Assessing the Utility of Hydrologic Model Diagnostics for Decision Support

by

Konhee Lee

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

Waterloo, Ontario, Canada, 2018

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

Theoretical, computational and experimental advances have led to easier access to more complex and robust hydrologic models.

These hydrologic models may be used to support decision making by water managers and stakeholders. Modeler may choose to utilize a various combination of model diagnostics on different hydrologic data available to describe the model performance. The "goodness" of a specific diagnostic may depend on multiple factors (hydrologic complexity of basin, data availability, data used for evaluation, resources spent on model, validation methods, and intended use of model). Through the DCT, which explicitly evaluates a model's skill at informing specific decisions, different model diagnostics are correlated to a model's decision-support capability.

In this thesis, a hydrologic model is used to evaluate three reservoir operation rule curves in the Lake of the Woods Watershed, based on ecological and economic impacts. Synthetic realities are generated through random sampling of parameters. Each synthetic reality is operated using all rule curves to determine the preferred rule curve for a given parameter set. Then, the model is calibrated to the synthetic realities' using various calibration formulations. For each calibration, the model is evaluated on whether the model prefers the same rule curve preferred by the synthetic reality. After many of parameter set realizations, each incremental value of calibration formulation is assigned a similarity score to describe the probability of informing the correct decision. Using the correlation, the model's capabilities and uncertainties may be more readily quantified and communicated to stakeholders. Results indicate specific calibration formulation may be beneficial to support specific decisions.

ACKNOWLEDGEMENTS

I would like to first and foremost thank my supervisors, Dr. James Craig and Dr. Bryan Tolson, who have inspired me and guided me to think and explore. Being co-supervised by such brilliant minds was truly an experience I will cherish for decades to come. Although there were times where my supervisors had conflicting good ideas, I was able to learn twice as much from my graduate experience.

I also would like to thank James Bomhof from Lake of the Woods Control Board for his generous support in Lake of the Woods – Rainy Lake model development. Without his help, this project surely would have not been possible.

I wish to thank you to numerous staff at Ontario Power Generation, namely Joan Frain, Mark Nussli, Michael McNiven, Connor Werstuck, and Kurt Kornelsen, who have provided support and input for model development.

Many good colleagues have helped me throughout my thesis. I would like to thank Hongli Liu and Robert Chlumsky who have helped me take the first steps in modelling. Dr. Juliane Mai have saved me countless hours with her help with Compute Canada, which I cannot thank enough. Thank you to Sarah Grass, Elise Devoise, and Ming Han who have helped with numerous modelling tips.

Thank you to Sarah Chun for making my endless nights of writing enjoyable.

Finally, I would like to thank all my family and friends who have provided continuous love and support.

Table of Contents

List of Figures ………………………………………………………………. vii

List of Tables ……………………………………………………………….. x

Chapter 1 Introduction ……………………………………..…………… 1

1.1 Goals and Objectives ……………………………………………….. 2

1.2 Thesis Organization ………………………………….……………… 3

Chapter 2 Background …………………………………………………... 5

2.1 Hydrologic Modelling ……………………………………………….. 5

2.1.1 What are Hydrologic Models ……………………………………….. 5

2.1.2 Model Classifications ……………………………………………... 6

2.1.3 Challenges in Modelling Canadian Shield Hydrology ………………. 7

2.2 Model Evaluation Methods ………………………………………………. 8

2.2.1 Model Calibration Techniques ……………………………………... 8

2.2.2 Commonly Used Metrics for Model Evaluation ……………………... 10

2.2.3 Model Validation Techniques ………………………………………. 14

2.2.4 Crash Testing Concepts in Model Validation ………………………... 18

2.3 Improved Model Validation Methods ……………………………………. 20

2.3.1 Issues with Current Model Validation Methods ……………………... 20

2.3.2 Using Models for Decision Making …………………………………. 21

Chapter 3 Methods and Results of Modelling Canadian Shield Hydrology ……. 23

3.1 Model Development ……………………………………………………. 23

3.1.1 Model Structure ……………………………………………………... 23

3.1.2 Explicit Representation of Canadian Shield Hydrology Characteristics ………………………………………………………. 24

3.1.3 Modelling Operation in Managed Reservoir ………………………… 26

3.2 Kaministiquia Watershed Model ……………………………………... 30

3.2.1 Kaministiquia Case Study Background ………………………………. 31

3.2.2 Kaministiquia Watershed Model Calibration Formulation ……….. 34

3.2.3 Kaministiquia Watershed Model Results …………………………….. 35

3.2.4 Additional Strategies to Improve Reservoir Simulation …………….. 36

3.3 Lake of the Woods Watershed Model ………………………………………. 43

3.3.1 Lake of the Woods Case Study Background …………………………. 44

3.3.2 Lake of the Woods Watershed Model Calibration Formulation …… 50

3.3.3 Lake of the Woods Watershed Model Results ……………………….. 51

Chapter 4 Assessing and Improving Hydrologic Models Used for Decision Making ………………………………………………………………... 57

4.1 Case Study – Selection of Operational Rule Curves ………………………. 57

4.1.1 Rule Curve Motivations in Rainy Lake ………………………………. 57

4.1.2 Flow Restrictions at Rainy Lake ……………………………………… 58

4.1.3 Rule Curve Options at Rainy Lake …………………………………… 60

4.2 Evaluation Criteria in Rule Curve Selection ………………………………. 62

4.2.1 Ecological Benefits …………………………………………………... 62

4.2.2 Economical Benefits …………………………………………………. 64

4.2.3 Flood damage reduction benefits …………………………………….. 64

4.3 Decision Crash Testing for Model Assessment …………………………….. 66

4.3.1 Generation of Synthetic Observations ……………………………….. 66
4.3.2 DCT Methodology …………………………………………………….. 69
4.4 DCT Results and Calibration Objective Formulation ……………………. 71
4.4.1 Results Based on Calibration to Different Objective Gauges ………. 72
4.4.2 DCT Applied on Different Calibration Objectives …………………... 77
4.4.3 DCT to Assess Error in Evaluation Criteria …………………………. 80
4.5 Case Study Conclusions ……………………………………………………. 94
Chapter 5 Conclusions ………………………………………………………... 96
5.1 Contribution to Literature ………………………………………………….. 96
5.2 Future Opportunities for DCT ……………………………………………... 96
References ……………………………………………………………………… 98
Appendix A – Raven Input File with Process Selection ………………………….. 103
Appendix B – Downstream hydrographs of Kaministiquia Watershed ………… 104
Appendix C – Hydrographs of Lake of the Woods Watershed ………………….. 111
Appendix D – Box Whisker Plots of Similarity Scores …………………………... 140

List of Figures

Figure 2.1 – Various value and performance rating of NSE and PBIAS across a select set of studies in the literature reporting calibration results ……………... 14
Figure 2.2 – Illustration of the GSST procedure (example with 18 years of historical data and 5 year windows) …………...…………...…………...……….. 17
Figure 2.3 – Discharge gauging stations and groundwater observation wells of the Karup catchment ……………………………………………………………….. 18
Figure 3.1 – Hydrologic process diagram of the Raven model adopted from R.Chlumsky (2017) ………………………………………………………………… 24
Figure 3.2 – Conceptual diagram of soil profiles in the model ………………….. 25
Figure 3.3 – Rectangular weir-like outlet for HEL ………………………………. 26
Figure 3.4 – Workflow to determine outflow to model reservoir operation ……. 30
Figure 3.5 – Watershed delineation of the Kaministiquia watershed …………… 31
Figure 3.6 – Plot of Dog Lake stage during calibration period when calibrated to average of stage NSE and dh/dt NSE …………………………………………… 37
Figure 3.7 – Stage plot of rain corrected Dog Lake model in comparison with initial validation results …………………………………………………………….. 38
Figure 3.8 – Plot of Dog Lake stage during validation period when calibrated to average of stage NSE and dh/dt NSE ……………………………………………… 39
Figure 3.9 - Plot of Kashabowie Lake stage during calibration period when calibrated to average of stage NSE and dh/dt NSE ………………………………. 40
Figure 3.10 - Plot of Kashabowie Lake stage during validation period when calibrated to average of stage NSE and dh/dt NSE ………………………………. 40
Figure 3.11 - Plot of Shebandowan Lake stage during calibration period when calibrated to average of stage NSE and dh/dt NSE ………………………………. 41
Figure 3.12 - Plot of Shebandowan Lake stage during validation period when calibrated to average of stage NSE and dh/dt NSE ………………………………. 41
Figure 3.13 – Inflow hydrograph of Dog Lake when calibrated to stage and inflow ………………………………………………………………………………… 42
Figure 3.14 – Stage plots of Dog Lake when calibrated to stage and inflow ……. 43
Figure 3.15 – Map of extent of LOWRL watershed ……………………………… 44
Figure 3.16 – Basin delineation used for Raven model …………………………... 46
Figure 3.17 – Assignment of basin number to each grid cell …………………….. 48
Figure 3.18 – Pareto front of non-dominated solutions in multi-objective calibration to flow NSE and inflow NSE ………………………………………….. 53
Figure 3.19 – NSEs of Pareto optimal solution evaluations during the validation period ………………………………………………………………………………... 54
Figure 4.1 – Drought line of Rainy Lake determined by IJC …………………… 59
Figure 4.2 – Stage-Discharge relationship of Rainy Lake at different gate configurations …………………………………………………………………….. 60
Figure 4.3 – Operational rule curve A used by LWCB with high and low targets ……………………………………………………………………………... 61
Figure 4.4 – Operational rule curve B modified from rule curve A with lower spring target to reduce flood risks ………………………………………………… 61

Figure 4.5 – Operational rule curve B modified from rule curve A with low winter drawdown to improve ecological benefits ………………………………… 62
Figure 4.7 – Stage records of Rainy Lake from five biggest historical floods …... 65
Figure 4.8 – Raven model simulated stage using rule curve A in comparison with observed peak during 2014 storm event ……………………………………... 66
Figure 4.9 – Example of synthetic calibration experiment to test optimization algorithm ……………………………………………………………………………. 68
Figure 4.10 – DCT setup workflow used for rule curve selection experiment ….. 70
Figure 4.11 – Box whisker plot of similarity score for economic benefit when calibrated to inflow NSE ………………………………………………………… 72
Figure 4.12 – Box whisker plot of similarity score for ecological benefit when calibrated to inflow NSE ………………………………………………………… 74
Figure 4.13 – Box whisker plot of similarity score for flood damage reduction benefit when calibrated to inflow NSE ……………………………………………. 74
Figure 4.14 – Comparison of mean similarity score for economic benefit based on calibration to inflow versus calibration to stream flow ………………………. 75
Figure 4.15 – Comparison of mean similarity score for ecological benefit based on calibration to inflow versus calibration to stream flow ………………………. 76
Figure 4.16 – Comparison of mean similarity score for flood damage reduction benefit based on calibration to inflow versus calibration to stream flow ……….. 76
Figure 4.17 – Comparison of mean similarity score for economic benefit based on calibration to different calibration objective functions ………………………. 78
Figure 4.18 – Comparison of mean similarity score for ecological benefit based on calibration to different calibration objective functions ………………………. 79
Figure 4.19 – Comparison of mean similarity score for flood damage reduction benefit based on calibration to different calibration objective functions ……….. 79
Figure 4.20 – Box whisker plot of calibrated model error in environmental benefit. X-axis shows gauges used in calibration and rule curve utilized ……….. 83
Figure 4.21 – Box whisker plot of calibrated model error in economic benefit. X-axis shows gauges used in calibration and rule curve utilized. ……………….. 84
Figure 4.22 – Box whisker plot of calibrated model error in flood damage reduction benefit. X-axis shows gauges used in calibration and rule curve utilized …………………………………………………………………………….. 85
Figure 4.23 – Box whisker plot of calibrated model error in environmental benefit and stopped when Rainy Lake NSE ≥ 0.7. X-axis shows gauges used in calibration and rule curve utilized ………………………………………………… 87
Figure 4.25 – Box whisker plot of calibrated model error in economic benefit and stopped when Rainy Lake NSE ≥ 0.7. X-axis shows gauges used in calibration and rule curve utilized ………………………………………………… 88
Figure 4.26 – Box whisker plot of calibrated model error in flood damage reduction benefit and stopped when Rainy Lake NSE ≥ 0.7. X-axis shows gauges used in calibration and rule curve utilized. ………………………………. 89
Figure 4.27 – Box whisker plot of calibrated model error in environmental benefit. X-axis shows diagnostic used in calibration and rule curve utilized. …... 91
Figure 4.28 – Box whisker plot of calibrated model error in economic benefit. X-axis shows diagnostic used in calibration and rule curve utilized. ……………. 92

Figure 4.29 – Box whisker plot of calibrated model error in flood damage reduction benefit. X-axis shows diagnostic used in calibration and rule curve utilized. ………………………………………………………………………………. 93

List of Tables

Table 3.1 – Summary of components involved in flow calculations for reservoir operations ……………………………………………………………………… 29
Table 3.2 – Summary of sub-basins and physical characteristics ……………….. 33
Table 3.3 – Summary of Kaministiquia Watershed model calibration results when calibrated to NSE with inflow forced gauges bolded ………………………. 35
Table 3.4 – Summary of Kaministiquia Watershed model validation results when calibrated to NSE with inflow forced gauges bolded ………………………. 36
Table 3.5 – Summary of Kaministiquia Watershed maximum stage constraint corrected model validation results when calibrated to NSE ……………………... 37
Table 3.6 – Cumulative drainage areas of WATFLOOD model and Raven model (km2) ………………………………………………………………………. 47
Table 3.7 – Summary of Raven parameters with equivalent WATFLOOD counterparts ……………………………………………………………………… 48
Table 3.8 – Summary of Raven model parameters calibrated …………………... 49
Table 3.9 – Calibration and validation results for calibration to flow NSE, with bolded NSE when model performs better by 0.05 or greater ……………………. 51
Table 3.10 – Calibration and validation results for calibration to reservoir inflow NSE, with bolded NSE when model performs better by 0.05 or greater ... 52
Table 3.11 – Average NSE of non-dominated solutions generated by multi-objective calibration during calibration and validation periods, with bolded NSE when model performs better by 0.05 or greater …………………………….. 55
Table 3.12 – Overall NSE comparison between Raven and WATFLOOD ……... 55
Table 4.1 – Summary of probability of survivability of species dependent on water level ……………………………………………………………………… 63
Table 4.2 – Similarity score across each evaluation criteria across four different calibration objective gauge formulations ………………………………………. 81
Table 4.3 – Similarity score across each evaluation criteria across four different calibration objective gauge formulations ………………………………………. 86
Table 4.4 – Similarity score across each evaluation criteria across four different calibration objective diagnostics: NSE, KGE, percent bias penalized NSE (NSEP), and percent bias, when calibrated with a budget of 50 model runs …… 90

Chapter 1

Introduction

Global warming continues to be one of the major problems of the 21st century. Climate change brings many challenges to the hydrology community, as it alters the hydrologic cycles and subsequently impacts the quantity and quality of regional water resources (Gleick, 1989). Hydrologic models are commonly used to predict the impact of climate change on water resources and to aid in decision making process to accommodate climate change. Interplay between theoretical, computational, and experimental advancements has led development of more complex and robust hydrologic models (Paniconi and Putti, 2015). However, due to inherent heterogeneity and lack of data, many parameters in hydrologic models cannot be measured. As a result, calibration has become a crucial component in hydrologic modelling. Calibration involves varying parameter values within reasonable ranges until the differences between modeled outputs of system response and the corresponding observations are minimized. The model is considered calibrated when it reproduces historical data within some subjectively acceptable level of coherence (Konikow and Bredehoeft, 1992). Every model serves a different purpose, and the subjective acceptable level of coherence differs from model to model and user to user. To satisfy the various needs of modelers, researchers have developed numerous model diagnostics to represent level of coherence. However, a value of a specific diagnostic may be deemed acceptable for one model application but unacceptable for another. Such discrepancies arise from multiple factors, including: hydrologic complexity of basin, data availability, resources spent on model, validation methods, and intended use of model. The value of a model comes from its ability to reliably synthesize data needed for decision support that is unavailable in the real world (Klemes, 1986). Accordingly, a model must demonstrate how well it can perform the kind of task for which it is intended. Since many model evaluation methods fail to test the model for its intended use, quantification of model performance still remains subjective. New model diagnostics continue to be published in literature, with no advancement in communicating the value of the new diagnostics in a practical applications.

This thesis utilizes the Decision Crash Testing (DCT) method (Chlumsky, 2017) to bridge the gap between traditional model diagnostics and the evaluation of a model's capabilities for its intended use. Since the goodness of a model diagnostic is heavily dependent on the model's intended use, it only makes sense to evaluate the goodness of a model diagnostic for a specific scenario of model application.

1.1 Goals and Objectives

This thesis has three main goals. The first goal is to hydrologic models suitable for DCT case study. The second goal of this thesis is to quantify the adequacy of commonly used model diagnostics in specific decision making contexts using the DCT framework. The third goal is to utilize the DCT framework to assess the optimal calibration objective formulation for a hydrologic model which will be used for a specific decision purpose. Achieving the goals required a realistic decision making scenario supported by a well-behaved hydrologic model.

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

1. To develop and implement novel approaches for modelling lakes and reservoir operations in the Canadian Shield
2. To present an implementation of DCT that can assess the utility of model diagnostics in a specific decision making scenario of reservoir operation rule curve selection

3. To quantify the effectiveness of different calibration objectives in a specific decision making scenario

1.2 Thesis Organization

This thesis is comprised of six chapters, first of which serves as an introduction to the thesis.

Chapter 2 provides relevant background about hydrologic modelling and challenges in the Canadian Shield. Then, it discusses both common and uncommon model evaluation methods and their inherent problems. This provides the justification of using the DCT framework and demonstrates how it may contribute to the state of modelling practice.

Chapter 3 discusses the approaches taken to better represent the hydrology in the Canadian Shield. Special emphasis is given in modelling both complex lake systems and human operation. The model structure developed is deployed in two Canadian Shield basins: the Kaministiquia watershed and Lake of the Woods watershed. Model inter comparison is performed with two traditional hydrologic models, the GR4J model and the WATFLOOD model.

Chapter 4 discusses how the developed Lake of the Woods model might be used in a specific decision making context of reservoir release decision making (specifically rule curve selection amongst three alternatives for Rainy Lake). Implementation of operational behavior and mapping of model output to decisions is discussed. The model utilizes real examples form the report '*Managing Water Levels and Flows in the Rainy River Basin'* (International Rainy and Namakan Lakes Rule Curve Study Board, 2017). Then, the key fundamentals of the DCT framework and how it can be implemented to quantify a model's capability in informing rule curve selection is discussed. Also, it demonstrates an objective comparison of different calibration objective formulations to support the rule curve decision making scenario.

Chapter 5 summarizes the thesis' contribution to literature. The results of the thesis brings new perspective to one of the biggest topics of debate in hydrology (Kirchener, 2006): *How important is it to get the right results for the right reasons?* Potential future improvements to the DCT framework is also presented.

Chapter 2

Background

This chapter provides the necessary background knowledge for understanding the significance of the research. General background regarding in hydrologic modelling and challenges are discussed. Then, different model evaluation methods are discussed along with their limitations. Finally, different requirements to overcome the shortcomings of traditional evaluation methods are discussed to illustrate the necessity of the thesis.

2.1 Hydrologic Modelling

In this section, the purpose and classification of hydrologic models are presented. Challenges in modelling Canadian Shield hydrology is also highlighted.

2.1.1 What are Hydrologic Models?

Hydrology is a science which treats movement of all phases of the earth’s water, with application in design and operation of hydraulic structures, water supply, wastewater treatment and disposal, irrigation, drainage, hydropower generation, flood control, navigation, erosion and sediment control, salinity control, pollution abatement, recreational use of water, and fish and wildlife protection. The role of applied hydrology is to help analyze the problem and aid in planning and management of water resources (Chow et al., 1988). Due to the heterogeneous nature of the natural systems and lack of resources, availability of hydrologic data is limited in space and time. To compensate and to make predictions regarding futures scenarios, researchers and hydrologists have developed mathematical models to simulate hydrology. Hydrologic models are used to synthesize a (continuous) record of some hydrologic variable Y, such as stream discharge, for a period T, from available concurrent records of other input variables X, Z, etc. A model may be used to simulate hydrologic variable Y for future period of T, under forecasts of input variables X, Z (such as weather forecasts), thus making the model a forecast model. A model output may be used for complex decision making problems where the output is a function of hypothetical input scenarios, typically for water-management decisions. Klemes (1986) argues that a useful model is a model capable of adequately synthesizing data to inform decision making process.

2.1.2 Model Classification

Singh (2002) classified hydrologic models based on (1) process description; (2) timescale; (3) space scale; (4) techniques of solutions; (5) land use; and (6) model use. Depending on the description of the processes, hydrologic models may be classified as conceptual or physically-based (Refsgaard, 1997).

In physically-based models, individual hydrologic processes are represented by individual physical representation of processes, driven by physically-meaningful and measureable parameters. Recent advancements in technology provide wider availability of spatially distributed parameter data, ranging from soil types and land use to radar rainfall, facilitating in production of simplified physically-meaningful distributed hydrologic models. Conceptual models, on the other hand, can be seen as data-driven models. Models attempt to transform model inputs (e.g., radiation, temperature, and precipitation) to appropriate model outputs (e.g., stream flow) through statistical and mathematical transfer functions. Conceptual models require large sets of observation data to adequately train the model to produce accurate outputs. Even with extensive training, conceptual models may have difficulty in predicting events beyond the conveyance of the training set (Todini, 2007).

The mathematical and physical equations used in hydrologic models are continuous in time and often space. However, analytical solutions are extremely difficult to obtain due to the complex nature of hydrologic systems. Therefore, numerical methods are used for most practical cases. General formulation involves partial differential equations in space and time. If the spatial derivatives are ignored, the models are called "lumped". In "distributed" models, the output is a function of space and time. Strictly speaking, for a model to be truly distributed, all aspects of the models, including initial and boundary conditions, parameters, forcing functions, and sources and sinks must be spatially distributed (Singh et al., 2002). Due to practical limitation of data and discrete descriptions of watershed geometry, modelers may use "semi-distributed" models. Semi-distributed models often use spatially distributed hydrologic response units (HRUs) to represent larger spatial areas as a single response unit with a unique response to a precipitation event. Properties within a single HRU are assumed to be homogenous.

Appropriate model complexity is heavily dependent on the intended application of the model and data availability. Models intended for forecasting may be better suited to use of data-driven models, as training sets become readily available after forecasts. Physically-based models may be more appropriate for application in what-if scenarios, such as land use change or reservoir operation change, in heavily instrumented basins.

2.1.3 Challenges in Modelling Canadian Shield Hydrology

The Canadian Shield occupies one-third of Canada's land area, comprising mainly of Precambrian rock that was glaciated by the Laurentide Ice Sheet to produce a rolling topography (Spence and Woo, 2008). The open water contained in wetland and lakes accounts for nearly 25% of the Shield Area. Typically water storage within the bedrock is small. Although dependent on soil depth, Spence and Woo (2008) estimated water storage of the terrain to be within 10 mm in Ontario. As a result, the lake system plays a critical role in water storage and runoff generation as response to precipitation events. Spence and Woo (2008) showed that the runoff generation from the catchments is dependent on the topography and connectivity of the lakes. Hydrology can be driven by the connectivity of the lakes where lakes become disconnected or connected depending on the season variation and elemental thresholds. Large lakes' storage and release functions can overwhelm the seasonality of the land phase runoff, resulting in streamflow signal dominated by the hydraulic dynamics of the lakes. The fill-and-spill response of the lakes is extremely difficult to model due to limitation in resources. Woo and Mielko (2008) utilized data on lake levels at half hour intervals, precipitation, ice fraction (from photography), flow into and out of lake, radiation, air temperature, and water temperature to model the fill-and-spill response of five lakes in the Northwest Territories. Such amount of hydrologic and forcing data is unavailable in most unmanaged reservoirs. Often, the impacts of the lake response are compensated through perturbation of physical and empirical parameters during the calibration period to match the hydrograph. Despite the difficulty, representation of the lake system is a crucial component in modelling the Canadian Shield hydrology.

2.2 Model Evaluation Methods

This section provides background knowledge in calibration techniques, common diagnostics, and validation techniques used in hydrologic modelling.

2.2.1 Model Calibration Techniques

Many hydrologic model parameters may be unavailable due to limited access to field data or empirical nature of the parameters. Even physically-based parameters may be a conceptual representations of abstract watershed characteristics depending on scale and discretization of the watershed. In such case, modelers are required to estimate model parameters to enable model to closely match the behavior of the real system it represents. A traditional method of parameter estimation is the “manual” calibration approach. A modeler with knowledge of the watershed and experience with the model would use trial-and-error procedure to adjust the parameters, while visually comparing the observations and simulated outputs using graphical plots (Gupta et al., 1999). Complicated interaction between model parameters can make manual calibration extremely time-consuming and frustrating. Nonetheless, manual calibration provides modelers doing the calibration with better understanding of parameter interaction and sensitivity of model outputs to model parameters.

To address the time-consuming and difficult nature of manual calibration, researchers have developed methods to speed up the estimation process through automatic calibration. Gupta et al. (1999) highlights the process of automatic calibration as follows:

1. A period of calibration data is selected
2. An initial guess is made as to the probable values (or range of values) for the parameters
3. The model is run using these values for the parameters
4. The “distance” between the model output and the observed data is measured using a mathematical equation called an objective function or model diagnostics
5. An automatic optimization procedure (called a search algorithm) is used to search for the parameter values that optimize the value of the objective function

An important choice made by the modeler in calibration is to choose the appropriate model diagnostic to be used. Past research has not proved possible to clearly demonstrate that a particular objective function is better suited for calibration of a model than some other (Gupta et al., 1998). Each objective function may be well suited for different parts of the hydrograph, and an optimal objective function may vary from one model application to another. Utilizing a single-objective function for calibration requires an erroneous assumption that all the available information regarding one hydrologic variable can be summarized using a single aggregate measure of model performance (Tang et al., 2005). In a multi-objective calibration experiment, a set of solutions that optimizes more than one objective function is found. The objective functions may be the same diagnostic for multiple observation data sets, different diagnostics for a single data set, or any combination of data sets and diagnostics. The set of solutions is also known as a Pareto front, with is comprised of Pareto optimal solutions. A solution X* is classified as Pareto optimal when there is no feasible solution X that has a better objective function value in one or more objectives without degrading performance for at least one other objective function value.

2.2.2 Commonly Used Metrics for Model Evaluation

In order to test the model's predictive abilities, modelers quantitatively assess the degree to which the model simulations match the observation data (Legates and McCabe, 1999). This quantitative assessment, also known as a model diagnostic, is the simplest form of model evaluation. Model diagnostics are often used to inform the model calibration process, where a range of model parameters are sampled to minimize the difference between the simulation results and the observation data, often expressed as a numerical diagnostic (Legates and McCabe, 1999). Moriasi et al. (2007) categorize these diagnostics into standard regression, dimensionless, error index, and graphical. Each diagnostic is designed to convey specific types of information, while inadequate with certain types of data. In most cases, these metrics are applied to comparison of an observed time series (e.g. hydrograph) to a modeled equivalent. In this section, four model diagnostics: root mean square error (RMSE), percent bias (PBIAS), Nash Sutcliffe Efficiency (NSE), and Kling Gupta Efficiency (KGE) are discussed in more detail.

RMSE, shown in equation 1, provides a mean error of the model error in the unit of interest.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i^{sim}-X_i^{obs}\right)^2}{n}} \tag{1}$$

where $x_i^{sim}$is the simulated value at time step i, $X_i^{obs}$is the observed value at time step i, and *n* is the total number of observations. A RMSE of zero indicates an error-free model, and it is commonly accepted that a lower RMSE indicates a better performance. Although researchers have made efforts to set a guideline to qualify what is considered a low RMSE (Singh et al., 2014), there is no widely-accepted standard threshold for adequate RMSE values. A RMSE-observation standard deviation ratio (RSR) was developed to normalize the RMSE by taking the ratio between the RMSE and the standard deviation of the observation data (Moriasi et al., 2007).

Percent bias, shown in equation 2, measures the average tendency of the simulated results to be larger or smaller than their observed counterparts.

$$PBIAS= \frac{\sum_{i=1}^{n}\left(X_i^{sim}-X_i^{obs}\right)*100}{\sum_{i=1}^{n}\left(X_i^{obs}\right)} \tag{2}$$

A positive percent bias indicates model overestimation, and a negative indicates underestimation. Percent bias is commonly used to assist in quantifying water balance error by calculating the percent deviation of streamflow volume.

NSE is one of the most widely used diagnostics in hydrologic modelling. It provides a normalized statistic that determines the relative magnitude of residual variance compared to the measured data variance. Computation of NSE is shown in equation 3.

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(X_i^{obs} - X_i^{sim}\right)^2}{\sum_{i=1}^{n}\left(X_i^{obs} - \overline{X^{obs}}\right)^2} \tag{3}$$

NSE ranges from -∞ to 1.0 with NSE = 1 being the optimal value. NSE has been recommended for use by the American Society of Civil Engineers (1995) and Lebates and McCabe (1999). Also, the extensive use of NSE in the hydrology community provides ample information on the reported values. Despite the convenience and popularity of the NSE, there have been numerous discussions about the suitability of the NSE (Gupta et al., 2009). NSE overestimates model performance for highly seasonal variables, such as snowmelt dominated basins. In some cases, low NSE may not necessarily indicate a poor model, but only that the observation data is very steady (Criss and Winston, 2008).

Weglarczyk (1998) showed a decomposition of NSE into measurements of three components: linear correlation, bias, and variability of the data. Subsequently, calibration of models using NSE can be viewed as optimizing a weighted objective function (and thus solving a multi-objective optimization problem). However, the bias term has a low "weight" when NSE is used with highly variable observation data (Gupta et al., 2009). Also, variability in flows is systematically underestimated so that the ratio of the simulated and observed data will tend to be equal to the correlation coefficient. This results in an underestimation of the peak flows when using NSE during the calibration process (Gupta et al., 2009). In contrast, the KGE metric, introduced as an alternative to NSE, incorporates equal weighting to correlation, bias, and variability of the data. The components of the KGE are shown in equation 4.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{4}$$

where

$$\alpha = \frac{\sigma_{sim}}{\sigma_{obs}},$$

$$\beta = \frac{\mu_{sim}}{\mu_{obs}}$$

$$r = \frac{Cov_{so}}{\sigma_{sim} \cdot \sigma_{obs}}$$

where σ represents standard deviation, μ represents mean, and Cov_{so} represents covariance between simulated and observed values. KGE ranges from -∞ to 1.0 with KGE = 1 being the optimal value.

Despite the effort to create different diagnostics to capture a variety of hydrologic signatures, the goodness of a diagnostic value remains highly subjective. Moriasi et al. (2007) has summarized various NSE and PBIAS values from multiple literature sources. The summarized values are shown in Figure 2.1. The relationship between the diagnostic values and performance ratings vary from paper to paper. Looking further into the papers referenced in Figure 2.1, it appears that determination of adequacy in performance rating is heavily dependent on author's experience and judgement. Motovillov et al. (1999) claimed *"Figure 5 shows the observed and simulated discharge values for a few basins in the NOPEX area for 2 years: one with 'satisfactory' agreement – 1986-87, and the other with 'the worst' agreement – 1988-89."* Author's claim on the NSE being 'satisfactory' did not have any reference to literature. The goodness of a particular diagnostic may be a function of multiple factors, including the hydrologic complexity of the watershed, data availability, resources spent on model, validation methods, and intended use of the model.

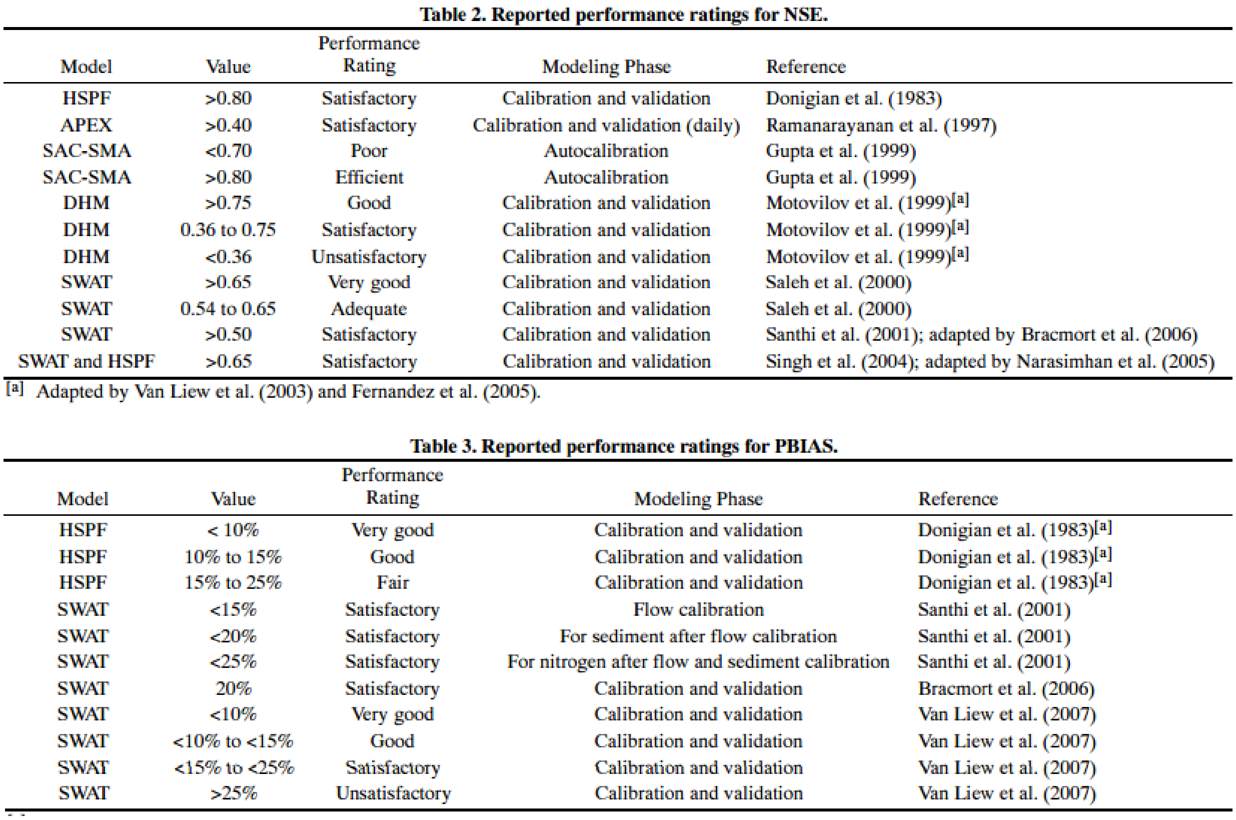


Table 2. Reported performance ratings for NSE.

Model	Value	Performance Rating	Modeling Phase	Reference
HSPF	>0.80	Satisfactory	Calibration and validation	Donigian et al. (1983)
APEX	>0.40	Satisfactory	Calibration and validation (daily)	Ramanarayanan et al. (1997)
SAC-SMA	<0.70	Poor	Autocalibration	Gupta et al. (1999)
SAC-SMA	>0.80	Efficient	Autocalibration	Gupta et al. (1999)
DHM	>0.75	Good	Calibration and validation	Motovilov et al. (1999)[a]
DHM	0.36 to 0.75	Satisfactory	Calibration and validation	Motovilov et al. (1999)[a]
DHM	<0.36	Unsatisfactory	Calibration and validation	Motovilov et al. (1999)[a]
SWAT	>0.65	Very good	Calibration and validation	Saleh et al. (2000)
SWAT	0.54 to 0.65	Adequate	Calibration and validation	Saleh et al. (2000)
SWAT	>0.50	Satisfactory	Calibration and validation	Santhi et al. (2001); adapted by Bracmort et al. (2006)
SWAT and HSPF	>0.65	Satisfactory	Calibration and validation	Singh et al. (2004); adapted by Narasimhan et al. (2005)

[a] Adapted by Van Liew et al. (2003) and Fernandez et al. (2005).

Table 3. Reported performance ratings for PBIAS.

Model	Value	Performance Rating	Modeling Phase	Reference
HSPF	< 10%	Very good	Calibration and validation	Donigian et al. (1983)[a]
HSPF	10% to 15%	Good	Calibration and validation	Donigian et al. (1983)[a]
HSPF	15% to 25%	Fair	Calibration and validation	Donigian et al. (1983)[a]
SWAT	<15%	Satisfactory	Flow calibration	Santhi et al. (2001)
SWAT	<20%	Satisfactory	For sediment after flow calibration	Santhi et al. (2001)
SWAT	<25%	Satisfactory	For nitrogen after flow and sediment calibration	Santhi et al. (2001)
SWAT	20%	Satisfactory	Calibration and validation	Bracmort et al. (2006)
SWAT	<10%	Very good	Calibration and validation	Van Liew et al. (2007)
SWAT	<10% to <15%	Good	Calibration and validation	Van Liew et al. (2007)
SWAT	<15% to <25%	Satisfactory	Calibration and validation	Van Liew et al. (2007)
SWAT	>25%	Unsatisfactory	Calibration and validation	Van Liew et al. (2007)

Figure 2.1 – Various value and performance rating of NSE and PBIAS across a select set of studies in the literature reporting calibration results (Moriasi et al., 2007)

2.2.3 Model Validation Techniques

A broad definition of validation includes any process that aims to verify the ability of a procedure to adequately accomplish a given task (Biondi, 2011). Model validation techniques are predicated upon the philosophy that a model must be tested for its intended use. Since no simulation model is intended merely to show how well it fits the data used for its development, performance characteristics during the calibration period are insufficient evidence for a model's satisfactory performance. Unfortunately, with the exception of forecasting, data for the model's intended use is unavailable to test the model's performance in its intended use– if it did, a simulation model would not be needed. Therefore, efforts need to be made to demonstrate a model’s ability to generate results for a situation *similar* to that of which the model is developed to be used for (Klemes, 1986). Klemes (1986) had proposed two major levels of categories to define model validation approach or tests:

(1) Stationary conditions (physical conditions do not change with time), and

(2) Nonstationary conditions (physical conditions change with time) - each of them being divided into hierarchical subgroups:

In each of the two categories, Klemes (1986) proposed testing the model utilizing two different basins:

(a) The same station (basin) which was used for calibration, and

(b) A different station (basin).

For each of the subgroups 1a, 1b, 2a, and 2b, an operational validation testing was proposed.

Split-Sample Testing (SST) (1a) is the most basic form of calibration-validation process and the full description of this test as proposed by Klemes (1986) is as follows. SST should be used to test models used for stationary climate and land use conditions within the same basin used for calibration. The model which passes a SST can be used for filling-in missing segment of, or extending, a streamflow record. SST involves, calibration of model using the first 70% of the observation data and validation using the remaining 30%. Next, the model is calibrated using the first 30% and validated using the remaining 70%. The model qualifies as acceptable if both validation results are similar and acceptable The Proxy-Basin Test (PBT) (1b) should be used to test models used for stationary climate and land use conditions within an ungauged basin. Passing the PBT demonstrates basic credibility in geographical transposability of the model. For example, in order to simulate streamflow data for an ungauged basin C, two gauged basins A and B are selected within the region. The model is then calibrated using basin A and validated using basin B and *vice versa*. The Differential Split-Sample Test (DSST) (2a) is used to evaluate models developed for non-stationary conditions within the same gauged basin. Typically, DSST is used to test if a model can simulate data for future change in land use and/or climate. For changes in climate, the modeler needs to identify two periods with different climate conditions. The model is calibrated using one period and validated using the other. In general, the model should demonstrate its ability to perform under the transition required (e.g., wetter to drier climate). Testing a change in land use requires finding a gauged basin with historical data before and after a change in land use. The model is calibrated using data before the land use change and validated using the other. The Proxy-Basin Differential Split-Sample Test (PBDSST) (2b) is a combination of PBT and DSST used to generate streamflow data for a nonstationary conditions for an ungauged basin. Test should be applied for models that need to be both geographically and climatically (or land-use-wise) transposable. Many researchers aim for such universal transposability of hydrologic models; yet such success may not be achieved in decades to come. For modelling an ungauged basin C, the modelers need to identify to gauged basins A and B, with characteristics similar to those of basin C. Calibration would be performed using one climatic condition (e.g. dry) of A and validated using a different climatic condition of B.

There are two limitations in the implementation of the DSST (Coron et al., 2014). First, it requires the modelers to identify in advance the climatic characteristics that will most likely play a key role in limiting the model transposability. Second, the number of transfer tests is usually small, limiting the ability to draw general conclusions and discovering the main drivers of model transposability from the results themselves. The General Split Sample Testing (GSST) tests both similar and contrasting climatic conditions (Coron et al., 2014). GSST requires a calibration using one window of the available historical data, and validation using all other windows of historical data that do not overlap with the calibration period. This process is repeated for all possible windows of historical data. The GSST is illustrated in Figure 2.2.

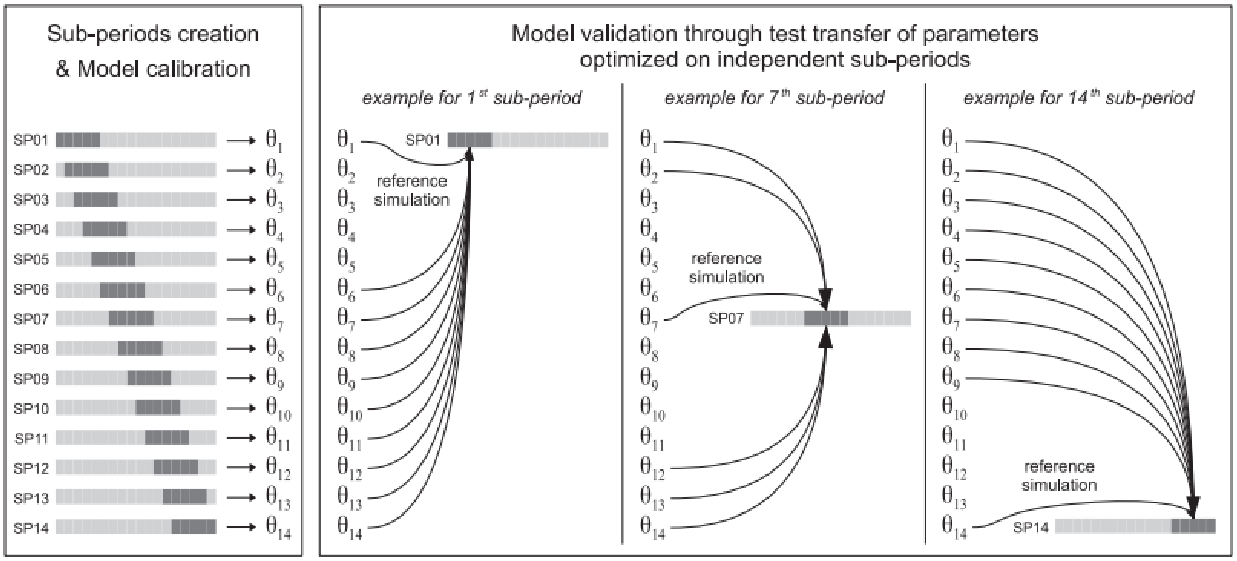

Sub-periods creation & Model calibration
SP01 → θ1
SP02 → θ2
SP03 → θ3
SP04 → θ4
SP05 → θ5
SP06 → θ6
SP07 → θ7
SP08 → θ8
SP09 → θ9
SP10 → θ10
SP11 → θ11
SP12 → θ12
SP13 → θ13
SP14 → θ14
Model validation through test transfer of parameters optimized on independent sub-periods
example for 1st sub-period
SP01
reference simulation
example for 7th sub-period
reference simulation
SP07
example for 14th sub-period
reference simulation
SP14

Figure 2.2 – Illustration of the GSST procedure (example with 18 years of historical data and 5 year windows) (Coron et al., 2014)

Refsgaard (1997) addressed the limitation in number of transfer tests by performing validation using spatially varying internal groundwater table levels in the Karup catchment in Denmark. Often, expectations are made that a successful split sample test on the outlet of a catchment and groundwater table indicate validity of simulation of internal flows and ground-water table levels (Refsgaard, 1997). The original model of the Karup catchment was calibrated and validated using a SST at station 20.05 and groundwater level simulations at wells 21, 44, 55, 8, 9, 11, and 12 shown in Figure 2.3.

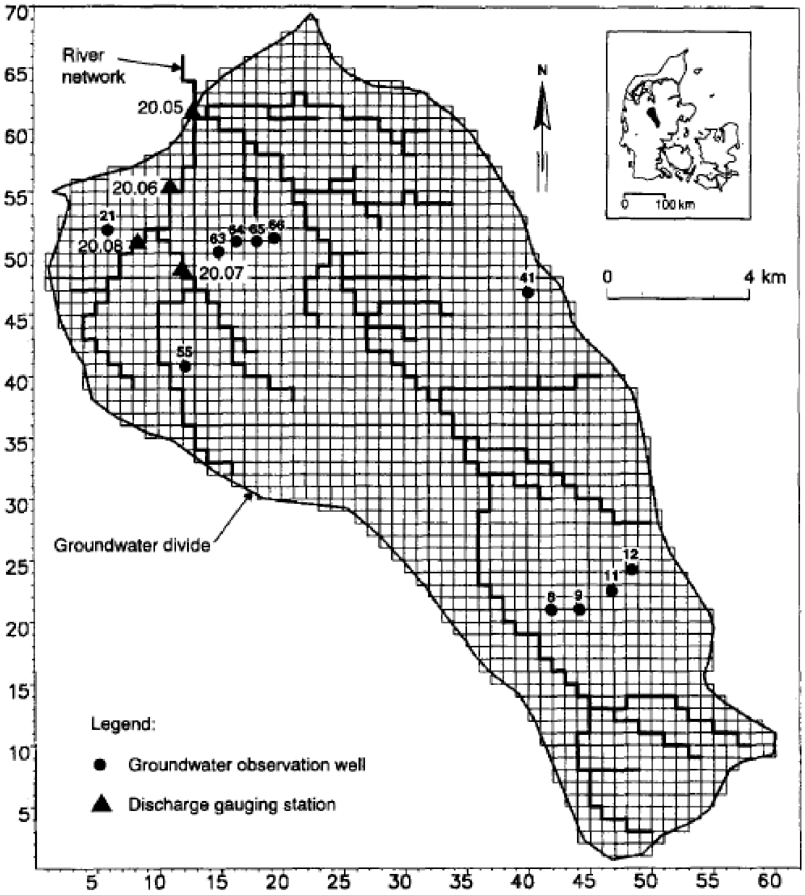


Figure 2.3 – Discharge gauging stations and groundwater observation wells of the Karup catchment (Refsgaard, 1997).

Validation of model using discharge stations not used for calibration resulted in poor performance. There was a clear underestimation of the baseflow level and total runoff. The multi-site validation not only showed inadequacy in the internal model validity, but also provided guidance in the reason for inadequacy – inaccurately simulated groundwater levels.

2.2.4 Crash Testing Concepts in Model Validation

The holy grail of hydrologic modelling has been achieving a degree of process understanding that enables development of model that provides physically realistic simulations across different hydrologic environments, and at multiple spatial and temporal scales (Gupta et al., 2014). The poor performance of models under proxy-basin tests further strengthens the difficulty in achieving this holy grail. Andressian (2006) illustrates the need to take advantage of the extensive data sets now available to make common a large-sample approach to hydrologic investigations. Large-sample hydrology tests can demonstrate robustness of the models, demonstrating the capabilities in regional and temporal transposability (Gupta et al., 2014). Andressian (2009) claims that hydrologic model testing should be similar to crash testing cars. During crash tests, cars are tested in conditions outside of intended use. The results are then interpreted by the end user of the car, allowing a choice in car based on the needs of the drivers. Only by testing hydrologic models under varying extreme conditions, can the model users fully understand the reliability, capabilities and limitations of the model. The rigorous nature of crash tests require hydrologic realism in the models. As a result, new model structures and processes can be identified during testing (Gupta et al., 2014).

Coron et al. (2012) performed a crash test on hydrologic models using contrasted climate conditions in 216 Australian catchments. Three models, GR4J, MORDOR6, and SIMHYD Plus Routing were crash tested in these catchments. Through the crash testing, the authors aimed to study the transfer of model parameter sets between climatically contrasted periods. By performing the GSST on all catchments, the impact of both the spatial and temporal variability in climate on parameter transposability was tested. Large-sample testing methods were shown to be effective in testing model transposability, a key requirement for models to be used to synthesize data for which data is unavailable. Although the benefits of large-sample hydrology are clear, many challenges exist for practical implementation of such evaluation methods. Large-sample hydrology requires extensive volumes of relevant data sets. Often, such extensive volume of data sets are difficult to acquire. Hydrologic data need to become more accessible through increase in more coherent reporting, storing, and sharing of data. Alternatively, depending on the required complexity of the catchment, hydrologic data may be synthesized (Mirus et al., 2012).

2.3 Improved Model Validation Methods

In this section the limitations and issues with the traditional and existing validation techniques presented above are discussed. Alternative validation techniques that address the limitations and issues are introduced.

2.3.1 Issues with Current Model Validation Methods

Philosophically speaking, Popper (1968) argued that models can never be truly validated, but only invalidated. Konikow and Bredehoeft (1992) demonstrated insufficiency in current validation practice using case studies of failed decision making by *validated* models. Konikow and Bredehoeft (1992) claimed that model validation is merely a process used to organize our thinking, test ideas for their reasonableness, and indicate which the sensitive parameters are. More rigorous evaluation methods, such as crash tests and structural adequacy tests, are no different. Passing these evaluation methods can give increased confidence in the model, but never give absolute confidence in the validity of the model. The issue with the current evaluation practice can be addressed by the basic concept introduced by Klemes (1986) 30 years ago: models need to be tested for their intended use. Despite the intention of models to aid in decision making, policy management, and water resources management, models are only tested rigorously in the ability to match historical observations. Large-sample testing methods are fundamentally no different. Large-sample testing methods are more rigorous ways to test a model’s ability to match historical data under varying conditions – not a tool to assess a model’s capability to support decision making, policy management, and water resources management.

2.3.2 Using Models for Decision Making

Despite the advancement in science, there still exists a gap between environmental science and decision making. Science and policy serve different purposes, resulting in different values, interests, concerns, and perspectives between the scientists and policy makers. These differences complicates the communication between the two parties, degrading the value of models in decision making process. One barrier between environmental modelers and policy makers is the results of scientific models not being available in the form required by the decision makers (Jacobs, 2002). Hydrologic model output variable Y may not be readily transformable into metric required by the decision makers. Hence, collaboration between scientists, decision makers, and stakeholders is crucial to transpose model output into clear and comprehendible metric. Another barrier is the lack of uncertainty analysis in environmental model applications. Accurate uncertainty analysis is required to effectively characterize errors and limitations of the model. Liu (2008) claims that model output uncertainty should be transferred over to decision making scenario analysis, to (1) understand impacts stemming from alternative conditions; (2) to assess potential risks and opportunities; and (3) to identify ways to respond to risks and opportunities, thus enabling improved decision making and assessment. With the two barriers in mind, Liu (2008) proposes a framework in linking science with environmental decision making. The framework is comprised of 9 steps: problem formulation; scenario definition; conceptual modelling; model development; verification, calibration, and validation; model simulation/scenario construction; scenario analysis and assessment; implementation/decision making; and monitoring and post audit. Few concepts from Liu’s framework highlights the key requirements for DCT. A clear outline of the decision context in both natural and human aspects is drawn, and stakeholders, scientists, and policy makers work together to develop a clear mapping of model output and decision making. During the verification, calibration, and validation process, Liu (2008) emphasizes that the performance criteria needs to be tailor-made to the specific decision context. A tailor-made performance criteria is desirable as it tests the model for its intended use.

Chapter 3

Methods and Results of Modelling Canadian Shield Hydrology

In this section, strategies used to address issues with modelling Canadian Shield hydrology are discussed. Model structure and modelling strategies are deployed to two sites: Kaministiquia Watershed and Lake of the Woods Watersheds. Model performance against an alternative and commonly employed model structure is presented.

3.1 Model Development

This section details of the overall model structure and explicit representation of Canadian Shield hydrology characteristics are discussed.

3.1.1 Model Structure

Raven is a hydrologic modeling framework that allows various model configurations, from conceptual to physically-based and from lumped to fully-distributed (Craig et al., 2018). Raven's modular design allows customization of hydrologic processes and forcing inputs for model development adequate for site and application. Raven's physical representation of lakes and various reservoir operation functions made Raven suitable as the modelling platform for a case study of reservoir rule curve selection. The hydrologic model structure follows closely, but not exactly, the multi-soil model developed by Robert Chlumsky at the University of Waterloo (Chlumsky 2017). The hydrologic process map is shown in Figure 3.1. Precipitation inputs are distributed into rainfall and snowfall based on temperature inputs. Then, precipitation is distributed across state variables, including lake storage, canopy, snow, ponded water, depression, surface storage, and soils. Water is redistributed across state variables through various hydrologic processes, then final flow is calculated through catchment routing.

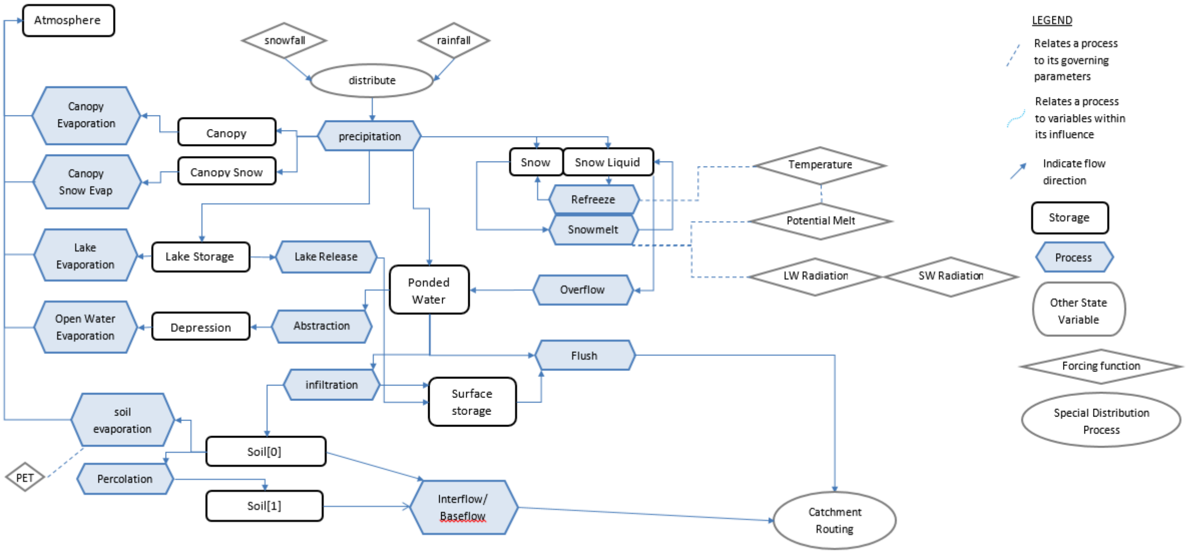


Figure 3.1 – Hydrologic process diagram of the Raven model adopted from R.Chlumsky (2017).

The model structure was setup to have either a bedrock outcrop (modelled with very thin soil) or deeper organic soil (with two soil layers) in a given sub-basin in order to accommodate the Canadian Shield landscape characterized by fractured bedrock layer under shallow soil layers. The depth of the fractured bedrock layer also acts as an extra calibration parameter, where extra storage in the fractured bedrock can help account for extra storage present in the landscape contributing to flow but not accounted for in the model, such as depressions and wetlands (Chlumsky, 2017). The conceptual soil profiles are shown in Figure 3.2. Model processes are relatively simple, a function of limited data availability of the case study sites. With additional data, such as measured radiation and snow depth and density, complex energy driven snow balance may be more adequate. Full input file with process description is in Appendix A.

3.1.2 Explicit Representation of Canadian Shield Hydrology Characteristics

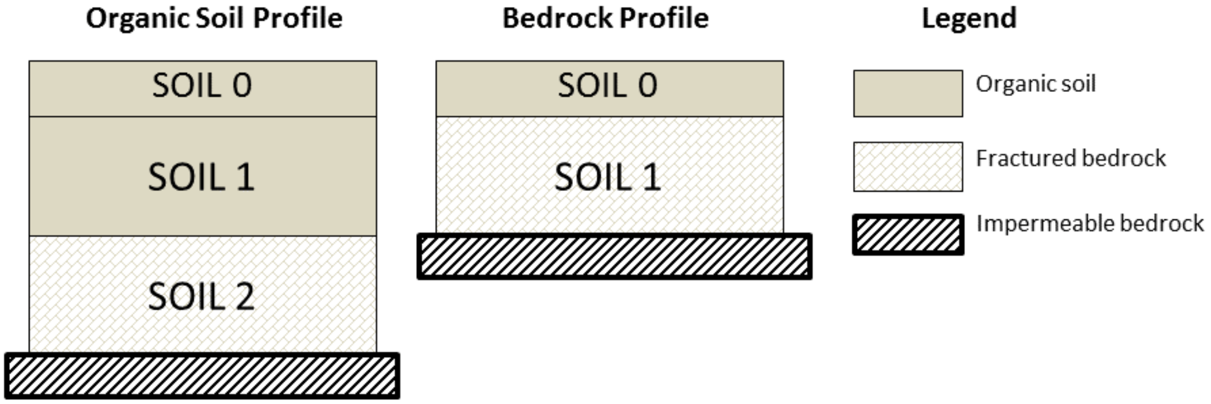


Figure 3.2 – Conceptual diagram of soil profiles in the model (Chlumsky, 2017)

In Section 2.1.3, the impact of lake systems on the hydrograph and the need for an explicit representation of the lake systems was discussed. However, an explicit representation of a complex system of hundreds of small and inter-connected lakes is often unrealistic due to the limitation in data. For a simplified representation of such a complex lake system, a hydrologically equivalent lake (HEL) concept was developed. The HEL is similar to the hydrologic equivalent wetland (HEW) concept in the SWAT model, which is a synthetic wetland module developed to mimic the conveyance and retention of wetland storage (Wang et al., 2008). HEL is a hypothetical lake that mimics the hydrologic response of the aggregate lake system. Whenever a flow gauge station is impacted by an upstream system of lakes based on GIS analysis and hydrograph analysis, a HEL was implemented directly upstream of the gauge. The GIS analysis involves inspecting for lakes above a threshold size connected to the gauge within proximity. The hydrograph analysis involves visual assessment of smoothness of the hydrographs, potentially caused by upstream lakes. Two user inputs are required for the HEL. The area of the HEL was calculated by summing up the area of the lakes above a threshold size

($10\ \mathrm{km}^2$). The outlet of the HEL was assumed to follow a rectangular weir-like structure shown in Figure 3.3. The rectangular weir equation is shown in equation 5.

$$Q = \frac{2}{3} CW\sqrt{2g} s^{1.5} \qquad (5)$$

where C is the weir coefficient (1.6 for rectangular weir), W is the weir width (m), g is the gravitational acceleration ($9.8\ \mathrm{m/s}^2$), and s is the height above weir crest (m). The weir width of the outlet for a HEL was set as a calibration parameter, restricted to reasonable range based on GIS analysis (?) of the largest lake represented in the HEL.

Figure 3.3 shows a schematic of the simulated variables associated with a HEL or an explicitly represented lake. ***Define all variables.

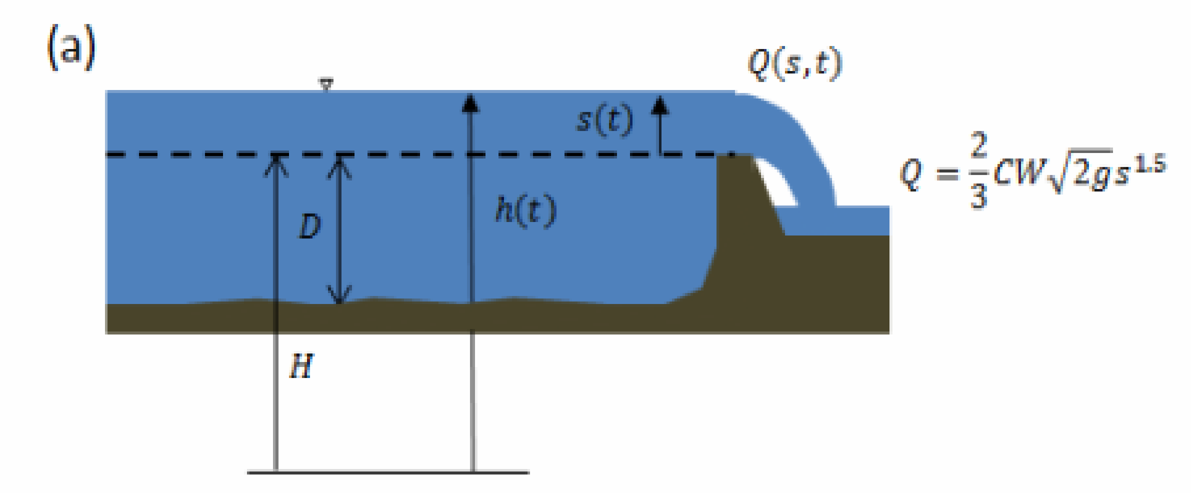


Figure 3.3 – Rectangular weir-like outlet for HEL

3.1.3 Modelling Operation in Managed Reservoir

A common application for hydrologic models by conservation authorities and hydropower companies is inflow forecasting for reservoir management. Based on short term and long term inflow forecasts, reservoir operators can determine how much water needs to be used for electric generation to maximize electric generation and minimize negative impacts such as flood risks. For typical managed reservoirs, outflow data is readily available. Since the outflow is ultimately determined by the operators through adjustment of outlet structure, modelling the outflow using hydrologic model is nearly impossible without target values and regulations, as it would require modelling of human judgement. In historical modelling of managed reservoirs, the outflow becomes a forced outflow to the reservoir (model generated outflow of the reservoir is overridden with measured outflow), and model is calibrated to inflow. In order to test various reservoir operation strategies, a model needs to be capable of modelling the human decision in reservoir operations. Modelling human decisions requires guidelines and regulations that are assumed to be followed by operators.

In this thesis, a set of rules were applied to emulate operator controls. First, a target reservoir level which the operator aims to follow is required. In many authorities, upper and lower limits of stage as a function of time of year are provided through operational rule curves. Within the maximum and minimum reservoir levels for any given time of the year, an operator may decide to target different stage within the range, depending on current conditions and short term forecasts. For example, after a snowy winter, an operator may target a lower part of the band before spring to accommodate high spring melt. For the case study in this thesis, it was assumed that the operator follows the mid-point of the band. For operational implementation, the model was supplied with a time series of target stage levels over the simulation period. Mass balance of reservoir follows equation 6.

$$\frac{\Delta V}{\Delta t} = \overline{Q_{in}} - \overline{Q_{out}} - \overline{ET} - \overline{Ext} \qquad (6)$$

where Q_{out} is flow (m^3/s), Q_{in} is reservoir inflow, ET is evapotranspiration (m/s), and Ext is reservoir extraction (m^3/s), averaged over time step. Target flow is calculated by utilizing target stage to determine necessary change in volume. At each time step *t*, the model calculates target flow $Q_{target}{}^{t+1}$ based on target stage for time step $t+1$ using equation 7.

$$Q_{target}{}^{t+1} = -2 * \frac{V_{target}{}^{t+1} - V^{t}}{\Delta t} + \left(-Q^{t} + (Qin^{t} + Qin^{t+1}) - ET * (A^{t} + A^{t+1}) - (Ext^{t} + Ext^{t+1})\right) \quad (7)$$

Target values are calculated based on target stage at time step $t+1$.

Then, the target flow is averaged over the time step through equation 8.

$$Q_{target}{}^{t+1} = \frac{Q_{target}{}^{t+1} + Q^{t}}{\Delta t} \quad (8)$$

In most managed reservoirs, regulations and structural limitations restrict maximum and minimum flows from the outlet. Adequate research needs to be conducted to understand the regulations and structural limitations that the operators will follow to properly implement reservoir operations in hydrologic models for specific reservoirs. Four restrictions have been implemented in the model: minimum flow, minimum flow during drought, maximum flow based on stage, and maximum increase in flow over a time step. Summary of components involved in flow calculation is shown in Table 3.1.

Table 3.1 – Summary of components involved in flow calculations for reservoir operations

Symbol	**Description**	**Determination**
Q_{target}^{t+1}	Flow required to reach target stage at *t+1*	Equation 7 and 8 based on modelled values and target stage time series
$Q_{MinDrought}$	Minimum flow when stage is below drought level	Regulations
Q_{Min}	Minimum flow when stage is above drought level	Regulations
Q_{Max}	Maximum flow	Hydraulic study of the outlet structures in case study
Q_{delta}	Maximum increase in flow over 1 day	Regulations

Calculation of flow for time step t+1 is determined by process shown in Figure 3.4.

In a case where the flow is not restricted by regulations or outlet structure, maximum stage restriction values may be required to model operator behavior to keep reservoir stage within a limit. With a maximum stage constraint, outflow is calculated using equation 9.

$$Q_{out}{}^{t+1} = -2 * \frac{V_{limit}{}^{t+1} - V^{t}}{\Delta t} + \left(-Q^{t} + \left(Qin^{t} + Qin^{t+1}\right) - ET * \left(A^{t} + Alimit^{t+1}\right) - \left(Ext^{t} + Ext^{t+1}\right)\right) \quad (9)$$

where volume and area is calculated using limiting stage.

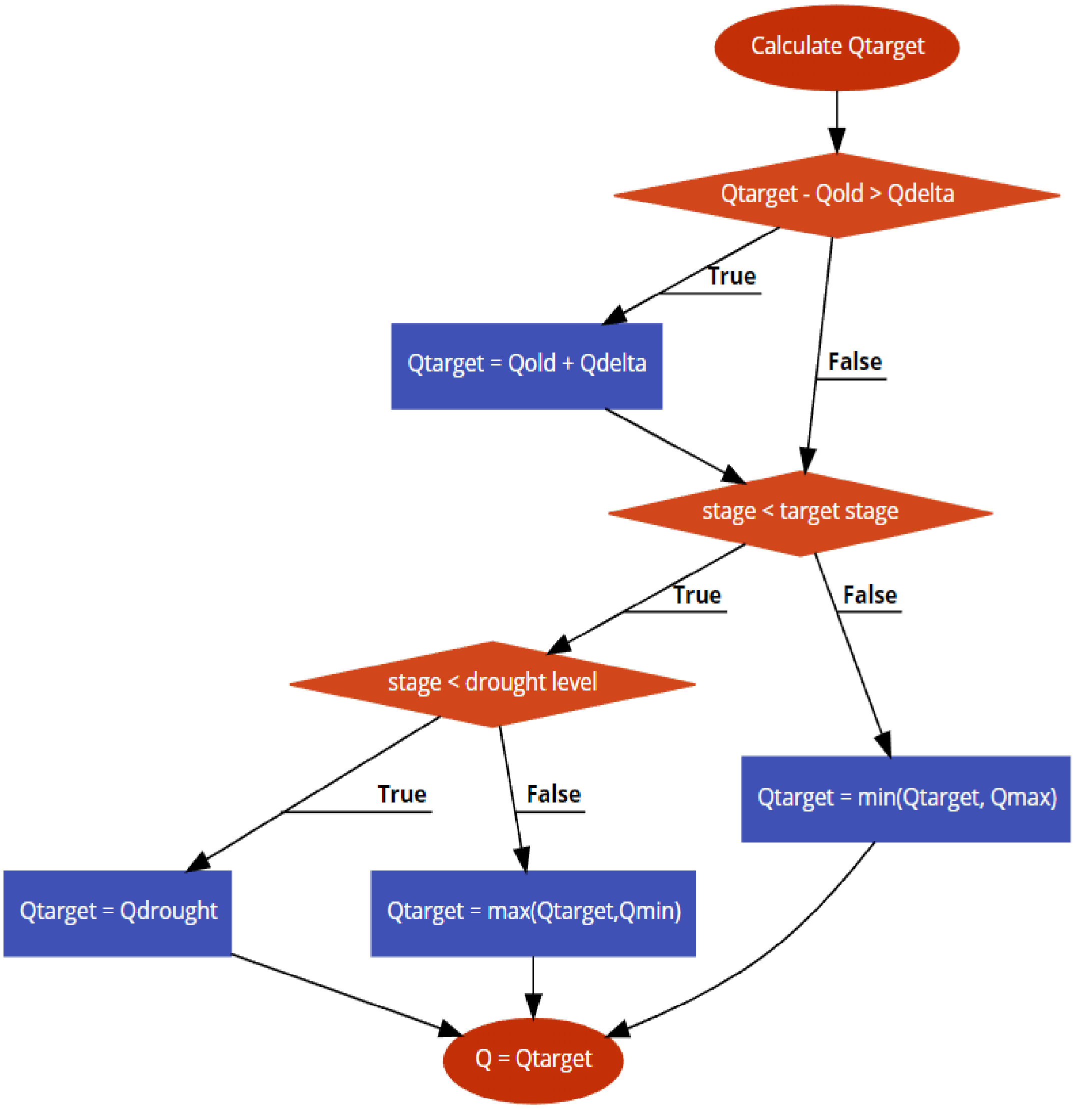


Figure 3.4 – Workflow to determine outflow to model reservoir operation

3.2 Kaministiquia Watershed Model

In this section, the first case study of Kaministiquia watershed is presented. During initial model development, impact of HEL representation was not tested in this watershed, as method was not fully developed during the initial model development. Later in the model development, HEL was implemented to improve stage simulation at Kashabowie Lake. Impact of reservoir operation modelling and calibration strategies is also discussed.

3.2.1 Kaministiquia Case Study Background

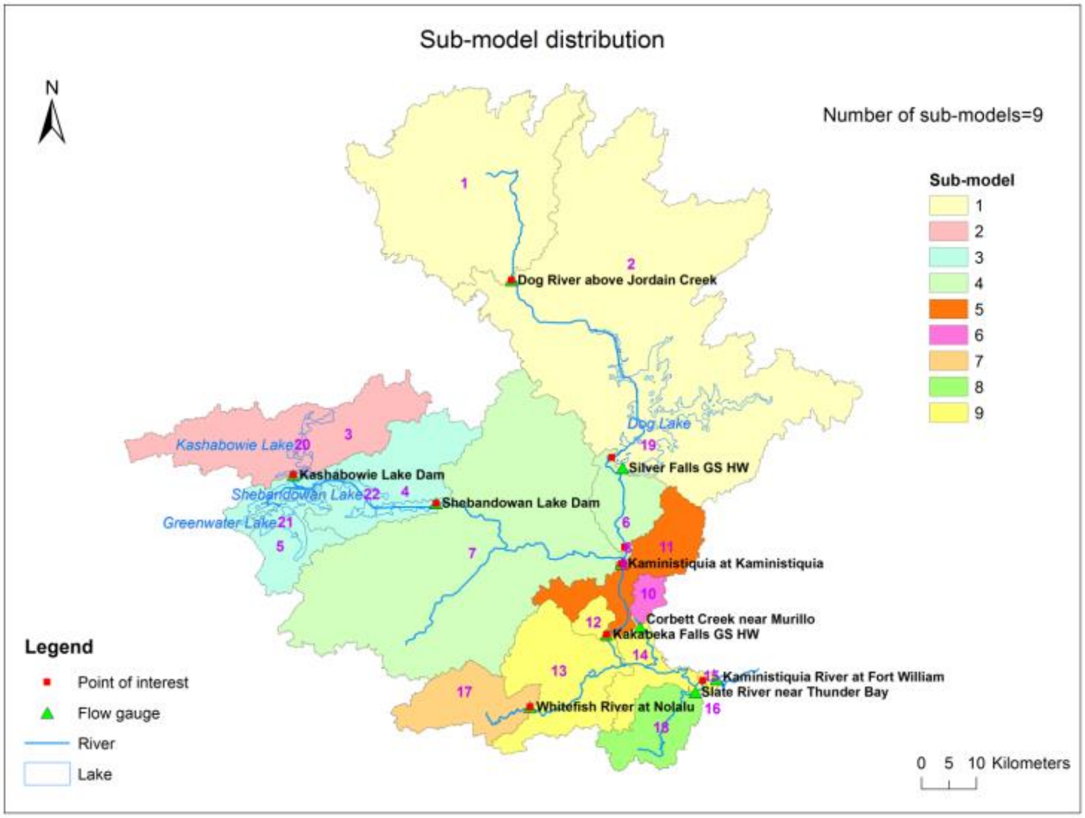
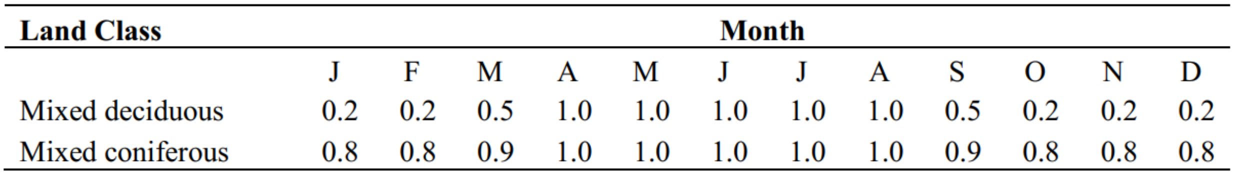


Figure 3.5 – Watershed delineation of the Kaministiquia watershed (Liu, 2017)

The Kaministiquia watershed is located west of Lake Superior, near Thunder Bay, Ontario. The watershed includes four dams and two generating stations managed by Ontario Power Generation (OPG). Initial Kaministiquia watershed model consists of 9 sub-models with varying number of sub-basins in each sub-model for a total of 27 sub-basins (Liu, 2017). During calibration and validation, outflow from upstream sub-basins were used as forced inflow for the sub-basin immediately downstream. The model was calibrated moving downstream at each of the 9 sub-model outlets. In forecast mode for use by OPG, the model will require predicted outflow data time series for the reservoirs (or a rule curve). Based on initial GIS work by Liu (2016) and Chlumsky (2017), the sub-basins were characterized by different soil layer type and vegetation. The two soil types are storage dominant (ABC2) and bedrock dominant (R1). Both soil types have a thin layer of soil at the top. The ABC2 profile is followed by a thicker layer of soil layer soil type with high permeability to allow water storage. The R1 type is followed by thick layer of soil layer with low impermeability to represent the bedrock layer. ABC2 profile is then followed by a thick layer of bedrock. Vegetation types are divided into deciduous forests and coniferous forests. The two vegetation classes have different seasonal leaf area index fraction for each month as shown in Figure 3.6. Model structure and hydrologic processes follow Section 3.1.1. Three sub-basins, Dog Lake Basin, Kashabowie Lake Basin, and Shebandowan Lake Basin, have managed reservoirs with outflow data. The model was calibrated to stage levels of the three reservoirs, with the measured outflows overriding modelled outflows. Model comparison against the GR4J model created by Liu (2017) is presented in the following section.



Land Class	**Month**											
	J	F	M	A	M	J	J	A	S	O	N	D
Mixed deciduous	0.2	0.2	0.5	1.0	1.0	1.0	1.0	1.0	0.5	0.2	0.2	0.2
Mixed coniferous	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.8	0.8	0.8

Figure 3.6 – Monthly leaf area index fraction (Chlumsky, 2016)

Since calibration was performed for each sub-model, land class and vegetation parameters differ from sub-model to sub-model. However, within a sub-model, basins with common land class or vegetation class share the same parameters. Table 3.2 summarizes each sub-basin with its vegetation class and soil profile.

Table 3.2 – Summary of sub-basins and physical characteristics

Submodel	Subbasin	Vegetation Class	Soil Profile
1	1	Mixed Deciduous	ABC2
	2	Mixed Deciduous	R1
	19	Lake	Lake
2	30	Mixed Coniferous	R1
	31	Mixed Coniferous	R1
	32	Mixed Coniferous	R1
	33	Mixed Coniferous	R1
	35	Lake	Lake
	36	Lake	Lake
	37	Lake	Lake
	39	Lake	Lake
3	4	Mixed Coniferous	R1
	5	Mixed Coniferous	R1
	21	Lake	Lake
	22	Lake	Lake
4	6	Mixed Deciduous	ABC2
	7	Mixed Deciduous	ABC2
	8	Mixed Deciduous	ABC2
	9	Mixed Deciduous	ABC2
5	11	Mixed Deciduous	ABC2
6	10	Mixed Deciduous	ABC2
7	17	Mixed Deciduous	ABC2
8	18	Mixed Deciduous	ABC2
9	12	Mixed Deciduous	ABC2
	13	Mixed Deciduous	ABC2
	14	Mixed Deciduous	ABC2
	15	Mixed Deciduous	ABC2

3.2.2 Kaministiquia Watershed Model Calibration Formulation

The model was calibrated from 2005-10-01 to 2012-10-01 with the first year as a warm-up period. In the Shebandowan Lake, the outlet structure was changed in 2009. As a result, the model calibration period was set from 2009-10-01 to 2012-10-01, with the first year as a warm-up period. Calibration was performed using the OSTICH Optimization Software Tool (Matott, 2017). Within OSTRICH, the Dynamically Dimensioned Search algorithm (Tolson and Shoemaker, 2007) was used for optimization, with a budget of 4,000 model runs which allowed convergence within reasonable computational cost.

Traditional calibration strategy involves calibration to estimated inflow data. Initial experiments showed poor performance in reservoir stage simulation when calibration to inflow. However, calibration to stage resulted in significantly improved stage simulation with a small deterioration in inflow simulation. As a result, calibration objective was formulated using reservoir stage to accurately capture both stage and inflow. Comparison of calibration to inflow is discussed further in Section 3.2.4. To capture fluctuation in stage during the calibration process, the NSE of change of reservoir stage at each time step (dh/dt) shown in equation 10 was incorporated into the objective function. Stage observations have much lower variance compared to flow observations. Incorporation of stage derivative into objective function resulted in better capturing of small fluctuations of stage. The objective function for calibration was set to maximizing the average of NSE of stage and NSE of change in stage over each time step. For sub-basins without reservoirs, model was calibrated to maximize the NSE of flow.

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(\frac{(h_{i+1}^{obs}-h_{i}^{obs})}{dt}-\frac{(h_{i+1}^{sim}-h_{i+1}^{sim})}{dt}\right)^2}{\sum_{i=1}^{n}\left(\frac{(h_{i+1}^{obs}-h_{i}^{obs})}{dt}-\overline{\frac{(h_{i+1}^{obs}-h_{i}^{obs})}{dt}}\right)^2} \tag{10}$$

3.2.3 Kaministiquia Watershed Model Results

Summary of model performance during calibration period is shown in Table 3.3.

Table 3.3 – Summary of Kaministiquia Watershed model calibration results when calibrated to NSE with inflow forced gauges bolded

Sub-model No.	Flow/Stage Gauge Name	Simulation Object	Calibration Period					
			GR4J			Multi Soil		
			NSE	NSE (dh/dt)	PCT_BIAS	NSE	NSE (dh/dt)	PCT_BIAS
1	Silver Falls GS HW - Dog Lake	Stage	-0.15	0.45	-1	0.85	0.52	0
2	Kashabowie Lake Dam	Stage	-3.1	-0.05	1	-0.47	-0.29	0
3	**Shebandowan Lake Dam**	**Stage**	**-0.81**	**0.35**	**-1**	**0.63**	**0.41**	**0**
4	**Kaministiquia at Kaministiquia**	**Flow**	**0.93**	**--**	**4**	**0.92**	**--**	**-1.6**
5	**Kakabeka Falls GS HW**	**Flow**	**0.98**	**--**	**3**	**0.98**	**--**	**2.8**
6	Corbett Creek near Murillo	Flow	0.66	--	19	0.63	--	9.9
7	Whitefish River at Nolalu	Flow	0.72	--	22	0.66	--	-27.7
8	Slate River near Thunder Bay	Flow	0.6	--	41	0.65	--	-10.8
9	**Kaministiquia River above Fort William**	**Flow**	**0.97**	**--**	**1**	**0.98**	**--**	**-1.4**

The physically-based multi soil model performed much better than the GR4J model in simulating stage during the calibration period. In calibration to flow gauges, the multi soil model performed slightly worse in Kaministiquia, Corbett Creek, and Whitefish, and performed better in Slate River and Fort William. Biggest difference occurred at Whitefish River at Nolalu (difference of 0.06 NSE).

Model validation period was set to 2012-10-01 to 2015-10-01. Summary of model performance during validation period is shown in Table 3.4.

Table 3.4 – Summary of Kaministiquia Watershed model validation results when calibrated to NSE with inflow forced gauges bolded

Sub-model No.	Flow/Stage Gauge Name	Simulation Object	Validation Period					
			GR4J			Multi Soil		
			NSE	NSE (dh/dt)	PCT_BIAS	NSE	NSE (dh/dt)	PCT_BIAS
1	Silver Falls GS HW - Dog Lake	Stage	-0.03	0.56	-14	-0.97	0.75	24
2	Kashabowie Lake Dam	Stage	-11.7	-0.73	-10	-60	-1.49	33
3	**Shebandowan Lake Dam**	**Stage**	**-426**	**-0.22**	**68**	**-57**	**0.41**	**23**
4	**Kaministiquia at Kaministiquia**	**Flow**	**0.9**	**--**	**7**	**0.92**	**--**	**-0.2**
5	**Kakabeka Falls GS HW**	**Flow**	**0.97**	**--**	**0**	**0.98**	**--**	**-0.2**
6	Corbett Creek near Murillo	Flow	0.79	--	30	0.80	--	27.5
7	Whitefish River at Nolalu	Flow	0.65	--	21	0.61	--	-17.9
8	Slate River near Thunder Bay	Flow	0.74	--	46	0.75	--	-1.6
9	**Kaministiquia River above Fort William**	**Flow**	**No Validation Data Available**					

Model showed poor performance in stage simulation at all three lakes during the validation period. Model showed aggregating volume error in stage simulation, increasing in high percent bias values. The multi soil model performed better than the GR4J model across all basins in simulating flow, except in the Whitefish River.

3.2.4 Additional Strategies to Improve Reservoir Simulation

To correct the poor performance in validation period, simulation of reservoir operation was implemented to the model by restricting the maximum stage. Corrected validation results are shown in Table 3.5.

Table 3.5 – Summary of Kaministiquia Watershed maximum stage constraint corrected model validation results when calibrated to NSE

Sub-model No.	Flow/Stage Gauge Name	Simulation Object	Validation Period					
			Maximum Stage Constraint Corrected			Initial Model		
			NSE	NSE (dh/dt)	PCT_BIAS	NSE	NSE (dh/dt)	PCT_BIAS
1	Silver Falls GS HW - Dog Lake	Stage	0.59	0.54	0	-0.97	0.75	24
2	Kashabowie Lake Dam	Stage	-0.76	-0.55	0	-60	-1.49	33
3	Shebandowan Lake Dam	Stage	-7	0.18	0	-57	0.12	23

Figure 3.6 shows significant improvement in NSE for the multi-soil model compared to the GR4J model during the calibration period at Dog Lake.

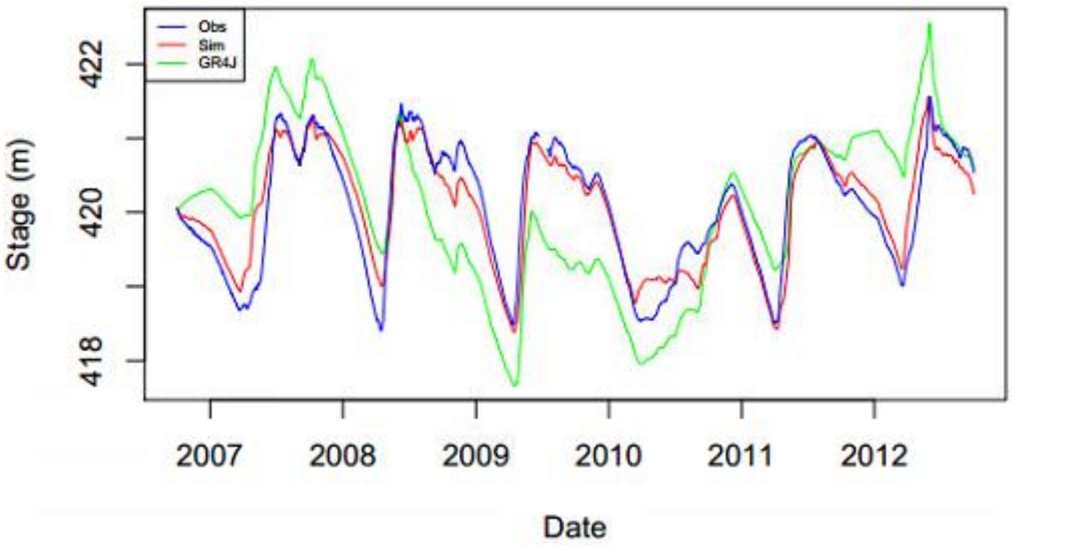


Figure 3.6 – Plot of Dog Lake stage during calibration period when calibrated to average of stage NSE and dh/dt NSE

However, initial validation results showed volume error of 24% and a NSE of -0.97. One approach to correct the error was to apply a precipitation correction of 0.93. Multiplying all precipitation by 0.93 produced good fit in stage graphs during the validation period, by reducing the total volume of water coming in to the basin, as shown in Figure 3.6.

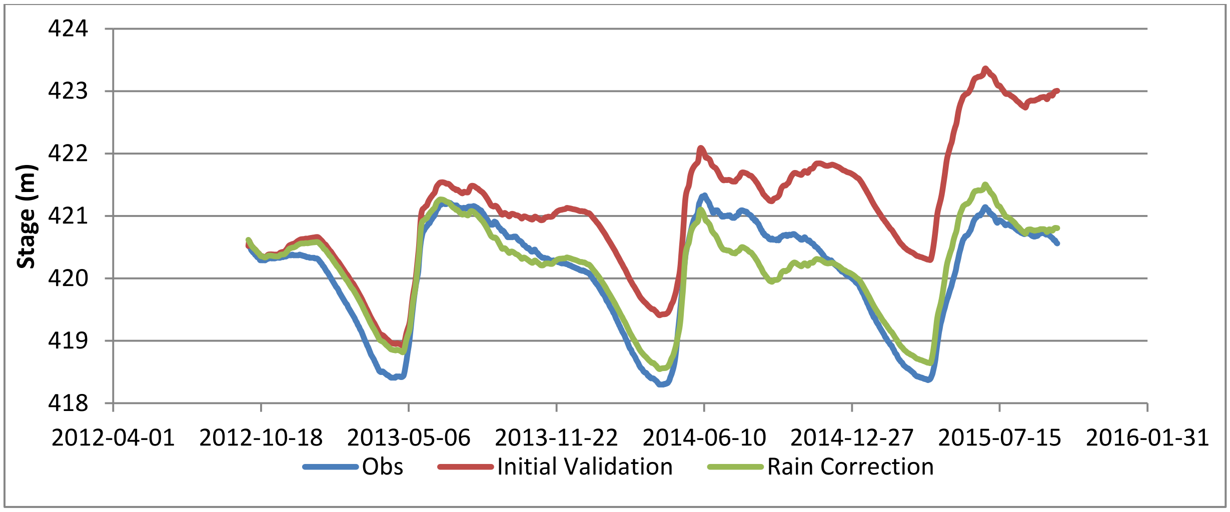


Figure 3.7 – Stage plot of rain corrected Dog Lake model in comparison with initial validation results

Second approach to correct the error was to model operator behavior using maximum stage constraint.

10 years of historical data showed annual maximum stage to be consistent near 421.56. Such result is most likely due to operational decision. To model the operator decisions, stage of the reservoir was set to 421.56 m. Figure 3.8 shows significant improvement in validation results.

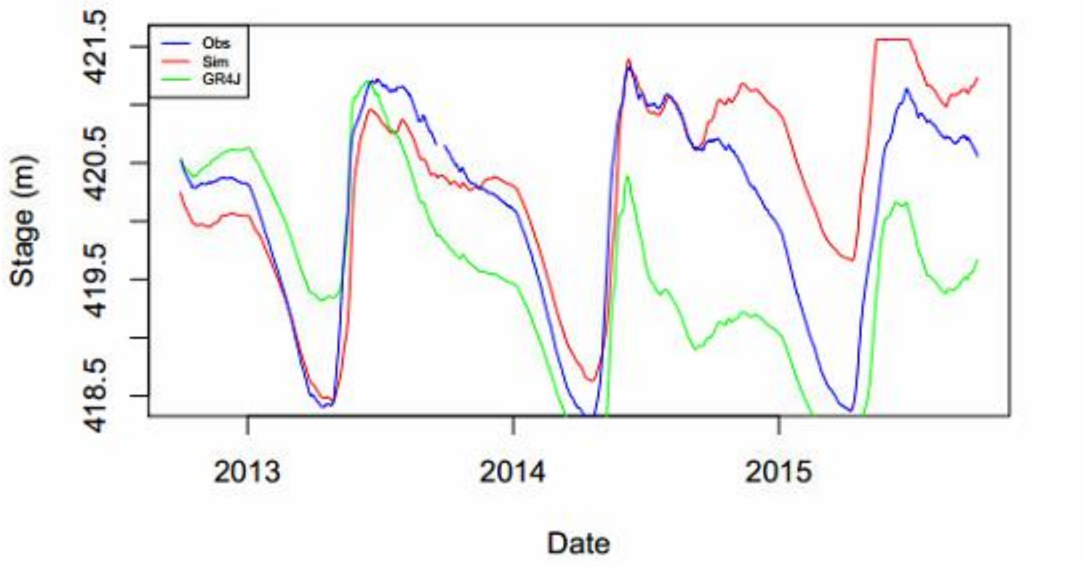


Figure 3.8 – Plot of Dog Lake stage during validation period when calibrated to average of stage NSE and dh/dt NSE

Similar to Dog Lake, a maximum stage constraint of 459.7 m was applied to Kashabowie Lake. Stage variance at Kashabowie Lake was less than 1 m. Modelling stage of reservoirs with low variance was a challenging task. A maximum stage constrain of 450.6 m was applied to Shebandowan Lake. Observation data during the validation period was flagged as possibly erroneous by OPG. Plots of stage simulation for Kashabowie Lake and Shebandowan Lake are shown in Figure 3.9 to 3.12. Kaministiquia River sub-basin was modelled with forced inflows coming from Shebandowan Lake and Dog Lake. Kakabeka Falls sub-basin used outflow from Kaministiquia river sub-basin as forced inflows. Kaministiquia River near Fort William sub-basin used outflows from Kakabeka Falls, Corbett Creek, Whitefish River, and Slate River as forced inflows. Hydrographs of individual downstream sub-basins are shown in Appendix B

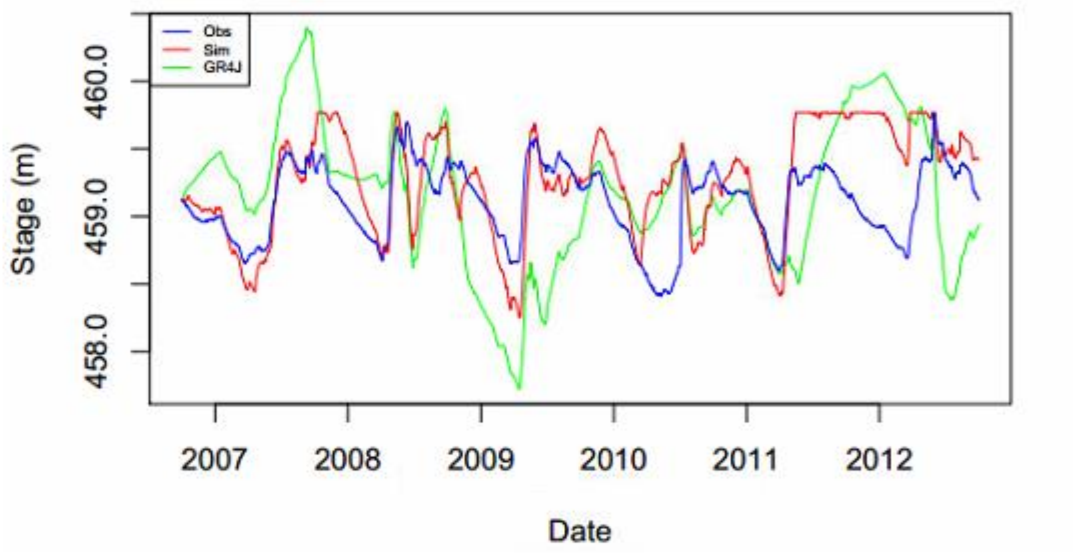


Figure 3.9 - Plot of Kashabowie Lake stage during calibration period when calibrated to average of stage NSE and dh/dt NSE

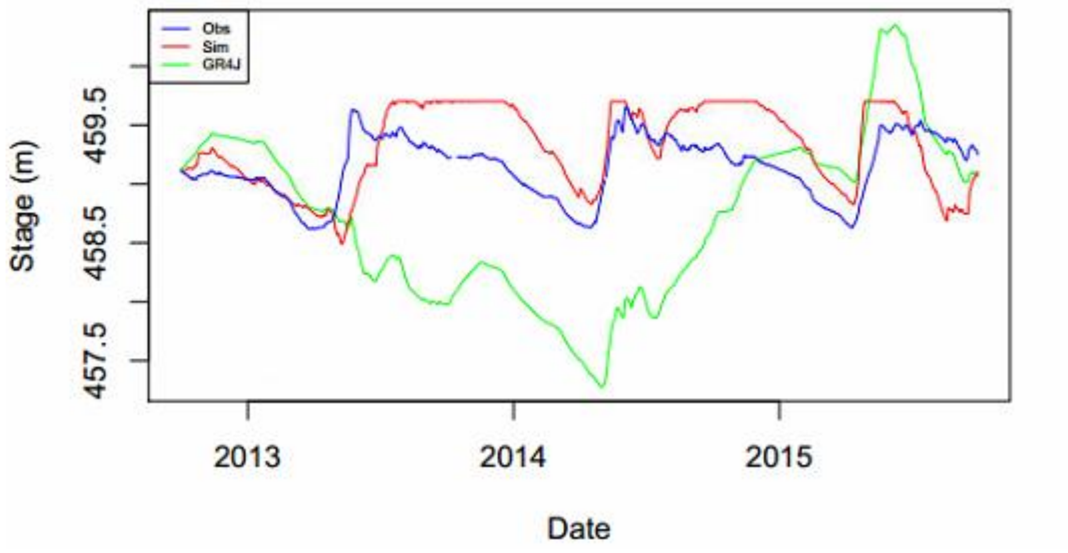


Figure 3.10 - Plot of Kashabowie Lake stage during validation period when calibrated to average of stage NSE and dh/dt NSE

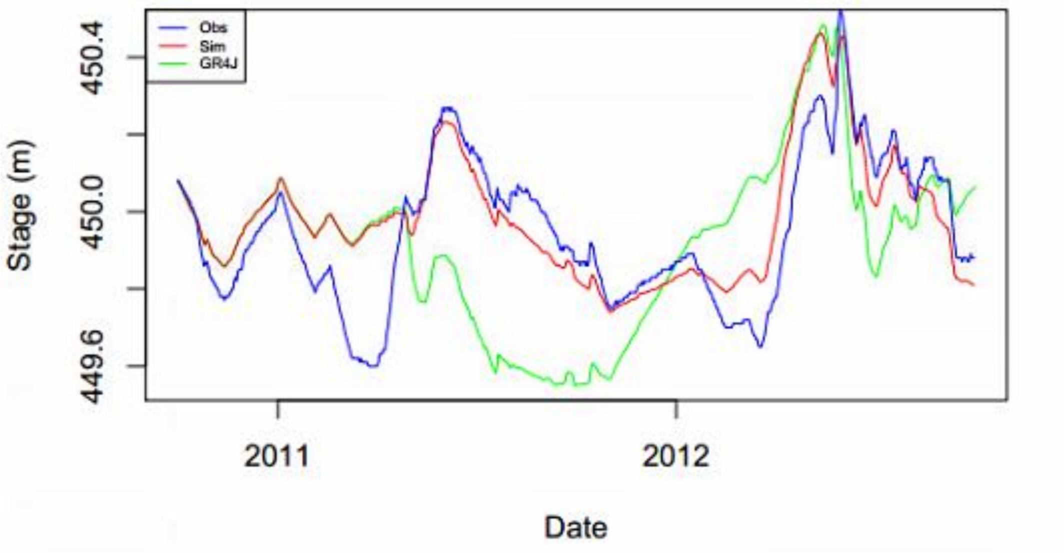


Figure 3.11 - Plot of Shebandowan Lake stage during calibration period when calibrated to average of stage NSE and dh/dt NSE

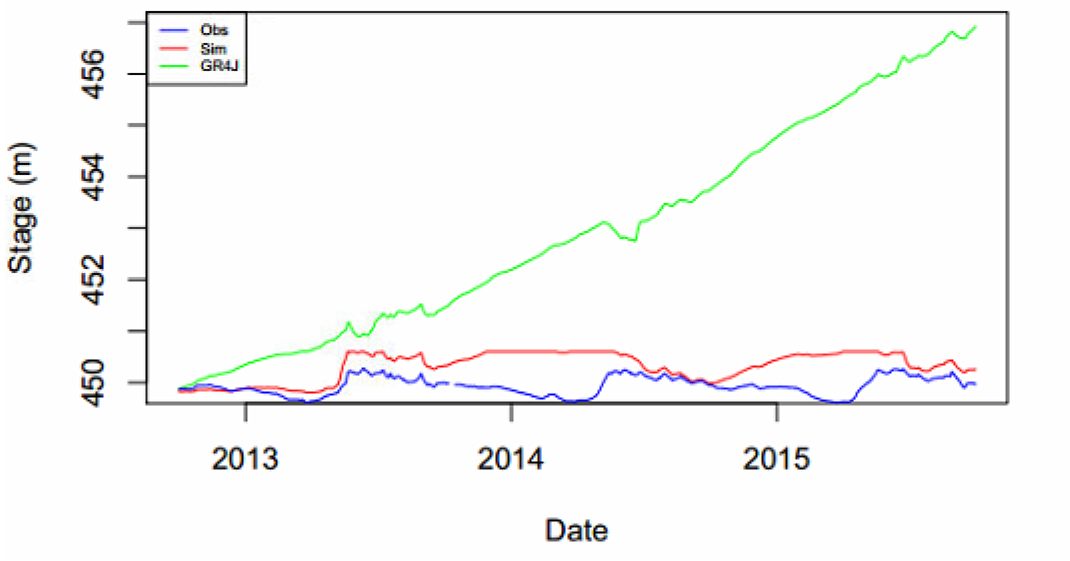


Figure 3.12 - Plot of Shebandowan Lake stage during validation period when calibrated to average of stage NSE and dh/dt NSE

Another calibration strategy for managed reservoir is to calibrate to measured inflows calculated from measured stage value. This may be an interest to parties utilizing model for inflow forecasts. Such strategy optimizes a model's ability to produce hydrologic variable of intended use. Reservoir inflow can be calculated using volume derived from stage using equation 11.

$$V_t = Q_{in} - Q_{out} - Ext + (P + ET) * A + V_{t-1} \tag{11}$$

Three sub-basins with explicit representation of reservoirs were calibrated to estimated inflows. Result shows that calibration to inflow show better inflow NSE, but significantly worse stage NSE. Figure 3.13 show inflow results when calibrated to stage and calibrated to inflow. Calibration to inflow had a NSE of 0.68 for inflow during the calibration period, where calibration to stage had a NSE of 0.55 for inflow.

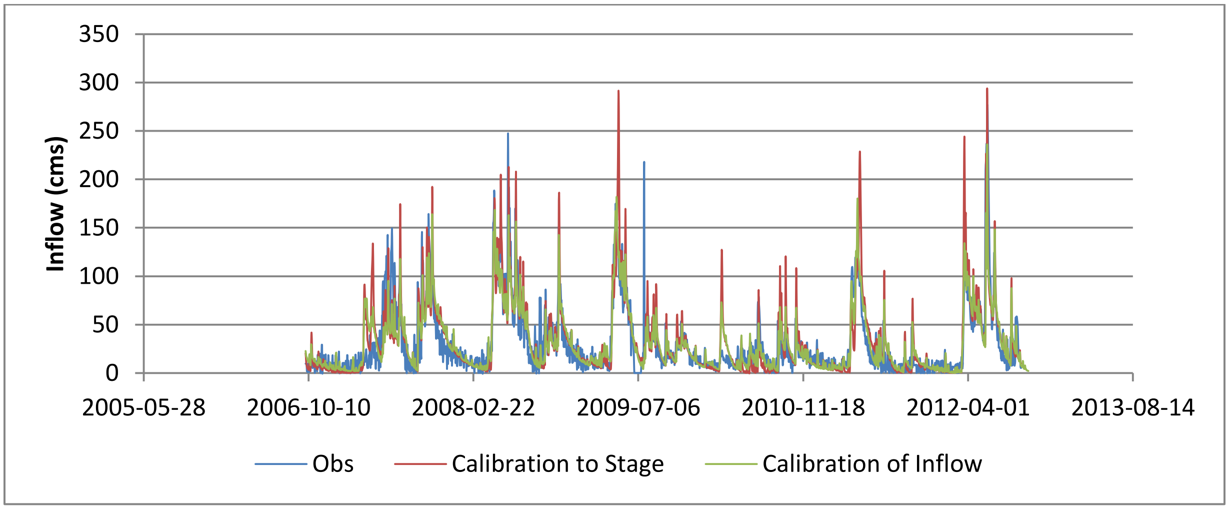


Figure 3.13 – Inflow hydrograph of Dog Lake when calibrated to stage and inflow

Inflow hydrograph generated from calibration to stage was much smoother during low flow seasons, with higher peak estimations compared to inflow hydrograph generated from calibration to stage. Inflow hydrograph generated from calibration to stage had better fit to observation inflow hydrograph during seasons with consistently high inflows, such as spring of 2008 and 2012. NSE for stage was 0.83 when calibrated to stage and -8.62 when calibrated to inflow. Figure 3.26 shows drastic underestimation of stage when calibrated to inflow

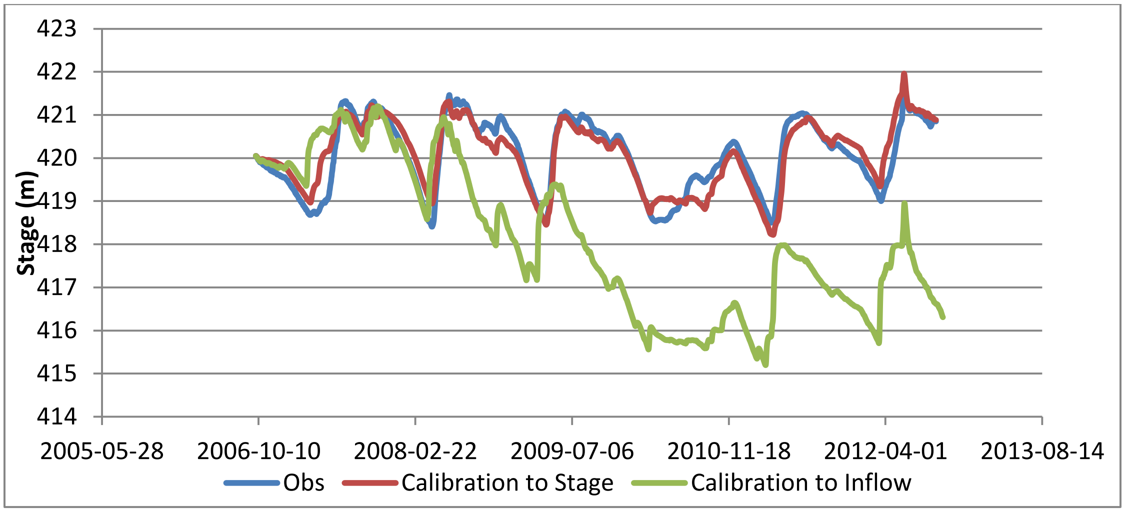


Figure 3.14 – Stage plots of Dog Lake when calibrated to stage and inflow

The calibration experiment shows calibrating to stage adds another level of complexity to the objective function, ensuring model is producing the right results for the right reason. In any model application where reservoir stage is important, such as flood prediction and land data assimilation systems, it would be extremely beneficial to utilize stage data in the calibration objective. However, if the sole intent of the model is to forecast inflows, it may be more beneficial to use an inflow calibrated model with stage adjustments made before each forecasts.

3.3 Lake of the Woods Watershed Model

In this section, the second case study of Lake of the Woods watershed is presented. Model utilized an existing WATFLOOD model as a base case. Model was transposed into Raven for model inter comparison and utilization in rule curve study.

3.3.1 Lake of the Woods Case Study Background

The Lake of the Woods – Rainy Lake (LOWRL) basin is located west of Lake Superior, bordering Manitoba and Minnesota. Three reservoirs: Lake of the Woods, Rainy Lake, and Namakan Lake, are managed by the Lake of the Woods Control Board (LWCB) under regulations and rule curves set by the International Joint Commission (IJC) to mandate water usage and watershed protection for the benefits of both Canada and the United States. As part of the daily reservoir operation, the LWCB (2016) has developed a hydrologic model using WATFLOOD to forecast inflows. The extent of the WATFLOOD model and the watershed location are shown in Figure 3.15.

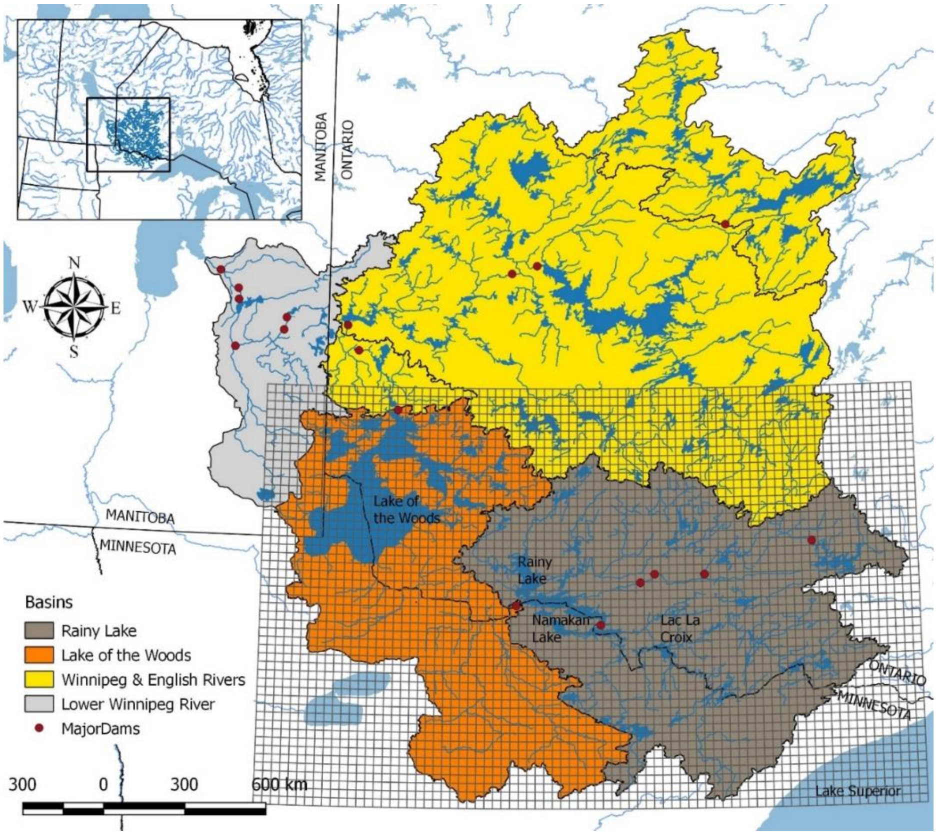


Figure 3.15 – Map of extent of LOWRL watershed

In this case study, the WATFLOOD model was transposed into a Raven model for utilization in rule curve study. A secondary objective of this process was to assess the ease of development and performance of Raven model created from geospatial data transposition of WATFLOOD inputs to Raven inputs. Other than the initial geospatial work to generate a semi-distributed watershed delineation, no additional geospatial data was required. The WATFLOOD model included 12 flow gauges and 4 inflow stations, with 34 reservoirs explicitly represented. Few flow gauges were located immediately downstream of the reservoirs. For the Raven model, sub-basins were explicitly delineated with outlets located at the flow gauges and reservoir outlets in the WATFLOOD model. The delineations are shown in Figure 3.16. To ensure geospatial consistency between the models, cumulative drainage areas at the 12 flow gauges were calculated. Table 3.6 shows that the areas between the two models are within a reasonable margin of error.

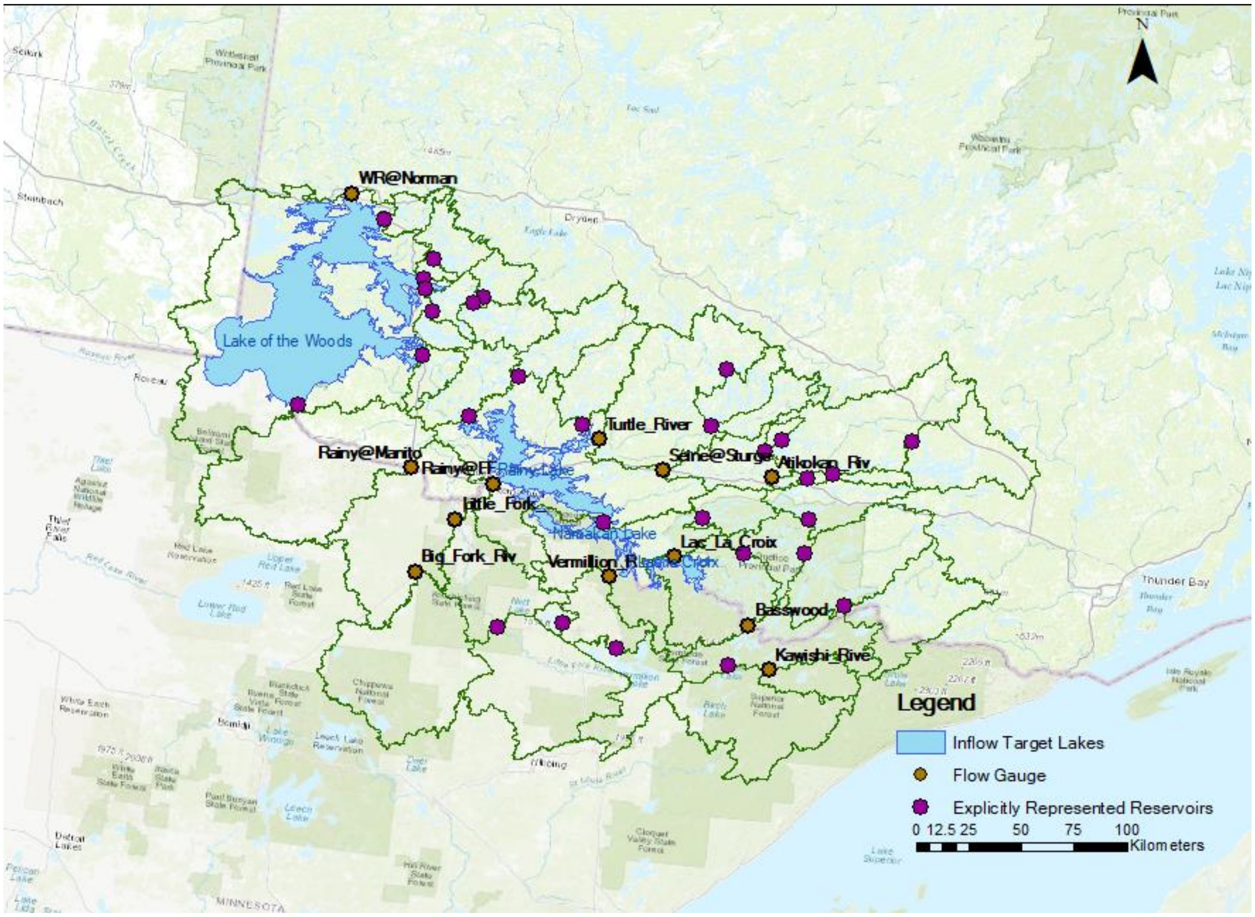


Figure 3.16 – Basin delineation used for Raven model

The original WATFLOOD model had 10 land class data in gridded format. Out of the 10, data on 9 of the land classes (agriculture, coniferous, deciduous, mixed, sparse, regenerating, wetland, water, and impervious) were used as Raven input. Mining was excluded as the fraction of land class area was nearly zero.

Table 3.6 – Cumulative drainage areas of WATFLOOD model and Raven model (km^2)

Station	WATFLOOD	Raven	Percent Error
Kawishi_Rive	611	638	4%
Basswood	4711	4475	-5%
Lac_La_Croix	12902	13220	2%
Vermillion_R	2413	2351	-3%
Atikokan_Riv	387	347	-10%
Seine@Sturge	5732	5899	3%
Turtle_River	4631	4742	2%
Rainy@FF	37249	38090	2%
Big_Fork_Riv	3661	3817	4%
Little_Fork_	4633	4667	1%
Rainy@Manito	48867	49948	2%
WR@Norman	67601	69311	3%

Each grid in WATFLOOD had fractions to represent the relative composition of each land class in each grid cell. Each grid cell was assigned a sub-basin number to match the delineation for the Raven model as shown in Figure 3.17. For a given Raven sub-basin, the total land class composition was calculated based on the WATFLOOD land class fraction data at corresponding grid cells with the sub-basin number. Similarly, bank-full area and channel slope data was available at each grid cells. For each sub-basin, grids with bank-full areas greater than 100 were assumed to be a part of the main channel in the sub-basin. The bank-full areas and channel slopes of the corresponding grid cells in the sub-basin were used to calculate the Raven channel profile at each sub-basin. Few parameters values required by Raven had equivalent counterparts in WATFLOOD. The summary of parameters with WATFLOOD counterparts that did not need additional calibration is shown in Table 3.7. Forcing functions for each sub-basin followed an average of all values with grids with the corresponding sub-basin number. Each reservoir in WATFLOOD was represented as an explicit lake sub-basin in Raven.

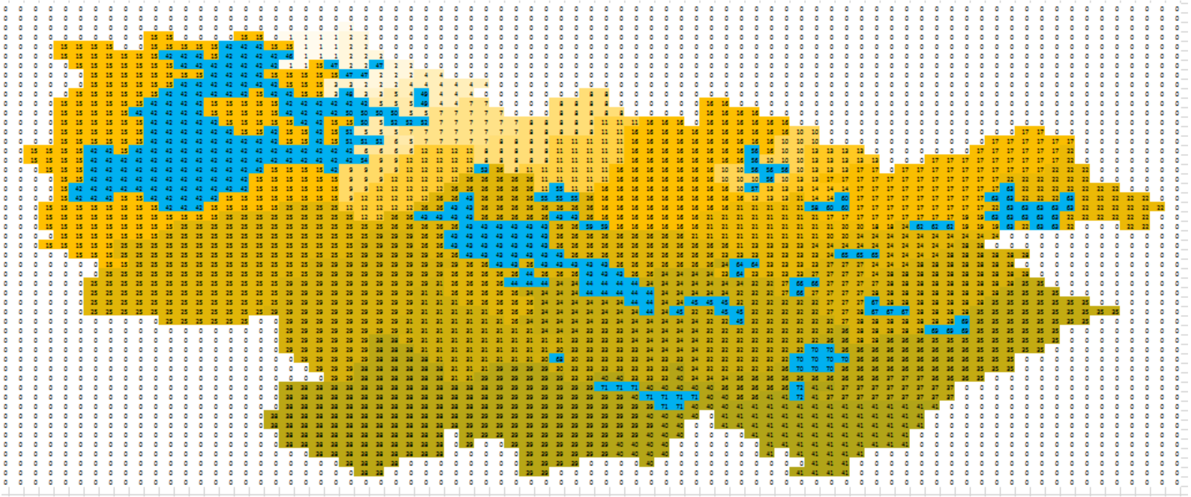


Figure 3.17 – Assignment of basin number to each grid cell

Table 3.7 – Summary of Raven parameters with equivalent WATFLOOD counterparts

Parameter Description	Raven	WATFLOOD
precipitation lapse rate mm/m	PrecipitationLapseRate	rlapse
temperature lapse rate dC/m	AdidabticLapseRate	tlapse
fraction of swe as water in ripe snow	Irreductible Snow Saturation	whcl
soil porosity	Porosity	spore
upper zone retention mm	Field_Capacity	fcap
wilting point - mm of water in uzs	Saturated Wilt	ffcap

Many additional parameters required calibration. The full summary of Raven calibrated parameters is shown in Table 3.8. Two inherent issues arise with such approach in model development. First is the loss in information from redundant rescaling of data. The initial geospatial data was scaled to small grid sizes for WATFLOOD. Then, geospatial data was re-averaged to fit larger sub-basins. Loss of geospatial information and corresponding increase of error would be inevitable in rescaling process. Error from initial geospatial data would result in decrease in performance during both the calibration and validation period. Increase in error and uncertainty could easily be avoided by performing required geospatial data using the Raven delineation.

Table 3.8 – Summary of Raven model parameters calibrated

Parameter	Description	Min	Max
rs_min	Rain snow transition temperature minimum	-1.00E+00	1.00E+00
rs_max	Rain snow transition temperature maximum	1.00E+00	2.00E+00
par_g_2	Irreducible snow saturation	0.00E+00	1.00E+00
beta_agr	HBV Beta parameter for infiltration (1 for each land class)	1.00E-01	2.00E+01
perc_agr_1	Max percolation rate (2 for each land class)	1.00E-01	5.00E+01
inter_agr	Interflow rate (1 for each land class)	5.00E-02	5.00E+01
dep_agr	Depression (1 for each land class)	1.00E+01	1.00E+03
basef_agr	Baseflow coefficient (1 for each land class)	1.00E-02	1.00E+01
basen_agr	Baseflow exponent n (1 for each land class)	5.00E-01	4.00E+00
petc_agr	PET correction (1 for each land class)	1.00E-01	1.20E+00
soild_agr_1	Top layer soil depth (1 for each land class)	1.00E-01	2.00E+02
soild_agr_2	Bottom layer soil depth (1 for each land class)	1.00E-03	1.00E+03
owpet_agr	Openwater PET correction (1 for each land class)	2.50E-01	1.00E+00
lrel_coef	Lake release coefficient	1.00E-02	1.00E+00
lpet_corr	Lake PET correction	1.00E-01	1.20E+00
mel_agr	Melt rate (1 for each land class)	2.50E-01	7.50E+00
max_ht_agr	Max vegetation height (1 for each land class)	0.00E+00	3.00E+00
max_lai_agr	Max leaf area index (1 for each land class)	0.00E+00	1.00E+01
max_lf_agr	Max leaf conductance (1 for each land class)	0.00E+00	1.00E+01
r01	Manning's coefficient (1 for each channel type)	0.0005	0.15
weir_01	Weir structure width/coefficient (1 for each reservoir)	1.00E+00	5.00E+02
tc_land	Time of concentration multiplier	1.00E-02	5.00E+01
tc_15	Time of concentration for wetland dominated basin	1.00E-02	5.00E+01

Second issue is the over parameterization in Raven model from utilizing geospatial data for WATFLOOD. Land class classification and number of reservoirs should be minimized unless supported by data. Raven required a great number of parameters without a WATFLOOD counterpart for each land class type. As a result, model became over-parameterized, requiring nearly 150 parameters to be calibrated. Such over-parameterization is likely to decrease model performance during validation period.

3.3.2 Lake of the Woods Watershed Model Calibration Formulation

The model was calibrated from 2004-10-01 to 2009-09-31 with the first year as a warm-up period. Validation was performed from 2009-10-01 to 2015-09-31. Single objective calibration was performed using the Dynamically Dimensioned Search algorithm (Tolson and Shoemaker, 2007) in OSTRICH (Matott, 2017), with a budget of 20,000 model runs to allow model convergence in all experiments. Multi-objective calibration was performed using the Pareto Archived Dynamically Dimensioned Search (Asadzadeh and Tolson, 2013), with a budget of 20,000 model runs. The output of the calibration period was used as initial conditions for the validation period. Similar to the Kaministiquia watershed model, outflows from the three managed reservoirs (Lake of the Woods, Rainy Lake, and Namakan Lake) were overridden with observation data. Model was calibrated three times to different objective functions:

1. The average NSE of 11 stream gauges
2. The average 7-day running average NSE of four reservoir inflows (Lake of the Woods, Rainy Lake, Lac-la-Croix, and Namakan Lake)
3. Multi-objective calibration to both stream gauges and inflows.

Each of the gauges was weighted differently based on yearly average flow. Flow gauge at Manitou was given a weight of 0 as majority of the flow is determined by the upstream Rainy Lake with overridden flow. The Flow gauge at Lake of the Woods was also given a weight of 0 as the outflow of Lake of the Woods is overridden by observation data.

3.3.3 Lake of the Woods Watershed Model Results

Table 3.9 shows the calibration and validation results when model was calibrated to flow NSE.

Table 3.9 – Calibration and validation results for calibration to flow NSE, with bolded NSE when model performs better by 0.05 or greater

Number	Station	Weight	Raven NSE Calibration	WATFLOOD NSE Calibration	Raven NSE Validation	WATFLOOD NSE Validation
1	Turtle	0.145	0.74	0.76	**0.8**	0.69
2	Atikokan	0.010	0.67	0.69	0.68	**0.77**
3	Seine	0.171	0.64	**0.72**	0.69	0.6
4	Manitou	0.000	0.92		0.86	
5	Little Fork	0.094	0.65	0.67	0.57	0.59
6	Big Fork	0.073	0.65	0.68	0.45	**0.7**
7	Vermillion	0.051	**0.78**	0.54	**0.72**	0.67
8	Basswood	0.105	0.52	**0.69**	0.63	0.6
9	Lac-la-Croix	0.336	0.78	**0.84**	0.76	**0.85**
10	Kawishiwi	0.015	0.58	**0.67**	0.62	**0.69**
11	Lake of the Woods	0.466	**0.82**	0.74	0.76	0.77
12	Rainy Lake	0.279	0.9	0.92	0.9	**0.96**
13	Lac La Croix	0.102	0.68	**0.82**	0.69	**0.84**
14	Namakan	0.153	0.76	0.75	0.79	0.82

Table 3.10 shows the calibration and validation results when model was calibrated to inflow NSE.

Table 3.10 – Calibration and validation results for calibration to reservoir inflow NSE, with bolded NSE when model performs better by 0.05 or greater

Number	Station	Weight	Raven NSE Calibration	WATFLOOD NSE Calibration	Raven NSE Validation	WATFLOOD NSE Validation
1	Turtle	0.145	0.56	**0.61**	**0.71**	0.56
2	Atikokan	0.010	**0.71**	0.46	**0.75**	0.69
3	Seine	0.171	0.42	0.45	0.49	**0.64**
4	Manitou	0.000	0.92		0.88	
5	Little Fork	0.094	**0.5**	0.4	**0.51**	0.29
6	Big Fork	0.073	-0.28	**0.55**	0.07	**0.67**
7	Vermillion	0.051	**0.67**	0.56	0.59	**0.7**
8	Basswood	0.105	0.35	**0.54**	**0.62**	0.29
9	Lac-la-Croix	0.336	0.79	**0.84**	0.81	0.8
10	Kawishiwi	0.015	0.42	**0.56**	0.66	0.66
11	Lake of the Woods	0.466	**0.92**	0.86	0.85	0.86
12	Rainy Lake	0.279	0.93	0.94	0.94	0.97
13	Lac La Croix	0.102	0.75	**0.81**	0.82	0.81
14	Namakan	0.153	0.79	0.79	0.82	0.81

Figure 3.18 shows the Pareto front of the multi-objective calibration experiment. Figure 3.19 shows the plots of Pareto optimal during the validation period. WATFLOOD’s Pareto front dominates Raven model’s Pareto front.

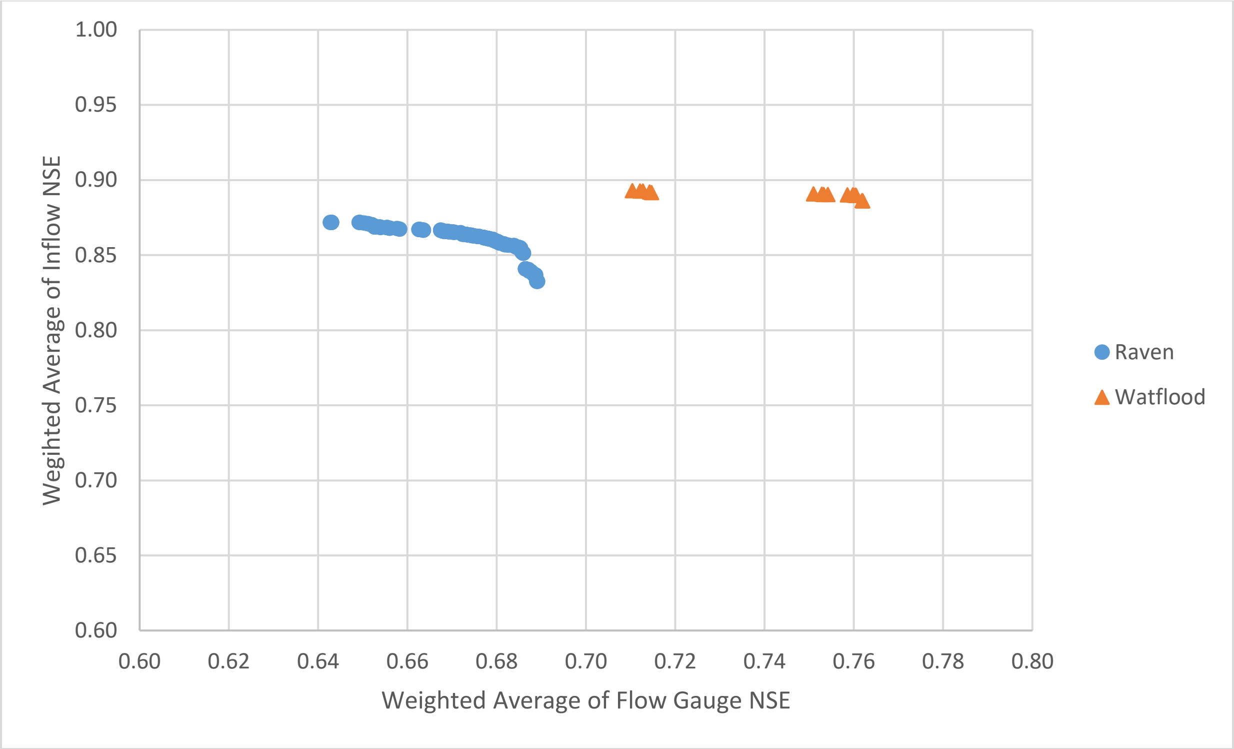


Figure 3.18 – Pareto front of non-dominated solutions in multi-objective calibration to flow NSE and inflow NSE

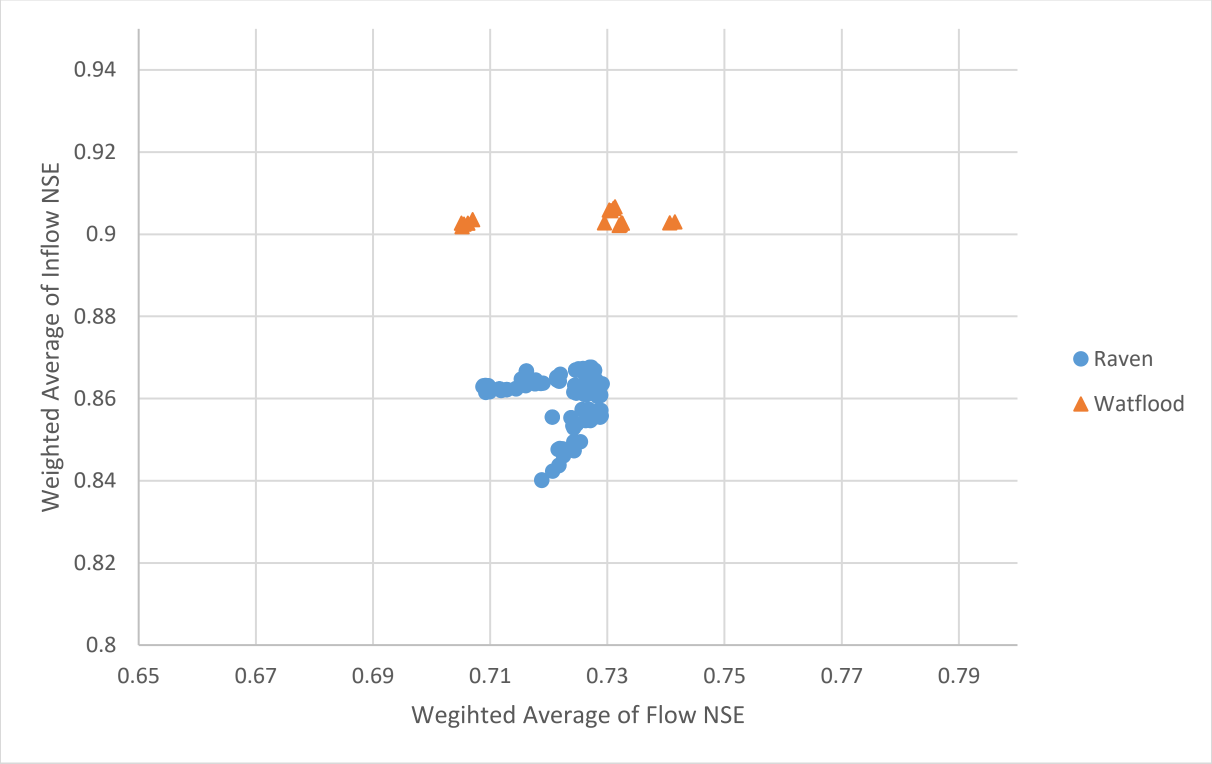


Figure 3.19 – NSEs of Pareto optimal solution evaluations during the validation period

Table 3.11 shows the calibration and validation results when model was calibrated using multi-objective calibration. The NSE values presented are averages of the non-dominated solutions.

Table 3.11 – Average NSE of non-dominated solutions generated by multi-objective calibration during calibration and validation periods, with bolded NSE when model performs better by 0.05 or greater

Number	Station	Weight	Raven NSE Calibration	WATFLOOD NSE Calibration	Raven NSE Validation	WATFLOOD NSE Validation
1	Turtle	0.145	0.72	0.76	**0.79**	0.71
2	Atikokan	0.010	0.69	0.66	0.71	0.71
3	Seine	0.171	**0.61**	0.45	0.63	**0.7**
4	Manitou	0.000	0.92		0.92	
5	Little Fork	0.094	0.62	0.64	**0.62**	0.51
6	Big Fork	0.073	0.55	**0.69**	0.55	**0.75**
7	Vermillion	0.051	**0.76**	0.59	**0.76**	0.68
8	Basswood	0.105	0.47	**0.72**	0.47	**0.7**
9	Lac-la-Croix	0.336	0.78	**0.87**	0.78	0.81
10	Kawishiwi	0.015	0.54	**0.71**	0.54	**0.72**
11	Lake of the Woods	0.466	0.89	0.89	0.89	0.9
12	Rainy Lake	0.279	0.91	0.94	0.91	**0.96**
13	Lac La Croix	0.102	0.7	**0.82**	0.7	**0.86**
14	Namakan	0.153	0.77	**0.86**	0.77	**0.82**

A summary of flow weighted average NSE for all experiments are shown in Table 3.12.

Table 3.12 – Overall NSE comparison between Raven and WATFLOOD

Calibration Objective	Flow NSE		Inflow NSE		Multi Objective	
Model	**Raven**	**WATFLOOD**	**Raven**	**WATFLOOD**	**Raven**	**WATFLOOD**
Flow NSE - Calibration	0.70	0.75	0.53	0.62	0.67	0.71
Inflow NSE - Calibration	0.82	0.80	0.89	0.87	0.86	0.89
Flow NSE - Validation	0.69	0.71	0.62	0.62	0.69	0.72
Inflow NSE - Validation	0.80	0.84	0.87	0.88	0.86	0.90

Multi-objective calibration showed to be beneficial in improving validation period results. In the Raven model, multi-objective calibration resulted in same flow NSE compared to flow calibrated experiment, with only 0.01 lower in inflow NSE compared to the inflow calibrated experiment. In the WATFLOOD model, model performed better in both flow and inflow NSE compared to all calibration experiments. Experiment shows multi-objective calibration can improve validation results through increase in physical realism in the model. Overall difference in weighted NSE between the WATFLOOD model and the Raven model was ~0.03. WATFLOOD performed better in south-eastern basins draining to Lac-la-Croix, including gauges at Basswood, Kawishiwi, and Lac-la-Croix. Raven performed better in northern basins draining to Rainy Lake, including gauges at Turtle, Seine, and Atikokan. Another major difference between the Raven model and the WATFLOOD model was the run time. The Raven model took approximately 12 seconds for a single model run during calibration, while the WATFLOOD model took approximately 4 minutes. The substantial reduction in runtime of Raven model makes Raven suitable for computationally expensive experiments, such as the DCT.

The hydrographs of the calibration experiments are presented in Appendix C.

Chapter 4

Assessing and Improving Hydrologic Models Used for Decision Making

In this chapter, the utility of the Lake of the Woods Raven model in a real-life decision making scenario is introduced. After an overview of the utility, an assessment of model's capabilities in the decision making scenario is performed using Decision Crash Testing (DCT). The DCT informs the limitations of the current model and preferred calibration objective function for a specific decision making scenario.

4.1 Case Study – Selection of Reservoir Operation Rule Curve

This section presents the problem background and key aspects for implementation of rule curves to the Lake of the Woods model.

4.1.1 Rule Curve Motivations in Rainy Lake

In 2015, a study was performed by the International Rainy and Namakan Lakes Rule Curves Study Board (IRNLRCSB) for the International Joint Commission (IJC) to reevaluate operating rule curves for the Rainy Lake based on (IRNLRCSB, 2015):

- Protecting shorelines from flood damage
- Ensuring water levels for hydroelectricity generation
- Protecting natural environments
- Recreational use of lakes
- Water quality

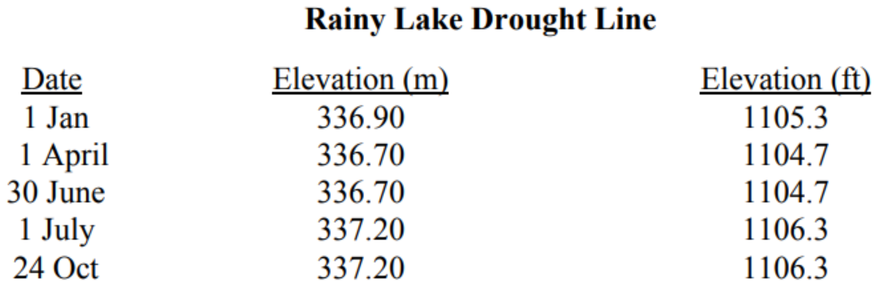
Hydrologic models, along with water quality models, hydraulic models, and global climate models were utilized to inform stakeholders the various impacts of different rule curve alternatives. The original study evaluated 6 rule curve alternatives among 7 key study themes: Fish, Wildlife, Economic Impacts, Archeological Resources, Vegetation, Invertebrates, and Water Quality. The 7 key study themes are broken down into 36 sub-categories, with many of the sub-categories requiring external studies, hydraulic models, global climate models, water quality models, and water temperature models. For this thesis, three rule curve alternatives with the greatest impacts on categories impacted by outputs of hydrologic model were selected for analysis and comparison via DCT. Since the goal of the DCT was to assess the utility of the hydrologic model, evaluation criteria impacted by the output of hydrologic model were necessary. The following three evaluation criteria that can be mapped from hydrologic model output were selected:

- Ecological benefit (based upon survivability of fish and wildlife determined by lake levels)
- Economic benefit (based upon volume of water over not used for electricity generation)
- Flood Damage Reduction (based upon stage of lake during a storm event during a spring snowmelt)

In summary, the DCT assesses the model’s utility in comparing three rule curves on their impact on three evaluation criteria relevant to model outputs.

4.1.2 Flow Restrictions at Rainy Lake

In Section 3.1.3, the importance of regulatory and hydraulic restrictions governing reservoir operation was discussed. The Rainy Lake has regulations on flow restrictions outlined by the IJC. Also, the natural features in the river channel between the lake outlet and the dam restricts the rate of flow out of Rainy Lake (CHC, 2010). When Rainy Lake level is below the drought line, the minimum flow is reduced to 65 m3/s. Otherwise, minimum flow is 100 m3/s. The drought levels outlined by IJC is shown in Figure 4.1.

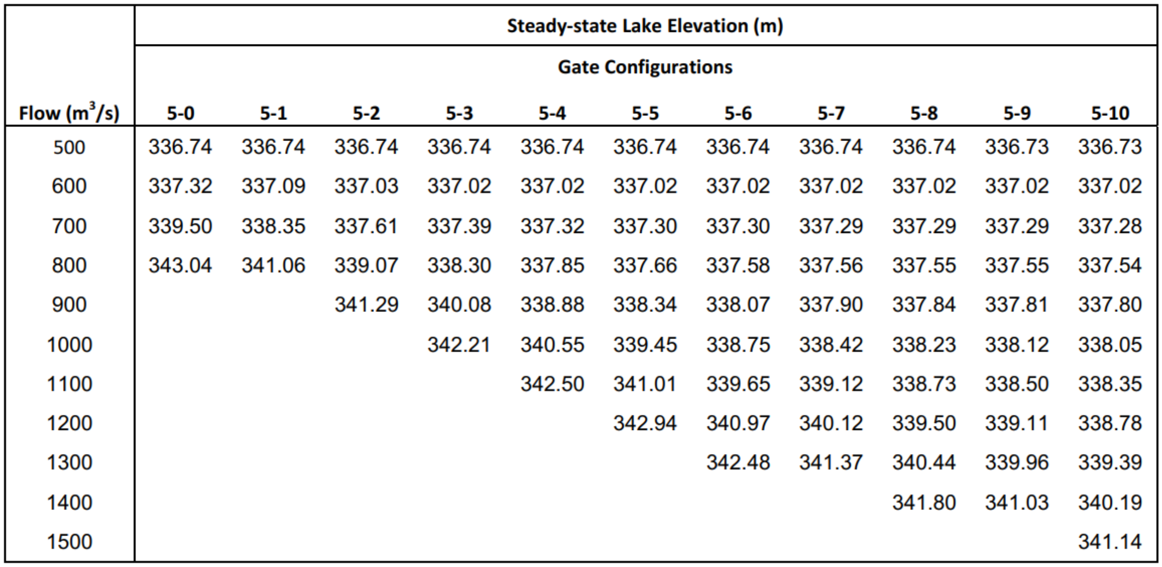


Rainy Lake Drought Line

Date	Elevation (m)	Elevation (ft)
1 Jan	336.90	1105.3
1 April	336.70	1104.7
30 June	336.70	1104.7
1 July	337.20	1106.3
24 Oct	337.20	1106.3

Figure 4.1 – Drought line of Rainy Lake determined by IJC (CHC, 2010)

Based on a hydraulic study the Natural Research Council Canada (2011), the stage - maximum discharge relationship of Rainy Lake is shown in Figure 4.2. For assessment of high lake level scenarios, the full gates open operation (5-10) was assumed. At each time step during model simulation, the maximum possible flow was calculated based on current lake levels.



	Steady-state Lake Elevation (m)										
	Gate Configurations										
Flow (m^3/s)	5-0	5-1	5-2	5-3	5-4	5-5	5-6	5-7	5-8	5-9	5-10
500	336.74	336.74	336.74	336.74	336.74	336.74	336.74	336.74	336.74	336.73	336.73
600	337.32	337.09	337.03	337.02	337.02	337.02	337.02	337.02	337.02	337.02	337.02
700	339.50	338.35	337.61	337.39	337.32	337.30	337.30	337.29	337.29	337.29	337.28
800	343.04	341.06	339.07	338.30	337.85	337.66	337.58	337.56	337.55	337.55	337.54
900			341.29	340.08	338.88	338.34	338.07	337.90	337.84	337.81	337.80
1000				342.21	340.55	339.45	338.75	338.42	338.23	338.12	338.05
1100					342.50	341.01	339.65	339.12	338.73	338.50	338.35
1200						342.94	340.97	340.12	339.50	339.11	338.78
1300							342.48	341.37	340.44	339.96	339.39
1400									341.80	341.03	340.19
1500											341.14

Figure 4.2 – Stage-Discharge relationship of Rainy Lake at different gate configurations (Natural Research Council Canada, 2011)

Limitations on increase of flow over a time step were not incorporated to the test as the limitation was not a constraint in the IRNLRCSB study.

4.1.3 Rule Curve Options at Rainy Lake

Three rule curves were evaluated in the case study. Rule Curve A is the operational rule curve at the time of study. Rule Curve B is a modified version of the Rule Curve A with a lower spring target for increase in spring flood damage reduction. Rule Curve C incorporates a lower winter drawdown for increase in ecological benefit with decrease in economic benefit. For model simulation, the target for reservoir operation was set to the median between the low and high stage values of the rule curve.

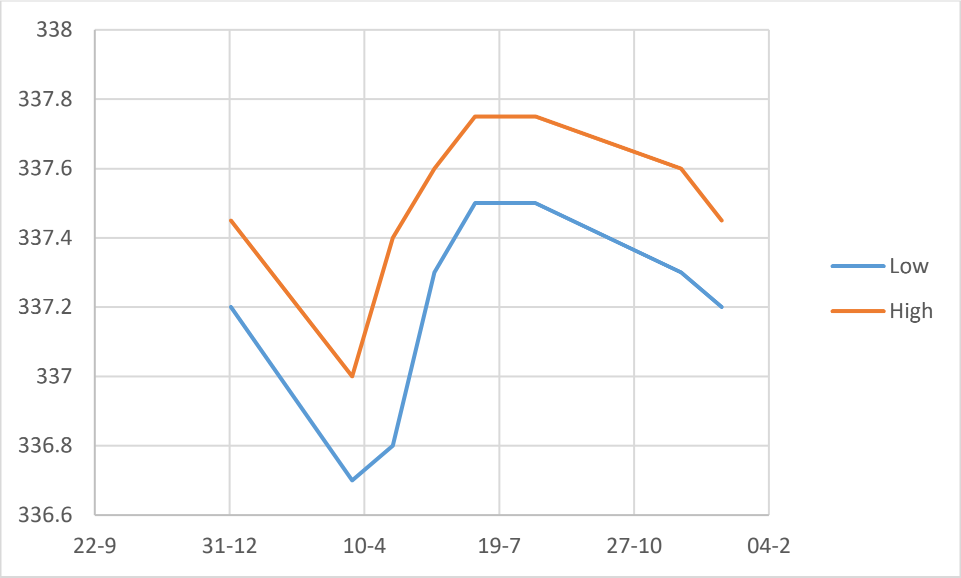


Figure 4.3 – Operational rule curve A used by LWCB with high and low targets

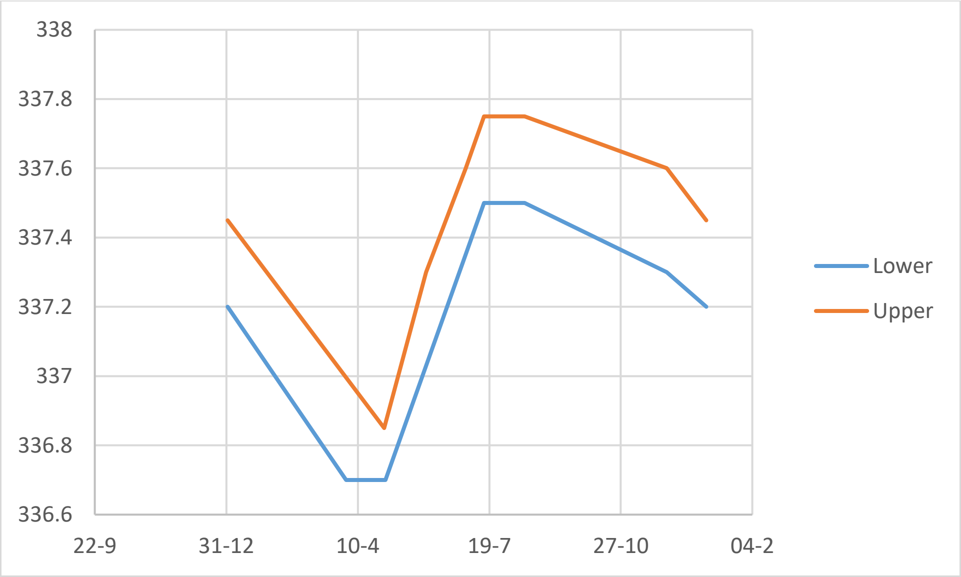


Figure 4.4 – Operational rule curve B modified from rule curve A with lower spring target to reduce flood risks

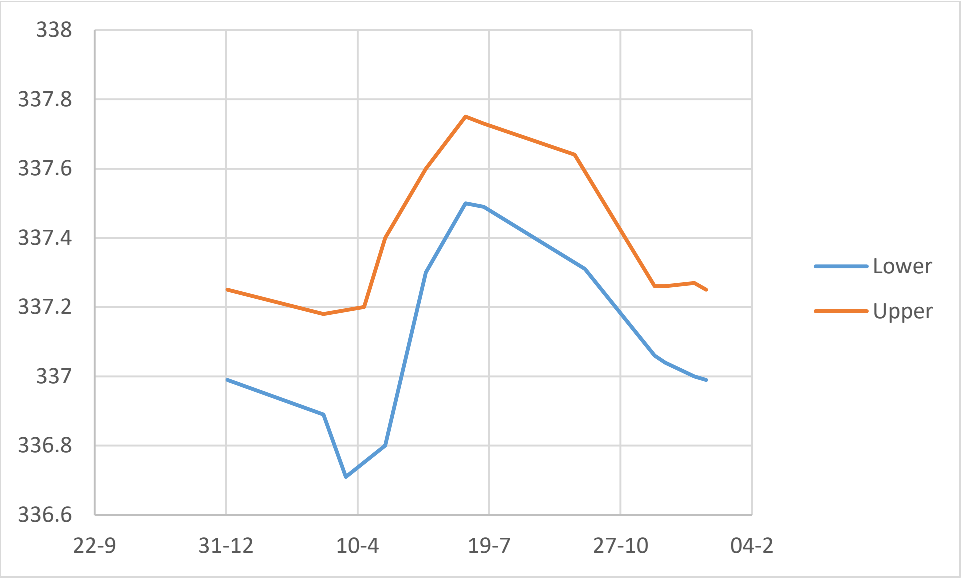


Figure 4.5 – Operational rule curve B modified from rule curve A with low winter drawdown to improve ecological benefits

4.2 Evaluation Criteria in Rule Curve Selection

In Chapter 2, the importance of clear transposition between hydrologic model outputs to evaluation criteria was discussed. In this section, the clear transposition of model outputs to evaluation criteria comprehendible to stakeholders is discussed.

4.2.1 Ecological Benefits

Ecological benefits are calculated based on probability of survivability of four species (Walleye, Whitefish Egg, Common Loon, and Muskrat) which is dependent of water level rise/fall over a specific period. The four species are impacted by water levels at different times of the year to assess model's capabilities in performing at different times of the years. The probability or survivability of the four species are summed for a final score, with the maximum score of 4. The probability of survivability dependent on water level rise and fall is summarized in Table 4.1 Corresponding Julian dates were estimated using Figure 4.6 For maximizing ecological benefit, the rule curve resulting in highest probability of survivability would be deemed as most desirable and selected as the "preferred" decision in DCT.

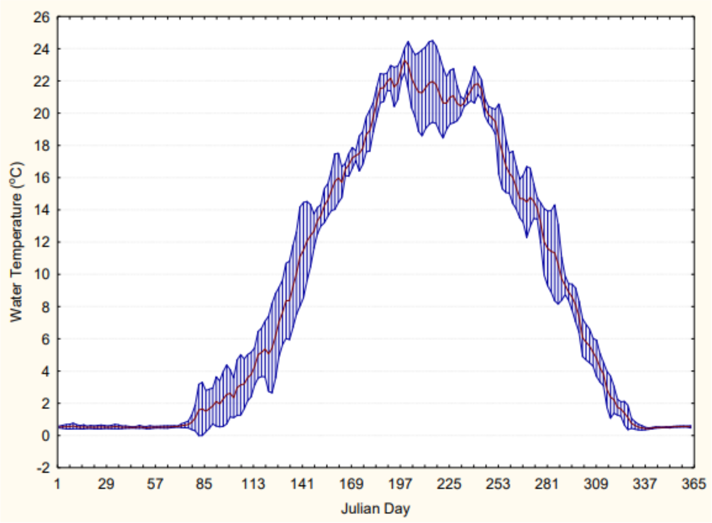


Figure 4.6 – Mean annual water temperature of Rainy Lake for the period of 2011-2014. The outer envelope represents the 95% confidence interval of values (Marshall and Foster, 2015)

Table 4.1 – Summary of probability of survivability of species dependent on water level

Specie:	Walleye	
Period Description:	Ice Out to Water Temp 11 deg C	
Julian Date Range:	71 - 140	
Drop/Rise	**Value (m)**	**PS**
Drop	< 0.1	1
Drop	> 1	0
Rise	< 0.5	1
Rise	> 2.5	0

Specie:	Whitefish Egg	
Period Description:	Mid November to Ice Out	
Julian Date Range:	319 - 71	
Drop/Rise	**Value (m)**	**PS**
Drop	< 0.5	1
Drop	> 2	0
Rise	< 0.5	1
Rise	> 2	0

Specie:	Common Loon	
Period Description:	3 Weeks before to after Ice Out	
Julian Date Range:	92-141	
Drop/Rise	**Value (m)**	**PS**
Drop	< 0.3	1
Drop	> 0.8	0
Rise	< 0.15	1
Rise	> 0.4	0

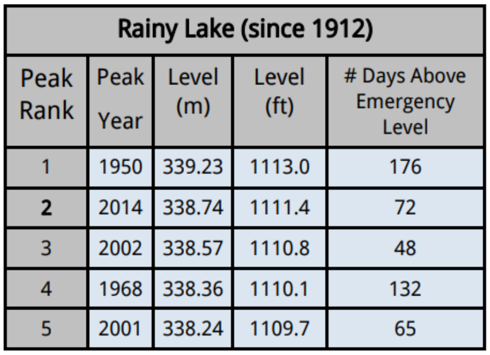
Specie:	Muskrat	
Period Description:	Winter (November to March)	
Julian Date Range:	319-90	
Drop/Rise	**Value (m)**	**PS**
Drop	< 0.15	1
Drop	> 0.6	0
Rise	< 0.15	1
Rise	> 0.33	0

4.2.2 Economic Benefits

There are two powerhouses (one Canadian and one American) that utilize water from Rainy Lake for electricity generation. The maximum flow for electricity generation is 150 m^3/s and 250 m^3/s for the Canadian and American powerhouses respectively. Therefore, any flow greater than 400 m^3/s would be potential energy wasted by releasing over the spillway. To calculate the economic benefit of each rule curve, a total sum of water over the spillway was calculated. Additional calculations would be required to convert volume of water to economic value, but volume was assumed to be sufficient for comparison between rule curves. For maximizing economic benefit, the rule curve with lowest volume of water released over the spillway would be deemed as most desirable and selected as the “preferred” decision in DCT.

4.2.3 Flood Damage Reduction

The IJC has defined 337.5 m as an emergency state level for Rainy Lake. Above the emergency level, shoreline erosion and property damage is likely to occur. In 2014, Rainy Lake recorded high lake level of 338.74 m as shown in Figure 4.7. Running the inflow calibrated Raven model with Rule Curve A produced maximum stage of 338.82 m. In other words, the inflow calibrated model predicted the flood stage within 0.1 m without utilizing forced outflow data. This demonstrates that the flow constraints and rule curve operation produce realistic stage values. Rule curves were evaluated on the reduction of peak stage from the 2014 flood event. For maximizing flood damage reduction benefit, the rule curve with the lowest peak stage during a storm event would be deemed as most desirable and selected as the “preferred” decision in DCT.



Rainy Lake (since 1912)				
Peak Rank	Peak Year	Level (m)	Level (ft)	# Days Above Emergency Level
1	1950	339.23	1113.0	176
2	2014	338.74	1111.4	72
3	2002	338.57	1110.8	48
4	1968	338.36	1110.1	132
5	2001	338.24	1109.7	65

Figure 4.7 – Stage records of Rainy Lake from five biggest historical floods (Water Levels Committee of the International Rainy-Lake of the Woods Watershed Board, 2015)

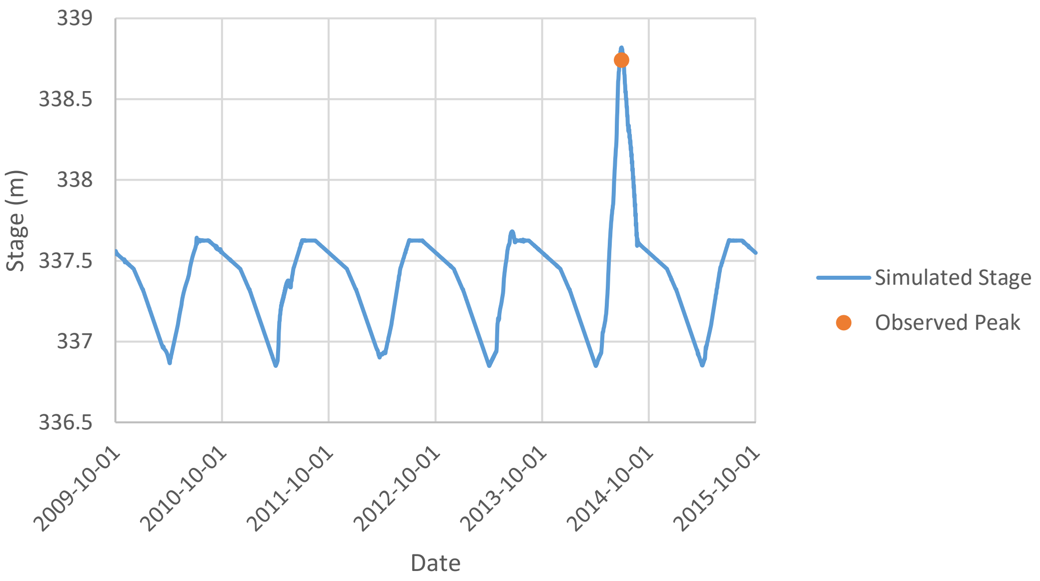


Figure 4.8 – Raven model simulated stage using rule curve A in comparison with observed peak during 2014 storm event

4.3 Decision Crash Testing for Model Assessment

In this section, the methodology of DCT and its application in model utility assessment are discussed.

4.3.1 Generation of Synthetic Observations

In hydrologic model evaluation techniques, one difficulty in testing the model for its intended use is the unavailability of the correct answer. For example, assume a model is utilized to make a decision on upgrading a bridge to accommodate a 100-year storm event that has not happened in the past. Before the storm happens, there would be no stage data responding to a 100-year storm event available. Traditional usage of model in the scenario would be to calibrate the model to inflow data, introduce a statistically generated storm-event, and assume the model is capable of producing accurate stage values. In this approach, a model is evaluated in its ability to generate inflow using storms smaller than a 100-year storm, and used to generate stage levels with a 100-year storm event. Similar problem occurs for rule curve selection. The historical data is generated using rule curve A for operation. When the model is used to evaluate rule curve B and rule curve C, simulation results do not have corresponding observation data, as the reservoir has never been operated using rule curve B or rule curve C. Traditional approach would calibrate the model to inflow data (using rule curve A), test the model with rule curve B and C, and assume the model is capable of informing rule curve performance based on various evaluation criteria (stage, volume, and peak stage).

A similar problem has been addressed in synthetic calibration experiments. If a researcher has developed a new optimization algorithm for calibration of hydrologic models, researcher may want to test the calibration approach using historical observation data. Once calibration experiment has been performed, the imperfection in hydrograph fit could be an attribution of two factors:

1. Error in model structure and observation data, resulting in impossible perfect fit between model simulation and historical data
2. Error in optimization algorithm or calibration strategy, incapable of finding the optimal parameter set

To assess which factor is impacting the performance, researchers can perform a synthetic calibration experiment. In a synthetic calibration experiment, the observation is generated from random sampling of model parameters. Then the same model is calibrated using the optimization algorithm. Since observation data is generated from the model used for calibration experiment, an optimal parameter set that generates perfect fit exists. Calibration may be repeated to ensure robustness. The synthetic calibration scheme is shown in Figure 4.9.

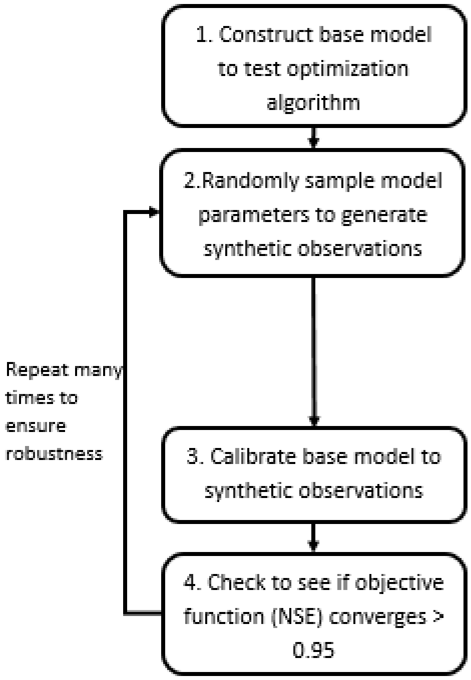

1. Construct base model to test optimization algorithm
2.Randomly sample model parameters to generate synthetic observations
Repeat many times to ensure robustness
3. Calibrate base model to synthetic observations
4. Check to see if objective function (NSE) converges > 0.95

Figure 4.9 – Example of synthetic calibration experiment to test optimization algorithm

Concepts from the synthetic calibration experiments can be utilized to address issue of data unavailability in decision making scenarios using a hydrologic model. Here, by random sampling of parameters, a “synthetic reality” is created using the model. In the synthetic reality, all hydrologic observation data is available, as it is assumed that the model with the randomly generated model parameters is the *truth*. Any decision criteria can readily be evaluated through model evaluation. To make synthetic reality consistent without observed hydrologic data, two safeguards were implemented. Without any safeguards, randomly generated parameters may result in unrealistic hydrologic data that would never be found in real life. Testing a hydrologic model’s capability in simulating unrealistic hydrologic data is unnecessary, simulation of unrealistic hydrologic data would not be an intended use of a hydrologic model. Random sampling of parameters was performed on a select number of parameters within a specified range. Range was determined from initial calibration of model to historical inflow data. Range was between +/- 50% from parameters calibrated to inflow. Next, the synthetic reality was tested to ensure it had a peak response useful for decision making experiment. A parameter set was rejected if the peak response to the 2014 flood was less than the emergency level of 337.5 m.

4.3.2 DCT Methodology

Decision Crash Testing (DCT) is a novel fit-for-purpose model evaluation method to rigorously test a model's capability in supporting decision making. The key concept behind the DCT is that if a hydrologic model is incapable of making the correct decisions when calibration to data guaranteed by a simplified synthetic reality, it would be naïve to believe that the same model is capable of making the correct decision in a much more complex actual reality. In DCT, the model to be used in real application is calibrated to the synthetic reality observation data using the intended method of calibration strategy for real model application. Then, the model is given a decision making scenario and tested on whether the correct decision is made compared to the decision made in synthetic reality. Since the synthetic reality is generated using a random sample of model parameters, with enough calibration budget, the model should be able to generate near perfect fit to the synthetic reality. With this near perfect fit, model is likely to make the correct decision. A calibration run of a model with a budget of 10,000 requires 10,000 model runs. At each model run, the model generates a different NSE and different hydrologic outputs. Utilizing the hydrologic outputs at each model run, the decision made in a given model run is determined. In a decision record file, results from each model run is archived by recording the NSE and the decision for each evaluation criteria. A decision record file for a calibration run with 10,000 models runs archives 10,000 records of NSE and three decisions, one for each evaluation criteria. Then, the decision record file is used to establish a correlation between NSE and decision making capabilities of the model for a specific evaluation criteria. The process can be repeated many number of times using additional realization of synthetic realities, *N*, allowing the model to be tested rigorously, similar to crash testing concepts. General DCT scheme for rule curve selection is illustrated in Figure 4.10

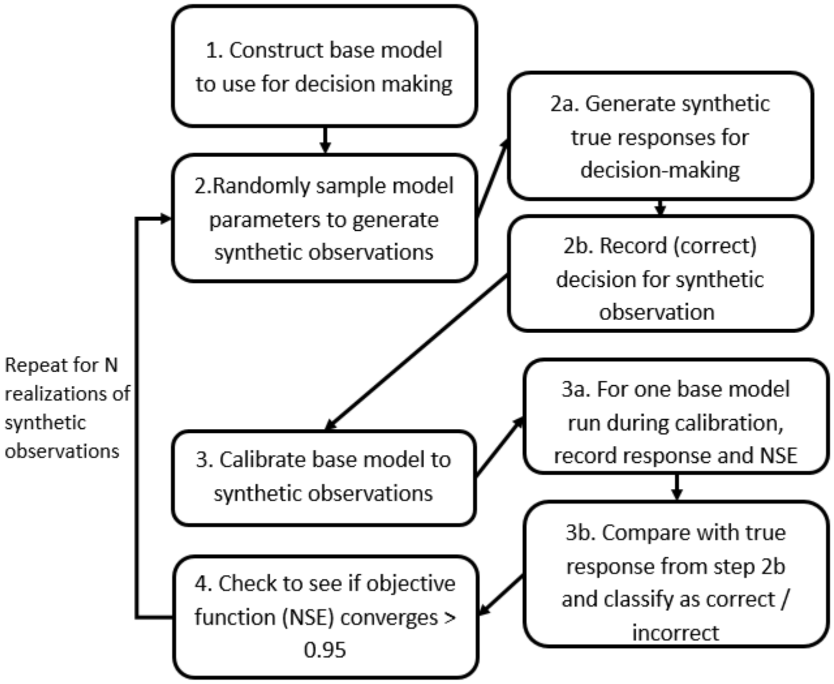

1. Construct base model to use for decision making
2.Randomly sample model parameters to generate synthetic observations
2a. Generate synthetic true responses for decision-making
2b. Record (correct) decision for synthetic observation
Repeat for N realizations of synthetic observations
3. Calibrate base model to synthetic observations
3a. For one base model run during calibration, record response and NSE
3b. Compare with true response from step 2b and classify as correct / incorrect
4. Check to see if objective function (NSE) converges > 0.95

Figure 4.10 – DCT setup workflow used for rule curve selection experiment

Detailed steps of DCT implemented for the case study are:

1. Randomly sample model parameters to generate synthetic reality
2. Operate synthetic reality model using Rule Curves A, B, C and record performance in ecological, economic, and flood damage reduction for synthetic reality
3. Determine rule curve rankings for each criteria for the synthetic reality (i.e. For synthetic reality *n*, the ranking for ecological benefits measured by probability of survivability is Rule Curve C > Rule Curve A > Rule Curve B)
4. Calibrate base model to synthetic reality generated inflows

a. For each model run *n'* in calibration, record the NSE for model run
b. Run the model with parameter set *n'* using Rule Curves A, B, C, and record performance in ecological, economic, and flood damage reduction
c. Determine rule curve rankings for each evaluation criteria for model run *n'*
d. Compare rule curve rankings from model run *n'* to rule curve rankings for synthetic reality *n* from step 3 for each evaluation criteria
e. For model run *n'*, record the NSE and a correct/incorrect for each evaluation criteria

5. Repeat steps 1-4 *N* times

At the end of a DCT experiment, *N* number of decision record files are populated. For each decision record file, the results are summarized into NSE bins. For example, out of the 10,000 model runs, all model runs with NSE between 0.5 and 0.55 were extracted. For each NSE bin, the similarity score was calculated for each evaluation criteria using equation 12.

$$\text{Similarity score} = \frac{\text{\# of Correct in NSE Bin}}{\text{Total \# of model runs in NSE Bin}} \tag{12}$$

A similarity score can be compared to a model's probability to inform the correct decision. With 100 decision record files, 100 similarity score values are populated for each NSE bin. Within a NSE bin, the mean and standard deviation in similarity score can be calculated.

4.4 DCT Results and Calibration Objective Formulation

In this section, DCT results from different calibration objective formulations are discussed. For the different calibration objective formulations, the same set of synthetic realities was utilized so that the only difference among the different experiments is the calibration objective.

4.4.1 Results Based on Calibration to Different Objective Gauges

The DCT experiment was performed to the two objective functions: the weighted average of NSE at 11 stream gauges and the weighted average of a 7-day running average NSE at four reservoir inflows. Figure 4.11 shows a box whisker plot of DCT experiment results with calibration to inflow NSE. Figure 4.11 shows the similarity score of rule curve ranking for economic benefit (volume of water over spillway). The box whisker plot shows the maximum, 25 percentile, median, 75 percentile, and minimum similarity score for each NSE bin. For example, the average similarity score value for model runs with a NSE range between 0.65 and 0.70 is 70%, suggesting that models within this NSE range have 70% likelihood of correctly informing the decision making.

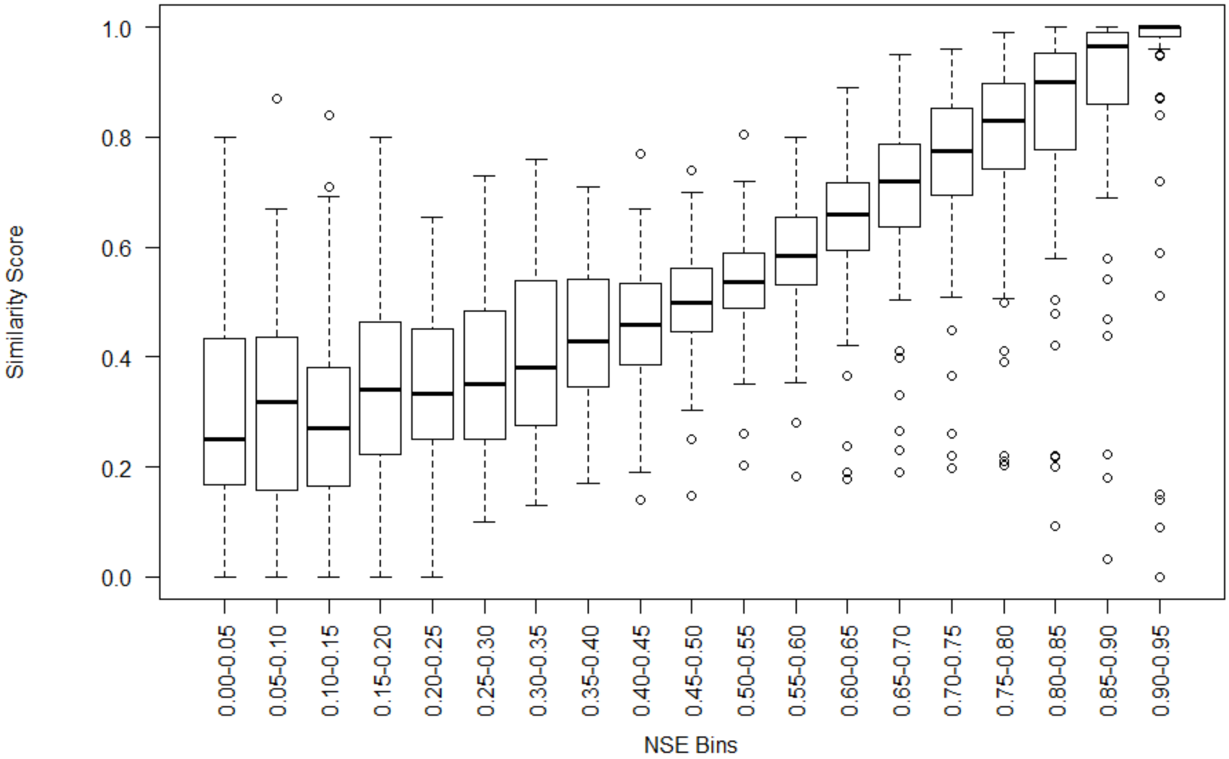


Figure 4.11 – Box whisker plot of similarity score for economic benefit when calibrated to inflow NSE

Another interpretation for the plot is the model can correctly rank the rule curves based on economic benefit greater than 70% of the times when calibrated to inflow NSE greater than 0.65. There exist synthetic realities where the similarity score are near 95% and synthetic realities where the similarity score are near 20% in the NSE bin of 0.60 to 0.65. Low similarity score values can be a result of a synthetic reality where the randomly sampled parameter set results in small differences in volumes of water spilled among the three curves. As a result, this becomes a *hard* decision for the model to make. Figure 4.12 and Figure 4.13 shows the box whisker plot for ecological benefit and flood damage reduction benefit, respectively. Both ecological benefit and flood damage reduction decisions are dependent upon stage levels. With models operating using the same rule curves under same regulations, similarity score consistently remain high despite the changes in NSE. Results show level dependent decision criteria are relatively insensitive to model performance compared to level independent decision criteria, given that the reservoir operating strategy remains consistent.

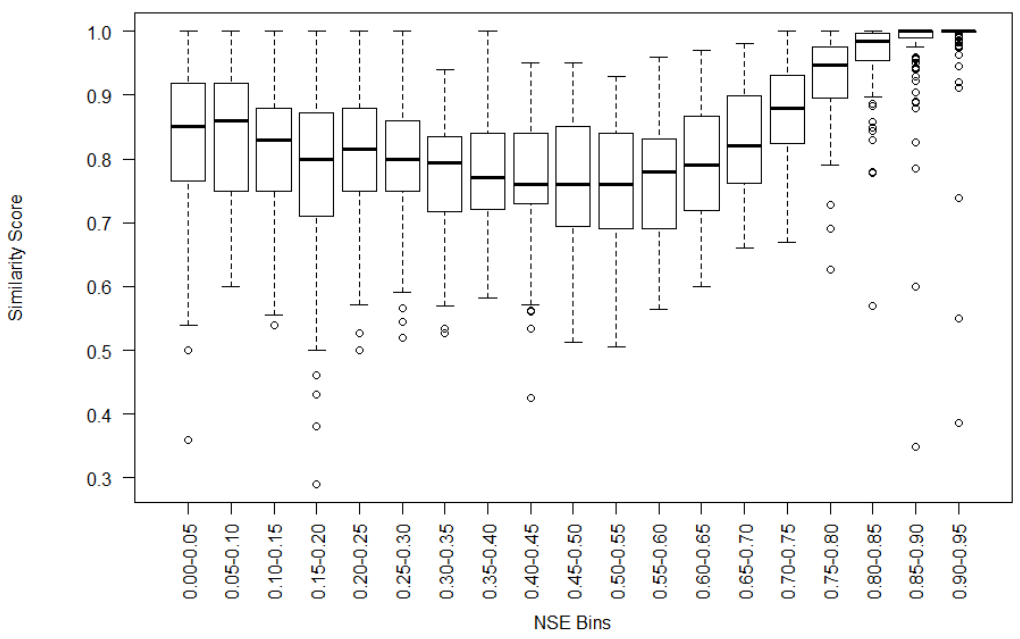


Figure 4.12 – Box whisker plot of similarity score for ecological benefit when calibrated to inflow NSE

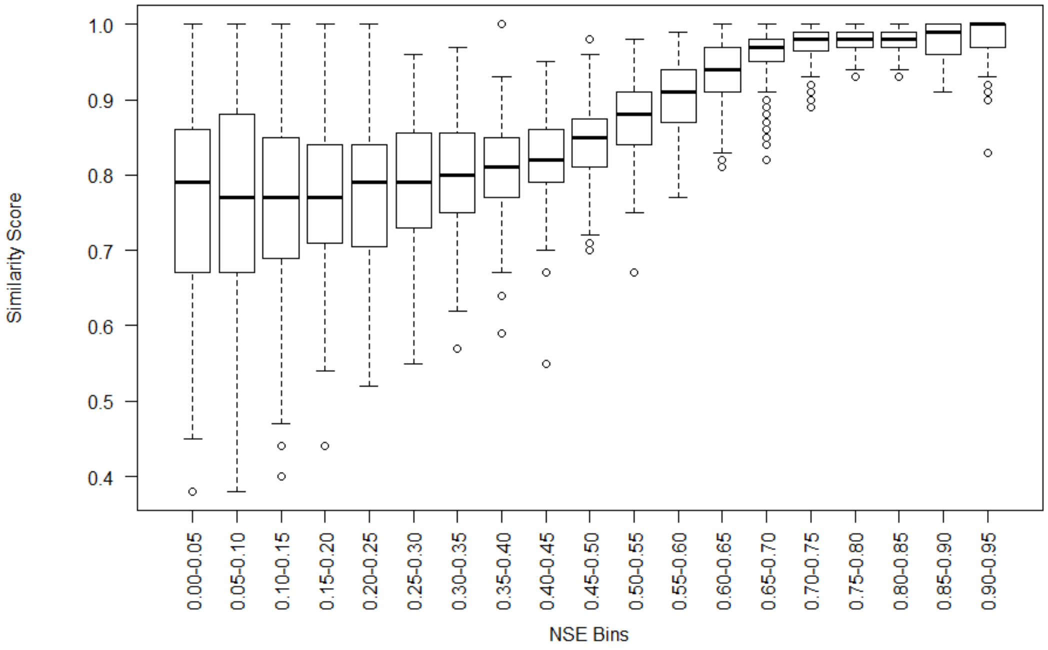


Figure 4.13 – Box whisker plot of similarity score for flood damage reduction benefit when calibrated to inflow NSE

In order to compare impact of gauge selection for calibration on decision making ability, the DCT was performed with calibration to NSEs of flow gauges. Individual box whisker plots can be found in Appendix D. Figures 4.14 to 4.16 show comparison of mean similarity score for calibration to each calibration objectives. Calibration to inflow showed better similarity score than calibration to flow gauges across all evaluation criteria when NSE > 0.65.

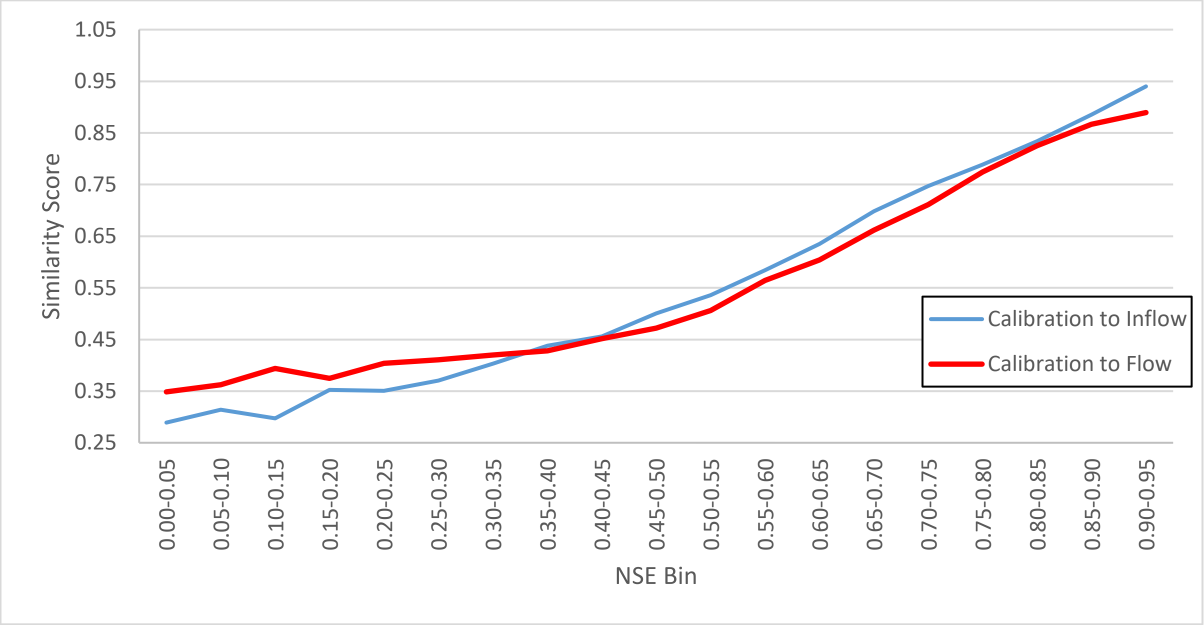


Figure 4.14 – Comparison of mean similarity score for economic benefit based on calibration to inflow versus calibration to stream flow

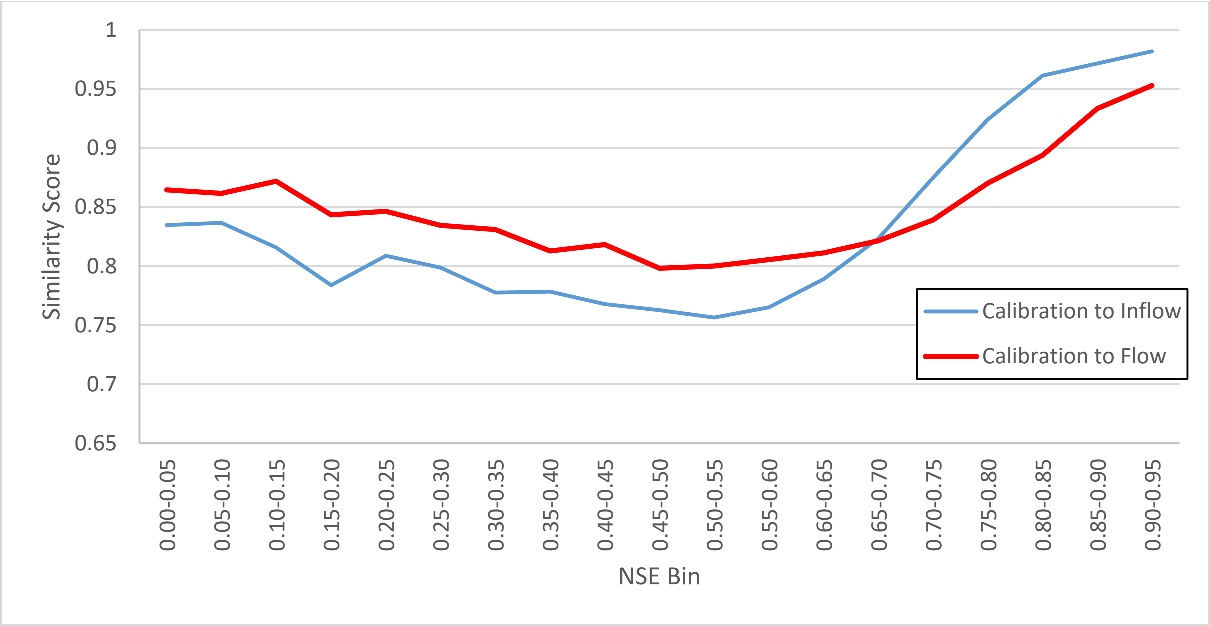


Figure 4.15 – Comparison of mean similarity score for ecological benefit based on calibration to inflow versus calibration to stream flow

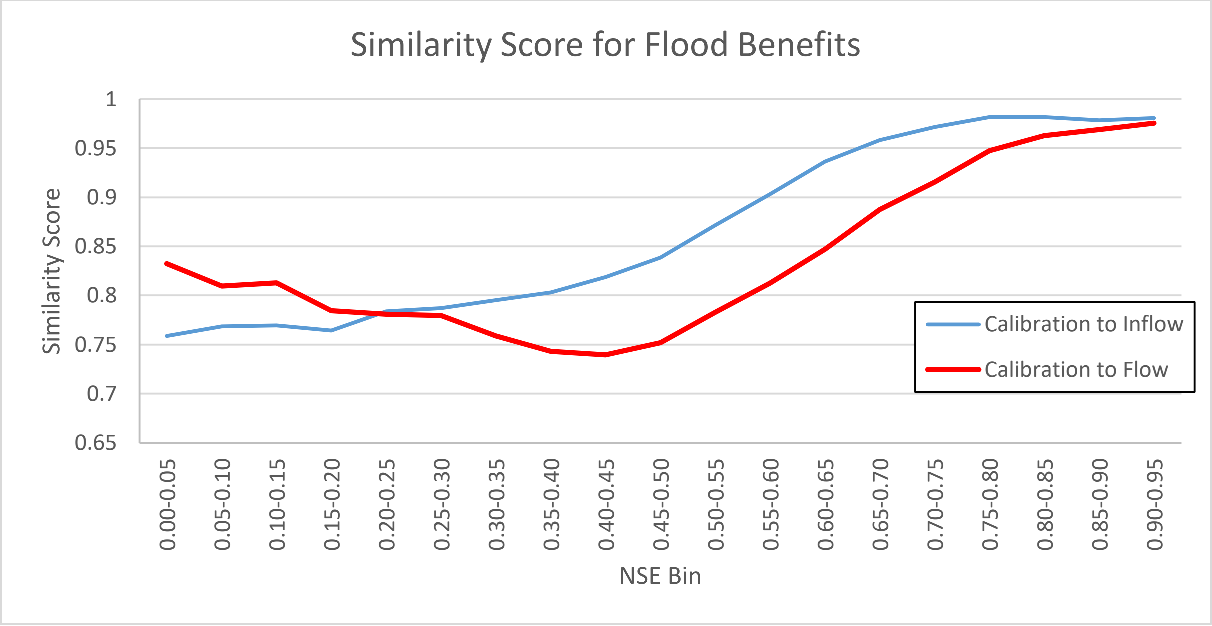


Figure 4.16 – Comparison of mean similarity score for flood damage reduction benefit based on calibration to inflow versus calibration to stream flow

Results from DCT can be utilized to understand the limitations of the initial case study of the Lake of the Woods watershed. The model was calibrated to a flow NSE of 0.70. Results from DCT would indicate that the probability of calibrated model correctly ranking rule curves for economic benefit is approximately 70%. This probability would likely be lower in reality, as prediction in reality is more difficult than prediction in synthetic reality. However, once the model has been calibrated to inflow, the probability increases, as the running average NSE of inflow for base mode is 0.89. The probability of the inflow calibrated model correctly ranking rule curves for economic benefits is nearly 90%, suggesting that inflow calibration strategy is preferable for this decision making application.

4.4.2 DCT Applied on Different Calibration Objectives

In order to determine optimal calibration objective for model application to inform decision making, DCT was performed on calibration to inflow to different objective function formulations. The objective functions are:

1. Calibration to inflow NSE
2. Calibration to inflow KGE
3. Calibration to inflow NSE penalized by Percent Bias, as shown in equation 13
4. Calibration to spring inflow NSE (March to June) where flow is highest

Mean similarity score of each objective function is shown in Figure 4.17 to 4.19.

$$\text{PBIAS penalized NSE} = \text{NSE} - \frac{\max(0,|\text{PBIAS}|-10)}{100} \qquad (13)$$

Results showed little difference between mean similarity score values across NSE, KGE, and Spring NSE. However, inclusion of percent bias into the objective function increased the mean similarity score by nearly 10% across all evaluation criteria until all objective functions converged near 0.75.

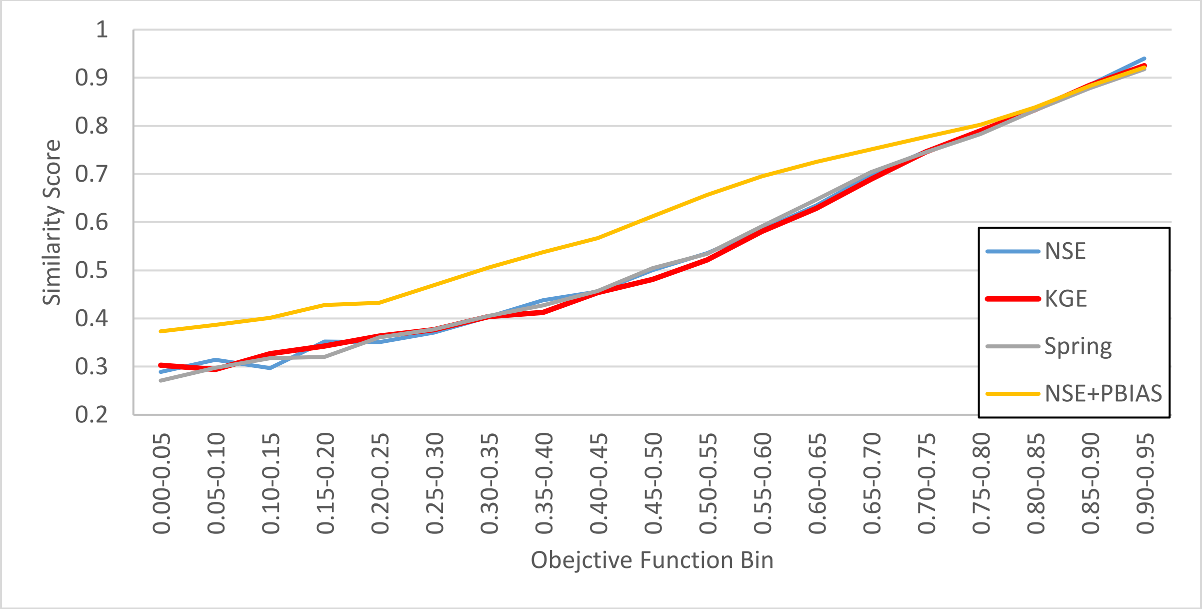


Figure 4.17 – Comparison of mean similarity score for economic benefit based on calibration to different calibration objective functions

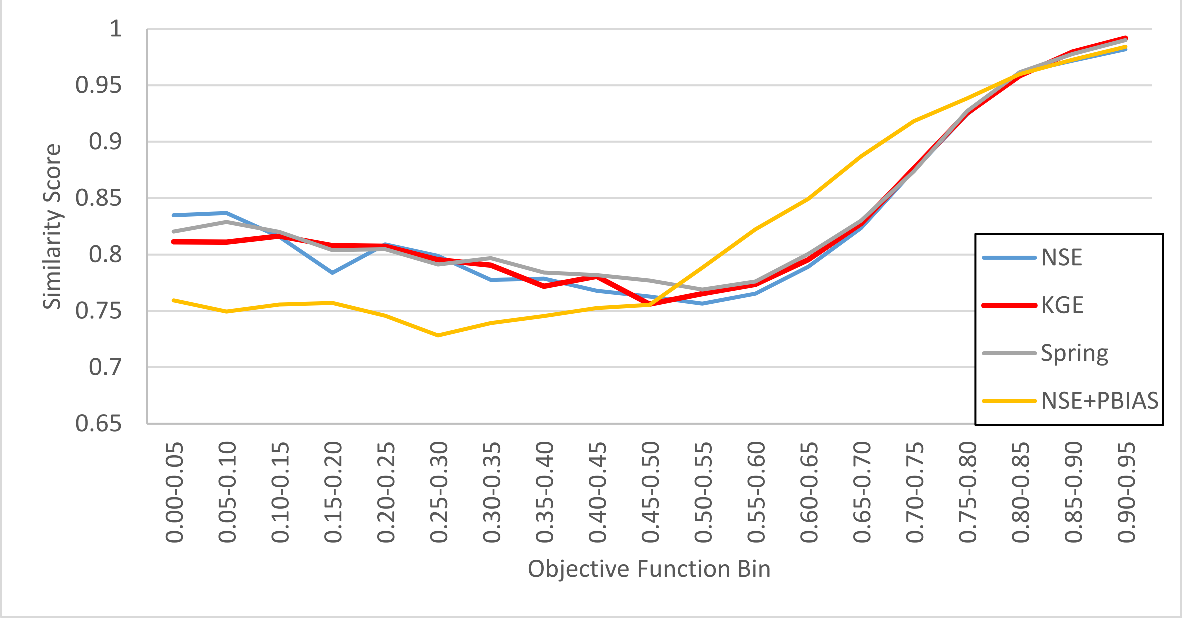


Figure 4.18 – Comparison of mean similarity score for ecological benefit based on calibration to different calibration objective functions

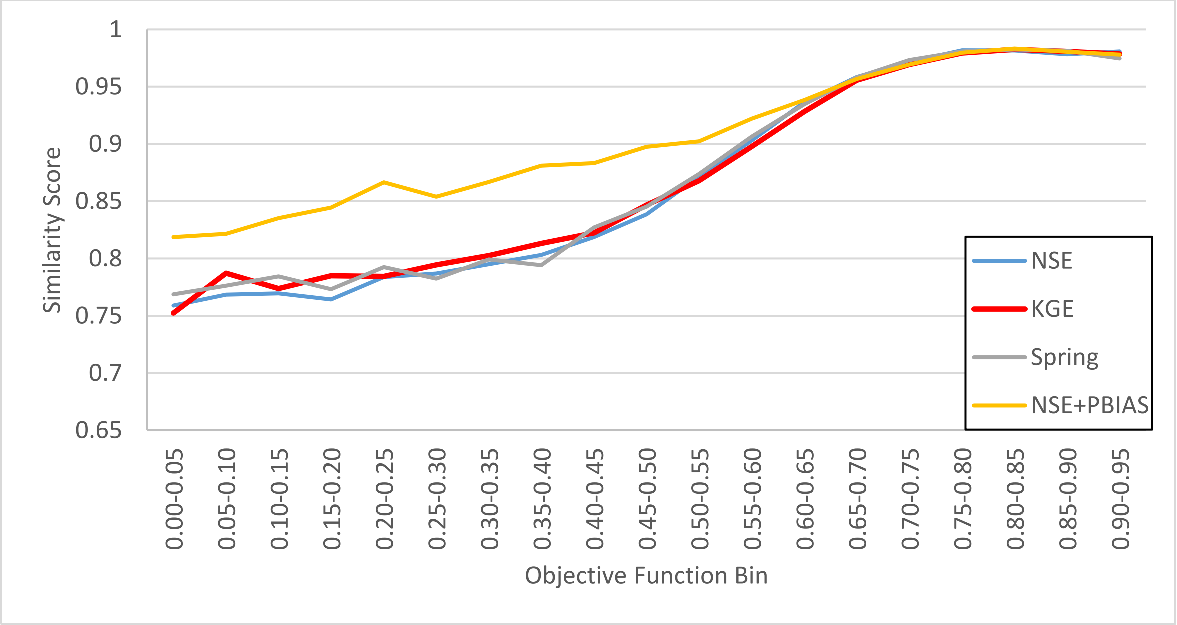


Figure 4.19 – Comparison of mean similarity score for flood damage reduction benefit based on calibration to different calibration objective functions

4.4.3 DCT to Assess Error in Evaluation Criteria

Results from Sections 4.4.2 and 4.4.3 showed that all calibration strategies converged to a near perfect fit with sufficient calibration budget, due to the chosen parameter sampling strategy. Such near perfect fit is extremely difficult to achieve in non-synthetic calibration, due to error in model structure, hydrologic data utilized in model, and observation data. To compensate for these errors, NSE and decision for each evaluation criteria was recorded at every model run to assess decision making capabilities across a wide range of NSE, as end results alone would not be beneficial. Another method to compensate is to restrict the calibration budget. Based on previous experiments, a calibration budget of 50 model runs resulted in a NSE of approximately 0.80 across the flow weighted average of 11 flow gauge NSEs and the flow weighted average of four inflow NSEs, similar to the LOWRL model used in the case study in Chapter 3. DCT was performed with 50 model runs as calibration budget, with combination of different gauges used in calibration objective. At the end of each calibration run, the model was evaluated on whether it made the correct rule curve choice across each evaluation criteria. Additionally the model was evaluated on whether it made the correct rule curve choices across all evaluation criteria. Next, the error of the model in each evaluation criteria was calculated. For each rule curve operation, the model error in ecological benefit was calculated by taking the difference in average probability of survivability between the calibrated model run and the synthetic parameter set driven model run. The model error in economic benefit was calculated by taking the difference in total volume of water released over the spillway (in %) between the calibrated model run and the synthetic parameter set driven model run. The model error in flood damage reduction was calculated by taking the difference between peak stage during a storm even in the calibrated model run and the synthetic parameter set driven model run. In these experiments, emphasis was given in DCT to evaluate selection of different gauges for calibration objective. DCT was performed with flow weighted average NSE across four different combination of flow gauges and inflows. The four gauge selections for calibration were:

1. All gauges and inflows in the LOWRL basin
2. All gauges and inflows upstream and including Rainy Lake
3. Rainy Lake inflow
4. Big Fork River flow gauge – a downstream flow gauge with shared parameters

Similarity score at the end of the DCT experiment for each evaluation criteria was calculated using equation 13.

$$\text{Similarity score} = \frac{\text{\# of Correct calibrated model runs}}{\text{Total \# of model runs}} \qquad (13)$$

Table 4.2 shows the similarity scores for each evaluation criteria across four different calibration objective formulations.

Table 4.2 – Similarity score across each evaluation criteria across four different calibration objective gauge formulations

	Environmental	Economic	Flood Damage Reduction	All
All Gauges	89	87	99	77
Upstream	91	94	100	87
Rainy	95	98	98	91
Big Fork	76	21	54	8

Results showed that not utilizing hydrologic data of Rainy Lake resulted in poor similarity score. Calibrating to Big Fork alone resulted in a similarity score of 21 for economic benefit, indicating model calibrated to Big Fork alone has a 21% probability of informing the correct decision based on economic benefits. However, similarity score for environmental benefits is relatively high. This is consistent with the findings in previous section where model performance has relatively low impact on decision making when it comes to criteria dictated by rule curve operations. Amongst the three calibration formulations that include Rainy Lake, calibrating to Rainy Lake alone resulted in highest similarity score across all evaluation criteria. The similarity score for making the correct decision in all evaluation criteria improved by 14 when calibrated to Rainy Lake alone, compared to calibrating to all gauges available. Figures 4.20 - 4.22 show error in calibrated model runs across each evaluation criteria. Figure 4.21 showed that the error in volume of water over spillway varies more significantly across the three gauge selections compared to the error in probability of survivability. Calibrating to Rainy Lake alone resulted in a mean error of -25% for when using rule curve A, while calibrating to all gauges resulted in a mean error of -40.5% when using rule curve A. Figure 4.21 showed that the error in volume of water over spillway varies more significantly across the three gauge selections compared to the error in probability of survivability. Calibrating to Rainy Lake alone resulted in a mean error of -25% for when using rule curve A, while calibrating to all gauges resulted in a mean error of -40.5% when using rule curve A. Figure 4.22 showed that the error in flood damage reduction varies more significantly across the three gauge selections as well. Calibrating to Rainy Lake alone resulted in a mean error of -0.7 m when using rule curve A, while calibrating to all gauges resulted in a mean error of -0.9 m when using rule curve A.

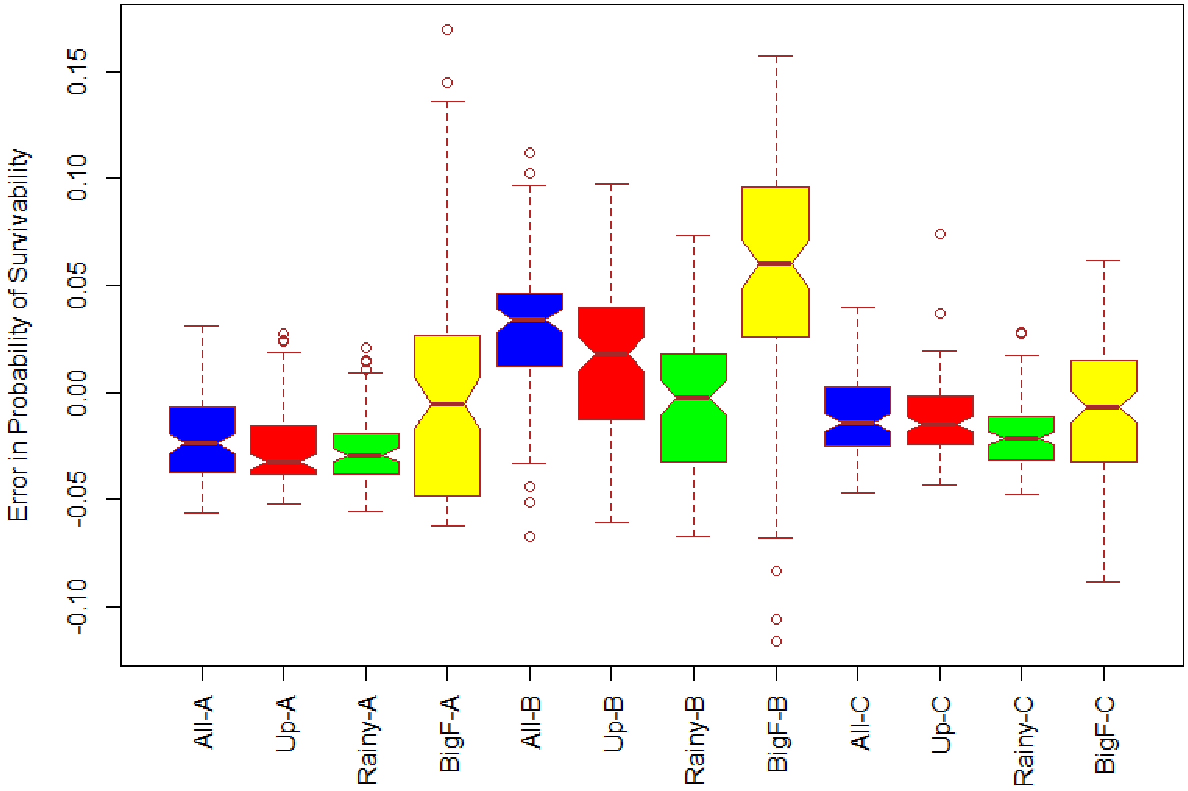


Figure 4.20 – Box whisker plot of calibrated model error in environmental benefit. X-axis shows gauges used in calibration and rule curve utilized.

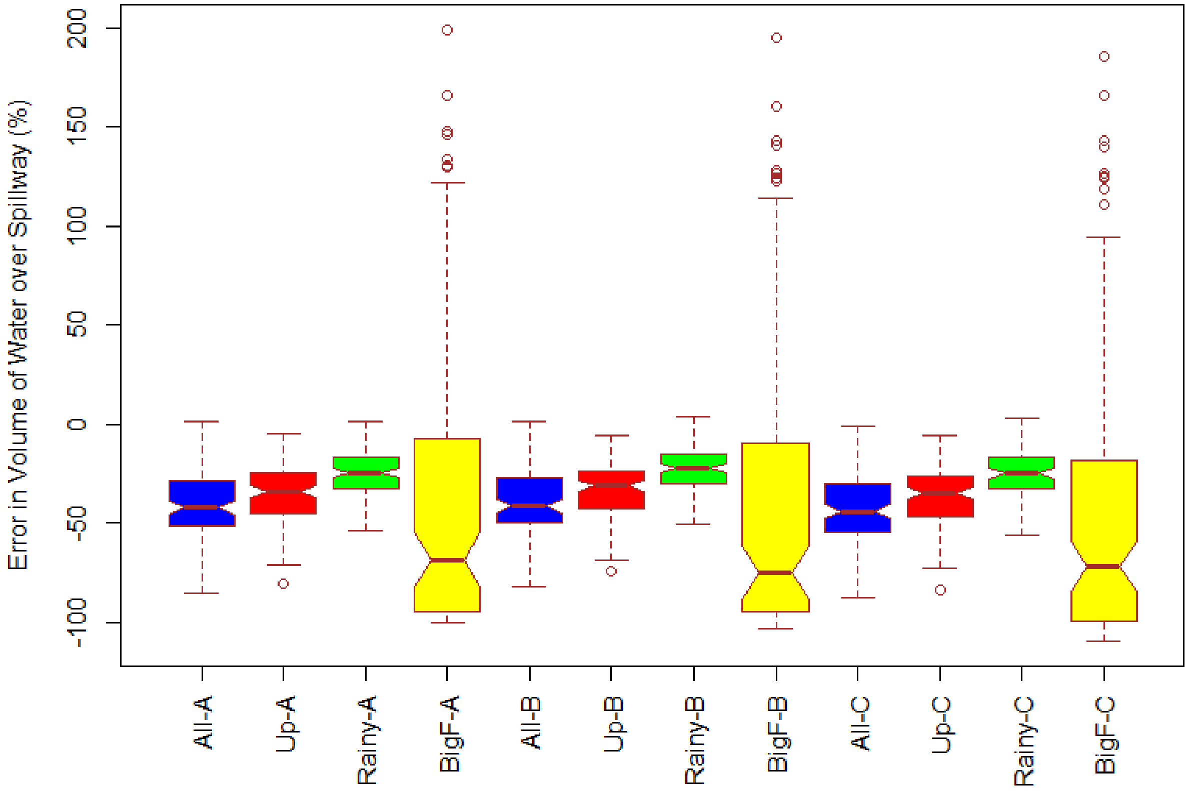


Figure 4.21 – Box whisker plot of calibrated model error in economic benefit. X-axis shows gauges used in calibration and rule curve utilized.

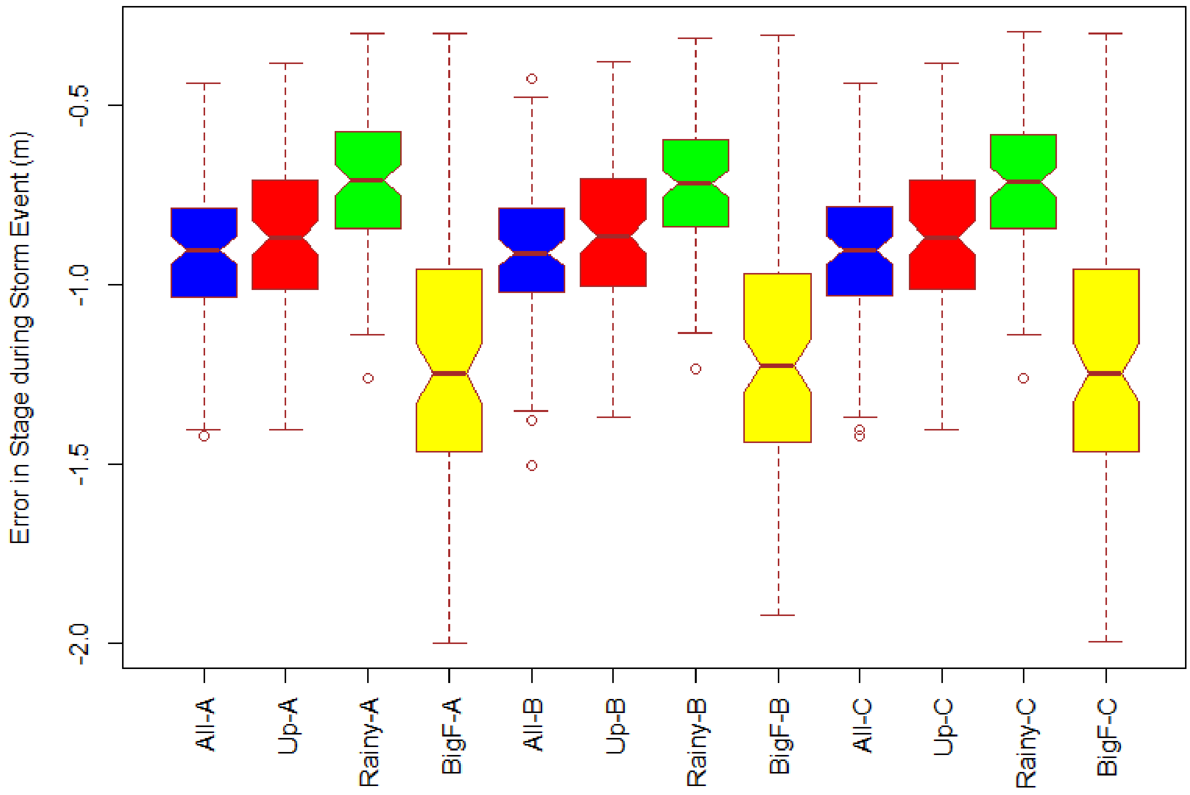


Figure 4.22 – Box whisker plot of calibrated model error in flood damage reduction benefit. X-axis shows gauges used in calibration and rule curve utilized.

When calibrating to all gauges, performance at Rainy Lake inflow would have been deterred to increase performance at other gauges. Worse similarity score in all-gauge calibration scenario may be due to lower performance in Rainy Lake inflows. In real life scenarios, there may be cases where similar performance at one gauge while achieving better performance at other gauges. This would require additional calibration budget, as more objective functions need to be optimized. However, if similar performance in Rainy Lake inflow can be achieved with better performance at other gauges, it may indicate that the right numbers are achieved at Rainy Lake for the right reasons. To test such scenario, DCT was performed across the three gauge combinations that include Rainy Lake inflow (all gauge, upstream, and Rainy Lake alone) with a calibration budget of 10,000. However, the calibration was set to terminate when Rainy Lake inflow NSE was equal or greater to 0.7. All calibrations in the DCT would have a Rainy Lake inflow NSE of approximately 0.7, but varying NSE across different gauges, depending on the calibration objective.

Table 4.3 shows the similarity scores for each evaluation criteria across the three different calibration objective formulations with calibration stopping when Rainy Lake inflow NSE ≥ 0.7.

Table 4.3 – Similarity score across each evaluation criteria across four different calibration objective gauge formulations

	Environmental	**Economic**	**Flood Damage Reduction**	**All**
All Gauges	80	80	94	65
Upstream	86	76	96	62
Rainy	85	71	96	62

Similarity score shows slight improvement overall when calibrated to all gauges, compared to calibrating to Rainy Lake alone. Figures 4.23 to 4.25 show that errors in evaluation criteria vary little across selection of gauges used for calibration.

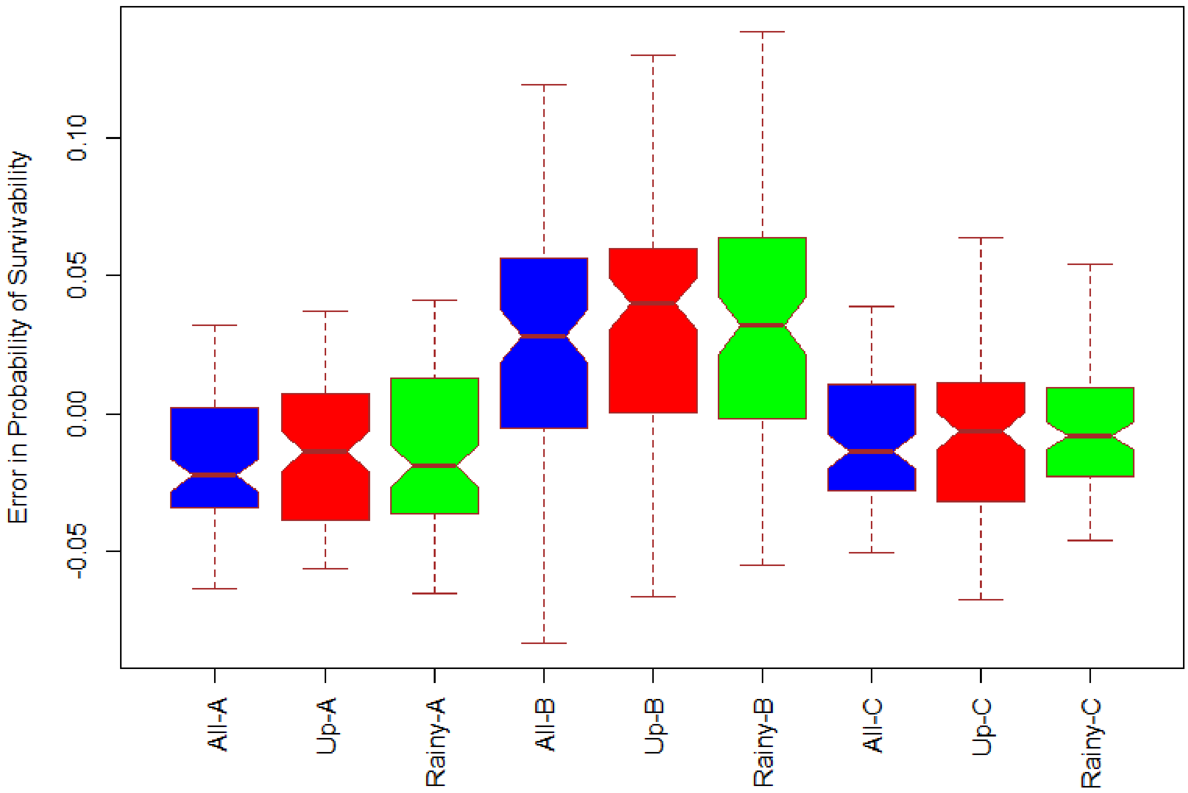


Figure 4.23 – Box whisker plot of calibrated model error in environmental benefit and stopped when Rainy Lake NSE ≥ 0.7. X-axis shows gauges used in calibration and rule curve utilized.

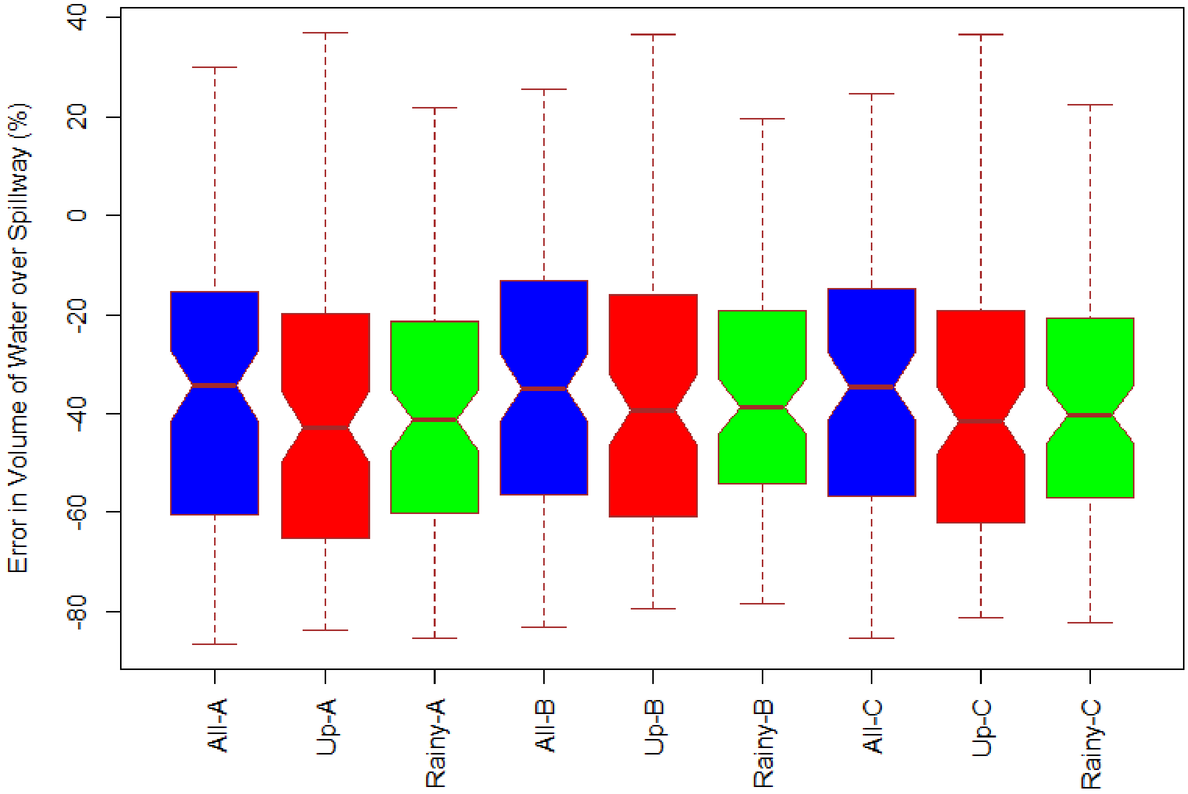


Figure 4.25 – Box whisker plot of calibrated model error in economic benefit and stopped when Rainy Lake NSE ≥ 0.7. X-axis shows gauges used in calibration and rule curve utilized.

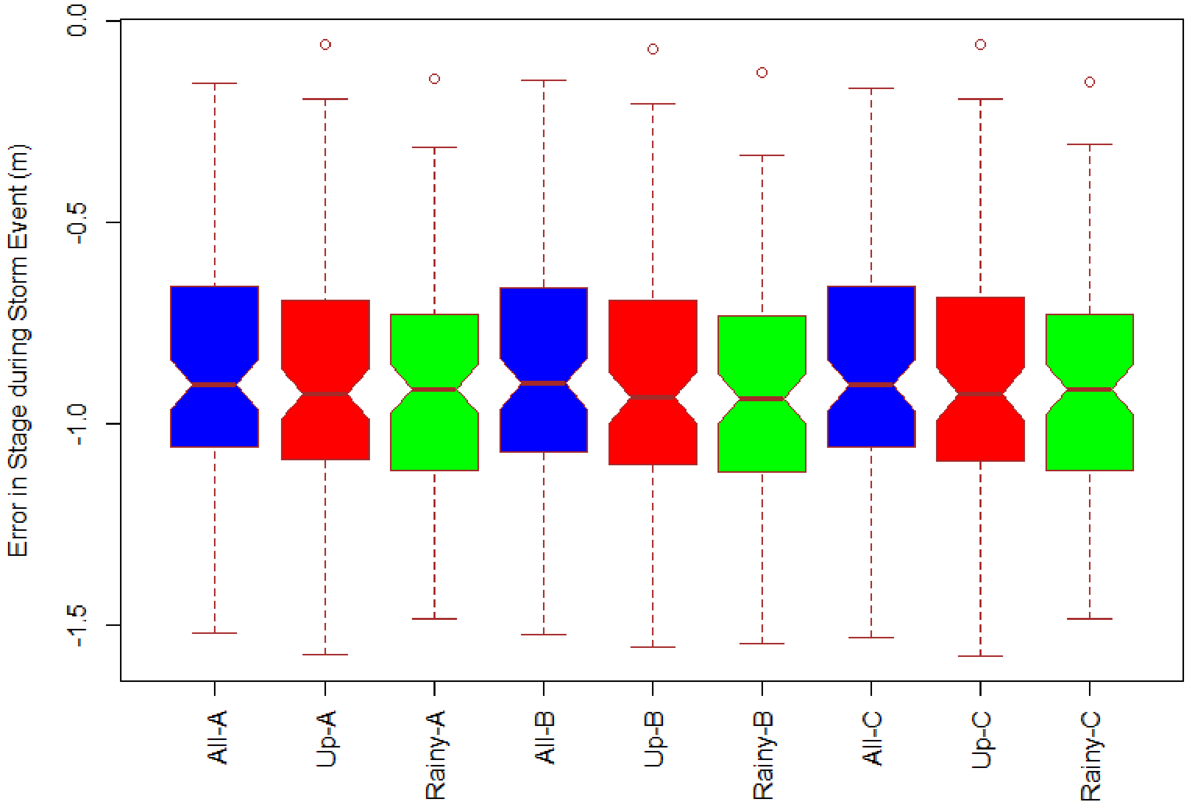


Figure 4.26 – Box whisker plot of calibrated model error in flood damage reduction benefit and stopped when Rainy Lake NSE ≥ 0.7. X-axis shows gauges used in calibration and rule curve utilized.

Next, DCT with 50 calibration budget was performed using different calibration diagnostics to compare calibrated model error when calibration was performed to different diagnostics. Model was calibrated to the flow weighted average of all 11 gauges and flow weighted average of all four inflows. The diagnostics used for the experiments were NSE, KGE, percent bias penalized NSE, and percent bias. Similarity score across the diagnostics are summarized in Table 4.4.

Table 4.4 – Similarity score across each evaluation criteria across four different calibration objective diagnostics: NSE, KGE, percent bias penalized NSE (NSEP), and percent bias, when calibrated with a budget of 50 model runs

	Environmental	Economic	Flood Damage Reduction	All
NSE	89	87	99	77
KGE	96	87	96	81
NSEP	90	83	94	75
PBIAS	77	63	80	42

Results remain consistent with findings in Section 4.4.2. Calibration to NSE, KGE, and NSEP resulted in similar similarity score when the objective function values were above 0.8. However, as shown in Figures 4.27 – 4.29, calibration to KGE showed significant reduction in errors in economic benefits and flood damage reduction benefits.

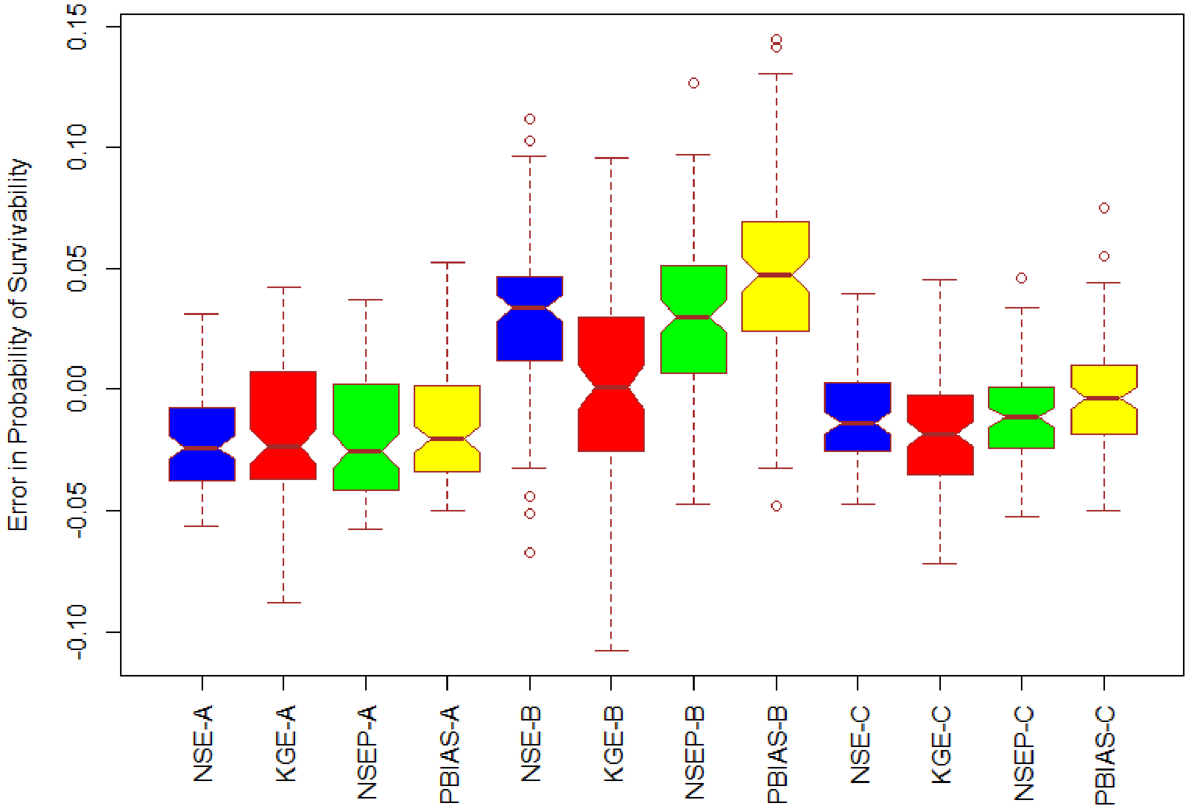


Figure 4.27 – Box whisker plot of calibrated model error in environmental benefit. X-axis shows diagnostic used in calibration and rule curve utilized.

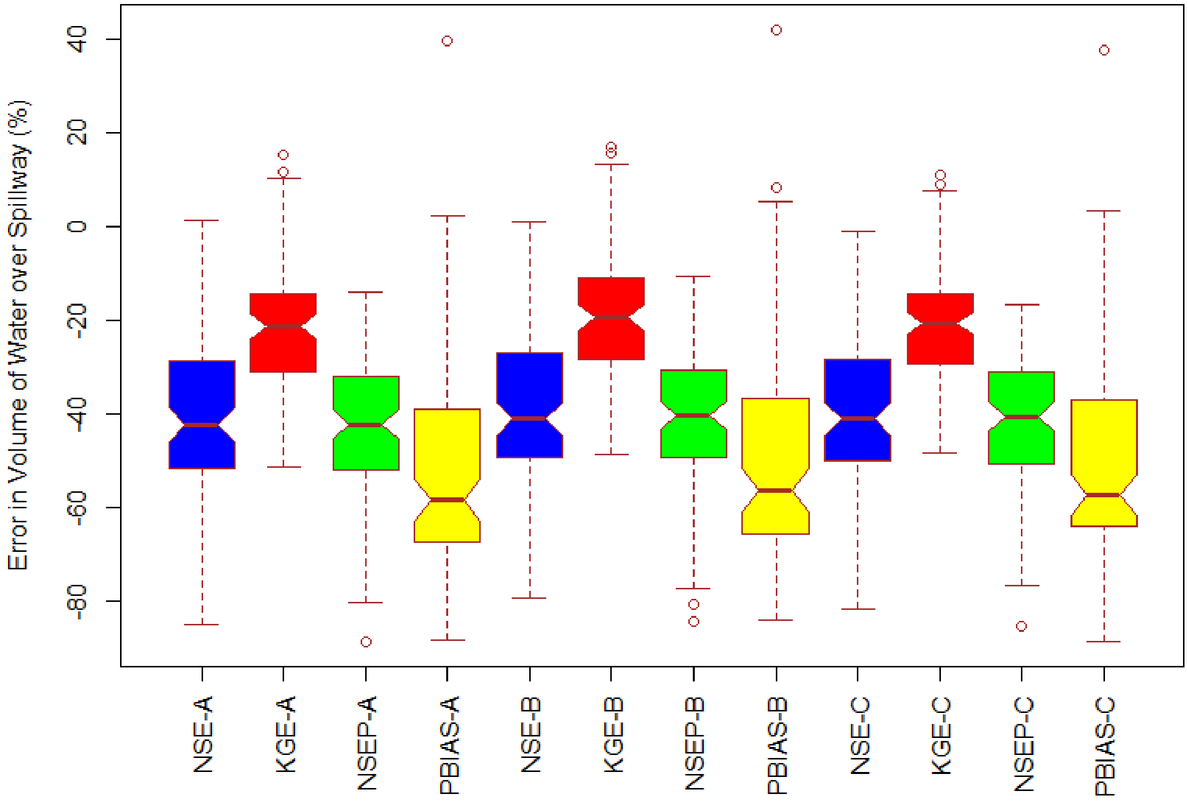


Figure 4.28 – Box whisker plot of calibrated model error in economic benefit. X-axis shows diagnostic used in calibration and rule curve utilized.

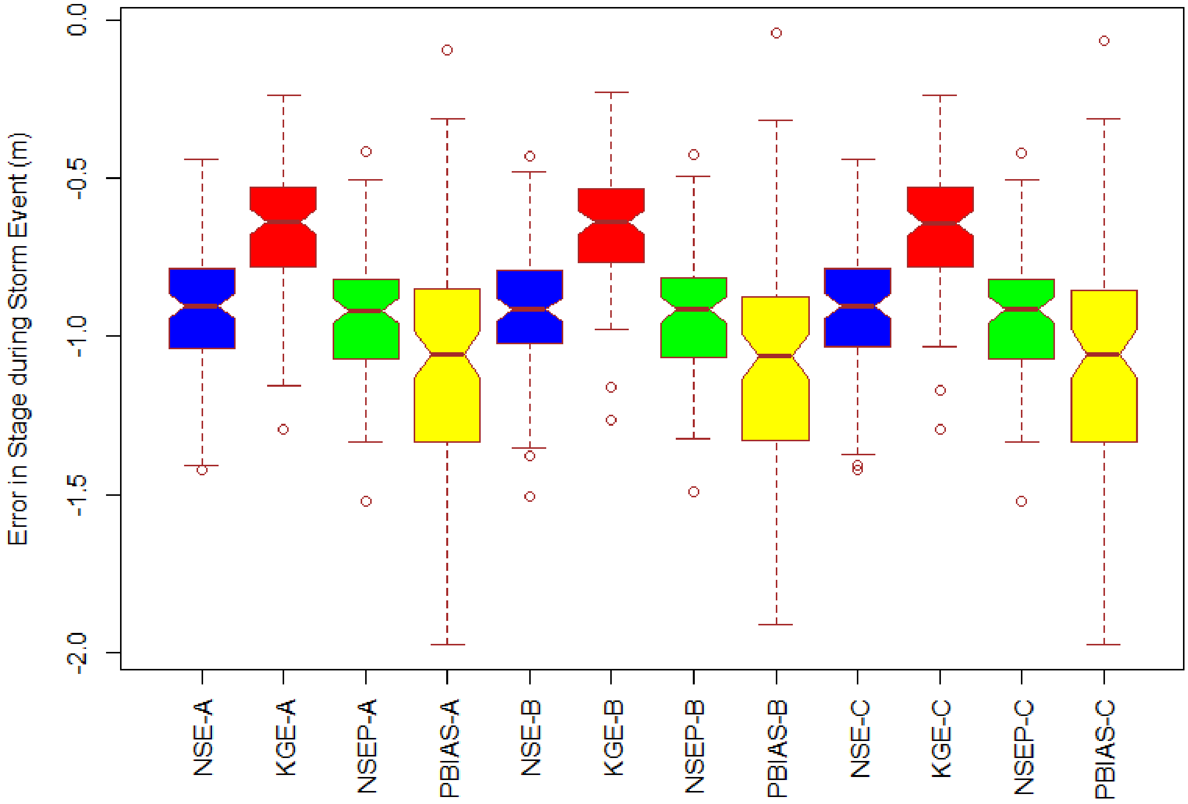


Figure 4.29 – Box whisker plot of calibrated model error in flood damage reduction benefit. X-axis shows diagnostic used in calibration and rule curve utilized.

Calibration to KGE resulted in noticeably lower error in economic benefit and flood damage reduction benefit. The mean error in volume of water over spillway (%) using rule curve A when calibrated to KGE was -21.5 %, where the mean error was -40.5 % when calibrated to NSE. The mean error in stage (m) during a flood event using rule curve A when calibrated to KGE was -0.66 m, where the mean error was -0.92 m when calibrated to NSE. Gupta et al. (2009) argue that calibrating to NSE results in a tendency of runoff peak to be systematically underestimated. The economic benefit and flood damage reduction benefit are closely related to variability of flow, which may benefit from improved performance in variability measures when calibrating to KGE. However, since calibration to NSE underestimated the error in all rule curve operations, the ranking of rule curves has minimal impact, resulting in minimal change in similarity score.

4.5 Case Study Conclusions

The result of Decision Crash Testing may provide valuable information to the limitations and usefulness of the models. Also, it provides a means of identifying the best formulation for calibration objective of the specific decision making scenario.

First, results of DCT show that the usefulness of model is heavily dependent on the evaluation criteria used to assess model quality. Usefulness of model in predicting hydrologic variables controlled by operation is limited. This is because stage related variables are largely controlled by implementation of reservoir operations, not model performance. Determination of what defines a *good* NSE requires careful assessment of intended use of the model, as a model with 0.1 NSE may be *good enough* in ranking rule curves for stage related evaluation criteria, while a model with a NSE of 0.8 may be insufficient for ranking rule curves for peak flows.

Second, calibration to observation data which is more closely related to the intended use of model may be beneficial for operational purposes. Calibration to 11 stream gauges was found to be a more difficult target to achieve compared to calibration to four inflow observations, as more spatially distributed gauges require the model to better represent reality (i.e., getting the right answers for the right reasons). When making decisions predicated on simulated reservoir stage values, utilizing models calibrated to reservoir inflow resulted in higher probability of making the correct decision. However, these calibration strategies do not need to be mutually exclusive. Two base models calibrated using each strategy can be used to rank rule curves, and the user can have a better understanding in the uncertainty of the rankings made by each model, without having to run additional DCT experiments. Section 4.4.3 demonstrated importance of prioritizing gauges to be optimized. Reservoir that directly impacts decision making should be prioritized. Improving performance at other gauges without sacrificing performance in high priority zone may result in slight increase in error reduction.

Third, inclusion of additional diagnostics, such as KGE and percent bias, into the calibration objective function can result in objectively better performance in decision making and error calculations in decision making. Various literatures emphasizes the importance of inclusion of multiple hydrologic diagnostics to calibration in order to assure model is getting the right answers for the right reasons (Gupta et al., 2012, Euser et al., 2013, Biondi et al., 2012). The DCT experiments in this thesis indicate showed that requirement for hydrologic realism may also have practical benefits in increasing similarity scores and reduction in calibrated model error specific to certain decision making scenarios.

The DCT experiments took a simplified approach in synthetic scenario sampling. The results from the experiments may be overestimating model performance in informing decision making, as a perfect solution to the parameter set exists for the calibration problem. This may also weaken the variability across performance of different calibration objective formulations. It may be beneficial to introduce uncertainty in synthetic observation, through change in model structure and introduction of noise to data.

Chapter 5

Conclusions

In this section, the thesis' contribution to literature and future opportunities are discussed.

5.1 Contribution to Literature

The result of this thesis clearly addresses the basic concept introduced by Klemes (1986): models need to be tested for their intended use. The case study in the thesis required a model for operational rule curve section. In order to model rule curve operation, novel approaches were developed to model lakes and reservoir operation in the Canadian Shield. Then, the Decision Crash Testing (DCT) method was utilized to establish correlation between traditional model diagnostics and model utility. The correlation helps illustrate the limitations and usefulness of a model in a clear manner understandable by most stakeholders. Furthermore, the DCT can be used to assess the results of different calibration formulations in an objective manner. Researchers and model users continuously debate on the importance of getting the right answer for the right reason. Through a novel approach on defining what the *right answer* is (more than simply a "high" NSE), results showed an increase in similarity score when additional diagnostics other than NSE were incorporated to enhance hydrological adequacy.

5.2 Future Opportunities for DCT

This thesis demonstrates an evaluation method to test model on its intended use, specifically applied to examine the appropriateness and utility of different objective function choices. Furthermore, it allows objective comparison between different calibration formulations. DCT should be implemented to test a wider variety of calibration objectives. The key requirements for DCT formulation in this thesis include:

1. Clear formulation of evaluation criteria based upon model output
2. Generation of synthetic reality using random sampling of parameters
3. Calibration objective formulation
4. Explicit decision making scenario (curve ranking)

In reality, model output may not easily be transformed into a decision making scenario. Often, decision involves human intervention and judgement, difficult concepts to incorporate into a model. It may be suitable to incorporate uncertainty in transforming model output to decision. A big hurdle during the DCT experiment was the quick convergence during calibration, relatively independent of calibration objective formulation. Due to random sampling of parameters without change in model structure, a perfect solution parameter set always exists for calibration. As a result, all calibration formulation showed to perform extremely well when provided enough budget to converge. In reality, such high performance is rarely achieved. Methods such as uncertainty in forcing data, adding complexity to model structures, and statistical methods in parameter sampling may address the issue by creating synthetic observations where no solution parameter set exists. In this thesis, each evaluation criteria was observed independently from each other. However, in reality, decision making process in hydrology is often a multi objective problem. Incorporation of multi-objective calibration into DCT may enrich the benefits of DCT in assessment of objective formulation.

References

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., … Valéry, A. (2009). Crash tests for a standardized evaluation of hydrological models. *Hydrology and Earth System Sciences*, *13*(10), 1757–1764. https://doi.org/10.5194/hess-13-1757-2009

Asadzadeh, M., & Tolson, B. (2013). Pareto archived dynamically dimensioned search with hypervolume-based selection for multi-objective optimization. *Engineering Optimization, 45*(12), 1489-1509. doi:10.1080/0305215x.2012.748046

Ascough, J. C., Maier, H. R., Ravalico, J. K., & Strudley, M. W. (2008). Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological Modelling*, *219*(3–4), 383–399. https://doi.org/10.1016/j.ecolmodel.2008.07.015

Beven, K. (2001). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, *5*(1), 1–12. https://doi.org/10.5194/hess-5-1-2001

Bierkens, F. P. B. (2015). Water Resources Research. *Global Hydrology 2015: State, Trends, and Directions Marc*, *51*, 4923–4947. https://doi.org/10.1002/2015WR017173.Received

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*(1–4), 242–266. https://doi.org/10.1016/j.jhydrol.2004.03.042

Canadian Hydraulics Centre National Research Council. (2010). *Rainy River 2D Hydrodynamic Model Conveyance Study*(Rep.).

Chlumsky, R. (2017). Rigorous Validation of Hydrologic Models in Support of Decision-Making, 128.

Chow, V. Te, Maidment, D. R., & Mays, L. W. (1988). Applied Hydrology. *Water Resources and Environmental Engineering*. https://doi.org/10.1016/j.soncn.2011.11.001

Cooper, M. (2010). Advanced Bash-Scripting Guide An in-depth exploration of the art of shell scripting Table of Contents. *Okt 2005 Abrufbar Uber Httpwww Tldp OrgLDPabsabsguide Pdf Zugriff 1112 2005*, *2274*(November 2008), 2267–2274. https://doi.org/10.1002/hyp

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., & Hendrickx, F. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research*, *48*(5), 1–17. https://doi.org/10.1029/2011WR011721

Craig, J.R., and the Raven Development Team, Raven user's and developer's manual (Version 2.8), URL: http://www.civil.uwaterloo.ca/jrcraig/Raven/Main.html (Accessed May, 2018).

D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, & T. L. Veith. (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*, *50*(3), 885–900. https://doi.org/10.13031/2013.23153

Daniela Biondi, Gabriele Freni, Vito Iacobellis, Giuseppe Mascaro, A. M. (2012). Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. *Physics and Chemistry of the Earth*, *42–44*.

Dong, F., Liu, Y., Su, H., Zou, R., & Guo, H. (2015). Reliability-oriented multi-objective optimal decision-making approach for uncertainty-based watershed load reduction. *Science of the Total Environment*, *515–516*, 39–48. https://doi.org/10.1016/j.scitotenv.2015.02.024

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, *17*(5), 1893–1912. https://doi.org/10.5194/hess-17-1893-2013

Guillaume, J. H. A., Kummu, M., Räsänen, T. A., & Jakeman, A. J. (2015). Prediction under uncertainty as a boundary problem: A general formulation using Iterative Closed Question Modelling. *Environmental Modelling and Software*, *70*, 97–112. https://doi.org/10.1016/j.envsoft.2015.04.004

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*(8), 1–16. https://doi.org/10.1029/2011WR011044

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

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

Haasnoot, M., Kwakkel, J. H., Walker, W. E., & ter Maat, J. (2013). Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. *Global Environmental Change*, *23*(2), 485–498. https://doi.org/10.1016/j.gloenvcha.2012.12.006

Hassan, A. E. (2004). Validation of Numerical Ground Water Models Used to Guide Decision Making. *Ground Water*. https://doi.org/10.1111/j.1745-6584.2004.tb02674.x

Herckenrath, D., Langevin, C. D., & Doherty, J. (2011). Predictive uncertainty analysis of a saltwater intrusion model using null-space Monte Carlo. *Water Resources Research*, *47*(5), 1–16. https://doi.org/10.1029/2010WR009342

International Rainy and Namakan Lakes Rule Curves Study Board. (2017). *Managing Water Levels and Flows in the Rainy River Basin*(Rep.).

Kavetski, D., & Kuczera, G. (2007). Model smoothing strategies to remove microscale discontinuities and spurious secondary optima im objective functions in hydrological calibration. *Water Resources Research*, *43*(3), 1–9. https://doi.org/10.1029/2006WR005195

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research, 42*(3). doi:10.1029/2005wr004362

KlemeŠ, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal, 31*(1), 13–24. https://doi.org/10.1080/02626668609491024

Krause, P., Boyle, D. P., & Bäse, F. (2005). Advances in Geosciences Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, *5*(89), 89–97. https://doi.org/10.5194/adgeo-5-89-2005

Lake of the Woods Control Board. (2016). *LOWRL Model Description*(Rep.).

Legates, D. R., & McCabe Jr., G. J. (2005). Evaluating the Use of "Goodness of Fit" Measures in Hydrologic and Hydroclimatic Model Validation. *Water Resources Research*, *35*(1), 233–241. https://doi.org/10.1029/1998WR900018

Liu, H. (2017). *Updated Kam Hydrologic Modelling Report 20180502, Prepared for Ontario Power Generations*(Rep.).

Liu, Y., Gupta, H., Springer, E., & Wagener, T. (2008). Linking science with environmental decision making: Experiences from an integrated modeling approach to supporting sustainable water resources management. *Environmental Modelling and Software*, *23*(7), 846–858. https://doi.org/10.1016/j.envsoft.2007.10.007\

Matott, LS. 2017. *OSTRICH: an Optimization Software Tool, Documentation and User's Guide, Version 17.12.19.* 79 pages, University at Buffalo Center for Computational Research, www.eng.buffalo.edu/~lsmatott/Ostrich/OstrichMain.html.

Matrosov, E. S., Woods, A. M., & Harou, J. J. (2013). Robust Decision Making and Info-Gap Decision Theory for water resource system planning. *Journal of Hydrology*, *494*, 43–58. https://doi.org/10.1016/j.jhydrol.2013.03.006

Marshall, T., & Foster, R. (2015). An analysis of water temperatures on the Rainy River in relation to critical fish spawning periods, with recommendations on peaking restrictions(Rep.).

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).

Pappenberger, F., & Beven, K. J. (2006). Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Research*, *42*(5), 1–8. https://doi.org/10.1029/2005WR004820

Ravalico, J. K., Maier, H. R., & Dandy, G. C. (2009). Sensitivity analysis for decision-making using the MORE method-A Pareto approach. *Reliability Engineering and System Safety*, *94*(7), 1229–1237. https://doi.org/10.1016/j.ress.2009.01.009

Refsgaard, J. C. (1997). Parameterisation, calibration and validation of distributed hydrological models. *Journal of Hydrology*. https://doi.org/10.1016/s0022-1694(96)03329-x

Refsgaard, J. C., & Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. *Water Resources Research*, *32*(7), 2189–2202. https://doi.org/10.1029/96WR00896

Shafii, M., & Tolson, B. A. (2015). Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research*, *51*(5), 3796–3814. https://doi.org/10.1002/2014WR016520

Singh, V. P., & Woolhiser, D. A. (2002). Mathematical Modeling of Watershed Hydrology. *Journal of Hydrologic Engineering*. https://doi.org/10.1061/(ASCE)1084-0699(2002)7:4(270)

Spence C., Woo M. (2008) Hydrology of the Northwestern Subarctic Canadian Shield. In: Woo M. (eds) Cold Region Atmospheric and Hydrologic Studies. The Mackenzie GEWEX Experience. Springer, Berlin, Heidelberg

Tang, Y., Reed, P., Wagener, T., Tang, Y., Reed, P., & How, T. W. (2005). How effective and efficient are multiobjective evolutionary algorithms at hydrologic model calibration ? To cite this version : HAL Id : hal-00298787 How effective and efficient are multiobjective evolutionary algorithms at hydrologic model calibration ?

Todini, E. (2007). Hydrological catchment modelling: Past, present and future. *Hydrology and Earth System Sciences*, *11*(1), 468–482. https://doi.org/10.5194/hess-11-468-2007

Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resour. Res.*, 43, W01413, doi:10.1029/2005WR004723.

Varouchakis, E. A., Palogos, I., & Karatzas, G. P. (2016). Application of Bayesian and cost benefit risk analysis in water resources management. *Journal of Hydrology*, *534*, 390–396. https://doi.org/10.1016/j.jhydrol.2016.01.007

Vijai, H., Sorooshian, S., & Yapo, P. O. (1999). Status of Automatic Calibration for Hydrologic Models: Comparison with Multilevel Expert Calibration, *4*(April), 135–143.

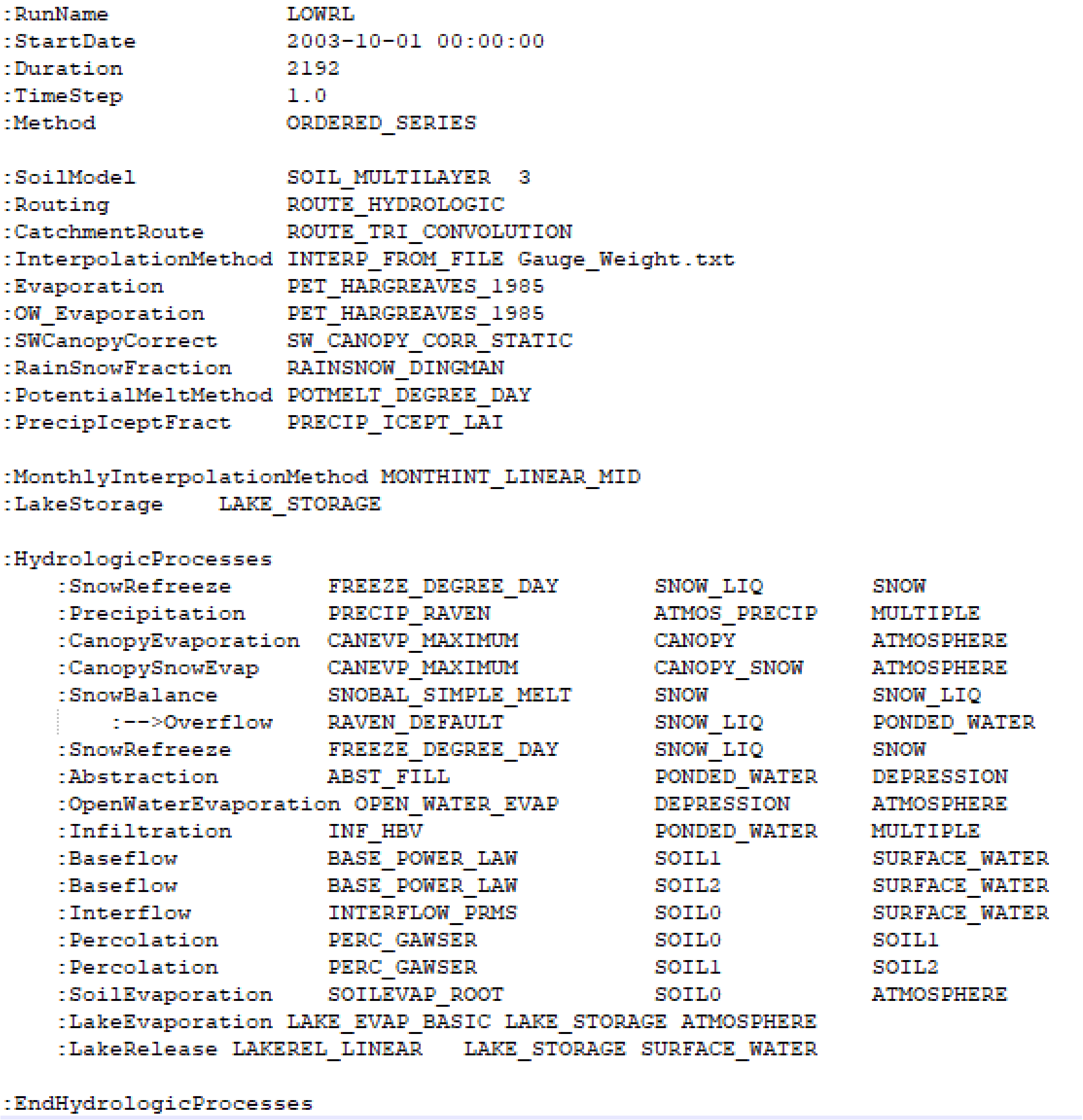
Weglarczyk, S. (1998). The interdependence and applicability of some statistical quality measures. *Journal of Hydrology*.

Xue, J., Gui, D., Zhao, Y., Lei, J., Zeng, F., Feng, X., Shareef, M. (2016). A decision-making framework to model environmental flow requirements in oasis areas using Bayesian networks. *Journal of Hydrology*, *540*, 1209–1222. https://doi.org/10.1016/j.jhydrol.2016.07.017

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), 1–18. https://doi.org/10.1029/2007WR006716

Yu, S., He, L., & Lu, H. (2016). A tempo-spatial-distributed multi-objective decision-making model for ecological restoration management of water-deficient rivers. *Journal of Hydrology*, *542*, 860–874. https://doi.org/10.1016/j.jhydrol.2016.09.055

Appendix A – Raven Input File with Process Selection



```
:RunName                LOWRL
:StartDate              2003-10-01 00:00:00
:Duration               2192
:TimeStep               1.0
:Method                 ORDERED_SERIES

:SoilModel              SOIL_MULTILAYER  3
:Routing                ROUTE_HYDROLOGIC
:CatchmentRoute         ROUTE_TRI_CONVOLUTION
:InterpolationMethod INTERP_FROM_FILE Gauge_Weight.txt
:Evaporation            PET_HARGREAVES_1985
:OW_Evaporation         PET_HARGREAVES_1985
:SWCanopyCorrect        SW_CANOPY_CORR_STATIC
:RainSnowFraction       RAINSNOW_DINGMAN
:PotentialMeltMethod POTMELT_DEGREE_DAY
:PrecipIceptFract       PRECIP_ICEPT_LAI

:MonthlyInterpolationMethod MONTHINT_LINEAR_MID
:LakeStorage     LAKE_STORAGE

:HydrologicProcesses
    :SnowRefreeze       FREEZE_DEGREE_DAY       SNOW_LIQ        SNOW
    :Precipitation      PRECIP_RAVEN            ATMOS_PRECIP    MULTIPLE
    :CanopyEvaporation  CANEVP_MAXIMUM          CANOPY          ATMOSPHERE
    :CanopySnowEvap     CANEVP_MAXIMUM          CANOPY_SNOW     ATMOSPHERE
    :SnowBalance        SNOBAL_SIMPLE_MELT      SNOW            SNOW_LIQ
        :-->Overflow    RAVEN_DEFAULT           SNOW_LIQ        PONDED_WATER
    :SnowRefreeze       FREEZE_DEGREE_DAY       SNOW_LIQ        SNOW
    :Abstraction        ABST_FILL               PONDED_WATER    DEPRESSION
    :OpenWaterEvaporation OPEN_WATER_EVAP       DEPRESSION      ATMOSPHERE
    :Infiltration       INF_HBV                 PONDED_WATER    MULTIPLE
    :Baseflow           BASE_POWER_LAW          SOIL1           SURFACE_WATER
    :Baseflow           BASE_POWER_LAW          SOIL2           SURFACE_WATER
    :Interflow          INTERFLOW_PRMS          SOIL0           SURFACE_WATER
    :Percolation        PERC_GAWSER             SOIL0           SOIL1
    :Percolation        PERC_GAWSER             SOIL1           SOIL2
    :SoilEvaporation    SOILEVAP_ROOT           SOIL0           ATMOSPHERE
    :LakeEvaporation LAKE_EVAP_BASIC LAKE_STORAGE ATMOSPHERE
    :LakeRelease LAKEREL_LINEAR   LAKE_STORAGE SURFACE_WATER

:EndHydrologicProcesses
```

Appendix B – Downstream hydrographs of Kaministiquia Watershed

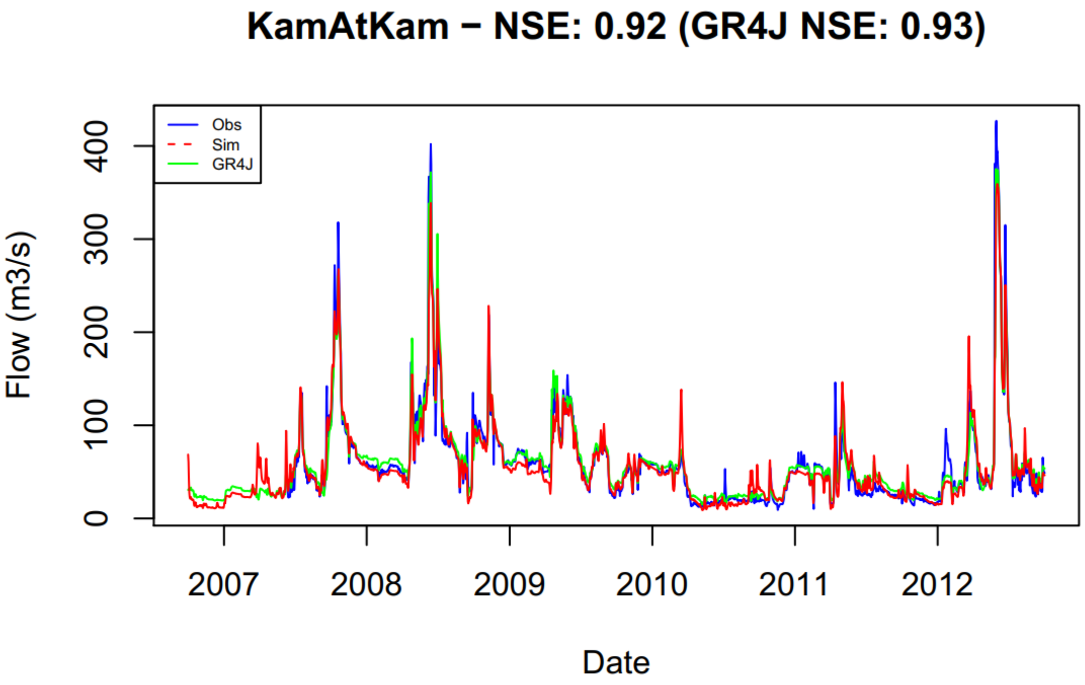

KamAtKam − NSE: 0.92 (GR4J NSE: 0.93)

Figure A.1 – Kaministiquia River calibration period flow results

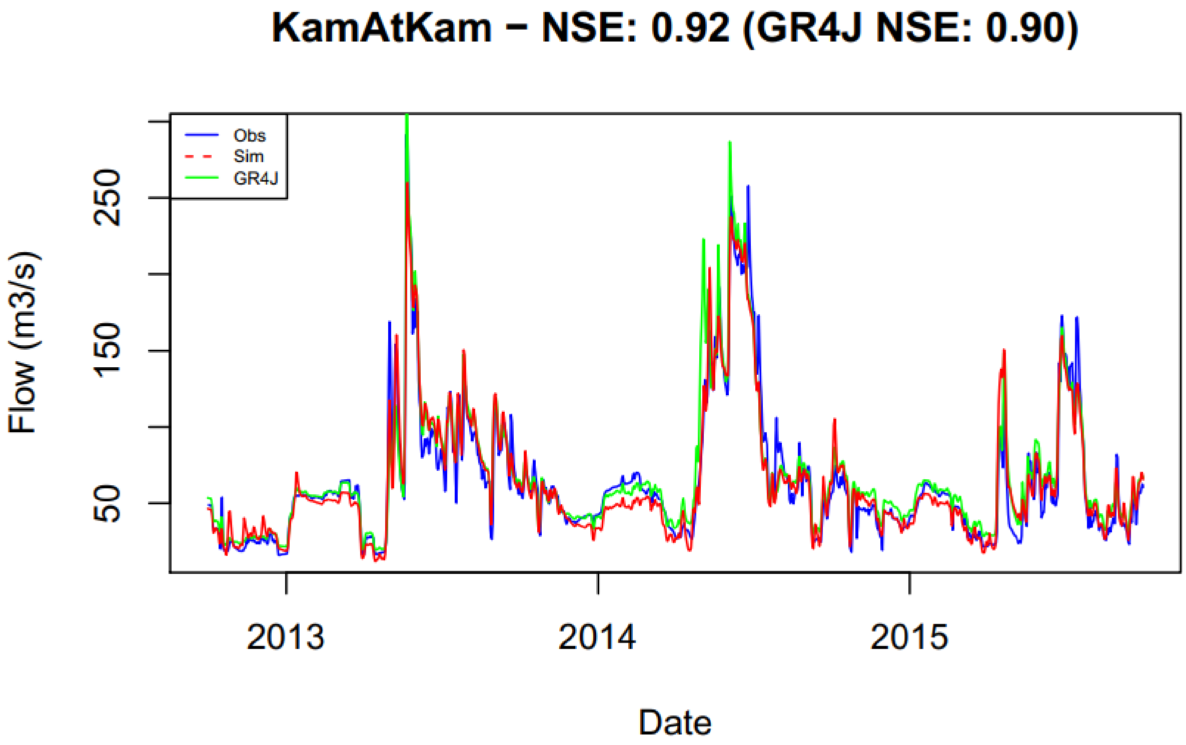
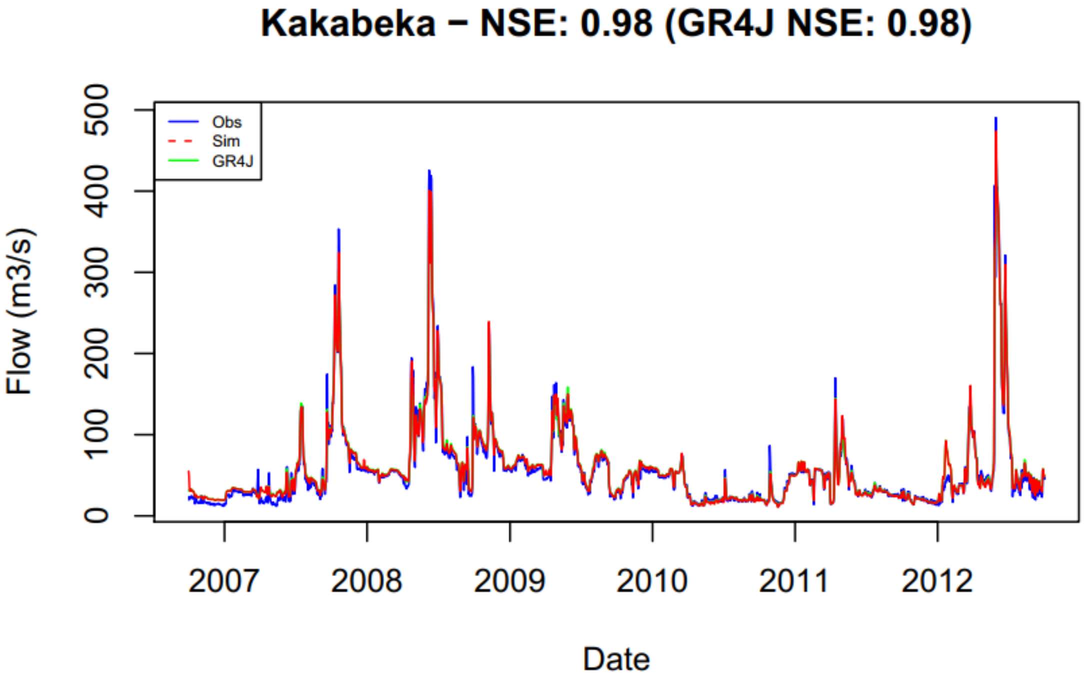

KamAtKam − NSE: 0.92 (GR4J NSE: 0.90)

Figure A.2– Kaministiquia River validation period flow results



A.3– Kakabeka Falls River calibration period flow results

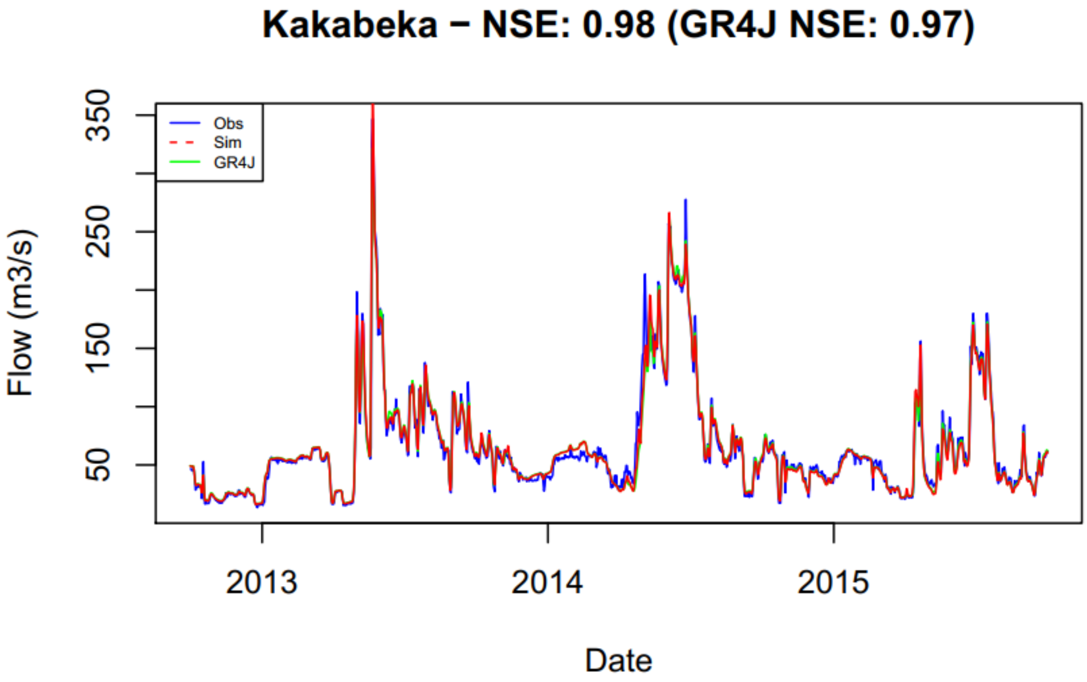


Figure A.4– Kakabeka Falls River validation period flow results

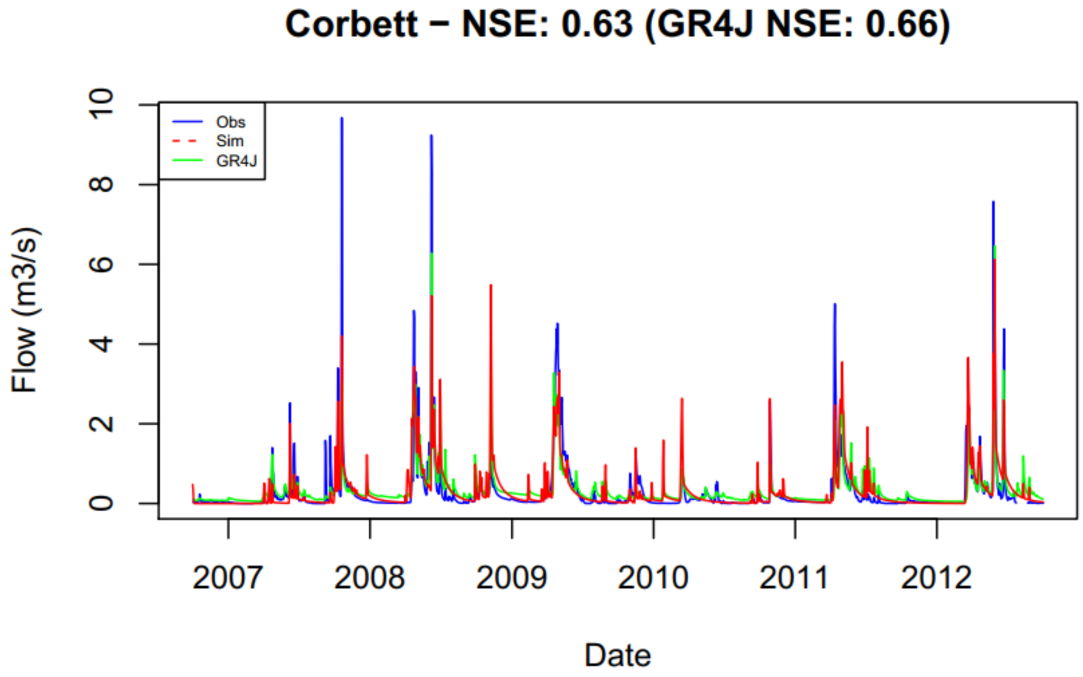


Figure A.5– Corbett River calibration period flow results

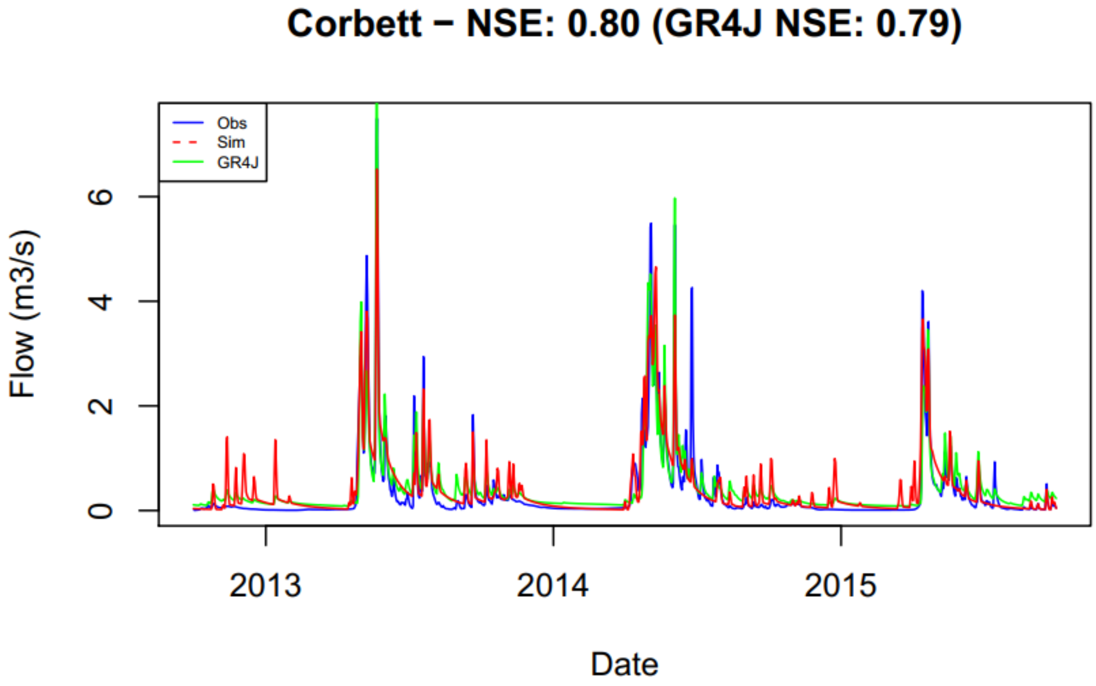


Figure A.6– Corbett River validation period flow results

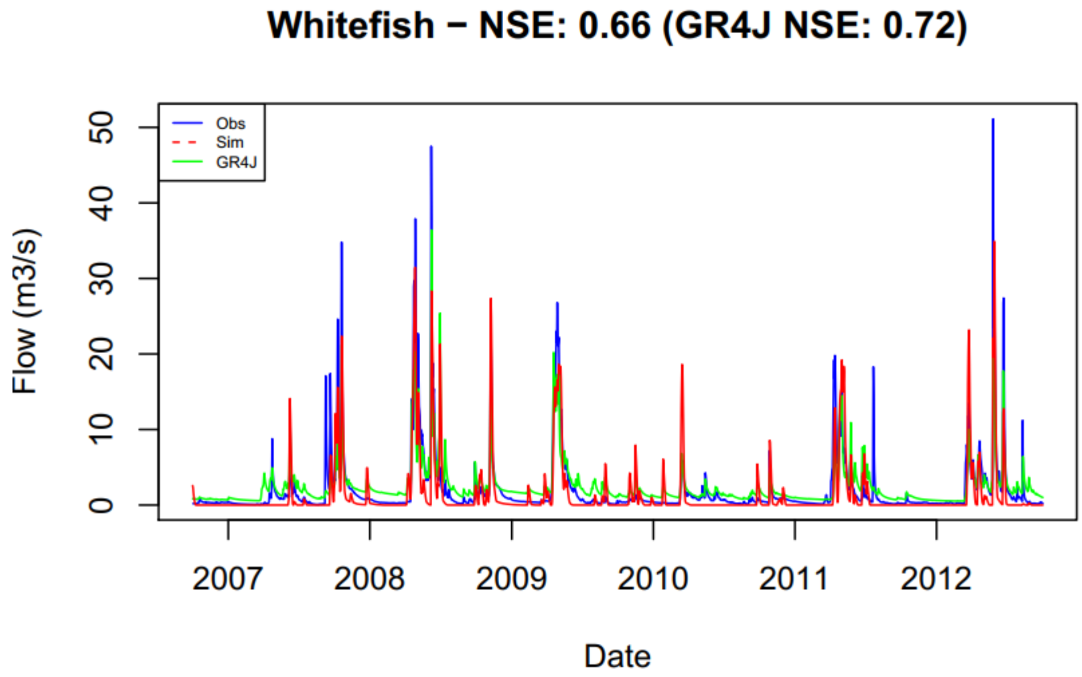


Figure A.7– Whitefish River calibration period flow results

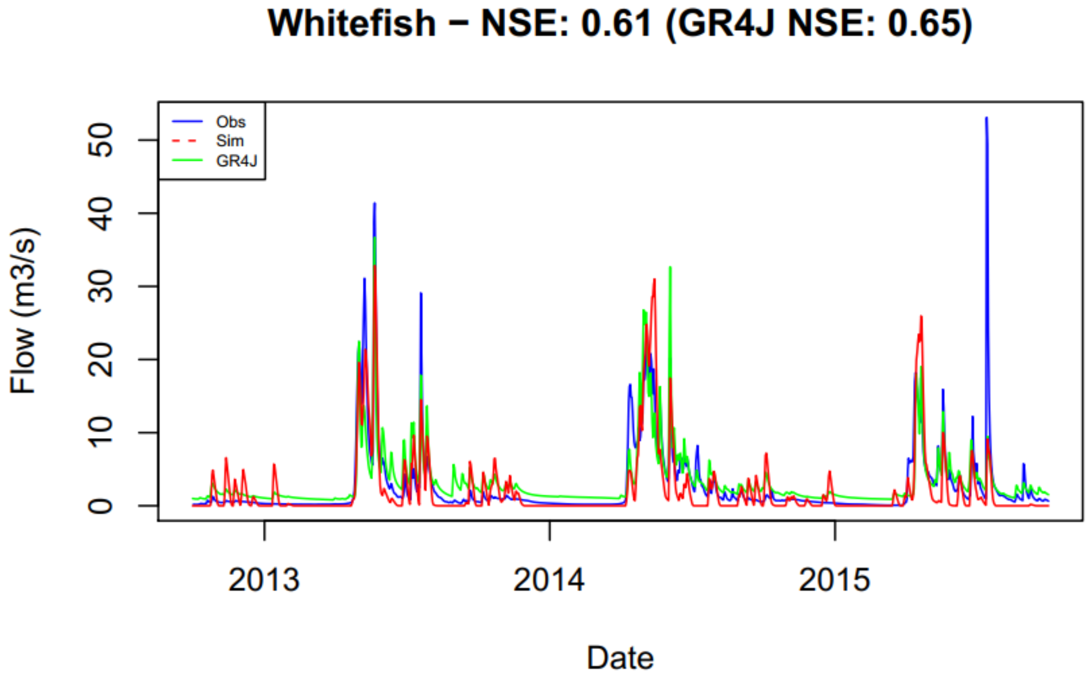


Figure A.8 – Whitefish River validation period flow results

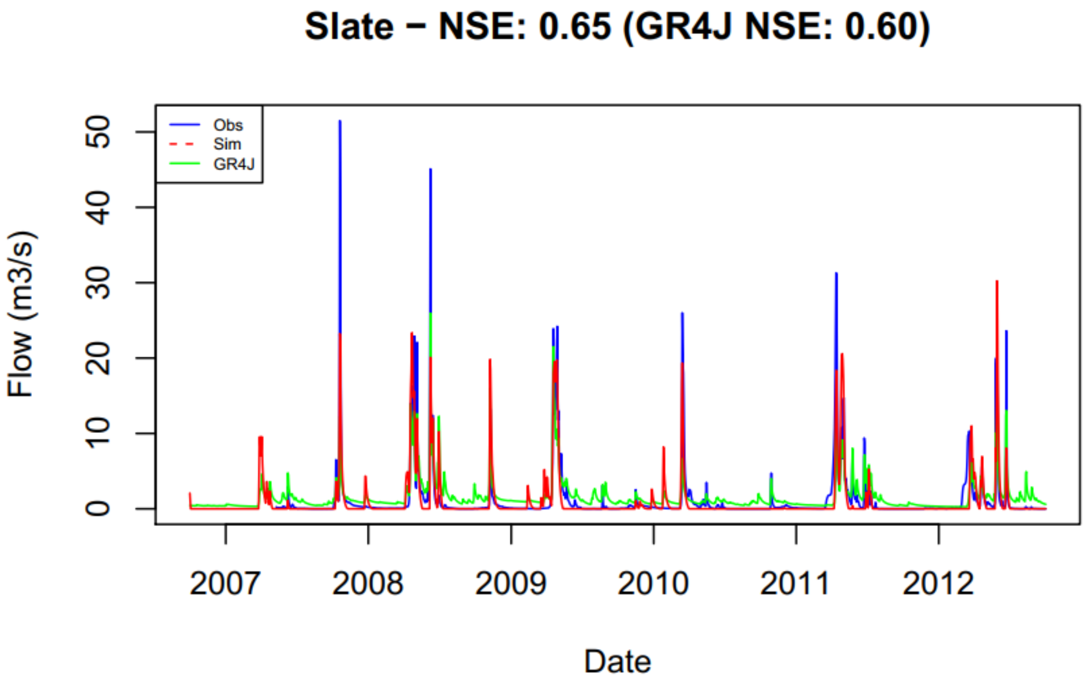


Figure A.9 – Slate River calibration period flow results

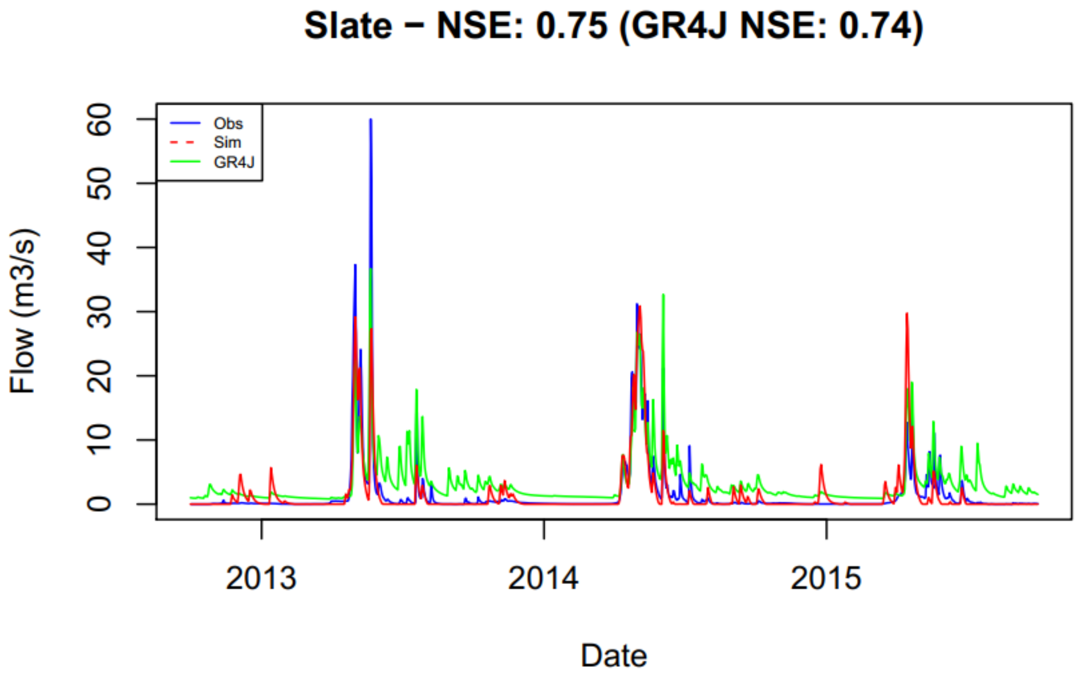


Figure A.10– Slate River validation period flow results

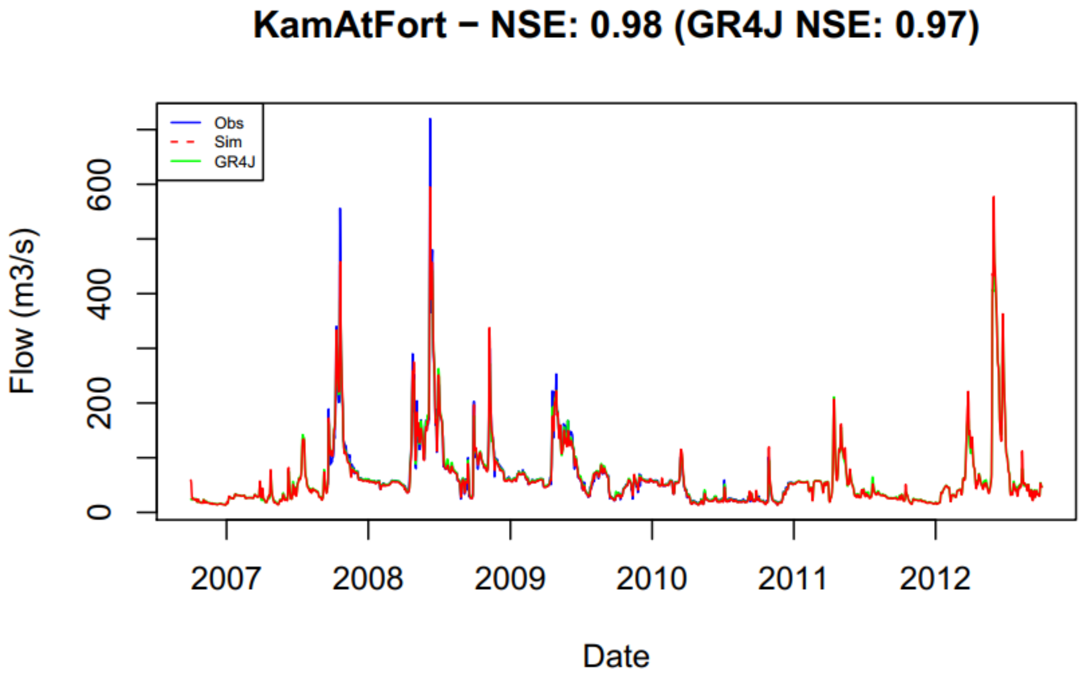


Figure A.11– Kaministiquia River at Fort Williams calibration period flow results

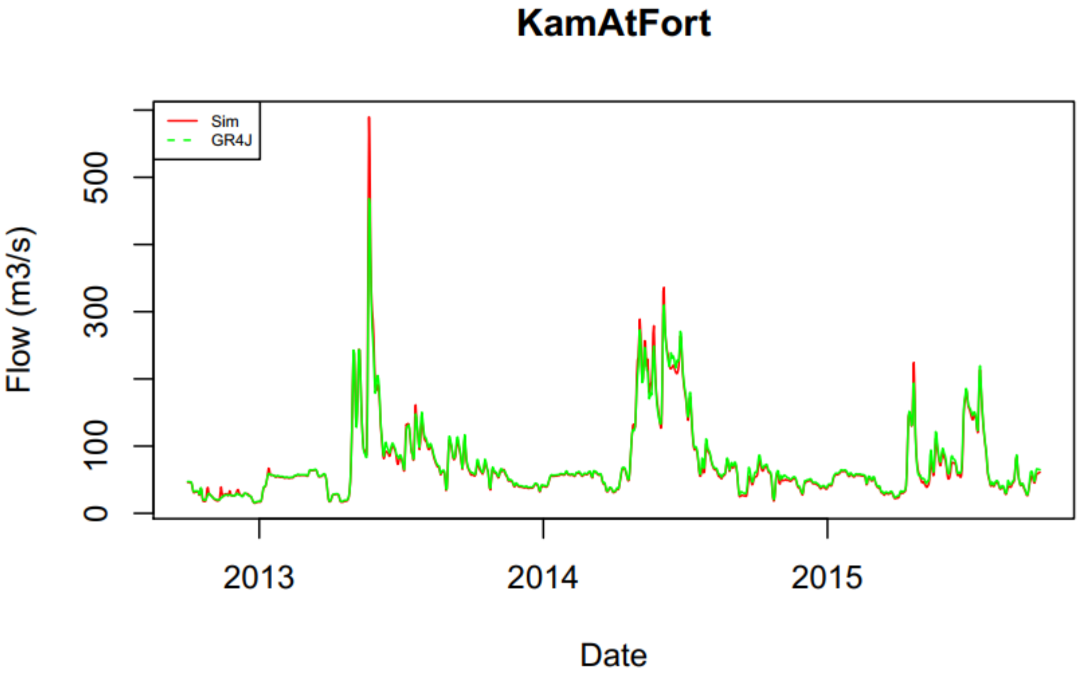
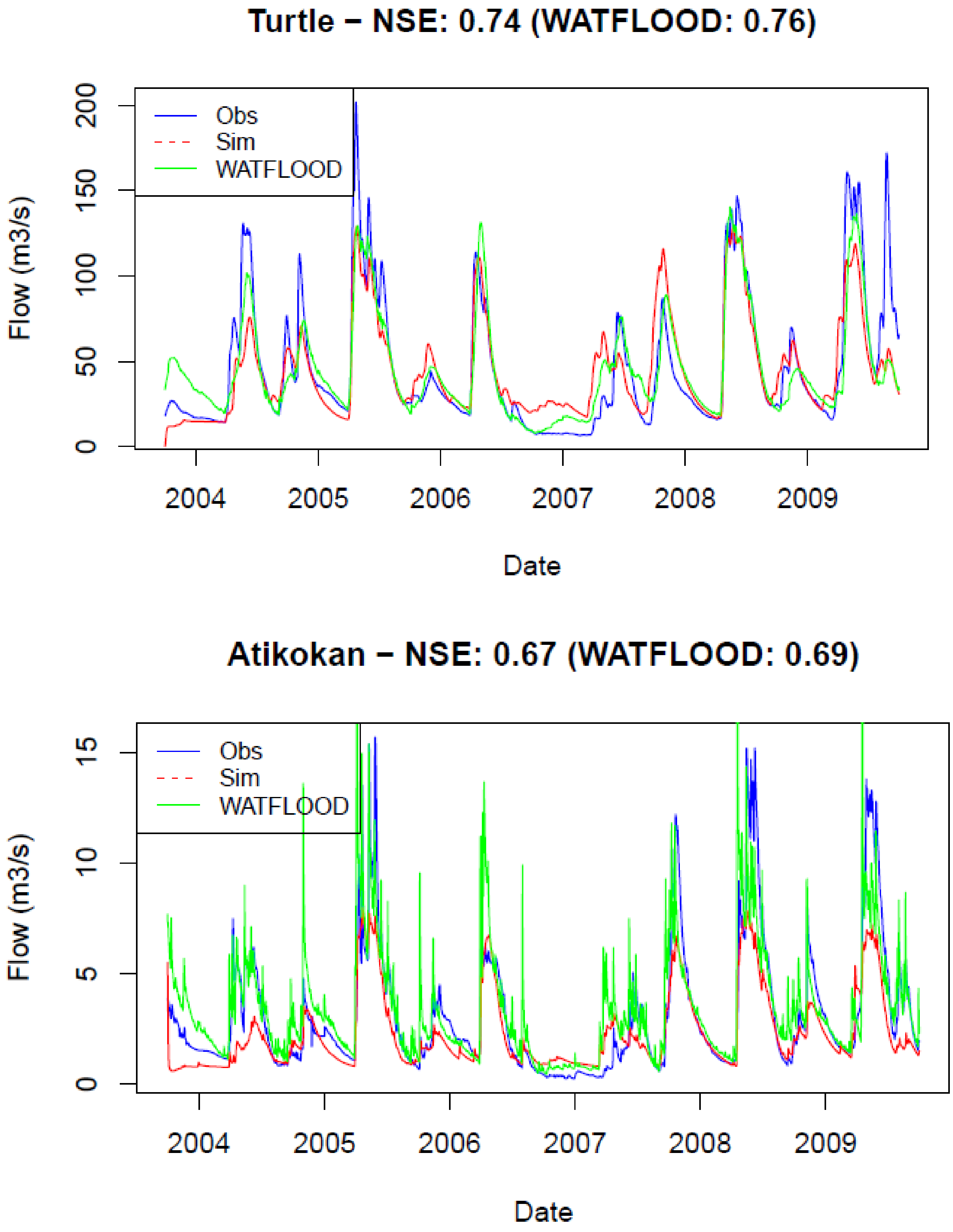


Figure A.12 – Kaministiquia River at Fort Williams validation period flow results

Appendix C – Hydrographs of Lake of the Woods Watershed

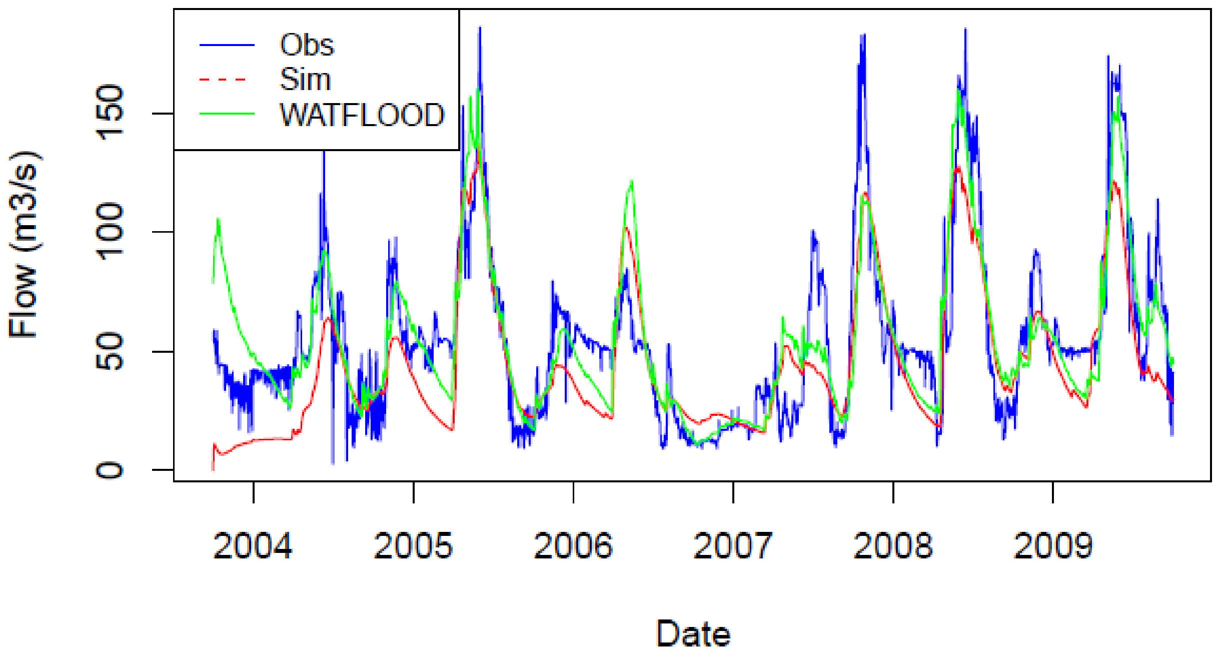
Figure B.1 – Hydrograph of stream flows for calibration period when calibrated to stream flow



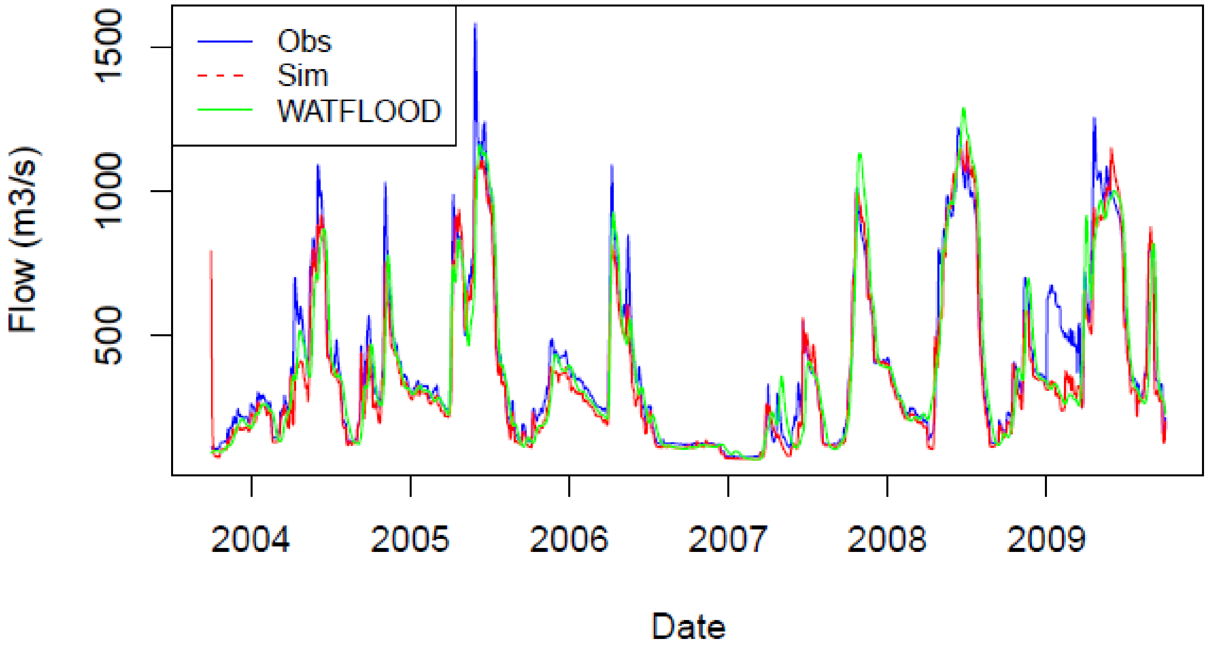
Turtle – NSE: 0.74 (WATFLOOD: 0.76)

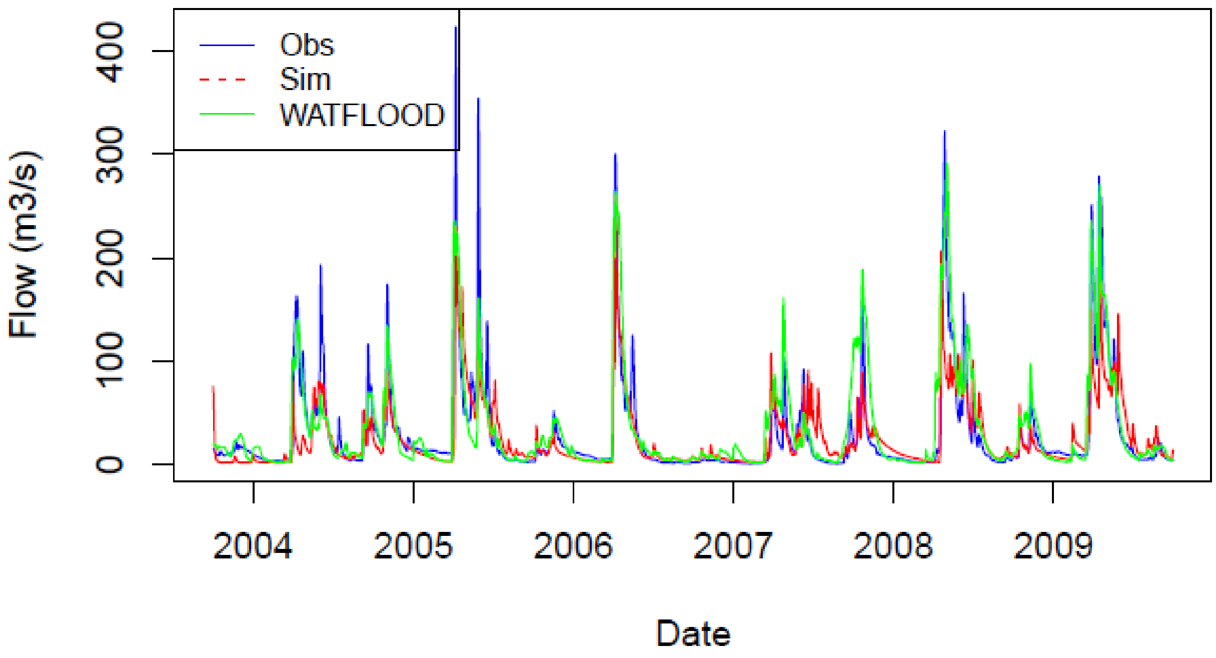
Atikokan – NSE: 0.67 (WATFLOOD: 0.69)

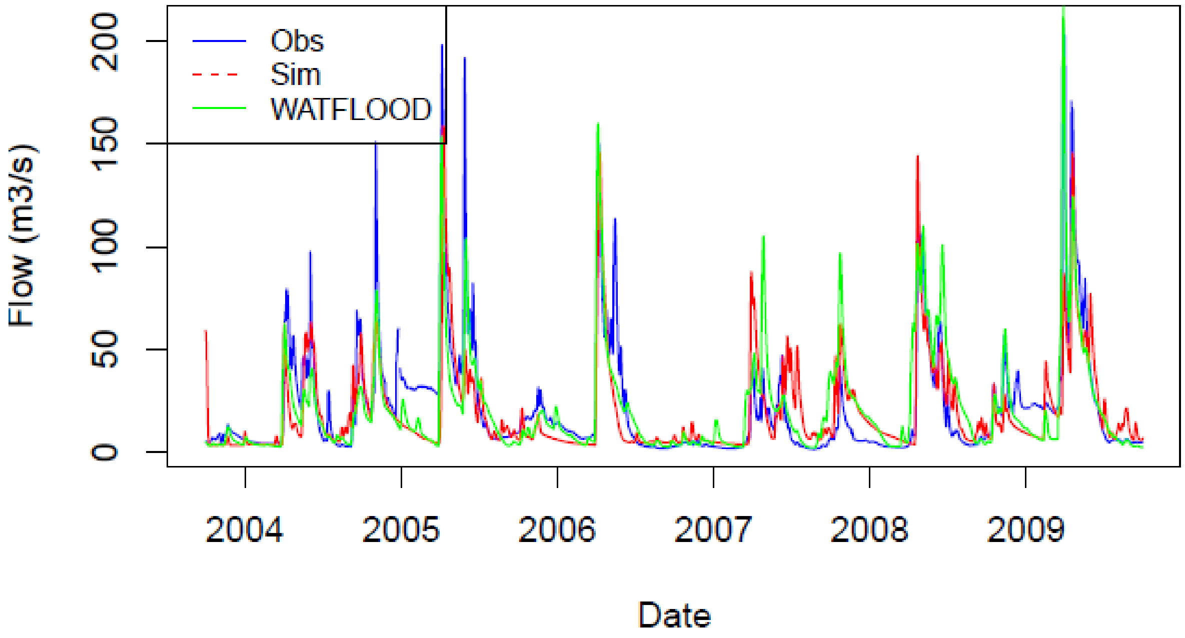
Seine_Sturgeon – NSE: 0.64 (WATFLOOD: 0.72)



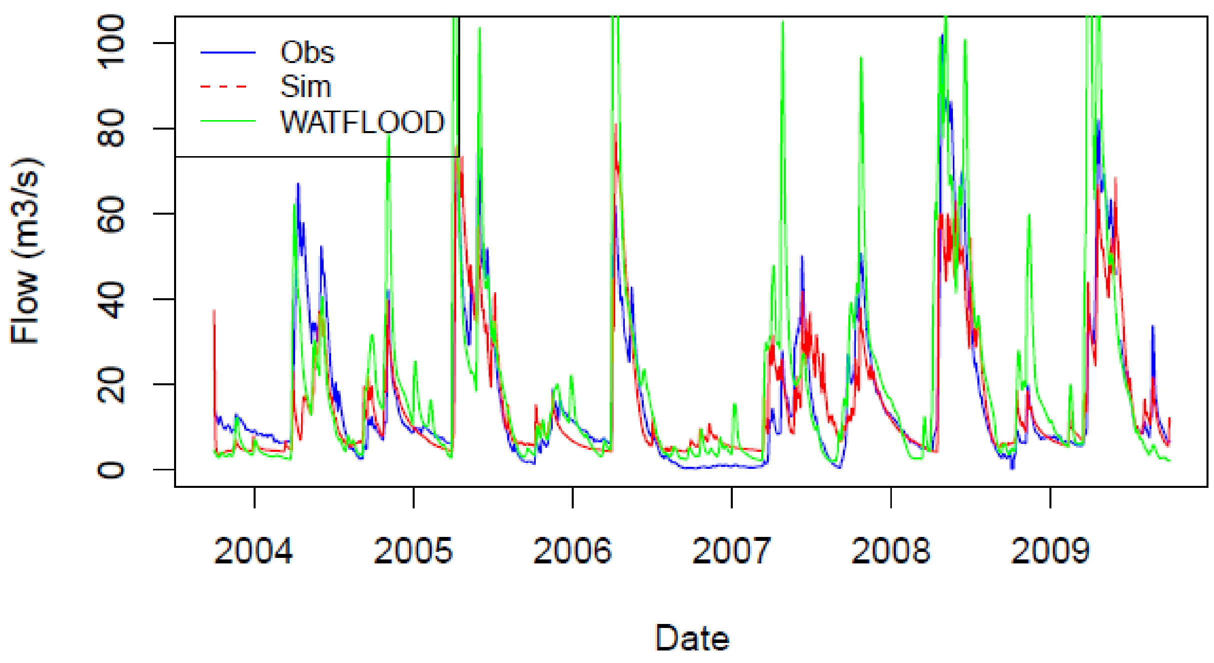
Rainy_Manitou – NSE: 0.92



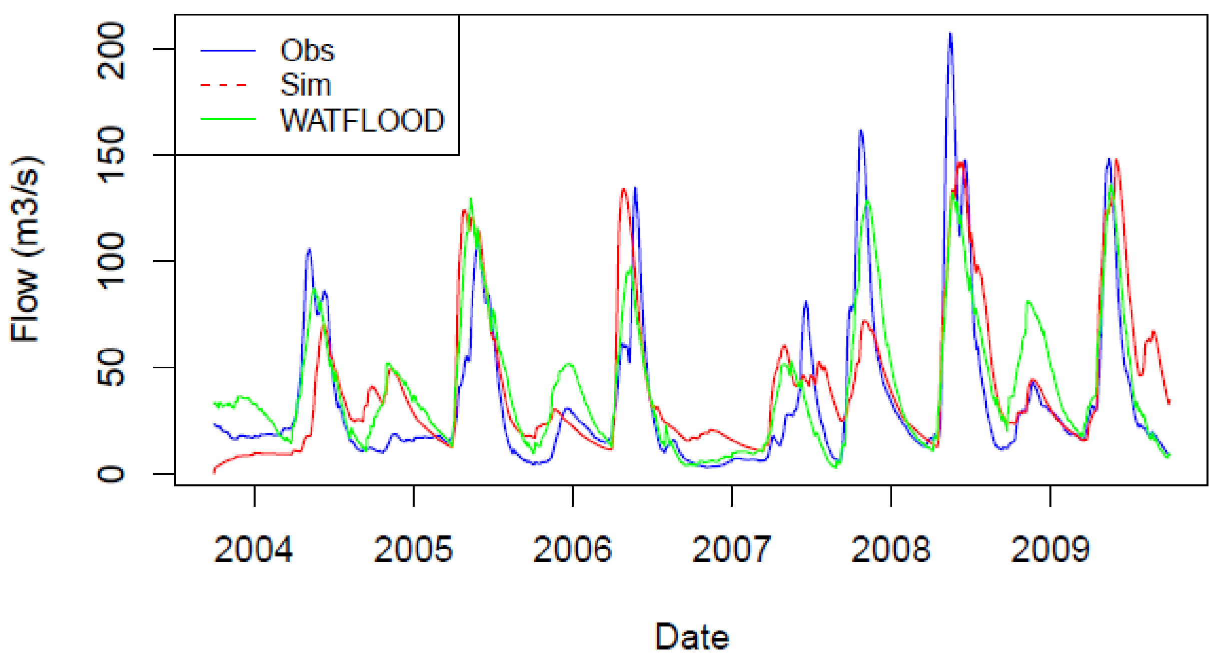

Little_Fork − NSE: 0.65 (WATFLOOD: 0.67)


Big_Fork − NSE: 0.65 (WATFLOOD: 0.68)

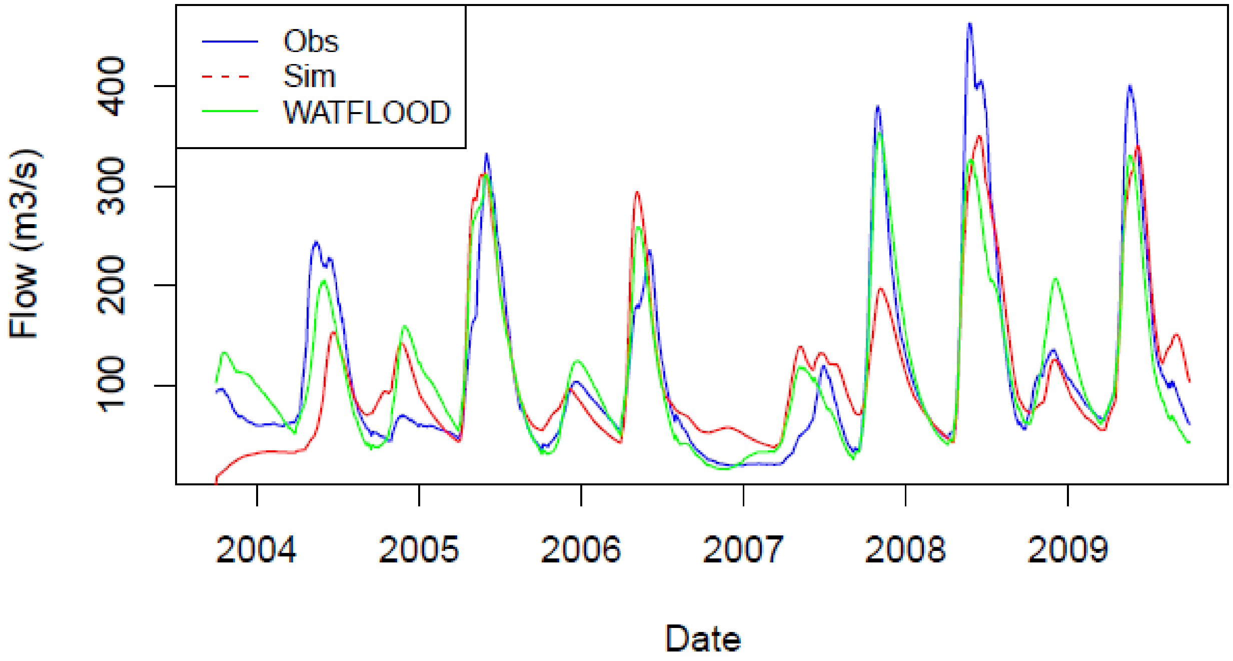
Vermillion – NSE: 0.78 (WATFLOOD: 0.54)



Basswood – NSE: 0.52 (WATFLOOD: 0.69)



LLC – NSE: 0.78 (WATFLOOD: 0.84)



Kawishi – NSE: 0.58 (WATFLOOD: 0.67)

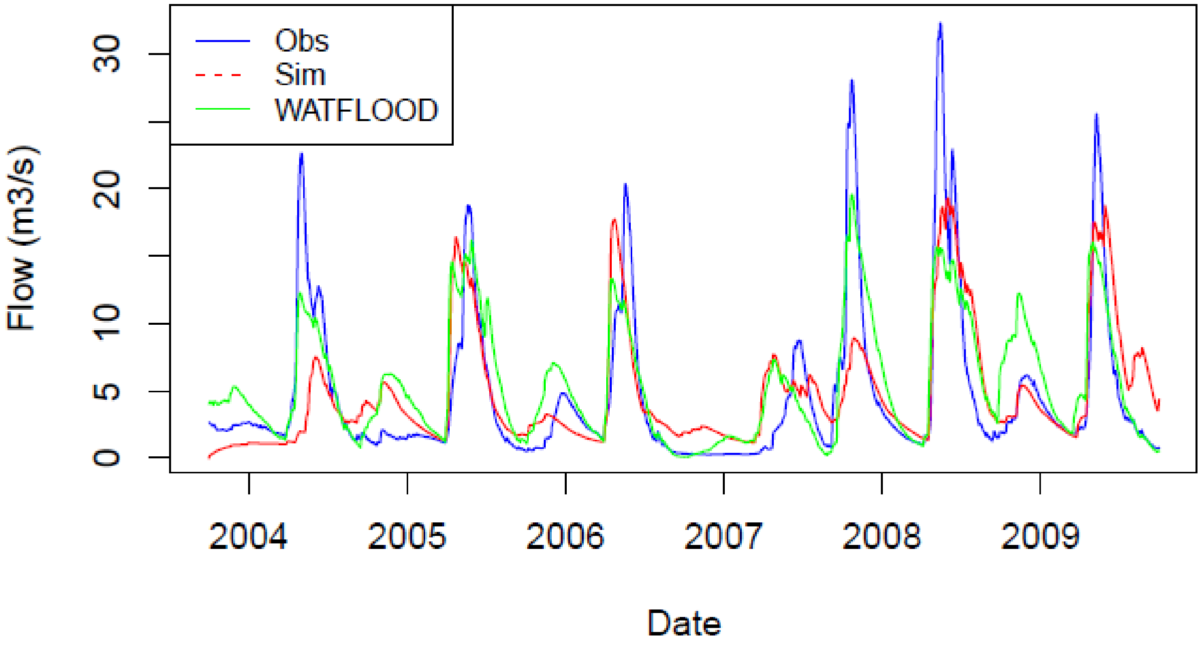
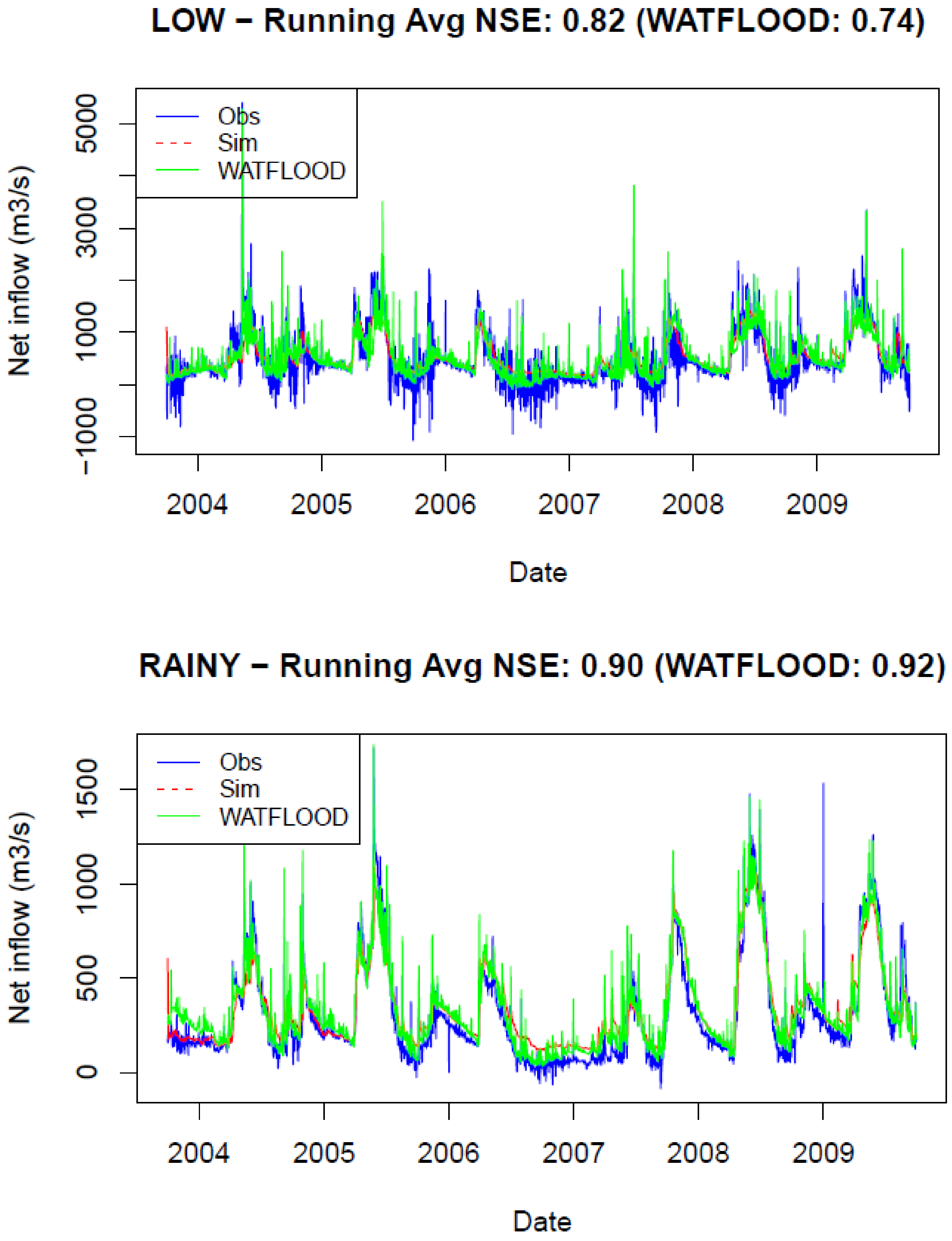
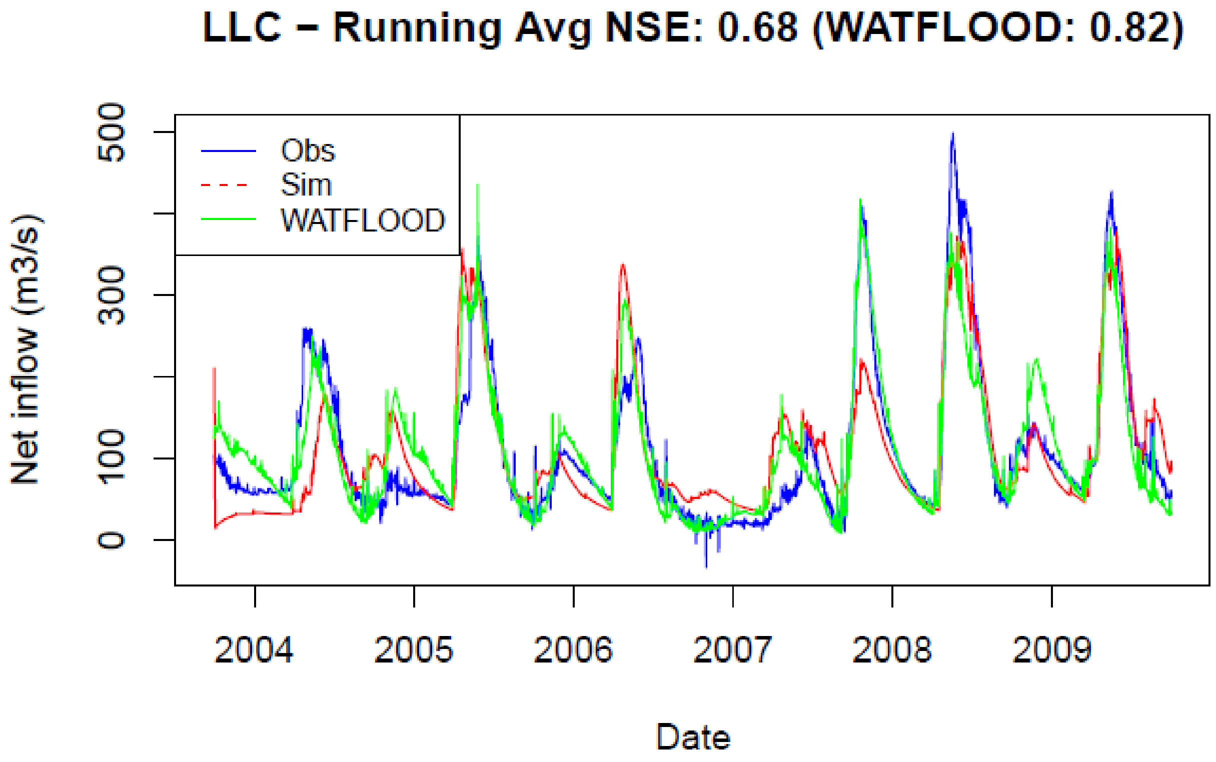


Figure B.2 – Inflow hydrograph for calibration period when calibrated to stream flow



LOW – Running Avg NSE: 0.82 (WATFLOOD: 0.74)

RAINY – Running Avg NSE: 0.90 (WATFLOOD: 0.92)


LLC – Running Avg NSE: 0.68 (WATFLOOD: 0.82)

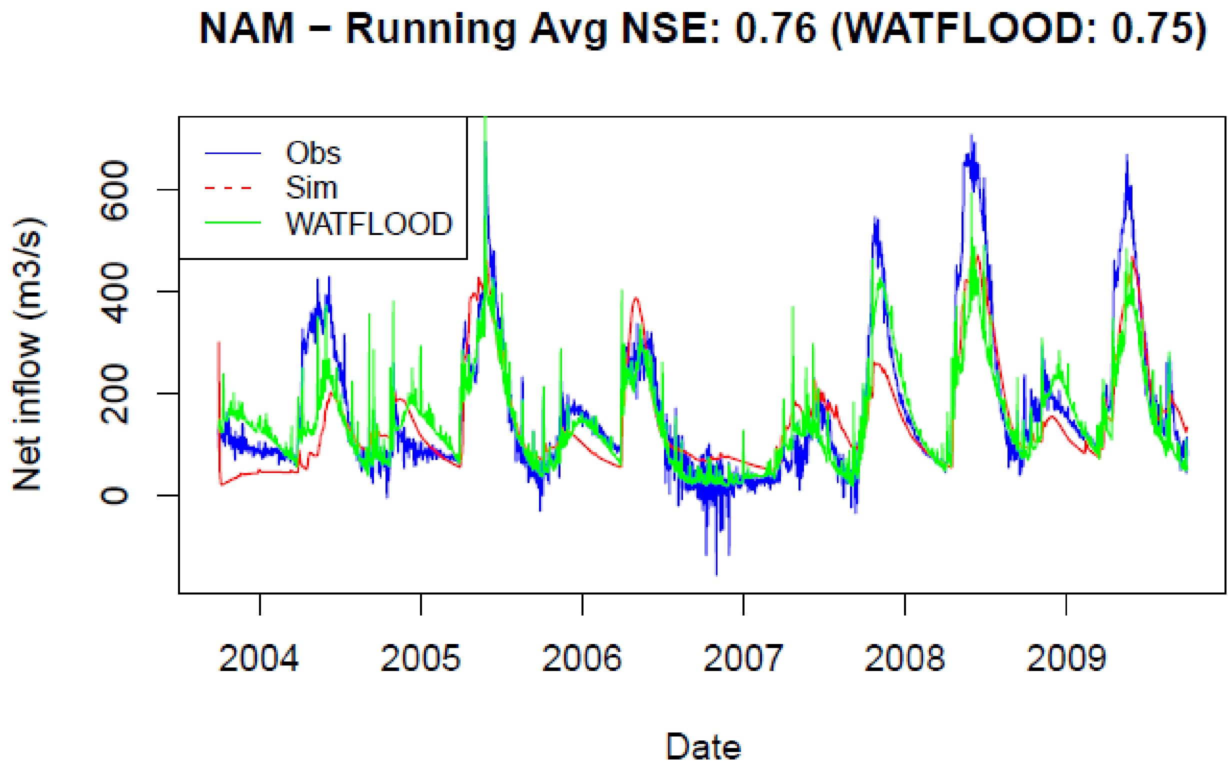
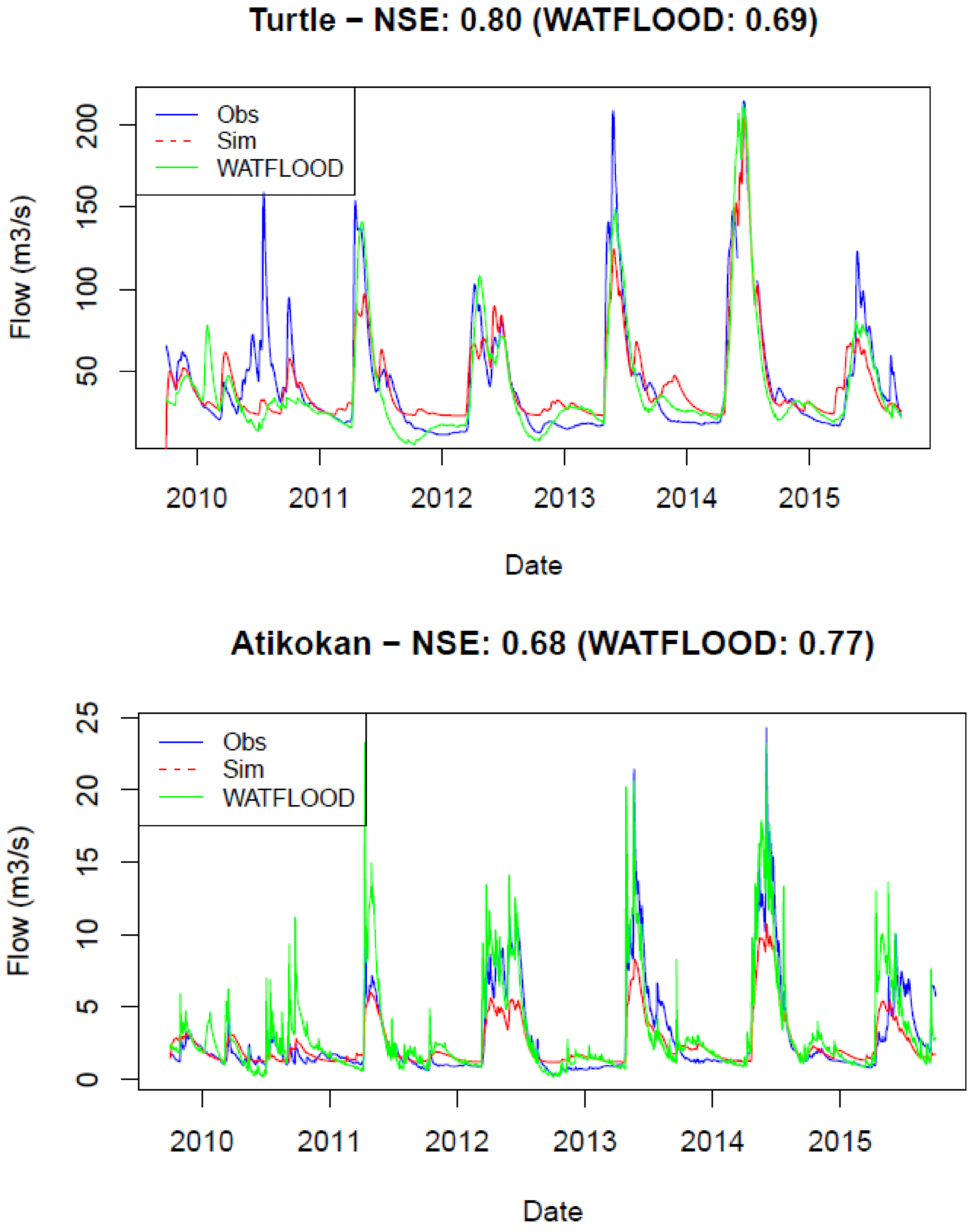
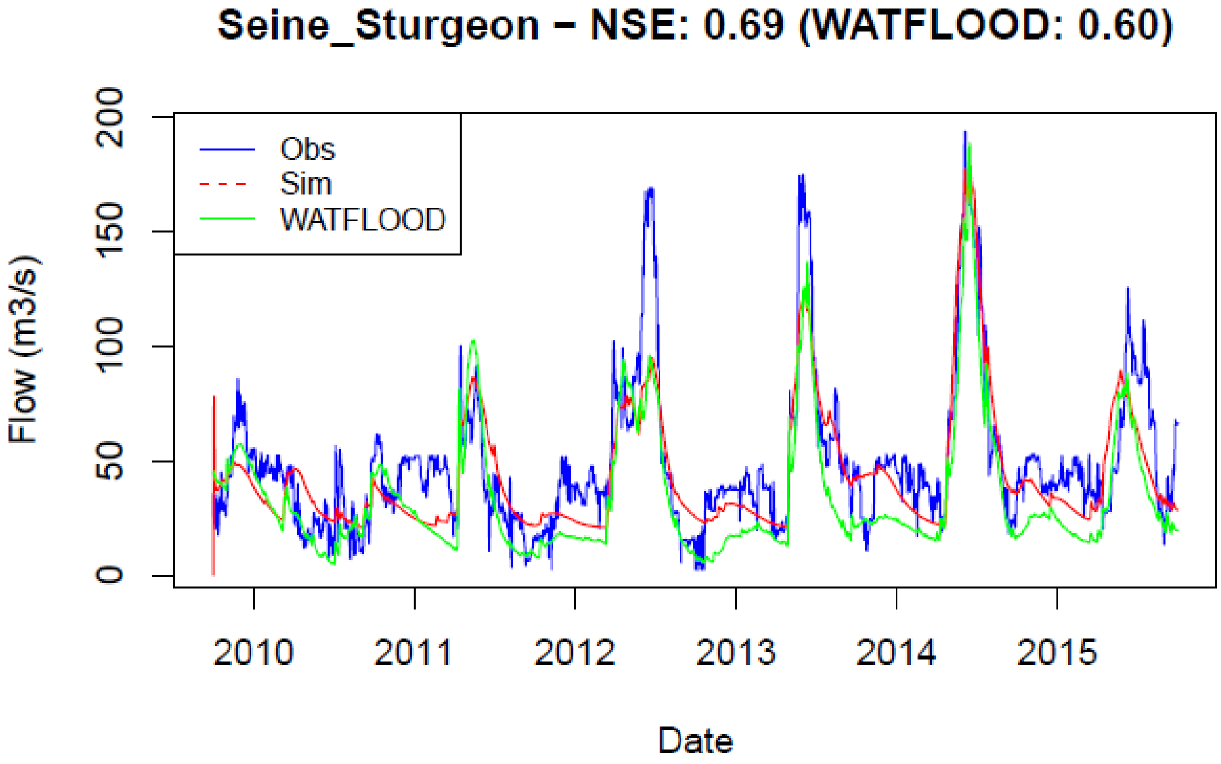

NAM – Running Avg NSE: 0.76 (WATFLOOD: 0.75)

Figure B.3 – Hydrograph of stream flows for validation period when calibrated to stream flow

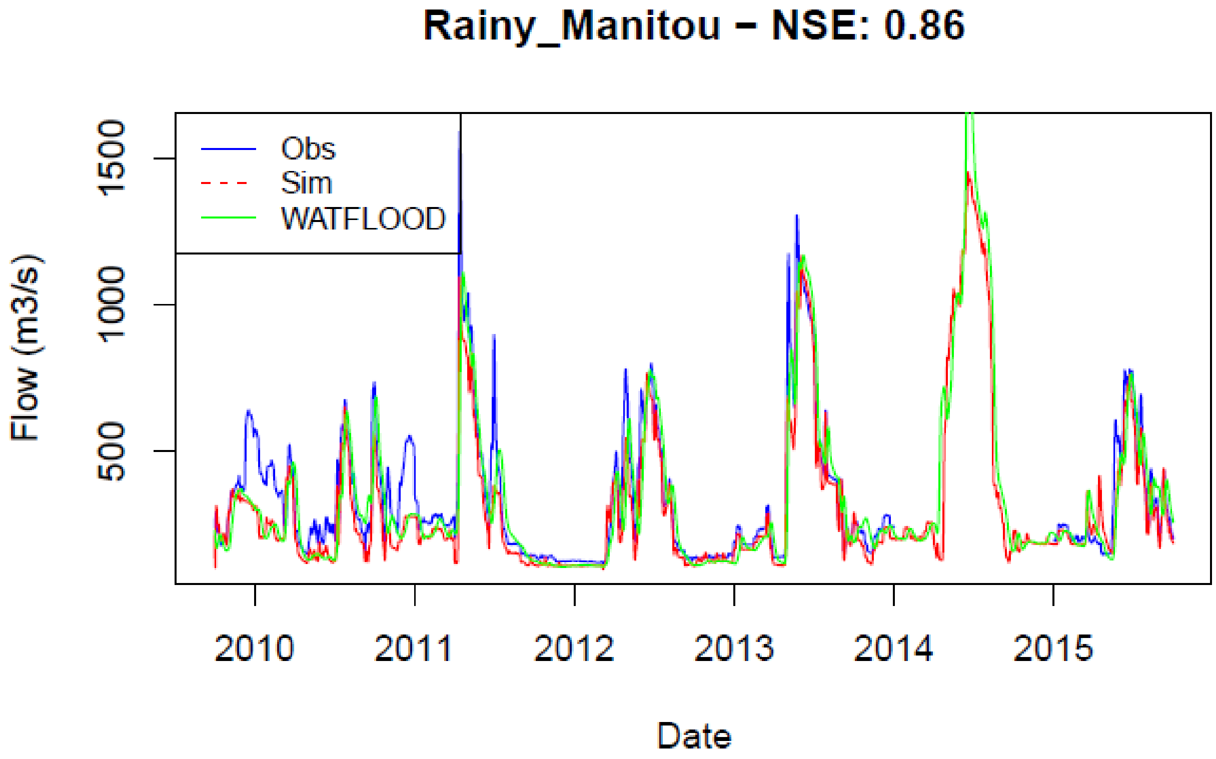


Turtle – NSE: 0.80 (WATFLOOD: 0.69)

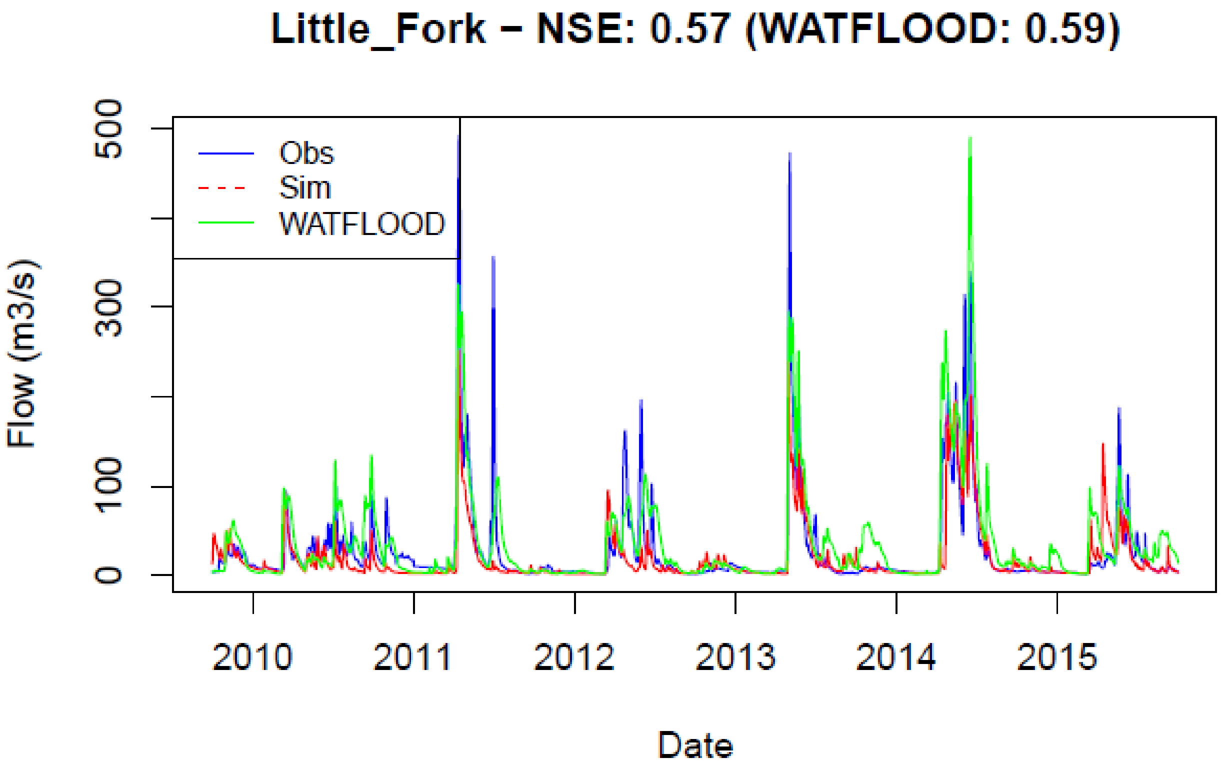
Atikokan – NSE: 0.68 (WATFLOOD: 0.77)

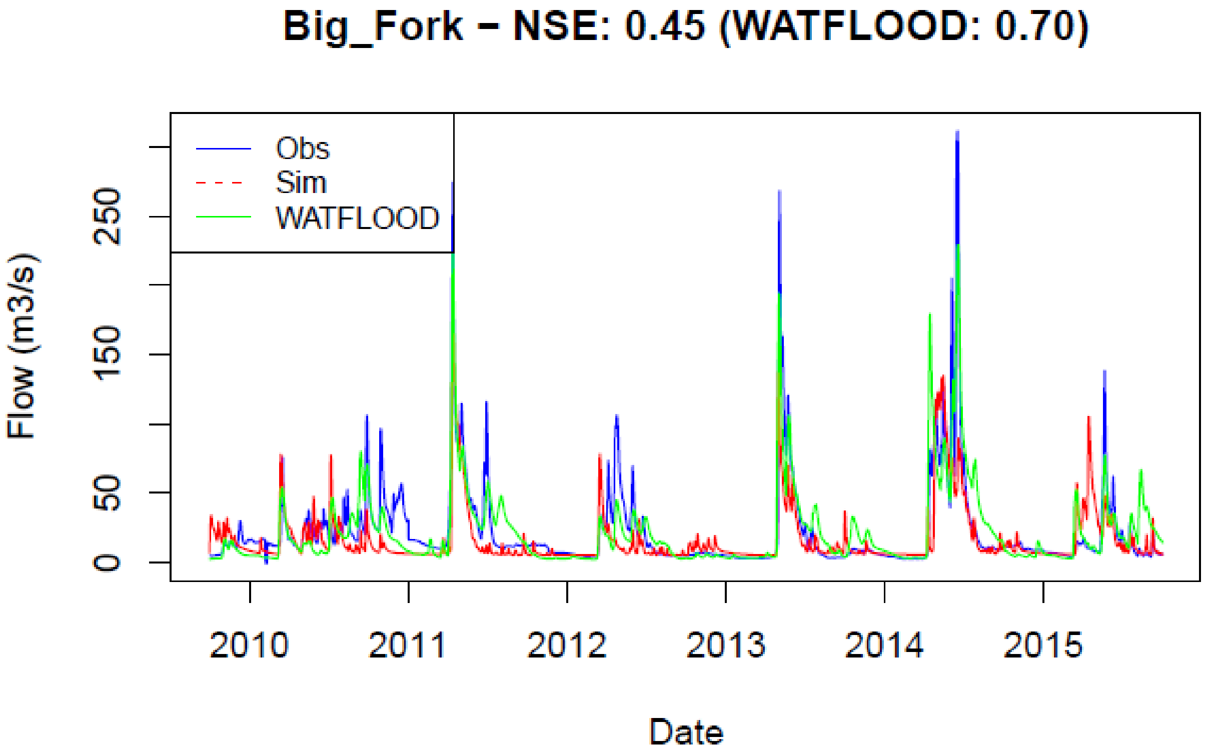


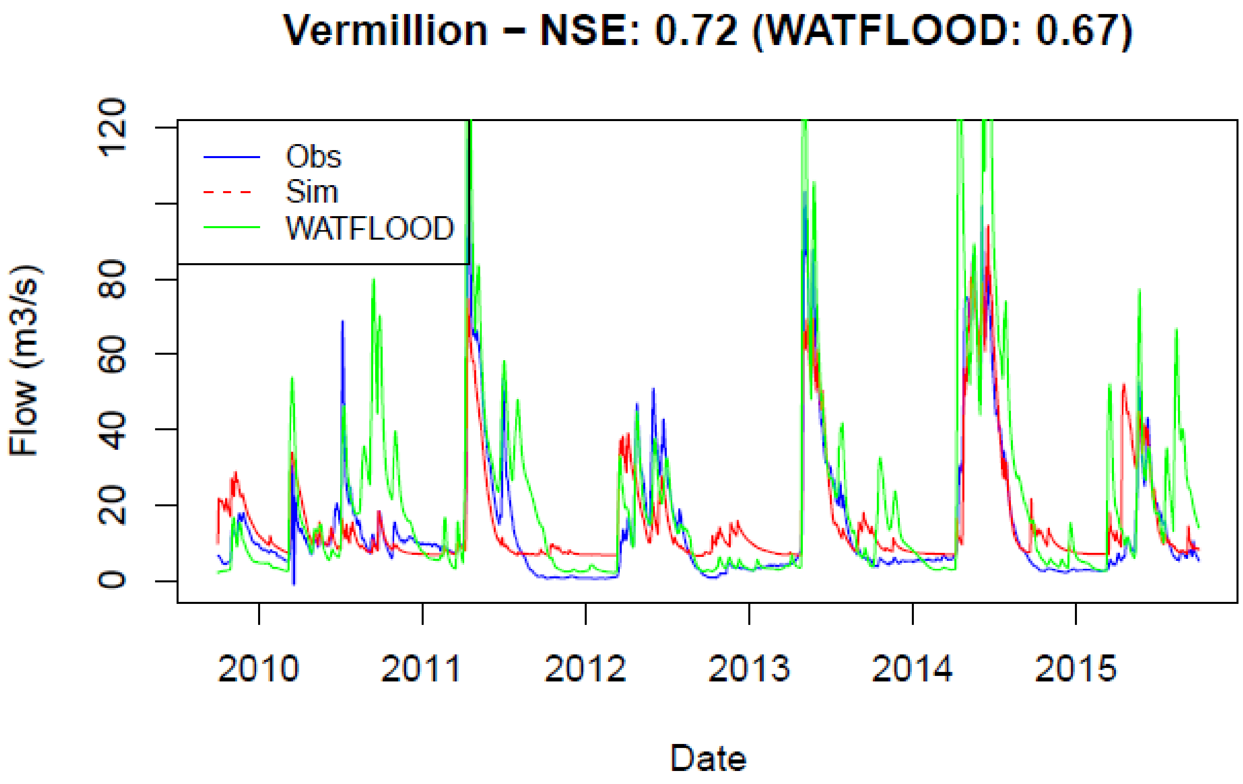
Seine_Sturgeon – NSE: 0.69 (WATFLOOD: 0.60)

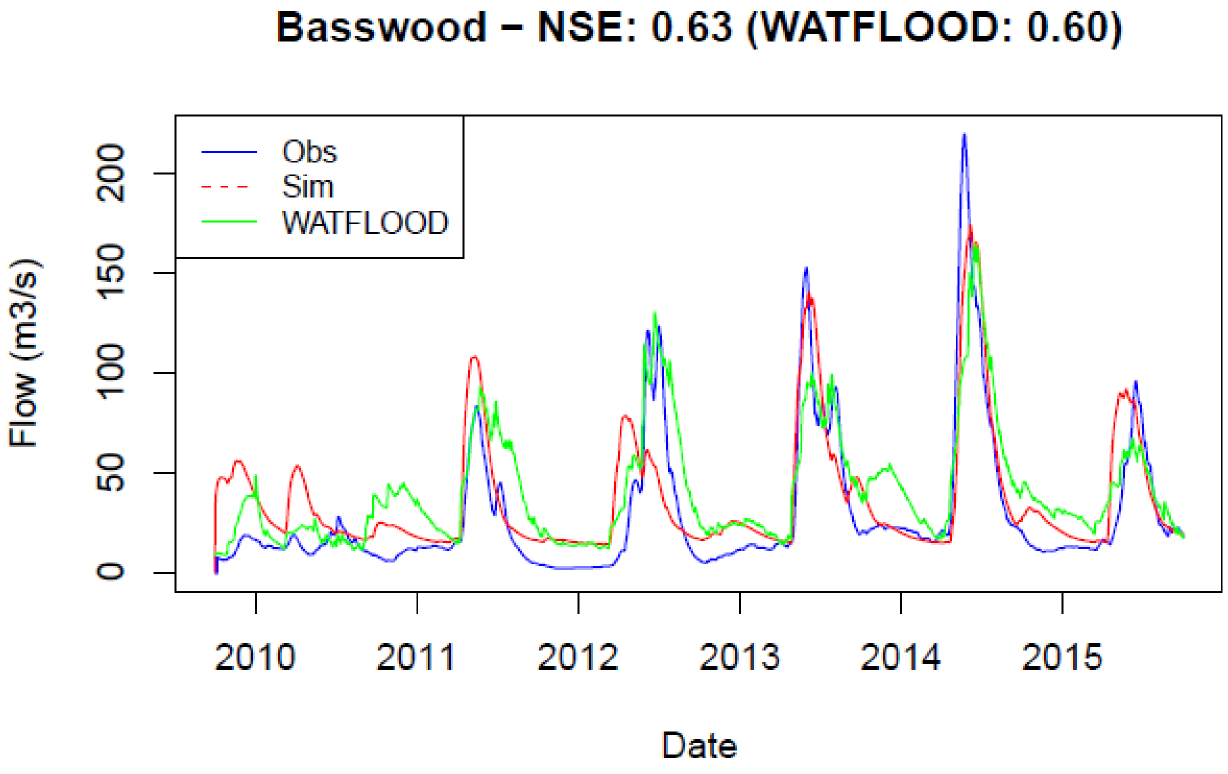


Rainy_Manitou – NSE: 0.86

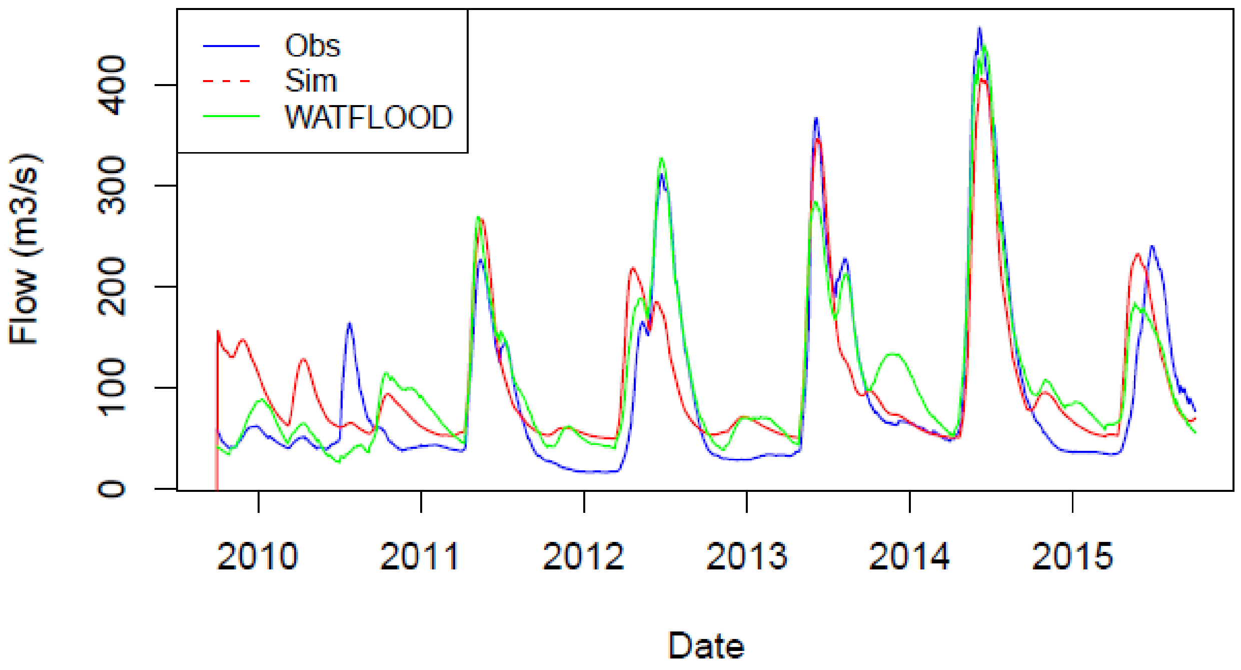

Little_Fork – NSE: 0.57 (WATFLOOD: 0.59)


Big_Fork – NSE: 0.45 (WATFLOOD: 0.70)


Vermillion – NSE: 0.72 (WATFLOOD: 0.67)
Obs
Sim
WATFLOOD
Flow (m3/s)
0
20
40
60
80
100
120
2010
2011
2012
2013
2014
2015
Date


Basswood – NSE: 0.63 (WATFLOOD: 0.60)
Obs
Sim
WATFLOOD
Flow (m3/s)
0
50
100
150
200
2010
2011
2012
2013
2014
2015
Date

LLC – NSE: 0.76 (WATFLOOD: 0.85)



Kawishi – NSE: 0.62 (WATFLOOD: 0.69)

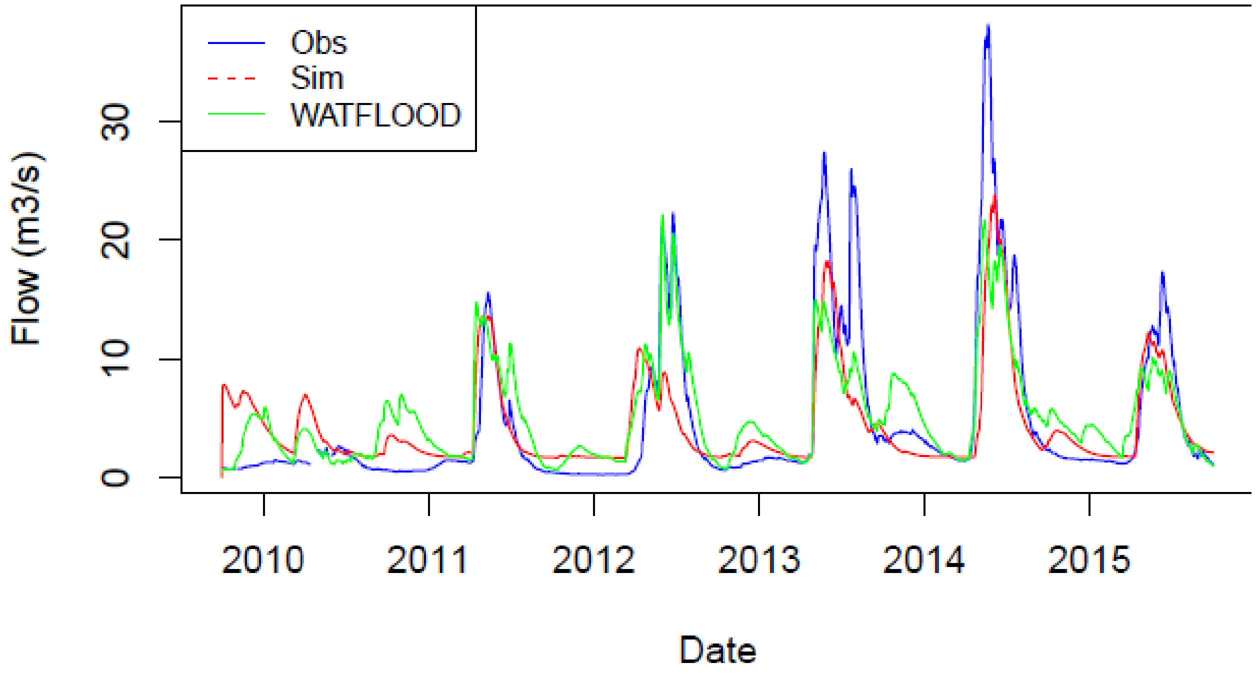
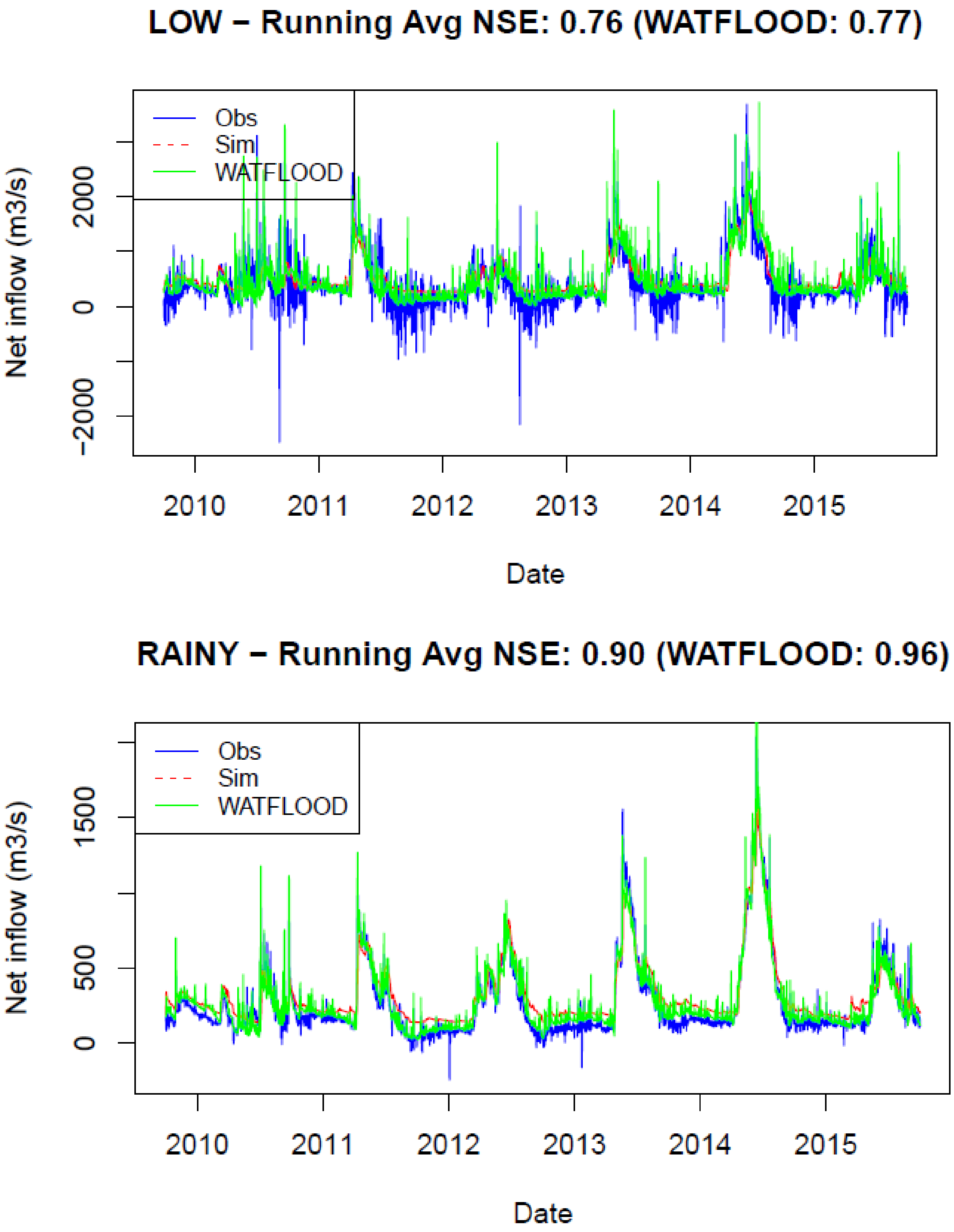
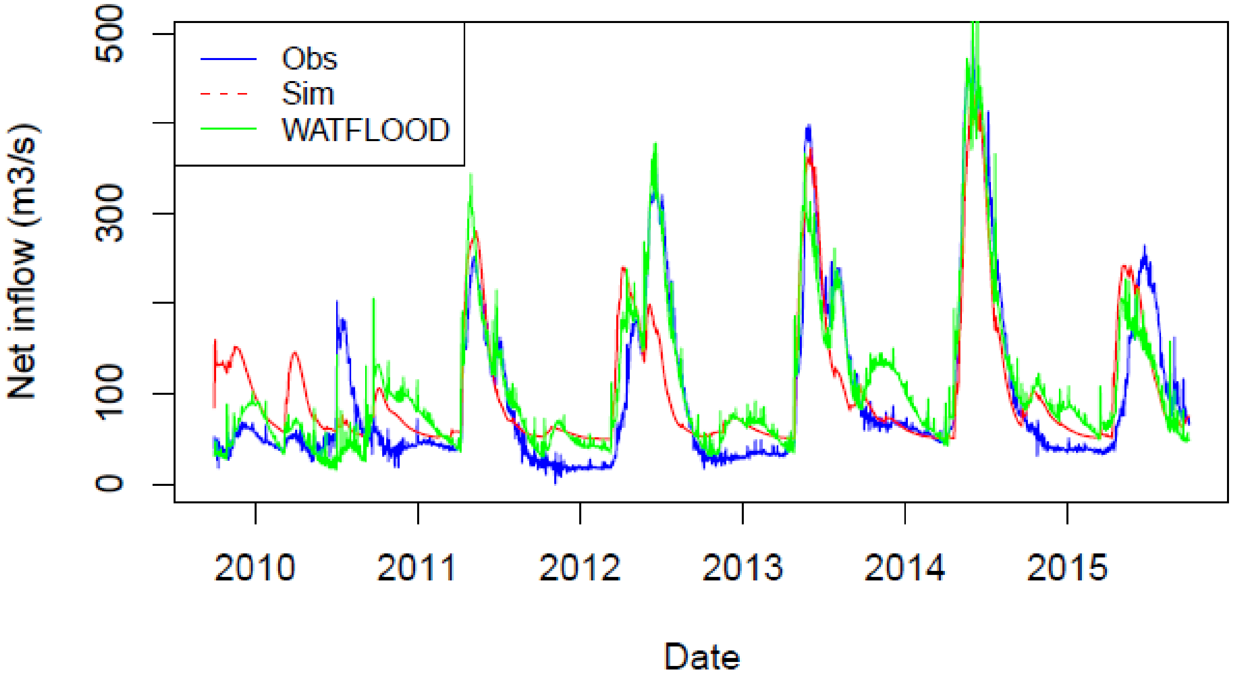


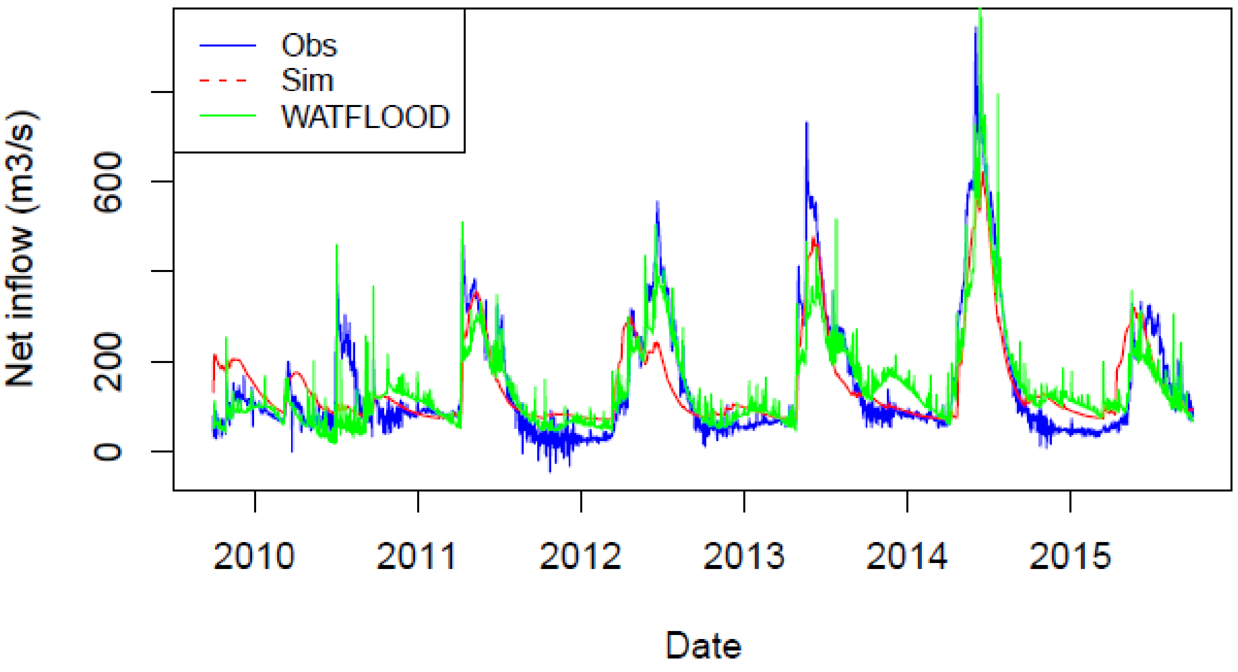
Figure B.4 – Inflow hydrograph for validation period when calibrated to stream flow


LOW – Running Avg NSE: 0.76 (WATFLOOD: 0.77)

RAINY – Running Avg NSE: 0.90 (WATFLOOD: 0.96)

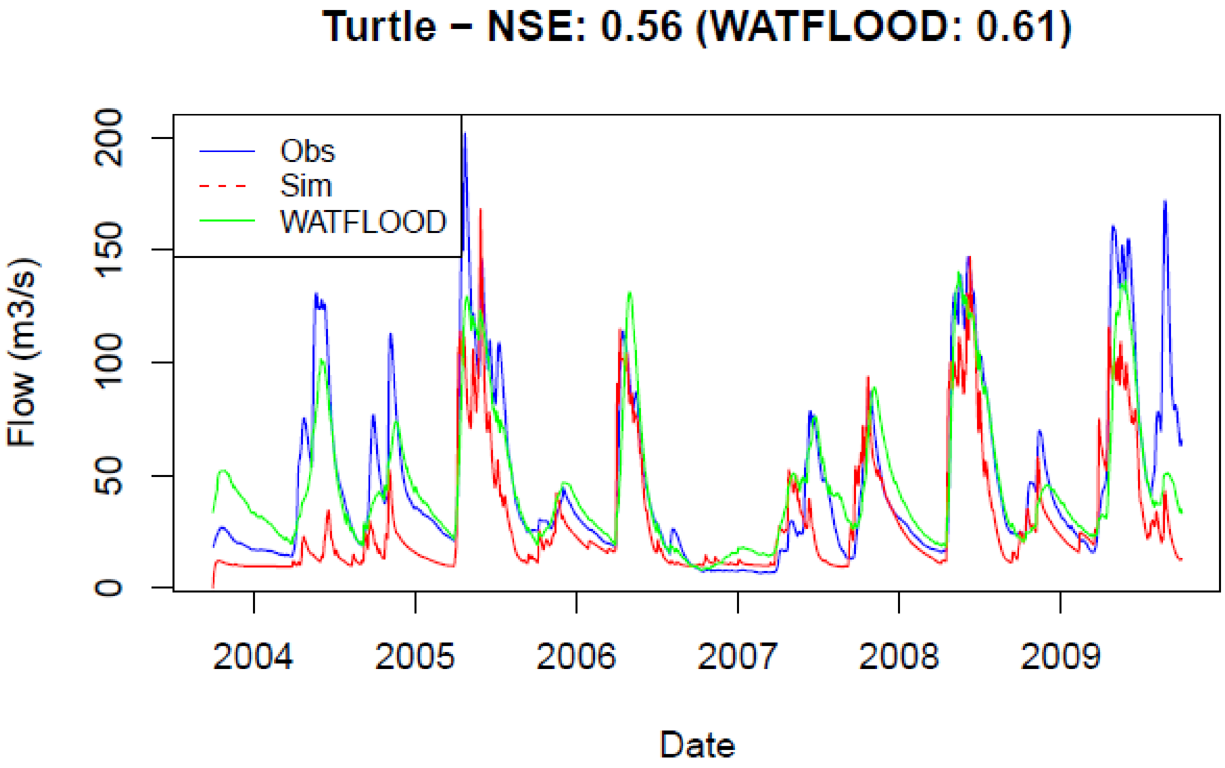


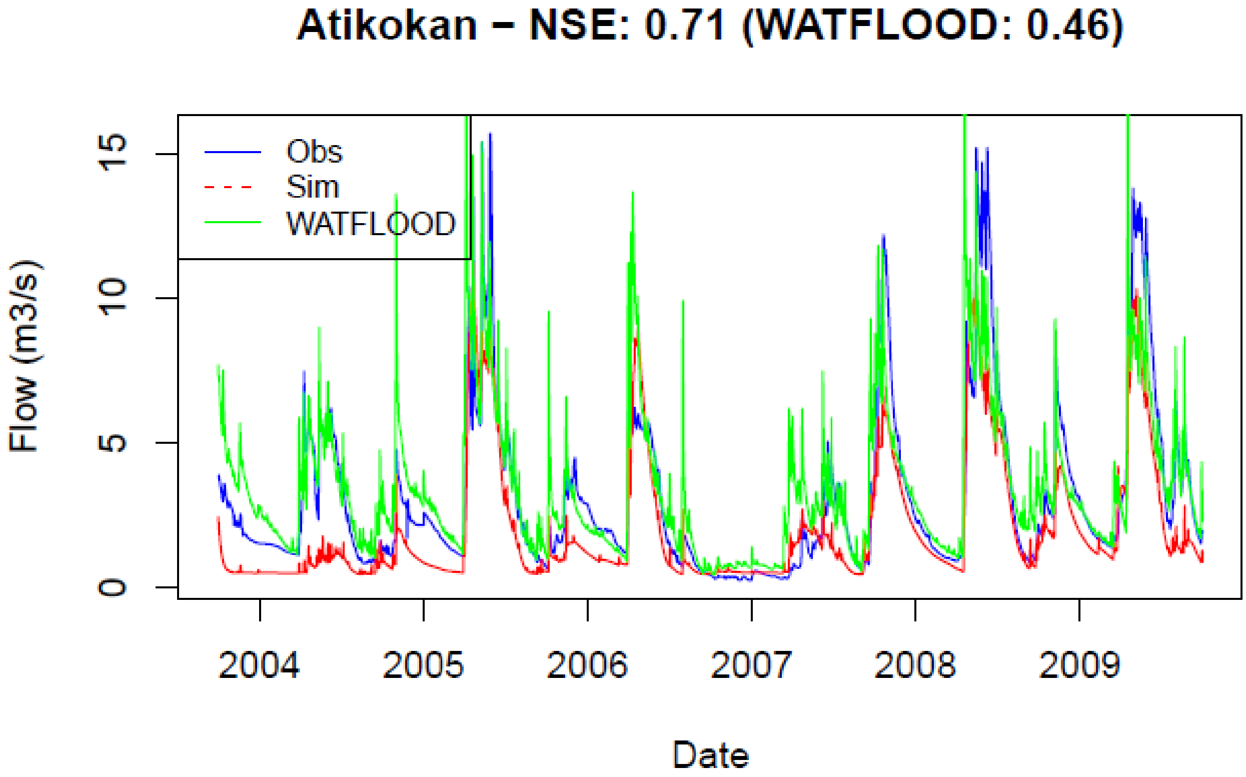
LLC − Running Avg NSE: 0.69 (WATFLOOD: 0.84)

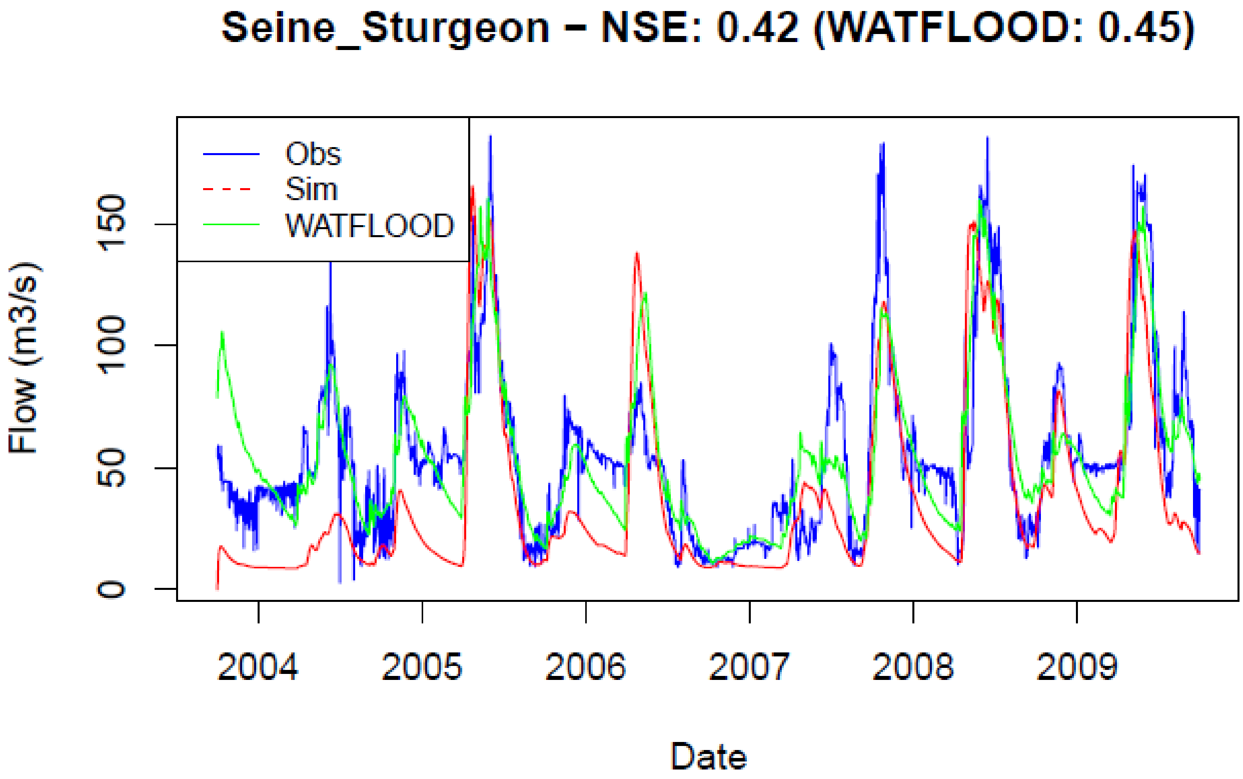


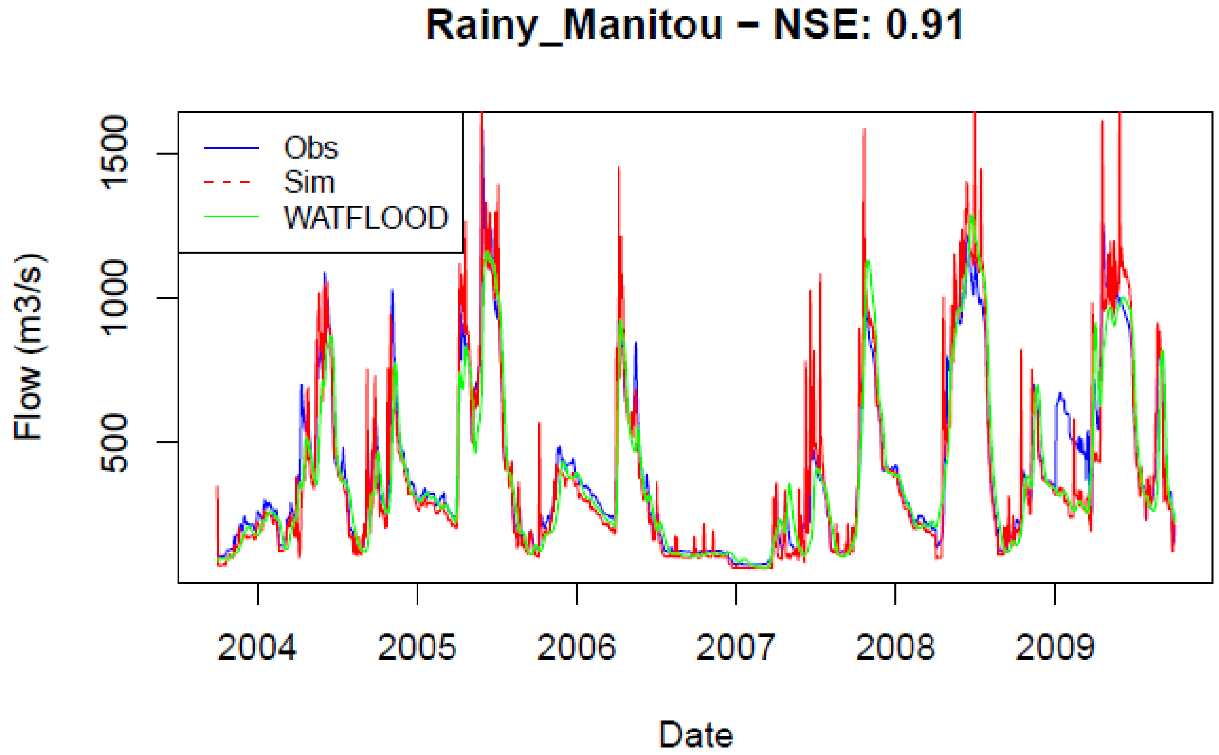
NAM − Running Avg NSE: 0.79 (WATFLOOD: 0.82)

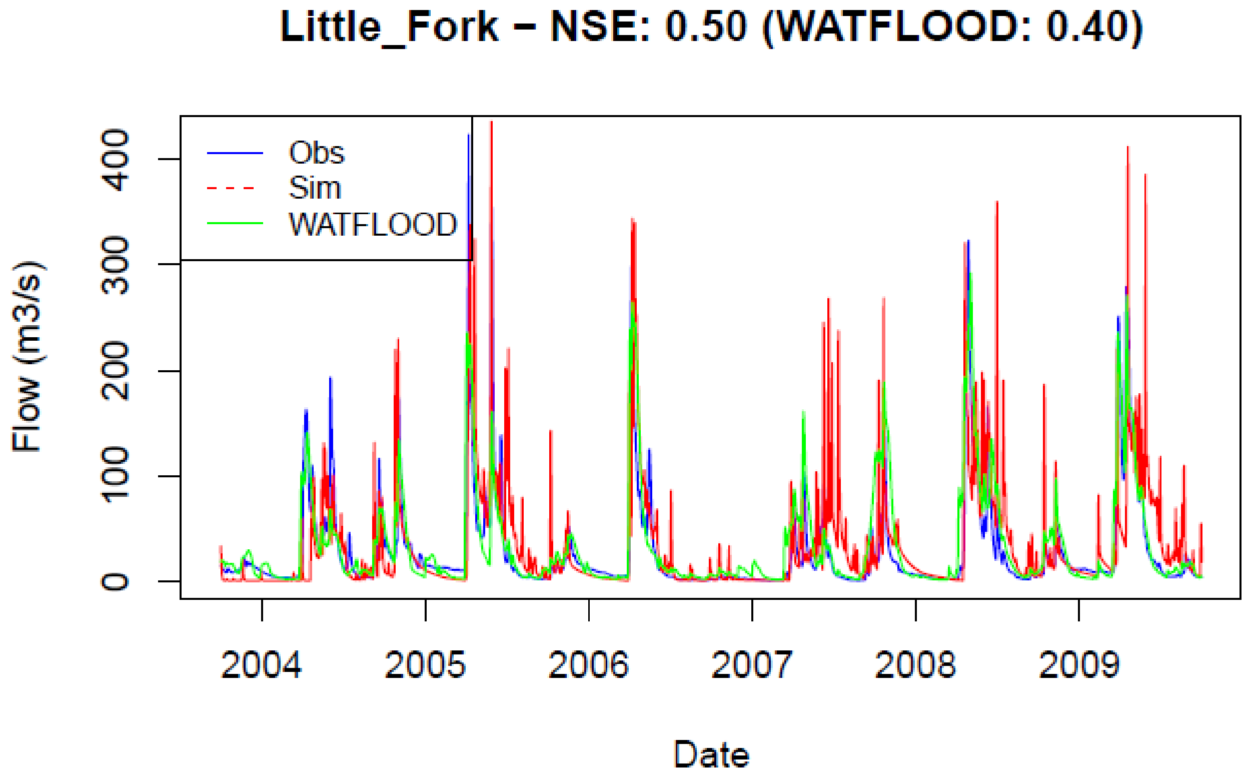
Figure B.5 – Hydrograph of stream flows for calibration period when calibrated to inflow

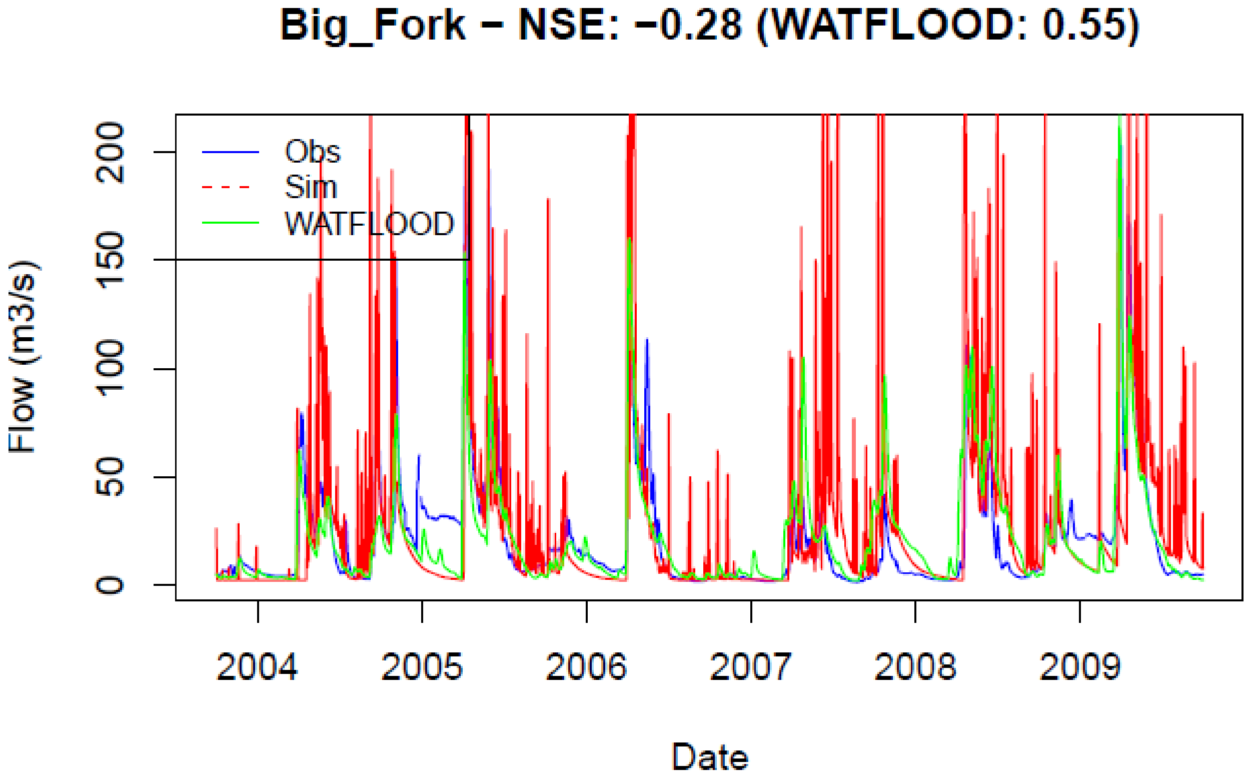


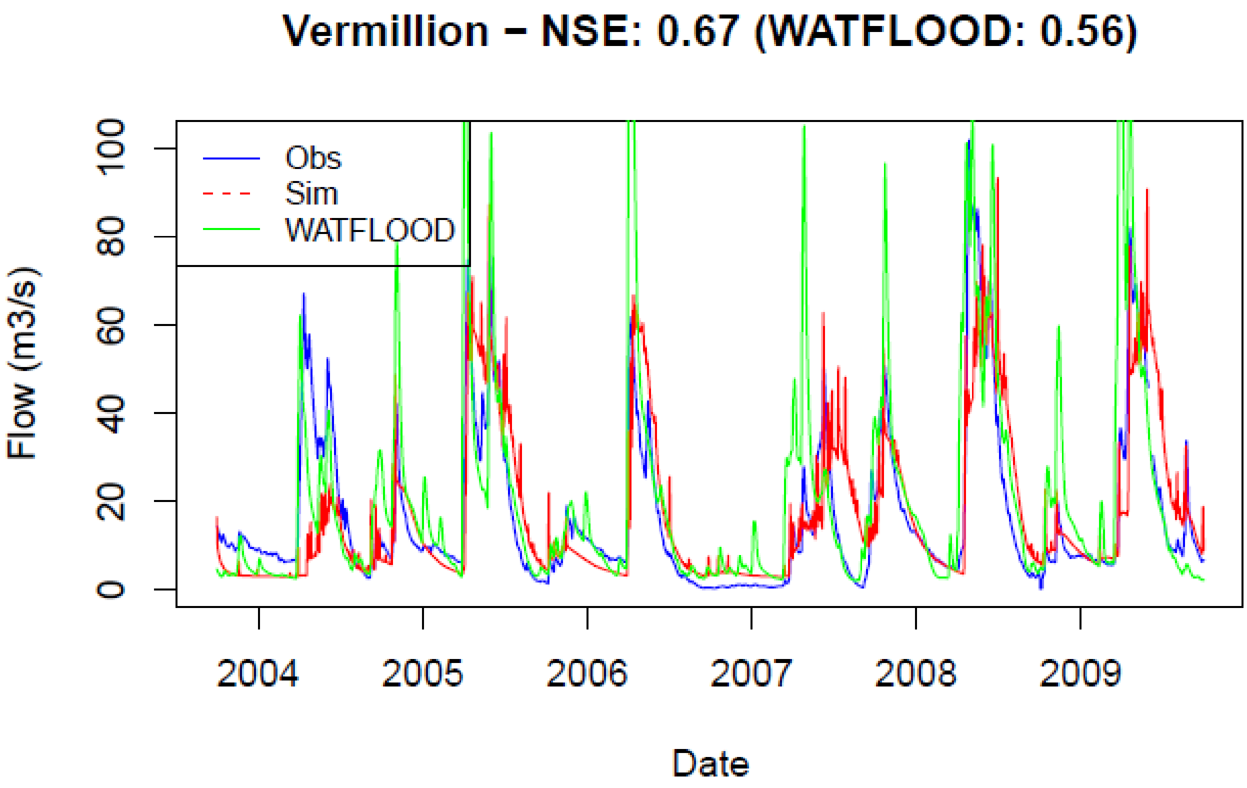


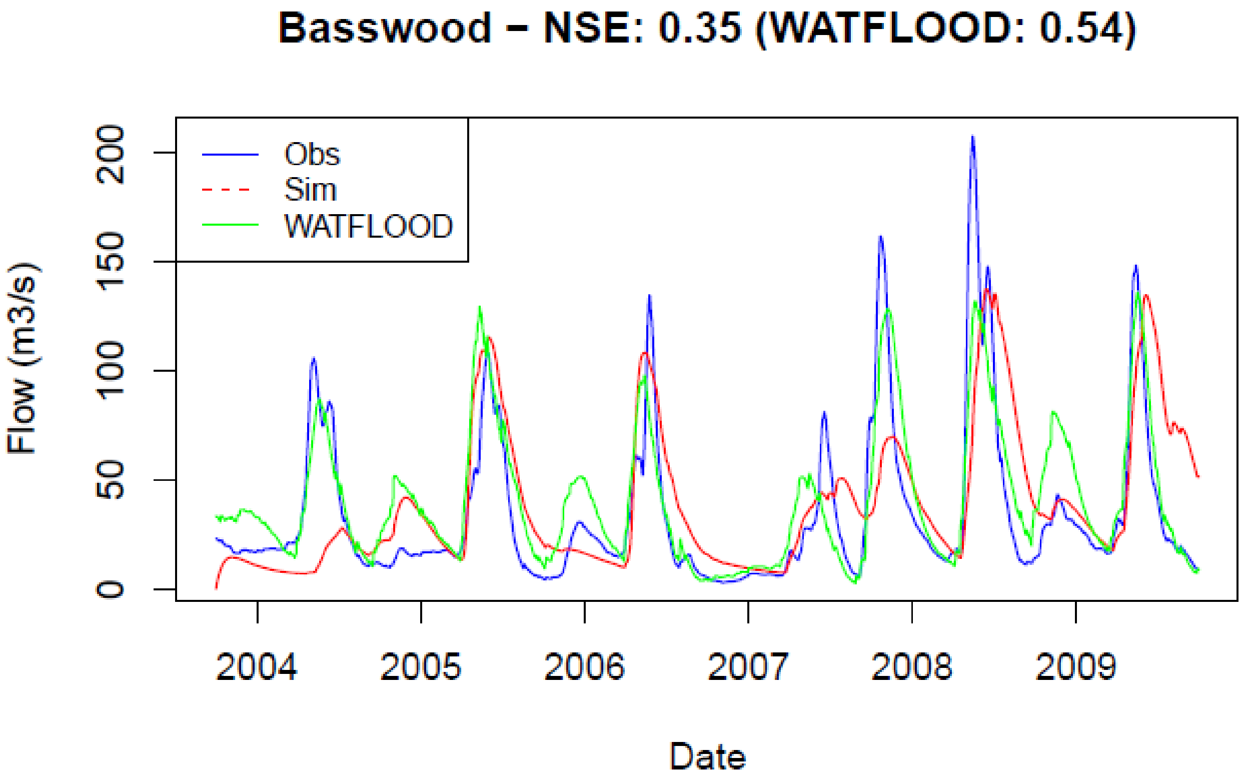

Seine_Sturgeon – NSE: 0.42 (WATFLOOD: 0.45)

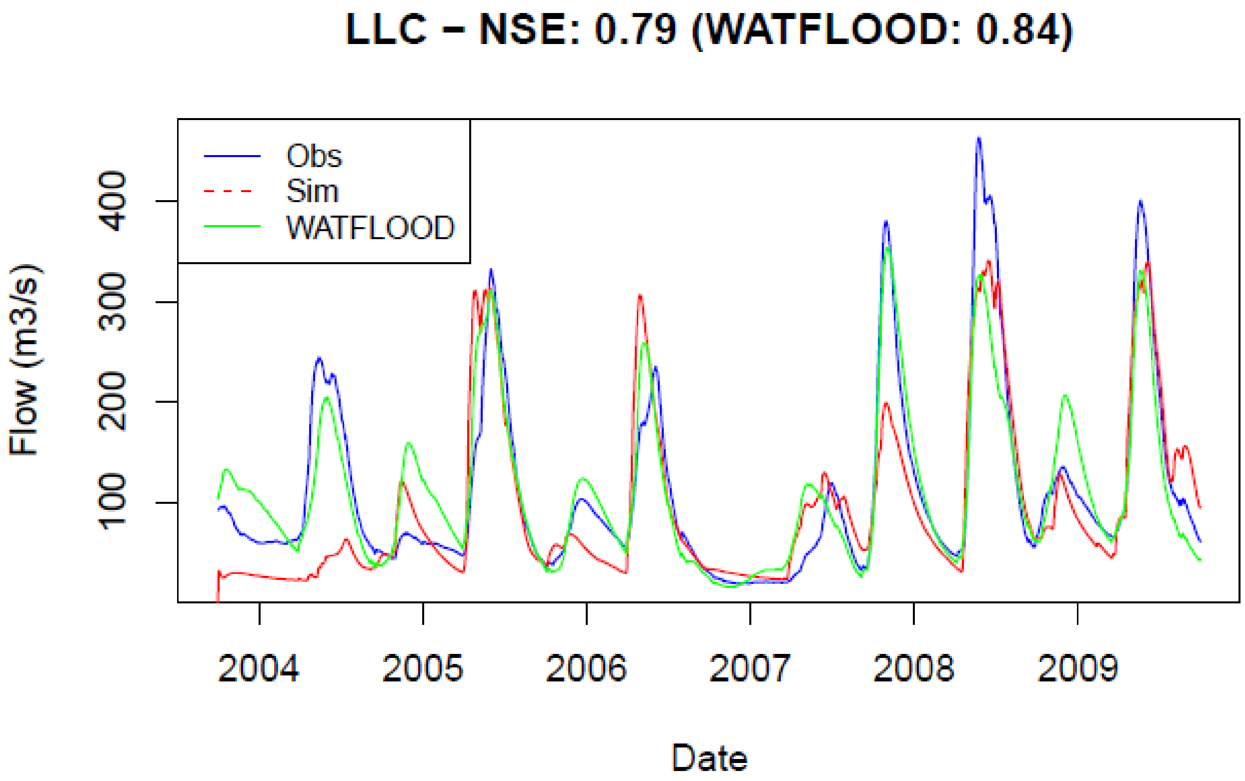

Rainy_Manitou – NSE: 0.91


Little_Fork − NSE: 0.50 (WATFLOOD: 0.40)
Obs
Sim
WATFLOOD
Flow (m3/s)
0
100
200
300
400
2004
2005
2006
2007
2008
2009
Date

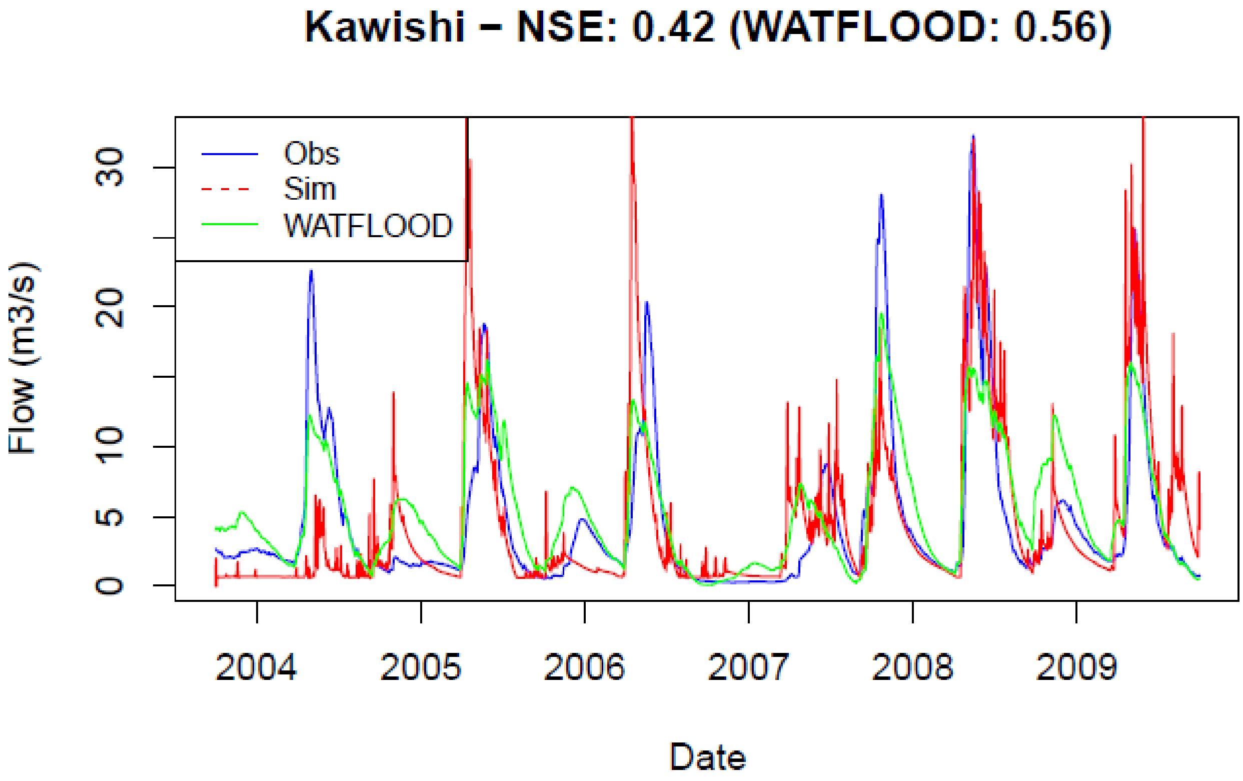

Big_Fork − NSE: −0.28 (WATFLOOD: 0.55)
Obs
Sim
WATFLOOD
Flow (m3/s)
0
50
100
150
200
2004
2005
2006
2007
2008
2009
Date


Vermillion – NSE: 0.67 (WATFLOOD: 0.56)
Obs
Sim
WATFLOOD
Flow (m3/s)
0
20
40
60
80
100
2004
2005
2006
2007
2008
2009
Date


Basswood – NSE: 0.35 (WATFLOOD: 0.54)
Obs
Sim
WATFLOOD
Flow (m3/s)
0
50
100
150
200
2004
2005
2006
2007
2008
2009
Date

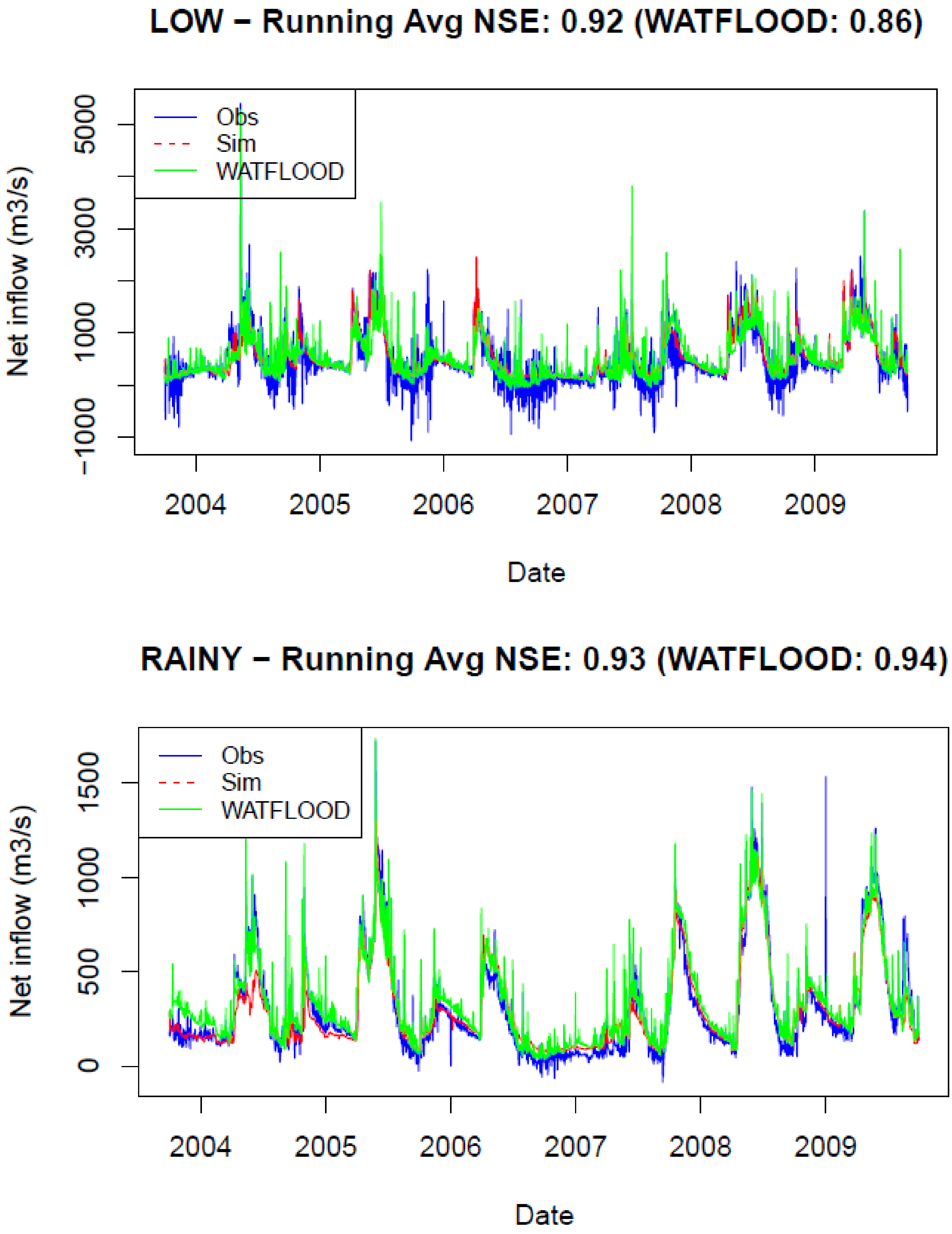


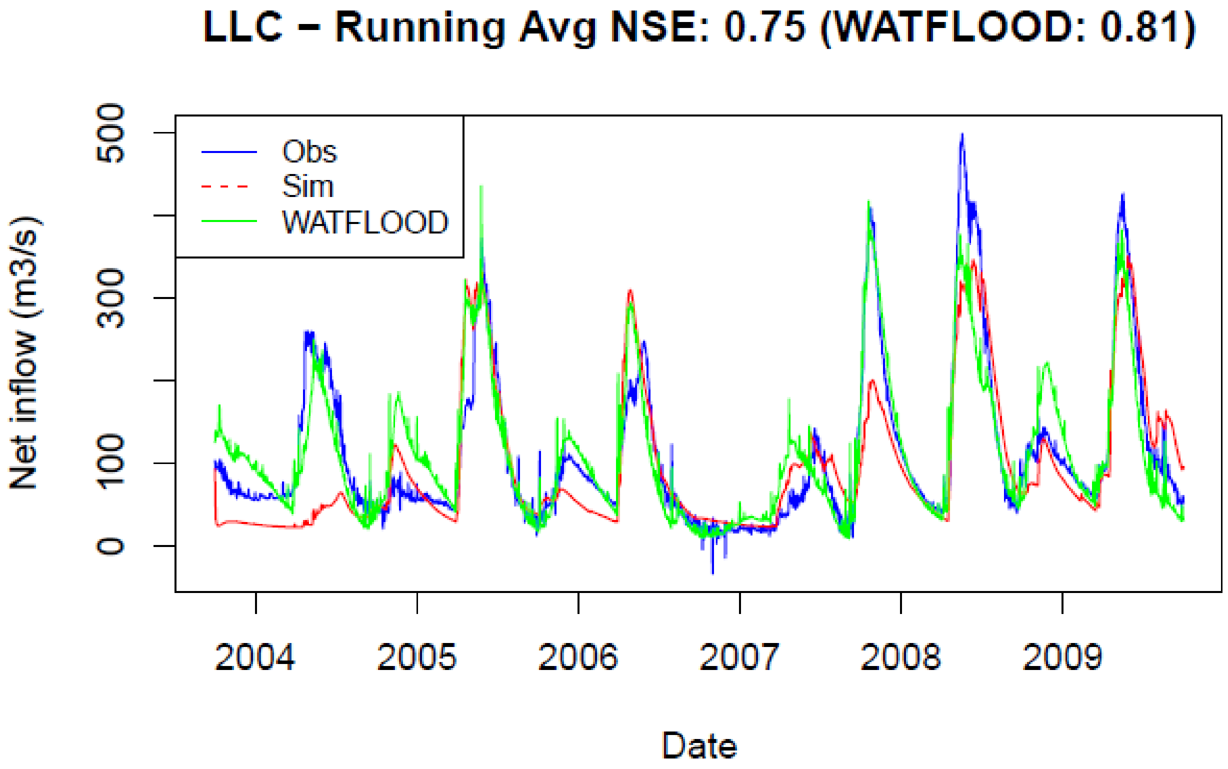
LLC – NSE: 0.79 (WATFLOOD: 0.84)



Kawishi – NSE: 0.42 (WATFLOOD: 0.56)

Figure B.6 – Inflow hydrograph for calibration period when calibrated to inflow




LLC – Running Avg NSE: 0.75 (WATFLOOD: 0.81)
Obs
Sim
WATFLOOD
Net inflow (m3/s)
0
100
300
500
2004
2005
2006
2007
2008
2009
Date

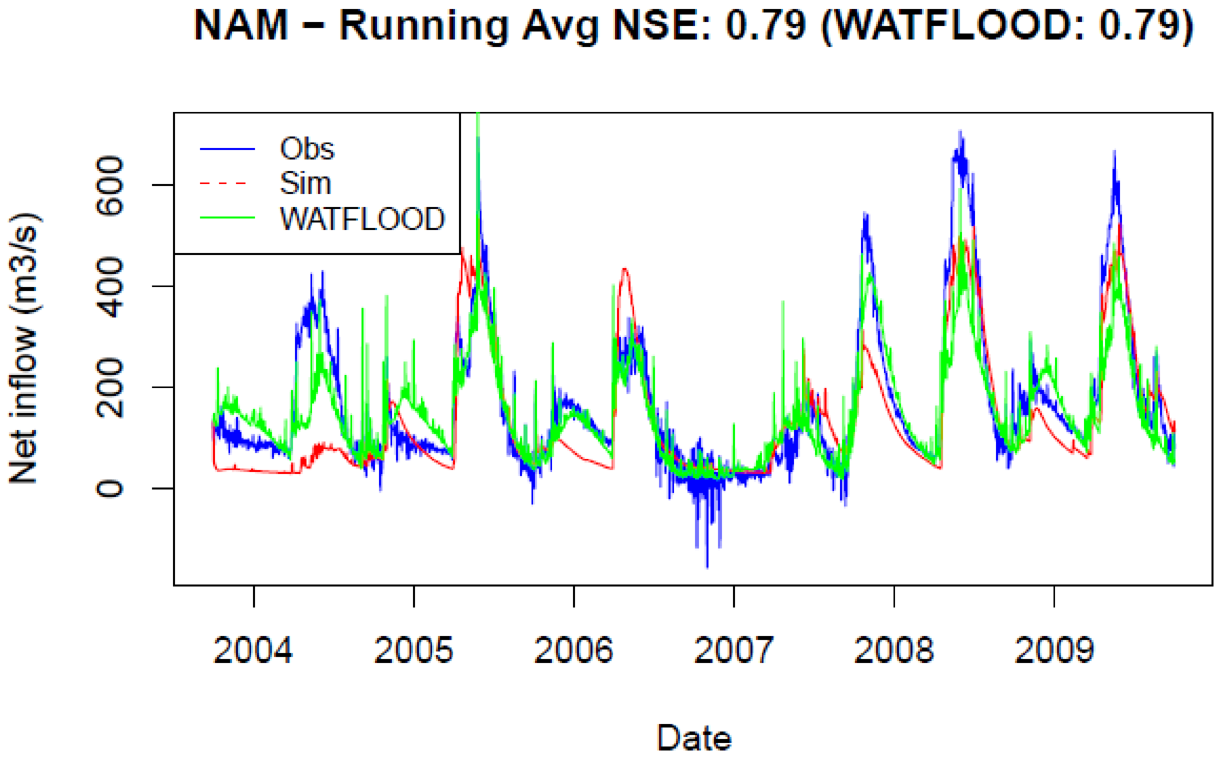
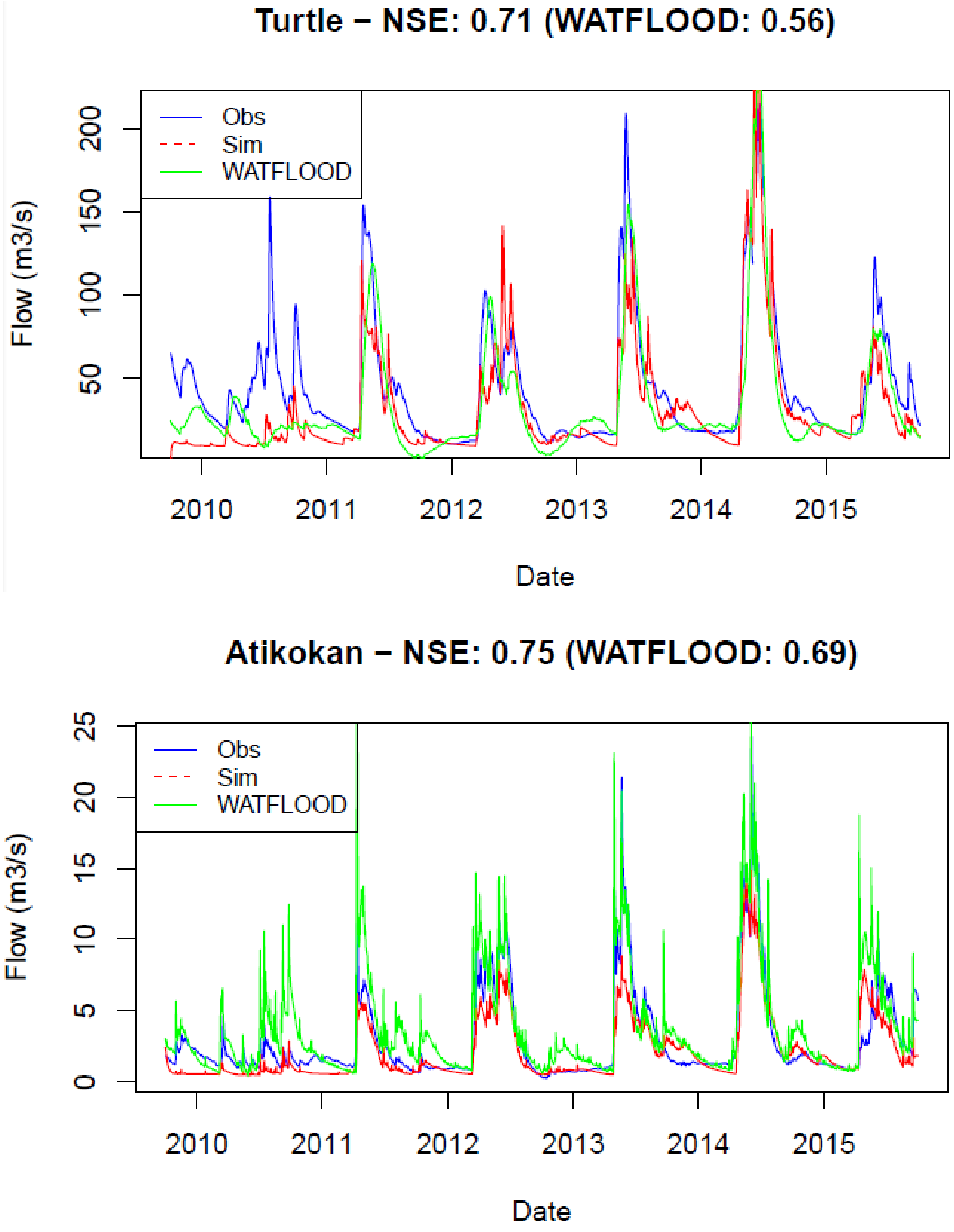

NAM – Running Avg NSE: 0.79 (WATFLOOD: 0.79)
Obs
Sim
WATFLOOD
Net inflow (m3/s)
0
200
400
600
2004
2005
2006
2007
2008
2009
Date

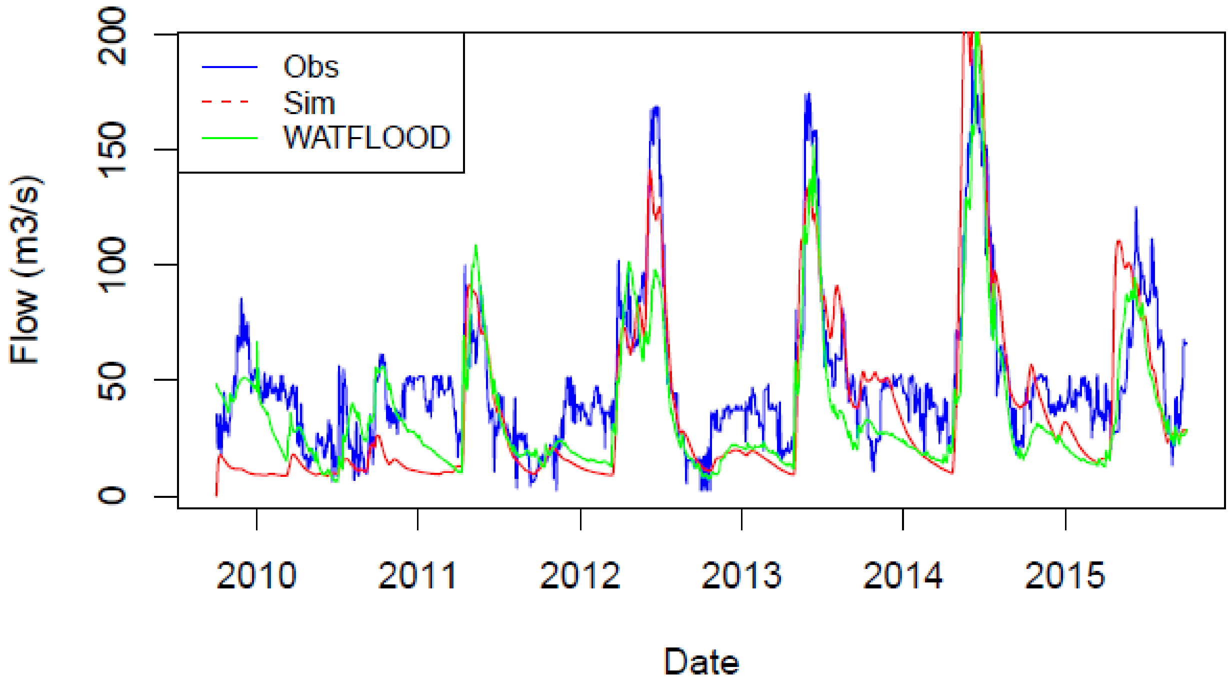
Figure B.7 – Hydrograph of stream flows for validation period when calibrated to inflow



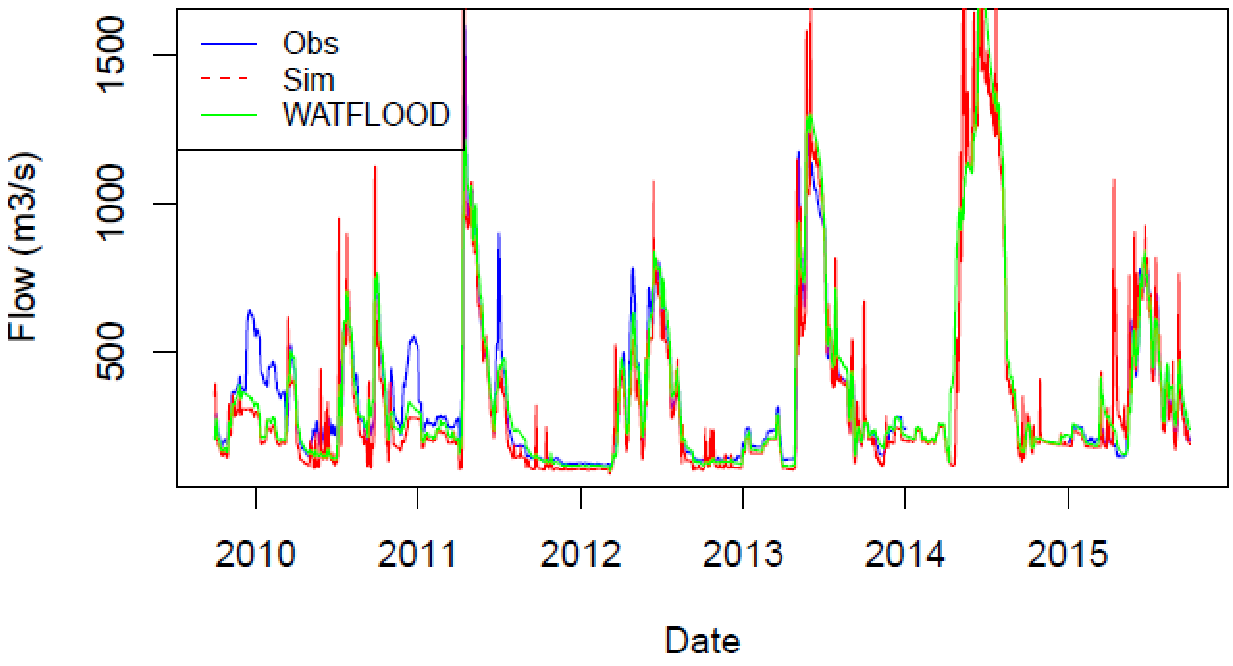
Turtle – NSE: 0.71 (WATFLOOD: 0.56)

Atikokan – NSE: 0.75 (WATFLOOD: 0.69)

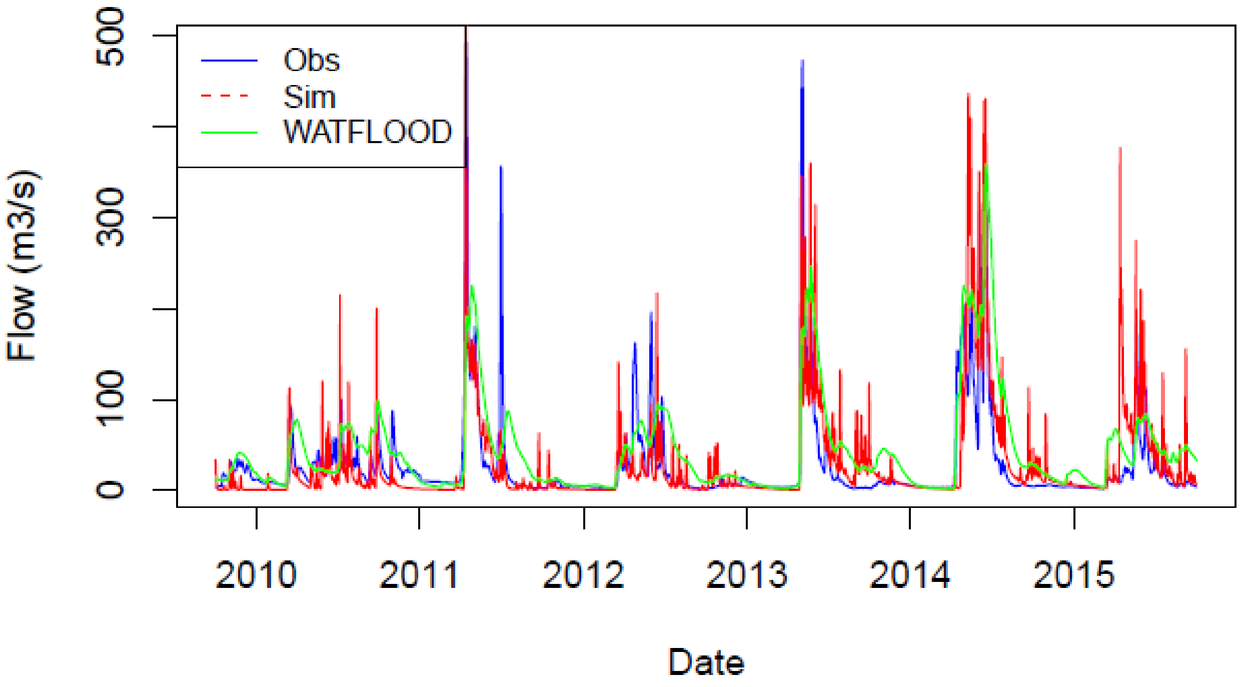
Seine_Sturgeon − NSE: 0.49 (WATFLOOD: 0.64)



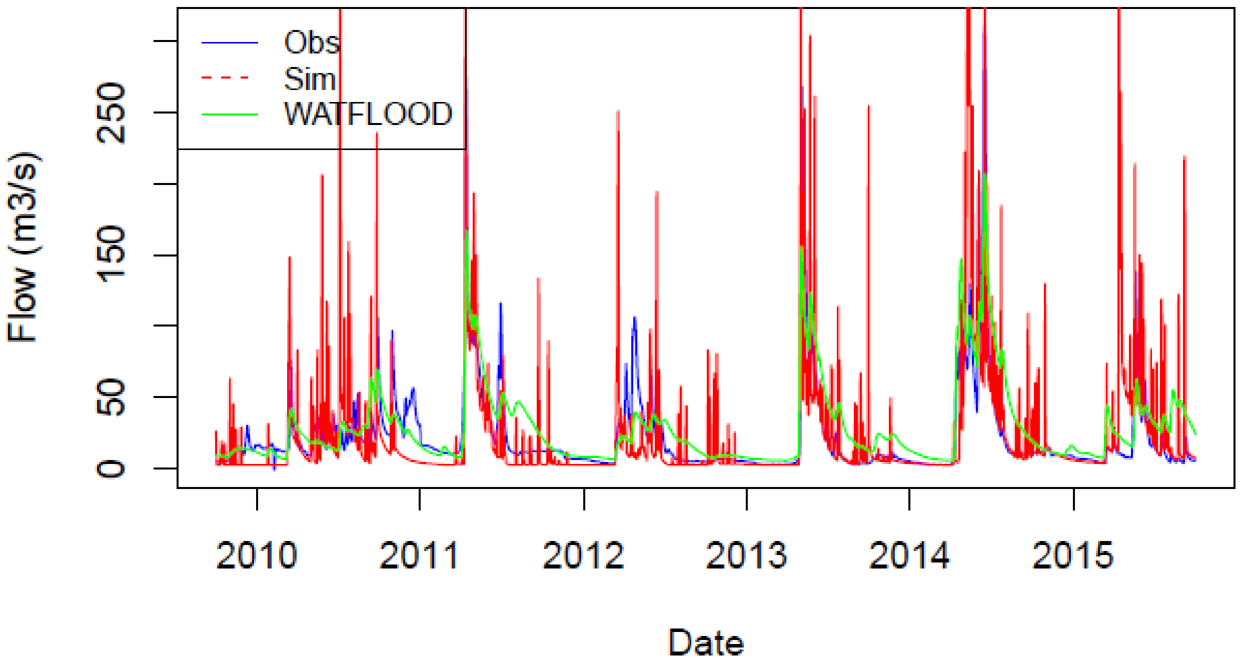
Rainy_Manitou − NSE: 0.88



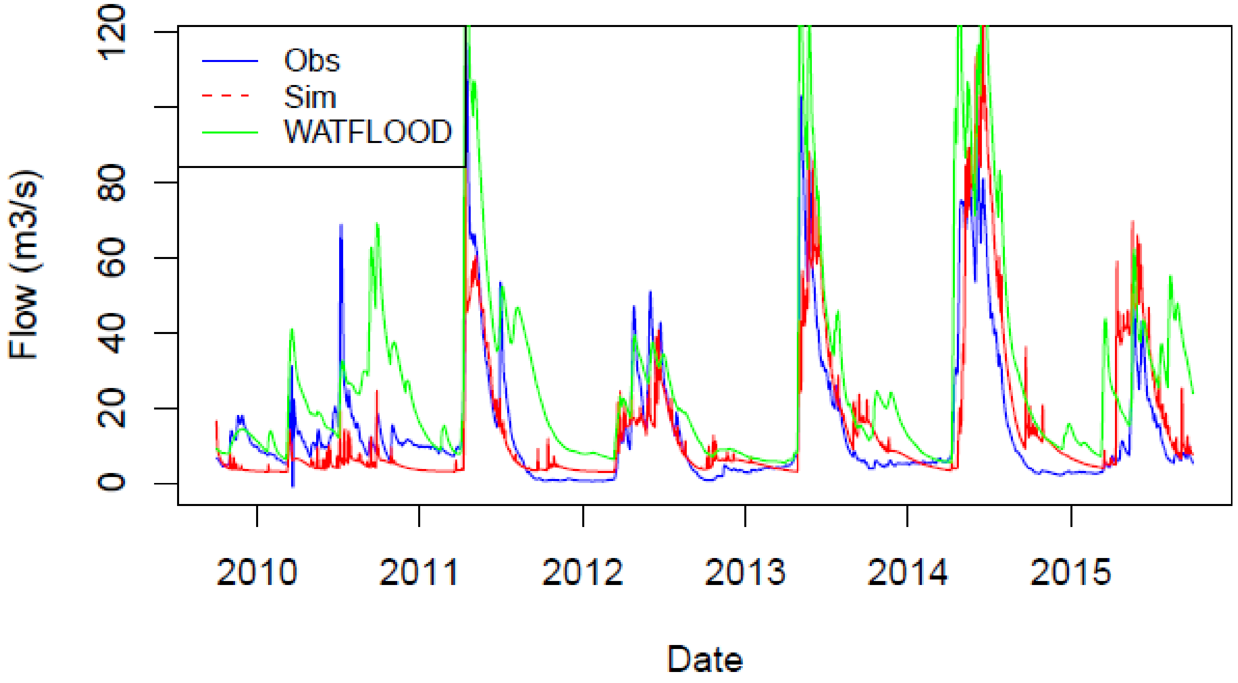
Little_Fork – NSE: 0.51 (WATFLOOD: 0.29)



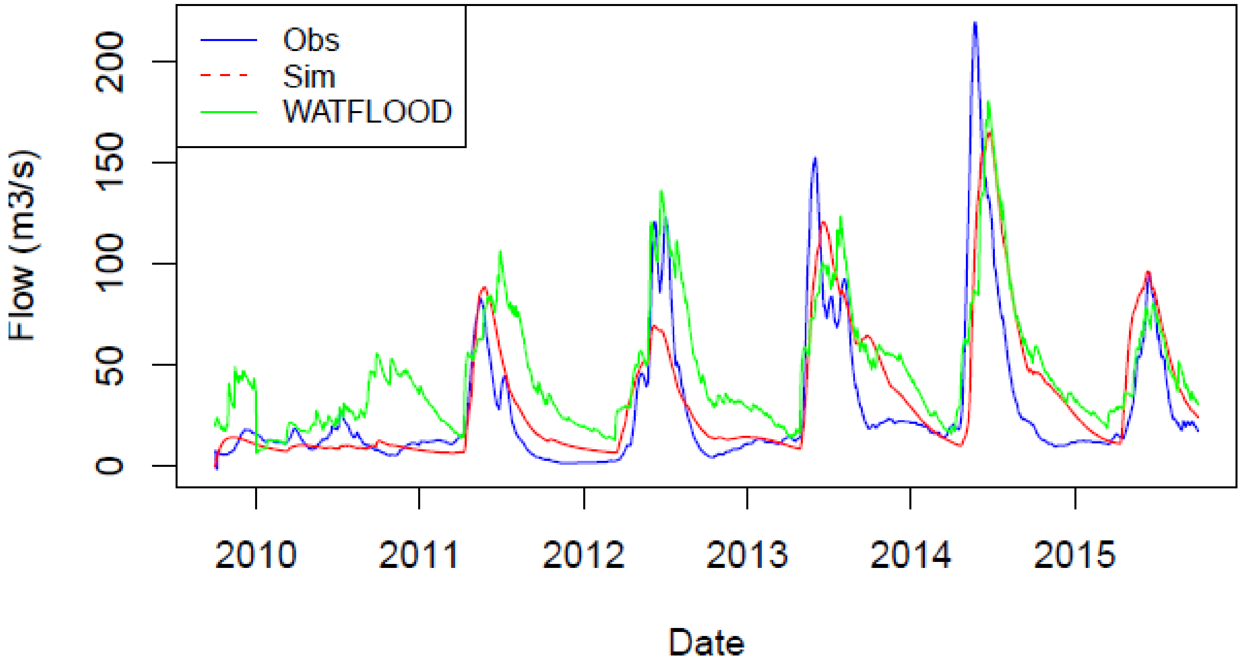
Big_Fork – NSE: 0.07 (WATFLOOD: 0.67)

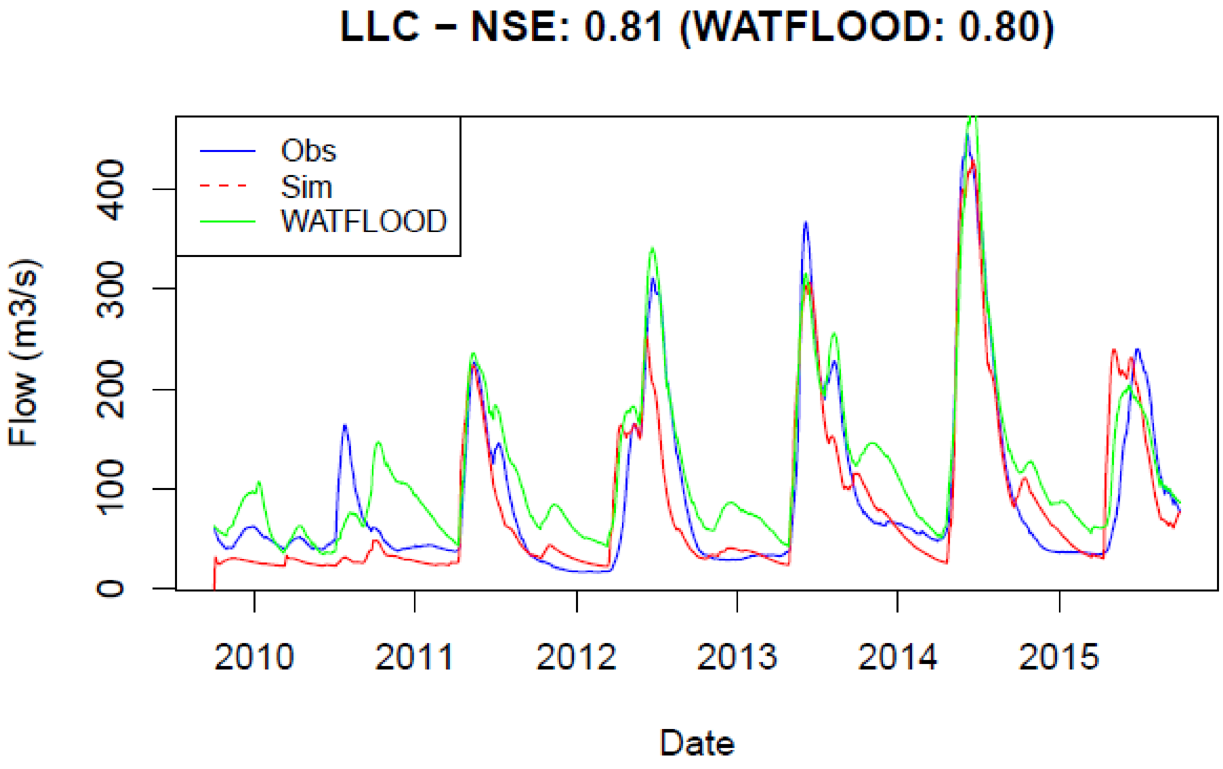


Vermillion − NSE: 0.59 (WATFLOOD: 0.70)

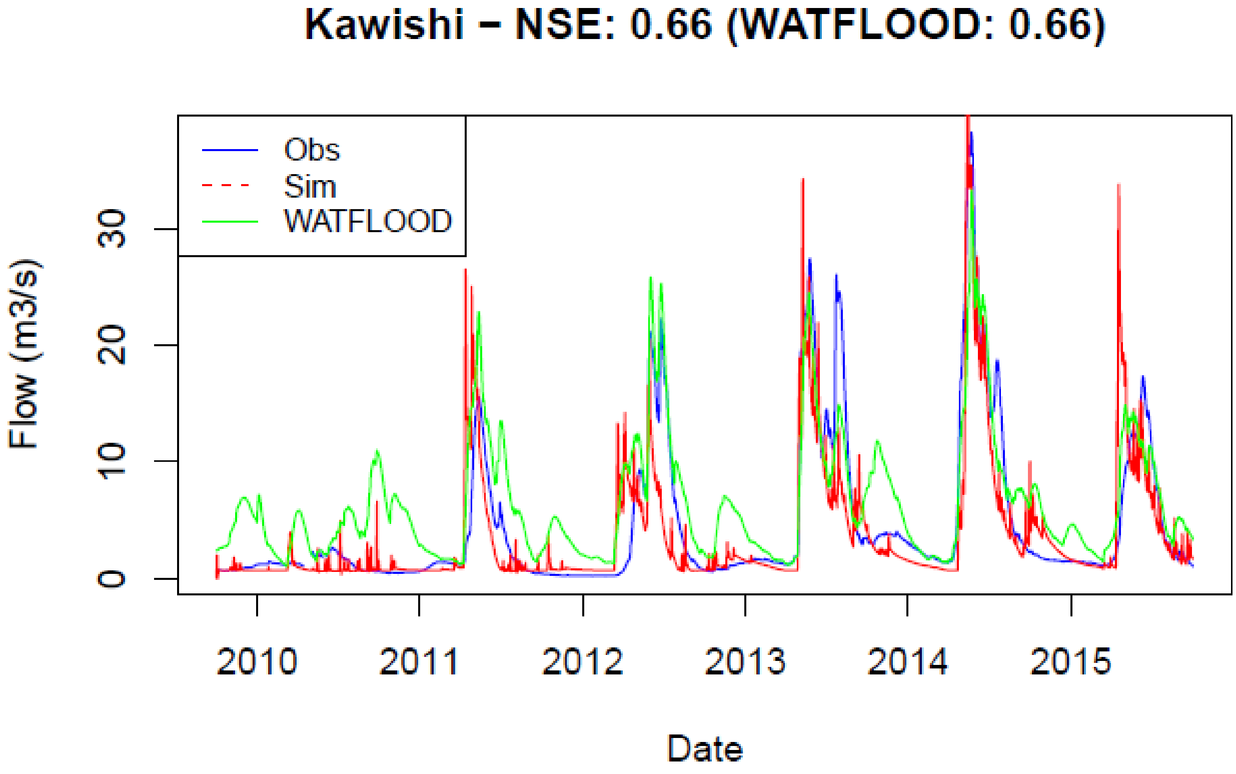


Basswood − NSE: 0.62 (WATFLOOD: 0.29)



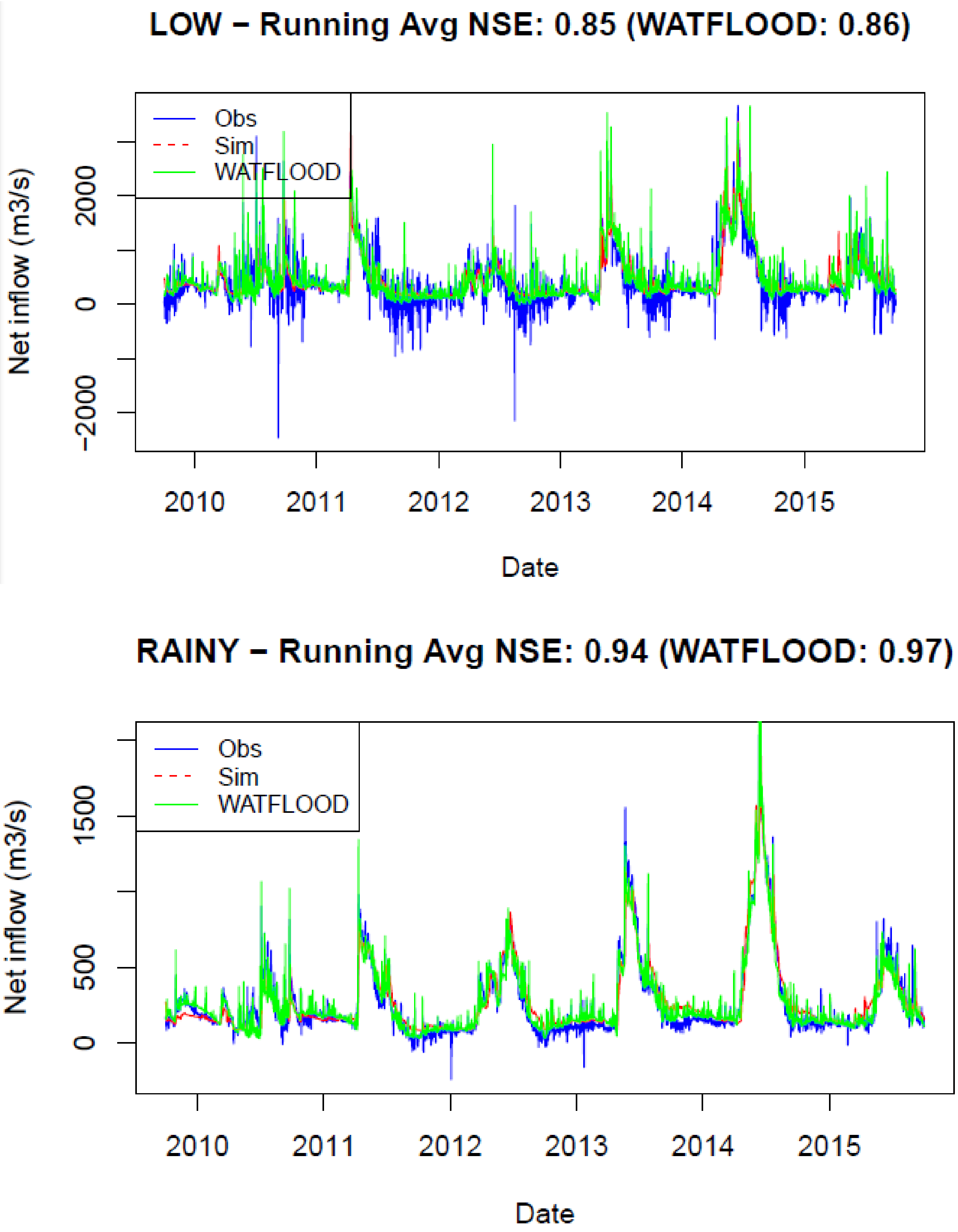


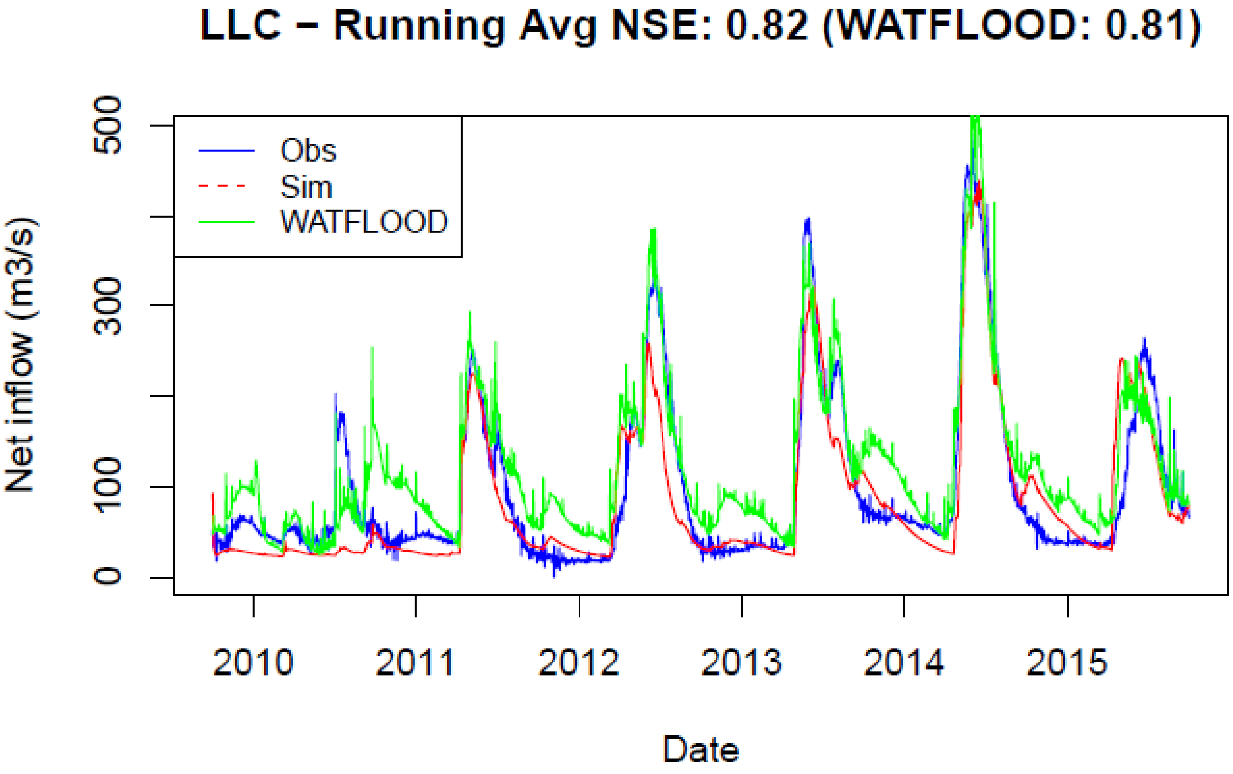
LLC – NSE: 0.81 (WATFLOOD: 0.80)
Obs
Sim
WATFLOOD
Flow (m3/s)
0
100
200
300
400
2010
2011
2012
2013
2014
2015
Date

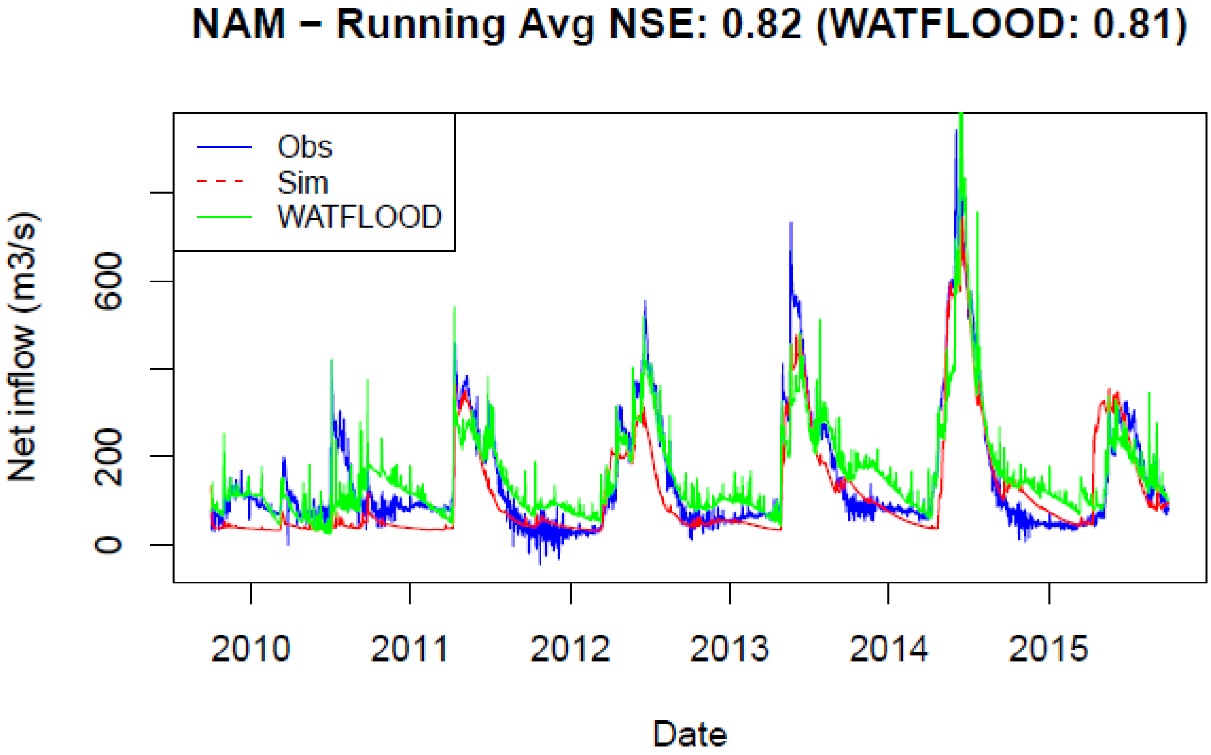


Kawishi – NSE: 0.66 (WATFLOOD: 0.66)
Obs
Sim
WATFLOOD
Flow (m3/s)
0
10
20
30
2010
2011
2012
2013
2014
2015
Date

Figure B.8 – Reservoir inflow hydrograph for validation period when calibrated to inflow

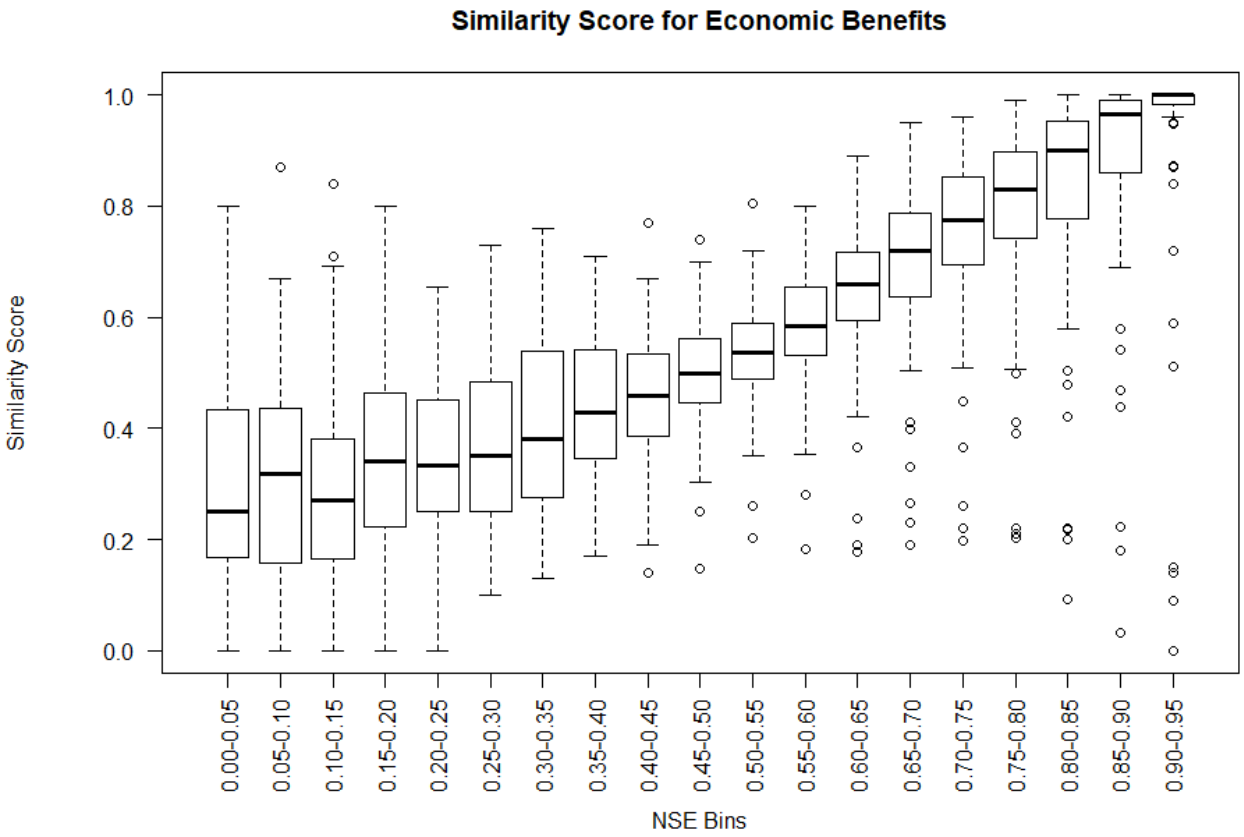


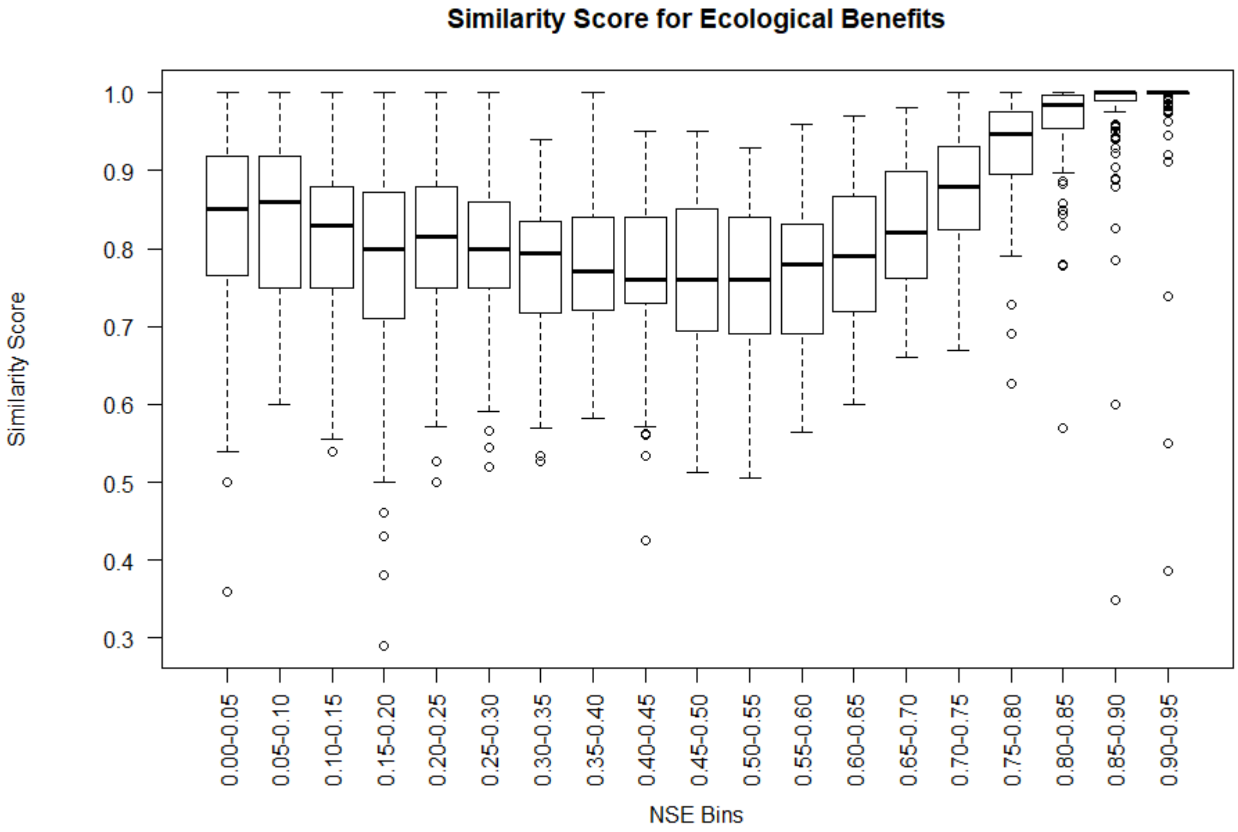

LLC – Running Avg NSE: 0.82 (WATFLOOD: 0.81)
Obs
Sim
WATFLOOD
Net inflow (m3/s)
0
100
300
500
2010
2011
2012
2013
2014
2015
Date


NAM – Running Avg NSE: 0.82 (WATFLOOD: 0.81)
Obs
Sim
WATFLOOD
Net inflow (m3/s)
0
200
600
2010
2011
2012
2013
2014
2015
Date

Appendix D – Box Whisker Plots of Similarity Scores

Figure C.1 – Box whisker plot of similarity score calibrated to inflow NSE




Similarity Score for Ecological Benefits
Similarity Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.00-0.05
0.05-0.10
0.10-0.15
0.15-0.20
0.20-0.25
0.25-0.30
0.30-0.35
0.35-0.40
0.40-0.45
0.45-0.50
0.50-0.55
0.55-0.60
0.60-0.65
0.65-0.70
0.70-0.75
0.75-0.80
0.80-0.85
0.85-0.90
0.90-0.95
NSE Bins

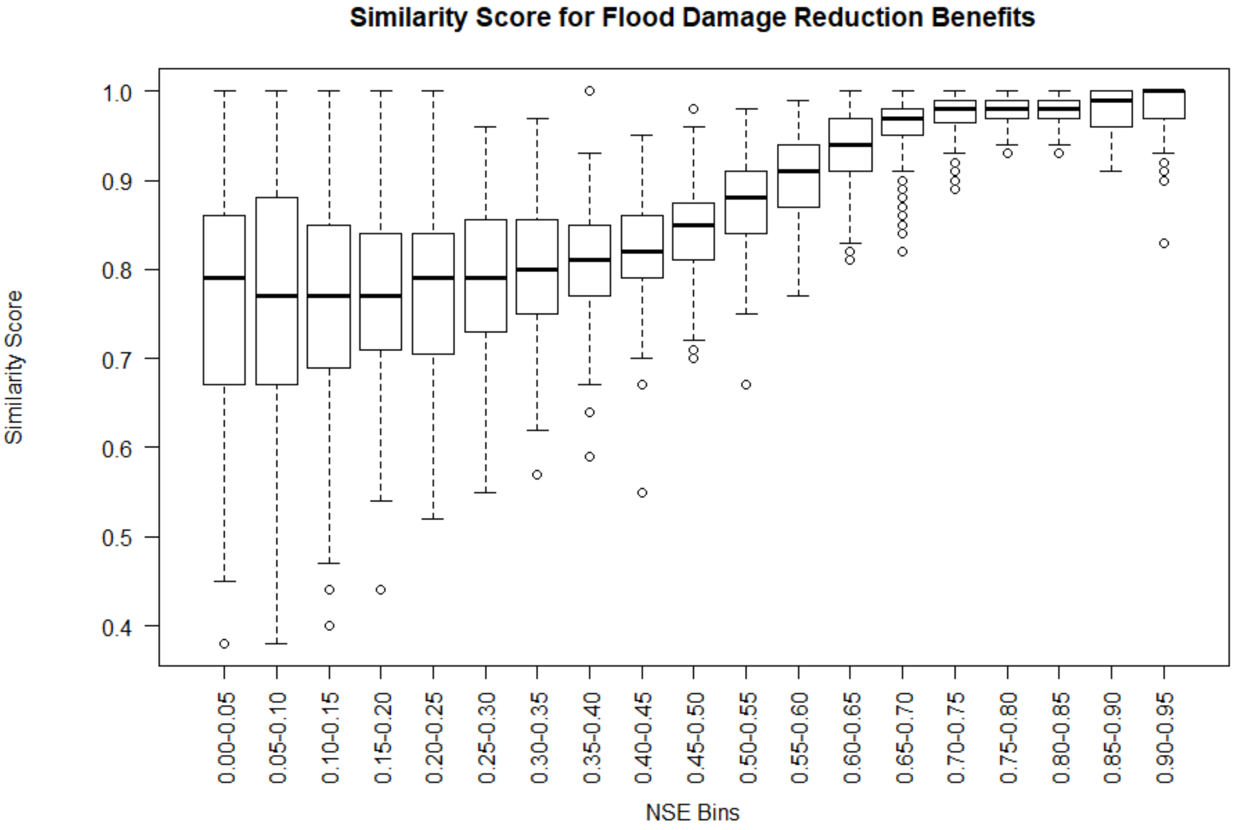
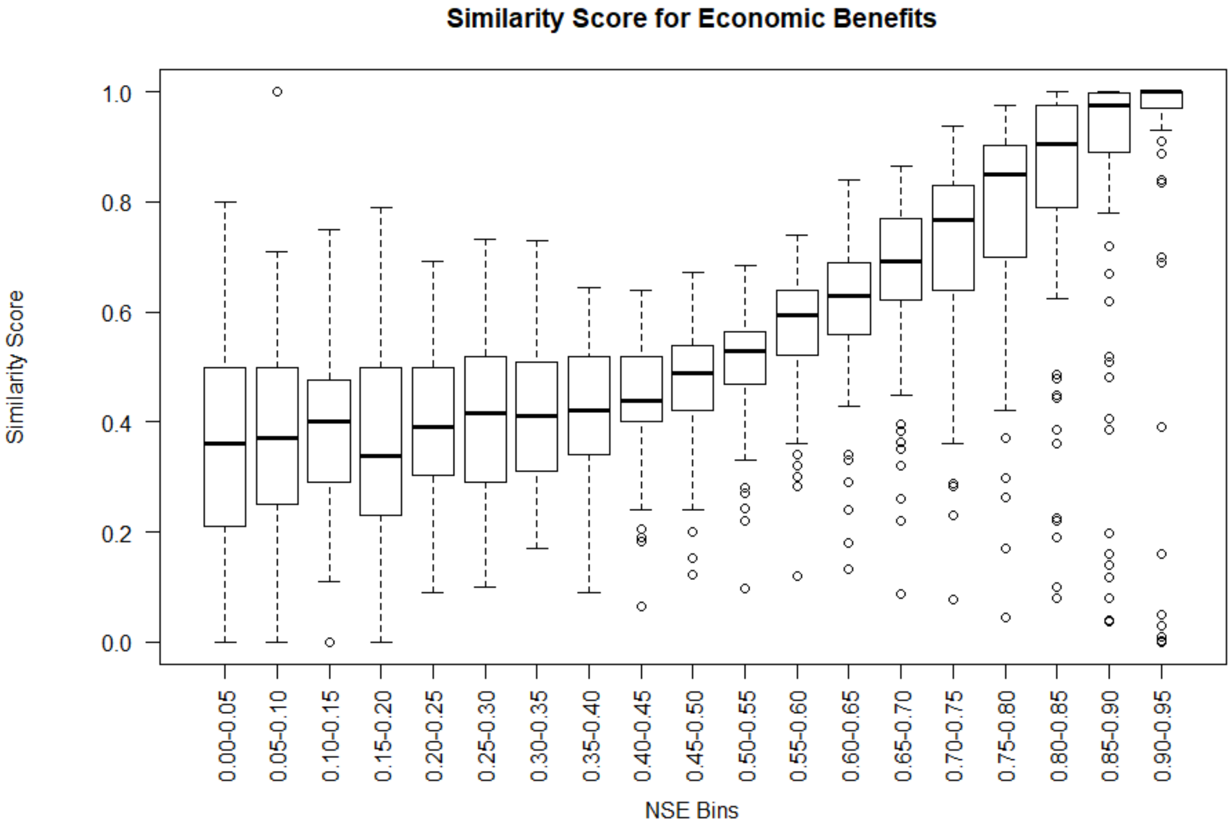
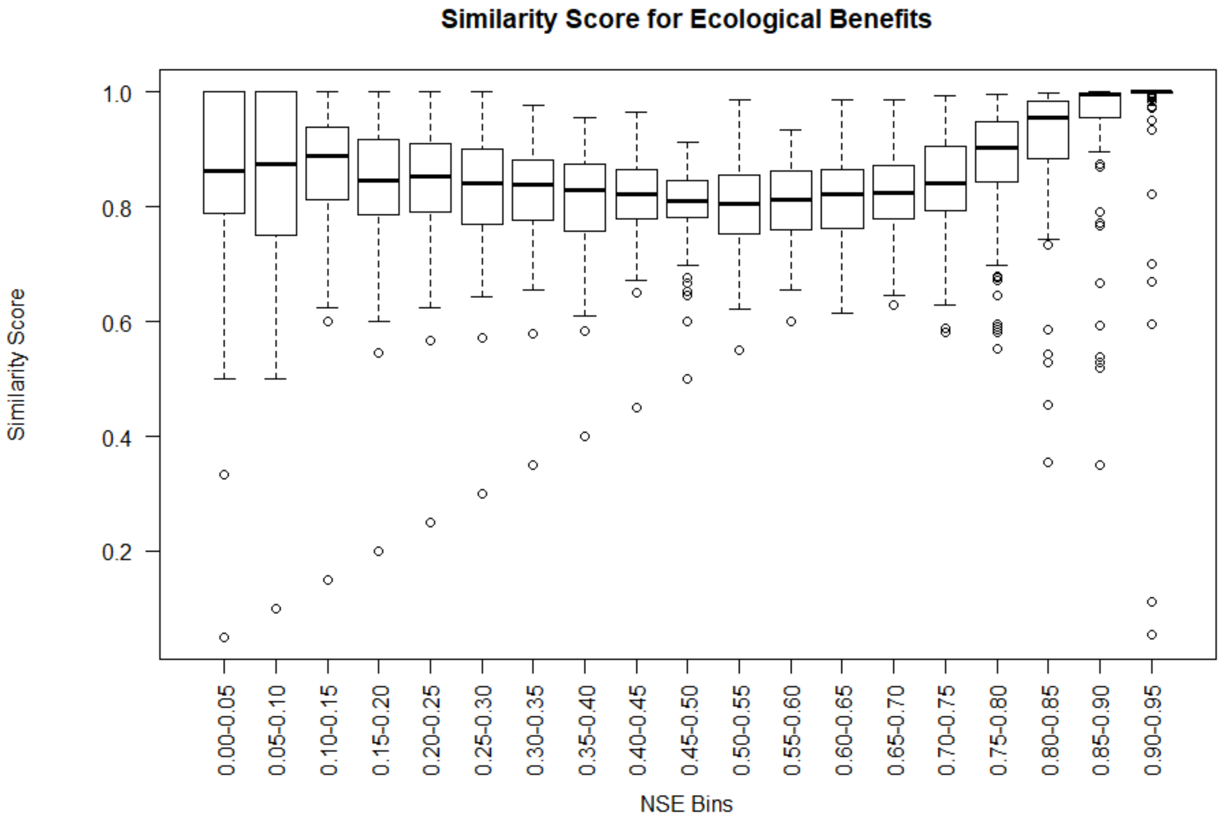
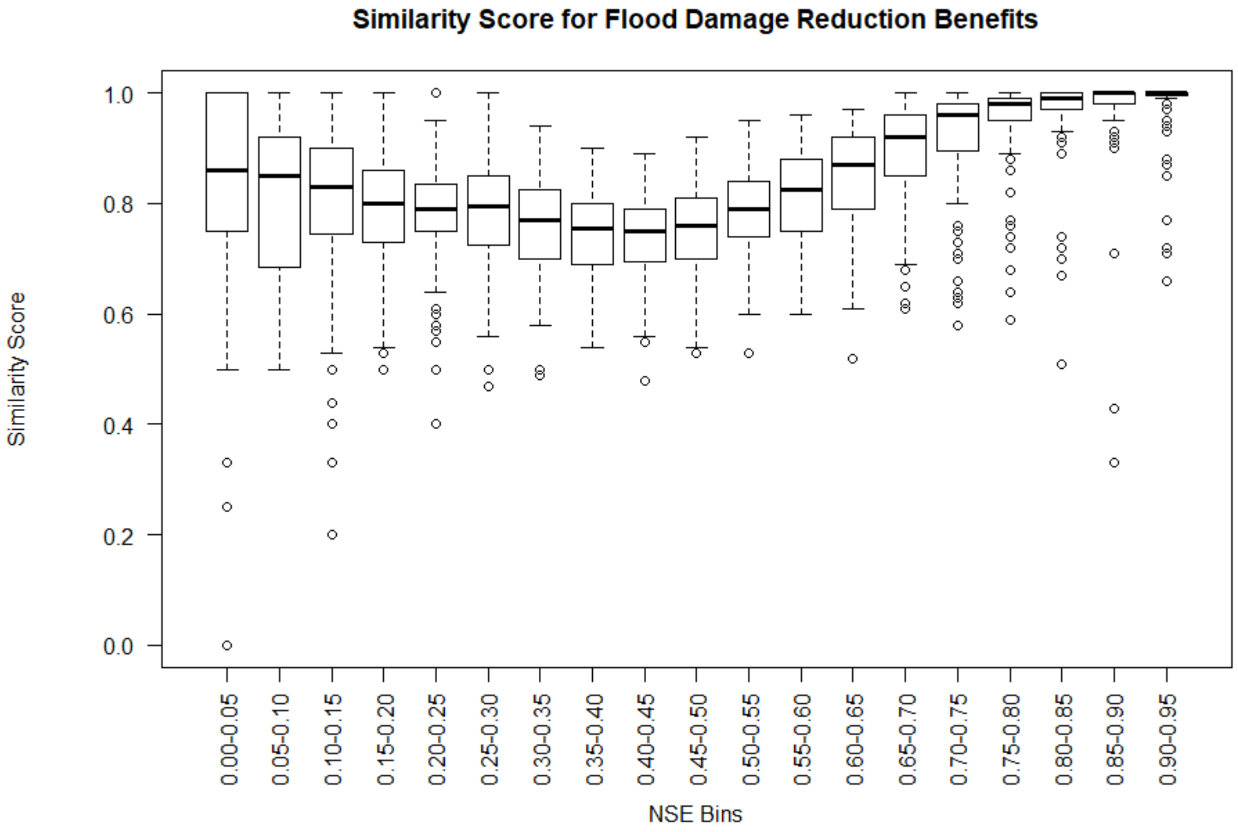

Similarity Score for Flood Damage Reduction Benefits
Similarity Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.00-0.05
0.05-0.10
0.10-0.15
0.15-0.20
0.20-0.25
0.25-0.30
0.30-0.35
0.35-0.40
0.40-0.45
0.45-0.50
0.50-0.55
0.55-0.60
0.60-0.65
0.65-0.70
0.70-0.75
0.75-0.80
0.80-0.85
0.85-0.90
0.90-0.95
NSE Bins

Figure C.2 – Box whisker plot of similarity score calibrated to stream flow NSE



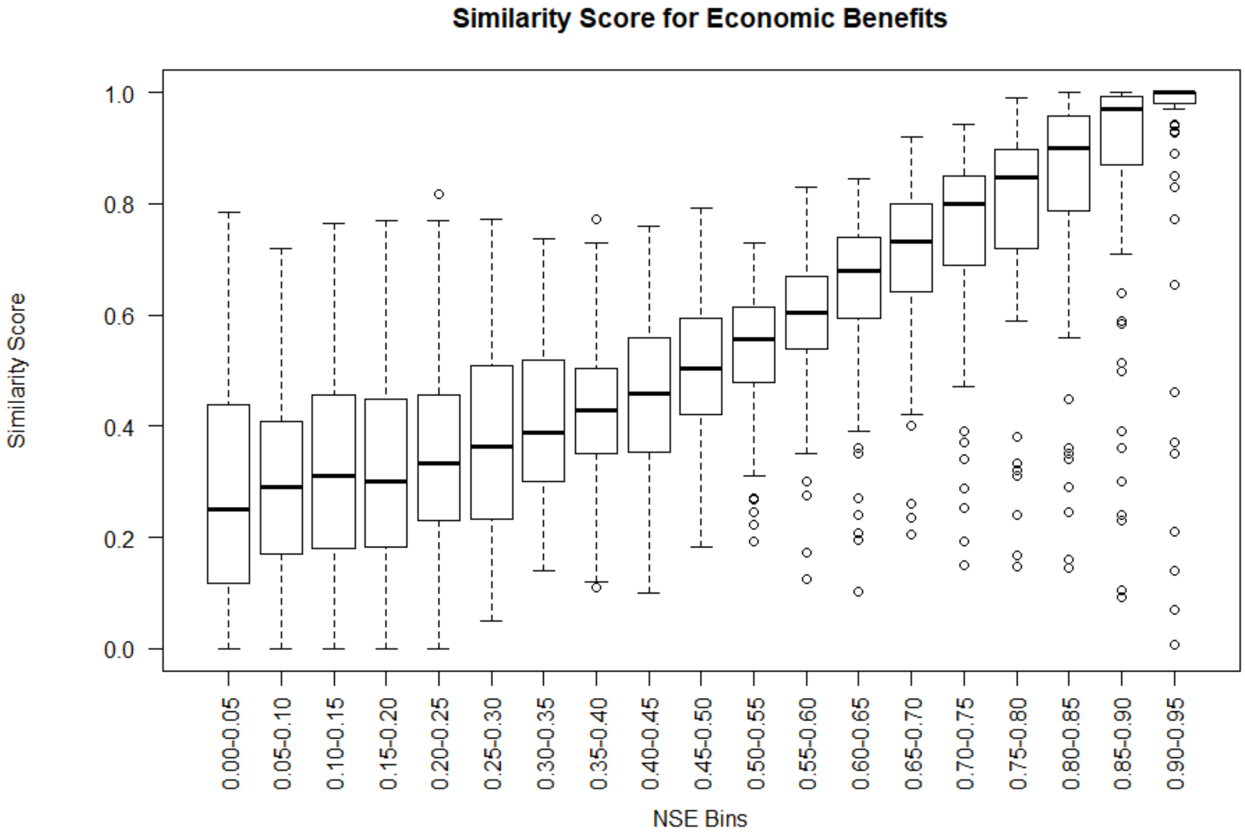


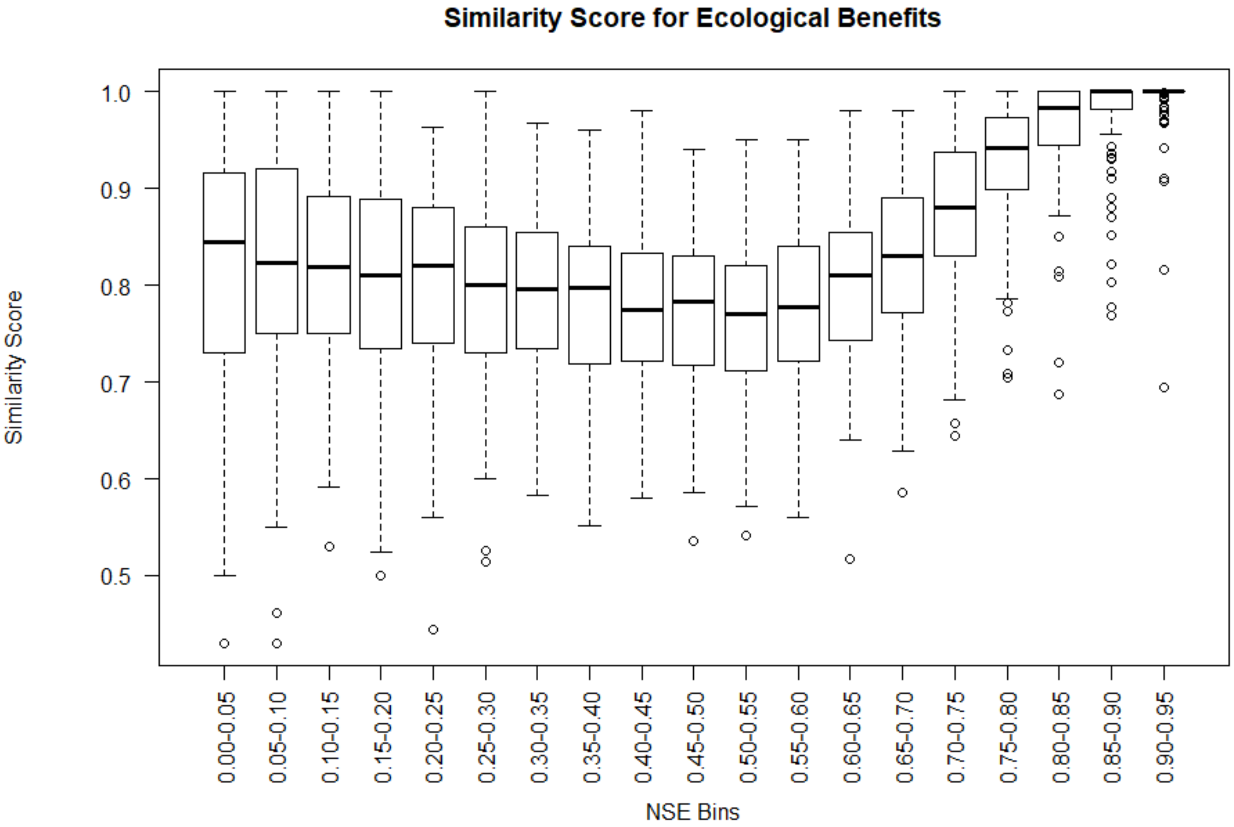
Similarity Score for Ecological Benefits



Similarity Score for Flood Damage Reduction Benefits

Figure C.3 – Box whisker plot of similarity score calibrated to spring inflow NSE




Similarity Score for Ecological Benefits
Similarity Score
1.0
0.9
0.8
0.7
0.6
0.5
0.00-0.05
0.05-0.10
0.10-0.15
0.15-0.20
0.20-0.25
0.25-0.30
0.30-0.35
0.35-0.40
0.40-0.45
0.45-0.50
0.50-0.55
0.55-0.60
0.60-0.65
0.65-0.70
0.70-0.75
0.75-0.80
0.80-0.85
0.85-0.90
0.90-0.95
NSE Bins

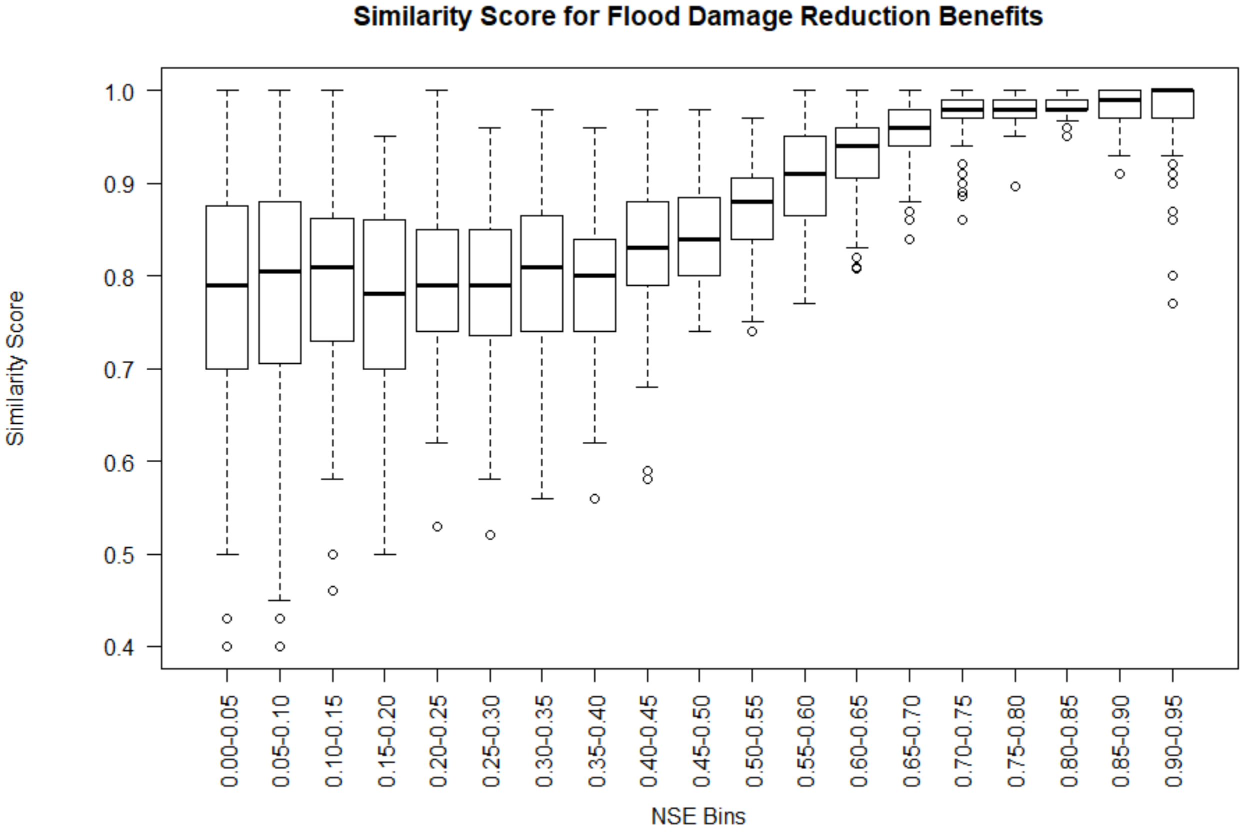
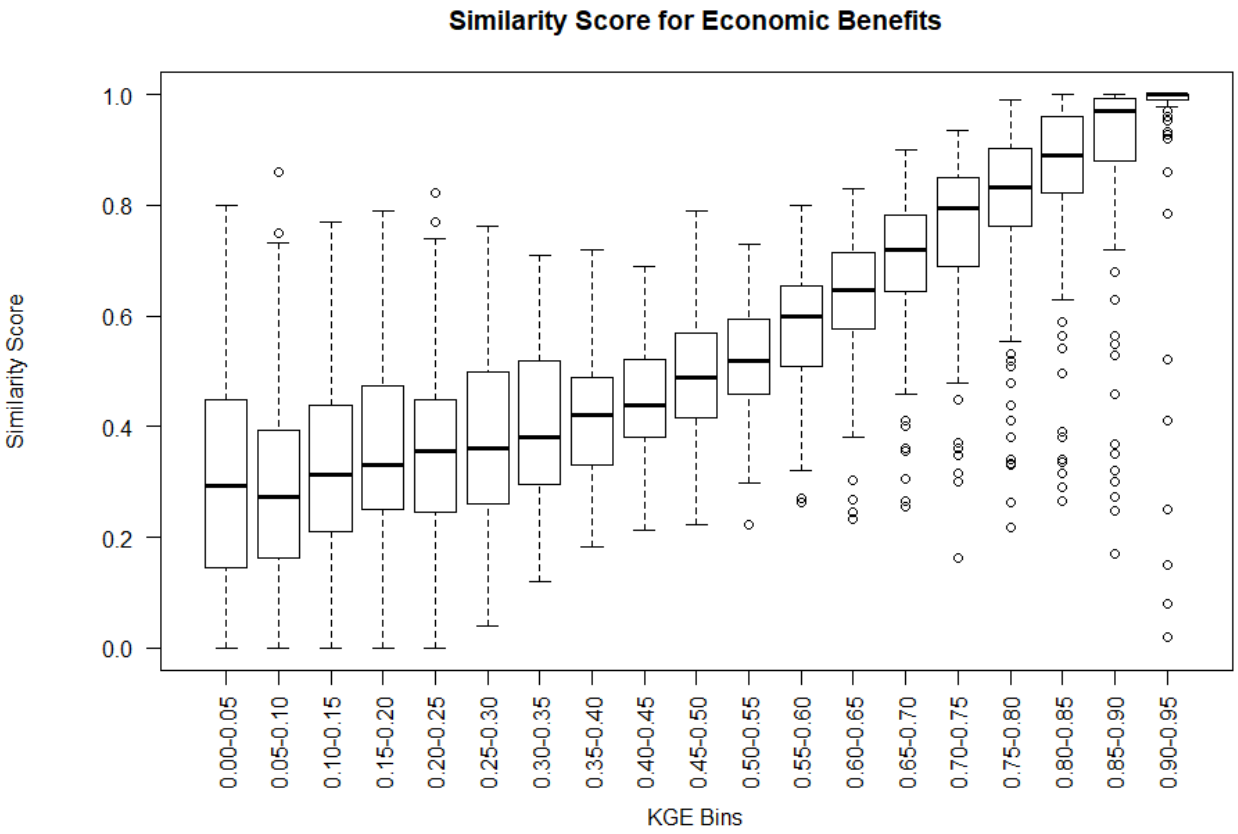
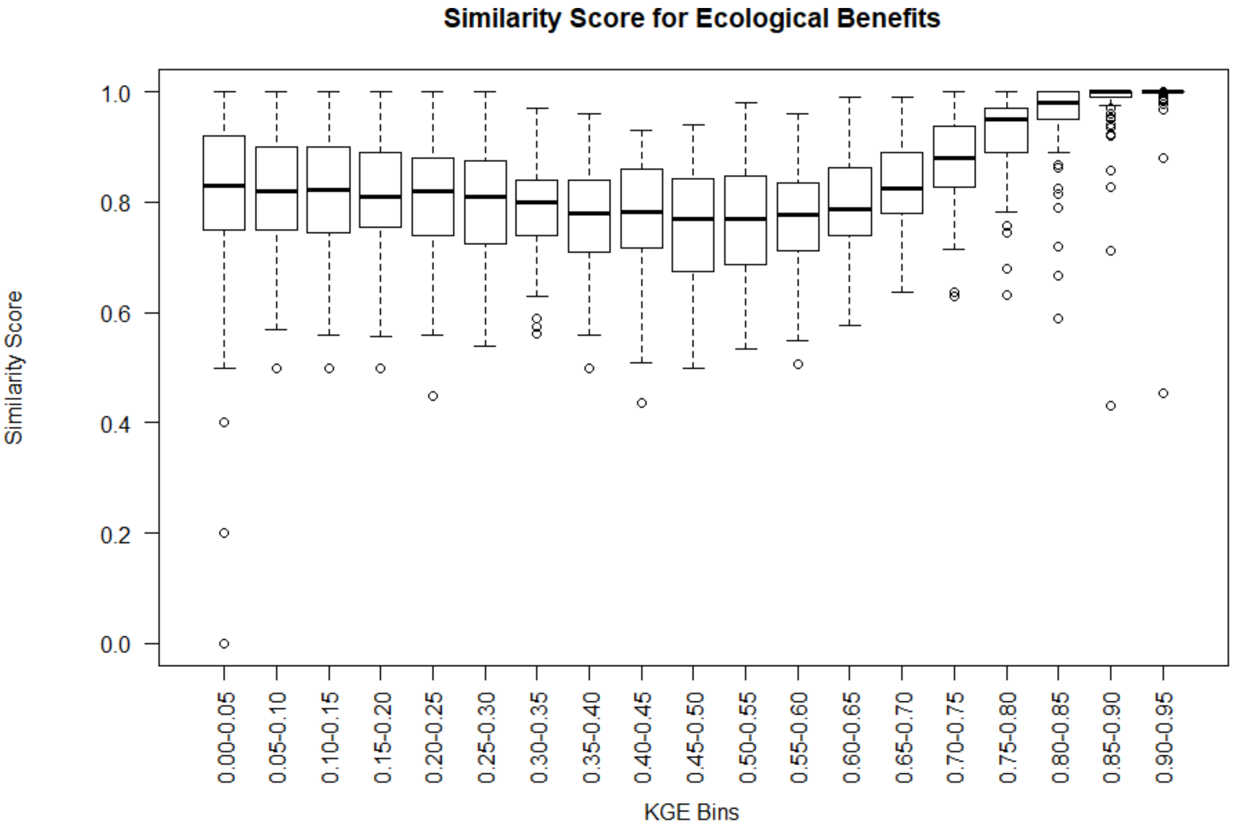

Similarity Score for Flood Damage Reduction Benefits
Similarity Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.00-0.05
0.05-0.10
0.10-0.15
0.15-0.20
0.20-0.25
0.25-0.30
0.30-0.35
0.35-0.40
0.40-0.45
0.45-0.50
0.50-0.55
0.55-0.60
0.60-0.65
0.65-0.70
0.70-0.75
0.75-0.80
0.80-0.85
0.85-0.90
0.90-0.95
NSE Bins

Figure C.4 – Box whisker plot of similarity score calibrated to inflow KGE




Similarity Score for Ecological Benefits
Similarity Score
1.0
0.8
0.6
0.4
0.2
0.0
0.00-0.05
0.05-0.10
0.10-0.15
0.15-0.20
0.20-0.25
0.25-0.30
0.30-0.35
0.35-0.40
0.40-0.45
0.45-0.50
0.50-0.55
0.55-0.60
0.60-0.65
0.65-0.70
0.70-0.75
0.75-0.80
0.80-0.85
0.85-0.90
0.90-0.95
KGE Bins

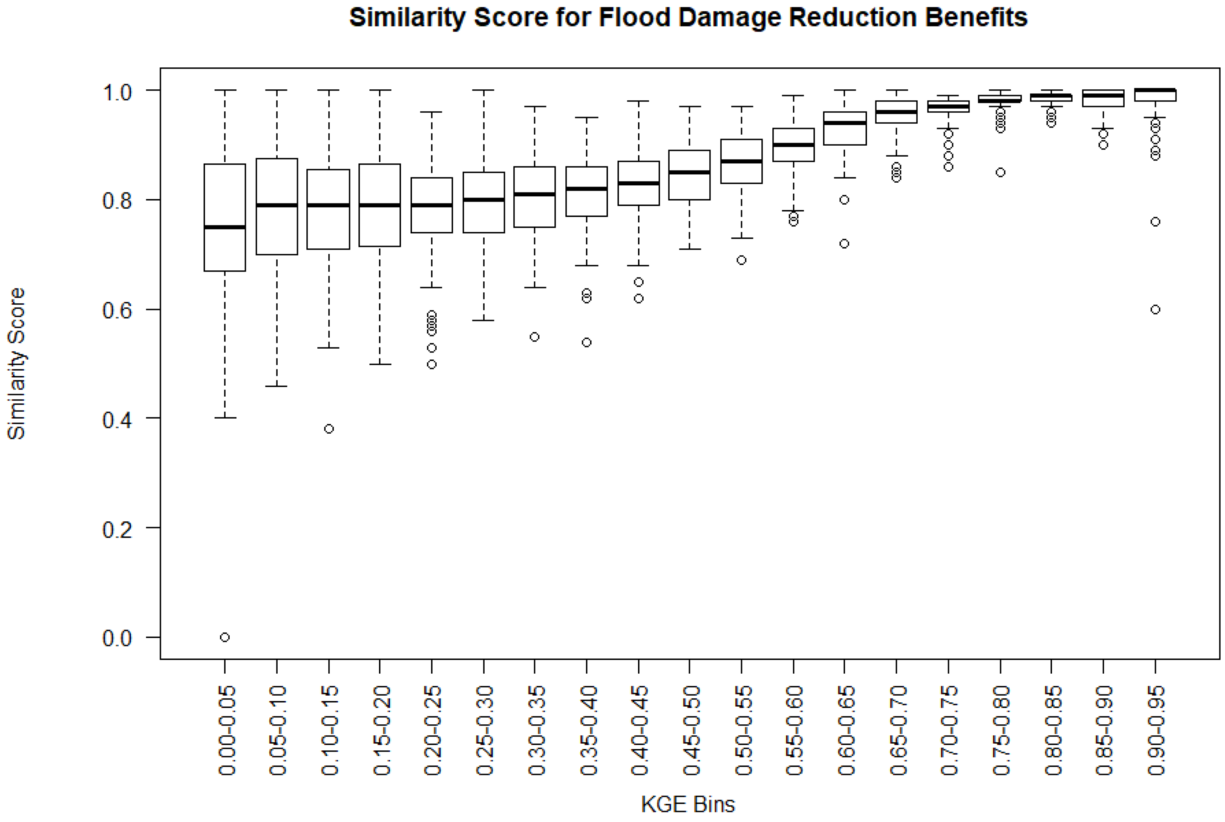
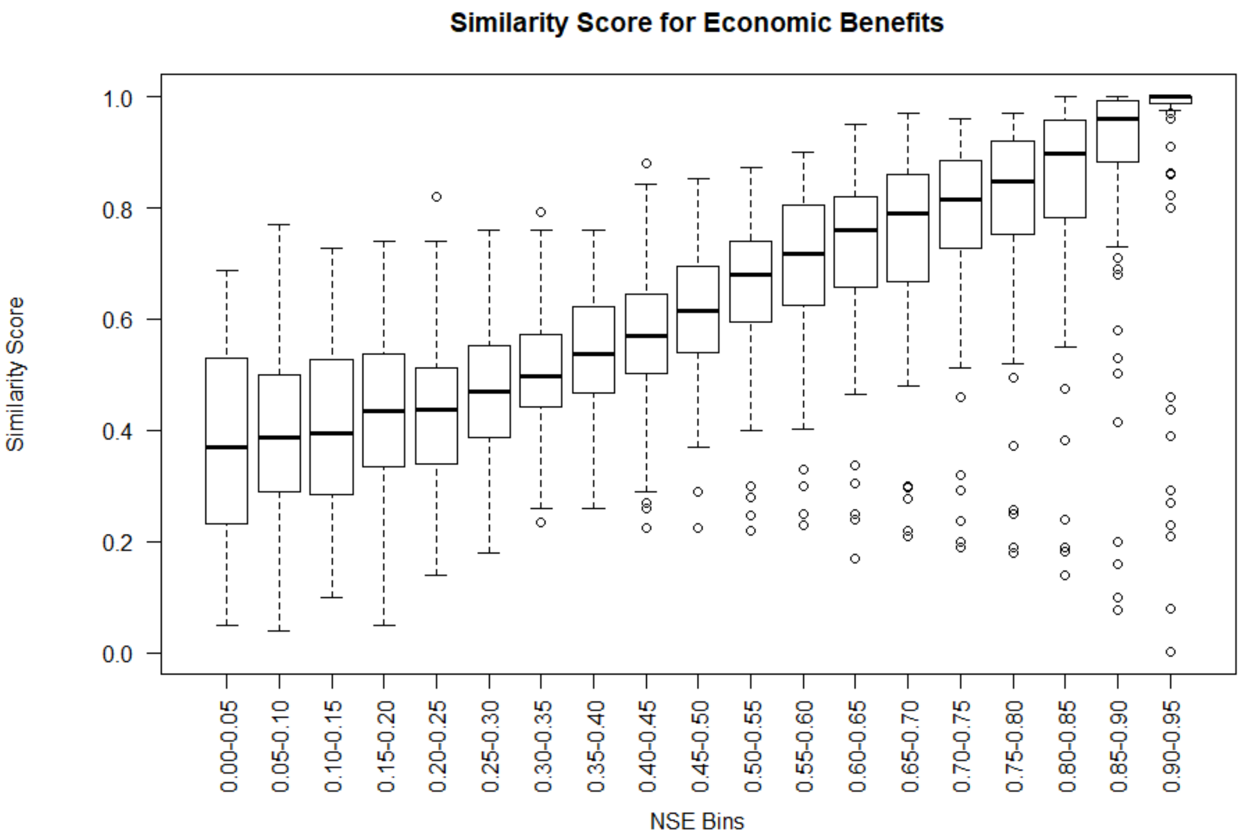
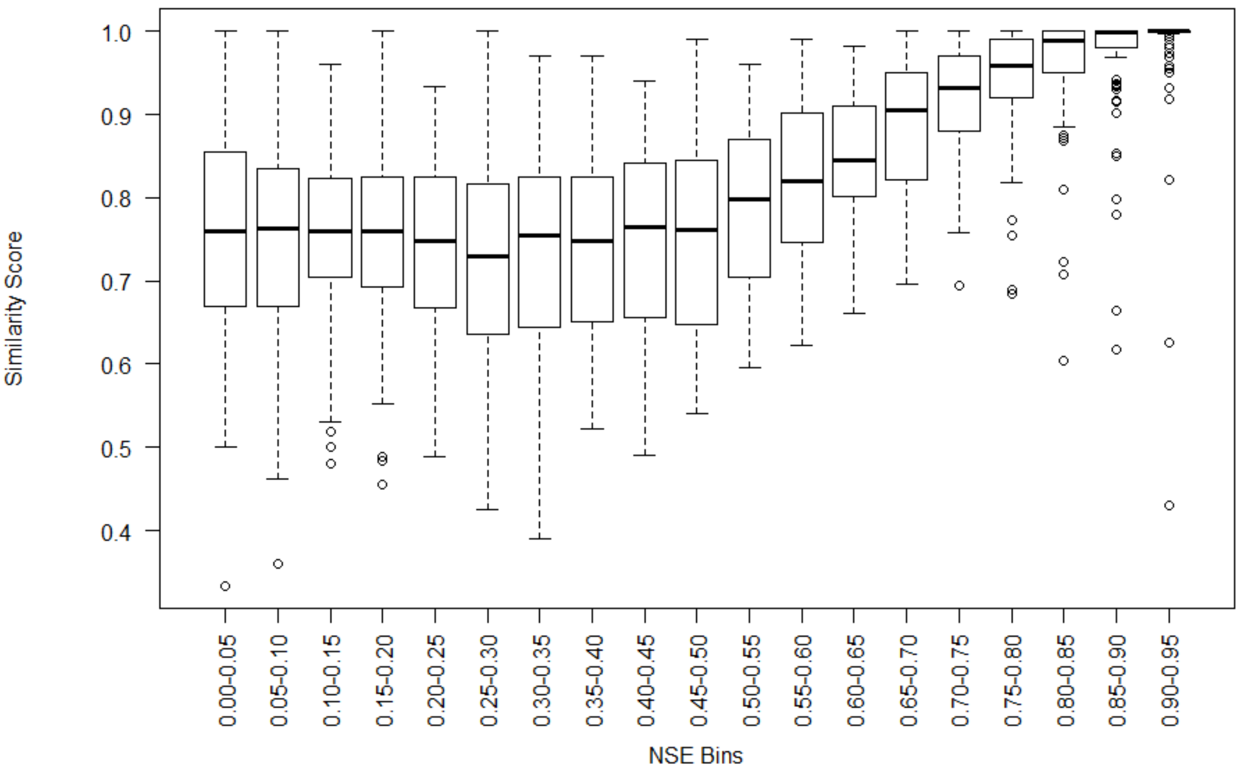

Similarity Score for Flood Damage Reduction Benefits
Similarity Score
1.0
0.8
0.6
0.4
0.2
0.0
0.00-0.05
0.05-0.10
0.10-0.15
0.15-0.20
0.20-0.25
0.25-0.30
0.30-0.35
0.35-0.40
0.40-0.45
0.45-0.50
0.50-0.55
0.55-0.60
0.60-0.65
0.65-0.70
0.70-0.75
0.75-0.80
0.80-0.85
0.85-0.90
0.90-0.95
KGE Bins

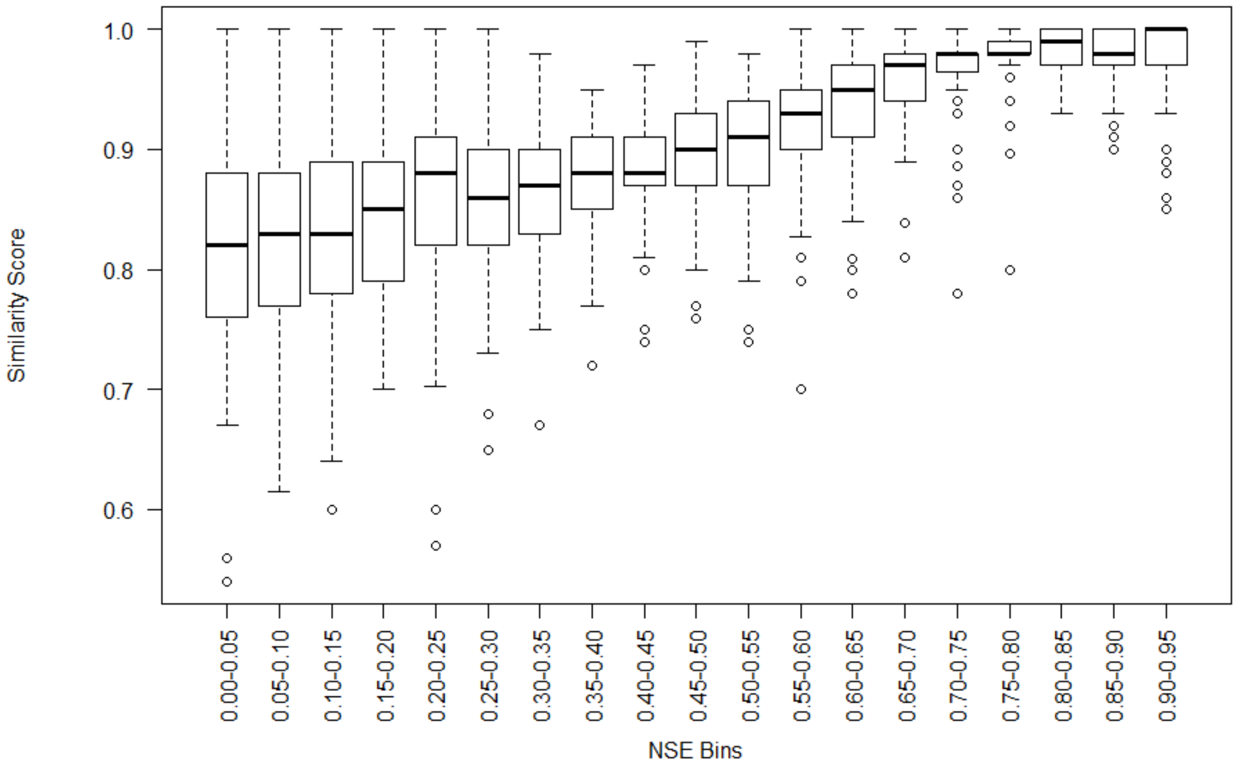
Figure C.4 – Box whisker plot of similarity score calibrated to inflow NSE penalized by PBIAS



Similarity Score for Ecological Benefits



Similarity Score for Flood Damage Reduction Benefits


Similarity Score
NSE Bins