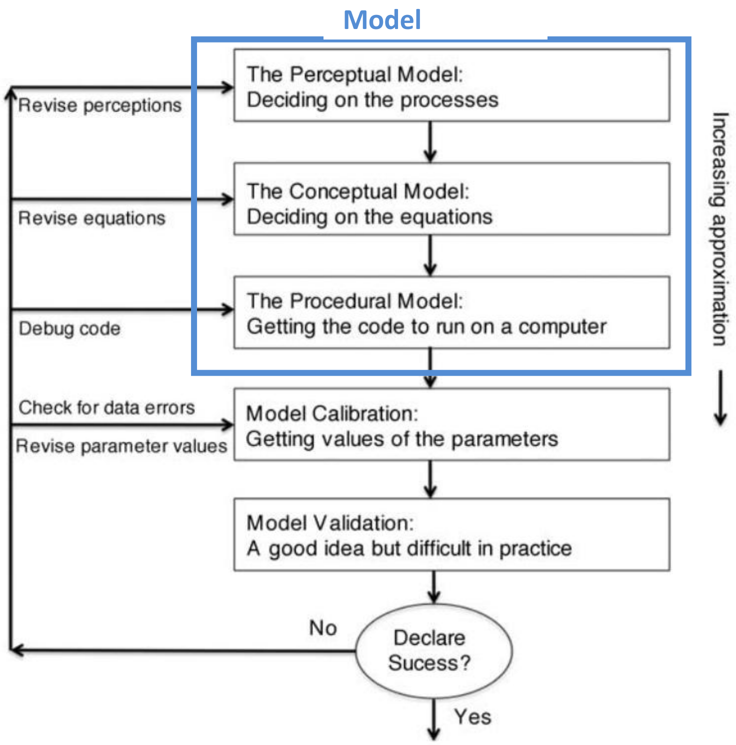

Figure 1: A schematic outline of the steps in the modelling procedure (Modified from K. J. Beven, 2011)

2.1.1 Model Assembly

The first three steps in the modelling procedure depicted in Figure 1 are together referred to as the model structure throughout this thesis. This includes everything from the theoretical understanding of the hydrology to the specific implementation in computer code.

The first step in the modelling procedure is to develop a perceptual model of the catchment, i.e. a qualitative idea of how the hydrology of the catchment functions and which processes are dominant. Hydrology involves many different processes occurring at different scales and interacting in complex ways, so the identification of the perceptual model can be difficult. The perceptual model is subjective, based on the understanding and experience of the modeller, so there can be many different perceptual models of the same site. The perceptual model is an important step because any mistake here will propagate through to the other steps and could result in large errors in the final model.

The next step in the modelling procedure is the development of a conceptual model. The conceptual model is the expression of the perceptual model as mathematical equations. This step often requires assumptions and simplifications to be made in order to convert the perceptual model into something that can be described mathematically. The simplification process is again subjective depending on many factors such as the scale of the model, data availability and the relative importance of each process. For example, a plot scale model with good soil data could use the Richards equation to model infiltration in detail, but a catchment scale model would likely not be able to use the more complex equation, and might instead use an empirical formula that is more feasible at the larger scale.

The third step, development of a procedural model, is the conversion of the conceptual model into computer code. Often in practice, an existing model code is selected following the first two steps, however if the conceptual model is not well represented by any existing code, a new procedural model may be developed (Butts, Payne, Kristensen, & Madsen, 2004). Depending on the set of differential equations that form the conceptual model, the equations may be solved analytically, or more often through the use of numerical approximations. The choice of numerical approximations (e.g., operator-splitting and implicit Euler schemes) can have a significant effect on model results (M. P. Clark & Kavetski, 2010).

2.1.2 Model Calibration

Parameters are variables that define the characteristics of a catchment or flow domain (K. J. Beven, 2011). Some parameters may be measurable (e.g., watershed area), others are measurable in theory but are impractical to measure at the catchment scale (e.g., hydraulic conductivity), and some parameters are purely conceptual and represent abstract characteristics that cannot be measured (e.g., maximum soil storage). The fourth step in the modelling procedure is model calibration, where parameters are adjusted so that the model closely matches the behaviour (H. V. Gupta, Sorooshian, & Yapo, 1998). Calibration can be either manual or automatic, though in practice it often is a combination of both.

Manual calibration involves the modeller adjusting the parameters by hand based on the understanding of the modeller. Parameters are adjusted by a trial and error approach where the evaluation of model performance at this step is typically a visual inspection of the modelled and observed hydrographs, but could also include other performance statistics. Manual calibration has the advantage of the modeller's expertise in guiding the calibration so that the physical realism of the model parameters can be maintained. The problem with manual calibration is that it can be very labour intensive for even a small number of parameters, and the results are based on the skill of the modeller (Boyle, Gupta, & Sorooshian, 2000). When the number of parameters is large, it is often more appropriate to use automatic calibration methods.

Automatic calibration involves the use of computer-based methods to fit a model to some period of observed data. Automatic calibration algorithms use various methods to search for optimal parameters much more efficiently than what is possible by manual calibration (Madsen, Wilson, & Ammentorp, 2002). The evaluation of model performance is based on statistical measures of fit, so the process is entirely objective, but care must be taken to ensure parameters maintain physical realism. Two major components of automatic calibration are the algorithm that is used to search the parameter space, and the objective function used to evaluate the quality of a parameter set.

Much research has been devoted to the development of automatic calibration algorithms, and there are many different options available. An extensive comparison of ten calibration methods was done by (Arsenault, Poulin, Côté, & Brissette, 2013). The study showed that the Dynamically Dimensioned Search (DDS) algorithm (Tolson & Shoemaker, 2007) used in this thesis was among the best, especially when calibrating larger numbers of parameters.

Objective functions are a numerical measure of the difference between modelled and observed values. Objectives functions reduce model fit to a single number that can be minimized during calibration. There are many different objective functions that are used regularly for hydrological modelling; (Moriasi et al., 2007) provides a review of many common objective functions. Typically a single objective function is used during calibration, but this leads to a bias towards whatever aspect the objective function measures (e.g., the commonly used Root Mean Square Error is biased towards high flows due to squaring the data). Recently multi-objective calibration has been increasingly used in order to balance the different aspects of the fit. A review of multi-objective calibration is provided in (Efstratiadis & Koutsoyiannis, 2010).

2.1.3 Model Validation

Model validation is the process of evaluating the calibrated model performance with independent data not used for calibration. The idea is that a model should be tested under conditions similar to how it will be used operationally, so in the case where the model will be used for predictions, the validation must use data from a time period outside the calibration period. Klemes (1986) proposed four validation tests of increasing difficulty.

The first test is the split-sample test, where the available record is split into a two separate periods used for calibration and validation. This is the basic form of validation which has become the standard approach used by hydrologist (Andréassian et al., 2009). The split-sample test is intended to test for overparameterization and overfitting, where the model is made to fit certain peculiarities or noise in the calibration period rather than the underlying processes being modelled. This test has been made more rigorous by using multiple sub-periods, each used for both calibration and validation. This more advanced test has been used for model comparisons in validation (e.g., C. Perrin, Michel, & Andréassian, 2001) and in calibration to find parameter sets that perform consistently through time (e.g., Gharari, Hrachowitz, Fenicia, & Savenije, 2013).

The other tests proposed by Klemes are designed to test specifically for model transposability in time and space by splitting the observations into periods with distinct hydrological conditions, and by using different basins for calibration and validation. These tests are intended for use when the model will be used to simulate flows under different conditions than what is available in the observations (such as with land use or climate change studies), and for simulating in an ungauged basin. Neither of these conditions are of interest in this thesis, so these tests are not used.

2.2 Classification of Models

There are many different ways that hydrological models can be classified, and the distinctions are not always clear; for example, models can occasionally be a hybrid of two different categories. Nevertheless, there are some overlapping characteristics that can be used to group models and provide a useful framework for comparing models and discussing their relative strengths and weaknesses. Jajarmizadeh, Harun, & Salarpour (2012) provide a good review of different classifications from the literature. The classifications used for this thesis are defined below and represent some of the most common ways of grouping models.

2.2.1 Empirical vs. Conceptual vs. Physically-Based Models

Hydrological models can be classified based on their representation of the hydrological processes.

Empirical models are black box models which describe the relationship between input and output data, without necessarily reflecting causation. These models ignore the underlying processes and instead represent the catchment with transfer functions to convert input data to an output prediction. Models of this type can range in complexity from simple linear regression or unit hydrographs, to much more complex artificial neural networks (Dawson, Abrahart, Shamseldin, & Wilby, 2006) and data-based mechanistic models (Young, 2003). Since these models have no physical meaning, they are only applicable at the location and under the conditions used to train the model.

Physically-based models divide the hydrological system into a group of component processes, with each process being represented by a governing equation based on physical laws, such as the conservation of mass, energy and momentum. Since these models are based on real physical laws, model parameters are theoretically all measurable and shouldn't require any calibration (K. J. Beven, 2011). The biggest problem with physically-based models is that while the parameters are measurable at a point scale, the high degree of heterogeneity makes it unfeasible to fully characterize the parameters at the catchment scale. Estimating parameters at the catchment scale requires the use of effective parameters, which are upscaled from local measurements. The use of upscaling assumes that the small scale physical laws apply to the larger scale model, so any emergent processes may be missed (Blöschl & Sivapalan, 1995). For example, Darcy's law can be used to describe groundwater flow, and hydraulic conductivity may be measured, but at a larger scale there could be preferential flow paths that often dominate the river response to groundwater. These effective parameters can be estimated based on characteristics such as soil type or land use, or they can be calibrated, but in either case a degree of uncertainty is introduced.

Conceptual models are similar to physically-based models in that they also use component processes to describe the hydrological system, however they rely on a simpler, idealized version of the processes. Conceptual models typically represent the catchment with a schematic of storages, with the fluxes between the storages based on relatively simple mathematical equations (Shaw, Beven, Chappell, & Lamb, 2010). The parameters do not always have a true physical meaning, so they cannot be measured directly and must be calibrated to observed data. The complexity of conceptual models can vary greatly, ranging from a few simple storages (e.g., The Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Bergström & Singh, 1995)) to complex systems with many storages and fluxes (e.g., the University of British Columbia Watershed Model (UBCWM) (Quick & Singh, 1995)). The appropriate level of complexity should be a balance between matching the complexity of the hydrology, and what is supported by the data (Wheater, 2002).

2.2.2 Lumped vs. Distributed vs. Semi-Distributed

Based on their spatial representation of the catchment, hydrological models can be classified into three groups: lumped, semi-distributed and distributed. Examples of possible spatial discretizations are shown in Figure 2.

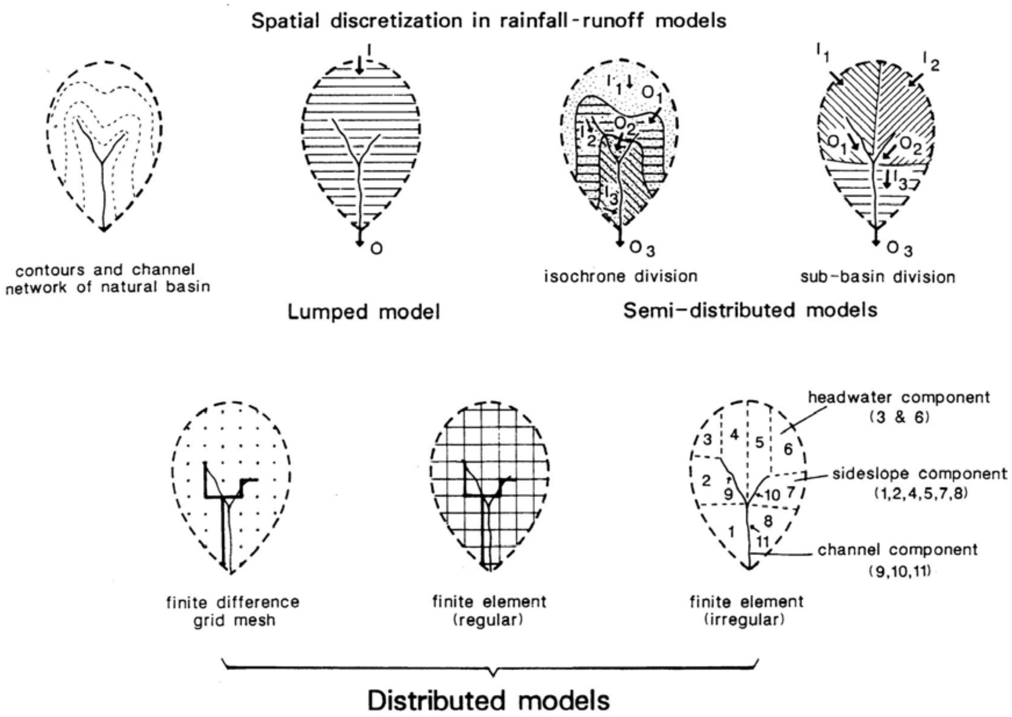
Spatial discretization in rainfall-runoff models
I
O
I1
I2
O1
O2
I3
O3
I1
I2
O1
O2
I3
O3
contours and channel network of natural basin
isochrone division
sub-basin division
Lumped model
Semi-distributed models
headwater component (3 & 6)
sideslope component (1,2,4,5,7,8)
channel component (9,10,11)
1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

Figure 2: Graphical representation of different spatial discretization (from Jones, 2014)

Lumped models treat the entire catchment as a single homogeneous unit where state variables represent averages across the entire catchment (K. J. Beven, 2011). These models typically have the advantage of fast computation time, low data requirements, and a smaller number of parameters, since the spatial distribution of forcings and parameters are not considered.

Distributed models divide the catchment into a large number of elements or grid cells. The equations are solved in each cell for the state variable representing the local average (Xu, 2002). This approach allows for the explicit representation of the heterogeneity of the catchment since forcings, parameters, and even the equations can be varied in each cell. While in theory distributed models are more realistic, they come at the cost of a much longer computation time, increased data requirements, and the large number of parameters lead to greater issues of equifinality (K. Beven, 2006).

Semi-distributed models are the middle ground between lumped and distributed models. In a semi-distributed model the catchment is divided into sub-units each of which are each assumed to be homogeneous. The sub-units in a semi-distributed model could be based on sub-basins, elevation bands, land use, or any other variable that could cause a difference in the hydrological response of the sub-units. The advantage of semi-distributed models is that they can incorporate the biggest sources of spatial heterogeneity while limiting the data requirements and number of parameters to a more manageable level than fully distributed models.

2.2.3 Stochastic vs. Deterministic Models

Hydrological models can be grouped into two broad categories, deterministic and stochastic (Shaw et al., 2010). Deterministic models always provide the same output for a given set of inputs and parameter values. Stochastic models attempt to capture the inherent randomness of hydrological systems by including some elements of uncertainty in the model inputs, boundary conditions and/or parameters. Generally, a model can be considered stochastic if the output includes some measure of predictive uncertainty, whereas a deterministic model provides a single value at each timestep (K. J. Beven, 2011).

The majority of models used in hydrology are deterministic in nature, although the lines are becoming blurred as more and more deterministic models are being driven by stochastic inputs to produce some uncertainty bounds (See section 2.6).

2.3 Models as system of Hypotheses

Both conceptual and physically-based hydrological models can be thought of as a collection of submodels where the output from one submodel is the input for the next submodel. This means that the model inputs (e.g., precipitation) get filtered through a variety of submodels before reaching an output suitable for evaluation purposes, typically in the form of a hydrograph. The fact that the submodels interact in complex nonlinear ways makes it difficult to determine whether a change to a given submodel actually increases performance, or if it simply gives a different output without any fundamental improvement of predictive ability.

Using the idea of a model as a collection of coupled hypotheses proposed by Clark, Kavetski, & Fenicia (2011), each submodel can be considered a separate hypothesis of process functions. For instance, the degree-day method and the more rigorous energy balance approach are two different methods of representing the snowmelt process, and may be considered distinct hypotheses. Decisions about the dominant processes, how the processes are represented mathematically, how the processes are connected, the selection of temporal and spatial discretization, and how to distribute the model forcings, are each considered to be separate hypotheses that can be subject to testing. In this thesis we are looking to compare competing hypotheses (in the form of different means of algorithmically representing distinct processes) to determine whether they are distinguishable given the available data and which is the most likely to improve performance.

2.4 Hypothesis testing

Typically during the development of hydrological models, decisions are made in an ad hoc fashion where a single option is chosen based on the modeller's understanding of the catchment or simply because the choices are hard coded in the selected model. In a way, this can be considered the most basic form of hypothesis testing, where options are considered and evaluated based on the past experience and expert judgment of the modeller. Although alternatives may be considered during the model development process, they are rarely reported in the literature and are likely not tested rigorously, though some recent exceptions are highlighted throughout this section.

Recently, hydrologists have recognized that the usual informal consideration of model hypotheses is insufficient (e.g., McMillan et al., 2011; Coxon et al., 2014; Sivakumar, 2008; M. P. Clark et al., 2011; Vache & McDonnell, 2006). Hydrologists have taken an interest in evaluating differing model structures in a more rigorous and scientifically defensible way, and there have been several approaches to handling the problem. The rest of this section will examine some of the current research into hypothesis testing of hydrological model structure.

2.4.1 Model Intercomparison Projects

Model intercomparison projects (e.g., Duan et al., 2006; Reed et al., 2004) are an example of organized hypothesis testing in hydrology. The idea behind these experiments is to have teams of hydrologist use many different models on large sets of catchments in order to look for trends in the results. They provide a good comparison of the performance of many models on a wide range of catchments, however there are some flaws in these experiments.

One problem with model intercomparison projects is that each of the models are typically run by different modellers. This means that model differences are confounded by multiple other factors such as calibration strategy, and the time commitment and skill of the individual modellers. As discussed by Clarke (2008), traditional hypothesis testing theory would eliminate operator bias by having each modeller use multiple models, but the time commitment needed to have researchers set up, calibrate and potentially learn several different models is prohibitive.

The biggest problem with current model intercomparison projects is that they are comparing entire models that have many structural differences, so even if a difference in performance can be identified it is impossible to attribute any improvement to a specific change (M. P. Clark et al., 2011). This means that results from these experiments cannot be used to assess any individual model component or to diagnose issues found with a model. The purpose of hypothesis testing should be to inform the development of better models, which is something that current model intercomparisons fail to do. By identifying entire models that are significantly better than the alternatives, model intercomparison projects could be used as the basis for more detailed studies investigating which components are most responsible for the difference.

2.4.2 Flexible Models

In order to be able to compare and test individual model components, the use of a flexible model architecture is required. These flexible models enable the user to vary the model structure within a single tool, allowing for easier implementation of multiple different hypotheses of model structure, as well as the ability to isolate changes to a single model component while keeping the rest of the model fixed. There have been several such architectures presented in the literature.

Examples of modelling frameworks include: the Framework for Understanding Structural Errors (FUSE) (M. P. Clark et al., 2008), which allows the user to select between different representations for soil stores based on four parent models; SuperFLEX (Fenicia, Kavetski, & Savenije, 2011), designed to allow for varying the model architecture as well as the individual process implementations by using combinations of generic components such as reservoirs and transfer functions; the Cold Regions Hydrological Model (CRHM) (Pomeroy et al., 2007), a modular modelling framework designed for use in cold regions; and the Structure for Unifying Multiple Modeling Alternatives (SUMMA; (M. P. Clark et al., 2015), a framework allowing for the use of different spatial representations, flux parameterizations, parameter values, and timestepping schemes.

The framework selected for use in this thesis was the Raven Hydrological Modelling Framework (J. R. Craig et al., 2016). Raven is a modular hydrological modelling framework capable of using different model structures, process algorithms, discretizations and numerical schemes. Raven has a wide variety of process algorithms available, ranging from simple conceptual algorithms to more physically based equations, and is able to adjust how these processes are interconnected. One of the best features of Raven is its ability to emulate several existing hydrological models, including the UBC watershed model used as the base model in this thesis.

2.5 Sources of uncertainty

Uncertainty is an important topic in hydrology and has been the subject of extensive research. There are many sources of uncertainty, but they can generally be grouped into four categories: input uncertainty, output uncertainty, parameter uncertainty and structural/model uncertainty (Renard, Kavetski, Kuczera, Thyer, & Franks, 2010) .

2.5.1 Input Uncertainty

Input uncertainty includes uncertainty in the inputs to a model, such as precipitation, temperature, and initial conditions.

The main causes of uncertainty in precipitation and temperature data are measurement errors and interpolation from a point scale to catchment scale observations (Tetzlaff & Uhlenbrook, 2005). Measurement errors are caused by inaccuracies in the instrumentation used to collect measurements. Measurement errors can also be caused by the placement of the gauges. For example, a rain gauge placed too near to trees could be affected by interception, or a temperature gauge placed near a body of water may read more moderate temperatures. Typically measurement errors are much smaller than errors from interpolation.

Interpolation uncertainty stems from the fact that model forcings are measured at only a handful of locations throughout the catchment, and may be representative of only a very small spatial extent (Singh, 1997). The spatial variability of precipitation especially is known to be quite large, so without a very dense network of gauges it is inevitable that the areal average precipitation will be subject to significant uncertainty (Tetzlaff & Uhlenbrook, 2005). In mountainous terrain, the impact of changing elevation on rainfall and temperature makes forcings even more variable, while at the same time the gauges are more sparse due to the difficulty of access. The spatial variability also increases with a higher temporal resolution. While the spatial differences in forcings may average out over the course of a day, the hourly precipitation and temperature used in this work will be more variable.

Uncertainty in initial conditions is caused by the fact that it is difficult or even impossible to measure many state variables used in hydrological models. The effects of initial conditions can be mitigated by using a spin-up period (Seck, Welty, & Maxwell, 2015), also known as a warm-up period. A spin-up period is the period for which a model is run and its performance is not evaluated, allowing the model to reach a state that is much less dependent on the initial conditions. Using a spin-up period means that the uncertainty of initial conditions is negligible, and thus is not considered in this work.

2.5.2 Output Uncertainty

Output uncertainty is the uncertainty found in the observations of the values used to evaluate a model, most often the river discharge. Discharge uncertainty can be caused by many different factors depending on the method used for measurement. The most common method of measuring discharge in practice is the rating curve method (H. McMillan, Krueger, & Freer, 2012). In this method, discharge is estimated based on the measured river stage through the use of a rating curve. The rating curve is produced by measuring the discharge and stage at multiple time points, and then interpolating between these points. This method is subject to uncertainty in the original measurements, interpolation error, changes to river flow regime since the original measurements (e.g., roughness changes or unsteady flow conditions), and perhaps most importantly, error in extrapolation (Baldassarre & Montanari, 2009). Since measurements are not typically taken during the most extreme river flows, the rating curves are extrapolated beyond real measurements, and thus the highest flow values have an increased uncertainty.

In the case study used for this work, the discharge is not measured using the rating curve method. In fact, the discharge is not measured at all; it is instead calculated using a water balance of a hydroelectric reservoir, based on measured outflow through the dam and the change in reservoir storage. This method is known to produce hydrographs with more uncertainty than observed at natural river channels (BC Hydro, 2009). The storage values are also not measured directly, but are calculated based on a stage-storage curve. The sources of uncertainty include errors in elevation readings, errors in outflow measurements and errors in the stage-storage curve. More information about the reservoir inflow error is provided in Section 3.2.

2.5.3 Structural Uncertainty

The complexity of hydrological systems requires that many simplifications and assumptions be made when developing a model. The imperfect representation of the true system is what causes structural uncertainty in a model (Refsgaard, Van der Sluijs, Jeroen P, Brown, & Van der Keur, 2006). Structural uncertainty is the most poorly understood and difficult to quantify, but incorrect or missing process representations can lead to substantial errors (Zhang, Hoermann, Gao, & Fohrer, 2011).

All models represent highly heterogeneous systems by aggregating to scales reasonable for modelling. While the system response can often be represented reasonably well despite the aggregation, the impact of spatial variability does introduce some degree of uncertainty to the model (Hublart, Ruelland, Dezetter, & Jourde, 2015). This problem is most significant with higher degrees of aggregation, such as in lumped models.

2.5.4 Parameter Uncertainty

Parameters in hydrological models are very rarely measured for any practical applications. Parameters are often spatially and temporally heterogeneous and cannot be measured at the catchment scale (K. Beven, 1993). Many other parameters are purely conceptual instead of an exact physical representation of a process. Because of the impossibility of measuring most model parameters, parameters are typically estimated through calibration.

Parameter uncertainty is the result of insufficient calibration data. Calibration data is limited to a portion of the recorded observations, and even for long records there may not be sufficient information to fully characterize all the parameters, especially for models with a large number of parameters. It has been shown that rainfall-runoff data supports the inference of only a few parameters without significant uncertainty (Jakeman & Hornberger, 1993). Despite this limitation, many models have significantly more parameters than what can be fully identified, and some distributed models can have more than 100 parameters.

Uncertainty in parameter values is also caused by the propagation of other errors through the model. Parameter definitions are obviously dependent on the model structure and so are influenced in part by structural errors. Since the calibration process is run using observed forcings, and compares model output to observations, input and output uncertainty also add some uncertainty to the parameter estimates.

2.6 Uncertainty Assessment Methods

The usual method of considering uncertainty in hydrological models is through some form of uncertainty assessment. Uncertainty assessment attempts to quantify the uncertainty in the model output by estimating the probability distribution or confidence limits on the output. The methods for estimating the output uncertainty can vary in complexity from relatively simple Monte Carlo methods, such as the Generalised Likelihood Uncertainty Estimation (GLUE) methodology (K. BEVEN & BINLEY, 1992), to more complex formal Bayesian methods, such as the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt et al., 2009).

The difference between the uncertainty estimation methods and the work done in this thesis is that they are trying to assess the amount of uncertainty in the modelled hydrograph, whereas this work is examining the uncertainty in performance of deterministic models. The uncertainty assessment methodologies typically focus on parameter uncertainty, since this often has the largest impact on the modelled hydrograph; however here parameter uncertainty is not included explicitly, and only exists as a byproduct of imperfect calibration.

While uncertainty assessment could be used to support model intercomparisons, the objectives of these methods are to minimize the uncertainty bounds while maximizing the coverage of the observations. If two models are compared, it is these two metrics that are used to distinguish between the models, rather than the performance of the individual model outputs. While reduced uncertainty bounds are preferred, they are not necessarily indicative of model improvements in a purely deterministic context.

Chapter 3
Methods

3.1 Study Area

The Alouette Lake basin, chosen for use in the case study of this thesis, is a coastal basin in the south-west of British Columbia. The watershed is shown in Figure 3. It is a relatively small basin with a drainage area of 202 km^2, and approximately 16 km^2 of that area is occupied by the Alouette lake reservoir. The lake lies in a narrow valley located in the Coast Mountain range extends across much of the length of the watershed, meaning rainfall has a relatively small travel distance to reach the lake.

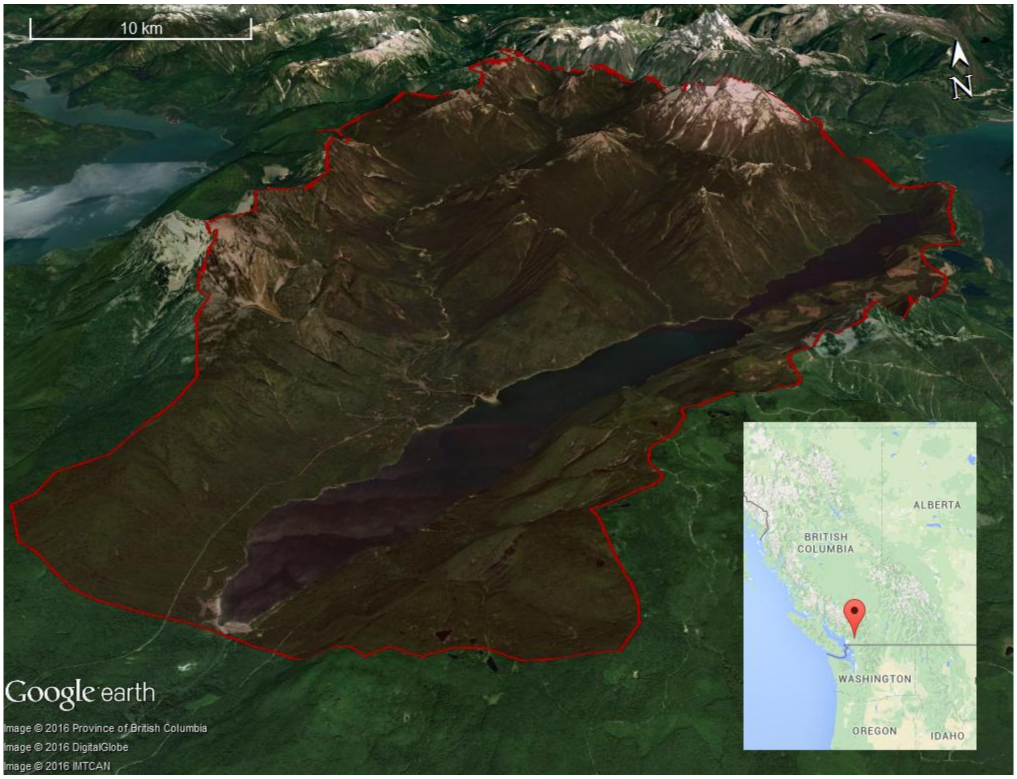

Figure 3: The Alouette Lake Basin terrain map

The basin is steep and mountainous with elevations ranging from 120 m to 1800 m, and a median basin elevation of 650 meters above sea level (masl). The hypsometric curve for the basin describing the percentage of the watershed above certain elevations is shown in Figure 4. The Alouette basin is steep, and combined with the short travel distance, this results in reservoir inflow rising rapidly in response to rainfall and receding equally fast after a storm.

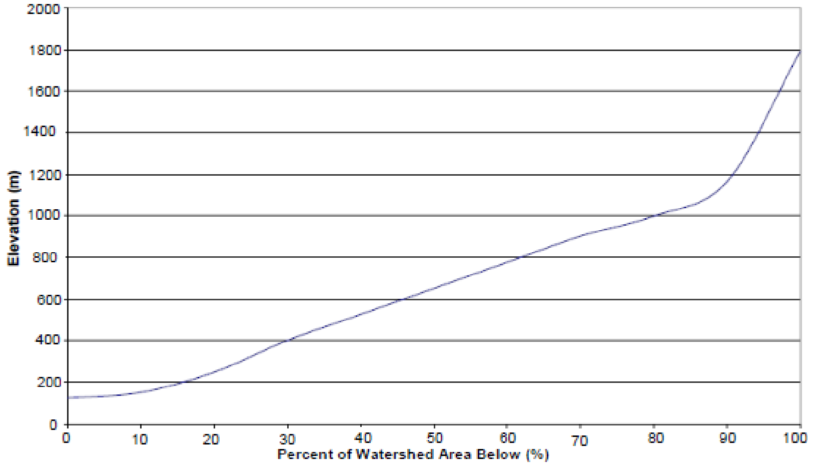

Figure 4: Hypsometric curve for the Alouette River Watershed (BC Hydro, 2009)

The basin is almost entirely comprised of Golden Ears Provincial Park and subject to logging restrictions and other conservation laws, meaning that much of the basin is well-forested. There is glacier cover present in approximately 1.5% of the basin area at the highest elevations, however glacier melt has minimal impact upon total reservoir inflows.

Runoff inflows to Alouette Lake follow an irregular temporal distribution, with high flows occurring in the spring due to snow melt, and in the fall from strong seasonal storms caused by the orographic effects of moist ocean air meeting the Coast Mountains. The summer months are generally dry, with only occasional convective rainfall providing minor peaks in flow. Calculated daily reservoir inflows were provided by BC Hydro for the period of 1987 to 2007, as well as hourly inflows from 2010 to 2016. The historical inflows are shown in Figure 5.

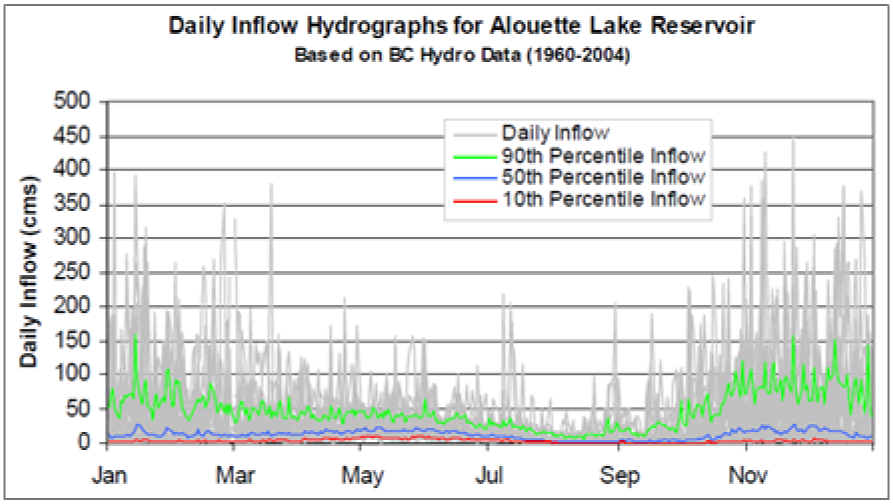

Figure 5: Historical Daily Inflows to Alouette Lake Reservoir (BC Hydro, 2009)

Meteorological data used in the model of the Alouette basin was also provided by BC Hydro for the same periods. The UBCWM implementation for the Alouette basin uses a single synthetic climate station to provide temperature and precipitation measurements. The synthetic station is a linear combination of three nearby climate stations using a weighted average, and the effective station elevation is a calibrated parameter. The weights were derived from a regression analysis performed by BC Hydro of all the surrounding gauging stations to find the optimal combination, and is the same combination used by BC Hydro for their operational forecasts. The use of a single synthetic station simplifies the problem of distributing the forcings throughout the basin, which can be difficult given the spatial heterogeneity of precipitation in particular, and is also less sensitive to missing or inaccurate measurements at the individual stations. The station names and the assigned weights are provided in Table 1, and the station locations are shown in Figure 6. The Coquitlam station is located outside of the Alouette basin, but it is near enough that it is still representative of the conditions in the basin, and the precipitation was found to correlate well with observed runoff.

Table 1: Meteorological Station used for modelling the Alouette basin

Station	ID	Elevation (masl)	Precipitation Weight	Temperature Weight
Alouette Lake Forebay	ALU	125	0.48	0.33
Coquitlam River above Coquitlam Lake	CQM	290	0.37	0.33
Gold Creek	GOC	794	0.15	0.33

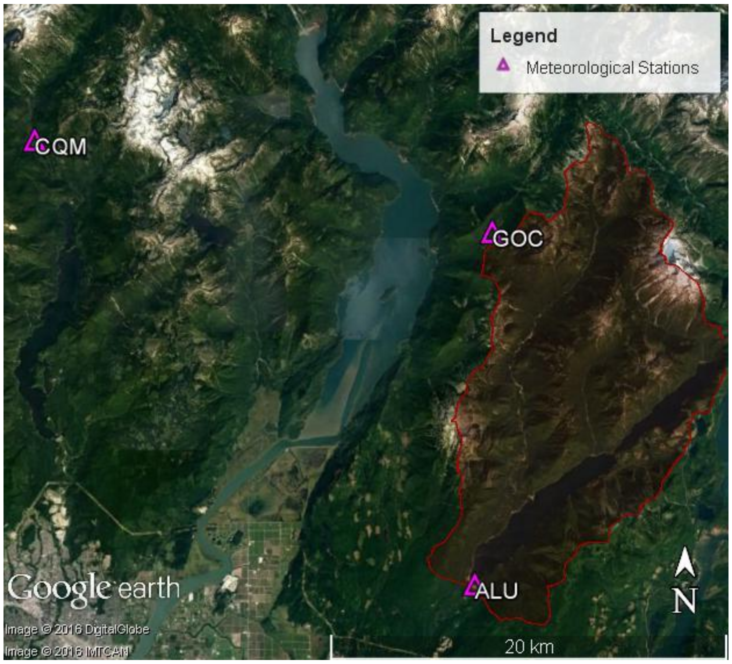

Figure 6: Alouette Meteorological Station Locations

3.2 Inflow Observation Errors

The observed streamflows used in the case studies for this thesis are actually estimated inflows to a reservoir used for hydroelectric generation, and so cannot be measured directly the way flow in rivers are measured. The detailed description of the inflow observation methods are provided here to explain the estimation of errors in this method, required for the application of the model evaluation process used later. The reservoir inflows are calculated using a basic mass balance approach using measured discharge through the Alouette dam, and the change in reservoir storage (BC Hydro, 2009). The general formula for the inflow calculation is given in equation (1).

$$I = O + \Delta S \tag{1}$$

where I is the average inflow to the reservoir over a time period (e.g., 1 hour), O is the average outflow through all dam outlets, and ΔS is the change in reservoir storage for the same period.

The reservoir storage is not measured directly, but derived from a stage-storage curve linking the measured reservoir elevation to a volume of water. This means that the change in storage is subject to several sources of uncertainty, such as, instrument error in the elevation readings, wind regime in the reservoir leading to non-representative elevation measurements and errors in the stage-storage curve which could be changing in time due to deposition and morphological changes. These reservoir storage errors are most pronounced in larger reservoirs, where a small error in elevation measurements can cause significant differences in the calculated volume. Even though Alouette lake is a relatively small reservoir, the storage errors are still the most significant source of uncertainty in the inflows.

Similarly, reservoir output is also estimated based on elevations, power generation output, operation of the spillway gates and rating curves and flow relationships for the various outlets. There are multiple outputs from the Alouette reservoir and each is subject to uncertainty, though the uncertainty for these engineered structures tend to be smaller than that found in natural systems. A diagram showing the various outputs from the Alouette lake reservoir is shown in Figure 7. The calculated outflows ignore other sources of water loss which adds to the uncertainty and means the calculated inflows potentially underestimate the true reservoir inflows. These other sources of water loss include seepage through or under the dam, infiltration to groundwater, and evaporation from the lake, which is seasonal and is most significant in the summer.

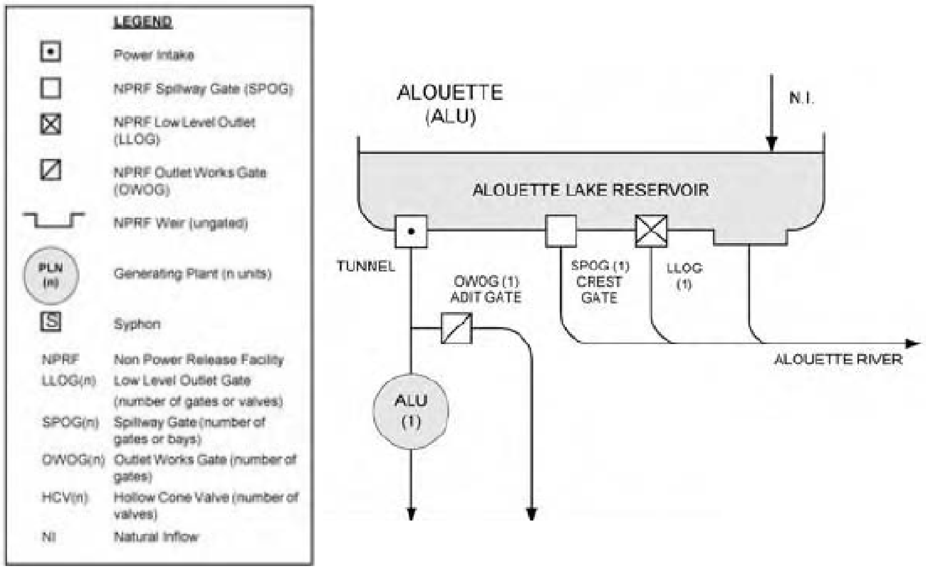
LEGEND
Power Intake
NPRF Spillway Gate (SPOG)
NPRF Low Level Outlet (LLOG)
NPRF Outlet Works Gate (OWOG)
NPRF Weir (ungated)
PLN (n)
Generating Plant (n units)
S
Syphon
NPRF Non Power Release Facility
LLOG(n) Low Level Outlet Gate (number of gates or valves)
SPOG(n) Spillway Gate (number of gates or bays)
OWOG(n) Outlet Works Gate (number of gates)
HCV(n) Hollow Cone Valve (number of valves)
NI Natural Inflow
ALOUETTE (ALU)
N.I.
ALOUETTE LAKE RESERVOIR
TUNNEL
OWOG (1) ADIT GATE
SPOG (1) CREST GATE
LLOG (1)
ALU (1)
ALOUETTE RIVER

Figure 7: Alouette Lake Reservoir output configuration (BC Hydro, 2009)

Inflows calculated using this method tend to generate hydrographs with a significant amount of random errors, i.e. noise. Figure 8 is an example of the calculated inflows to the Alouette lake reservoir in January 2002, showing significant noise with some inflows reaching negative values which has no physical meaning without the presence of errors. While the errors are especially noticeable during periods of lower inflows when there is a lower signal-to-noise ratio, the errors appear to be relatively constant throughout the year (Weston, 2011).

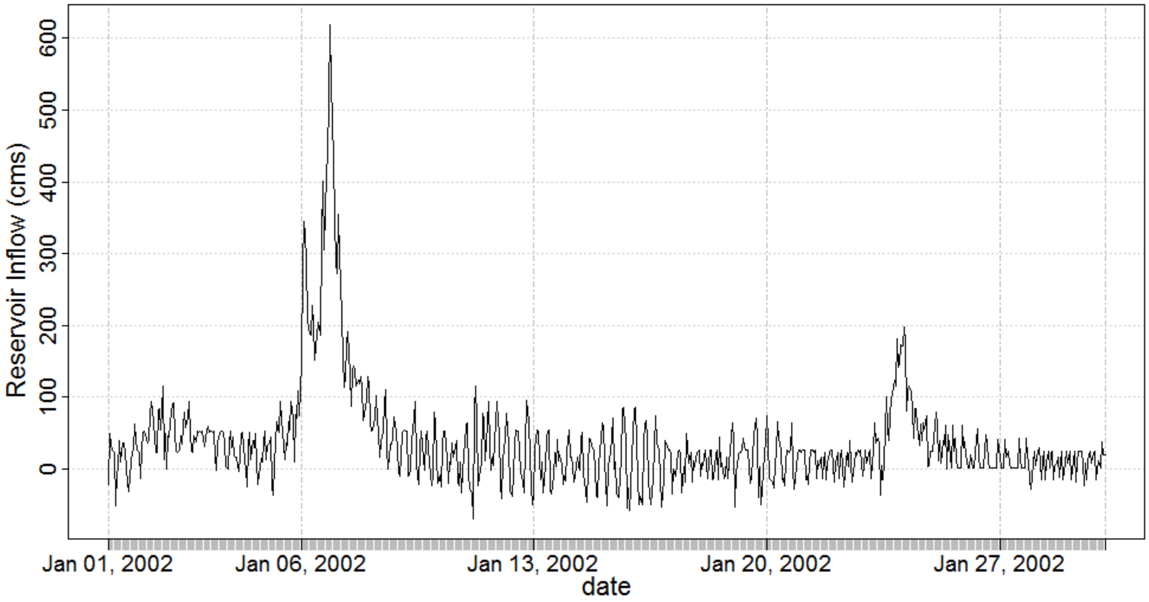

Figure 8: Example of noise in the Alouette inflows January 2002

Good inflow records are essential for the calibration of hydrological models, and for assessing their quality in validation, so steps must be taken to account for the noise present in the raw data. Currently the daily inflow records are quality controlled by BC Hydro by comparing them to nearby Water Survey of Canada streamflow gauges which have much less noise. Errors are corrected while attempting to respect the long term inflow volume as much as possible. This process cannot be done in real time, so the data are quality controlled a few years later.

BC Hydro has recently begun a preliminary automatic quality control process for hourly inflows by smoothing the hourly inflow data. Reservoir inflow rates should change relatively slowly, especially during low flow periods which aren't driven directly by rainfall (Deng, Liu, Guo, Wang, & Wang, 2015). The observed data fluctuates around local mean values that should be near to the true inflows due to the slow rate of change. By smoothing the data noise can be reduced while maintaining a good approximation of the relevant data.

Several of the methods suggested by Berrada, Bennis, & Gagnon (1996) were tested by BC Hydro and a quadratic moving average was selected as the best smoothing filter (Weston, 2011). The quadratic moving average uses linear least squares regression to fit a second degree polynomial through subsets of the data centered on each point. This is a specific version of the commonly used Savitzky-Golay filter (Savitzky & Golay, 1964) which reduces least squares polynomial smoothing of any order and sampling window to a weighted moving average whose weights only need to be calculated once. The benefit of the quadratic moving average over the simple linear moving average is that it is able to remove much of the noise from the data while better maintaining the magnitude of the natural extreme peaks and valleys, which are modulated by the simpler process.

A two-tiered approach was adopted by BC Hydro to smooth the inflows whereby a longer window is used for the moving average during periods of low flows. This allows for the higher flow periods, which tend to change more rapidly to be smoothed using a shorter moving average that is able to maintain sharp natural peaks in the hydrograph, and a longer moving average during low flows to produce smooth slow changing base flows. The lengths of the windows used for the smoothing vary for each reservoir. For the Alouette lake reservoir, the length is 19 hours for high flows above 25 cms and 49 hours for low flows. The smoothed version of the January 2002 hydrograph is shown in Figure 9. The figure shows that much of the noise is removed using this procedure, but the magnitude of the peak on January 7th is significantly reduced. It is difficult to determine whether the true inflow during these high flow events is being underestimated, or if the peak in the raw data is exaggerated due to the noise. It is this uncertainty that is examined throughout this chapter.

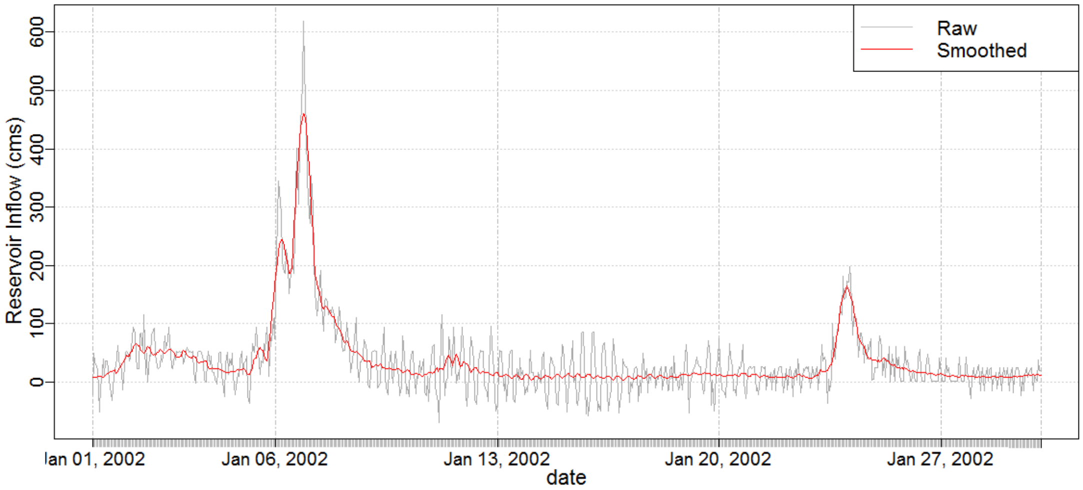

Figure 9: Smoothed Alouette Inflow Hydrograph January 2002

One benefit of using the quadratic moving average is that it is fit using linear regression, and thus it is easy to calculate a confidence interval for this fit. The confidence interval from the smoothing can provide a rough estimate of the uncertainty in the observations. The 90% confidence interval from the quadratic moving average is shown in Figure 10. It is important to note that this is not the true uncertainty of the inflows, only a rough estimate based on the statistical properties of the selected smoothing algorithm. Nevertheless, the confidence interval provides a useful rough approximation of the uncertainty, and appears to have the expected characteristics, such as increased uncertainty around peak flows.

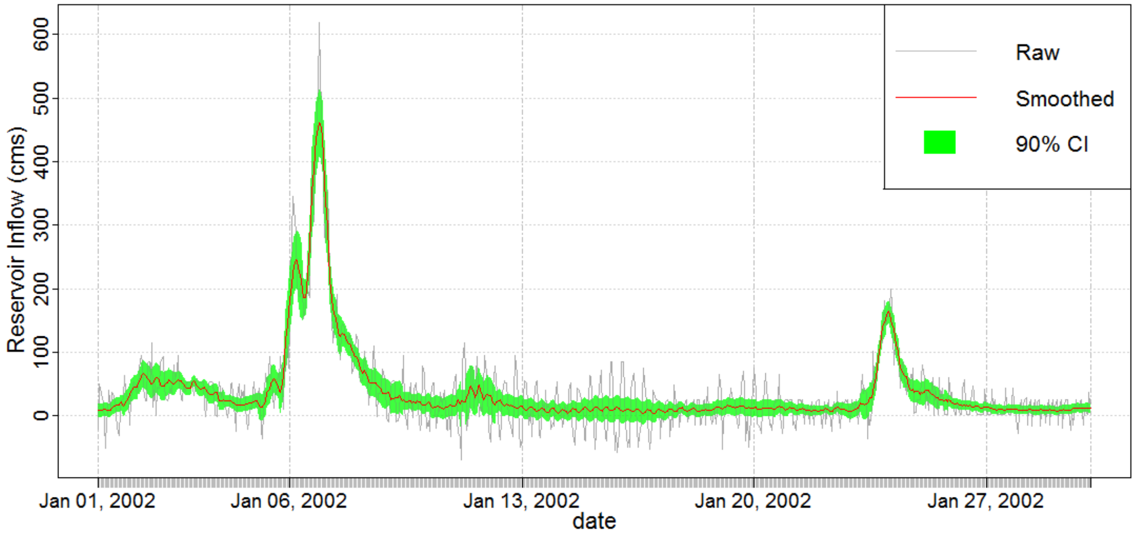

Figure 10: 90% confidence interval for the Alouette lake reservoir inflows

In order to test the effects of the inflow uncertainty on the calibration of a hydrological model and on the comparison of different submodels, it is necessary to generate different plausible deterministic inflow hydrographs rather than simply a confidence interval. One possible method would be to randomly sample each point from within the confidence interval, but such an approach produces hydrographs that fluctuate wildly, similar to the raw data.

The method used to generate plausible inflow hydrographs within the expected uncertainty bounds is to randomly generate weights for each point in the raw data, then use a weighted version of the quadratic moving averages to smooth the raw data following BC Hydro's two-tiered approach. Introducing the uncertainty before smoothing allows for more realistic continuous hydrographs to be generated while maintaining an appropriate amount of variance in the results. An example of 100 different hydrographs generated using this process are shown in Figure 11 along with the 90% confidence interval. The weights are drawn from uniform distribution with a range of 0.5 to 1.5. The distribution of weights was selected arbitrarily, but was found to generate sufficiently different hydrographs while remaining within the confidence interval with few exceptions. It is important to note that these realizations are not expected to fully characterize the expected uncertainty distribution, but rather provide a distribution of plausible inflow hydrographs which respect the original data.

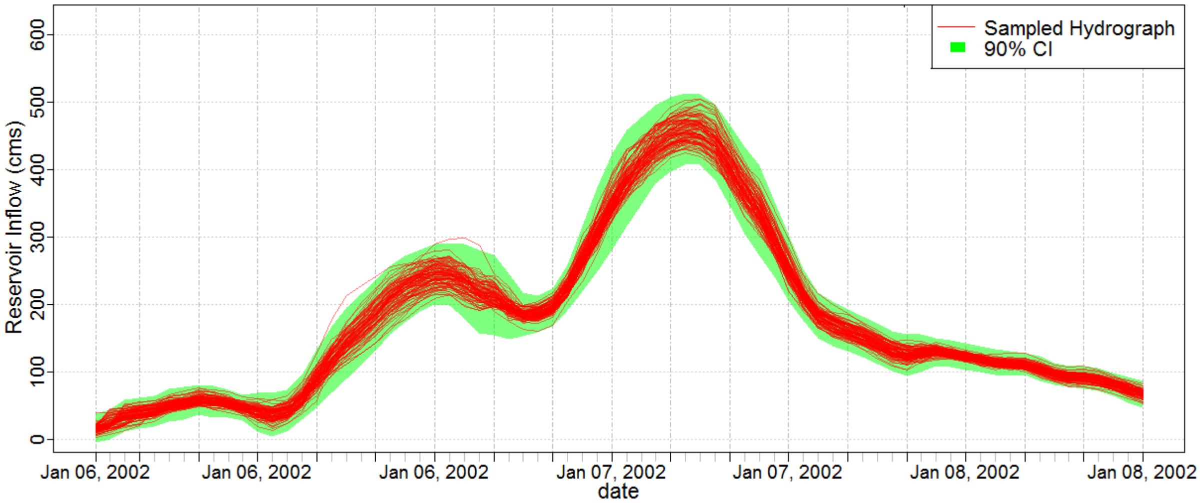

Figure 11: Sampled deterministic hydrographs using a randomly weighted quadratic moving average smoothing process

3.3 University of British Columbia Watershed Model

The UBCWM is a semi-distributed, deterministic, conceptual hydrological model. It was designed for forecasting the runoff from mountainous catchments (Quick & Pipes, 1977), and has been applied to various regions around the world, from coastal mountains in BC, to the Himalayas , and throughout

Europe (Micovic & Quick, 2009). Since the model was developed for use in mountainous areas where data tends to be relatively sparse, the data requirements for the UBCWM are only the daily minimum and maximum air temperature, and the daily precipitation. A flowchart of the general workings of the UBCWM are provided in Figure 12, and a complete description of the processes can be found in (Quick & Singh, 1995).

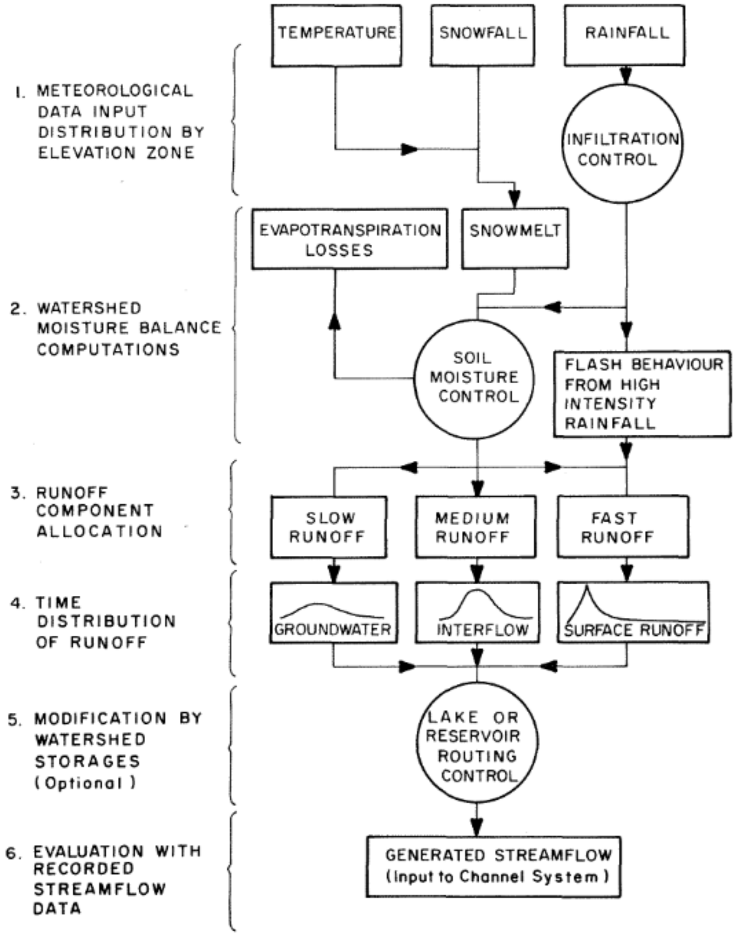
TEMPERATURE
SNOWFALL
RAINFALL
1. METEOROLOGICAL DATA INPUT DISTRIBUTION BY ELEVATION ZONE
INFILTRATION CONTROL
EVAPOTRANSPIRATION LOSSES
SNOWMELT
2. WATERSHED MOISTURE BALANCE COMPUTATIONS
SOIL MOISTURE CONTROL
FLASH BEHAVIOUR FROM HIGH INTENSITY RAINFALL
3. RUNOFF COMPONENT ALLOCATION
SLOW RUNOFF
MEDIUM RUNOFF
FAST RUNOFF
4. TIME DISTRIBUTION OF RUNOFF
GROUNDWATER
INTERFLOW
SURFACE RUNOFF
5. MODIFICATION BY WATERSHED STORAGES (Optional)
LAKE OR RESERVOIR ROUTING CONTROL
6. EVALUATION WITH RECORDED STREAMFLOW DATA
GENERATED STREAMFLOW (Input to Channel System)

Figure 12: UBCWM Generalized flow chart (from Quick & Pipes, 1977)

The semi-distributed nature of the UBCWM comes from division of the watershed into several units based on elevation. These elevation bands are used to distribute the temperature and precipitation forcings after accounting for the orographic gradients that play an important role in mountain hydrology. Each band is characterized by its own physical parameters, such as impermeable area and forest cover, and runoff is calculated separately in each band.

The UBCWM has been ported to the Raven hydrological modelling framework (Craig et al., 2016) which is able to emulate the original UBCWM nearly perfectly. The Raven implementation of the UBCWM is what is now being used operationally at BC Hydro, and is used throughout this thesis as the baseline model. The use of Raven allows for more flexibility to be applied in the use of the UBCWM, and potential modifications to the model.

A version of the UBCWM running at an hourly timestep was developed for the Alouette basin based on the current daily operational model of BC Hydro. The base UBCWM and all modified versions of the model in this thesis were calibrated using the DDS optimization algorithm (Tolson & Shoemaker, 2007) with a computational budget of 1000 model runs. The objective function used was the root mean square error (RMSE). Optimizing root mean square error is equivalent to optimizing the Nash-Sutcliffe efficiency (NSE) metric, since the Nash-Sutcliffe value is simply a normalized version of the MSE. Hourly precipitation and runoff data from a two year period, 2011-2013, was used for calibration following a one year warm up period used to minimize the effects of initial conditions. Another two year period from 2014-2015 was used as the validation period, with the calibration period acting as the warm up. The base UBCWM had 15 calibrated parameter summarized in Table 2.

Table 2: Calibrated parameters of the base UBCWM

Parameter	Description	Processes involved
A0TLZP	Temperature lapse when precipitation > A0PPTP rate (°C / 1000m)	Lapse rate
A0TLXM	Lapse rate of maximum daily temperature (°C / 1000m)	Lapse rate
A0TLNM	Lapse rate of minimum daily temperature (°C / 1000m)	Lapse rate
P0TEDL	Lapse rate of maximum temperature range (°C / 1000m)	Lapse rate
P0ALBMIN	Albedo of very deep and aged snowpack and of glacier	Potential Melt
P0PERC	Groundwater percolation (mm/day)	Infiltration/Percolation
P0RREP	Correction factor for rainfall (%)	
P0SREP	Correction factor for snowfall (%)	
C0ELPT	Elevation of the meteorological station (m)	Lapse rate
P0FRTK	Fast runoff rate constant (1/days)	Routing
P0IRTK	Interflow rate constant (1/days)	Routing
P0UGTK	Upper groundwater runoff rate constant (1/days)	Routing
P0DZTK	Deep groundwater runoff rate constant (1/days)	Routing
V0FLAS	Flash flood threshold (mm)	Infiltration/Percolation
P0DZSH	Deep zone share (%)	Infiltration/Percolation

3.4 Model Hypotheses

To demonstrate the proposed testing methodology, several different changes to the UBCWM were tested and compared to a base model in the context of uncertainty in inflows. Each of the changes are described below, with more detailed descriptions of the UBCWM algorithms available in (Quick & Singh, 1995).

3.4.1 Infiltration/Percolation/Routing

Infiltration is the process by which water enters the top soil, and an infiltration algorithm must partition the rainfall and snowmelt into infiltrated water and surface runoff. Percolation refers to the downward movement of water through the soil. The soil model used in this thesis has a single layer representing the unsaturated topsoil, and two layers representing deeper groundwater stores. The percolation algorithms are used to calculate the rate of water movement from the topsoil to the groundwater storage.

Routing is the process that involves the delay and redistribution of the runoff before it reaches the outlet of the watershed, or in this case, the reservoir. This is one of the most important processes since it controls the timing of the peak flows following rainfall.

Two different conceptualizations of the infiltration and percolation to groundwater were implemented using Raven. In the original UBCWM, infiltration, percolation and , and the partitioning of the water into different routing components are all handled using a single algorithm which is why the three processes are tested together here.

3.4.1.1 UBCWM

The infiltration algorithm for the UBCWM uses a hierarchical system to subdivide rain and snowmelt inputs into several storages representing evaporation loss, and fast, medium, slow and very slow runoff. The storages are prioritized with excess water from each priority overflowing to the next lower priority.

The highest priority for the rain and snowmelt is the fast runoff representing surface runoff from impermeable areas. Each elevation band has a parameter defining the impermeable fraction, which can be further modified based on the current soil moisture, with effective runoff increasing with moisture content. The impermeable fraction determines the amount of water going directly to the fast runoff storage, with the remaining water passing to the lower priorities. In addition to the runoff from the impermeable area, during times of high intensity rainfall the precipitation falls faster than it can infiltrate into the soil and so a higher percentage of water goes to fast runoff.

The second priority is the soil moisture deficit. Rather than specifying a total soil moisture capacity, the UBCWM tracks the moisture deficit in the soil caused by evaporation from the soil. The soil moisture deficit must be completely satisfied before any more water is available for runoff.

The third priority is the slow and very slow runoff storages. Input to these storages is a conceptualization of percolation to groundwater. The very slow runoff represents deep groundwater storage which is typically nearly constant with output constituting the river baseflow. The slow runoff is an upper groundwater storage that tends to vary more quickly and represents a seasonal component to the baseflow. The percolation to these two storages accepts any water up to a fixed limit. This limit is most often a calibrated parameter. The subdivision of the percolation between the two groundwater stores is fixed ratio that is also typically calibrated.

The last priority is the medium runoff store representing interflow. Despite having the lowest priority, the interflow is often the most significant component to runoff during times of high volume rainfall and snowmelt.

Each of the runoff storages is routed using a Nash unit hydrograph (J. Nash, 1957), where the inflows are passed through a series of linear reservoirs to produce the appropriate timing for the flow. The two groundwater stores each use a single reservoir for routing, while the fast and medium runoff use two and three reservoirs, respectively.

The combined infiltration, percolation and routing involves a total of seven calibrated parameters, which is almost half the calibrated parameters in the base model. Of these seven parameters, four are involved with routing the different runoff components, and three involve the infiltration and percolation processes.

3.4.1.2 HBV

The second implementation of the infiltration and percolation processes is referred to throughout this thesis as HBV infiltration. The infiltration algorithm used is the same as what is used in the HBV model (Bergström & Singh, 1995), but the percolation and groundwater processes are handled differently.

In this conceptualization of infiltration, the infiltration rate is a nonlinear relationship between the amount of snowmelt and rainfall, and the current saturation of the topsoil. The infiltration formula is given in equation (2).

$$M_{inf} = R\left(1 - \left(\frac{\phi_{soil}}{\phi_{max}}\right)^{\beta}\right) \tag{2}$$

where M_{inf} is the infiltration rate (mm/d), R is the rainfall/snowmelt rate (mm/d), β is the shape parameter controlling the nonlinearity, ϕ_{soil} is the current soil moisture content (mm), and ϕ_{max} is the maximum soil water storage (mm).

The percolation from the topsoil to the groundwater stores is calculated using an algorithm based on the Guelph Agricultural Watershed Storm-Event Runoff (GAWSER) model (Schroeter, 1989). The percolation rate is dependent on the saturation above the natural soil moisture after drainage, i.e. field capacity. The percolation formula is given in equation (3).

$$M_{perc} = M_{max}\left(\frac{\phi_{soil} - \phi_{fc}}{\phi_{max} - \phi_{fc}}\right) \tag{3}$$

where M_{perc} is the percolation rate (mm/d), M_{max} is the maximum percolation rate, and ϕ_{fc} is the soil field capacity.

The percolated water is split into slow and very slow groundwater reservoirs using a fixed fraction, exactly the same as with the UBCWM infiltration. The routing is also done in the same way using linear reservoirs, but in this case runoff uses only a single reservoir.

This algorithm has seven calibrated parameters, the same number as the base UBCWM. For this algorithm the P0IRTK and V0FLAS parameters are not used since there is no interflow component and no flash flood correction used. Two new parameters are introduced: the β shape parameter for the HBV infiltration, and the field capacity controlling the amount of percolation.

3.4.2 Lapse Rate

The orographic effects on temperature and precipitation are an important aspect of modelling mountainous regions. In this model lapse rates are used to adjust the temperature and precipitation observations from the elevations of the gauges to the elevation of each band.

Two different methods of calculating the temperature and precipitation lapse rates were tested: the fairly complex UBCWM method, and a very simple linear lapse rate.

3.4.2.1 UBCWM

The UBCWM uses a series of lapse rates and inflection points to describe the orographic correction profile. Normally four lapse rates are calculated for the temperature orographic correction, separate lapse rates for the minimum and maximum temperatures above and below 2000 m elevation, but in this case the watershed doesn't reach the 2000 m threshold, so only two lapse rates are used. The temperature lapse rates depend in part on the precipitation to control the transition from a wet to a dry adiabatic lapse rate. The precipitation correction factor is calculated as

$$V = \begin{cases} \min\left(\frac{P}{A0PPTP}, 1\right), & if\ A0PPTP > 0 \\ 0, & if\ A0PPTP \leq 0 \end{cases} \tag{4}$$

where $V$is the rainfall correction factor, P is the precipitation rate, and $A0PPTP$ is the threshold precipitation for temperature lapse rate. In this thesis $A0PPTP$ is fixed at 5 mm/d. A corrected adiabatic lapse rate is determined by providing a weighted average between the specified dry adiabatic lapse rate and the wet adiabatic lapse rate, calculated as

$$\alpha_c = V\alpha_w + \ (1 - V)\alpha_d \tag{5}$$

where α_c is the corrected adiabatic lapse rate, α_w is the wet adiabatic lapse rate, and α_d is the dry adiabatic lapse rate. The temperature lapse rate is further modified by a daily temperature range factor (w_t), calculated as the current daily temperature range divided by the maximum temperature range parameter A0TERM (fixed at a value of 20°C for this work), as shown in Equation (6).

$$w_t = \frac{T_{max} - T_{min}}{A0TERM} \tag{6}$$

The lapse rates for the maximum and minimum daily temperatures are then calculated as shown in Equations (7) and (8) respectively. As the daily temperature range approaches zero the lapse rates approach the corrected adiabatic lapse rate.

$$\alpha_{max} = w_t A0TLXM + (1 - w_t)\alpha_c \tag{7}$$

$$\alpha_{min} = w_t A0TLNM + (1 - w_t)\alpha_c \tag{8}$$

where α_{max} is the maximum temperature lapse rate, α_{min} is the minimum temperature lapse rate, and $A0TLXM$ and $A0TLNM$ are both calibrated lapse rate parameters. Since the model used is run with an hourly timestep instead of the daily model this algorithm was originally designed for, the lapse rate for each hourly temperature is calculated as a linear interpolation between the minimum and maximum temperature lapse rates as shown in Equation (9)

$$\alpha_h = \alpha_{min} + \frac{T_h - T_{min}}{T_{max} - T_{min}}(\alpha_{max} - \alpha_{min}) \tag{9}$$

where α_h is the lapse rate for the hourly temperature, and $T_h$is the hourly temperature. Finally, the corrected temperature for each elevation band in the model is calculated as shown in Equation (10).

$$T = T_g - \alpha(z - z_g) \tag{10}$$

where T is the estimated temperature, T_g is the measured temperature at the gauge, α is the calculated lapse rate, z is the elevation at the center of each elevation band, and z_g is the elevation of the gauge.

The orographic precipitation adjustment uses a temperature-corrected lapse rate with two inflection points, as shown in Figure 13. The base orographic correction equation is shown in Equation (11)

$$P = P_g \cdot (1 + \alpha F_t)^{\frac{z - z_g}{100}} \tag{11}$$

where P is the corrected precipitation rate, P_g is the measure gauge precipitation, and F_t is a temperature correction factor calculated as shown in Equation (12)

$$F_t = \begin{cases} 1, & if\ t_{band} \le 0\ C \\ 1 - A0STAB(t_{band})^2, & if\ t_{band} > 0\ C \end{cases} \tag{12}$$

where $A0STAB$ is the precipitation gradient modification factor, and t_{band} is the temperature at the first listed elevation band in the model. $A0STAB$ is fixed at a value of zero in this thesis, so the temperature correction is not used.

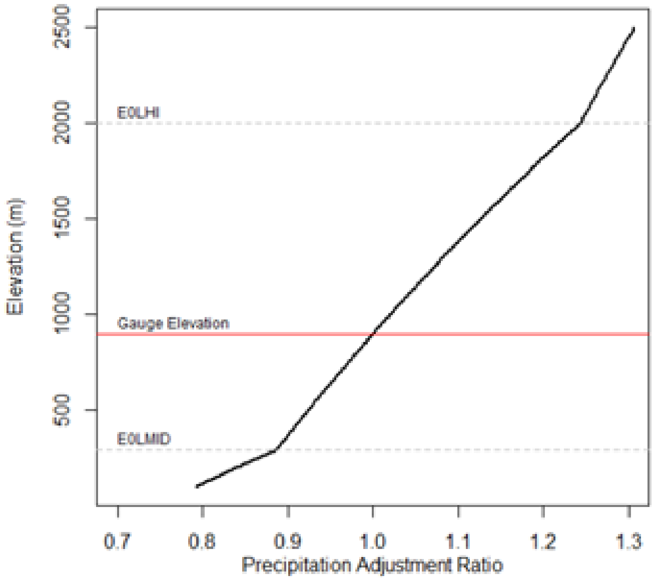

Figure 13: UBC Watershed Model Orographic Correction

This algorithm has five calibrated parameters in this work, four involved with the temperature lapse rates and the last being the gauge elevation. The precipitation lapse rate has no parameter being calibrated, but that lapse has 6 fixed parameters that have previously been calibrated by BC Hydro and provide a good result. The gauge elevation has an effect on both temperature and precipitation corrections, and together with the gauge correction factors for rain and snow provides enough flexibility to the precipitation for adequate calibration.

3.4.2.2 Simple Linear Lapse

As a much simpler alternative to the UBCWM lapse algorithms, a linear lapse rate was used for both precipitation and temperature. The temperature correction assumes completely adiabatic expansion is the cause of the temperature gradient and can be calculated using equation (10) with the lapse rate now being a calibrated parameter representing the adiabatic lapse rate.

The precipitation lapse rate is calculated using a very similar formula given in equation (13). In this case the equation is not based on a simplified physical law, but is an empirical formula coming from the observation that precipitation increases at higher elevations.

$$P = P_g - \alpha(z - z_g) \tag{13}$$

where P is the precipitation at a given elevation band, and P_g is the measured temperature at a gauge.

This algorithm has four calibrated parameters, one less than the base UBCWM algorithm. The three parameters from the base model involving temperature lapse rates are replaced with only one for this algorithm. The P0TEDL parameter was also calibrated for this algorithm since it is involved in the UBCWMs estimation of wind speed, not the orographic correction of temperature. One parameter was added to control the precipitation lapse rate, and the gauge elevation is also calibrated here.

3.4.3 Interception

Interception is a process where precipitation is blocked by vegetation before reaching the ground. The vegetation layer holds water in what is known as canopy storage, and the amount of storage depends on the forest coverage and the amount of leaves. Water exits canopy storage either through evaporation or by eventually dripping off of the trees to the land surface. Two different methods of handling canopy interception were tested.

3.4.3.1 UBCWM

The UBCWM doesn't model the canopy layer, but reduces the amount of precipitation to account for interception. Using this method, a fixed percentage of the precipitation is removed up to a maximum amount. The formula is given in equation (14).

$$P_{thru} = I_{\%} \cdot min(\mathrm{P}, \phi_{\mathrm{cap}}) \cdot F_c \tag{14}$$

where P_{thru} is the is the canopy throughfall, i.e. the precipitation not intercepted, $I_{\%}$ is the percentage of precipitation being intercepted, P is the above canopy precipitation, ϕ_{cap} is the maximum canopy interception rate, and F_c is the forest coverage.

No parameters from this algorithm were used in the calibration, with the interception percentage being fixed at 12%, and the maximum interception rate fixed at 10 mm/d. Preliminary testing showed inclusion of these parameters in the calibration had little impact on the results.

3.4.3.2 Canopy

An alternative to the UBCWM approach is to explicitly model the canopy layer. This set up is more complex, since in addition to the interception process, the canopy storage needs to be tracked, accounting for evaporation and dripping from the canopy. The amount of interception for this method uses a similar formula to the UBCWM and is given in equation (15).

$$P_{int} = min(I_{\%} \cdot P,\ \phi_{cap} - \phi_{can}) \cdot F_c \tag{15}$$

where P_{int} is the amount of intercepted precipitation, $I_{\%}$ is the percentage of precipitation being intercepted, P is the above canopy precipitation, ϕ_{cap} is the maximum capacity for canopy storage, ϕ_{can} is the current canopy water storage, and F_c is the forest coverage.

This method assumes water evaporates from the canopy at the rate of potential evaporation, as calculated in the UBCWM. The formula used for canopy evaporation is given in equation (16).

$$M_{evap} = PET \cdot F_c \tag{16}$$

where M_{evap} is the rate of evaporation from the canopy, PET is the potential evapotranspiration, and F_c is the forest cover.

Water also leaves the canopy through the canopy drip process, with the drips being added to the precipitation throughfall. The amount of canopy dripping is proportional to the current storage and is calculated using equation (17).

$$M_{drip} = \alpha \left(\frac{\phi_{can}}{\phi_{cap}} \right) \tag{17}$$

where M_{drip} is the amount of canopy drip, and α the drip proportion.

This algorithm adds five new calibrated parameters and is the greatest increase in model complexity of all the algorithms tested. The added parameters are the interception percentages for rain and snow, the maximum canopy storage for rain and snow, and the drip proportion.

3.4.4 Potential Melt

Potential snow melt is a model forcing that is often estimated in models because it is difficult to measure. Two different potential melt algorithms were tested here, both of which are physically based using an energy balance approach.

3.4.4.1 UBCWM

The total potential melt is an accumulation of separate melt components, as shown in Equation (18).

$$M_{melt} = \frac{1}{\lambda_f \rho_w} \big((1 - \alpha_s) S + L_n + Q_c + Q_a + Q_r \big) \tag{18}$$

where M_{melt} is the total potential melt rate, λ_f is the latent heat of fusion, ρ_w is the density of water, α_s is the snow albedo, S is the incoming shortwave radiation, L_n is the net longwave radiation, Q_c is the convective melt energy, Q_a is the advective melt energy, and Q_r is the melt energy due to rainfall. The convective and advective melt energy are calculated using Equations (19) and (20), respectively.

$$Q_c = 0.113 \cdot p \cdot T_a \cdot V \cdot R_M \tag{19}$$

$$Q_a = 0.44 \cdot T_{min} \cdot V \cdot R_M \cdot [(1 - f_c)p + f_c] \tag{20}$$

where p is the air pressure, T_a is the average daily air temperature, T_{min} is the minimum daily temperature, V is the wind velocity, f_c is the fraction of forest cover, and R_M is a reduction factor calculated using Equation (21)

$$\begin{gathered} R_M = 1 - 7.7 \cdot R_I \\ 0 \leq \; R_M \; \leq 1.6 \end{gathered} \tag{21}$$

where R_I is a liberalized estimate of Richardson's number

$$R_I = \frac{0.095 \cdot T_a}{V^2} \tag{22}$$

The melt due to rainfall energy input is calculated using Equation (23)

$$Q_r = \frac{c_w}{\lambda_f} T_a \cdot P_r \tag{23}$$

where c_w is the specific heat capacity of water, and P_r is the daily total rainfall.

The only parameter related to the potential melt is the minimum albedo. The minimum albedo is not a direct parameter of this algorithm, but influences the calculation of the estimated snow albedo, and so ultimately effects the potential melt.

3.4.4.2 Energy Balance

This algorithm is based on the same energy balance principal used in the base UBCWM potential melt algorithm shown in Equation (18), but the melt components are calculated using the methods from (Dingman, 2015). The biggest difference is that the UBCWM algorithm is based on daily values of temperature and precipitation, but this algorithm uses the timestep of the model, in this case hourly values. The shortwave and longwave radiation, as well as the rain energy input are the same for both algorithms, but the convective and advective melt components are calculated differently. The convective melt is calculated using Equation (24).

$$Q_c = \rho_a c_a V \frac{\kappa^2}{log\left(\frac{h}{z_0}\right)^2} (T_a - T_s) \tag{24}$$

where ρ_a is the density of air, c_a is the specific heat capacity of air, κ is the von Karman constant, h is the reference height for wind, z_0 is the surface roughness height, T_a is the air temperature, and T_s is the temperature of the snow surface. The advective melt component is calculated using Equation (25).

$$Q_a = \frac{\lambda_v \rho_a M_a V}{p M_w} \cdot \frac{\kappa^2}{log\left(\frac{h}{z_0}\right)^2} (P_a - P_s) \tag{25}$$

where λ_v is the latent heat of vaporization, M_a and M_w are the molecular weights of air and water, respectively, P_a is the vapor pressures of air at height h, and P_s is the vapor pressure at the snow surface. This method adds no additional calibrated parameters, but maintains the same minimum snow albedo parameter as the base model.

3.5 Statistical Methods

This section briefly describes the statistical tests used later in this thesis for hypothesis testing and model comparisons.

3.5.1 Parametric vs. Nonparametric Methods

Parametric tests are statistical procedures that rely on assumptions about the shape of the distribution from which the sample is being drawn, most often assuming a normal distribution. A nonparametric test makes no or few assumptions about the underlying population distribution (Hoskin, 2012). By making assumptions about the distribution, parametric tests have more information available leading to increased statistical power. Statistical power is the likelihood of rejecting the null hypothesis when it is false, meaning nonparametric tests will generally require larger sample sizes to detect model differences. When the assumptions of a parametric test are violated, the test may give incorrect results, so a nonparametric alternative should generally be used, especially when distributions are non-normal.

3.5.2 Wilcoxon-Mann-Whitney

The Wilcoxon–Mann–Whitney test is a nonparametric hypothesis test to check the statistical significance of the difference between two distributions (Montgomery & Runger, 1994). The test was developed to check for whether or not two samples came from the same population, especially cases of stochastic dominance, i.e. whether one sample is more likely to be greater than the other (Mann & Whitney, 1947). Here, this test can be used to assess whether one population of model quality metrics is statistically distinguishable from another.

The Wilcoxon–Mann–Whitney is performed by first ranking all values, with rank 1 being the lowest value. The test statistic (U) can then be calculated using equation (26).

$$U_i = R_i - \frac{n_i(n_i + 1)}{2} \tag{26}$$

where $U_i$is the test statistic for the ith sample, R_i is the sum of the ranks from the ith sample, and n_i is the sample size of the ith sample. The equation is used to find the U value for the two samples, and the lower of the two is compared to critical values to check for significance. The test was performed using the "wilcox.test" function from R, an open source statistical programming language (R Team, 2015).

3.5.3 ANOVA

Analysis of variance (ANOVA) is a method of partitioning the total variation found in a dataset into meaningful components attributable to different sources of variation. The ANOVA method can be used to test for significant differences between means of multiple groups. There are three types of ANOVAs: fixed effects models, where the factors being compared are controlled by the experimenter, random effects models, where the factors are a random variable sampled from a larger population, and mixed effects models, a combination of the other two models, where the experiment has both controlled and random factors.

The mixed effects model is what is used in this thesis. The full explanation of the ANOVA calculations for mixed effects models is too long to include here, but details can be found in most statistics textbooks (e.g. Montgomery & Runger, 1994). The calculations of the ANOVA statistics was done using the package 'ez' (Lawrence, 2015) available in R.

3.5.4 Bootstrapping

Bootstrapping is a method of estimating the accuracy of any sample statistic (Efron & Tibshirani, 1994). Any statistic based on a sample from a larger population will contain some amount of error, but sampling multiple times from the population multiple times to get an estimate may be too costly. The basic idea behind bootstrapping is that a sample can be treated as an approximation of the total population, i.e. a pseudo population. New samples can be drawn from this pseudo population with replacement in a process called resampling. Since every value in the pseudo population is known, resampling can be done very easily, so in bootstrapping the data is resampled many times (100,000 times in this thesis), and the error in the statistic of the resamples can be estimated. In this thesis the error is shown using a confidence interval found using the percentile interval method, where the 90% confidence interval is defined as the 5th and 95th percentiles of all the resampled statistics. The bootstrap method is nonparametric and can be used for any distribution, as long as the original sample is representative of the population.

3.5.5 Effect Sizes

Hypothesis tests, such as the Wilcoxon-Mann-Whitney and ANOVA are designed to test for statistically significant differences, but with a large sample size even very small differences can be detected as statistically significant. Hypothesis tests alone are not sufficient to see if model differences produce performance improvements large enough to be meaningful for any practical purpose. A common method of quantifying the difference between groups in statistics is the use of effect size indices, which not only give a measure the size of the difference, but are also standardized so they can be compared between different studies. Unlike with hypothesis tests, there is no standard effect sizes that can be considered significant, it is up to each user to decide if an effect is large enough for practical purposes.

3.5.5.1 Cohen's d

Cohen's d is an effect size measure used when comparing two means. It is one of the most widely used effect size statistics, which is why it is included in this thesis. Cohen's d is calculated as

$$d = \frac{\overline{x_1} - \overline{x_2}}{s} \tag{27}$$

where d is the effect size, $\bar{x}$ is the sample mean, and s is the pooled standard deviation, i.e., the average spread of all data points about their group means. Cohen's d is used to describe the standardized mean difference of an effect (Lakens, 2013). When paired samples are used, s should instead be the standard deviation of the distribution of paired differences. Cohen's d could be useful for comparing the effects of modelling decisions between watersheds, since each watershed may have different degrees of uncertainty, which would likely result in different standard deviations for the performance measure distributions.

3.5.5.2 Generalized Eta Squared

The eta squared (η^2) effect size statistic is a measure of the proportion of the observed variation that can be attributed to each effect in an ANOVA. Eta squared can be calculated as

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \tag{28}$$

where η^2 is the eta squared effect size, SS_{effect} is the sum of squares of the effect, and SS_{total} is the total sum of squares, i.e., the total variability of the data. The sum of squares values are calculated as part of the ANOVA process, so this is an easy statistic to use following an ANOVA. Eta squared has a fairly intuitive interpretation as the percentage of variance attributed to the effect, meaning an η^2 of 0.2 implies that 20% of the total variance is due to the effect.

The eta squared value calculated in equation (28) is not able to be compared between studies since the total variability of each study will be different. Olejnik & Algina (2003) proposed a generalized eta squared statistic (η_G^2) which corrects for differences between studies that have different experimental designs, though by correcting for these differences, the effect sizes no longer sum to one, so it is slightly less intuitive. It is this statistic that is used in this thesis, as calculated by the ezANOVA function from the 'ez' package (Lawrence, 2015) available in R.

3.5.5.3 Probability of Superiority

The final effect size measure used in this thesis is the probability of superiority, sometimes called the 'common language effect size' (McGraw & Wong, 1992). As the name implies, the probability of superiority (PS) is the percent chance that a random value taken from the one population, will be greater than a random value taken from a second population. PS can be easily calculated from the results of a Wilcoxon–Mann–Whitney test as

$$PS = \frac{U}{n_1 \times n_2} \tag{29}$$

where PS is the probability of superiority, U is the Wilcoxon–Mann–Whitney test statistic, and n_1 and n_2 are the sample sizes of the first and second samples, respectively.

When paired samples can be used, PS can be calculated as

$$PS = \frac{N_{x-y>0}}{n} \tag{30}$$

where $N_{x-y>0}$ is the number of times the paired differences are greater than 0, and n is the total number of paired differences.

Chapter 4
Model Comparisons

4.1 Introduction

This section details a method of rigorously testing the performance of two competing modelling decisions to see if they can be differentiated despite being obscured by the presence of observation uncertainty. This method is contrasted with conventional ad hoc approaches which evaluate submodel supremacy via simple metrics in the absence of consideration of uncertainty. The method uses a Monte Carlo method to sample from the uncertain observations, and then treats each sample as a single deterministic set of plausible observations. The sampled observations are used to calibrate the models and evaluate them using as a validation data set. The results from the many different calibrations and validations are assembled into distributions of performance statistics which are then compared using statistical hypothesis testing. Three approaches of increasing severity are tested:

1. Fixed observation data in calibration and uncertainty in validation data (see section 4.2)
2. Uncertainty in validation and calibration data (see section 4.3)
3. Uncertainty in forcings and inflow data (see section 4.5)

Compared to the basic method which ignores uncertainty, these methods provide a more rigorous test, at the expense of increased computation and analysis.

4.2 Uncertainty in Validation Period Only

Many hydrological models, including the models being tested here, have a large number of calibrated parameters, which allow considerable flexibility in the models. The performance of a model during calibration is in part a measure of the models flexibility in fitting a curve, not necessarily its ability to simulate and predict the hydrology. For this reason, the calibration performance is disregarded and the comparison of two different model hypotheses is here done under validation conditions only. Uncertainty in the observed inflows used to evaluate the models introduces uncertainty in the performance statistics.

Applying uncertainty to the validation data to generate distributions of performance metrics requires minimal computational time and modeller effort. In this first test, two different model configurations are calibrated using the original deterministic observations ignoring the effects of uncertainty in the calibration. These two calibrated models are then evaluated under uncertain conditions by comparing them to the many different versions of the validation hydrograph generated stochastically using the process described in section 3.2. Note that the cost of calculating the evaluation metrics is negligible compared to the cost of calibration, so this test is no more expensive than the typical ad hoc comparisons currently being used.

This methodology was applied to two different possible changes to determine whether or not they can be shown to produce improved results considering the uncertainty in the validation data. The two model configurations evaluated were the infiltration and potential melt algorithms described in sections 3.4.1 and 3.4.4, respectively.

Applying the uncertainty to the validation data and using Nash-Sutcliffe efficiency as the performance statistic, the results using M=100 sampled validation hydrographs for the two different infiltration algorithms are shown in Figure 14. In this case the difference in performance is greater than the uncertainty introduced to the validation data (i.e., there is no overlap in the distributions), so the HBV infiltration algorithm is clearly an improvement given this level of uncertainty, and no further analysis is necessary.

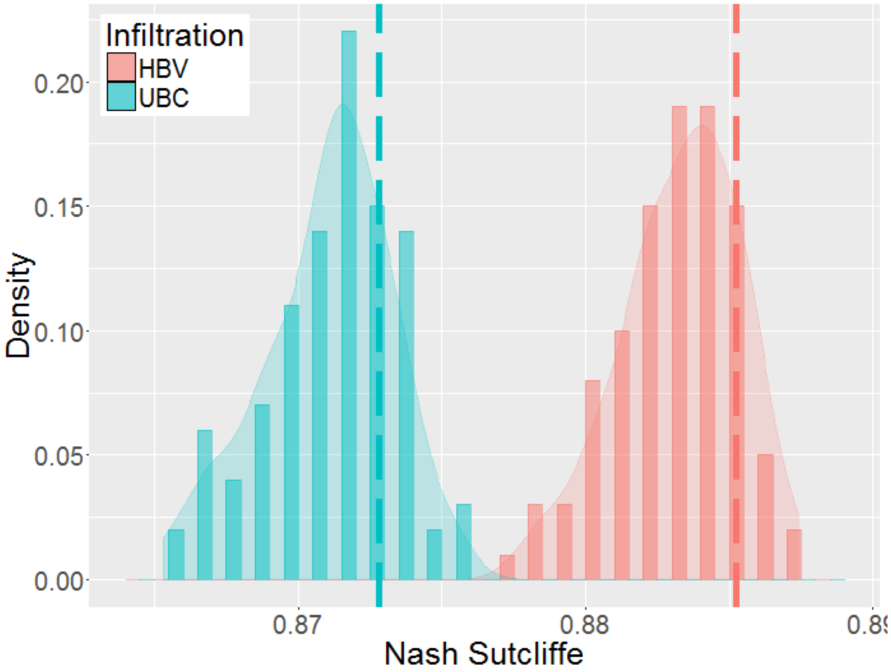

Figure 14: Distributions of Nash Sutcliffe values for two different infiltration algorithms generated by considering validation data uncertainty. Lines show results from a single calibration to original observations.

Once again using M=100 samples of validation data, and the Nash-Sutcliffe performance statistic, the results of the two different potential melt algorithms are shown in Figure 15. Here there is a considerable overlap between the two distributions, and it is not immediately clear whether or not the energy balance algorithm results in a fundamental improvement in model quality. Further analysis is needed to compare the two distributions using statistical hypothesis testing to determine if the difference is significant.

The results of the two models were compared using the Wilcoxon–Mann–Whitney two-sample rank-sum test to check the statistical significance of the difference in distributions. The median Nash-Sutcliffe values for the UBCWM and energy balance potential melts were 0.8709 and 0.8720 respectively; the difference in performance distributions of the two algorithms was found to be statistically significant at the 95% confidence level (Mann-Whitney U = 3832, p=0.004 < 0.05).

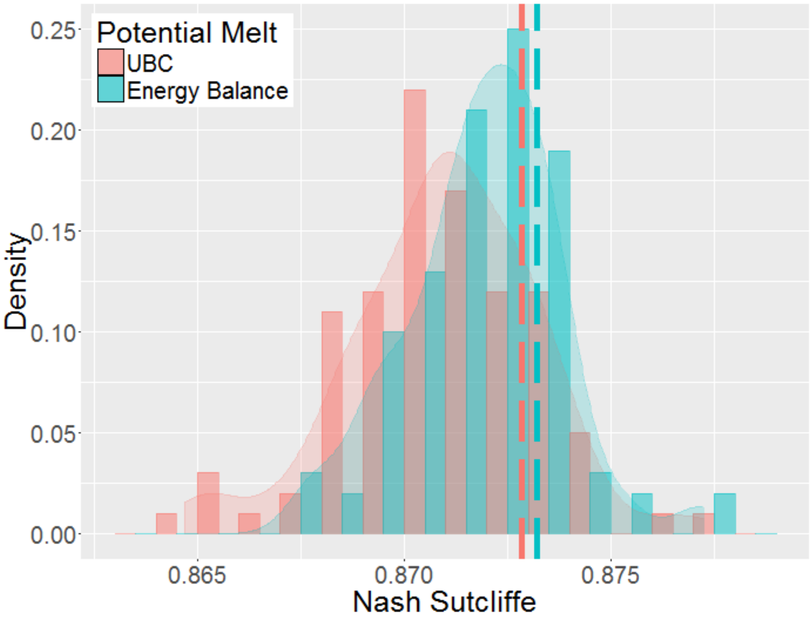

Figure 15:Distributions of Nash Sutcliffe values for two different potential melt algorithms generated by considering validation data uncertainty. Lines show results from a single calibration to original observations

While the hypothesis test above shows that there is a significant difference in performance between the models, it is not sufficient to see if the difference is enough to be meaningful for any practical purpose. This hypothesis test is checking for a difference in medians, so a significant result only implies that at least 50% of the time the new model is outperforming the base model, which isn't a very high bar to reach. The effect size measures are more effective in communicating the degree of the difference. This example has a Cohen's d of 0.4, which is usually considered a relatively small effect, which agrees with a qualitative comparison of the distribution in Figure 15, where the distributions have considerable overlap.

Cohen's d is not the most intuitive measure of the practical difference between modelling decisions. I believe the more useful measure for the comparison of two distributions is the probability of superiority. The probability of superiority in this case is the probability that a random sample from the energy balance model will outperform a random sample from the base UBCWM. The probability of superiority was calculated with confidence intervals estimated using a simple bootstrap method recommended in (Ruscio & Mullen, 2012). This example has a PS of 0.61, with a 95% CI of [0.54, 0.69]. That means that for any given validation data set, there is still a 39% chance that the original UBCWM potential melt algorithm will outperform the energy balance algorithm, and that chance could actually be as high as 46%.

The previous analysis was performed using independent samples for each model, but the statistical power of the comparison can be increased by using paired data. To pair the data, the same set of sampled observations can be used to evaluate the two models, then a comparison can be made between the performance of the different models given the same information. The distributions of results for the potential melt algorithms are shown in Figure 16, this time using the same set of observations for both models.

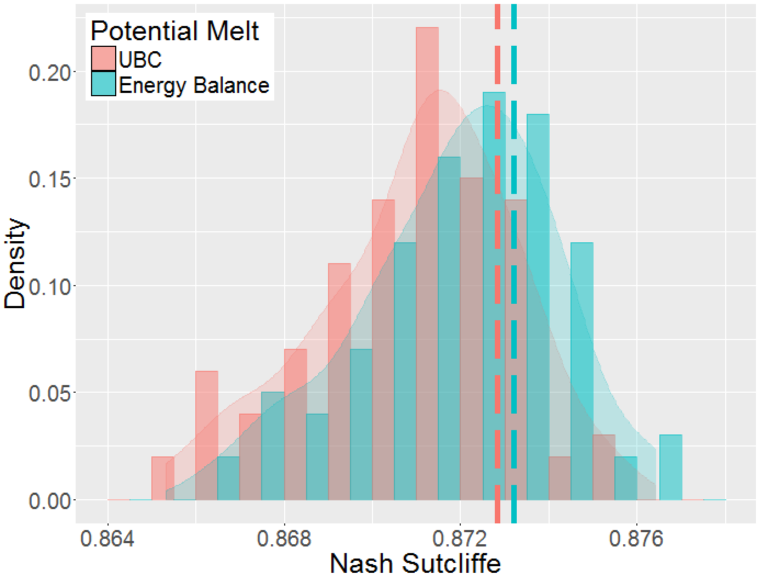

Figure 16: Distributions of Nash Sutcliffe values for two different potential melt algorithms generated using paired samples of uncertain observation hydrographs

The distributions from the paired observations look similar to the unpaired results, but since the results are now paired, the distribution of the difference in performance on each of the 100 sampled hydrographs can now be plotted. The distribution of paired differences in Nash-Sutcliffe performance between the two potential melt algorithms is shown in Figure 17.

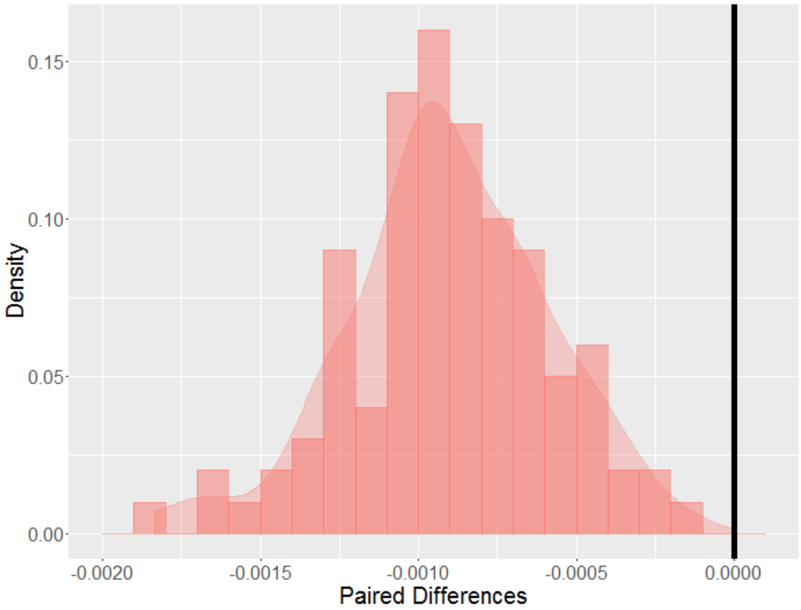

Figure 17: Distribution of paired differences in Nash-Sutcliffe performance

This plot shows that despite the significant overlap in the two distribution, the energy balance consistently outperforms the default UBCWM potential melt algorithm, producing larger Nash-Sutcliffe values for each version of the observation hydrograph sampled. By pairing the results, probability of superiority increased from 61% to 100%. The difference in performance is not very substantial (the mean difference is -0.0009), but this test still provides a more robust comparison with stronger evidence of an improvement than simply using a single deterministic calibration and validation period.

One issue of concern with this methodology is that when the models produce results that are outside the range of the observation uncertainty, then the model that is closer to the observation will produce better results for every realization. This is not necessarily a problem if the model is closer to the observations because it is a better representation of the watershed, but if the difference is caused by an error, such as incorrect precipitation data, then this introduces a biased advantage for one of the models. Care must be taken to ensure that large discrepancies do not unduly influence the results, though the same could be said about the standard single validation method.

When considering changes to individual process algorithms in a model, the results can tend to be quite similar. In some cases, the changes may only affect a specific time period or response to specific conditions, leaving the rest of the hydrograph unchanged, and leads to similar performance scores on integrated measures such as Nash Sutcliffe. For instance, the potential melt differences tested will only have an impact while there is snow to be melted, so the results for the summer months are unaffected. The similarity in the hydrographs results in a strong correlation between the results of the models when paired by validation hydrograph, because if the uncertainty brings the observed hydrograph closer to one of the models, it also tends to move it closer to the other. In cases such as this, the hydrograph uncertainty plays a more important role in overall performance than the model decision. The correlation in model results for the potential melt and infiltration changes are shown in Figure 18. The correlation is stronger in the potential melt case because that change produces a less significant change in the output, especially since the differences are mainly isolated to the winter months. This correlation is related to the problem previously discussed, and may be an indication of bias in the results.

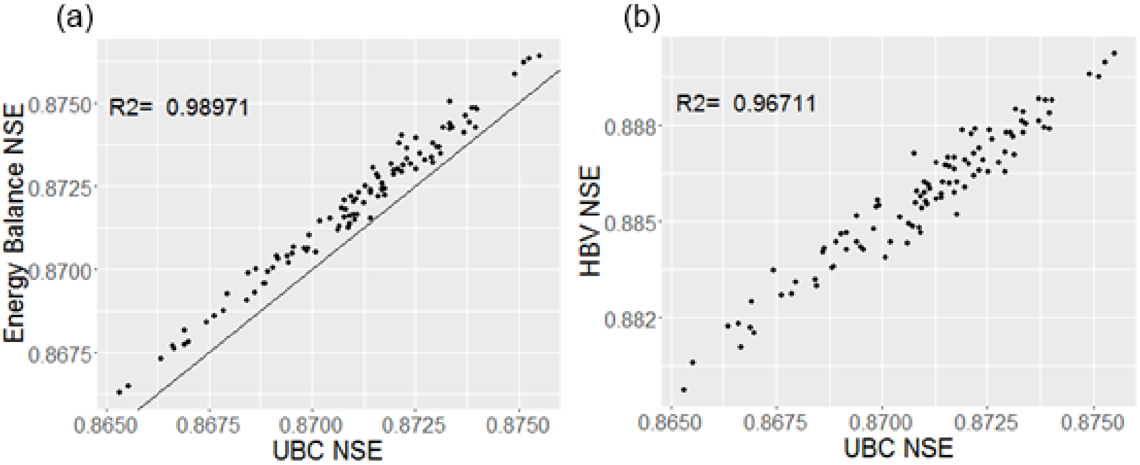

Figure 18: Scatter plot showing different degrees of correlation in paired results from potential melt (a) and infiltration (b) models

To demonstrate that the correlation of the results is due to the similarities between the models, a comparison between the base UBCWM and a completely different model was used. For this example the Raven version of the GR4J (C. Perrin, Michel, & Andréassian, 2003) model was calibrated for the Alouette basin. The GR4J model was set up to use the same elevation band configuration as the UBCWM, and the same gauge corrections, but the processes are modelled very differently. The purpose was to examine the correlation in performance between the models based on the validation data, so not much effort was put into the calibration of the GR4J model and the performance is considerably worse. Nevertheless, the results shown in Figure 19 effectively show that there is much less correlation when there are fewer similarities between the models being compared. The R^2 correlation value of 0.755 between the GR4J and UBCWM models is much lower than the value of 0.967 found for the two different infiltration models.

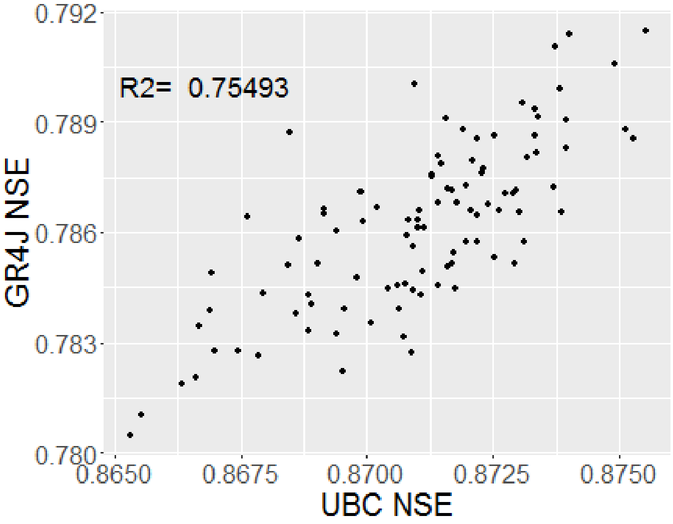

Figure 19: : Scatter plot showing the correlation in paired results between the UBCWM and GR4J model.

4.3 Including Uncertainty in Calibration Data

The previous section showed that while adding uncertainty to the validation data allows for hypothesis testing to be performed and algorithms to be distinguished more effectively than using purely deterministic models, the results are highly correlated and depend on the single deterministic modelled hydrographs being compared. As discussed in section 2.5.4, the parameters of hydrological models are subject to uncertainty, and there are many different combinations of parameter values that may produce similar results. Automatic calibration procedures are often not able to find a true global optimum because of the complexity of the optimization process, and calibrations can give different results if run using different starting values, or if there is a random component to the calibration algorithm. For this reason, comparing single calibrations of models may not be sufficient to say with confidence whether or not a modelling decision resulted in improved performance, even when using hypothesis testing and considering the observation uncertainty in the validation data as in the previous section. The effects of individual model changes may be small enough that the added uncertainty coming from the parameters may still make the two different models effectively indistinguishable.

To account for the parameter uncertainty inherent in the calibration process, models being compared in the section are each calibrated multiple times to different realizations of the inflow hydrograph, and the distributions of performance statistics resulting from the different calibrations are compared using formal hypothesis testing. Each model was calibrated N=100 times using the DDS algorithm with random starting values for each parameter. Importantly, only a single validation data set was used. The results shown in this section summarize the validation performance of the different calibrated parameter sets and algorithms. Note that this test is much more computationally intensive than the previous test requiring N times more model evaluations, where N is the number of calibrations performed per model.

Results from the two different infiltration algorithms are shown in Figure 20. The HBV infiltration model is still clearly outperforming the UBC infiltration ($PS = 0.99$), though the added uncertainty has caused a slight overlap between the distributions. Note that the single calibration is near the center of the distribution for the UBC model, but on the low end for the HBV model, so the difference shown in the previous section is lower than the average difference in distributions.

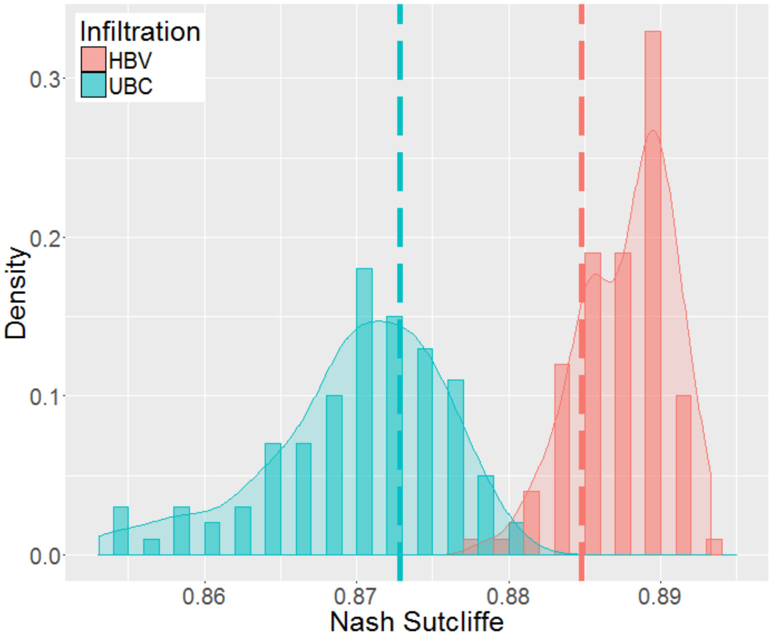

Figure 20: Distributions of validation results from 100 calibrations of the different infiltration models. Lines represent the single calibration result used in the previous section.

So far all the tests performed have had distributions of similar shapes roughly following a normal distribution. Here a test is performed to show the operation of the test under conditions where the models have very different distributions. The distributions of Nash-Sutcliffe results from N=100 calibrations of two different interception algorithms, along with the results from the HBV interception model are shown in Figure 21. The canopy interception clearly outperforms the default UBCWM, with a mean result comparable to the HBV interception model, but the distribution for the canopy model is much more spread out. The canopy model adds four extra parameters, and there appears to be a higher degree of parameter uncertainty.

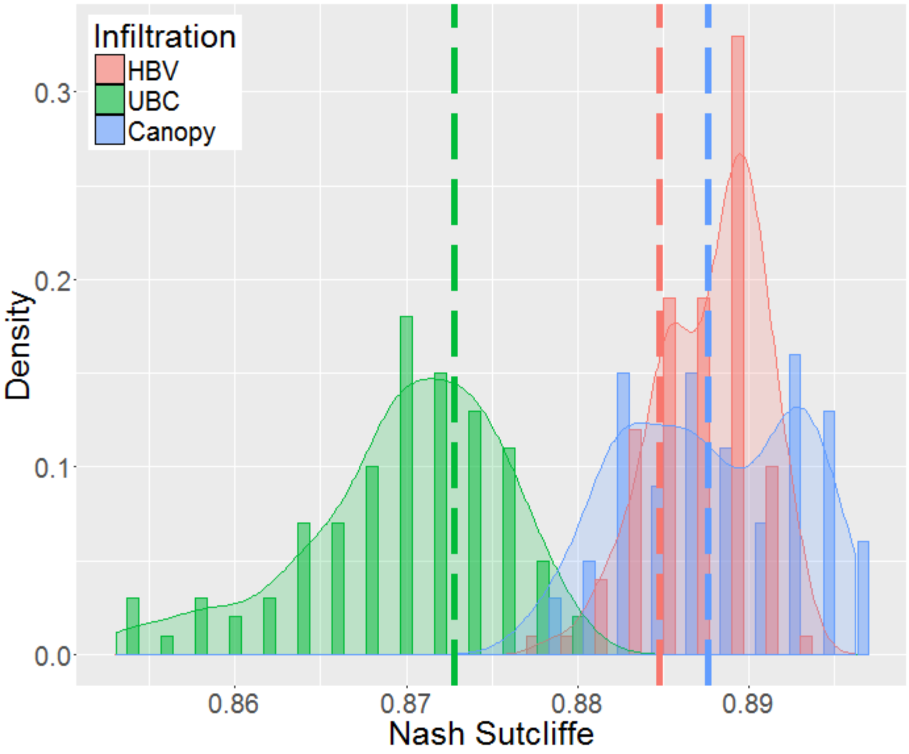

Figure 21: Distributions of validation results from 100 calibrations of the default UBCWM, HBV interception model, and the canopy interception model. Lines represent the single calibration result used in the previous section.

Using the same unpaired Wilcoxon–Mann–Whitney two-sample rank-sum test as in the previous section, a hypothesis test between the HBV and canopy models can be used to determine if one model change tends to outperform the other. The difference in performance distributions of the two algorithms was not found to be statistically significant at the 95% confidence level (Mann-Whitney U 4967, p=0.231 >0.5). While the canopy model has some of the best performing results, this is balanced by the fact that it also has many calibrations that perform worse than the HBV model, so no clear conclusion can be made. It is arguable that the model with the best performing calibration should be used, but the best performance in the selected validation period does not ensure that it will also give the best performance for all periods. In fact, for the calibration period the HBV model has the best performance, as shown in Figure 22.

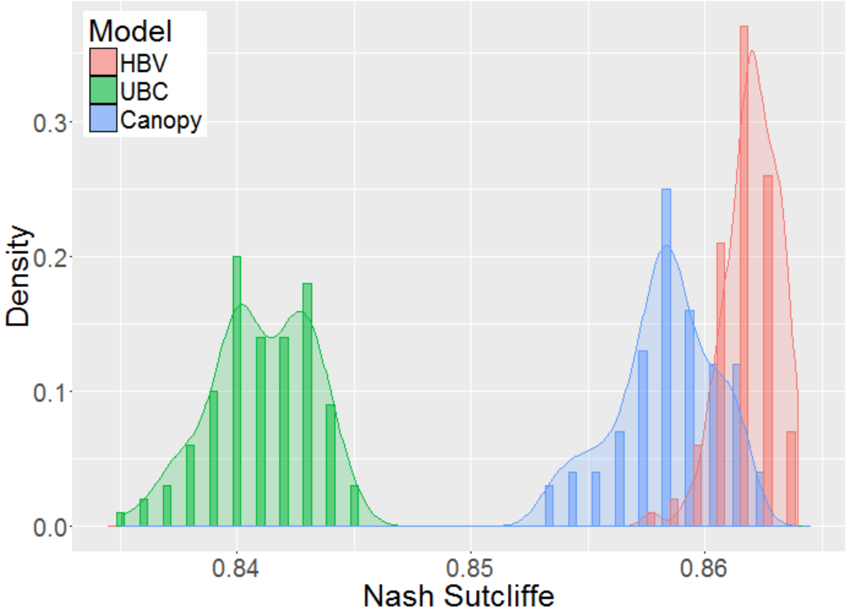

Figure 22: Distributions of calibration results from 100 calibrations of the default UBCWM, HBV infiltration, and the canopy interception models.

Looking once again at the different potential melt models, this time with multiple different calibrations, we see that there remains a large overlap between the distributions shown in Figure 23. We are not able to pair any of the data since these are all independent calibrations, so the unpaired Wilcoxon–Mann–Whitney is again used for the hypothesis testing. With the uncertainty coming from the repeated calibration, these two models no longer have a significant difference (Mann-Whitney U 4490, p=0.213 > 0.05). It seems that the difference found in section 4.2 was in fact an artifact of the individual calibrations, highlighting the problems described at the end of that section.

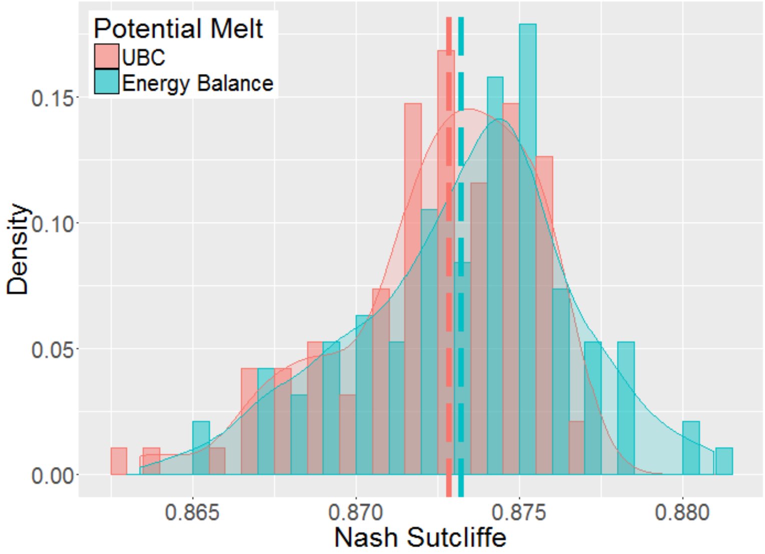

Figure 23: Distributions of validation results from 100 calibrations of the different potential melt models. Lines represent the single calibration result used in the previous section.

4.4 Multiple Calibrations with Uncertain Validation Data

Even though the method of hypothesis testing using uncertain validation data was shown to have problems, the uncertainty introduced by the error in the validation should not necessarily be ignored. The multiple results from the validation uncertainty can be generated much easier than the multiple calibrations used for this section, so the two methods may be used together for more powerful hypothesis tests.

An advantage of including the validation uncertainty is that it helps make the data more normally distributed, which allows for the use of parametric hypothesis testing methods. The distribution of results from individual calibrations is not necessarily normally distributed, which is why a nonparametric test was used in the previous section. Figure 24 shows the distribution of results from N=100 calibrations of the standard UBCWM, along with a Q-Q plot showing that the data does not follow a normal distribution.   The distribution of results from a single calibration evaluated against multiple validation data sets does tend to be normally distributed, and the combination of multiple calibrations and validation uncertainty produces results that appear much closer to a normal distribution, as shown in Figure 25.

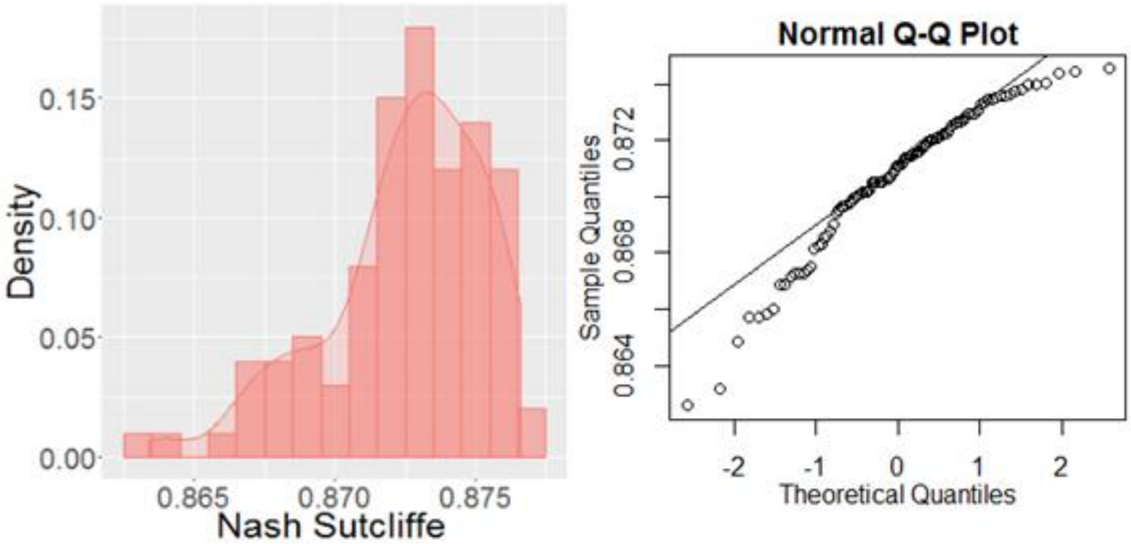

Figure 24: Normality testing on the results from 100 calibration of the UBCWM

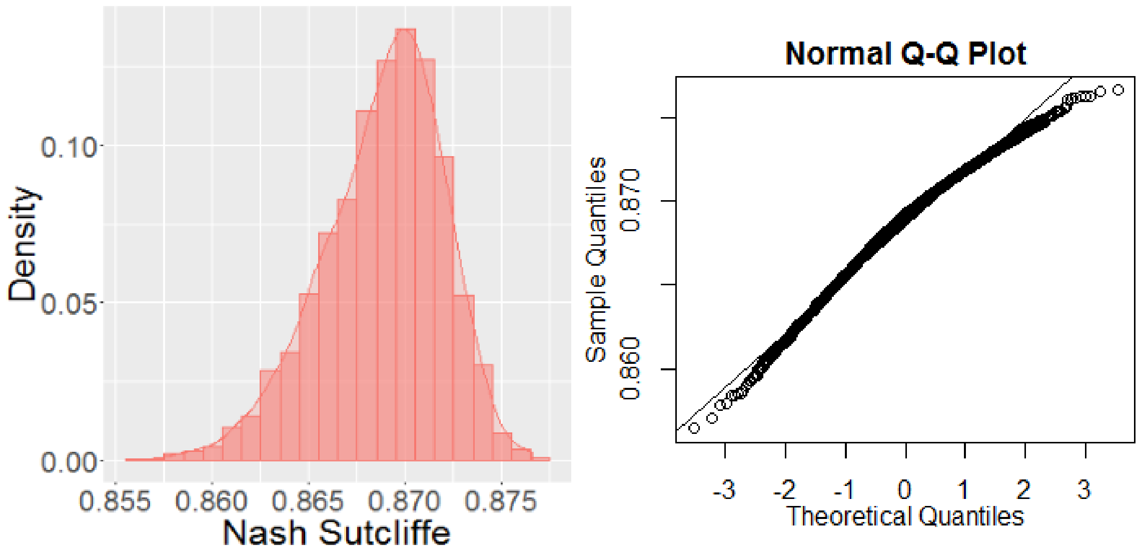

Figure 25: Normality testing on the results from 100 calibration of the UBCWM evaluated against 100 sampled validation hydrographs

A mixed-design analysis of variance (ANOVA) can be used to test for differences between the mean performance of the two models. For this test, the individual calibrations are treated as different subjects, the validation data sets are treated as repeated measures (i.e., a within-subject variable), and the different models are the fixed effects factor (i.e., a between-subject variable). The mixed-design ANOVA is able to able to simultaneously test for significant differences in the means cause by both the within-subject and between-subject variables, as well as any interactions between them. This means that the extra information from the validation data error that can be generated much easier than the multiple calibrations is used to increase the statistical power, and the ANOVA model also takes into account the effect of the different validation data to factor out the potential bias discussed at the end of section 4.2. Increasing the statistical power means that fewer samples are needed to detect differences, so less time needs to be spent on multiple calibrations which can be computationally expensive.

To illustrate the increased ability to detect difference gained by including the validation data uncertainty, tests were run comparing the two different methods of calculating the temperature and precipitation lapse rates. The first test is using 25 separate calibrations of each model, and is not including any uncertainty in the validation data. As can be seen in Figure 26, the data does not appear to be normally distributed, especially in the case of the simple lapse rate. For this reason, the hypothesis testing was performed once again using the nonparametric unpaired Wilcoxon–Mann–Whitney test. This test fails to find a statistically significant difference between the two distributions (Mann-Whitney U 417, p=0.152).

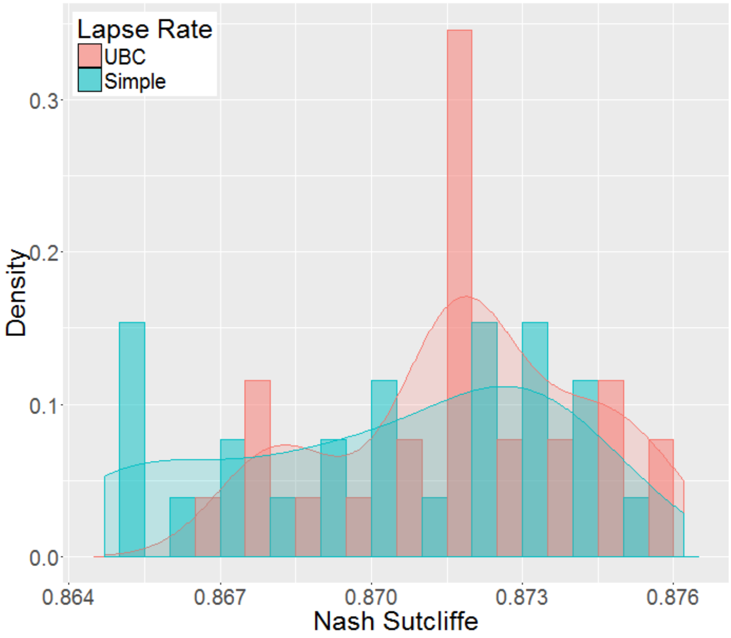

Figure 26: Distributions of validation results from 25 calibrations of the different lapse rate models

The second test takes advantage of the easy-to-generate validation uncertainty results to increase the statistical power. Figure 27 shows the distributions of Nash-Sutcliffe values for the combined multiple calibrations and validation data uncertainty. These distributions appear to be more normally distributed, though the simple lapse rate distribution is not perfectly normal. Despite this deviation from normality, given the larger sample size, it was decided that the data is near enough to normally distributed to justify using the parametric mixed-design ANOVA. The results from the ANOVA analysis show that there is a statistically significant difference between the two models (F(1,48)=3.914, p=0.0488). Even though it is statistically significant, the difference is once again not very substantial with a probability of superiority, $PS = 0.61$ and a 95% CI of [0.59,0.63].

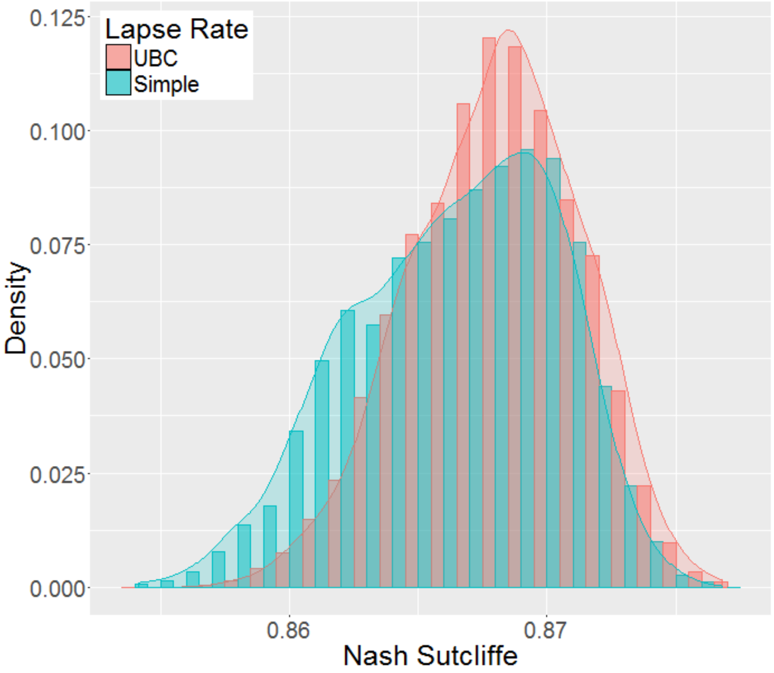

Figure 27: Distributions of validation results from 25 calibrations of the different lapse rate models evaluated using 100 sampled validation data sets

By simply including the validation uncertainty in the results we were able to detect a statistically significant difference that went undetected when using the multiple calibrations alone. This shows that the inclusion of the repeated measures in the form of the validation data uncertainty does increase the statistical power of the test at the expense of only a very modest increase in computation. When the number of separate calibrations is increased to N=100, the multiple calibration test also finds a significant difference (Mann-Whitney U 6346, p=0.001), confirming that there is in fact a significant difference in the models. Although these tests were able to detect a difference, the actual difference in the means is quite small (approximately 0.0016), and as can be seen from Figure 27 there is considerable overlap in the distributions and the probability of superiority is only 0.63 [0.56, 0.70].

4.5 Uncertainty in Forcing and Observation Data

The tests so far have all focused of the uncertainty in the observed hydrographs, but there are many other sources of uncertainty that could be included in this hypothesis testing methodology. For this test, in addition to the inflow observation uncertainty, uncertainty in the precipitation data was also included during calibration.

The precipitation uncertainty was introduced by adding uncertainty to the weights associated with the three precipitation gauges used when estimating the areal average rainfall as was done in (Montanari & Di Baldassarre, 2013). At each timestep the individual weights were selected from a uniform probability distribution in the range of $\pm 20\%$ of their usual values, then the weights were rescaled so that they sum to 1. This method of precipitation uncertainty assumes that the error at each individual gauge is negligible, and only the distribution of the uncertainty throughout the basin is uncertain. The assumption seem reasonable, especially in a mountainous catchment where rainfall distribution is highly heterogeneous.

The results of the two infiltration algorithms for N=100 calibrations each using different realizations of the inflow and precipitation observation data and evaluated against M=100 different validation data sets are shown in Figure 28 (c). The results from the validation uncertainty only, and the calibration and validation uncertainty test are included for comparison in Figure 28 (a) and (b), respectively.

It is clear from the figure that the addition of new sources of uncertainty increases the variance of the distributions, leading to increased overlap making the models more difficult to distinguish clearly. The standard deviation of the HBV performance distributions increases from 0.0016 for the validation uncertainty to 0.0037 for the calibration uncertainty, and finally to 0.0066 for the precipitation uncertainty case.

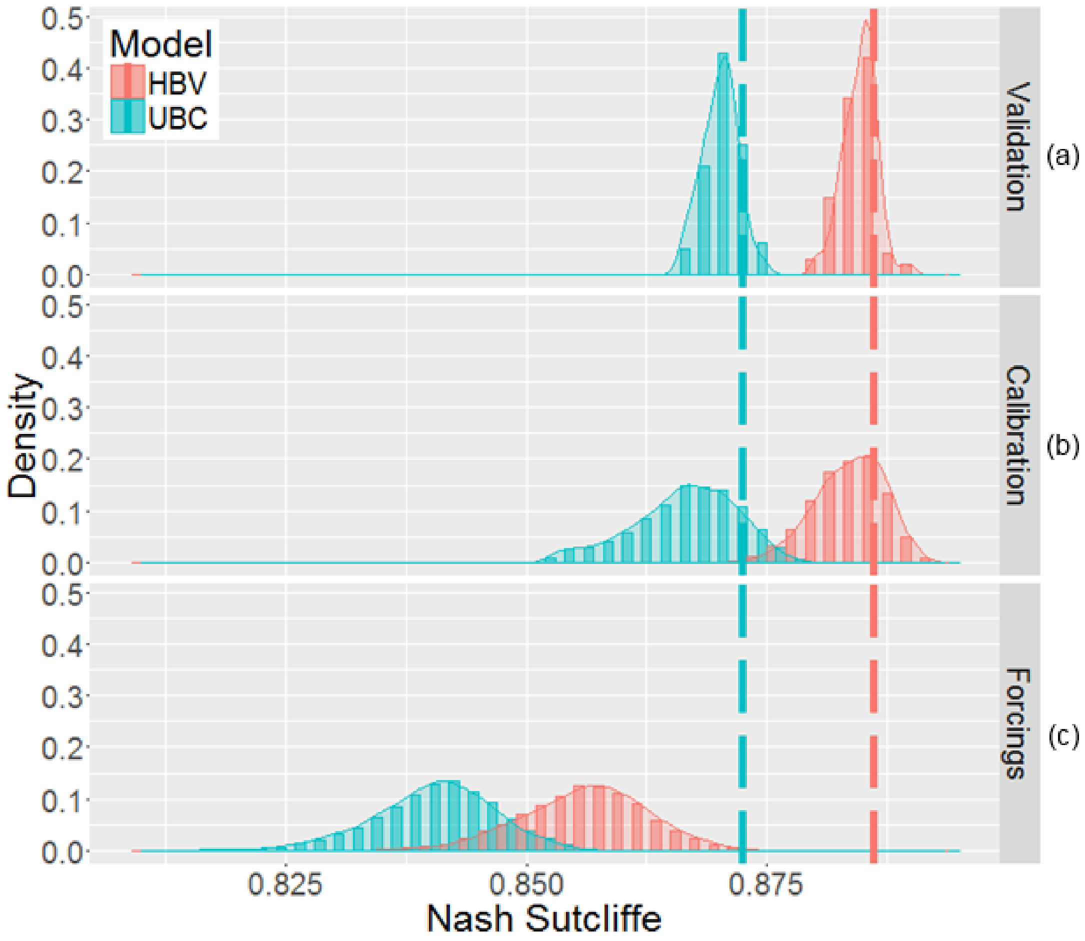

Figure 28: Comparison of performance distributions for the two infiltration algorithms considering (a) validation uncertainty only, (b) calibration and validation uncertainty, and (c) calibration, validation and precipitation uncertainty. The dashed lines show results from a single deterministic calibration to the true observed data.

The absolute performance also decreases with each addition of uncertainty, with the mean Nash-Sutcliffe efficiency of the HBV model going from 0.886 for the single calibration to 0.884 with the validation uncertainty, 0.883 with calibration uncertainty, and finally 0.856 with the inclusion of precipitation uncertainty. The added uncertainty to the inflow hydrographs are relatively small and some realizations perform better than the single calibration, so it seems the method of adding uncertainty is reasonable. The significant drop in performance with the added precipitation uncertainty suggests that the method of introducing uncertainty to the precipitation may not be appropriate, or at least may be too large in this case. Theoretically, the uncertainty should leave the mean unchanged and only increase the variance, unless there is reason to believe the observations are biased and that is reflected in the uncertainty. If the uncertainty method is centered on the observations and is an accurate reflection of the real level uncertainty, you would expect that at least some realizations would bring the observations nearer to the true value resulting in improved performance, so the uncertainty in this case is clearly deficient somewhere.

Despite the problems with the precipitation uncertainty generation methods used here, the purpose was simply to demonstrate that including more uncertainty into the models results in increased variance and therefore a decrease in the ability to distinguish between the models. This objective was accomplished, as demonstrated by the probability of superiority decreasing from $PS = 1$ in the case of only validation uncertainty, to $PS = 0.99$ with the inclusion of calibration data uncertainty, and finally $PS = 0.95$ for the models with precipitation uncertainty. In this case the difference in performance was enough that they are still clearly distinguishable, but if the models were beginning nearer in performance then you would expect them to eventually become indistinguishable given enough uncertainty.

Chapter 5
Diagnostic Model Evaluation

5.1 Introduction

All the previous tests have focused on comparing the overall performance of individual model changes, as measured by the Nash-Sutcliffe efficiency. The goal of this chapter is to expand the hypothesis testing methods from Chapter 4 to use multiple performance measures in order to understand how a modelling decision affects specific aspects of model performance, and to cases involving multiple modelling options to examine the impact of interactions between the decisions.

5.2 Distributions of Multiple Performance Statistics

The Nash-Sutcliffe efficiency was the only model evaluation statistic used in all the tests of Chapter 4, and was also used as the calibration objective function. The use of only a single criterion potentially introduces bias in the results since each performance statistic will tend to put more weight on certain aspects of the hydrographs (Krause, Boyle, & Bäse, 2005). For example, the NSE has an emphasis on higher flow values because the error is squared. Recent work has suggested the need for multiple evaluation criteria, and specifically criteria that are able to diagnose deficiencies in the model structure (Legates & McCabe, 1999). Many studies have begun using hydrological signatures as a method of model of evaluating the ability of a model to match specific aspects of the watershed behaviour (e.g., Coxon et al., 2014; Euser et al., 2013; Ley, Hellebrand, Casper, & Fenicia, 2016; Shafii & Tolson, 2015).

The methods from this thesis should not be seen as an alternative to diagnostic model evaluation methods. In fact, this work complements the use of hydrological signatures by quantifying the significant uncertainty found in each of these signatures, meaning more confidence can be placed in model comparisons that show clear improvements. Since the model signatures can be calculated from the same calibrations of the previous sections, the cost of including multiple evaluation criteria is negligible. This section examines the distributions of several common hydrological performance measures and signatures found in the literature. A brief description of the performance measures is given in Table 3, and the equations can be found in Appendix B. The four flow duration curve based indices are shown in Figure 29.

Table 3: Model performance measures and hydrological signatures

Short Name	Description
NSE	Nash-Sutcliffe efficiency (J. E. Nash & Sutcliffe, 1970) is equivalent to the mean square error normalized by the deviation of the observations from their mean. It is one of the most common hydrological performance measures, and has an optimum value of 1.
PBIAS	Percent bias (Yapo, Gupta, & Sorooshian, 1996) is a measure of the models ability to match the overall volume of flow. The optimal value is 0, with negative and positive values indicating a tendency to underestimate and overestimate flows, respectively.
R2	The coefficient of determination (R^2) is a measure of the dispersion of how much of the observed variance is explained by the model. The optimal value is 1.
Monthly bR2	The coefficient of determination applied to mean monthly flows to test the ability of the model to capture the general seasonal trends.
Variance	Percent bias in the modelled vs. observed flow variance.
AR1	Percent bias in the lag-1 autocorrelation coefficient (Winsemius, Schaefli, Montanari, & Savenije, 2009). This is a measure of the persistence of the watershed, related to the recession curve characteristics.
FHV	Percent bias in the flow duration curve (FDC) high-segment volume (< 2% exceedance probability) (Yilmaz, Gupta, & Wagener, 2008). Tests the reaction of the watershed to the most extreme precipitation.
FMV	Percent bias in the FDC medium high-segment volume (2-20% exceedance probability) (Ley et al., 2016). Tests the variability of the reaction to heavy rainfall
FMS	Percent bias in the FDC mid-segment slope (20-70% exceedance probability) (Yilmaz, Gupta, & Wagener, 2008). Test the general reactivity of the watershed.
FLV	Percent bias in the FDC low-segment volume (>70% exceedance probability) (Yilmaz, Gupta, & Wagener, 2008). Tests the ability to match the baseflow.

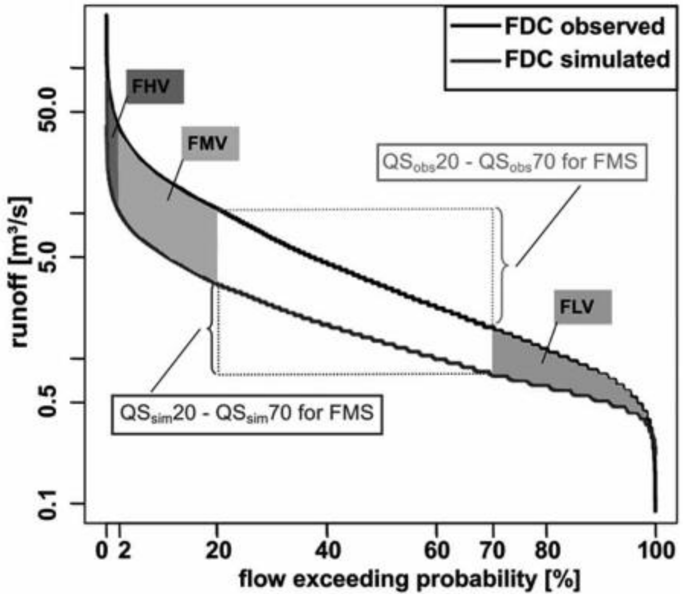

Figure 29: Summary of the FDC based indices (from Ley et al., 2016)

The distributions of the various performance statistics are shown in Figure 30. The results were generated using calibration data uncertainty only, following the same procedure as Section 4.3. A complete detailed analysis of the different performance measures and what they can tell us about model deficiencies is difficult, and not the purpose of this work, but there are several points of interest in the results. The most important thing to note is that despite the fact that the HBV model seems like a large improvement when looking at the Nash-Sutcliffe efficiency alone, the inclusion of other metrics shows that the improvement may not be so clear. In the case of FMS, (which characterizes the skill at replicating the FDC for mid-range flows), the HBV model actually performs worse than the base model, so the change of infiltration algorithm may not be an improvement in all aspects of the model fit. Some other metrics, such as the PBIAS and FMV have distributions with significant overlap where the models cannot be clearly distinguished. Interestingly, for several factors the base UBCWM has a larger variance than the HBV model, suggesting that model may be subject to greater parameter uncertainty. The difference is most pronounced in the AR1 statistic where the UBCWM distribution is nearly flat, while the HBV distribution is much closer to normally distributed, meaning there appears to be much greater uncertainty in the UBCWM, most likely related to the routing parameters.

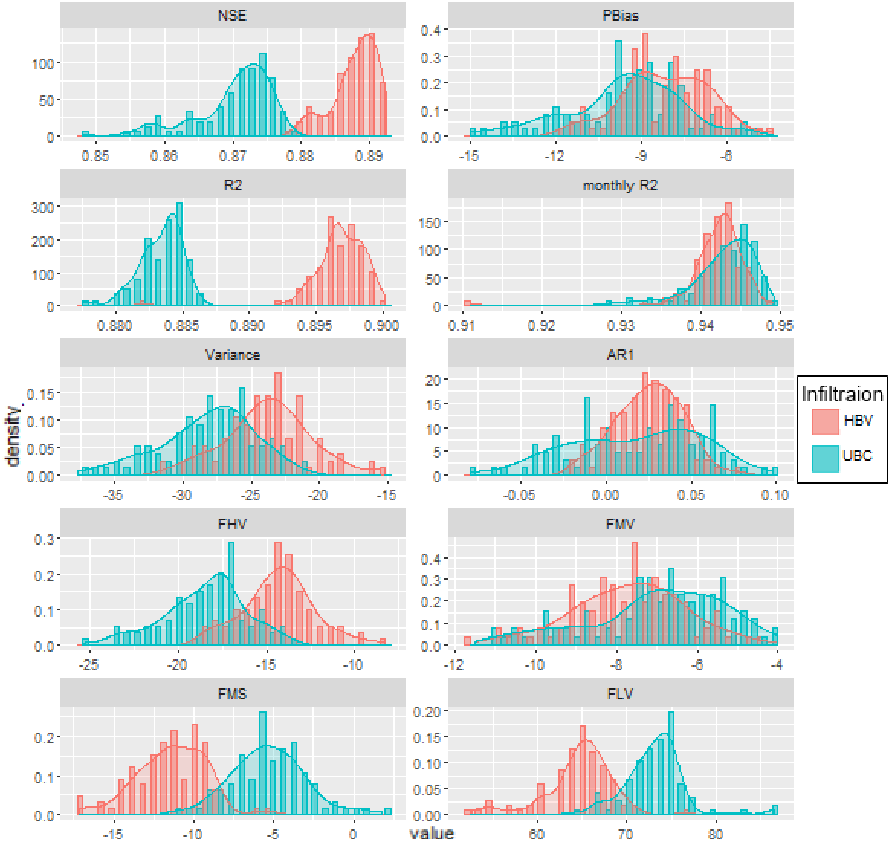

Figure 30: Distributions of multiple performance statistics for different infiltration algorithms

The base model used for this work comes from BC Hydro, and so the model changes here are ostensibly being tested to improve operations in a hydroelectric context. For hydroelectric companies, predicting the total volume of water available for generation is important, typically even more important than accurately modelling the timing of the flow. The mean PBIAS of nearly 10% for both infiltration models may be more error in the volume than what is acceptable, and with significant overlap it is not clear which model should be preferred in terms of their ability to predict the total volume. To examine the impact of the calibration objective function on the distributions of the performance measures, N=100 new calibrations were run using the sum of the root mean square error and the PBIAS as the objective function. To minimize the randomness of the comparison, the same realizations of the calibration data were used in these calibrations as were used in the previous calibrations. The results of the new calibrations with the modified objective function are shown in Figure 31.

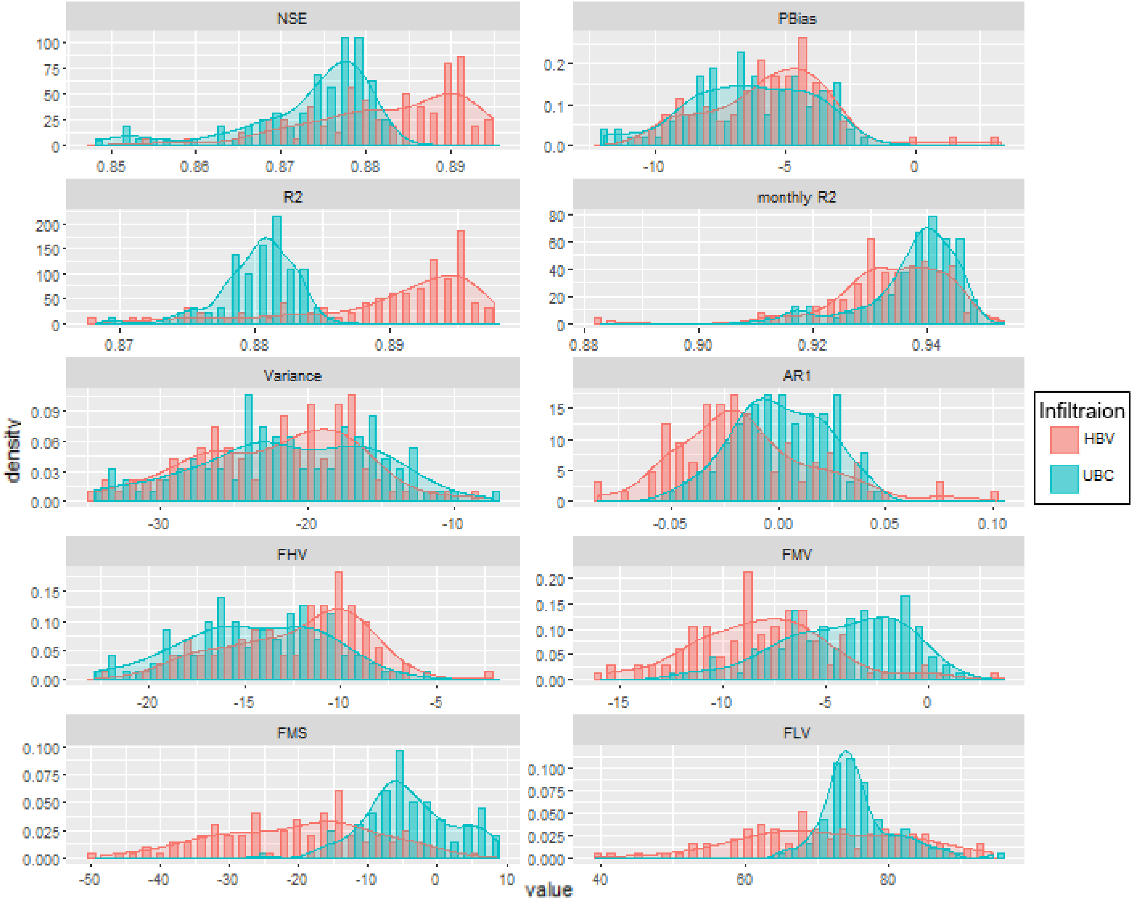

Figure 31: Distributions of multiple performance statistics for different infiltration algorithms, now including PBIAS in the calibration objective function

There are several interesting differences in the results now that PBIAS is being used as a component of the calibration. The first thing to note is both models show a significant performance increase with regards to PBIAS, with median bias increasing from -8.2 to -5.4 and from -9.4 to -6.3 for the HBV and UBC models, respectively. The new calibrations also show a decrease in the variability of the PBIAS result, so including the statistic in the calibration makes its performance more consistent.

The change in the PBIAS statistics are unsurprising and make sense intuitively, but an effect that is less intuitive the change in the Nash-Sutcliffe performance distributions. The median performance for both models is mostly unchanged, with a slight increase for the base UBC model and a slight decrease for the HBV model, but the variance of both models is significantly increase, nearly doubling for the UBC model and increasing by a factor of seven for the HBV model. The increase in parameter uncertainty (represented here by the increased variance in performance) when objectives are added is a known characteristic of multi-objective calibration, depending on how the multi objective calibration is implemented (Efstratiadis & Koutsoyiannis, 2010). This is because the added objectives act as additional degrees of freedom rather than constraints, i.e., overall performance can be improved by improving either objective, so there are more possible parameter sets able to have adequate performance. The interesting thing about these results is that the change in the variance is so much more pronounced for the HBV model compare to the UBC model, implying that there is a large tradeoff between performance in the NSE and PBIAS for the HBV model in many cases.

The change in performance is not only different for the two models, but also differs for each of the performance measures. In many cases the extra calibration objective increases the variability significantly for both models (e.g., NSE and Variance), but in other cases the variability can be greatly reduced, for example the distribution of the AR1 signature for the UBC model becomes much clearer and is centered around 0 showing very good performance. The four FDC signatures give clues about why the variability in the HBV Nash-Sutcliffe distribution is so high; the FMS and FLV signatures show the largest increase in variability, so the biggest differences seem to be in the lower flow values.

The significant uncertainty in the majority of the performance measures makes it difficult to make a clear choice between the two models being considered in this section. If some metrics show an improvement, and others are indistinguishable, then the model can be considered an overall improvement. However, in cases where multiple metrics are distinguishable, but the results are inconsistent in which model performs best, then the selection of the "best" model becomes subjective. Some consideration of the purpose and goals of the models must be done when deciding on which performance measures are most important, and what to use as the objective function. Including too many performance measures will increase the difficulty in distinguishing between models and will result in many models being considered equifinal, though a more advanced multi-objective calibration may reduce this problem.

5.3 Process Interactions

All the previous tests have been methods of comparing any two individual modelling options, but often when examining a model for potential improvements there will be multiple different hypotheses of potential changes which could improve results. These different hypotheses may not be able to be compared individually because changes to any of the model processes could and typically will interact in complex and non linear ways. This section presents a method of simultaneously testing multiple potential model changes to determine which produce significant differences in performance, and assess the possible interactions between the changes.

The method used is the 2^k factorial design of experiment (Montgomery & Runger, 1994). The 2^k factorial design is an experiment with $k$factors where each factor has two levels, in this case the factors are model processes (infiltration, canopy and lapse) and the levels are either the base UBCWM process or the proposed alternative process description. To run this type of experiment each possible combination of the different factor levels needs to be run. Since there is considerable variance in the results of each model caused by uncertainty, each combination should be replicated multiple times to get an accurate estimate of the effects. This type of experiment can be computationally very expensive when considering many potential model changes. In cases where this design is too expensive, there exist fractional factorial designs (Montgomery & Runger, 1994) which require fewer model evaluations at the expense of confounding higher order interactions with other effects.

Here a full 2^3 factorial design was used to examine the effects of changes to the infiltration, interception and lapse rate algorithms, as well as any interactions between the changes. Each of the 8 possible combinations of these algorithms was calibrated N=100 times using different samples from the uncertain calibration hydrographs, each evaluated with M=100 different validation data sets for a total of 10,000 evaluations for each model, as done in Section 4.5. The experimental design is summarized in Table 4, along with the model names for each combination that are used throughout this section. The distributions of Nash-Sutcliffe results from the 8 models are shown in Figure 32. The distributions all follow a similar pattern of being approximately normal but with a slight left skew. There is considerable overlap between many of the distributions and the effects of the individual changes are not obvious, so further analysis is necessary.

Table 4: Design for 2^3 factorial experiment. '-' and '+' denote the default UBCWM algorithm, and the proposed new algorithm, respectively.

Model Name	Infiltration	Interception	Lapse Rate
Base	-	-	-
HBV	+	-	-
Canopy	-	+	-
HBV_Canopy	+	+	-
Lapse	-	-	+
HBV_Lapse	+	-	+
Canopy_Lapse	-	+	+
HBV_Canopy_Lapse	+	+	+

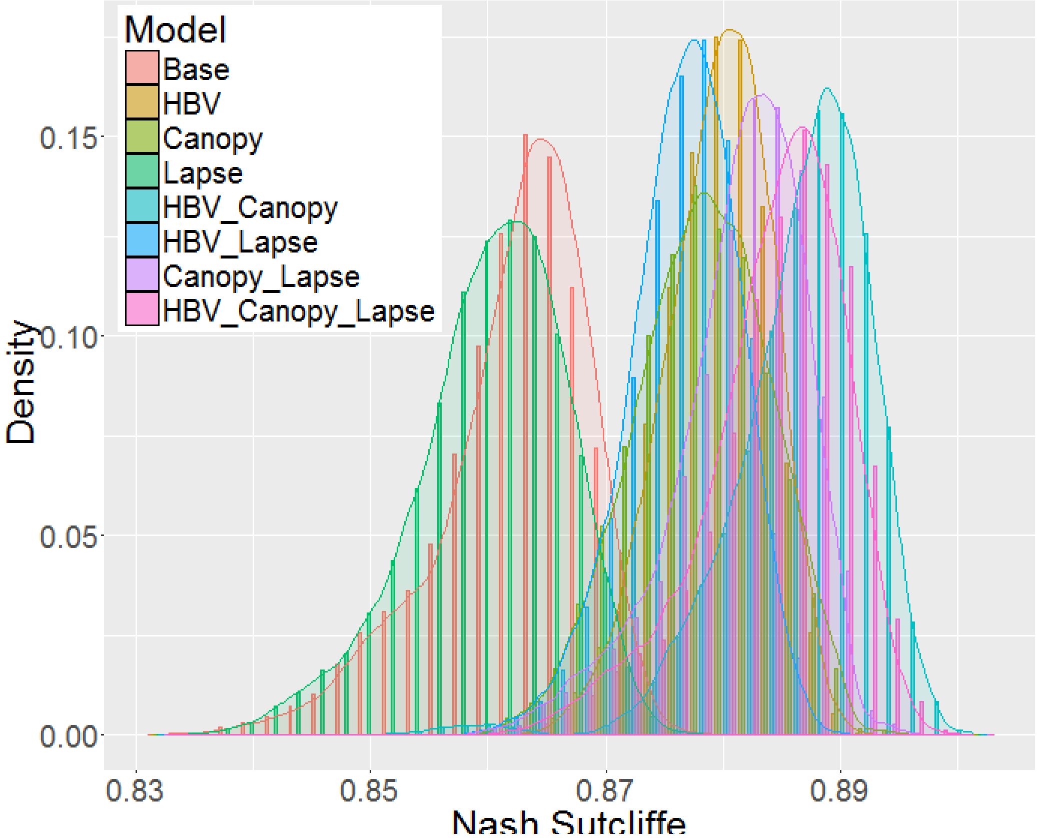

Figure 32: Distributions of Nash-Sutcliffe values for the 8 models of the factorial experiment

An ANOVA was calculated on the results of the factorial experiment with the results shown in Table 5. The ANOVA shows statistically significant differences for all model changes and interactions, but given the sample size used, even a very small difference will be significant. A more useful measure of the differences in performance is the generalized eta squared effect size (η_G^2) which gives a measure of the proportion of the observed variation that can be attributed to each effect. Comparing the effect sizes from the ANOVA, the infiltration and interception have the greatest impact on model results, and there is a notable interaction between the two changes. The effect size of these three factors is an order of magnitude larger than the other effects, so clearly they are the most significant. The modified lapse rate algorithm has very little effect overall, but there are some minor interaction effects with the other two changes.

Table 5: ANOVA results for the factorial experiment

Effect	Degrees of Freedom	Sum of Squares	F	p-value	η_G^2
Infiltration (A)	1	0.02666604	1140.735	1.372E-155	0.590
Interception (B)	1	0.03456317	1478.563	2.58E-183	0.651
Lapse (C)	1	0.00022268	9.525825	0.00209655	0.012
AB	1	0.00624563	267.1791	5.7174E-52	0.252
AC	1	0.00054761	23.42607	1.5607E-06	0.029
BC	1	0.00040032	17.12532	3.8741E-05	0.021
ABC	1	0.00027615	11.81321	0.00061879	0.015
Residuals	792	0.01851			

The ANOVA is useful because it can identify significant factors, and the effect size gives a quantitative measure that can be compared across different studies. When the data is not normally distributed there are nonparametric alternatives. One simple alternative is to use bootstrapping to get estimates of the mean value for each model along with a confidence interval which can be compared to get an idea of the effects of the factors. Here bootstrapping was performed with 100,000 resamples to get mean estimates which are compared graphically. A graphical comparison is more intuitive than the effect size measure, and is a good method of understanding your data even though it does not provide a formal statistic. The statistical significance of an effect can be roughly approximated by comparing the confidence intervals. If there is an overlap in the confidence intervals, then the difference in means is likely not statistically significant. The sample size is large enough here that the confidence interval is small and there is little overlap.

The main effects of each individual modelling decision are shown in Figure 33. It is clear from the plots that both infiltration and interception have a strong effect on the model performance and the magnitude of the effect is similar. The lapse rate does have a small impact on the results, but the difference is much less than the effects of the other two changes. Note that the conclusions from the graphical analysis match the results of the ANOVA.

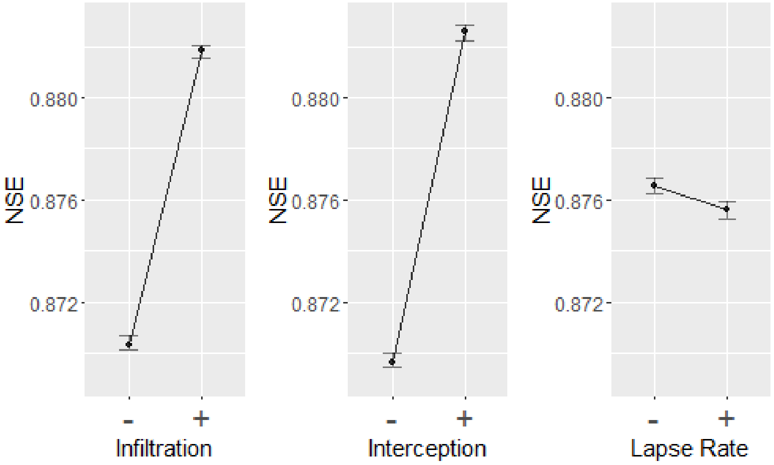

Figure 33: Main effects of the factorial experiment

The two-way interactions are plotted in Figure 34. A two-way interaction effect is the change in the effect of one variable over the levels of the other variable. The interaction effects can be interpreted from the plot as the difference in slope between the two lines. There appears to be a significant interaction between infiltration and interception since the increase in performance from changing the infiltration algorithm is considerably smaller when using the canopy interception rather than the base UBCWM. The other two interaction show a small difference, but overall the line appear quite similar. Once again the results from the graphical analysis match the ANOVA results.

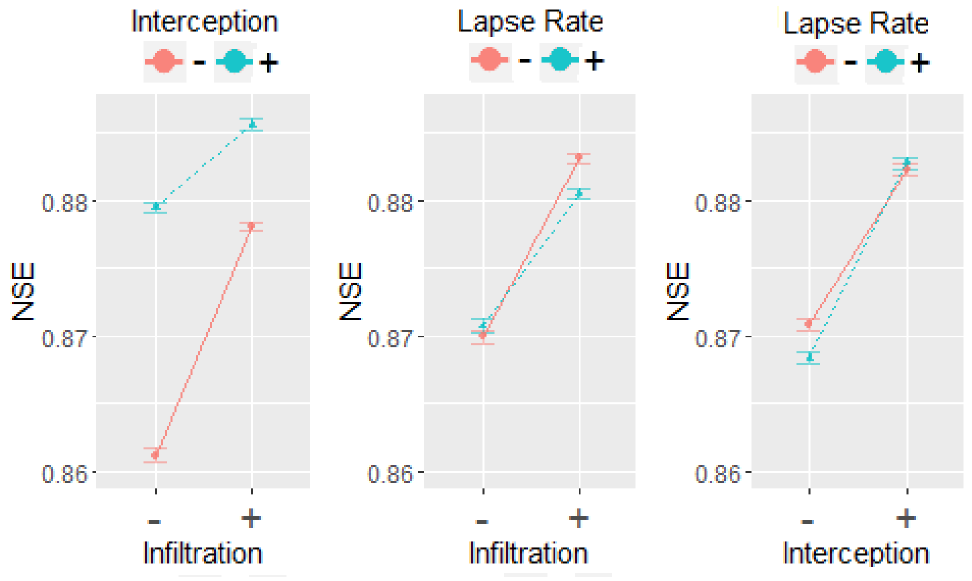

Figure 34: Two way interaction effects of the factorial experiment

The result of the three-way interaction is shown in Figure 35. There is a three-way interaction if the two-way interactions differ depending on the level of a third variable. The two plots in Figure 35 appear somewhat similar, so the three-way interaction effect is likely not very large, which is again in agreement with the ANOVA.

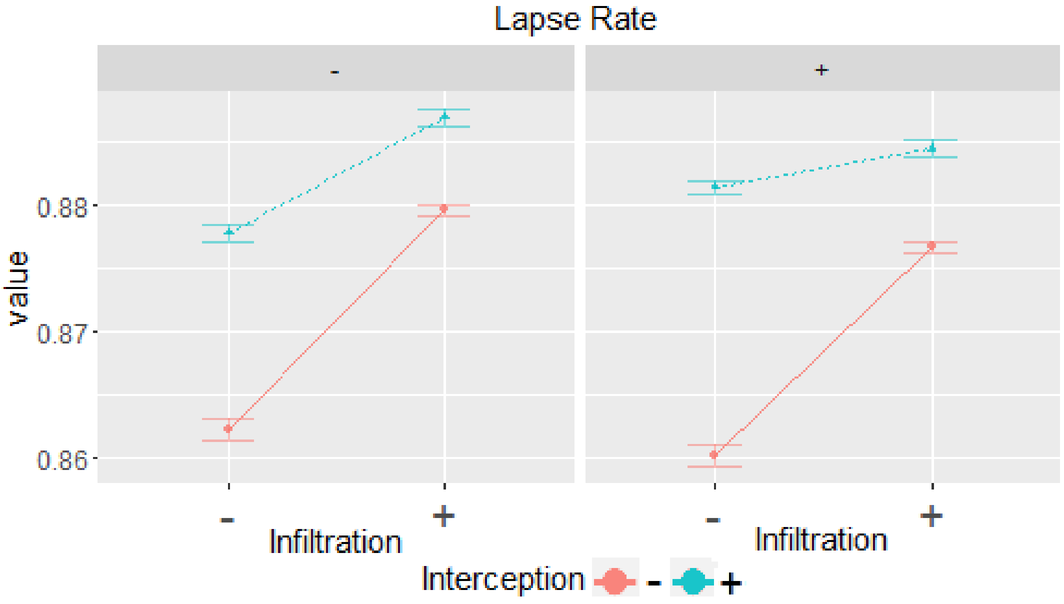

Figure 35: three way interaction effects of the factorial experiment

The ANOVA results can identify which model changes have the largest effect, but this is focused on the means of the model performance distributions, so once again we can follow this with a calculation of the probability of superiority. PS is limited to comparing two models, so a pair-wise comparison was done for the effects found to be important from the ANOVA. First, the HBV and Canopy models were each compared to the base model, resulting in a PS of 0.993 and 0.977, respectively, so both appear to be clear improvements to the base model. Canopy model was then compared against the HBV model with a result of $PS = 0.425$, so the HBV model seems to have a slight advantage over the Canopy model, but the difference is not large. Finally, the HBV_Canopy model was compared against the HBV model to see if the combination of both changes is distinguishable from the individual change despite the interaction. The comparison had a result of $PS = 0.759$, so the model with both changes appears to be an improvement, though it didn't outright dominate the HBV model.

The pair-wise comparisons show which models have the best performance, and the degree to which they can be distinguished, but a series of pair-wise comparisons could be done without the ANOVA results. The real benefit of the factorial ANOVA is that understanding why model changes interact can lead to a deeper understanding of model deficiencies. For example, in cases where the combined effects of two model changes is less than would be expected from the two individual effects it implies that the both model changes are improving a similar aspect of the hydrograph.

In the case of the infiltration and interception model changes, both model changes add a new water storage that must be filled before there is a large runoff produced, and both storages drain slowly over time. Introducing this kind of threshold behaviour means that when there is a small amount of precipitation following a dry period, the modelled flow will be reduced. The base model has a tendency to over predict flow from small summer storms, and to under predict the larger flows in the fall and spring, as shown in Figure 36. By reducing the flow in cases where the base model over predicts flows, the overall runoff can be increased so that the larger events produce more flow. Both model changes help reduce the same model deficiency (as shown in Figure 37), so the interaction effect between the two is understandable.

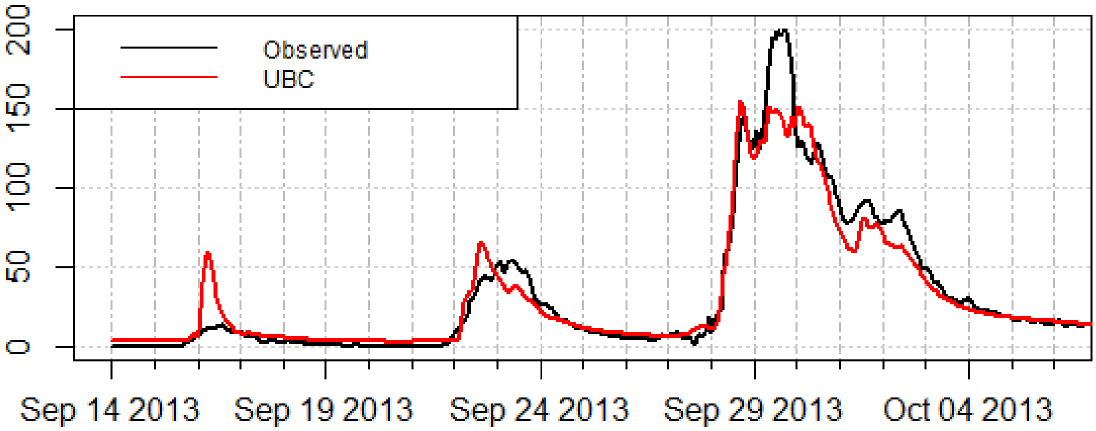

Figure 36: Hydrographs showing over predicted flow Sep. 16, and under prediction Sep. 29

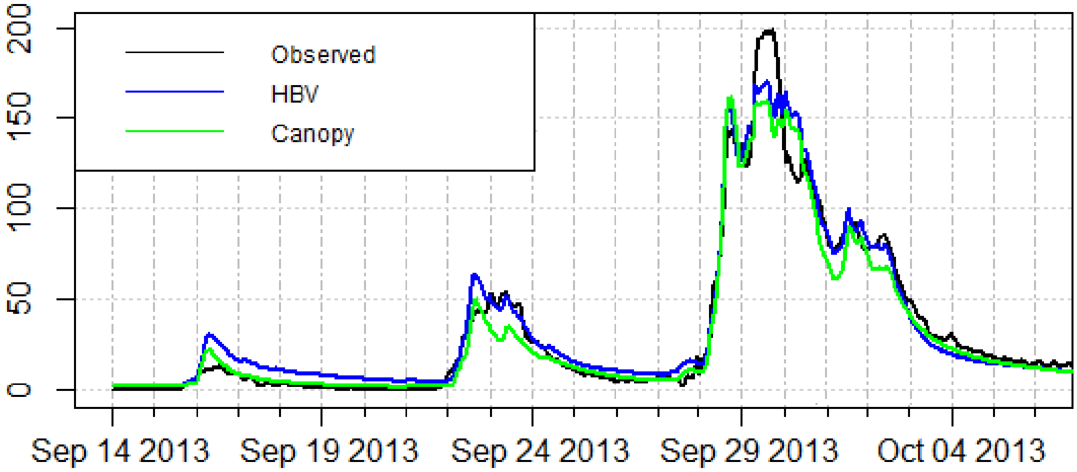

Figure 37: Hydrograph showing significant improvements, especially for the Sep. 16 event

The fact that including both model changes still results in an improvement seems to suggest that the neither change is able to fully correct the problem with the model alone. The change to the infiltration algorithms also has other effects, such as different baseflow, so it's possible a part of the combined improvement is due to improvements in other aspects of the model. This kind of analysis can be supported by the multiple performance statistics of Section 5.2. The result distributions for the variance and FDC low-segment volume are shown in Figure 38. The variance is used here as a proxy for threshold behaviour in the runoff, since increased high flows and decreased low flows should result in a greater variance. The FLV statistic is used to show the effects of the model changes on the baseflow.

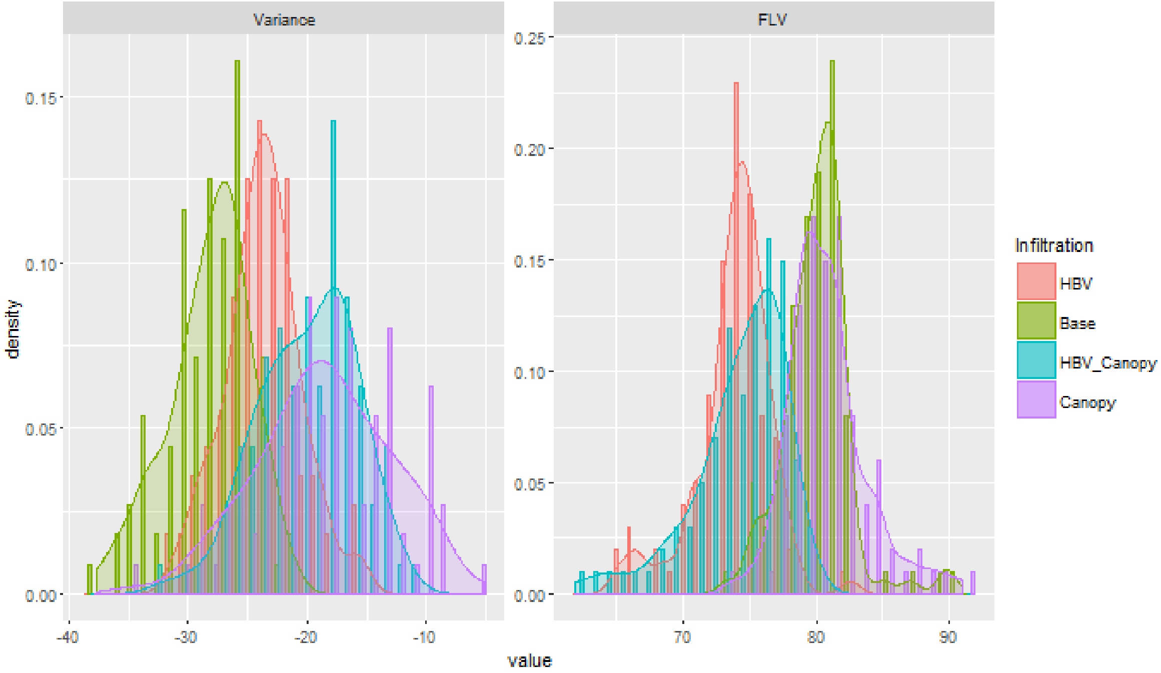

Figure 38: Distributions of Variance and Low Flow Volume for changes to infiltration and interception algorithms, and their interaction

The results in Figure 38 show that the changes to infiltration and interception both increase the variance, with the Canopy model showing the larger increase. In all cases the variance is too low and combining both model changes doesn't lead to further increases, and in fact causes a slight decrease compare to the Canopy model alone. The FLV statistic shows little difference between the base model and the Canopy model, but the inclusion of the change in infiltration algorithm improves model performance. The improved baseflow could in part be the cause of the increase in performance of the Nash-Sutcliffe efficiency when including both model changes.

Chapter 6
Conclusion

The main goal of this study was to demonstrate a new methodology to compare individual modelling decisions that is more rigorous than the current practice. This work provides a method that takes the considerable uncertainty found in hydrological modelling into account by generating probability distributions of performance statistics which are able to be compared using statistical hypothesis testing. The user can decide on the acceptable size of the difference to consider models distinguishable, whether it is any statistically significant difference in the mean performance, or some probability of superiority.

It was also shown that increasing the amount of uncertainty being considered results in more variance in the performance of the models, resulting in greater overlap in the performance distributions and thus a greater difficulty in distinguishing between modelling decisions. The appropriate level of uncertainty to be considered in the methodology will depend on the watershed and model being used, but ideally all significant sources of uncertainty in observations and model forcings should be included.

This study went on to demonstrate that the new testing procedure can be used in conjunction with diagnostic model evaluation methods. A diagnostic approach to model evaluation is becoming increasingly popular for its ability to examine specific aspects of a hydrological model, but it was shown that many of the diagnostic signatures are subject to a considerable uncertainty. The procedure presented in this work allows for the uncertainty in the signatures to be estimated and therefore more confidence can be placed on model comparisons.

Finally, it was shown how the testing procedure could be extended to consider multiple modelling decisions simultaneously to assess the effects of interactions between modelling components. The interactions can be examined using the diagnostic methods to understand the cause of model differences and which areas of the model are in need of improvement.

6.1 Limitations

Despite the benefits of the proposed procedure, there are some limitations present in this work.

First, the procedure proposed here can be computationally expensive since each model being test needs to be calibrated many times. If there are multiple potential modelling decisions being tested for interactions using the procedure presented, the number of model evaluations required can quickly become unmanageable; however, this problem could be alleviated in several ways. One thing that can help solve the issue of computational cost is to reduce the number of calibrations used for each model. The low-cost method of evaluating the models with validation data uncertainty was shown to increase the statistical power of the model comparisons, potentially allowing for fewer calibrations to be used. Another method to reduce the computational cost is to reduce the cost of each calibration by using a more efficient automatic calibration algorithm, or simply reducing the number of model evaluations used in each calibration. For example, the DDS algorithm used for calibration in this work is designed to scale the search based on the number of available model evaluations, so this number could potentially be reduced with a limited loss of performance. If alternatives are being tested for practical applications, the calibration procedure used should be comparable to what will be used in practice.

Another potential limitation of the proposed procedure is that it relies on the assumption that the method of generating different realization of the observations is an accurate representation of the true uncertainty. If this assumption is violated, the calculated distributions are no longer able to give a realistic comparison of results. This means that the procedure is limited to use in watersheds where the observation uncertainty can be characterized with a reasonable degree of confidence.

Another issue, not with the proposed procedure, but with the work presented in this thesis is related to the limited data used. All the results were based on relatively short calibration and validation periods. Only five years of data was available for use at an hourly timestep in the Alouette basin. A short period means that results will be influenced more heavily by individual events, and the uncertainty will of those events have a larger effect. In addition, the results of only a single watershed were included. It is likely that results from other watershed will vary in the performance of the tested model hypothesis, and in the response of the models to the uncertainty. Therefore, the results presented here may only be representative of the Alouette basin from the years 2011-2015. However, the methodology will still be valid for other sites and model types.

6.2 Future work

Future work for this research should focus on further development of the diagnostic model assessment techniques together with the examination of the interaction effects between modelling decisions. The purpose of this study was to examine individual modelling decisions in isolation, and the majority of this work focused on overall performance using only a single performance measure, but Chapter 5 showed that additional diagnostic performance measures and model interaction are able to give a better understanding of why the performance may differ. The diagnostic evaluation and model interactions add much more complexity to the problem and need to be studied in more detail than what was done in this work.

More work should also be done on exploring other sources of uncertainty and how they could be included in the procedure presented in this thesis. Potentially major sources of uncertainty are the time periods used for calibration and validation. This work could be expanded to generate performance distributions from multiple calibration and validation periods, and potentially also considering multiple calibration objective functions. Each additional source of uncertainty would likely make it more difficult to distinguish between modelling decisions, but also adds more confidence to the results when an improvement is found.

Bibliography

Andréassian, V., Valéry, A., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., et al. (2009). Crash tests for a standardized evaluation of hydrological models. *Hydrology and Earth System Sciences, 13*(10), 1757-1764.

Arsenault, R., Poulin, A., Côté, P., & Brissette, F. (2013). Comparison of stochastic optimization algorithms in hydrological model calibration. *Journal of Hydrologic Engineering, 19*(7), 1374-1384.

Baldassarre, G. D., & Montanari, A. (2009). Uncertainty in river discharge observations: A quantitative analysis. *Hydrology and Earth System Sciences, 13*(6), 913-921.

BC Hydro. (2009). Alouette Project Water Use Plan. Retrieved from https://bchydro.com/content/dam/hydro/medialib/internet/documents/environment/pdf/wup_-_alouette_wup.pdf

Bergström, S., & Singh, V. (1995). The HBV model. *Computer Models of Watershed Hydrology., ,* 443-476.

Berrada, F., Bennis, S., & Gagnon, L. (1996). Validation des données hydrométriques par des techniques univariées de filtrage. *Canadian Journal of Civil Engineering, 23*(4), 872-892.

Beven, K. J. (2011). *Rainfall-runoff modelling: The primer* John Wiley & Sons.

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources, 16*(1), 41-51.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology, 320*(1-2), 18-36.

Beven, K., & Binley, A. (1992). The future of distributed models - model calibration and uncertainty prediction. *Hydrological Processes, 6*(3), 279-298.

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes, 9*(3-4), 251-290.

Boyle, D. P., Gupta, H. V., & Sorooshian, S. (2000). Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research, 36*(12), 3663-3674.

Butts, M. B., Payne, J. T., Kristensen, M., & Madsen, H. (2004). An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. *Journal of Hydrology, 298*(1), 242-266.

Clark, M. P., & Kavetski, D. (2010). Ancient numerical daemons of conceptual hydrological modeling: 1. fidelity and efficiency of time stepping schemes. *Water Resources Research, 46*(10)

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research; Water Resour.Res., 47*

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015). A unified approach for process-based hydrologic modeling: 1. modeling concept. *Water Resources Research, 51*(4), 2498-2514.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research, 44*, W00B02.

Clarke, R. T. (2008). Issues of experimental design for comparing the performance of hydrologic models. *Water Resources Research, 44*(1)

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., & Clark, M. (2014). Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. *Hydrological Processes, 28*(25), 6135-6150.

Craig, J. R., W. Jenkinson, G. Jost, M. Serrer, A. P. Snowdon, N. Sgro, M. Shafii, and B. A. Tolson. Flexible watershed simulation with the Raven hydrological modeling framework. Submitted April 2016 to Environmental Modelling and Software.

Dawson, C. W., Abrahart, R. J., Shamseldin, A. Y., & Wilby, R. L. (2006). Flood estimation at ungauged sites using artificial neural networks. *Journal of Hydrology, 319*(1), 391-409.

Deng, C., Liu, P., Guo, S., Wang, H., & Wang, D. (2015). Estimation of nonfluctuating reservoir inflow from water level observations using methods based on flow continuity. *Journal of Hydrology, 529*, 1198-1210.

Dingman, S. L. (2015). *Physical hydrology* Waveland press.

Duan, Q., Schaake, J., Andreassian, V., Franks, S., Goteti, G., Gupta, H., et al. (2006). Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology, 320*(1-2), 3-17.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap* CRC press.

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: A review. *Hydrological Sciences Journal–Journal Des Sciences Hydrologiques, 55*(1), 58-78.

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. motivation and theoretical development. *Water Resources Research, 47*, W11510.

Gharari, S., Hrachowitz, M., Fenicia, F., & Savenije, H. H. G. (2013). An approach to identify time consistent model parameters: Sub-period calibration. *Hydrology and Earth System Sciences, 17*(1), 149-161.

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research, 34*(4), 751-763.

Gupta, H. V., Wagener, T., & Liu, Y. Q. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes, 22*(18), 3802-3813.

Hoskin, T. (2012). Parametric and nonparametric: Demystifying the terms. *Mayo Clinic,*

Hublart, P., Ruelland, D., Dezetter, A., & Jourde, H. (2015). Reducing structural uncertainty in conceptual hydrological modelling in the semi-arid andes. *Hydrology and Earth System Sciences, 19*(5), 2295-2314.

Jajarmizadeh, M., Harun, S., & Salarpour, M. (2012). A review on theoretical consideration and types of models in hydrology. *Journal of Environmental Science and Technology, 5*(5), 249.

Jakeman, A., & Hornberger, G. (1993). How much complexity is warranted in a rainfall-runoff model? *Water Resources Research, 29*(8), 2637-2649.

Jones, J. A. A. (2014). *Global hydrology: Processes, resources and environmental management* Taylor \& Francis.

KLEMES, V. (1986). Dilettantism in hydrology - transition or destiny. *Water Resources Research, 22*(9), S177-S188.

Krause, P., Boyle, D., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences, 5*, 89-97.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863.

Lawrence, M. A. (2015). *Ez: Easy analysis and visualization of factorial experiments*

Legates, David R., and Gregory J. McCabe. "Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation."Water resources research 35.1 (1999): 233-241.

Ley, R., Hellebrand, H., Casper, M. C., & Fenicia, F. (2016). Comparing classical performance measures with signature indices derived from flow duration curves to assess model structures as tools for catchment classification. *Hydrology Research, 47*(1), 1-14.

Madsen, H., Wilson, G., & Ammentorp, H. C. (2002). Comparison of different automated strategies for calibration of rainfall-runoff models. *Journal of Hydrology, 261*(1), 48-59.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics,* , 50-60.

McGraw, K. O., & Wong, S. (1992). A common language effect size statistic. *Psychological Bulletin, 111*(2), 361.

McMillan, H. K., Clark, M. P., Bowden, W. B., Duncan, M., & Woods, R. A. (2011). Hydrological field data from a modeller's perspective: Part 1. diagnostic tests for model structure. *Hydrological Processes, 25*(4), 511-522.

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes, 26*(26), 4078-4111.

Micovic, Z., & Quick, M. C. (2009). Investigation of the model complexity required in runoff simulation at different time scales/Etude de la complexité de modélisation requise pour la simulation d'écoulement à différentes échelles temporelles. *Hydrological Sciences Journal, 54*(5), 872-885.

Montanari, A., & Di Baldassarre, G. (2013). Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty. *Advances in Water Resources, 51*, 498-504.

Montgomery, D. C., & Runger, G. C. (1994). *Applied statistics and probability for engineers* Wiley.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE, 50*(3), 885-900.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology, 10*(3), 282-290.

Nash, J. (1957). The form of the instantaneous unit hydrograph. *International Association of Scientific Hydrology, Publ, 3*, 114-121.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434.

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology, 279*(1), 275-289.

Perrin, C., Michel, C., & Andréassian, V. (2001). Does a large number of parameters enhance model performances. *Journal of Hydrology, 242*(3-4), 275-301.

Pomeroy, J. W., Gray, D. M., Brown, T., Hedstrom, N. R., Quinton, W. L., Granger, R. J., et al. (2007). The cold regions hydrological process representation. *Hydrological Processes, 21*(19), 2650-2667.

Quick, M., & Pipes, A. (1977). UBC WATERSHED MODEL/Le modèle du bassin versant UCB. *Hydrological Sciences Journal, 22*(1), 153-161.

Quick, M., & Singh, V. (1995). The UBC watershed model. *Computer Models of Watershed Hydrology., ,* 233-280.

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*(1-4), 27-60.

Refsgaard, J. C., Van der Sluijs, Jeroen P, Brown, J., & Van der Keur, P. (2006). A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources, 29*(11), 1586-1597.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research, 46*(5)

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research, 47*(2), 201-223.

Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627-1639.

Schroeter, H. (1989). GAWSER training guide and reference manual. *School of Engineering, University of Guelph, Ontario, Canada,*

Seck, A., Welty, C., & Maxwell, R. M. (2015). Spin-up behavior and effects of initial conditions for an integrated hydrologic model. *Water Resources Research, 51*(4), 2188-2210.

Shaw, E. M., Beven, K. J., Chappell, N. A., & Lamb, R. (2010). *Hydrology in practice, fourth edition* Taylor \& Francis.

Singh, V. (1997). Effect of spatial and temporal variability in rainfall and watershed characteristics on stream flow hydrograph. *Hydrological Processes, 11*(12), 1649-1669.

Sivakumar, B. (2008). Dominant processes concept, model simplification and classification framework in catchment hydrology. *Stochastic Environmental Research and Risk Assessment, 22*(6), 737-748.

Team, R. C. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tetzlaff, D., & Uhlenbrook, S. (2005). Significance of spatial variability in precipitation for process-oriented modelling: Results from two nested catchments using radar and ground station data. *Hydrology and Earth System Sciences Discussions, 9*(1/2), 29-41.

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research, 43*(1)

Vache, K., & McDonnell, J. (2006). A process-based rejectionist framework for evaluating catchment runoff model structure. *Water Resources Research, 42*(2), W02409.

Vrugt, J. A., Ter Braak, C., Diks, C., Robinson, B. A., Hyman, J. M., & Higdon, D. (2009). Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation, 10*(3), 273-290.

Weston, Scott. (2011). Technical Report - A Review of Inflow Quality Control Procedures at BC Hydro. *Internal BC Hydro report*.

Wheater, H. S. (2002). Progress in and prospects for fluvial flood modelling. *Philosophical Transactions.Series A, Mathematical, Physical, and Engineering Sciences, 360*(1796), 1409-1431.

Winsemius, H., Schaefli, B., Montanari, A., & Savenije, H. (2009). On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research, 45*(12)

Xu, C. (2002). Hydrologic models.

Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1996). Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. *Journal of Hydrology, 181*(1), 23-48.

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research, 44*(9)

Young, P. (2003). Top-down and data-based mechanistic modelling of rainfall–flow dynamics at the catchment scale. *Hydrological Processes, 17*(11), 2195-2217.

Zhang, X., Hoermann, G., Gao, J., & Fohrer, N. (2011). Structural uncertainty assessment in a discharge simulation model. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques, 56*(5), 854-869.

Appendix A
Calibrated Parameters and Ranges

This appendix provides a summary of all the parameters included in the model calibrations, along with the calibration ranges used, and which models include the parameters.

Parameter	Description	Max	Min	Models
A0TLZP	Temperature lapse with precipitation	0	7	All except simple lapse
A0TLXM	Lapse rate of maximum daily temperature	5	20	All except simple lapse
A0TLNM	Lapse rate of minimum daily temperature	0	10	All except simple lapse
P0TEDL	Lapse rate of maximum temperature range	0	20	All
P0ALBMIN	Albedo of very deep and aged snowpack	0.1	0.3	All
P0PERC	Groundwater percolation (mm/day)	2	40	All
P0RREP	Correction factor for rainfall (%)	1	2	All
P0SREP	Correction factor for snowfall (%)	1	2	All
C0ELPT	Elevation of the meteorological station (m)	500	800	All
P0FRTK	Fast runoff rate constant (1/days)	0	24	All
P0IRTK	Interflow rate constant (1/days)	0	24	All except HBV
P0UGTK	Upper groundwater runoff rate constant	0	0.5	All
P0DZTK	Deep groundwater runoff rate constant	0	0.07	All
V0FLAS	Flash flood threshold (mm)	20	80	All except HBV
P0DZSH	Deep zone share (%)	0.1	0.9	All
HBV_Beta	HBV infiltration exponent β	0	100	HBV only
field_cap	Topsoil field capacity	0.5	1	HBV only
R_Icept	Rain interception (%)	0	1	Canopy only
S_Icept	Snow interception (%)	0	1	Canopy only
Drip	Canopy drip proportion	0	0.6	Canopy only
Max_R_Cap	Maximum canopy rain storage capacity	0	75	Canopy only
Max_S_Cap	Maximum canopy snow storage capacity	0	75	Canopy only
Temp_Lapse	Simple linear temperature lapse rate	0	10	Simple lapse only
Precip_Lapse	Simple linear precipitation lapse rate	0	7	Simple lapse only

Appendix B
Performance Measure Equations

This appendix provides the mathematical formulation of the performance measures and hydrological signatures used for model evaluation.

Nash-Sutcliffe efficiency:

$$NSE = 1 - \frac{\sum(Q_o - Q_s)^2}{\sum(Q_o - \overline{Q_o})^2}$$

where Q_o is the observed flow value, and Q_s is the simulated flow value.

Percent Bias:

$$PBIAS = \frac{\sum(Q_s - Q_o)}{\sum Q_o} \times 100$$

Coefficient of Determination:

$$r^2 = \left(\frac{\sum(Q_o - \overline{Q_o})(Q_s - \overline{Q_s})}{\sqrt{(Q_o - \overline{Q_o})^2}\sqrt{(Q_s - \overline{Q_s})^2}}\right)^2$$

Bias in the Variance:

$$Variance = \frac{s_s^2 - s_o^2}{s_o^2} \times 100$$

where s_o^2 is the variance of the observed flow values, and $s_s^2$is the variance of the simulated values.

Bias in the Lag-1 Autocorrelation Coefficient:

$$AR1 = \frac{ar_s - ar_o}{ar_o} \times 100$$

where ar is the lag-1 autocorrelation coefficient calculated as

$$ar = \frac{\sum(Q_t - \bar{Q})(Q_{t+1} - \bar{Q})}{\sum(Q_t - \bar{Q})^2}$$

Bias in the High-Segment Volume:

$$FHV = \frac{\sum_{h=1}^{H}(Q_{sh} - Q_{oh})}{\sum_{h=1}^{H} Q_{oh}} \times 100$$

where h=1,2,...H are the flow indices for flows with exceedance probabilities less than 0.02.

Bias in the Medium Segment Volume:

$$FMV = \frac{\sum_{m=1}^{M}(Q_{sm} - Q_{om})}{\sum_{h=1}^{H} Q_{om}} \times 100$$

where m=1,2,...M are the flow indices for flows with exceedance probabilities between 0.2 and 0.02.

Bias in the Mid-Segment Slope:

$$FMS = \frac{[log(Q_{sm1}) - log(Q_{sm2})] - [log(Q_{om1}) - log(Q_{om2})]}{[log(Q_{om1}) - log(Q_{om2})]} \times 100$$

where $m1$ and $m2$ are the lowest and highest flow exceedance probabilities (0.2 and 0.7 respectively) within the mid-segment of the flow duration curve.

Bias in the Low-Segment Volume:

$$FLV = -1\frac{\sum_{l=1}^{L}[log(Q_{sl}) - log(Q_{sL})] - \sum_{l=1}^{L}[log(Q_{ol}) - log(Q_{oL})]}{\sum_{l=1}^{L}[log(Q_{ol}) - log(Q_{oL})]} \times 100$$

where l=1,2,...L are the flow indices for flows with exceedance probabilities greater than 0.7, L being the index of the minimum flow value.