Assessing the Relative Contributions of Input, Structural, Parameter, and Output Uncertainties to Total Uncertainty in Hydrologic Modeling

By

Scott Pokorny

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Civil Engineering

University of Manitoba

Winnipeg, Manitoba

Abstract

The simulation of physical environments by hydrologic models has become common as computational power has increased. It is well known that, to simulate the hydrology of a physical environment, simplifications of that environment are needed. The simplified versions of hydrologic processes generate uncertainty, in addition to ingesting uncertainty from input data. The uncertainty from one modeling step affects the next through propagation. Although computational power has increased through time, the computational demand for uncertainty analysis still remains a common limiting factor on the level of detail an uncertainty analysis can be conducted with. This thesis generates an estimate of total uncertainty propagated from input, structural, and parameter uncertainties for the Nelson River in the Lower Nelson River Basin near the outlet to Hudson Bay, as part of the BaySys project. Each source of uncertainty was relatively partitioned for determination of the most valuable source of uncertainty for consideration in an operational environment with a limited computational budget. The results of this thesis show the complex spatial and temporal variation present in gridded climate data. This thesis also presents an ensemble-based methodology to account for the input uncertainty associated with gridded climate data subject to propagation. The ensemble of input data was propagated through an ensemble of hydrologic models. Relative sensitivities of model parameters were shown to vary temporally and based on performance metrics, suggesting that aggregated performance metrics obscure information. Lastly, relative partitions of uncertainty were compared through cumulative distribution functions. Accounting for all sources of uncertainty appeared valuable towards improving streamflow predictability, however, structural uncertainty may be the most valuable in an operational environment with a limited computational budget followed by input, and parameter uncertainty.

Acknowledgments

Thank you to the University of Manitoba, Manitoba Hydro and partners in funding through the Natural Sciences and Engineering Research Council of Canada through funding of the BaySys project. Thank you to my committee, Drs. Stadnyk, Ali, and Déry for all their help and support. Thank you to the province of Manitoba for providing fellowship funding.

Thank you to Kristina Koenig, Mike Veira, Phil Slota, Mark Gervais, Shane Wruth, and Kevin Sagan from Manitoba Hydro who were always willing to meet and provide technical support and guidance for all aspects of my master's degree. The extra help and time taken from busy schedules was always greatly appreciated.

Thank you to all the water resources graduate students for the help, good times, and great discussions that we had, especially Joey Simoes, Andrew Tefs, Tegan Holmes, Chelsea Nguyen, Kevin Lees, and Andrew Murray. I am happy to have had the pleasure of spending my time in grad school with all of you.

A very special thank you to my mom, Joan Pokorny, my dad, Roland Pokorny, and my sister, Kelly Pokorny (whom I put up with). Thank you so much for helping and supporting me with every aspect of my time in grad school I hope to make you proud. Thanks to Lexi (my dog) for always offering hugs and love.

A much deserved thanks to Dr. Tricia Stadnyk and Dr. Genevieve Ali for their complete support and continual excellent advice. I appreciate the continued support; even though my thesis saw three great advisors and friends leave the University of Manitoba, which must be some kind of record. I

can really clear a room. A special thanks Dr. Peter Rasmussen, who was unable to see the results of my thesis, but I know would have loved it.

Table of Contents

List of Tables

List of Figures

Contributions of Co-Authors

Aid and guidance was provided by co-authors. They are noted by their contributions to the two manuscripts in chapters 3 and 4. The analysis, results, and discussions are primarily my own work.

Chapter 3: Assessment of Ensemble-Based Gridded Climate Data and Evaluation of uncertainty in hydrologic modeling arising from input data selection. (Pokorny, S., Stadnyk, T., Ali, G., Lilhare, R., Déry, S., Koenig, K.)

Dr. Stadnyk, Dr. Déry, and Kristina Koenig contributed their excellent insight to the selection of the gridded climate products utilized in the generation of the gridded climate data ensemble. Dr. Stadnyk, Dr. Ali, Dr. Déry, and Raj Lilhare were instrumental in the development and evolution of the methodology by providing a much needed insight into the preliminary results. Dr. Stadnyk and Dr. Ali provided continual guidance to the development of figures that presented results with concision. All co-authors contributed to the editing and revisions of the manuscript text.

Chapter 4: Cumulative effects of multiple uncertainty sources on flow predictability in a hydrologic modeling environment. (Pokorny, S., Stadnyk, T., Ali, G., Déry, S., Tefs, A., Holmes, T., Lilhare, R., Koenig, K.)

Andrew Tefs and Dr. MacDonald provided access and training for the operation and calibration of the HYPE model. Tegan Holmes provided access, training, and continual insight into the operation, calibration, and setup of the WATFLOOD model. Kristina Koenig and Manitoba Hydro provided access and support to the development and calibration of the HEC-HMS model. Drs. Stadnyk and Ali provided guidance on the development and revision of the methodology of the study as well as having provided continued guidance to the development of figures, which presented results with concision. All co-authors contributed to the editing and revisions of the manuscript.

Chapter 1 Introduction

Water is an essential consideration for much of Canada's economic activity, from flood forecasting and flood damage mitigation to water management and resource optimization for hydropower generation. Almost all Canadian industries are invested in the future outlook of water resources in Canada. Precipitation-runoff modeling (hereafter referred to as hydrologic modeling) offers valuable insight into water management, particularly in regions of limited data availability, such as Northern Canada (Mekis and Vincent, 2011). Hydrologic models are numerical representations of physical environments used to simulate environmental processes, including runoff and streamflow (Pechlivanidis et al., 2011). Hydrologic models are simplified representations of real environments driven by inputs simplified to match the hydrologic model spatial and temporal resolutions; these simplifications generate uncertainty at every stage in the modeling process (Ajami et al., 2007; Demargne et al., 2014). Uncertainty is defined here as the realistic range of a certain value for which an exact value cannot be determined (Uusitalo et al., 2015). There are many purposes for which a hydrologic model can be utilized; however, the model development process is similar regardless of the final application. The simulated output variable may be used to fill missing data in an existing dataset, or to produce new data where no observed product is available, such as a forecast. Uncertainty will affect the accuracy and reliability of simulations and predictions, and potentially impact the significance of statistical tests, such as trend analysis (e.g. McLeod, 2005). Understanding the effects of the various sources of uncertainty is important for defining the limits of understanding gained from hydrologic modeling and the models' usability (Dams et al., 2015). Uncertainty analysis, however, is generally considered to have an unattainably high computational demand in operational environments with limited computational budgets (Ajami et al., 2007).

Hydrologic modeling uncertainty can be considered as arising from four broad sources: input, structural, parameter, and output uncertainties (Matott et al., 2009; Pechlivanidis, 2011; Uusitalo et al., 2015). Many studies consider only the uncertainty in hydrologic model parameters (Ajami et al., 2007). Without considering all sources of uncertainty, an assumption is made that unaccounted for uncertainty, such as from input data, is negligible (Demargne et al., 2014). To facilitate more complex uncertainty assessments, frameworks have been developed to consider multiple sources of uncertainty, such as the Bayesian Total Error Analysis (BATEA) framework (Kavetski et al., 2006a, 2006b) and the Integrated Bayesian Uncertainty Estimator (IBUNE) (Ajami et al., 2007). Uncertainty in meteorological input is generally focused on the magnitude and timing of precipitation data. Other meteorological data (i.e. temperature) are often considered to be of higher quality and to have lower spatial variability (e.g. Rapaić et al., 2015; Essou et al., 2016). Therefore, it is common in the literature to consider precipitation alone to represent meteorological input uncertainty (e.g. Kavetski et al., 2006a, 2006b). Other meteorological data are sometimes also used in hydrologic simulation, such as wind data, but they are less common. Input precipitation uncertainty is usually assessed through comparison with observed benchmarks, often ground-based climate station data (e.g. Eum et al., 2014). Perturbation of meteorological inputs and variation of model structure are commonly done to sample input and structural uncertainties, respectively (Ajami et al., 2007; Tasdighi et al., 2018). Output uncertainty is associated with potential error in the observed record of the target timeseries for simulation (streamflow in this study), and is often considered by assigning subjective bounds of uncertainty such as the limits of acceptability (LOA) (Beven, 2006). Some studies go further, for instance through variation in rating curve fit, to estimate output uncertainty (McMillan et al., 2010). The various sources of uncertainty are interconnected through propagation, which refers to how uncertainty from one modeling step affects the next, moving downstream; propagation will be discussed in more detail in Chapter 4.

The type and depth of an uncertainty analysis is likely to be guided by computational resource availability. Uncertainty analysis is generally based on sampling from a likelihood distribution at some point in the modeling process. Frameworks like BATEA assign many variables for each source of uncertainty to sample from. For instance, BATEA samples the uncertainty associated with individual precipitation events by varying each event separately by assigning variables to storm events. Ajami et al., (2007) suggest this may lead to dimensionality problems. With the high volume of model runs required for nearly all uncertainty analysis methods, it is desirable to suggest which source of uncertainty is most valuable to include for the improvement of streamflow estimation, while maintaining operationally feasible computational requirements.

This study is part of the larger Hudson Bay System (BaySys) project, which is a regulation impact study on Hudson Bay (see Barber, 2014). The goal of the BaySys project is to partition the relative effects of hydropower regulation and climate change on the net export of freshwater to Hudson Bay under future regimes. The BaySys project is supported by industrial partners in the hydropower industry that are interested in furthering their understanding of climate change impacts on their system and future water supply. The Nelson River is the single largest freshwater source (by volume) to Hudson Bay (Déry et al., 2016) and therefore, it is selected as the focus for an uncertainty analysis on estimated streamflow (Barber, 2014). Understanding the impact of modeling uncertainty on our predictive capability (for streamflow) is important for this project to ascertain the practical limits of understanding that can be gained for hydropower operators from the models of the physical environment generated under the BaySys project.

This thesis considers the uncertainty about our predictions of the modeled physical environment, and the impacts that it may have on our ability to partition the relative impacts of hydropower regulation and climate forcing for the Hudson Bay system. Analyzing uncertainty in historical time periods informs

confidence in the predictive capacity of the models by allowing for the generation of uncertainty envelopes. These uncertainty envelopes provide insight into expected predictability based on the ability of the models to capture historical events within a simulated uncertainty envelope. Historical analysis contributes to an improved understanding of the relative partitions of the total uncertainty generated by the models and, therefore, reliability of estimated streamflow for Hudson Bay.

To partition the relative impacts of climate change from those of hydropower regulation on the freshwater input to Hudson Bay, climate projections that attempt to simulate future climate scenarios are used. The future climate projections are created by meteorological models referred to as global climate models (GCMs) (IPCC, 2014). Multiple realizations of projected future climate scenarios were generated for the BaySys project using varied climate model forcing, such as the amount of greenhouse gases in the atmosphere at different temporal points in the future, and various technological responses to rising greenhouse gas concentrations (Van Vuuren et al., 2011). The representation of uncertainty within BaySys will first utilize uncertainty envelopes derived from the historical period. Modeling uncertainty will then be propagated into the future using streamflow response and uncertainty (from the historic period) about that response to each future climate scenarios. Once all climate scenario responses are combined, a total future uncertainty envelope can be derived. This thesis represents the first step in this work, which is the derivation of the historical uncertainty envelopes.

1.1. Objectives

The main objective of this thesis is to make a first attempt at generating a total uncertainty estimate for the simulation of Nelson River streamflow. The main objective is achieved through three intermediate objectives: 1) exploration of the partitioned relative effects of each source of uncertainty; 2) determination of the most valuable source of uncertainty to consider for streamflow predictability improvement for an operational environment; and 3) quantification of the total uncertainty from the

Nelson River for the purpose of providing uncertainty bounds around modeled freshwater flows that will be used for the propagation of uncertainty in the BaySys project.

The first objective is achieved by selecting an ensemble of input datasets and an ensemble of hydrologic models. Sampling from the input ensemble, model structures, and model parameters is then done to assess streamflow predictability within the range of output uncertainty. The second objective is achieved by assessing the partitioned relative effects of each source of uncertainty through cumulative distribution functions (CDF) of simulated streamflow with respect to observed streamflow. The third objective is achieved by generating the full range of uncertainty produced by the uncertainty sources considered in this thesis.

Chapter 2 : Background on Hydrologic Modeling Uncertainty

This chapter reviews some of the general concepts pertaining to hydrologic models and the uncertainty they generate. Further review of individual, previously published uncertainty studies is provided in Chapters 3 and 4.

Hydrologic models are designed to represent real hydrologic processes occurring in a physical environment through mathematical equations. These equations are simplified versions of real mass and energy fluxes occurring in nature. Natural hydrological processes are complex and variable in space and time, and these processes cannot be simulated for all scales by a single hydrologic model at the level of detail they generally occur (Blöschl, 2001). Therefore, hydrologic models make broad simplifications, or conceptualizations, of hydrologic processes acting on smaller scales through spatial and temporal aggregation. By considering areas of relatively large spatial extent at time steps of one hour or longer, hydrologic models are generally able to reproduce average hydrologic conditions over a spatially aggregated area.

2.1. Input Uncertainty

Hydrologic models require meteorological inputs, at a minimum, precipitation. Hydrologic models are data intensive, often requiring high resolution meteorological input. Ground-based climate stations are often not sufficient to meet the data requirements of hydrologic models (e.g. Mekis and Vincent, 2011); ground-based climate stations are, however, considered to be the highest quality meteorological data

source (e.g. Wong et al., 2017). Gridded climate data are a commonly considered alternative when higher resolution climate data are required. There are numerous studies in the literature comparing ground-based climate station data to gridded climate data alternatives (e.g. Pavelsky and Smith, 2006; Bukovsky and Karoly, 2007; Becker et al., 2009; Choi et al., 2009; Eum et al., 2014; Rapaić et al., 2015; Kluver et al., 2016; Essou et al., 2016; Gbambie et al., 2017; Wong et al., 2017; Boluwade et al., 2018; Fortin et al., 2018). Comparisons between data sources for the purpose of inferring a measure of product quality are challenging as all data sources have their inherent uncertainties. Ground-based climate station precipitation data suffer from four main types of known uncertainty: undercatch (Goodison et al., 1998), evaporation loss (Mekis and Hogg, 1999), wetting loss (Yang et al., 1998), and trace precipitation (Yang et al., 1998). These uncertainties in the measurement of precipitation are accompanied by measurement accuracy limitation associated with the type of precipitation gauge used (Michaelides et al., 2009). An important distinction is the difference between error and uncertainty: gauge measurements that are clearly incorrect due to equipment malfunction or other such measurement challenges are errors. Adjustment of uncertainties, such as undercatch (Mekis and Vincent, 2011), cannot be considered corrections, as correction would require the true event magnitude to be known. Uncertainty in gauge measurements can only be reduced, but never eliminated. The uncertainty can be reduced to a more realistic range using reference gauges (e.g. Sevruk et al., 2009) or metadata-based adjustments, such as utilizing notes about station maintenance occurrence (Mekis and Vincent, 2011). Mekis and Vincent (2011) also utilized other relevant data to correct for systematic biases, such as utilizing snow on ground measurements to aid in the development of undercatch adjustment factors for solid precipitation. The inclusion of other types of data used by Mekis and Vincent (2011) narrows the uncertain range by providing guidance as to what can be considered a realistic measurement.

Gridded climate data products are generated through a combination of spatial and temporal interpolation, and atmospheric modeling. Products involving the ingestion of observed data for atmospheric modeling are termed reanalysis products (e.g. Mesinger et al., 2006). Similar to hydrologic models, reanalysis products rely on climate models that simplify reality, which, therefore, introduces uncertainty. The existing literature often focuses on the attempted determination of a single best product (e.g. Eum et al., 2014; Wong et al., 2017) but generally does not fully explore the effect that subjective decision making during product evaluation has on the representation of uncertainty (Rapaić et al., 2015). Uncertainty from the evaluation process is further explored and discussed in Chapter 3.

Hydrologic models require meteorological input at concomitant spatial and temporal resolution. This often requires inputs to be spatially interpolated or spatially aggregated, and temporally aggregated or disaggregated. The effect of, or uncertainty introduced by, spatial interpolation and spatial aggregation is often not discussed. However, some studies, such as Wong et al., (2017), consider this by limiting the grid size for selected datasets. Tustison et al., (2001) showed how error can be lowered through interpolation to increase proximity to a comparative point of observation, suggesting that interpolation may shift uncertainty but not necessarily change its magnitude when generating data at point locations. Gridded products on the same grid will have the same representativeness error, meaning that if the comparison is between the gridded datasets, the same error will be present in all datasets. Therefore, representativeness error is considered to cancel if the comparison is only between gridded products.

Regardless of the source of meteorological input, uncertainty will be introduced to the model ingesting the inputs. The study of uncertainty in meteorological data input is usually focused on the choice of one (historical) gridded dataset, such as Mei et al., (2016) and Nikolopoulos et al., (2010), who evaluated the effects of satellite and radar input data uncertainty, respectfully. Li et al., (2018) used a multiplier with varied mean and standard deviations for input data to show how input uncertainty must be accounted

for to correctly determine the range of parameter uncertainty. Ndiritu (2013) explored a method for estimating uncertainty in ground-based climate station data by iteratively excluding sets of climate stations. Frameworks like BATEA (Kavetski et al., 2006a and 2006b) and IBUNE (Ajami et al., 2007) utilize storm multipliers, with BATEA being more computationally intensive by assigning individual multipliers to each storm event.

The utilization of meteorological input ensembles is more common in climate change studies than in historical studies. It is assumed that future climate projections are inherently more uncertain than the simulations of historic climate generated by reanalysis products such as the North American Regional Reanalysis (NARR) product (Mesinger et al., 2006). Some studies compare global climate models (GCMs) to historic climate datasets from a performance perspective, such as Wong et al., (2017). Others studies use historic climate datasets as a benchmark for comparison (e.g. Sillmann et al., 2013). The high uncertainty associated with climate change projections suggests that projection of a historical uncertainty envelope assuming stationarity is not possible. Studies such as Dams et al., (2015) show the relative dominance of climate change projection uncertainty for the contribution towards total uncertainty. The fundamental idea of climate change precludes non-stationarity in climate data, which therefore promotes the use of ensembles to represent the range of uncertainty in future projections. Historical data, however, also come with uncertainty, which is less often studied (e.g. Sillmann et al., 2013). For example, Wi et al., (2015) used a 36 GCM ensemble but did not utilize a comparative historical meteorological ensemble.

2.2. Parameter Uncertainty

Parameter uncertainty is generated by the simplification of reality. Hydrologic processes are complex in nature but are represented by simplified algorithms. These algorithms require parameter estimation: however, the spatial aggregation underlying the model structure requires parameters to be spatially

averaged estimates. Some models tie parameters to physical characteristics of an environment, such as the Soil and Water Assessment Tool (SWAT) model (Neitsch et al., 2011) with its estimation of Green & Ampt infiltration parameters. The most popular method for evaluating parameter uncertainty is the Generalized Likelihood Uncertainty Estimation (GLUE) method (Beven and Binley, 1992). GLUE implementations sample parameters in search of parameter combinations that produce a "behavioral" simulation. A behavioral hydrologic model simulation – or realization – is a particular model structure and parameterization capable of producing output sufficiently similar to a selected observed record with a given input. Sufficient model-to-observed record similarity is determined by an often subjectively set performance metric threshold (e.g. Shafii et al., 2015). The principle of equifinality suggests that there exists a high number of parameter combinations capable of meeting a particular set of performance criteria; therefore, parameters are more realistically represented by a likelihood rather than a specific value. The subjectivity of model performance and alternatives to subjective behavioral model simulation selection criteria will be further discussed in Chapter 4.

2.3. Structural Uncertainty

Spatial and temporal discretization, as well as the hydrologic methods selected to estimate real hydrologic processes, cumulatively represent the hydrologic model structure. Spatial discretization exists on a continuum: models are classified as either lumped or gridded (USACE, 2016; Kouwen, 2018). Lumped models may be fully lumped, meaning a desired physical environment is represented by a single grid cell. Semi-lumped models discretize a physical environment into multiple sub-basins, each delineated by a drainage divide (USACE, 2016). Gridded models may be semi-distributed, meaning that grid cells utilize grouped response units (Kouwen, 2018). Fully distributed models account for the position of physical features like land class locations, not just as a percentage within a grid cell. Spatial discretization may be developed based on available data or other limitations such as runtime

considerations. The hydrologic methods selected to estimate physically-based hydrologic processes likely require parameter estimation, which represents the average characteristics across a spatially discretized region. Parameters may also be empirical and have no physical meaning, requiring optimization to ascertain their values. Therefore, parameter and structural uncertainty are implicitly tied through the spatial aggregation of the model structure.

There is no single best structure for representing a physical environment, and moreover, many structures may be sufficient to produce a behavioral simulation. By simplifying a physical environment, the numerical representation of that physical environment becomes computationally feasible but introduces uncertainty. Model structural uncertainty is often considered through the selection of multiple hydrologic models; however, some studies consider variations of model structure within a single hydrologic model framework (e.g. Wi et al., 2015; Muhammad et al., 2019). Ajami et al., (2007) used Bayesian model averaging to generate uncertainty limits on estimated flow; however, many ensemble-based methods exist. Wilks (2006) provides a summary of some methods commonly used in the literature. Dams et al., (2015) utilized the range of monthly runoff from multiple models to examine the relative contributions of uncertainty by model structure and input. The use of hydrologic model ensembles is becoming more common in the literature (e.g. Karlsson et al., 2016). Multiple models are often utilized to vary the simplifying assumptions made by a hydrologic model and determine the effect imposed by the assumptions. Shafii et al., (2015) used multiple models to compare behavioral selection criteria; the use of multiple models allowed for a less subjective comparison with the goal of generating a comparison less reliant on a single model's assumptions. Studies like Wi et al., (2015) and Muhammad et al., (2019) varied the spatial discretization of model structure in an attempt to quantify the impacts of assuming a set spatial aggregation structure. Studies like Tasdighi et al., (2018) vary the hydrologic process representation in a single model to quantify the performance impact of hydrologic process representation.

2.4. Output Uncertainty

Output uncertainty is defined as the uncertainty associated with the measurement of observed data, which are streamflow data in this study. Streamflow data are estimated using a rating curve, which relates water depth to streamflow, through an empirically derived function developed using multiple, site-specific field measurements. Rating curves are ideally generated at straight, morphologically stable river reaches; however, real environments rarely meet these criteria and, therefore, are subject to measurement error (e.g. McMillan et al., 2010). In reality, actual rating curves are more typically looped, with a lower storage (or stage) for the rising limb of the hydrograph, and higher storage (stage) for the same discharge on the falling limb (Dingman, 2015); the looped shaped is referred to as the hysteresis effect (Braca, 2008). This looped shape deviates from operationalized rating curves that give discharge as a unique function of stage, therefore any deviation from idealized rating curves adds uncertainty to flow estimation. The closer to the desired straight and stable channel criteria a location of measurement is, the narrower the rating curve loop, and the lower the uncertainty associated with measurements at that location. Dingman (2015) suggests uncertainty as high as $\pm$ 10%; however, the Water Survey of Canada (WSC) suggests $\pm$ 5% for WSC gauges (Environment Canada, 1980). A rating curve is no longer valid for ice-on conditions, as the presence of ice introduces changes to reach shape and roughness; ice jamming may also introduce backwater effects. Reliable and reusable ice-affected rating curves cannot be developed, as the physical characteristics of an ice cover evolve through the ice-on period. Ice covers in different years are unlikely to be similar enough to generate a reusable rating curve. The increase in uncertainty is highly variable based on the characteristics of the ice cover, but will be higher than ice-off conditions (Environment Canada, 1980). Output uncertainty is further discussed in Chapter 4.

2.5. Uncertainty Propagation

Input data uncertainty is connected to overall modeling uncertainty through propagation, as is each source of uncertainty. Propagation is the transfer of uncertainty from one modeling step to the next, and the effect of that introduced uncertainty on the selected output variables (e.g. streamflow). Few studies focus on the cumulative effects of propagation, such as Demargne et al., (2014). Demargne et al., (2014) generated hydrologic forecasts, which accounted for the propagation of uncertainty. Multiple sources of uncertainty were sampled for propagation; for example, multiple meteorological forecast datasets were ingested after they were bias corrected for ensemble generation. Few studies consider the propagation of uncertainty due to the high computational demand (e.g. Ajami et al., 2007), as well as the data intensive nature of uncertainty propagation. The multiple sources of forecasted meteorological data utilized by Demargne et al., (2014) for the generation of ensemble realizations required the management of large volumes of data, added processing time, and increased data processing complexity. Similar challenges exist for each source of uncertainty considered. A hydrologic forecast, like a historical simulation, cannot produce perfect results. Accounting for uncertainty propagated from each major source of uncertainty no longer makes the implicit assumption that data are near truth or that the model used is accurate. Multiple realizations are generated and can be represented as a distribution. This increases the probability that events will be captured by modeling uncertainty envelopes (Demargne et al., 2014).

The process of calibration varies model parameters with the goal of finding parameter sets that produce behavioral simulations. Output uncertainty means the observed record that performance is quantified with is simply a realization considered to be of high likelihood (e.g. McMillan et al., 2010). It may be mathematically plausible to produce a near perfect representation of an observed record; however, once a calibration produces output within the uncertainty limits of the observed data, increasing performance loses meaning. Values within the range of output uncertainty are considered realistic

realizations of unequal likelihood (Beven and Binley, 1992). Uncertainty in the simulated variable will exceed that of output uncertainty, as uncertainty generated in the modeling process propagates to the simulated variable (e.g. McMillan et al., 2010). The relevancy of propagation in the literature is further discussed in Chapter 4.

The focus of this thesis is hydrologic modeling uncertainty targeting the output of simulated streamflow. The goal of BaySys is to partition the effects of regulation and climate change on freshwater input to Hudson Bay (Barber, 2014). A diverse set of models are utilized in the BaySys project, for example the hydrologic models in this study in addition to ocean models, water quality models and biogeochemical models, among others. All models used ingest, generate, and propagate uncertainty. The hydrologic models in this study produce simulations of freshwater streamflow, which is used as input to all other models within the BaySys project. Therefore the propagation of uncertainty in streamflow in this study refers to uncertainty propagated through the climate scenarios and into the freshwater system, representing total uncertainty about streamflow prediction; which also represents the first tier of uncertainty in the modeling component of BaySys. The selected study region is the Lower Nelson River Basin (LNRB) where the Nelson River discharges into Hudson Bay. The Nelson River is the largest river by streamflow magnitude contributing to Hudson Bay (Déry et al., 2016). It is assumed that other rivers contributing freshwater to Hudson Bay will behave within the total uncertainty for the largest freshwater source to Husdon Bay (i.e. the Nelson River). The uncertainty estimate generated by this study will then be used to provide streamflow with uncertainty bounds as input uncertainty propagation through other BaySys project models.

2.6. Gaps in the Current Literature

There are many gaps in the existing literature within the topic of uncertainty, four of which are highlighted here. The current literature largely does not address the spatial and temporal variation in

gridded climate data and how it affects uncertainty. Studies often utilize a single spatial aggregation for the comparison of gridded climate data products within long temporal periods for analysis (e.g. Wong et al., 2017). This leaves a gap in the knowledge of climate data input uncertainty subject to propagation within hydrologic models of various spatial and temporal aggregations. The present thesis performs analysis of gridded climate data in moving temporal windows at multiple spatial aggregations to improve understanding of the temporally dependent nature of gridded climate data uncertainty. The use of multiple spatial aggregations informs on the effect that spatial aggregation has on model structural development. This literature gap is addressed in Chapter 3 as part of the first objective of this thesis.

The existing literature shows dichotomy in gridded climate data studies. Studies either suggest that a single best product could be determined with sufficient analysis (e.g. Wong et al., 2017), or that multiple products should be selected (e.g. Gbambie et al., 2017). Few studies investigate the information gained from selecting an ensemble of gridded climate data products from the perspective of uncertainty (e.g. Dams et al., 2015). Studies generally do not find a single best gridded climate data product, although, some studies used ensemble techniques to generate an ensemble realization which performed best (e.g. Yao et al., 2014). The suggested treatment of gridded climate data ensembles for the representation of uncertainty in hydrologic modeling remains a gap in the literature. The present thesis examines how an ensemble of gridded climate datasets can be used to estimate the total uncertainty subject to propagation in a hydrologic model. This literature gap is addressed in Chapter 3 as part of the first objective of this thesis.

The cumulative effect of uncertainty propagation on streamflow predictability represents another gap in the literature. Few studies in the current literature include all four broad sources of uncertainty. Even in studies that analyze all four sources of uncertainty, results are usually presented as either a total

representation of uncertainty (e.g. Demargne et al., 2014), or a relative partition for the purpose of determining a dominant source of uncertainty (e.g. Dams et al., 2015). The present thesis generates relative partitions for each source of uncertainty to determine the most valuable source of uncertainty to consider towards improving streamflow predictability. This literature gap is addressed in Chapter 4 as part of the second objective of this thesis. To address this gap, the results of the input data uncertainty analysis from Chapter 3 are utilized to provide the range of input uncertainty subject to propagation. Conclusions drawn from the analysis in Chapter 4 are reliant on the results of Chapter 3.

Lastly, the existing literature lacks long temporal scale applications (e.g. 30 year climatic period or longer) of uncertainty propagation in hydrologic modeling. Studies generally focus on smaller unregulated basins (e.g. near the regional scale of 1000km^2 (Blöschl and Sivapalan, 1995)) over short temporal periods (e.g. usually less than 10 years (Nikolopoulos et al., 2010)) and rarely include gauges effected by regulation. The present thesis addresses this gap by generating a total uncertainty estimate for the Nelson River and comparing it with the smaller unregulated total uncertainty estimates for gauges within the LNRB for the 1981-2010 period. This gap is addressed in Chapter 4 as part of the third objective of this thesis. The generation of total uncertainty bounds for the Nelson River is reliant on the results generated in Chapters 3 and 4.

Chapter 3 : Assessment of Ensemble-Based Gridded Climate Data and Evaluation of Uncertainty in Hydrologic Modeling Arising from Input Data Selection.

(Pokorny, S., Stadnyk, T., Ali, G., Lilhare, R., Déry, S., Koenig, K.)

3.1. Abstract

The spatial and temporal performance of an ensemble of five gridded climate datasets (precipitation and temperature) (NARR, ERA-Interim, WFDEI, GFD-Hydro, and NRCAN's ANUSPLIN) was evaluated over the Nelson Churchill River Basin for annual and seasonal moving time windows in the 1981-2010 period. Only results for precipitation were presented as temperature data were found to be of higher relative quality, suggesting lower uncertainty. Evaluation of performance was based on comparison to Environment and Climate Change Canada (ECCC) and Adjusted and Homogenized Climate Change Data (AHCCD) ground-based climate stations at three spatial discretizations: distributed, major basin aggregations, and full watershed aggregation. All gridded datasets showed spatial variation, most notable in year to year total precipitation bias, with no clear best performing dataset with respect to observed ground-based climate records. The gridded datasets generally showed the worst performance in the representation of extreme events. The absolute minimum, mean, and absolute maximum ensemble realizations were generated from an ensemble that included all five gridded climate datasets. Ensemble realizations performed similarly in quantitative metrics but with consistently high negative and positive year to year total precipitation bias for the ensemble minimum and maximum, respectively. Minimum and maximum realizations were assumed to represent uncertainty boundaries of the ensemble to quantitatively measure the uncertainty envelope overlap with observed events for the selected spatial resolutions. The uncertainty analysis showed that high magnitude precipitation events were often outside the uncertainty envelope. Increasing spatial aggregation (i.e. major basin to full watershed), however, reduced the occurrence of extreme events and suggested lower uncertainty was introduced into the hydrologic model.

3.2. Introduction

High-quality climate data are essential for accurate representation of physical environments by hydrological models. In Canada, mountainous and northern regions suffer from data sparsity and paucity issues, with significant gaps in existing long-term records (e.g. Price et al., 2000; Mekis and Vincent, 2011; Eum et al., 2014; Wong et al., 2017). This makes observed ground-based climate station data inadequate for data-intensive applications (e.g. forcing hydrologic models), establishing a need for reliable alternatives (e.g. gridded climate data) and for estimates of the uncertainty associated with those alternatives. Uncertainty is defined here as the range of realistic values that a climate variable can be considered to represent observed conditions (ISO and OIML, 1995). In hydrologic modeling, while uncertainty can be introduced through input data, model structure, parameters, and output data (Matott et al., 2009), input climate data uncertainty is less often studied; instead, input uncertainty is used as a data selection criterion (e.g. Eum et al., 2014). Estimating the uncertainty subject to propagation within a hydrologic model is important for estimating the total uncertainty bounds of hydrologic model simulated streamflow; many studies actually find input uncertainty to be the largest source of uncertainty contributing to total uncertainty (e.g. Dams et al., 2015). Propagation is defined as the transfer of uncertainty from one modeling step to the next, and how the cumulative effects of uncertainty impacts model output. Therefore, to produce the total uncertainty estimate required for the BaySys project, the current study (Chapter 3) focuses on gridded data uncertainty for hydrologic modeling. This study also addresses several challenges that have been identified in the literature with regards to performance metrics, data product selection, spatial aggregation, and temporal period selection.

Gridded climate data are used for various applications; with many options available, assessing performance is a significant issue (e.g. Pavelsky and Smith, 2006; Bukovsky and Karoly, 2007; Becker et al., 2009; Choi et al., 2009; Eum et al., 2014; Rapaić et al., 2015; Kluver et al., 2016; Essou et al., 2016; Gbambie et al., 2017; Wong et al., 2017; Boluwade et al., 2018; Fortin et al., 2018). Performance metrics are used to compare a data product to observed conditions; the most common types of performance metrics are continuous statistics (e.g. root mean square error (RMSE)) (e.g. Wong et al., 2017), categorical statistics (e.g. equitable threat score (ETS)) (e.g. Lespinas et al., 2015), extreme analysis (e.g. comparison of one-day maximum precipitation (RX1) between observed and simulated datasets) (e.g. Asong et al., 2017), and proxy validation by hydrologic models (e.g. Gbambie et al., 2017), among others. Continuous statistics measure agreement between simulated and observed timeseries. Categorical statistics measure agreement between simulated and observed (precipitation) events binned by magnitude. Extreme analysis compares the occurrence and magnitude of extreme events between simulated and observed timeseries. Proxy validation applies climate data as input forcing to a hydrologic model; simulated flows are then compared to observed timeseries and scored using statistics such as the Nash Sutcliffe Efficiency score (NSE score) (Nash and Sutcliffe, 1970), and inferences are subsequently made about the quality of the climate data. The observed data used for comparison are often ground-based climate station data as they are often the highest-quality observed records (Mekis and Vincent, 2011; Vincent et al., 2012). Comprehensive assessment should consider three broad aspects of performance: timing, magnitude, and extremes occurrence. Performance metrics are often used in the literature to suggest an optimal dataset; yet performance metric selection is often limited to only two of the three desired performance aspects, therefore limiting result interpretations. For example, Wong et al., (2017) concluded that additional analysis of extreme event representation by gridded climate datasets was needed before an optimal dataset could be suggested based on their analysis.

Regarding data product selection, it is common in climate change studies to select multiple global climate models (GCMs) to represent uncertainty (e.g. Masson and Knutti, 2011; Knutti et al., 2013; Dams et al., 2015; Sanderson et al., 2015; Leduc et al., 2016); the selected datasets are referred to as an ensemble. Historical studies, however, rarely select more than one data product, which misrepresents historical climate data uncertainty – if it is considered at all. The reason meteorological input data ensembles are more common in climate change studies than in historical studies is the widely agreed upon assumption that future climate projections have higher uncertainty than estimates of historic climate. The inherent future variability in climate change suggests that historical uncertainty envelopes cannot simply be projected into future periods by assuming stationarity (e.g. Sillmann et al., 2013; Dams et al., 2015). This does not, however, suggest that historical climate data are free of uncertainty, as many studies found input uncertainty to be a dominant source contributing to overall uncertainty (e.g. Wi et al., 2015). An increasing number of gridded climate datasets covering various spatial extents have been developed in the past decade, not only across Canada (e.g. Hutchinson et al., 2009; Fortin et al., 2018), but across North America (e.g. Mesinger et al., 2006) and globally (e.g. Berg et al., 2018). Gridded climate data comparisons generally focus on precipitation due to the tendency toward higher uncertainty in estimation of both occurrence and magnitude, relative to temperature (e.g. Wong et al., 2017). Results vary depending on the spatial and temporal extent of a study, and the performance metrics used. Uncertainty in station and gridded data products is often discussed, owing to the paucity and sparsity of observations, but remains difficult to quantify and, therefore, many studies do not offer a clear choice of a single best product, instead concluding to proceed with caution (e.g. Gbambie et al., 2017). Bukovsky and Karoly (2007) compared a suite of gridded climate datasets for the period of 1991-2000 and found the North American Regional Reanalysis (NARR) to perform better over the continental U.S. but to have more uncertainty outside the U.S. due to lower observation network density. Becker et al., (2009) further examined NARR over the continental U.S and found systematic bias towards

overestimation of light precipitation, and slight underestimation of extremes. Essou et al., (2016) compared the hydrologic performance of NARR with other commonly used gridded products, including ERA-Interim, WATCH Forcing Data Era-Interim (WFDEI), and others: NARR generally produced higher NSE scores through proxy validation and WFDEI performed best of the global products. Choi et al., (2009) assessed NARR for hydrologic model forcing viability in Manitoba and found, similar to Becker et al., (2009), more uncertainty outside the U.S. Eum et al., (2014) compared NARR, Natural Resources Canada's (NRCAN) ANUSPLIN, and the Canadian Precipitation Analysis (CaPA) to three climate stations in the Athabasca River Basin. NARR was found to have a statistical break that occurred in January 2004, coinciding with the time when Canadian climate station data were no longer assimilated into NARR. Wong et al., (2017) compared a suite of gridded datasets in ecodistricts and found WFDEI and CaPA to perform well, but both showed spatial variability. Rapaić et al., (2015) compared a large suite of observation-based and reanalysis datasets for the Canadian Arctic and found no dataset was "best"; they instead concluded multiple datasets should be considered. Gbambie et al., (2017) compared gridded datasets that involved interpolation of climate station data in Quebec and similarly concluded that multiple datasets should be used, as each contained useful information.

Studies often focus on gridded data performance at a set spatial aggregation (the spatial averaging of multiple grid points) for varied temporal periods, spanning many years. Gridded data are generally aggregated spatially to simplify the comparison with climate station data (e.g. Wong et al., 2017). Aggregation in other studies has generally been dictated by the spatial discretization scheme used by a single hydrological model structure, i.e. distributed, semi-distributed, or lumped; Khakbaz et al., (2012) summarize some notable differences in these structures. Performance assessment is, therefore, limited because a set aggregation leaves a gap in our knowledge of how input uncertainty propagates into hydrologic models of varied spatial structures. Spatial variability is a common factor complicating dataset intercomparison (e.g. Becker et al., 2009; Eum et al., 2014; Rapaić et al., 2015; Essou et al., 2016;

Gbambie et al., 2017; Wong et al., 2017; Fortin et al., 2018). Mekis and Vincent (2011) and Wong et al., (2017) summarized the limitations associated with meteorological station records, and Choi et al., (2009) state that due to data sparsity, spatially aggregated comparisons may create misleading results due to varied resolutions. Therefore, regardless of aggregation, perfect performance metric values are not an attainable goal since station data also has uncertainty.

Understanding the uncertainty introduced into hydrological modeling by gridded climate data, and how to better account for this uncertainty, remains a major gap in the literature. The goal of this study is to assess input data uncertainty in hydrological modeling via the dataset selection, temporal period of analysis, and spatial aggregation of the model. Multiple datasets are selected to generate an ensemble to assess representation of station climate data uncertainty. Results are presented at multiple spatial aggregations (lumped, semi-lumped, and distributed) to better reflect how spatial aggregation affects uncertainty. The goal is to estimate and propagate total uncertainty generated by choice of input data onto simulated flows for the Nelson River. It is recognized that, in operational contexts, this represents an unrealistic range of uncertainty considered, or needing to be considered, for hydrologic analysis and infrastructure design.

3.3. Study Area

The Nelson Churchill Watershed (NCW) covers approximately 1.4 million km^2 of the North American landmass and can be broken into several major drainage basins (NRCAN, 2018; Figure 1a).

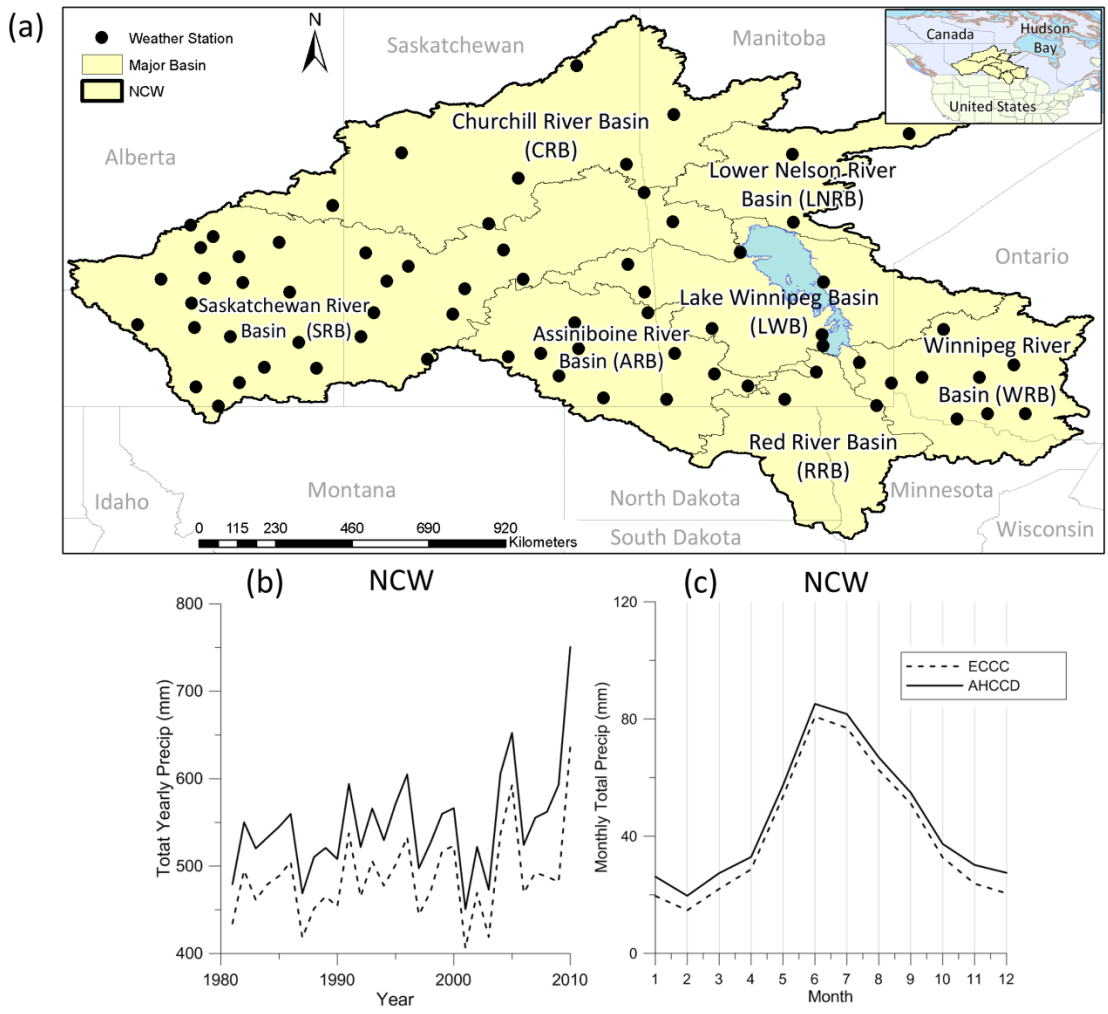

Figure 1: (a) Map of the Nelson Churchill Watershed including major basin delineations and 71 selected observed station locations. (b) Spatially averaged yearly total precipitation timeseries including all stations shown in (a). (c) Period mean monthly precipitation including all stations shown in (a) for the period 1981-2010.

It covers a diverse set of ecodistricts and climatic regions, which generally increases the uncertainty in gridded climate data. The NCW includes low relief prairies in southern Manitoba, Saskatchewan, and Alberta; a mountainous region on the western edge of the Saskatchewan River Basin (SRB), data-sparse cold climate northern regions (i.e., Sub-Arctic, and the Northwestern forest), water bodies (e.g. Lake Winnipeg, and wetlands), and the U.S.-Canada border; each introducing unique climate data challenges. Data assimilation challenges specifically occur at the US-Canada border region resulting from changes in

observation station density (Bukovsky and Karoly, 2007). For an in-depth description of the NCW, see Benke and Cushing (2011).

3.4. Climate Data Sets

The climate variables included in this study are daily precipitation (Pr), daily minimum temperature (Tmin), daily mean temperature (Tmean), and daily maximum temperature (Tmax).

3.4.1. Observed Station Data

Two sets of ground-based climate station data were selected for comparison to gridded climate data. The first are near real-time daily unadjusted Environment and Climate Change Canada (ECCC) observations. The second set of ground-based climate station data are the daily Adjusted and Homogenized Climate Change Data (AHCCD) (Mekis and Vincent, 2011; Vincent et al., 2012). AHCCD use data retrieved from the National Climate Data Archive of Environment and Climate Change Canada: 464 ground-based precipitation timeseries and 338 ground-based temperature timeseries. Data have been adjusted for common precipitation measurement issues such as wind undercatch, evaporation, wetting losses, and trace precipitation. The adjustments used in the AHCCD dataset generally added water to reduce the effects of known biases in ECCC data, such as wind undercatch, that reduced recorded precipitation. Inconsistent methods for the handling of trace precipitation through time and between stations created inhomogeneities in ECCC data. Therefore, AHCCD utilized varied trace precipitation adjustments based on available metadata to better homogenize the data; however, Mekis and Vincent (2011) state that it is possible inhomogeneities likely still exist. For temperature, data are adjusted for changes in observation time – ensuring values are associated with the day they occurred – and discontinuities – such as non-climatic shifts. Mekis and Vincent (2011) state that significant uncertainty

remains in these data for extremes and high spatial and temporal variability in the snow water equivalent adjustment, therefore, they are not recommended for short-term timeseries applications or for extreme events, such as blizzards. AHCCD is used by most comparison studies focused on climate data over Canada, as it is considered to be the best representation of observed climate (e.g. Wong et al., 2017). In the current thesis chapter, all comparisons are made to AHCCD unless stated as different (i.e. AHCCD is used as the station input for performance metrics). AHCCD is the preferred choice as it is expected to have less error and similar uncertainty to ECCC data.

In this study, station selection was based on climate variable availability for both AHCCD and ECCC data for data quality control and assurance reasons. Stations were selected that had minimal missing data in both the AHCCD and ECCC station records for a particular location. A total of 71 stations were selected (Figure 1a): an average of all 71 stations from 1981-2010 across the NCW domain was used to produce the average yearly total precipitation timeseries (Figure 1b) and monthly average annual precipitation (Figure 1c). AHCCD generally has higher precipitation values than ECCC, with the largest difference among products often associated with solid precipitation events (Figure 1c), an observation that is consistent with the findings of Mekis and Vincent (2011). We only show precipitation data, as differences in temperature among data products are much smaller and generally insignificant (figure not shown). Station data were not included for the U.S. portion of the watershed due to differences in data collection and processing techniques; similarly, Lake Winnipeg buoy data was also excluded for the same reasons. To prevent the introduction of unknown changes to station data uncertainty, only Canadian land based stations were considered. The reduction in spatial coverage (e.g. in the RRB) will, however, also increase uncertainty. This was preferable since the uncertainty is increase due to a known reduction in ground based climate station coverage. The U.S. portions of the SRB, ARB, and WRB were similar in size or smaller than ungauged areas on the Canadian side of those basins, therefore the effect of excluding U.S. gauges is assumed to only affect the RRB.

3.4.2. Gridded Data Products

A total of five gridded climate datasets were selected. The criteria for gridded dataset selection included the selection of products that have been suggested in the literature as high-performing, have data availability from 1981 to 2010, and have a daily temporal resolution or finer.

Table 1: Main characteristics of the five gridded climate datasets selected for the current study. Variables presented include daily precipitation (Pr), daily minimum temperature (Tmin), daily mean temperature (Tmean), and daily maximum temperature (Tmax).

Name	Period (Temporal resolution)	Domain (Spatial resolution)	Variables	Reference	Product Description
ANUSPLIN	1950-2013 (daily)	Canada (~0.1°)	Pr, Tmin, Tmax	Hutchinson et al., (2009)	Interpolated ECCC dataset using trivariate thin-plate smoothing spline between latitude, longitude, and elevation. The version updated to cover 1950-2013 was used although a version extending up to 2016 was released after the completion of this study.
North American Regional Reanalysis (NARR)	1979- ~present (3 hourly)	North America (~0.32°)	Pr, Tmin, Tmean, Tmax	Mesinger et al., (2006)	A reanalysis dataset with many sources of assimilated data, such as the global reanalysis product GR2, gauge observations, and others. NARR stopped assimilating Canadian station data in 2004, which introduced a detectable statistical break (Eum et al., 2014). In 2015 the period of April 2009-January 2015 (and thereafter) was updated to address some data processing issues, which improved border effects along US-Canada international border particularly focused on southern Ontario.
ERA-Interim (ERA-I)	1979- ~ present (3-hourly)	Global (0.75°)	Pr, Tmin, Tmean, Tmax	Dee et al., (2011)	ERA-I is a reanalysis dataset that assimilates a large number of data sources, such as the Integrated Forecast System (IFS) cy31r2, satellite data, and others. ERA-I is a replacement for the previous ERA-40 dataset and features 4D-VAR data assimilation among other improvements to the original ERA-40 data set, which only had data up to 2002.
Watch Forcing data ERA-Interim (WFDEI)	1979-2013 (3-hourly)	Global (0.5°)	Pr, Tmin, Tmean, Tmax	Weedon et al., (2014)	WFDEI is an adjusted version of ERA-I using the European Union Water and Global Change (WATCH) Forcing Data (WFD) methodology, which includes various adjustments and bias corrections. These data are a replacement for the original ERA-40-based WFD dataset. The version updated to cover 1979-2013 was used although a version extending up to 2016 was released after the completion of this study.
Global Forcing Data – Hydro (GFD-HYDRO)	1979- ~present (3-hourly)	Global (0.5°)	Pr, Tmin, Tmean, Tmax	Berg et al., (2018)	GFD-Hydro closely mimics the methodology of WFDEI with updates to current versions of observed data networks. GFD-HYDRO is meant to be a global product similar to WFDEI, but available at near real time. Notable differences between WFDEI and GFD-HYDRO exist for precipitation, due to a reduction in undercatch adjustments, but offer more agreement in temperature.

The gridded datasets have a variety of temporal resolutions; therefore, each data set was aggregated to the largest time step feasible for hydrologic modeling (i.e., daily). For intercomparison, the datasets

must also be on a common spatial grid. Wong et al., (2017) chose to upscale and limit data selection to those with spatial resolutions of 0.5^{o} and finer, while Rapaić et al., (2015) chose to interpolate to match ECCC's CANGRD. There is no commonly accepted best practice, as all methods will introduce some degree of uncertainty. In this study, we chose bilinear interpolation to a common 10 km grid assuming precipitation values occurred at the center of a grid point, which was deemed suitable for distributed hydrologic modeling in the NCW (Lilhare et al., Accepted). The uncertainty associated with interpolation would be considered input uncertainty ingested by a hydrologic model.

Additional gridded climate datasets were considered for analysis but rejected based on the findings of other studies or by BaySys project criteria. The CaPA dataset was found to be a well performing dataset in many recent studies over Canada (e.g. Wong et al., 2017; Gbambie et al., 2017). The CaPA dataset only extends to 2002 historically, which fails to meet the time period requirement of the BaySys project of 1981-2010. Other products, such as MERRA (Rienecker et al., 2011), were found to be generally outperformed by products such as NARR in a study of the Canadian Arctic by Rapaić et al., (2015). Wong et al., (2017) compared the performance of the Princeton dataset developed by Sheffield et al., (2006), but found it was generally outperformed by the other selected datasets in the study.

The ERA-Interim and WFDEI products were included in the study as they were global products, which was of interest for the ocean modelers in the BaySys project (Barber, 2014). NRCAN's ANUSPLIN was selected for bias correction of future climate projections in an earlier stage of the BaySys project. NARR was included as it was generally considered to be a well performing dataset in the literature and because it offers wind data for use in the VIC hydrologic model (Liang et al., 1994) used by Lilhare et al., (Accepted) as part of the BaySys project. The HYPE model (SMHI, 2018), used for the generation of streamflow as part of the BaySys project, was initially calibrated using an early generation of GFD-Hydro

data provided by SMHI. Therefore, the current generation of GFD-Hydro (Berg et al., 2018) was included in this study.

3.5. Methodology

3.5.1. Performance Assessment

Analyses of individual datasets were conducted to ensure each product is a reasonable representation of ground based climate station data. Months with missing data were excluded from the performance assessment following the World Meteorological Organization standards, in which a month is considered missing if more than five days or three consecutive days in a month are missing (World Meteorological Organization. (1989)).

Continuous Statistics

Three continuous statistics were used to evaluate gridded dataset performance: root-mean-squared-error (RMSE), percent bias (PBIAS), and Spearman's rank correlation coefficient (Cor) as shown by equations (3.1) to (3.3):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Sim_i - Obs_i)^2} \quad (3.1)$$

$$PBIAS = \frac{\sum_{i=1}^{N}(Sim_i - Obs_i)}{\sum_{i=1}^{N}(Obs_i)} \times 100 \quad (3.2)$$

$$Cor = 1 - \frac{6\sum_{i=1}^{N}(Sim_i - Obs_i)^2}{N(N^2-1)} \quad (3.3)$$

in which N is the number of time steps of observed station data and corresponding gridded dataset data, and Sim and Obs are gridded and station timeseries, respectively. RMSE measures the magnitude of the squared difference between a gridded and station timeseries (smaller values are desired), PBIAS

measures the tendency of a gridded dataset to over- or under-predict an observed timeseries (smaller values are desired), and Cor represents a gridded dataset's ability to correctly reproduce the timing of observed climate station values (values close to 1 are desired). Spearman's rank correlation coefficient was selected as it is non-parametric and weights large differences higher than small: such was desirable in the current study as small differences may be within the uncertainty of the observations.

Categorical Statistics

Categorical statistics measure precipitation events captured within binned ranges (Jolliffe and Stephenson 2003), which are used to evaluate event occurrence partitioned by magnitude. This provides more information on uncertainty contributions associated with event types (Lespinas et al., 2015; Asong et al., 2017). Categorical statistics were also a standard metric used in the development and assessment of CaPA (Fortin et al., 2018). Events captured are measured by a contingency table (Table 2).

Table 2: Contingency table to assess when an event is correctly represented by a gridded data set for categorical statistics

	Observed	
Simulated	Obs = 1	Obs = 0
Sim = 1	Hit (H)	False Positive (F)
Sim = 0	Miss (M)	Correct Negative (C)

A value of 1 represents an event occurring within a bin, and a value of 0 represents an event not occurring within that bin. Bin size selection was adopted from the World Meteorological Organization (WMO) standards: [0, 0.2); [0.2, 1); [1, 2); [2, 5); [5, 10); [10, 25); [25, 50); and [50, inf); all in mm, where square brackets are inclusive and curved brackets are exclusive. Similar to Asong et al., (2015) and

Lespinas et al., (2015), two categorical statistics were used that are ECCC standard evaluation metrics (Fortin et al., 2018): the equitable threat score (ETS, equation 3.4 & 3.5) and the frequency bias (FBIAS, equation 3.6):

$$ETS = \frac{H-H_R}{H+F+M-H_R} - 1 \quad (3.4)$$

$$H_R = \frac{(H+F)(H+M)}{N} \quad (3.5)$$

$$FBIAS = \frac{H+F}{H+M} - 1 \quad (3.6)$$

in which N represents the total number of hits, false positives, misses, and correct negatives, and H_R is the number of correct forecasts assuming completely random forecasts (Lespinas et al., 2015).

Extreme Indices

Extreme indices measure the occurrence and magnitude of extreme events (Sillmann et al., 2013). Each extreme index was evaluated for seasonal, annual, and the full temporal periods. Five temperature extremes were selected: frost days (FD) when Tmin < 0^{o}C, summer days (SU) when Tmax > 25^{o}C, icing days (ID) when Tmax < 0^{o}C, and tropical nights (TR) when Tmin > 25^{o}C; as well as daily temperature range (DTR, equation 3.7).

$$DTR = \frac{\sum_{i=1}^{N} Tmax_i - Tmin_i}{N} \quad (3.7)$$

Seven precipitation extreme indices were selected: 1 day maximum precipitation (RX1), 5 day consecutive maximum precipitation (RX5), simple precipitation intensity index (SDII, equation 3.8), number of days above 10 mm (R10), number of days above 20 mm (R20), dry spell length (CDD) measuring consecutive days < 1 mm, and wet spell length (CWD) measuring consecutive days ≥ 1 mm.

$$SDII = \frac{\sum_{i=1}^{N} Pr_{1mm}}{W} \quad (3.8)$$

in which Pr_{1mm} represents events $\geq$ 1 mm and W is the count of days $\geq$ 1 mm.

The choice of precipitation extreme indices was made to allow the examination of persistence patterns (i.e. CDD, CWD, and RX5), the effects of spatial aggregation on absolute extremes (i.e. RX1), the relative occurrence of higher intensity precipitation events (i.e. R10 and R20), and the tendency to over or under simulate low precipitation days (i.e. SDII). A simpler subset of temperature extreme indices was selected, mostly made up of temperature threshold indices. Temperature threshold indices suggest a model's ability to reproduce the number of occurrences of events beyond the threshold value, as opposed to percentile measures that provide more information on magnitude than occurrence quantities. Occurrence quantities were considered the focus based on the potential effect of utilizing varied spatial aggregations for comparison. Additionally, the daily temperature range was selected to reflect a dataset's ability to mimic daily temperature variations.

3.5.2. Ensemble Creation

Ensemble ranges for each variable were generated by selecting the minimum and maximum value in a time step for each grid point. Since each member is considered an acceptable representation of the observed environment, the minimum and maximum ensemble members represent the total range of uncertainty among gridded datasets for that time step and location. For this study, the ensemble mean was computed using equal weighting of the five datasets. The purpose for including the ensemble mean was to represent a high likelihood realization. Other methods to generate a high likelihood realization, such as using the median, would not be expected to significantly improve results, but rather, offer a different high likelihood realization. Ideally, all observed values would occur within the ensemble range of uncertainty; therefore, the uncertainty represented is estimated by the number of values that fell within the ensemble range.

The goal of this study within the BaySys project was to generate an estimate of total uncertainty associated with simulating flows for the Nelson River. This thesis provides an estimation of the total uncertainty bounds for flow from the Nelson River that is required for propagation into other models used within the BaySys project. Therefore, the BaySys project requires an estimate of the possible range of total uncertainty, which includes low likelihood realizations that could increase in probability under a changing hydroclimate. Therefore, this guided the decision to utilize the absolute maximum and minimum ensemble realizations to better align with the goal of providing a possible range of total uncertainty. Since the gridded datasets were generated using climate models, it would be desirable to evaluate the input, structural, and parameter uncertainty associated with those models to determine a reasonable estimate of the bounds of uncertainty. The calibrated climate models used for the gridded datasets are not publically available. Therefore, the absolute minimum and maximum ensemble members are selected to intentionally widen the uncertain range of the total ensemble. Using other ensemble methods to generate dry and wet realizations would only account for structural uncertainty in the gridded climate datasets. Ideally, the intentionally widened uncertainty range would be more representative of the uncertainty bounds that could have been generated if climate model input and parameter uncertainties were also sampled. Without actually sampling from the climate model input and parameter uncertainties, the likelihood associated with the minimum and maximum precipitation realizations is not known. It is assumed that they are at an extreme low likelihood, but no quintile can be reported, therefore, it is possible that the intentionally wide uncertainty range may overestimate the actual uncertain range of the gridded climate data.

3.5.3. Spatial Aggregation

Seasonal, annual, and study period performance analyses were conducted using three spatial aggregations: fully aggregated over the NCW (lumped), aggregated by major basin (semi-lumped), and a

station comparison to the nearest grid cell (distributed). Spatial aggregation was done using a simple arithmetic mean of points falling inside and along a delineation. Ground based climate station aggregated data were only assumed missing if all stations in a basin were missing data for a particular day; this assumption was more impactful in basins with low spatial coverage. When some, but not all, stations are missing data for a timestep, the spatially aggregated timeseries had lower spatial coverage for that timestep, potentially lowering performance metrics. Precipitation uncertainty is often higher than that of other meteorological variables, such as temperature (e.g. Rapaić et al., 2015), in part, due to the spatial positioning and movement of storms. By spatially aggregating, the reliance on grid point storm positioning is less emphasized (i.e. the comparison of a grid of data to a point gauge); this minimized the contribution of spatial positioning to uncertainty given the relatively coarse resolution of grids in this study (10 km). Some hydrologic models, however, ingest meteorological data at the grid point scale; therefore, grid point comparisons were also generated for comparison with results from aggregations. Grid point uncertainty is more reflective of total uncertainty associated with a meteorological data ensemble; this provides insight into uncertainty subject to propagation when comparing models of different spatial structures.

A challenge with utilizing the absolute maximum and minimum ensemble realizations was the spatial variability of storms. If the spatial positioning of a storm missed a grid cell used for comparison with a ground-based climate station, the precipitation at that grid cell is likely to be zero. Therefore, the spatial aggregation to basin scales provided average conditions in the aggregated area rather than a single point. This better accounted for the spatial variation in storm representation by the individual datasets. By conducting analysis on both the basin scale aggregations and individual points, additional information as to the quantitative occurrence of variations in the spatial positioning of storms was generated through comparison of results for the selected spatial aggregations.

The Red River Basin (RRB) was different than other basins for the semi-lumped comparison. All gridded datasets with the exception of ANUSPLIN have data for the full RRB extent; therefore, spatial aggregation in the RRB utilized all data available. This meant that NARR, ERA-I, WFDEI, and GFD-Hydro's RRB spatial averages considered grid points in the U.S., while ANUSPLIN's RRB spatial average only considered grid points in the Canadian portion of the basin. Extreme events are less detectable for spatial averages as events are averaged with cells not affected by the same event. Therefore, the inclusion of a distributed comparison offers additional information on extremes and their spatial positioning, which was more valuable in the RRB.

3.6. Results

3.6.1. Gridded Dataset Analysis

Period mean monthly plots provide a generalized dataset comparison (Figure 2). ERA-Interim is generally the wettest of the gridded datasets and is often wetter than AHCCD, while NARR and ANUSPLIN are often the driest datasets, usually drier than ECCC. The largest differences are in basins with fewer climate stations and poor climate station coverage (i.e., Lake Winnipeg Basin (LWB) and Lower Nelson River Basin (LNRB)), as well as basins near the Canada-USA border (i.e., Winnipeg River Basin (WRB) and Red River Basin (RRB)). NARR performs worse in the WRB than in other basins; the 2015 update cited performance in southern Ontario as a focus for improvement. Results in the RRB must be interpreted with care given that only stations in Canada were considered. Temperatures among all gridded datasets and both AHCCD and ECCC were in close agreement, with the exception of NARR. NARR was generally warmer in all basins by up to ~1^{o}C (shown in Appendix A, Figure A.1); the warm bias in NARR has been documented and explored in the literature (e.g. Choi et al., 2009; Bennington et al., 2010, Lilhare et al., Accepted).

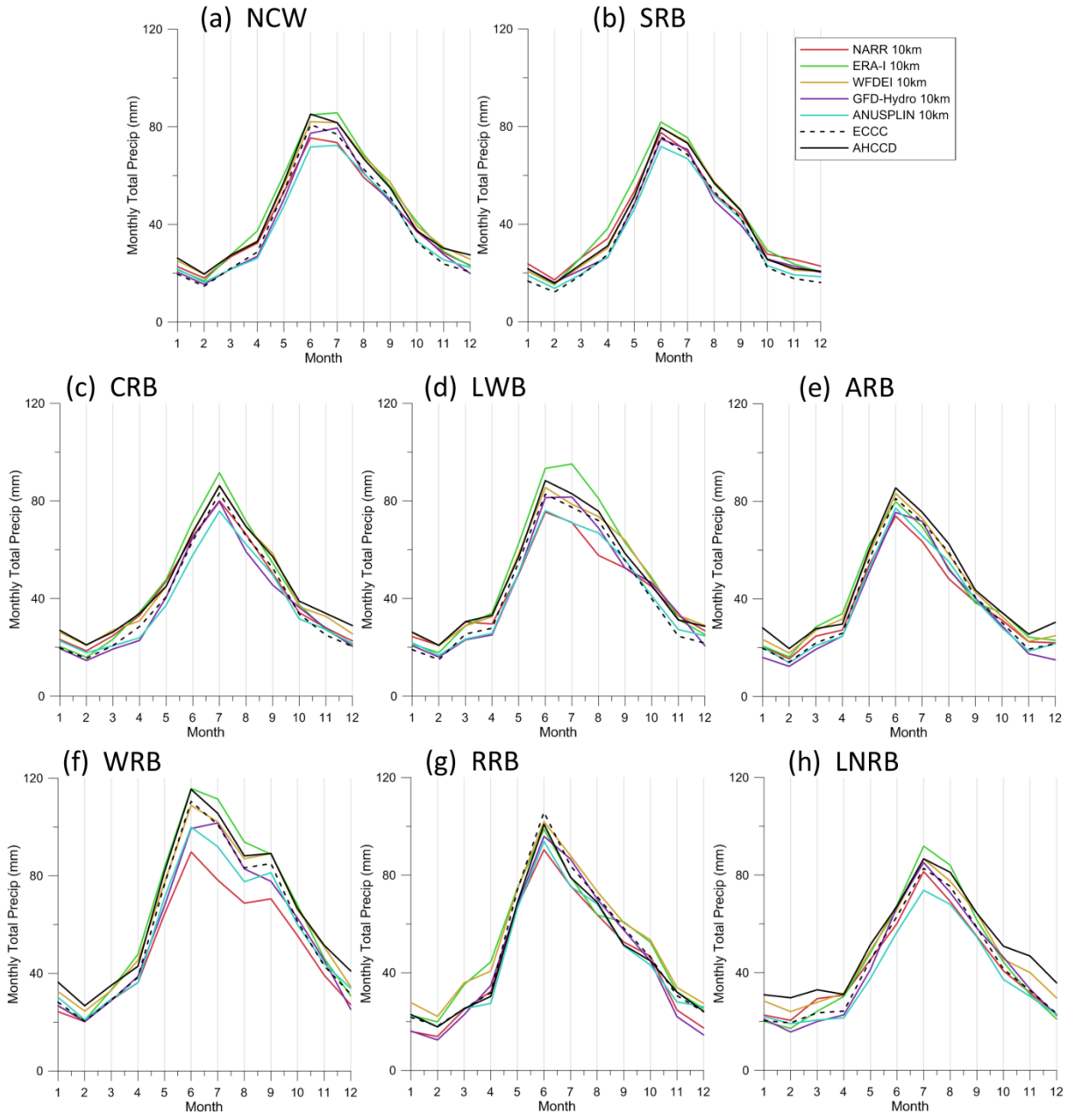

Figure 2: Period mean monthly precipitation for the period of 1981-2010, spatially averaged over each of the major basins. Aggregations include the lumped Nelson Churchill Watershed (NCW), and the semi lumped major basin aggregations that include: the Saskatchewan River Basin (SRB), the Churchill River Basin (CRB), the Lake Winnipeg Basin (LWB), the Assiniboine River Basin (ARB), the Winnipeg River Basin (WRB), the Red River Basin (RRB), and the Lower Nelson River Basin (LNRB).

The relative performance of each dataset, temporally partitioned into a moving annual temporal window (Figure 3), is less consistent, indicating performance between years can be highly variable. Hydrologic models generally assess performance on a daily or monthly temporal scale, therefore,

making the information in Figure 3 more representative of general calibration methodologies in hydrologic modeling. Correlations are generally higher in larger basins with better climate station coverage, as opposed to lower correlations in the RRB and LWB that both had poor climate station coverage. Years that showed lower correlations in all datasets were generally due to some missing data in the AHCCD datasets. Basin averages were only considered missing if all stations were missing for a timestep, therefore, when some stations had missing data the data coverage was reduced for those timesteps which reduced correlations if it occurred frequently in a year. All correlation values are statistically significant at the 99% level (Figure 3a).

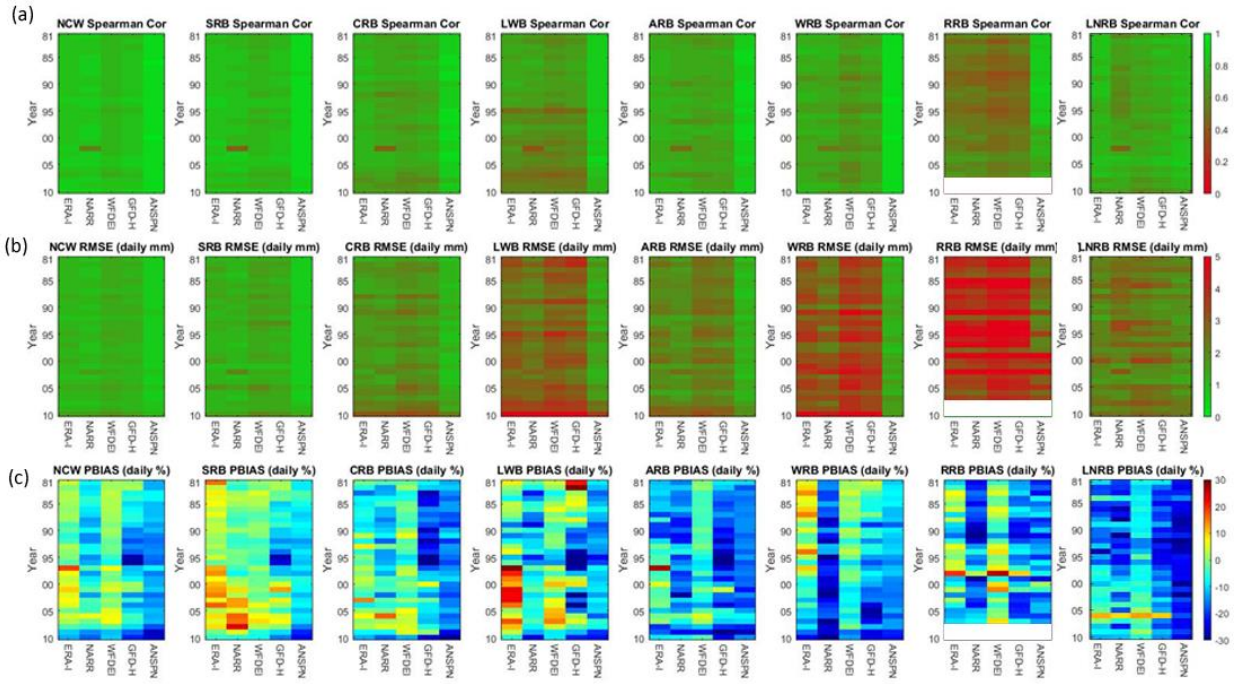

Figure 3: Daily precipitation spatially aggregated annual continuous statistics with reference to the AHCCD observed data set in each major basin (1981-2010). (a) daily Spearman correlation (b) daily RMSE, and (c) daily PBIAS. White is used to represent periods with no available data.

RMSE values in Figure 3b were reflective of the information from Figure 2, in which the disagreements between gridded datasets and station data are most notable in the LWB, WRB, RRB, and the LNRB. PBIAS (Figure 3c) was critical in highlighting temporal inconsistencies obscured by the longer-term temporal averaging period used in Figure 2. While Figure 2 suggested the existence of a strong wet bias

for ERA-Interim in the LWB, Figure 3 revealed that this positive bias was mainly influenced by the 1997-2004 period. Similar temporal inconsistencies existed for each dataset, as performance varied from year to year. Performance was generally worse for the distributed comparisons (not shown), in which yearly correlations ranged from near zero to 0.83 for all datasets with the exception of ANUSPLIN whose correlations were generally above 0.7 and in some cases approached 1.0. Since ANUSPLIN was interpolated from the selected ECCC station locations, the distributed comparison is expected to perform well since the two datasets are not independent. It is also important to note that ground based climate station data was also ingested by the reanalysis products making them also not completely independent. Yearly distributed RMSE ranged from 0.2 to 11 mm, and yearly PBIAS showed similar amounts of disagreement and temporal variability to those from Figure 3c. Performance was generally worst in summer (Appendix A, Figure A.4) or winter (Appendix A, Figure A.2); and best in spring (Appendix A, Figure A.3) or autumn (Appendix A, Figure A.5) for all three spatial aggregations. Although precipitation events were generally smaller in winter, RMSE values were low in winter relative to the other seasons. Comparisons with ECCC were similar with the exception that PBIAS values more often suggested a wet bias. This result was expected, as AHCCD adjustments generally added precipitation; the largest differences occurred in winter.

Categorical statistics, similar to continuous statistics, performed better when gridded datasets were lumped or semi-lumped. ETS scores were generally highest for the 0-0.2 mm bin, the 10-25 mm bin, and the 25-50 mm bin; performance was generally worst for the 0.2-1 mm bin and the 1-2 mm bin. FBIAS scores were generally negative for events below 0.2 mm and those above 2 mm; events between 1 and 10 mm generally had the smallest FBIAS scores, suggesting the gridded datasets did not show a tendency to consistently over or underestimate events of those magnitudes. A notable trend was the absence of high-magnitude events in larger sub-basins. The NCW, SRB, and CRB often had no events larger than 25 mm occur as opposed to the distributed comparison, which often had events occur above

50 mm: spatial aggregation had a smoothing effect on extreme events, reducing their frequency. Extreme event frequency was, however, better preserved in smaller spatial aggregations (i.e. LNRB versus NCW aggregations).

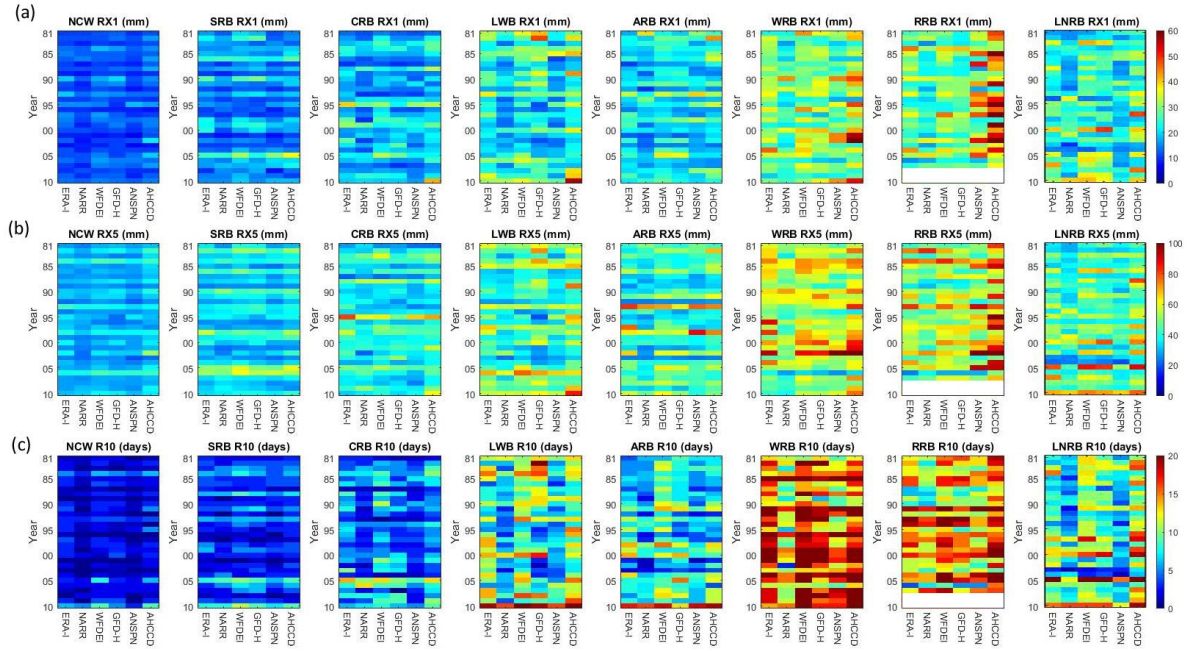

Figure 4: Daily precipitation spatially aggregated extreme yearly indexes in each major basin (1981-2010). (a) RX1 (b) RX5, and (c) R10. White is used to represent periods with no available station data

Results from the yearly extremes analysis are presented on Figure 4. Larger aggregated basins (i.e. the NCW versus the LNRB) showed lower extreme values. The gridded datasets often had lower RX1 and RX5 values than those of AHCCD; the largest difference between the gridded datasets and AHCCD was often in years with the largest extreme events (Figure 4a and b). The distributed comparison showed more variation, as increased reliance on spatial variability of storm location representation was introduced. The distributed comparison similarly showed the greatest difference between the gridded datasets and AHCCD among the largest events. Basin-aggregated yearly R20 counts were similar to R10, for which counts varied between years (Figure 4c). Overall, the gridded datasets had higher R10 yearly counts by an average over all years of 0.2 days. CDD lengths were generally higher for all gridded

datasets for lumped and semi-lumped aggregations. Yearly R10 and R20 counts were generally closer to those of AHCCD for the distributed comparison for all gridded datasets. CDD lengths showed larger differences for the distributed comparison, in which NARR and ERA-Interim generally produced lower CDD lengths than those of AHCCD. WFDEI performed similarly to the basin aggregated CDD lengths and GFD-Hydro and ANUSPLIN generally had higher CDD lengths for the distributed comparison. RX1 and RX5 values were generally closer to those of AHCCD in winter (Appendix A, Figure A.6) and autumn (Appendix A, Figure A.9), and had the largest differences in summer (Appendix A, Figure A.8), indicating gridded datasets were less likely to accurately estimate the magnitude of convective storms. R10 and R20 counts were less consistent in bias, but the average across all summers and datasets was less than a 1-day difference, although counts were more often underestimations. This suggested that, like bias (Figure 3c), R10 values vary year to year, but, the average across all years cancels out the over or underestimate of extremes. Temperature extremes showed significantly less variation than those from precipitation (not shown), therefore, only precipitation performance was considered for the ensemble selection criteria. Comparison with ECCC showed fewer wet days than AHCCD, which suggested better extreme value representation by the gridded datasets.

3.6.2. Ensemble Analysis

The ensemble ranges of yearly total precipitation are presented in Figure 5. If the ensemble envelope is considered to be representative of total uncertainty, then wet conditions, as well as the low spatial coverage of climate stations, appeared to increase uncertainty (Figure 5d, f, and h). The ensemble mean was often similar to ECCC, although larger sub-basins tended to approach AHCCD; performance plots (similar to Figure 3) are presented for the ensemble members in Figure 6. This relationship varied seasonally, as the ensemble mean generally approached AHCCD in winter, spring and autumn for large

basin aggregations but was similar to ECCC in summer. The envelope created by the ensemble minimum and maximum realizations was widest in summer and narrowest in winter.

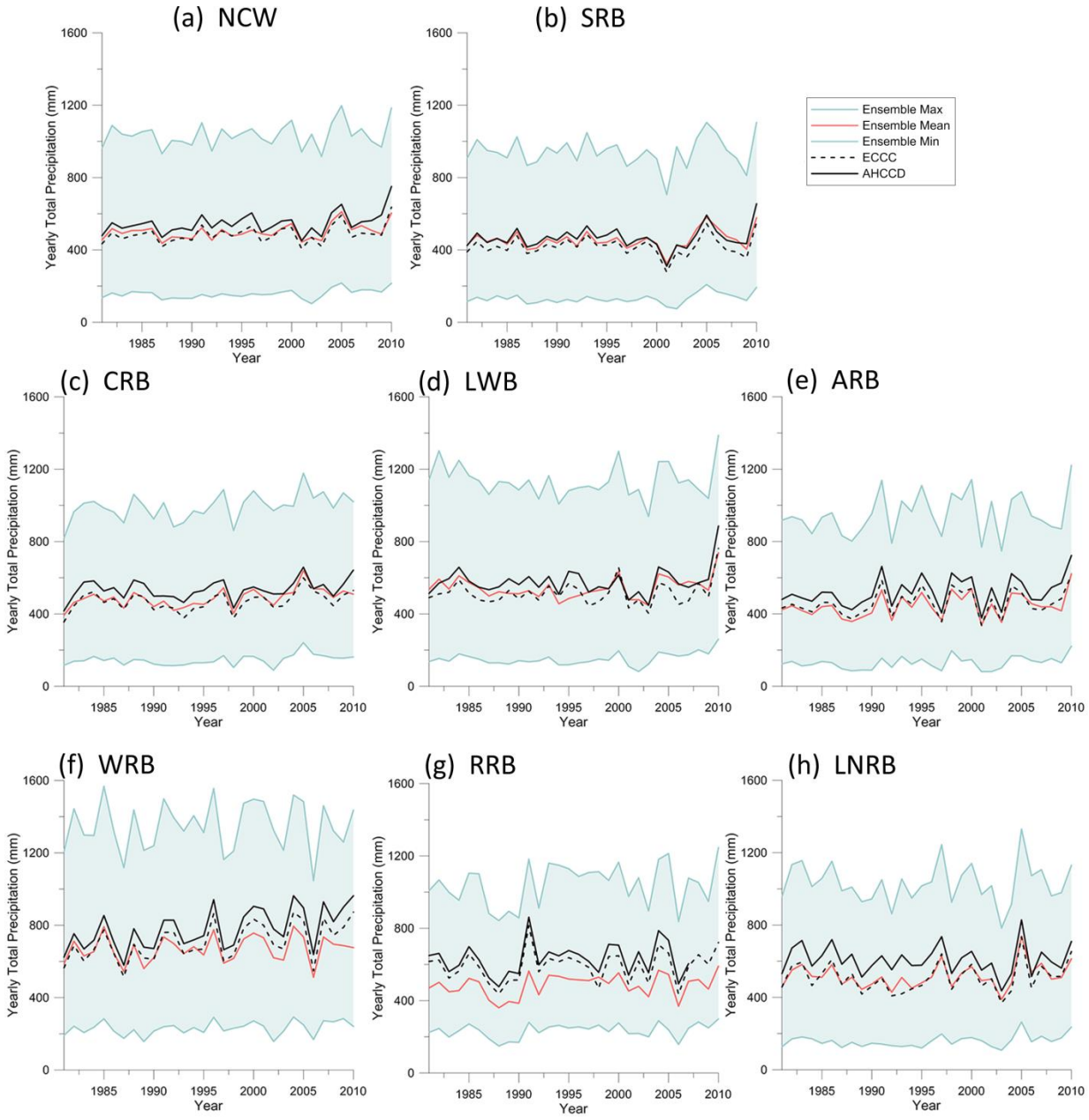

Figure 5: Basin-averaged yearly total precipitation timeseries showing the ensemble minimum, mean, and maximum, as well as ECCC and AHCCD for each major basin in the NCW (1981-2010).

Figure 6a shows correlations between the ensemble minimum, mean, and maximum with respect to AHCCD generally as good as, or better than, any individual dataset correlation with respect to AHCCD (Figure 3a). The ensemble mean generally produced the highest correlations of the three ensemble

realizations with respect to AHCCD. Similar to the individual dataset correlations, ensemble realization correlations are lower in the LWB and RRB, but showed improvement in the WRB and LNRB. Correlations are generally lower for the ensemble minimum as Spearman's rank correlation coefficient is more sensitive to large differences between the timeseries used to calculate correlations (i.e. the ensemble minimum and AHCCD). The ensemble members tended to underestimate extreme events, so using the ensemble minimum tended to increase differences more than using the ensemble maximum did. Including the ensemble minimum and maximum was valuable to the BaySys project, as the ensemble minimum and maximum present the extreme range of uncertainty subject to propagation through to simulated streamflow. As a contributor to total uncertainty, providing wide total uncertainty bounds is useful for testing model sensitivity to the lowest and highest flows within the simulated streamflow uncertainty bounds, and though they may represent low likelihood scenarios today, in the future, the probability of occurrence of these scenarios could increase.

Yearly PBIAS values (Figure 6c) for the ensemble mean were generally negative, with respect to AHCCD, but showed less temporal variation compared to any individual dataset PBIAS values (Figure 3c). RMSE values (Figure 6b) for the ensemble mean showed the most improvement over Individual dataset RMSE values presented in Figure 3b for the larger basins such as the NCW, SRB, CRB, and ARB. RMSE values were similar for the WRB, LWB, and LNRB, although some years in the LNRB showed small improvement. Scores were similar or worse in the RRB, which is a reflection on the discontinuity created by the partial coverage of climate stations and ANUSPLIN. PBIAS values for the ensemble minimum and maximum suggest a significant underestimation and overestimation of AHCCD, respectively (Figure 6c).

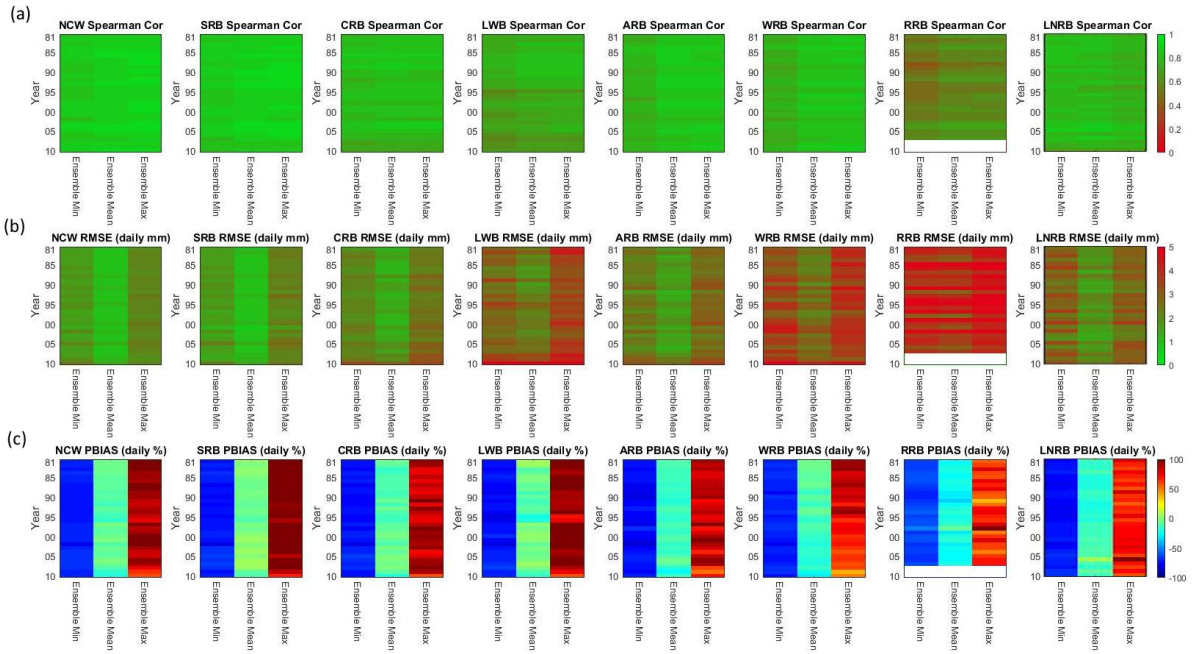

Figure 6: Basin-averaged daily precipitation continuous yearly statistics with reference to the AHCCD observed data set in each major basin (1981-2010) for (a) daily Spearman correlation (b) daily RMSE, and (c) daily PBIAS for the ensemble minimum, mean, and maximum.

Performance for the distributed comparison was similar to Figure 6 for the ensemble mean, in which yearly correlations ranged from near zero to 0.85. Yearly RMSE ranged from 0.5 to 9.3 mm, and PBIAS on average across all years and stations ranged from -18% to +9%. The ensemble minimum and maximum both had worse correlations with respect to AHCCD than the ensemble mean for the distributed comparison, with the minimum being the worse of the two. Yearly correlations for the ensemble minimum and maximum ranged from near zero to 0.8 and near zero to 0.83, respectively. Yearly RMSE values spanned 0.1-10.9 mm and 0.8-10.7 mm, respectively, and PBIAS values for the ensemble minimum were generally near -70% while ensemble maximum PBIAS values were generally near 90%. The large negative PBIAS of the ensemble minimum was sourced from a consistent underestimation of daily precipitation values rather than missing the occurrence of a large number of events. The ensemble maximum has a large positive bias for the same reason, but in the opposite direction. This is supported by the ensemble minimum and maximum RMSE values (Figure 6b) being

similar to RMSE values of individual datasets (Figure 3b), with respect to AHCCD. This suggested that, on average, the difference between the minimum or maximum ensemble realization and AHCCD on any particular day was generally similar to the individual gridded datasets on any particular day. The exception was in summers (appendix A, Figure A.12) where RMSE values for the minimum and maximum ensemble realizations were slightly larger than the RMSE values for individual datasets (Appendix A, Figure A.4), with respect to AHCCD.

Seasonal performance for all three spatial aggregations was reflective of the gridded dataset weaknesses. Performance was often worst in either winter (Appendix A, Figure A.10) or summer (Appendix A, Figure A.12), and often best in either spring (Appendix A, Figure A.11) or autumn (Appendix A, Figure A.13) (with the exception of winter RMSE values that were often smallest due to fewer large events). Differences in ensemble performance between seasons were smaller and often showed improvements over individual datasets. The distributed comparison (not shown) of the ensemble mean to AHCCD stations showed more improvement than the basin aggregation comparison, with respect to the individual datasets. Performance metrics in winter, including correlation, were similar to spring and autumn; summer was generally still the lowest performing season. The ensemble minimum and maximum similarly showed strong negative and positive biases respectively.

For basin aggregations, ETS scores for the ensemble minimum and mean, similar to the individual datasets, were generally highest for the 0-0.2 mm bin, the 10-25 mm bin, or the 25-50 mm bin; performance was still generally worst for the 0.2-1 mm bin or the 1-2 mm bin. The ensemble maximum generally had worse performance than the minimum and mean in the <0.2 mm bin, but had higher scores for events larger than 10 mm. Most notable were non-zero scores for the small number >50 mm events in the WRB, meaning that the ensemble maximum did produce events larger than 50 mm even when spatially aggregated. Ensemble minimum FBIAS values were positive for events below 0.2 mm and

negative for all others. Ensemble mean FBIAS values were generally negative for events smaller than 0.2 mm and for events larger than 10 mm; other event magnitudes fluctuated between positive and negative values. Ensemble maximum FBIAS values were generally positive for events larger than 0.2. Ensemble minimum and maximum FBIAS values were smallest between 0.2 and 2 mm, however, the ensemble mean showed no trend in which magnitude of events produced the smallest FBIAS values. These relationships were consistent with those of the distributed comparison; the higher frequency of events larger than 25 mm was notably better represented by the ensemble maximum than by any individual gridded dataset.

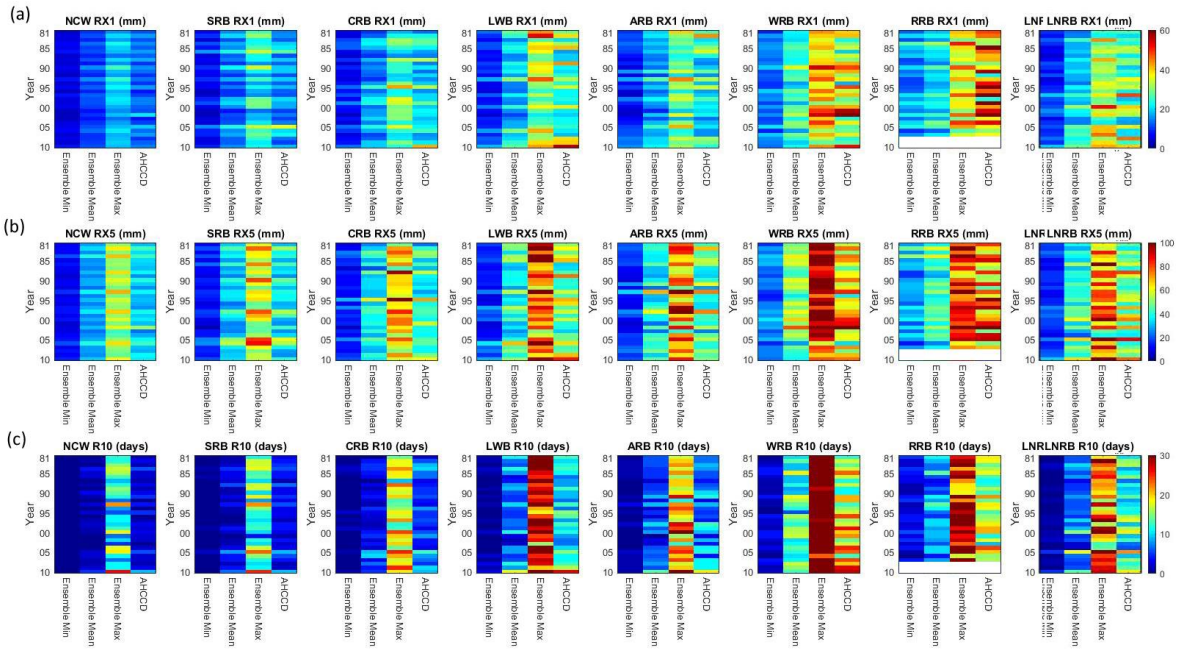

Figure 7: Basin-averaged daily precipitation annual extreme indexes in each major basin (1981-2010) for (a) RX1 (b) RX5, and (c) R10 for the ensemble minimum, mean, and maximum.

Results for RX1, RX5, and R10 for the selected ensemble realizations are presented in Figure 7. The ensemble minimum generally performed poorly for all extreme metrics. The individual gridded datasets often underestimated extreme values; similarly, the ensemble mean did as well. The ensemble minimum was generated to be a low likelihood ensemble realization for the purpose of estimating the

possible range of uncertainty for the Nelson River as part of the BaySys project. Since the ensemble mean generally underestimated extremes, the ensemble minimum must also underestimate extremes, but to a greater degree. The ensemble maximum generally overestimated extreme events; however, it performed well in comparison with AHCCD for RX1 values (Figure 7a). Yearly RX1 values for basin aggregations were generally underestimated by the ensemble mean and slightly overestimated by the ensemble maximum, with notable temporal variation between years. Similar results are seen for RX5 and R10 values (Figure 7b and c); however, the overestimation by the ensemble maximum was larger. CWD values were often slightly overestimated by the ensemble mean and significantly overestimated by the ensemble maximum.

Extremes showed more variation for the distributed comparison, the ensemble mean still generally underestimated the values of RX1 and RX5 but also produced some years of overestimation. Similar to the basin aggregations, the ensemble maximum produced small overestimations of RX1 but significantly overestimated values for RX5. Distributed R10 and R20 values were generally underestimated by the ensemble mean and overestimated by the ensemble maximum, although differences were generally smaller than those of the basin-aggregated comparison. Similar to the individual datasets, the largest differences in extremes occurred during summer (Appendix A, Figure A.16), and the smallest differences were often in winter (Appendix A, Figure A.14) or autumn (Appendix A, Figure A.17).

Spatial aggregation was the most influential aspect when assessing events captured by the ensemble (Figure 8). Figure 8a shows the ensemble mean yearly average precipitation spatially aggregated over the NCW to be ~500 mm year^{-1} and indicates that only 8% of events above the 50th percentile were not captured by the ensemble (Figure 8b). Roughly three quarters of all events missed were above the ensemble maximum; however, of the nearly 500 missed events above the 50th percentile, only 20 events were below the ensemble minimum. For events below the ensemble minimum, the average difference

was 0.2 mm and the largest differences occurred between the 80th and 90th percentiles. For events above the ensemble maximum, the average difference was 0.56 mm. Differences increased with event magnitude, being lowest near the 50th percentile (0.12 mm) and highest above the 99th percentile (2 mm).

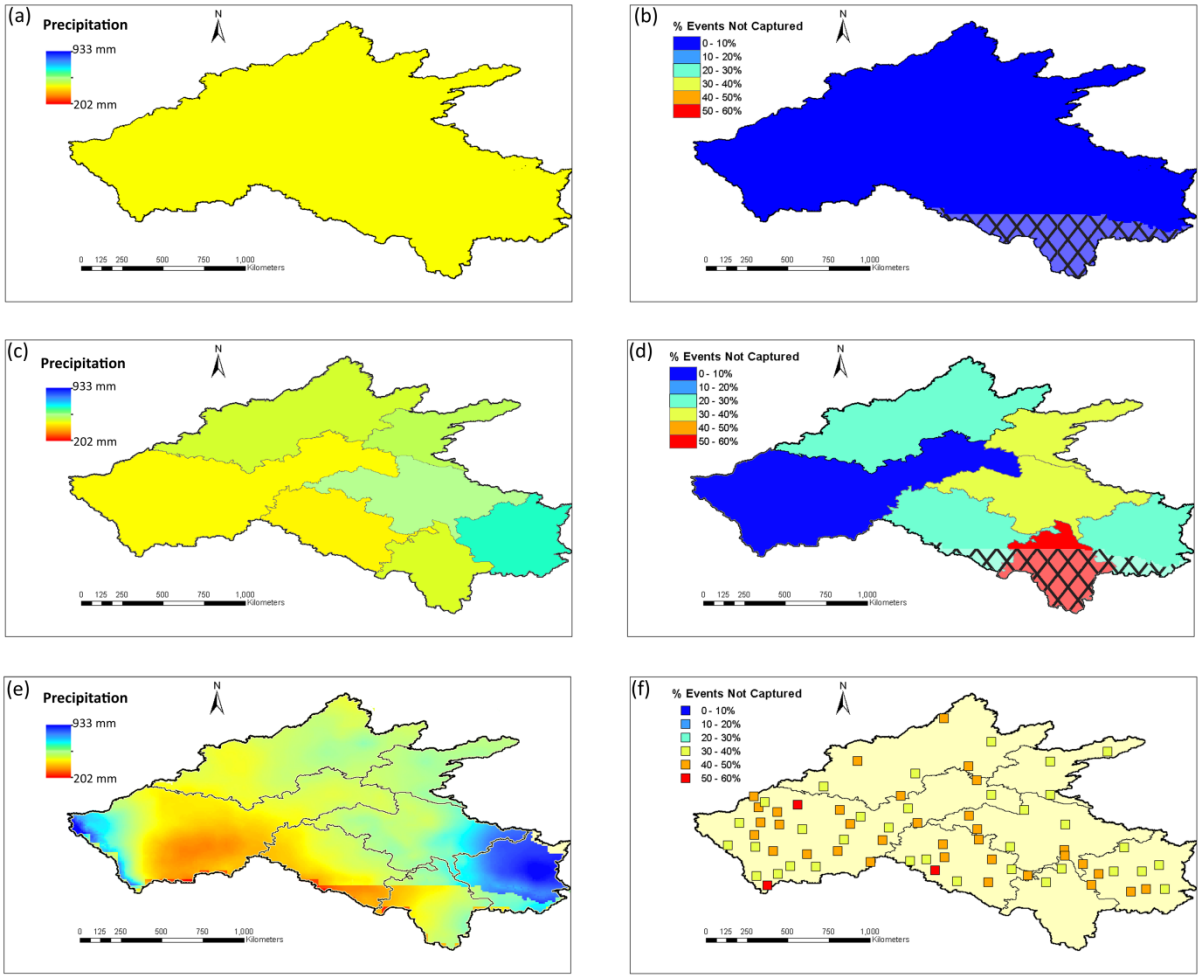

Figure 8: Period mean yearly sum precipitation (1981-2010) according to different spatial aggregation schemes: fully lumped (a), semi lumped (c), and fully distributed (e). The % of AHCCD events missed by the ensemble range above the 50th percentile when using different spatial aggregation schemes: fully lumped (b), semi lumped (d), and fully distributed (f). The black and white crosshatch pattern seen in (b) and (d) covers areas outside of Canada that may not be correctly represented due to dataset limitations.

Semi-lumped aggregation indicated more diversity in yearly average precipitation and the number of events not captured (Figure 8c and d). The worst performing basin was the RRB, missing 53% of events above the 50th percentile. This is an expectedly poor performance given that gridded data over the full

basin were considered, while only climate stations in Canada were considered. Excluding the RRB, the percentage of events below the ensemble ranged from 6% to 18% above the 50th percentile, while the RRB had 32% of missed events below the ensemble minimum (Figure 8d). Events below the ensemble minimum were missed by an average of 0.82 mm, and events above the ensemble maximum were missed by an average of 1.64 mm; misses both below the ensemble minimum and above the ensemble maximum had larger differences as event magnitude increased.

Distributed comparisons were the worst, ranging from 31% to 52% of events missed above the 50th percentile (Figure 8f). Events were below the ensemble minimum an average of 22% across all stations, and events below the ensemble minimum were missed by an average of 1 mm. Events above the ensemble maximum were missed by an average of 2.2 mm, although above the 95th percentile misses were as large as 12 mm. Similar to events captured, a comparison between Figure 8a, c, and e shows how spatial variation in precipitation quantity is lost with progressively larger spatial aggregations.

3.7. Discussion and Conclusion

3.7.1. Uncertainty from Spatial Aggregation

To inter-compare gridded datasets, they must be on a common grid; however, interpolation adds uncertainty to the input data comparison. The Kolmogorov-Smirnov test detected statistically significant differences between each of the gridded datasets pre and post bilinear interpolation for all spatial aggregations at the 5% significance level. One exception was NARR when lumped or semi-lumped aggregations were used: this is likely due to NARR having the smallest change in grid spacing. The interpolated datasets produced performance metrics as good as, or better than, their original grids, which is similar to what has been reported in other studies (e.g. Tustison et al., 2001).

The comparison of individual gridded datasets (Figure 2) provides no indication of potential performance evolution through time, which is a known issue for at least the NARR data (Eum et al., 2014). Additionally, short term events that effect wet/dry and warm/cool years such as the impacts of El Niño-Southern Oscillation (ENSO) years are smoothed out over long temporal periods (Shabbar and Khandekar, 1996; Trenberth, 1997). No information is gained towards understanding a datasets ability to reproduce events such as ENSO events by performance metrics calculated over long temporal periods (e.g. 30 or more years). Therefore, yearly and seasonal temporal periods are more appropriate for understanding dataset characteristics (Figure 3). The gridded datasets were weakest in the representation of extreme events (in which extreme events are considered to be quantified by the selected extreme indices), although spatial aggregation was the most significant factor influencing event magnitude (Figure 4). This conclusion is supported in the literature: Sun and Barros (2010) examined the representation of extremes by NARR and found it to generally underestimate extremes. Sillmann et al., (2013) compared ERA-Interim with several other reanalysis and GCM products and showed ERA-Interim to slightly underestimate precipitation extremes for a large region over central and eastern Canada. The most extreme events within any given year generally always occur in summer (e.g. when using extreme indices focused values such as RX1) due to convective storms and fronts (Dingman, 2014). Uncertainty in the spatial positioning of storm events by a gridded dataset may lead to an event being reasonably represented in magnitude, but not at the location of a ground-based climate station. This would suggest that the distributed comparison was dependent on spatial positioning and magnitude of a storm event, while a spatially aggregated basin (e.g. SRB) would be more sensitive to storm magnitude rather than positioning. This implies that the general uncertainty in gridded datasets is lower when spatially aggregated, which agrees with the literature (e.g. Carpenter and Georgakakos, 2006). There is a general gradient in extremes magnitude in Figure 4: extremes are lower for the gridded datasets and AHCCD in larger spatial aggregations (e.g. NCW) and higher in smaller spatial aggregations (e.g. LNRB). This is an

anticipated result of averaging the spatial variability in precipitation extremes (e.g. Fischer et al., 2013; Pendergrass et al., 2017). A storm occurring in a particular location is averaged with all other grid points where the storm did not occur within a spatially aggregated area. Similarly, ground-based climate stations capturing the storm are averaged with ground-based climate stations that did not capture that storm within an aggregated area. When climate station placement is relatively dense and uniform in an area, the spatial average of the gridded datasets and the ground-based climate stations are comparable (e.g. SRB). If climate station coverage is sparse (e.g. LNRB and RRB), storm events may be entirely missed by a sparse climate station network, but well represented by a gridded dataset. This suggests added-value to assessing multiple spatial aggregation approaches rather than attempting to determine an optimal spatial aggregation approach (e.g. Fischer et al., 2013; Pendergrass et al., 2017).

The lower CDD lengths of NARR and ERA-Interim were an expected result. Becker et al., (2009) reports a slight systematic bias towards more frequent light precipitation in NARR; this is further supported by Choi et al., (2009). ERA-Interim generally produced wet conditions (Figure 3c), suggesting CDD length to be generally shorter, a wet bias was also reported by Rapaić et al., (2015). The longer CDD lengths of ANUSPLIN were also expected, as ANUSPLIN generally showed a dry bias (Figure 3c), which was similarly found by Gbambie et al., (2017). Similar to the conclusions of Wong et al., (2017), WFDEI was often similar to AHCCD, with some notable temporal variation (Figure 3c).

3.7.2. On the Value of Climate Data Ensembles

Similar to the conclusions of Gbambie et al., (2017), the present study shows that each dataset provides valuable information but also has temporally and spatially variable strengths and weaknesses. These results contradict those of studies similar to Eum et al., (2014) and Wong et al., (2017) that suggest a "best" product may be potentially identifiable with enough analysis. It is important to note that performance metrics calculated for the full 1981-2010 period more closely agree with results from Eum

et al., (2014) and Wong et al., (2017), as high and low performance years average out but obscure temporally dependent information (Rapaić et al., 2015). Comparing the gridded datasets with ECCC produced similar results to those of the comparison with AHCCD. The ECCC comparison generally showed gridded precipitation dataset bias to be wet, with reference to the AHCCD comparison; the difference was most significant in winter.

It appears beneficial to combine the individual datasets into an ensemble in which the ensemble range represents the uncertainty associated with estimating station data, rather than attempting to identify the dataset that has the least uncertainty associated with it (Gbambie et al., 2017). Yearly precipitation values provide an overview of the ensemble range (Figure 5). Yao et al., (2014) used a linear weighting method to combine various data sets and found it to improve results when compared with ground-based products. Other methods such as Bayesian model averaging (Yang et al., 2012) have been implemented successfully. Weighting methods may be able to account for spatial and temporal complexities but are limited by uncertainty in station data measurement and spatial coverage; therefore, a simple average of the gridded datasets was used in the current study to avoid making assumptions about the importance of each dataset.

The ARB, WRB, and RRB all appear drier when compared to ECCC (Figure 5): this is partly caused by border effects and ANUSPLIN lacking data for the USA. The LNRB had a wide ensemble range due to poor spatial coverage of climate stations and the complexity introduced by the influence of Hudson Bay on local climate (Rouse, 1991). The LWB may also have had a larger range due to the complexities associated with simulating the climatic effects of Lake Winnipeg.

The difference between individual datasets (Appendix A, Figure A.4) and ensemble realizations (Appendix A, Figure A.12) RMSE values in summer with respect to AHCCD suggested that events such as convective storms may have been missed, or they had their spatial extent increased by taking the

absolute minimum and maximum, respectively. The absolute minimum and maximum ensemble realizations were more reasonable in winter, spring, and autumn. This is also reflected in the seasonal ensemble envelopes, similar to those in Figure 5, in which the summer envelope was on average twice as wide as spring or autumn, and the winter envelope was generally the narrowest.

3.7.3. Generating Uncertainty Envelopes

Since the ensemble range can be interpreted as the (input data) uncertainty associated with the ensemble's representation of station conditions, reducing the ensemble range without increasing the number of events outside the ensemble range could be viewed as improving the quality of the ensemble. The ensemble could be generated in many ways that would affect this range, such as the weighting methods previously discussed (e.g. Ruosteenoja et al., 2007). In data sparse environments, an event that did not occur at a station, but was well represented in a gridded dataset, would produce a reduction in performance when spatially aggregated. Similarly, distributed comparisons are subject to event spatial positioning: an event in a coarse grid may not be spatially positioned to correlate correctly to a climate station without some degree of spatial aggregation. In data sparse environments, a storm event (such as a convective thunderstorm) of small spatial extent, captured by a ground-based climate station, would be over represented by ECCC and AHCCD datasets for semi-lumped and lumped aggregations. A storm event of large spatial extent, which occurs between ground-based climate stations, would receive no representation by ECCC and AHCCD, but would appear over represented by the gridded datasets if compared to ECCC or AHCCD. Therefore, it is necessary to further examine the performance of the ensemble and how aggregation affects the propagation of input data uncertainty into a hydrological model. It is, perhaps, more useful to consider the absolute minimum and maximum ensemble realizations for the LNRB than would be for the SRB. When gauge networks are sparse, it is more likely that storms occurred in an un-gauged area: this increases the uncertainty in the spatially

aggregated AHCCD, or ECCC, timeseries. Therefore, to represent that uncertainty, lower likelihood ensemble realizations increase the ensembles ability to envelope the observed uncertainty. Selecting the absolute maximum and minimum ensemble realizations may over represent uncertainty in the SRB, as the relatively denser climate station network, in comparison to the LNRB. It is, however, not recommended to utilize only the mean ensemble realization. While it may show generally high performance (e.g. Lilhare et al., Accepted), performance metrics assume uncertainty free observed data. It is known that observed climate data contain uncertainty (e.g. Mekis and Vincent, 2011). Attempting to select a single ensemble realization, which is most similar to observed climate data, would be an attempt to match a single realization of observed data uncertainty and assume that realization is true. Therefore, the inclusion of additional ensemble realizations is important for representing the range of uncertainty for an area.

RMSE values for the ensemble minimum and maximum (Figure 6b) reveal that error was not significantly larger on any particular day than that of lower-performing years for each gridded dataset. Therefore, while performance is low for the ensemble minimum and maximum, it is sourced from a consistent under or overestimation of observed conditions, rather than temporally dependent bias changes. The exception was summer, where higher RMSE values suggest storm events may have been extended or reduced. It is desirable to generate an uncertainty envelope that does not introduce error. Error would include clearly incorrect events in the ensemble realizations; for each day, the ensemble minimum consistently underestimated precipitation by an average 2-4 mm and the ensemble maximum consistently overestimated precipitation by an average of 2-4 mm. This represents uncertainty in the simulation of precipitation. It would, however, be an error for the ensemble minimum to suggest no convective storms had ever occurred. Similarly, it would be an error for the ensemble maximum to suggest large basin scale convective storms persisted for the duration of the summer seasons. The reasonable RMSE values presented in Figure 6b suggest that the uncertainty envelope, while not

probable, is possible at low likelihoods. The ensemble minimum and maximum highlighted issues surrounding timing and spatial consistency. If a storm event is not correctly timed by a gridded dataset, each representation of that event will be captured by the maximum, but missed by the minimum. This method for generating minimum and maximum ensemble members represents the extreme uncertain range of the ensemble, although it would represent the lowest likelihood realizations of the uncertain range. Similarly, small differences in spatial positioning of storm events would cause the maximum to overestimate an event's spatial extent and the minimum to underestimate an event's spatial extent; these issues contribute to the extreme PBIAS values seen in Figure 6c. Distributed comparisons generally show more improvement than the basin comparison (i.e. Figures 3 to 6).

Small overestimation of the RX1 values of the ensemble maximum at lumped and semi-lumped aggregations suggests differences in spatial positioning of events (Figure 7). Gridded datasets disagreeing in the spatial positioning of an event would cause the ensemble maximum to take all representations of the event at a single time step, which may cause an unrealistic increase in the spatial extent of an event. This result is supported by the improvement in the representation of RX1 values by the distributed comparison, which is not sensitive to an increase in the spatial extent of an event beyond the closest grid point. Large overestimation in RX5 values consistent between lumped, semi-lumped and distributed comparisons suggests that differences in event timing did exist. This caused the ensemble maximum to represent a particular event for a longer temporal period than what actually occurred.

Ensemble minimum, mean, and maximum should ideally compare well with observations; however, the goal of the ensemble is to represent the uncertainty generated by simulating precipitation, near and between climate gauges. The ensemble minimum should be consistently drier than the minimum of the observations, and the ensemble maximum wetter, so that the uncertainty envelope brackets all possible values from station records. An ideal ensemble would bracket the observed value for each day for the

full temporal period selected without having an unrealistically wide envelope (Figure 7). This concept is common in the literature for hydrologic modeling: Shafii et al., (2015) discussed these qualities in terms of reliability and sharpness. Reliability is defined here as the number of observed values within the ensemble envelope with respect to the total number of values. Sharpness is defined as the width of the envelope. Shafii et al., (2015) presents a methodology that attempts to maximize reliability with the narrowest envelope possible. In terms of reliability and sharpness as defined by Shafii et al., (2015), selecting the absolute maximum and minimum ensemble realizations was equivalent to maximizing reliability without consideration of sharpness. There are many studies that define an ideal uncertainty envelope similarly to Shafii et al., (2015) (e.g. Zhou et al., 2016). An argument against this definition is that a wide envelope is undesirable. Observed data are not uncertainty free, therefore, an uncertainty envelope should not always be considered desirable because it is narrow. This concept of quantifying the quality of an ensemble uncertainty envelope is common in hydrologic modeling uncertainty studies, but less so for the generation of meteorological data ensembles.

3.7.4. Ensemble Reliability

The ensemble generally bracketed small events well with the exception of zero precipitation days, which were sometimes not captured by the ensemble. For the ensemble to bracket a day of zero precipitation, the ensemble minimum must have zero precipitation occur at all grid points within an aggregated area. If any grid cell produced precipitation, the ensemble would be considered not to bracket that day. This is an undesirable condition; ground-based climate stations did not have sufficient coverage to ensure that there would be no precipitation event that would not have been observed by at least one station. It is more likely for the three stations in the LNRB to all observe zero precipitation than would be for zero precipitation to occur at any of the thousands of grid points near basin boundaries where no observations were collected. Considering the enveloping of zero precipitation days is likely to lead to a

penalty being applied to the ensemble for reasonably representing precipitation occurring between the sparse ground-based climate stations. Events above the 50th percentile were considered the focus of the ensemble, as this was generally a percentile that zero precipitation days were not considered for the lumped and semi-lumped aggregations; this is why Figure 8b, d, and f presents events above the 50th percentile captured by the uncertainty envelope of the ensemble. Events above the 50th percentile represent 95% of total precipitation in an average year; most zero precipitation days were not included. An important note is that the representation of zero precipitation events by the ensemble was better for the distributed comparison: since the ensemble envelope was generated from a single grid point in that case, there were no assumptions made about the existence of precipitation beyond a ground-based climate station. An average of the seven major basin aggregations showed nearly 900 zero precipitation days outside the ensemble envelope, whereas an average across all climate stations in the distributed comparison showed only 150 zero precipitation days outside the ensemble.

The SRB had the lowest uncertainty (Figure 8d), while the RRB had the highest. This was due in part to the SRB having the highest ground based climate station density. The RRB, however, only had three ground based climate stations at the northern side of the basin. The smoothing effect of spatial aggregation that generally reduced uncertainty is also dependent on data density. This suggests that the chosen spatial aggregation of a hydrologic model should also then depend in part on data density.

One weakness of the ensemble was a tendency to underestimate larger events even at the lumped spatial aggregation (Figure 8b). The performance of the ensemble was most affected by spatial aggregation; performance improved for lumped comparisons but was limited to extreme event magnitudes. The ensemble minimum was generally dry and the ensemble maximum wet for the distributed, semi-lumped, and lumped spatial aggregations with respect to AHCCD. The ensemble minimum and maximum represented the extreme range of uncertainty for the ensemble, although they

are the least likely realizations of the ensemble. The minimum and maximum ensemble realizations, however, are best for providing the maximum possible uncertainty for input propagation, since the goal of the BaySys project is to quantify the total possible uncertainty rather than the most likely realizations. The range of ensemble uncertainty, however, still did not bracket a relatively large number of events above the 50th percentile, especially for the distributed comparison. A possible way to improve the ensemble minimum and maximum, meaning narrow the range without reducing the number of events represented, would be to use a similar methodology to the reliability and sharpness metrics sometimes used in hydrologic uncertainty studies (e.g. Shafii et al., 2015; Zhou et al., 2016; Tweldebrahn et al., 2018). Some modification to the reliability and sharpness methodology is, however, suggested: since there is no adjustment made for the expected uncertainty in observed data with this methodology, it implicitly assumes that the observed data are near true, which is known to be incorrect. Additionally, the reliability and sharpness methodology is used to select or reject simulations from thousands of total simulations, as each selected simulation is associated with a single product. Therefore, additional consideration of ensemble methods such as Bayesian model averaging (e.g. Yang et al., 2012) or storm multipliers (e.g. Ajami et al., 2007) would offer benefits to generating ensemble realizations, as opposed to the ensemble in the present study that would have had only five members for selection if reliability and sharpness were applied without augmentation.

Overall, the ensemble mean, and the absolute ensemble maximum and minimum ensemble realizations represent the average and the extreme range of uncertainty possible for propagation through other models in the BaySys project. This methodology provided two low likelihood ensemble realizations at the wet and dry sides of the ensemble, as well as a higher likelihood ensemble mean. The ensemble mean is expected to produce the best performing simulations when used as input to models in the BaySys project. However, the ensemble mean should not be the only ensemble realization given focus. Including the ensemble minimum and maximum realizations is important for the representation of

uncertainty. While best for meeting requirements for the BaySys project, it is likely desirable to replace the absolute minimum and maximum ensemble realizations with a narrower uncertainty envelope. The minimum and maximum could be generated based on a methodology incorporating elements similar to those from the reliability and sharpness metrics.

3.8. Acknowledgments

Thanks to University of Manitoba, Manitoba Hydro, and partners funding through the Natural Sciences and Engineering Research Council of Canada through funding of the BaySys project. Thanks to makers of the gridded datasets used in this study. Thanks to Environment and Climate Change Canada for providing access the ANSUPLIN dataset and the AHCCD dataset, and SMHI for providing access to the Hydro-GFD dataset.

Chapter 4 : Cumulative Effects of Multiple Uncertainty Sources on Flow Predictability in a Hydrologic Modeling Environment

(Pokorny, S., Stadnyk, T., Ali, G., Déry, S., Tefs, A., Holmes, T., Lilhare, R., Koenig, K.)

4.1. Abstract

The treatment of uncertainty in rainfall-runoff models is generally centered around the concept of producing the narrowest possible uncertainty envelope that still brackets all observed values. It is common in the literature to not consider all sources of uncertainty, namely input, structural, parameter, and output uncertainties, simultaneously. Even when hydrologic modeling studies do consider multiple uncertainty sources, they often do not quantify the relative contributions of each source to total uncertainty. This paper presents a methodology for the relative partitioning of uncertainty towards the estimation of total streamflow uncertainty, and the estimation of the cumulative effects of uncertainty propagation towards streamflow predictability. Sources of uncertainty were assessed with respect to an expected improvement in reproducing observed data based on computational time invested. An ensemble of three hydrologic models was forced by three realizations from an input data ensemble that was derived from five gridded climate datasets. Input and structural uncertainty were found to be the largest relative sources of uncertainty. Considering all sources of uncertainty was often the best for representing observed data; however, that approach was computationally intensive. If a choice needs to be made regarding which source of uncertainty to deal with in priority, structural uncertainty would be most valuable for consideration in an operational environment with access to limited computational resources, followed by input and parameter uncertainties.

4.2. Introduction

Hydrologic models often target the prediction of streamflow through a simplified representation of a physical environment that is driven by inputs, which are also simplifications of observed meteorological conditions (Pechlivanidis, 2011). Hydrologic regimes are complex and rapidly changing; therefore, hydrologic models are often relied upon to make operational predictions. Hydrologic model output, however, suffers from uncertainty generated by simplifying the selected physical environment (Uusitalo et al., 2015). Uncertainty is defined as the realistic range to which an exact value for a given variable cannot be determined but can be represented by a likelihood (Uusitalo et al., 2015). By simplifying the physical environment, the assumptions made in a hydrologic model are wrong at any particular location, but can be representative of average conditions. Since processes that occur in a physical environment are approximate in spatially aggregated areas, uncertainty is introduced to the model via simplifying assumptions. Without addressing hydrologic modeling uncertainty, the amount of valuable information that can be gained from hydrologic simulation may be compromised. Demargne et al., (2014) show that forecasted streamflow would be poorly represented by a single simulation, but more often captured observed streamflow within the envelope of uncertainty. Therefore, Demargne et al., (2014) increase the value of a forecast by more often representing observed occurrences within an uncertainty envelope. Hydrologic modeling involves four broad sources of uncertainty: input, structural, parameter, and output uncertainties (Matott et al., 2009; Pechlivanidis, 2011; Uusitalo et al., 2015). In the specific context of precipitation-runoff models (hereafter referred to as hydrologic models), input uncertainty is the uncertainty associated with model forcing (e.g. precipitation). Output uncertainty is the uncertainty associated with observed data a model is calibrated (or validated) against. Parameter uncertainty refers to hydrologic parameter selection for the mathematical equations used to represent hydrological

processes. Structural uncertainty refers to model spatial aggregation: lumped, semi-lumped, or distributed (e.g. Khakbaz et al., 2012), and hydrologic process representation or simplification (e.g. Priestley-Taylor equation for evapotranspiration (Priestley and Taylor, 1972)). Parameter and structural uncertainties are inherently co-dependent but also tied to input uncertainty as well. There are numerous methods reported in the literature to quantify uncertainty in hydrologic modeling studies; Matott et al., (2009) provide a summary of some of these methods and associated tools described in the literature, and they discuss their relative popularity. Regardless of the quantification methods used, one important element to consider is that the four broad sources of uncertainty are interconnected through propagation. Propagation is how uncertainty from one modeling step affects the next, moving downstream. As a result, uncertainty sources can only be analyzed in a relative sense (e.g. Dams et al., 2015). Propagation of a particular source of uncertainty can be considered by sampling from the likelihood distribution of that source (Brown and Heuvelink, 2006). The current literature, however, largely does not address the cumulative effects of uncertainty and how cumulative effects impact the predictive capability of a hydrologic model (i.e., the ability of the model to adequately simulate streamflow within reasonable prediction or error bounds) (e.g. Ajami et al., 2007; Tweldebrahn et al., 2018).

Input data uncertainty is perhaps the simplest to analyze in a hydrologic modeling context, given that input data uncertainty can be analyzed independently from a hydrologic model (e.g. Pokorny et al., in prep). Hydrologic models are data intensive; observed ground-based climate station data often suffer from sparsity and paucity issues, therefore, hydrologic studies may consider interpolated or reanalysis gridded climate datasets as model input (e.g. Choi et al., 2009). In historical studies, a single dataset is often suggested/considered: it is selected based on an often subjectively set evaluation of error with respect to an observed data benchmark, usually ground-based climate station data (e.g. Choi et al., 2009; Eum et al., 2014, Wong et al., 2017). Some studies, however, suggest input data ensembles, which

are more representative of input uncertainty (e.g. Rapaić et al., 2015, Gbambie et al., 2017, Lilhare et al., Accepted). An ensemble is defined here as the selection of multiple models, whether climate models (e.g. Semenov et al., 2010), hydrologic (e.g. Ajami et al., 2007), or other model types such as statistical or physically-based hydrologic models (e.g. Demargne et al., 2014). The argument favoring the use of input data ensembles is supported by Pokorny et al., (in prep), who showed how a single gridded dataset would misrepresent input uncertainty when used as input to a hydrologic model. Pokorny et al., (in prep) further show how input uncertainty is connected to model spatial aggregation, which affects the magnitude of uncertainty subject to propagation. There are many studies that consider model input uncertainty, more often for climate change studies (e.g. Dams et al., 2015; Karlsson et al., 2016). These studies generally find future climate projections to be a dominant source of uncertainty over model structure. However, fewer studies consider historical input uncertainty (e.g. Nikolopoulos et al., 2010).

Parameter uncertainty is assessed by sampling parameter values in a user-defined range. Studies that focus on parameter uncertainty often consider uncertainty through parameter likelihood estimation (Beven and Binley, 1992), which selects behavioral parameter sets for the development of likelihood distributions. Behavioral parameter sets are those that meet or exceed an often subjectively defined model performance criterion (e.g. Stedinger et al., 2008; Li et al., 2010; Li and Xu, 2014; Shafii et al., 2015). Parameters that are sensitive (e.g. Razavi and Gupta, 2015) but do not have defined distributions may suffer from identifiability issues. If the parameter distribution varies temporally, the overall distribution for the full temporal period will appear poorly defined. The creation of a parameter distribution is based on the definition of a likelihood function, which suggests a higher likelihood for when a model simulation is more similar to the observations the model is attempting to replicate (Beven and Binley, 1992). An overall parameter distribution is generated through binned likelihoods of selected behavioral parameter sets, in which the likelihood is quantified utilizing the full modeled temporal period. The literature shows a variety of likelihood and pseudo-likelihood functions that have been

implemented (e.g. Shafii et al., 2015). Identifiability-focused studies utilize uncertainty frameworks, such as the Dynamic Identifiability Analysis (DYNIA) framework, to detect parameters with temporally dependent distributions (e.g. Wagener et al., 2003; Abebe et al., 2010; Merz et al., 2011). Rather than generating a single likelihood distribution, the DYNIA framework generates likelihood distributions in moving temporal windows. This allows time periods of higher and lower identifiability of a parameter to be detected. If a parameter has a well-defined parameter distribution under some circumstance, such as high flows, but is not well defined under other conditions, the parameter has identifiability issues.

Structural uncertainty is assessed by varying model structure (e.g. Dams et al., 2015). Studies such as Muhammad et al., (2018) examine structural uncertainty within the same model by varying the spatial aggregation of the study area. Some studies also compare different numerical representations of a hydrologic process within the same model such as Tasdighi et al., (2018), who compared two infiltration methods with the same model framework. Another approach to considering structural uncertainty is to select an ensemble of models. Wi et al., (2015) tested the effects of lumped, semi-lumped, and distributed structures and found uncertainty was not significantly increased by the added complexity of a distributed model, as opposed to a lumped or semi-lumped model.

Output uncertainty is often mentioned in theoretical papers, but rarely quantified in hydrologic modeling studies. McMillan et al., (2010) used rating curve deviation to estimate flow record uncertainty by varying the rating curve, constructing a cumulative distribution function (CDF) based on the range of estimated flows, and representing uncertainty using the 5^{th} and 95^{th} quantiles. The Water Survey of Canada specifies $\pm$ 5% uncertainty in flow estimates for open water conditions (Environment Canada, 1980); however, Dingman (2014) cites errors up to $\pm$ 10%. Therefore, even if a perfect model were possible, its performance should ideally be limited to the uncertainty of the observed flow against which it is calibrated. Output uncertainty studies similar to McMillan et al., (2010) account for uncertainty in

the measurement of output data; however, more common in the literature is the limits of acceptability (LOA) approach (Beven, 2006). The LOA method selects behavioral parameter sets by choosing those within the uncertainty limits of the observed timeseries. The LOA criteria often need to be relaxed, as producing a simulation entirely within an output uncertainty envelope has rarely been reported in the literature. It is not appropriate to expect a model to be less uncertain than the data it is calibrated against. Therefore, studies in the literature often find few, if any, models that produce simulations entirely within the LOA, even when the LOA is widened slightly beyond the output uncertainty envelope (e.g. Teweldebrhan et al., 2018). Rather than reject those models that only exceed the LOA on a limited number of time steps, the LOA is relaxed to allow a subjectively set number of simulated time steps to exceed the LOA so model realizations may still be accepted as behavioral (e.g. Teweldebrhan et al., 2018). Similar to the LOA are the reliability and sharpness metrics (Li et al., 2010; Shafii et al., 2015; Zhou et al., 2016). Reliability is calculated by selecting a model realization, then adding additional model realizations that increase the number of observed events bracketed. A penalty is applied to widening the envelope of simulated output, which is referred to as sharpness. Model realizations are accepted if they improve reliability more than they decrease sharpness. A common problem with output uncertainty studies is deciding on limits of uncertainty: wider envelopes should not necessarily be considered a penalty (i.e. sharpness), and uncertainty propagated through the model must exceed that of the observed flow record for the period of calibration. There is no general consensus in the literature suggesting using the reliability and sharpness metrics over the LOA method, as both require some subjective decision making.

Few studies consider all sources of uncertainty (e.g. Demargne et al., 2014), or focus on propagation (e.g. Brown and Heuvelink, 2006; Ajami et al., (2007); Nikolopoulos et al., 2010; Mei et al., 2016). If, for example, only a single model structure is considered, an assumption is made that the selected model is sufficient to accurately represent the hydrological processes in the region of study. The present study

attempts to address the aforementioned gaps by using both hydrologic model and input data ensembles to: 1) evaluate if parameter uncertainty and identifiability are seasonally dependent; 2) assess if model performance, quantified using traditional metrics, scales up or down with sub-basin area; and 3) quantify uncertainty sourced from input data, parameter, and model structure uncertainties as they propagate into overall uncertainty. We relatively partition the effects of each source of uncertainty and assess their relative cumulative impact on streamflow prediction with reference to an estimated envelope of output uncertainty. A challenge with uncertainty analysis is high computational demand; computational budgets are often limited in an operational environment. Therefore, an auxiliary goal of the present study is also to establish a hierarchy - from most to least valuable - of the sources of uncertainty to consider for the improvement of predictability, in case of a limited computational budget.

4.3. Study Area

The region of study is the Lower Nelson River Basin (LNRB) in Northern Manitoba, Canada. The LNRB is a basin of approximately 90,500 km^2 at the downstream end of the approximately 1,400,000 km^2 Nelson Churchill Watershed (NCW) (Figure 9). The LNRB is characterized by a sub-arctic climate and low-relief terrain largely covered by forest and wetlands, with many lakes due to the low topographic relief. The LNRB is a data-sparse basin with limited climate stations. Water from Lake Winnipeg enters the LNRB via the Nelson River on the southwest edge of the basin, flows through the Jenpeg Generating Station (Figure 9, Water Survey of Canada gauge ID 05UB009) and the free flowing East Channel 05UB008, and is augmented by flow diverted from the Churchill River (Notigi control structure) via the Burntwood River on the northwestern edge of the basin (Figure 9, Water Survey of Canada gauge ID 05TF710). There are six Manitoba Hydro-owned generating stations in operation within the LNRB: the Jenpeg (Water Survey of Canada gauge ID 05UB009), Kelsey (Water Survey of Canada gauge ID 05UE005), Kettle (Water Survey of Canada gauge ID 05UE006), Longspruce (Water Survey of Canada gauge ID 05UE007),

Limestone (no public streamflow gauge available), and Wuskwatim (no streamflow gauge available) generating stations act as points of regulation. Additionally, there is the Keeyask Generating Station (currently under construction), which has impacted the streamflow record at the Keeyask location (not publically available and not used in this study) during construction but is not yet operational. The focus of this study is on a historical period, therefore, only stations with historical hydrometric data were considered. The data recorded at the stations presented in Figure 9 are publically available with the exception of the proprietary Notigi control structure and Limestone Generating Station hydrometric gauges (Water Survey of Canada Notigi gauge ID 05TF710), which are owned and operated by Manitoba Hydro. LNRB hydrometric gauge locations are presented in Appendix B, Table B. 1.

The LNRB has a sub-arctic climate with cold winters: January is the coldest month with a 30-year average temperature (1981-2010) of -23.9^{o}C and -24.4^{o}C at Thompson and Gillam, respectively (Appendix B, Figure B.1). Summers are cool, with July being the warmest month with a 30-year average temperature (1981-2010) of 16.2^{o}C and 15.8^{o}C at Thompson and Gillam, respectively. Precipitation (1981-2010) is highest in July at 80.9 mm and 78.6 mm for Thompson and Gillam, respectively. Total annual precipitation for Thompson and Gillam was 509.0 mm and 496.4 mm, respectively, in the 1981-2010 period, of which 43% occurred in summer (JJA).

The largest river in the LNRB is the Nelson River, which is subject to regulation at multiple locations. Within the LNRB, the Nelson River accumulates flow from several smaller unregulated rivers, as well as flow diverted through the Notigi control structure from the Churchill River. Peak flows on unregulated rivers in the LNRB are generated by snowmelt runoff in late spring or early summer, often occurring in May or early June. The Grass River, however, generally has later peak flows due to flow attenuation in wetlands and lakes within the basin. Regulated flows on the Burntwood River downstream of the Notigi control structure and on the Nelson River are generally more uniform throughout a year, as regulation

increases or decreases flow based on factors such as energy demand, reservoir elevation constraints, optimization of resources, and others. For an in-depth review of the LNRB, see Holmes (2016).

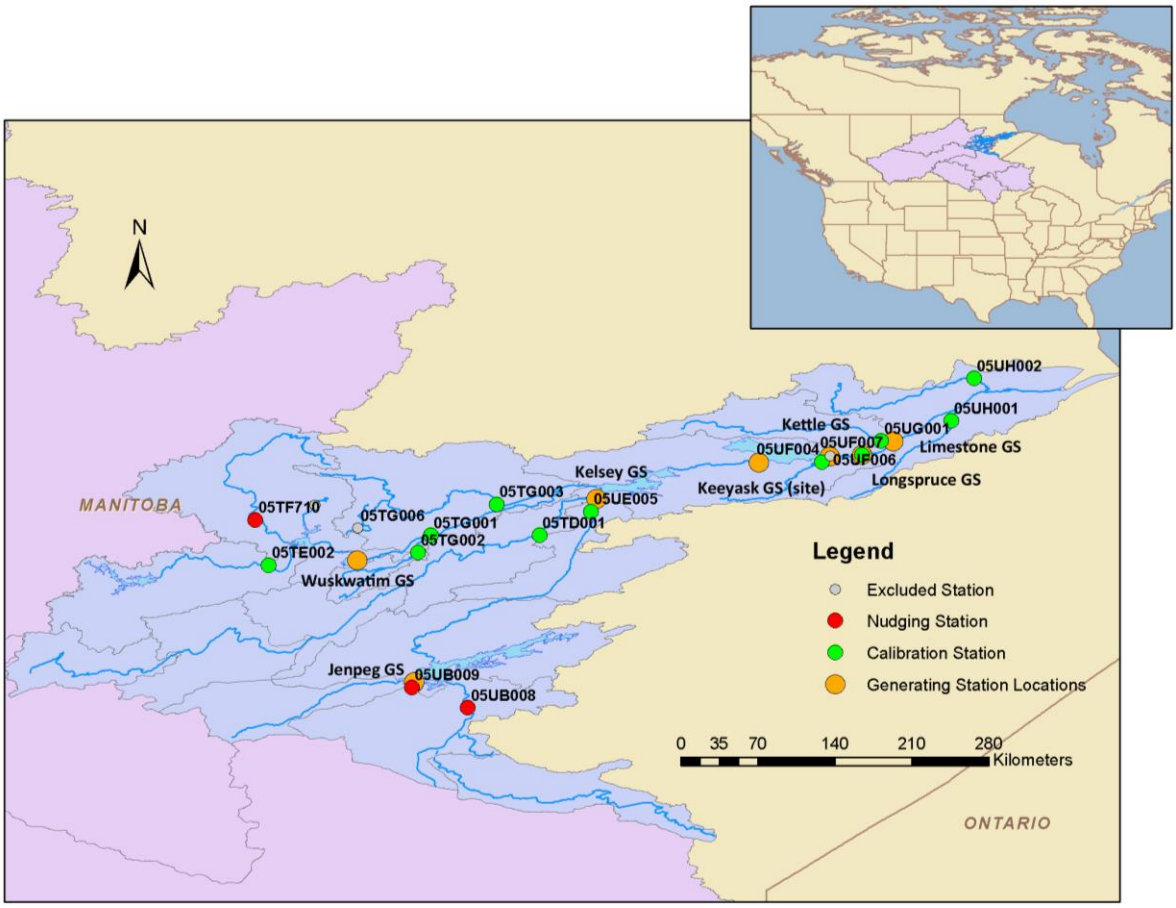

Figure 9: The Lower Nelson River Basin with available hydrometric stations with historical data available, including the locations of generating stations, and the currently under construction Keeyask generating station.

4.4. Input Data

Model meteorologic input data for this study consists of daily timeseries of precipitation and temperature from 1981-2010. Precipitation and temperature data are gathered from the ensemble of gridded climate datasets presented in Pokorny et al., (in prep). The ensemble consists of five gridded datasets: the North American Regional Reanalysis (Mesinger et al., 2006), ERA-Interim (Dee et al., 2011), Watch Forcing Data ERA-Interim (Weedon et al., 2014), GFD-Hydro (Berg et al., 2018), and Natural

Resources Canada's ANUSPLIN (Hutchinson et al., 2009). The first four gridded datasets were bi-linearly interpolated to a 10 km grid based on ANUSPLIN. Three precipitation realizations were considered: minimum, mean, and maximum. Minimum (maximum) precipitation was generated by selecting the minimum (maximum) value of all five gridded datasets for each grid cell at each time-step (i.e., daily). The mean precipitation realization was generated by arithmetic mean of all five datasets with an equal weighting. Regarding other temperature inputs, only the mean temperature realization (daily minimum and maximum temperatures, and daily mean temperature) was considered as uncertainty was noticeably lower in the estimation of temperatures, relative to the uncertainty in precipitation (Pokorny et al., in prep). For an in-depth review of the climatology and physiography of the LNRB, see Smith (2015), Holmes (2016), and Lilhare et al., (Accepted).

An important consideration regarding the minimum and maximum precipitation realizations is their likelihood of occurrence (Pokorny et al., in prep). There is an underlying likelihood distribution of precipitation magnitude for each grid cell at each time step. By selecting the minimum (maximum) value for each grid cell at each time step, a low likelihood precipitation realization is always considered at all grid cells for all time steps. Nikolopoulos et al., (2010) utilized 20 precipitation realizations and stated that further sampling was not applied to hydrologic simulation due to computational limits. Nikolopoulos et al., (2010) selected precipitation realizations that ranged from the 5^{th} to 95^{th} percentile of precipitation volume generated from perturbing the precipitation products. The present study considered a larger model domain, multiple hydrologic models, and more than 65 times more time steps than Nikolopoulos et al., (2010). Therefore, the present study aims to adjust precipitation realization sampling to be representative of sampling a dry, low likelihood (e.g. at or less than the 5^{th} percentile), a high likelihood (e.g. 50^{th} percentile), and a wet low likelihood (e.g. at or greater than the 95^{th} percentile) set of precipitation realizations; although, as Pokorny et al., (in prep), the wet and dry realizations utilized may be more extreme than the 5ht and 95 percentiles. This method of sampling offers insight

into hydrological modeling input uncertainty subject to propagation, although the minimum and maximum precipitation realizations are of a low likelihood of occurrence. The BaySys project requires these low likelihood precipitation realizations for propagation of uncertainty towards total possible uncertainty bounds on flow in the Nelson River. The low and high flow cases are relevant for assessing the sensitivity of ocean models receiving modeled freshwater streamflow as input. The assessment of sensitivity for the BaySys project is concerned with sensitivity of a wide range of possible flows within the envelope of total uncertainty, regardless of their likelihood of occurrence. This decision also provides information towards understanding the required sampling needed to determine the underlying likelihood distributions for each grid cell at each time step. A similar method for representing the range of uncertainty was used by Dams et al., (2015), in which the range between the minimum and maximum monthly discharge was used in part to suggest a dominant source of uncertainty.

Since the LNRB is a downstream basin in the NCW, the flow from upstream basins must be added to the basin as an external flow source. The flow records from the regulated Jenpeg generating station (05UB009) and regulation effected East Channel (05UB008), and the Notigi control structure (05TF710) are, therefore, added as model forcings to the LNRB hydrologic models. The uncertain range of input flows is considered through output uncertainty, which is discussed later in the methodology section.

4.5. Hydrologic Models

An ensemble of hydrologic models was selected to account for structural uncertainty by varying both spatial aggregation and hydrologic process representation. The models selected to represent the LNRB included the HEC-HMS, HYPE, and WATFLOOD hydrologic models (Table 3).

Table 3: Summary of key structural differences between the selected hydrologic models

Hydrologic Model	Hydrologic Processes	Selected methods	Comments
HEC-HMS (Sagan, 2017)	Infiltration	Soil Moisture Accounting	Semi-lumped sub basin model with basin sizes ranging from 360 – 12,000 km^2
	Evapotranspiration	Priestly Taylor	
	Snowmelt	Temperature Index	
	Routing	Muskingum	
HYPE (MacDonald et al., under revision)	Infiltration	HYPE default infiltration	Semi-lumped sub basin model with basin sizes generally around 400 km^2
	Evapotranspiration	Priestly Taylor	
	Snowmelt	Temperature + Radiation Index	
	Routing	Lag, Recession, and Attenuation	
WATFLOOD (Holmes, 2016)	Infiltration	Phillips Formula	Gridded model with 10 km grid spacing
	Evapotranspiration	Hargreaves	
	Snowmelt	Temperature Index	
	Routing	Storage routing	

A total of 23 HYPE parameters, 51 WATFLOOD parameters, and 63 HEC-HMS model parameters were selected for sensitivity analysis and parameter uncertainty assessment; parameter selection was made based on the sensitivity analyses conducted by recent model developers and available computational resources (Holmes, 2016; Sagan, 2017; MacDonald et al., under revision). Some parameter grouping was done which is expected to affect parameter sensitivity, however, the parameter groupings were required to ensure computational feasibility; parameter groupings were considered acceptable by the recent model developers (Holmes, 2016; Sagan, 2017; MacDonald et al., under revision). Some adjustments were made to the HEC-HMS model as Sagan (2017) was focused on seasonal extremes by adjusting parameter ranges (widening) beyond what Sagan (2017) presented to account for more variability across all seasons. Additionally, methods for evapotranspiration and routing were updated to the Priestly Taylor and Muskingum routing methods respectively. Parameters associated with hydrologic method changes which were not part of Sagan (2017) were included in this study. In addition to sampling done for sensitivity analysis, which will be further discussed in Section 4.6.2, 2,300, 5,100, and 6,300 parameter samples were generated for each of the HYPE, WATFLOOD, and HEC-HMS models, respectively. The additional sampling quantity was informed by the VARS sensitivity analysis, which

showed that each model required increasing sampling quantities based on the selected number of selected model parameters. These additional samples were held constant for each precipitation input. A total of 27,700, 61,300, and 75,700 hydrologic model runs were completed for HYPE, WATFLOOD, and HEC-HMS, respectively.

4.6. Methodology

4.6.1. Data Preparation

Three precipitation input datasets were utilized, including the minimum, mean, and maximum precipitation ensemble realizations produced by Pokorny et al., (in prep). Only the mean temperature ensemble realization and the unaltered observed flow records for the Jenpeg generating station and the Notigi control structure were used in combination with each precipitation realization. HEC-HMS required daily dewpoint temperatures for the Priestly-Taylor method; dewpoint temperature timeseries were estimated using Equation 4.1, developed by Hubbard et al., (2003):

$$T_d = \alpha(T_n) + \beta(T_x - T_n) + \gamma\left(P_{Daily}\right) + \lambda \qquad [4.1]$$

where T_d is the dewpoint temperature (^{o}C); T_n is the daily minimum temperature (^{o}C); T_x is the daily maximum temperature (^{o}C); P_{Daily} is the daily precipitation (mm); and α, β, γ, and λ are regression constants.

For the LNRB, the HEC-HMS model, originally developed by Manitoba Hydro and further developed by Sagan (2017), uses point gauge climate data generated by aggregating climate data grid points within, and along, basin delineations. Input data for the HYPE model, recently developed in part by Macdonald et al., (under revision), is generated by assigning nearest climate data grid points to sub basins (SMHI,

2018). The input data for the WATFLOOD LNRB model, recently developed in part by Holmes (2016), match the 10 km gridded ensemble data.

4.6.2. Sensitivity Analysis

To address the first research question, a sensitivity analysis on varied temporal periods was conducted. A general sensitivity analysis utilizing the variogram analysis of response surfaces (VARS) framework was applied to the hydrologic model ensemble (Razavi and Gupta, 2015; Razavi and Gupta, 2016 a and b; Gupta and Razavi, 2017). VARS sensitivity analysis was applied seasonally and over the full period; seasons were defined as winter (December, January, and February), spring (March, April, and May), summer (June, July, and August), and autumn (September, October, and November). If parameters suffered from identifiability issues, then they would vary temporally in relative sensitivity and reliability. In the VARS framework, sensitivity is defined by the relative magnitude of change in a response variable associated with varying a parameter value. Reliability is defined by bootstrap resampling subsets of model runs and counting the number of times the same relative sensitivity order was generated. If a parameter is consistently the most sensitive in the subsets of model runs, then the relative sensitivity is reliable. Sensitivity is presented as a unit-less value relative to other parameters included in the sensitivity analysis; parameters are ranked from least to most sensitive based on their relative sensitivity. Sensitivity values are not comparable between analyses; reliability, however, is comparable as it is a percentage of occurrences of the same relative sensitivity rank during bootstrap resampling.

Parameter sampling utilized 100 star centers for generating VARS sensitivity rankings for all models (Razavi and Gupta, 2016b). A variogram resolution of 0.1 was selected, to be consistent with examples provided by Razavi and Gupta (2016b). Due to the computational demand of the VARS framework, VARS was run with a single precipitation ensemble realization of NARR, selected at random. Only the results for “iVARS 5” are presented (Haghnegahdar et al., 2017). The “iVARS” variable is generated by

integrating over varied ranges of the directional variograms. For example, iVARs 1, with a variogram having a resolution of 0.1, would represent the performance metrics sensitivity to changes in parameter values by comparing parameter values 0.1-0.9 to parameter values 0.2-1 (parameter values are scaled by values ranging from 0-1 based on their respective parameter search ranges). The value for iVARS 5 would compare ranges 0.1-0.5 and 0.6-1.0. Values for iVARS beyond iVARS 5 start losing information, for example iVARS 6 compares the ranges 0.1-0.4 to 0.7-0.1; the values for 0.5 and 0.6 would no longer be included in the sensitivity estimate. Therefore, iVARS values can be generated past iVARS 5, but it is not recommended as it does not utilize all parameter information. The iVARS values represent a parameter's sensitivity to changes of different magnitudes, in which iVARS 1 represents sensitivity to small parameter value changes, and iVARS 2-5 each represent sensitivity to progressively larger changes in parameter values. The values for iVARS 5 are presented in this study as they are suggested to be the most representative of parameter sensitivity (Razavi and Gupta, 2016b; Haghnegahdar et al., 2017).

Each model had a different number of total samples determined by the 100 star centers, with HYPE having 20,800 samples, WATFLOOD having 46,000 samples, and HEC-HMS having 56,800 samples. VARS star sampling generates directional variograms in the dimension of each parameter, meaning that once a parameter's directional variogram is sampled for a star center, it is held constant until being varied for the next star center. Therefore, while each model had a different number of samples, they had a consistent number of star centers. The VARS star sampling creates a high density of sampling at one parameter value per star center. Therefore, additional sampling using orthogonal Latin hypercube sampling was included, which offers efficient uniform sampling (Tang, 1993). Uninformed priors were used for all hydrologic model parameters in this study; additionally, the use of more advanced sampling methods, such as the dynamically dimensioned search (DDS) (Tolson and Shoemaker, 2007), require a predefinition of success (i.e. a performance metric to optimize). The same set of parameter samples was applied to all three precipitation inputs, which allowed for insights into the evolution of parameter

distributions based on sampling input uncertainty. Therefore, if a bias in the sampling was to exist, such as less variation in the samples of a particular parameter's interactions with another, the same bias would be present for all three inputs, making the relative comparison more robust to oddities produced by lower sampling quantities. Advanced algorithms such as DDS do not guarantee that the parameter space will be explored in the same way for each input dataset. Therefore, orthogonal Latin hypercube sampling allowed for efficient sampling of uniform priors, while making the comparison of the effect of changing input datasets more robust. The limitation of using uniformed priors was that lower performance will be achieved by the models as more samples will be taken at low likelihood regions of the parameter space. The focus of this study is not on achieving a high quality calibration, but rather to ensure sufficient samples are taken to generate an estimate of parameter uncertainty in each model.

To generate a likelihood distribution for a parameter, repetitive samples, while other parameter variograms were generated, were filtered out, leaving only 1,000 effective samples per parameter. For this study, a non-uniform parameter distribution was considered to be evidence of a parameter having a likelihood distribution that showed notable deviation from the uniform prior that it was sampled from, thus suggesting a possible fit by a statistical distribution; no distribution fitting was conducted and the ability to fit a distribution was based on visual inspection. VARS star-based sampling generates a star center, parameters are then varied across the range of their directional variogram at the specified resolution (i.e. 0.1 in this study). While a parameter is being varied across its directional variogram, all other parameters are fixed at the star center. If 23 parameters are analyzed, after the first parameter is varied, it is held constant at the star center value while the other 22 parameters are varied. The high parameter sample concentrations at star centers creates a significant inequality in parameter sample weighting, this prevents the fitting of a statistical distribution (an example of a non-uniform sample created by VARS star concentration can be seen in Appendix B, Figure B.2).

Parameters were classified using four broad types: meteorological, land use, routing, and lake-based parameters. The parameter groupings were somewhat subjective; snowmelt based parameters such as HYPE's "fpsno corr", which is a broad correction factor for the fraction of potential evaporation used for snow sublimation in all land classes, may have also fit appropriately as a meteorological parameter. Primary parameter classifications were determined as suggested by recent model developers.

4.6.3. Uncertainty Analysis

Input and structural uncertainty were considered by varying the selected precipitation ensemble realization and the hydrologic model, respectively. It is important to note that flow through the Jenpeg generating station and the Notigi control structure were added to the ensemble of hydrologic models as an input. Similar to temperature input, the flow input was not sampled. There is, of course, uncertainty associated with both temperature and flow inputs, but it is expected to be lower than that of precipitation input. Dingman (2014) suggests that uncertainty is likely lower at points of regulation as the regulatory structure serves as a stable point for measurement and is less subject to morphological changes than a natural section of channel would be. Parameter uncertainty was assessed using the generalized likelihood uncertainty estimation (GLUE) methodology (Beven and Binley, 1992). Parameters were sampled from uniform prior distributions; behavioral parameter sets were used to generate parameter likelihood distributions. Selection of behavioral parameter sets is often subjective, however, less subjective selection criteria are a common topic of discussion in the literature (e.g. Stedinger et al., 2008; Li et al., 2010; Li and Xu, 2014; Shafii et al., 2015). The minimum and maximum precipitation realizations were selected to reflect the extreme range of uncertainty, representing the lowest likelihoods. Traditional methods for selecting behavioral parameter sets would reject simulations at the extreme range of likelihood, meaning the lowest likelihood, but highest risk samples are rejected, therefore, the top 10% best performing runs were selected for each realization (Shafii et al., 2015). The

pseudo-likelihood function was computed using the Kling-Gupta Efficiency (KGE) (Gupta et al., 2009) to assess model performance; the KGE was also used to assess VARS sensitivity. KGE scores were calculated for an equal weight average of all gauges and for each individual gauge while ignoring the others.

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_s}{\sigma_o} - 1\right)^2 + \left(\frac{\mu_s}{\mu_o} - 1\right)^2} \quad [4.2]$$

where r is the Pearson correlation coefficient between the simulated and observed timeseries, σ_s and σ_o represent the simulated and observed standard deviations, and μ_s and μ_o are the simulated and observed mean. The KGE has a range of [1 to -∞), with a value of 1 being a perfect score, a value of 0 meaning performance no better than using the timeseries mean value across all time steps, and a value less than 0 being worse than using the timeseries mean.

Finally, output uncertainty was considered by assuming the standard ice-off measurement uncertainty of ±5% suggested by the WSC (Environment Canada, 1980). Ice-on measurement uncertainty is difficult to quantify, as ice conditions vary throughout the ice-on period, and from year to year. Therefore, while it is likely a significant underestimation of uncertainty for ice-on conditions, the standard ice-off measurement uncertainty is applied to ice-on measurements.

The full historic climatological period from 1981-2010 was used for model calibration. Uncertainty methods were applied to each hydrometric gauge in the LNRB, presented in Figure 9. Unregulated gauged upstream areas range in size from 886 km^2 (Taylor River near Thompson, 05TG002) to 15,400 km^2 (Grass River above Standing Stone Falls, 05TD001). Regulated gauged upstream areas range from 18,500 km^2 (Burntwood River near Thompson, 05TG001) to approximately 1,400,00 km^2 (Nelson River at Long Spruce Generating Station, 05UF007). The variation in basin scale provides insight into the second research question by generating information on the relationship that basin scale has with model performance in the LNRB, as represented by the KGE.

4.6.4. Evaluation of Model Predictability

To determine a hierarchy for the broad sources of uncertainty, from most to least valuable, we assessed the relative effects on flow predictability for each partitioned source of uncertainty, propagated through to a simulated flow timeseries. The relative effect of each source of uncertainty was partitioned by holding other sources of uncertainty constant, for example, by varying precipitation realization for a single model structure (Dams et al., 2015). Cumulative distribution functions (CDF) were generated using the top 10% of model realizations, determined by the KGE score. Uncertainty was considered for the full 1981-2010 period. To generate CDFs, flows were averaged by day across the 30-year study period to generate a 30-year average annual hydrograph, which illustrated average conditions for each gauged area. The predictability was evaluated based on the CDF range of flows within the 10^{th} and 90^{th} quantiles, which were selected based on the operational flood forecasting practices of Manitoba Infrastructure (Personal communications with Fisaha Unduche of Manitoba Infrastructure), and the ±5% observed flow uncertainty.

4.7. Results

4.7.1. Temporal Dependence of Parameter Identifiability

VARS sensitivity results are presented in Figure 10 and indicated that relative sensitivity rankings changed seasonally. More relevant, however, were the reliabilities of those rankings, which are comparable across temporal periods and models. No parameter type was consistently dominant in the relative sensitivity rankings of any model (Figure 10), with the exception of HEC-HMS. HEC-HMS routing parameters were generally less relatively sensitive based on the KGE metric applied with VARS.

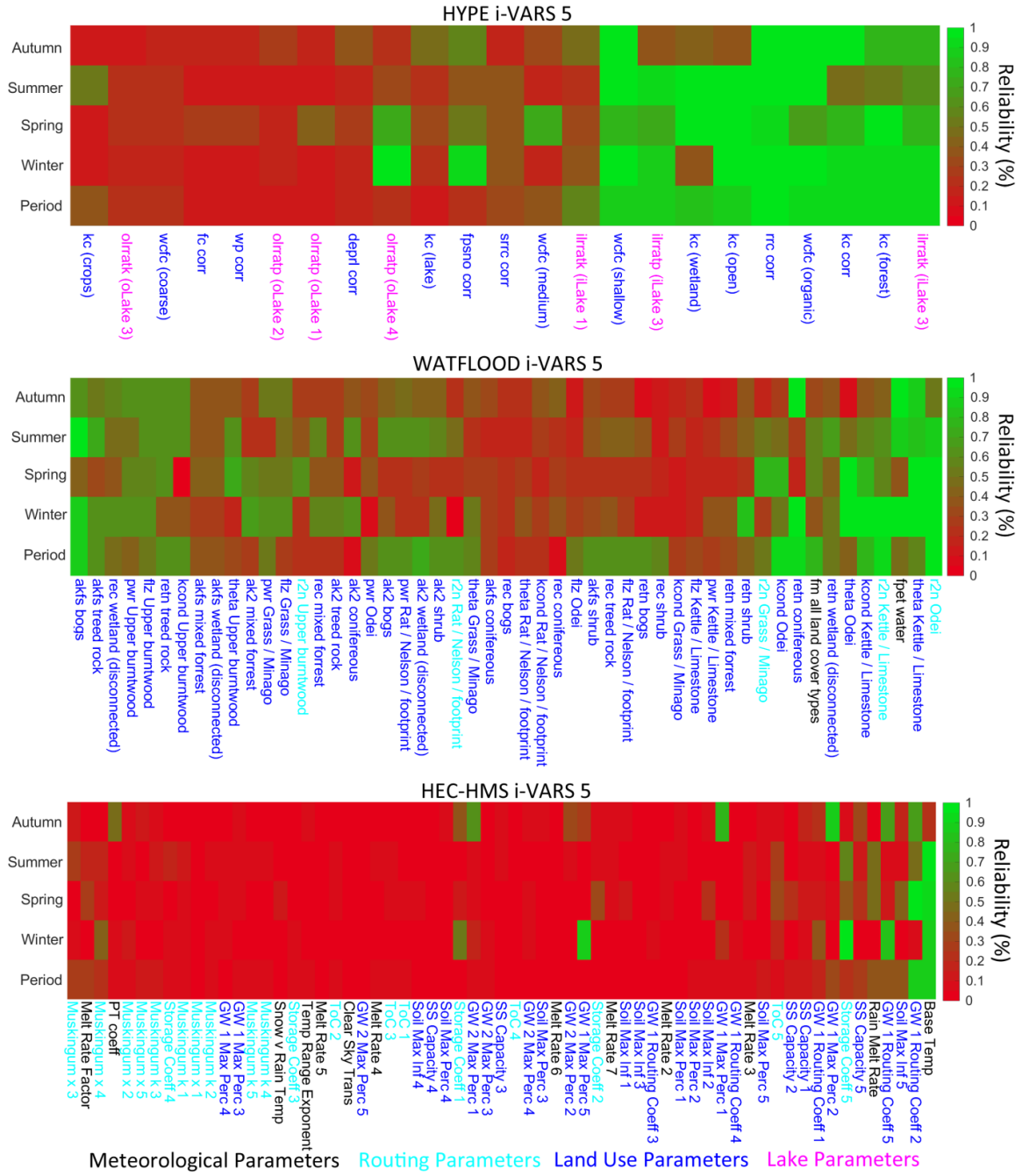

Figure 10: VARS parameter sensitivity reliabilities were ordered from least (left) to most sensitive (right), based on the period sensitivity (1981 - 2010). Variables are color-coded to reflect their category within the hydrologic model (i.e. parameters assigned based on land use classifications are labeled as land use parameters). Parameters are presented in Table B. 2, Table B. 3, and Table B. 4.

Even the most reliable parameters from the VARS analysis resulted in roughly uniform distributions, or distributions that could not be fit with an analytical statistical distribution (an example of the same distribution seen in Figure B.2 is presented in Appendix B, Figure B.3, but with duplicate samples filtered

out). Without analytical parameter distributions, parameter uncertainty can potentially be misrepresented due to under sampling or lack of identifiability. Figure 10 showed that sensitivity reliability was often seasonally dependent, which suggested that parameter identifiability was an issue for all models. This is not unexpected given that parameters are tied to processes, at least conceptually, with processes being relatively more (or less) important at different times of the year (e.g. snowmelt vs. evapotranspiration), as our VARS analysis revealed. This may prevent well-defined parameter distributions from being generated, even with increased sampling. HEC-HMS had notably less reliable VARS sensitivity rankings. WATFLOOD had many low reliability parameter rankings, but a few with reliability near 100%. HYPE had the largest number of high reliability parameter rankings. The most sensitive and least sensitive parameters were generally ranked reliably, as the relative sensitivity varied significantly more between the most and second most sensitive parameter, as opposed to parameters in the mid-ranks, where the relative difference in sensitivity between adjacent parameters was lower. More samples would allow VARS to better map the parameter interactions and separate variables with similar relative sensitivity. The more parameters analyzed by VARS, the more parameter interactions that needed to be explored. Therefore, HYPE often had the most reliable VARS results because it had fewer parameters to assess. The relative parameter sensitivities and reliabilities varied with performance metric choice and season. For example, Appendix B, Figure B.4 to Figure B.8 present VARS relative sensitivities and reliabilities for iVARS 1-5 produced by HYPE, where it is noted that changes in reliability are tied to the seasonal changes in relative sensitivities. The “fpsno corr” parameter was (as expected) more sensitive in winter, when snow is present, than in summer. Relative sensitivities for summer most resemble the period sensitivities, suggesting that summers were most influential in the calculation of KGE scores. VARS results for HYPE evaluated with the PBIAS score are presented in Appendix B, Figure B.9 to Figure B.13. Unlike the KGE, PBIAS can be both positive and negative without necessarily decreasing performance (i.e. a -1% and +1% PBIAS can be seen as equal performance with an

opposite simulation tendency), which means iVARS 5 is not necessarily higher than iVARS 1, which was true for the KGE. The “kc corr” parameter is suggested to be insensitive by the period PBIAS, yet, while not equally sensitive, is sensitive in all seasons. This behavior can be seen in Appendix B, Figure B.14, in which absolute period PBIAS VARS results are presented. The absolute PBIAS does not differentiate between positive and negative bias, and therefore, shows “kc corr” as being sensitive. This means that the direction of bias differed by season, and opposite seasonal directions canceled out annually. Finally, the log NSE score was utilized with VARS (Appendix B, Figure B.15). Using a log transform gives a higher weighting to low flows, which changed the relative order of parameter sensitivity and increased reliability. This suggests that HYPE parameters were more reliable in relative sensitivity for the estimate low flows rather than high flows, which is expected due to higher input uncertainty with summer high flows. Different performance metrics implicitly weight types of events (i.e., peak flow, baseflow, etc.) higher than others; for example, the NSE emphasizes peak flow performance, and log transformed NSE emphasizes low flows. Some parameters were also more sensitive to certain types of events, such as Muskingum routing parameters in HEC-HMS for low flows. This further supports that parameter identifiability is generally an issue for all three hydrologic models as sensitivity varies with the emphasis on the fit of high versus low flows. However, this also argues against combined performance metrics, instead, suggesting that multiple performance metrics be utilized but investigated separately in multiple time periods.

4.7.2. Scale Dependence of Model Performance

To improve representation of parameter uncertainty, all models were run using orthogonal Latin hypercube sampling (OLHS) with the number of samples scaled by the number of parameters selected for each model. Results from the top 10% of period KGE scored OLHS model runs have been summarized (Table 4).

Table 4: 5th and 95th percentile KGE scores for all models, precipitation inputs, and gauged areas; gauges are ordered from smallest to largest (by gauged area). Gauges affected by regulation are highlighted in bold*.

Gauge[1]	Model	Precip Min		Precip Mean		Precip Max	
		KGE (95th)	KGE (5th)	KGE (95th)	KGE (5th)	KGE (95th)	KGE (5th)
Taylor River (05TG002)	HEC-HMS	0.17	0.04	0.59	0.49	-0.56	-1.04
	WATFLOOD	-0.48	-0.59	0.79	0.63	-1.89	-2.01
	HYPE	-0.39	-0.49	0.79	0.78	-0.81	-1.14
Kettle River (05UF004)	HEC-HMS	-0.15	-0.23	0.49	0.30	0.74	0.64
	WATFLOOD	-0.34	-0.44	0.84	0.79	-1.31	-1.47
	HYPE	-0.31	-0.39	0.70	0.66	0.01	-0.13
Angling River (05UH001)	HEC-HMS	0.00	-0.13	0.70	0.50	0.42	0.18
	WATFLOOD	-0.45	-0.58	0.61	0.42	-1.33	-1.41
	HYPE	-0.31	-0.38	0.59	0.49	0.12	-0.03
Weir River (05UH002)	HEC-HMS	0.02	-0.11	0.74	0.58	0.39	0.14
	WATFLOOD	-0.31	-0.43	0.79	0.74	-1.71	-1.85
	HYPE	-0.15	-0.28	0.74	0.69	-1.51	-1.80
Limestone River (05UG001)	HEC-HMS	0.06	-0.07	0.74	0.57	0.22	-0.10
	WATFLOOD	-0.32	-0.40	0.83	0.78	-1.68	-1.81
	HYPE	-0.22	-0.36	0.72	0.70	-1.50	-1.74
Burntwood Above Leaf Rapids (05TE002)	HEC-HMS	-0.23	-0.31	0.26	0.13	0.15	0.02
	WATFLOOD	-0.45	-0.55	0.81	0.78	-2.82	-3.04
	HYPE	-0.46	-0.51	0.77	0.73	-1.31	-1.91
Odei River (05TG003)	HEC-HMS	0.12	-0.02	0.69	0.60	-0.08	-0.54
	WATFLOOD	-0.43	-0.55	0.84	0.72	-1.70	-1.82
	HYPE	-0.32	-0.43	0.80	0.75	-2.12	-2.49
Grass River (05TD001)	HEC-HMS	0.49	0.29	0.59	0.50	-0.97	-1.62
	WATFLOOD	-0.54	-0.59	0.86	0.80	-2.91	-3.23
	HYPE	-0.60	-0.62	0.87	0.71	-2.16	-3.15
Burntwood near Thompson (05TG001)	HEC-HMS	0.85	0.84	0.86	0.85	0.81	0.75
	WATFLOOD	0.84	0.84	0.90	0.89	0.73	0.68
	HYPE	0.82	0.82	0.87	0.85	0.85	0.75
Nelson River at Kelsey GS (05UE005)	HEC-HMS	0.95	0.95	0.93	0.92	0.87	0.84
	WATFLOOD	0.88	0.87	0.84	0.84	0.71	0.70
	HYPE	0.81	0.80	0.94	0.93	0.71	0.64
Nelson River at Longspruce GS (05UF007)	HEC-HMS	0.87	0.86	0.92	0.91	0.80	0.75
	WATFLOOD	0.89	0.88	0.95	0.94	0.58	0.55
	HYPE	0.68	0.67	0.92	0.90	0.60	0.49

For non-regulated basins, the mean precipitation input, not surprisingly, produced the highest KGE scores for all basins and models, with the exception of the Kettle River simulated by HEC-HMS. Precipitation minimum and maximum realizations generally produced negative KGE scores (Table 4). For regulated basins, KGE scores were higher, which is a reflection of the high magnitude inputs. HEC-HMS showed the most variation in performance, between basins, but the least variation between precipitation realizations. For all models, the most sensitive parameters were represented by non-

[1] Some unregulated gauges may be affected by regulation but only minimally, such that the effects of regulation

uniform likelihood distributions using the higher number of OLHS samples, as opposed to the VARS sampling (Table 4).

A linear regression of the 95th percentile KGE scores with basin area (APPENDIX B, Figure B.16) shows that there were no significant correlations between spatial scale and model performance when utilizing the mean precipitation realization for any model (p-value of 0.05) (Table 5). WATFLOOD showed a significant (p-value of 0.05) decrease in KGE scores for the maximum precipitation realization as basin size increased (APPENDIX B, Figure B.18). HYPE had a significant decrease in KGE scores for the minimum precipitation realization as basin size increased (APPENDIX B, Figure B.17). There was no common behavior, with respect to scale-dependent performance, among the models. HEC-HMS had no significant linear regressions in the range of KGE values between the 5th and 95th percentiles. As basin size increased, WATFLOOD had a significant decrease in KGE range, respectively, for the minimum and maximum precipitation realizations (Appendix B, Figure B.20 and Figure B.21). HYPE, however, had a significant increase in KGE range as basin size increased for the mean and maximum precipitation realizations (APPENDIX B, Figure B.19).

Table 5: R^2 and p-values for a linear regression analysis of KGE scores (1981-2010 daily data) and 5th-95th KGE score spread with respect to basin area. Significant linear regressions are presented in bold based on 0.05 p-value significance.

Model (input)	95th percentile KGE r^2	95th percentile KGE p-value	5th - 95th percentile KGE range r^2	5th - 95th percentile KGE range p-value
HEC-HMS (Max precip)	0.50	0.05	0.42	0.08
HEC-HMS (Mean Precip)	0.02	0.74	0.37	0.11
HEC-HMS (Min Precip)	0.46	0.07	0.48	0.06
WATFLOOD (Max precip)	**0.61**	**0.02**	**0.75**	**0.01**
WATFLOOD (Mean Precip)	0.19	0.28	0.10	0.46
WATFLOOD (Min Precip)	0.37	0.11	**0.57**	**0.03**
HYPE (Max precip)	0.48	0.06	**0.88**	**0.00**
HYPE (Mean Precip)	0.50	0.05	**0.57**	**0.03**
HYPE (Min Precip)	**0.56**	**0.03**	0.41	0.09

4.7.3. Relative Contributions to Overall Uncertainty

The partitioned relative uncertainties generated by the input, structure, and parameters propagated into streamflow at the basin outlets vary in magnitude. Figure 11 provides results for the Angling River gauge; the Angling River example was selected for display in this thesis chapter because it showed the largest 5th to 95th percentile spread in KGE scores for HEC-HMS and WATFLOOD, and the second largest for HYPE (behind the Grass River). The Angling River was also not subject to regulation (although similar plots for all the other gauges are available for consultation in Appendix B, Figure B.22 - Figure B.31). In Figure 11, the width of the flow envelope in any one panel is a representation of uncertainty propagated through a hydrologic model. The flow envelope for any one panel represents parameter uncertainty for that model and precipitation realization (e.g. Figure 11a). Input uncertainty propagation is represented by the difference among streamflow envelopes generated for a single model (e.g. HYPE, Figure 11abc). Structural uncertainty can be assessed by comparing the streamflow envelopes for a single precipitation realization in all three models (e.g Figure 11aei). The last row (Figure 11mno) combines all models to examine input data uncertainty; the last column (Figure 11dhi) combines all precipitation realizations to assess model structural uncertainty; and the bottom-right plot (Figure 11p) shows total uncertainty propagated to streamflow.

Parameter uncertainty was most significant in HEC-HMS: a comparison of flow data envelopes in Figure 11bfj shows the widest envelope for HEC-HMS. Input uncertainty was most significant in WATFLOOD, where a comparison of Figure 11dhi reveals the widest envelope for WATFLOOD. In general, precipitation input generated the greatest amount of uncertainty in streamflow simulations for all basins. Model structure generated more uncertainty for the minimum and maximum precipitation realizations than it did for the mean precipitation. Areas of zero density (e.g. the white space between the blue lines from Figure 11e and 11g) have no simulated hydrographs; these are a result of the way the input data ensemble was created. Since only the mean, absolute minimum, and maximum input

data ensemble realizations were utilized, areas of zero density were created due to the large input range. Generating more precipitation input realizations of varied likelihoods would fill in the missing flow density, but the goal was to generate total possible bounds for predicted streamflow here. More precipitation input sampling would not change the maximum width of the flow envelope generated by precipitation-based uncertainty.

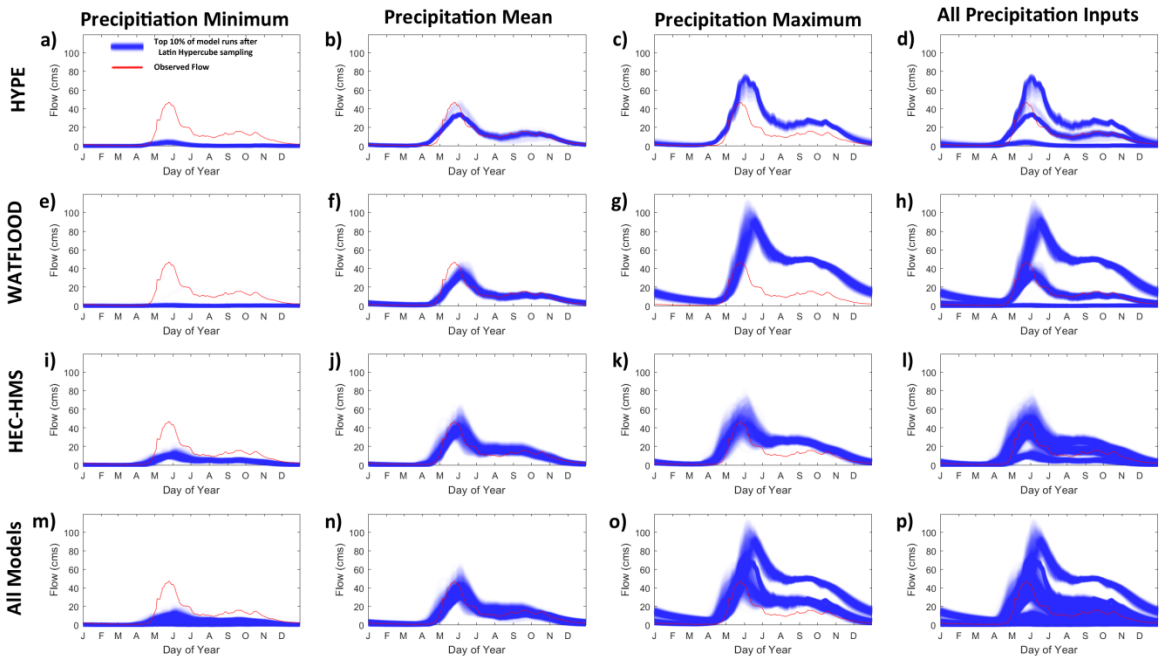

Figure 11: 30-year average hydrographs (1981-2010) for the Angling River gauge, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

Operationally, high sampling quantities for all sources of uncertainty may not be possible, therefore, the effect each source of uncertainty has on streamflow predictability is important when determining computational budget priorities (Figure 12). It is desirable to have the simulated flow cumulative distribution function (CDF) for a particular day intersect the observed value near the 50% non-exceedance probability; crossing higher (lower) than 50% suggests a tendency to underestimate (overestimate). A range of $\pm$ 5% is suggested for observed flow (output data) uncertainty (Environment

Canada, 1980; Dingman, 2014). In all unregulated basins, the ensemble of hydrologic models and precipitation inputs CDF (e.g. Figure 12a, red line) intersected the observed 30-year average within the 10^{th} and 90^{th} uncertainty bounds. Output uncertainty limits are representative of realistic interpretations of observed data, however with unequal likelihoods. The ensemble of hydrologic models and precipitation realizations CDF (Figure 12a, red line) often intersected the observed uncertainty envelope near the 50^{th} percentile of the uncertainty bounds and was the highest performing for the minimum flow case in all basins. The maximum 30-year average flow case showed more variability. For the maximum flow case, the full hydrologic ensemble using either the mean or maximum precipitation realizations was often comparable or higher performing than the ensemble of hydrologic models and precipitation realization CDF (e.g. Figure 12b, red line).

Regulated gauges were more challenging for models to accurately simulate (Figure 13). Simulated flow timeseries for the Nelson at Kelsey (05UE005) performed poorly compared with observed flows. Performance generally improved when considering output uncertainty, given the magnitude of flow for regulated gauges. The best performance generally occurred at the extreme likelihood range of output uncertainty. There is a notable drop in the observed flow record for the Nelson at Longspruce (05UF007) in the end of December; this low flow corresponds to December 25^{th}.

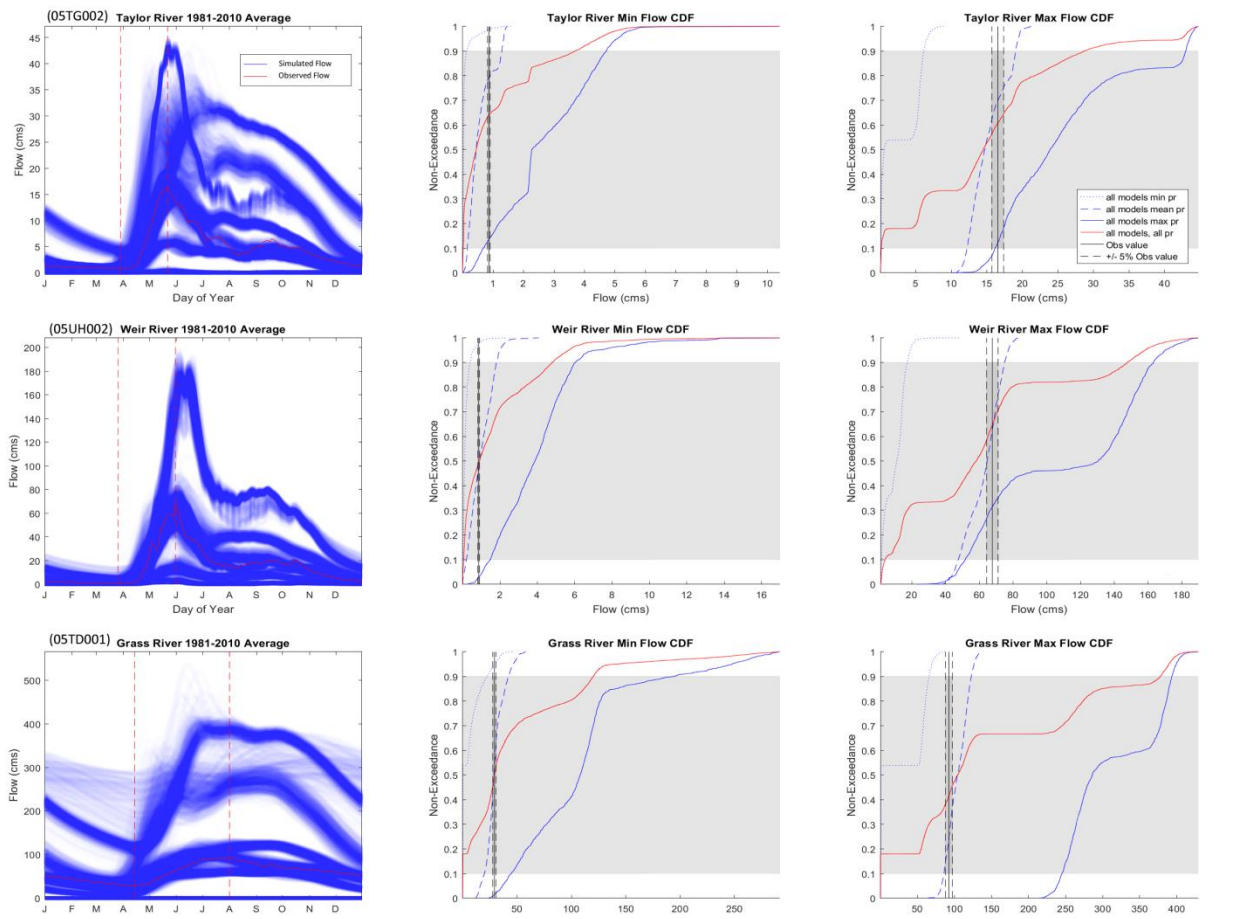

Figure 12: Un-regulated 30-year average hydrographs (1981-2010) for selected hydrometric gauges, with CDF plots for the minimum (Left CDF) and maximum (Right CDF) 30 year average flows. 30-year hydrographs include the top 10% of runs for all hydrologic models and precipitation inputs.

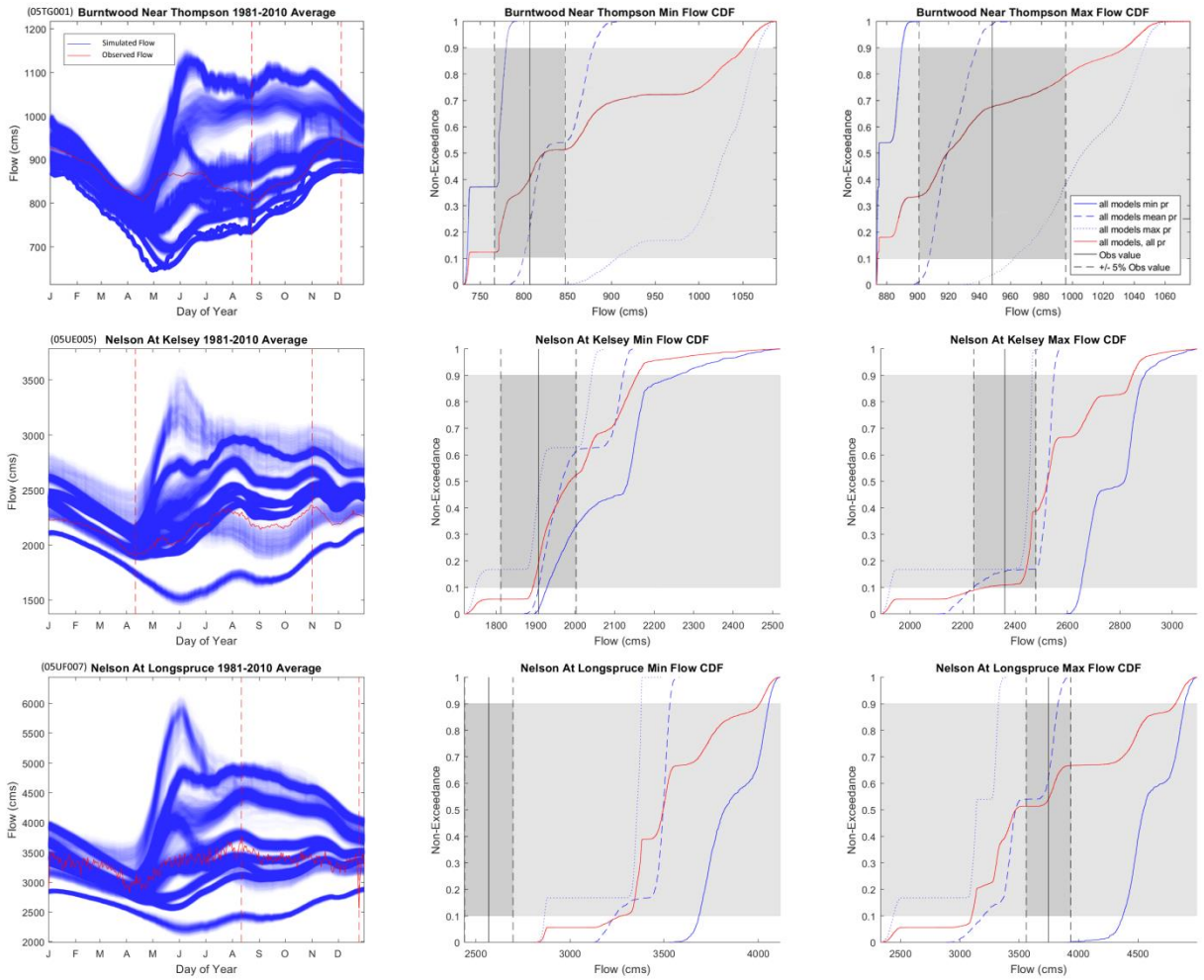

Figure 13: Regulated 30-year average hydrographs (1981-2010) for selected hydrometric gauges, with CDF plots for the minimum (Left CDF) and maximum (Right CDF) 30 year average flows. 30-year hydrographs include the top 10% of runs for all hydrologic models and precipitation.

4.8. Discussion and Conclusion

4.8.1. Parameter Identifiability

VARS sensitivity results show that models with more selected parameters (e.g. HEC-HMS) generate lower VARS sensitivity reliabilities for those parameters; this was also shown by Razavi and Gupta (2016b). The seasonality of sensitivity rank reliabilities offered insight into the presence of identifiability

issues. HYPE's "fpsno corr" parameter is most identifiable in winter, and least identifiable in summer (Figure 10). It is understandable that a snowmelt parameter would have low identifiability in a season that rarely has snow. Other parameters that show that sensitivity reliability was seasonally dependent suggest identifiability issues; this would not have been quantified if only the period scores were utilized. Using the full period generates near uniform parameter likelihood distributions for parameters with noticeable identifiability issues. Identifiability issues are sometimes related to structural deficiencies, however, to assess the presence of potential structural deficiencies, further examination of soft data such as evapotranspiration would have been needed. Frameworks such as DYNIA would be well suited to determine temporally variable parameter likelihood distributions (e.g. Wagener et al., 2003; Abebe et al., 2010; Merz et al., 2011). DYNIA would have provided the parameter distributions for a moving temporal window, in which parameters with identifiability issues would take on different distributions under different hydrologic conditions. The VARS analysis suggested which parameters would produce different distributions, but did not provide those distributions. Our results indicate that many model parameters have lower reliability when the analysis is conducted over an entire performance period (annual or longer term) than they do when evaluated for an individual season (Choi and Beven, 2007). A further notable finding was the dependence of parameter sensitivity on performance metric and temporal period of assessment. Razavi and Gupta (2016b), and many other studies, conducted a sensitivity analysis on annual periods, however, results show that this may obscure true parameter sensitivity. Similarly, the choice to utilize either a single performance metric, or an aggregated score, is not advised, as it fails to generate specific information gained from viewing metrics individually.

The choice to only use VARS with NARR is unlikely to have affected the uniformity of likelihood distributions. Pokorny et al., (in prep) showed that high and low performing years averaged when considered in a long temporal period. Since the sensitivities were determined based on period and seasonal KGE score, it is expected that a similar averaging would occur with any high likelihood

realization. Therefore, it is expected that some variation in sensitivity would occur by using other precipitation datasets; however, the distributions produced are unlikely to have been non-uniform for a different high likelihood precipitation dataset. Assessing the skewness of parameter distributions between with different precipitation datasets would have provided additional information about hydrologic process sensitivity to differences in precipitation input. Since the goal was to generate an estimate of total uncertainty, the full range input uncertainty was considered regardless of differences in parameter sensitivity among high likelihood precipitation inputs (e.g. NARR vs. WFDEI) (e.g. Li and Xu, 2014).

VARS only generated an effective 1,000 samples for parameter likelihood distributions. Even the most sensitive HYPE parameters resulted in approximately uniform distributions (or in distributions which were difficult to generate a statistical fit with), even though HYPE had the fewest parameters. An example distribution of HYPE's "kc corr" parameter in the Taylor River can be found in Appendix B, Figure B.3. The "kc corr" parameter indicated consistent reliability across all seasons and was highly sensitive. The likelihood distribution, however, included gaps that resulted from under sampling the parameter range. More than 1,000 samples are usually always seen in the literature: for instance, Shafii et al., (2015) utilized multiple sampling quantities ranging up to 100,000. The models used by Shafii et al., (2015) had fewer parameters than those in this study. Similarly, higher sampling than the 1,000 samples generated by VARS is seen throughout the literature, which suggested under sampling was likely an issue in the present study. Utilizing only 1,000 parameter samples was the result of selecting 100 VARS star centers for sampling. To achieve 6,300 parameter samples for HEC-HMS using VARS, 630 star centers would have been needed, which would have required more than 350,000 HEC-HMS runs. Therefore, to reduce computation times, OLHS was utilized separately from VARS for parameter distribution development; however, assessing the 1,000 samples produced by VARS was used to aid in the OLHS quantity choice.

The goal of parameter sampling is to generate sufficient samples to approximate the uncertainty generated from parameter estimation. The production of non-uniform distributions capable of statistical distribution fitting was considered here as an indicator of a quantity of parameter sampling that is useful for estimating the parameter uncertainty. Information on parameter identifiability generated by VARS in combination with the OLHS was used to determine if a sensitive parameter with a near-uniform distribution was near-uniform due to under sampling, or as a result of identifiability issues. After the additional Orthogonal Latin hypercube sampling, many of the most sensitive parameters showed non-uniform distributions. Using HEC-HMS as an example, the base temperature (Appendix B, Figure B.34) and the groundwater routing coefficient for layer 1 (parameter group 5) (Appendix B, Figure B.35) parameter distributions would be possible to fit using statistical distributions, which was expected given their high relative sensitivities. The parameter distribution for clear sky transmittance, however, remained relatively uniform (Appendix B, Figure B.36); this suggests that likely, further sampling would have been needed to generate distributions for less sensitive parameters. This, however, assumes that identifiability issues would not prevent defined distributions from being generated, which is difficult to determine with the VARS analysis given the low reliabilities.

4.8.2. Model Selection

The selection of the top 10% highest performing models was based on the period KGE values (Table 4). The presence of scale-dependent performance was not consistent between models for either the 95th percentile KGE values or the 5th to 95th percentile spread of KGE values. The minimum and maximum precipitation realizations were sampled as low likelihood precipitation scenarios from the gridded data ensemble produced by Pokorny et al., (in prep). The resulting KGE score from the minimum and maximum precipitation realizations is understandably low. A similar study was conducted by Nikolopoulos et al., (2010), who found a strong relationship between performance and basin scale when

considering input uncertainty propagation through a hydrologic model. Nikolopoulos et al., (2010) did not, however, consider structural and parameter uncertainty, which Ajami et al., (2007) suggested to be significant. If only the HYPE model was used, a conclusion similar to Nikolopoulos et al., (2010) could have been presented. This suggests that the results of Nikolopoulos et al., (2010) may require additional consideration of all sources of uncertainty to confirm a strong consistent relationship between model performance and basin scale. HEC-HMS and WATFLOOD had lower p-values for the linear regression of KGE score and basin size with the minimum and maximum precipitation realizations. Lower p-values for the minimum and maximum precipitation realizations suggested that input uncertainty affected the significance of the relationship between KGE performance and basin scale. In addition to input uncertainty, model structure was, in part, responsible for the direction of the relationship. Suggesting a source of uncertainty responsible for a linear regression characteristic, however, must consider that the sources are only relatively partitioned, but are still interconnected by propagation. Pokorny et al., (in prep) showed that the maximum precipitation realization increased the spatial extent of summer storms. Therefore, it is expected that larger basins would decrease in performance, as they are more subject to the increased spatial extents of a storm: this is supported by the agreement in linear regression direction (significant for WATFLOOD and HYPE) in APPENDIX B, Figure B.18. The mean and minimum precipitation realizations did not produce consistent linear regression directions for all models.

4.8.3. Parameter Selection

Negative KGE scores are a reflection of parameter search ranges that did not allow parameters to take on high performing values when using low likelihood precipitation realizations. Parameter likelihood distributions for these precipitation realizations resulted in high parameter likelihoods at the extremes of the parameter ranges (e.g. Appendix B, Figure B.37 and Figure B.38). The distributions, however, do

make sense; for example, the base temperature parameter in HEC-HMS is responsible for determining the spring melt timing. Since the unregulated basin peak flows are driven by snowmelt, selecting a more negative base temperature would artificially reduce spring peak flow by increasing snowmelt early in spring. Similarly, high positive base temperatures would artificially increase the spring peak flow by having more snowmelt occur later in spring without reduction in snowpack volume earlier in spring. The maximum precipitation realization was wet, therefore, the base temperature shifted to negative values to artificially lower the spring peak; the minimum precipitation realization was dry, therefore, a positive shift in base temperature artificially raised the spring peak. Parameter ranges are generally developed to reflect realistic hydrologic response under normal, or average, hydrologic conditions, and are developed or recommended by model developers. These recommended parameter ranges are not designed to reflect extreme scenarios, such as the absolute minimum and absolute maximum precipitation realizations. The parameters showed likelihoods at the extreme range of the search ranges in response to the low likelihood precipitation realizations, which were selected for the purpose of generating total possible uncertainty bounds as part of the BaySys project as well as to show the evolution of parameter/structural uncertainty as precipitation becomes very wet/dry.

There were fewer regulated gauges in the LNRB than there were unregulated gauges. The contributing basins upstream of the three regulated gauges included in this study received flow forcing from upstream, un-modeled areas. Performance metrics were generally high for the regulated basins as they receive a significant portion of total flow from forced (observed) inputs. Scores for the selected regulated gauges were less reliant on model performance since the upstream flow forcing relied on observed records. To determine any basin size dependency for regulated regions, the full upstream area would need to be modeled.

The spread of the top 10% of period KGE scores is also a reflection of model structure (Table 4). KGE scores for WATFLOOD varied the most with precipitation input, and the least for HEC-HMS. This result is not surprising given that HEC-HMS utilizes spatially averaged inputs, while WATFLOOD utilizes distributed precipitation inputs (i.e., every grid point is assigned a value of precipitation). Spatial averaging appears to reduce the effect of input data uncertainty, which is in agreement with the results of Pokorny et al., (in prep) and also suggests that lumped models are capable of reasonable performance in data sparse regions (e.g. Das et al., 2008). HYPE's input structure did not utilize all grid points, but also did not spatially average the inputs, and therefore shows larger variations in top 10% period KGE scores than HEC-HMS, but smaller than WATFLOOD.

4.8.4. Uncertainty Extremes

The structural and parameter uncertainty envelopes were generally wider when utilizing the minimum or maximum precipitation realizations. This is an expected result, the results of the VARS sensitivity analysis showed that relative parameter sensitivity changes based on performance metrics due to different flow event types (i.e. peak versus low flows) being emphasized. The minimum and maximum precipitation realizations create extreme scenarios (i.e. the minimum precipitation realization may be more appropriate in an extreme drought). Therefore utilizing a dry or wet input would be expected to shift sensitivity; this was also seen in the non-uniform distributions generated by the minimum and maximum precipitation realizations (e.g. Appendix B, Figure B.37 and B.37). Different parameter sensitivities would affect the uncertainty envelopes of parameter and structural uncertainty, in this case, by widening them. A wider flow envelope suggests that more uncertainty among model simulations is generated by varying precipitation realizations than is generated by varying model structure (or parameters). It was possible to sample the extreme range of input uncertainty; however, it is difficult to sample the extreme range of model structure.

It is important to define the extreme range of likelihood to the BaySys project. Studies often suggest climate data to be the most dominant source of uncertainty in future projections, in which uncertainty is larger in future periods than in historic periods (e.g. Dams et al., 2015). With a higher uncertainty in future projections, low likelihood historical events would be expected to be of higher likelihood in in the future, such as large floods (Kay et al., 2009). Therefore presenting a wider envelope of uncertainty is desirable as an events likelihood is also dependent on the time period examined.

4.8.5. Predictive Capability

The effect on model predictive capability resulting from precipitation input uncertainty shows that the minimum precipitation input is consistently too dry (Figure 12), however, the maximum precipitation input produces some useful estimates of observed streamflow, such as for the Angling River (Figure 12).

Excluding structural uncertainty (i.e. any single model forced with all three precipitation ensembles (not shown)) revealed that no one model was best. Considering structural uncertainty with a single precipitation realization (Figure 12) often intersected the observed uncertainty envelope within the 10th and 90th percentile uncertainty bounds when utilizing the mean or maximum precipitation realization. This suggested that based on the precipitation sampling conducted in this study, structural uncertainty often improved observed streamflow representation more than input uncertainty. The consideration of all three sources of uncertainty often produced a CDF that intersected the observed uncertainty envelope nearest the 50th percentile (red lines in Figure 12). There are two considerations for determining which CDF best represented the observed uncertainty envelope: intersection near the 50th percentile, and the percentile range within the observed uncertainty envelope. For example, the Weir River (Figure 12, maximum flow) shows that when considering all hydrologic models with the mean precipitation as input, the CDFs are within the observed flow uncertainty envelope between approximately the 45th and 80th percentiles, while total uncertainty was only within the observed

uncertainty between the 60th and 70th percentiles. This means that the mean precipitation realization with all hydrologic models is more likely to produce a flow within the observed uncertainty than total uncertainty, based on the sampling conducted in this study. Considering all sources of uncertainty generally intersected the observed uncertainty envelope near the 50th percentile, but with a small percentile range. Further sampling of precipitation input uncertainty would be required to potentially improve the models' representation of average observed streamflow, by generating more high likelihood realizations. However, the goal of the BaySys project was to generate a total possible uncertainty envelope for the Nelson River flow, which did not require further input sampling. Therefore, the inclusion of structural uncertainty was often more valuable for improving the representation of the average observed streamflow than the consideration of input or parameter uncertainty with the sampling done in this study.

Regulated gauge results (Figure 13), however, showed that there were events generated by regulation that were not reproducible with the selected models. The minimum flow case for the Longspruce generating station corresponds to low flows on Christmas Day, a trend documented by Déry et al., (2016). This is a regulation decision driven by energy demand, which requires full regulation modeling to be captured accurately by a hydrologic model (Tefs, 2018). The HYPE model version used in this study utilized an early representation of regulation implemented by MacDonald et al., (under revision); however, HYPE best represented the observed flow uncertainty envelope for regulated gauges. The regulated gauges were less valuable for determining the hierarchy uncertainty source, for consideration within a limited computation budget. The results show that implementing regulation into the hydrologic models would be most effective for improving representation of the observed uncertainty envelope (Tefs, 2018).

4.8.6. Computational Demands

Runtimes for a single HYPE, WATFLOOD, and HEC-HMS run were 45 minutes, 15 minutes, and 4 minutes, respectively. These runtimes represent one complete model run, including the updating of parameter files and model inputs, the model run, and the export of results. For WATFLOOD and HYPE, output was stored for later post processing, however, HEC-HMS included an additional interfacing with the HEC-DSSVUE data viewer to extract output for the selected locations. The domain of the HEC-HMS model and the WATFLOOD model is the LNRB, while the HYPE model was run for the full NCW domain. The HYPE model has upstream dependencies that prevented the LNRB from being run alone. The total serial computational time for all runs would have been approximately 4.7 years; parallel computing allowed all runs to be completed in approximately 6 months. In an operational setting, such runtimes are not generally feasible. If additional computational resources had been available, more parameters, higher parameter sampling quantities, and additional structures would have been utilized; although, the sampling conducted was sufficient to generate an estimate of total uncertainty. Additional precipitation sampling would have filled in areas of zero density in Figure 12 and 13, and the CDF within observed flow uncertainty for the Weir River (Figure 12, maximum flow) would be expected to increase beyond the 60^{th} and 70^{th} percentiles. As a result, the total uncertainty CDF would be more likely to simulate flow within the observed uncertainty envelope, but at the expense of increased computational time. Kavetski et al., (2006a and b) assigned a variable to each storm event, and varied them individually; however, Ajami et al., (2007) adjusted this process to be more computationally feasible by varying the mean and standard deviations of applied precipitation distributions; both these methodologies are sampling intensive. Varying model structure, however, requires the development of multiple models. Therefore, the context of an operational environment dictates which source of uncertainty would be most valuable to consider. Based on the sampling done in the present study, structural uncertainty was generally the most valuable, followed by input uncertainty. Parameter uncertainty when considered alone (e.g. Figure

11a) was generally less valuable that input and structural uncertainty as parameter uncertainty bounds were generally narrow. The narrow parameter uncertainty bounds did not envelope many observed events, but was computationally expensive. If sub-daily flow forecasts were required, input uncertainty may not be computationally possible to consider. While it would require initial development time, additional model structures could be utilized to improve forecast quality while only adding the runtime of the additional models for the forecast process. If additional model development is not accessible, including input uncertainty would be valuable to consider. If sufficient computational time is available, as suggested by Ajami et al., (2007), inclusion of all sources of uncertainty is still recommended.

The most downstream publically available hydrometric gauge on the Nelson River is the Nelson River at Longspruce Generating Station gauge (05UF007). The total uncertainty associated with simulating flow for the Nelson River with reference to output uncertainty is, therefore, reasonably represented by the Long Spruce gauge (Figure 13). The total uncertainty generally envelopes output uncertainty as well as presenting a wide uncertainty boundary, which represented low likelihood realizations of freshwater flow for use in other models in the BaySys project. The effects imposed by regulation were, however, not all well represented. Therefore, the inclusion of regulations modeling presented by Tefs (2018) would offer improvement to representation of the uncertainty associated with regulation based changes to flow (e.g. the low flow seen on December 25th for the Nelson River at Long Spruce Generating Station gauge (05UF007)). The total uncertainty envelope showed that simulation predictability was generally within the suggested 10th and 90th percentiles utilized by Manitoba Infrastructure, however, the 25th and 75th percentile range was often sufficient to represent observed uncertainty.

4.9. Acknowledgments

Thanks to University of Manitoba, Manitoba Hydro, and partners funding through the Natural Sciences and Engineering Research Council of Canada through funding of the BaySys project. Thanks to

developers of the gridded datasets and hydrologic models used in this study. Thanks to Environment Canada for providing access the AHCCD dataset, NRCAN for providing access to the ANSUPLIN dataset, and SMHI for providing access to the Hydro-GFD dataset. Thanks to Dr. Saman Razavi for providing access to the VARS sensitivity framework.

Chapter 5 : Conclusions and Recommendations

5.1. Summary of Major Findings

This thesis contributes to the understanding of hydrologic modeling uncertainty and how input, parameter, and structural uncertainty propagates towards total uncertainty. All model inputs and numerical steps within a model introduce uncertainty into the hydrologic model output. Similar to other studies in the literature, it was found that of the meteorological input, temperature generally showed less uncertainty than precipitation; therefore, precipitation was selected as the focus for input uncertainty assessment. It was found that performance metrics should be selected to assess a precipitation dataset's representation of precipitation event timing and magnitude, as well as the occurrence and magnitude of extremes, to ensure a robust assessment of dataset quality. The spatial aggregation was found to have a significant effect on the input uncertainty subject to propagation. It was found that gridded climate datasets generally struggled to reproduce the frequency of occurrence and magnitude of extreme precipitation events. Spatial aggregation, however, reduced input uncertainty in the representation of extremes by reducing the dependence on accurate storm positioning by averaging grid cells where an event occurred with grid cells where no event occurred. Overall, it was found that as spatial aggregation increased, performance metrics suggested improved dataset performance, although, sparse observed climate station density still negatively impacted performance metrics.

Gridded climate dataset performance was temporally variant for the gridded climate datasets examined in this thesis. Bias in a moving temporal window was not consistent in any of the gridded datasets examined, suggesting that a single generalized "best" gridded climate dataset cannot be determined. Instead, it is recommended that multiple gridded climate datasets appropriate for a region and/or model be selected and applied in an ensemble, rather than attempting to determine a single best product. If multiple products are selected and treated as an ensemble, then the individual datasets can be seen as realizations of that ensemble. However, an ensemble realization does not need to be associated with a single product. It was found that the ensemble mean generally outperformed individual datasets used to generate the ensemble. Absolute minimum and maximum ensemble realizations were generated to represent the total uncertainty associated with the gridded climate data ensemble. It was found that, while not probable, the minimum and maximum precipitation realizations were valuable for estimating total precipitation uncertainty subject to propagation through a hydrologic model. Despite the maximum (minimum) precipitation realizations being generated to represent the extreme range of likelihood, many events of higher (lower) magnitude were not bracketed by the ensemble envelope. While the goal of this thesis was to generate an estimate of total uncertainty, if additional realizations of higher likelihood were desired, a modified version of the reliability and sharpness metrics from other hydrologic studies could be utilized (e.g. Shafii et al., 2015). A modified version of the reliability and sharpness metrics would then include uncertainty as a factor in realization generation.

Overall, it was found that input data ensembles must be utilized to better account for input uncertainty. The temporal dependence of individual gridded climate dataset performance prevents a single best product from being determined. Additionally, considering model performance at a single spatial aggregation, or over long temporal periods, obscures the source or reason for good or bad dataset performance.

A sensitivity analysis of an ensemble of hydrologic models was used to detect the presence of parameter identifiability issues, as well as to guide parameter sampling quantities for OLHS simulations. It was found that all three hydrologic models included in this study showed seasonally dependent parameter sensitivity. It was also found that, similar to the temporal period of evaluation, the selection of performance metrics for model evaluation was of high importance. A single performance metric, such as KGE, or a single weighted performance metric, obscures the information gained from examining multiple performance metrics separately. Identifiability issues were found in all models for all performance metrics assessed.

The selection of behavioral models is a common topic in the hydrologic modeling literature. It was found that much of the literature supporting behavioral selection criteria would have, in fact, excluded all simulations produced by the models used in the present study when forced with the minimum and maximum precipitation realizations. Behavioral selection criteria should select behavioral simulations while accounting for output uncertainty and allowing low likelihood, high risk simulations to be selected. Selection of the top 10% of simulations produced by each meteorological input mostly addressed these challenges.

Parameter distributions generated from the VARS sensitivity sampling did not show notable deviation from their uniform priors to suggest a statistical distribution fit. Additional samples were generated utilizing OLHS; sampling quantities were determined in part by the VARS sensitivity analysis. It was found that higher sampling quantities produced parameter distributions that showed notable deviation from the uniform prior distributions. Similar to results found in the literature, higher sampling quantities were not necessarily able to generate non-uniform parameter distributions for low sensitivity parameters, or those that suffer from notable parameter identifiability issues. Parameter distributions were different when considering the minimum and maximum precipitation realizations representing

extreme scenarios. Similar to results found in the literature, parameter distributions are impacted by input uncertainty.

The cumulative effects of propagation were explored through the relative partitioning of the broad sources of uncertainty. It was found that, similar to the results of the input uncertainty analysis, input uncertainty was lower in HEC-HMS as it utilized larger spatial aggregation than the other two models. Structural uncertainty was also found to be a significant source of uncertainty, and it varied with input uncertainty. The value of accounting for each source of uncertainty was assessed by estimating the predictive capability by CDFs generated from different combinations of relatively partitioned uncertainty sources. It was found that there is benefit in including all sources of uncertainty. Uncertainty analysis has a high computational demand associated with it. It was found that if computational resources are available, it is valuable to the predictive capability to account for all sources of uncertainty. However, in environments with limited computational resources, structural uncertainty was found to be most valuable, followed by input, and parameter uncertainty. The importance of each relative partition is, however, context sensitive; an operational flow forecasting environment may find the most value in additional model structures. If streamflow is not the target output, the most valuable relative source of uncertainty may be different. Similarly, in a research environment a certain relative source may be the focus, therefore, the importance of a relative source of uncertainty selected may be determined by the goal of the research (e.g. historic vs. future input uncertainty within an existing model framework).

An estimate of total uncertainty associated with simulating flow for the Nelson River was generated. Some regulation based events, such as low flow through the Long Spruce Generating Station on December 25th, were not well represented and would require regulations modeling for improvement. The total uncertainty estimate generally enveloped the observed flow data within the 10th and 90th percentiles but the 25th and 75th percentiles were often sufficient.

5.2. Communication of Uncertainty

The general goal of hydrologic modeling is to produce output of the highest reasonably attainable quality. Studies that include uncertainty generally consider a narrow envelope of uncertainty to be desirable. This was a feature of Bayesian inference as opposed to GLUE, which often produced wider likelihood distributions. The consideration of wide likelihood distributions that do not exclude low likelihood realizations should be viewed as increasing confidence in modeled output (Juston et al., 2013). A single high likelihood parameter set and structure is likely to misrepresent events as it is a single model realization generated to best match a single realization of uncertain hydrometric data. The inclusion of uncertainty suggests more events would occur within the uncertain range. Perhaps most important is the recognition that in a changing climate, narrow likelihood distributions that were artificially constrained may no longer represent future events, such as increasing flood magnitudes (Kay et al., 2009). A wider range of likelihood with well-defined distribution tails is likely to be valid for future climate realizations, although, the likelihood of a particular event will not be stationary. A better way to communicate this uncertainty is to move away from presenting a single realization and instead present a likelihood range, as a new standard.

5.3. Recommendations and Future Research

This research has identified specific areas that require further investigation, including:

1) Input uncertainty:

- The development of the input data ensemble in this thesis was generated based on five gridded precipitation products. The inclusion of these products was based on the performance of the individual ensemble members, and data availability for the time period of analysis. The effect of each gridded dataset being added or removed from the ensemble should be investigated, as

well as the possible inclusion of other well-known gridded datasets (e.g. CaPA) that are available in more recent time periods.

- Ensemble realizations were generated by selecting the extreme range of the ensemble and the mean. Inclusion of other ensemble members should be studied utilizing methods such as Bayesian model averaging (or other ensemble generation methods), weighting of ensemble members, and consideration of multiple criteria.
- The full ensemble range was used to assess the uncertainty of the ensemble with respect to observed ground-based climate data. Further examination of methods for representing the ensemble range should be considered, such as adapting methods like the reliability and sharpness concept for selecting behavioral output and narrowing the uncertain range.

2) Structural and parameter uncertainty:

- Each model structure utilized only a single set of mathematical methods to represent the hydrologic processes occurring in the physical LNRB environment. This made the implicit assumption that the uncertainty associated with a single model was sufficiently represented. Combinations of other methods (e.g., for computing evapotranspiration in WATFLOOD, one could use Hargreaves, Priestley-Taylor or Evaporation pan data) are available within a single model, and should also be considered to allow for the relative partitioning of uncertainty generated by the selection of mathematical representation of a physical process.
- Increasing the number of samples when exploring parameter space should be considered. Orthogonal Latin hypercube sampling was used here as it generated samples without predefining a performance threshold. Further sampling would enable parameter distribution fitting, which could be used in combination with variations on the mathematical representation

of hydrologic processes to examine the interactions between parameters through distribution comparison. Additionally, the use of search algorithms other than OLHS could be utilized.

- Sensitivity analysis was conducted on the full study period, as well as seasonally. Considering input data performance on finer temporal periods (e.g., year-by-year) revealed there is temporal variation within a single dataset. A similar procedure, like that of the DYNIA framework, should be considered with VARS by generating sensitivity and reliability information in a temporally variant moving window approach.
- Uncertainty was considered utilizing 30-year daily average annual hydrographs. Additional analyses should be conducted focusing separately on wet and dry years from the historical record.
- Additional model structures could be included to further sample the range of structural uncertainty.

3) Output uncertainty:

- Output uncertainty was considered only through the range of predictability in CDF analyses. Output uncertainty representation should be considered as a criterion during the selection of behavioral simulations.
- Ice-on conditions were assumed to have the same uncertainty as ice-off conditions, which is likely a significant underestimation of uncertainty during ice-affected periods. Consideration of ice-affected measurement uncertainty should be introduced for a reasonable estimation of output uncertainty. Since ice-affected rating curves cannot be reliably generated, a range of estimated ice-affected output uncertainty could be considered instead.

4) Uncertainty propagation:

- Propagation was based on a single performance metric of KGE. Additional performance metrics could be considered.
- Propagation was based only on streamflow. Additional data sources, such as stable water isotopes, could be considered to determine impacts of uncertainty on the estimation of transit time and residence time of water in storage that may not be detectable from total streamflow alone.

References

Abebe, N. A., Ogden, F. L., & Pradhan, N. R. (2010). Sensitivity and uncertainty analysis of the conceptual HBV rainfall–runoff model: Implications for parameter estimation. Journal of Hydrology, 389(3-4), 301-310.

Ajami, N. K., Duan, Q., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resources Research, 43(1).

Asong, Z. E., Razavi, S., Wheater, H. S., & Wong, J. S. (2017). Evaluation of integrated multisatellite retrievals for GPM (IMERG) over southern Canada against ground precipitation observations: A preliminary assessment. Journal of Hydrometeorology, 18(4), 1033-1050.

Barber, D.G. (2014). BaySys – Contributions of climate change and hydroelectric regulation to the variability and change of freshwater-marine coupling in the Hudson Bay System. Retrieved from: http://umanitoba.ca/faculties/environment/departments/ceos/media/BaySys_PROJECT_DESCRIPTION.pdf

Becker, E. J., Berbery, E. H., & Higgins, R. W. (2009). Understanding the characteristics of daily precipitation over the United States using the North American Regional Reanalysis. Journal of Climate, 22(23), 6268-6286.

Bennington, V., McKinley, G. A., Kimura, N., & Wu, C. H. (2010). General circulation of Lake Superior: Mean, variability, and trends from 1979 to 2006. Journal of Geophysical Research: Oceans, 115(C12).

Benke, A. C., & Cushing, C. E. (Eds.). (2011). Rivers of North America. Academic Press.

Berg, P., Donnelly, C., & Gustafsson, D. (2018). Near-real-time adjusted reanalysis forcing data for hydrology. Hydrology and Earth System Sciences, 22(2), 989.

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

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

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

Boluwade, A., Zhao, K. Y., Stadnyk, T. A., & Rasmussen, P. (2018). Towards validation of the Canadian Precipitation Analysis (CaPA) for hydrologic modeling applications in the Canadian Prairies. Journal of Hydrology, 556, 1244-1255.

Braca, G. (2008). Stage-discharge relationships in open channels: Practices and problems. Univ. degli Studi di Trento, Dipartimento di Ingegneria Civile e Ambientale.

Brown, J. D., & Heuvelink, G. B. (2006). Assessing uncertainty propagation through physically based models of soil water flow and solute transport. Encyclopedia of hydrological sciences, 1181–1195

Bukovsky, M. S., & Karoly, D. J. (2007). A brief evaluation of precipitation from the North American Regional Reanalysis. Journal of Hydrometeorology, 8(4), 837-846.

Carpenter, T. M., & Georgakakos, K. P. (2006). Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. Journal of hydrology, 329(1-2), 174-185.

Choi, W., Kim, S. J., Rasmussen, P. F., & Moore, A. R. (2009). Use of the North American Regional Reanalysis for hydrological modelling in Manitoba. Canadian Water Resources Journal, 34(1), 17-36.

Dams, J., Nossent, J., Senbeta, T. B., Willems, P., & Batelaan, O. (2015). Multi-model approach to assess the impact of climate change on runoff. Journal of Hydrology, 529, 1601-1616.

Das, T., Bárdossy, A., Zehe, E., & He, Y. (2008). Comparison of conceptual model performance using different representations of spatial variability. Journal of Hydrology, 356(1-2), 106-118.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., ... & Bechtold, P. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the royal meteorological society, 137(656), 553-597.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., ... & Schaake, J. (2014). The science of NOAA's operational hydrologic ensemble forecast service. Bulletin of the American Meteorological Society, 95(1), 79-98.

Déry, S. J., Stadnyk, T. A., MacDonald, M. K., & Gauli-Sharma, B. (2016). Recent trends and variability in river discharge across northern Canada. Hydrology and Earth System Sciences, 20(12), 4801-4818.

Environment Canada (1980), Manual of Hydrometric Data Computation and Publication Procedures, Fifth Edition, Inland Waters Directorate, Water Resources Branch, Ottawa.

Essou, G. R., Sabarly, F., Lucas-Picher, P., Brissette, F., & Poulin, A. (2016). Can precipitation and temperature from meteorological reanalyses be used for hydrological modeling?. Journal of Hydrometeorology, 17(7), 1929-1950.

Eum, H. I., Dibike, Y., Prowse, T., & Bonsal, B. (2014). Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the Athabasca Watershed, Canada. Hydrological Processes, 28(14), 4250-4271.

Fischer, E. M., Beyerle, U., & Knutti, R. (2013). Robust spatially aggregated projections of climate extremes. Nature Climate Change, 3(12), 1033.

Fortin, V., Roy, G., Stadnyk, T., Koenig, K., Gasset, N., & Mahidjiba, A. (2018). Ten years of science based on the Canadian precipitation analysis: A CaPA system overview and literature review. Atmosphere-Ocean, 56(3), 178-196.

Gbambie, A. S. B., Poulin, A., Boucher, M. A., & Arsenault, R. (2017). Added value of alternative information in interpolated precipitation datasets for hydrology. Journal of Hydrometeorology, 18(1), 247-264.

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.

Gupta, H., & Razavi, S. (2017). Challenges and future outlook of sensitivity analysis. In Sensitivity analysis in earth observation modelling (pp. 397-415).

Haghnegahdar, A., Razavi, S., Yassin, F., & Wheater, H. (2017). Multicriteria sensitivity analysis as a diagnostic tool for understanding model behaviour and characterizing model uncertainty. Hydrological Processes, 31(25), 4462-4476.

Historical Climate Data - Climate - Environment and Climate Change Canada. (2018). Retrieved from http://climate.weather.gc.ca/

Holmes, T. L. (2016). Assessing the Value of Stable Water Isotopes in Hydrologic Modeling: A Dual-Isotope Approach (MSc Thesis, Department of Civil Engineering, University of Manitoba, Manitoba).

Hubbard, K. G., Mahmood, R., & Carlson, C. (2003). Estimating daily dew point temperature for the northern Great Plains using maximum and minimum temperature. Agronomy Journal, 95(2), 323-328.

Hutchinson, M. F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska, E., & Papadopol, P. (2009). Development and testing of Canada-wide interpolated spatial models of daily minimum–maximum temperature and precipitation for 1961–2003. Journal of Applied Meteorology and Climatology, 48(4), 725-741.

ISO, I., & OIML, B. (1995). Guide to the Expression of Uncertainty in Measurement. Geneva, Switzerland.

IPCC. (2014). Fifth Assessment Report. Cambridge: Cambridge University Press.

Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2003). Forecast verification: a practitioner's guide in atmospheric science. John Wiley & Sons.

Juston, J., Kauffeldt, A., Quesada Montano, B., Seibert, J., Beven, K., & Westerberg, I. (2013). Smiling in the rain: Seven reasons to be positive about uncertainty in hydrological modelling. Hydrological Processes, 27(7), 1117-1122.

Karlsson, I. B., Sonnenborg, T. O., Refsgaard, J. C., Trolle, D., Børgesen, C. D., Olesen, J. E., ... & Jensen, K. H. (2016). Combined effects of climate models, hydrological model structures and land use scenarios on hydrological impacts of climate change. Journal of Hydrology, 535, 301-317.

Kavetski, D., Kuczera, G., & Franks, S. W. (2006a). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research, 42(3). Article. No. W03407.

Kavetski, D., Kuczera, G., & Franks, S. W. (2006b). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. Water Resources Research, 42(3). Article. No. W03408.

Kay, A. L., Davies, H. N., Bell, V. A., & Jones, R. G. (2009). Comparison of uncertainty sources for climate change impacts: flood frequency in England. Climatic change, 92(1-2), 41-63.

Khakbaz, B., Imam, B., Hsu, K., & Sorooshian, S. (2012). From lumped to distributed via semi-distributed: Calibration strategies for semi-distributed hydrologic models. Journal of Hydrology, 418, 61-77.

Kluver, D., Mote, T., Leathers, D., Henderson, G. R., Chan, W., & Robinson, D. A. (2016). Creation and validation of a comprehensive 1^{o} by 1^{o} daily gridded North American dataset for 1900–2009: Snowfall. Journal of Atmospheric and Oceanic Technology, 33(5), 857-871.

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. Geophysical Research Letters, 40(6), 1194-1199.

Kouwen, N. (2018). WATFLOOD users manual. Water Resources Group, University of Waterloo.

Leduc, M., Laprise, R., De Elia, R., & Šeparović, L. (2016). Is institutional democracy a good proxy for model independence?. Journal of Climate, 29(23), 8301-8316.

Lespinas, F., Fortin, V., Roy, G., Rasmussen, P., & Stadnyk, T. (2015). Performance evaluation of the Canadian precipitation analysis (CaPA). Journal of Hydrometeorology, 16(5), 2045-2064.

Li, L., Xia, J., Xu, C. Y., & Singh, V. P. (2010). Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models. Journal of Hydrology, 390(3-4), 210-221.

Li, L., & Xu, C. Y. (2014). The comparison of sensitivity analysis of hydrological uncertainty estimates by GLUE and Bayesian method under the impact of precipitation errors. Stochastic environmental research and risk assessment, 28(3), 491-504.

Li, B., He, Y., & Ren, L. (2018). Multisource hydrologic modeling uncertainty analysis using the IBUNE framework in a humid catchment. Stochastic Environmental Research and Risk Assessment, 32(1), 37-50.

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. Journal of Geophysical Research: Atmospheres, 99(D7), 14415-14428.

Lilhare, R., Déry, S.J., Pokorny, S., Stadnyk, T.A., and Koenig, K.A. (Accepted). Inter-comparison of multiple hydro-climatic datasets across the Lower Nelson River Basin, Manitoba, Canada. Journal of Atmosphere Ocean. Accepted.

Masson, D., & Knutti, R. (2011). Climate model genealogy. Geophysical Research Letters, 38(8).

Matott, L. S., Babendreier, J. E., & Purucker, S. T. (2009). Evaluating uncertainty in integrated environmental models: a review of concepts and tools. Water Resources Research, 45(6) Article No. W06421.

MacDonald, M., Stadnyk, T., Déry, S., Gustafsson, D., Isberg, K., Arheimer, B. (under revision). Improved high-latitude water storage for hydrological modelling of the Hudson Bay Drainage Basin.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., & Clark, M. (2010). Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. Hydrological Processes, 24(10), 1270-1284.

Mekis, É., & Vincent, L. A. (2011). An overview of the second generation adjusted daily precipitation dataset for trend analysis in Canada. Atmosphere-Ocean, 49(2), 163-177.

Mei, Y., Nikolopoulos, E. I., Anagnostou, E. N., & Borga, M. (2016). Evaluating satellite precipitation error propagation in runoff simulations of mountainous basins. Journal of Hydrometeorology, 17(5), 1407-1423.

Merz, R., Parajka, J., & Blöschl, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. Water Resources Research, 47(2). Article No. W02531.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jovic, D., Woollen, J., Rogers, E., Berbery, E. H., Ek, M. B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., and Shi, W. (2006). North American regional reanalysis, Bulletin of the American Meteorological Society, 87(3), 343-360.

Michaelides, S., Levizzani, V., Anagnostou, E., Bauer, P., Kasparis, T., & Lane, J. E. (2009). Precipitation: Measurement, remote sensing, climatology and modeling. Atmospheric Research, 94(4), 512-533.

Milewska, E. J., Hopkinson, R. F., & Niitsoo, A. (2005). Evaluation of geo-referenced grids of 1961–1990 Canadian temperature and precipitation normals. Atmosphere-Ocean, 43(1), 49-75.

Ndiritu, J. G. (2013). Incorporating rainfall uncertainty into catchment modelling. Journal of the South African Institution of Civil Engineering, 55(3), 36-46.

Neitsch, S. L., Arnold, J. G., Kiniry, J. R., & Williams, J. R. (2011). Soil and water assessment tool theoretical documentation version 2009. Texas Water Resources Institute.

Nikolopoulos, E. I., Anagnostou, E. N., Hossain, F., Gebremichael, M., & Borga, M. (2010). Understanding the scale relationships of uncertainty propagation of satellite rainfall through a distributed hydrologic model. Journal of Hydrometeorology, 11(2), 520-532.

Nrcan.gc.ca. (2018). Natural Resources Canada | Natural Resources Canada. [online] Available at: https://www.nrcan.gc.ca/home.

Pavelsky, T. M., & Smith, L. C. (2006). Intercomparison of four global precipitation data sets and their correlation with increased Eurasian river discharge to the Arctic Ocean. Journal of Geophysical Research: Atmospheres, 111, D21112, doi:10.1029/2006JD007230.

Pechlivanidis, I. G., Jackson, B. M., McIntyre, N. R., & Wheater, H. S. (2011). Catchment scale hydrological modelling: a review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications. Global NEST journal, 13(3), 193-214.

Pendergrass, A. G., Knutti, R., Lehner, F., Deser, C., & Sanderson, B. M. (2017). Precipitation variability increases in a warmer climate. Scientific Reports, 7(1), 17966.

Pokorny, S., Stadnky, T., Ali, G., Déry, S., & Koenig, K. (in preparation). Assessment of Ensemble-Based Gridded Climate Data and Evaluation of uncertainty in hydrologic modeling arising from input data selection.

Price, D. T., McKenney, D. W., Nalder, I. A., Hutchinson, M. F., & Kesteven, J. L. (2000). A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. Agricultural and Forest meteorology, 101(2-3), 81-94.

Priestley, C.H.B. and R.J. Taylor. 1972. On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. Monthly Weather Review, vol 100, pp81-92.

Rapaić, M., Brown, R., Markovic, M., & Chaumont, D. (2015). An evaluation of temperature and precipitation surface-based and reanalysis datasets for the Canadian Arctic, 1950–2010. Atmosphere-Ocean, 53(3), 283-303.

Razavi, S., & Gupta, H. V. (2015). What do we mean by sensitivity analysis? The need for comprehensive characterization of “global” sensitivity in Earth and Environmental systems models. Water Resources Research, 51(5), 3070-3092.

Razavi, S., & Gupta, H. V. (2016 a). A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory. Water Resources Research, 52(1), 423-439.

Razavi, S., & Gupta, H. V. (2016 b). A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application. Water Resources Research, 52(1), 440-455.

Reusser, D. E., Blume, T., Schaefli, B., & Zehe, E. (2009). Analysing the temporal dynamics of model performance for hydrological models. Hydrology and Earth System Sciences, 13(7), 999-1018.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., ... & Bloom, S. (2011). MERRA: NASA’s modern-era retrospective analysis for research and applications. Journal of Climate, 24(14), 3624-3648.

Rouse, W. R. (1991). Impacts of Hudson Bay on the terrestrial climate of the Hudson Bay Lowlands. Arctic and Alpine Research, 23(1), 24-30.

Ruosteenoja, K., Tuomenvirta, H., & Jylhä, K. (2007). GCM-based regional temperature and precipitation change estimates for Europe under four SRES scenarios applying a super-ensemble pattern-scaling method. Climatic Change, 81(1), 193-208.

Sagan, K. A. B. (2017). Sensitivity of probable maximum flood estimates in the Lower Nelson River Basin (MSc Thesis, Department of Civil Engineering, University of Manitoba, Manitoba).

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015). A representative democracy to reduce interdependency in a multimodel ensemble. Journal of Climate, 28(13), 5171-5194.

Semenov, M. A., & Stratonovitch, P. (2010). Use of multi-model ensembles from global climate models for assessment of climate change impacts. Climate Research, 41(1), 1-14.

Shabbar, A., & Khandekar, M. (1996). The impact of El Niño-Southern Oscillation on the temperature field over Canada: Research note. Atmosphere-Ocean, 34(2), 401-416.

Shafii, M., Tolson, B., & Matott, L. S. (2015). Addressing subjective decision-making inherent in GLUE-based multi-criteria rainfall–runoff model calibration. Journal of Hydrology, 523, 693-705.

Sheffield, J., Goteti, G., & Wood, E. F. (2006). Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling. Journal of Climate, 19(13), 3088-3111.

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. Journal of Geophysical Research: Atmospheres, 118(4), 1716-1733.

SMHI. (2018). HYPE model documentation. Retrieved from: http://www.smhi.net/hype/wiki/doku.php

Smith, A. (2015). Utilizing lumped coupled tracer-aided modelling to identify temporal trends in basin-scale evapotranspiration partitioning.

Stedinger, J. R., Vogel, R. M., Lee, S. U., & Batchelder, R. (2008). Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. Water Resources Research, 44(12) W00B06, doi:10.1029/2008WR006822.

Tang, B. (1993). Orthogonal array-based Latin hypercubes. Journal of the American statistical association, 88(424), 1392-1397.

Tasdighi, A., Arabi, M., & Harmel, D. (2018). A probabilistic appraisal of rainfall-runoff modeling approaches within SWAT in mixed land use watersheds. Journal of Hydrology, 564, 476-489.

Tefs, A. A. (2018). Simulating hydroelectric regulation and climate change in the Hudson Bay drainage basin (MSc Thesis, Department of Civil Engineering, University of Manitoba, Manitoba).

Teweldebrhan, A. T., Burkhart, J. F., & Schuler, T. V. (2018). Parameter uncertainty analysis for an operational hydrological model using residual-based and limits of acceptability approaches. Hydrology and Earth System Sciences, 22(9), 5021-5039.

Trenberth, K. E. (1997). The definition of El Niño. Bulletin of the American Meteorological Society, 78(12), 2771-2778.

Tustison, B., Harris, D., & Foufoula-Georgiou, E. (2001). Scale issues in verification of precipitation forecasts. Journal of Geophysical Research: Atmospheres, 106(D11), 11775-11784.

USACE. (2016). Hydrologic modeling system HEC-HMS. User's manual.

Uusitalo, L., Lehikoinen, A., Helle, I., & Myrberg, K. (2015). An overview of methods to evaluate uncertainty of deterministic models in decision support. Environmental Modelling & Software, 63, 24-31.

Van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., ... & Masui, T. (2011). The representative concentration pathways: an overview. Climatic change, 109(1-2), 5.

Vincent, L. A., Wang, X. L., Milewska, E. J., Wan, H., Yang, F., & Swail, V. (2012). A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis. Journal of Geophysical Research: Atmospheres, 117(D18).

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., & Gupta, H. V. (2003). Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. Hydrological Processes, 17(2), 455-476.

Weedon, G. P., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., ... & Best, M. (2011). Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century. Journal of Hydrometeorology, 12(5), 823-848.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. Water Resources Research, 50(9), 7505-7514.

Wilks, D. S. (2006). Comparison of ensemble-MOS methods in the Lorenz'96 setting. Meteorological Applications, 13(3), 243-256.

World Meteorological Organization. (1989). Calculation of Monthly and Annual 30-Year Standard Normals. WCDP-No. 10, WMO-TD/No. 341, World Meteoro-logical Organization.

Wong, J. S., Razavi, S., Bonsal, B. R., Wheater, H. S., & Asong, Z. E. (2017). Inter-comparison of daily precipitation products for large-scale hydro-climatic applications over Canada. Hydrology and Earth System Sciences, 21(4), 2163.

Yang, C., Yan, Z., & Shao, Y. (2012). Probabilistic precipitation forecasting based on ensemble output using generalized additive models and Bayesian model averaging. Acta Meteorologica Sinica, 26(1), 1-12.

Yao, Y., Liang, S., Xie, X., Cheng, J., Jia, K., Li, Y., & Liu, R. (2014). Estimation of the terrestrial water budget over northern China by merging multiple datasets. Journal of hydrology, 519, 50-68.

Zhou, R., Li, Y., Lu, D., Liu, H., & Zhou, H. (2016). An optimization based sampling approach for multiple metrics uncertainty analysis using generalized likelihood uncertainty estimation. Journal of Hydrology, 540, 274-286.

Appendix A

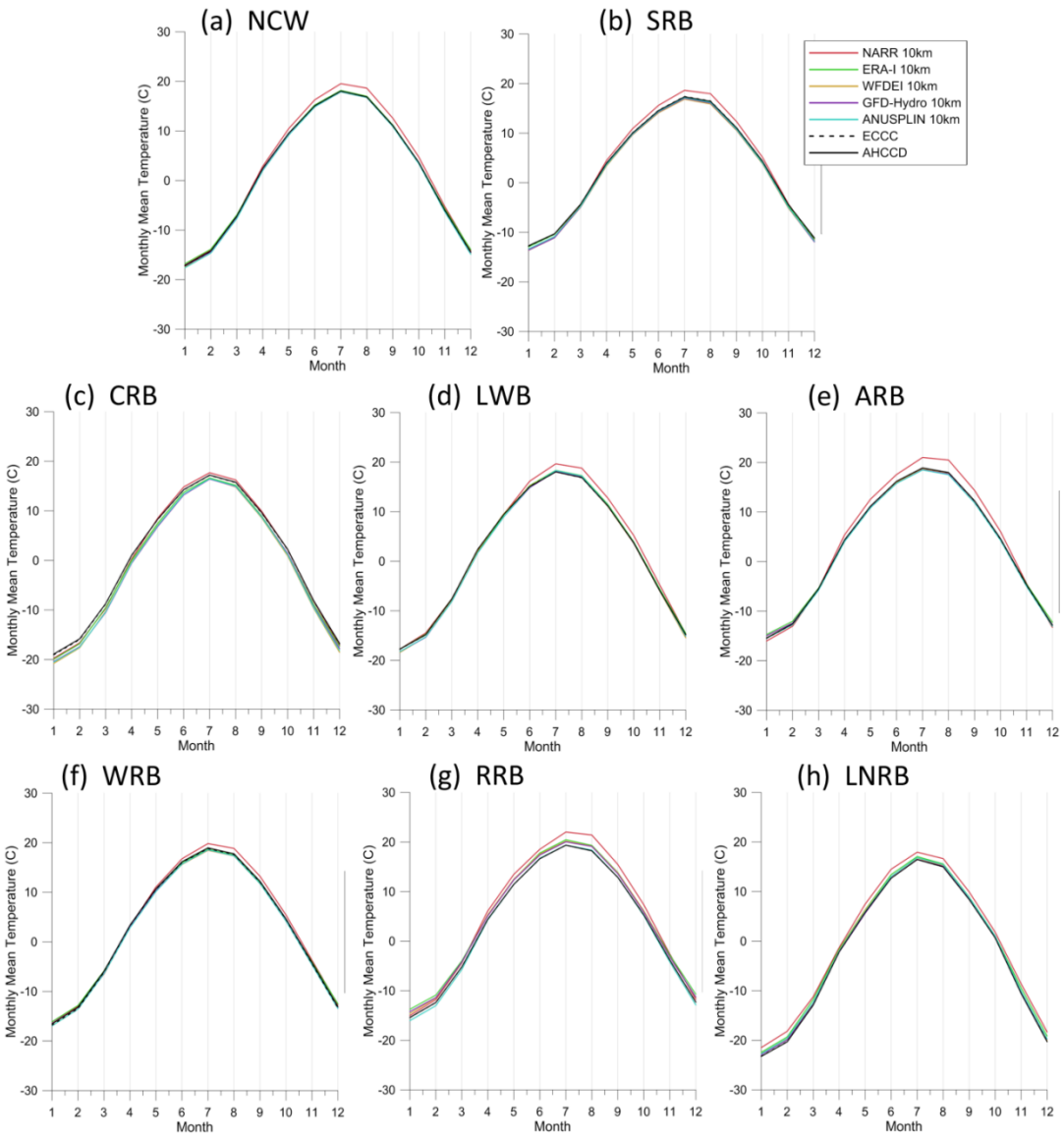

Figure A.1: Period mean monthly mean temperatures for the period of 1981-2010, spatially averaged over each of the major basins.

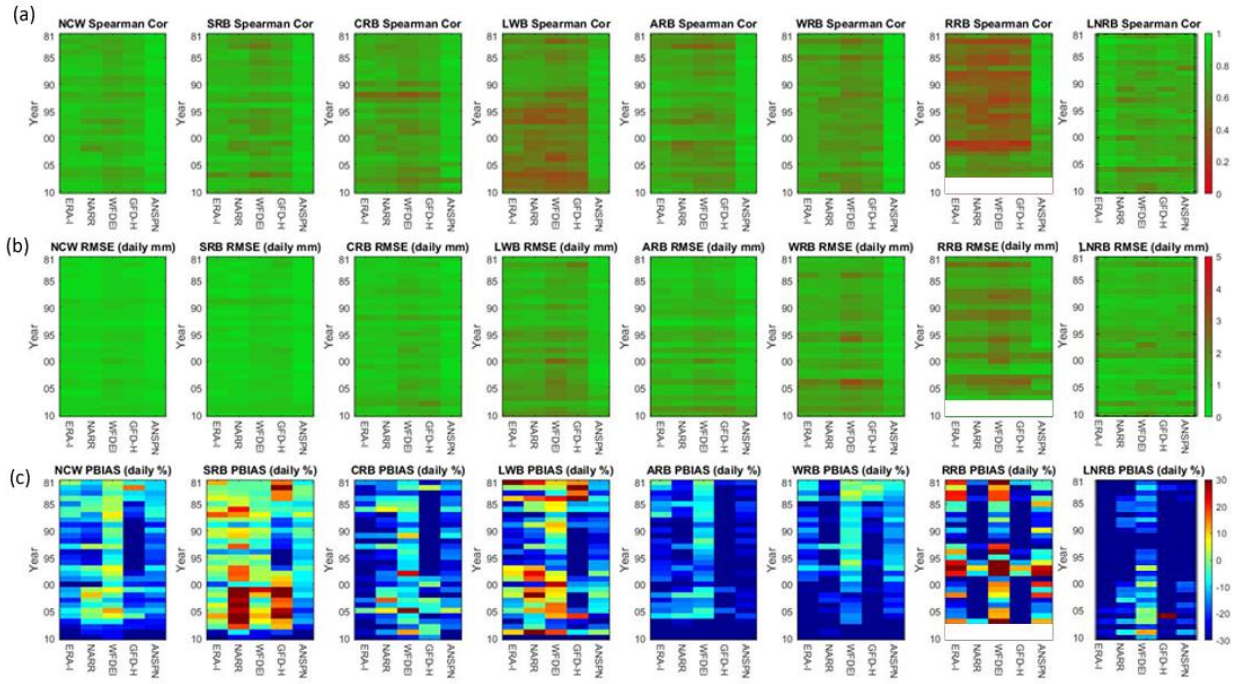

Figure A.2: Daily precipitation spatially aggregated continuous winter statistics with reference to the AHCCD observed dataset in each major basin. (a) daily Spearman correlation (b) daily RMSE, and (c) daily PBIAS. White is used to represent periods with no available data.

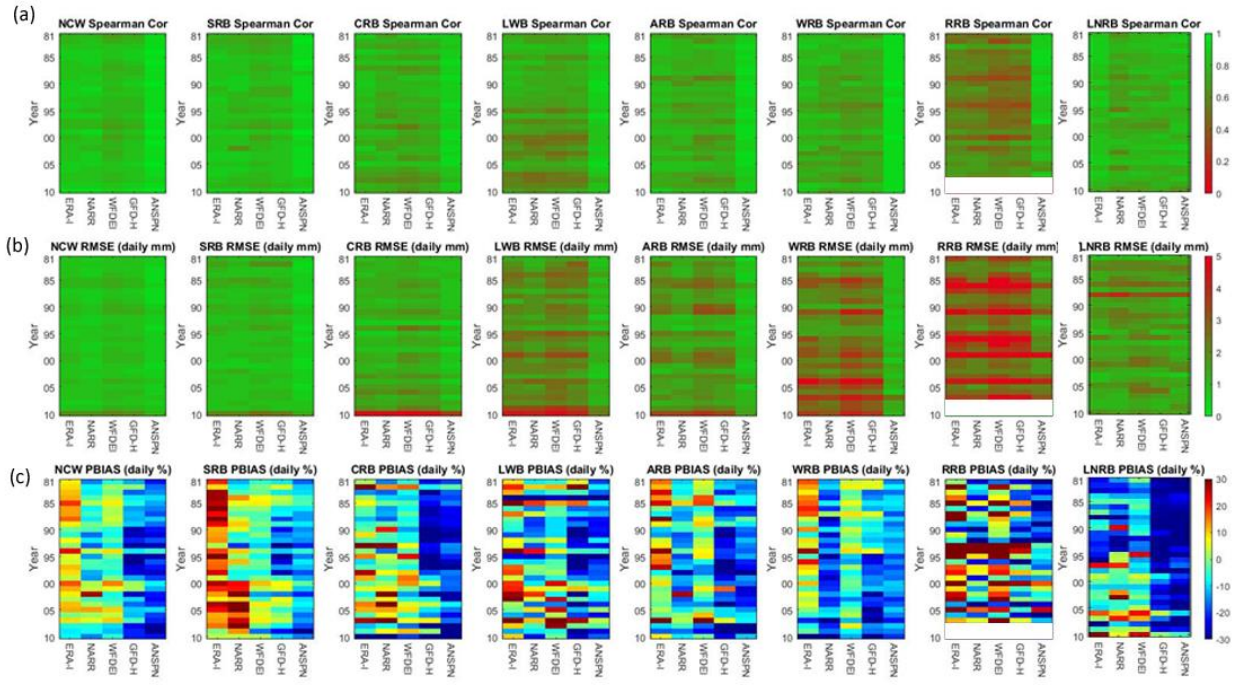

Figure A.3: Daily precipitation spatially aggregated continuous spring statistics with reference to the AHCCD observed dataset in each major basin. (a) daily Spearman correlation (b) daily RMSE, and (c) daily PBIAS. White is used to represent periods with no available data.

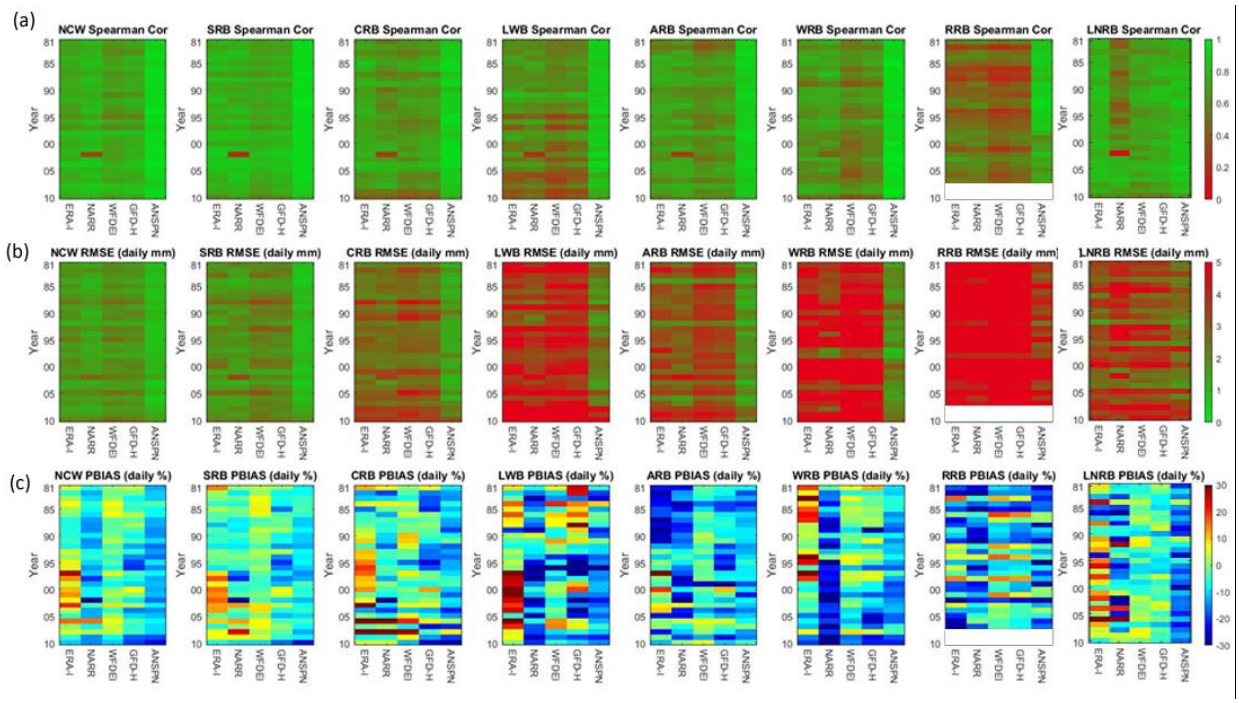

Figure A.4: Daily precipitation spatially aggregated continuous summer statistics with reference to the AHCCD observed dataset in each major basin. (a) daily Spearman correlation (b) daily RMSE, and (c) daily PBIAS. White is used to represent periods with no available data.

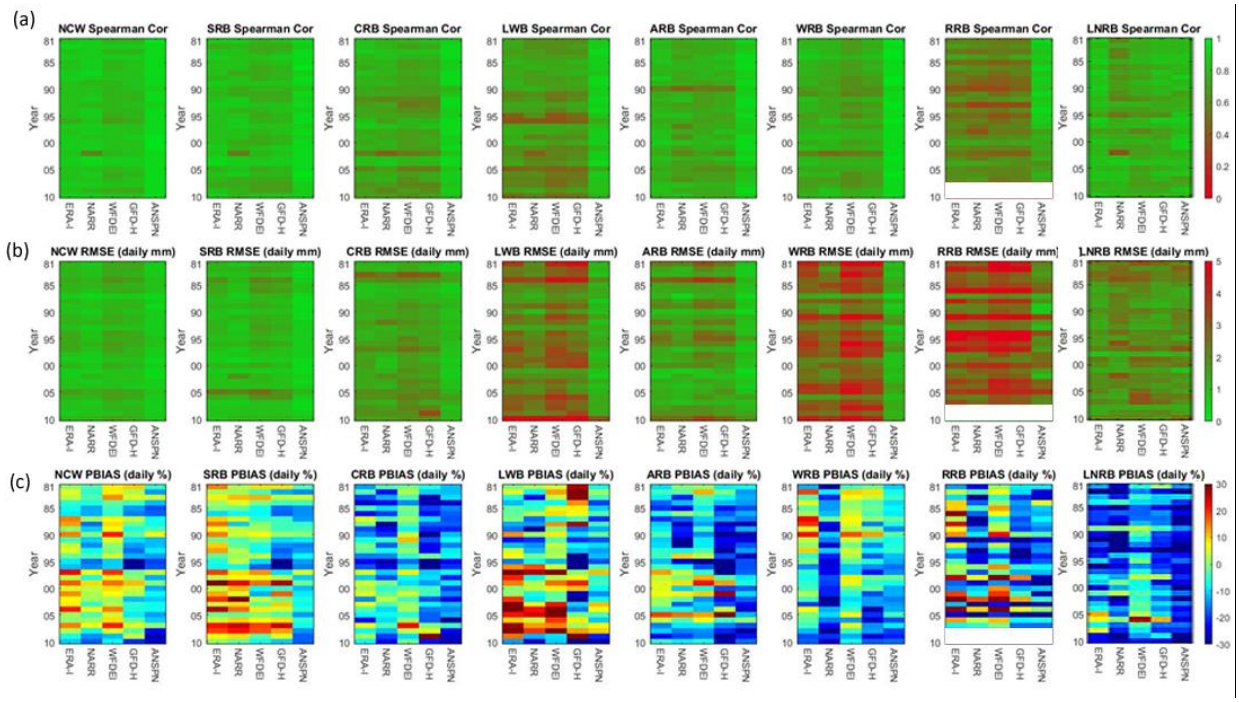

Figure A.5: Daily precipitation spatially aggregated continuous autumn statistics with reference to the AHCCD observed dataset in each major basin. (a) daily Spearman correlation (b) daily RMSE, and (c) daily PBIAS. White is used to represent periods with no available data.

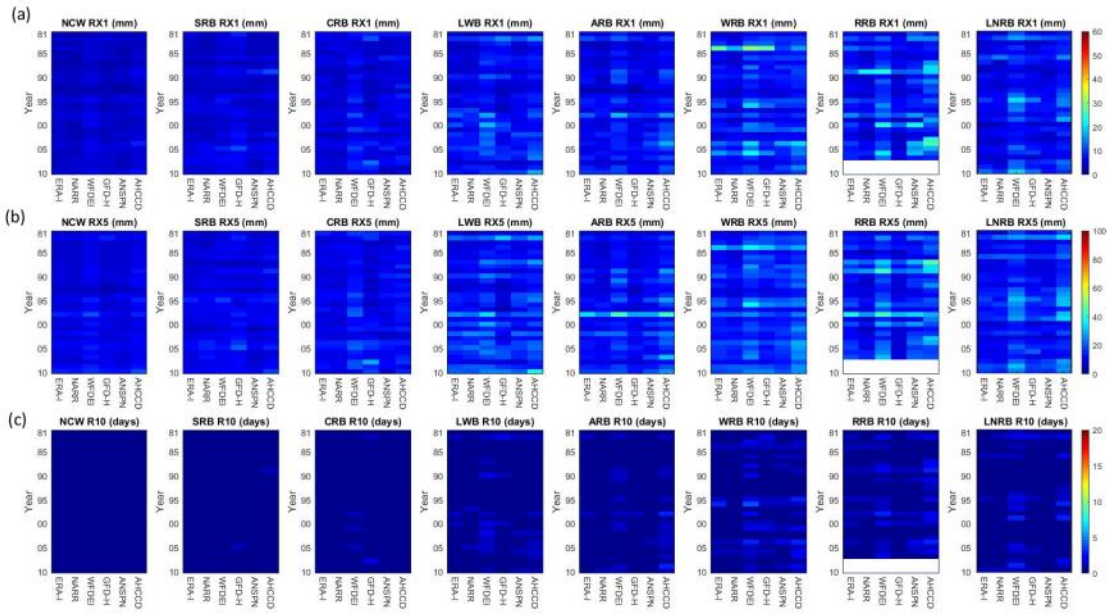

Figure A.6: Daily precipitation spatially aggregated winter extreme indexes in each major basin. (a) RX1 (b) RX5, and (c) R10. White is used to represent periods with no available station data

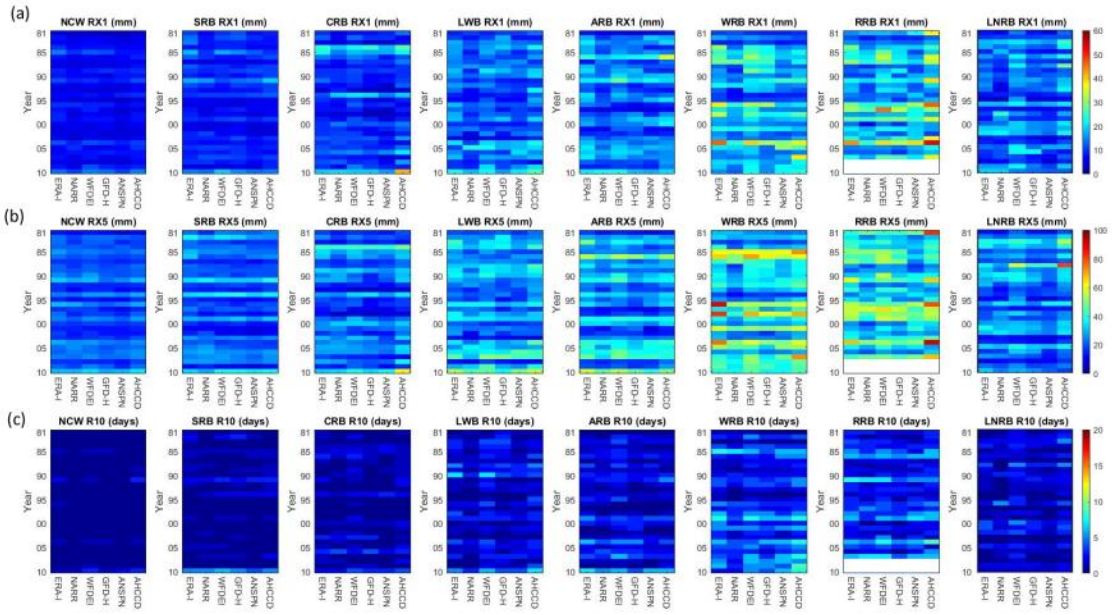

Figure A.7: Daily precipitation spatially aggregated spring extreme indexes in each major basin. (a) RX1 (b) RX5, and (c) R10. White is used to represent periods with no available station data

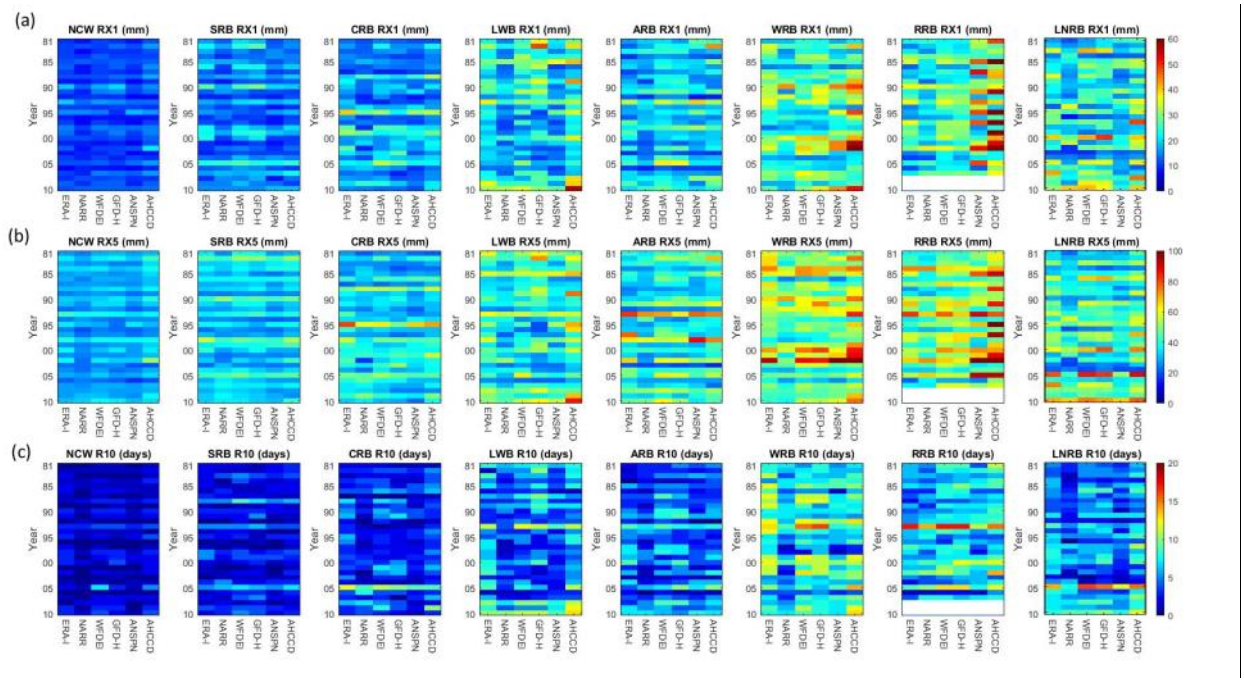

Figure A.8: Daily precipitation spatially aggregated summer extreme indexes in each major basin. (a) RX1 (b) RX5, and (c) R10. White is used to represent periods with no available station data

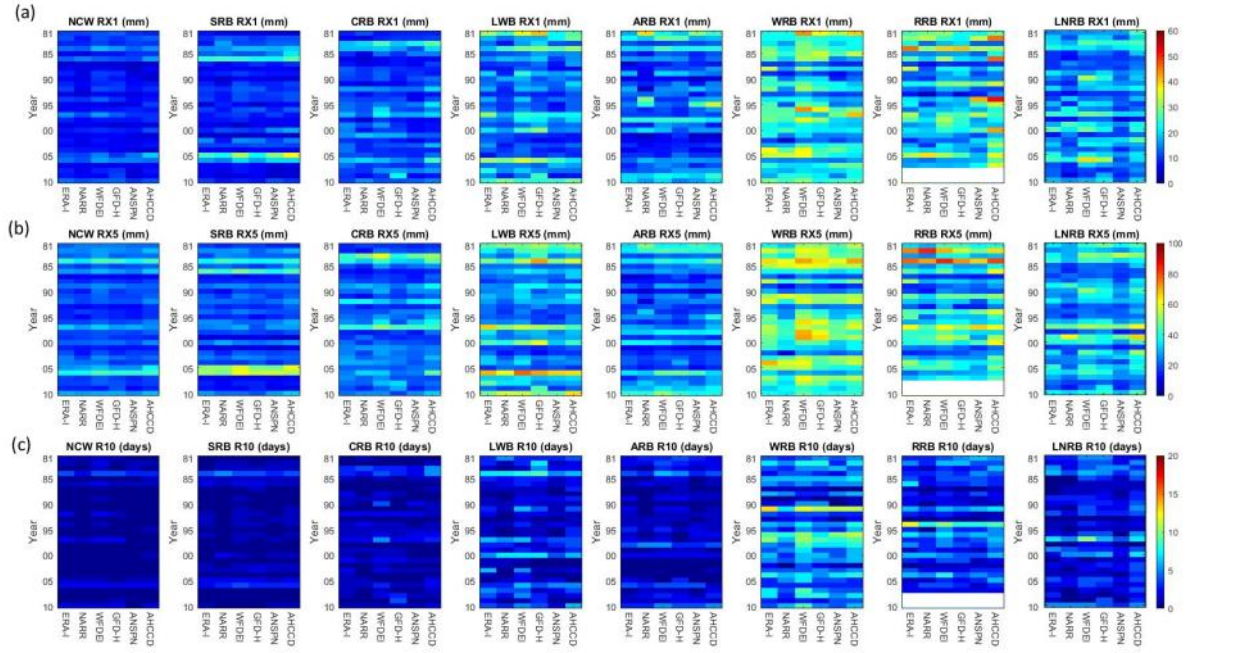

Figure A.9: Daily precipitation spatially aggregated autumn extreme indexes in each major basin. (a) RX1 (b) RX5, and (c) R10. White is used to represent periods with no available station data

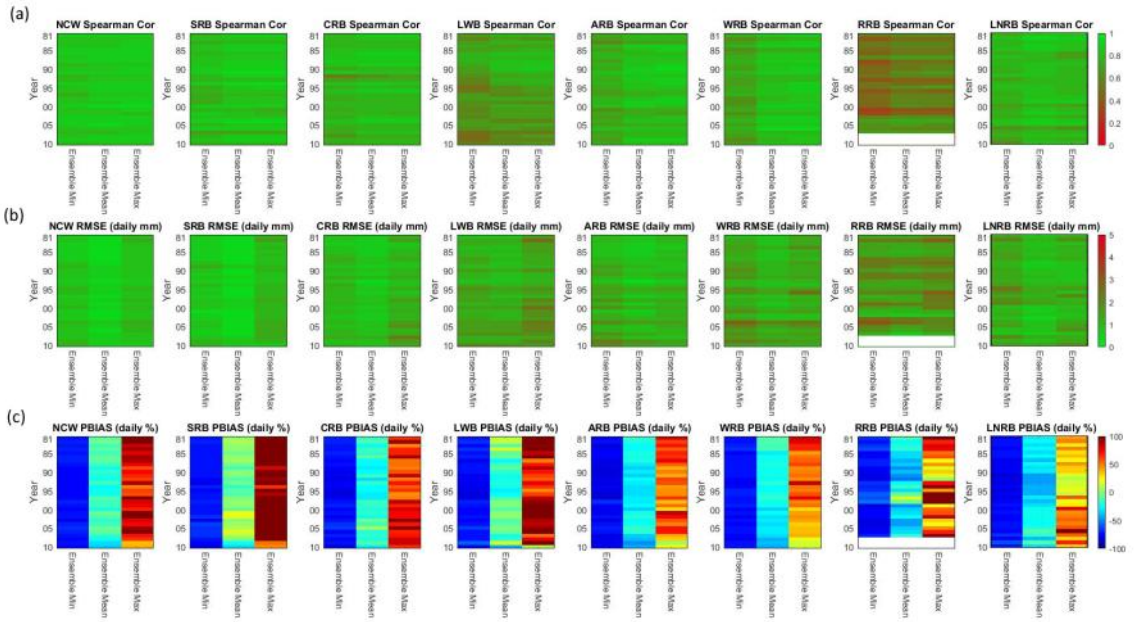

Figure A.10: Basin-averaged daily precipitation continuous winter statistics with reference to the AHCCD observed data set in each major basin for (a) daily Spearman correlation (b) daily RMSE, and (c) daily PBIAS for the ensemble minimum, mean, and maximum.

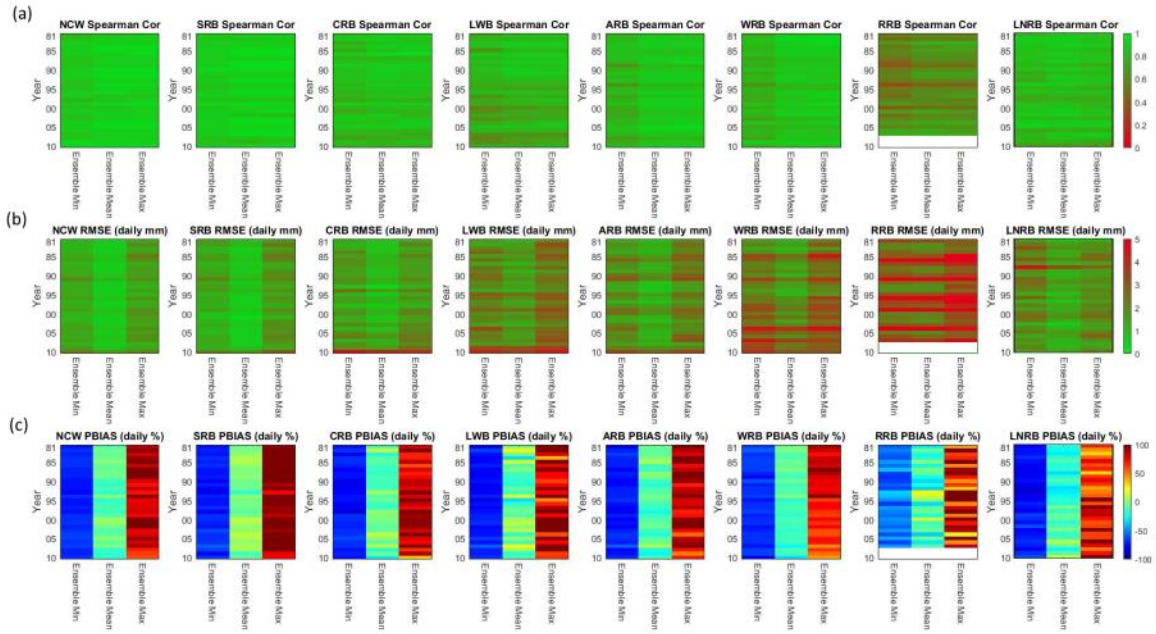

Figure A.11: Basin-averaged daily precipitation continuous spring statistics with reference to the AHCCD observed data set in each major basin for (a) daily Spearman correlation (b) daily RMSE, and (c) daily PBIAS for the ensemble minimum, mean, and maximum.

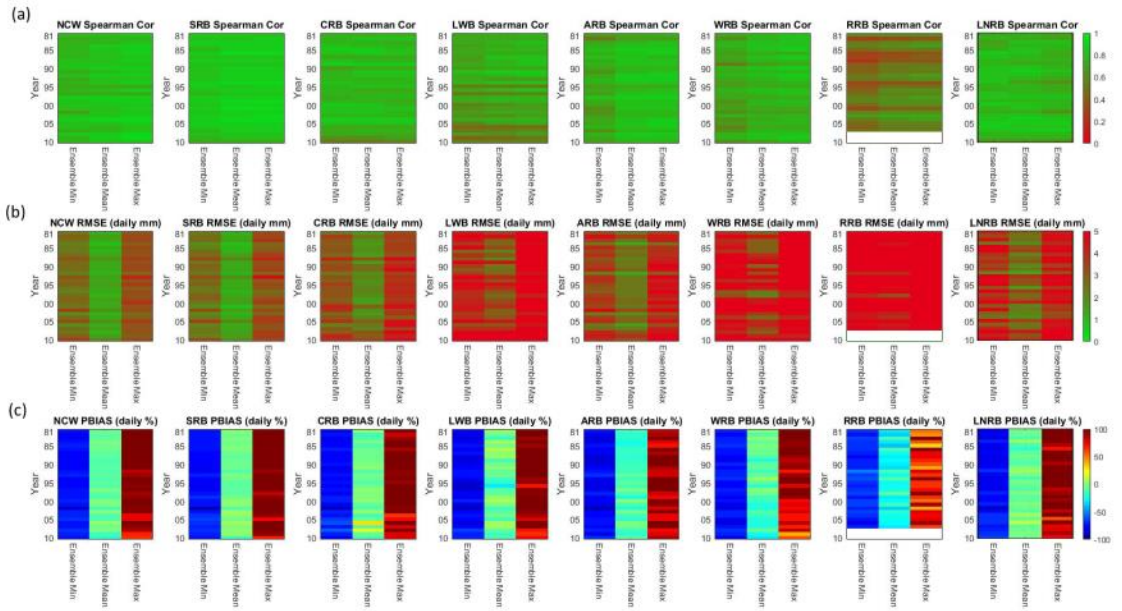

Figure A.12: Basin-averaged daily precipitation continuous summer statistics with reference to the AHCCD observed data set in each major basin for (a) daily Spearman correlation (b) daily RMSE, and (c) daily PBIAS for the ensemble minimum, mean, and maximum.

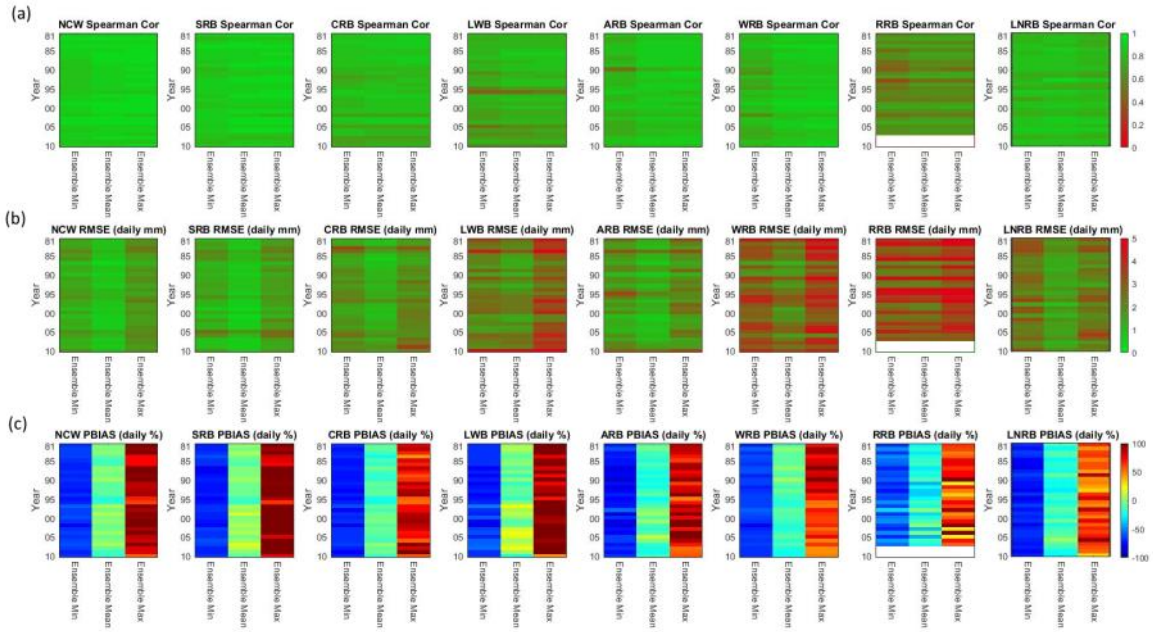

Figure A.13: Basin-averaged daily precipitation continuous autumn statistics with reference to the AHCCD observed data set in each major basin for (a) daily Spearman correlation (b) daily RMSE, and (c) daily PBIAS for the ensemble minimum, mean, and maximum.

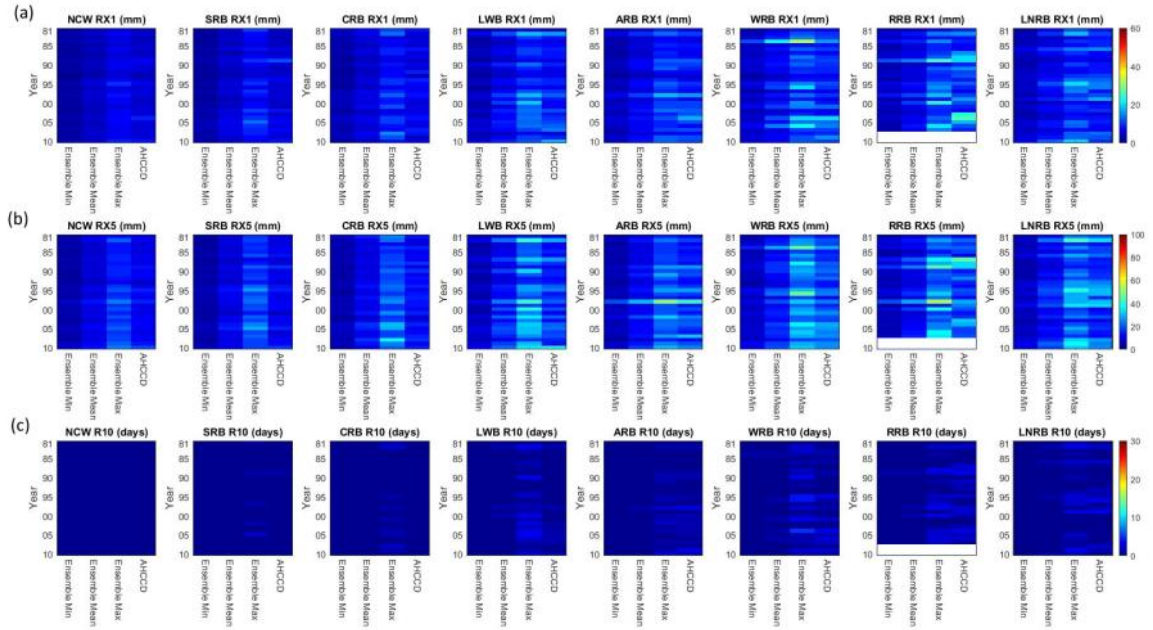

Figure A.14: Basin-averaged daily precipitation winter extreme indexes in each major basin for (a) RX1 (b) RX5, and (c) R10 for the ensemble minimum, mean, and maximum.

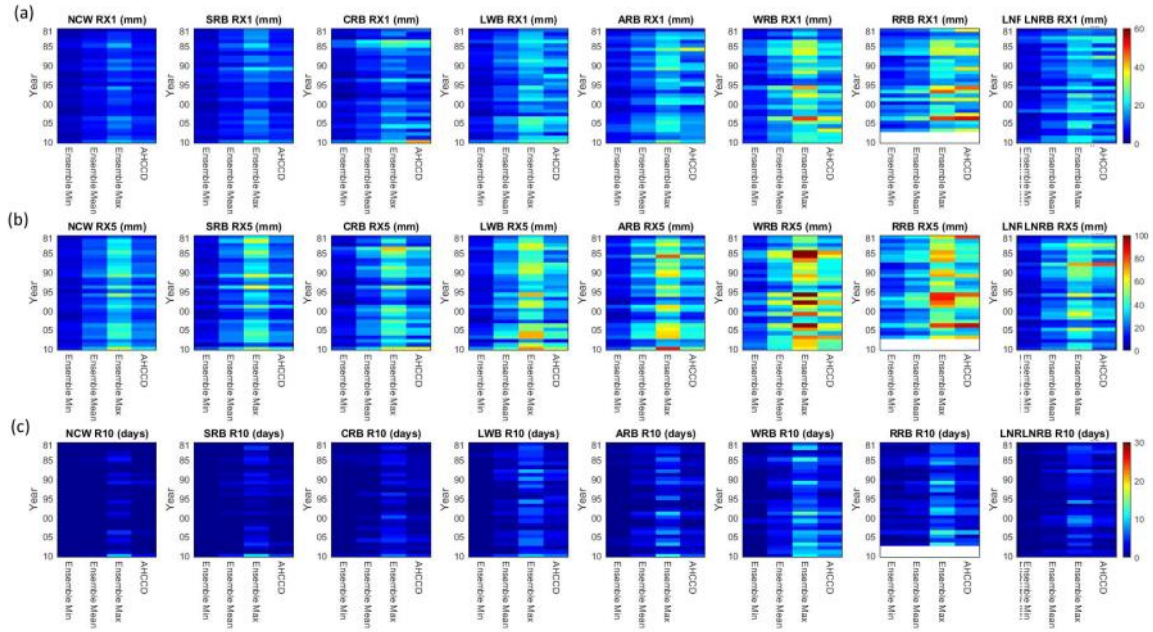

Figure A.15: Basin-averaged daily precipitation spring extreme indexes in each major basin for (a) RX1 (b) RX5, and (c) R10 for the ensemble minimum, mean, and maximum.

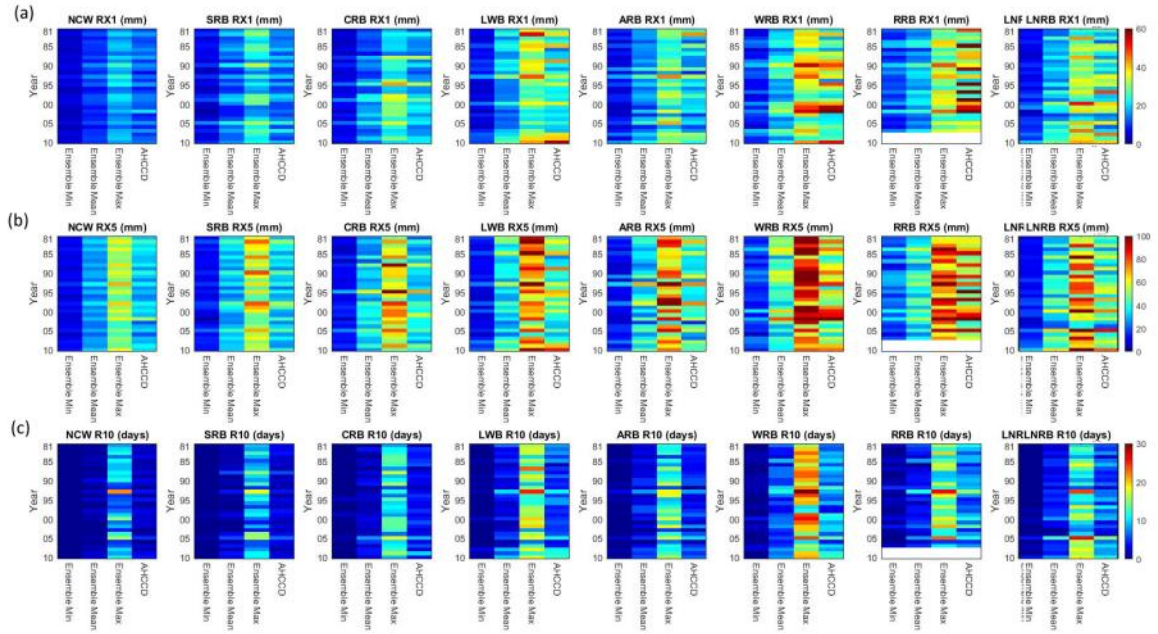

Figure A.16: Basin-averaged daily precipitation summer extreme indexes in each major basin for (a) RX1 (b) RX5, and (c) R10 for the ensemble minimum, mean, and maximum.

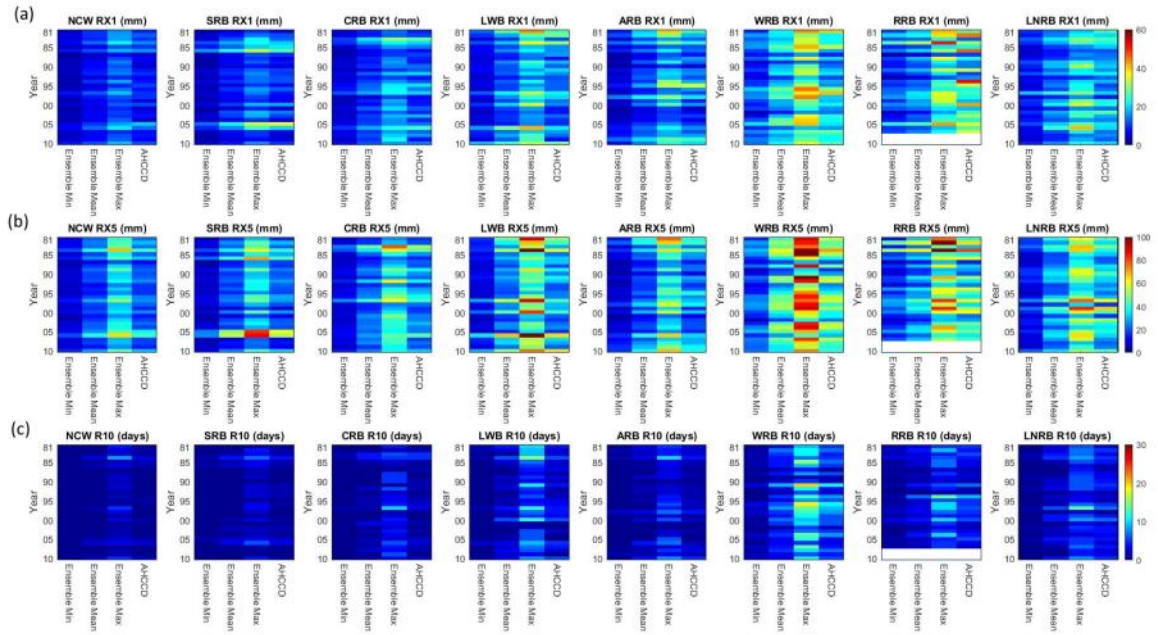

Figure A.17: Basin-averaged daily precipitation autumn extreme indexes in each major basin for (a) RX1 (b) RX5, and (c) R10 for the ensemble minimum, mean, and maximum.

Appendix B

Table B. 1: Hydrometric gauges In the LNRB.

Station Name	Station ID	Longitude	Latitude	Data Availability	Gauged Area
Sapochi River Near Nelson House	05TG006	-98.49	55.80	1993-2016	391
Footprint River Above Footprint Lake	05TF002	-98.84	55.99	1977-2017	643
Taylor River Near Thompson	05TG002	-98.00	55.60	1970-2017	886
Kettle River Near Gillam	05UF004	-94.69	56.34	1963-2017	1090
Angling River Near Bird	05UH001	-93.64	56.67	1979-2017	1560
Weir River Above The Mouth	05UH002	-93.45	57.02	1977-2017	2190
Limestone River Near Bird	05UG001	-94.21	56.51	1963-2016	3270
Burntwood River Above Leaf Rapids	05TE002	-99.22	55.49	1985-2017	5810
Odie River Near Thompson	05TG003	-97.35	55.99	1979-2017	6110
Rat River Below Notigi	05TF710	-99.33	55.86	1979-2019	6140
Grass River above Standing Stone Falls	05TD001	-97.00	55.74	1915-2017	15400
Burntwood River Near Thompson	05TG001	-97.89	55.74	1956-2018	18500
Nelson River At Kelsey GS	05UE005	-96.58	55.93	1960-2016	1050000
Nelson River At Kettle Generating Station	05UF006	-94.63	56.39	1987-2016	1100000
Nelson River At Long Spruce Generating Station	05UF007	-94.36	56.39	1987-2016	1100000
Nelson River (East Channel) Below Sea River Falls	05UB008	-97.59	54.33	1967-2017	
Nelson River (West Channel) At Jenpeg	05UB009	-98.04	54.49	1967-2016	

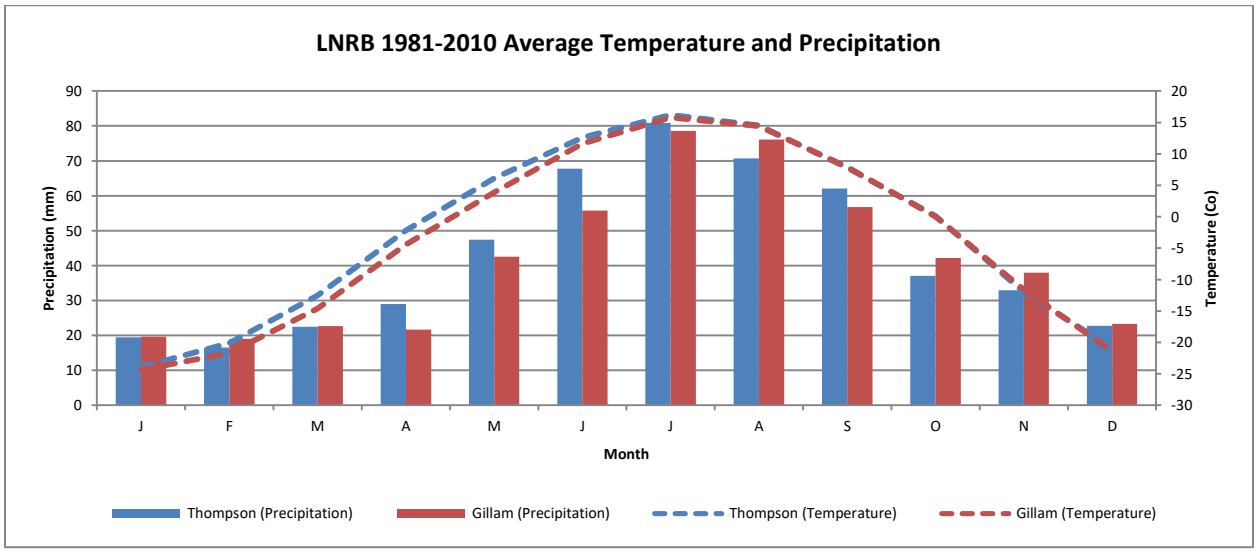

Figure B.1: 30-year (1981-2010) average monthly precipitation and temperature for the Gillam and Thompson climate gauges

Table B. 2: Selected HEC-HMS parameters with accompanied parameter search ranges. Parameters grouped in five groups and varied together in their respective groups. The groups include 1) Upstream of Burntwood below Thompson gauge and The Odie River basin; 2) The Grass River basin; 3) Tylor, Upper Burntwood, Footprint, and Wuskatim local basins; 4) The Nelson River contributing area upstream of the Kelsey GS but downstream of Jenpeg; 5) All remaining LNRB area.

HEC-HMS Parameters[2]	Minimum value	Maximum Value
Surface Storage	5	75
Maximum Soil Infiltration	0.6	4
Maximum Soil Percolation	0.25	1.25
Ground Water 1 Routing Coefficient	400	2400
Ground Water 1 Maximum Percolation	0.45	1.5
Ground Water 2 Maximum Percolation	0.45	1.5
Time of Concentration	100	680
Storage Coefficient	170	920
Muskingum K	10	100
Muskingum X	0.1	0.4
Base Temperature	-5	5
Melt Rate ATI Decay Factor	0.9	0.99
Rain Melt Rate	0	10
Snow vs Rain Temperature	-2	2
Clear Sky Transmittance	0.5	0.9
Temperature-Range Exponent	1.8	3
Priestley-Taylor Coefficient	1.1	1.5

Melt rate	Min	0.01	0.05	0.1	0.5	1	2
	Max	6	6	6	6	6	6

[2] More information is available on the function of each parameter in Sagan (2017) and USACE. (2016).

Table B. 3: WATFLOOD selected parameters and accompanied search ranges

WATFLOOD landclass parameters[3]		conifereous	mixed forrest	treed rock	shrub	bogs	wetland (disconnected)	Water
Infiltration Coeficient (akfs)	Min	1	1	1	1	0.04	0.04	
	Max	5	5	5	5	500	500	
Interflow coefficient (rec)	Min	0.5	0.5	0.5	0.5	0.5	0.5	
	Max	4	4	4	4	4	4	
Interception evaporation factor (fpet)	Min							0.5
	Max							1.1
Upper Zone retention (retn)	Min	20	20	10	10	30	30	
	Max	150	150	150	150	400	400	
Recharge coefficient (bare ground) (ak2)	Min	0.001	0.001	0.001	0.001	0.001	0.001	
	Max	0.2	0.2	0.2	0.2	0.2	0.2	
Recharge coefficient (Snow Covered) [4](ak2fs)	Min	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	
	Max	0.02	0.02	0.02	0.02	0.02	0.02	
Melt factor (fm)	Min	0.05[5]	0.05	0.05	0.05	0.05	0.05	0.05
	Max	0.25	0.25	0.25	0.25	0.25	0.25	0.25
WATFLOOD basin parameters		Rat / Nelson / footprint	Upper burntwood	Grass / Minago	Kettle / Limestone	Odei		
Lower zone Coefficient (flz)	Min	0.00001	0.00001	0.00001	0.00001	0.00001		
	Max	0.005	0.005	0.005	0.005	0.005		
Lower zone exponent (pwr)	Min	2	2	2	2	2		
	Max	4	4	4	4	4		
Channel Mannings n (r2n)	Min	0.001	0.001	0.001	0.001	0.001		
	Max	0.02	0.02	0.02	0.02	0.02		
Wetland or bank porosity (theta)	Min	0.1	0.1	0.1	0.1	0.1		
	Max	0.95	0.95	0.95	0.95	0.95		
Wetland/bank lateral conductivity (kcond)	Min	0.1	0.1	0.1	0.1	0.1		
	Max	0.25	0.25	0.25	0.25	0.25		

[3] More information is available from Holmes (2016) and Kouwen (2018)

[4] Parameters in blue are varied together. Individual ak2 values are generated for each landclass and are applied to ak2fs by: ak2fs = 0.01*ak2

[5] Parameters in red are varied as one

Table B. 4: HYPE selected parameters and accompanied search ranges

HYPE Parameters[6]	Min	Max
srrc_corr	0.8	1.2
rrc_corr	0.9	1.2
kc_corr	0.9	1.4
fc_corr	0.8	1.3
wp_corr	0.8	1.1
deprl_corr	0.6	1.6
kc (lake)	0.7	1.3
kc (wetland)	0.4	0.9
kc (crops)	0.7	1.3
kc (forest)	0.4	0.9
kc (open)	0.7	1.3
wcfc (coarse)	0.05	0.25
wcfc (medium)	0.1	0.3
wcfc (organic)	0.3	0.5
wcfc (shallow)	0.05	0.15
ilrratp (iLake 3)	1	2
ilrratk (iLake 1)	2.1	70
ilrratk (iLake 3)	2	100
olrratp (oLake 1)	1	5
olrratp (oLake 2)	1	5
olrratp (oLake 4)	1	4.5
olrratk (oLake 3)	1.2	97
fpsno_corr	0.75	1.15

[6] More information available in Tefs (2018), Macdonald (under revision), and SMHI (2018)

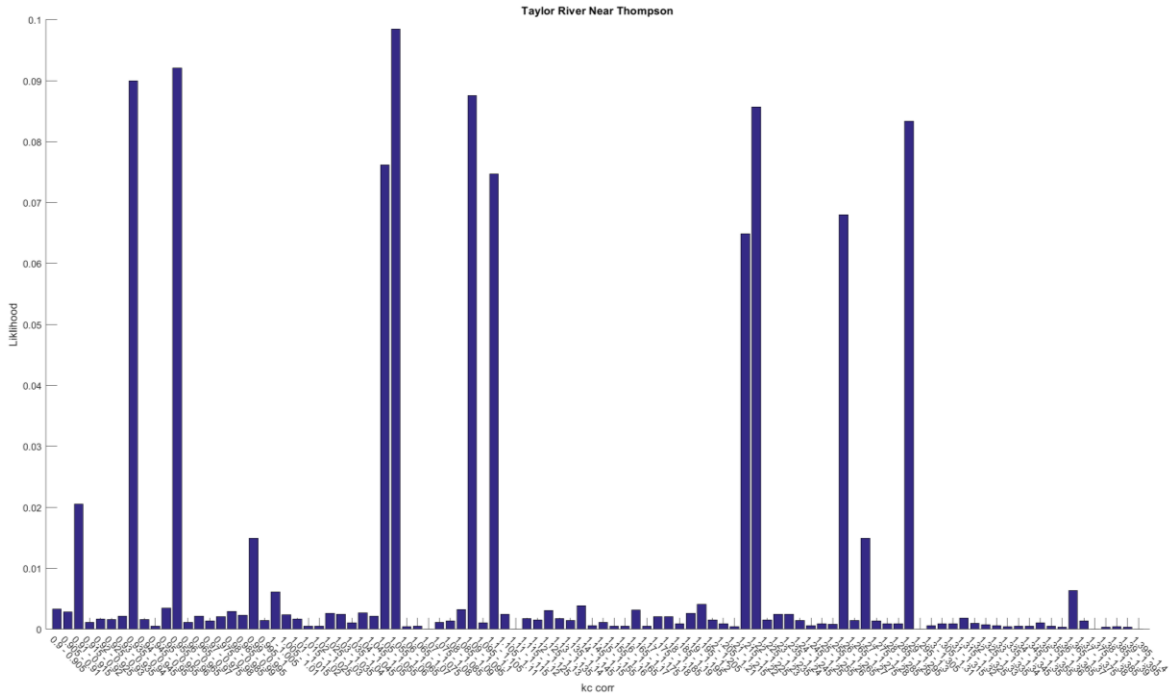

Figure B.2: Parameter distribution for the "kc corr" HYPE parameter without filtering duplicate samples created by the VARS star based sampling.

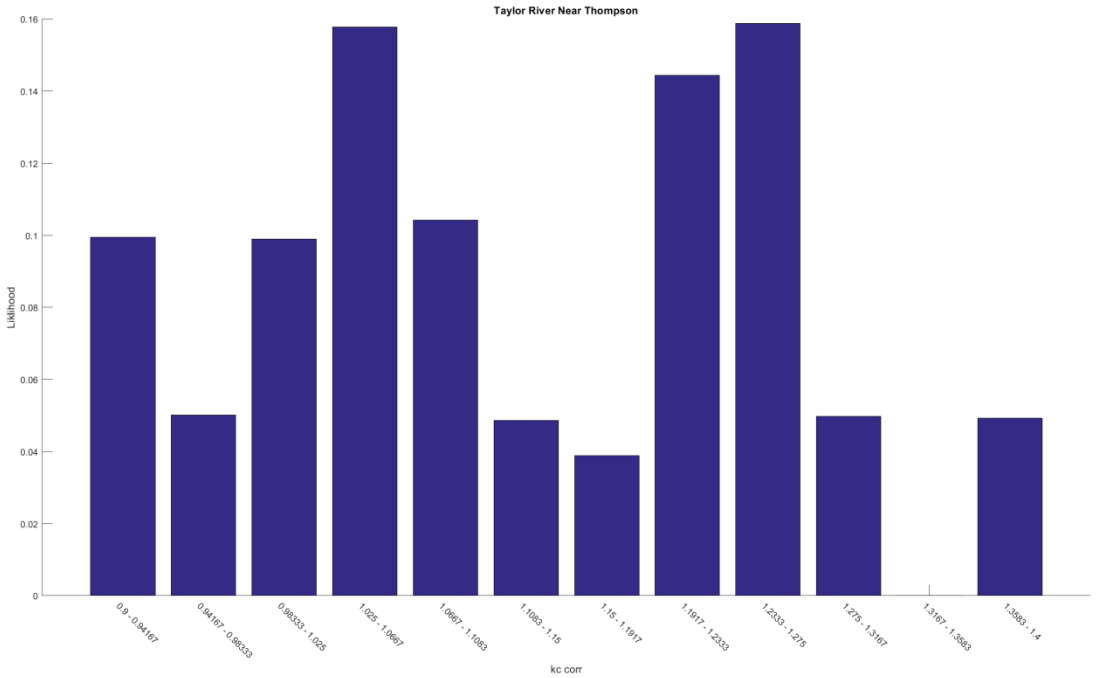

Figure B.3: Parameter distribution for the "kc corr" HYPE parameter with filtering duplicate samples created by the VARS star based sampling

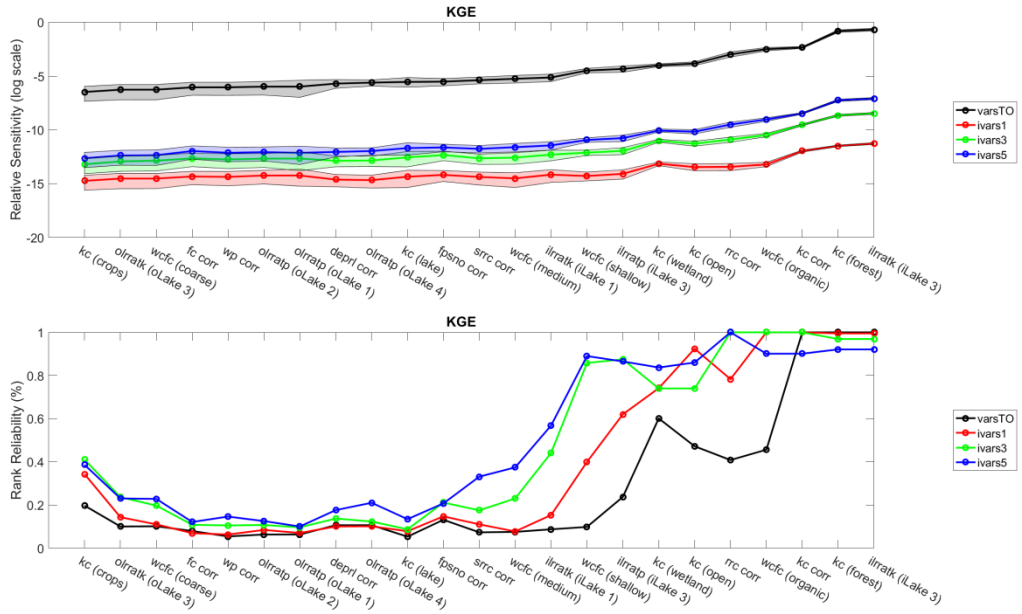

Figure B.4: VARS results for period KGE scores produced by HYPE.

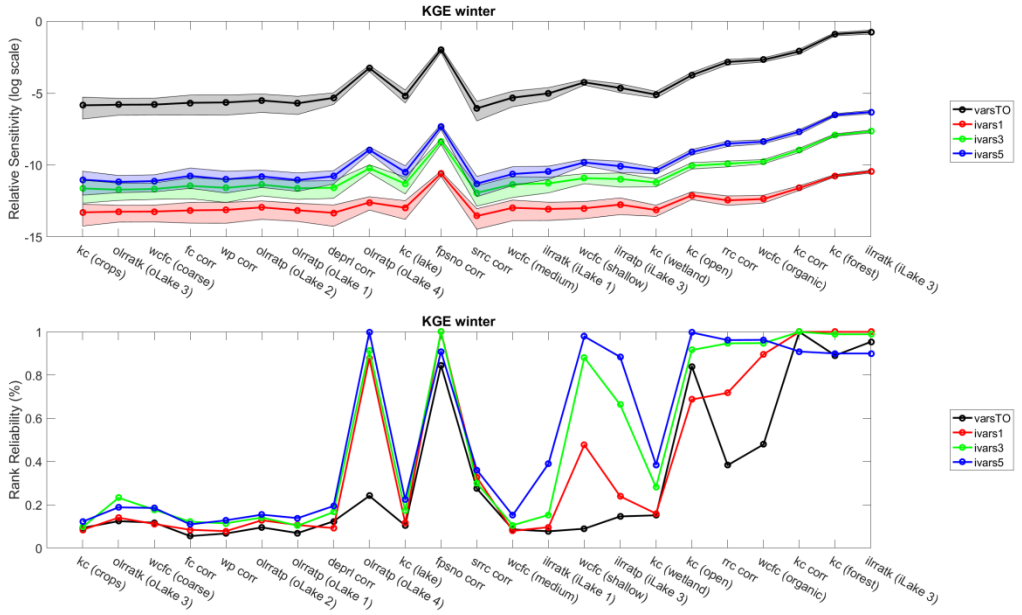

Figure B.5: VARS results for winter KGE scores produced by HYPE.

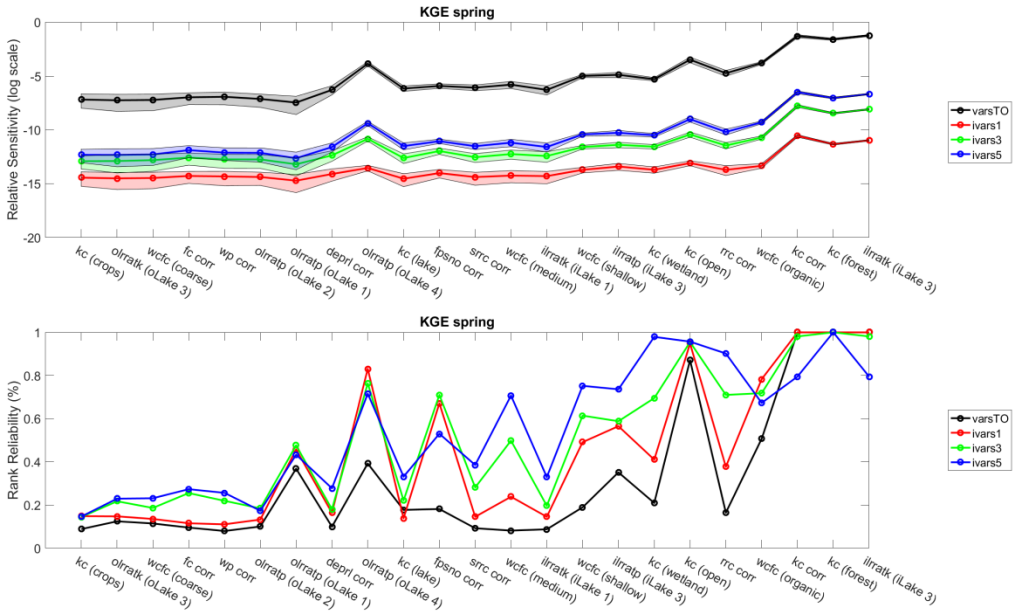

Figure B.6: VARS results for spring KGE scores produced by HYPE.

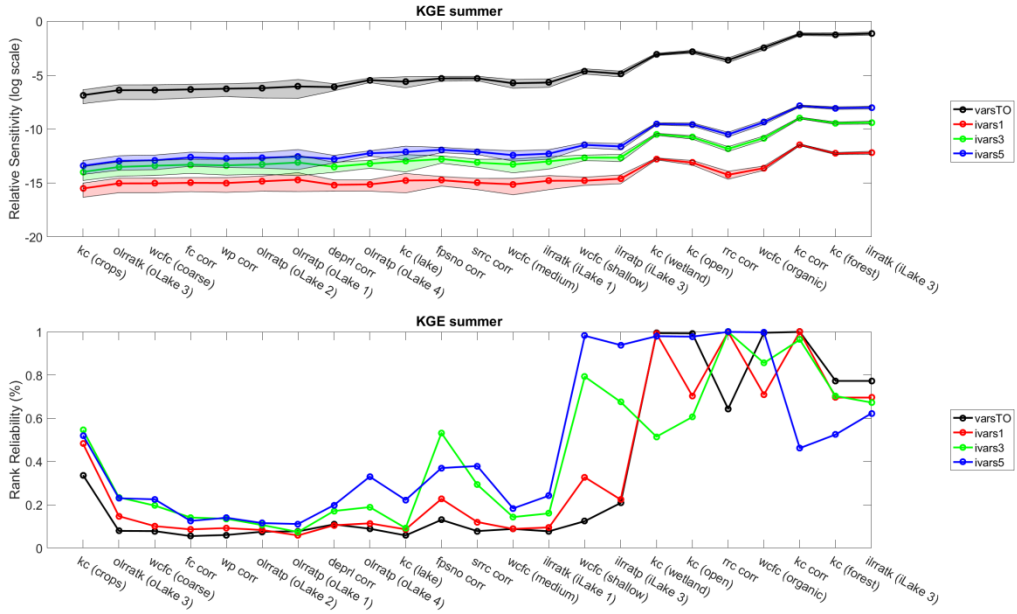

Figure B.7: VARS results for summer KGE scores produced by HYPE.

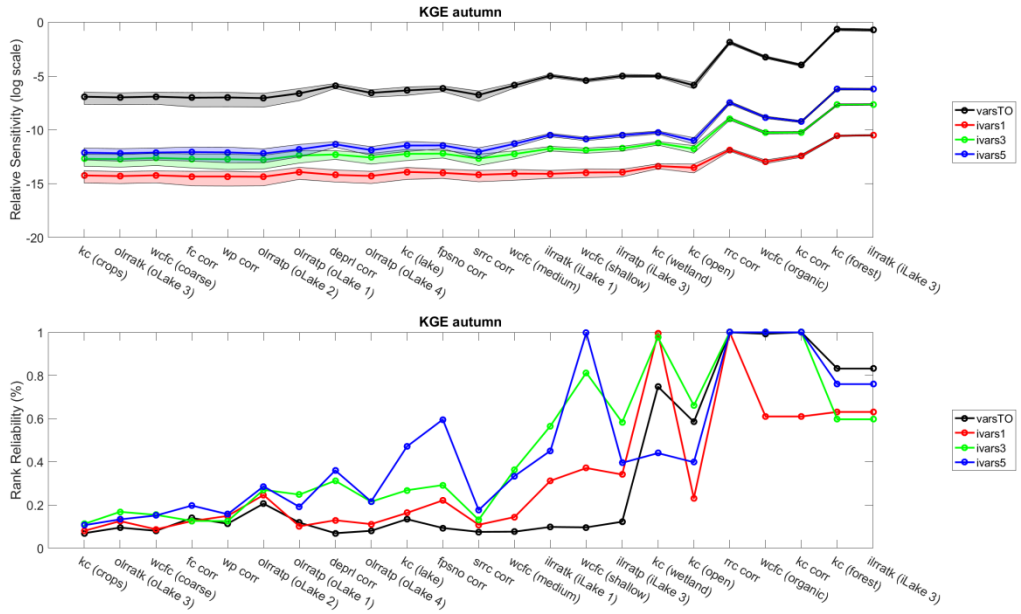

Figure B.8: VARS results for autumn KGE scores produced by HYPE.

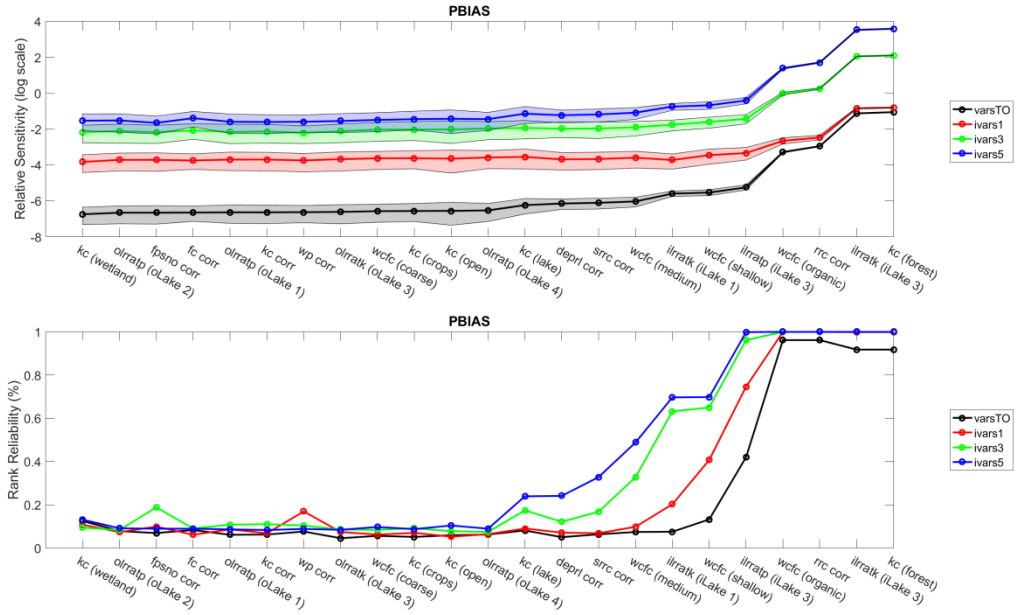

Figure B.9: VARS results for period PBIAS scores produced by HYPE.

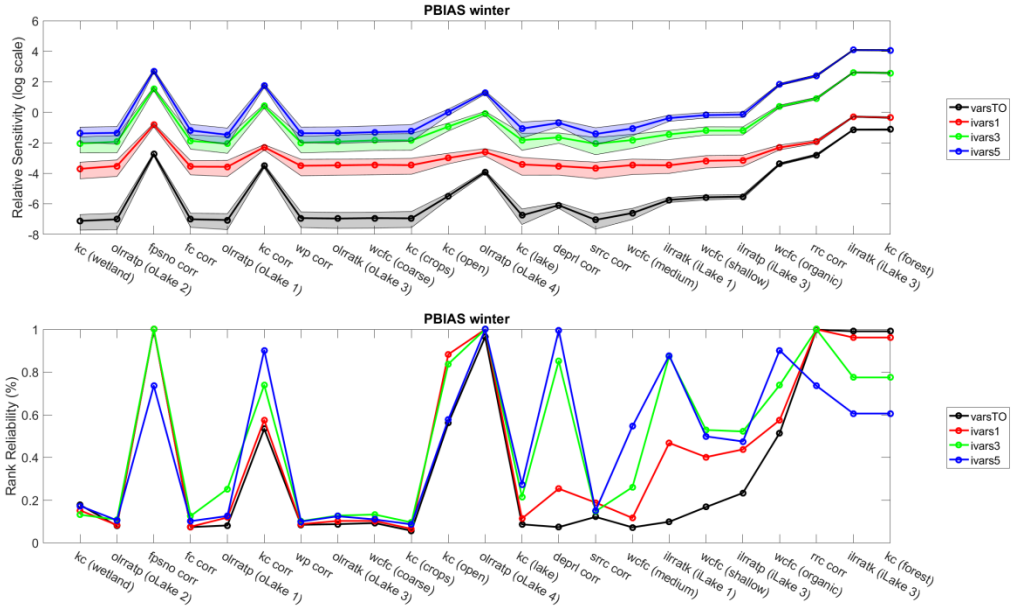

Figure B.10: VARS results for winter PBIAS scores produced by HYPE.

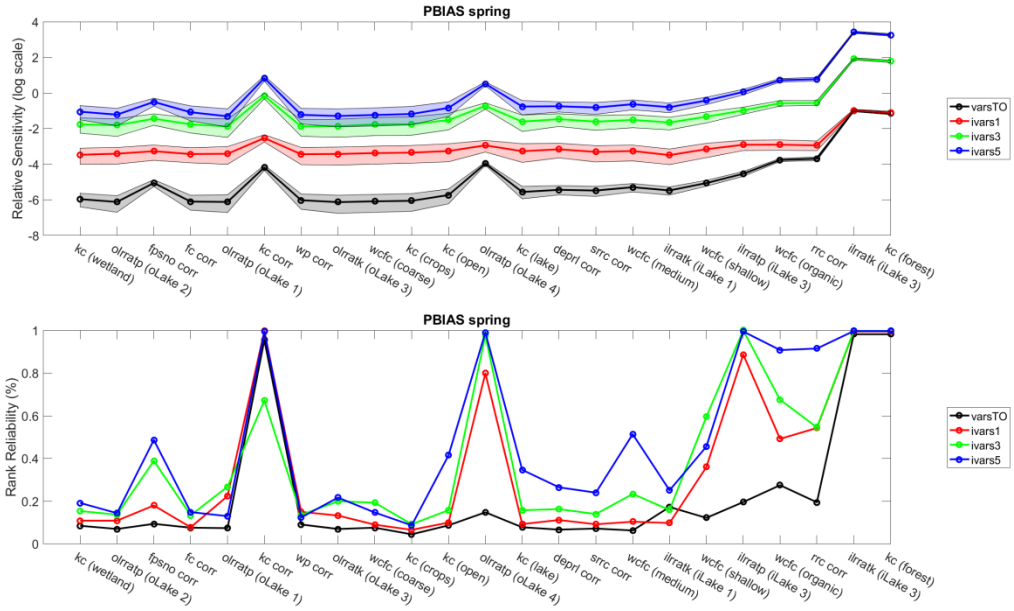

Figure B.11: VARS results for spring PBIAS scores produced by HYPE.

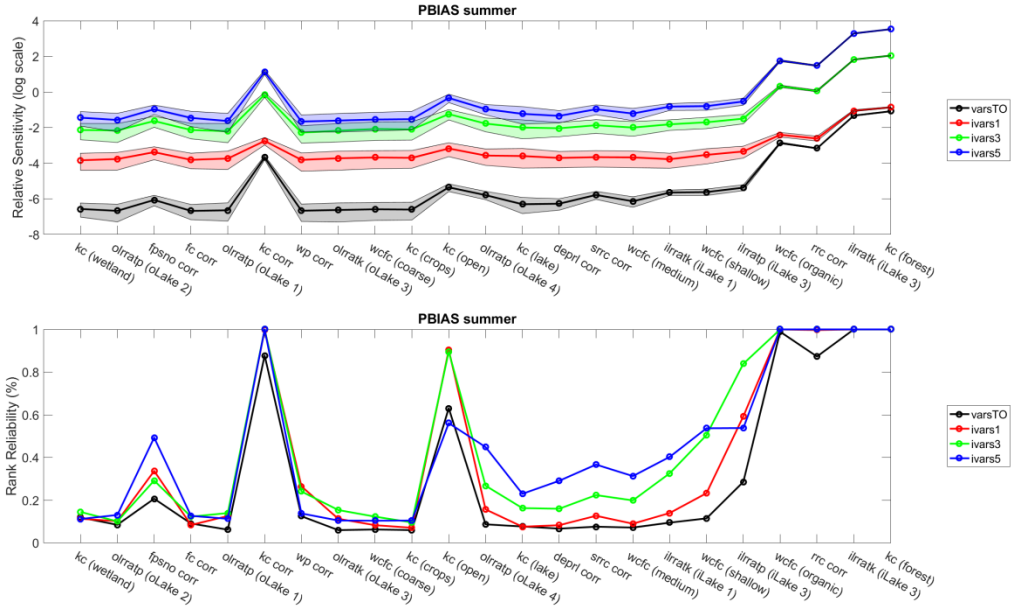

Figure B.12: VARS results for summer PBIAS scores produced by HYPE.

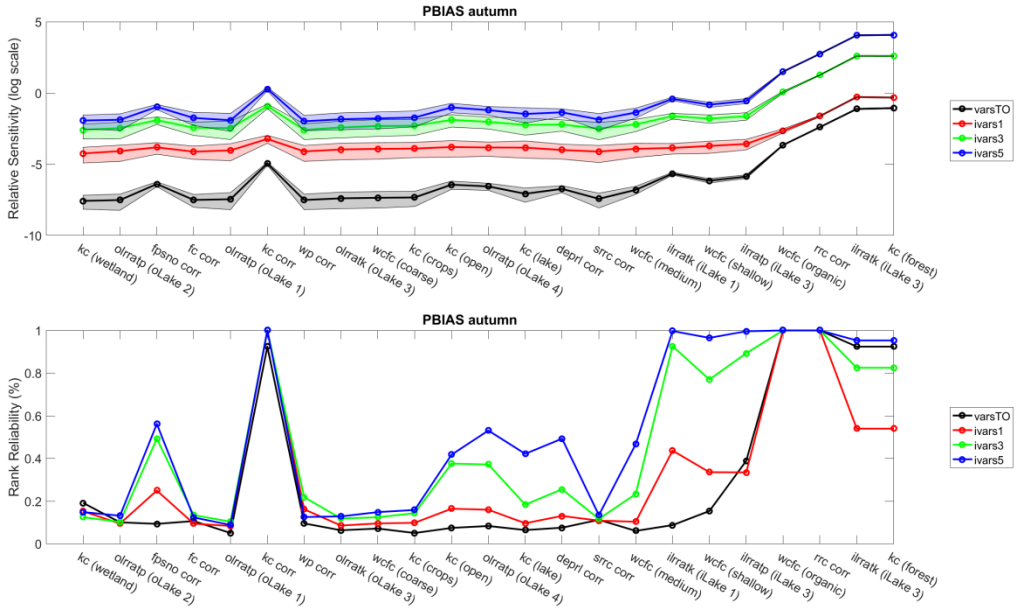

Figure B.13: VARS results for autumn PBIAS scores produced by HYPE.

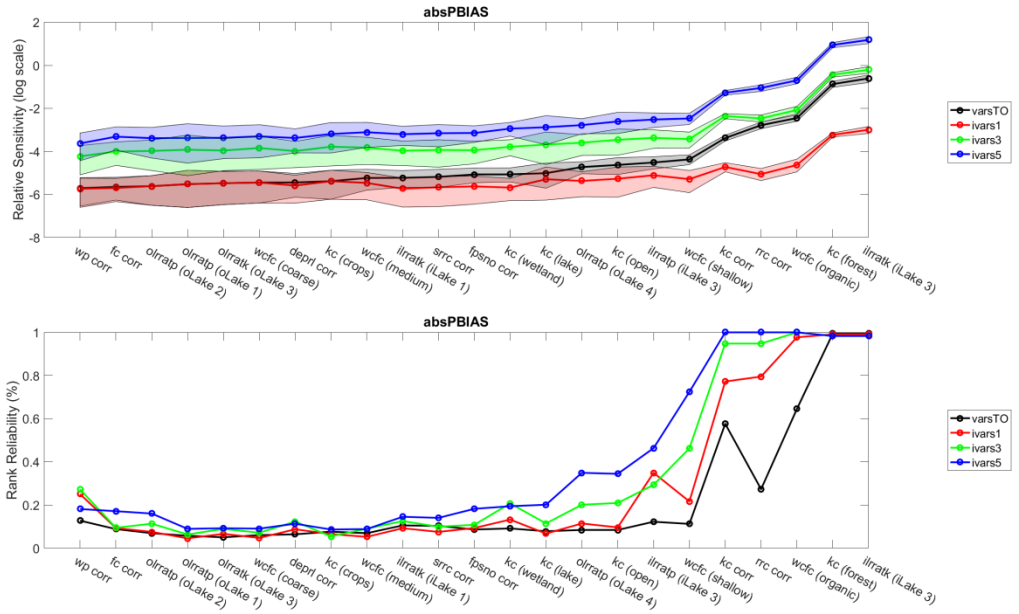

Figure B.14: VARS results for period absolute PBIAS scores produced by HYPE.

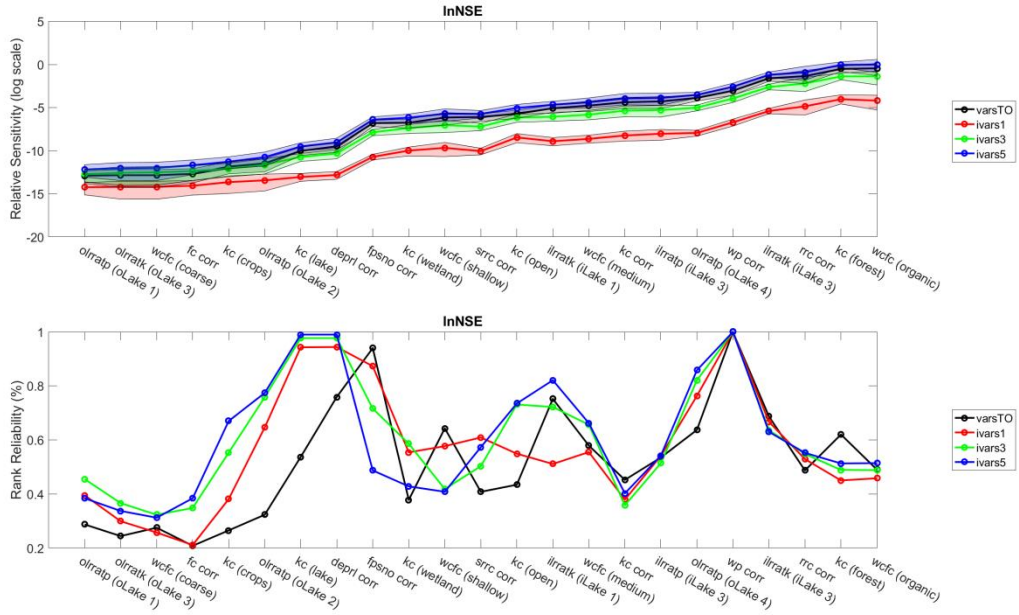

Figure B.15: VARS results for period log NSE scores produced by HYPE.

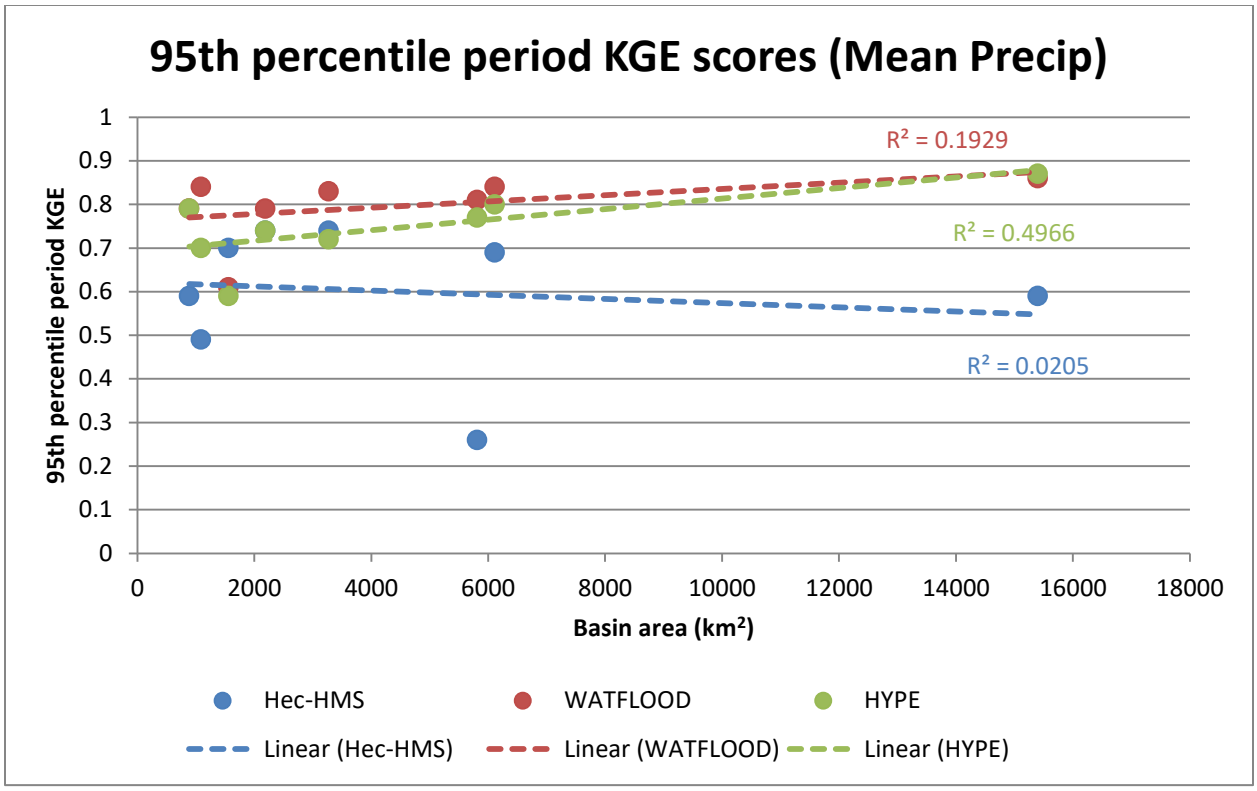

Figure B.16: Trend analysis for the 95th percentile period KGE scores for all models forced with the mean precipitation realization.

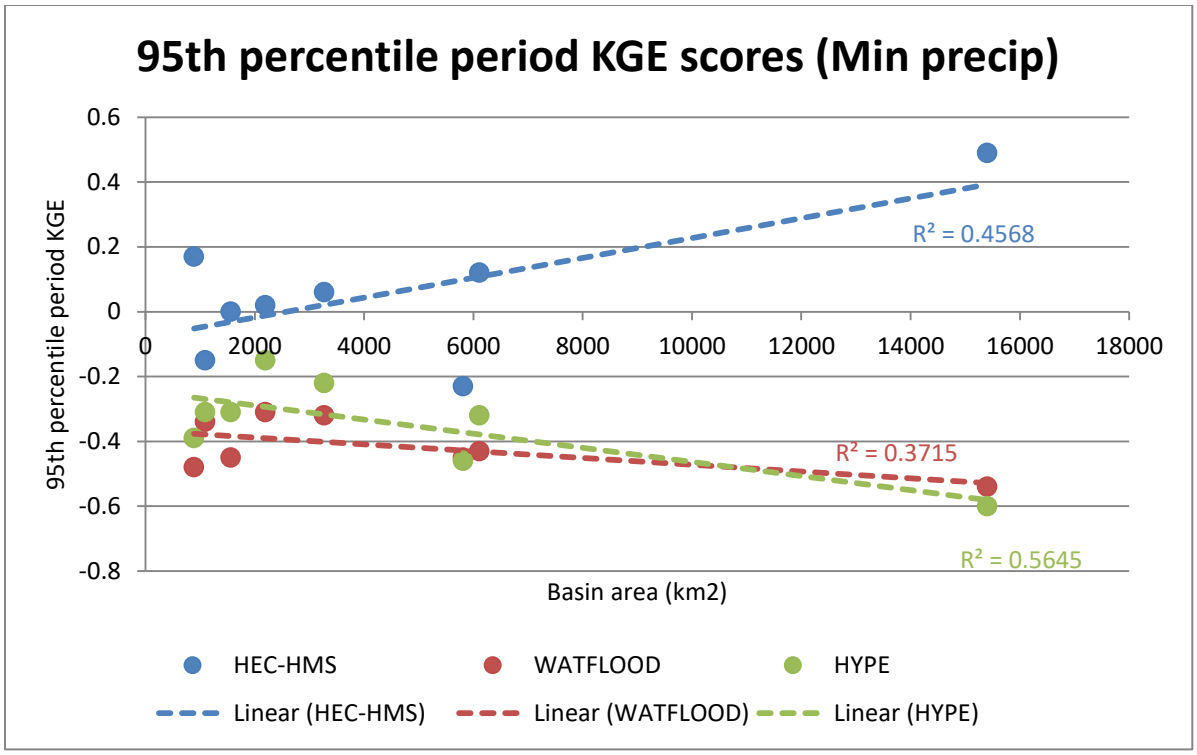

Figure B.17: Trend analysis for the 95th percentile period KGE scores for all models forced with the minimum precipitation realization.

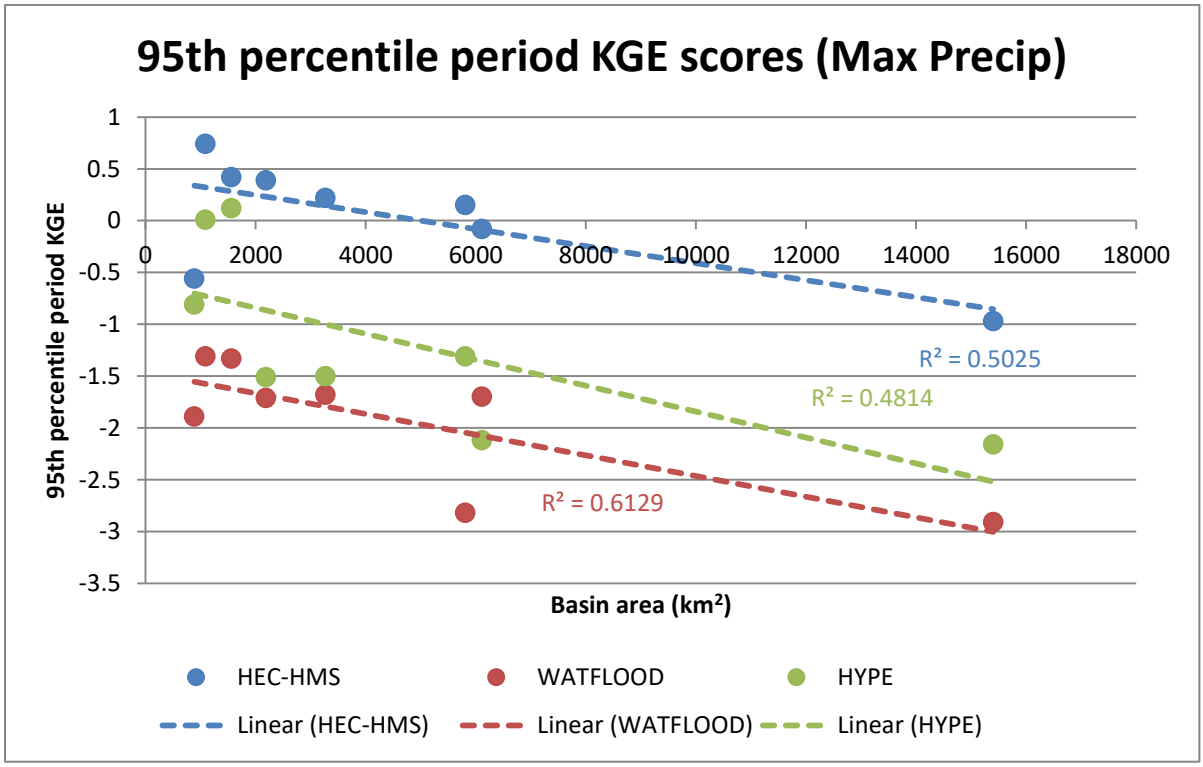

Figure B.18: Trend analysis for the 95th percentile period KGE scores for all models forced with the maximum precipitation realization.

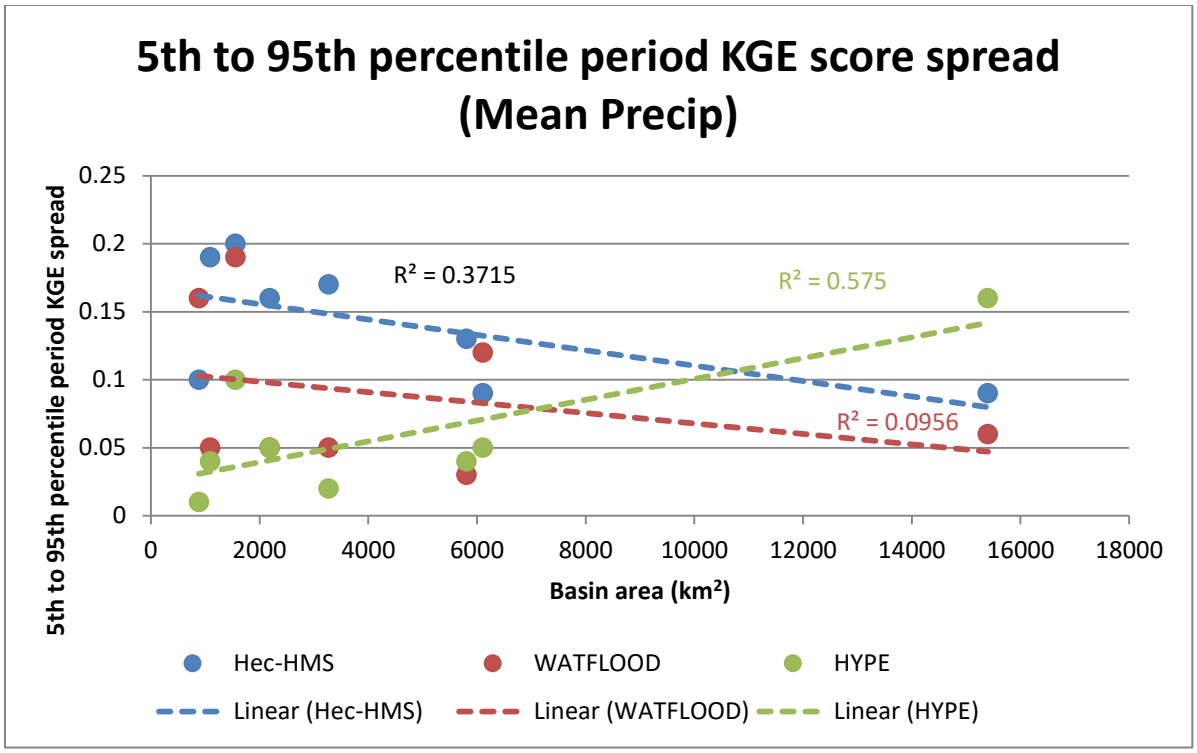

Figure B.19: Trend analysis for the 5-95th percentile period KGE score spread for all models forced with the mean precipitation realization.

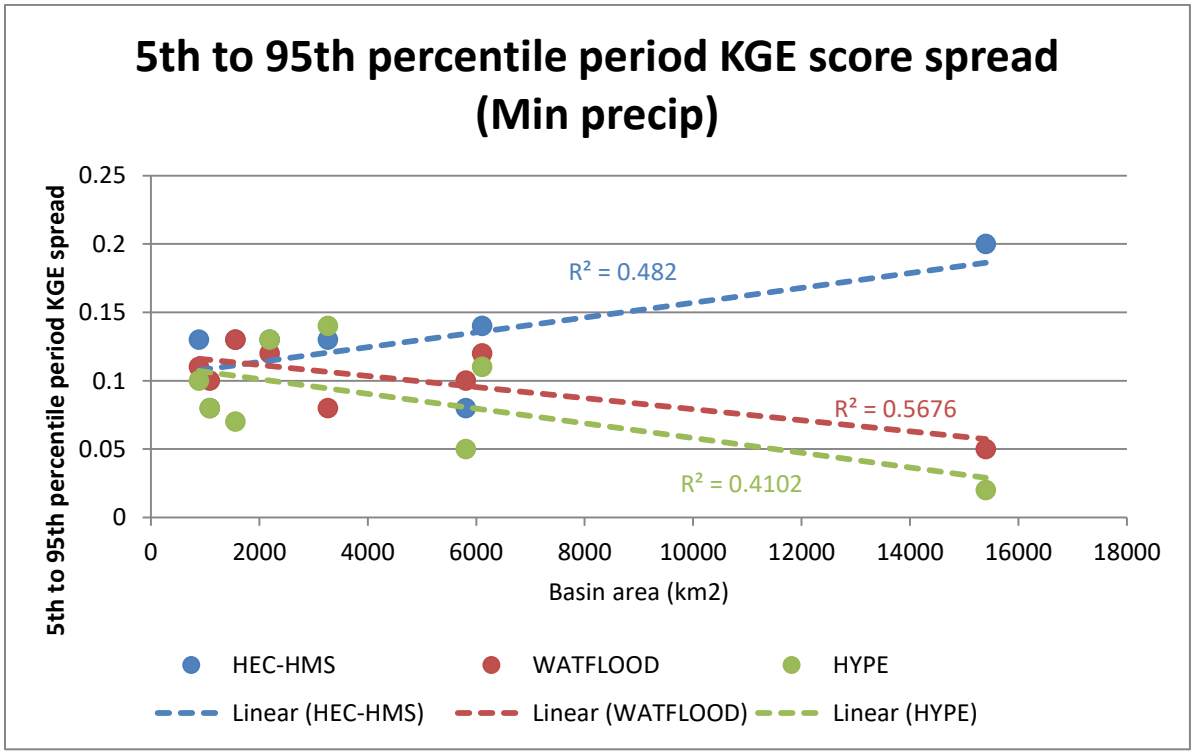

Figure B.20: Trend analysis for the 5-95th percentile period KGE score spread for all models forced with the minimum precipitation realization.

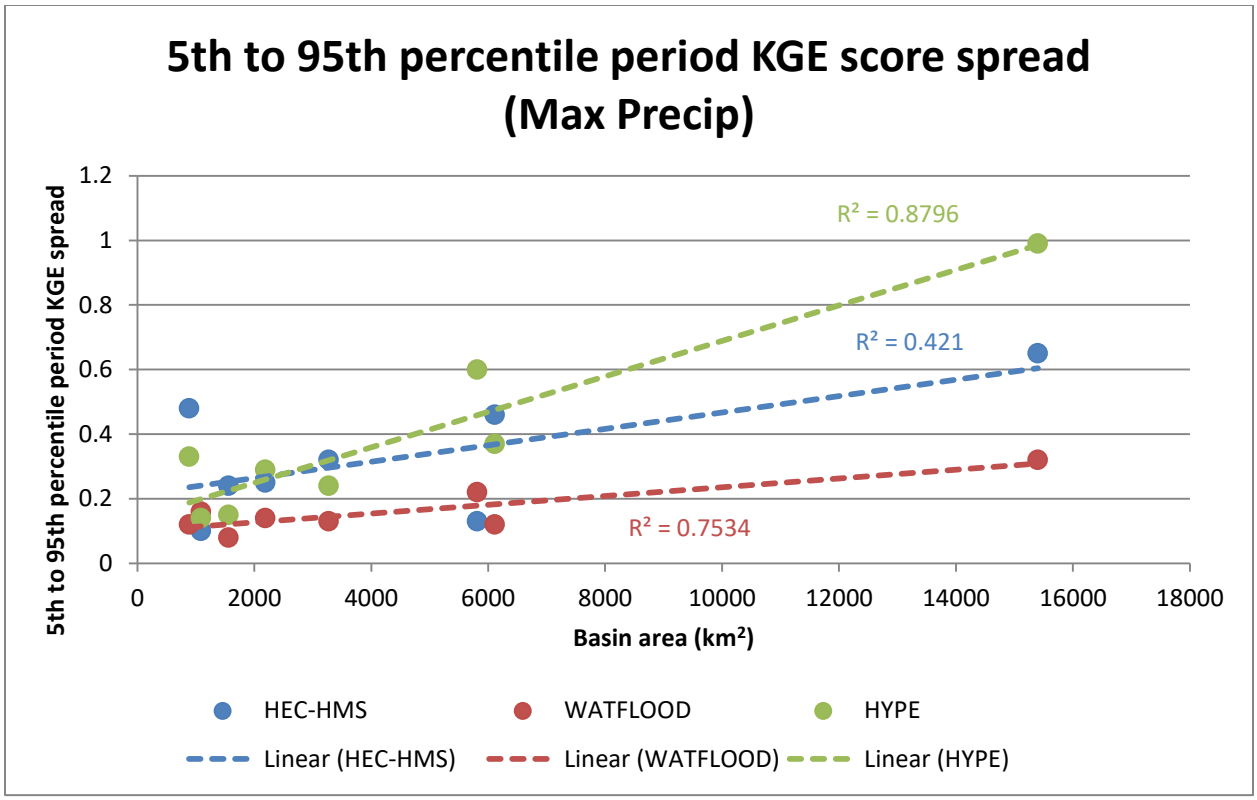

Figure B.21: Trend analysis for the 5-95th percentile period KGE score spread for all models forced with the maximum precipitation realization.

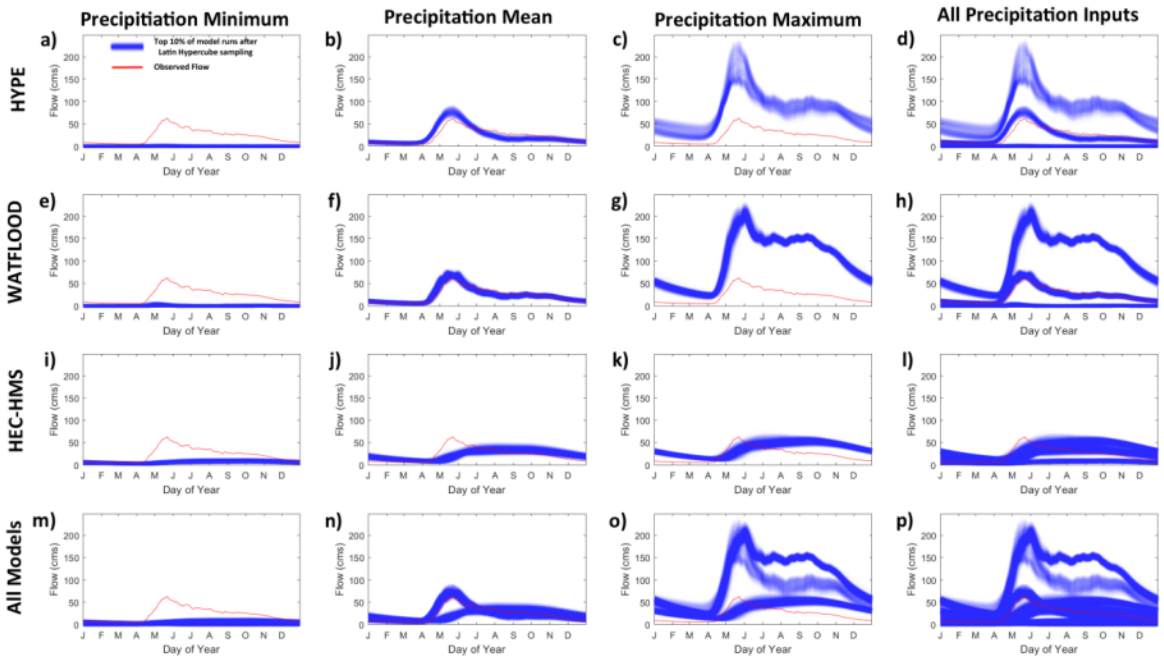

Figure B.22: 30-year average hydrographs for the Burntwood River above Leaf Rapids, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

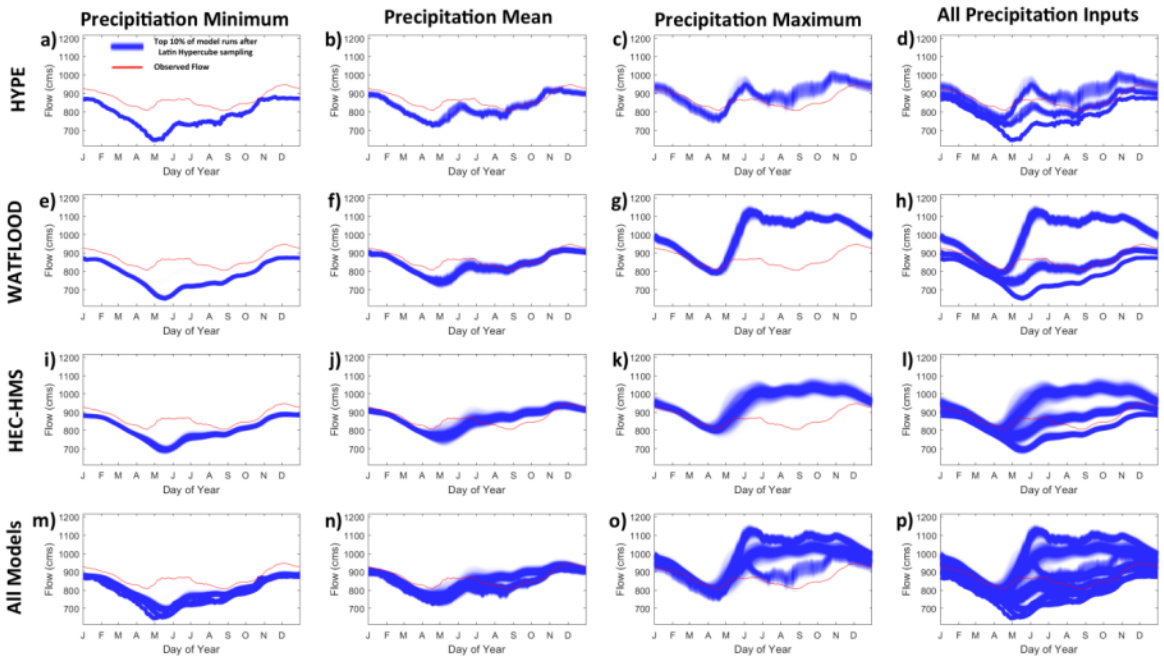

Figure B.23: 30-year average hydrographs for the Burntwood near Thompson, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

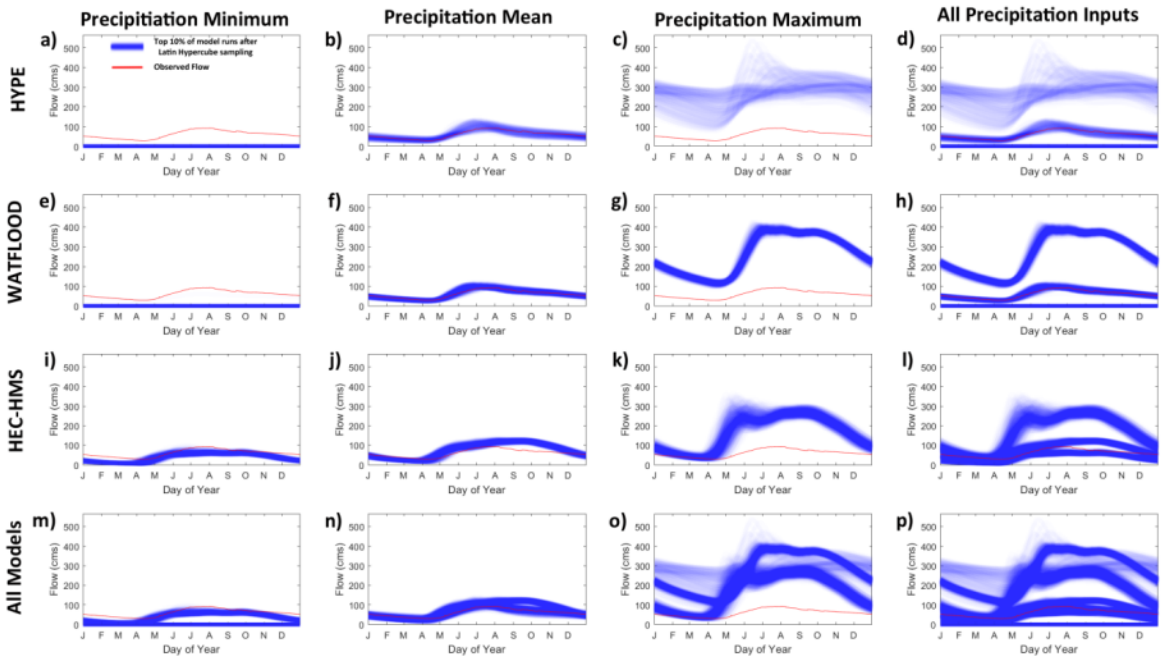

Figure B.24: 30-year average hydrographs for the Grass River, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

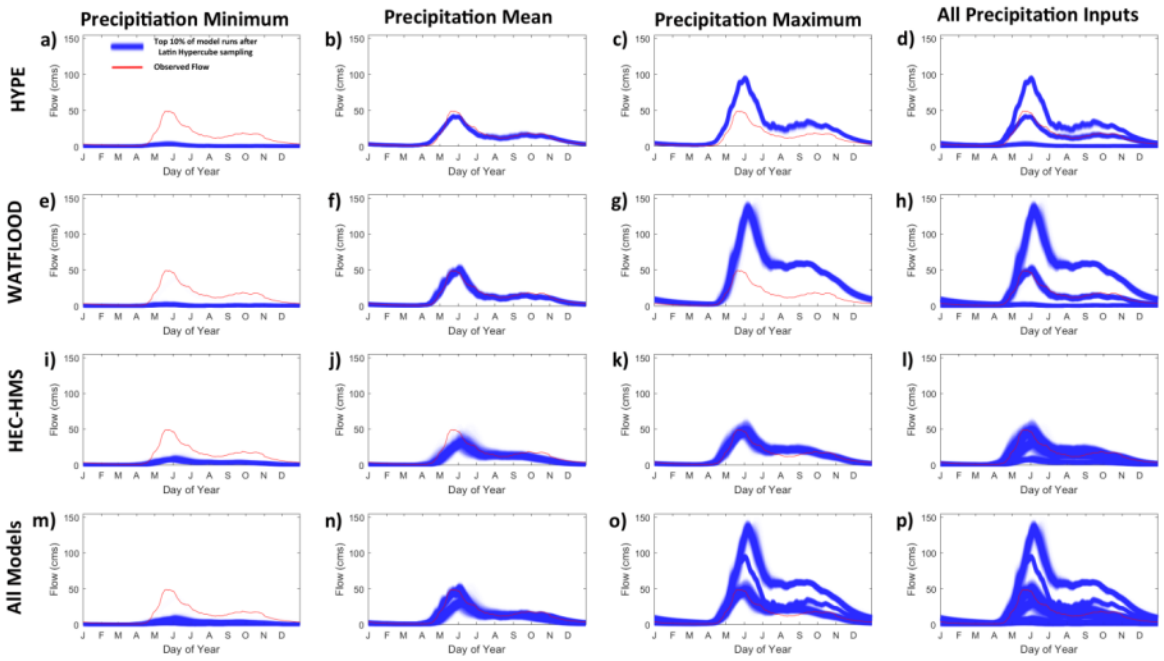

Figure B.25: 30-year average hydrographs for the Kettle River, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

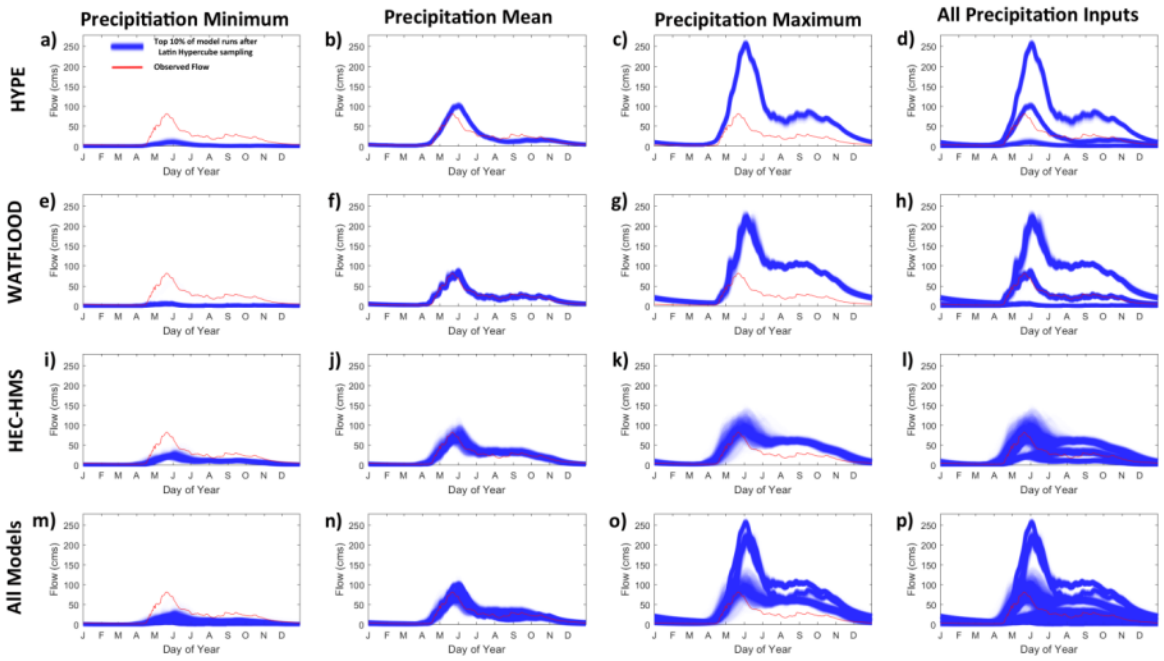

Figure B.26: 30-year average hydrographs for the Limestone River, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

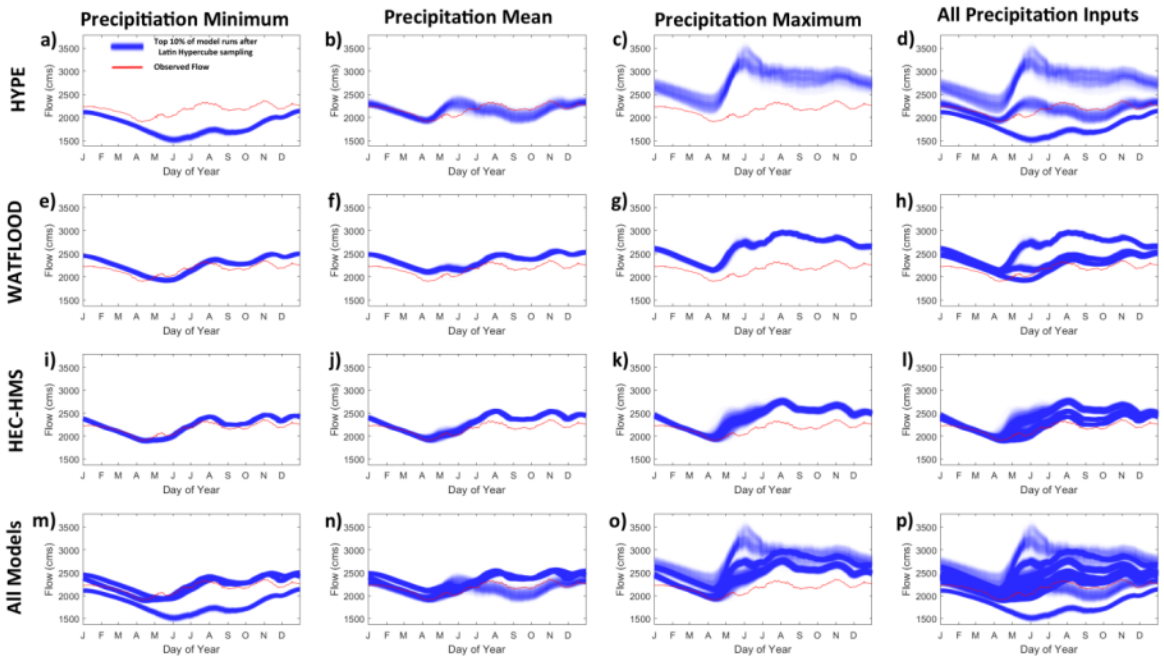

Figure B.27: 30-year average hydrographs for the Nelson River at Kelsey, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

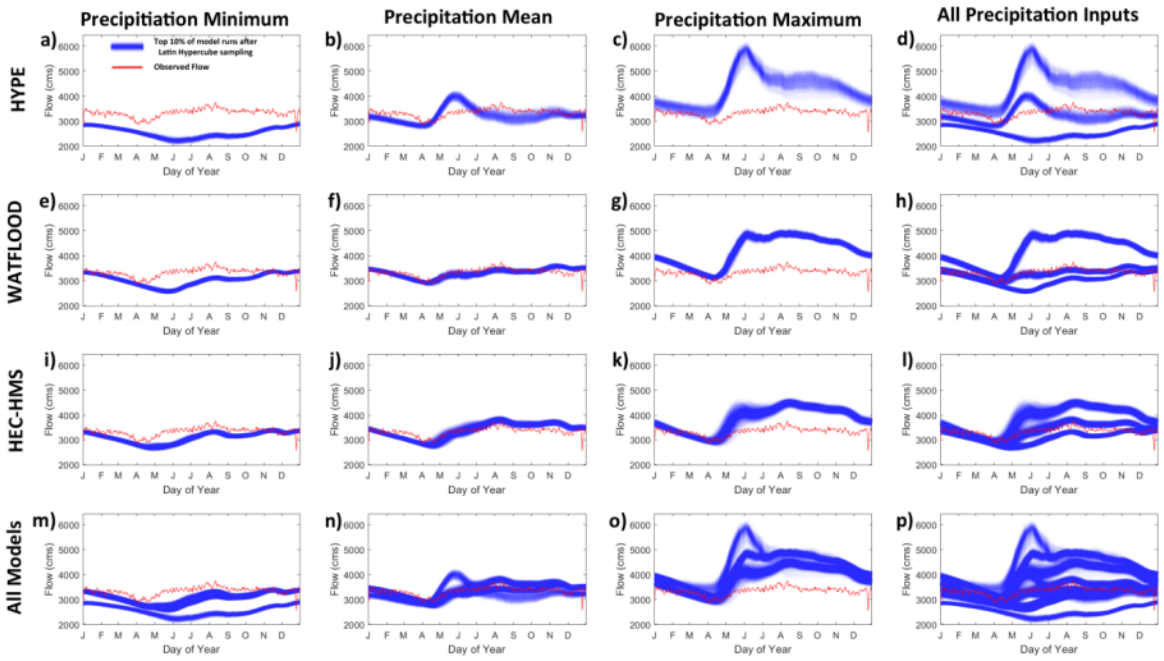

Figure B.28: 30-year average hydrographs for the Nelson River at Longspruce, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

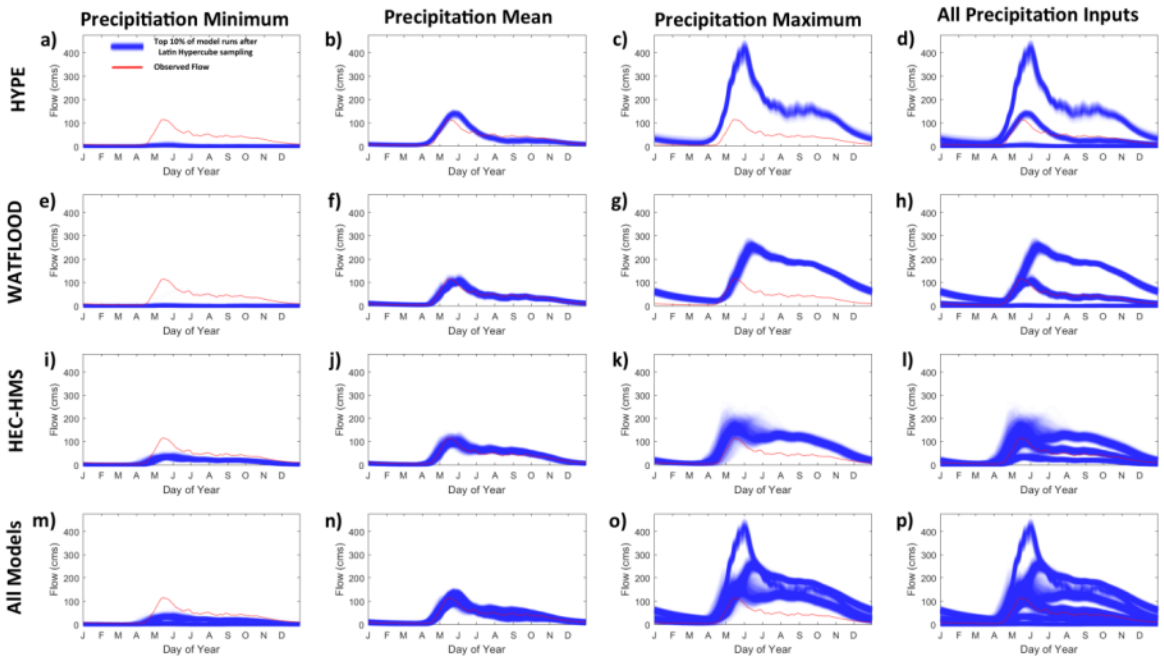

Figure B.29: 30-year average hydrographs for the Odei River, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

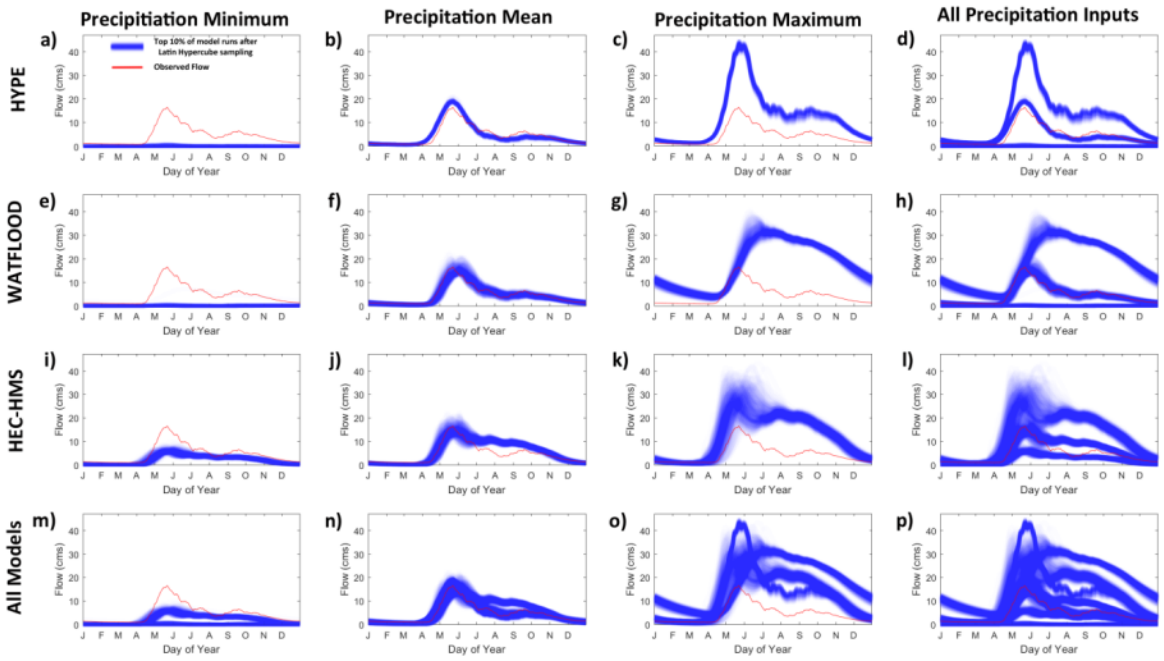

Figure B.30: 30-year average hydrographs for the Taylor River, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

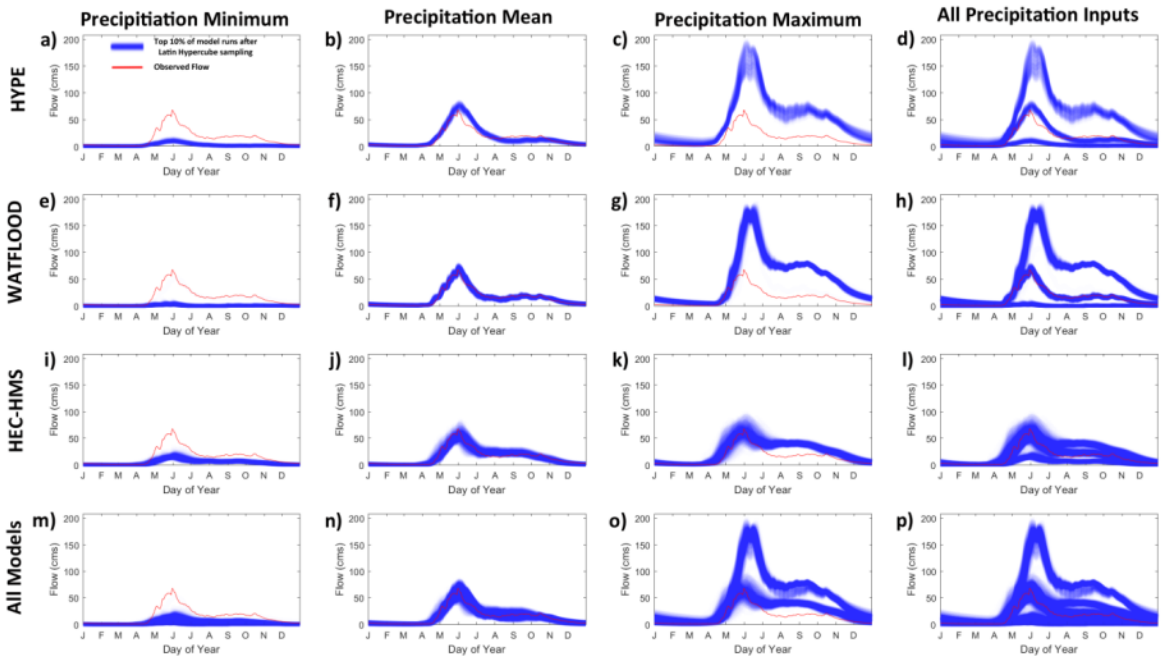

Figure B.31: 30-year average hydrographs for the Weir River, generated by selecting the top 10% of orthogonal Latin Hypercube sampled runs for each hydrologic model and precipitation realization. Simulated hydrographs are darker blue when there was higher density of simulated flows.

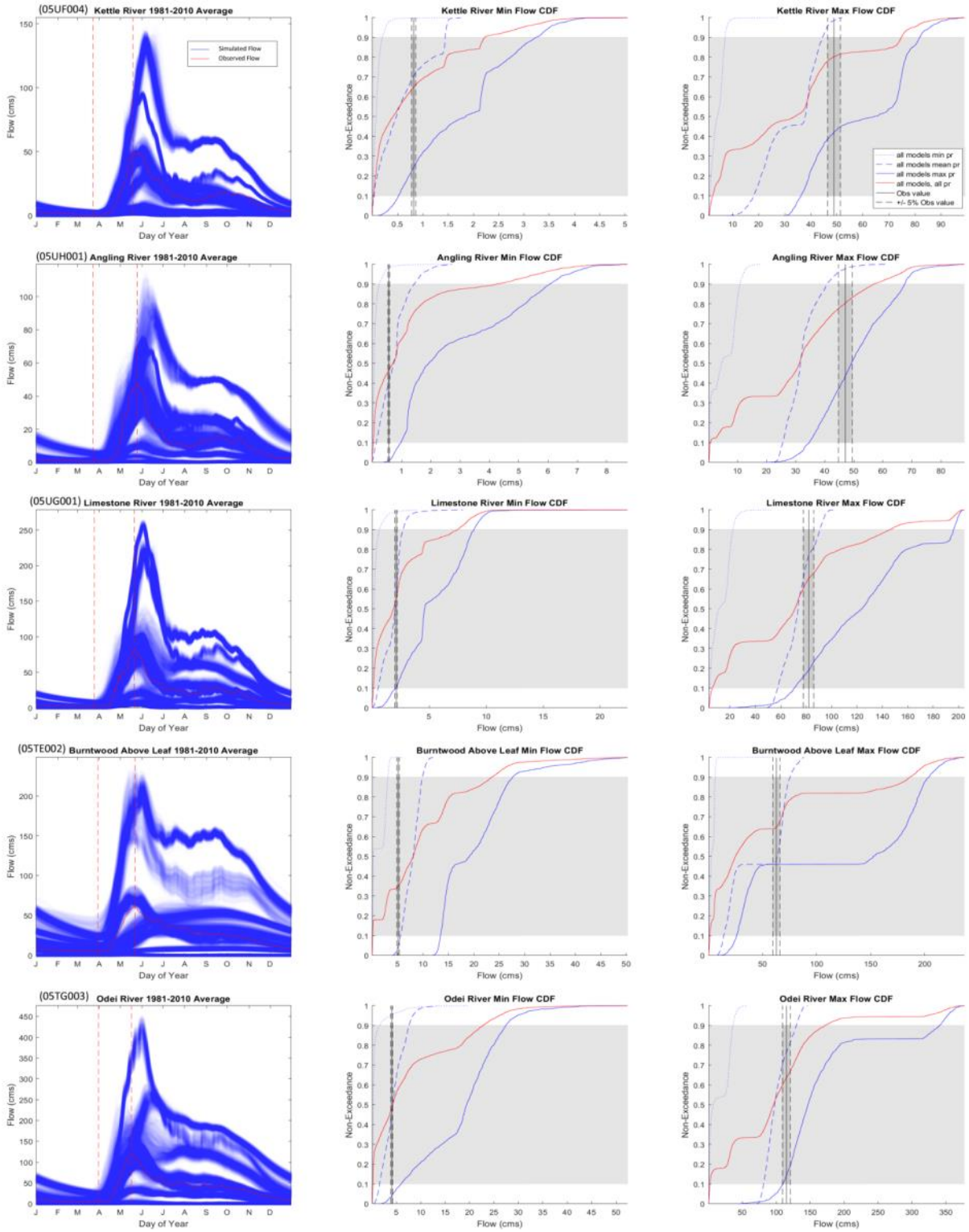

Figure B.32: Un-regulated 30-year average hydrographs for selected hydrometric gauges, with CDF plots for the minimum (Left CDF) and maximum (Right CDF) 30 year average flows. 30-year hydrographs include the top 10% of runs for all hydrologic models and precipitation.

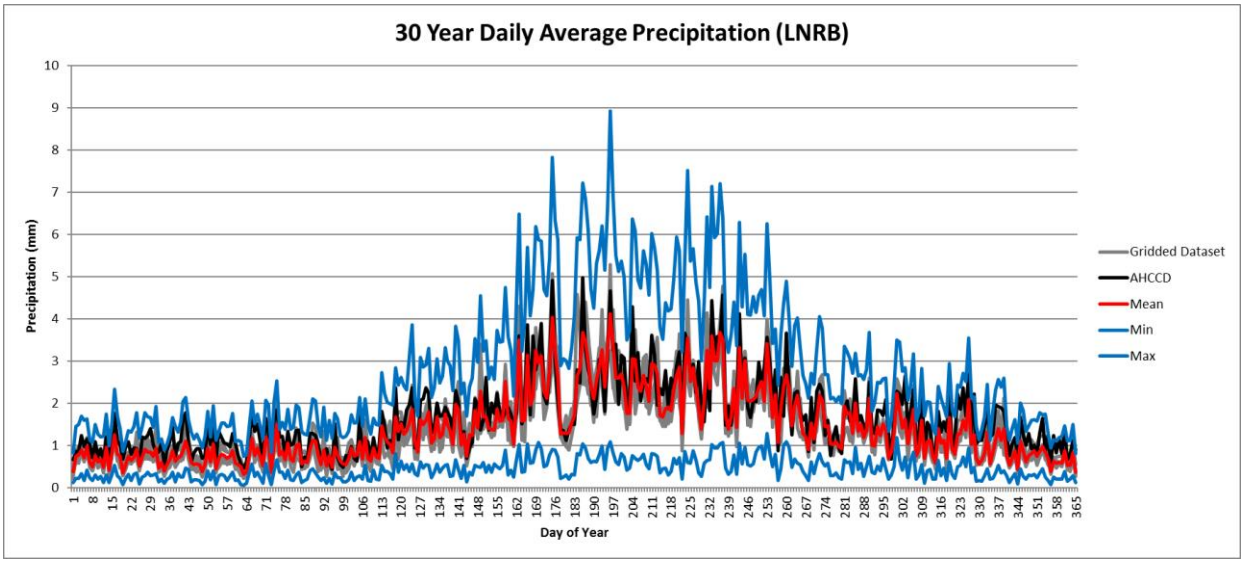

Figure B. 33: Period (1981-2010) daily precipitation averages for each of the five individual gridded data products, AHCCD, and the min, mean, and max ensemble realizations.

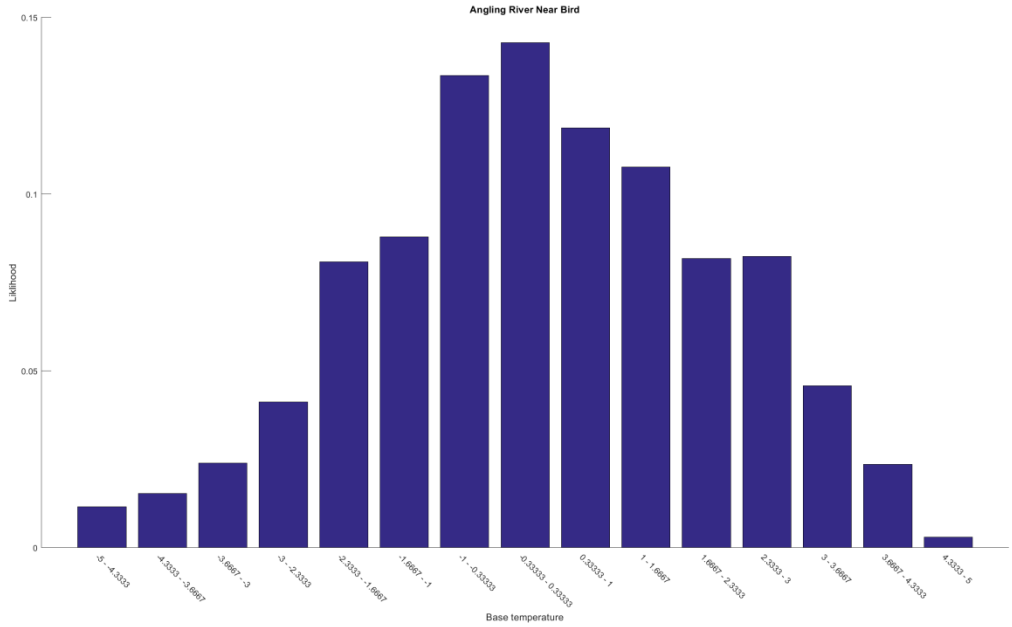

Figure B.34: HEC-HMS base temperature parameter likelihood distribution utilizing the mean precipitation realization.

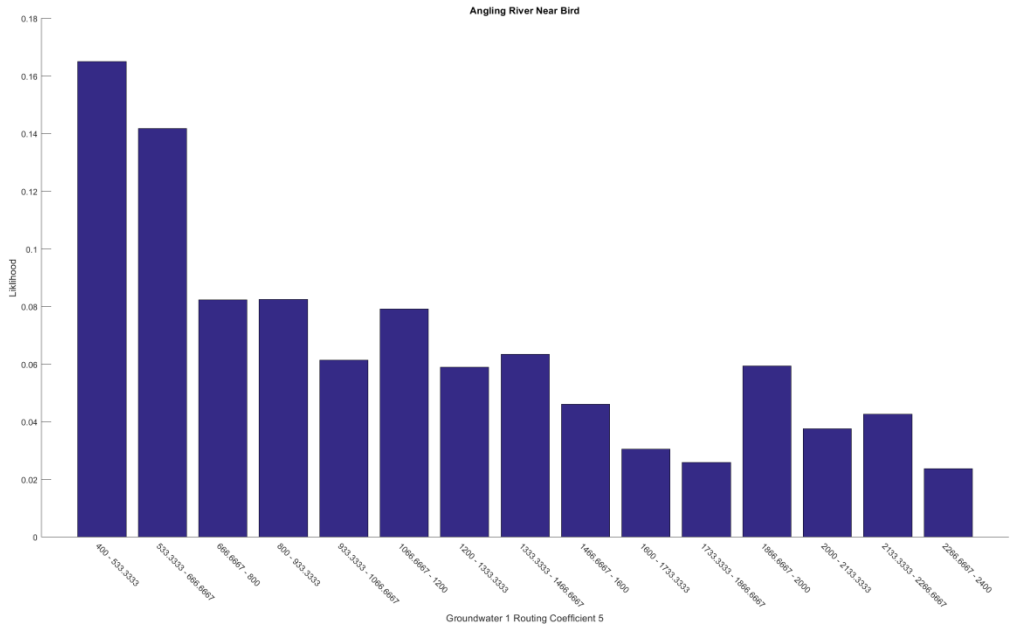

Figure B.35: HEC-HMS groundwater 1 routing (parameter group 5) parameter likelihood distribution utilizing the mean precipitation realization.

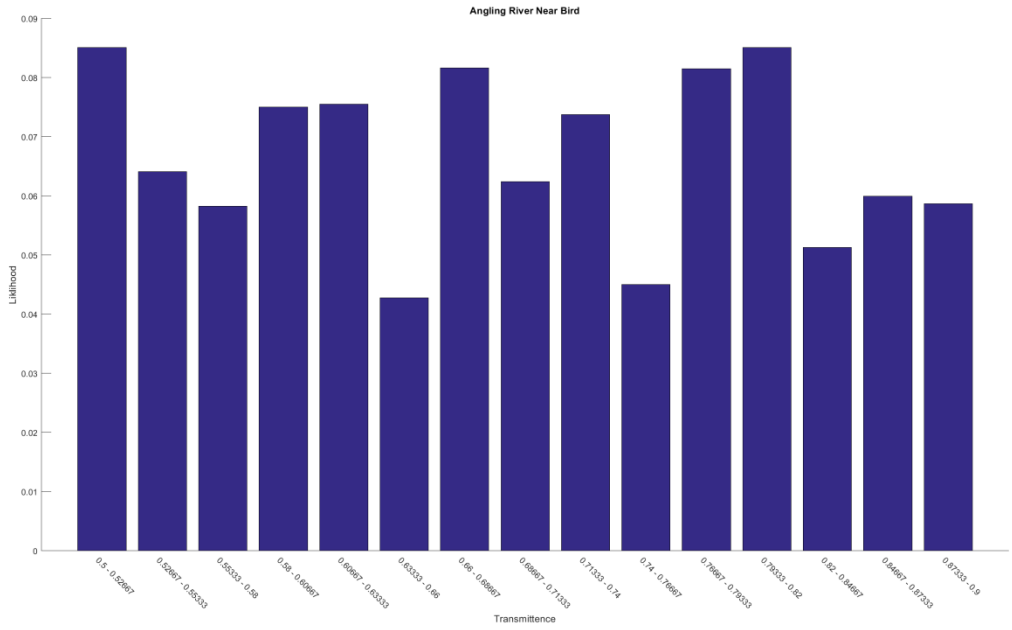

Figure B.36: HEC-HMS clear sky transmittance parameter likelihood distribution utilizing the mean precipitation realization.

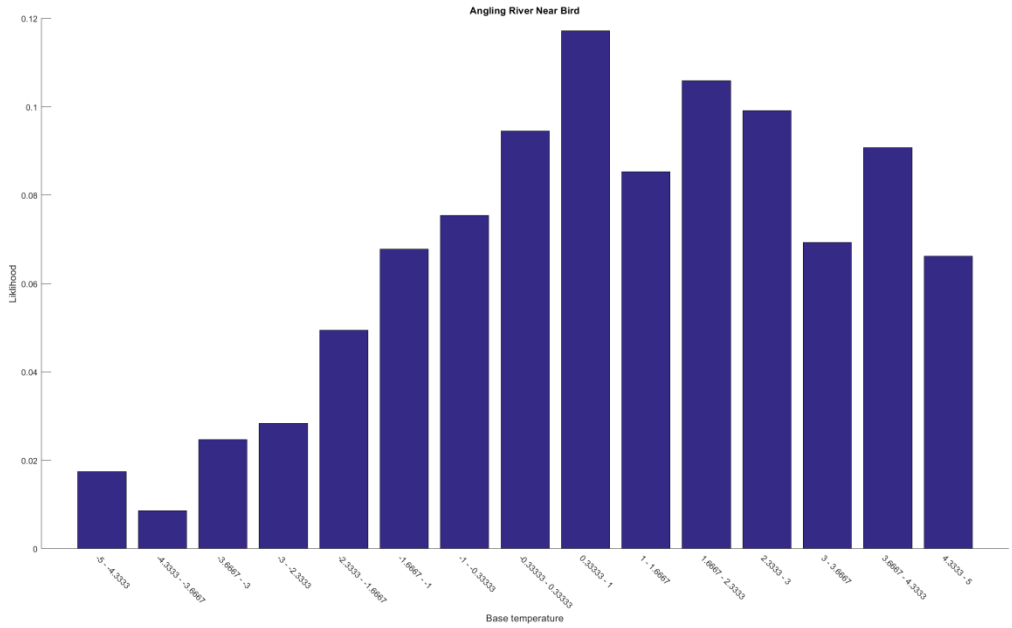

Figure B.37: HEC-HMS base temperature parameter likelihood distribution utilizing the minimum precipitation realization.

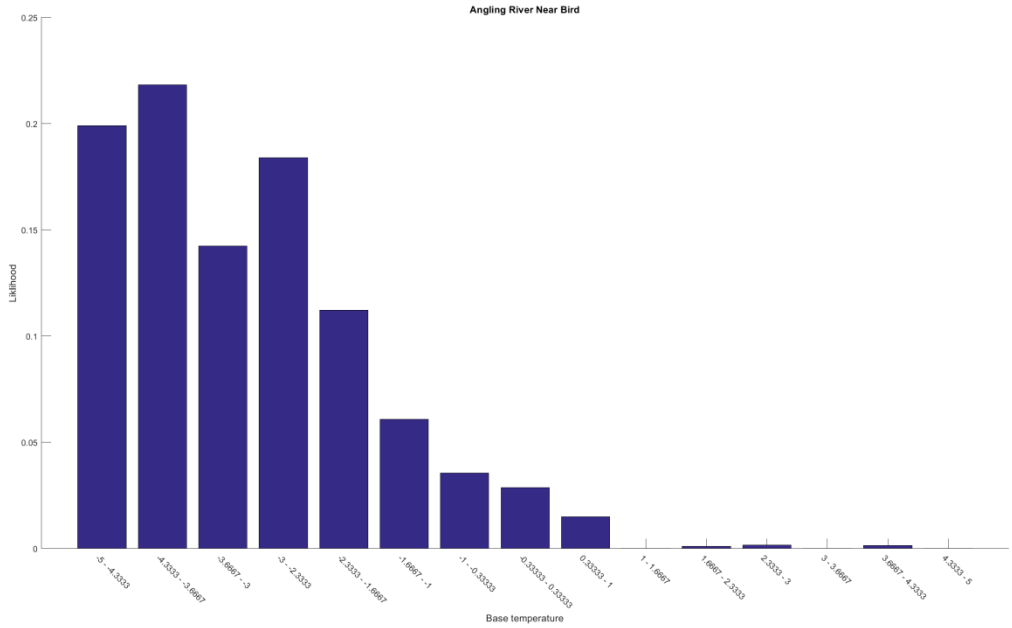

Figure B.38: HEC-HMS base temperature parameter likelihood distribution utilizing the maximum precipitation realization.