Improved Data Uncertainty Handling in Hydrologic Modeling and Forecasting Applications

by

Hongli Liu

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Civil Engineering (Water)

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Hamid Moradkhani Professor Department of Civil, Construction and Environmental Engineering University of Alabama
Supervisor(s)	Bryan Tolson Associate Professor Department of Civil and Environmental Engineering University of Waterloo
Internal Member	James Craig Associate Professor Department of Civil and Environmental Engineering University of Waterloo
Internal Member	Nandita Basu Associate Professor Department of Civil and Environmental Engineering Department of Earth and Environmental Sciences University of Waterloo
Internal-external Member	Walter Illman Professor Department of Earth and Environmental Sciences University of Waterloo

Author's Declaration

This thesis consists of materials all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

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

Statement of Contributions

Chapter 2 of this thesis consists of a published paper (Liu et al., 2016) that was co-authored by myself, my supervisor (Dr. Tolson), Dr. Craig and Dr. Shafii. In terms of contributions, Dr. Toslon, Dr. Craig, and I developed the discretization error metrics. Dr. Shafii and I developed the hydrologic model of the case study watershed. I implemented the method, analyzed results, and wrote the paper with the great assistance of Dr. Tolson and Dr. Craig. All authors contributed to revision of the paper.

Chapter 4 of this thesis consists of a published paper (Liu et al., 2019) that was co-authored by myself, Dr. Thiboult, my supervisor (Dr. Tolson), Dr. Anctil, and Dr. Mai. In terms of contributions, Dr. Toslon, Dr. Thiboult, and Dr. Anctil and I designed the comparison study and analyzed the study results. Dr. Thiboult and I implemented the ensemble Kalman filter based flow forecasting for 20 case studies. Dr. Mai transferred the shared data among coauthors and assisted me to set up the experiments on SHARCNET high performance computers. I wrote the paper with the great assistance of Dr. Thiboult. All authors contributed to revising the paper.

Abstract

In hydrologic modeling and forecasting applications, many steps are needed. The steps that are relevant to this thesis include watershed discretization, model calibration, and data assimilation. Watershed discretization separates a watershed into homogeneous computational units for depiction in a distributed hydrologic model. Objective identification of an appropriate discretization scheme remains challenging in part because of the lack of quantitative measures for assessing discretization quality, particularly prior to simulation. To solve this problem, this thesis contributes to develop an *a priori* discretization error metrics that can quantify the information loss induced by watershed discretization without running a hydrologic model. Informed by the error metrics, a two-step discretization decision-making approach is proposed with the advantages of reducing extreme errors and meeting user-specified discretization error targets.

In hydrologic model calibration, several uncertainty-based calibration frameworks have been developed to explicitly consider different hydrologic modeling errors, such as parameter errors, forcing and response data errors, and model structure errors. This thesis focuses on climate and flow data errors. The common way of handling climate and flow data uncertainty in the existing calibration studies is perturbing observations with assumed statistical error models (e.g., addictive or multiplicative Gaussian error model) and incorporating them into parameter estimation by integration or repetition with multiple climate and (or) flow realizations. Given the existence of advanced climate and flow data uncertainty estimation methods, this thesis proposes replacing assumed statistical error models with physically-based (and more realistic and convenient) climate and flow ensembles. Accordingly, this thesis contributes developing a climate-flow ensemble based hydrologic model calibration framework. The framework is developed through two stages. The first stage only considers climate data uncertainty, leading to the climate ensemble based hydrologic calibration framework. The framework is parsimonious and can utilize any sources of historical climate ensembles. This thesis demonstrates the method of using the Gridded Ensemble Precipitation and Temperature Estimates dataset (Newman et al., 2015), referred to as N15 here, to derive precipitation and temperature ensembles. Assessment of this framework is conducted using 30 synthetic experiments and 20 real case studies. Results show that the framework generates more robust parameter estimates, reduces the inaccuracy of flow predictions caused by poor quality climate data, and improves the reliability of flow predictions.

The second stage adds flow ensemble to the previously developed framework to explicitly consider flow data uncertainty and thus completes the climate-flow ensemble based calibration framework. The complete framework can work with likelihood-free calibration methods. This thesis demonstrates the method of using the hydraulics-based Bayesian rating curve uncertainty estimation method (BaRatin) (Le Coz et al., 2014) to generate flow ensemble. The continuous ranked probability score (CRPS) is taken as

an objective function of the framework to compare the scalar model prediction with the measured flow ensemble. The framework performance is assessed based on 10 case studies. Results show that explicit consideration of flow data uncertainty maintains the accuracy and slightly improves the reliability of flow predictions, but compared with climate data uncertainty, flow data uncertainty plays a minor role of improving flow predictions.

Regarding streamflow forecasting applications, this thesis contributes by improving the treatment of measured climate data uncertainty in the ensemble Kalman filter (EnKF) data assimilation. Similar as in model calibration, past studies usually use assumed statistical error models to perturb climate data in the EnKF. In data assimilation, the hyper-parameters of the statistical error models are often estimated by a trial-and-error tuning process, requiring significant analyst and computational time. To improve the efficiency of climate data uncertainty estimation in the EnKF, this thesis proposes the direct use of existing climate ensemble products to derive climate ensembles. The N15 dataset is used here to generate 100-member precipitation and temperature ensembles. The N15 generated climate ensembles are compared with the carefully tuned hyper-parameter generated climate ensembles in ensemble flow forecasting over 20 catchments. Results show that the N15 generated climate ensemble yields improved or similar flow forecasts than hyper-parameter generated climate ensembles. Therefore, it is possible to eliminate the time-consuming climate relevant hyper-parameter tuning from the EnKF by using existing ensemble climate products without losing flow forecast performance.

After finishing the above research, a robust hydrologic modeling approach is built by using the thesis developed model calibration and data assimilation methods. The last contribution of this thesis is validating such a robust hydrologic model in ensemble flow forecasting via comparison with the use of traditional multiple hydrologic models. The robust single-model forecasting system considers parameter and climate data uncertainty and uses the N15 dataset to perturb historical climate in the EnKF. In contrast, the traditional multi-model forecasting system does not consider parameter and climate data uncertainty and uses assumed statistical error models to perturb historical climate in the EnKF. The comparison study is conducted on 20 catchments and reveal that the robust single hydrologic model generates improved ensemble high flow forecasts. Therefore, robust single model is definitely an advantage for ensemble high flow forecasts. The robust single hydrologic model relieves modelers from developing multiple (and often distributed) hydrologic models for each watershed in their operational ensemble prediction system.

Acknowledgements

I would like to thank Dr. Bryan Tolson, my PhD supervisor, for his academic guidance throughout my PhD studies. It is my fortunate to be a student of such a knowledgeable, supportive, and responsible supervisor. Also, I have always been grateful to Bryan for his support and advice on things out of research, especially for raising me up throughout tough times and providing heartfelt advices on decisions.

I would like to thank my thesis committee guiding me through all these years. Thank you, Dr. Hamid Moradkhani from University of Alabama, and Drs. James Craig, Walter Illman, and Nandita Basu from University of Waterloo. I would also thank the great support and collaboration of Drs. François Anctil, Antoine Thiboult, and Mahyar Shafi for case study sharing, Dr. Juliane Mai for experimental deployment on supercomputers, Dr. Andrew Newman for his help in utilizing the ensemble climate dataset, Drs. Renard Benjamin and Jérôme Le Coz for their help in applying the BaRatin method, and Dr. Ming Han for his help in using the bankful widths and depths database. I would also thank the guidance of Drs. Andrew Newman, Andrew Wood, and Martyn Clark on my research during my visit to the National Center for Atmospheric Research (NCAR).

Part of the thesis contents were peer-reviewed in the form of technical journal papers before thesis submission. I would like to thank the editors and anonymous reviewers who reviewed the papers associated with Chapter 2 and Chapter 4. Their constructive comments helped improve my work.

During my PhD study, I was supported by the NSERC FloodNet project, Collaborative Water program RBC Water Scholars Graduate Entrance Scholarship, University of Waterloo International Doctoral Student Awards, University of Waterloo Graduate Merit Scholarship, and Dr. T.E. Unny Memorial Award which are hereby acknowledged. I am also very thankful to my friends and colleagues who offer me lasting support as real friends.

I would like to express my sincere gratefulness to my parents Xuepei and Zhen and my brother Hexin for providing me with love and encouragement, without whom I would never complete my PhD studies so smoothly.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Hydrologic Modeling and Forecasting Overview

A hydrologic model is a simplified representation of the natural system (i.e., the hydrologic cycle or parts of it) (Refsgaard, 1990). It can be expressed by a set of mathematical descriptions converting natural climate forcings (e.g., rainfall) into a system response (e.g., runoff). In these mathematical expressions, variables are quantities related to measurements and can vary in space and time, such as the series of inputs to and outputs from the model and the state variables/conditions of some part of the model (e.g., snow water equivalent), whereas model parameters are coefficients/constants describing a theoretical relationship or variables hard to measure (Altman and Bland, 1999). Hydrologic models are mainly used for two purposes. The first is to help human beings to understand the natural hydrologic processes and explore the implications of certain assumptions behind the nature hydrologic processes while the second is to predict/ forecast the system responses to a designed event/ an anticipated event (Beven, 1989).

To build a hydrologic model, many steps are required as summarized by Refsgaard (1996). The steps related to this thesis include model construction, performance criteria definition, calibration, and validation. In model construction, watershed discretization is an important component designed to separate a watershed into homogeneous computational units in order to represent hydrologic processes (Singh and Frevert, 2005); process representation decides which physical processes need to be represented and which mathematical approaches are used to represent the hydrologic processes (Clark et al., 2015); in addition, model construction involves setting boundary and initial conditions and parameterization. The next step is to define the performance criteria that need to be used in the subsequent calibration and validation. In this step, the indicator of model performance needs to be firstly defined to assess the agreement of model outputs with measured system responses (Klemeš, 1986). Then the performance criteria are identified considering the desired modeling accuracy as well as realistic limits (e.g., data availability and computational capacity) (Refsgaard, 1990). Model calibration is the process of adjusting model parameters so that model outputs sufficiently meet the specified performance criteria (Duan et al., 2003). Model validation is applying the calibrated model to a different time period and/or region without changing the parameter values (e.g., see Klemeš, 1986). The model can be said to be "validated" if the accuracy in the validation period meets the specified performance criteria. Note that the validation in this thesis is actually conditional validation because it is dependent on certain performance criteria and validation data set (Beven and Young, 2013; Young, 2001).

Once built, a hydrologic model can be run and used in two modes: simulation and forecasting. Being consistent with the terminology guidance in Beven and Young (2013), this thesis defines simulation as the quantitative reproduction of the behavior of the natural system with given inputs to the given time but without direct reference to any observed responses. Without reference implies that observations are not

used during the simulation to, for example, update model state variables. Running a hydrologic model in simulation mode, called hydrologic simulation, is useful in building up a hydrologic model (e.g., calibration and validation), understanding hydrologic processes, and conducting scenario/impact analysis. Scenario analysis means simulating system responses under alternative possible future conditions, for example, land use changes, design storms, and climate changes. Furthermore, forecasting is defined as the quantitative reproduction of the behavior of the natural system ahead of time driven by forecasted climate forcings and with reference to observed states (where applicable) and responses up to the current time (e.g., the time the forecast is made). Forecasting applications typically develop and test a forecasting system on past data (past forecasts of climate and past observations of system response).

Given the nature of hydrologic systems and the uncertainties in the modeling and application processes, hydrologic models should be applied with proper considerations of relevant uncertainties (Beven and Young, 2013; Liu and Gupta, 2007). A source of uncertainty can be characterized as either aleatory or epistemic depending on the problem context (Kiureghian and Ditlevsen, 2009). Uncertainties are characterized as epistemic if the modeler sees a possibility to reduce them by gathering more data or by refining models (e.g., lack of knowledge or data). Uncertainties are categorized as aleatory if the modeler does not foresee the possibility of reducing them (e.g., intrinsic randomness of a phenomenon). The advantage of separating the uncertainties into aleatory and epistemic is that we thereby make clear which uncertainties can be reduced and which uncertainties are less prone to reduction, at least in the near future. This categorization helps in determining the allocation of resources while developing models (Kiureghian and Ditlevsen, 2009).

Table 1-1 provides a summary of model building stages and model applications investigated in this thesis and the corresponding uncertainty sources in each stage/application. Note that for each uncertainty source studied in this thesis, the source of uncertainty is only deemed epistemic if, in the context of the thesis, the investigation involves attempting to reduce the degree of uncertainty of the corresponding source. The remainder of the chapter is organized around Table 1-1. Each modeling stage or model application in the table is first expanded upon with a high-level description and brief literature overview in the remainder of this chapter. Note that since Chapter 2 through Chapter 6 of the thesis are organized as a series of papers, each contains a more detailed review of the specific relevant literature.

Table 1-1. Summary of model building stages and model applications investigated in the thesis. Uncertainty sources corresponding to each stage/application are listed and any sources considered uncertain in the thesis are *italicized*. Note that NA under uncertainty type implies that source of uncertainty is not addressed in this thesis.

Modeling stage/application		**Uncertainty source**	**Uncertainty type**
Model structure identification	Watershed discretization	• Spatial data measurement/classification	NA
		• *Information loss due to discretization*	Epistemic
	Process representation	• Model complexity decision	NA
		• Numerical approximation	NA
Model parameter estimation	Model calibration (including model validation)	• *Model parameter*	Epistemic
		• *Forcing data*	Epistemic
		• *Response data*	Aleatory
		• Model structure	NA
		• *Initial and boundary condition*	Aleatory
Hydrologic Forecasting	Streamflow forecasting (including data assimilation)	• *Model parameter*	Epistemic
		• *Forcing data*	Aleatory
		• *Response data*	Aleatory
		• Model structure	NA
		• *Initial and boundary condition*	Aleatory
		• *Forecast forcing data*	Aleatory
		• Operation	NA

1.1.1. Model Structure Identification

Watershed discretization

A watershed, also known as a basin or a catchment, is a region of physical space where precipitation collects and drains off into a common outlet. Depending on the treatment of the spatial variations in variables and parameters, a watershed can be regarded as one unit or more than one unit. When the watershed is treated as only one unit in a hydrologic model, the variables and parameters represent the average values for the entire watershed, and the model is lumped. Conversely, when the watershed is treated as more than one unit in a hydrologic model, the variables and parameters represent spatial heterogeneity of hydrologic properties or responses across the watershed, and the model is distributed.

Watershed discretization is a process of separating a watershed into a number of effectively homogeneous computational units to characterize the spatial heterogeneity of hydrologic properties. Therefore, watershed discretization is only needed for distributed hydrologic modeling. The shape of the resultant units can be regular (e.g., grid squares) or irregular (hydrologic response units (HRUs)) depending on the modeler's requirement. In hydrologic modeling, watershed discretization commonly relies on the factors that determine rainfall-runoff responses, essentially topographic-based characteristics, such as elevation, land use, and soil. Therefore, these factors corresponding spatial data are needed in watershed discretization. Discretization can also depend on the spatial pattern of available climate data.

Process representation

After discretizing the watershed, modelers need to determine the appropriate hydrologic model complexity and process representation. The choice of model complexity decides which physical processes should be represented explicitly, and which processes can be ignored or simplified. The choice of process representation decides which mathematical approaches are used to represent the dominant hydrologic processes (Clark et al., 2015).

The basic hydrologic physical processes are interception, evapotranspiration, infiltration, groundwater flow, surface flow and routing, and snowmelt. The basic laws of classical physics that are used in various hydrologic process representations include the conservations of mass, energy, and momentum. The conservation of mass is relied upon most commonly. The conservation of energy is involved in evapotranspiration and snowmelt. The conservation of momentum is important in the analysis of fluid flow, specifically open channel routing of turbulence flows (Dingman, 2015).

Depending on the treatment of model complexity and process representation, hydrologic models can be divided into three types or classes: empirical, conceptual, and physically-based. The characteristics of the three model classes are compared in Table 1-2. Some frequently used hydrologic models are listed in Table 1-3.

- Empirical models, also called black box models, do not consider hydrologic processes (zero model complexity) but use approximate relationships developed from observations to simulate flows and storages (Dingman, 2015). For example, the unit hydrograph of a watershed is constructed from the rainfall and runoff observations of several storms of approximately equal durations within the modeled watershed (Dunne and Leopold, 1978). Empirical models rely on simplification.
- Conceptual models consider different hydrologic processes and represent them with conceptual storages. These conceptual storages are configured in series and/or in parallel to form a conceptual scheme. The conceptual scheme can be interpreted as a reasonably realistic representation of the functioning of the catchment, but there is no way of assigning physical locations to the storages in catchment (Beven, 2012). Conceptual models utilize empirical or semi-empirical equations to describe flow addition or subtraction within and between conceptual stores (Clark and Kavetski, 2010). Empirical equations are derived only from observations, while semi-empirical equations are based on a combination of observations and theoretical considerations (e.g., Darcy's law, Horton model for infiltration (Horton, 1941), or the degree-day model for snowmelt (Bergstrom, 1975)).
- Physically-based models also take into account different hydrologic processes but use the equations derived from basic physics (e.g., the conservations of mass, momentum, energy) to simulate flows and storages (Dingman, 2015). A rigorous physically-based hydrologic model may use the complete partial differential equations derived from physics to describe each hydrologic process. However, more

practical physically-based models use simplification formulae of the complete partial differential equations. For example, the St. Venant equations are the complete partial differential equations of open-channel flow, but the kinematic wave model (a simplification of the St. Venant equations) is more commonly used in physically-based hydrologic models.

Researchers' desire for physically-based models has to be tempered with practical limitations (Beven and Young, 2013; Freeze and Harlan, 1969). For example, the physical principles on which physically-based models are based are small scale physics of a homogenous system, while model grid is usually orders of magnitude larger than the theoretically applicable area. Lumping the small scale physical equations to model scale with the same parameters introduces conceptual elements (Beven, 1989). Also, the application of some physically-based equations can have massive data requirements that are not readily available and economically infeasible to acquire. As such, the existing so-called physically-based models are essentially a hybrid of physically-based equations and empirical equations (empirical or semi-empirical), in which some processes are described by physics derived equations, and others are expressed with empirical equations. The more hydrologic processes are understood and quantified with physically-based equations, the more the hydrologic model is physically-based.

Here it is worth clarifying the relation between the model structure identification and the number of model parameters. Hydrologic model parameters are constants in physically-based equations and empirical equations or the physical properties that are hard to measure in the field (e.g., hydraulic conductivity and porosity at the model spatial unit scale). The number of hydrologic model parameters depends on the hydrologic process representation as well as the number of computational units from watershed discretization. Increasing the number of computational units in a model leads to an increase in the number of hydrologic model parameters that need to be estimated.

Table 1-2. Characteristics of empirical, conceptual, and physically-based models

Comparison aspects	**Empirical model**	**Conceptual model**	**Physically-based model**
Consider hydrologic processes	No	Yes	Yes
Function equations	Empirical equations	Empirical and semi-empirical equations	Partial differential equations & Empirical and semi-empirical equations
Spatial representation	Lumped	Lumped	Distributed or semi-distributed
Parameter derivation	Calibrate based on observations	Estimate from measurements or calibrate based on observations	Estimate from measurements or calibrate based on observations
Advantages	• Easy to implement • Computationally efficient	• Easy to understand and implement	• Adequate descriptions of hydrologic processes • Good extrapolation capacity
Disadvantages	• Lowest explanatory depth • Lowest extrapolation capacity	• Limited physical interpretations • Limited extrapolation capacity	• Hard to understand and implement • Large number of parameters that require massive measurements and observations to estimate • Computationally inefficient

Table 1-3. Examples of empirical, conceptual and physically-based hydrologic models

Category	Model Acronym	Model Full Name	References
Empirical model	Unit hydrograph	Unit hydrograph	(Sherman, 1932)
	Transfer function model	Transfer function model	(Beven, 2012)
	ABC model	ABC model	(Kavetski et al., 2002)
	ARIMA	Autoregressive Integrated Moving Average model	(Box et al., 1994)
	CLS	Constrained Linear Systems model	(Todini and Wallis, 1977)
	API	Antecedent Precipitation Index model	(World Meteorological Organisation, 1994)
Conceptual model	SWM	Stanford Watershed Model	(Crawford and Linsley, 1966)
	GR4J	Ge'nie Rural a' 4 Parame'tres Journalier	(Perrin et al., 2003)
	HBV	Hydrologiska Byrans Vattenbalansavdelning	(Bergström, 1976)
	HYMOD	Hydrological MODel	(Boyle et al., 2011)
	Tank	Tank	(Sugawara, 1979)
	VIC	Variable Infiltration Capacity model	(Zhao, 1992)
Physically-based model	UBCWM	University of British Columbia Watershed Model	(Quick, 1995)
	WATFLOOD	Waterloo Flood Forecasting Model	(Kouwen, 1988)
	SWAT	Soil and Water Assessment Tool	(Arnold et al., 1998)
	TOPMODEL	TOPography based hydrological MODEL	(Beven and Kirkby, 1979)
	HRCDHM	Hydrologic Research Center Distributed Hydrologic Model	(Carpenter et al., 2001)
	TOPNET	TOPNET	(Bandaragoda et al., 2004)
	SHE	European Hydrologic System	(Abbott et al., 1986)
	MISBA	Modified Interaction Soil Biosphere Atmosphere	(Kerkhoven and Gan, 2006)
	IHDM	Institute of Hydrology Distributed Model	(Beven et al., 1987)
	tRIBS	TIN-based Real-time Integrated Basin Simulator	(Ivanov et al., 2004)
	HydroGeoSphere	HydroGeoSphere	(Therrien et al., 2010)
	PAWS	Process-based Adaptive Watershed Simulator	(Shen and Phanikumar, 2010)
	HYDRUS	HYDRUS	(Šimůnek et al., 2016)
	MODFLOW	MODular three-dimensional finite-difference ground-water FLOW	(McDonald and Harbaugh, 1988)

1.1.2. Model Parameter Estimation

Model calibration

Model calibration is the process of adjusting model parameters such that the model simulation results adequately predict the measured system response data (Duan et al., 2003). Hydrologic model calibration was originally implemented in a manual way in which experienced hydrologists use a trial-and-error procedure to adjust the parameters while comparing the observed responses and the simulated model outputs with graphical plots. Manual calibration was initially feasible because the hydrologic model design was simple and enables modelers to learn that the adjustment of certain parameters in certain directions has predictable effects on the model outputs (Gupta et al., 1999). However, as the hydrologic model structures become more nonlinear, it is difficult to predict the effects of model parameter adjustment on simulated model outputs. Moreover, the subjective nature of the manual calibration leads to different parameter results from different modelers (Pechlivanidis and Jackson, 2011).

Due to the time-consuming, unpredictive, and subjective nature of manual calibration, automatic calibration has been developed since the 1960s and early 1970s (Gupta et al., 1999). Automatic calibration iteratively samples parameter sets from the feasible parameter space, produces the simulated model outputs, and then, depending on the category of automatic procedure, calculates a metric that represents the quality of the parameter set such as an objective function, likelihood function, or pseudo-likelihood function. The iterative sampling-simulation-evaluation procedure is terminated when the optimal parameter solution is found, or the parameter distributions have converged, or a user-specified number of behavioral parameter sets is obtained, or the maximum number of user-specified model runs is reached.

Automatic calibration methods can be classified into different categories. Four classification methods are detailed below.

Depending on the calibration objective, calibration methods can be classified as time-domain and signature-domain. Taking the rainfall-runoff model as an example, the time-domain calibration seeks to match model outputs with observed streamflow time series. The signature-domain calibration seeks to match model outputs with specific features, or 'signature', of observed streamflow time series. Typical signatures used in hydrologic calibration include the flow duration curve, master recession curves, baseflow indices, and the other streamflow characteristics (Kavetski et al., 2018).

Depending on the number of the calibration objectives, calibration methods can be classified as single-objective and multi-objective. Single-objective calibration only optimizes one objective function. Multi-objective calibration optimizes more than one objective function simultaneously and is applied when optimal decisions need to be taken in the presence of trade-offs between two or more conflicting objectives. Multi-objective calibration employs multiple evaluation criteria (e.g., RMSE and reliability) or/and multiple sets of observations (e.g., low flows and high flows, water quantity and water quality) (Shafii et

al., 2015). These objectives are expressed in a numerical form and can be used to obtain the Pareto solutions, reflecting various trade-offs between the conflicting objectives. A solution is called Pareto optimal (or nondominated) if none of its objective function values can be improved without degrading some of the other objective function values (Asadzadeh and Tolson, 2013). Weighting factors can be used to obtain a composite objective function if a "globally optimal" single parameter set is required.

Depending on the treatment of parameter set equivalence, model calibration methods can be categorized into optimization-based and uncertainty-based. Optimization-based calibration aims to find a single "optimal" parameter set without recognizing parameter set equivalence (Tolson and Shoemaker, 2007) and thus views calibration as a deterministic estimation of model parameters. Uncertainty-based calibration identifies numerous 'plausible' parameter sets and acknowledges the correctness of different combinations of hydrologic response mechanisms and parameters (Tolson and Shoemaker, 2008). The generalized likelihood uncertainty estimation (GLUE) (Beven and Binley, 1992) first popularized the calibration of hydrologic models considering the model parameter uncertainty.

For uncertainty-based calibration approaches, depending on whether the likelihood function is formulated using probability theory, the calibration algorithms can be classified into two types: Bayesian inference and likelihood-free inference. Bayesian inference specifies the likelihood function on the basis of probability theory, while the likelihood-free inference does not need to specify a likelihood function on the basis of probability theory. From this point forward, the word 'likelihood' is used in situations where the likelihood function is defined strictly based on probability theory. The word 'pseudo-likelihood' is used in situations where the likelihood function is not defined based on probability theory. Some examples of the pseudo-likelihood functions are the Nash-Sutcliffe coefficient of efficiency (NSE) (Nash and Sutcliffe, 1970) and Kling-Gupta efficiency (Gupta et al., 2009).

Below we will provide some further introduction to Bayesian inference and two typical likelihood-free inference methods: GLUE (Beven and Binley, 1992) and dynamically dimensioned search-approximation of uncertainty (DDS-AU) (Tolson and Shoemaker, 2008). For each inference method, introductions cover its optimization algorithm, parameter sampling method, and prediction uncertainty characterization.

(1) **Bayesian inference**

Assume a deterministic hydrologic model is expressed as:

$$\widehat{\mathbf{Q}} = h(\boldsymbol{\theta}, \mathbf{x}) \tag{1-1}$$

where $\widehat{\mathbf{Q}}$ is the simulated streamflow time series, h is the hydrologic model, $\boldsymbol{\theta}$ is the hydrologic model parameter vector, $\mathbf{x}$ is the input data.

Assume the stochastic model output is expressed as:

$$\mathbf{Q}(\boldsymbol{\theta}, \mathbf{x}) = h(\boldsymbol{\theta}, \mathbf{x}) + \varepsilon \tag{1-2}$$

where deterministic model is corrupted by an error ε that accounts for all the hydrologic modeling errors.

For the time-domain calibration, the calibration objective $\mathbf{Y}(\boldsymbol{\theta}, \mathbf{x})$ is the streamflow time series:

$$\mathbf{Y}(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{Q}(\boldsymbol{\theta}, \mathbf{x}) \tag{1-3}$$

For the signature-domain calibration, the calibration objective $\mathbf{Y}(\boldsymbol{\theta}, \mathbf{x})$ is the signature $g(\mathbf{Q}(\boldsymbol{\theta}, \mathbf{x}))$ that is computed from THE streamflow time series $\mathbf{Q}(\boldsymbol{\theta}, \mathbf{x})$:

$$\mathbf{Y}(\boldsymbol{\theta}, \mathbf{x}) = g(\mathbf{Q}(\boldsymbol{\theta}, \mathbf{x})) \tag{1-4}$$

Given a stochastic model $\mathbf{Y}(\boldsymbol{\theta}, \mathbf{x})$, a prior distribution of model parameters $p(\boldsymbol{\theta})$ and observed data $\tilde{\mathbf{Y}}$, the posterior distribution of model parameters $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ can be estimated from the Bayes theorem by:

$$p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) = \frac{p(\tilde{\mathbf{Y}}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\tilde{\mathbf{Y}})} \propto p(\tilde{\mathbf{Y}}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) \tag{1-5}$$

where $p(\tilde{\mathbf{Y}}|\boldsymbol{\theta})$ is the likelihood of the measured objective (i.e., streamflow time series or signature) given the sampled parameters $\boldsymbol{\theta}$. $p(\tilde{\mathbf{Y}})$ is the distribution of the measured objective, which can be ignored if one is only interested in the relative posterior parameter probability.

There are two main methods for approximating the posterior parameter distributions. When the probability density function of the likelihood is determined, the Monte Carlo Markov Chain (MCMC) is a standard sampling method for sampling from the Bayesian posteriors. In MCMC, a Markovian random walk is constructed in the parameter space in such a way that the resulting steps are samples from the parameter distribution. The sampling is terminated when the sampling sequences converge to the target distributions, thereby the parameter inferences from them are reliable (Gelman, 1996). The two most common MCMC sampling methods in hydrologic model calibration are Metropolis-Hastings algorithm (Metropolis et al., 1953) and Gibbs sampling (Geman and Geman, 1984). When the probability density function of the likelihood is easier to sample from than to determine, the approximate Bayesian computation (ABC) sampling algorithm is used to sample from the Bayesian posteriors (Kavetski et al., 2018). The ABC accepts a parameter sample if its corresponding distance metric (e.g., RMSE between observed and simulated model output) is no bigger than the user defined small tolerance. The sampling is terminated when the required number of parameter samples is met.

After model calibration, the streamflow prediction probability distribution can be constructed by using the entire posterior parameter distribution along with the input data and model error (Huard and Mailhot, 2008). However, in practice, the prediction probability is approximated using a collection of sampled posterior parameter sets. For each sampled posterior parameter set, run the deterministic model (Equation (1-1)) to compute $h(\boldsymbol{\theta}, \mathbf{x})$, draw an error from the model error distribution (i.e., $\varepsilon \sim N(0, \delta_\varepsilon^2)$, δ_ε is known), and compute the streamflow time series $h(\boldsymbol{\theta}, \mathbf{x}) + \varepsilon$ (Equation (1-2)). Repeat the above process for a

sufficient number of parameter sets sampled from the joint posterior parameter distributions. The collection of all the computed streamflow realizations is used to approximate the posterior distribution of streamflow.

A prediction interval is an interval in the domain of the posterior prediction distribution. It is associated with a specified probability of the random variable (i.e., streamflow) lying within the bounds. For example, given the 95% prediction bounds of streamflow, the actual streamflow should lie within the bounds with probability 0.95. The lower and upper bounds of the 95% prediction intervals correspond to the 2.5^{th} and 97.5^{th} quantiles of the prediction probability distribution. Prediction bounds for a time series are determined independently for each time step.

(2) **GLUE**

GLUE is a likelihood-free inference, though the GLUE likelihood function can be, but is not typically, probability theory based (Tolson and Shoemaker, 2008). NSE is the most frequently used pseudo-likelihood function of GLUE (Stedinger et al., 2008). GLUE samples parameters from the prior parameter distributions (uniform distributions at most time), produces a deterministic model output (Equation (1-1)), and computes the pseudo-likelihood value between the observed and the deterministic model output. The parameter set is identified as behavioral when its pseudo-likelihood is better than the specified pseudo-likelihood criteria.

GLUE uses a Monte Carlo approach to sample parameters and is terminated when the specified number of behavioral parameter sets is met. The prediction uncertainty estimation of GLUE is conditioned on the pseudo-likelihood value. In detail, all the behavioral parameter sets are assigned rescaled weights based on their likelihood measures. The sum of all the rescaled weights is equal to one. The model outputs are ranked in order of magnitude and, using the rescaled likelihood weights associated with each model output, a distribution of the prediction is calculated (Beven and Binley, 1992).

There are two important aspects worth noting about GLUE. First, the core difference between Bayesian inference and GLUE lies in the treatment of the model error in the likelihood function formulation. Bayesian inference uses the probabilistic model output that explicitly incorporates the model error (ε) in the likelihood function formulation, while GLUE uses the deterministic model output in the pseudo-likelihood calculation. Second, when the GLUE likelihood is not formulated using probability theory (the case at most time), its parameter inference and streamflow prediction cannot be expected to provide a probabilistic description of uncertainty (Kavetski et al., 2018). For instance, Montanari (2005) reported a GLUE case study where the 95% GLUE prediction intervals include only 62% of the observed data. The reason behind this is that the prediction uncertainty is associated with the likelihood measure, and the likelihood magnitude is conditioned on the GLUE sampling efficiency. If the GLUE parameter sample approximates the maximum likelihood parameter solution, the likelihood value will be good, and thus its associated model output will play an important role of forming the prediction uncertainty. However, the

typical GLUE uniform random sampling is very inefficient, so it is hard for GLUE to identify high pseudo-likelihood parameter sets and thus characterize streamflow prediction uncertainty well (Kuczera and Parent, 1998; Tolson and Shoemaker, 2008).

(3) **DDS-AU**

DDS-AU is a likelihood-free inference method. DDS-AU method is based on multiple independent optimization trials of the dynamically dimensioned search (DDS) stochastic global optimization algorithm (Tolson and Shoemaker, 2007). The DDS algorithm searches globally at the beginning of the search and becomes a more local search as the number of iterations approaches the maximum allowable number of model runs. The search adjustment from global to local is realized by dynamically and probabilistically reducing the number of parameters that are to be perturbed in the next sampling. When moving to the next sampling, a parameter set is created by only perturbing the previously selected parameters, and the perturbation magnitudes are randomly sampled from a normal distribution with zero mean. The DDS algorithm terminates when the maximum allowable number of model runs is reached.

In DDS-AU, multiple DDS optimization trials are initialized to different initial solutions and/or different random seeds, so multiple DDS optimization trials will follow different search paths and can terminate at different final solutions. The final best parameter solutions (one per DDS optimization trial) are collected to identify the behavioral parameter sets based on the behavioral threshold (i.e., a specified pseudo-likelihood criteria). A parameter set that has a better pseudo-likelihood value than the behavioral threshold is identified as behavioral. All the behavioral parameter sets together constitute the uncertainty estimates of parameters.

DDS-AU uses all the behavioral parameter sets to construct the prediction intervals. All the model outputs of the behavioral parameter sets are treated equally. Like GLUE, the likelihood function of DDS-AU is not formulated based on probability theory, so its prediction uncertainty estimation should not be interpreted as a statistically rigorous description of predictive uncertainty. However, in contrast with GLUE, the DDS-AU behavioral threshold is based on the modeler's belief of what is acceptable but not conditioned on the inefficient Monte Carlo sampling results, so DDS-AU should generate more meaningful prediction intervals (Tolson and Shoemaker, 2008).

Given the importance of the behavioral threshold in inference and prediction, researchers explored ways of reducing the subjectivity of determination of the behavioral threshold for likelihood-free inference (e.g., GLUE and DDS-AU). For instance, Shafii et al. (2015) developed the Pareto rank-based behavioral solution identification approach and the optimal criteria-aggregation-based behavioral threshold identification approach. The idea of the two approaches is to simultaneously consider multiple aspects of the probabilistic prediction interval and to choose among the several groups of candidate parameter sets,

the group that achieves the best overall performance of the prediction intervals is taken as the set of behavioral parameter sets.

Model validation

Model validation is the process of applying a calibrated model to a different time period and/or region without changing the parameter values (Klemeš, 1986). Model validation is an essential step of examining the effectiveness of parameter inference. The model can be said to be validated if the indicator of model performance in the validation period meets the specified performance criteria.

Surprisingly, when comparing various model calibration method results, the focus of comparisons is often on calibration period rather than validation period performance. Specifically, many studies do not report on model validation results, such as Balin et al. (2010), Henn et al. (2016), Kavetski et al. (2002), Kavetski et al. (2006b), Kuczera (1983), and Sorooshian and Dracup (1980). In these studies, only the calibration period results were shown and used to evaluate the impacts of their developed calibration methods on parameter estimates and model outputs. Note that there are also many studies comparison methods based on the validation period performance and some examples include Honti et al. (2013), Shafii et al. (2014), Montanari (2005), Renard et al. (2011), and Vrugt et al. (2005).

Model parameter estimation process has four main sources of uncertainty. The first is model parameter uncertainty due to model conceptualization and the equifinality effect. The second is data uncertainty from the measurement and spatiotemporal representation of forcing data and response data. The third is model structure uncertainty from the imperfect representation of hydrologic processes. The fourth is initial and boundary condition uncertainty (Abbaszadeh et al., 2018).

1.1.3. Hydrologic Forecasting

A common reason to build, calibrate, and validate an operational hydrologic model is to use it within a hydrologic forecasting system. Hydrologic forecasting generates the prior estimates of future states of hydrologic phenomena (WMO, 2003). Lead time, also known as forecast period, is a key characteristic of hydrologic forecasts. According to lead time, the World Meteorological Organization (1994) classifies hydrologic forecasts into three categories: short-term (up to two days), medium-range (2-10 days), and long-range (exceeding 10 days).

Hydrologic forecasting includes three phases: simulation, data assimilation, and forecasting (Thiboult et al., 2016). In simulation, the hydrologic model runs with measured climate until the first lead time of the forecast. Simulation here is essentially a type of hindcasting to warm up the hydrologic model with historical data. In data assimilation, the model state estimates (e.g., reservoir water level, soil moisture) are updated by combining the observed and simulated states (if applicable) or system response (e.g., streamflow) using a certain data assimilation algorithm (Reichle, 2008). In forecasting, the hydrologic model runs with the forecasted climate and generates the future flow estimates until the end of the lead time.

Data assimilation

As can be seen from the above description, data assimilation combines the complementary information from measurements and models to provide the optimal estimates of system states. Data assimilation methods can be categorized into three types. The first type is simplistic data assimilation, such as replacing the model estimate with the observation (direct insertion). The simplistic data assimilation method can be useful, for example, for the assimilation of snow cover observations (Rodell et al., 2004). The second type is sequential assimilation, which uses observations as they become available to correct the current state of the model. The introduction of the Kalman filter (KF) (Kalman, 1960) was undoubtedly a landmark. In KF, the prior state estimates and their uncertainties are first predicted, and when a corresponding measurement is observed, the state estimates are updated using a weighted average, with more weight being given to the estimates with higher certainty. A limitation of KF-based methods is its Gaussian approximation to the posterior distributions of state variables. Particle filtering (PF) is specifically developed to overcome the Gaussian assumption of KF-based methods (Pasetto et al., 2012). PF is an ensemble-based sequential assimilation, and its posterior distributions of state variables are characterized by a set of discrete random particles (samples) with associated importance weights (Liu and Gupta, 2007). The important weights are normalized so that they sum up to one. A limitation of PF is that it is not applicable to high dimensional state variables due to the curse of dimensionality issue. Curse of dimensionality means that when the state dimension increases, the largest importance weight converges to one, thus the state posterior distribution is represented by a single point in the state space (Moradkhani et al., 2012; Stordal et al., 2011). The third type of data assimilation methods is variational data assimilation (VDA), which updates states over a pre-determined window of time in order to optimize the initial state of the model. The VDA approach primarily relies on optimization theory, such as minimization of a cost function aggregating the error over a pre-determined assimilation window (Abaza et al., 2014b).

In addition to upgrading model state estimates, data assimilation has been extended to estimate model parameters in recent years. This extension arises because parameter estimation (model calibration) and data assimilation both help correcting model prediction biases. Model calibration usually focuses on parameter uncertainty, and the estimated model parameters and model states are time invariant. In contrast, data assimilation has capacity of considering various hydrologic modeling uncertainties (typically accounting for forcing and response data uncertainties), and the estimated model parameters and/or model states can be time variant. It is desirable to combine model calibration with data assimilation to consider all kinds of uncertainty in parameter estimation (Liu and Gupta, 2007). In the literature, several approaches have been established to jointly estimating state and model parameters. For example, Moradkhani et al. (2012, 2005) developed two dual state-parameter estimation methods based on ensemble Kalman filter (EnKF) and PF, respectively. In these methods, states and parameters are recursively estimated based on Monte Carlo

sampling and sequentially updated by EnKF or PF method at each time step. Smith et al. (2013) augmented state vector with model parameters to form an extended state vector and updated the extended state vector using four-dimensional variational (4D-Var) assimilation.

Streamflow forecasting

The hydrologic forecasting paradigm has shifted from deterministic to probabilistic for pursuing a better, more detailed evaluation and description of the hydrologic forecast uncertainty. Deterministic forecasting gives only one possible forecast value at each time step, while probabilistic forecasting produces an ensemble of forecast values at each time step from which one can estimate the probability distribution of the system response.

Historically, ensemble streamflow forecasting was implemented based on historical observations, called extended streamflow prediction, assuming that past meteorological events can represent future events (Day, 1985). Nowadays ensemble flow forecasting uses ensemble weather forecasts to predict flows and often achieves reduced spread compared to extended streamflow prediction (Cloke and Pappenberger, 2009). The increasing accessibility of ensemble meteorological prediction benefits the hydrology community in issuing ensemble streamflow forecast. At the beginning of the 1990s, meteorologists pioneered the operational use of ensembles by constructing the meteorological ensemble prediction systems (MEPSs), mostly to take into account imperfect initial conditions that are a prime uncertainty source considering the chaotic nature of the atmospheric physics (Thiboult et al., 2016).

In addition to the weather forecast uncertainty, ensemble streamflow forecasting is also subject to uncertainties from operation and hydrologic modeling. Operational uncertainty is caused by erroneous or missing data, human processing errors, and unpredictable interventions. Hydrologic modeling uncertainty includes all the errors related to hydrologic modeling (e.g., model parameter, data, model structure, and initial and boundary condition) (Krzysztofowicz, 1999).

1.2. Data Uncertainty Overview

This thesis specifically investigates the data uncertainty handling in hydrologic model calibration and data assimilation. Data uncertainty includes errors from the forcing data used to drive the model and the response data used to evaluate the model. Below we will review the uncertainty sources and uncertainty treatments in model calibration and data assimilation for climate forcing and streamflow data, respectively.

1.2.1. Data Uncertainty Source and Estimation

The forcing data of a hydrologic model includes time series of precipitation and temperature for all physically-based hydrologic models, as well as shortwave/longwave radiation, air density and pressure, wind velocity, etc. in some of the more complex hydrologic models. This thesis focuses on precipitation and temperature because they are most common inputs to physically-base hydrologic models and can cause

substantial errors to other forcing data when the other forcing variables (e.g., radiation, evapotranspiration) are derived from precipitation and temperature with empirical algorithms (Newman et al., 2015).

Climate data uncertainty source and estimation

Measurement is a main source of uncertainty of climate data uncertainty. For gauge measurement, for example, the tipping bucket rainfall measurement, errors arise from wind, wetting, evaporation, and splashing losses (Humphrey et al., 1997). For remote sensing measurement, satellite and radar measure radiances or electronic radio signals which are used in subsequent mathematical inversion, errors arise from the radiance or electronic radio signal measurements.

Climate data uncertainty is also related to sampling mechanism and estimation algorithm. For example, in the tipping bucket rainfall measurement, errors arise when the bucket size and the recording frequency are not representative (Habib et al., 2002). Errors also appear when the rain gauge network and the choice of interpolation method are not appropriate for all rainfall events. In remote sensing based measurement, errors are associated with the choice of the inversion algorithm (e.g., the retrieval algorithm, radar reflectivity-rain rate relationship) (Nijssen, 2004; Schiemann et al., 2011). Moreover, biases arise from upscaling or downscaling meteorological variables to the effective modeling scale.

In order to estimate climate data uncertainty, the classical and straightforward way is to represent the real climate variable in terms of a probability distribution or an ensemble of realizations. The way of generating space-time uncertain climate data has been extensively explored in the atmospheric and oceanic sciences, and precipitation (rainfall mainly) is the most studied data type. These methods can be roughly categorized into two types based on their application scales and computational costs. One is stochastic weather generator (WG) and another is numerical weather prediction (NWP) (Ailliot et al., 2015; Semenov and Stratonovitch, 2010).

WG focuses on small spatial scales, typically a few sites within a region extending over few kilometers and is characterized by providing numerous random realizations of the weather variables, mainly at the daily or sub-daily scales. Ailliot et al. (2015) divided the current stochastic WGs into four groups: resampling methods, the Box-Jenkins methodology, point process models, and hierarchical models, in which the hierarchical model is the most versatile approach for building multisite and multivariate WGs. WGs are not designed to produce past or forecast values of the weather variables but provide multiple realizations of the weather variables by perturbing observations (Gaborit et al., 2013; Wilks and Wilby, 1999). Therefore, the WG generated synthetic sequences of the weather variables are widely used in impact studies (Ailliot et al., 2015; Chandler and Wheater, 2002; Kyriakidis et al., 2004). For example, in the hydrologic impact analysis of climate change, the WG generated precipitation and temperature synthetic realizations can represent the characteristics of certain climate change scenarios and be used in rainfall-runoff models to estimate the impacts of precipitation and temperature changes on runoff, soil moisture

regime, and groundwater recharge (Refsgaard and Abbott, 1990). In addition, the WG generated rainfall consequences can also represent rainfall data uncertainty and have been used in some model calibrations (Balin et al., 2010; Blazkova and Beven, 2009; Del Giudice et al., 2016).

In contrast to WG, NWP features generating the behavior of the whole atmosphere and its interactions with other components (vegetation, oceans, etc.) at an extensive scale and for a long time period (~2-15 days) (Ailliot et al., 2015). NWP models use the initial state of the atmosphere to simulate meteorological variables. Initial states include surface observations, upper air observations, vertical wind profilers, remote sensing satellite data, airplane measurements, etc. Model outputs typically contain precipitation, radiation, temperature, wind, humidity, and pressure (Al-Yahyai et al., 2010). NWP models comply with fundamental physical principles such as conservation of mass, momentum, and energy with the model system. NWP can be operated at global or regional scale. Global climate models are characterized by extensive coverage but limited by extremely expensive computational costs and low spatial resolution (e.g., 30-50 km) (Dietrich et al., 2009). Limited area models cover the domain of the country and the surrounding areas at a high resolution (less than 10km) (Semenov and Stratonovitch, 2010).

Streamflow data uncertainty source and estimation

Continuous streamflow is not measured directly by any instrument but derived from water levels based on the level-discharge relationship, called a rating curve. In order to get the rating curve, the continuous records of stage (i.e., the height of water level) and discharges need collected at a location along a stream. Discharge is calculated by multiplying the area of water in a channel cross section by the average velocity of the water in the cross section in an aggregation way, so the height and width of the subsection and the velocity of the streamflow at each subsection also need to be measured. The rating curve can be defined based on the instantaneous stage and discharge measurements (gauging data). The fitted rating curve is then used to convert the measured stage to the estimate of streamflow (Olson and Norris, 2007).

The usual expression of the rating curve in a uniform channel and conventional section controls can be expressed as a power function (ISO 1100, 2010):

$$Q = a(h - b)^c \tag{1-6}$$

where Q is discharge, h is stage, a is a scaling coefficient, b is the reference level, and c is an exponent and can be related to the type of hydraulic control. Developing an accurate rating curve relation requires numerous instantaneous measurements at all ranges of stage and discharge. Rating curves differ from site to site depending on the site hydraulic condition such as the shape, size, slope, and roughness of the channel and flow stability. Moreover, for each site, the rating curve needs continuously checked and modified due to channel geometry changes from erosion, deposition, flood, seasonal vegetation growth, debris, or ice.

According to Coxon et al. (2015), there are four main sources of uncertainty in discharge estimations: (1) gauging data, (2) natural conditions, (3) rating-curve approximation, and (4) human alterations. Gauging

data errors arise from point stage, cross section height, weight, and velocity measurements, and insufficient sampling. Natural processes can lead to channel geometry and cross section changes through erosion, sedimentation, hysteresis, and weed growth. Rating curve fitting is also a significant source of uncertainty depending on the prior knowledge of the rating curve relation, the number and coverage of points over the flow range, in particular at high flow, and the multi-section breaks due to different channel or hydraulic controls. Finally, human regulation and intervention affect the discharge uncertainty through either changing the gauging station, or the use of multiple weirs at a station where the rating curve changes depending on which weir is in operation.

Rating curve uncertainty estimation is the main mechanism for probabilistically estimating discharge in the literature. Traditional statistical approaches exploit the residual variance of the rating curve regression to determine the discharge uncertainty bounds but fail to explicitly incorporate the measurement uncertainty in gauging data (Petersen-Øverleir, 2006; Venetis, 1970). Bayesian methods are advantageous to using the hydraulic knowledge to set the prior distributions of rating curve parameters and deriving a likelihood function that accounts for gauging data measurement uncertainties (Le Coz et al., 2014; Moyeed and Clarke, 2005). Bayesian methods have been applied to numerous gauging stations and multiple types of rating curve relationships and have become the preferred choice to estimate flow data uncertainty (Coxon et al., 2015; Petersen-Øverleir, 2006).

1.2.2. Data Uncertainty Handling in Model Calibration

Climate data uncertainty handling in model calibration

In most model calibration studies, climate data uncertainty is ignored or assumed negligible, and response data uncertainty is lumped with the model error into a single random independent Gaussian error with a constant variance (Kuczera et al., 2006). However, numerous studies have revealed that ignoring climate data uncertainty and assembling all errors into a white noise term lead to biased parameter estimates and unreliable simulation bounds (Kavetski et al., 2002; Kuczera et al., 2006). Therefore, researchers have increasingly paid attention to explicitly consider data uncertainty during model calibration.

Hydrologic model calibration studies usually use two data error models, additive and multiplicative, to quantify climate data errors. In additive error models, the error term is usually assumed Gaussian with zero mean and a constant or proportional variance depending on the variable (Beven and Freer, 2001). In multiplicative error models, the error term is usually assumed Gaussian with a mean of one and a heteroscedastic variance proportional to the magnitude of observations (Moradkhani et al., 2005). Hydrologic model calibration deals the above error models in two different ways. The first way is inferring the climate relevant parameters with model parameters through model calibration (see Blazkova and Beven, 2009; Kavetski et al., 2002). The second way is characterizing the forcing data uncertainty prior to

hydrologic model calibration and unconditionally sampling from this characterization to estimate hydrologic model parameters (e.g., Balin et al., 2010).

Streamflow data uncertainty handling in model calibration

The classical response data error model is quantified with an additive Gaussian error that actually combines model structure and response data errors and sometimes forcing data errors together. In order to distinguish model structure errors from response data errors, Kuczera et al. (2006) assigned an error term for the response data error and the structural error, respectively. The structural error is characterized by the storm-dependent variation of one or two model parameters. Therefore, the two errors can be directly represented and propagated within Bayesian inference. Although this method is a huge advance in disaggregating response data errors, its application is rare in the literature. One possible reason is that distinction between different storms is practically difficult (Vrugt et al., 2008), so the storm-dependent sensitivity analysis is limited. Another reason is that it is difficult to integrate these errors into a formal likelihood function in the Bayesian analysis.

GLUE also explored the explicit consideration of response data uncertainty in model calibration. Beven (2006) proposed using the triangular, trapezoidal, or Beta distributions to define the acceptable error around the system response observation, thereby generating the acceptable range of the model simulated response. In model evaluation, given a sampled parameter set, if the model output is within the acceptable range, a higher weight is assigned to the prediction. All the weights for the individual data points (e.g., streamflow at each time step) are combined to provide a single weight associated with a particular model.

1.2.3. Data Uncertainty Handling in Data Assimilation

Data assimilation improves the state estimates of a system by accounting for the errors of the forcing and response data and the errors of the model itself (Abaza et al., 2014b). Successful data assimilation relies on accurate estimates of forcing and output data uncertainty (Moradkhani et al., 2005). The most commonly used data error model in data assimilation is the additive Gaussian error with zero mean. The variance is either a constant value or a proportion of the variable measurement. The Gaussian assumption is prevalent because some data assimilation methods such as the Extended Kalman filter are based on the assumption that the model and output observation errors are normally distributed with zero mean. Despite the Gaussian error assumption prevalence, some other assumed distributions have shown promise. For example, the multiplicative error and the gamma distribution have worked as precipitation error descriptors in some EnKF and VDA applications (Abaza et al., 2014b; Moradkhani et al., 2005; Thiboult et al., 2016).

1.3. Research Gaps

A key difficulty in the watershed discretization process is maintaining a balance between the aggregation-induced information loss and the computational cost increase caused by the inclusion of additional

computational units. Most modelers make discretization decisions in an ad hoc way or use a cumbersome trial-and-error approach by building and possibly calibrating multiple models with discretization candidates (Haghnegahdar et al., 2015). In order to quantify the information loss as well as to avoid time-consuming model runs, a few studies have designed *a priori* discretization error metrics. However, these *a priori* error metrics fail to: directly target hydrologically useful indicators; refer to the original spatial input data in a complete way (i.e., cell-by-cell comparison); and evaluate discretization quality from the distributed perspective (Booij, 2003).

Regarding handling data uncertainty, it is found that the common way that model calibration and data assimilation explicitly consider data uncertainty is adding or multiplying an error term to observation data. The error term follows a certain subjectively assumed probability distribution and can be conveniently encoded in calibration and assimilation algorithms. However, the emerging data uncertainty estimation approaches in other disciplines are advantageous over these subjectively assumed statistical error models because they respect the conservation equations of the system and take into account physical processes or relevant physical variables. Therefore, it is wise to adopt these advanced data uncertainty estimates to represent data errors in model calibration and data assimilation methods.

1.4. Aim and Scope

The aims of this thesis are to (1) find a quantitative solution to measure discretization-induced information loss of input data and facilitate discretization decision making, and (2) apply advanced climate and flow data uncertainty estimation products and methods to hydrologic model calibration and data assimilation processes to improve hydrologic modeling and streamflow forecasting.

Limits to this thesis are noted here. The research does not examine model structure uncertainty. Climate data uncertainty is narrowed down to the precipitation and temperature variables only. Hydrologic model calibration relies upon likelihood-free calibration methods and so residual error models of hydrologic model predictions are not constructed. All ensemble model predictions are based on assuming all members of the ensemble (e.g., behavioral parameter sets) are equally likely and hence do not need to be weighted as a function of their calibration period performance. All hydrologic forecasting experiments are based on the EnKF. Model calibration and data assimilation processes are independent, and model parameters are estimated before and are fixed in data assimilation. Model parameters are treated as time-invariant.

1.5. Thesis Contributions

This thesis has four contributions. The four contributions relate to the four hydrologic modeling and application processes shown in the following figure.

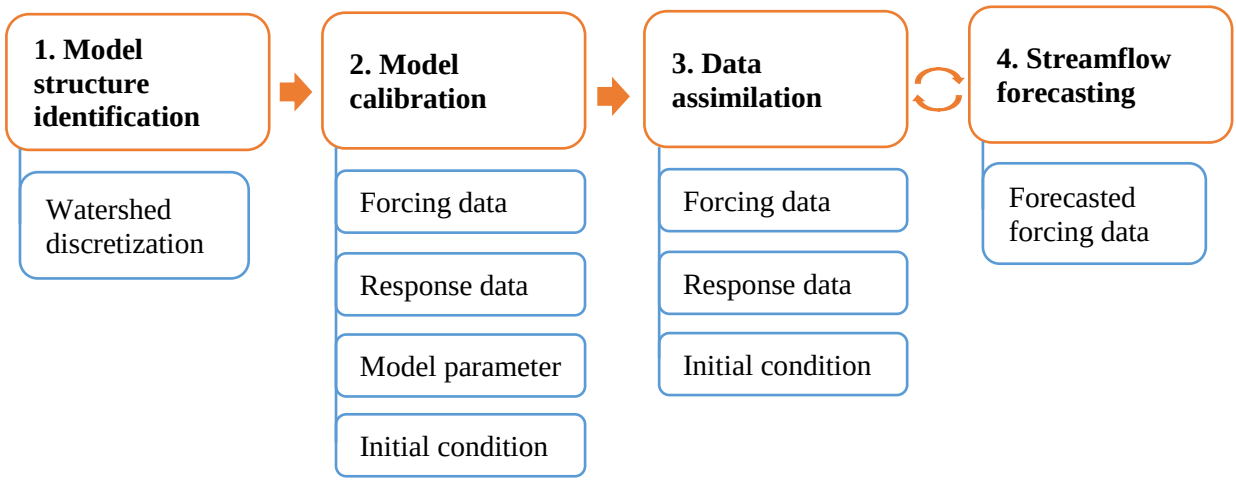

Figure 1-1. Diagram of four thesis contributions relevant hydrologic modeling and application processes followed by the uncertainty sources considered in this thesis.

1. Develop and apply an *a priori* discretization error metrics that can quantify the watershed discretization induced information loss. These metrics do not require hydrologic simulation, are independent of any specific modeling software, provide modelers with directly interpretable information on discretization quality, and allow for multi-site and multi-variable discretization evaluations prior to hydrologic model development. Informed by the error metrics, a two-step discretization decision-making approach is proposed. The discretization decision-making approach facilitates non-uniform discretization and meeting user-specified discretization error targets.
2. Propose a climate-flow ensemble based hydrologic model calibration framework. This framework avoids the use of assumed statistical error models, replacing them with more advanced and realistic climate and flow uncertainty estimation methods and products. It provides a simple way to explicitly consider climate and flow data uncertainty in model calibration. The framework is developed in two steps: the first step builds the climate ensemble based calibration framework, and the second step adds flow ensemble and accomplishes the climate-flow ensemble based calibration framework.
3. For the first time, compare the Newman et al. (2015) dataset (referred to as N15 in this thesis) derived climate ensemble with two carefully tuned hyper-parameter generated climate ensembles in the EnKF based ensemble streamflow forecasting. The carefully tuned hyper-parameters and their corresponding flow forecasting results are taken from Thiboult et al. (2016). Results show that the N15 generated climate ensemble yields improved or similar flow forecasts relative to the two traditional climate ensembles. Therefore, it is possible to eliminate the time-consuming climate relevant hyper-parameter tuning from the EnKF by using existing ensemble climate products without losing flow forecast performance.

4. Develop a robust single hydrologic model based ensemble streamflow forecasting system by using the parameter and climate data uncertainty based model calibration and the more realistic climate ensemble based EnKF. The robust single-model forecasting system is compared with a traditional multi-model forecasting system. The multi-model flow forecasts are taken from Thiboult et al. (2016). Results show that the robust single model generates improved high flow forecasts relative to multiple models. Therefore, it is an advantage to use a robust single hydrologic model for ensemble high flow forecast. The robust single model releases modelers from building multiple models in operational streamflow prediction.

1.6. Thesis Structure

After the introduction and the background in Chapter 1 (this chapter), Chapter 2 presents the *a priori* discretization error metrics published in *Journal of Hydrology*. Chapter 3 details the climate ensemble based hydrologic model calibration framework based on a manuscript submitted to *Advances in Water Resources*. Chapter 4 compares the N15 derived climate ensemble with the hyper-parameter derived climate ensembles in the EnKF based ensemble streamflow forecasting, which is published in *Journal of Hydrology*. Chapter 5 compares the robust single-model forecasting system with a traditional multi-model forecasting system in ensemble streamflow forecasting. Chapter 6 completes the climate-flow ensemble based calibration framework by adding flow ensemble to the climate ensemble based calibration framework (Chapter 3). Chapter 7 ends this thesis with research conclusions, limitations, and future work directions.

Chapter 2.
A Priori Discretization Error Metrics for Distributed Hydrologic Modeling Applications

This chapter is a mirror of the following published article with minor changes to increase its consistency with the body of the thesis. Changes were only made in the Summary (abstract). References are unified at the end of the thesis.

Liu, H., Tolson, B.A., Craig, J.R., Shafii, M., 2016. A priori discretization error metrics for distributed hydrologic modeling applications. J. Hydrol. 543, 873–891. https://doi.org/10.1016/j.jhydrol.2016.11.008

Summary

Objective identification of an appropriate discretization scheme remains a challenge in part because of the lack of quantitative measures for assessing discretization quality, particularly prior to simulation. This thesis proposes *a priori* discretization error metrics to quantify the information loss of any candidate discretization scheme without having to run and calibrate a hydrologic model. Informed by the error metrics, a two-step discretization decision-making approach is proposed with the advantage of reducing extreme errors and meeting the user-specified discretization error targets. The metrics and decision-making approach are applied to the discretization of the Grand River watershed in Ontario, Canada.

Section 2.1 provides the definitions of watershed discretization and relevant concepts (e.g., subbasin, hydrologic response unit), and gives a critical review of existing techniques of watershed discretization. Section 2.2 describes the thesis proposed *a priori* discretization error metrics and the two-step discretization decision-making approach. Section 2.3 explains the error metric applications to the Grand River watershed discretization. Section 2.4 provides an effective discussion of the proposed methods. Section 2.5 ends this chapter with a brief conclusion.

2.1. Introduction

In distributed hydrologic modeling, a watershed is treated as a number of small homogeneous units to address the spatial heterogeneity which results from variability of physical processes and physical character across a watershed (Singh and Frevert, 2005). This spatial heterogeneity is often attributed to the uneven distribution of a hydrologic properties across a watershed (Anselin, 2010). The spatial discretization process, whereby we separate a watershed into homogeneous computational units for depiction in a hydrologic model, is really the effort of determining how to characterize the inherent spatial heterogeneity found in a watershed. In general, spatial discretization should be detailed enough to capture the dominant processes and natural variability, while it also needs to be as concise as possible to save computation time and respect data availability (Booij, 2005). Excessively detailed spatial discretization increases model complexity (i.e.,

number of computational units) and thus increases model computation time. However, an overly coarse aggregation can lead to substantial information losses and give rise to increased model structural uncertainty, whose impacts on hydrologic predictions are far more adverse than those of parameter and data uncertainty (Fortin et al., 2010; Liu and Gupta, 2007). Therefore, defining an appropriate level of discretization is a critical task in distributed hydrologic modeling.

In order to investigate spatial discretization, it is necessary to first clarify the components of watershed discretization. This thesis will be examining the common subbasin-HRU discretization approach. In this approach, a watershed is discretized into a set of one or more subbasins, which can be further discretized into a number of contiguous or non-contiguous hydrologic response units (HRUs), defined as areas with hydrologically unique response to meteorologic events. Subbasins are referred to by different names in the literature, including grid cell, subcatchment, and subwatershed (Tuppad, 2006). Here we recursively define a subbasin as the drainage area of a location on a stream network minus the drainage areas of one or more upstream subbasins which flow directly into the subbasin. Headwater subbasins are those which do not have any subbasins upstream, i.e., those whose drainage areas are equal to their subbasin area. An HRU is the basic computational unit of hydrologic simulation and typically defined as a unique combination of hydrologic response determinants such as soil, land cover, terrain type, and management policy (Flügel, 1995), often generated from readily available mapping products. The HRU is conceptually similar to other computational units such as the Representative Elementary Area (REA), Representative Elementary Watershed (REW), Grouped Response Unit (GRU), hydro-landscape unit, and field (Dehotin and Braud, 2008; Fenicia et al., 2016; Kouwen et al., 1993; Reggiani et al., 1998; Wood et al., 1988), and therefore the approach developed here will port over to models which are discretized using these alternative definitions of the smallest computational unit. In recent decades, the traditional approach for watershed discretization has been to use Geographic Information Systems (GISs) such as ESRI's ArcGIS software or ArcGIS-based toolkits such as Arc HYDRO, ArcSWAT, and HEC-GeoHMS (ESRI, 2014; Fleming and Doan, 2009; Maidment, 2002; Winchell et al., 2007). While such automatic techniques make watershed discretization easy to practically implement, they do not have an explicit mechanism to account for, or assess, spatial input data information losses due to discretization choices. Here, information loss refers to the content change between candidate discretization schemes and the original, fully detailed, input data layers. Instead, modelers can only explicitly assess the model complexity under candidate discretization schemes based on the number of modelled homogeneous areas (subbasin or HRU computational units).

Haghnegahdar et al. (2015) claim that most modelers make discretization decisions in an ad hoc fashion. This approach is often based on the past experience of the modeler, rules of thumb or default discretization settings in specialized ArcGIS-based toolkits for creating a distributed hydrologic model (e.g., ArcSWAT (Winchell et al., 2007)). The shortcoming with all ad hoc approaches is that there is no

quantitative or formal justification of the selected discretization over other potential discretization choices. More sophisticated discretization approaches found in the literature use a cumbersome trial-and-error approach of building and then possibly calibrating multiple candidate models with different discretization levels in order to identify the most appropriate choice. For example, Arnold et al. (2010) compared the calibration and validation period flow simulation results of an enhanced SWAT model with four landscape delineations, and Petrucci and Bonhomme (2014) tested the calibration and validation period water quantity and water quality simulation results of six different discretization scenarios of the Stormwater Management Model. Haghnegahdar et al. (2015) followed a similarly intensive but improved process except that they took into account the computational time spent for calibrating (calibration budget) and focused on the model performance in ungauged basins under four discretization schemes for a land-surface hydrologic modeling application. All of these approaches require model calibration in order to assess the quality of a given discretization scheme.

Given the above limitations, other studies have instead focused on designing *a priori* discretization error metrics to quantify the information loss incurred from spatial discretization. Such metrics are advantageous in that they do not require model runs. Haverkamp et al. (2002) provided an entropy based statistical tool, the Subwatershed Spatial Analysis Toll (SUSAT), to estimate the information loss for subwatershed and HRU discretization, respectively. Booij (2003) utilized the bias of the variance of aerially averaged variables under different correlation lengths to decide the appropriate modeling scale. Dehotin and Braud (2008) used Manhattan distance to measure the composition descriptor (e.g., histogram, mean, standard deviation, or matrix of co-occurrence) similarity between each mapping cell and the reference zones. There are three main shortcomings of the existing *a priori* discretization error metrics. First, the metrics do not directly correlate to the information required by hydrologic modeling applications, in particular for semi-distributed modeling. For example, entropy represents spatial disorder from the systematic perspective, but spatial heterogeneity essentially describes spatial pattern variability (Journel and Deutsch, 1993). Changes in system disorder cannot fully reflect the (more hydrologically important) changes in spatial heterogeneity and hence entropy is not a directly interpretable indicator for hydrologic modeling. Second, their property change identification process fails to refer to the original spatial input data in a complete way (i.e., cell-by-cell comparison). Instead, they use the overall heterogeneity statistics difference between a candidate discretization scheme and the original spatial input data as the information loss, which may lead to the equifinality problem. Finally, the existing *a priori* approaches are all aggregated (e.g., over the entire study watershed) and do not provide spatially distributed evaluations of candidate discretizations. The importance of evaluating distributed model behaviors rather than an integrated value (e.g., runoff at the watershed outlet) for distributed models has been highlighted by numerous researchers (Beven and Binley, 1992; Grayson et al., 1995; Refsgaard, 1997; Shrestha and Rode, 2008). Just like multi-

site calibration provides an efficient framework for spatially distributed evaluations (Madsen, 2003), multi-site discretization quality assessment is intrinsically valuable to reduce the prevalence of aggregation or compensation effects in distributed hydrologic modeling. With such shortcomings in mind, this chapter is focused on developing *a priori* discretization error metrics that are directly interpretable, spatially distributed, and hydrologically relevant, providing a direct measurement of information loss relative to the original spatial input data, where the original spatial data is presumed to have the highest information content.

In addition to the information loss induced by the extensively studied HRU discretization, another type of information loss occurs due to subbasin discretization which affects the routing processes of semi-distributed and distributed models, hereinafter called routing information loss. In a finely discretized fully distributed model, channel structure, channel roughness, and therefore network travel times can be well-respected. As the watershed is discretized into subbasins, stream network branches are implicitly merged, replaced, and shortened. As far as we know, in the published literature, the routing information loss has never been quantified though its significance has been highlighted by many studies. For example, Haverkamp et al. (2002) indicated that the influences of the routing structure through subbasins to the watershed outlet should be considered in discretization evaluations when the effect of the routing on model results is not negligible. Dehotin and Braud (2008) emphasized the prospect of inclusion of linear discontinuities, including river reaches, hedges, ditches, and dikes, in order to properly describe networks in spatial discretization. Here, we address this need through the introduction of additional error metrics to estimate the routing information loss due to subbasin discretization.

The specific goals of this chapter are to (1) introduce *a priori* discretization error metrics to quantify the information loss due to subbasin and HRU discretization, respectively; (2) propose a two-step decision-making approach to identify an appropriate discretization scheme; (3) apply the error metrics and decision-making approach to the discretization of the Grand River watershed in Ontario, Canada. The simplicity of the error metrics allows for easy recoding and adoption into the preprocessing of a wide range of distributed models, including all semi-distributed models, such as HBV (Bergström, 1992), TOPMODEL (Beven and Kirkby, 1979), WATFLOOD (Kouwen, 1988), the Soil and Water Assessment Tool (SWAT) (Arnold et al., 1998), and Modélisation Environmentale–Surface et Hydrologie (MESH) (Pietroniro et al., 2007). The error metrics may also be useful for fully distributed models, e.g., System Hydrologique Europeen (SHE) (Abbott et al., 1986), TOPKAPI (Ciarapica and Todini, 2002), and Soil Moisture Distributed and Routing (SMDR) (Srinivasan et al., 2007) when the model cell scales are greater than the resolution of original spatial input data.

2.2. Methods

2.2.1. Discretization Error Metrics

The proposed *a priori* discretization error metrics provide a novel and simple quantitative measurement of the information loss in the process of spatial discretization. They are introduced for the purpose of assessing candidate discretization schemes and finding an appropriate discretization level in data preprocessing without having to rely on computationally intensive hydrologic model building exercises. For each candidate discretization scheme, the metrics are designed to compare the user-defined key model input variable properties with that of a reference discretization scheme. The reference scheme is defined as a scheme that fully retains the information of the original spatial input data or, in special cases, the finest plausible discretization. Both a subbasin discretization error metric and a HRU discretization error metric are proposed.

2.2.1.1. Subbasin Discretization Error Metric

In general, the routing process has two components: in-catchment routing and in-channel routing. In-catchment routing occurs within a subbasin and refers to the means of handling the delayed release of water from runoff, interflow, and baseflow to a subbasin outlet. This time delay is typically described by a unit hydrograph. In contrast, in-channel routing is the means by which water is exchanged downstream between subbasins and within the main channel of each subbasin. These definitions are applied by other models like ArcSWAT and HEC-GeoHMS (Fleming and Doan, 2009; Winchell et al., 2007). In this thesis, the subbasin discretization assessment focuses on the influences of discretization only on in-channel routing. The approach assumes that in-channel routing is unidirectional (i.e., water moves downstream only through a branching stream network), each subbasin has one outlet and one main channel, headwater subbasins have no main channel for routing, and non-headwater subbasins have upstream subbasin flows added to the beginning of their respective main channels. Should any of these assumptions not hold in other modeling case studies, the error metric procedures detailed below would need to be adjusted accordingly.

Calculation of subbasin discretization errors requires a high-resolution reference subbasin discretization scheme. For the drainage area upstream of a subbasin outlet, the in-channel routing length error (ΔL_s) equals to the in-channel routing length difference between the reference scheme (scheme 0) and the evaluated discretization (scheme s) as shown below.

$$\Delta L_s = L_0 - L_s = \frac{\sum_{i=1}^{n} A_{i0} L_{i0}}{\sum_{i=1}^{n} A_{i0}} - \frac{\sum_{j=1}^{m} A_{js} L_{js}}{\sum_{j=1}^{m} A_{js}} \tag{2-1}$$

where $L_0$and $L_s$are respectively the area-weighted in-channel routing length of scheme 0 and scheme s. For scheme 0, there are n subbasins within the evaluated drainage area and $i = 1,2,\cdots,n$ represents subbasin indices. A_{i0} is the area of subbasin i in scheme 0, and $\frac{\sum_{i=1}^{n} A_{i0} L_{i0}}{\sum_{i=1}^{n} A_{i0}}$ is the area-weighted sum of the in-channel

routing length of subbasin i from the subbasin i outlet to the drainage area outlet of interest. For scheme s, there are m subbasins within the evaluated drainage area and $j = 1,2,\cdots,m$ represents subbasin indices. A_{js} is the area of subbasin j in scheme s, and $\frac{\sum_{j=1}^{m} A_{js}L_{js}}{\sum_{j=1}^{m} A_{js}}$ is the area-weighted sum of the in-channel routing length of subbasin j from the subbasin j outlet to the drainage area outlet of interest. The total area of the drainage area is $A = \sum_{i=1}^{n} A_{i0} = \sum_{j=1}^{m} A_{js}$.

The calculation of the in-channel routing length difference between schemes is best described in Figure 2-1 below with a visual example. The example in Figure 2-1 demonstrates the in-channel routing length difference (ΔL_s) between scheme 0 and scheme s as the difference in the thick routing arrows between the two discretization options. For example, in scheme 0, flows from headwater subbasins 1, 2 and 3 are all routed in the main channel of subbasin 7 for 2 km. In comparison, with the coarser discretization scheme s, the flows from this region of the watershed (subbasins 1, 2 and 3 in scheme 0) are no longer routed in-channel for this distance and thus treated as a discretization error. A similar error occurs for the subarea including subbasins 4, 5 and 6. In our metric, in-channel routing length errors are computed for subbasin outlets of interest and in this example, the 'outlet' is the site of interest in Figure 2-1. If all flows reaching the outlet had a 2 km shorter in-channel routing length in scheme s versus scheme 0, then ΔL_swould be 2 km at the outlet. This is not typically the case and so the representative change in routing length, ΔL_s, must account for this using area-weighting.

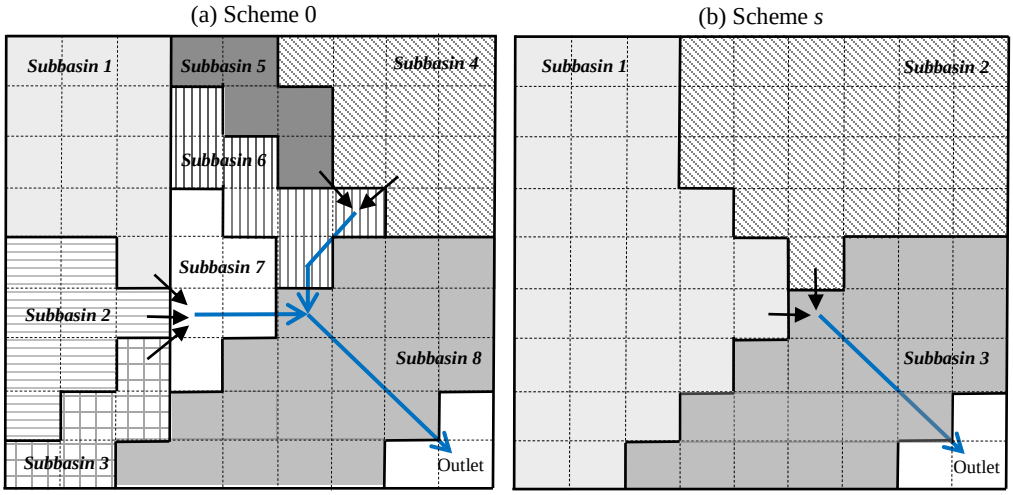

Figure 2-1. Subbasin discretization error depiction for a candidate discretization scheme (Scheme s) compared to the reference (most finely detailed) scheme (Scheme 0). Each cell is 1 km × 1 km. Small arrows indicate flow directions out of headwater subbasins, while large thicker arrows indicate in-channel routing and routing direction in other basins.

2.2.1.2. HRU Discretization Error Metric

As explained before, the information loss from spatial discretization is due to the diminished representation of spatial data content between a candidate discretization scheme and the original, fully detailed, input data

layer. To quantify the relevant (case study specific) information loss derived from HRU discretization, the dominant hydrologic processes should first be identified by considering the modeling purpose, physiographic characteristics and management measures within the watershed. These dominant processes can be linked to dominant hydrologic model input variables derived from map inputs which will be used to evaluate information losses. For example, in rainfall-runoff modeling, if infiltration is identified as a critical process then the most relevant variables to compute information losses for can be hydraulic conductivity and/or available water content.

For a drainage area above an outlet, assume there are *n* HRUs in the reference scheme (scheme 0), and *m* HRUs in the evaluated discretization (scheme *s*), and thus $n \geqslant m$. In order to effectively consider the spatial pattern changes between the two schemes, the evaluated scheme layer needs to be overlaid with the reference scheme layer using vector overlay tools (e.g., union) for vector maps or raster overlay tools (i.e., weighted overlay) for raster maps in ArcGIS (ESRI, 2014). After overly, each polygon or cell of the output possesses both the evaluated and reference scheme HRU properties. Assume there are v polygons (cells) ($u = 1, \dots, v$) of the output. HRU discretization error metrics are designed to go through each polygon (cell) and measure the relative error of variable change between scheme *s* and scheme 0. Two different *a priori* discretization error metrics corresponding to nominal (categorized) and quantitative (continuous) data are developed.

Figure 2-2 shows an example of the overlay comparison process required for computing HRU discretization errors for an example subbasin, corresponding to subbasin 1 of scheme 0 in Figure 2-1, and uses a nominal variable (land cover) as an example. In scheme 0, there are four different land covers scattered over the entire subbasin (Figure 2-2a), but only two land covers remain in the coarser scheme *s* (Figure 2-2b). After overlay and property comparison, four cells show a property change as highlighted in Figure 2-2c, in which one cell of coniferous forest turns into deciduous forest, and one cell of coniferous forest and two cells of pasture turn into crop. The information loss due to recategorization is considered as a discretization error, expressed in terms of recategorized area (i.e., 4 km^2 in this example). For quantitative variables, the only difference is the absolute values of the property changes are utilized as shown in the following equations.

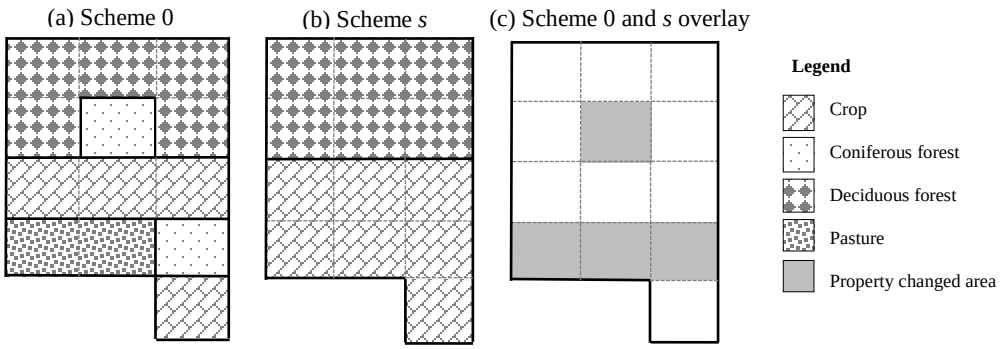

Figure 2-2. Example overlay process required for computing HRU discretization errors within a delineated subbasin. Above example uses land cover as a nominal variable of interest.

For nominal input variables (e.g., soil and land cover), the relative error equals to the sum of areas with property change from scheme 0 to scheme s divided by the total drainage area.

$$\delta_{HRUs} = \frac{\sum_{u=1}^{v} \Delta_u A_u}{\sum_{u=1}^{v} A_u} \tag{2-2}$$

$$\Delta_u = \begin{cases} 0, & if\ variable\ prorpty\ is\ unchanged\ relative\ to\ the\ reference\ scheme \\ 1, & if\ variable\ property\ is\ changed\ relative\ to\ the\ reference\ scheme \end{cases} \tag{2-3}$$

where δ_{HRUs} is relative error (0-1) of the evaluated scheme *s* describing the proportion of the drainage area where the variable property is changed and thus incorrect relative to the original spatial data. A_u is the area of the u^{th} polygon (cell) of the overlay output.

For quantitative input variables (e.g., hydraulic conductivity and available water content), the relative error equals to the area-weighted sum of the absolute values of input variable differences of all polygons (cells) between scheme s and scheme 0 divided by the area-weighted mean input variable value of scheme 0 within the drainage area. It is expressed as:

$$\delta_{HRUs} = \frac{\sum_{u=1}^{v} w_u |x_{us} - x_{u0}|}{\sum_{i=1}^{n} w_i x_{i0}} \tag{2-4}$$

where δ_{HRUs} is relative error (0-1) of the evaluated scheme s indicating the level of absolute input variable value change relative to the mean value of scheme 0. x_{us} and x_{u0} are the input variable values of the u^{th} polygon (cell) in scheme s and scheme 0, respectively. w_u is the area weight of the u^{th} polygon (cell) of the total drainage area. It is calculated by:

$$w_u = \frac{A_u}{\sum_{u=1}^{v} A_u} \tag{2-5}$$

where $\sum_{u=1}^{v} A_u$ is the total area of the evaluated drainage area.

The absolute value operation utilized in Equation (2-4) is to properly track all spatial heterogeneity changes once the input variable property differs from the original spatial input data. In other words, compensation effects (two errors cancelling each other) are not allowed.

2.2.2. Sensitivity of Hydrologic Model Simulation Results to Discretization Error Metrics

To validate the impact of the *a priori* error metrics on hydrologic model simulation results, multiple hydrologic models were built (one for each candidate discretization scheme). The only difference between these models exists in discretization. We chose to build our simulation models for different discretization levels in the Raven hydrologic modeling framework (Craig and Raven Development Team, 2018)**.** All the models are semi-distributed with two buckets and simulate water transfer between soil (upper and lower layers) and atmosphere through a series of hydrologic processes. The models simulate on an hourly time step and in-channel routing is based on a non-linear level pool routing approach using Manning's equation. Specific details of the hydrologic model are provided in Appendix A2-1 of this thesis.

Similar to discretization error metrics, hydrologic simulation results are assessed relative to a reference simulation result. The reference simulation result is obtained by running the hydrologic model with the reference discretization scheme (scheme 0). Recall that the reference discretization scheme fully retains the information of the original spatial input data or has the finest plausible discretization. All other model simulation results are compared relative to the reference result using error indices such as the peak flow rate error, the peak flow timing error, and the cumulative flow volume error. The peak flow rate error is computed as the absolute peak flow rate difference between scheme s and scheme 0 divided by the peak flow rate of scheme 0. The peak flow timing error is the time of peak flow occurrence with scheme 0 minus the time of peak flow occurrence with scheme s. The cumulative flow volume error is the absolute cumulative flow volume difference between scheme s and scheme 0 divided by the cumulative flow volume of scheme 0. Non-zero values for these indices are the direct result of different discretization choices.

The relationship between discretization errors and model errors is estimated by the Spearman's rank correlation coefficient (r_s) which ranges from -1 to +1. The objective of this analysis is to validate that changes in our proposed error metrics indeed impact hydrologic model simulation results. Note that our analysis necessarily avoids the issue of model calibration and validation decisions confounding the analysis. A future larger scale, multi-basin study would be required to properly validate the role discretization errors have in terms of their net impact on model predictive accuracy.

2.2.3. Discretization Decision-Making Approach

This research demonstrates one of many ways modelers can utilize the proposed *a priori* error metrics by using them within a structured two-step approach to watershed discretization decision-making. The two-step approach is applicable to both subbasin and HRU discretization decisions and is described in the following two sections.

2.2.3.1. Subbasin Discretization Decision-Making Approach

- Step 1: Select a subbasin scheme from candidate discretization schemes (Candidacy step).

Candidate subbasin schemes would typically first be generated by placing subbasin outlets at the sites of interest within the watershed (e.g., gauge stations and/or reservoirs) and at stream junctions, with subbasin boundaries determined using standard terrain analysis algorithms. The subbasin boundaries will vary depending on stream network resolution. Here, we generate the stream network and junctions based on a flow accumulation threshold as done in ArcSWAT (Winchell, et al., 2007). Other approaches to junction generation could be used, for example, truncating the stream network based upon Strahler stream order. The relationships between the flow accumulation threshold and the coarseness of the stream network are monotonic – as the accumulation threshold increases, stream network becomes less detailed and fewer subbasins are included. In this step, typically users should vary the spatially consistent flow accumulation threshold (uniformly applied for the entire watershed) and assess the resulting routing length errors.

A routing length error threshold (referred to as the preliminary error threshold) is then specified to select a subbasin scheme from candidates. The selected scheme meets the criteria that all sites of interest satisfy the preliminary error threshold at the minimum discretization complexity cost (i.e., the total number of subbasins) among all candidate schemes. Setting the preliminary error threshold to a very large value would function to select the most coarsely defined candidate scheme among the candidates.

- Step 2: Refine subbasin discretization for the areas with extreme discretization errors (Polishing step).

This step is used to refine the candidate subbasin discretization selected in Step 1 for the areas with the most extreme discretization errors. It can also be used to focus on minimizing discretization errors at modeler-specified critical sites of interest where smaller discretization errors are desired for some reason. Functionally speaking, this step is optional. If utilized, this step involves specifying a second, stricter routing length error threshold (referred to as extreme error threshold) and requires the stream junction locations of other finer resolution candidate schemes. Given a subbasin scheme from Step 1, the complete process of Step 2 is demonstrated by Figure 2-3.

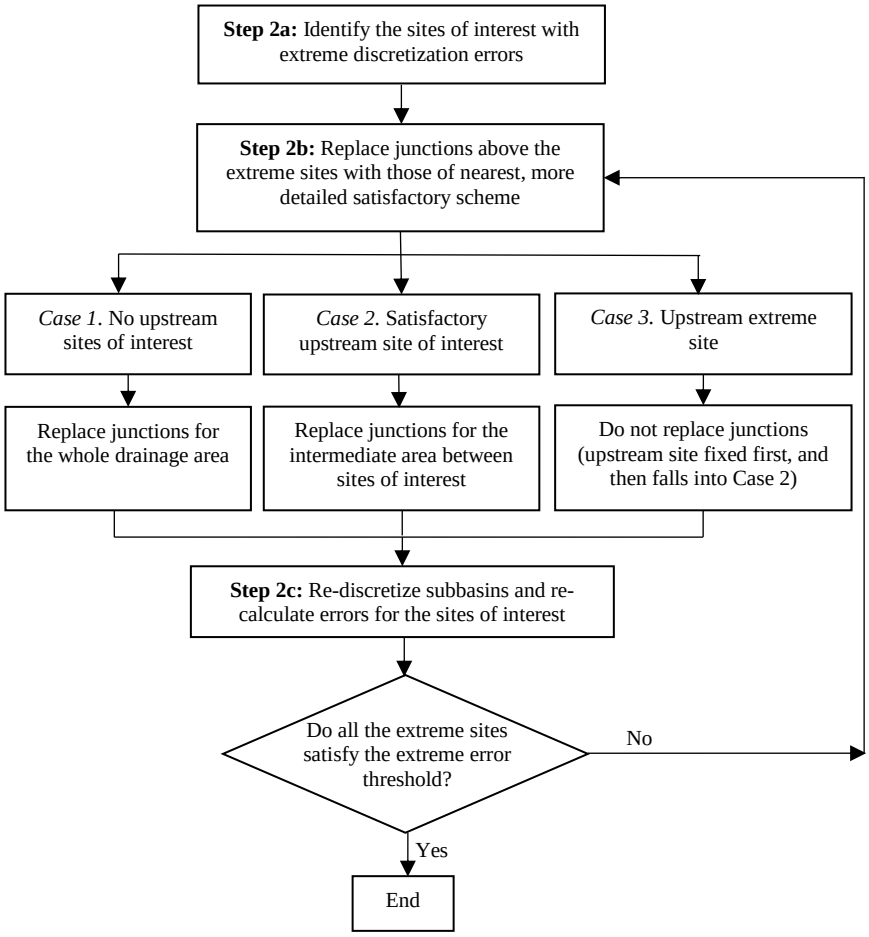

Figure 2-3. Flow chart of subbasin discretization decision-making approach Step 2: Polishing. Note that this flow chart applies to HRU discretization decisions as well as described in Section 2.2.1.2.

- Step 2a: Identify the sites of interest with extreme discretization errors.

  Sites of interest with discretization errors not satisfying the extreme error threshold are identified. These sites are referred to as extreme sites.

- Step 2b: Replace junctions in the upstream refinement areas of extreme sites with those of the nearest, more detailed satisfactory discretization scheme.

  There are three cases in identifying the upstream refinement area for each extreme site:

  - Case 1. If the extreme site has no upstream sites of interest, increased resolution of the stream network is applied to the whole drainage area above the extreme site.
  - Case 2. If the extreme site has satisfactory (non-extreme) upstream site(s) of interest, increased network resolution is only applied to the intermediate area between upstream site(s) of interest and the extreme site.
  - Case 3. If the extreme site has an upstream extreme site(s), the network is not refined. What will happen in this case is that the upstream extreme site(s) will first be refined (under Case 1) and then

in a future discretization refinement iteration, the intermediate area(s) will only be refined if the new discretization error(s) for the site in question remains extreme (the extreme site will be re-categorized into Case 2).

Once the upstream refinement area is determined, replace the junctions within it with those of the nearest more detailed satisfactory scheme. More detailed alternate candidate schemes would typically be available from the candidacy selection step (Step 1) but if not, the modeler would be required to generate one or more detailed schemes (e.g., by decreasing the flow accumulation threshold). It is worth explaining the reason why there is no need to replace junctions for the extreme sites of Case 3. In Case 3, the influence of the upstream refinement on the downstream error metric result is unclear unless the new errors are recalculated. If the extreme site located downstream can take the advantage of upstream refinement and obtain a satisfactory error result without junction replacements, this will be the most cost-effective solution in terms of model complexity.

- Step2c: Re-discretize subbasins and re-calculate errors for the sites of interest.

  In order to get the systematic upstream-downstream flow path relation among subbasins, re-discretize the watershed with the updated junctions and re-calculate the error metric results. The detailed re-discretization processes are provided in the Appendix A2-2 of this thesis.

  - If the new error metric results in all the previously extreme sites are satisfactory (less than the extreme error threshold), adopt these junctions. *Step 2* ends.
  - If some extreme sites do not satisfy the extreme error threshold, return to *Step 2b*.

  Iterate *Step 2b* and *Step 2c* until all the extreme sites are satisfactory.

Because the polishing step introduces non-uniformity to the discretization scheme (i.e., the refined areas have finer subbasin discretization than the non-refined areas), we refer to this discretization scheme as a non-uniform scheme.

2.2.3.2. HRU Discretization Decision-Making Approach

Similar to subbasin discretization decision-making, modelers can also choose an appropriate HRU discretization following the two-step decision-making approach outlined in Section 2.2.3.1. *Step 1* is selecting a uniform HRU scheme from candidates based on some predefined uniform HRU discretization preliminary error threshold(s). As with subbasin discretization, the candidate HRU discretization schemes should each be based on some uniform level of detail across the watershed. As described in Section 2.2.3.1, we identified candidate HRU schemes by varying an HRU size threshold, below which the small HRUs in that subbasin are merged and replaced with more dominant HRU types. Again, the relationship between this size threshold and the model complexity is monotonic. Unlike the subbasin discretization step, there

may be multiple hydrologic model input variables for which a modeler wishes to compute HRU discretization errors. In this case, the metric results of multiple input variables can be treated equally or assigned different weights based on their importance in decision-making.

Step 2 is polishing HRU discretization. The only difference from subbasin discretization refinement is that, in *Step 2b*, HRUs can be directly replaced without junction replacement. *Step 2c* simply involves merging all resultant HRUs into an output layer and re-calculating errors for the sites of interest.

2.3. Results

This study is conducted in the Grand River watershed in southwestern Ontario, Canada. With drainage area of 6704 km^2, the Grand River flows south to Lake Erie and is mainly covered by agricultural land. The applications are presented in two sections. Section 2.3.1 shows the application of the subbasin discretization error metric, and Section 2.3.2 shows the application of the HRU discretization error metric.

2.3.1. Subbasin Discretization Error Metric Application

2.3.1.1. Candidate Subbasin Discretization Schemes

In this study, subbasins were represented in subwatershed format and derived from $10m \times 10m$ digital elevation model (DEM) data. Subbasins were discretized based on the ArcSWAT (Winchell et al., 2007) flow accumulation threshold approach as described in Section 2.2.3.1. Research shows that, reducing the flow accumulation threshold below 0.5% of the maximum flow accumulation does not improve model performance but complicates remaining preprocessing, whereas increasing it significantly above 1% might lead to performance ramifications (Djokic, 2008). According to these findings, we took the percentage of the maximum flow accumulation across the entire watershed as the subbasin discretization threshold and assumed 0.5% as the minimum flow accumulation threshold value. Therefore, twelve candidate subbasin schemes were generated corresponding to twelve successively increasing flow accumulation thresholds. The detailed subbasin discretization results are listed in Table 2-1.

Table 2-1. Candidate subbasin discretization schemes

Scheme	Flow accumulation threshold (%)	Number of subbasins
0	0.5	130
1	0.6	110
2	0.7	100
3	0.8	94
4	0.9	92
5	1.0	90
6	2.0	60
7	3.0	46
8	5.0	44
9	6.0	40
10	10.0	38
Max	Only sites of interest	32

Scheme 0 was defined as the reference scheme because subbasin discretization with threshold 0.5% is the finest scheme of all the candidates and we assume the channel information loss between the real full channel scheme (i.e., one channel for each DEM cell) and scheme 0 is irremediable. Scheme Max only used the 32 sites of interest as subbasin outlets. The 32 sites include 24 gauge stations, 7 dams, and the watershed outlet, and their detailed information has been listed in Table 2-2.

Table 2-2. Details of 32 sites of interest and their drainage areas

Site of interest	**Site name**	**Drainage area (km^2)**	**Site of interest**	**Site name**	**Drainage area (km^2)**
1	02GA041	66	17	02GA015	565
2	Luther Dam	45	18	02GA038	313
3	02GA014	654	19	Laurel Creek Dam	31
4	02GA039	272	20	02GA024	59
5	Shand Dam	775	21	02GA047	757
6	02GA016	776	22	02GA048	2477
7	Conestogo Dam	559	23	Shades Mill Dam	96
8	02GA028	564	24	02GA018	536
9	02GA040	178	25	02GA003	3490
10	Woolwich Dam	60	26	02GA010	1028
11	Guelph Dam	241	27	02GB006	157
12	02GA034	1148	28	02GB007	384
13	02GA031	40	29	02GB001	4784
14	02GA023	113	30	02GB008	378
15	02GA029	226	31	02GB010	170
16	02GA006	769	32	Watershed outlet	6704

This chapter considers all 32 sites of interest as locations where the discretization error metrics will be assessed. Each drainage area is the combined total upstream area draining to the site as illustrated in Figure 2-4. For instance, drainage area 3 is defined to include subbasins 1, 2 and 3.

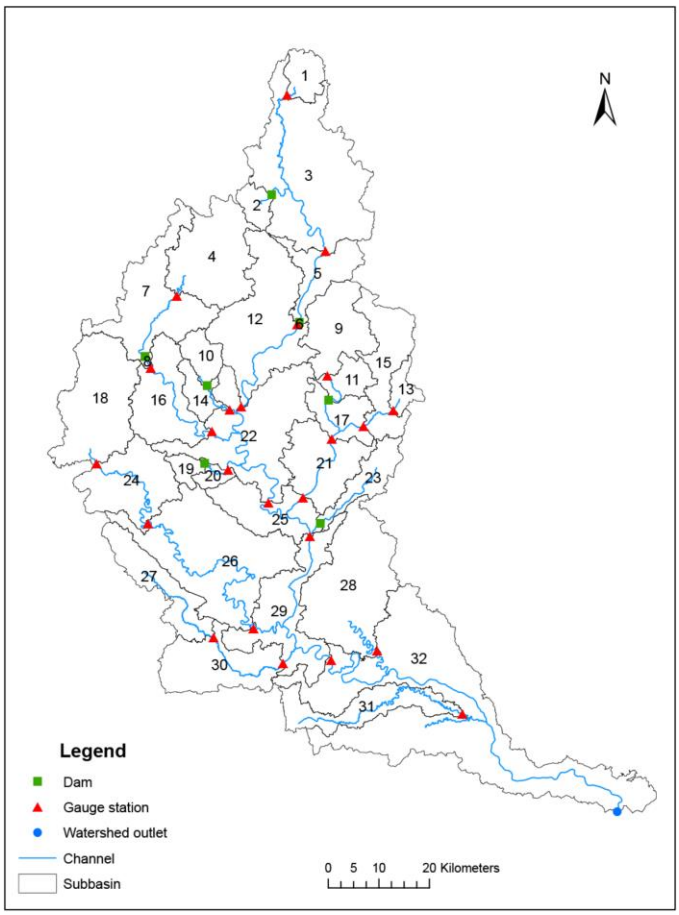

Figure 2-4. The 32 sites of interest in the Grand River watershed defining the subwatersheds where discretization errors are computed. The map also corresponds to the coarsest subbasin discretization (scheme Max) mentioned later.

2.3.1.2. Subbasin Discretization Error Metric Results

In distributed hydrologic modeling applications, most of the time modelers will only pay attention to the information loss at the sites of interest. Moreover, it is unnecessary to analyze error metric results for the sites above which candidate subbasin discretizations are always as fine as the reference one because in this situation the error metric result is always zero. As a result, this work limited the subbasin error metric results analysis to the 32 sites as introduced in Table 2-2 and then excluded the 13 sites whose upstream subbasins do not change from scheme 0 to scheme Max. The remaining 19 sites for analysis are sites 3, 5, 6, 7, 8, 9, 11, 12, 16, 17, 18, 21, 22, 24, 25, 26, 28, 29, 32, and their error metric results for the twelve discretization schemes were computed. For brevity, only the results from nine representative subbasin schemes are shown in Figure 2-5.

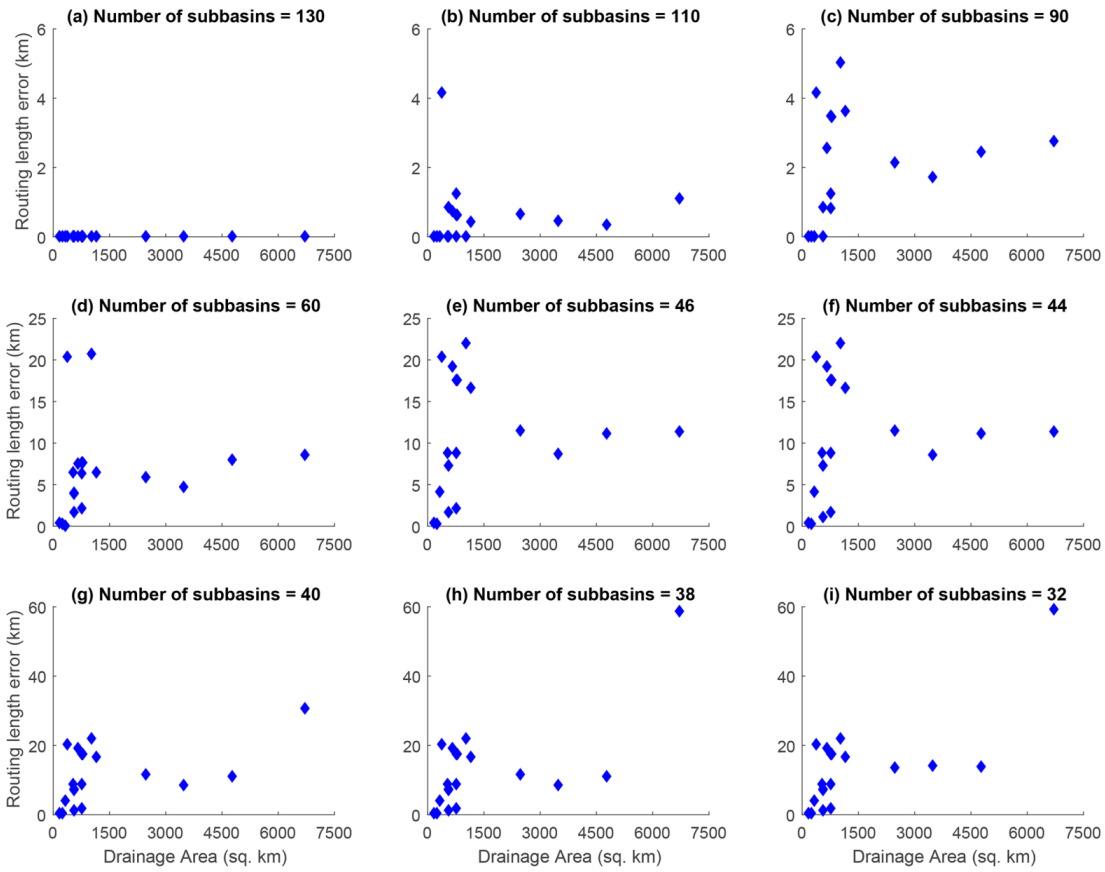

Figure 2-5. Subbasin error metric results of the 19 sites under nine representative subbasin schemes from Table 2-1 (with 130 to 32 subbasins). Each scheme is identified by the number of subbasins in subtitle and the errors are computed for the subwatershed areas draining to the 19 sites of interest being analyzed.

In each subplot of Figure 2-5, the routing length errors of the 19 sites are plotted versus their drainage areas. Figure 2-5a shows that when discretization is detailed at the reference scheme level, no error exists. Then in Figure 2-5(b-i), as subbasin discretization gets coarser, the number of subbasins within a drainage area decreases, and the routing length error increases. This is reflected by the ranges of error values of Figure 2-5(b-i). Moreover, in each subplot, the downstream sites with the largest drainage areas typically have intermediate error values rather than the maximum value of all the errors at the 19 sites of interest. For example, moving downstream in the Grand River watershed, site 22 (drainage area 2477 km^2), site 25 (drainage area 3490 km^2), and site 29 (drainage area 4784 km^2) all get intermediate error values for all the subbasin schemes. This trend can be explained by the fact that, for in-channel routing, the downstream error integrates all its upstream errors in an area-weighted fashion (see Equation (2-1)), so the drainage area outlet is not necessarily the point that has the largest information loss. This implies that if modelers are concerned about the multi-site discretization quality or the multi-site hydrologic model performance,

multiple sites rather than just the watershed outlet are worth considering in subbasin discretization evaluation.

2.3.1.3. Sensitivity of Hydrologic Model Simulation Results to Subbasin Discretization Error Metric

To access the sensitivity of model simulation results to the proposed subbasin error metric, we built twelve hydrologic models corresponding to all the subbasin schemes of Table 2-1, in which their only difference is subbasin discretization and the connectivity between subbasins. We focused the analysis on a short period (Jan 4 – Jan 20, 2008) of peak or near peak measured flows over the last ~15 year period across the Grand River watershed. The reference simulation result corresponds to the model using the reference discretization scheme (scheme 0 of Table 2-1) and all simulation model results were compared relative to the reference result using the peak flow rate error and peak flow timing error.

Taking the watershed outlet as an example, Figure 2-6 summarizes the relationship between the *a priori* routing length error metric and the hydrologic model error indices where each data corresponds to one of the eleven candidate subbasin discretization schemes. As the routing length error increases, both model error indices increase (almost monotonically) to practically significant levels. Correlation (r_s) between the routing length error and the peak flow rate error is 0.99, and correlation (r_s) between the routing length error and the peak flow timing error is also 0.99. This strong correlation is observed for the majority of sites of interest (e.g., considering the correlation between the routing length error and the peak flow rate error, 15 sites show r_s values of 0.8 or more).

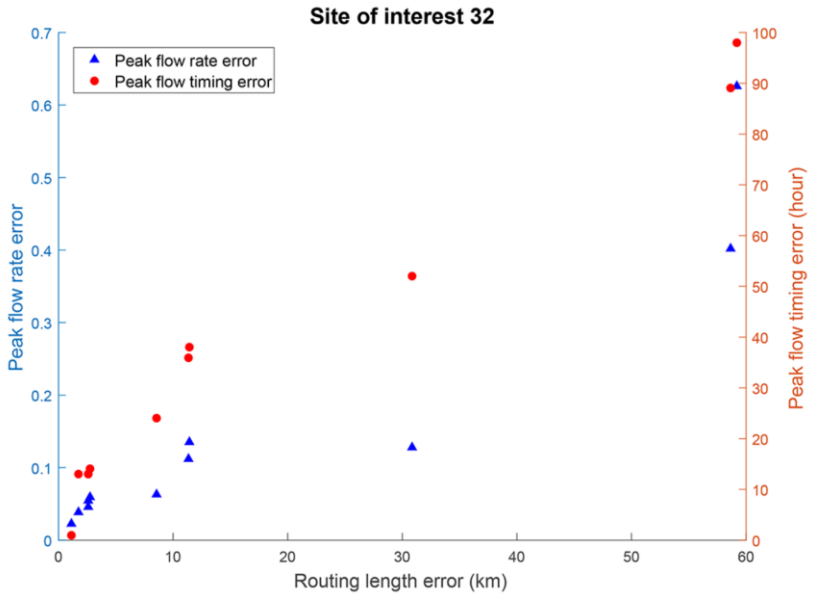

Figure 2-6. Peak flow rate errors (left y-axis) and peak flow timing errors (right y-axis) versus subbasin discretization error metric results under eleven candidate subbasin schemes at the watershed outlet.

2.3.1.4. Subbasin Discretization Decision-making

Based on the error metric results of all candidate subbasin discretization schemes, we applied the two-step decision-making approach to get an appropriate subbasin discretization scheme. It was assumed that all of the 19 sites of interest are equally important, and 21 km is selected as the preliminary routing length error threshold. The subjective value of 21 km was selected for demonstration purposes and based on balancing travel time error implications (assuming a reference velocity of 1 m/s) and computational complexity (limiting number of subbasins).

- Step 1: Select a subbasin scheme from candidate discretization schemes

  Scheme 6 (number of subbasins=60) was chosen as the uniform threshold subbasin scheme because the error metric values of all the 19 sites of scheme 6 are satisfactory (less than 21 km) and the number of subbasins is the minimum among all the satisfactory schemes (schemes 1-6).

- Step 2: Refine subbasin discretization for the areas with extreme discretization errors
  - Step 2a: Identify the sites of interest with extreme discretization errors (extreme sites).

    The extreme error threshold was 10.7 km, defined as the 90^{th} percentile of the error distribution of scheme 6, and the resultant extreme sites that have the highest 10% errors were sites 26 and 28, which are highlighted in Figure 2-7a and Figure 2-7b.

  - Step 2b: Replace junctions in the upstream refinement areas of extreme sites with those of the nearest, more detailed satisfactory discretization scheme. Specifically, different sites have different upstream refinement areas:

    Case 1: Site 28 has no upstream sites of interest, thus junction replacement is applicable to the whole drainage area above site 28. Since the error of site 28 in scheme 5 is 4.2 km (less than 10.7 km), scheme 5 is the nearest satisfactory scheme compared with scheme 6.

    Case 2: Site 26 has a satisfactory upstream site of interest, site 24, so junction replacement only takes place in the intermediate area between the site 24 and site 26. Since the error of site 26 in scheme 5 is 5.0 km (less than 10.7 km), scheme 5 is also the nearest satisfactory scheme relative to scheme 6.

  - Step2c: Re-discretize subbasins and re-calculate errors for the sites of interest.

After re-discretization, the subbasin compositions within the upstream refinement areas were changed to the new more detailed subbasins as shown in Figure 2-7c. Meanwhile, the total number of subbasins for the entire Grand River watershed increased from 60 to 66. The routing length errors of sites 26 and 28 became satisfactory (less than 10.7 km as shown in Table 2-1).

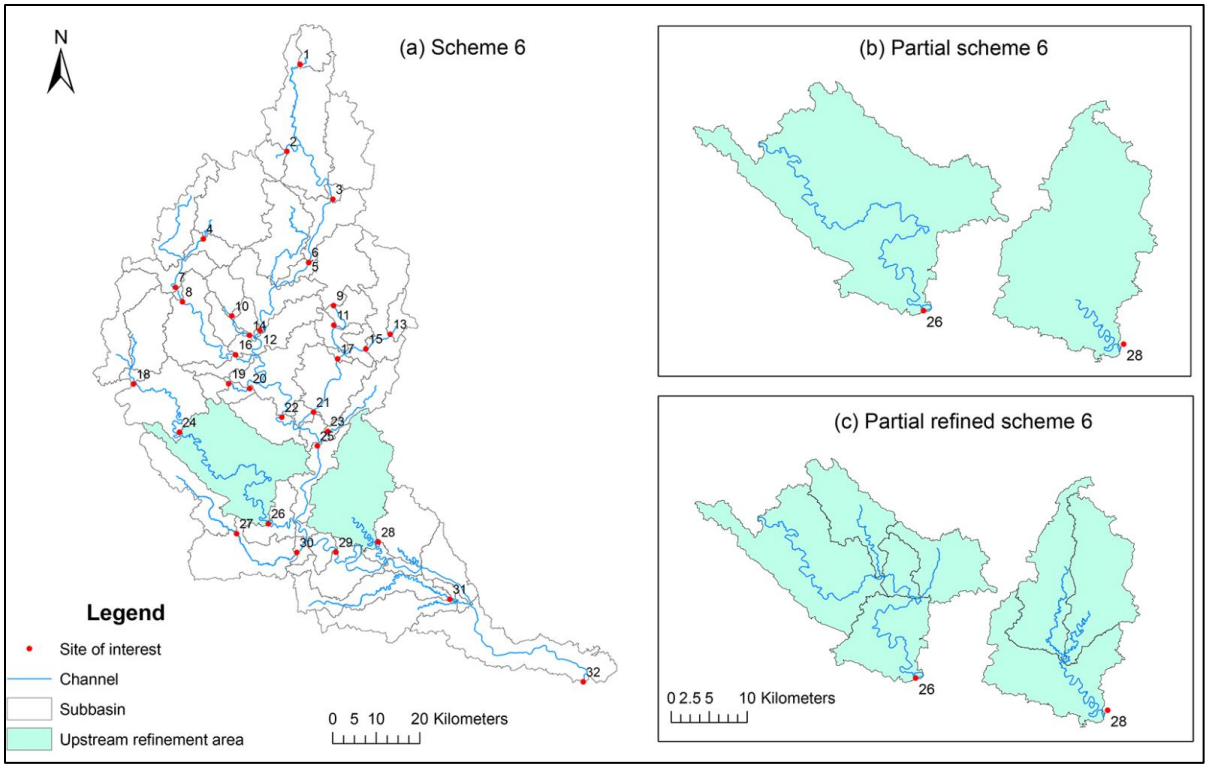

Figure 2-7. Example subbasin discretization refinement. Each site of interest is labelled with its site number from Table 2-2. Panel a shows subbasin discretization of scheme 6 where extreme errors are observed in sites 26 and 28. For the identified junction replacement areas, panel b and panel c show the detailed subbasin discretizations under scheme 6 and the refined scheme 6, respectively.

Table 2-3 shows the routing length errors of scheme 6, refined scheme 6, and scheme*. Scheme* has the same number of subbasins as refined scheme 6 but was generated with a uniform flow accumulation threshold of 1.55%. In addition to the purposeful reduction of routing errors at the two extreme sites, Table 2-3 also shows the substantial error decrease of all the associated downstream sites (e.g., sites 29 and 32) in refined scheme 6. Moreover, comparing refined scheme 6 with scheme*, the error mean and standard deviation of refined scheme 6 are lower than those of scheme*. This indicates that the refined subbasin discretization better represents the in-channel routing structure than the uniform discretization under the same number of computational (subbasin) units.

Table 2-3**.** Subbasin discretization error metric results for three subbasin discretization schemes. Scheme 6 is based on a flow accumulation threshold of 2.0%, while Scheme* is based on a threshold of 1.55%. Sites of interest that are discretized the same way under all three schemes are not included. Highlighted errors for refined scheme 6 are lower than corresponding errors in one or both of Scheme 6 and Scheme*. Note that site 32 corresponds to the watershed outlet.

Site of interest	Scheme 6		Refined scheme 6		Scheme*	
	Number of subbasins	Error (km)	Number of subbasins	Error (km)	Number of subbasins	Error (km)
7	4	4.0	4	4.0	6	0.9
8	5	4.0	5	4.0	7	0.9
16	6	6.4	6	6.4	2	3.9
17	7	1.7	7	1.7	8	4.1
18	3	0.0	3	0.0	7	1.7
22	25	5.9	25	5.9	27	5.2
25	37	4.7	37	4.7	39	4.2
26[e]	5	20.7	9	5.0	5	20.7
28[e]	1	20.4	3	4.2	3	4.2
29	47	8.0	51	4.6	49	7.6
32	60	8.5	66	5.2	66	7.0
Error mean		7.7		4.2		5.5
Error St. deviation		6.8		1.8		5.5
Error 90th percentile [f]		10.7				

[e] denotes an extreme site under scheme 6 based on exceeding the 90th percentiles of the error metric. The subbasin discretization within this site's drainage area is refined based on *Step2*.
[f] The 90th percentile computed based on errors across all 19 sites considered (see Section 2.3.1.2).

2.3.2. HRU Discretization Error Metric Application

2.3.2.1. Candidate HRU Discretization Schemes

In this chapter, HRU is discretized after subbasin, and an HRU is defined as the unique combination of subbasin and soil and land cover categories. Subbasin input was one of the candidate subbasin schemes generated in Section 2.3.1 (Table 2-1). Soil spatial input data was from the Canadian Soil Information Service (CANSIS) available from Agriculture and Agri-Food Canada (2013) and subdivided into fourteen classes in terms of soil profile. Each soil profile except water is built up by a unique soil horizon combination from three mineral horizons A, B, C, and an organic horizon O. Soil profile A-B-C covers more than 70% of the Grand River watershed (Table 2-4a). Land cover spatial input data was from Canada's National Land Cover Database available from Natural Resources Canada (2014) and subdivided in seven classes, in which cropland is dominant across the watershed (Table 2-4b). Soil and land cover inputs used here are vector coverages derived from 1:20,000 to 1:60,000 scale county-level soil maps attained from CANSIS and 1:50,000 scale land cover maps from Canada's National Land Cover Database.

Table 2-4. Grand River watershed (a) Soil classes (b) Land cover classes and their percent coverage of the watershed.

Soil class	Area percentage (%)
A B C	72.29
Water	8.17
A B BC C	7.76
O B	3.40
A B	3.32
A AB B C	2.51
A B AB B C	1.13
AB	0.64
C	0.27
A C	0.25
O C	0.12
A	0.09
C A C	0.04
A AB C	0.03

Land cover class	Area percentage (%)
Annual Cropland	40.70
Perennial Cropland and Pasture	33.91
Deciduous Forest	14.74
Urban	5.43
Mixed Forest	2.98
Wetland	1.24
Water	1.00

Table 2-5. Candidate HRU discretization schemes with two subbasin discretization schemes (90 and 32 subbasins)

HRU Scheme	HRU size threshold (% of subbasin area)	Number of HRUs	
		Number of subbasins=90	Number of subbasins=32
0	0	2706	1232
1	1	852	333
2	2	625	234
3	3	502	190
4	4	433	156
5	5	385	135
6	6	346	121
7	7	318	109
8	8	290	99
9	9	252	90
10	10	234	84
Max	One HRU per subbasin	90	32

The map obtained by the overlay (union) of the above subbasin, soil, and land cover layers defines the reference HRU scheme (scheme 0). Since the map algebra union operation usually leads to a very fragmented set of sliver HRUs, these sliver HRUs can be suppressed for aggregation based on certain HRU size threshold. Here, the HRU size threshold was defined as the HRU area percentage of its affiliated subbasin. The HRU whose area percentage is less than the size threshold was merged with its neighboring HRU sharing the longest border within the same subbasin. In order to investigate the influence of the subbasin discretization input on HRU discretization, we chose two representative subbasin schemes (scheme 5 and scheme Max) as subbasin inputs to discretize HRUs, respectively. The generated candidate HRU schemes are listed in Table 2-5. For the HRU candidates under 90 subbasins, HRU scheme 0 (number of HRUs=2706) is the reference scheme; while for the HRU candidates under 32 subbasins, HRU scheme 0 (number of HRUs=1232) is the reference scheme. Each reference scheme retains 100% of land cover and soil data as the reference scheme does not eliminate/aggregate sliver HRUs. In HRU scheme Max, each

subbasin is represented by the dominant HRU. Table 2-5 shows that subbasin discretization choice significantly affects HRU discretization complexity (i.e., the number of HRUs) because under the same HRU size threshold, the number of HRUs with 90 subbasins input is always two to three times more than that with 32 subbasins input.

2.3.2.2. HRU Discretization Error Metric Results

In this chapter, infiltration and evapotranspiration were identified as the two dominant hydrologic processes, thus vertical hydraulic conductivity (Kz), available water content (AWC), and land cover were defined as the key hydrologic model input variables of interest. For each soil class of Table 2-4a, Kz and AWC are the weighted harmonic mean values of the Kz and AWC of its soil horizon components. The detailed soil horizon information is available from Agriculture and Agri-Food Canada (2013). The area-weighted mean values of Kz and AWC of the entire watershed are 0.9 cm/h and 12.6% (except the soil class water), respectively. Figure 2-8 demonstrates the discretization error metric results of Kz, AWC, and land cover at the watershed outlet versus HRU size thresholds. As the HRU size threshold increases, discretization gets coarser, meanwhile the relative errors of all the three variables increase. However, the same HRU size threshold imposes different impacts on the information losses of different variables. For example, under the same HRU schemes (before HRU scheme Max), the relative errors of Kz and land cover are always similar in magnitude (Figure 2-8a, Figure 2-8c), but the relative errors of AWC are comparatively smaller (less than 0.05 in Figure 2-8b). In HRU scheme Max, land cover error jumps to 0.55, while Kz and AWC errors are 0.15 and 0.05, respectively. Land cover errors jump to much higher values compared to Kz and AWC because some merged HRUs only experience land cover changes but no change in soil properties. The results show that, unsurprisingly, relative discretization errors are positively correlated with HRU size threshold.

The subbasin discretization decision between 90 or 32 subbasins has a substantial influence on HRU discretization complexity (100%-200% increase in number of HRUs seen in Table 2-5). However, this decision does not make a big difference for information loss as Figure 2-8 indicates that two error metric results (AWC and land cover) of the three variables are almost identical and only one variable (Kz) obtains slightly different error metric results under different subbasin inputs.

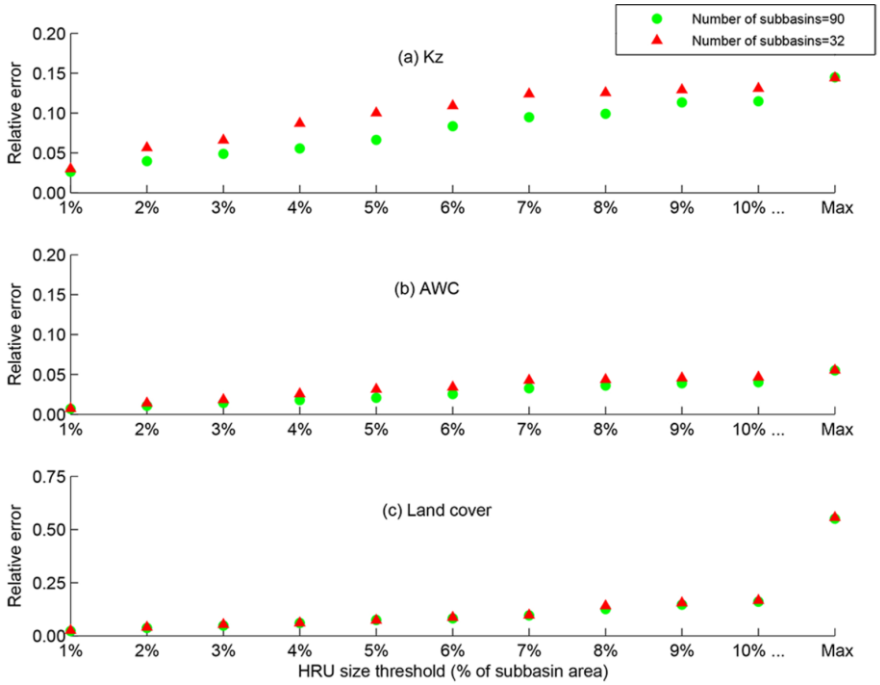

Figure 2-8. HRU discretization error metric results of three variables under all the candidate HRU schemes at the watershed outlet. HRU scheme is identified by the HRU size threshold (1%, 2%… Max). Subbasin scheme is identified by the number of subbasins (subbasin scheme 5 has 90 subbasins, and subbasin scheme Max has 32 subbasins).

Figure 2-8 supports how a modeler might make decisions based on a single watershed outlet. However, in distributed or semi-distributed modeling applications where distributed watershed responses are of interest, discretization errors should be assessed at multiple sites beyond just the outlet. Figure 2-9 is a more robust comparative approach than Figure 2-8 as it compares discretization errors at all the 32 sites of interest across the Grand River watershed under subbasin scheme 5 (number of subbasins = 90). The interesting pattern in Figure 2-9 is that for all the three variables of interest (Kz, AWC, and Land cover), the largest discretization errors (and the highest variance) appear in the relatively small drainage areas, and as drainage area increases, errors approach some constant level. Therefore, while errors for the watershed outlet might be sufficiently small, they can be unacceptably large in some small upstream subbasins. Although results are not shown, this pattern persists across all HRU discretization levels.

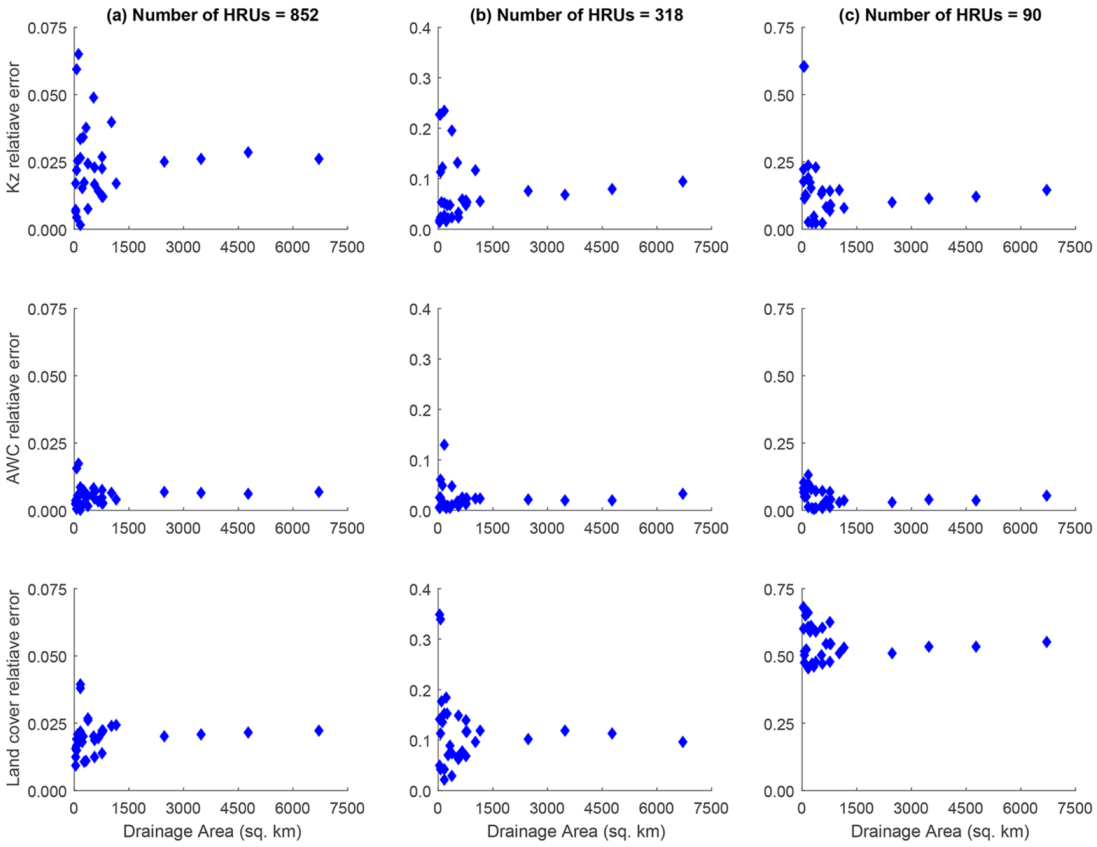

Figure 2-9. HRU error metric results of all the 32 sites of interest under three HRU discretization schemes for 90 subbasins. Each HRU discretization scheme is identified by the number of HRUs in subtitle, and each site is identified by its drainage area.

2.3.2.3. Sensitivity of Hydrologic Model Simulation Results to HRU Discretization Error Metrics

Similar to the sensitivity analysis in Section 2.3.1.3, we checked the sensitivity of hydrologic model simulation results to the proposed HRU error metrics based on twelve hydrologic models. These models correspond to all the HRU schemes under 90 subbasins of Table 2-5, and the only difference between these models is the property of HRUs. The model output with scheme 0 (Number of HRUs =2706) is the reference simulation result in model errors calculation. The peak flow rate error and cumulative flow volume error were computed.

Figure 2-10 presents the relationship between the *a priori* HRU discretization error metrics and the model error indices where each data corresponds to one of the eleven candidate HRU discretization schemes at the watershed outlet (subbasin 32 outlet). The two model errors are plotted versus the HRU discretization errors of Kz, AWC, and land cover. Clearly, both model errors indices monotonically increase with the HRU discretization errors of the three variables. Correlations (r_s) between the three HRU discretization errors (Kz, AWC, and land cover) and the peak flow rate error are all 0.99. Similarly, correlations (r_s) between the three HRU discretization errors and the cumulative flow volume error are also 0.99. This strong

correlation also shows up in most sites of interest (e.g., considering the correlation between the land cover error metric and the peak flow rate error, 23 sites show r_s values of 0.8 or more).

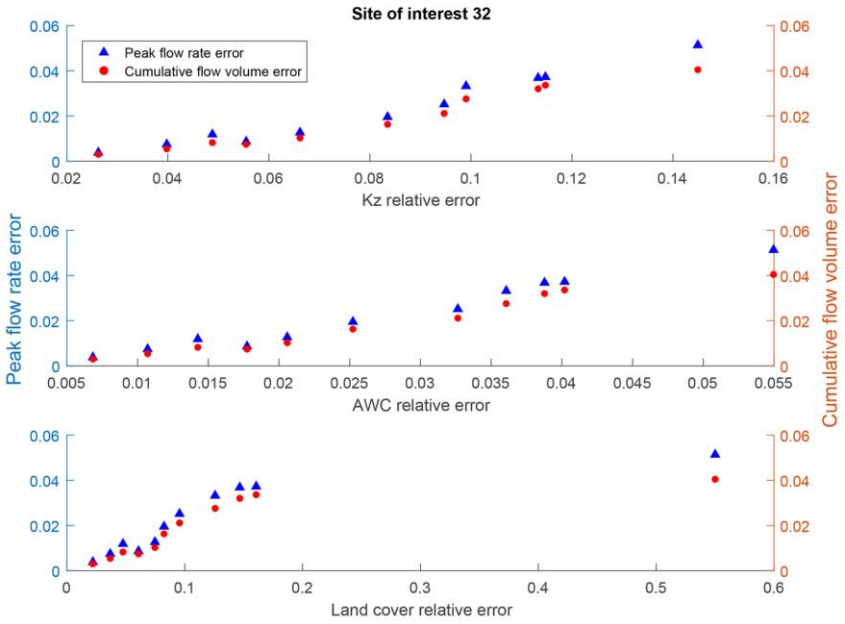

Figure 2-10. Relative errors of peak flow rate (left y-axis) and relative errors of cumulative flow volume (right y-axis) versus HRU discretization error metric results of three variables under all the candidate HRU schemes at the watershed outlet.

Figure 2-11 provides a more complete description of hydrologic simulation responses by plotting all sites of interest model errors against their drainage areas under the same three representative HRU schemes of Figure 2-9. The upstream sites with relatively small drainage areas obtain a high variance of model errors, in which some of them have three or more times errors than their downstream sites. This observation appears in both the peak flow rate error and the cumulative flow volume error and is consistent with results from Figure 2-9 (indicating the largest HRU discretization errors are also associated with small drainage areas).

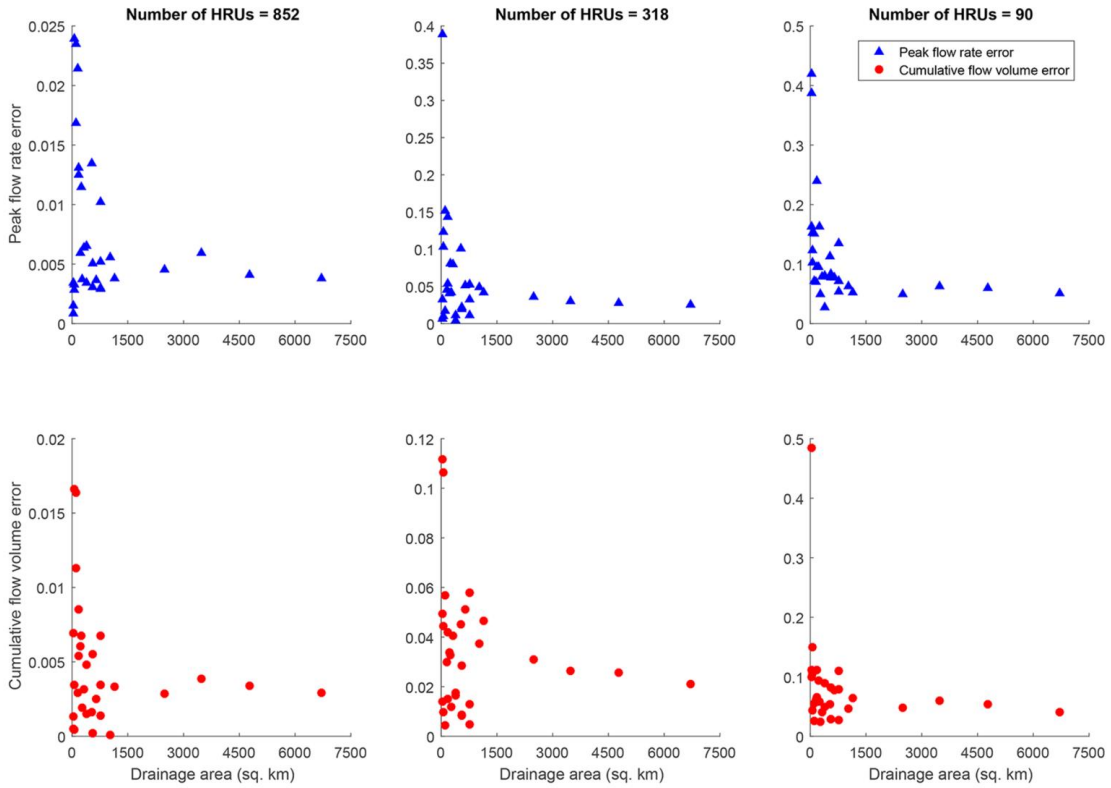

Figure 2-11. Peak flow rate errors (upper panels) and cumulative flow volume errors (lower panels) of all 32 sites of interest under the three representative HRU discretization schemes. Each scheme is identified by the number of subbasins in subtitle, and each site is identified by its drainage area.

2.3.2.4. HRU Discretization Decision-making

An alternative to the commonly applied uniform discretization framework demonstrated above is to make discretization decisions differently in different parts of the watershed, in response to excessively high error metric values. This relies on the two-step HRU discretization decision-making approach (see Section 2.2.3.2) where different subareas can use different HRU delineation thresholds. To demonstrate, assume subbasin scheme 5 (number of subbasins=90) is the subbasin input for HRU discretization; all the 32 sites of interest and all the three hydrologic model input variables of interest are equivalently important in HRU scheme decision-making; and 0.40 is the preliminary error threshold for all the three variables. The subjective value of 0.40 was selected for demonstration purposes only and selected with the goal of generating a modest number of HRUs relative to the range of candidate HRU discretizations.

- Step 1: Select an HRU scheme from candidate discretization schemes.

  HRU scheme 10 (number of HRUs=234) was chosen as the uniform HRU scheme because the relative errors of all the sites of interest in scheme 10 are satisfactory (less than 0.40) and the number of HRUs is minimum among all the satisfactory schemes (schemes 1-10).

- Step 2: Refine HRU discretization for the areas with extreme discretization errors.
  - Step 2a: Identify the sites of interest with extreme discretization metrics (extreme sites).

    The extreme error thresholds were defined as the 90th percentiles of the error distributions of the three variables (0.13, 0.06, and 0.31 for Kz, AWC, and land cover, respectively). As a result, the sites having the highest 10% Kz, AWC, or land cover errors were identified as the extreme sites of interest to have their discretization refined (i.e., sites 10, 13, 15, 19, 20, 23, 28, and 31). The drainage areas above sites 19 and 20 are highlighted for discretization refinement demonstration in Figure 2-12.

  - Step 2b: Replace HRUs in the upstream refinement areas of extreme sites with those of the nearest, more detailed satisfactory discretization scheme, in which different extreme sites have different upstream refinement areas.

    Case 1: Extreme sites 10, 13, 19, 23, 28, and 31 have no upstream sites of interest, so the HRU refinement areas cover the whole drainage areas above these sites.
    Case 3: Extreme sites 15 and 20 have extreme upstream sites of interest (site 13 and 19, respectively), and it is unnecessary to replace HRUs across their entire drainage area in the first refinement iteration.

    Then, only Case 1 sites had the HRUs within them replaced with those of the nearest more detailed satisfactory HRU scheme relative to HRU scheme 10. This replacement step was applied independently for each of the refined extreme sites. The detailed HRU replacement results are summarized in Table 2-6.

Table 2-6. HRU replacement results for the extreme sites in the first HRU discretization refinement iteration

HRU replacement area	**Original number of HRUs**	**Nearest satisfactory HRU Scheme**	**New number of HRUs**
Drainage area above site 10	2	HRU scheme 1	10
Drainage area above site 13	4	HRU scheme 9	5
Drainage area above site 19	4	HRU scheme 5	8
Drainage area above site 23	2	HRU scheme 7	5
Drainage area above site 28	10	HRU scheme 5	16
Drainage area above site 31	3	HRU scheme 3	7

  - Step 2c: Merge all resultant HRUs into an output layer and re-calculate discretization errors for the sites of interest.

    After the first refinement iteration, it was found that all Case 1 sites became satisfactory, but some errors of sites 15 and 20 were still extreme. Therefore, *Step 2b* needed to be repeated to replace HRUs for sites 15 and 20.

- Step 2b: Replace some or all HRUs in the upstream refinement areas of extreme sites with those of the nearest, more detailed satisfactory discretization scheme.

  Case 2: Extreme sites 15 and 20 became Case 2 after the first refinement iteration, and thus the intermediate area between sites 13 and 15 and the intermediate area between sites 19 and 20 were identified as the HRU replacement areas in the second refinement iteration. The HRUs within these intermediate areas were replaced with those of the nearest satisfactory schemes relative to HRU scheme 10. The detailed HRU replacement results are summarized in Table 2-7.

Table 2-7. HRU replacement results for the extreme sites in the second HRU discretization refinement iteration

HRU replacement area	Original number of HRUs	Nearest satisfactory HRU Scheme	New number of HRUs
Intermediate area 13-15[i]	10	HRU scheme 8	16
Intermediate area 19-20	1	HRU scheme 4	3

[i] Intermediate area 13-15 means the intermediate area between site 13 and site 15. This also applies to intermediate area 19-20.

- Step 2c: Merge all resultant HRUs into an output layer and re-calculate errors across the watershed.

  After the second iteration, the errors of all the originally identified extreme sites became satisfactory, thus the refinement process ended and this scheme was the refined HRU scheme. Figure 2-12 provides a visual comparison for the HRUs before and after refinement of the drainage areas of sites 19 and 20 under subbasin scheme 5 (90 subbasins).

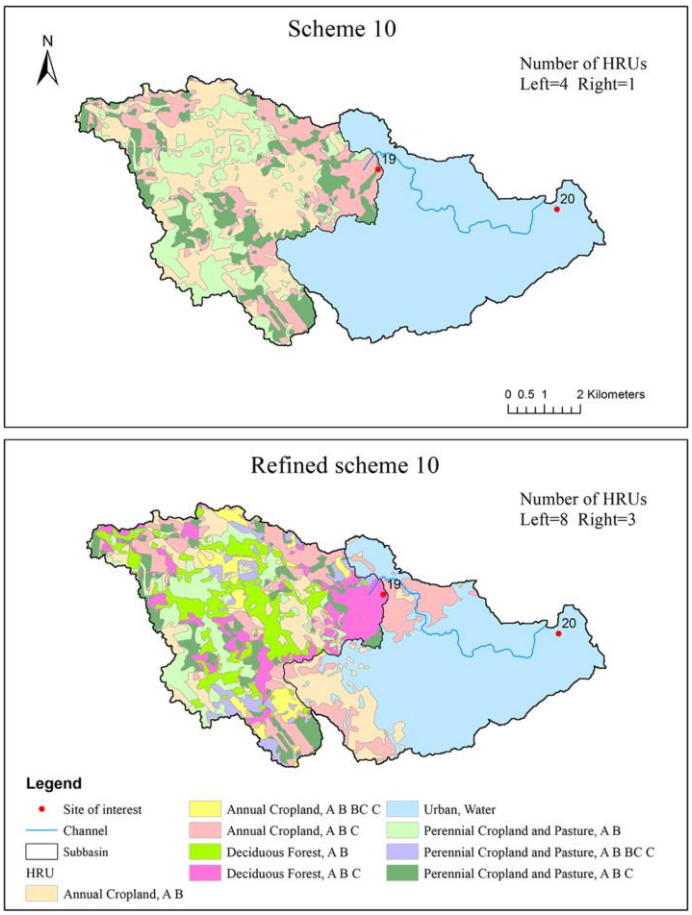

Figure 2-12. Example HRU discretization refinement. For the identified junction replacement areas of sites 19 and 20, upper and lower panels show the detailed HRU discretizations under scheme 10 and the refined scheme 10 (using 90 subbasins), respectively.

Table 2-8 shows the error metric results of HRU scheme 10, refined scheme 10, and scheme*. Scheme* has the same number of HRUs as refined scheme 10 (number of HRUs=271) but was generated with a uniform HRU size threshold of 8.4%. Through discretization refinement, the extreme errors identified in *Step 2b* above are reduced, and the discretization quality for site 32 representing the entire watershed is also improved for all the three variables. Moreover, the discretization error means and standard deviations of the three variables across all the sites of interest of the refined scheme also decrease in contrast with scheme*. Therefore, the non-uniform discretization functions to retain more input data information than the uniform discretization under the same discretization complexity. In addition, to get a sense of how non-uniform the HRU discretization is in refined scheme 10, the average HRU sizes of the uniformly discretized areas and the non-uniformly discretized areas within the watershed were respectively calculated as 28.8 km^2 and 13.6 km^2. The latter is more than 50% smaller than the former, which means the HRUs within the refined areas are obviously finer than those of the uniformly discretized areas.

Table 2-8**.** Discretization error metric results for three HRU discretization schemes (using 90 subbasins). Scheme 10 is based on an HRU size threshold of 10%, while Scheme* is based on a threshold of 8.4%. Note that site 32 corresponds to the watershed outlet and sites of interest 1, 9, 11, 18, 24, 27 and 30 are not included because they are discretized the same way under all three schemes. Highlighted errors for Refined scheme 10 are lower than corresponding errors in one or both of Scheme 10 and Scheme*.

Site of interest	Scheme 10				Refined scheme 10				Scheme*			
	Number of HRUs	Kz	AWC	Land cover	Number of HRUs	Kz	AWC	Land cover	Number of HRUs	Kz	AWC	Land cover
2	3	0.06	0.02	0.25	3	0.06	0.02	0.25	4	0.02	0.01	0.22
3	17	0.06	0.03	0.20	17	0.06	0.03	0.20	22	0.06	0.02	0.14
4	7	0.02	0.01	0.12	7	0.02	0.01	0.12	8	0.02	0.01	0.07
5	20	0.06	0.03	0.22	20	0.06	0.03	0.22	25	0.05	0.02	0.17
6	22	0.06	0.03	0.22	22	0.06	0.03	0.22	28	0.05	0.02	0.17
7	15	0.02	0.01	0.10	15	0.02	0.01	0.10	16	0.02	0.01	0.07
8	18	0.02	0.01	0.10	18	0.02	0.01	0.10	19	0.02	0.01	0.07
10[e]	2	0.11	0.06	0.11	10	0.06	0.02	0.02	2	0.11	0.06	0.11
12	29	0.06	0.03	0.19	29	0.06	0.03	0.19	35	0.06	0.02	0.15
13[e]	4	0.02	0.00	0.31	5	0.01	0.01	0.22	6	0.02	0.00	0.14
14	4	0.12	0.05	0.14	12	0.10	0.03	0.08	4	0.12	0.05	0.14
15[e]	14	0.03	0.01	0.36	21	0.03	0.01	0.26	20	0.03	0.01	0.30
16	24	0.07	0.01	0.09	24	0.07	0.01	0.09	26	0.06	0.01	0.08
17	33	0.05	0.02	0.27	40	0.05	0.02	0.23	40	0.05	0.02	0.23
19[e]	4	0.23	0.03	0.35	8	0.13	0.01	0.17	4	0.23	0.03	0.35
20[e]	5	0.23	0.03	0.34	11	0.13	0.02	0.17	5	0.23	0.03	0.34
21	36	0.07	0.03	0.26	43	0.07	0.03	0.23	43	0.07	0.03	0.23
22	89	0.08	0.02	0.15	103	0.08	0.02	0.15	102	0.08	0.02	0.12
23[e]	2	0.13	0.07	0.39	8	0.03	0.01	0.07	5	0.05	0.02	0.18
25	141	0.08	0.03	0.19	168	0.08	0.02	0.17	166	0.08	0.02	0.16
26	19	0.12	0.03	0.16	19	0.12	0.03	0.16	20	0.12	0.02	0.16
28[e]	10	0.21	0.06	0.13	16	0.12	0.03	0.08	11	0.20	0.05	0.13
29	176	0.09	0.03	0.19	203	0.09	0.02	0.17	207	0.09	0.02	0.16
31[e]	3	0.24	0.13	0.04	7	0.06	0.04	0.05	3	0.24	0.13	0.04
32	234	0.12	0.04	0.16	271	0.10	0.03	0.15	271	0.11	0.04	0.14
Error mean		0.09	0.03	0.20		0.07	0.02	0.15		0.09	0.03	0.16
Error Std. deviation		0.07	0.03	0.09		0.04	0.01	0.07		0.07	0.03	0.08
Error 90th percentile [f]		0.13	0.06	0.31								

[e] denotes an extreme site under scheme 10 based on exceeding the 90th percentiles of the error metrics. The HRU discretization within this site's drainage area is refined based on *Step2*.

[f] The 90th percentile computed based on errors across all 32 sites of interest.

2.4. Discussion

2.4.1. Reference Discretization Scheme Determination

The reference scheme is defined as a scheme that fully retains the information of the original spatial input data or, in special cases, the finest plausible discretization. The implication is that the modeler is interested in quantifying how much information is lost relative to the reference scheme.

Our subbasin reference scheme was defined based on a subjective flow accumulation threshold to determine reference main channel lengths. Alternatively, the real full flow path information (i.e., flow path of each cell in the DEM) can be obtained, for example, by the flow length tool of ArcGIS, and the corresponding discretization could be used as the reference scheme. For HRU discretization, our reference scheme retained all raw input spatial data and thus avoided any subjective decisions. Alternatively, the reference HRU scheme could be subjectively defined as a discretization that addresses some numerical and topological problems if this discretization is the one that modelers will practically apply and want relative errors computed against (Sanzana et al., 2013). While absolute discretization error metric values will be impacted by what can be a subjective reference scheme choice, the relative error values among candidate discretization choices should not change significantly.

2.4.2. Discretization Error Metrics

The subbasin discretization error metric estimates the in-channel routing length difference relative to the reference scheme. An improved approach would be to instead consider travel time error by using a reference flow velocity. Assuming the reference flow velocity as a constant is an easy and common method for practical purposes (De Lavenne et al., 2015; Rigon et al., 2016; Sanzana et al., 2013). However, this assumption is questionable as typical watersheds have faster velocity upstream reaches compared to lower velocity downstream reaches. As such, a travel time error could instead be based on a spatially variable reference flow velocity. In addition, other available roughness, geometry, channel slope information can be incorporated into the routing information loss estimation by being linked to flow velocity (e.g., with Manning's equation).

The *a priori* metrics (both nominal and quantitative) are able to provide directly meaningful descriptions on information loss because they explicitly characterize how much area or value of the hydrologic model input variable is changed after discretization. Moreover, they are unique as compared to the existing *a priori* metrics in the way they identify the property change. Haverkamp et al. (2002) , Booij (2003), and Dehotin and Braud (2008) all define discretization information loss as the overall statistics difference between the candidate scheme and the reference scheme, failing to conduct the cell-by-cell comparison with the original spatial input data. In contrast, the metrics proposed here correspond one-to-one with information loss during discretization. The overlay comparison process is a straightforward

technique and is feasible for both raster and vector spatial input data. This enables other hydrologic variables of interest such as land surface slope and aspect to be analyzed with similar *a priori* discretization error metrics.

Figure 2-5 and Figure 2-9 show that discretization errors are highest in smaller upstream subbasins while Figure 2-11 shows that hydrologic model error indices (for peak flow and cumulative volume) are also highest in smaller upstream subbasins. This observation explains to some extent the modeling difficulties associated with small upstream subbasins in semi-distributed modeling. Although past studies such as Andersen (2001) and Tuppad (2006) have attributed the poor relative performance of calibrated upstream gauges to calibrated downstream locations to factors like more uncertain rainfall, our observation reveals that relatively poor performance in the smaller upstream subbasins of our case study can be expected since the discretization errors (and hydrologic model error indices) of these subbasins have a high variance and can be three times larger than the corresponding errors of the downstream larger drainage areas in the uniform threshold discretization framework. This demonstrates the utility of multi-site discretization evaluation in distributed modeling applications, and also suggests that a non-homogenous approach to watershed discretization decision-making would be beneficial.

Our HRU error metric approach for nominal input data did not disaggregate the individual area changes of the different categories of nominal data. For example, the area changes of the crop land or the deciduous forest land. However, modelers may only care about the area change of a certain category in their watershed (e.g., the change from forest to suburban may be of consequence but the change from wetland to swamp may be immaterial). Although not demonstrated in this work, the error metric for nominal variables (Equation (2-2)) can be readily modified to assess the relative error of a specific category of nominal input data.

2.4.3. Variations to Discretization Approach

Our approach first generated all candidate subbasin schemes by ArcSWAT, and then generated candidate HRU schemes by sliver area aggregation. Any other candidate schemes generated by different discretization methods can also be evaluated by our proposed *a priori* discretization error metrics. Additional checks could be added to the subbasin discretization step, for example, checking the reference scheme against additional data such as orthophotos or hydrographic survey maps. Another variation is related to handling the small but potentially important sliver HRUs in HRU discretization simplification. For instance, in periurban areas where the land cover is very heterogeneous, some small HRUs can be meaningful in terms of hydrology, thus such HRUs should be protected from merging in discretization simplification. One approach to preserve these key HRUs is to introduce an importance factor that would artificially increase the areas of key HRUs so that they would exceed the HRU discretization threshold used to aggregate small HRUs.

2.5. Conclusions

This chapter proposed *a priori* discretization error metrics that can estimate the information loss for any candidate discretization scheme. These metrics do not require model simulation, are independent of any specific modeling software, provide modelers with directly interpretable information on discretization quality, and allow for multi-site and multi-variable discretization evaluations prior to model development. In particular, the subbasin error metric provides the first attempt at quantifying the routing information loss from discretization; the HRU error metrics improve upon the existing *a priori* metrics in variable property change identification by the overlay comparison process. The proposed error metrics are straightforward to understand and easy to recode into the preprocessing of any semi-distributed hydrologic models and the fully distributed models using spatial input data aggregation. As a potential application of the proposed *a priori* discretization error metrics, a two-step decision-making approach was formulated to help modelers to get the appropriate subbasin and HRU discretization schemes, respectively. The approach does not only allow choosing a traditional spatially uniform-threshold discretization scheme based on the modeler-defined error threshold(s), but also enables compressing extreme errors to satisfy the modeler-specified discretization error targets.

These *a priori* discretization error metrics were applied to the discretization of the Grand River watershed. Results indicated that the discretization-induced information loss as measured by our discretization error metrics monotonically increases as discretization gets coarser. Hydrologic modeling under candidate discretization schemes validates the strong correlation between our discretization error metrics and model predictions (peak flow rate, cumulative flow and peak flow timing). Discretization evaluation results show that model accuracy moving from larger downstream locations to smaller upstream locations would be expected to increase since the largest discretization errors and highest error variability occur in smaller upstream locations. This pattern is also evident when changes in hydrologic model outputs were used in place of HRU discretization error metrics. Finally, results show that the common and convenient approach of applying uniform discretization across the watershed domain performs worse compared with the metrics-informed non-uniform discretization approach as the latter is able to preserve more input data information using the same number of computational units. However, the influence of non-uniform discretization on hydrologic model outputs should be further studied using a number of hydrologic models and case studies.

In applying the proposed *a priori* discretization error metrics to discretization decision-making, accounting for input forcing data (e.g., precipitation and temperature) resolution is also an important future consideration. This will require comparing the spatial and temporal distributions of the forcing input data under candidate schemes and those under the reference scheme. Beyond the application in discretization decision-making, future studies can utilize the discretization error metrics in other ways. For instance, the

discretization error metrics may be useful in trying to account for the uncertainty induced by watershed discretization decisions which is commonly ignored. Furthermore, the discretization error metrics should prove useful even when they are not calculated *a priori* in that they could serve an important role in diagnosing the causes of model prediction errors in distributed modeling applications.

Chapter 3.
Climate Ensemble-Based Calibration Framework for Optimizing Prediction Bound Quality

This chapter is a replicate of a submitted manuscript to *Advances in Water Resources* in with minor changes to increase its consistency with the body of the thesis and to avoid redundant material. Changes were only made in the Summary (abstract). References are unified at the end of the thesis.

Liu, H., Tolson, B.A., 2019. Climate ensemble-based calibration framework for optimizing prediction bound quality. Submitted to Adv. Water Resour. (Under Review).

Summary

Several methods have been proposed to explicitly consider climate data uncertainty in hydrologic modeling. These methods usually use assumed statistical error models to perturb climate variables. Statistical error models are easy to construct but may not reflect real characteristics of the true climate. This work proposes the direct use of existing physically-based climate ensemble products in model calibration. Based on climate ensemble, this work formalizes a climate ensemble-based hydrologic model calibration framework. The framework uses existing climate ensembles as the informative and convenient prior estimates of the true climate and functions to filter out the poor-quality climate ensemble members during calibration. The framework can utilize any source of historical climate ensembles. For demonstration, the Gridded Ensemble Precipitation and Temperature Estimates dataset (Newman et al., 2015) is used to represent precipitation and temperature data uncertainty in this chapter. The framework performance is compared with an optimization-based calibration and an uncertainty-based calibration, both using measured climate data in calibration, for 30 synthetic experiments and 20 real case studies. Results show that the framework generates more robust parameter estimates, refines climate uncertainty estimates, reduces inaccurate flow predictions caused by poor-quality climate measurements, and improves the reliability of flow predictions. Results also show that the highest quality model predictions are only possible when the characterization of climate uncertainty is consistent between the calibration and validation periods (e.g., assuming calibration and validation period climates are either both uncertain or both deterministic).

Section 3.1 reviews existing model calibration approaches that explicitly incorporate forcing data uncertainty. Section 3.2 describes in detail the proposed climate ensemble based model calibration framework. Section 3.3 reviews the utilized climate ensemble product (i.e., Newman et al. (2015)) and explains the comparative calibration setups of the synthetic experiment and real case study. Section 3.4 shows the results of the synthetic experiment and real case study. Section 0 provides conclusions about the proposed calibration framework.

3.1. Introduction

Given the nature of hydrologic systems and the uncertainty in modeling processes, it has been broadly recognized that hydrologic models need to be applied with proper considerations of relevant hydrologic uncertainties (Beven and Young, 2013; Krzysztofowicz, 1999). There are four main sources of uncertainty in hydrologic modeling: data, model parameter, model structure, and initial and boundary condition (Abbaszadeh et al., 2018; Liu and Gupta, 2007). For hydrologic models, data can refer to climate data (e.g., precipitation, temperature, radiation, and wind speed), soil and land cover data, or system response data. Data uncertainty is caused by a myriad of factors, including measurement errors, sampling, interpolation and inversion from a sparse observation network (Beven and Young, 2013). Data uncertainty can be represented by a probability distribution function or an ensemble of the variable of interest. This chapter will focus on climate data uncertainty in hydrologic modeling and in particular will use an existing historical climate ensemble dataset to efficiently represent precipitation and temperature data uncertainty in model calibration. Since climate data is usually used as forcings or input to hydrologic models, it is also referred to as forcing data or input data in the literature.

In model calibration studies, it has been revealed that ignoring climate data uncertainty and assembling all hydrologic uncertainties (e.g., parameter, data, and model structure) into a single error term lead to biased parameter estimates and unreliable simulation bounds (Kavetski et al., 2002; Kuczera et al., 2006). As a result, many approaches have been proposed to explicitly account for climate data uncertainty in model calibration. These approaches typically use assumed statistical models to describe climate data uncertainty and determine the statistical model relevant parameters prior to or within model calibration. Most studies rely on climate observations and perturb the observation with a statistical error model to characterize data uncertainty. The typical error model is additive or multiplicative Gaussian with a predefined constant or proportional variance (Fuentes-Andino et al., 2017; Kavetski et al., 2002; Salamon and Feyen, 2009), while other probability distributions are also used to depict climate errors, such as the lognormal distribution and Gamma distribution (Del Giudice et al., 2016; Montanari, 2005; Renard et al., 2010, 2011). The other studies rely on the true climate and derive the uncertain climate from the truth. The true climate is usually described with empirical statistical models or stochastic process models, for example, the exponential function and hierarchical model. Then the uncertain climate is simply sampled from the true climate or derived by adding an assumed statistical error model to the truth (Balin et al., 2010; Blazkova and Beven, 2009; Huard and Mailhot, 2008).

Given the above climate data uncertainty estimation models, there are two main ways to determine the climate relevant parameters. The first and the most common way is to estimate them simultaneously with model parameters or at least in conjunction with model parameter estimation. For example, in Bayesian inference, the likelihood function is adjusted, and the climate relevant parameters are inferred together with

model parameters (Del Giudice et al., 2016; Huard and Mailhot, 2006, 2008, Kavetski et al., 2002, 2006b, Renard et al., 2010, 2011). In the generalized likelihood uncertainty estimation (GLUE), the climate relevant parameters are estimated with model parameters by being sampled in each model run (Blazkova and Beven, 2009; Montanari, 2005) or estimated conditioned on each behavioral model parameter set (Fuentes-Andino et al., 2017). The second way is to determine the climate relevant parameters before model calibration. For instance, Balin et al., (2010) performed a two-stage Bayesian inference where they first estimated stochastic process model parameters for the true rainfall by Bayesian inference and then generated realizations of observed rainfall by adding a pre-defined random error to the true rainfall. Hydrologic model parameters are estimated in Balin et al., (2010) using Bayesian inference conditioned on each rainfall realization and the approach accepts all rainfall realizations as plausible and then in prediction mode, the posterior parameter sets from each realization are aggregated and sampled from randomly to produce uncertainty bounds on the hydrograph. This unconditional use of all rainfall realizations could be problematic where the rainfall realization is substantially inconsistent with the actual rainfall measurements (or the true but unknown rainfall).

Although the assumed climate data uncertainty estimation models have been applied in many hydrologic model calibration studies, they have limitations from the model structure and parameter estimation perspectives. From the model structure perspective, these assumed statistical error models may not reflect the best estimates of the true climate. For example, the commonly used multiplicative stochastic error model has the inherent deficiency to quantify measurement uncertainty when no rainfall is recorded, which can be especially important in poorly gauged areas (Wright et al., 2017). Furthermore, these climate data uncertainty estimation models typically lack the consideration of the spatiotemporal correlation in generating climate data errors (e.g., see Fuentes-Andino et al., 2017; Kavetski et al., 2002; Renard et al., 2010; Salamon and Feyen, 2010, 2009, Vrugt et al., 2008, 2005). This simplified assumption may be invalid, especially in studies with dense rain gauge networks. From the parameter estimation perspective, estimating the climate relevant parameters with model parameters increases the dimensionality and thus the difficulty of the calibration problem (Ajami et al., 2007; Henn et al., 2015; Huard and Mailhot, 2008; Yen et al., 2015). Determining the climate relevant parameters before model calibration is often subjective and based on the literature or the modeler's experience (Huard and Mailhot, 2006; Salamon and Feyen, 2010).

The literature on likelihood-free or non-Bayesian model calibration techniques considering uncertainty is dominated by the GLUE methodology (Beven and Binley, 1992) and the more recent limits of acceptability variation to GLUE (Beven, 2006) which requires an explicit characterization of measured flow (or more generally, system response) uncertainty. As noted by Kavetski et al. (2018), the predictive uncertainty described by GLUE is virtually always a function of only model parameter uncertainty. The few GLUE rainfall-runoff model calibration studies considering calibration period climate input uncertainty

include Fuentes-Andino et al. (2017), Blazkova and Beven (2009), and Montanari (2005). Although it is simple to imagine using a climate ensemble within the GLUE framework, none of these studies do so and each of these studies instead utilizes an explicit but simplified climate uncertainty/error model. Fuentes-Andino et al. (2017) added rainfall multipliers for selected large flood events but first sampled behavioral hydrologic model parameters based on model runs using the deterministic climate observations in order to save computational time. This approach is problematic because it eliminates hydrologic model parameter sets from consideration that could have otherwise been behavioral under one or more of the sampled behavioral rainfall multipliers. Blazkova and Beven (2009) calibrated their model under a limits of acceptability approach comparing model outputs against summary information of the flow duration curve and the frequency characteristics of flood discharges, snow water equivalent and hourly and daily rainfall frequencies. They reported substantial difficulty in identifying an adequate number of behavioral parameter sets despite conducting over 600,000 model runs and they did not consider time series/residuals-based goodness-of-fit measures. The investigation of Montanari (2005) is limited to a synthetic calibration study. Montanari (2005) presented the rainfall uncertainty by five rain gauge weights and these weights were assigned pre-defined uniform distributions. The gauge weights, like the hydrologic model parameters, were randomly sampled for each hydrologic model run to find the behavioral hydrologic model parameter sets plus gauge weights.

Given these limitations of the climate data uncertainty estimation in model calibration, we propose the direct use of existing historical climate ensemble products to efficiently represent climate data uncertainty. The climate ensemble specifically refers to a geo-physically informed and more realistic climate ensemble representing climate data errors. We rely on the Newman et al. (2015) ensemble historical climate dataset as an example. Our purpose is to make use of the advances of meteorology and other relevant fields to improve the handling of climate data uncertainty in hydrologic model calibration, especially to avoid the construction of climate data uncertainty estimation models as part of the hydrologic model building process. For instance, in the context of Bayesian inference, climate ensemble products can replace assumed statistical error models and work as an informative and convenient prior estimate of the climate variables. Based on the climate ensemble, a hydrologic model calibration framework is proposed to explicitly account for climate data uncertainty in model calibration. Note that this research only considers model parameter and climate data uncertainty in model calibration: it does not include a statistical error model for hydrologic model prediction error and so the proposed calibration framework needs adjustments if it is to be adapted for use in calibration studies relying on formal Bayesian inference.

The specific goals of this chapter are to: (1) formalize a climate ensemble based calibration framework that explicitly accounts for climate data uncertainty in model calibration; (2) demonstrate how to use an existing ensemble dataset of historical climate to represent the uncertainty of precipitation and temperature;

and (3) use a large number of synthetic experiments and real calibration case studies to demonstrate the strengths of the proposed calibration framework in flow prediction and parameter estimation.

For demonstration, the Newman et al. (2015) ensemble dataset was used to represent the precipitation and temperature data uncertainty in this chapter. The Newman et al. (2015) dataset contains 100 historical realizations of precipitation and temperature for the contiguous United States, northern Mexico, and southern Canada. It provides spatially and temporally correlated climate uncertainty estimates and is straightforward to use for both lumped and distributed hydrologic models due to its gridded and ensemble attributes. This dataset is more complex but has much more realistic characterization of climate uncertainty compared to characterizations and error models in the various studies noted above that were built for rainfall-runoff modeling purposes only. An overview of the Newman et al. (2015) dataset is in Section 3.3.3. Note that historical ensemble climate products are inherently limited to the region they were developed for. Alone, they cannot be transferred or applied to a new region unless the ensemble climate generation methods are available for application in the new region.

Unlike the nearly standard approach with synthetic experimentation in past calibration studies where a single synthetic experiment is employed to test the performance of the proposed method (e.g., Kavetski et al., 2002; Renard et al., 2010; Vrugt et al., 2005; Montanari, 2005), we will repeat the synthetic experiment 30 times and base conclusions on a sample of synthetic tests. Real case study results will be presented for a sample of 20 different catchments. The synthetic and real case study results will show that the proposed climate ensemble based calibration framework outperforms the observed climate based calibration in parameter estimation and flow prediction without extra computational demand.

3.2. Methods

The climate ensemble based calibration framework was developed for the purpose of explicitly considering climate data uncertainty in model calibration process. Our development of this framework was motivated by an example of McIntyre et al. (2002). In their study, the system response observation data is assumed uncertain and replaced with an ensemble of response realizations. Each response realization is taken as the target response in Bayesian inference and generates maximum likelihood parameters. All the maximum likelihood parameters conditioned on all the response realizations are combined to get the converged posterior parameter distribution. In contrast, our framework focuses on climate data uncertainty.

3.2.1. Climate Ensemble Based Model Calibration Framework

Assume a deterministic hydrologic model is expressed as $\widehat{\mathbf{Y}} = H(\boldsymbol{\theta}, \mathbf{X})$, where $\widehat{\mathbf{Y}}$ is the simulated system response, H is the hydrologic model, $\boldsymbol{\theta}$ is the hydrologic model parameter vector, $\mathbf{X}$ is the input data. The input data only refers to climate data in this chapter. Both model parameters and input are uncertain. All

the other errors, including model structure errors and unaccounted data errors (e.g., streamflow) are not explicitly considered in this framework.

A climate ensemble can be used to describe the uncertainty in all climate inputs. Each ensemble member forms a complete set of climate forcings to drive the hydrologic model, so each member is at minimum a time series of multiple variables, but it can also be a spatiotemporal dataset with multiple variables. Like model parameters, there can be a prior and posterior distribution for the climate input to the model. The prior climate ensemble, $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M\}$ has a large number of ensemble members, M, which can be derived from existing climate ensemble products and can also include the observed climate. ϕ is the climate ensemble member number ($\phi = 1,2, \dots, M$) and the calibration problem now involves determining ϕ in addition to the hydrologic model parameters. The climate input for running a hydrologic model is defined by $\mathbf{X}_\phi$. In the simplest approach (as in our study), the prior distribution for ϕ is defined by assuming each member is equally probable, thus ϕ satisfies a discrete uniform distribution, $\phi \sim U(1, M)$. The posterior distribution of ϕ is estimated simultaneously with model parameters, and its corresponding climate ensemble members are used to approximate the posterior distribution of the climate variable. The climate ensemble member number (ϕ) is introduced to make the incorporation of climate uncertainty as parsimonious as possible.

The climate ensemble based calibration framework seeks to jointly identify behavioral model parameter sets and their corresponding behavioral climate realizations. A behavioral solution in this framework is defined by a vector of hydrologic model parameter values and a climate ensemble member number. The detailed joint identification process includes two steps.

Step 1. Sample model parameters conditioned on each sampled climate ensemble member.

Assume that the prior climate uncertainty is represented by M climate ensemble members. A climate ensemble member $\mathbf{X}_\phi$is randomly sampled from the given prior climate ensemble. Based on the sampled climate, $\mathbf{X}_\phi$, implement a model parameter estimation procedure and get C candidate model parameter solutions. Included in this step is the selection of a user-specified calibration objective function (which can also be a likelihood or pseudo-likelihood function). The minimum value of C is one. The above process is repeated for multiple random climate ensemble members (e.g., K climate samples) and this generates a total of $C_{total} = \sum_{k=1}^{K} C_k$ candidate behavioral solutions. Given that some of the members in the prior climate ensemble may be too inconsistent with the observed climate and thus generate streamflow simulations inconsistent with observations, the candidate behavioral solutions are the subjected to a filtering step to potentially eliminate candidate behavioral solutions.

Step 2. Identify the behavioral model parameter solutions (the filtering step).

The proposed calibration framework is designed to objectively identify behavioral solutions following Shafii et al. (2015). As such, the criteria to classify a candidate behavioral solution as behavioral (e.g., determination of a behavioral threshold) is objectively determined as part of the model calibration process. Shafii et al. (2015) demonstrated two objective approaches for behavioral solution identification and here we utilize their optimal criteria-aggregation-based approach. The idea of this approach is to simultaneously consider multiple aspects of the probabilistic prediction bounds (e.g., reliability and sharpness) and to choose among the several groups of candidate behavioral solutions. The candidate group of behavioral solutions that achieves the best overall performance of the prediction intervals in the model calibration period constitutes the behavioral solution ensemble.

The optimal criteria-aggregation-based approach first demonstrated in Shafii et al. (2015) aggregates multiple performance metrics into one aggregate metric but the approach is equivalent when only a single metric is defined as the objective function (e.g., RMSE between the simulated and the observed streamflow as utilized in this study). The detailed behavioral solution identification approach, assuming the objective function is to be minimized, is explained as below.

a) Sort all the sampled candidate behavioral solutions form Step 1 from the best to the worst according to their objection function values.
b) Determine candidate behavioral threshold values and their corresponding groups of solutions. Assume ten candidate behavioral thresholds are determined as the 10^{th} to 100^{th} percentiles of the objective function distribution at a 10% interval. Correspondingly, ten solution groups are formed. Each group includes all the candidate behavioral solutions whose objective functions are less than or equal to a candidate behavioral threshold. For example, when using an RMSE objective function, the first group includes all the candidate behavioral solutions whose RMSEs are strictly less than the 10^{th} percentile of the RMSE distribution. Therefore, the higher the behavioral threshold is, the more solutions the group includes.
c) For each candidate solution group, calculate the reliability and sharpness of its flow prediction intervals.

   Reliability is calculated as the percentage of the observed flow within the 95% prediction bounds. Sharpness, first introduced in Yadav et al. (2007), is calculated as the ratio of the 95% prediction bounds width to the 95% prediction bounds width obtained with the hydrologic model parameters ($\boldsymbol{\theta}$) and climate ensemble member number (ϕ) uniformly sampled from the feasible parameter ranges.

d) For each candidate solution group, calculate its distance-to-ideal (DTI) value.

DTI is calculated as the Euclidean distance between (R, S) and (R_{best}, S_{best}). R_{best} and S_{best} are the individual best reliability and sharpness values, respectively, of all the candidate solution groups and define the ideal point in objective space.

e) The candidate solution group with the minimum DTI is identified as the behavioral solution ensemble and the corresponding objective function threshold is the behavioral threshold. Given the objectivity of the behavioral identification process, the number of behavioral solutions and even the behavioral threshold can vary between calibration experiments.

Since both the model parameters and the climate input are uncertain, the non-behavioral parameter solutions can be due to either a poor-quality hydrologic model parameter set (far from the most likely parameter set due to a poor calibration algorithm result) or a poor-quality climate ensemble member (far from the true climate) or a combination of both. As such, the filtering step eliminates non-behavioral model parameter sets as well as poor-quality prior climate ensemble members from consideration in uncertainty propagation to model outputs. The number of the behavioral model parameter sets typically vary among different sampled climate inputs.

The behavioral model parameter sets are used to approximate the posterior distributions of model parameters. The corresponding climate ensemble member numbers (ϕ) in the behavioral solutions constitute the posterior distribution of ϕ, thereby the posterior distributions of the climate variables.

3.2.2. Uncertainty Propagation

The behavioral model parameter sets and their corresponding climate ensemble members are used to propagate uncertainty to model output in both the calibration and validation periods. Specifically, each behavioral model parameter set is applied only with the climate ensemble member that is used to infer the parameter solution and generates a deterministic model output for the prediction period. This operation is appropriate considering the dependency of parameter uncertainty on input data. Moreover, this operation maintains the pattern and temporal correlation of the climate variables and makes the mapping of climate uncertainty characterization from the calibration period to the validation period simple.

A collection of the deterministic model outputs based on all the behavioral parameter sets and climate ensemble members constitute the ensemble of model outputs. The distribution of the model output can be easily computed by performing multiple deterministic simulations $H(\boldsymbol{\theta}, \mathbf{X}_{\phi})$ using the behavioral model parameter sets and their corresponding climate ensemble members. Like in the dynamically dimensioned search-approximation of uncertainty (DDS-AU) (Tolson and Shoemaker, 2008) and other multi-model forecasting studies such as Dietrich et al. (2009), Thiboult et al. (2016), and Velázquez et al. (2011), each behavioral ensemble member is assumed to have the same likelihood and thus equally probable, and the prediction intervals for a time series are determined independently for each time step.

3.3. Data and Experimental Design

The proposed climate ensemble based calibration framework is implemented in 30 synthetic experiments and 20 real case studies. This section describes in detail the research area, hydrologic model, the Newman et al. (2015) dataset, and the comparative calibration setups of the synthetic experiment and real case study.

3.3.1. Research Area and Data

The real case study is conducted in the same 20 catchments as in Thiboult et al. (2016). These catchments are in southern Québec, Canada with different physiographic characteristics and hydrologic responses. Table 3-1 lists some main characteristics of the 20 catchments. The synthetic experiment is conducted in the first catchment of Table 3-1. Figure 3-1 displays the distribution of the 20 catchments and the hydrometric stations at catchment outlets.

Table 3-1. Main characteristics of the 20 Québec catchments. Q and P are the observed streamflow and precipitation, respectively (based on Table 1 of Thiboult et al. (2016)).

No.	River name	Area (km^2)	Average slope (%)	Mean ann. Q (m^3/s)	Coeff. of variation of Q^*	Mean ann. P (mm)	Mean ann. Snow (cm)
1	Trois Pistoles	923	0.52	18	1.81	1109	382
2	Du Loup	512	0.78	10	1.47	1050	378
3	Gatineau	6796	0.12	127	1.08	1023	332
4	Dumoine	3743	0.13	50	0.81	968	297
5	Kinojevis	2572	0.12	39	1.12	921	324
6	Matawin	1383	0.29	24	1.11	1025	328
7	Croche	1551	0.33	29	1.24	996	360
8	Vermillon	2650	0.20	39	1.10	957	312
9	Batiscan	4483	0.45	96	1.03	1162	381
10	Sainte Anne	1539	0.81	51	1.20	1412	502
11	Bras du Nord	643	0.82	19	1.21	1385	499
12	Du loup	767	0.78	12	1.27	1020	332
13	Aux Ecorces	1107	1.04	28	1.09	1236	450
14	Metabetchouane	2202	0.43	48	1.19	1168	420
15	Peribonka	1010	0.50	19	1.16	1000	376
16	Ashuapmushuan	15342	0.16	300	0.92	984	379
17	Ashuapmushuan	11200	0.12	227	0.88	1001	394
18	Au Saumon	586	0.65	8	1.36	877	334
19	Mistassini	9534	0.20	200	1.08	1004	409
20	Valin	761	1.06	24	1.13	1123	453

* Coefficient of variation is the ratio of the standard deviation to the mean. It shows the extent of variability relative to the mean for daily flow.

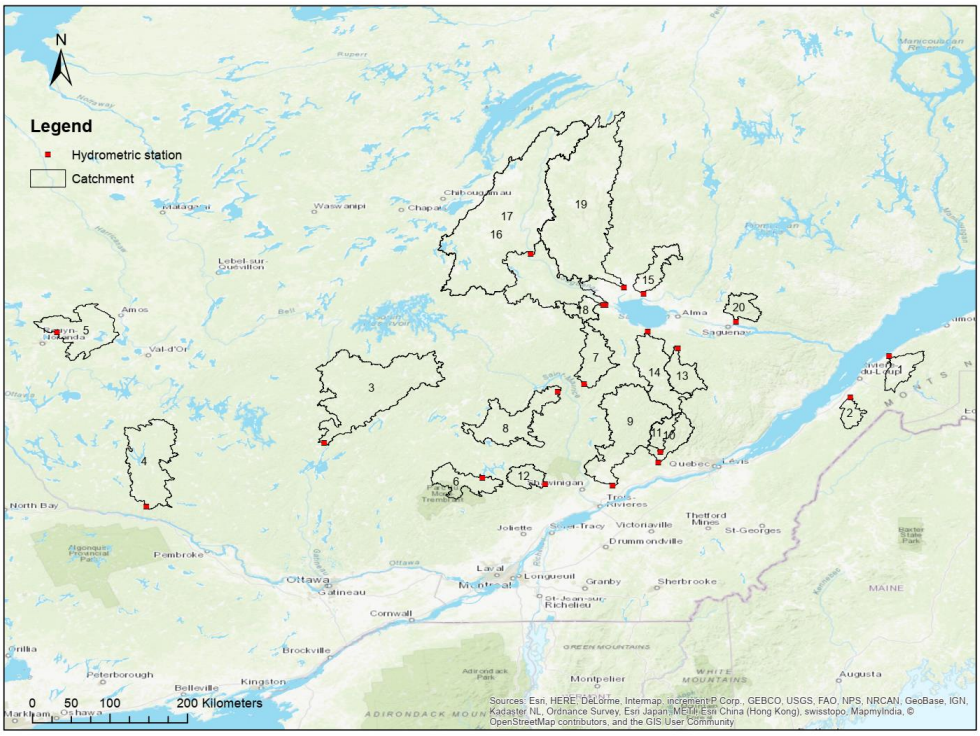

Figure 3-1. Distribution of the 20 Québec catchments and hydrometric stations at catchment outlets

Measurements of streamflow, precipitation and maximum and minimum temperatures are provided by the Direction de l'Expertise Hydrique. The measured climate data are on the 0.1° grid and are generated by the Kriging interpolation given station measurements. In model application, the climate is lumped to the catchment scale by calculating the area-weighted average of the all the grid cells within the catchment.

3.3.2. Hydrologic Model

This chapter used the GR4J (Génie Rural à 4 paramètres Journalier) model that was introduced by Perrin et al. (2003) in both the synthetic and real case study. The GR4J model is a daily time-step, lumped four-parameter rainfall-runoff model and has been applied in numerous studies (e.g., Demirel et al., 2013; McInerney et al., 2017; Renard et al., 2010). In GR4J, basin processes are described by a production store and a routing store. The model includes a conceptual representation of the main hydrologic processes such as percolation, routing, and groundwater exchange. Detailed descriptions of the GR4J involved hydrologic processes are in Appendix A3-1 of this thesis.

GR4J takes precipitation depth and potential evapotranspiration as input. Therefore, two methods are additionally employed to provide necessary driving forces for GR4J. Precipitation depth is calculated by the two-parameter snow accounting routine Cemaneige (Valéry et al., 2014) that is driven by daily precipitation and air temperature and generates the amounts of rain and snowmelt of the catchment.

Potential evapotranspiration is estimated by a conceptual formula proposed by Oudin et al. (2005) based on air temperature and calculated radiation.

There are a total of six parameters in the hydrologic model - four in GR4J and two in Cemaneige. The six parameters' meanings and ranges are listed in Table 3-2.

Table 3-2. Hydrologic model parameters to be optimized in the synthetic and real case study

Parameter	Description	Unit	Minimum	Maximum
X1	Maximum capacity of the production store	mm	10	2500
X2	The groundwater exchange coefficient	mm	-15	10
X3	One day ahead maximum capacity of the routing store	mm	10	700
X4	Time base of unit hydrograph UH1	day	0	7
X5	Snowmelt factor	--	0	1
X6	Cold content factor	mm/d	0	20

3.3.3. Newman et al. (2015) Dataset (N15)

A main focus of this research is to demonstrate how to use an existing ensemble dataset of the historical climate to represent climate data uncertainty in model calibration. The existing climate ensemble dataset used here is the Gridded Ensemble Precipitation and Temperature Estimates (Newman et al., 2015). For short, it is referred to as the N15 dataset in the remaining of this chapter. The description of the N15 dataset is minimized here and closely follows the brief description of the dataset given in Liu et al. (2019). Readers should refer to Newman et al. (2015) for a more complete description of the N15 dataset.

The N15 dataset has 100 historical realizations of daily total precipitation, mean temperature, and daily temperature range for the period 1980-2012. The dataset covers the contiguous United States, northern Mexico, and southern Canada at $1/8^{\circ}$ resolution. It is free for download at https://www.earthsystemgrid.org/dataset/gridded_precip_and_temp.html.

Newman et al. (2015) Dataset Generation

The N15 dataset is generated from the observation based spatial interpolation and ensemble generation. In spatial interpolation, the conditional cumulative distribution functions (CDF) of precipitation and temperature at each grid cell are estimated by locally weighted regressions. The regressions use the elevation, latitude, and longitude of neighboring stations as explanatory variables. In ensemble generation, the spatiotemporally correlated random fields are generated to sample realizations from the CDFs. The spatial and temporal correlations of the random fields are accounted for by using the nested grid approach and the first-order autocorrelation, respectively (Fang and Tacher, 2003; Newman et al., 2015). Note that each climate ensemble member should not be substituted or combined with other members across time and space because of the spatial-temporal correlation and the cross correlation between precipitation and temperature variables in random field generation.

Liu et al. (2019) reports three advantages that the N15 dataset has over other climate data uncertainty estimates. First, it quantifies the probability of precipitation based on zero-to-one probability, not zero-or-one probability, even for zero precipitation observations. Second, it considers the spatial and temporal correlations of climate variables in random field generation. Third, the dataset covers an extensive area at a high spatial resolution and can be directly used for catchments within the contiguous United States, northern Mexico, and southern Canada.

Bias Correction

It is always necessary to check the quality of the ensemble dataset before utilizing it in operation because it is found that meteorological ensembles can be inconsistent with observations, and hydrologically important variables need to be adjusted to be realistic before being used (Graham et al., 2007; Hay et al., 2000; Lenderink et al., 2007; Piani et al., 2010; Yang et al., 2010). This chapter corrected the biases of precipitation, mean temperature, and daily temperature range at the catchment scale. The detailed operation of bias correction is described as follows.

Bias refers to the difference between the ensemble mean and the true value of the variable being evaluated. Since the truth is unknown, the observation is taken as the truth here by assuming the climate measurement is more reliable than any of the N15 climate ensemble members. Taking precipitation as an example, we first calculated the lumped precipitation of each catchment based on the ensemble mean of the N15 dataset and the data considered to be the observed climate, as provided by the Direction de l'Expertise Hydrique, respectively. Then we compared the catchment ensemble mean and the observed precipitation over time. Based on the above analysis, we found that the precipitation ensemble has a systematic bias in contrast with the observation, while the temperature ensembles do not show systematic bias. The precipitation bias correction was carried out by shifting all the ensemble members at the catchment scale by the same magnitude that is equal to the difference between the ensemble mean and the observation at each time step. Bias correction is applied independently to each time step, so the correction magnitude varies with time. In addition, the temperature ensemble members were shifted one day ahead due to the time lag of the ensemble versus the observation (this correction was deemed appropriate based on personal communications with the N15 dataset developer). With these settings, 100 bias-corrected realizations of daily precipitation, and 100 corrected realizations of mean temperature and daily temperature range were defined as the 100-member bias-corrected N15 climate ensemble.

Note that this research used the Newman et al. (2015) dataset version 1.0 for all the synthetic and real case studies considering the dataset availability when we conducted this study. However, in version 1.0, five ensemble members were corrupted at some point in the final formatting, uploading, or file transferring process. These corrupt ensemble members are members 1, 2, 26, 51, and 76 which deviate from most of the ensemble members. This issue has been fixed in version 1.1. However, by using the dataset version 1.0,

our research will show that the proposed calibration framework is able to filter out these incorrectly corrupted data so that their climate uncertainty was not propagated to the model output.

3.3.4. True, Measured, and Prior Climate

The synthetic evaluation of the framework involves 30 experimental replicates conducted for the same catchment (the first catchment of Table 3-1) where each of the 30 experiments uses a different, randomly sampled synthetic measured climate and the synthetic true parameter set. In each synthetic experiment, the prior climate is represented by the same 100-member bias-corrected N15 climate ensemble. The synthetic measured climate is appointed as a different random member of the prior climate ensemble. For example, in the first experiment, the synthetic measured climate is the first member of the prior climate ensemble. In the second experiment, the synthetic measured climate is the second member of the prior climate ensemble, and so on until the 30^{th} experiment.

The synthetic true parameter is needed in order to compare the parameter estimate to the true parameter value (Huard and Mailhot, 2006; Kavetski et al., 2002). In each synthetic experiment, the true hydrologic model parameter set is appointed as a different random parameter set sampled from the feasible parameter ranges. As the model evaluation objective, the synthetic true flow is equal to the hydrologic model simulated flow with the true climate and the true parameter set, in which the synthetic true climate is subjectively selected as the 90^{th} member of the prior climate ensemble for demonstration purpose only and is the same for all 30 synthetic experiments. By doing this, we assume no model structure uncertainty and no other data uncertainties in our hydrologic modeling.

The 20 real case studies are conducted in the 20 catchments of Table 3-1. For each catchment, the prior climate ensemble is composed of the 100-member bias-corrected N15 climate ensemble plus the measured climate of the catchment (totally, 101 members). The measured climate and flow are provided by the Direction de l'Expertise Hydrique. No model structure uncertainty and no other data uncertainties are considered in our hydrologic modeling.

3.3.5. Comparative Model Calibration Setup

Each synthetic and real case study has three calibration approaches. The first two calibration approaches are benchmarks, the third calibration approach follows our proposed climate ensemble based calibration framework. The two benchmarks are selected because they represent two common and efficient model calibration practices. The first calibration approach does not consider any source of uncertainty and generates a single parameter solution. The second calibration approach considers only parameter uncertainty and generates multiple parameter solutions. The three calibration approaches and their setups are shown in Table 3-3 and Table 3-4 for the synthetic experiment and real case study, respectively.

Table 3-3. Three comparative calibration setups for the synthetic experiment

Calib. approach	Calib. period climate data	Calib. algorithm	Climate ensemble members	DDS optim. trials per climate ensemble member	Model evaluations per DDS optim. trial	Total computation cost	Valid. period climate data
1	Synthetic Measured Climate	DDS	1	1	500	500	Synthetic Measured Climate
2	Synthetic Measured Climate	DDS-AU	1	400	50	20,000	Synthetic Measured Climate
3	Prior Climate Ensemble	DDS-AU	100	4	50	20,000	Posterior Climate Ensemble

Table 3-4. Three comparative calibration setups for the real case study

Calib. approach	Calib. period climate data	Calib. algorithm	Climate ensemble members	DDS optim. trials per climate ensemble member	Model evaluations per DDS optim. trial	Total computation cost	Valid. period climate data
1	Measured Climate	DDS	1	1	5,000	5,000	Measured Climate
2	Measured Climate	DDS-AU	1	404	400	161,600	Measured Climate
3	Prior Climate Ensemble	DDS-AU	101	4	400	161,600	Posterior Climate Ensemble

In both synthetic and real case studies, the warm-up, calibration and validation periods are from October 1, 1997 to September 30, 1998, from October 1, 1998 to September 30, 2005, and from October 1, 2005 to September 30, 2010, respectively. For the synthetic experiments, the objective function to be optimized is the root mean square error (RMSE) between the simulated flow and the true flow expressed as:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{Y}_t - Y_t)^2} \tag{3-1}$$

where $\hat{Y}_t$ and Y_t are the simulated and true flows at time t $(t = 1,2, \dots, T)$. The real case study uses the same objective function except replacing the true flow, Y_t with the measured flow, $\tilde{Y}_t$.

(1) Calibration Approach 1: Deterministic

Calibration approach 1 uses the measured climate as model input and calibrates the hydrologic model by minimizing the RMSE between the simulated flow and the true flow (for the synthetic cases) or the observed flow (for the real cases) with the dynamically dimensioned search (DDS) algorithm (Tolson and

Shoemaker, 2007). The computational budget (in terms of maximum allowable model evaluations) was set 500 for the synthetic cases and 5,000 for the real cases because our trials with different numbers of model evaluations showed that our six-parameter hydrologic model would not gain significant performance improvements beyond these limits. In model validation, the obtained single parameter set is validated with the deterministic measured climate, generating a deterministic flow prediction.

(2) Calibration Approach 2: Parameter Uncertainty

Calibration approach 2 uses the same measured climate and objective function as approach 1 but identifies numerous behavioral parameter sets by the DDS-AU algorithm (Tolson and Shoemaker, 2008). DDS-AU describes hydrologic model parameter uncertainty by identifying multiple behavioral parameter sets in separate optimization trials. Separate DDS optimization trials initialized to different initial solutions and/or different random seed will follow different search paths and can terminate at different final solutions.

In the synthetic experiments, the DDS-AU calibration performs 400 separate DDS optimization trials and each trial uses 50 model evaluations. In the real case studies, the DDS-AU calibration performs 404 separate DDS optimization trials and each trial uses 400 model evaluations. In both synthetic and real case studies, each DDS optimization trial is initialized to a different initial solution and a different random seed. Only the final best DDS solution (one per trial) is considered as a possible behavioral parameter set for filtering.

The total number of model evaluations or total computational cost is 20,000 ($1 \times 400 \times 50 = 20{,}000$) for each synthetic experiment and 161,600 ($1 \times 404 \times 400 = 161{,}600$) for each real case study. The total number of candidate parameter solutions is 400 for each synthetic experiment and 404 for each real case study.

After parameter sampling, the DDS-AU algorithm uses the equivalent objective behavioral solution identification process as described earlier in Section 3.2.1. The only difference for calibration approach 2 is that the candidate behavioral solutions do not include any climate uncertainty and are only based on model parameter uncertainty.

In model validation, each of the obtained behavioral parameter sets is validated with the deterministic measured climate. All model outputs of interest (streamflow) are assumed equally likely and then combined to constitute the ensemble flow predictions.

(3) Calibration Approach 3: Parameter and Climate Uncertainty

Calibration approach 3 uses each of the 100 climate ensemble members for model calibration in the synthetic experiment and uses each of the 101 climate ensemble members (i.e., 100 climate ensemble members plus the observed climate) for model calibration in the real case study. Note that a random sampling of climate ensemble members is unnecessary in this application due to the small size of the

available climate ensemble. Each climate ensemble member is used as input to the hydrologic model and is subject to calibration with DDS-AU.

In both the synthetic and real case studies, given a climate input, the DDS-AU calibration performs 4 separate DDS optimization trials. Each DDS optimization trial uses 50 model evaluations for the synthetic experiment and 400 model evaluations for the real case study. Each DDS optimization trial is initialized to a different initial solution and a different random seed. Only the final best DDS solution (one per trial) is considered as a possible behavioral parameter set for filtering.

The total number of model evaluations or total computational cost is 20,000 ($100 \times 4 \times 50 = 20{,}000$) for each synthetic experiment and 161,600 ($101 \times 4 \times 400 = 161{,}600$) for each real case study. The total number of candidate parameter solutions is 400 for the synthetic cases and 404 for the real cases. Note that the total computational cost is set the same for approaches 2 and 3 in order to fairly compare how each approach performs. The number of model evaluations per DDS optimization trial is also set the same for approaches 2 and 3 to ensure the same optimization effort per DDS optimization trial.

After parameter sampling, the candidate behavioral solutions are filtered to identify the behavioral solutions based on the same behavioral solution identification approach used in approach 2 (and described fully in Section 3.2.1). The behavioral parameter sets and their corresponding climate ensemble members are then used to propagate uncertainty to flow prediction.

Uncertainty propagation to flow prediction involves all the behavioral model parameter sets and their corresponding climate ensemble member numbers (behavioral solutions). In our experiments, the maximum number of the behavioral solutions is 400 for the synthetic experiments and 404 for the real cases and each behavioral solution is utilized in flow prediction. Considering the dependency of parameter estimates on climate input as well as the temporal correlation of climate data, each behavioral solution of approach 3 is only applied with its corresponding climate ensemble member as used in model calibration. Each of these predicted flows is assumed equally likely and used to create the ensemble of flow predictions.

3.3.6. Evaluation of Flow Prediction and Parameter Estimation

The flow prediction and parameter estimation results of each synthetic experiment are evaluated by comparing with the true flows and true parameter values, respectively. The flow prediction of each real case study is evaluated by comparing with the observed flow. The flow prediction is assessed from two perspectives: deterministic and probabilistic. When only a deterministic flow prediction is required, the deterministic flow prediction is calculated as the mean of the ensemble prediction, and the calibration approach performance is evaluated by comparing the ensemble mean flow prediction with the true flow for synthetic experiment or the observed flow for real case study. The Kling-Gupta efficiency (KGE) is used

as the deterministic prediction evaluation metric (Gupta et al., 2009). The KGE is a variant of the Nash-Sutcliffe efficiency coefficient and is expressed as:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{3-2}$$

where r is the linear correlation coefficient between the simulated and true (observed) flows, α is the ratio between the simulated flow standard deviation and the true (observed) flow standard deviation, and β is the ratio between the mean simulated and mean true (observed) flows. The range of KGE is between negative infinity to one with the optimal value of one.

When the probabilistic flow prediction is required, the calibration approach performance is evaluated by comparing the ensemble flow prediction with the true flow for synthetic experiment or the observed flow for real case study. When modelers evaluate the probabilistic prediction, reliability is often the first concern. Recall that reliability here is defined as the percentage of the true (observed) flow values falling into the 95% flow prediction intervals. However, a great reliability at the cost of excessively wide prediction intervals is not desired. The ensemble flow prediction is expected to contain as many true (observed) flow points as possible but also be as narrow as possible. Therefore, the spread is introduced as a second metric to assess the prediction width. In this thesis, the spread is calculated as the square root of the average ensemble prediction variance over the evaluation period (Fortin et al., 2014). This spread metric has been used in many studies to evaluate the prediction width (e.g., Abaza et al., 2017a; Fernández-González et al., 2017; Green et al., 2017; Junk et al., 2015; Thiboult et al., 2016).

$$spread = \sqrt{\frac{1}{T}\sum_{t=1}^{T} Var\left(\hat{Y}_t\right)} \tag{3-3}$$

$$Var(\hat{Y}_t) = \frac{1}{N-1}\sum_{n=1}^{N}\left(\hat{Y}_{t,n} - Y_t\right)^2 \tag{3-4}$$

where T is the total time steps of the evaluation period ($t = 1, \dots, T$). $Var(\hat{Y}_t)$ is the variance of the simulated flow at time t relative to the true (observed) flow. $\hat{Y}_{t,n}$ is the n^{th}simulated flow of the ensemble flow prediction at time t. N is the prediction ensemble size ($n = 1, \dots, N$). The range of spread is non-negative with the optimum of zero and shares the unit of the flow.

In addition to the reliability and spread, the reliability diagram is used to examine the probabilistic prediction from a more comprehensive perspective. The reliability diagram is a graph of the observed frequency of an event plotted against the predicted probability of an event (Hartmann et al., 2016). In theory, a perfect prediction system will result in the prediction with a probability of X% being consistent with the observation X% of the time. Hence, when plotting a reliability diagram, comparisons are made against the diagonal. A curve above the diagonal line denotes an over-dispersion, an under-dispersion is in the opposite case (Thiboult et al., 2016).

As to the evaluation of the parameter estimates, only the DDS-AU algorithm based calibration (i.e., approaches 2 and 3 in Table 3-3 and Table 3-4) results are evaluated as they attempt to describe parameter uncertainty. The true parameter coverage is used as a quantitative evaluation metric and for a given calibration approach is calculated as the relative frequency across all 30 synthetic experiments that the behavioral parameter ensemble contains the true parameter value between its min-max bounds. In addition, the number of behavioral parameter sets is evaluated for both the synthetic and real case studies. Given the same model performance metric levels for two calibration approaches, the one with a smaller number of behavioral parameter sets is preferred (assuming modelers prefer quicker uncertainty propagation experiments).

3.4. Results

The comparison results of the three calibration approaches are presented in two sections. Section 3.4.1 is for the synthetic experiments. Section 3.4.2 is for the real case studies. Each section compares the flow predictions and parameter estimates for the calibration approaches of Table 3-3 or Table 3-4, and also compares the prior and posterior climate ensembles of calibration approach 3. After evaluating the proposed framework, Section 3.4.3 shows two additional findings based on the real case studies by investigating: (1) the impacts of the calibration- and the validation- period climate uncertainty on flow predictions, and (2) the impacts of the dependency relation between model parameters and model input on flow predictions. Note that all the reported evaluation results on flow predictions are for the validation period.

3.4.1. Synthetic Experiment

3.4.1.1. Evaluation of Flow Predictions

Figure 3-2 details the flow prediction evaluation metrics results of calibration approaches 1, 2, and 3 for all 30 synthetic experiments. For the ensemble mean flow prediction (Figure 3-2a), an interesting result is that, approach 2 does not provide much better KGE results than approach 1, though parameter uncertainty has been considered in approach 2. However, by accounting for climate data uncertainty in model calibration, approach 3 (the proposed framework) substantially improves the KGE for experiments showing poor performance with approach 1 or 2 (e.g., KGEs in the 1st, 2nd, 10th, and 26th experiments). Figure 3-2a also indicates that for the experiments where the benchmark configurations perform well, the framework maintains their good performance. Therefore, the framework substantially improves poor predictions and maintains good predictions in the context of the ensemble mean flow prediction.

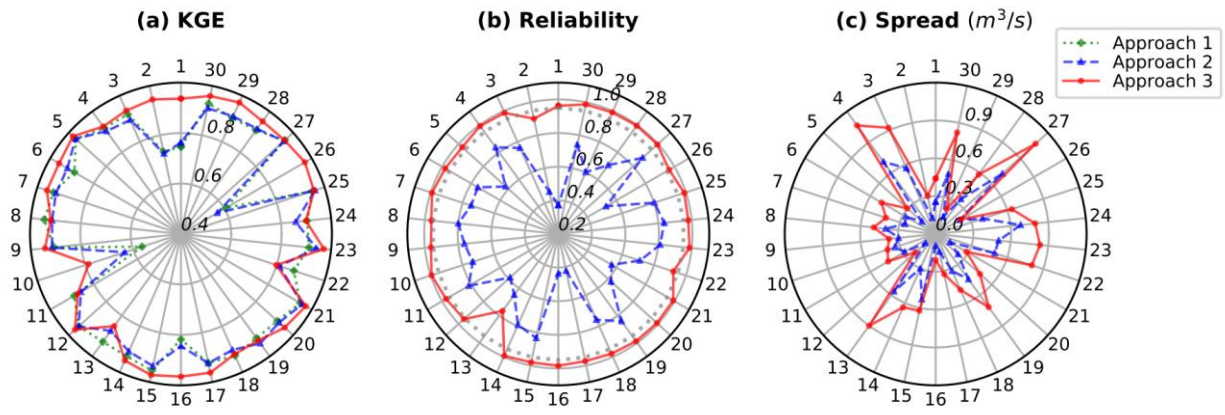

Figure 3-2. Flow predication evaluation metrics results of calibration approaches 1, 2, and 3 for all 30 synthetic experiments. See Table 3-3 for the descriptions of approaches 1, 2, and 3. Since approach 1 generates a deterministic flow output to which the reliability and spread are not applicable, it is not shown in panels (b-c). Each synthetic experiment is identified by the label on the outer edge of the wheel. The metric value of each synthetic experiment is represented by the value on the synthetic experiment corresponding spoke. The dotted line of panel (b) represents the reliability value of 0.95.

As to the probabilistic flow prediction, the proposed framework is advantageous in improving the reliability. As shown in Figure 3-2b, approach 3 achieves much higher reliability than approach 2 in all 30 experiments. The average absolute reliability improvement of 30 synthetic experiments is 0.31. More importantly, the reliability of the 95% prediction intervals is closer to the 95% observed probability (i.e., the dash line in Figure 3-2b) in approach 3 than in approach 2 in all experiments. In terms of the prediction width (Figure 3-2c), the spread increases due to the use of climate ensemble in flow prediction. The average absolute spread increase of 30 synthetic experiments is 0.34 mm/d. Considering the magnitude of the reliability increase, the wider prediction intervals may be welcome among water resources managers and flood forecasters. Based on other synthetic experiments, not reported on in detail here, using other random members of the bias-corrected N15 climate ensemble as the true climate, similar relative performance levels were observed.

3.4.1.2. Evaluation of Hydrologic Model Parameter Estimates

Among the 180 parameter estimate ensembles (6 parameters × 30 synthetic experiments), the true parameter coverages of approaches 2 and 3 are 95.0% and 99.4%, respectively. The ensemble mean parameter estimate of approach 3 is closer to the true parameter value than that of approach 2 in 62.8% of the 180 estimates. This result is consistent with the findings of many studies that ignoring the forcing data uncertainty leads to biased estimates of model parameters (Kavetski et al., 2002; Kuczera et al., 2006).

To visualize the parameter estimation improvements brought by the proposed framework, Figure 3-3 plots the histograms of the posterior parameter estimates of approaches 2 and 3 for two synthetic experiments, the 1st and 18th experiments. In the 1st experiment, approach 2 parameter estimates barely

cover or fail to cover the true values of parameters X2, X3, X5, and X6, but approach 3 yields parameter ensembles that do cover all the true parameter values. Recall that in the 1st synthetic experiment, the first member of the N15 climate ensemble is used as the synthetic measured climate for model calibration, but this climate member is incorrectly corrupted and deviates from most of the 100 climate ensemble members. Considering this, Figure 3-3a shows that the poor-quality measured climate leads to biased parameter estimates.

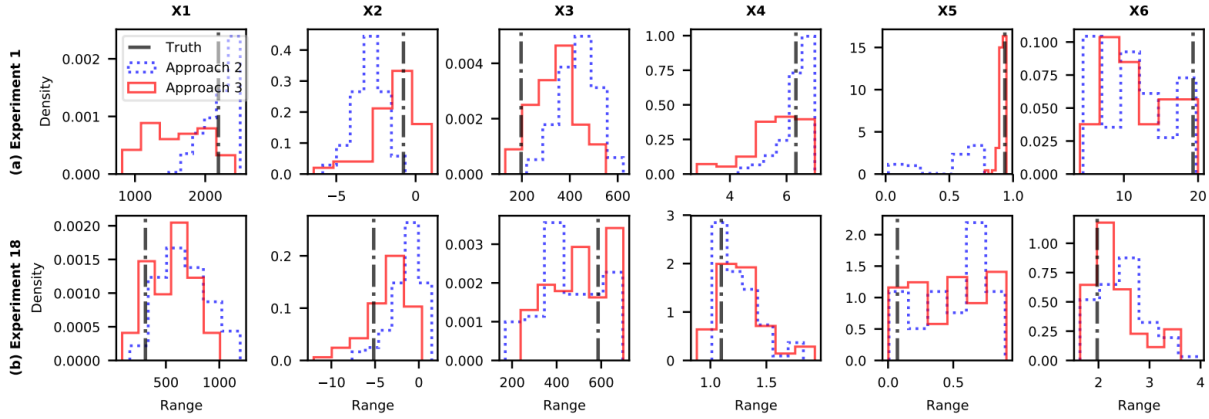

Figure 3-3. Comparative histograms of the six hydrologic model parameters of approaches 2 and 3 for the 1st and 18th synthetic experiments. The vertical dash-dotted line represents the synthetic true parameter value of the synthetic experiment.

Moreover, comparing the posterior parameter histograms generated by the synthetic measured climate (approach 2) and by the ensemble climate (approach 3) in Figure 3-3, especially all parameters in the 1st experiment and X2, X3, X5, and X6 in the 18th experiment, the ensemble climate generates different posterior parameter distributions from the synthetic measured climate. This implies that parameter estimates change when calibrating the hydrologic model with different input data.

In addition to the superior predictive performance of the proposed framework shown above in the synthetic tests, the proposed framework yields such improved predictions with a much lower number of behavioral parameter sets than approach 2 in all 30 synthetic experiments. The average number of behavioral parameter sets of 30 experiments is 184 for approach 2 and 52 for approach 3. The reduction of the number of behavioral parameter sets is especially helpful for complex and large-scale hydrologic model applications. For example, the computational costs (e.g., total model runtimes in serial computing scenario) associated with producing ensemble flow forecasts would be lowered by $(184 - 52)/184 = 71.7\%$ using the parameters produced by the proposed framework compared to using the parameters produced by approach 2.

3.4.1.3. Evaluation of Climate Ensemble Member Number Estimates

After analyzing the hydrologic model parameter estimation results, we look at the latent parameter - climate ensemble member number (ϕ) estimation results of calibration approach 3. Figure 3-4 shows the prior and posterior histograms of the climate ensemble member number (ϕ) for the same two synthetic experiments as in Figure 3-3. Recall that in the synthetic experiments, approach 3 uses the 100-member climate ensemble as model inputs. Each climate ensemble member has equal prior probability in the DDS-AU based calibration. After approach 3 calibration, the unique climate ensemble member numbers are reduced from 100 to only 57 and 58 for the 1st and 18th synthetic experiments, respectively. The posterior climate ensemble members have different posterior probabilities. Noticeably, the five incorrectly corrupted climate ensemble members of the N15 dataset (version 1.0) (i.e., members 1, 2, 26, 51, and 76) are all excluded from the posterior climate ensemble members in both experiments.

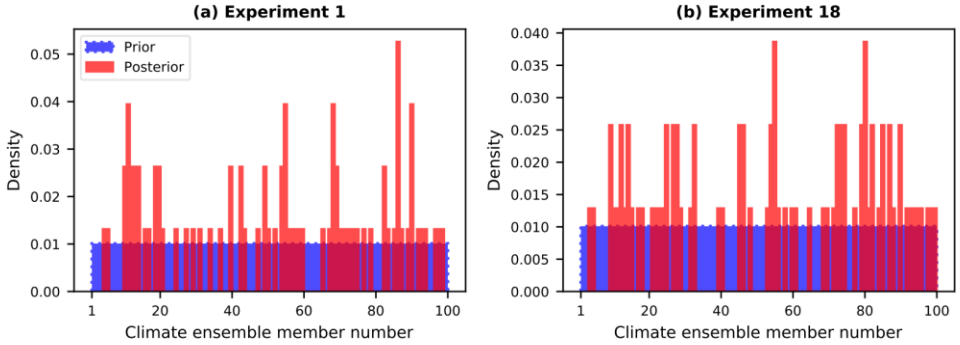

Figure 3-4. Comparative histograms of the climate ensemble member number (ϕ) of calibration approach 3 for the 1st and 18th synthetic experiments.

To look at the impacts of the filtering function of the framework on climate uncertainty estimates, Figure 3-5 takes the 1st synthetic experiment as an example and compares the synthetic truth, synthetic measurement, prior and posterior histograms of the precipitation and daily mean temperature on a wet day (September 9, 2004) and a dry day (March 1, 2010), respectively. First, we look at the impact of the filtering function on the measured climate. In Figure 3-5, the synthetic measured precipitation and temperature deviate from the true values. When such a bad quality climate measurement is used in calibration approaches 1 and 2, it is hard for both approaches to find good parameter solutions. However, in approach 3, the impacts of this poor-quality climate measurement were eliminated because, in addition to the synthetic measured climate, other climate ensemble members were also used to search good parameter solutions. The bad quality climate ensemble members were filtered out (see Figure 3-4a) because they did not generate any behavioral parameter set in approach 3. In model validation, this bad quality climate ensemble member was kept in approaches 1 and 2, just like most modelers would do in practice given that they often have no other option but to utilize the deterministic climate data available. In contrast, in

approach 3, the synthetic measured climate was automatically excluded and was not used to estimate model prediction uncertainty.

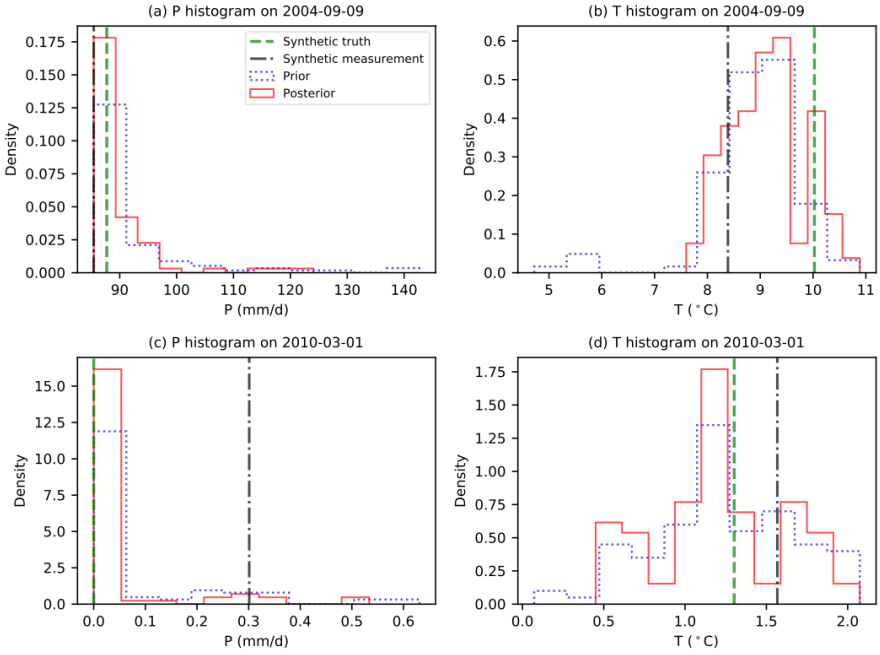

Figure 3-5. Comparative histograms of the precipitation and daily mean temperature of calibration approach 3 for the 1st synthetic experiment on a wet day (September 9, 2004) and a dry day (March 1, 2010).

In addition to filtering out the poor-quality measured climate, the framework is also able to refine the uncertainty estimates of the climate input. As shown in Figure 3-5, for both precipitation and temperature, the posterior ensembles are less spread than the prior ensembles by excluding many extreme values, and the posterior ensemble means appear to be closer to the synthetic true values than the prior ensemble means. Similar trends also appear for other time steps (though not shown here).

3.4.2. Real Case Study

3.4.2.1. Evaluation of Flow Predictions

Figure 3-6 details the flow prediction evaluation metrics results of calibration approaches 1, 2, and 3 for all 20 catchments. For the deterministic prediction, the observed climate based calibration approaches 1 and 2 already achieve practically good KGEs (i.e., higher than 0.75). The ensemble climate based calibration maintains these good performances for KGEs within the range of 0.79 to 0.93. For the probabilistic prediction, approach 3 improves the average reliability of 20 catchments from 0.4 to 0.7 relative to approach 2.

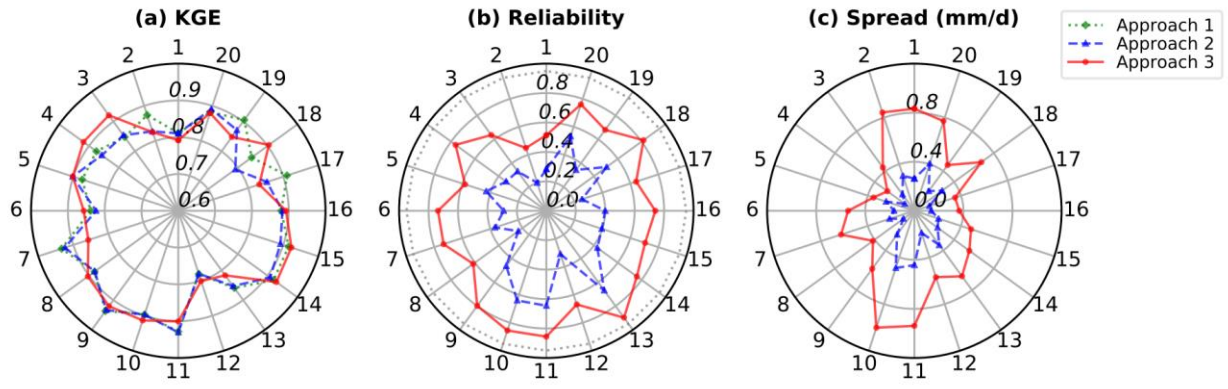

Figure 3-6. Flow predication evaluation metrics results of calibration approaches 1, 2, and 3 for all 20 catchments. See Table 3-4 for the descriptions of approaches 1, 2, and 3. Approach 1 is deterministic and thus is not shown in panels (b-c). Each catchment is identified by the label on the outer edge of the wheel. The metric value of each catchment is represented by the value on the catchment corresponding spoke. The dotted line of panel (b) represents the reliability value of 0.95.

A more complete reliability comparison between approaches 2 and 3 can be seen through the reliability diagram (Figure 3-7). Both approaches 2 and 3 are under dispersion and are therefore over confident. Although the system reliability is not perfect for approach 3, approach 3 reduces the difference from the perfect reliability (i.e., diagonal) to a much greater extent than approach 2.

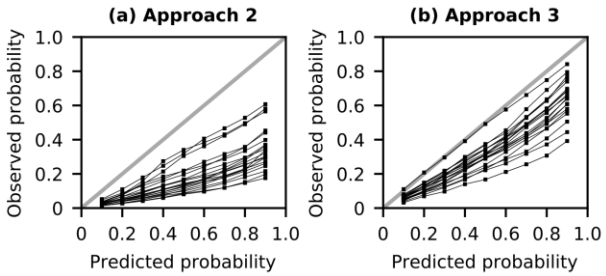

Figure 3-7. Reliability diagrams of calibration approaches 2 and 3 for all 20 catchments. Each curve of a reliability diagram refers to a catchment. The diagonal represents the perfect reliable prediction.

The reason for the reliability increases of approach 3 is that the proposed framework uses the behavioral parameter sets corresponding climate ensemble (i.e., the posterior climate ensemble) in flow prediction. The increased spread of climate inputs, relative to the single observed climate, is propagated to model output and broadens the prediction width. As shown in Figure 3-6c, the average absolute spread increase of 20 catchments is 0.34 mm/d. Therefore, more flow observations can be included in the prediction intervals.

3.4.2.2. Evaluation of Hydrologic Model Parameter Estimates

The Wilcoxon rank-sum hypothesis test (Helsel and Hirsch, 2002) is conducted to compare the posterior parameter distributions of approaches 2 and 3 for the real case study. Results show that approaches 2 and 3

generate different marginal parameter distributions for 84.2% of the 120 parameter estimations (6 parameters × 20 real cases) with a 5% level of significance. This finding is consistent with the results of the synthetic experiments (Section 3.4.1.2). It confirms that the perturbed climate data generates different posterior parameter distributions than those generated by calibrating with only the observed climate, and parameter uncertainty estimation is dependent on climate input in practical applications.

To visualize the parameter estimate differences between approaches 2 and 3, Figure 3-8 displays their posterior parameter histograms of all six parameters for two catchments. Based on the Wilcoxon rank-sum test, parameters X2, X3, X5, and X6 for Trois Pistoles, and parameters X1, X2, X3, X5, and X6 for Sainte Anne are significantly different between approaches 2 and 3 (5% level of significance).

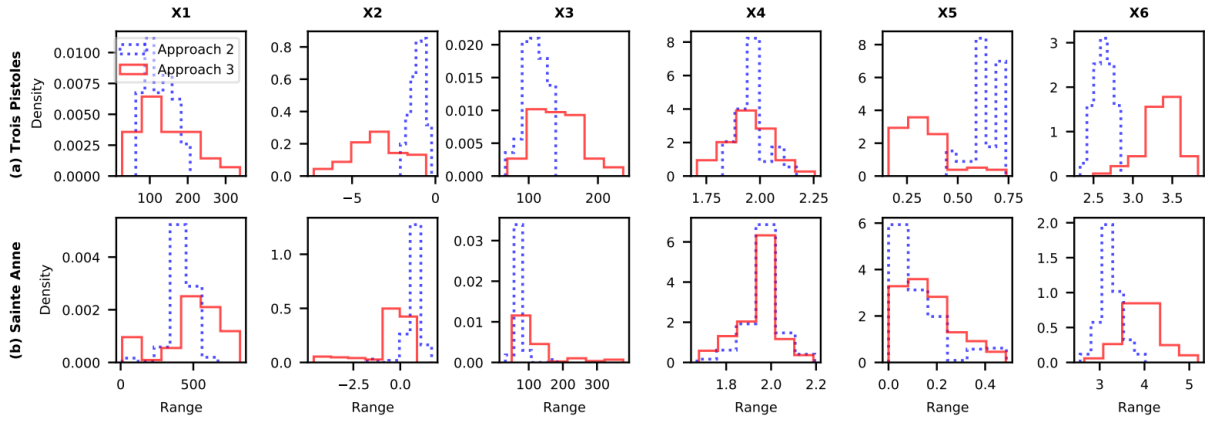

Figure 3-8. Comparative histograms of all six parameter estimates of approaches 2 and 3 for the Trois Pistoles and Sainte Anne catchments (the 1st and 10th catchments in Table 3-1).

In addition, calibration approach 3 reduces the average number of the behavioral parameter solutions from 162 to 123 relative to approach 2. Therefore, the computational cost of applying the hydrologic model is reduced by $(162 - 123)/162 = 24.1\%$ using approach 3 produced parameter sets compared to using the approach 2 produced parameter sets. The reduction in the number of the behavioral parameter sets is consistent with the result of the synthetic experiments.

3.4.2.3. Evaluation of Climate Ensemble Member Number Estimates

In terms of the climate ensemble member number (ϕ) estimates of approach 3, Figure 3-9 shows the prior and posterior histograms of the climate ensemble member number (ϕ) for the same two catchments as in Figure 3-8. In the real case studies, approach 3 uses the 101-member climate ensemble as model inputs, and each climate ensemble member has equal prior probability in the DDS-AU based calibration. After approach 3 calibration, the unique posterior climate ensemble member numbers are reduced from 101 to only 34 and 57 for the Trois Pistoles and Sainte Anne catchments, respectively. The posterior climate ensemble members have different posterior probabilities. The measured climate (i.e., climate ensemble member number 101) remains in the posterior climate ensemble for both catchments.

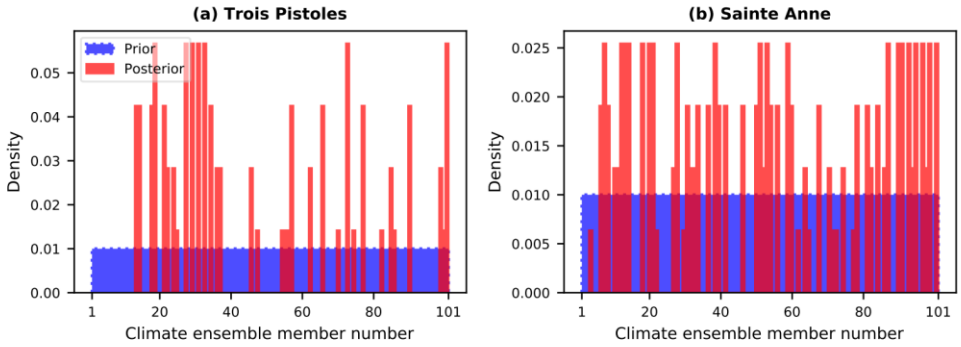

Figure 3-9. Comparative histograms of the climate ensemble member number (ϕ) of approach 3 for the Trois Pistoles and Sainte Anne catchments (the 1st and 10th catchments in Table 3-1).

Taking the 10th experiment as an example, Figure 3-10 compares the prior and posterior histograms of the precipitation and daily mean temperature on a wet day (September 9, 2004) and a dry day (March 1, 2010). Similar as in the synthetic experiments, for both precipitation and temperature, the posterior climate ensemble is slightly more restrictive than the prior climate ensemble by excluding some extreme values.

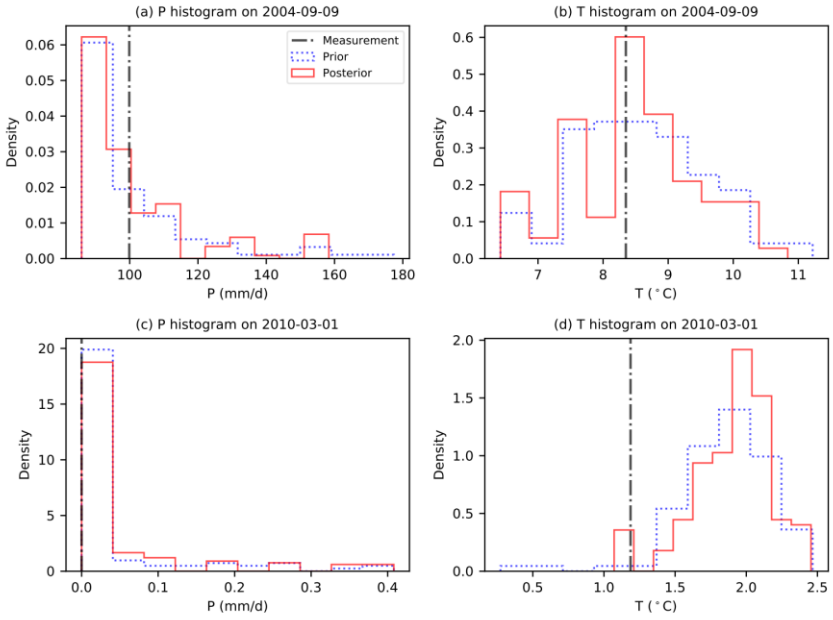

Figure 3-10. Comparative histograms of the precipitation and daily mean temperature of approach 3 for the Sainte Anne catchment (the 10th catchment of Table 3-1) on a wet day (September 9, 2004) and a dry day (March 1, 2010).

3.4.3. Additional Findings Based on Real Case Study

3.4.3.1. Climate Uncertainty Accounted in the Calibration or Validation Period

The reader by now might ask: the results in Figure 3-6 are promising, but how about considering climate data uncertainty in only the calibration period or only the validation period? In the attempt to account for climate data uncertainty, some studies used perturbed climates in model calibration but did not validate

their climate uncertainty estimates in a region or for a time period that was not explicitly utilized to estimate model parameters (e.g., see Balin et al., 2010; Huard & Mailhot, 2006; Kavetski et al., 2002; Moradkhani et al., 2005). Another common practice, especially in ensemble streamflow forecasting, is calibrating hydrologic models with observed climates but applying the inferred parameter set with ensemble climate forecasts (see e.g., Cameron et al., 1999; Carpenter and Georgakakos, 2004; McMillan et al., 2013; Reichle et al., 2002; Thiboult et al., 2016). Both procedures fail to follow a consistent climate data uncertainty assumption throughout the modeling process. Despite this inconsistency, it is unclear if these procedures from the literature are sufficient to get similar or more accurate and reliable predictions than the proposed framework. To answer this question, we compared the proposed framework with the two practices.

The two practices are named scenario C and scenario V here. Scenario C only accounts for the calibration period climate uncertainty. Scenario V only accounts for the validation period climate uncertainty. Recall that approach 2 does not consider climate uncertainty at all, while approach 3 accounts for climate data uncertainty in both the calibration and validation periods (see Table 3-4). In scenario C, the inferred hydrologic model parameters are the same as those of approach 3 but they are validated with the measured climate, so climate uncertainty is only accounted in the calibration period. In scenario V, the parameters solutions are the same as approach 2's, estimated based on the observed climate, but each behavioral parameter is validated with all the 101-member climate ensemble. Therefore, climate uncertainty is only accounted in the validation period.

Figure 3-11 shows the flow prediction evaluation metrics results of approaches 2 and 3 and scenarios C and V over all 20 catchments. From the ensemble mean prediction perspective (Figure 3-11a), approaches 2 and 3 get better KGE than scenarios C and V in 15 of the 20 catchments. This result implies that it is beneficial to use a consistent climate uncertainty assumption, deterministic or uncertain, throughout the calibration process for accurate deterministic predictions.

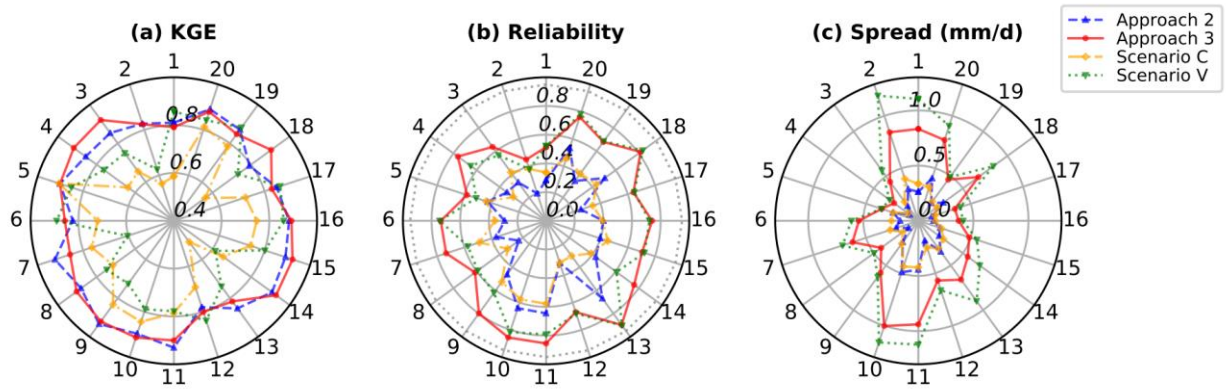

Figure 3-11. Flow predication evaluation metrics results of calibration approaches 2 and 3 and scenarios C and V over all 20 catchments. Scenario C only accounts for the calibration period climate uncertainty, while scenario V only accounts for the validation period climate uncertainty. Each catchment is identified by the label on the outer edge of the wheel. The metric value of each catchment is represented by the value on the catchment corresponding spoke. The dotted line of panel (b) represents the reliability value of 0.95. The results of approaches 2 and 3 are the same as the results reported in Figure 3-6.

From the probabilistic prediction perspective (Figure 3-11(b-c)), scenario C gets similar reliability and spread results with approach 2. Both approach 2 and scenario C have inferior reliabilities to approach 3 in all 20 catchments. Therefore, only considering the calibration period climate uncertainty (scenario C) does not significantly increase model output uncertainty relative to the approach that does not explicitly consider climate uncertainty (approach 2). This finding is consistent with the results of several calibration studies showing the limited impact of input uncertainty on model output (Balin et al., 2010; Montanari, 2005; Montanari and Di Baldassarre, 2013).

Most importantly, Figure 3-11(b-c) show that scenario V is demonstrably worse than approach 3 in 11 of 20 catchments where scenario V reliabilities are equal to or lower than approach 3 reliabilities (when rounding to two decimal places) despite having a larger spread of predictions. In the other nine catchments, minor reliability improvements of scenario V over approach 3 are gained at the expense of larger spreads. There are no catchments where scenario V is better than approach 3 in terms of both spread and reliability.

Taking the Trois Pistoles catchment (the 1st catchment in Table 3-1) as an example, Figure 3-12 compares the predicted flows of approach 3 and scenarios C and V which all account for climate uncertainty but in different ways. Scenario C prediction intervals are poor because they do not contain a number of flow observations, particularly in late April and May. Although scenario V and approach 3 both have the same reliability for the entire validation period (~0.90), the width of the scenario V prediction intervals for the April-June period is often two times or more the width of the approach 3 prediction intervals. The numerical differences of spread from Figure 3-11c therefore translate into practically significant differences in the 95% prediction interval quality.

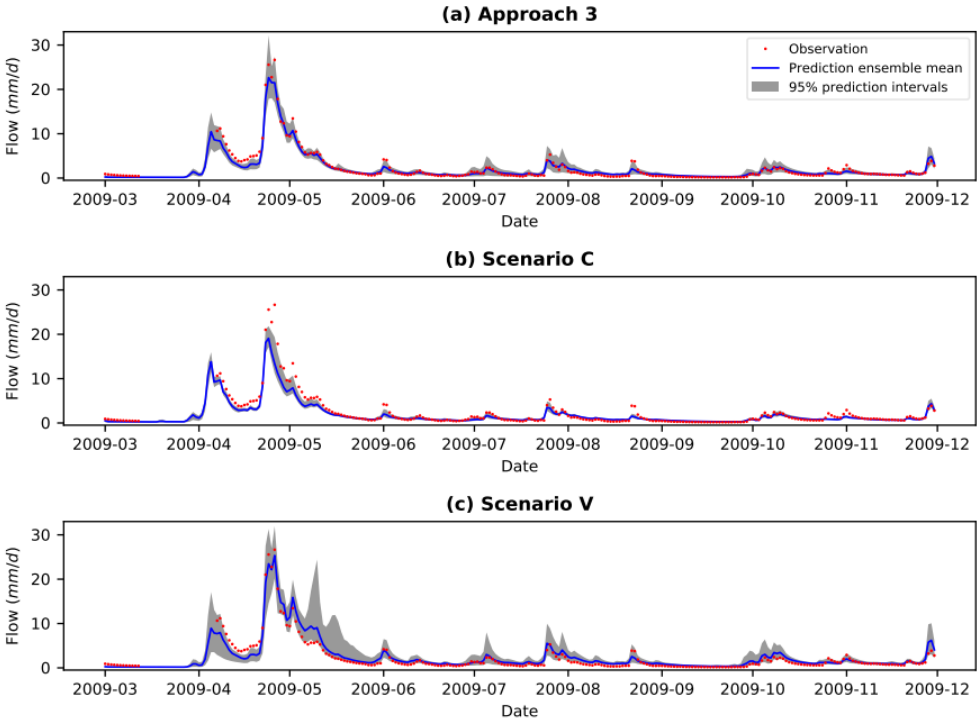

Figure 3-12. Flow predictions of calibration approach 3 and scenarios C and V of the Trois Pistoles catchment (the 1st catchment in Table 3-1) for a portion of the validation period.

All these results indicate that it is insufficient to only account for climate uncertainty in the calibration period or the validation period (or similarly, the forecasting period). If a climate variable is assumed uncertain, it should be ideally treated uncertain in both the calibration and validation phases.

3.4.3.2. Parameter Uncertainty Dependency on Input Data in Uncertainty Propagation

In theory, hydrological model parameters do not depend on the particular values but the statistics of the input when the input data is sufficient in model calibration (Montanari and Koutsoyiannis, 2012). However, in practice, input data is seldom sufficient. Parameters can change when calibrating the hydrological model with different input data. Such a dependency relation between model parameters and input should be recognized and respected in uncertainty propagation to model output. In the literature, some studies used the same climate data uncertainty estimates generated prior to model calibration throughout the calibration and validation processes (e.g., see Montanari, 2005; Montanari & Di Baldassarre, 2013). The climate realizations that generate non-behavioral parameter sets (e.g., model predictions that are inconsistent with measured response data) are nonetheless utilized in uncertainty propagation. This operation discards the dependency of parameters on climate inputs in uncertainty propagation.

To test the impacts of the dependency relation between model parameters and model input on flow predictions, we compared approach 3 flow predictions with those using the prior climate ensemble in model

validation. The latter are not subject to the filtering of the climate ensemble and thus disregard the dependency of parameter uncertainty on climate input in uncertainty propagation.

Given the behavioral parameters of approach 3, the prior climate ensemble based validation can be implemented in two ways. As shown in Figure 3-13, one is validating each behavioral parameter with all the 100 prior climate ensemble members, referred to as scenario 1:N. Another is applying each behavioral parameter set with a random member sampled from the prior climate ensemble with replacement, referred to as scenario 1:1. The latter is repeated 100 times to reduce the impact of sampling errors on model performance and the mean evaluation metrics results of the 100 scenario 1:1 validations are reported.

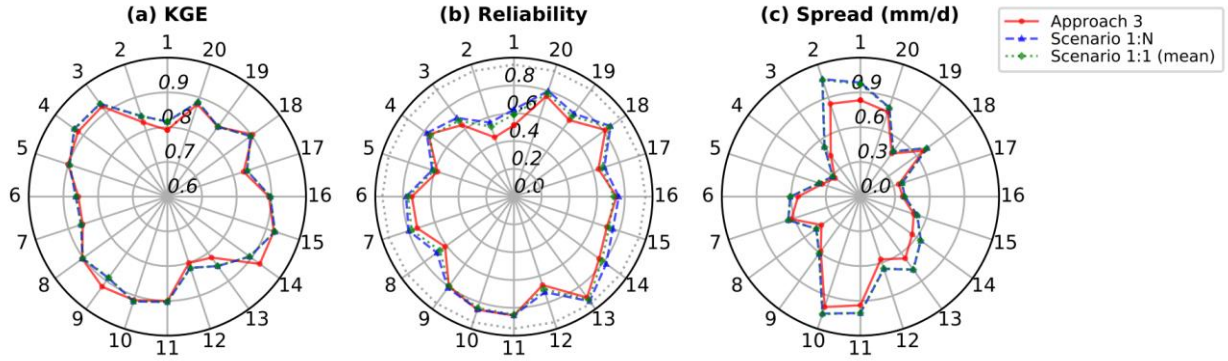

Figure 3-13. Flow predication evaluation metrics results of calibration approach 3 and scenarios 1:N and 1:1 for all 20 catchments. Scenario 1:N validates each parameter with all the 100 prior climate ensemble members. Scenario 1:1 validates each parameter with a random member of the prior climate ensemble. Scenario 1:1 is repeated 100 times and the mean metrics results are reported. Each catchment is identified by the label on the outer edge of the wheel. The metric value of each catchment is represented by the value on the catchment corresponding spoke. The dotted line of panel (b) represents the reliability value of 0.95. The results of approach 3 are the same as the results reported in Figure 3-6.

Compared with the results of approach 3, the KGE and reliability changes of scenarios 1:N and 1:1 are marginal, and the spreads of scenarios 1:N and 1:1 are increased by 0.07 mm/d and 0.06 mm/d on average of the 20 catchments, respectively. The spread increases are particularly substantial for the 1st, 2nd, 13th, and 14th catchments. Hence the posterior climate ensemble of approach 3 generates a tighter ensemble of flow predictions that are no less accurate (KGE) and no less reliable than the prior climate ensemble (i.e., scenarios 1:N and 1:1). Moreover, the proposed calibration framework saves 100-fold computational cost than scenario 1:N since each behavioral parameter set is only applied with one climate ensemble member.

3.5. Conclusions and Future Work

This chapter formalized a climate ensemble based hydrologic model calibration framework to explicitly account for climate data uncertainty in model calibration. The framework is simple and parsimonious because it has only one latent climate input parameter to estimate. Moreover, this work proposed the direct use of existing climate ensemble products as the prior climate ensemble. The direct use of existing climate

ensembles is not only convenient but also provides informative prior information about climate data uncertainty.

For demonstration, 100-member precipitation and temperature ensembles were utilized from the Gridded Ensemble Precipitation and Temperature Estimates dataset (Newman et al., 2015), referred to as N15, in this thesis. The N15 dataset is demonstrated as a promising resource of historical climate ensemble. This work is the first demonstration of using the N15 dataset to account for the calibration period climate input uncertainty in hydrologic model calibration. It provides realistic, spatially and temporally correlated precipitation and temperature uncertainty estimates for the contiguous United States, northern Mexico, and southern Canada. The ensemble is straightforward to use for both lumped and distributed hydrologic models. Moreover, its climate ensemble generation method enables producing more than 100 members by sampling more random fields from the standard normal distribution. The climate ensemble generation can also be extended to areas out of its current coverage where there are gauge measurements (readers can contact the N15 dataset developers to get the climate ensemble generation code). Future work should investigate the use of other climate ensemble products or ensemble generation methods according to research area and data availability.

A large number of synthetic and real case studies presented here assessed the framework against two benchmark calibration approaches. One is the optimization-based DDS calibration that does not consider any source of uncertainty, and another is the uncertainty-based DDS-AU calibration that considers parameter uncertainty. Assessment of the framework against the benchmark approaches focused on validation period performance. Unlike most other synthetic calibration studies involving climate data uncertainty, this research is based on large-sized sample of experiments, so the results provide a better assessment of the average performance of the framework.

Synthetic and real case study results demonstrated that the ensemble climate based calibration framework outperforms the traditional calibration approaches in both flow predictions and parameter estimates. As to flow predictions, the framework effectively reduces the inaccurate flow predictions caused by the poor-quality climate ensemble members and improves the accuracy (KGE) and reliability of the ensemble flow predictions. For parameter estimates, accounting for climate data uncertainty in model calibration generates more robust parameter estimates since the synthetic experiments show true parameter values are more frequently recovered compared to calibrations using deterministic measured climate. Moreover, the flow predictions improvements are generated using the same (or less) computational effort for model calibration and occur despite the fact that the proposed calibration framework generates far fewer behavioral parameter sets (72% and 24% less for synthetic and real cases, respectively, on average) than the benchmark calibration approach. Propagating uncertainty with a smaller parameter ensemble size is a definite advantage. In terms of the climate uncertainty estimation, the synthetic and real cases demonstrate

the positive impacts of the proposed framework filtering out the poor-quality climate measurement and refining the climate uncertainty estimation.

The experiments also reveal two interesting findings based on the real case studies. The case studies demonstrate that it is insufficient to only account for climate data uncertainty in the calibration period or in the validation period (or similarly, the forecasting period). For example, in 11 of 20 case studies, the proposed framework generated superior quality ensemble flow predictions (equal or better reliability with smaller spread in predictions) compared to accounting for climate uncertainty in only the validation period. Importantly, in the other nine case studies, the proposed framework was never inferior to any other calibration approaches (e.g., there were no results showing an alternative approach generated better reliability with a smaller spread). Moreover, respecting parameter uncertainty dependency on input data in uncertainty propagation generates tighter, no less accurate (KGE), and no less reliable prediction ensembles than the propagation without the respect of parameter uncertainty dependency on input data.

The proposed climate ensemble based calibration framework is flexible enough to work with a variety of sampling algorithms and both lumped and distributed hydrologic modeling applications that deal with climate data uncertainty. It is also easily implemented on supercomputing clusters and parallelized where each model calibration process can be implemented independently (e.g., calibrating while climate inputs are fixed to a sampled prior climate ensemble member).

In the future, more experiments are needed to validate the proposed framework on distributed models and high dimensional parameter estimation problems. Distributed hydrologic models have the capacity to incorporate a variety of spatial configurations and realistically represent spatial heterogeneity, so distributed models potentially have greater predictive performance improvements than lumped hydrologic models when using the climate ensemble based model calibration framework. Moreover, it is worth investigating the effectiveness of the proposed framework in high dimensional problems. A foreseeable challenge is handling the increased computational cost induced from using ensemble climate inputs. If users do not want to sacrifice the parameter estimation quality behind each sampled climate input, a currently viable solution is using parallel computation.

Chapter 4.
Efficient Treatment of Climate Data Uncertainty in Ensemble Kalman Filter (EnKF) based on an Existing Historical Climate Ensemble Dataset

This chapter is a mirror of the following published article with minor changes to increase its consistency with the body of the thesis and to avoid redundant material. Changes were only made in the Summary (abstract). References are unified at the end of the thesis.

Liu, H., Thiboult, A., Tolson, B., Anctil, F., Mai, J., 2019. Efficient treatment of climate data uncertainty in ensemble Kalman filter (EnKF) based on an existing historical climate ensemble dataset. J. Hydrol. 568, 985–996. https://doi.org/10.1016/J.JHYDROL.2018.11.047

Summary

Successful data assimilation depends on the accurate estimation of forcing data uncertainty. Forcing data uncertainty is typically estimated based on statistical error models. In practice, the hyper-parameters of statistical error models are often estimated by a trial-and-error tuning process, requiring significant analyst and computational time. To improve the efficiency of forcing data uncertainty estimation, this chapter proposes the direct use of existing ensemble climate products to represent climate data uncertainty in the ensemble Kalman filter (EnKF) of flow forecasting. Specifically, the Newman et al. (2015) dataset (N15 for short), covering the contiguous United States, northern Mexico, and southern Canada, is used here to generate the precipitation and temperature ensemble in the EnKF application. This chapter for the first time compares the N15 generated climate ensemble with the carefully tuned hyper-parameters generated climate ensemble in a real flow forecasting framework. The forecast performance comparison of 20 Québec catchments shows that the N15 generated climate ensemble yields improved or similar deterministic and probabilistic flow forecasts relative to the carefully tuned hyper-parameters generated climate ensemble. Improvements are most evident for short lead times (i.e., 1-3 days) when the influence of data assimilation dominates. However, the analysis and computational time required to use N15 is much less compared to the typical trial-and-error hyper-parameter tuning process.

Section 4.1 gives the definition of hyper-parameter tuning and reviews its implementations, advantages, and disadvantages in the EnKF applications. Section 4.2 describes in detail the EnKF method, forecasting experiments, and two comparative approaches of generating climate ensembles. Section 4.3 presents the comparison results and discussion of the climate ensembles and the flow forecasts over 20 Québec catchments. Conclusions and future work are found in Section 4.4.

4.1. Introduction

Ensemble Kalman filter (EnKF) is a sequential data assimilation technique that was proposed by Evensen (1994) as an alternative to the extended Kalman filter. The EnKF uses an ensemble of simulations to

represent the distribution of the system state and replaces the covariance matrix by the sample covariance. It is hence well suited to highly nonlinear models (catchment hydrologic models in our case) as noted in other studies (McMillan et al., 2013; Reichle and Koster, 2003). The EnKF has many variants, such as using a pair of ensemble Kalman filters (Houtekamer and Mitchell, 1998), a hybrid techniques that combine the EnKF with the 3D variational method (Hamill and Snyder, 2000) or with the variance redactor (Heemink et al., 2001), an ensemble square root filter (Tippett et al., 2003), and a bias-aware retrospective EnKF (Pauwels et al., 2006). These methods all need an ensemble of simulations to represent the state ensemble and model error covariance.

The background error of the state ensemble, before updating, includes internal error and external error. Both are worth considering in the state ensemble generation of the data assimilation. The internal error is introduced by the use of imperfect initial conditions, and the external error refers to the model deficiency (Evensen, 1994). In the EnKF, an ensemble of initial conditions can be generated by adding random noise to the best guess initial conditions or by repeating the warm-up procedure with changed forcing data and/or model parameters (Evensen, 1994; Reichle and Koster, 2003). In the simplest form, the EnKF only accounts for the background error associated with initial conditions (Evensen, 2003; Hamill and Snyder, 2000; Tippett et al., 2003). The external error can be incorporated by either treating the model deficiency as a whole and adding random noise to deterministic model outputs, or explicitly accounting for different sources of model errors, such as model parameter errors, data errors, and structure errors (Del Giudice et al., 2015; Kumar et al., 2016; Liu and Gupta, 2007). This paper focuses on explicitly accounting for measured climate data uncertainty in the EnKF, which is one part of the external error. For the remainder of this paper, climate data uncertainty and climate ensemble will both refer to the measured historical climate. In contrast, ensemble flow forecasting can also involve climate forecast uncertainty and a corresponding forecasted climate ensemble.

In the EnKF applications of flow forecasting, the most common way of explicitly accounting for climate data uncertainty is treating climate variables as random variables and perturbing climate variables with stochastic errors. In most studies, the error is additive or multiplicative and is assumed to be Gaussian with a predefined constant or proportional variance (Khaki et al., 2017; Rasmussen et al., 2015; Weerts and El Serafy, 2006). Some other probability distributions and stochastic processes are also utilized to generate climate errors (Abaza et al., 2014b; Dunne and Entekhabi, 2006; Eicker et al., 2014; Leisenring and Moradkhani, 2011). Although the predefined error models are easy to construct, they may not reflect the best estimates of the true climate. For example, the common multiplicative stochastic error model approach (e.g., Kavetski et al., 2006a) has the inherent deficiency that it is unable to quantify measurement uncertainty when no rainfall is recorded, which can be especially important in poorly gauged areas (Wright et al., 2017). Another problem is that the distribution variances and stochastic model parameters are often

subjectively determined based on the order of magnitude or user's experience of uncertainty (Rasmussen et al., 2015; Reichle et al., 2002b).

A solution to reducing the subjectivity in determining climate errors is hyper-parameter tuning, also known in the literature as filter calibration, filter tuning, and EnKF optimization (e.g., Khaki et al., 2017; Reichle and Koster, 2003; Thiboult et al., 2016). In the context of the EnKF, hyper-parameters refer to the parameters of the prior error distributions. Hyper-parameter tuning is a process that recursively tries various sets of hyper-parameter values until the optimal filter performance or forecast performance is found. A typical example is in Reichle and Koster (2003) where a lognormal distributed error is used to perturb measured precipitation values. The standard deviation of the error distribution is determined by trying a selection of standard deviation values until the best filter performance is achieved. The filter performance is assessed by the root mean square error (RMSE) of the aggregated difference between the true state and its EnKF estimate over all catchments. In addition, many studies conduct climate relevant hyper-parameter tuning with other processes to improve the characterization of the background error, such as adjusting the hyper-parameters of system response observation errors (e.g., streamflow errors) (Clark et al., 2008; Reichle and Koster, 2003; Wang et al., 2017), and choosing ensemble size and state variables (Thiboult et al., 2016; Wang et al., 2017).

In addition to the manual hyper-parameter tuning, there are some advanced approaches to reduce the subjectivity of climate uncertainty estimation in flow prediction. The main idea behind these advanced approaches is to infer the climate relevant hyper-parameters with hydrologic model parameters based on automatic calibration algorithms. For example, in Bayesian inference, the input error is expressed by an error model. The likelihood function is adjusted to incorporate the input error so that the hyper-parameters can be inferred with hydrologic model parameters via Bayesian inference (Del Giudice et al., 2016; Kavetski et al., 2006b, 2006a, Renard et al., 2011, 2010; Sikorska et al., 2012). Another example is to simultaneously conduct model calibration and data assimilation (e.g., EnKF and particle filter). The hyper-parameters and hydrologic model parameters are updated either simultaneously with the states within the assimilation (Moradkhani et al., 2005; Salamon and Feyen, 2010, 2009) or out of each assimilation loop (Vrugt et al., 2005).

These advanced approaches are essentially calibration algorithms that explicitly consider climate uncertainty in parameter inference. In contrast, the previously introduced hyper-parameter tuning is separate from model calibration and is implemented with fixed hydrologic model parameters. To our knowledge, the manual hyper-parameter tuning is more popular than any advanced approaches in the EnKF based flow forecasting. The main reason is that it is still uncommon for people to explicitly consider climate data uncertainty in model calibration, so the advanced approaches have not been widely applied in practice. Moreover, very few of these advanced approaches have been set up and validated in the real flow

forecasting with forecast climate and data assimilation. For instance, the Bayesian inferred climate hyper-parameters are rarely used to generate the climate ensemble for the EnKF. In contrast, in the literature, there are numerous studies adopting the hyper-parameter tuning in EnKF based flow forecasting (e.g., see Abaza et al., 2014a; Li et al., 2014; McMillan et al., 2013; Noh et al., 2014; Reichle et al., 2002; Reichle and Koster, 2003; Thiboult et al., 2016). For example, Thiboult et al. (2016) used hyper-parameter tuning to estimate climate uncertainty because they determined hydrologic model parameters before data assimilation without considering climate uncertainty. Therefore, the tuned hyper-parameter approach is taken as the baseline approach to compare our new approach with in this chapter.

Although hyper-parameter tuning largely solves the subjectivity problem of determining climate errors, it has three limitations. The first is the intensive time and computational cost that users have to spend in the iterative application of data assimilation and forecasting to evaluate filter or forecast performance and find the optimal hyper-parameters for each case study (McMillan et al., 2013; Noh et al., 2014; Slater and Clark, 2006). This issue is due to the ad hoc nature of the hyper-parameter tuning operation. The second limitation is that hyper-parameter tuning mixes the climate uncertainty estimation with the data assimilation and flow forecasting processes. Climate data uncertainty is mostly caused by measurement errors, so its uncertainty estimation depends on measurement errors. However, hyper-parameter tuning determines climate uncertainty after running data assimilation and flow forecasting, the resultant hyper-parameter values may vary with the factors, such as the tuning and data assimilation method, and the climate forecast, which are irrelevant to the climate variable measurement. A consequence of this issue is that sometimes the climate errors are overestimated to compensate other model or initial condition errors and to eventually ensure good filtering and forecast performance (Clark et al., 2008; Evensen, 2007; Thiboult and Anctil, 2015). This is essentially getting the right results for wrong reasons. The third issue is the lack of consideration of the spatiotemporal correlation in generating climate errors. It is easy to understand that accounting for the spatiotemporal correlation gives a better description of the true climate. A better climate ensemble gives a better description of background error and thereby a better filter performance (McMillan et al., 2013; Rasmussen et al., 2015; Reichle and Koster, 2003). In practice, most EnKF applications neglect the climate spatiotemporal correlation (e.g., Abaza et al., 2014; Eicker et al., 2014; Rasmussen et al., 2015; Thiboult and Anctil, 2015; Whitaker and Hamill, 2002). Part of the reason is for simplicity, but the more important reason is that the spatiotemporal correlation characteristics of the uncertain climates are unknown beforehand and hard to quantify (Rasmussen et al., 2015). In the very few studies that account for the correlation(s), the temporal correlation is typically modelled with an autoregressive model of order one, and the spatial correlation is computed by the nested grid approach or the Fourier transform (Clark et al., 2008; Reichle and Koster, 2003; Tangdamrongsub et al., 2015).

Given these limitations of hyper-parameter tuning, a small number of studies have tried to avoid it by generating a climate ensemble (and a system response observation ensemble) before the EnKF phase. Slater and Clark (2006) and Clark et al. (2006) generated precipitation and temperature ensembles prior to the EnKF, based on a geo-statistical method introduced by Clark and Slater (2006). Huang et al. (2017) directly used the 100 members of a historical ensemble climate dataset developed by Newman et al. (2015) to force their hydrologic model in the EnKF. The latter dataset is generated by following the geo-statistical method of Clark and Slater (2006) but making several modifications, among which the foremost is incorporating the temporal correlation in the spatially correlated random field generation (Newman et al., 2015). For short, the Newman et al. (2015) dataset is referred to as N15 in the remainder of this paper. Although N15 has been used in some applications, more precisely, seasonal streamflow simulations (not forecasting), it has not yet been established that it yields better streamflow forecasting than the carefully tuned hyper-parameters based climate ensemble. The answer to this question has wide implications for the applications of the EnKF and its variants. If N15 produces the practically same forecasting results as the carefully tuned hyper-parameters do, then the subjective and arduous hyper-parameter tuning practice can be eliminated. The saved time can instead be used to further enhance forecast performance in other ways, such as improving the model parameters and structure, and taking into account the other model errors in filter calibration.

Therefore, the objectives of this chapter are to: (1) compare the climate ensemble generated by N15 with the climate ensemble generated by carefully tuned hyper-parameters, and (2) compare their flow forecast results over a large number of catchments in the EnKF based ensemble flow forecasting. This is the first study to compare the N15 generated ensemble with the carefully tuned hyper-parameters generated climate ensemble in a real flow forecasting framework. The tuned hyper-parameters and corresponding flow forecasting results are taken from Thiboult et al. (2016), and relevant details are provided here in Section 4.2 and Section 4.3.

4.2. Methods and Data

4.2.1. Ensemble Kalman Filter

This section provides a brief summary of the EnKF algorithm. More information about the EnKF equations and mathematical background can be found in Evensen (2003) and Houtekamer and Mitchell (2001). The state vector $\mathbf{X}$ evolves according to:

$$\mathbf{X}_t^- = \mathbf{M}_t(\mathbf{X}_{t-1}^+, \mathbf{U}_t, \theta) + \boldsymbol{\eta}_t \tag{4-1}$$

where $\mathbf{X}^-$ and $\mathbf{X}^+$ represent the prior and posterior estimates of the state, respectively. $\mathbf{M}$ is the non-linear forward operator forced by the previous state, the climate input $\mathbf{U}$, and the model parameter θ. $\boldsymbol{\eta}$ is the model error due to uncertainties in model structure, model parameters, initial conditions, and input data.

The state is transformed to the system response observation $\mathbf{Y}$ by

$$\mathbf{Y}_t = \mathbf{H}_t(\mathbf{X}_t^-) + \boldsymbol{\epsilon}_t \tag{4-2}$$

where $\mathbf{H}$ is the observation operator that converts the model state to the observation. $\mathbf{H}(\mathbf{X}^-)$ is the prior estimate of the system response. $\boldsymbol{\epsilon}$ is the response observation error.

When a response observation is available, the model state can be updated as a weighted average between the prior state and the difference between the prior estimate and observation of the system response:

$$\mathbf{X}_t^+ = \mathbf{X}_t^- + \mathbf{K}_t(\mathbf{Y}_t - \mathbf{H}_t(\mathbf{X}_t^-)) \tag{4-3}$$

where $\mathbf{K}$ is the Kalman gain. $\mathbf{K}$ functions as the weight in a state update and is calculated by:

$$\mathbf{K}_t = \mathbf{P}_t\mathbf{H}_t^T(\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T + \mathbf{R}_t)^{-1} \tag{4-4}$$

where $\mathbf{P}$ is the covariance of the state error, and $\mathbf{R}$ is the response observation error covariance.

When solving the Kalman gain, Houtekamer and Mitchell (2001) propose calculating $\mathbf{PH}^T$ and $\mathbf{HPH}^T$ directly from the ensemble members, rather than calculating each element of Equation (4-4).

$$\mathbf{P}_t\mathbf{H}_t^T = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{X}_{t,i}^- - \overline{\mathbf{X}_t^-})[\mathbf{H}(\mathbf{X}_{t,i}^-) - \overline{\mathbf{H}(\mathbf{X}_t^-)}]^T \tag{4-5}$$

$$\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T = \frac{1}{N-1}\sum_{i=1}^{N}[\mathbf{H}(\mathbf{X}_{t,i}^-) - \overline{\mathbf{H}(\mathbf{X}_t^-)}][\mathbf{H}(\mathbf{X}_{t,i}^-) - \overline{\mathbf{H}(\mathbf{X}_t^-)}]^T \tag{4-6}$$

where N is the ensemble size ($i = 1, \dots, N$), and

$$\overline{\mathbf{X}_t^-} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_{t,i}^- \tag{4-7}$$

$$\overline{\mathbf{H}(\mathbf{X}_t^-)} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{H}(\mathbf{X}_{t,i}^-) \tag{4-8}$$

Since the focus of this research is investigating the influence of different climate ensembles on flow forecasting, it is important to clarify how the change of climate inputs affects the EnKF. Changes in the input $\mathbf{U}$ will be firstly passed to the prior state $\mathbf{X}^-$ according to Equation (4-1). Change of state estimate $\mathbf{X}^-$ will then affect the prior estimate of system response $\mathbf{H}(\mathbf{X}_t^-)$ based on Equation (4-2) and finally affect the state update based on Equation (4-3).

4.2.2. Hydrologic Model

Thiboult et al. (2016) used 20 hydrologic models to account for model structure uncertainty in ensemble flow forecasting, arguing that different hydrologic models may compensate each other in terms of the overall forecast performance. Since the focus of this chapter is on the climate ensemble, the multi-model approach is not retained to ensure that the different sources of uncertainty are disaggregated. A single hydrologic model, GR4J (Génie Rural à 4 paramètres Journalier) model, corresponding to the 5th model of Thiboult et al. (2016), is chosen to provide a fair comparison of the different climate ensembles.

Here the GR4J model is same as in Chapter 3. Descriptions of the hydrologic processes and parameterizations of the GR4J model can be found in Section 3.3.2 of Chapter 3 and Appendix A3-1. There is a total of six parameters in the hydrologic model - four in GR4J and two in Cemaneige. The parameter

values for each catchment are taken from Thiboult et al. (2016). In the latter, the optimal parameter set is derived by minimizing the root mean square error between the simulated and observed flows in 1990-2000 with the shuffled complex evolution calibration method (Duan et al., 1992).

4.2.3. Research Area and Data

This research is conducted on the same 20 catchments as in Chapter 3. The list of main characteristics of the 20 catchments can be found in Table 3-1.

Same as Chapter 3, the measurements of streamflow, precipitation and maximum and minimum temperatures are provided by the Direction de l'Expertise Hydrique. Precipitation and temperature data are gridded and generated by Kriging interpolation over a 0.1° grid cell given station measurements. According to the Québec climate monitoring program (Bergeron, 2016), precipitation is measured by tipping bucket gauge, heated tipping bucket gauge, or weighting rain gauge. The daily precipitation and temperature measurements are interpolated based on 392 meteorological stations over the domain from $43^{\circ}N$ to $53^{\circ}N$ and from $55^{\circ}W$ to $81.5^{\circ}W$. The 20 catchments of Table 3-1 are situated in the region where the station density is relatively high. The Kriging error is less than $0.1\ mm/d$ for precipitation and is around $0.1℃$ for both the maximum and the minimum temperatures.

Forecast precipitation and maximum and minimum temperatures are retrieved from the THORPEX interactive grand global ensemble (TIGGE) database. The raw forecast data are at 0.5° and 6-hour resolution. They are downscaled from 0.5° to 0.1° by a bilinear interpolation and aggregated from 6-hour to daily time step to improve spatial resolution and meet the time step requirement of the hydrologic model. The forecast climate has 50 members, and the forecast lead time is 9 days. In model application, the measured and forecast climates are lumped to the catchment scale by calculating the average of the all the grid cells within the catchment.

4.2.4. Forecasting Experiment

In the forecasting phase, simulation, data assimilation, and forecasting alternate as follows: (1) the model is forced with the measured climates up to the first day *t* of the forecast, (2) the state estimates are updated based on the measured flow with the EnKF, and (3) the model is forced with meteorological forecasts to generate hydrologic ensemble flow forecasts until *t*+9 days.

In Thiboult et al. (2016), the simulation and forecasting periods are November 1, 2003 to October 31, 2008 and November 1, 2008 to December 1, 2010, respectively. In the simulation period, the model is started with the same initial states (e.g., water levels of two stores) on November 1, 2003 and evolves until October 31, 2008 with the measured climate. In simulation, the EnKF is implemented to generate multiple state conditions as a consideration of the initial condition error for forecasting. In data assimilation, given a hydrologic model, Thiboult et al. (2016) explicitly addressed the model errors from initial conditions and

climate input data (i.e., precipitation and temperature) as well as the flow observation error in the EnKF, while they did not account for the model structure and parameter errors in an explicit manner. The EnKF ensemble size is 50. State variables are daily updated.

A forecasting system can be identified by the climate forecast and a collection of settings for the hydrologic model and the data assimilation technique. Since the EnKF is the focus of this chapter, the settings for the climate forecast and hydrologic model are identical from one forecasting system to another, and only the EnKF implementation is varied to compare the performances of three forecasting systems. The three systems are differentiated by their climate ensembles used in the EnKF. Two climate ensembles are generated by the carefully tuned hyper-parameters, while the third climate ensemble is generated from the N15 dataset. The detailed climate ensemble generation processes and their corresponding forecasting systems are explained below.

4.2.4.1. Traditional Ensemble Generation

This study used two forecasting systems of Thiboult et al. (2016): system H and system H'. More precisely, the subsets of systems H and H' since a single hydrologic model, not 20 hydrologic models, is used for forecasting. The two forecasting systems have been chosen because systems H and H' differ only in hyper-parameter magnitudes and are both examples of the hyper-parameter tuning approach. The two approaches represent two commonly used hyper-parameter tuning strategies in practice. Hyper-parameter tuning of system H requires more attention and computational costs, and hyper-parameter tuning of system H' only involves estimating hyper-parameters from the literature.

Systems H and H' are henceforth referred as the statistical specific (Ss) system and the statistical uniform (Su) system, respectively, in this research.

- The Ss system: its hyper-parameters are chosen to optimize forecast performance. As such, the hyper-parameters values here are artificially overestimated and compensate for other sources of uncertainty that are not explicitly accounted for in the EnKF. The optimal hyper-parameters of system Ss are specific to each catchment.
- The Su system: its hyper-parameters describe a more realistic estimate of climate and flow data uncertainties and exhibit more reasonable perturbation magnitudes. The hyper-parameters of system Su are uniform for all catchments.

A brief overview of their hyper-parameter tuning processes is provided below. More details can be found in Thiboult and Anctil (2015). In both systems, precipitation is perturbed by a Gamma distribution with the mean being the observation and standard deviation being a proportion of the observation. The proportion has three options (25%, 50%, and 75%). Temperature is perturbed by an additive error that follows a normal distribution with zero mean and two standard deviation options (2°C and 5°C). Maximum

and minimum temperatures are both perturbed by the same additive error random variable. Flow is also perturbed by a normally distributed additive error where the mean error is zero and the standard deviation is proportional to the observed flow at each time step, and the proportion has two options (10% and 25%). The temporal and spatial correlations of errors are not considered in the ensemble generation. All error distributions and their variances constitute the hyper-parameters of the streamflow forecasting experiment. In total, $3 \times 2 \times 2 = 12$ combinations of hyper-parameter values are tested in the hyper-parameter tuning experiments.

Thiboult et al. (2016) also identified the optimal state variables in the process of hyper-parameter tuning to further improve forecast performance. The GR4J model has two potential state variables to be updated in the EnKF. One is the water level of the production store (S), another is the water level of the routing store (R). Updating these states affects the model in different ways, especially regarding the time lag between state updating and the effect on simulated streamflow. Three state combinations (S, R, and both S and R) are tested with the hyper-parameter tuning.

In the hyper-parameter tuning of system Ss, each of the 12 hyper-parameters and state variables are tested. Forecast results are evaluated in terms of reliability and bias in Thiboult and Anctil (2015). Reliability is measured by the Normalized Root-mean-square error Ratio (NRR). Details of NRR are provided in Appendix A4-1. Bias is measured by the Nash Sutcliffe efficiency coefficient (NSE) between the observed flow and the forecasted flow ensemble median. The optimal combination of hyper-parameters and state variables is taken as the one that achieves the best NSE among the three best NRRs.

Table 4-1 summarizes the optimal hyper-parameters and state variables of 20 catchments of system Ss. Here it is worth clarifying the workload of tuning hyper-parameters for system Ss. As mentioned before, there are 12 combinations of hyper-parameters, and three possible choices for the state variables (R, S, and both R and S). This means that there are $12 \times 3 = 36$ combinations for each catchment to be tested. Since we worked on 20 catchments, this requires $36 \times 20 = 720$ ensemble flow forecasting experiments. This is still tolerable as we work with GR4J which only has two state variables, but in the case where one uses a hydrologic model with more state variables, the number of flow forecasting experiments increases dramatically as the number of combinations per catchment model is given by $2^r - 1$, where r is the number of state variables.

Table 4-1. Optimal hyper-parameters and state variables of 20 catchments of the statistical specific Ss system (from Thiboult et al. (2016)).

Catchment No.	Precipitation distribution standard deviation proportion	Temperature distribution standard deviation (℃)	Flow distribution standard deviation proportion	State variables
1	0.75	2	0.1	S-R*
2	0.75	2	0.1	S-R
3	0.25	5	0.1	R
4	0.75	2	0.1	R
5	0.75	2	0.1	S-R
6	0.75	2	0.1	R
7	0.75	2	0.1	S-R
8	0.75	5	0.1	R
9	0.75	2	0.1	S-R
10	0.75	2	0.1	S-R
11	0.75	5	0.1	S-R
12	0.5	2	0.1	S-R
13	0.75	2	0.1	S-R
14	0.5	2	0.1	S-R
15	0.75	2	0.25	S-R
16	0.75	2	0.1	S-R
17	0.75	2	0.1	R
18	0.75	2	0.25	S-R
19	0.75	2	0.1	S-R
20	0.75	2	0.1	S-R

* S-R refers to both S and R.

The hyper-parameters of system Su are a simpler version of the hyper-parameters of system Ss and are used to describe climate uncertainty more realistically (Thiboult et al., 2016). The standard deviation proportion of precipitation distribution is 25%, the standard deviation of temperature distribution is 2℃, the standard deviation proportion of flow distribution is 10%. The state variables S and R are both updated for every catchment. The hyper-parameter magnitudes of Su are lower than or equal to those of Ss given in Table 4-1. The climate ensembles generated by the tuned hyper-parameters of Ss and Su are called the traditional ensemble to distinguish from the N15 generated climate ensemble.

4.2.4.2. Newman et al. (2015) Ensemble Generation

This section presents how to use N15 to generate the precipitation and temperature ensemble. The descriptions of N15 and its bias correction processes are the same as in Section 3.3.3 of Chapter 3 except that the this research used the Newman et al. (2015) dataset version 1.1, rather than version 1.0 in applications. Version 1.1 fixes the data corruption issues of the five ensemble members occurred in version 1.0.

In Thiboult et al. (2016), the EnKF ensemble size is 50, while N15 has 100 members. To provide a fair comparison of the systems, 50 members are randomly selected from the 100-member bias corrected N15 ensemble without replacement. The selected 50 members are referred to as the ensemble N and the corresponding forecasting system is called system N. Except the climate ensemble, system N uses the identical flow perturbations and state variables with system Su in the EnKF phase. Table 4-2 summaries the EnKF settings of the three systems Su, Ss, and N regarding data perturbations and state variables.

Table 4-2. EnKF data perturbations and state variables of systems Su, Ss, and N. $\tilde{P}, \tilde{T}$, and $\tilde{Q}$ represent the measured precipitation, temperature, and flow, respectively. c_P and c_Q are the standard deviation proportions of the precipitation and flow distributions, respectively. σ is the standard deviation of the temperature distribution. Note that system Ss contains all the optimum hyper-parameters and state variables of 20 catchments. See Table 4-1 for each catchment's optimum configurations within system Ss.

System	Statistical uniform (Su)	Statistical specific (Ss)	N
Precipitation perturbation	$\mathrm{P}\sim\mathrm{Gamma}(\frac{1}{{c_P}^2}, {c_P}^2 \cdot \tilde{P})$		Derived from N15
	$c_P = 0.25$	$c_P = 0.25, 0.5, 0.75$	
Temperature perturbation	$\mathrm{T}\sim\mathrm{Normal}(\tilde{T}, \sigma^2)$		
	$\sigma = 2$	$\sigma = 2, 5$	
Flow perturbation	$\mathrm{Q}\sim\mathrm{Normal}(\tilde{Q}, {c_Q}^2 \cdot \tilde{Q}^2)$		
	$c_Q = 0.1$	$c_Q = 0.1, 0.25$	$c_Q = 0.1$
State variables	S-R*	S, R, S-R	S-R

* S-R refers to both S and R.

4.2.5. Evaluation of Ensemble Flow Forecasts

The flow forecasts of each forecasting system are assessed from two perspectives: deterministic and probabilistic. The metrics below are used to measure flow forecast performances for systems Su, Ss, and N and are selected to be consistent with the metrics used in Thiboult et al. (2016).

When a deterministic flow forecast is required, the mean of the forecast ensemble is evaluated. The RMSE is used as the deterministic forecast evaluation metric.

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\bar{\hat{y}}_t - y_t)^2} \tag{4-9}$$

where T is the number of time steps of the evaluation period ($t = 1, \dots, T$). $\bar{\hat{y}}_t$ and y_t are the forecasted ensemble mean flow and the observed flow at time t, respectively. The RMSE is non-negative with the optimum of zero. Units of RMSE are the units of the flows.

When probabilistic flow forecasts are required, the ensemble flow forecast results are evaluated by comparing the ensemble flow forecasts with the measured flows. Being consistent with Thiboult et al. (2016), the mean continuous ranked probability score (MCRPS) is adopted as the ensemble forecast

evaluation metric. The continuous ranked probability score (CRPS) measures the proximity of the forecast distribution and the measurement distribution at a single time step. When the measurement is deterministic, the CRPS is calculated as (Gneiting and Raftery, 2007):

$$CRPS(F_t(\hat{y}), y_t) = \int_{-\infty}^{+\infty} (F_t(\hat{y}) - H\{\hat{y} - y_t\})^2 d\hat{y} \tag{4-10}$$

where F_t is the cumulative distribution function of ensemble flow forecast at time t. $\hat{y}$ is the forecasted flow. $H\{\hat{y} - y_t\}$ is the Heaviside function expressed as:

$$H\{\hat{y} - y_t\} = \begin{cases} 0, for\ \hat{y} - y_t < 0 \\ 1, for\ \hat{y} - y_t \geq 0 \end{cases} \tag{4-11}$$

The range of CRPS is non-negative with the best value of 0. Units of CRPS are the units of the flows.

MCRPS is the average CRPS over the entire evaluation period. MCRPS is also non-negative with the optimum of zero.

$$MCRPS = \frac{1}{T} \sum_{t=1}^{T} CRPS(F_t(\hat{y}), y_t) \tag{4-12}$$

In addition, reliability and spread are assessed for probabilistic flow forecasts. As in Thiboult et al. (2016), the mean absolute error of the reliability diagram (MaeRD) is used to estimate reliability. The reliability diagram is a graph of the observed frequency of an event plotted against the forecast probability of an event (Hartmann et al., 2016). In theory, a perfect forecast system will result in forecasts with a probability of X% being consistent with the observations X% of the time. Hence when plotting a reliability diagram, comparisons are made against the diagonal. A curve above the diagonal line denotes an over-dispersion, an under-dispersion is in the opposite case (Thiboult et al., 2016).

The MaeRD measures the average distance between the forecast probability and the observation probability over all quantiles of interest (Thiboult et al., 2016).

$$MaeRD = \frac{1}{K} \sum_{k=1}^{K} |P_{fcst,k} - P_{obs,k}| \tag{4-13}$$

where K is the number of quantiles of interest ($k = 1, \dots, K$). $P_{fcst,k}$ and $P_{obs,k}$ are the forecast probability and observed probability at the k^{th} quantile of interest, respectively. K equals to nine in this chapter. The MaeRD is dimensionless and non-negative with the optimal value of zero.

Spread is an indicator that should be considered along with reliability because a perfectly reliable forecast at the cost of excessively high dispersion is not desired. The spread equals to the square root of the average ensemble variance over the evaluation period (Fortin et al., 2014).

$$spread = \sqrt{\frac{1}{T} \sum_{t=1}^{T} Var(\hat{y}_t)} \tag{4-14}$$

$$Var(\hat{y}_t) = \frac{1}{N-1} \sum_{n=1}^{N} \left(\hat{y}_{t,n} - y_t\right)^2 \tag{4-15}$$

where N is the forecast ensemble size ($n = 1, \dots, N$). $\hat{y}_{t,n}$ is the n^{th} forecasted flow at time t. The spread is non-negative with an optimum of zero. Units of spread are the units of the flows.

4.3. Results and Discussion

The comparison results are presented in two sections. Section 4.3.1 compares the climate ensembles of systems Su, Ss, and N. Section 4.3.2 compares the flow forecasts of systems Su, Ss, and N.

4.3.1. Climate Ensemble Comparison

Taking precipitation as an example, Figure 4-1 compares the 50-member climate ensembles of systems Su, Ss, and N for the Aux Ecorces catchment (the 13th catchment of Table 3-1) in the period of July 15-31, 2009. The center line in each box represents the median of the ensemble (q_2), the top and bottom edges of the box are the first and third quartiles of the ensemble (q_1 and q_3). The whiskers above and below the box are the maximum and minimum of the ensemble excluding the outliers. Members are considered as outliers if their value is either greater than $q_3 + w(q_3 - q_1)$ or less than $q_1 - w(q_3 - q_1)$. w equals to 2.7 times of the ensemble standard deviation and corresponds to a 99.3% percent coverage for normally distributed data. Outliers are represented by the small cross outside the whiskers.

Figure 4-1 demonstrates that the three climate ensembles follow different distributions, and there is a stronger temporal correlation in the N ensemble than in other two ensembles, especially for the measured low and no precipitation events near high ones. For instance, a heavy rain event occurs on July 26, system N is the only system to observe a significant probability of rainfall for the preceding and following days (e.g., July 25, 27). Considering the temporal correlation, it is likely that the traditional ensembles miss parts of this rainfall event.

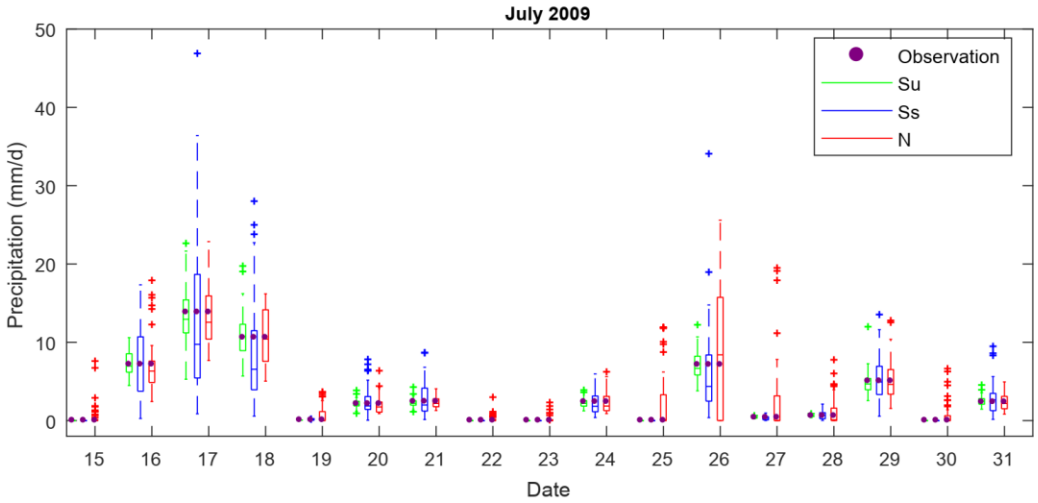

Figure 4-1. 50-member climate precipitation ensembles of systems Su, Ss, and N for the Aux Ecorces catchment (the 13th catchment of Table 3-1) and the period of July 15-31, 2009. See Table 4-2 for the descriptions of systems Su, Ss, and N.

In fact, there are a large number of non-zero precipitation estimates in ensemble N when the precipitations are zero in all members of the Su and Ss ensembles. For the entire forecasting period (761 days), a total of $761 \times 2 \times 20 = 30{,}440$ daily climate ensembles are generated for systems Su and Ss over 20 catchments. Among them, 5293 are such zero precipitation ensembles and account for 17.4% of the total ensembles. However, in system N, 94% of the corresponding 5293 N ensembles have at least one ensemble member with a non-zero precipitation. This demonstrates that ensemble N generates a non-zero probability of precipitation when ensembles Su and Ss estimate no precipitation (zero probability). As explained in Section 4.2.4.2, this capacity is due to the use of zero-to-one probability, not zero-or-one probability, to quantify the probability of precipitation, even for days with precipitation observations equal to zero. This is an appreciable feature of the N ensemble and is beneficial to generate more reliable predictions. Studies have shown that forcing data uncertainty dominate model errors (Carpenter et al., 2001; Slater and Clark, 2006). Climate is the most important driving factor for hydrologic models to generate a spread in the state variables, and this spread is essential to map the state variable space in the EnKF (Reichle et al., 2002b). With ensembles Su and Ss, when no precipitation is observed, the climate spread is zero. Therefore, the state variable spread would tend to collapse. However, ensemble N preserves a greater state variable spread, even during meteorological periods where precipitation is less likely to occur.

To investigate the overall ensemble spread cross the forecasting period, we compared climate ensemble N with climate ensembles Su and Ss, respectively, in terms of the metric spread. Recall that spread is calculated as the square root of the mean ensemble variance over the evaluation period (Equations (4-14) and (4-15)). Based on an analysis of spread across the 50 forecast climate members and 20 catchments for the entire forecasting period, and so a sample size of 50×20=1000, the precipitation spread of ensemble N is always greater than that of Su but smaller than that of Ss for 95% of the 1000 ensembles, the temperature spread of ensemble N is smaller than the spreads of Su and Ss for all the 1000 ensembles.

4.3.2. Flow Forecast Comparison

RMSE and MCRPS

Figure 4-2 shows the RMSE and MCRPS evaluation results of the forecasts issued by systems Su, Ss, and N for 20 catchments. The RMSE and MCRPS are used to demonstrate the deterministic and probabilistic forecast performances, respectively. For brevity, the 2nd, 4th, 5th, and 7th lead day results are not shown. The center of each radar plot corresponds to the optimal metric value, while the outer circle indicates the worse metric value for a given lead time. In Figure 4-2, system N yields improved or similar results relative to systems Su and Ss over all lead times. This is partly explained by the ability of system N to account for the uncertainty of low and no rainfall events, as mentioned in Section 4.3.1.

Moreover, the improvements of system N are notable for the 1st and 3rd lead days and diminish with increasing forecast horizons. This phenomenon indicates that the N15 generated climate ensemble is beneficial for improving short-term flow forecast. On the other hand, the decreasing relative improvement as the forecast horizon increases is understandable because data assimilation is known for its dominant influence on shorter horizons, while meteorological ensemble forecasts typically dominate longer ones (Thiboult et al. 2016).

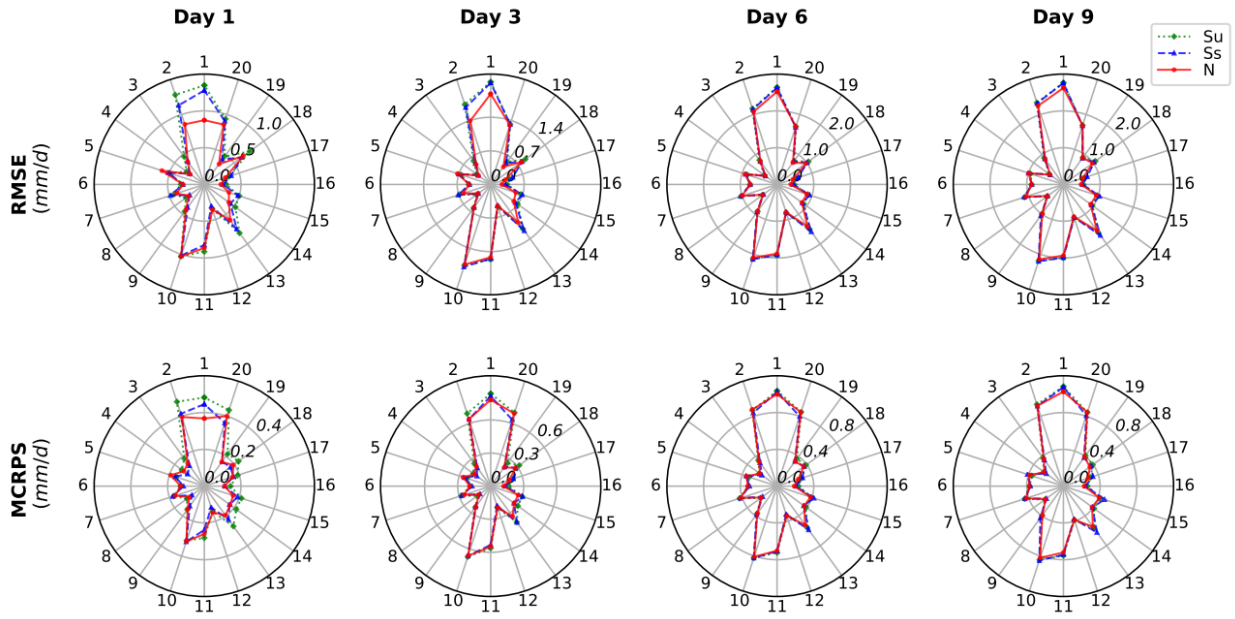

Figure 4-2. RMSE and MCRPS of systems Su, Ss, and N of 20 catchments for the 1st, 3rd, 6th, and 9th lead days flow forecasts. The RMSE is the root mean square error between the forecasted ensemble mean and the observed flows. The MCRPS is the mean continuous ranked probability score between the forecasted ensemble and the observed flows. See Table 4-2 for the descriptions of systems Su, Ss, and N. Each catchment is identified by the label on the outer edge of the circle and the catchment metric result is represented by the value on the corresponding spoke.

Reliability and Spread

Figure 4-3 details the reliability diagrams of systems Su, Ss, and N. All three systems are under-dispersion and are therefore over confident for all lead times. Possible reasons include inaccurate or biased meteorological forecasts or poorly calibrated model parameters. Another potential reason for only systems Su and N is the lack of full consideration of model errors in the EnKF, so the state variable space is not fully explored. To achieve reliability with a given forecast system, a post-processing of the meteorological and/or hydrologic forecasts would be necessary. Some operational guidance can be found in Abaza et al, (2017a), Boucher et al. (2015), Raftery et al. (2005), and Rana et al. (2014). This study did not conduct post-processing because post-processing would complicate the comparison of the effects of climate ensemble on flow forecasting.

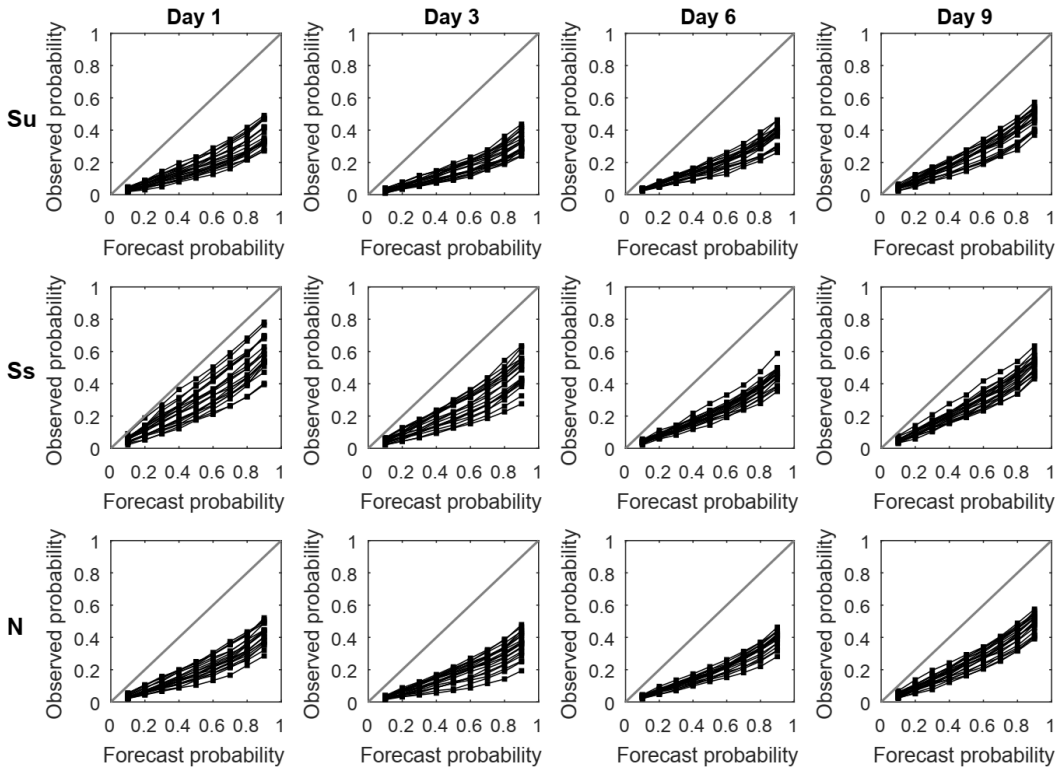

Figure 4-3. Reliability diagrams of systems Su, Ss and N for all 20 catchments and the 1st, 3rd, 6th, and 9th lead days flow forecasts. Each curve of a reliability diagram refers to a catchment. The diagonal line represents the perfectly reliable forecast. See Table 4-2 for the descriptions of systems Su, Ss, and N.

Larger differences among the systems are observed from the MaeRD and spread results as depicted in Figure 4-4. Compared with system Su, system N successfully reduces the MaeRD for more than 70% of the 20 catchments without significant spread changes. Thus, ensemble N produces more reliable flow forecasts than ensemble Su. Compared with system Ss, system N globally worsens the MaeRD. Nonetheless, this decrease of performance needs to be qualified since the reliability of system Ss is achieved through inflated hyper-parameters in order to get more reliable hydrologic ensembles by indirectly accounting for additional sources of uncertainty. Like in Figure 4-2, the metric differences between all three systems diminish with increasing forecast horizons as the data assimilation dominates short-term forecasts (Thiboult et al. 2016).

It is not unexpected that system Ss gets wider prediction spread than system N because system Ss uses higher and catchment specific perturbation magnitudes to account for other model errors in the EnKF. High climate uncertainty can be propagated to model outputs generating wide prediction intervals that contain more flow observations. From this point of view, the under-dispersion problem of system N can be fixed by taking into account other model errors that are specific to the catchment and the hydrologic model in the EnKF.

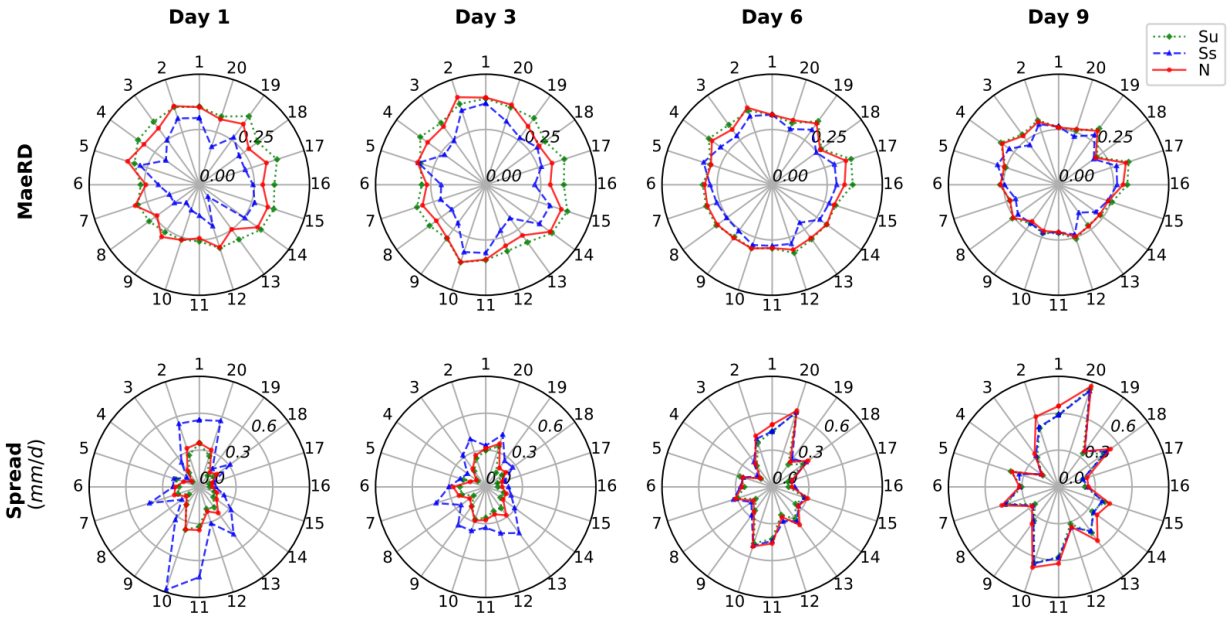

Figure 4-4. MaeRD and spread of systems Su, Ss and N of 20 catchments for the 1st, 3rd, 6th, and 9th lead days flow forecasts. The MaeRD is the average distance between the forecast probability and the observation probability over nine quantiles. The spread is the square root of the average variance of forecasted flow ensemble. See Table 4-2 for the descriptions of systems Su, Ss, and N. Each catchment is identified by the label on the outer edge of the circle and the catchment metric result is represented by the value on the corresponding spoke.

In terms of incorporating other model errors in the EnKF, one can consider the errors of other climate inputs, model parameter and model structure. For example, Reichle et al. (2002) add errors to wind speed, short- and long-wave radiative flux, and surface pressure in addition to precipitation and temperature. Clark et al. (2006) consider the model parameter uncertainty in the EnKF by generating 100 parameter sets corresponding to 100 climate ensemble members via the Monte Carlo Markov Chains.

Hydrograph

To visualize the flow forecast improvements due to N15, Figure 4-5 illustrates the first lead day flow forecasts of systems Su, Ss, and N for the Aux Ecorces catchment (the 13th catchment of Table 3-1) and a portion of the forecasting period. The ensemble mean and the 95% prediction intervals of forecasted flows are shown. Figure 4-5 confirms the forecast performance upgrade by using ensemble N. The deterministic flow forecasts are more consistent with the observations in system N than in systems Su and Ss. Also, more observations fall into the 95% prediction envelope of system N than that of systems Su and Ss. Specifically, looking at the peak flow forecasts in May 2009, the ensemble means of systems Su and Ss always underestimate the peaks. The 95% prediction intervals of systems Su and Ss cover only a small part of the observations. In comparison, system N improves the consistency between the ensemble mean and the peak, and its 95% forecast intervals contain almost all the peak flows.

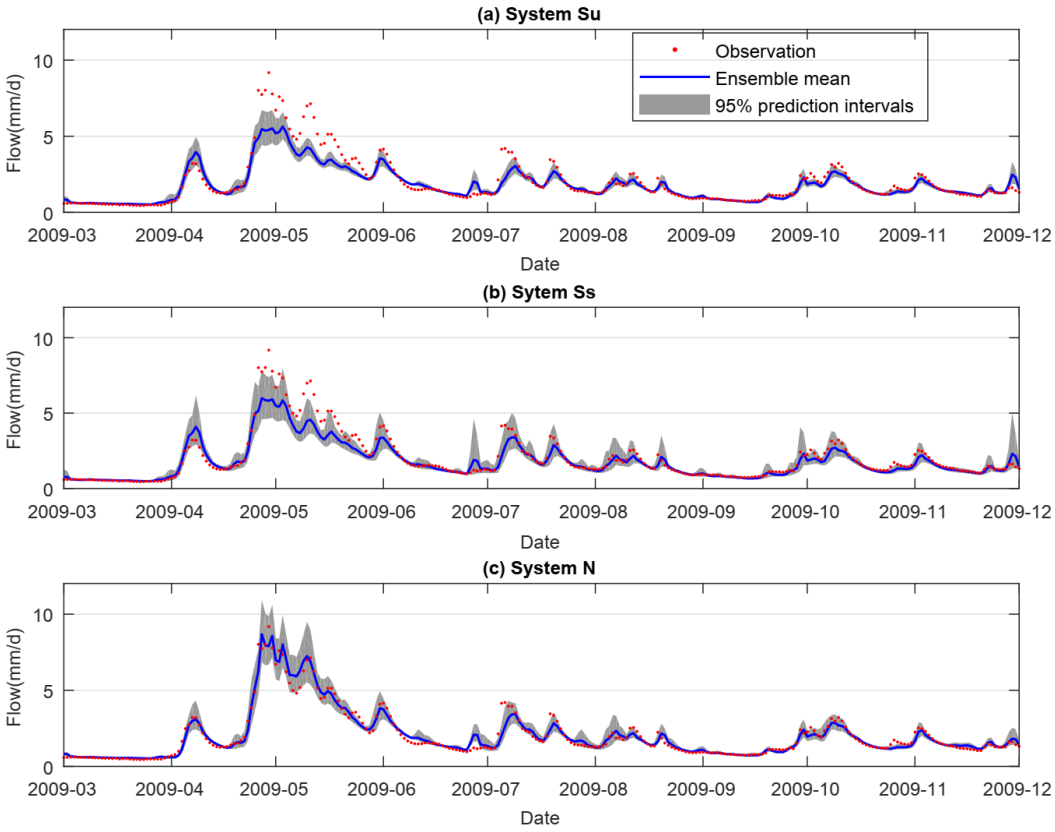

Figure 4-5. Flow forecasts of systems Su, Ss, and N for the Aux Ecorces catchment (the 13th catchment of Table 3-1) and a portion of the forecasting period. See Table 4-2 for the descriptions of systems Su, Ss, and N.

4.4. Conclusions and Future Work

This chapter proposed using an existing ensemble climate product to efficiently represent climate data uncertainty in the EnKF of short-term ensemble flow forecasting. Specifically, this chapter used the Newman et al. (2015) dataset, referred to here as N15, to represent measured precipitation and temperature uncertainties. To our knowledge, this is the first study to compare the N15 generated climate ensemble with the carefully tuned hyper-parameter generated climate ensemble in real ensemble flow forecasting. The tuned hyper-parameters are from two ensemble forecasting systems developed in Thiboult et al. (2016). One is system Su, a subset of system H' in Thiboult et al. (2016), that uses somewhat realistic perturbation magnitudes to represent climate uncertainty. Another is system Ss, a subset of system H in Thiboult et al. (2016), that uses unrealistically inflated hyper-parameters to implicitly compensate unaccounted model errors and optimize forecast performance. The hyper-parameters of the two systems are carefully tuned by Thiboult and Anctil (2015). Another highlight of this chapter is that a large number of experiments (20 Québec catchments) are explored to draw general conclusions.

The climate ensemble comparison showed that there is a stronger temporal correlation in the N15 generated climate ensemble than in the traditional ensembles Su and Ss, especially for the measured low and no precipitation events near high precipitations. N15 estimates a non-zero probability of precipitation when both traditional ensembles predict zero probability of precipitation. The ensemble flow forecast comparison of 20 catchments demonstrated that the N15 generated ensemble yields improved or similar deterministic and probabilistic flow forecasts relative to both traditional climate ensembles as measured by a variety of performance metrics (i.e., RMSE, MCRP, reliability diagram, MaeRD, and spread). The relative improvement of N15 derived forecasts is especially significant for short lead times (i.e., 1-3 days in our case) when the influence of the data assimilation dominates.

Our comparison study suggests that it is possible to eliminate the need for precipitation and temperature relevant hyper-parameter tuning from the EnKF by using an example historical ensemble climate product without losing flow forecast performance. Moreover, two gains are: (1) saving a great amount of time and computational cost from hyper-parameter tuning, and (2) partly disaggregating sources of uncertainty in the EnKF by explicitly addressing precipitation and temperature uncertainties. The saved efforts can be used to incorporate unaccounted model errors (e.g., model parameter and structure errors). Explicit consideration of uncertainty sources in the EnKF is critical in the pursuit of the right results for the right reasons.

N15 is an effective resource of historical climate ensemble for the contiguous United States, northern Mexico, and southern Canada. This resource provides realistic and spatiotemporally correlated precipitation and temperature uncertainty estimates. It is straightforward to be used for both lumped and distributed hydrologic model applications. Moreover, its climate ensemble generation method is applicable to real-time flow forecasting as its real-time ensemble generation code is available (readers can contact the N15 dataset developers to get the code). Future work should investigate the use of other historical climate ensemble products according to research area and data availability.

There are two areas for improving our research. First, our findings depend on the fact that we considered the initial condition uncertainty and the forcing and response data uncertainty in the EnKF. The forecasting performance improvements may not be as significant as they are presented here after adding other model errors (e.g., model structure and parameter errors) due to the compensation effect. In the future, more experiments are needed to explicitly incorporate model parameter and model structure uncertainties in the EnKF. This could be achieved by, for example, using multiple parameter sets per model and multiple hydrologic models in the state update, respectively. After taking into account other uncertainties, the authors will explore if the N15 based forecast broadens streamflow prediction intervals without deteriorating the overall forecast performance. The second future research area is to investigate the transferability of the current findings to other hydrologic models. Our research is built on a single hydrologic model (i.e., GR4J),

and though the results are promising, the N15 based forecasting needs to be validated for a variety of hydrologic models, in particular, distributed hydrologic models.

Chapter 5. Single-model and Multi-model Flow Forecasting Comparison

Summary

After finishing Chapter 3 and Chapter 4, a robust hydrologic model that explicitly and efficiently handles measured climate data uncertainty has been developed. This chapter will apply such a robust hydrologic model to ensemble flow forecasting and compare its forecast results with previously published multi-model forecast results. Current ensemble flow forecasting operation generally lacks the explicit consideration of different hydrologic modeling errors, especially parameter and data errors. However, disaggregating different hydrologic errors has been proved important by many studies from both parameter estimation and flow prediction perspectives. This chapter applies a single hydrologic model that is calibrated using the ensemble climate calibration framework of Chapter 3, thus explicitly considering parameter and calibration period climate data uncertainty, to an ensemble flow forecasting framework. Moreover, this chapter uses the Newman et al. (2015) dataset (N15) derived climate ensemble in the single-model based ensemble Kalman filter (EnKF) and compares the robust single model flow forecast with the traditional multi-model flow forecast. The reason for such a comparison is that, in the literature, the traditional multiple models (without uncertainty considerations in model calibration) have been tested and demonstrated higher skill and reliability than the single model in ensemble flow forecasting (Georgakakos et al., 2004; Hopson et al., 2010; Thiboult et al., 2016; Velázquez et al., 2011). This chapter investigates if such a conclusion holds when the uncertainty-based model calibration and the more realistic climate ensemble based data assimilation approaches are applied to improve the performance of the single-model ensemble forecast. The multi-model forecast results are taken from Thiboult et al. (2016). Comparison results show that the robust single hydrologic model generates improved high flow forecasts relative to the traditional multiple models. High flows here refer to the flows that are equal to or greater than the 90th percentile of the flow records. The results suggest that it is an advantage to use a robust single hydrologic model for high flow forecasts. Using a robust single hydrologic model also relieves modelers from calibrating multiple hydrologic models for operational flow prediction.

Section 5.1 reviews the treatments of the different uncertainty sources of ensemble flow forecasting. Section 5.2 describes the comparative hydrologic models, calibration methods, data assimilations, and flow forecasting configurations of the single-model and multi-model forecasting systems. Section 5.3 presents the comparison results for all flows and high flows, respectively, over 20 Québec catchments. Conclusions and future work are found in Section 5.4.

5.1. Introduction

The hydrologic forecasting paradigm has shifted from deterministic to probabilistic for pursuing a better, more detailed evaluation and description of the uncertainty linked to hydrologic forecasts. Historically, the

ensemble flow forecast was implemented based on historical observations, called extended streamflow prediction, assuming that past meteorological events can represent future events (Day, 1985). Nowadays the ensemble flow forecast uses forecasted climates to predict flows and often achieves reduced spread compared to the extended streamflow prediction (Cloke and Pappenberger, 2009). Moreover, the increasingly accessible ensemble meteorological forecast products benefit the hydrology community in issuing ensemble flow forecasts (Thiboult et al., 2016).

According to Dietrich et al. (2009), ensemble flow forecasts can be produced by three different ways. The first is single system ensemble. It is generated based on the variation of initial and boundary conditions, different input forcing data, different hydrologic rainfall-runoff model structures, and the variation of model parameters. The second is multi-model ensemble. It is the combination of the deterministic forecasts from different forecast models. The last is lagged average ensemble that combines current forecasts with forecasts from earlier model runs. This thesis focuses on the single system ensemble (i.e., each hydrologic model generates multiple forecasts by using ensemble climate forecast).

As to the uncertainty of the ensemble flow forecast, Krzysztofowicz (1999) decomposed the total uncertainty into three sources: operational, input, and hydrologic modeling. Operational errors are caused by erroneous or missing data, human processing errors, and unpredictable interventions and are usually not considered in practice. Input uncertainty is associated with the random errors in meteorological variables that are used to forecast flows with a hydrologic model and is represented by ensemble meteorological forecasts. Hydrologic modeling uncertainty includes all the errors related to hydrologic modeling processes (e.g., initial and boundary condition, model structure, parameter, and measurement data) and is represented by an ensemble of differently structured hydrologic models or an ensemble of parameter solutions. However, the two methods for expressing the hydrologic modeling uncertainty are both problematic. The hydrologic model ensemble has typically accounted for model structure errors only; and the parameter ensemble has not explicitly accounted for the various errors related to hydrologic modeling (Dietrich et al., 2009; Thiboult et al., 2016) .

Researchers determining model parameters prior to data assimilation have used a few ways to generate and utilize parameter ensembles in flow forecasting. Georgakakos et al. (2004) sampled 100 parameter sets from the feasible parameter space using a Monte Carlo approach and each parameter set is individually used to generate flow forecasts. All the flow forecasts conditioned on the 100 sampled parameter sets constitute the ensemble flow forecasts. Clark et al. (2006) generated a total of 100 parameter sets (one for each of the 100 climate ensemble members) using the Monte Carlo Markov Chain method (maximum 400 iterations). This procedure is advanced in terms of concurrently accounting for uncertainties in forcing data and parameters in parameter ensemble generation. In flow simulation (not forecasting), Clark et al. (2006) randomly sampled one forcing input and one parameter set from their 100-member ensembles, respectively,

to save computational costs from calculating all possible combinations of forcing and parameter ensembles (i.e., 100×100). Dietrich et al. (2009) sampled 20 parameters sets from the parameter ranges defined by optimization, stochastic methods, and expert knowledge and introduced real-time likelihood in constructing flow forecast bounds. A likelihood is assigned for each flow forecast ensemble member at each time step and is calculated according to the principle of Bayesian inference in comparison with the observed flow. The flow ensemble members with relatively high likelihood values are used to construct prediction uncertainty bounds. Georgakakos et al. (2004) and Dietrich et al. (2009) assumed parameter ensembles representing model parameter uncertainty only and did not consider other hydrologic modeling errors (e.g., calibration period data errors), while Clark et al. (2006) experiments are only limited to flow simulation. Note that data assimilation based joint state and parameter estimation is an option of considering various hydrologic modeling errors in parameter ensemble generation (e.g., see Moradkhani et al., 2012, 2005; Smith et al., 2013; Vrugt and Robinson, 2007), but simultaneous parameter estimation and data assimilation is beyond our scope here.

This chapter works in an ensemble flow forecasting framework (i.e., using forecasted climates) and focuses on generating parameter ensembles with the consideration of both parameter and calibration period climate data uncertainties. The latter can be realized by using the climate ensemble based calibration method developed in Chapter 3. Moreover, the Newman et al. (2015) dataset (referred to as N15) based EnKF, developed in Chapter 4, will also be used in single-model forecasting. Combining the above two methods leads to a robust single-model ensemble flow forecasting system. This robust single-model system will be compared with a traditional multi-model ensemble flow forecasting system. In the latter, each hydrologic model is assigned one parameter solution, and climate data are perturbed by assumed statistical error models in the EnKF. These or similar multiple models have been tested and demonstrated higher skill and reliability than any of the multiple models in past single-model and multi-model forecast comparison studies (e.g., Georgakakos et al., 2004; Hopson et al., 2010; Hsu et al., 2009; Madadgar and Moradkhani, 2014; Najafi and Moradkhani, 2015; Sharma et al., 2019; Thiboult et al., 2016; Velázquez et al., 2011). This chapter will investigate if such a conclusion holds when the uncertainty-based model calibration and the more realistic climate ensemble based data assimilation approaches are applied to improve the performance of single-model forecasts. The comparison study results will have wide implications for operational flow forecasting. If the robust single-model forecasts show similar or better performance than the multi-model forecasts, the robust single model can be deemed sufficient for ensemble flow forecasting in practice. This result will be welcome among water resources managers and flood forecasters because it is practically difficult to develop multiple accurate and often distributed hydrologic models for a watershed in their formal hydrologic ensemble prediction system.

The objectives of this research are to: (1) apply a robust single hydrologic model that is calibrated with an explicit consideration of parameter and climate data uncertainty to ensemble flow forecasting, and (2) compare the robust single-model and multiple-model flow forecasts over a large number of catchments. To investigate the separate and joint impacts of the uncertainty-based calibration and the N15 based data assimilation on forecasting, three single-model forecasting systems are constructed and compared with the multi-model forecasting system. The first single-model system only applies the parameter and climate data uncertainty based calibration. The second single-model system only applies the N15 based data assimilation. The third single-model system applies both the uncertainty-based calibration and the N15 based data assimilation and is specifically referred to as the robust single-model system in this chapter. The multi-model forecast results for comparison are taken from Thiboult et al. (2016), and relevant details are provided here in Section 5.2.4.

5.2. Methods and Data

This section describes the setups of the single-model and multi-model forecasting systems, respectively, including hydrologic models and data, model calibration and forecasting configurations, and evaluation metrics of flow forecasts. The multi-model forecasting descriptions are based on Thiboult et al. (2016).

5.2.1. Hydrologic Models and Data

The multi-model forecasting includes 20 different lumped hydrologic models to account for model structure uncertainty in forecasting. Information of the 20 hydrologic models is listed in Table 5-1. Note that Thiboult et al. (2016) used the same snow accumulation, melting module, and evapotranspiration calculation methods for all the 20 hydrologic models. The precipitation is calculated by the two-parameter snow accounting routine Cemaneige (Valéry et al., 2014). The potential evapotranspiration is estimated by a conceptual formula proposed by Oudin et al. (2005) based on air temperature and calculated radiation.

The single-model forecasting is built on the GR4J (Génie Rural à 4 paramètres Journalier) model, corresponding to the 5th hydrologic model of Table 5-1. Descriptions of the hydrologic processes and parameterizations of the GR4J model can be found in Section 3.3.2 of Chapter 3 and Appendix A3-1 of this thesis.

5.2.2. Research Area and Data

The single-model and multi-model forecasting systems are applied to the same 20 catchments as Thiboult et al. (2016) which are also the same as in Chapter 3 and Chapter 4. The list of main characteristics of the 20 catchments can be found in Table 3-1 of Chapter 3.

Table 5-1. Information of the 20 hydrologic models (from Thiboult et al. (2016))

Number	Model name	Number of optimized parameters	Number of reservoirs	Number of state variables
1	BUCKET	6	3	2
2	CEQUEAU	9	2	1
3	CREC	6	3	2
4	GARDENIA	6	3	3
5	GR4J	4	2	2
6	HBV	9	3	2
7	HYMOD	6	5	2
8	IHACRES	7	3	3
9	MARTINE	7	4	3
10	MOHYSE	7	2	1
11	MORDOR	6	4	3
12	NAM	10	7	5
13	PDM	8	4	3
14	SACRAMENTO	9	5	2
15	SIMHYD	8	3	2
16	SMAR	8	3	1
17	TANK	7	4	1
18	TOPMODEL	7	3	3
19	WAGENINGEN	8	3	1
20	XINANJIANG	8	4	4

The measured climate data (precipitation, and maximum and minimum temperatures) and flow data are provided by the Direction de l'Expertise Hydrique. Climate data are on the 0.1° grid and are generated by the Kriging interpolation given station measurements. In addition, the measured climate ensemble is derived from the Newman et al. (2015) dataset (version 1.1) and will be used in the climate ensemble based model calibration and the N15 based EnKF of the single-model forecasting. The descriptions of N15 climate ensemble generation method and its bias correction process are the same as in Section 3.3.3 of Chapter 3.

The forecasted precipitation and temperature data are retrieved from the THORPEX interactive grand global ensemble (TIGGE) database. The raw forecast data are at 0.5° and 6-hour resolution. They are downscaled from 0.5° to 0.1° by a bilinear interpolation and aggregated from 6-hour to daily time step to improve spatial resolution and meet the time step requirement of the hydrologic model. The forecast climate has 50 members, and the forecast lead time is 9 days. In model application, both the measured and forecast climates are lumped to the catchment scale by calculating the average of the all the grid cells within the catchment.

5.2.3. Model Calibration

In the multi-model forecasting, Thiboult et al. (2016) determined the optimal parameter values for each hydrologic model and each catchment by minimizing the root mean square error (RMSE) between the simulated and observed flows from January 1, 1990 to January 1, 2000 based on the shuffled complex evolution (SCE) calibration method (Duan et al., 1992). The number of optimized parameters for each of the 20 hydrologic models is listed in Table 5-1 (except the two common parameters in Cemaneige).

In the robust single-model forecasting, model calibration follows the same calibration period and objective function as in the multi-model forecasting. There is a total of six parameters in the single hydrologic model - four in GR4J and two in Cemaneige. Their parameter ranges and initial values are the same as in Thiboult et al. (2016) (Table 5-2).

Table 5-2. Parameters of the GR4J based single hydrologic model

Parameter	Description	Unit	Initial	Minimum	Maximum
X1	Maximum capacity of the production store	mm	100	10	1000
X2	The groundwater exchange coefficient	mm	0	-5	3
X3	One day ahead maximum capacity of the routing store	mm	90	5	300
X4	Time base of unit hydrograph UH1	day	8	1	16
X5	Snowmelt factor	--	0.5	0	1
X6	Cold content factor	mm/d	10	0	20

The difference between the multi-model calibration and the robust single-model calibration is in the treatment of parameter and measured climate data uncertainty in model calibration. The multi-model calibration does not consider the parameter and calibration period measured climate data uncertainties, while the robust single-model calibration considers both uncertainty sources. Specifically, the parameter uncertainty is considered by using the uncertainty-based calibration algorithm - dynamically dimensioned search-approximation of uncertainty (DDS-AU) (Tolson and Shoemaker, 2008). The climate data uncertainty is considered by using the historical climate ensemble as forcing data. The climate ensemble based model calibration framework that is proposed in Chapter 3 is adopted here to explicitly consider both uncertainty sources. Detailed implementations are explained below. Note that the framework is re-applied in this chapter, with assumptions precisely matching those in Thiboult et al. (2016), and thus calibration results in this Chapter are different than those in Chapter 3.

Step 1. Sample model parameters conditioned on each sampled climate ensemble member.

Assume the prior climate uncertainty is represented by the 101-member bias corrected climate ensemble (100 from the N15 climate ensemble and one from observation). Use each of the 101 climate ensemble members as input to the hydrologic model and implement the DDS-AU based parameter sampling. In DDS-AU, each climate ensemble member is used in four separate DDS optimization trials to get different

parameter solutions. As such, the calibration experiments yield a maximum of $101 \times 4 = 404$ candidate behavioral parameter sets. To be consistent with the initialization setup of the multi-model calibration of Thiboult et al. (2016), each DDS optimization trial is initialized with the same parameter values as shown in Table 5-2. However, each DDS optimization trial is assigned with a different random seed to ensure a different search pats, and thus a different final solution in DDS-AU. Each DDS optimization trial uses 400 model evaluations.

Recall that, in Chapter 3, the final best solution of each DDS optimization trial constituted the candidate behavioral parameter set and obtained under-dispersion flow predictions across the validation period (e.g., see Figure 3-7). The same under-dispersion results also occur here in ensemble flow forecasting. For demonstration, Figure 5-1a shows the first lead day flow forecast results of the Du Loup catchment (the 2nd catchment of Table 3-1) by using the final best solution per each DDS optimization trial to constitute the candidate behavioral parameter sets followed with the behavioral parameter selection. Table 5-3 summarizes the KGE, MCRPS, reliability, and spread evaluation results of the forecasts in Figure 5-1. As seen from Figure 5-1, the ensemble means match the observations well (KGE=0.86), but the 95% prediction intervals are quite narrow and closely surround the observations (spread=0.27 mm/d, reliability=0.26).

To explore the possible boundaries of the ensemble flow forecast given the single hydrologic model, small preliminary comparative experiments were conducted based on 20 random parameter sets from the parameter feasible ranges in ensemble flow forecasting. Using random parameters is not expected to be appropriate for flow forecasting but can give us a view of the possible forecast results without any model calibration. Using 20 random parameter sets is to ensure that the single-model system has the same total computational cost (i.e., number of model runs) as the 20 hydrologic models based multi-model system. Figure 5-1b shows the ensemble flow forecast results based on 20 random parameter sets for the same catchment of Figure 5-1a. The random parameter sets generate sufficiently wide (in fact, over-dispersion predictions) and reliable prediction intervals at the cost of decreasing the prediction accuracy (spread=1.42 mm/d, reliability=0.70, KGE=0.44). This experiment indicates that there must be some model calibration/behavioral parameter solution identification strategy that can produce better quality prediction bounds; something between Chapter 3 approach (Figure 5-1a) and the random solutions (Figure 5-1b) that overcomes the under-dispersion issue while retaining practically good prediction accuracy.

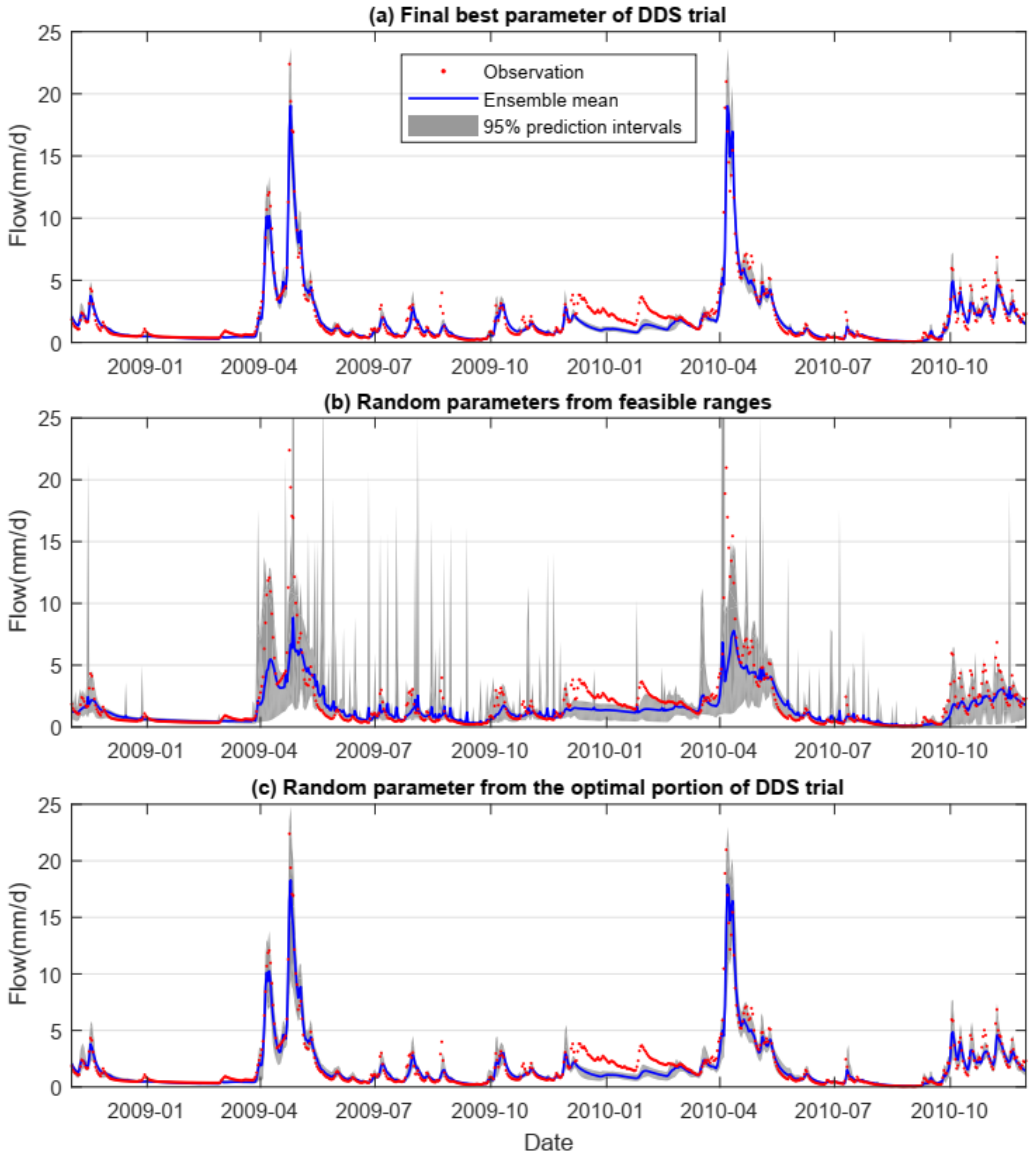

Figure 5-1. First lead day flow forecasts of the single-model forecasting system for the Du Loup catchment (the 2nd catchment of Table 3-1) of the entire forecasting period. The three panels represent different ways of selecting behavioral parameter sets, each with a parameter ensemble size of 20.

Table 5-3. KGE, MCRPS, reliability and spread evaluation results of the single-model forecasting systems for the Du Loup catchment (the 2nd catchment of Table 3-1) and for the first lead day.

Forecast system	KGE	MCRPS (mm/d)	Reliability	Spread (mm/d)
(a) Final best parameter of DDS trial	0.86	0.26	0.27	0.34
(b) Random parameters from feasible ranges	0.44	0.50	0.70	1.42
(c) Random parameter from the optimal portion of DDS trial	0.84	0.28	0.63	0.50

The under-dispersion of flow forecast ensemble indicates that the behavioral parameter solutions are not well distributed throughout the behavioral parameter space. To increase the distribution of the behavioral parameter solutions, we can use a different strategy to sample a candidate behavioral parameter

set from each DDS trial (not the final best per DDS trial). The premise here is that no additional model evaluations are needed and the number of model evaluations per DDS trial is unchanged (400). To solve this problem, Tolson and Shoemaker (2008) suggested randomly sampling a single behavioral parameter solution from the search history of each DDS optimization trial. However, to identify the behavioral parameter sets per DDS trial, a behavioral threshold needs to be subjectively determined for each DDS trial. This approach is contrary to the goal of reducing the subjectivity of model calibration process. Therefore, an alternative DDS optimization trial search history sampling strategy is defined here. Since the goal of increasing the behavioral parameter solution spread is to increase the flow prediction spread and reliability, we consistently used the optimal criteria-aggregation-based approach (Shafii et al., 2015) to identify the ideal parameter sampling portion for the DDS optimization trial. Detailed processes are explained below.

a) Divide each DDS optimization trial search history into ten portions. Assume a DDS search trial has a total of *N* parameter sets. The first portion includes the best 10% of the *N* parameter sets according to their calibration period objective function values (RMSE in our case). The second portion includes the best 20% of the *N* parameter sets, and so on, until the tenth portion which has all the *N* parameter sets.
b) For each portion, randomly select a parameter set. The sampled parameter sets from all the DDS trials and all the climate ensemble inputs are combined to constitute the candidate behavioral solutions. This process is conducted for each of the ten portions followed by the behavioral parameter selection (described in Step2 below) and the calculation of the calibration period flow prediction reliability and spread. The reliability and spread are calculated in the same way as in Step2 below.
c) For each portion, calculate its behavioral parameter sets corresponding distance-to-ideal (DTI) value. DTI is calculated as the Euclidean distance between (R, S) and (R_{best}, S_{best}). R_{best} and S_{best} are the individual best reliability and sharpness values, respectively, of the ten behavioral parameter sets of the ten portions and define the ideal point in objective space.
d) The portion with the minimum DTI is identified as the ideal parameter sampling portion for the DDS optimization trial.

Step 2. Identify the behavioral model parameter solutions (the filtering step).

Given the candidate parameter solutions conditioned on all the climate ensemble members, the second step is to identify the behavioral parameter sets using the optimal criteria-aggregation-based approach (Shafii et al., 2015) as described in Section 3.2.1 of Chapter 3.

In forecasting application, to ensure that the single-model system has the same total computational cost as the 20 hydrologic models based multi-model system, only 20 behavioral parameter solutions were applied to each catchment. The 20 parameter solutions were randomly selected from all the behavioral solutions with replacement.

Figure 5-1c shows the ensemble flow forecasts using the above described strategy for the same catchment of Figure 5-1 (a-b). In Figure 5-1c, the 95% prediction intervals effectively increase compared with those of Figure 5-1a, and the ensemble means match the observations very well (spread=0.5 mm/d, reliability=0.63, KGE=0.84). The same trend is evident in all the other catchments, though results are not shown here. These demonstrate the effectiveness of the proposed strategy. Our flow forecasting results with the robust single model (shown later) have been generated using this strategy.

5.2.4. Forecasting Experiment

In the forecasting phase, simulation, data assimilation, and forecasting alternate as follows: (1) the model is forced with the measured climate up to the first day *t* of the forecast, (2) the state estimates are updated based on the measured flow with the EnKF, and (3) the model is forced with forecast climate to generate hydrologic ensemble flow forecasts until *t*+9 days (Thiboult et al., 2016).

For all the comparison forecasting systems, the simulation period is from November 1, 2003 to October 31, 2008, and the forecasting period is from November 1, 2008 to December 1, 2010. In the simulation period, the model is started with the same initial states on November 1, 2003 and evolves until October 31, 2008 with the measured climate. Moreover, the EnKF is implemented in the simulation period to generate multiple state conditions as a consideration of the initial condition error for forecasting. In the forecasting period, data assimilation and forecasting processes alternate at each time step. The EnKF is conducted to update the state variable estimates at the daily base. The EnKF ensemble size is 50. The 50-member TIGGE forecast climate is used as model input to generate 9-day ensemble flow forecast.

The multi-model system and the robust single-model system use different strategies to perturb measured climate data in the EnKF. The multi-model forecasting system, corresponding to system H' of Thiboult et al. (2016), uses assumed statistical error models to perturb climate and flow data. The hyper-parameters are estimated from the literature to be consistent with the uncertainty estimates of observed climate and streamflow at catchment scale. The hyper-parameter values are uniform across all the 20 catchments. In detail, precipitation is perturbed by a Gamma distribution with the mean being the observation and standard deviation being 25% of the observation. Temperature is perturbed by an additive error that follows a normal distribution with zero mean and standard deviation of 2℃. Flow is perturbed by a normally distributed additive error with zero mean and standard deviation as 10% of the observed flow at each time step. Moreover, the state variables to be updated are also uniform across all the 20 catchments for each hydrologic model. The number of state variables to be updated in data assimilation is listed in Table 5-1. A detailed description of the state variable is provided in Thiboult and Anctil (2015) and Thiboult et al. (2016).

The robust single-model forecasting system uses the same flow error models and state variables as the multi-model forecasting for each hydrologic model and each catchment in the EnKF, but its climate

ensemble is directly derived from the N15 dataset. The N15 dataset is derived in the same way as described in Section 3.3.3 of Chapter 3. Compared with using assumed statistical error models, using the N15 dataset saves modelers from tuning hyper-parameters catchment by catchment and provides more realistic precipitation and temperature uncertainty estimates (e.g., zero-to-one probability of precipitation, spatiotemporal correlation).

In summary, there are two differences between the multi-model forecasting system and the robust single-model forecasting system. One is the incorporation of the calibration period hydrologic modeling uncertainty in forecasting, another is the generation of climate ensembles in the EnKF. The multi-model system considers the model structure error by using 20 hydrologic models, while the robust single-model system considers the model parameter and measured climate data errors by using 20 parameter solutions (for the same number of model runs as the multi-model system in forecasting). The multi-model forecasting uses the tuned hyper-parameters to generate climate ensembles in the EnKF, but the robust single-model forecasting uses the N15 dataset to generate climate ensembles in the EnKF.

To investigate the separate and joint impacts of the uncertainty based calibration and the N15 based data assimilation on the overall robust single-model forecasting performance, three different single-model forecasting systems are constructed. Table 5-4 summarizes the model calibration and the EnKF setups that distinguish the multi-model and the three single-model forecasting systems.

- Single-model (Calib): only the parameter and climate uncertainty based calibration is incorporated. The EnKF is implemented in the same way as the multi-model forecasting system that uses assumed statistical error models to perturb measured precipitation and temperature data.
- Single-model (EnKF): only the N15 based EnKF data assimilation is incorporated. The hydrologic model parameters are the same as in the multi-model forecasting system and are estimated without the consideration of parameter and climate data uncertainty.
- Single-model (Calib & EnKF): both the parameter and climate uncertainty based calibration and the N15 based data assimilation are incorporated. This single-model (Calib & EnKF) forecasting system is referred to as the robust single-model forecasting system.

Table 5-4. Model calibration and EnKF setups of the multi-model and three single-model forecasting systems

Forecasting system	**Model calibration**			**EnKF**
	Calib. algorithm	**Calib. period climate data**	**Parameter solution**	
Multi-model	SCE	Measured Climate	Single	Statistical Error Model
Single-model (Calib)	DDS-AU	Climate Ensemble	Multiple	Statistical Error Model
Single-model (EnKF)	SCE	Measured Climate	Single	Climate Ensemble
Single-model (Calib & EnKF)	DDS-AU	Climate Ensemble	Multiple	Climate Ensemble

5.2.5. Evaluation of Ensemble Flow Forecasts

Ensemble flow forecasts of each forecasting system are assessed from two perspectives: deterministic and probabilistic. When a deterministic flow forecast is required, the mean of the forecast ensemble is evaluated. The KGE is used as the deterministic forecast evaluation metric.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{5-1}$$

where r is the linear correlation coefficient between the simulated and true (observed) flows, α is the ratio between the simulated flow standard deviation and the observed flow standard deviation, and β is the ratio between the mean simulated and mean observed flows. The range of KGE is between negative infinity to one with the optimal value of one.

When probabilistic flow forecasts are required, the ensemble flow forecast results are evaluated by comparing the ensemble flow forecast with the measured flow. Being consistent with Thiboult et al. (2016), the mean continuous ranked probability score (MCRPS) is adopted as the ensemble forecasts evaluation metric. The continuous ranked probability score (CRPS) measures the proximity of the forecast distribution and the measurement distribution at a single time step. When the flow measurement is deterministic, the CRPS is calculated as (Gneiting and Raftery, 2007):

$$CRPS(F_t(\hat{y}), y_t) = \int_{-\infty}^{+\infty} (F_t(\hat{y}) - H\{\hat{y} - y_t\})^2 d\hat{y} \tag{5-2}$$

where F_t is the cumulative distribution function of ensemble flow forecasts at time t. $\hat{y}$ is the forecasted flow. y_t is the measured flow at time t. $H\{\hat{y} - y_t\}$ is the Heaviside function expressed as:

$$H\{\hat{y} - y_t\} = \begin{cases} 0, for\ \hat{y} - y_t < 0 \\ 1, for\ \hat{y} - y_t \geq 0 \end{cases} \tag{5-3}$$

The range of CRPS is non-negative with the best value of 0. Units of CRPS are the units of the flows.

MCRPS is the average CRPS over the entire evaluation period. MCRPS is also non-negative with the optimum of zero and shares the unit of the flow.

$$MCRPS = \frac{1}{T}\sum_{t=1}^{T} CRPS(F_t(\hat{y}), y_t) \tag{5-4}$$

In addition, reliability and spread are assessed for probabilistic flow forecasts. Being consistent with Thiboult et al. (2016), the reliability diagram is used to depict reliability. Reliability diagram is a graph of the observed frequency of an event plotted against the forecast probability of an event (Hartmann et al., 2016). In theory, a perfect forecast system will result in forecasts with a probability of X% being consistent with observations X% of the time. Hence when plotting a reliability diagram, comparisons are made against the diagonal. A curve above the diagonal line denotes an over-dispersion, an under-dispersion is in the opposite case. Specifically, the reliability of the 95% flow prediction intervals will be shown in this chapter to give modelers a hydrologically meaningful understanding of the reliability.

Spread is an indicator that should be considered along with reliability because a perfectly reliable forecast at the cost of excessively high dispersion is not desired. The spread equals to the square root of the average ensemble variance over the evaluation period (Fortin et al., 2014).

$$spread = \sqrt{\frac{1}{T}\sum_{t=1}^{T} Var(\hat{y}_t)} \tag{5-5}$$

$$Var(\hat{y}_t) = \frac{1}{N-1}\sum_{n=1}^{N}\left(\hat{y}_{t,n} - y_t\right)^2 \tag{5-6}$$

where N is the forecast ensemble size ($n = 1, \dots, N$). $\hat{y}_{t,n}$ is the n^{th} forecasted flow at time t. The range of spread is non-negative with the optimum of zero. Units of spread are the units of the flows.

5.3. Results and Discussion

The comparison results are presented in two sections. Section 5.3.1 compares the multi-model and single-model forecasts for all the flows. Section 5.3.2 compares the multi-model and single-model forecasts only for the high flows that are equal to or greater than the 90th percentile of the flow records.

5.3.1. Flow Forecast Comparison

Evaluation Metric Results

Figure 5-2 shows the RMSE, MCRPS, reliability, and spread evaluation results of the multi-model and three single-model forecasting systems of 20 catchments for the 1st, 3rd, 6th and 9th lead days flow forecasts. The RMSE and MCRPS are used to demonstrate the deterministic and overall probabilistic forecast performances, respectively. The reliability is for the 95% flow forecast intervals. The spread measures the flow ensemble width. For brevity, the 2nd, 4th, 5th, 7th, and 8th lead day results are not shown. Table 5-5 lists the average absolute differences of KGE, MCRSP, reliability and spread between each of the three single-model forecasting systems and the multi-model forecasting system over 20 catchments on each lead day. The absolute metric difference for each catchment is calculated as the single-model system metric result minus the multi-model system metric result. The average absolute metric difference is the mean of the absolute metric differences of 20 catchments.

In both Figure 5-2 and Table 5-5, the comparison between the single-model (Calib & EnKF) system and the multi-model system shows that, by incorporating both the parameter and measured climate data uncertainty and using the more realistic climate ensemble based EnKF, the robust single model produces improved or at least no worse deterministic and overall probabilistic forecasts than the multiple models for all lead times. The forecast reliability of the robust single-model system is worse than that of the multi-model system due to the limited forecast width. However, for most catchments, the prediction reliability of the robust single-model is practically good. The mean reliability of 20 catchments is in the range of 0.68 and 0.73 for all horizons.

Another interesting finding is that, in Figure 5-2, as lead time advances, the prediction spread increases in all the three single-model systems, whereas the prediction spread decreases in the multi-model system. For example, for the 1st catchment, from the first lead day to the ninth lead day, the robust single-model forecast spread rises from 0.41 mm/d to 0.61 mm/d, while the multi-model forecast spread drops from 1.01 mm/d to 0.78 mm/d.

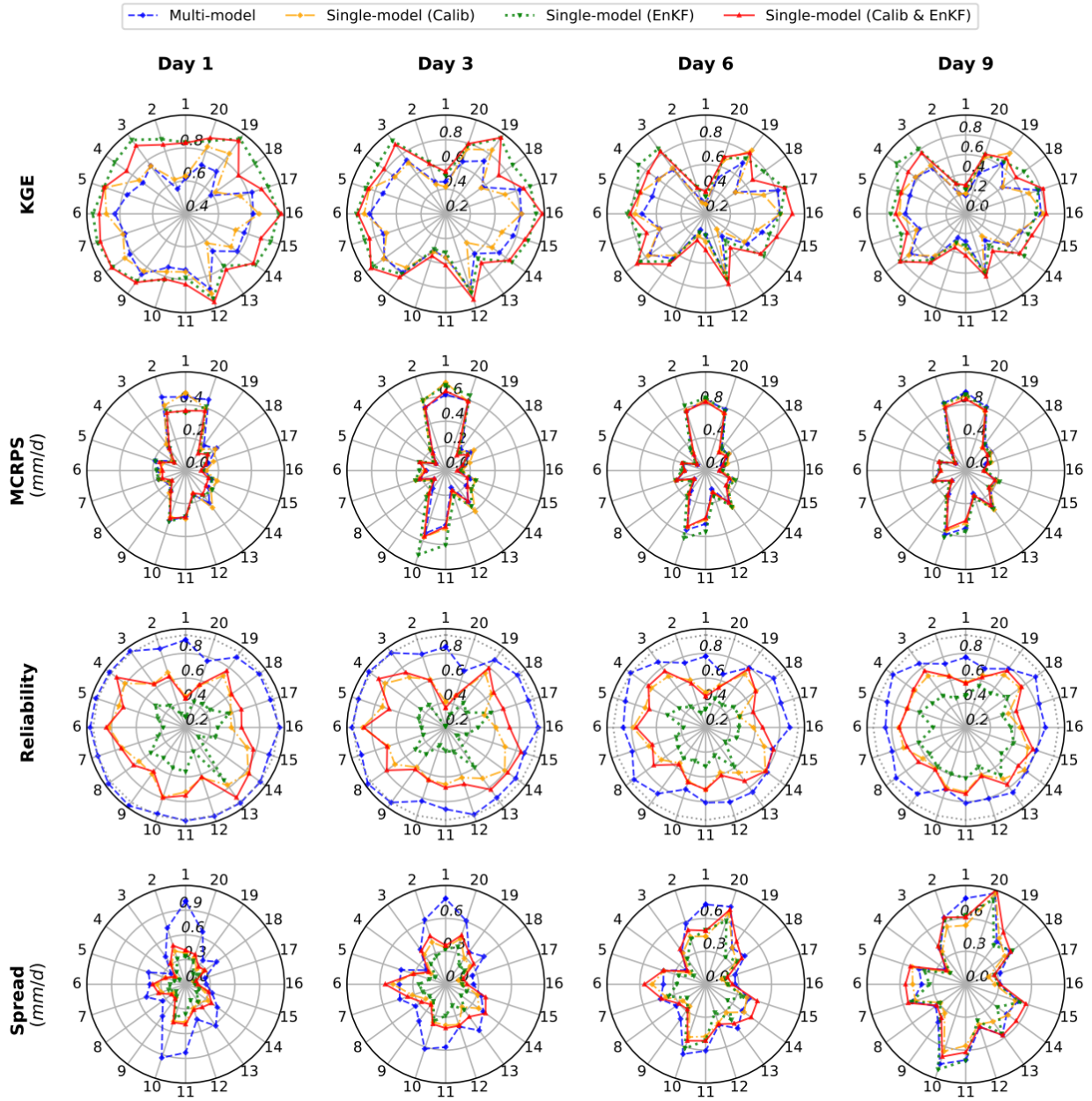

Figure 5-2. KGE, MCRPS, reliability and spread evaluation results of the multi-model and three single-model forecasting systems of 20 catchments for the 1st, 3rd, 6th and 9th lead day flow forecasts. See Table 5-4 for the descriptions of the multi-model and single-model forecasting systems. Each catchment is identified by the label on the outer edge of the circle and the catchment metric result is represented by the value on the corresponding spoke. The dotted line of the reliability panel represents the value of 0.95.

Table 5-5. Average absolute differences of KGE, MCRPS, reliability and spread between the single-model forecasting systems and the multi-model forecasting system over 20 catchments. The absolute metric difference for each catchment is calculated as the single-model system metric result minus the multi-model system metric result. Each column represents a single-model forecasting system of Table 5-4.

Lead day	Average absolute KGE difference			Average absolute MCRPS difference (mm/d)		
	Calib	EnKF	Calib & EnKF	Calib	EnKF	Calib & EnKF
1	0.02	0.16	0.14	0.00	-0.02	-0.03
2	0.00	0.13	0.12	0.03	0.02	-0.01
3	-0.01	0.11	0.11	0.03	0.05	0.00
4	0.00	0.11	0.11	0.03	0.05	0.00
5	0.00	0.10	0.11	0.02	0.05	-0.01
6	0.01	0.10	0.11	0.01	0.04	-0.01
7	0.02	0.10	0.11	0.00	0.03	-0.02
8	0.02	0.10	0.11	0.00	0.02	-0.02
9	0.02	0.11	0.11	-0.01	0.01	-0.03
Lead day	**Average absolute reliability difference**			**Average absolute spread difference (mm/d)**		
	Calib	**EnKF**	**Calib & EnKF**	**Calib**	**EnKF**	**Calib & EnKF**
1	-0.25	-0.48	-0.22	-0.23	-0.30	-0.19
2	-0.24	-0.49	-0.20	-0.18	-0.25	-0.14
3	-0.23	-0.49	-0.19	-0.14	-0.22	-0.09
4	-0.21	-0.46	-0.18	-0.11	-0.18	-0.06
5	-0.19	-0.42	-0.16	-0.09	-0.14	-0.04
6	-0.17	-0.37	-0.15	-0.07	-0.11	-0.02
7	-0.16	-0.33	-0.14	-0.06	-0.08	-0.01
8	-0.15	-0.30	-0.13	-0.06	-0.05	0.00
9	-0.14	-0.26	-0.13	-0.05	-0.03	0.01

Relative Contributions of the Two Methods

The relative contributions of the two methods (the parameter and climate data uncertainty based calibration and the N15 based EnKF) to the overall accuracy and reliability improvements of the single-model (Calib & EnKF) system are assessed based on Figure 5-2 and Table 5-5. First, the parameter and climate data uncertainty based calibration is the main contributor to the prediction spread increase of the single-model (Calib & EnKF) system. In Figure 5-2, the single-model (Calib) system generates higher spread than the single-model (EnKF) system for short term forecasts (i.e., 1-3 lead days), and the single-model (Calib) system spread is closer to the single-model (Calib & EnKF) system spread than the single-model (Calib) system spread in most catchments. Second, the more realistic climate ensemble based data assimilation is the main contributor of the accuracy improvement of the single-model (Calib & EnKF) system. In Figure 5-2, the single-model (EnKF) system produces more accurate deterministic forecasts than the single-model (Calib) system. In sum, considering parameter and climate data uncertainty is an effective

way of increasing prediction width and using the more realistic climate ensemble in data assimilation is beneficial to improve prediction accuracy.

With the joint contribution of the two methods, the forecast accuracy and spread of the single model forecast system increase. As the result, the reliability of the single-model (Calib & EnKF) system becomes practically good (as mentioned above, the mean reliability of 20 catchments is in the range of 0.68 and 0.73 for all horizons). The long-term forecast results particularly reveal that increasing forecast width alone does not necessarily improve forecast reliability. The improvement of forecast reliability also requires the improvement of forecast accuracy. In Figure 5-2, on the 6^{th} and 9^{th} lead days, although the single-model (Calib) spread is similar with the single-model (EnKF) spread, the reliability of the former is higher than that of the latter. An important reason for the reliability difference is that the accuracy of the single-model (Calib) system is worse than that of the single-model (EnKF) system.

The evaluation metric differences of Table 5-5 demonstrate that the overall performance change of the single-model (Calib & EnKF) system relative to the multi-model system is not a simple addition of the individual changes of the single-model (Calib) and single-model (EnKF) systems. This is because compensation effects occur between the two methods in forecasting process. Specifically, for KGE and MCRPS, the combination of the two methods always outperforms any of the two methods used alone (except the 1^{st} and 2^{nd} lead days for KGE). This is consistent with the findings of several studies showing that the combinational consideration of different uncertainty sources of ensemble flow forecast generates the best forecasts than the consideration of any single uncertainty sources (Thiboult et al., 2016; Velázquez et al., 2011).

Reliability Diagram

To get a comprehensive understanding of the forecast reliability, reliability diagram is assessed. Figure 5-3 displays the reliability diagrams of the multi-model and three single-model forecasting systems for 20 catchments of the 1^{st}, 3^{rd}, 6^{th} and 9^{th} lead days. The multi-model system generates over-dispersion flow ensembles on the 1^{st} lead day, almost perfect reliability on the 3^{rd} lead day, and under-dispersion flow ensembles for the longer horizons. In contrast, the three single-model systems are all under-dispersion for all percentiles of interest and all horizons.

Note that the results in Figure 5-2 and Figure 5-3 would not significantly change if more than 20 behavioral parameter sets are used for each catchment (randomly selected from the behavioral parameter ensemble).

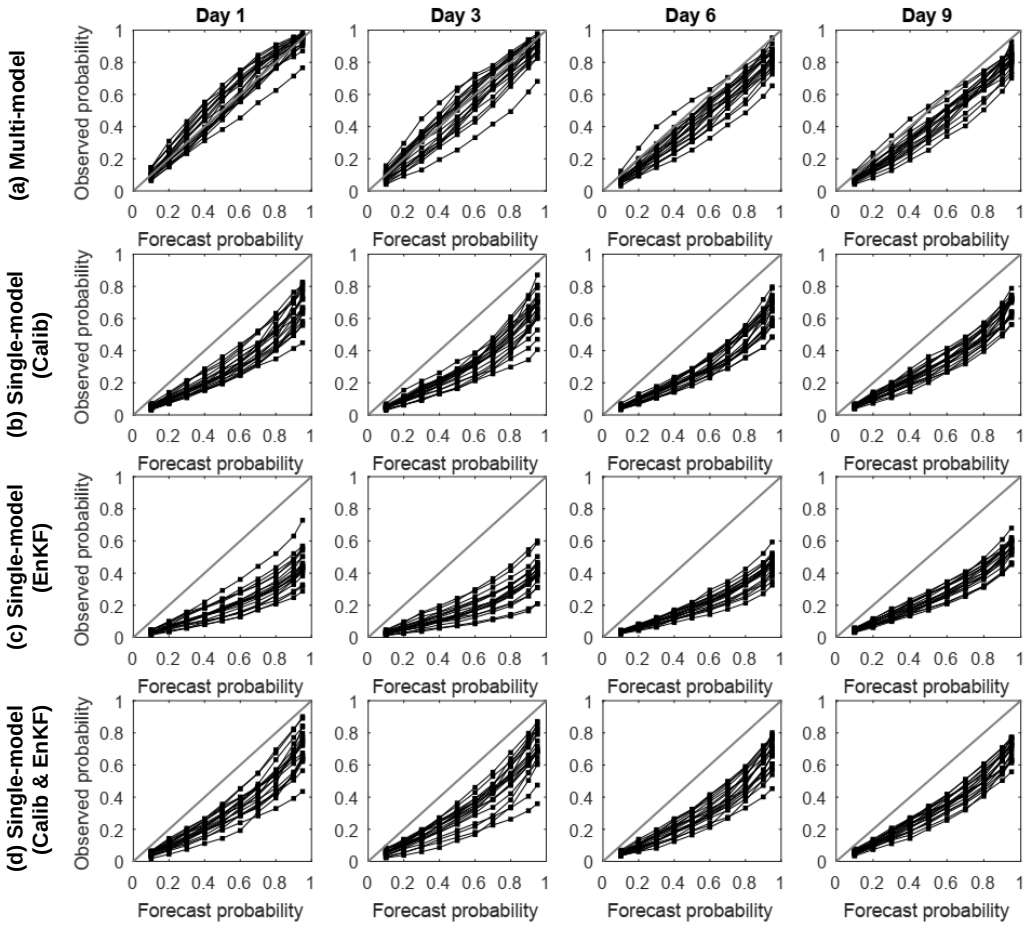

Figure 5-3. Reliability diagrams of the multi-model and three single-model forecasting systems for all 20 catchments and for the 1st, 3rd, 6th, and 9th lead days flow forecasts. See Table 5-4 for the descriptions of the multi-model and single-model forecasting systems. Each curve of a reliability diagram refers to a catchment. The diagonal line represents the perfectly reliable forecast.

5.3.2. High Flow Forecast Comparison

Considering that the main purpose of ensemble flow forecasting is often for flood forecasting, the single-model and multi-model flow forecast comparison is repeated for high flow only. High flow here refers to the flow that is equal to or greater than the 90th percentile of the flow records.

Evaluation Metric Results

Figure 5-4 shows the RMSE, MCRPS, reliability, and spread evaluation results for the high flow forecasts of the multi-model and three single-model systems of 20 catchments for the 1st, 3rd, 6th and 9th lead days. Table 5-6 summarizes the high flow average absolute differences of KGE, MCRSP, reliability and spread between each of the three single-model forecasting systems and the multi-model forecasting system over 20 catchments. Same as in Section 5.3.1, the absolute metric difference for each catchment is

calculated as the single-model system metric result minus the multi-model system metric result. The average absolute metric difference is the mean of the absolute metric differences of 20 catchments.

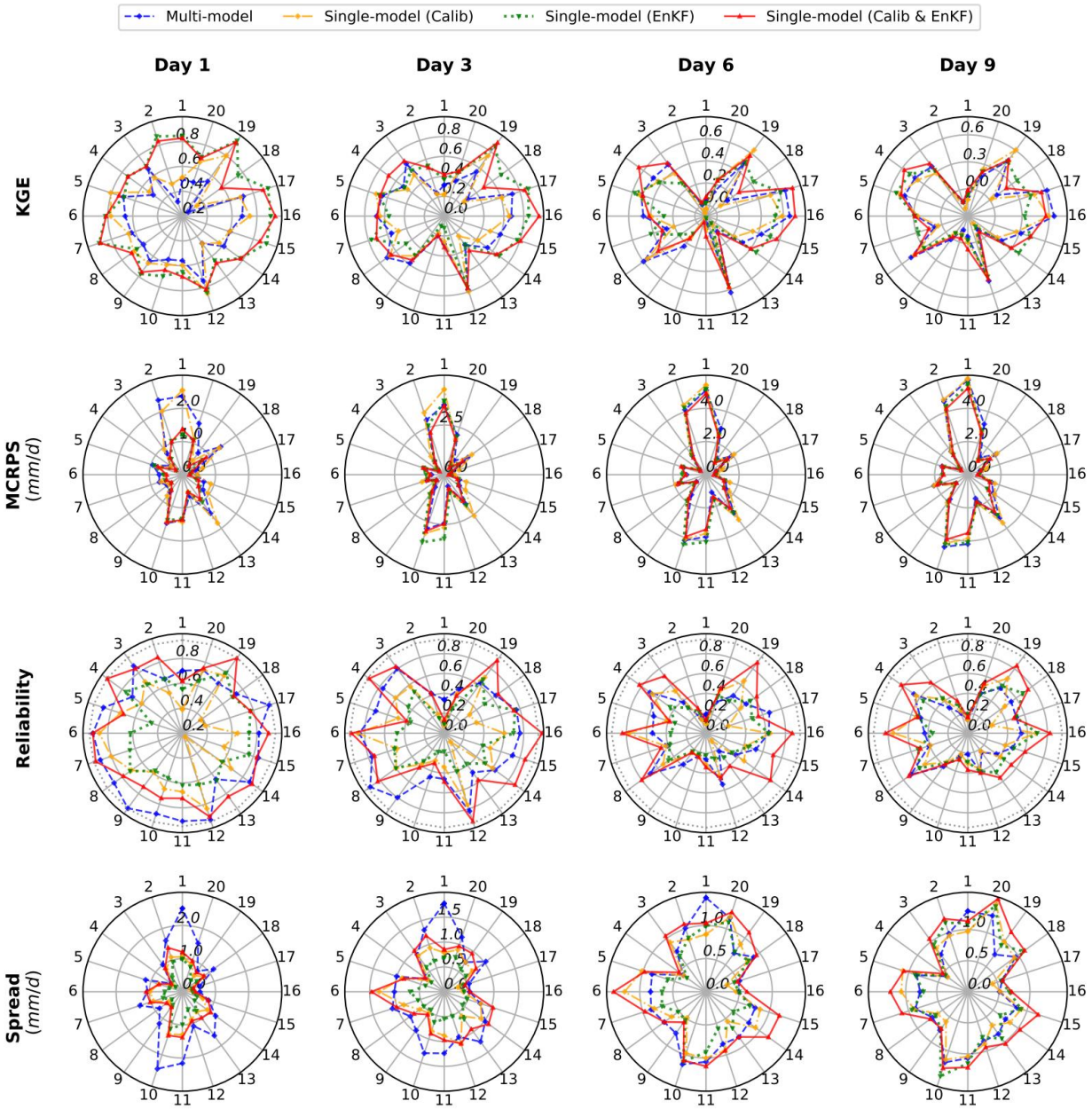

Figure 5-4. KGE, MCRPS, reliability and spread evaluation results of the multi-model and three single-model forecasting systems of 20 catchments for the 1st, 3rd, 6th and 9th lead day high flow forecasts. See Table 5-4 for the descriptions of the multi-model and single-model forecasting systems. Each catchment is identified by the label on the outer edge of the circle and the catchment metric result is represented by the value on the corresponding spoke. The dotted line of the reliability panel represents the value of 0.95.

Table 5-6. High flow average absolute differences of KGE, MCRPS, reliability and spread between the single-model forecasting systems and the multi-model forecasting system over 20 catchments. The absolute metric difference for each catchment is calculated as the single-model system metric result minus the multi-model system metric result. Each column represents a single-model forecasting system of Table 5-4.

Lead day	Average absolute KGE difference			Average absolute MCRPS difference (mm/d)		
	Calib	**EnKF**	**Calib & EnKF**	**Calib**	**EnKF**	**Calib & EnKF**
1	0.06	0.22	0.20	-0.02	-0.32	-0.32
2	0.00	0.14	0.16	0.17	-0.08	-0.18
3	-0.04	0.06	0.11	0.24	0.06	-0.10
4	-0.04	0.02	0.08	0.22	0.10	-0.08
5	-0.04	-0.01	0.05	0.18	0.10	-0.09
6	-0.04	-0.03	0.04	0.13	0.05	-0.12
7	-0.04	-0.03	0.03	0.09	-0.03	-0.17
8	-0.03	-0.01	0.03	0.06	-0.09	-0.20
9	-0.03	0.01	0.04	0.03	-0.16	-0.24
Lead day	**Average absolute reliability difference**			**Average absolute spread difference (mm/d)**		
	Calib	**EnKF**	**Calib & EnKF**	**Calib**	**EnKF**	**Calib & EnKF**
1	-0.20	-0.16	-0.01	-0.47	-0.66	-0.37
2	-0.22	-0.23	0.00	-0.34	-0.55	-0.21
3	-0.20	-0.24	0.01	-0.23	-0.45	-0.10
4	-0.14	-0.20	0.04	-0.14	-0.34	-0.01
5	-0.08	-0.15	0.08	-0.08	-0.24	0.06
6	-0.03	-0.08	0.12	-0.04	-0.15	0.10
7	-0.02	-0.03	0.12	-0.02	-0.07	0.14
8	0.00	0.00	0.14	0.00	0.00	0.17
9	0.01	0.04	0.14	0.01	0.05	0.19

Similar with the comparison results of all flows in Section 5.3.1, the single-model (Calib & EnKF) system yields improved deterministic and overall probabilistic flow forecasts than the multi-model system. However, a big difference from Section 5.3.1 is that the single-model (Calib & EnKF) system generates more reliable high flow forecasts than the multi-model system as the lead time advances (Figure 5-4). On the 1st lead day, there are 8 of 20 catchments where the single-model (Calib & EnKF) system achieves more reliable high flow forecasts than the multi-model system, and the mean absolute reliability improvement of 20 catchments is -0.01 (Table 5-6). The reliability improvement becomes more obvious for the forecast after the 3rd lead day. On the 9th lead day, there are 18 of 20 catchments where the single-model (Calib & EnKF) system achieves more reliable high flow forecasts than the multi-model system, and the mean absolute reliability improvement of 20 catchments is 0.14 (Table 5-6).

Reliability Diagram

Figure 5-5 gives a more comprehensive exam of the reliability by illustrating the reliability diagrams of the four forecasting systems for high flow forecasts of 20 catchments on the 1st, 3rd, 6th and 9th lead days. In Figure 5-5d, on the 1st lead day, the reliability of the single-model (Calib & EnKF) system is near perfect. When the forecast horizon is more than 3 days, the single-model (Calib & EnKF) system produces better reliability than the multi-model system over all percentiles of interest. At the same time, considering the prediction accuracy enhancement in Figure 5-4 and Table 5-6, it is concluded that the robust single hydrologic model is definitely a better choice than the multiple hydrologic models in operational high flow forecasting (i.e., flood forecasting).

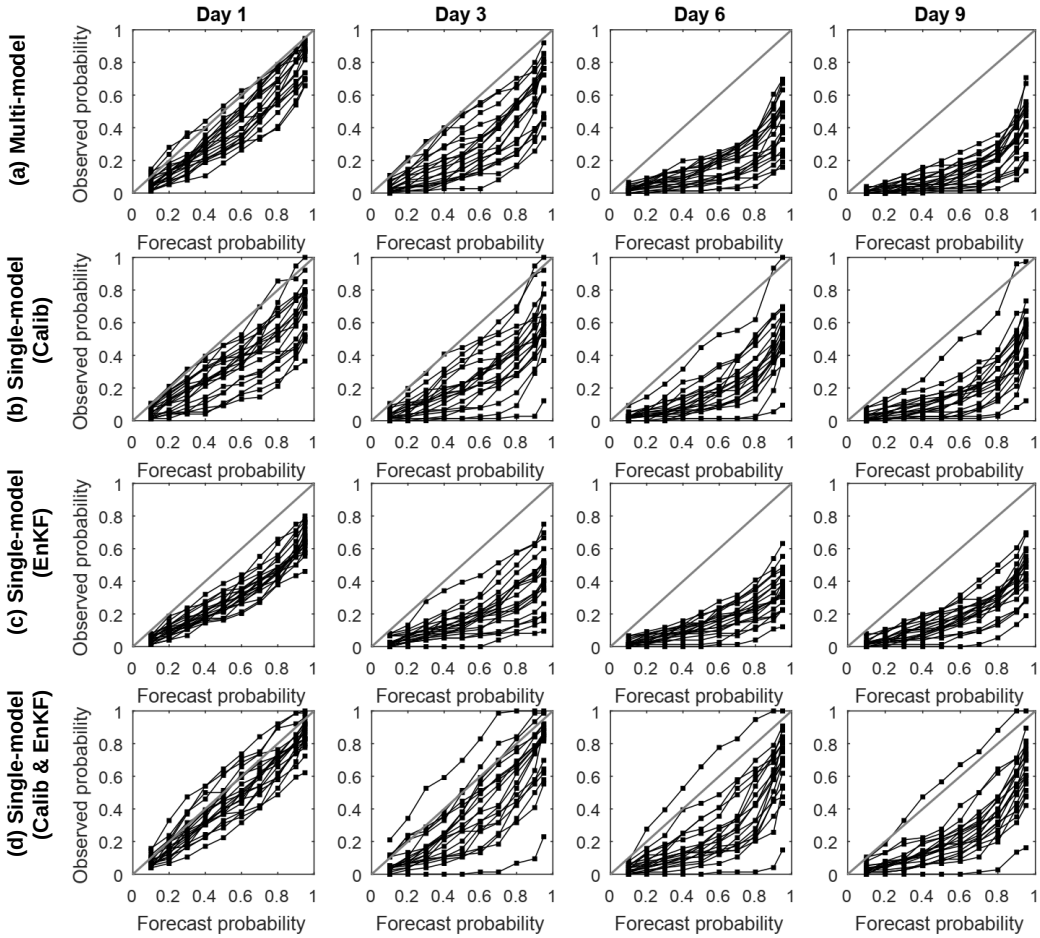

Figure 5-5. Reliability diagrams of the multi-model and three single-model forecasting systems for all 20 catchments and the 1st, 3rd, 6th, and 9th lead days high flow forecasts. See Table 5-4 for the descriptions of the multi-model and single-model forecasting systems. Each curve of a reliability diagram refers to a catchment. The diagonal line represents the perfectly reliable forecast.

Hydrograph

To visualize the different characteristics of the multi-model and three single-model flow forecasting systems, Figure 5-6 displays the first lead day forecast hydrographs of the four systems for the Peribonka catchment (the 15th catchment of Table 3-1) and the entire forecasting period. The ensemble means and the 95% prediction intervals of forecasted flows are plotted. For the comparison hydrographs of all other 19 catchments, please refer to Appendix 5-1.

Figure 5-6 confirms the forecast performance upgrade by using the robust single hydrologic model that considers parameter and climate data uncertainty and uses more realistic climate ensemble in the EnKF. The deterministic flow forecasts (ensemble means) are more consistent with the observations in the robust single-model system (Figure 5-6d) than in the multi-model system (Figure 5-6a). Also, for the high flow forecasts over the 90th percentile of the flow observation, more observations fall into the 95% prediction envelope of the robust single-model system than that of the multi-model system. Specifically, looking at the peak flow forecast in May 2009, the ensemble means of the multi-model system always underestimate the peak flows, and the 95% prediction intervals of the multi-model system cover only a part of the observations. In comparison, the robust single-model system of Figure 5-6d improves the consistency between the ensemble means and the peak flows, and its 95% forecast intervals contain almost all the peak flows.

Moreover, Figure 5-6 demonstrates the relative contributions of the two methods (the parameter and climate data uncertainty based calibration and the N15 based EnKF) to the overall accuracy and reliability improvements of the single-model (Calib & EnKF) system. The parameter and climate data uncertainty based calibration is the main contributor to the flow spread increase, because the flow width of the single-model (Calib) system is much higher than that of the single-model (EnKF) system and is similar with the width of the single-model (Calib & EnKF) system. Moreover, the more realistic climate ensemble (N15) based data assimilation produces more accurate ensemble mean flows than the single-model (Calib) system and is the main contributor of the overall single-model accuracy improvement. With the joint contributions of the two methods, the reliability of the single-model (Calib & EnKF) system reaches a satisfactory level in Figure 5-6d.

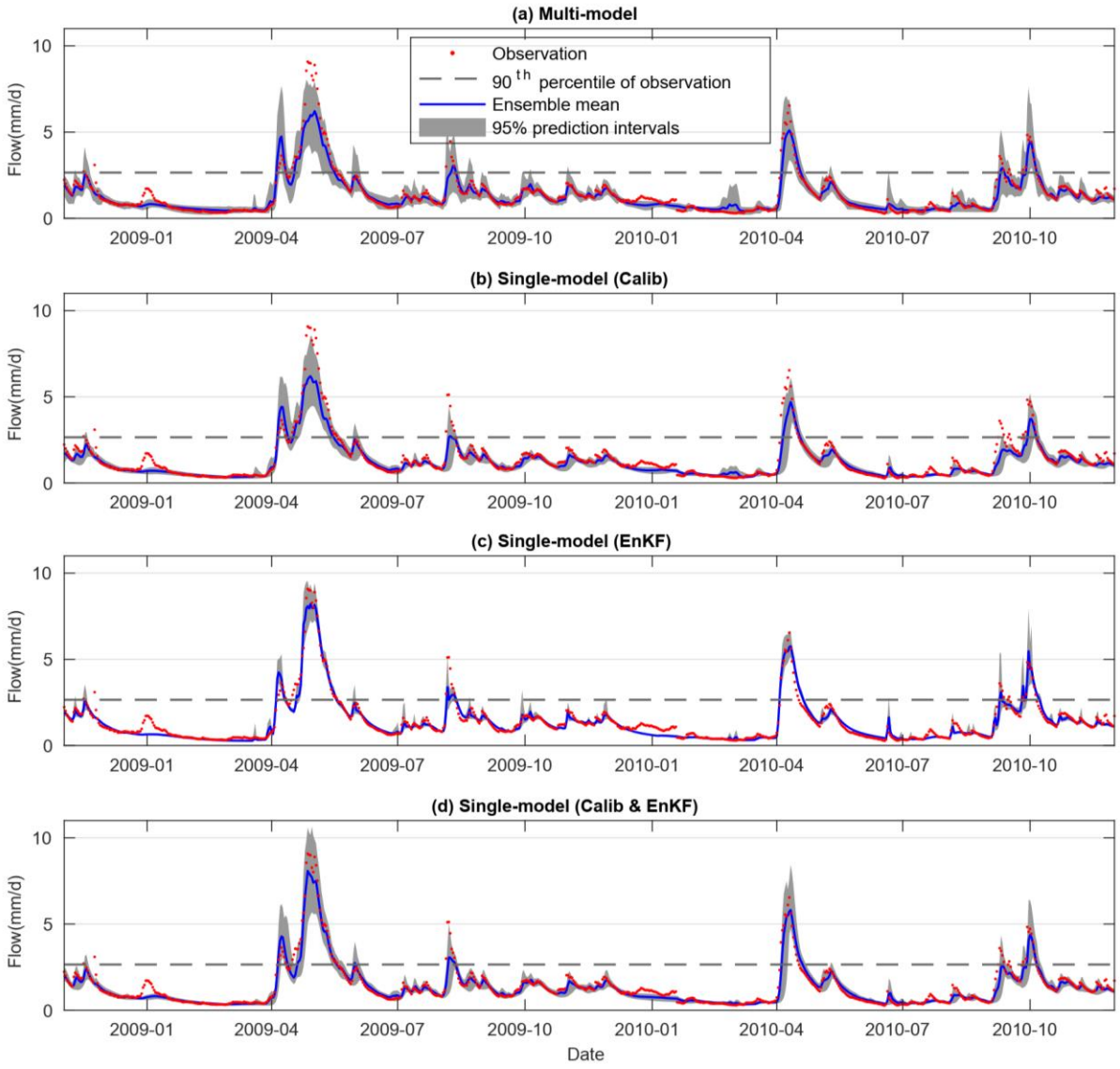

Figure 5-6. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Peribonka catchment (the 15th catchment of Table 3-1) of the entire forecasting period. See Table 5-4 for the descriptions of the multi-model and single-model forecasting systems.

5.4. Conclusions and Future Work

This chapter built a robust single hydrologic model for ensemble streamflow forecasting. The robust single hydrologic model was calibrated with an explicit consideration of parameter and climate data uncertainty and used the Newman et al. (2015) dataset (referred to as N15) to represent measured precipitation and temperature data uncertainty in the EnKF. In order to investigate the separate and joint influences of the two methods on ensemble flow forecast, three different single-model forecasting systems were constructed. The first single-model system incorporates parameter and measured climate data uncertainty in forecasting by running the parameter and climate data uncertainty based model calibration and getting multiple parameter solutions. The second single-model system uses the N15 based EnKF data assimilation. The third

single-model system applies the two methods together (the parameter and climate data uncertainty based calibration and the N15 based EnKF). In this chapter, the robust single-model system specifically refers to the third single-model system. Each single-model system was compared with a 20 hydrologic models based forecasting system (multi-model system) over 20 catchments. The multi-model system considers model structure errors by using 20 different hydrologic models and uses assumed statistical error models to perturb measured climate data in the EnKF. The multi-model forecast results are taken from Thiboult et al. (2016).

Our comparison study showed that the robust single hydrologic model that applies the two methods together generates improved or similar deterministic and overall probabilistic flow forecasts than the multiple hydrologic models. Moreover, the robust single model generates practically reliable all flow forecasts and more reliable high flow forecasts than the multiple models. High flow here refers to the flow no less than the 90^{th} percentile of the flow records. The reliability improvements for high flow forecasts are most evident for lead times after 3 days (Figure 5-4 and Table 5-6). All these results suggest that it is an advantage to use a robust single hydrologic model for deterministic flow forecasts and probabilistic high flow forecasts. Moreover, using a hydrologic single model relieves modelers from calibrating multiple hydrologic models for each watershed in operational flow prediction. The direct use of the N15 derived climate ensemble frees modelers from tuning hyper-parameters or checking the literature to set suitable precipitation and climate perturbations for the EnKF if utilized in data assimilation.

The forecast performance comparison among the three single-model systems showed that the combination of the two methods outperforms any of the two methods used alone. Relatively speaking, the parameter and measured climate data uncertainty is the main contributor to broaden flow spread, and the more realistic climate ensemble based data assimilation is the main contributor to ensure forecast accuracy. The joint contribution of the two methods brings in practically good reliability of the robust single-model forecasting system.

The current single-model and multi-model comparison research is only built on lumped hydrologic models. Future work can use distributed hydrologic model in the single-model forecasting system and compare the robust distributed hydrologic model with multiple lumped hydrologic models in ensemble flow forecasting.

Chapter 6.
Hydrologic Model Calibration under Uncertainty using Flow Ensembles

Summary

After investigating the treatment of climate data uncertainty in model calibration and the ensemble Kalman filter (EnKF), this thesis intended to explore flow data uncertainty in the same order of our previous studies on climate data uncertainty (i.e., calibration, data assimilation, and flow forecasting). In the literature, flow data uncertainty is rarely explicitly considered in hydrologic model calibration either because its effect is considered negligible or because it is difficult to disaggregate flow data errors from other errors. This chapter addresses the explicit consideration of flow data uncertainty by using flow ensemble and forms the climate-flow ensemble based hydrologic model calibration framework. The framework is built on Chapter 3 proposed climate ensemble based model calibration framework. This chapter demonstrates the method of using the hydraulics-based Bayesian rating curve uncertainty estimation method (BaRatin) (Le Coz et al., 2014) to generate flow ensemble. The continuous ranked probability score (CRPS) is taken as an objective function of the framework to compare the scalar model prediction with the measured flow ensemble. The framework performance is compared with two observed flow based calibrations, one considers parameter uncertainty, another considers parameter and climate data uncertainty. The assessments over 10 catchments show that the framework maintains the flow prediction accuracy and slightly improves the prediction reliability compared with the two observed flow based calibrations. However, results show that flow data uncertainty plays a less important role than climate data uncertainty in flow prediction. This work also demonstrates the dependency of parameter uncertainty on forcing and response data.

Section 6.1 reviews the flow data uncertainty estimation methods and the model calibration methods that explicitly consider flow data uncertainty. Section 6.2 describes the proposed climate-flow ensemble based calibration framework. Section 6.3 reviews the BaRatin framework and explains the comparative calibration setups of the case study. Section 6.4 shows the case study results. Section 6.5 provides conclusions about the proposed calibration framework.

6.1. Introduction

Continuous streamflow is not measured directly by any instrument but derived from water levels based on the level-discharge relationship, called a rating curve. To get the rating curve, the continuous records of stage (i.e., the height of water level) and discharges need collected at a location along a stream. Discharge is calculated by multiplying the area of water in a channel cross section by the average velocity of the water in the cross section in an aggregation way, so the height and width of the subsection and the velocity of the streamflow at each subsection also need to be measured. Based on the instantaneous stage and discharge

measurements (also called gauging data), the rating curve can be defined and used to convert the measured stage to the estimate of streamflow in operation (Olson and Norris, 2007).

Several equations have been established to express the rating curve, amongst them the most commonly used is the power equation (ISO 1100, 2010):

$$Q = a(h - b)^c \tag{6-1}$$

where Q is discharge, h is stage, a is a scaling coefficient, b is the reference level, and c is an exponent and can be related to the type of the hydraulic control. Developing an accurate rating curve relation requires numerous instantaneous measurements at all ranges of stage and discharge. The rating curve differs from site to site depending on the site hydraulic condition such as the shape, size, slope, and roughness of the channel and flow stability. Moreover, for each site, the rating curve needs continuously checked and modified due to channel geometry changes from erosion, deposition, flood, seasonal vegetation growth, debris, or ice (Olson and Norris, 2007).

According to Coxon et al. (2015), measured flow data errors mainly arise from four sources: (1) gauging data, (2) natural conditions, (3) rating-curve approximation, and (4) human alterations. Gauging data errors can be caused from the point stage, cross section height, weight, and velocity measurements, and insufficient sampling. Natural conditions can lead to the channel geometry and cross section changes through erosion, sedimentation, hysteresis, and weed growth. The fit of the rating curve is also a significant source of uncertainty depending on the prior knowledge of the rating curve relationship, the number and coverage of points over the flow range, in particular at high flow, and the multi-section breaks due to different channel or hydraulic controls. Finally, human regulation and intervention affect the discharge uncertainty through either changing the gauging station, or the use of multiple weirs at a station where the rating curve changes depending on which weir is in operation.

Rating curve uncertainty estimation is the main access to getting probabilistic discharge estimates in the literature. Traditional statistical approaches exploit the residual variance of the rating curve regression to determine the discharge uncertainty bounds but fail to explicitly incorporate the measurement uncertainty in gauging data (Petersen-Øverleir, 2006; Venetis, 1970). There are also approaches that rely on the stage-discharge measurement uncertainty to estimate rating curve uncertainty. For example, the fuzzy regression method fits multiple rating curves to subsets of stage-discharge measurements or a large number of stage-discharge samples that are randomly selected from the assumed measurement errors (McMillan et al., 2010; Pappenberger et al., 2006). To explicitly handle the aleatory and epistemic errors of the gauging data, the likelihood function based rating curve uncertainty estimation is developed to match the interpretation of the range of gauging data (Mcmillan and Westerberg, 2015). To assess the time variability of the gauging data, the stage-discharge measurements are grouped based on the valid date range of rating curve and are fitted separately using the nonparametric regression (Coxon et al., 2015). In addition to the gauging data errors,

Bayesian methods are advantageous in using hydraulic knowledge to set the prior distributions of rating curve parameters and dealing with the biased uncertainty estimation from non-stationary error assumptions of other methods (Le Coz et al., 2014; Moyeed and Clarke, 2005). Therefore, Bayesian methods have become the preferred choice to estimate flow data uncertainty (Coxon et al., 2015; Mcmillan and Westerberg, 2015; Petersen-Øverleir, 2006).

Although a variety of rating curve uncertainty estimation methods are available to determine flow data uncertainty, these methods have not been widely applied to hydrologic model calibration to account for flow data uncertainty. This is because the effect of flow data uncertainty is usually considered negligible or because it is difficult to disaggregate flow data errors from other errors. Among the existing uncertainty based calibration frameworks in the literature, the Bayesian total error analysis methodology (BATEA) is a good framework that supports the explicit consideration of flow data uncertainty (Kavetski et al., 2002). BATEA uses a probabilistic error model to explicitly describe flow data uncertainty and incorporates the assumed error model into the likelihood function by integration. Initially, Kavetski et al. (2006a, 2002) used an additive Gaussian error to describe response and model structure errors together. Kuczera et al. (2006) made a progress of distinguishing structure errors from response errors by assigning them separate error terms, and the flow error is assumed additive and Gaussian. To better represent flow data uncertainty, Thyer et al., (2009) used rating curve to estimate flow measurement errors before Bayesian inference. The flow measurement error (ε) was calculated as the difference between the given rating curve estimated flow ($\hat{Q}$) and the measured flow ($\tilde{Q}$) and was fitted to an assumed normal distribution. The fitted measurement error (ε) was then added to the hydrologic model simulated flow to build the probabilistic true flow for Bayesian inference. In sum, the focus of the formal likelihood based calibration studies that explicitly consider flow data uncertainty is developing an accurate and precise probabilistic error model to represent flow measurement errors, though only the addictive Gaussian error has been tried so far. The assumed or fitted statistical error models are easy to use to construct the formal likelihood, but they do not necessarily have the inherent probabilistic properties of flow measurement errors (Vrugt and Sadegh, 2013).

When the explicit errors are difficult or impossible to quantify, the informal likelihood is more robust than the formal likelihood in parameter estimation (Mcmillan and Westerberg, 2015). In the likelihood-free calibration studies, two main strategies have been explored to estimate and account for flow data uncertainty in model calibration. The first strategy still resorts to assumed statistical error models but utilizes flow ensemble to represent flow data errors in the likelihood function. For example, in environmental model calibrtion, McIntyre et al. (2002) assumed the organic carbon and dissolved oxygen data follow normal distributions and sampled multiple realizations from the assumed distributions. Each response realization was taken as the target response to generate the maximum likelihood parameters. All the maximum likelihood parameters conditioned on all the response realizations were combined to get the converged

posterior parameter distribution. The second strategy uses multiple fitted rating curves and compares model output with the uncertainty boundary or empirical distribution of flow (Krueger et al., 2009; McMillan et al., 2010; Pappenberger et al., 2006). For example, McMillan et al. (2010) fitted multiple possible rating curves using a number of stage-discharge samples from the assumed data error distributions. Based on the multiple fitted rating curves, they produced an empirical probability density function (PDF) for observed flow at each required stage and calculated the conditional probability of the modelled flow at each time step. All the evaluation period conditional probabilities were integrated to get the likelihood in the maximum likelihood estimation.

Enlighted by the above two methods, this chapter focuses on the likelihood-free calibration and proposes a flow ensemble based calibration framework to explicitly consider flow data uncertainty in parameter estimation. This chapter demonstrated using the Bayesian rating curve framework (BaRatin) (Le Coz et al., 2014) to generate flow ensemble. The BaRatin framework combines hydraulics knowledge and stage-discharge measurement errors in rating curve uncertainty estimation. Our purpose is to avoid the flow error models that are assumed for mathematical convenience. Instead, we use more physically-based flow uncertainty estimates to represent low data errors in hydrologic model calibration. The BaRatin approach has been shown providing reliable and physically sound flow uncertainty estimates for various flow conditions (Le Coz et al., 2014). An overview of the BaRatin framework is in Section 6.3.2 of this thesis.

Given the flow ensemble, the model evaluation metrics need to be selected to be able to assess the scalar flow simulation versus the measured flow ensemble at each time step. Model calibration studies have developed many evaluation metrics to assess model output against uncertain response. Among them, the most straightforward way is to compare model output with the upper-lower bounds of the response measurement. For instance, Pappenberger and Beven (2004) divided the flow uncertainty boundaries constrained hydrograph area into a number of boxes and calculated the performance within boxes according to the distances to the observation per time step. Krueger et al. (2009) modified goodness-of-fit indicators based on the uncertainty bounds. When model output is within the uncertainty bounds of the observation at a time, the deviation is defined as zero; otherwise, the deviation is defined as the ratio of the distance between the model output and the nearest uncertainty boundary to the width of the uncertainty boundary. In addition, the meteorology field has developed a number of evaluation metrics that compare probabilistic weather forecast with climate event or value, such as the Brier score (Brier, 1950), the continuous ranked probability score (CRPS) (Matheson and Winkler, 1976), the Kullback-Leibler divergence (Weijs et al., 2010) and others (Gneiting and Raftery, 2007). This research took advantage of these existing uncertainty based evaluation metrics and added the CRPS to the objective function of the proposed climate-flow ensemble based calibration framework.

The specific goals of this chapter are to: (1) form the climate-flow ensemble based calibration framework by adding flow ensemble to Chapter 3 proposed climate ensemble based calibration framework; (2) demonstrate how to use the BaRatin method to generate flow ensemble; (3) use a large number of case studies to test the additional value of the explicit consideration of flow data uncertainty in flow prediction and parameter estimation. The formed climate-flow ensemble based calibration framework is flexible enough to work with all likelihood-free calibration methods to explicitly consider climate and flow data uncertainty.

6.2. Methods

Assuming flow data is uncertain and can be represented by a flow ensemble, the flow ensemble is added to the climate ensemble based calibration framework that is proposed by Chapter 3 to explicitly account for flow data uncertainty. The completed framework is named the climate-flow ensemble based hydrologic model calibration framework. The only difference between the former framework (Chapter 3) and the current one is in the objective function construction. In the climate ensemble based calibration, the objective function compares the simulated flow with the observed flow using the absolute error based goodness-of-fit indicators (e.g., root mean square error). In the climate-flow ensemble based calibration, the probabilistic error based goodness-of-fit indicators are added to the objective function to compare the simulated flow with the measured flow ensemble. CRPS is the probabilistic error based goodness-of-fit indicator that is used in this chapter.

CRPS generalizes the traditional deviation between two deterministic data to the difference between two distributions at a time step. CRPS is reduced to the absolute error when both the predicted and observed flows are deterministic. CPRS has been widely used in whether and flow forecast studies (e.g., Abaza et al., 2017a; Hersbach and Hersbach, 2000; Leutbecher and Palmer, 2008; Zamo and Naveau, 2018). In model calibration, CPRS is defined as the quadratic discrepancy between the observed flow cumulative distribution function (CDF) and the empirical CDF of the scalar prediction (Gneiting and Raftery, 2007).

$$CRPS(F_t(y), \hat{y}_t) = \int_{-\infty}^{+\infty} (F_t(Y) - H\{y - \hat{y}_t\})^2 dY \quad (6\text{-}2)$$

where $F_t(y)$ is the measured flow CDF at time t and is approximated from the measured flow ensemble. $\hat{y}_t$ is the simulated flow at time t. $H\{y - \hat{y}_t\}$ is the Heaviside function and represents the predicted flow CDF at time t.

$$H\{y - \hat{y}_t\} = \begin{cases} 0, for\ y - \hat{y}_t < 0 \\ 1, for\ y - \hat{y}_t \geq 0 \end{cases} \quad (6\text{-}3)$$

The range of CRPS is non-negative with the best value of 0. Units of CRPS are the units of the flows.

The mean continuous ranked probability score (MCRPS) is calculated as the average CRPS over the entire evaluation period.

$$MCRPS = \frac{1}{T}\sum_{t=1}^{T} CRPS(F_t(y), \hat{y}_t) \quad (6\text{-}4)$$

MCRPS is also non-negative with the optimum of zero and is in the same unit as the flow.

By adding MCRPS to the objective function, the climate-flow ensemble based calibration becomes a multi-objective optimization problem. One type of objective function is based on the traditional absolute error to ensure the deterministic consistency with observation. Another type of objective function is based on the probabilistic error to ensure the similarity with observation or measurement ensemble. The different objectives can be treated equally or assigned different weights according to their importance in calibration.

6.3. Data and Experimental Design

The climate-flow ensemble based calibration framework is applied to 10 Québec catchments which are subset of the 20 catchments of Chapter 3. This section describes the research area, hydrologic model, flow ensemble generation, and the comparative calibration setups of the case study.

6.3.1. Research Area, Data and Hydrologic Model

Table 6-1 lists the 10 Québec catchments and their hydrometric stations at catchment outlets. Figure 6-1 displays the distribution of the 10 catchments and their hydrometric stations. This research uses the same meteorological and flow measurements, the Newman et al. (2015) dataset (N15) derived climate ensemble, and the hydrologic model as in Section 3.3 of Chapter 3. The same six parameters (four in GR4J and two in Cemaneige) are to be calibrated with their meanings and ranges shown in Table 3-2 of Chapter 3.

Table 6-1. Hydrometric station information of the 10 Québec catchments.

No.	River name	Station number*	Latitude	Longitude
1	Trois Pistoles	022301	48.09	-69.20
2	Sainte Anne	050408	46.85	-71.88
3	Bras du Nord	050409	46.97	-71.85
4	Du loup	052805	46.60	-73.19
5	Aux Ecorces	061020	48.38	-71.99
6	Metabetchouane	061502	48.69	-72.49
7	Ashuapmushuan	061901	49.28	-73.36
8	Ashuapmushuan	061905	48.68	-72.51
9	Metabetchouane	062102	48.89	-72.26
10	Valin	062701	48.49	-70.97

* Station number follows the numbering regulation of the Direction de l'Expertise Hydrique.

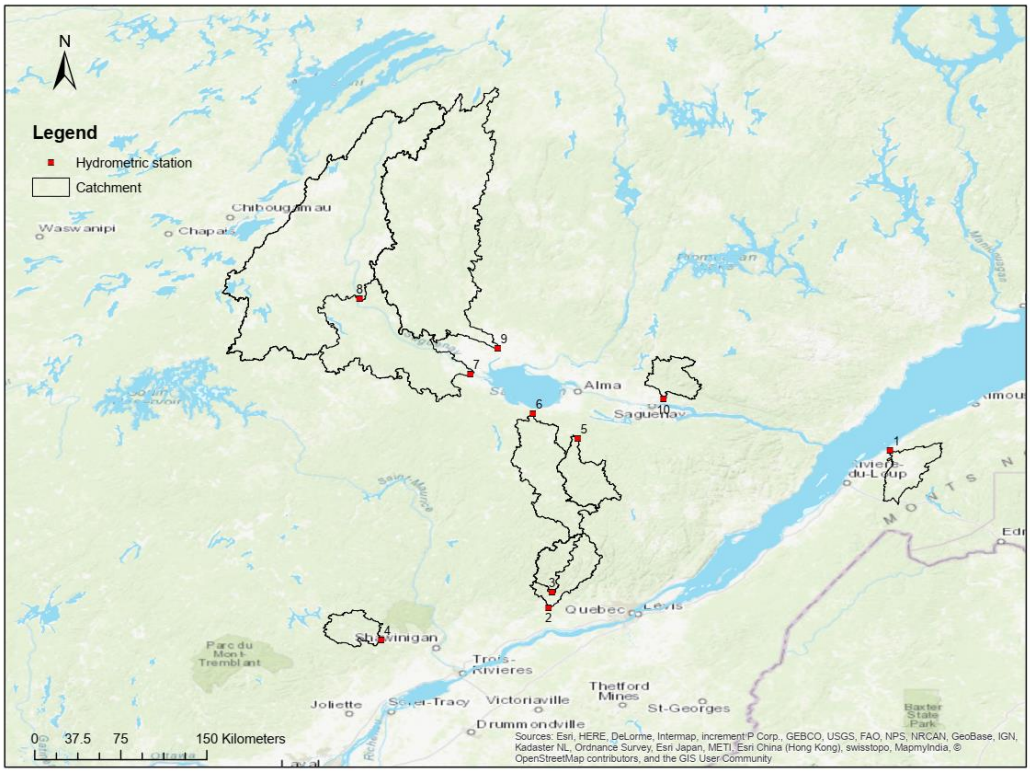

Figure 6-1. Distribution of the 10 Québec catchments and the hydrometric stations at catchment outlets.

6.3.2. Flow Ensemble Generation

This research uses the Bayesian rating curve (BaRatin) framework that is proposed by Le Coz et al. (2014) to estimate rating curve uncertainty. Section 6.3.2.1 provides a brief summary of the BaRatin framework by following Le Coz et al. (2014). Section 6.3.2.2 describes in detail the flow ensemble generation processes and results for 10 case studies.

6.3.2.1. Bayesian Rating Curve (BaRatin) Framework

Bayesian Inference

The BaRatin framework formalizes the overall relationship between stage-discharge measurements as:

$$\tilde{Q}_i = f\left(\tilde{H}_i|\boldsymbol{\theta}\right) + \epsilon_i^Q + \epsilon_i^f \quad (6\text{-}5)$$

$$\text{with } \epsilon_i^Q \sim N\left(0, \sigma_{Q_i}\right), \epsilon_i^f \sim N(0, \sigma_f)$$

where $\tilde{H}_i$ and $\tilde{Q}_i$ are the i^{th} measured water level and discharge ($i = 1,2, \dots, N$). N is the total number of gauging measurements ($\tilde{H}_i, \tilde{Q}_i$). Each gauging measurement is assumed independent. Stage measurements are assumed negligible in this research. $\boldsymbol{\theta}$ includes all the rating curve parameters, such as parameters a, b, and c in Equation (6-1). $f(h|\boldsymbol{\theta})$ is the rating curve equation in form of Equation (6-1). $f(h|\boldsymbol{\theta})$ can be given by a single flow control or a piecewise function of several flow controls (see Equation (6-10) for details).

ϵ_i^Q is the error of discharge measurements and assumed Gaussian distributed with zero mean and a known standard deviation σ_{Q_i} in this research. ϵ_i^f is the remnant error used to describe the rating curve function (f) uncertainty that cannot perfectly represent the stage-discharge relationship. ϵ_i^f is also assumed Gaussian distributed with zero mean, but its standard deviation σ_f is unknown and can be treated as a function of the rating curve discharge (Q).

$$\sigma_f = \gamma_1 + \gamma_2 Q \tag{6-6}$$

where γ_1 and γ_2 are two unknown parameters and need to be inferred with rating curve parameter $\boldsymbol{\theta}$. The discharge measurement error ϵ_i^Q and the remnant error ϵ_i^f are assumed independent.

The measured discharge $\tilde{Q}_i$ is essentially Gaussian distributed with mean $f(\tilde{H}_i|\boldsymbol{\theta})$ and standard deviation $\sqrt{{\sigma_{Q_i}}^2 + {\sigma_f}^2}$, with $\sigma_f = \gamma_1 + \gamma_2 Q$.

In Bayesian inference, the unknown parameters are $\theta, \gamma_1, \gamma_2$. Their posterior distribution is estimated by the Markov-Chain Monte Carlo (MCMC) sampling following:

$$p(\boldsymbol{\theta}, \gamma_1, \gamma_2|\tilde{\boldsymbol{H}}, \tilde{\boldsymbol{Q}}) \propto p(\tilde{\boldsymbol{Q}}|\theta, \gamma_1, \gamma_2, \tilde{\boldsymbol{H}}) \cdot p(\boldsymbol{\theta}, \gamma_1, \gamma_2) \tag{6-7}$$

where $p(\boldsymbol{\theta}, \gamma_1, \gamma_2|\tilde{\boldsymbol{H}}, \tilde{\boldsymbol{Q}})$ is the posterior distribution of all the parameters given the measured stage and discharge. $p(\tilde{\boldsymbol{Q}}|\boldsymbol{\theta}, \gamma_1, \gamma_2, \tilde{\boldsymbol{H}})$ is the likelihood of the measured discharge given the parameters and measured stage. $p(\tilde{\boldsymbol{Q}}|\boldsymbol{\theta}, \gamma_1, \gamma_2, \tilde{\boldsymbol{H}})$ is expressed as:

$$p(\tilde{\boldsymbol{Q}}|\boldsymbol{\theta}, \gamma_1, \gamma_2, \tilde{\boldsymbol{H}}) = \prod_{i=1}^{N} p_G\left(\tilde{Q}_i \middle| f(\tilde{H}_i|\boldsymbol{\theta}), \sqrt{{\sigma_{Q_i}}^2 + {\sigma_f}^2}\right) \tag{6-8}$$

where $\tilde{\boldsymbol{Q}} = (\tilde{Q}_1, \dots, \tilde{Q}_N)$ are N independent discharge measurements, $p_G(z|m, s)$ denotes the probability density function of a Gaussian distribution with mean m and standard deviation s, evaluated at the value z.

$p(\boldsymbol{\theta}, \gamma_1, \gamma_2)$ is the prior distribution of all the parameters. Assuming the prior of each parameter is independent, $p(\boldsymbol{\theta}, \gamma_1, \gamma_2)$ can be calculated by:

$$p(\boldsymbol{\theta}, \gamma_1, \gamma_2) = p(\gamma_1)p(\gamma_2)\prod_{j=1}^{N_{par}} p(\theta_j) \tag{6-9}$$

where $p(\theta_j)$ is the prior distribution of the j^{th} rating curve parameter $\boldsymbol{\theta}$ ($j = 1, \dots, N_{par}$). For example, if the rating curve is in the form of Equation (6-1), $\boldsymbol{\theta}$ include three parameters (a, b, c) and $N_{par} = 3$. The prior knowledge about $\boldsymbol{\theta}$ comes from the hydraulic analysis of the hydrometric gauge. Each parameter of $\boldsymbol{\theta}$ is assumed Gaussian distributed with the given prior mean and standard deviation. $p(\gamma_1)$ and $p(\gamma_2)$ are the prior distributions of the remnant error parameters. According to Le Coz et al. (2014), γ_1 and γ_2 are assumed to be uniformly distributed within [0,10000] by default.

Flow Controls

The significance of the BaRatin framework is using the hydraulic knowledge at the hydrometric station to constitute the prior knowledge of Bayesian inference. Therefore, the main step of the BaRatin method is conducting hydraulic analysis, including identifying flow controls and specifying the hydraulics informed priors for the parameters of each flow control.

A flow control refers to any feature that imposes a relationship between the flow depth and discharge in a channel (Osman Akan, 2006). For example, a critical flow section is a flow control because at this section, the Froude number is equal to one and $Q = \sqrt{gA^3/T}$, where Q is discharge, g is the gravitational acceleration, A is flow area, and T is top width. The normal flow section is also a flow control because at this section, the stage-discharge relation follows the Manning equation $Q = \frac{k_n}{n} A R^{2/3} S_0^{1/2}$, where k_n is a constant, n is the Manning roughness factor, R is the hydraulic radius, and S_0 is the longitudinal bottom slope of the channel. Likewise, various hydraulic structures that control flow, such as weirs and gates, are also flow controls.

Flow control is determined by the physical characteristic of the channel. For example, when the flow is low, the water level may be controlled by a small riffle. As the flow rises before overbank flow, the water level is controlled by the main river channel. When overbank flow happens, the water level is under both the main channel control and the floodplain control. Therefore, different flow controls are valid at different ranges of stage (named different segments), and multiple flow controls may be active within the same stage range. The total amount of flow is the sum of discharges within the active stage. It can be written using the following versatile equation for $N_{segment}$ stage ranges and $N_{control}$ flow controls.

$$Q = \sum_{s=1}^{N_{segment}} \mathbf{1}_{[k_{s-1}:k_s]} \times \sum_{j=1}^{N_{control}} M(s,j) \times a_j (h - b_j)^{c_j} \quad (6\text{-}10)$$

where k_{s-1} and k_s are the minimum and maximum stages for the s^{th}segment between which the s^{th} segment is valid. When water level is within the range of $[k_{s-1}: k_s]$, $\mathbf{1}_{[k_{s-1}:k_s]} = 1$; otherwise, $\mathbf{1}_{[k_{s-1}:k_s]} = 0$. k_{s-1} is also called the transition water level for the s^{th} segment above which the s^{th} segment is active. M is the flow control matrix. $M(s,j) = 1$ if the j^{th}flow control is active in the stage range $[k_{s-1}: k_s]$; otherwise, $M(r,j) = 0$.

Recall that b is the reference level. Users do not need to specify the priors for b. For the first segment, the reference water level is equal to the transition water level ($b_1 = k_1$). For the segment higher than the first one, b is calculated to maintain the continuity of the stage-discharge function at the transition water level. For example, there are two continuous segments 1 and 2, each segment has only one flow control. The first segment rating curve is known, and a_2 and c_2 of the second segment rating curve are also known, calculate b_2. At the transient water level ($h = k$), discharge is continuous, so $a_1(k - b_1)^{c_1} = a_2(k - b_2)^{c_2}$. b_2 can be calculated from the continuous equation.

Different from the reference level b, users need to define the priors for the transition water level k. The transition water level defines the stage above which the segment is active. Other than the first segment, the transition water level is different from the reference level ($k \leq b$).

Hydraulics Informed Priors

In the BaRatin framework, each flow control is associated with a stage-discharge relation in the form of Equation (6-1). Each of these parameters (a, b, c) can be related to the physical characteristics of the flow control. Because of this, it is possible to define the approximate values for each parameter of the rating curve and their uncertainties. Table 6-2 lists the stage-discharge relationships of the most typical flow controls and their correspondence to the rating curve parameter a of Equation (6-1) (Renard et al., 2016). All the physically meaningful parameters of Table 6-2 are assumed following the Gaussian distribution. For the parameters that have default priors in Table 6-2, the uncertainty is in the same unit as the parameter it is assigned to.

The transition water level k is not included in Table 6-2 because it is not specified for flow control but for segment. Each segment requires a prior estimation for k depending on the site stage-discharge condition. The gravitational acceleration g is also not included in Table 6-2 because it is assigned the same prior uncertainty in the BaRatin framework whenever applicable. $g \sim N(9.81, 0.005^2)$.

To estimate the parameters without the default priors, users do not have to have expertise in hydraulics. Useful information could come from observation of channel geometry and roughness, observation of flow patterns over a range of discharge values, expert knowledge of station managers, field surveys of cross-sections, photos of flows, topographic maps, flood marks, and GIS products such as Google Earth. In addition, numerical hydraulic models are helpful to identify the sequence of controls, including their zones of influence relative to stage and their association within complex channels.

To facilitate the use of the BaRatin method, Renard et al. (2016) developed a graphical interface - Baratin Advanced Graphical Environment (BaRatinAGE). BaRatinAGE allows users to enter the results of hydraulic analysis and the prior estimates, runs the MCMC sampling, and generates the parameter posterior distributions.

Table 6-2. Stage-discharge relationship of the most typical flow controls and their correspondence to the rating curve parameter a

Flow control type	**Assumptions**	**Stage-discharge relationship**	**Parameters without default priors**	**Parameters with default priors**	**Rating curve parameter a**
Rectangular channel (wide)	Wide rectangle channel (i.e., $H-b \ll B_w$), steady and uniform flow	$Q(H)=K_s\sqrt{S}B_w(H-b)^c$	• $K_s=$ flow resistance parameter $\left(m^{1/3}/s\right)=1/\mathrm{n}$ (n is Manning's n) • $S=$ longitudinal slope of the river channel • $B_w=$ width of the channel (m)	• $C_r=$ discharge coefficient $\sim N(0.4,0.05^2)$ • $c=$ exponent for a rectangular critical cross section $\sim N(1.67,0.025^2)$	$K_s\sqrt{S}B_w$
Parabolic channel (wide)	Wide parabolic channel (i.e., $H-b \ll \frac{3}{8}\frac{B_p^2}{H_p}$), steady and uniform flow	$Q(H)=K_s\sqrt{S}(\frac{2}{3})^{\frac{5}{3}}\frac{B_p}{\sqrt{H_p}}(H-b)^c$	• $K_s=$ flow resistance parameter $\left(m^{1/3}/s\right)=1/\mathrm{n}$ (n is Manning's n) • $S=$ longitudinal slope of the river channel • $B_p=$ width of the parabola (m) • $H_p=$ height of the parabola (m)	• $c=$ exponent for a parabolic channel control $\sim N(2.17,0.025^2)$	$K_s\sqrt{S}(\frac{2}{3})^{\frac{5}{3}}\frac{B_p}{\sqrt{H_p}}$
Triangular channel	Triangular channel, steady and uniform flow	$Q(H)=K_s\sqrt{S}\tan(\frac{v}{2})(\frac{\sin(v/2)}{2})^{2/3}(H-b)^c$	• $K_s=$ flow resistance parameter $\left(m^{1/3}/s\right)=1/\mathrm{n}$ (n *is* Manning's n) • $S=$ longitudinal slope of the river channel • $v=$ triangle open angle (degree)	• $c=$ exponent for a triangular channel control $\sim N(2.67,\ 0.025^2)$	$K_s\sqrt{S}\tan(v/2)(\frac{\sin(v/2)}{2})^{2/3}$
Rectangular weir	Perpendicular to mean flow, no backwater	$Q(H)=C_r\sqrt{2g}B_w(H-b)^c$	• $B_w=$ width of the spillway (m)	• $C_r=$ discharge coefficient $\sim N(0.4,0.05^2)$ • $c=$ expnent for a rectangular critical cross section $\sim N(1.67,0.025^2)$	$C_r\sqrt{2g}B_w$
Parabolic weir	Perpendicular to mean flow, no backwater	$Q(H)=C_p\sqrt{2g}\frac{B_p}{\sqrt{H_p}}(H-b)^c$	• $B_p=$ width of the parabola (m) • $H_p=$ height of the parabola (m)	• $C_p=$ discharge coefficient $\sim N(0.22,0.02^2)$ • $c=$ exponent for a parabolic critical cross section$\sim N(2.0,0.025^2)$	$C_p\sqrt{2g}\frac{B_p}{\sqrt{H_p}}$
Triangular weir	Perpendicular to mean flow, no backwater	$Q(H)=C_t\sqrt{2g}\tan(\frac{v}{2})(H-b)^c$	• $K_s=$ flow resistance parameter $\left(m^{1/3}/s\right)=1/\mathrm{n}$ (n is Manning's n) • $S=$ longitudinal slope of the river channel • $v=$ triangle open angle (degree)	• $C_t=$ discharge coefficient $\sim N(0.31,0.025^2)$ • $c=$ exponent for a parabolic critical cross section $\sim N(2.5,0.025^2)$	$C_t\sqrt{2g}\tan(\frac{v}{2})$
Free-flowing orifice	Perpendicular to mean flow, no backwater	$Q(H)=C_o\sqrt{2g}A_w(H-b)^c$	• $A_w=$ cross section area of the orifice (m^2)	• $C_o=$ discharge coefficient $\sim N(0.6,0.05^2)$ • $c=$ exponent for a parabolic critical cross section $\sim N(0.5,0.025^2)$	$C_o\sqrt{2g}A_w$

Backwater Effect

A shortcoming of the BaRatin framework is that currently it does not deal with the rating procedures when hydrometric station is subject to backwater. Backwater refers to water being backed in the course. When backwater occurs, stage becomes higher than the smallest stage required for passing the same amount of discharge.

There are two kinds of backwater and both complicate rating curve development. One is constant backwater. It is caused by a temporary modification of the control, such as the debris, vegetation, or beaver dam obstructing the channel or riffle. Constant backwater effect can be solved by estimating a shifted rating curve valid for a given period. Another is variable backwater. Variable backwater occurs when the physical properties of the downstream course that contains the gauging station change. Changes can be caused by confluent streams, lakes, dams with movable gates, tide or water returning from overbank flow (Petersen-Øverleir and Reitan, 2009). In cold regions, variable backwater also happens when (a) no thaw occurs, (b) ice or frail ice dam forms upstream, (c) heavy snow increases the pressure on the upstream ice cover, or (d) spring flood rises after the ice breakup. Variable backwater effect can be solved when there are two gauges available for rating curve estimation (e.g., see Mansanarez et al., 2016; Petersen-Øverleir and Reitan, 2009).

In this thesis, since the BaRatin method does not handle backwater effect, its rating curve uncertainty estimate is only valid for the non-backwater affected data. As a result, flow ensemble is only generated for the non-backwater period, and the proposed climate-flow ensemble based calibration framework is also only applied to the non-backwater period.

6.3.2.2. Case Study Flow Ensemble Generation

This research took the advantage of BaRatinAGE and conducted five steps to generate flow ensemble. The first three steps are used to prepare the necessary inputs for the BaRatinAGE software: (1) identify flow controls for a hydrometric station; (2) specify hydraulics informed priors for flow control parameters; (3) collect instantaneous stage-discharge measurements. With the hydraulics informed priors and measurement data, BaRatinAGE will generate MCMC parameter samples. With the BaRatinAGE generated parameters, two more steps are needed to generate flow ensemble: (4) test parameter convergence and (5) propagate parameter uncertainty to flow.

These steps were accomplished with the aids of the Google maps, the near-global bankful widths and depths database that is developed by Andreadis et al. (2013), and the rating curve development relevant data from the Direction de l'Expertise Hydrique. Detailed processes are explained below.

(1) Flow Controls

Based on the observation of the Google maps views of the 10 hydrometric stations, we divided all their water levels into three segments and chose 1-2 flow controls to represent each segment's stage-

discharge relationship (Table 6-3): (a) low flow segment controlled by a natural riffle, (b) low flow segment controlled by a main channel, (c) high flow segment controlled by a main channel and a floodplain channel. For demonstration, Figure 6-2 illustrates the Google maps views of two hydrometrics stations (the 2nd and 7th hydrometric stations of Table 6-1).

Table 6-3. Segments and flow control configuration for all the hydrometric station

	Control 1: Riffle	Control 2: Main channel	Control 3: Floodplain
Segment 1	√		
Segment 2		√	
Segment 3		√	√

(2) Hydraulics Informed Priors

To determine the stage-discharge relationship, the riffle control is approximated as a rectangular weir, the main channel and floodplain channel are both approximated as a rectangular channel (wide). According to Table 6-2, the stage-discharge relationship of the rectangular weir follows $Q(H) = C_r\sqrt{2g}B_w(H-b)^c$, the unknown parameter is B_w. The stage-discharge relationship of the rectangular channel (wide) follows $(H) = K_s\sqrt{S}B_w(H-b)^c$, the unknown parameters are K_s, S and B_w. The reasoning of each unknown parameter prior, including the transition water level, is explained below.

- k is the transition water level above which the segment becomes active.
    - For the first segment, the mean value of the transition water level is assumed as the minimum stage of the operational stage-discharge check table. The operational stage-discharge check table essentially includes the empirical stage-discharge correspondence values at an extensive stage-discharge range. Our check tables are provided by the Direction de l'Expertise Hydrique.
    - For the second segment, the mean value of the transition water level is calculated as the mean stage of the historical 15-min stage records (https://www.cehq.gouv.qc.ca/hydrometrie/historique_donnees/default.asp).
    - For the third segment, the mean value of the transition water level is calculated as the first segment transition water level plus the bankful flow depth. The bankful flow depth is derived from the bankful width and depth database (Andreadis et al., 2013).
- B_w is the width of the spillway or channel.
    - For the riffle control and main channel control, their mean width values are assumed the same for the lack of the width information. The mean width is estimated by the measure distance tool of the Google maps. The standard deviation is subjectively decided. In most cases, the standard deviation is assumed 10m and 15m for the riffle control and main channel control, respectively.

- For the floodplain control, the mean channel width is assumed equal to the bankful channel width and is derived from the bankful width and depth database (Andreadis et al., 2013). The standard deviation is assumed 5m.

(a) Hydrometric station 050408

(b) Hydrometric station 061901

Figure 6-2. Google maps views of hydrometrics stations 050408 and 061901 (the 2nd and 7th hydrometric stations of Table 6-1).

- K_s is the flow resistance parameter (Strickler coefficient) and can be estimated based on $K_s = 1/\mathrm{n}$, where n is the Manning's n.
    - For the main channel control, the channel is assumed clean, straight, full stage, no rifts or deep pools. According to Chow (1959), such a channel's Manning's n mean is equal to 0.03. Its standard deviation is assumed 0.005.
    - For the floodplain control, the channel is assumed having heavy stand of timber, a few down trees, little undergrowth, flood stage below branches. According to Chow (1959), such a channel's Manning's n mean is equal to 0.1. The standard deviation is assumed 0.1.
- S is the longitude slope of the channel. The main channel control and the floodplain control are assumed sharing the same prior of S. The mean channel slope is calculated as the ratio of the elevation difference between the most upstream and the most downstream points of the channel to the channel length. The channel elevations and lengths are obtained from the bankful width and depth database (Andreadis et al., 2013). Since all the 10 hydrometric stations are close to the St. Lawrence river, their channel slope standard deviation is assumed very small.

Table 6-4 lists the hydraulic priors of the three controls for all the 10 hydrometric stations. Note that K_s is not included in the table as its priors are assumed the same across all the hydrometric stations. K_s is calculated based on $K_s = 1/\mathrm{n}$. For the main channel control, $\mathrm{n} \sim \mathrm{N}(0.03, 0.005^2)$; for the floodplain control, $\mathrm{n} \sim \mathrm{N}(0.1, 0.05^2)$. Due to the lack of precise cross-section configurations, most properties of the flow controls are not known for sure, so their standard deviations are set high. In contrast, the properties that are derived from the bankful width and depth database have relatively low uncertainty because of the reference values from the bankful width and depth database (Andreadis et al., 2013). Even if the parameter uncertainty is high, it is not a problem for Bayesian inference as long as there are sufficient stage-discharge measurement data (Le Coz et al., 2014).

Table 6-4. Hydraulics informed priors of the 10 Québec hydrometric stations. Mean value followed by standard deviation in brackets. k is the transition water level. B_w is the width of the spillway or channel. K_s is the flow resistance parameter. S is the longitude slope of the channel.

No.	**Hydrometric station**	**Segment 1**	**Segment 2**	**Segment 3**	**Control 1: Riffle**	**Control 2: Main channel**		**Control 3: Floodplain**	
		k (m)	k (m)	k (m)	B_w (m)	B_w (m)	S (-)	B_w (m)	S (-)
1	022301	27.08 (0.25)	27.58 (0.50)	28.63 (0.05)	40 (10)	40 (15)	0.0119 (0.005)	68 (5)	0.0119 (0.005)
2	050408	26.45 (0.25)	27.25 (0.50)	27.96 (0.05)	55 (10)	55 (15)	0.0119 (0.001)	66 (5)	0.0119 (0.001)
3	050409	27.58 (0.25)	28.05 (0.50)	28.73 (0.05)	22 (10)	22 (5)	0.0048 (0.002)	46 (5)	0.0048 (0.002)
4	052805	27.11 (0.25)	27.52 (0.50)	29.50 (1.00)	30 (10)	30 (10)	0.0048 (0.001)	39 (10)	0.0048 (0.002)
5	061020	27.49 (0.25)	27.75 (0.50)	29.18 (0.25)	73.5 (10)	74 (15)	0.0024 (0.001)	76 (5)	0.0024 (0.001)
6	061502	27.02 (0.25)	27.61 (0.50)	29.15 (0.05)	100 (10)	100 (15)	0.0075 (0.0035)	102 (5)	0.0075 (0.0035)
7	061901	22.58 (0.25)	24.05 (0.50)	26.18 (0.05)	195 (10)	195 (15)	0.0019 (0.001)	199 (5)	0.0019 (0.001)
8	061905	28.69 (0.25)	29.39 (0.50)	31.88 (0.05)	136 (10)	136 (15)	0.0041 (0.002)	170 (5)	0.0041 (0.002)
9	062102	25.90 (0.25)	26.87 (0.50)	29.15 (0.05)	110 (10)	110 (15)	0.0036 (0.0015)	175 (5)	0.0036 (0.0015)
10	062701	28.57 (0.25)	29.13 (0.50)	33.00 (1.00)	25 (5)	25 (8)	0.0140 (0.005)	35 (10)	0.0140 (0.005)

(3) Stage-Discharge Measurements

Stage-discharge measurements are expected to be representative of the stage-discharge relation for the study period and cover all the possible stage ranges, though extremely low and high flow conditions rarely happen. In this research, the instantaneous stage and discharge measurements are provided by the Direction de l'Expertise Hydrique. The number of stage-discharge measurements in each station varies from 7 to 58 (Table 6-5).

Table 6-5. Number of stage-discharge measurements of the 10 hydrometric stations

No.	Hydrometric station	Number of measurements	No.	Hydrometric station	Number of measurements
1	022301	42	6	061502	48
2	050408	14	7	061901	47
3	050409	16	8	061905	31
4	052805	58	9	062102	41
5	061020	23	10	062701	7

In this research, the stage measurement errors are assumed negligible. The discharge measurement errors are assumed normally distributed with zero mean and standard deviation 2.5% of the rating curve estimated discharge, given that the discharge at the Québec sites is measured by the acoustic dropper current profiler (ADCP) (Le Coz et al., 2014).

(4) Convergence Test

Although BaRatinAGE facilitates the implementation of the BaRatin method, it does not assess parameter convergence before utilization. By default, BaRatinAGE discards the first half of MCMC iterations for burning. Since this empirical setup may not work for all cases, we implemented a convergence test prior to uncertainty propagation. Here the Gelman-Rubin convergence diagnostic (Gelman, 1996) was adopted to test the convergence of the BaRatinAGE MCMC sampling results. Five parallel MCMC samplings were run with widely dispersed initial parameter values. In each MCMC sampling, the number of cycles was set to 100, the number of jump variance adaptions was set to 1,000, so a total of $100 \times 1{,}000 = 100{,}000$ parameter set samples were generated. The Gelman-Rubin convergence diagnostic R was calculated for each parameter based on the 100,000 parameter samples of each of the five parallel samplings. Results show that for all the parameters and all the 10 stations, their parameter samples are converged after the first half of iterations.

After the convergence test, 5,000 parameter sets were selected from the second half of a MCMC sampling sequence (one of every 10 parameter sets), forming 5,000 rating curves. Based on the 5,000 rating curves, BaRatinAGE computed the 95% rating curve uncertainty intervals and identified the most likely rating curve among the 5,000 rating curves. Figure 6-3 illustrates the 95% rating curve uncertainty intervals

and the most likely rating curve in comparison with the operational rating curve for stations 022301 and 062701. Only the range with discharges less than 300 m^3/s is plotted here for a fair comparison, though the rating curve uncertainty boundaries are extrapolated over all possible stage-discharge ranges which differ from station to station.

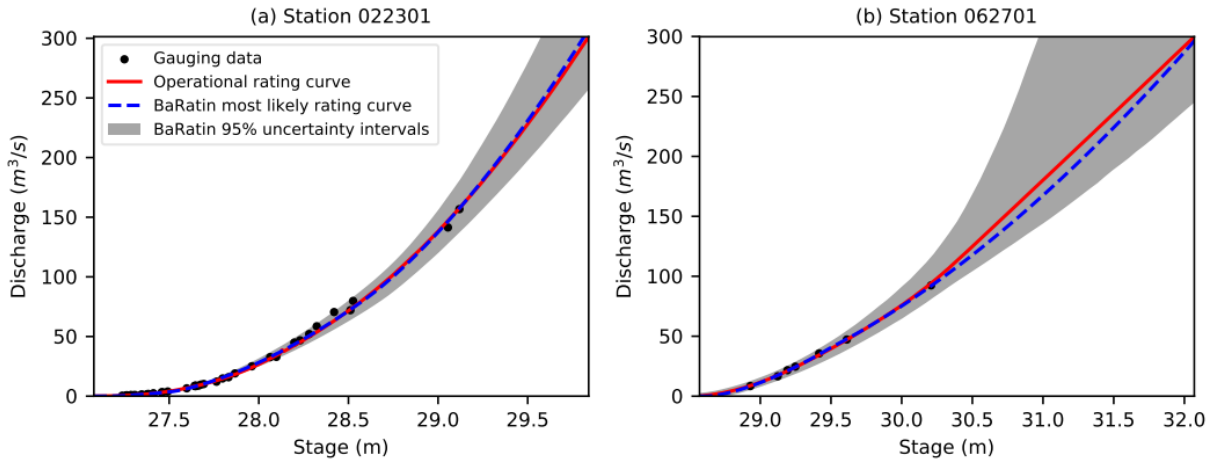

Figure 6-3. Comparison between the BaRatin rating curve estimates and the operational rating curve of gauge stations 022301 and 062701 (the 1st and 10th stations of Table 6-1). Gauging data are the stage-discharge measurements that are used in Bayesian inference.

In Figure 6-3, the 95% rating curve uncertainty intervals always cover the operational rating curve with variable widths. The uncertainty bounds are narrow when there is a large number of measurements and become wide when there is a small number of measurements. Similarly, the BaRatin generated most likely rating curve matches the operational rating curve well when there are many measurements but deviates from the operational curve when there are sparse or no measurements. This phenomenon is more obvious for station 062701 when its discharge is larger than 100 m^3/s and no gauging data is available, its rating curve uncertainty boundaries drastically spread out, and the most likely rating curve stays away from the operational curve. Therefore, by using more gauging data, station 022301 gets more accurate and precise rating curve estimates than station 062701. The similar consistency and boundary change trend are also found in all other hydrometric stations (not shown here), which demonstrate the robustness of the BaRatin method.

(5) Uncertainty Propagation to Flow

Given a total of M rating curve parameter sets from the parameter posterior distributions, M daily flow realizations can be computed. Taking one rating curve parameter set as an example, its corresponding flow realization is calculated as:

$$Q_i = f(h|\boldsymbol{\theta}_i) + \epsilon_i^f \tag{6-11}$$

$$\text{where } \epsilon_i^f \sim N(0, \gamma_{1,i} + \gamma_{2,i} \cdot f(h|\boldsymbol{\theta}_i))$$

where h is the stage measurement, $\boldsymbol{\theta}_i$ is the i^{th} rating curve parameter set ($i = 1,2, \dots, M$), $f(h|\boldsymbol{\theta}_i)$ is the rating curve computed discharge, ϵ_i^f is the i^{th} remnant error and assumed uncorrelated in time. The remnant error is assumed Gaussian with zero mean and standard deviation proportional to the rating curve calculated discharge. $\gamma_{1,i}$ and $\gamma_{2,i}$ are the standard deviation parameters of ϵ_i^f. Each realization of the remnant error depends on sampling a normally distributed random variable.

The instantaneous flow realization is calculated based on Equation (6-11). The calculation needs the instantaneous stage measurements (every 15 minutes in our case) and one realization of the remnant error. The instantaneous stage measurements are provided by the Direction de l'Expertise Hydrique in this study. Based on the calculated instantaneous flow, the daily mean flow realization conditioned on the i^{th} rating curve is computed as the average of all the realizations of the calculated instantaneous flows on the same day. Finally, the overall flow measurement uncertainty is described empirically by all the *M* daily mean flow realizations.

To reduce the computational cost of assessing model performance and keep the representativeness of our rating curve uncertainty estimation, this research used 100 of the 5,000 rating curves to generate flow ensemble for each hydrometric station. The 100 rating curves were randomly selected without replacement. Figure 6-4 compares the generated 100-member flow ensemble with the observed flow for stations 022301 and 062701. The backwater period is blank (mainly from January to March and the late December).

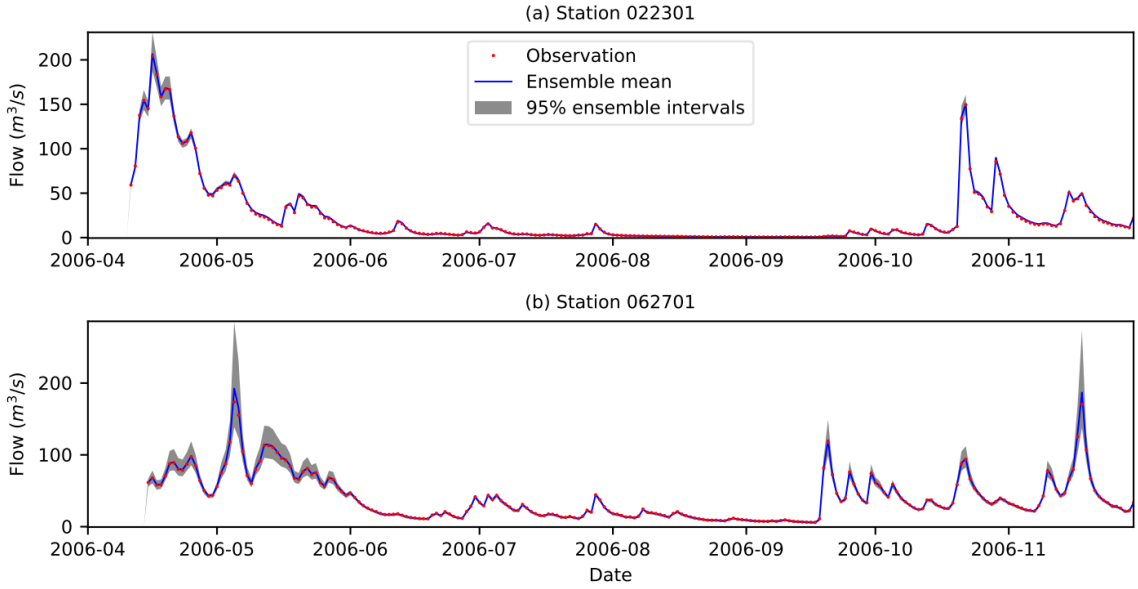

Figure 6-4. 100-member flow ensembles and the flow observations of hydrometric stations 022301 and 062701 (the 1st and 10th stations of Table 6-1) in April-November 2006.

For both hydrometric stations, the ensemble mean flow matches the observed flow very well, and the 95% flow intervals are narrow for low flow and wide for high flow. These result are consistent with the

rating curve estimation results of Figure 6-3. In Figure 6-3, when flow is less than 150 m^3/s for 022301 and 100 m^3/s for 062701, all the 5,000 rating curves are concentrated around the operational rating curve, and the 95% rating curve uncertainty intervals are relatively small. As a result, in Figure 6-4, though only 100 rating curves are used, the 100 flow ensemble members are highly concentrated around the operational rating curve derived flow record (observation). When flow is higher than 150 m^3/s for 022301 and 100 m^3/s for 062701, the flow uncertainty intervals rapidly increase proportional to the flow magnitude. Moreover, since station 062701 rating curve is more uncertain than station 022301 (Figure 6-3), the flow uncertainty bounds of station 062701 are wider than those of station 022301 (Figure 6-4).

6.3.3. Comparative Model Calibration Setup

In order to compare the impact of the explicit consideration of flow data uncertainty on model calibration, three comparative model calibration approaches are built for each case study. The numbering of calibration approaches is in line with Chapter 3. Approach 2 is a benchmark without considering any climate and flow data uncertainty. Approach 3 is the climate ensemble based calibration. Approach 4 is the climate-flow ensemble based calibration. Some main calibration setups are listed in Table 6-6. Comparisons between approaches 2 and 3 will reveal the impact of the explicit climate data uncertainty. Comparisons between approaches 3 and 4 will reveal the impact of the explicit flow data uncertainty. The magnitudes of the climate and flow data uncertainty impacts indicate the relative influence of the climate data uncertainty and the flow data uncertainty on model calibration.

Table 6-6. Three comparative calibration setups for the case study

Calib. approach	Calib. algorithm	Calib. period climate data	Calib. period flow data	Objective function
2	DDS-AU	Deterministic	Deterministic	KGE
3	DDS-AU	Ensemble	Deterministic	KGE
4	PADDS	Ensemble	Ensemble	KGE & MCRPS

Same as in Chapter 3, the warm-up, calibration and validation periods are from October 1, 1997 to September 30, 1998, from October 1, 1998 to September 30, 2005, and from October 1, 2005 to September 30, 2010, respectively. Recall that due to the BaRatin method limit to backwater effect, our flow ensemble was only generated for the non-backwater period. Therefore, the three calibration approaches are conducted only on the non-backwater affected days to ensure a fair comparison.

(1) Calibration Approach 2: Parameter Uncertainty

Calibration approach 2 uses the measured climate as model input and identifies numerous behavioral parameter sets by the dynamically dimensioned search - approximation of uncertainty (DDS-AU) algorithm (Tolson and Shoemaker, 2008). The DDS-AU calibration is based on the results of 404 independent DDS optimization trials. Each DDS optimization trial uses 400 model evaluations, and each trial is initialized to

a different initial solution and a different random seed. Only the final best DDS solution (one per trial) is considered as a possible behavioral parameter set for filtering. Therefore, for each case study, the total number of model evaluations or total computational cost is 161,600 ($404 \times 400 = 161{,}600$), and the total number of candidate parameter solutions is 404.

Since approach 2 does not explicitly consider flow data uncertainty, the objective function compares the simulated flow with the observed flow based on the KGE.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{6-12}$$

where r is the linear correlation coefficient between the simulated and observed flows, α is the ratio between the simulated flow standard deviation and the observed flow standard deviation, and β is the ratio between the mean simulated flow and the mean observed flow. The range of KGE is between negative infinity to one with the optimal value of one.

With all the candidate behavioral parameter sets, the behavioral parameter sets are selected based on the optimal criteria-aggregation-based approach (Shafii et al., 2015) as described in Section 3.2.1 of Chapter 3. Note that the objective function (KGE) is to be maximized here.

In model validation, each of the obtained behavioral parameter sets is validated with the measured climate. All model outputs are assumed equally likely and combined to constitute ensemble flow predictions.

(2) Calibration Approach 3: Parameter and Climate Uncertainty

In addition to parameter uncertainty, approach 3 explicitly considers climate data uncertainty by using each of the 101 climate ensemble members (100 from the N15 derived climate ensemble and one from the observation) for model calibration. A random sampling of the climate ensemble is unnecessary in this application because of the small size of the available ensemble. The calibration algorithm and objective function are the same as in approach 2. For each climate ensemble member, DDS-AU runs 4 independent DDS optimization trials to get candidate parameter samples. Each DDS optimization trial is initialized to a different initial solution and a different random seed. Only the final best DDS solution (one per trial) is considered as a possible behavioral parameter set for filtering. The number of model evaluations per DDS optimization trial (400) is set the same as in approach 2 to ensure the same optimization effort per DDS optimization trial. As such, the total computational cost ($404 \times 400 = 161{,}600$) and the total number of candidate parameter sets (404) are the same for approaches 2 and 3 to fairly compare how each calibration approach performs.

The candidate parameter sets are filtered to identify the behavioral parameter sets based on the same behavioral solution identification approach as used in approach 2. The behavioral parameter sets and their corresponding climate ensemble members are used to propagate uncertainty to flow prediction. Each of the predicted flows is assumed equally likely and used to constitute ensemble of streamflow predictions.

(3) Calibration Approach 4: Parameter and Climate-Flow Uncertainty

On the basis of approach 3, approach 4 adds the explicit consideration of flow data uncertainty. The objective functions include KGE and MCRPS. KGE is used to ensure the consistency with observed flow and is computed based on Equation (6-12). MCRPS is used to ensure the similarity with observed flow ensemble and is computed based on Equations (6-2), (6-3), and (6-4). The two objectives are assumed equally important in our experiments. The Pareto archived dynamically dimensioned search algorithm (PADDS) (Asadzadeh and Tolson, 2013) is used to sample parameters based on the two objectives.

Further, the Pareto rank-based behavioral solution identification approach (Shafii et al., 2015) is used to select behavioral parameter solutions. The Pareto rank-based behavioral solution identification approach is for multi-objective calibration and is different from the optimal criteria-aggregation-based approach used in calibration approaches 2 and 3. Details are as follows.

a) Sort all the candidate parameter sets based on the concept of non-domination and assign a Pareto rank to each parameter set. The total number of the Pareto ranks is r_{max}. r_{max} varies with case study.

b) Create rank-to-r candidate behavioral parameter groups.

   Each group includes all the candidate parameter sets whose Pareto ranks are less than or equal to r. In total, r_{max} parameter groups are formed corresponding to r_{max} Pareto ranks.

c) For each candidate behavioral parameter group, evaluate its calibration period flow prediction results by calculating reliability and sharpness (R, S).

d) Remove reliability and sharpness (R, S) outliers.

   Among the r_{max} evaluation results (R, S), remove the outliers that perform very poor for one of the two metrics before the calculation of the distance-to-ideal (DTI). To identify the outliers of (R, S), the Mahalanobis distance is adopted because it is suitable to measure the distance between correlated variables, such as reliability and sharpness (Cunderlik and Burn, 2006). In detail, calculate the Mahalanobis distance between each (R, S) and the mean reliability and sharpness $(\bar{R}, \bar{S})$; identify the outlier as the (R, S) whose Mahalanobis distance is more than two times standard deviation away from the mean Mahalanobis distance; the remaining (R, S) pairs and their corresponding parameter groups are taken to the next step.

e) For the candidate behavioral parameter groups with outliers removed, calculate their distance-to-ideal (DTI) values.

   DTI is calculated as the Euclidean distance between (R, S) and (R_{best}, S_{best}). R_{best} and S_{best} are the individual best reliability and sharpness values, respectively, of all the remaining parameter groups and define the ideal point in the objective space.

f) The candidate behavioral parameter group with the minimum DTI is identified as the behavioral parameter group.

In the PADDS based parameter sampling, the number of model evaluations per DDS optimization trial (400) is set the same as in calibration approaches 2 and 3 to ensure the same optimization effort per DDS optimization trial. The computational cost of approach 4 is increased relative to approaches 2 and 3 due to the calculation of the additional objective function - MCRPS.

In model validation, the behavioral parameter sets and their corresponding climate ensemble members are used to propagate uncertainty to flow prediction. Each of the predicted flows is assumed equally likely and used to constitute the ensemble of streamflow predictions.

6.3.4. Evaluation of Flow Prediction and Parameter Estimation

The flow prediction is assessed from two perspectives: deterministic and probabilistic. The evaluation metrics are selected to be consistent with Chapter 3, including KGE, reliability, spread, and reliability. Their calculation equations are the same as in Section 3.3.6 of Chapter 3. In addition, the mean absolute error of the reliability diagram (MaeRD) is computed to assess the average distance between the forecast probability and the observation probability over all quantiles of interest (Thiboult et al., 2016). The MaeRD calculation equation is the same as Equation (4-13) in Section 4.2.5 of Chapter 4.

The parameter estimates are evaluated from two aspects: parameter histogram and the number of behavioral parameter sets. The evaluation methods are also the same as the parameter evaluation of Section 3.3.6 of Chapter 3. Given the same model performance metric levels for two calibration approaches, the one with a smaller number of behavioral parameter sets is preferred (assuming modelers prefer quicker uncertainty propagation experiments).

6.4. Results

The comparison results of the three calibration approaches are presented in two sections. Section 6.4.1 is the evaluation of flow predictions. Section 6.4.2 is the evaluation of parameter estimates. Note that all the reported evaluation results on flow predictions are for the non-backwater days of the validation period.

6.4.1. Evaluation of Flow Predictions

Figure 6-5 details the flow prediction evaluation metrics results of calibration approaches 2, 3, and 4 for all 10 catchments. For the deterministic flow prediction, approach 2 already achieves practically good KGEs (i.e., higher than 0.80). The climate ensemble based calibration (approach 3) and the climate-flow ensemble based calibration (approach 4) maintain these good performances.

For the probabilistic prediction, on average, approach 3 improves the reliability by 91% over approach 2 for all catchments, approach 4 improves the reliability by 5.7% over approach 3 for 7 catchments (the 4th-

10th catchments). The reliability increase of approach 3 over approach 2 benefits from the explicit consideration of climate data uncertainty, and the reliability increase of approach 4 over approach 3 benefits from the explicit consideration of flow data uncertainty.

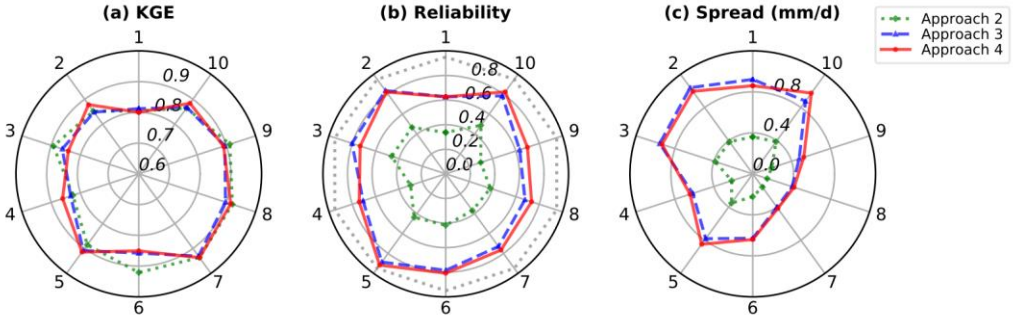

Figure 6-5. Flow predication evaluation metrics results of calibration approaches 2, 3, and 4 for all 10 catchments. See Table 6-6 for the descriptions of approaches 2, 3, and 4. Each catchment is identified by the label on the outer edge of the wheel. The metric value of each catchment is represented by the value on the catchment corresponding spoke. The dotted line of panel (b) represents the reliability value of 0.95.

A more complete reliability comparison between approaches 2, 3, and 4 can be seen from the reliability diagram (Figure 6-6). All approaches are under-dispersion and over confident. Although the reliability is not perfect for approaches 3 and 4, both approaches reduce the difference from the perfect reliability (i.e., diagonal) to a great extent compared with approach 2. In particular, in Figure 6-6c, the reliability of the Aux Ecorces catchment (the 5th catchment of Table 6-1) is almost perfect for all percentiles of interest with approach 4. The MaeRD evaluation results provide a quantitative measurement of the difference from the perfect reliability (Figure 6-6d). On average, approach 3 reduces the MaeRD by 56% over approach 2 for all catchments, approach 4 reduces the MaeRD by 12% over approach 3 for the 4th -10th catchments.

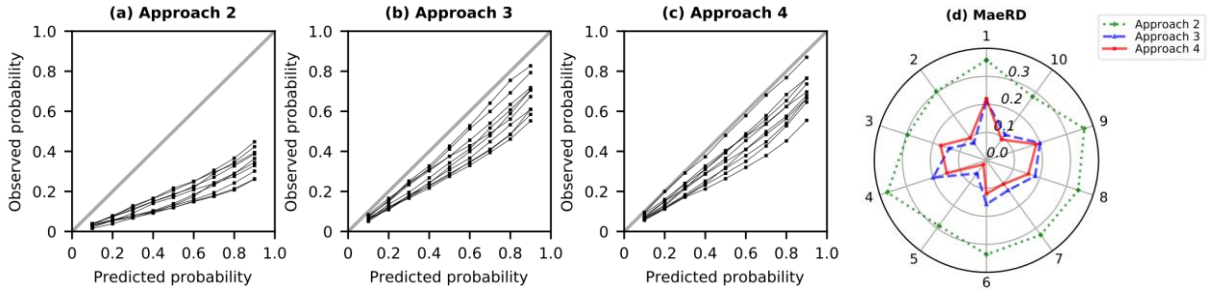

Figure 6-6. Reliability diagrams and MaeRD evaluation results of calibration approaches 2, 3, and 4 for all 10 catchments. Each curve of a reliability diagram refers to a catchment. The diagonal represents the perfect reliable prediction.

The reliability improvement is at the cost of the increased prediction spread. As shown in Figure 6-5c, on average, approach 3 raises the spread by 156% over approach 2 for all catchments, approach 4 raises the spread by 6% over approach 3 for the 4th -10th catchments. Compared with the magnitude of reliability

increases (56% and 12% for approaches 3 and 4, respectively), the raised prediction intervals may be welcome for decision makers.

The probabilistic prediction results also suggest that given a hydrologic model, the climate data uncertainty has greater impacts on flow predictions than the flow data uncertainty. The flow prediction changes due to the explicit recognition of flow data uncertainty are minor compared with the flow prediction changes due to the explicit recognition of climate data uncertainty. This finding is consistent with the result of Kuczera et al. (2006) showing the dominated impact of the forcing data uncertainty on overall flow prediction uncertainty relative to the flow data uncertainty.

6.4.2. Evaluation of Parameter Estimates

The Wilcoxon rank-sum hypothesis test (Helsel and Hirsch, 2002) is conducted to compare the posterior parameter distributions of approaches 2, 3, and 4 for 10 case studies. Result shows that approaches 2 and 3 generate different parameter distributions for 63.3% of the 60 parameter estimations (6 parameters $\times$10 catchments); approaches 3 and 4 generate different parameter distributions for 78.3% of the 60 parameter estimations with a 5% level of significance. This finding reveals that perturbed climate data generate different parameter distributions from observed climate, and perturbed flow data generate different parameter distributions from observed flow. Therefore, parameter uncertainty estimation is dependent on the forcing and response data that are used to drive and evaluate hydrologic models, respectively.

To visualize the estimated parameter differences between approaches 2, 3, and 4, Figure 6-7 compares the posterior parameter histograms of all six parameters for the Trois Pistoles and Aux Ecorces catchments (the 1st and 5th catchments of Table 6-1). Based on the Wilcoxon rank-sum test, parameters X2, X3, X5, and X6 for Trois Pistoles, and parameters X2, X3, X4, X5, and X6 for Aux Ecorces are significantly different between approaches 2 and 3. Parameters X1, X2, X3, and X6 for Trois Pistoles, and parameters X1, X2, and X4 for Aux Ecorces are significantly different between approaches 3 and 4.

In terms of the number of parameter solutions, approaches 2, 3, and 4 generate on average 122, 95, and 337 behavioral parameter sets per catchment, respectively. The reason why approach 4 gets more parameter solutions than approaches 2 and 3 is that approaches 2 and 3 work on a single-objective optimization question and take only the best parameter set per DDS optimization trial as the candidate parameter solution. However, approach 4 solves a multi-objective optimization problem and takes at least one non-dominated parameter solution per DDS optimization trial as the candidate parameter solution. Although the total number of solutions of approach 4 is big, its flow predictions are not substantially expanded (as shown in Figure 6-5). This is actually because many sets of the approach 4 solutions differ in only one or two parameter values given the nature of the DDS algorithm that is used in the PADDS.

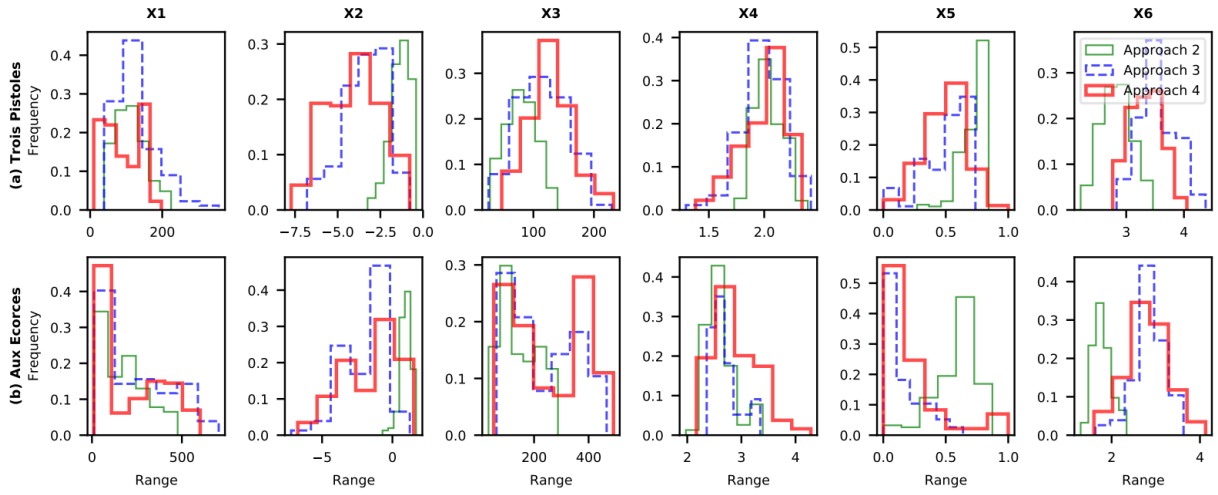

Figure 6-7. Comparative histograms of all six parameter estimates of approaches 2, 3, and 4 for the Trois Pistoles and Aux Ecorces catchments (the 1st and 5th catchments of Table 6-1).

6.5. Conclusions and Future Work

This chapter addressed the explicit consideration of flow data uncertainty by using flow ensemble and formalized the climate-flow ensemble based hydrologic model calibration framework. The framework was built on Chapter 3 proposed climate ensemble based model calibration framework. The flow ensemble was generated by the hydraulics-based Bayesian rating curve uncertainty estimation method (BaRatin) (Le Coz et al., 2014). The continuous ranked probability score (CRPS) was taken as an objective function of the framework to compare the scalar model prediction with the measured flow ensemble.

The proposed framework provides a simple way to explicitly consider flow data uncertainty in likelihood-free calibration methods. Compared with other likelihood-free calibration methods that explicitly consider flow data uncertainty, this framework features in: (1) using the physically-based flow uncertainty estimates to represent flow data uncertainty, and (2) using the uncertainty-based metric to evaluation the calibration period model performance. Our employed BaRatin method has been demonstrated providing realistic flow uncertainty estimates and is easy to use based on its supportive software BaRatinAGE.

The framework was assessed against two calibration approaches. One considers parameter uncertainty, another considers parameter and climate data uncertainty. Assessment of the framework is based on 10 case studies, so the results provide a better assessment of the average performance of the framework. The assessment focused on the validation period performance. Results show that explicitly considering the flow data uncertainty maintains the flow prediction accuracy and slightly improves the flow prediction reliability. Compared with the climate data uncertainty, the flow data uncertainty plays a minor role of improving predictions. Results also demonstrate the dependency of parameter uncertainty on forcing and response data.

The conclusion that flow data uncertainty plays a minor role of improving predictions is drawn on the fact that our case study catchments have good operational rating curve management and flow records, so the estimated flow uncertainty is small and its impact on parameter estimation and flow prediction is limited. More experiments need to be conducted in poorly gauged areas in order to: (1) fairly evaluate the importance of flow data uncertainty on flow prediction, and (2) figure out the applicability of the flow uncertainty based model calibration.

The proposed calibration framework is flexible enough to work with a variety of likelihood-free calibration methods and both lumped and distributed hydrologic modeling applications that deal with flow data uncertainty. In the future, the climate-flow ensemble based calibration framework needs to be tested with distributed hydrologic models, especially the models that are sensitive to river channel or reservoir routing and have relatively detailed routing process descriptions.

Chapter 7.
Conclusions and Future Work

This thesis presented a quantitative solution to measure the discretization-induced information loss of input data and applied advanced climate and flow data uncertainty estimation products and methods to hydrologic model calibration and data assimilation. Conclusions from each chapter (except Chapter 2) below are based on no less than ten case studies. Therefore, the findings should be able to extend to other cases studies. The following sections outline the conclusions, limitations, and recommendations for future works.

7.1. Conclusions

Chapter 2 proposed *a priori* discretization error metrics that can estimate the information loss for candidate discretization scheme. As a potential application of the proposed error metrics, Chapter 2 also formulated a two-step discretization decision-making approach to help modelers to decide the appropriate subbasin and HRU (hydrologic response unit) discretization schemes, respectively. The following conclusions are drawn:

- There is a strong monotonic correlation between the discretization induced information loss and the model performance. As watershed discretization gets coarser, the discretization-induced information losses increase, and the hydrologic model predictions become worse.
- The watershed outlet is not necessarily the point that has the largest information loss and worst model performance because the largest discretization errors occur in some upstream small drainage areas rather than downstream large drainage areas.
- The discretization error metrics-informed non-uniform discretization across watershed is a better choice than the common and convenient uniform discretization because the former preserves more input data information than the latter while using the same number of computational units.

Chapter 3 proposed a climate ensemble based hydrologic model calibration framework to explicitly account for measured climate data uncertainty in model calibration. 30 synthetic and 20 real case studies were used to assess the framework against two benchmark calibration approaches. One is the optimization-based calibration that does not consider any source of uncertainty, and another is the uncertainty-based calibration that considers parameter uncertainty. The following conclusions are drawn:

- Explicitly accounting for climate data uncertainty in model calibration significantly improves the accuracy and reliability of ensemble flow predictions, especially for the areas with poor quality climate measurements.
- Explicitly accounting for climate data uncertainty in model calibration produces more robust and a smaller number of parameter solutions than using deterministic observed climate.
- Parameter uncertainty is dependent on model forcing data because the climate ensemble generates statistically different posterior parameter distributions from the observed climate.

- It is insufficient to only account for climate data uncertainty in the calibration period or in the validation period (or similarly, the forecasting period). It is important to use consistent climate data uncertainty assumption and treatment throughout the hydrologic modeling process.

Chapter 4 used the Newman et al. (2015) dataset (referred to as N15 in this thesis) to efficiently represent measured precipitation and temperature data uncertainty in the EnKF. The N15 derived climate ensemble was compared with the carefully tuned hyper-parameter generated climate ensembles in an ensemble flow forecasting framework. The carefully tuned hyper-parameters are from two ensemble flow forecasting systems of Thiboult et al. (2016): one uses realistic perturbation magnitudes, another uses unrealistically inflated hyper-parameters. The comparison study was conducted on 20 catchments. The following conclusions are drawn:

- The N15 dataset features in estimating non-zero probability of precipitation when no rainfall is recorded. This feature can be especially important in poorly gauged areas.
- The N15 generated climate ensemble yields improved or similar flow forecasts relative to the two traditional climate ensembles. The forecast improvement of N15 is especially significant for short lead times (i.e., 1-3 days in our case) when the influence of data assimilation dominates.
- It is possible to eliminate the need for precipitation and temperature relevant hyper-parameter tuning from the EnKF by using existing historical climate ensemble products without losing flow forecast performance.

Chapter 5 applied a robust single hydrologic model to ensemble flow forecasting and compared its flow forecast results with the multi-model flow forecast results of Thiboult et al. (2016). The robust single model was calibrated with an explicit consideration of parameter and climate data uncertainty and used the N15 dataset derived climate ensemble in the EnKF. The comparison study was conducted on 20 catchments. The following conclusions are drawn:

- The robust single model generates improved or at least similar deterministic flow forecasts than the multiple models.
- The robust single model generates improved reliable flow ensembles for high flows that are equal to or greater than the 90^{th} percentile of the flow records.
- It is better to use a robust single hydrologic model for the deterministic flow forecast and the high flow ensemble forecast.

Chapter 6 added the explicit consideration of flow data uncertainty to model calibration and formed the climate-flow ensemble based hydrologic model calibration framework. The framework was built on Chapter 3 proposed climate ensemble based model calibration framework. The flow ensemble was generated by the hydraulics-based Bayesian rating curve uncertainty estimation method (BaRatin) (Le Coz

et al., 2014). 10 case studies were used to assess the framework against two calibration approaches. One considers parameter uncertainty, another considers parameter and climate data uncertainty. The following conclusions are drawn:

- Explicitly accounting for flow data uncertainty in model calibration maintains the accuracy of the deterministic flow prediction and slightly improves the reliability of the probabilistic prediction.
- Flow data uncertainty plays a minor role of improving ensemble flow predictions compare with climate data uncertainty.
- Parameter uncertainty is dependent on both forcing and response data because the climate and flow ensembles generate statistically different posterior parameter distributions from the observed climate and flow.

7.2. Limitations

Some notable limitations in this thesis that limit the generalizability of the findings and should be considered in any future work on these topics include:

Firstly, the use of lumped conceptual hydrologic models to test the proposed methods (except Chapter 2). Lumped hydrologic models have no representation of spatial variability in meteorology across the domain, so predictive performance improvements from using gridded climate ensemble products are probably limited. Moreover, conceptual hydrologic models simplify hydrologic processes, for example, routing process, thus predictive performance improvements from using flow ensembles may be limited. Using conceptual hydrologic models also makes merging conceptual state variables with actual state variable measurements (should they become available) through data assimilation difficult.

Secondly, the choice of performance criteria/evaluation metrics for probabilistic flow predictions. This thesis used reliability, spread, MCRPS, and the reliability diagram to assess probabilistic predictions. There are many alternative evaluation metrics that have been extensively used in the literature to assess the prediction quality in a comprehensive way. Some examples include reliability and sharpness of Renard et al. (2010), normalized root mean square error ratio (NRR) and 95% exceedance ratio of DeChant and Moradkhani (2012), and rank histogram of Marzban et al. (2011). Using different evaluation metrics may result in different parameter estimation or flow prediction results or provide a new perspective to analyze the method performance.

Thirdly, the choice of model calibration algorithms. This thesis used the dynamically dimensioned search (DDS) algorithm and the DDS derived algorithms, such as DDS-approximation of uncertainty (DDS-AU) and Pareto archived DDS (PADDS) to estimate model parameters. Other parameter estimation algorithms are also encouraged to be applied within the thesis proposed climate-flow ensemble based calibration framework (if applicable). In addition, parameter estimation and state variable updates can be

performed simultaneously by some data assimilation algorithms that are not considered in the thesis (e.g., Moradkhani et al., 2012, 2005; Smith et al., 2013; Vrugt and Robinson, 2007). Such a concurrent procedure is particularly applicable to low dimensional parameter estimation problems.

7.3. Future Work

Chapter 2 proposed *a priori* discretization error metrics that are easy to recode into the preprocessing of any semi-distributed hydrologic models and the fully distributed models using spatial input data aggregation. There are three valuable future works on it. First, given the great value of the non-uniform discretization in preserving input data information, future work can explore the influence of non-uniform discretization on hydrologic model outputs using a number of hydrologic models and case studies. Another useful direction of applying the proposed discretization error metrics is to use them to account for the input forcing data (e.g., precipitation and temperature) information loss. This will require comparing the spatial and temporal distributions of the forcing data under candidate schemes and those under the reference scheme. Beyond the application in discretization decision-making, future studies can explore the value of using the proposed discretization error metrics, even when they are not calculated *a priori*, to diagnose the causes of model prediction errors in distributed modeling applications.

Chapter 3 generated 100-member precipitation and temperature ensembles based on the N15 dataset and demonstrated it as a good resource to represent uncertain historical climates. Given the promising performance of the N15 dataset in improving hydrologic predictions, it will be valuable to apply the ensemble generation method of the Newman et al. (2015) to other deterministic climate data networks, such as gauged observation or gridded deterministic prediction. Moreover, a limitation of this chapter study is the use of lumped hydrologic model which has no representation of spatial variability in meteorology across the domain. More experiments are needed to test the framework in distributed model calibration problems. Another limitation of this chapter is that the proposed framework has only been applied to 6-parameter estimation problems. More experiments are needed to test the proposed framework in higher dimensional problems to test the feasibility of the proposed model calibration method.

Chapter 4 demonstrated the N15 dataset as a good resource to represent uncertain historical climates in the EnKF application. Future work can investigate the transferability of the current findings to other climate ensemble products and other hydrologic models, in particular, distributed hydrologic models. In addition, this work has considered initial condition uncertainty and forcing and response data uncertainty in the EnKF. In the future, more experiments can be done to explicitly incorporate model parameter and model structure uncertainty in the EnKF. This gives a comprehensive consideration of difference hydrologic modeling errors in the explicit manner. Meanwhile, it will help to explore the individual impact of each modeling error on flow forecasting.

Chapter 5 explicitly incorporated parameter and measured climate data uncertainty in ensemble flow forecasting based on a single hydrologic model. The flow predictions of the single-model system are under-dispersion. Future work needs to explore solutions to improving the prediction spread while maintaining the prediction accuracy. Moreover, the current single-model and multi-model comparison is only built on lumped hydrologic models. Using a more physically-based single-model forecasting system is appealing. Future work can use a physically-based hydrologic model in the single-model forecasting system and compare the physically-based robust single hydrologic model with multiple lumped hydrologic models in ensemble flow forecasting.

Chapter 6 explicitly considered flow data uncertainty in model calibration and concluded that flow data uncertainty plays a minor role of improving flow predictions in 10 case study catchments. This conclusion is drawn on the fact that our case study catchments have good quality operational rating curve management and flow records. Future work needs to conduct the flow data uncertainty analysis and the flow uncertainty based calibration to poorly gauged areas to fairly evaluate the importance of flow data uncertainty on flow prediction and to figure out the applicability of the flow uncertainty based calibration. Moreover, future work needs to apply the proposed framework to distributed physically-based hydrologic models, especially the models that are sensitive to river channel or reservoir routing and have relatively detailed routing process descriptions.

References

Abaza, M., Anctil, F., Fortin, V., Perreault, L., 2017a. Hydrological Evaluation of the Canadian Meteorological Ensemble Reforecast Product. Atmosphere-Ocean 55, 195–211. https://doi.org/10.1080/07055900.2017.1341384

Abaza, M., Anctil, F., Fortin, V., Perreault, L., 2017b. On the incidence of meteorological and hydrological processors: Effect of resolution, sharpness and reliability of hydrological ensemble forecasts. J. Hydrol. 555, 371–384. https://doi.org/10.1016/j.jhydrol.2017.10.038

Abaza, M., Anctil, F., Fortin, V., Turcotte, R., 2014a. Sequential streamflow assimilation for short-term hydrological ensemble forecasting. J. Hydrol. 519, 2692–2706. https://doi.org/10.1016/J.JHYDROL.2014.08.038

Abaza, M., Garneau, C., Anctil, F., 2014b. Comparison of Sequential and Variational Streamflow Assimilation Techniques for Short-Term Hydrological Forecasting. J. Hydrol. Eng. 20, 04014042. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001013

Abbaszadeh, P., Moradkhani, H., Yan, H., 2018. Enhancing hydrologic data assimilation by evolutionary Particle Filter and Markov Chain Monte Carlo. Adv. Water Resour. 111, 192–204. https://doi.org/10.1016/j.advwatres.2017.11.011

Abbott, M.B., Bathurst, J.C., Cunge, J.A., O'Connell, P.E., Rasmussen, J., 1986. An introduction to the European Hydrological System—Systeme Hydrologique Europeen,"SHE", 1: History and philosophy of a physically-based, distributed modelling system. J. Hydrol. 87, 45–59.

Agriculture and Agri-Food Canada, 2013. Detailed Soil Survey (DSS) Compilations. URL http://sis.agr.gc.ca/cansis/nsdb/dss/v3/index.html

Ailliot, P., Allard, D., Monbet, V., Naveau, P., 2015. Stochastic weather generators: an overview of weather type models. J. la Société Française Stat. 156, 101–113.

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 Resour. Res. 43, W01403. https://doi.org/10.1029/2005WR004745

Al-Yahyai, S., Charabi, Y., Gastli, A., 2010. Review of the use of numerical weather prediction (NWP) models for wind energy assessment. Renew. Sustain. Energy Rev. https://doi.org/10.1016/j.rser.2010.07.001

Altman, D.G., Bland, J.M., 1999. Statistics notes: variables and parameters. BMJ 318, 1667. https://doi.org/10.1136/bmj.318.7199.1667

Andreadis, K.M., Schumann, G.J.P., Pavelsky, T., 2013. A simple global river bankfull width and depth database. Water Resour. Res. 49, 7164–7168. https://doi.org/10.1002/wrcr.20440

Anselin, L., 2010. Thirty years of spatial econometrics. Pap. Reg. Sci. 89, 3–25.

Arnold, J.G., Allen, P.M., Volk, M., Williams, J.R., Bosch, D.D., 2010. Assessment of different representations of spatial variability on SWAT model performance. Trans. ASABE 53, 1433–1443. https://doi.org/10.13031/2013.34913

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment part I: model development. J. Am. Water Resour. Assoc. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x

Asadzadeh, M., Tolson, B., 2013. Pareto archived dynamically dimensioned search with hypervolume-

based selection for multi-objective optimization. Eng. Optim. 45, 1489–1509. https://doi.org/10.1080/0305215X.2012.748046

Balin, D., Lee, H., Rode, M., 2010. Is point uncertain rainfall likely to have a great impact on distributed complex hydrological modeling? Water Resour. Res. 46, W11520. https://doi.org/10.1029/2009WR007848

Bandaragoda, C., Tarboton, D.G., Woods, R., 2004. Application of TOPNET in the distributed model intercomparison project. J. Hydrol. 298, 178–201. https://doi.org/10.1016/j.jhydrol.2004.03.038

Bergeron, O., 2016. Guide d'utilisation 2016 - Grilles climatiques quotidiennes du Programme de surveillance du climat du Québec (Version 1.2). Québec.

Bergstrom, S., 1975. The development of a snow routine for the HBV-2. Nord. Hydrol. 6, 73–92.

Bergström, S., 1992. The HBV model: Its structure and applications. Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden.

Bergström, S., 1976. Development and application of a conceptual runoff model for Scandinavian catchments. Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden.

Beven, K., 2012. Rainfall-Runoff Modelling: The Primer: Second Edition, Rainfall-Runoff Modelling: The Primer: Second Edition. Wiley. https://doi.org/10.1002/9781119951001

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320, 18–36. https://doi.org/10.1016/J.JHYDROL.2005.07.007

Beven, K., 1989. Changing ideas in hydrology - The case of physically-based models. J. Hydrol. 105, 157–172. https://doi.org/10.1016/0022-1694(89)90101-7

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Process. 6, 279–298. https://doi.org/10.1002/hyp.3360060305

Beven, K., Calver, A., Morris, E.. M., 1987. Institute of Hydrology Distributed Model.

Beven, K., Freer, J., 2001. Equifinality , data assimilation , and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249, 11–29. https://doi.org/https://doi.org/10.1016/S0022-1694(01)00421-8

Beven, K., Young, P., 2013. A guide to good practice in modeling semantics for authors and referees. Water Resour. Res. 49, 5092–5098. https://doi.org/10.1002/wrcr.20393

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. J. 24, 43–69. https://doi.org/10.1080/02626667909491834

Blazkova, S., Beven, K., 2009. A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. Water Resour. Res. 45, W00B16. https://doi.org/10.1029/2007WR006726

Booij, M.J., 2005. Impact of climate change on river flooding assessed with different spatial model resolutions. J. Hydrol. 303, 176–198. https://doi.org/10.1016/j.jhydrol.2004.07.013

Booij, M.J., 2003. Determination and integration of appropriate spatial scales for river basin modelling. Hydrol. Process. 17, 2581–2598. https://doi.org/10.1002/hyp.1268

Boucher, M.-A., Perreault, L., Anctil, F., Favre, A.-C., 2015. Exploratory analysis of statistical post-processing methods for hydrological ensemble forecasts. Hydrol. Process. 29, 1141–1155. https://doi.org/10.1002/hyp.10234

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time series analysis : forecasting and control. Prentice Hall.

Boyle, D.P., Gupta, H. V., Sorooshian, S., 2011. Multicriteria calibration of hydrologic models, in: Duan, Q., Gupta, H. V., Sorooshian, S., Rousseau, A.N., Turcotte, R. (Eds.), Calibration of Watershed Models, Volume 6. American Geophysical Union, pp. 185–196. https://doi.org/10.1029/ws006p0185

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Cameron, D.S., Beven, K.J., Tawn, J., Blazkova, S., Naden, P., 1999. Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty). J. Hydrol. 219, 169–187. https://doi.org/10.1016/S0022-1694(99)00057-8

Carpenter, T.M., Georgakakos, K.P., 2004. Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model. J. Hydrol. 298, 202–221. https://doi.org/10.1016/J.JHYDROL.2004.03.036

Carpenter, T.M., Georgakakos, K.P., Sperfslagea, J.A., 2001. On the parametric and NEXRAD-radar sensitivities of a distributed hydrologic model suitable for operational use. J. Hydrol. 253, 169–193. https://doi.org/10.1016/S0022-1694(01)00476-0

Chandler, R.E., Wheater, H.S., 2002. Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. Water Resour. Res. 38, 1192. https://doi.org/10.1029/2001wr000906

Chow, V.T., 1959. Open-channel hydraulics. McGraw-Hill, New York.

Ciarapica, L., Todini, E., 2002. TOPKAPI: a model for the representation of the rainfall-runoff process at different scales. Hydrol. Process. 16, 207–229. https://doi.org/10.1002/hyp.342

Clark, M.P., Kavetski, D., 2010. Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes. Water Resour. Res. 46, W10510. https://doi.org/10.1029/2009WR008894

Clark, M.P., Nijssen, B., Lundquist, J.D., Kavetski, D., Rupp, D.E., Woods, R.A., Freer, J.E., Gutmann, E.D., Wood, A.W., Brekke, L.D., Arnold, J.R., Gochis, D.J., Rasmussen, R.M., 2015. A unified approach for process-based hydrologic modeling: 1. Modeling concept. Water Resour. Res. 51, 2498–2514. https://doi.org/10.1002/2015WR017198

Clark, M.P., Rupp, D.E., Woods, R.A., Zheng, X., Ibbitt, R.P., Slater, A.G., Schmidt, J., Uddstrom, M.J., 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Adv. Water Resour. 31, 1309–1324. https://doi.org/10.1016/j.advwatres.2008.06.005

Clark, M.P., Slater, A.G., 2006. Probabilistic quantitative precipitation estimation in complex terrain. J. Hydrometeorol. 7, 3–22. https://doi.org/10.1175/JHM474.1

Clark, M.P., Slater, A.G., Barrett, A.P., Hay, L.E., McCabe, G.J., Rajagopalan, B., Leavesley, G.H., 2006. Assimilation of snow covered area information into hydrologic and land-surface models. Adv. Water Resour. 29, 1209–1221. https://doi.org/10.1016/j.advwatres.2005.10.001

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. J. Hydrol. 375, 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Coxon, G., Freer, J., Westerberg, I.K., Wagener, T., Woods, R., Smith, P.J., 2015. A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. Water Resour. Res. 51, 5531–5546. https://doi.org/10.1002/2014WR016532

Craig, J.R., Raven Development Team, 2018. Raven user's and developer's manual (Version 2.8).

Crawford, N., Linsley, R., 1966. Digital Simulation in Hydrology'Stanford Watershed Model 4.

Cunderlik, J.M., Burn, D.H., 2006. Switching the pooling similarity distances: Mahalanobis for Euclidean. Water Resour. Res. 42, W03409. https://doi.org/10.1029/2005WR004245

Day, G.N., 1985. Extended Streamflow Forecasting Using NWSRFS. J. Water Resour. Plan. Manag. 111, 157–170. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)

De Lavenne, A., Boudhraâ, H., Cudennec, C., 2015. Streamflow prediction in ungauged basins through geomorphology-based hydrograph transposition. Hydrol. Res. 46, 291–302. https://doi.org/10.2166/nh.2013.099

DeChant, C.M., Moradkhani, H., 2012. Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting. Water Resour. Res. 48, W04518. https://doi.org/10.1029/2011WR011011

Dehotin, J., Braud, I., 2008. Which spatial discretization for distributed hydrological models? Proposition of a methodology and illustration for medium to large-scale catchments. Hydrol. Earth Syst. Sci. 12, 769–796. https://doi.org/https://doi.org/10.5194/hess-12-769-2008

Del Giudice, D., Albert, C., Rieckermann, J., Reichert, P., 2016. Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation. Water Resour. Res. 52, 3162–3186. https://doi.org/10.1002/2015WR017871

Del Giudice, D., Löwe, R., Madsen, H., Mikkelsen, P.S., Rieckermann, J., 2015. Comparison of two stochastic techniques for reliable urban runoff prediction by modeling systematic errors. Water Resour. Res. 51, 5004–5022. https://doi.org/10.1002/2014WR016678

Demirel, M.C., Booij, M.J., Hoekstra, A.Y., 2013. Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models. Water Resour. Res. 49, 4035–4053. https://doi.org/10.1002/wrcr.20294

Dietrich, J., Schumann, A.H., Redetzky, M., Walther, J., Denhard, M., Wang, Y., Utzner, B.P., Uttner, U., 2009. Assessing uncertainties in flood forecasts for decision making: prototype of an operational flood management system integrating ensemble predictions. Nat. Hazards Earth Syst. Sci 9, 1529–1540. https://doi.org/10.5194/nhess-9-1529-2009

Dingman, S.L., 2015. Physical hydrology, 3rd ed, Progress in Physical Geography. Waveland Press, Inc., Long Grove, Illinois.

Djokic, D., 2008. Comprehensive terrain preprocessing using arc hydro tools. US ESRI.

Duan, Q., Gupta, H. V., Sorooshian, S., Rousseau, A.N., Turcotte, R., 2003. Calibration of Watershed Models, American Geophysical Union. https://doi.org/10.1029/WS006

Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res. 28, 1015–1031. https://doi.org/10.1029/91WR02985

Dunne, S., Entekhabi, D., 2006. Land surface state and flux estimation using the ensemble Kalman smoother during the Southern Great Plains 1997 field experiment. Water Resour. Res. 42, W01407. https://doi.org/10.1029/2005WR004334

Dunne, T., Leopold, L., 1978. Water in environmental planning, 1st edition. ed. W. H. Freeman.

Eicker, A., Schumacher, M., Kusche, J., Döll, P., Schmied, H.M., 2014. Calibration/Data Assimilation Approach for Integrating GRACE Data into the WaterGAP Global Hydrology Model (WGHM) Using an Ensemble Kalman Filter: First Results. Surv. Geophys. 35, 1285–1309.

https://doi.org/10.1007/s10712-014-9309-8

ESRI, 2014. ArcGIS 10.2.2 for Desktop.

Evensen, G., 2007. Data Assimilation, The Ensemble Kalman Filter, Springer. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-38301-7

Evensen, G., 2003. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367. https://doi.org/10.1007/s10236-003-0036-9

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143. https://doi.org/10.1029/94JC00572

Fang, J., Tacher, L., 2003. An efficient and accurate algorithm for generating spatially-correlated random fields. Commun. Numer. Methods Eng. 19, 801–808. https://doi.org/10.1002/cnm.621

Fenicia, F., Kavetski, D., Savenije, H.H.G., Pfister, L., 2016. From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions. Water Resour. Res. 52, 954–989. https://doi.org/10.1002/2015WR017398

Fernández-González, S., Martín, M.L., Merino, A., Sánchez, J.L., Valero, F., 2017. Uncertainty quantification and predictability of wind speed over the Iberian Peninsula. J. Geophys. Res. Atmos. 122, 3877–3890. https://doi.org/10.1002/2017JD026533

Fleming, M.J., Doan, J.H., 2009. Geospatial Hydrologic Modeling Extension Version 4.2.

Flügel, W. -A, 1995. Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Bröl, Germany. Hydrol. Process. 9, 423–436. https://doi.org/10.1002/hyp.3360090313

Fortin, L.-G., Ludwig, R., Braun, M., Cyr, J.-F., Biner, S., Mauser, W., Turcotte, R., May, I., Caya, D., Chartier, I., Vescovi, L., Chaumont, D., 2010. The role of hydrological model complexity and uncertainty in climate change impact assessment. Adv. Geosci. 21, 63–71. https://doi.org/10.5194/adgeo-21-63-2009

Fortin, V., Abaza, M., Anctil, F., Turcotte, R., Fortin, V., Abaza, M., Anctil, F., Turcotte, R., 2014. Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? J. Hydrometeorol. 15, 1708–1713. https://doi.org/10.1175/JHM-D-14-0008.1

Freeze, R.A., Harlan, R.L., 1969. Blueprint for a physically-based, digitally-simulated hydrologic response model. J. Hydrol. 9, 237–258. https://doi.org/10.1016/0022-1694(69)90020-1

Fuentes-Andino, D., Beven, K., Kauffeldt, A., Xu, C.-Y., Halldin, S., Di Baldassarre, G., 2017. Event and model dependent rainfall adjustments to improve discharge predictions. Hydrol. Sci. J. 62, 232–245. https://doi.org/10.1080/02626667.2016.1183775

Gaborit, É., Anctil, F., Pelletier, G., Fortin, V., 2013. On the reliability of spatially disaggregated global ensemble rainfall forecasts. Hydrol. Process. 27, 45–56. https://doi.org/10.1002/hyp.9509

Gelman, A., 1996. Inference and Monitoring Convergence in Markov Chain Monte Carlo in Practice. WR Gilks, S. Richardson, and DJ Spiegelhalter, eds. London: Chapman & Hall.

Geman, S., Geman, D., 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6, 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Georgakakos, K.P., Seo, D.J., Gupta, H., Schaake, J., Butts, M.B., 2004. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. J. Hydrol. 298, 222–241.

https://doi.org/10.1016/j.jhydrol.2004.03.037

Gneiting, T., Raftery, A.E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. J. Am. Stat. Assoc. 102, 359–378. https://doi.org/10.1198/016214506000001437

Graham, L.P., Hagemann, S., Jaun, S., Beniston, M., 2007. On interpreting hydrological change from regional climate models. Clim. Change 81, 97–122. https://doi.org/10.1007/s10584-006-9217-0

Grayson, R., Blöschl, G., Moore, I.D., Singh, V.P., 1995. Distributed parameter hydrologic modelling using vector elevation data: THALES and TAPES-C. Comput. Model. watershed Hydrol. 669–696.

Green, B.W., Sun, S., Bleck, R., Benjamin, S.G., Grell, G.A., Green, B.W., Sun, S., Bleck, R., Benjamin, S.G., Grell, G.A., 2017. Evaluation of MJO Predictive Skill in Multiphysics and Multimodel Global Ensembles. Mon. Weather Rev. 145, 2555–2574. https://doi.org/10.1175/MWR-D-16-0419.1

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1999. Status of Automatic Calibration for Hydrologic Models: Comparison with Multilevel Expert Calibration. J. Hydrol. Eng. 4, 135–143. https://doi.org/10.1061/(ASCE)1084-0699(1999)4:2(135)

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. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Habib, E., Krajewski, W.F., Kruger, A., 2002. Sampling Errors of Tipping-Bucket Rain Gauge Measurements. J. Hydrol. Eng. 6, 159–166. https://doi.org/10.1061/(asce)1084-0699(2001)6:2(159)

Haghnegahdar, A., Tolson, B.A., Craig, J.R., Paya, K.T., 2015. Assessing the performance of a semi-distributed hydrological model under various watershed discretization schemes. Hydrol. Process. 29, 4018–4031. https://doi.org/10.1002/hyp.10550

Hamill, T.M., Snyder, C., 2000. A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme. Mon. Weather Rev. 128, 2905–2919. https://doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2

Hartmann, H.C., Pagano, T.C., Sorooshian, S., Bales, R., 2016. Confidence builders: evaluating seasonal climate forecasts from user perspectives. Kementeri. Kesehat. RI 1–168. https://doi.org/10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2

Haverkamp, S., Srinivasan, R., Frede, H.G., Santhi, C., 2002. Subwatershed spatial analysis tool: Discretization of a distributed hydrologic model by statistical criteria. J. Am. Water Resour. Assoc. 38, 1723–1733. https://doi.org/https://doi.org/10.1111/j.1752-1688.2002.tb04377.x

Hay, L.E., Wilby, R.L., Leavesley, G.H., 2000. A comparison of delta change and downscaled GCM scenarios for three mountainous basins in the United States. J. Am. Water Resour. Assoc. 36, 387–397. https://doi.org/10.1111/j.1752-1688.2000.tb04276.x

Heemink, A.W., Verlaan, M., Segers, A.J., 2001. Variance reduced ensemble Kalman filtering. Mon. Weather Rev. 129, 1718–1728. https://doi.org/10.1175/1520-0493(2001)129<1718:VREKF>2.0.CO;2

Helsel, D.R., Hirsch, R.M., 2002. Hydrologic Analysis and Interpretation, in: Helsen, D., Hirsch, R. (Eds.), Techniques of Water-Resources Investigations of the United States Geological Survey. United States Geological Survey, Reston, VA, USA, pp. 117–124.

Henn, B., Clark, M.P., Kavetski, D., Lundquist, J.D., 2015. Estimating mountain basin-mean precipitation from streamflow using Bayesian inference. Water Resour. Res. 51, 8012–8033. https://doi.org/10.1002/2014WR016736

Henn, B., Clark, M.P., Kavetski, D., Newman, A.J., Hughes, M., McGurk, B., Lundquist, J.D., 2018. Spatiotemporal patterns of precipitation inferred from streamflow observations across the Sierra Nevada mountain range. J. Hydrol. 556, 993–1012. https://doi.org/10.1016/j.jhydrol.2016.08.009

Hersbach, H., Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather Forecast. 15, 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Honti, M., Stamm, C., Reichert, P., 2013. Integrated uncertainty assessment of discharge predictions with a statistical error model. Water Resour. Res. 49, 4866–4884. https://doi.org/10.1002/wrcr.20374

Hopson, T.M., Webster, P.J., Hopson, T.M., Webster, P.J., 2010. A 1–10-Day Ensemble Forecasting Scheme for the Major River Basins of Bangladesh: Forecasting Severe Floods of 2003–07. J. Hydrometeorol. 11, 618–641. https://doi.org/10.1175/2009JHM1006.1

Horton, R.E., 1941. An Approach Toward a Physical Interpretation of Infiltration-Capacity1. Soil Sci. Soc. Am. J. 5, 399. https://doi.org/10.2136/sssaj1941.036159950005000C0075x

Houtekamer, P.L., Mitchell, H.L., 2001. A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation. Mon. Weather Rev. 129, 123–137. https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2

Houtekamer, P.L., Mitchell, H.L., 1998. Data Assimilation Using an Ensemble Kalman Filter Technique. Mon. Weather Rev. 126, 796–811. https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2

Hsu, K.L., Moradkhani, H., Sorooshian, S., 2009. A sequential Bayesian approach for hydrologic model selection and prediction. Water Resour. Res. 45, W00B12. https://doi.org/10.1029/2008WR006824

Huang, C., Newman, A.J., Clark, M.P., Wood, A.W., Zheng, X., 2017. Evaluation of snow data assimilation using the ensemble Kalman filter for seasonal streamflow prediction in the western United States. Hydrol. Earth Syst. Sci. 21, 635–650. https://doi.org/10.5194/hess-21-635-2017

Huard, D., Mailhot, A., 2008. Calibration of hydrological model GR2M using Bayesian uncertainty analysis. Water Resour. Res. 44, W02424. https://doi.org/10.1029/2007WR005949

Huard, D., Mailhot, A., 2006. A Bayesian perspective on input uncertainty in model calibration: Application to hydrological model “abc.” Water Resour. Res. 42, 1–14. https://doi.org/10.1029/2005WR004661

Humphrey, M.D., Istok, J.D., Lee, J.Y., Hevesi, J.A., Flint, A.L., 1997. A new method for automated dynamic calibration of tipping-bucket rain gauges. J. Atmos. Ocean. Technol. 14, 1513–1519. https://doi.org/10.1175/1520-0426(1997)014<1513:ANMFAD>2.0.CO;2

ISO 1100, 2010. Hydrometry — Measurement of liquid flow in open channels — Part 2: Determination of the stage-discharge relationship.

Ivanov, V.Y., Vivoni, E.R., Bras, R.L., Entekhabi, D., 2004. Catchment hydrologic response with a fully distributed triangulated irregular network model. Water Resour. Res. 40, W11102. https://doi.org/10.1029/2004WR003218

Journel, A.G., Deutsch, C. V, 1993. Entropy and spatial disorder. Math. Geol. 25, 329–355. https://doi.org/https://doi.org/10.1007/BF00901422

Junk, C., Monache, L.D., Alessandrini, S., Cervone, G., Von Bremen, L., 2015. Predictor-weighting strategies for probabilistic wind power forecasting with an analog ensemble. Meteorol. Zeitschrift 24, 361–379. https://doi.org/10.1127/metz/2015/0659

Kalman, R.E., 1960. A New Approach to Linear Filtering and Prediction Problems. J. Basic Eng. 82, 35. https://doi.org/10.1115/1.3662552

Kavetski, D., Fenicia, F., Reichert, P., Albert, C., 2018. Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Theory and Comparison to Existing Applications. Water Resour. Res. 54, 4059–4083. https://doi.org/10.1002/2017WR020528

Kavetski, D., Franks, S.W., Kuczera, G., 2002. Confronting input uncertainty in environmental modelling, in: Duan, Q., Gupta, H. V., Sorooshian, S., Rousseau, A.N., Turcotte, R. (Eds.), Calibration of Watershed Models. American Geophysical Union, Washington, D.C., pp. 49–68. https://doi.org/10.1029/WS006p0049

Kavetski, D., Kuczera, G., Franks, S.W., 2006a. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. Water Resour. Res. 42, W03408. https://doi.org/10.1029/2005WR004376

Kavetski, D., Kuczera, G., Franks, S.W., 2006b. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resour. Res. 42, W03407. https://doi.org/10.1029/2005WR004368

Kerkhoven, E., Gan, T.Y., 2006. A modified ISBA surface scheme for modeling the hydrology of Athabasca River Basin with GCM-scale data. Adv. Water Resour. 29, 808–826. https://doi.org/10.1016/j.advwatres.2005.07.016

Khaki, M., Schumacher, M., Forootan, E., Kuhn, M., Awange, J.L., van Dijk, A.I.J.M., 2017. Accounting for spatial correlation errors in the assimilation of GRACE into hydrological models through localization. Adv. Water Resour. 108, 99–112. https://doi.org/10.1016/j.advwatres.2017.07.024

Kiureghian, A. Der, Ditlevsen, O., 2009. Aleatory or epistemic? Does it matter? Struct. Saf. 31, 105–112. https://doi.org/10.1016/j.strusafe.2008.06.020

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J. 31, 13–24. https://doi.org/10.1080/02626668609491024

Kouwen, N., 1988. WATFLOOD: a micro-computer based flood forecasting system based on real-time weather radar. Can. Water Resour. J. 13, 62–77. https://doi.org/10.4296/cwrj1301062

Kouwen, N., Soulis, E.D., Pietroniro, A., Donald, J., Harrington, R.A., 1993. Grouped response units for distributed hydrologic modeling. J. Water Resour. Plan. Manag. 119, 289–305. https://doi.org/10.1061/(ASCE)0733-9496(1993)119:3(289)

Krueger, T., Quinton, J.N., Freer, J., Macleod, C.J.A., Bilotta, G.S., Brazier, R.E., Butler, P., Haygarth, P.M., 2009. Uncertainties in Data and Models to Describe Event Dynamics of Agricultural Sediment and Phosphorus Transfer. J. Environ. Qual. 38, 1137. https://doi.org/10.2134/jeq2008.0179

Krzysztofowicz, R., 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resour. Res. 35, 2739–2750. https://doi.org/10.1029/1999WR900099

Kuczera, G., 1983. Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. Water Resour. Res. 19, 1151–1162. https://doi.org/10.1029/WR019i005p01151

Kuczera, G., Kavetski, D., Franks, S., Thyer, M., 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. J. Hydrol. 331, 161–177. https://doi.org/10.1016/j.jhydrol.2006.05.010

Kuczera, G., Parent, E., 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm. J. Hydrol. 211, 69–85. https://doi.org/10.1016/S0022-1694(98)00198-X

Kumar, S. V., Dong, J., Peters-Lidard, C.D., Mocko, D., Gómez, B., 2016. Role of forcing uncertainty

and model error background characterization in snow data assimilation. Hydrol. Earth Syst. Sci. Discuss. 1–24. https://doi.org/10.5194/hess-2016-581

Kyriakidis, P.C., Miller, N.L., Kim, J., 2004. A spatial time series framework for simulating daily precipitation at regional scales. J. Hydrol. 297, 236–255. https://doi.org/10.1016/j.jhydrol.2004.04.022

Le Coz, J., Renard, B., Bonnifait, L., Branger, F., Le Boursicaud, R., 2014. Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach. J. Hydrol. 509, 573–587. https://doi.org/10.1016/j.jhydrol.2013.11.016

Leisenring, M., Moradkhani, H., 2011. Snow water equivalent prediction using Bayesian data assimilation methods. Stoch. Environ. Res. Risk Assess. 25, 253–270. https://doi.org/10.1007/s00477-010-0445-5

Lenderink, G., Buishand, A., van Deursen, W., 2007. Estimates of future discharges of the river Rhine using two scenario methodologies: direct versus delta approach. Hydrol. Earth Syst. Sci. 11, 1145–1159. https://doi.org/10.5194/hess-11-1145-2007

Leutbecher, M., Palmer, T.N., 2008. Ensemble forecasting. J. Comput. Phys. 227, 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014

Li, Y., Ryu, D., Western, A.W., Wang, Q.J., Robertson, D.E., Crow, W.T., 2014. An integrated error parameter estimation and lag-aware data assimilation scheme for real-time flood forecasting. J. Hydrol. 519, 2722–2736. https://doi.org/10.1016/J.JHYDROL.2014.08.009

Liu, H., Thiboult, A., Tolson, B.A., Anctil, F., Mai, J., 2019. Efficient treatment of climate data uncertainty in ensemble Kalman filter (EnKF) based on an existing historical climate ensemble dataset. J. Hydrol. 568, 985–996. https://doi.org/10.1016/J.JHYDROL.2018.11.047

Liu, H., Tolson, B.A., Craig, J.R., Shafii, M., 2016. A priori discretization error metrics for distributed hydrologic modeling applications. J. Hydrol. 543, 873–891. https://doi.org/10.1016/j.jhydrol.2016.11.008

Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resour. Res. 43, W07401. https://doi.org/10.1029/2006WR005756

Madadgar, S., Moradkhani, H., 2014. Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. Water Resour. Res. 50, 9586–9603. https://doi.org/10.1002/2014WR015965

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv. Water Resour. 26, 205–216. https://doi.org/10.1016/S0309-1708(02)00092-1

Maidment, D.R., 2002. Arc Hydro: GIS for water resources. ESRI, Inc.

Mansanarez, V., Le Coz, J., Renard, B., Lang, M., Pierrefeu, G., Vauchel, P., 2016. Bayesian analysis of stage-fall-discharge rating curves and their uncertainties. https://doi.org/10.1002/2016WR018916

Marzban, C., Wang, R., Kong, F., Leyton, S., Marzban, C., Wang, R., Kong, F., Leyton, S., 2011. On the Effect of Correlations on Rank Histograms: Reliability of Temperature and Wind Speed Forecasts from Finescale Ensemble Reforecasts. Mon. Weather Rev. 139, 295–310. https://doi.org/10.1175/2010MWR3129.1

Matheson, J.E., Winkler, R.L., 1976. Scoring Rules for Continuous Probability Distributions. Manage. Sci. 22, 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087

McDonald, J.M., Harbaugh, A.W., 1988. A modular three-dimensional finite-difference flow model., in: Techniques of Water Resources Investigations of the U.S. Geological Survey, Book 6. p. 586. https://doi.org/10.1016/0022-1694(86)90106-X

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., Kuczera, G., 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resour. Res. 53, 2199–2239. https://doi.org/10.1002/2016WR019168

McIntyre, N., Wheater, H., Lees, M., 2018. Estimation and propagation of parametric uncertainty in environmental models. J. Hydroinformatics 4, 177–198. https://doi.org/10.2166/hydro.2002.0018

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. Hydrol. Process. 24, 1270–1284. https://doi.org/10.1002/hyp.7587

McMillan, H.K., Hreinsson, E.Ö., Clark, M.P., Singh, S.K., Zammit, C., Uddstrom, M.J., 2013. Operational hydrological data assimilation with the recursive ensemble Kalman filter. Hydrol. Earth Syst. Sci. 17, 21–38. https://doi.org/10.5194/hess-17-21-2013

Mcmillan, H.K., Westerberg, I.K., 2015. Rating curve estimation under epistemic uncertainty. Hydrol. Process. 29, 1873–1882. https://doi.org/10.1002/hyp.10419

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092. https://doi.org/http://dx.doi.org/10.1063/1.1699114

Montanari, A., 2005. Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. Water Resour. Res. 41, W08406. https://doi.org/10.1029/2004WR003826

Montanari, A., Di Baldassarre, G., 2013. Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty. Adv. Water Resour. 51, 498–504. https://doi.org/10.1016/j.advwatres.2012.09.007

Montanari, A., Koutsoyiannis, D., 2012. A blueprint for process-based modeling of uncertain hydrological systems. Water Resour. Res. 48, 1–15. https://doi.org/10.1029/2011WR011412

Moradkhani, H., Dechant, C.M., Sorooshian, S., 2012. Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method. Water Resour. Res. 48, W12520. https://doi.org/10.1029/2012WR012144

Moradkhani, H., Sorooshian, S., Gupta, H. V., Houser, P.R., 2005. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour. 28, 135–147. https://doi.org/10.1016/j.advwatres.2004.09.002

Moyeed, R.A., Clarke, R.T., 2005. The use of Bayesian methods for fitting rating curves, with case studies. Adv. Water Resour. 28, 807–818. https://doi.org/10.1016/j.advwatres.2005.02.005

Najafi, M.R., Moradkhani, H., 2015. Ensemble Combination of Seasonal Streamflow Forecasts. J. Hydrol. Eng. 21, 04015043. https://doi.org/10.1061/(asce)he.1943-5584.0001250

Nash, J.E., Sutcliffe, J. V., 1970. River flow forecasting through conceptual models part I - A discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Natural Resources Canada, 2014. Land Cover, circa 2000-Vector. URL http://www.geobase.ca/

Newman, A.J., Clark, M.P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., Mizukami, N., Brekke, L., Arnold, J.R., 2015. Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous

United States. J. Hydrometeorol. 16, 2481–2500. https://doi.org/10.1175/jhm-d-15-0026.1

Nijssen, B., 2004. Effect of precipitation sampling error on simulated hydrological fluxes and states: Anticipating the Global Precipitation Measurement satellites. J. Geophys. Res. 109, 1–15. https://doi.org/10.1029/2003JD003497

Noh, S.J., Rakovec, O., Weerts, A.H., Tachikawa, Y., 2014. On noise specification in data assimilation schemes for improved flood forecasting using distributed hydrological models. J. Hydrol. 519, 2707–2721. https://doi.org/10.1016/J.JHYDROL.2014.07.049

Olson, S.A., Norris, J.M., 2007. U.S. Geological Survey Streamgaging - from the National Streamflow Information Program. URL http://water.usgs.gov/edu/measureflow.html.

Osman Akan, A., 2006. Open channel hydraulics. Elsevier/BH.

Oudin, L., Michel, C., Anctil, F., 2005. Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 1 - Can rainfall-runoff models effectively handle detailed potential evapotranspiration inputs? J. Hydrol. 303, 275–289. https://doi.org/10.1016/j.jhydrol.2004.08.025

Pappenberger, F., Beven, K.J., 2004. Functional classification and evaluation of hydrographs based on multicomponent mapping (Mx). Int. J. River Basin Manag. 2, 89–100. https://doi.org/10.1080/15715124.2004.9635224

Pappenberger, F., Matgen, P., Beven, K.J., Henry, J.-B., Pfister, L., Fraipont, P., 2006. Influence of uncertain boundary conditions and model structure on flood inundation predictions. Adv. Water Resour. 29, 1430–1449. https://doi.org/10.1016/J.ADVWATRES.2005.11.012

Pasetto, D., Camporese, M., Putti, M., 2012. Ensemble Kalman filter versus particle filter for a physically-based coupled surface–subsurface model. Adv. Water Resour. 47, 1–13. https://doi.org/10.1016/J.ADVWATRES.2012.06.009

Pauwels, V.R.N., De Lannoy, G.J.M., Pauwels, V.R.N., Lannoy, G.J.M. De, 2006. Improvement of Modeled Soil Wetness Conditions and Turbulent Fluxes through the Assimilation of Observed Discharge. J. Hydrometeorol. 7, 458–477. https://doi.org/10.1175/JHM490.1

Pechlivanidis, I., Jackson, B., 2011. Catchment scale hydrological modelling: a review of model types, calibration approaches and uncertainty analysis methods in eth context of recent developments in technology and applications. Glob. NEST 13, 193–214. https://doi.org/10.1002/hyp

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. 279, 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Petersen-Øverleir, A., 2006. Modelling stage—discharge relationships affected by hysteresis using the Jones formula and nonlinear regression. Hydrol. Sci. J. https://doi.org/10.1623/hysj.51.3.365

Petersen-Øverleir, A., Reitan, T., 2009. Bayesian analysis of stage-fall-discharge models for gauging stations affected by variable backwater. Hydrol. Process. 23, 3057–3074. https://doi.org/10.1002/hyp.7417

Petrucci, G., Bonhomme, C., 2014. The dilemma of spatial representation for urban hydrology semi-distributed modelling: Trade-offs among complexity, calibration and geographical data. J. Hydrol. 517, 997–1007. https://doi.org/10.1016/j.jhydrol.2014.06.019

Piani, C., Haerter, J.O., Coppola, E., 2010. Statistical bias correction for daily precipitation in regional climate models over Europe. Theor. Appl. Climatol. 99, 187–192. https://doi.org/10.1007/s00704-009-0134-9

Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., Verseghy, D., Soulis, E.D.,

Caldwell, R., Evora, N., Pellerin, P., 2007. Using the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale. Hydrol. Earth Syst. Sci. Discuss. 11, 1279–1294. https://doi.org/https://doi.org/10.5194/hess-11-1279-2007

Quick, M.C., 1995. The UBC Watershed Model, in: Singh, V. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Colorado, pp. 233–280.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Mon. Weather Rev. 133, 1155–1174. https://doi.org/10.1175/MWR2906.1

Rana, A., Foster, K., Bosshard, T., Olsson, J., Bengtsson, L., 2014. Impact of climate change on rainfall over Mumbai using Distribution-based Scaling of Global Climate Model projections. J. Hydrol. Reg. Stud. 1, 107–128. https://doi.org/10.1016/J.EJRH.2014.06.005

Rasmussen, J., Madsen, H., Jensen, K.H., Refsgaard, J.C., 2015. Data assimilation in integrated hydrological modeling using ensemble Kalman filtering: evaluating the effect of ensemble size and localization on filter performance. Hydrol. Earth Syst. Sci. 19, 2999–3013. https://doi.org/10.5194/hess-19-2999-2015

Refsgaard, J., 1997. Parameterisation, calibration and validation of distributed hydrological models. J. Hydrol. 198, 69–97. https://doi.org/10.1016/S0022-1694(96)03329-X

Refsgaard, J., 1990. Terminology, modelling protocol and classification of hydrological model codes, in: Abbott, M.B., Refsgaard, J.C. (Eds.), Distributed Hydrological Modelling. Water Science and Technology Library, Vol 22. Springer, Dordrecht, pp. 17–39.

Refsgaard, J., Abbott, M., 1990. The role of distributed hydrological modelling in water resources management, in: Abbott, M.B., Refsgaard, J.C. (Eds.), Distributed Hydrological Modelling. Water Science and Technology Library, Vol 22. Springer, Dordrecht, pp. 1–16.

Reggiani, P., Sivapalan, M., Hassanizadeh, S.M., 1998. A unifying framework for watershed thermodynamics: balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics. Adv. Water Resour. 22, 367–398. https://doi.org/https://doi.org/10.1016/S0309-1708(98)00012-8

Reichle, R.H., 2008. Data assimilation methods in the Earth sciences. Adv. Water Resour. 31, 1411–1418. https://doi.org/10.1016/j.advwatres.2008.01.001

Reichle, R.H., Koster, R.D., 2003. Assessing the Impact of Horizontal Error Correlations in Background Fields on Soil Moisture Estimation. J. Hydrometeorol. 4, 1229–1242. https://doi.org/10.1175/1525-7541(2003)004<1229:ATIOHE>2.0.CO;2

Reichle, R.H., McLaughlin, D.B., Entekhabi, D., 2002a. Hydrologic data assimilation with the ensemble Kalman filter. Mon. Weather Rev. 130, 103–114. https://doi.org/10.1175/1520-0493(2002)130<0103:HDAWTE>2.0.CO;2

Reichle, R.H., Walker, J.P., Koster, R.D., Houser, P.R., 2002b. Extended versus Ensemble Kalman Filtering for Land Data Assimilation. J. Hydrometeorol. 3, 728–740. https://doi.org/10.1175/1525-7541(2002)003<0728:EVEKFF>2.0.CO;2

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resour. Res. 46, W05521. https://doi.org/10.1029/2009WR008328

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., Franks, S.W., 2011. Toward a reliable

decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. Water Resour. Res. 47, W11516. https://doi.org/10.1029/2011WR010643

Renard, B., Laurent, B., Jerome, L.C., Branger, F., Dmitri, K., 2016. Software BaRatin (Version 2-1).

Rigon, R., Bancheri, M., Formetta, G., de Lavenne, A., 2016. The geomorphological unit hydrograph from a historical-critical perspective. Earth Surf. Process. Landforms 41, 27–37. https://doi.org/10.1002/esp.3855 State

Rodell, M., Houser, P.R., Rodell, M., Houser, P.R., 2004. Updating a Land Surface Model with MODIS-Derived Snow Cover. J. Hydrometeorol. 5, 1064–1075. https://doi.org/10.1175/JHM-395.1

Salamon, P., Feyen, L., 2010. Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation. Water Resour. Res. 46, W12501. https://doi.org/10.1029/2009WR009022

Salamon, P., Feyen, L., 2009. Assessing parameter, precipitation, and predictive uncertainty in a distributed hydrological model using sequential data assimilation with the particle filter. J. Hydrol. 376, 428–442. https://doi.org/10.1016/j.jhydrol.2009.07.051

Sanzana, P., Jankowfsky, S., Branger, F., Braud, I., Vargas, X., Hitschfeld, N., Gironás, J., 2013. Computer-assisted mesh generation based on hydrological response units for distributed hydrological modeling. Comput. Geosci. 57, 32–43. https://doi.org/10.1016/j.cageo.2013.02.006

Schiemann, R., Erdin, R., Willi, M., Frei, C., Berenguer, M., Sempere-Torres, D., 2011. Geostatistical radar-raingauge combination with nonparametric correlograms: methodological considerations and application in Switzerland. Hydrol. Earth Syst. Sci. 15, 1515–1536. https://doi.org/10.5194/hess-15-1515-2011

Semenov, M., Stratonovitch, P., 2010. Use of multi-model ensembles from global climate models for assessment of climate change impacts. Clim. Res. 41, 1–14. https://doi.org/10.3354/cr00836

Shafii, M., Tolson, B., Matott, L.S., 2014. Uncertainty-based multi-criteria calibration of rainfall-runoff models: A comparative study. Stoch. Environ. Res. Risk Assess. 28, 1493–1510. https://doi.org/10.1007/s00477-014-0855-x

Shafii, M., Tolson, B., Shawn Matott, L., 2015. Addressing subjective decision-making inherent in GLUE-based multi-criteria rainfall-runoff model calibration. J. Hydrol. 523, 693–705. https://doi.org/10.1016/j.jhydrol.2015.01.051

Sharma, S., Reed, S., Mejia, A., Siddique, R., Ahnert, P., 2019. Hydrological Model Diversity Enhances Streamflow Forecast Skill at Short- to Medium-Range Timescales. Water Resour. Res. 55, 1–21. https://doi.org/10.1029/2018wr023197

Shen, C., Phanikumar, M.S., 2010. Author ' s personal copy A process-based , distributed hydrologic model based on a large-scale method for surface – subsurface coupling. Adv. Water Resour. 33, 1524–1541. https://doi.org/10.1016/j.advwatres.2010.09.002

Sherman, L.K., 1932. Streamflow from rainfall by the unit hydrograph method. Eng. News-Record 501–505.

Shrestha, R.R., Rode, M., 2008. Multi-objective calibration and fuzzy preference selection of a distributed hydrological model. Environ. Model. Softw. 23, 1384–1395. https://doi.org/10.1016/j.envsoft.2008.04.001

Sikorska, A.E., Scheidegger, A., Banasik, K., Rieckermann, J., 2012. Bayesian uncertainty assessment of

flood predictions in ungauged urban basins for conceptual rainfall-runoff models. Hydrol. Earth Syst. Sci. 16, 1221–1236. https://doi.org/10.5194/hess-16-1221-2012

Šimůnek, J., van Genuchten, M.T., Šejna, M., 2016. Recent Developments and Applications of the HYDRUS Computer Software Packages. Vadose Zo. J. 15, 1–25. https://doi.org/10.2136/vzj2016.04.0033

Singh, V.P., Frevert, D.K., 2005. Watershed models. CRC Press.

Slater, A.G., Clark, M.P., 2006. Snow Data Assimilation via an Ensemble Kalman Filter. J. Hydrometeorol. 7, 478–493. https://doi.org/10.1175/JHM505.1

Smith, P.J., Thornhill, G.D., Dance, S.L., Lawless, A.S., Mason, D.C., Nichols, N.K., 2013. Data assimilation for state and parameter estimation: application to morphodynamic modelling. Q. J. R. Meteorol. Soc. 139, 314–327. https://doi.org/10.1002/qj.1944

Sorooshian, S., Dracup, J.A., 1980. Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases. Water Resour. Res. 16, 430–442. https://doi.org/10.1029/WR016i002p00430

Srinivasan, M.S., Steenhuis, T.S., Gérard-Marchant, P., Veith, T.L., Gburek, W.J., 2007. Watershed Scale Modeling of Critical Source Areas of Runoff Generation and Phosphorus Transport. J. Am. Water Resour. Assoc. 41, 361–377. https://doi.org/10.1111/j.1752-1688.2005.tb03741.x

Stedinger, J.R., Vogel, R.M., Lee, S.U., Batchelder, R., 2008. Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. Water Resour. Res. 44, W00B06. https://doi.org/10.1029/2008WR006822

Stordal, A.S., Karlsen, H.A., Naevdal, G., Skaug, H.J., Vallès, B., Karlsen, H.A., Skaug, · H J, 2011. Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter. Comput Geosci 15, 293–305. https://doi.org/10.1007/s10596-010-9207-1

Sugawara, M., 1979. Automatic calibration of the tank model / L'étalonnage automatique d'un modèle à cisterne. Hydrol. Sci. Bull. 24, 375–388. https://doi.org/10.1080/02626667909491876

Tangdamrongsub, N., Steele-Dunne, S.C., Gunter, B.C., Ditmar, P.G., Weerts, A.H., 2015. Data assimilation of GRACE terrestrial water storage estimates into a regional hydrological model of the Rhine River basin. Hydrol. Earth Syst. Sci. 19, 2079–2100. https://doi.org/10.5194/hess-19-2079-2015

Therrien, R., Sudicky, E.A., Park, Y.J., McLaren, R.G., 2010. HydroGeoSphere: A Three-Dimensional Numerical Modelling Describing Fully-Integrated Subsurface and Surface Flow and Transport. Groundwater Simulations Group, University of Waterloo, Waterloo, ON.

Thiboult, A., Anctil, F., 2015. On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments. J. Hydrol. 529, 1147–1160. https://doi.org/10.1016/J.JHYDROL.2015.09.036

Thiboult, A., Anctil, F., Boucher, M.A., 2016. Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrol. Earth Syst. Sci. 20, 1809–1825. https://doi.org/10.5194/hess-20-1809-2016

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S.W., Srikanthan, S., 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. Water Resour. Res. 45, W00B14. https://doi.org/10.1029/2008WR006825

Tippett, M.K., Anderson, J.L., Bishop, C.H., Hamill, T.M., Whitaker, J.S., 2003. Ensemble Square Root Filters. Mon. Weather Rev. 131, 1485–1490. https://doi.org/10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2

Todini, E., Wallis, J., 1977. Using the CLS for daily or longer period rainfall-runoff modelling, in: Mathematical Models for Surface Water Hydrology: Proceedings of the Workshop. Wiley, Pisa, Italy, p. 423.

Tolson, B.A., Shoemaker, C.A., 2008. Efficient prediction uncertainty approximation in the calibration of environmental simulation models. Water Resour. Res. 44, W04411. https://doi.org/10.1029/2007WR005869

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

Tuppad, P., 2006. Hydrologic modeling response to NEXRAD and raingage spatial variability and strategic watershed management. Ph.D. Thesis. Kansas State University, Manhattan, Kansas, USA.

Valéry, A., Andréassian, V., Perrin, C., 2014. "As simple as possible but not simpler": What is useful in a temperature-based snow-accounting routine? Part 1 - Comparison of six snow accounting routines on 380 catchments. J. Hydrol. 517, 1176–1187. https://doi.org/10.1016/j.jhydrol.2014.04.059

Velázquez, J.A., Anctil, F., Ramos, M.H., Perrin, C., 2011. Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. Adv. Geosci. 29, 33–42. https://doi.org/10.5194/adgeo-29-33-2011

Venetis, C., 1970. A Note on the Estimation of the Parameters in Logarithmic Stage-Discharge Relationships With Estimates of Their Error. Int. Assoc. Sci. Hydrol. Bull. 15, 105–111. https://doi.org/10.1080/02626667009493957

Vrugt, J.A., Diks, C.G.H., Gupta, H. V, Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. Water Resour. Res. 41, W01017. https://doi.org/10.1029/2004WR003059

Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43, W01411. https://doi.org/10.1029/2005WR004838

Vrugt, J.A., Sadegh, M., 2013. Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. Water Resour. Res. 49, 4335–4345. https://doi.org/10.1002/wrcr.20354

Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour. Res. 44, W00B09. https://doi.org/10.1029/2007WR006720

Wang, S., Huang, G.H., Baetz, B.W., Cai, X.M., Ancell, B.C., Fan, Y.R., 2017. Examining dynamic interactions among experimental factors influencing hydrologic data assimilation with the ensemble Kalman filter. J. Hydrol. 554, 743–757. https://doi.org/10.1016/j.jhydrol.2017.09.052

Weerts, A.H., El Serafy, G.Y.H., 2006. Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. Water Resour. Res. 42, W09403. https://doi.org/10.1029/2005WR004093

Weijs, S. V., van Nooijen, R., van de Giesen, N., Weijs, S. V., Nooijen, R. van, Giesen, N. van de, 2010. Kullback–Leibler Divergence as a Forecast Skill Score with Classic Reliability–Resolution–

Uncertainty Decomposition. Mon. Weather Rev. 138, 3387–3399. https://doi.org/10.1175/2010MWR3229.1

Whitaker, J.S., Hamill, T.M., 2002. Ensemble Data Assimilation without Perturbed Observations. Mon. Weather Rev. 130, 1913–1924. https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2

Wilks, D.S.S., Wilby, R.L.L., 1999. The weather generation game: a review of stochastic weather models. Prog. Phys. Geogr. 23, 329–357. https://doi.org/10.1177/030913339902300302

Winchell, M., Srinivasan, R., Di Luzio, M., Arnold, J., 2007. ArcSWAT interface for SWAT2005 User's guide.

WMO, 2003. Hydrological forecasting practices. World Meteorol. Organ. Libr. 41, 25–28. https://doi.org/10.1016/b978-0-444-86236-5.50111-5

Wood, E.F., Sivapalan, M., Beven, K., Band, L., 1988. Effects of spatial variability and scale with implications to hydrologic modeling. J. Hydrol. 102, 29–47. https://doi.org/https://doi.org/10.1016/0022-1694(88)90090-X

World Meteorological Organisation, 1994. Guide to Hydrological Practices, Hydrological Sciences Journal. https://doi.org/10.1080/02626667.2011.546602

Wright, A.J., Walker, J.P., Pauwels, V.R.N., 2017. Estimating rainfall time series and model parameter distributions using model data reduction and inversion techniques. Water Resour. Res. 53, 6407–6424. https://doi.org/10.1002/2017WR020442

Yadav, M., Wagener, T., Gupta, H., 2007. Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. Adv. Water Resour. 30, 1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005

Yang, W., Andréasson, J., Phil Graham, L., Olsson, J., Rosberg, J., Wetterhall, F., 2010. Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. Hydrol. Res. 41, 211–229. https://doi.org/10.2166/nh.2010.004

Yen, H., Su, Y.W., Wolfe, J.E., Chen, S.T., Hsu, Y.C., Tseng, W.H., Brady, D.M., Jeong, J., Arnold, J.G., 2015. Assessment of input uncertainty by seasonally categorized latent variables using SWAT. J. Hydrol. 531, 685–695. https://doi.org/10.1016/j.jhydrol.2015.10.058

Young, P., 2001. Data-based mechanistic modelling and validation of rainfall-flow processes, in: Model Validation: Perspectives in Hydrological Science. pp. 117–161.

Zamo, M., Naveau, P., 2018. Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. Math. Geosci. 50, 209–234. https://doi.org/10.1007/s11004-017-9709-7

Zhao, R.-J., 1992. The Xinanjiang model applied in China. J. Hydrol. 135, 371–381. https://doi.org/10.1016/0022-1694(92)90096-E

Appendix

A2-1. Description of the hydrologic model of Chapter 2

This hydrologic model first partitions precipitation into rain and snow, and each term is partly intercepted (and possibly evaporated) by rain and snow canopy. Snow storage is transformed to snowmelt using a simple degree-day approach. Excess rainfall and snowmelt are then combined, and a part of this combination is stored in depression areas, which are also subject to evaporation. The remainder of excess water is then partitioned into infiltration to the upper soil layer and overland runoff; the latter is added to the surface water storage. Next, the evapotranspiration process is applied to the upper layer soil moisture, and water is subsequently drained both vertically (i.e., percolation to the lower soil layer) and horizontally (i.e., interflow to surface water storage). Evapotranspiration is also applied to the lower layer soil moisture and then baseflow occurs to the surface water storage. Finally, surface water (composed of overland runoff, interflow and baseflow) is routed to the sub-watershed's outlet via a unit hydrograph routing scheme.

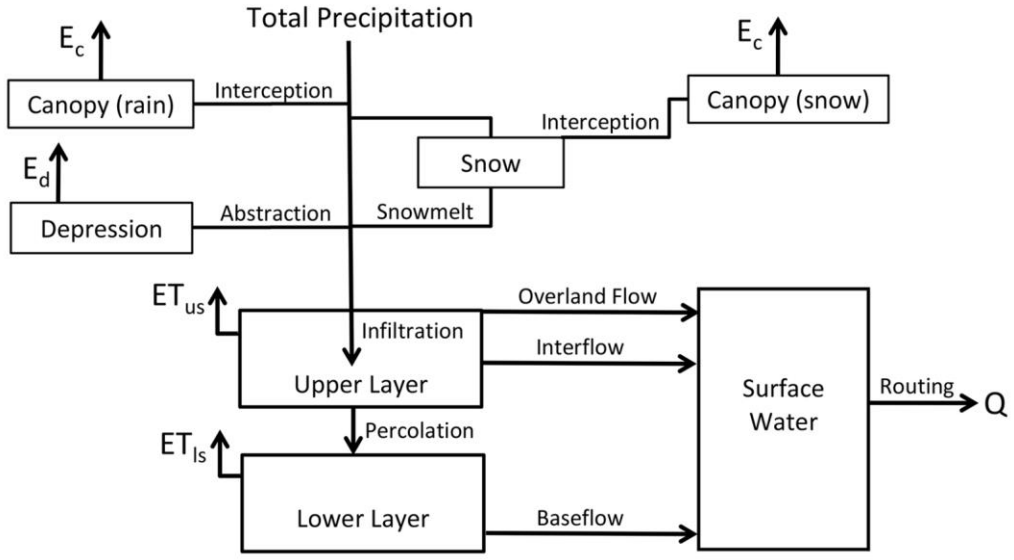

Figure A2-1. Hydrologic processes of the two-bucket simulation model. Ec and Ed represent evaporation from canopy and depression, respectively. ETus and ETls represent evapotranspiration from the upper layer and the lower layer of soil. Q is discharge from surface water routing.

A2-2. Detailed processes of subbasin re-discretization in ArcSWAT

This appendix provides the detailed processes of subbasin re-discretization of *Step 2c* in Chapter 2 (see Section 2.2.3.1). The descriptions are based on the ArcGIS and ArcSWAT manuals (ESRI, 2014; Winchell et al., 2007). Software used in Chapter 2 includes:

- ArcGIS 10.2.2 (http://www.esri.com/software/arcgis/arcgis-for-desktop)
- ArcSWAT 2012.10_2.17 (http://swat.tamu.edu/software/arcswat/).

Readers are suggested to refer to the official manuals of ArcGIS and ArcSWAT for the detailed descriptions based on their software version conditions.

1. Spatial reference setup

   ArcSWAT| Layers| Data Frame Properties| Projected Coordinate Systems

   Select the appropriate projection system based on case study requirement. This step does not only set the spatial reference but also determine the DEM horizontal (X-Y) unit.

2. DEM setup

   ArcSWAT| Watershed Delineator| Automatic Watershed Delineation| DEM Setup

   2.1. Load DEM grid map

   2.2. Click *DEM projection setup* button and change the vertical (Z) unit based on DEM data properties.

   2.3. Import mask (optional)

   Only the proportion of the DEM covered by the mask will be processed, thus it helps to reduce processing time.

   2.4. Import burn-in stream network (optional)

   The polyline shapefile of stream network is superimposed onto the DEM to improve hydrographic segmentation and sub-watershed boundary delineation.

3. Stream definition

   ArcSWAT| Watershed Delineator| Automatic Watershed Delineation| Stream Definition

   3.1. Flow direction and accumulation

   - Click *Flow direction and accumulation* button

     This step pre-processes the DEM by filling sinks and calculating the flow direction and flow accumulation grids. It can take a very long time when using high resolution DEMs (30-m or higher) over large areas, but it is once and for all because its results can be kept and reused in the future discretization. Therefore, we kindly remind readers that, in order to save operating time, please keep all the above settings once set up.

Note that, if all the above settings as well as flow direction and flow accumulation results are already kept from subbasin candidate schemes generation (*Step 1* of Chapter 2), the whole steps above can be omitted.

3.2. Type in the flow accumulation threshold (area in hectares)

After step 3.1, users will get a minimum, maximum, and suggested sub-watershed area in hectares (here understood as flow accumulation area). The minimum flow accumulation threshold was used to create the stream network in our case study.

3.3. Click *stream network* button

Step 3.3 will generate two layers: Reach (the current synthetic drainage network) and Monitoring Point (the respective stream junction points).

The reason of using the minimum flow accumulation threshold is to ensure that the later manually added sites of interest can be placed onto the existing reaches, and the required stream junctions of the selected satisfactory discretization scheme (from *Step 2b* of Chapter 2) can be covered by the stream junctions corresponding to the finest flow accumulation threshold. Users can change the flow accumulation threshold of Step 3.2 based on the locations of their interesting sites as well as the discretization level of the identified satisfactory schemes from *Step 2b* of Chapter 2.

3.4. Delete useless stream junctions for the refinement area based on the satisfactory discretization scheme identified from *Step 2b* of Chapter 2.

Assume users have the stream junctions of all candidate discretization schemes from subbasin candidate schemes generation (*Step 1* of Chapter 2). Now for each refinement area, take the junctions of the identified satisfactory scheme as bases; compare them with the stream junctions obtained from Step 3.3; delete the junctions that do not exist in the identified satisfactory scheme from the stream junctions obtained from Step 3.3 by the Edit tool of ArcGIS.

3.5. Delete useless stream junctions for the non-refinement area based on the uniform discretization scheme selected from *Step 1* of Chapter 2.

In the non-refinement area, compare the stream junctions of the selected uniform scheme from *Step 1* of Chapter 2 with the stream junctions obtained from Step 3.3, and delete the points that do not exist in the *Step 1* selected uniform scheme area from the stream junctions obtained from Step 3.3 by the Edit tool of ArcGIS.

4. Outlet and inlet definition

ArcSWAT| Watershed Delineator| Automatic Watershed Delineation| Outlet and Inlet Definition

4.1. Choose *subbasin outlet* option

4.2. Edit manually

Click *Add* and manually add the interesting sites by clicking the mouse over the map on the screen. These points will be stored in the Monitoring Point layer.

Note that users can also add inlet of draining watershed and point source input based on research requirement. These additional points can also be imported by Table.

5. Watershed outlet selection and definition

ArcSWAT| Watershed Delineator| Automatic Watershed Delineation| Watershed Outlet(s) Selection and Definition

5.1. Select the whole watershed outlet

Click *Select* button, and select the watershed outlet of the Grand River.

5.2. Delineate watershed

Click *Delineate watershed* button. The watershed delineation process will run.

Compared with the traditional subbasin discretization processes with a uniform threshold throughout the whole watershed, the special processes of subbasin re-discretization (*Step 2c* of Chapter 2) only exist in Step 3.2-3.5. Their purpose is assigning different detail levels of stream junctions to different areas according to their satisfactory subbasin schemes. For the refinement area, finer stream junctions are applied, while for the non-refinement area, the stream junctions of the selected uniform discretization scheme from *Step 1* of Chapter 2 are used. We reiterate once again that applying the minimum flow accumulation threshold in Step 3.2 is unnecessary. Users are free to employ the appropriate discretization threshold as long as all satisfactory levels of stream junctions can be obtained and the interesting sites can be added onto the reach lines.

Moreover, note that the HRU discretization refinement does not need re-discretization. The original intention of conducting subbasin re-discretization is to get a systematic upstream-downstream routing structure throughout the entire watershed with different detail levels of stream junction. Such a systematic upstream-downstream routing structure is needed in in-channel routing length calculation. However, HRU error metrics do not depend upon the routing path, therefore re-discretization is not needed in HRU discretization refinement, and different detail levels of HRUs can be simply merged into an output layer.

A3-1. Description of the GR4J hydrologic model

The description of the GR4J model is minimized here and closely follows Perrin et al. (2003). Readers are suggested referring to Perrin et al. (2003) for a more complete description of the GR4J model. As shown in Figure A3-1, GR4J model takes precipitation and evapotranspiration as model inputs (forcing data). The model state variables are the water levels of the production store and the routing store. Water quantities are all in unit of mm. The following operation description is for a single time step.

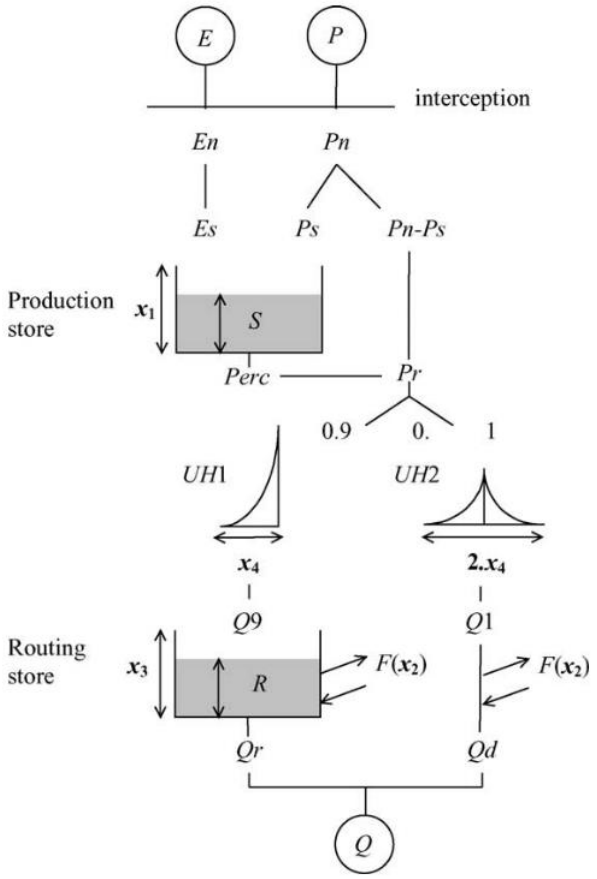

Figure A3-1. Hydrologic processes of the GR4J model (from Figure 1 of Perrin et al. (2003)).

Step 1. Net rainfall and evapotranspiration determination

When rainfall (P) is no less than evapotranspiration (E), net rainfall (P_n) is equal to P minus E, and net evapotranspiration capacity (E_n) is zero. Otherwise, P_n is zero, and E_n is equal to E minus P.

$$\begin{aligned} P_n &= P - E, and\ E_n = 0, when\ P \geq E \\ P_n &= 0, and\ E_n = E - P, when\ P < E \end{aligned} \quad \text{(A3-1)}$$

Step 2. Production store water level update

When P_n is not zero, a part P_s of P_n fills the production store. P_s is determined by:

$$P_s = \frac{x_1\left(1-(\frac{S}{x_1})^2\right)\tanh(\frac{P_n}{x_1})}{1+\frac{S}{x_1}\tanh(\frac{P_n}{x_1})} \quad \text{(A3-2)}$$

where $x_1$is the maximum capacity of the production store (mm).

When E_n is not zero, a quantity of water is evaporated from the production store. The actual evaporation is calculated by:

$$E_s = \frac{S\left(2-\frac{S}{x_1}\right)\tanh(\frac{E_n}{x_1})}{1+(1-\frac{S}{x_1})\tanh(\frac{E_n}{x_1})} \quad \text{(A3-3)}$$

Given the filled rainfall and actual evaporation, the water level (S) of the production store is updated with:

$$S = S - E_s + P_s \quad \text{(A3-4)}$$

A percolation leak from the production store is calculated as:

$$Perc = S\left\{1 - [1 + (\frac{4}{9}\frac{S}{x_1})^4]^{-\frac{1}{4}}\right\} \quad \text{(A3-5)}$$

Percolation is always less than the water level of the production store. After percolation, the water level of the production store is updated with:

$$S = S - Perc \quad \text{(A3-6)}$$

Step 3. Streamflow routing with unit hydrograph

The total amount of water to be routed is equal to:

$$P_r = Perc + (P_n - P_s) \quad \text{(A3-7)}$$

P_r is divided into two flow components. 90% of $P_r$is routed by a unit hydrograph UH1 and then a non-linear routing store. The remaining 10% of P_r is routed by a unit hydrograph UH2. UH1 has a time base of x_4 days, and UH2 has a time base of $2x_4$ days. With UH1 and UH2, one can simulate the time lag between the rainfall event and the resulting streamflow peak.

The ordinates of unit hydrograph UH1 and UH2 are used to spread effective rainfall over several successive time steps. Assume UH1 and UH2 have *n* and *m* ordinates, respectively. *n* and *m* are the smallest integers exceeding x_4 and $2x_4$, respectively. The ordinates of both unit hydrographs are derived from the S-curves (cumulative proportion of the input with time) denoted by SH1 and SH2, respectively. SH1 and SH2 are defined along time t:

$$SH1(t) = \begin{cases} 0, & for\ t \le 0 \\ (\frac{t}{x_4})^{\frac{5}{2}}, & for\ 0 < t < x_4 \\ 1, & for\ t \ge x_4 \end{cases} \quad \text{(A3-8)}$$

$$SH2(t) = \begin{cases} 0, & for\ t \le 0 \\ \frac{1}{2}(\frac{t}{x_4})^{\frac{5}{2}}, & for\ 0 < t \le x_4 \\ 1 - \frac{1}{2}(2 - \frac{t}{x_4})^{\frac{5}{2}}, & for\ x_4 < t < 2x_4 \\ 1, & for\ t \ge 2x_4 \end{cases} \quad \text{(A3-9)}$$

Then UH1 and UH2 ordinates are calculated by:

$$UH1(j) = SH1(j) - SH1(j-1) \quad \text{(A3-10)}$$

$$UH2(j) = SH2(j) - SH2(j-1) \quad \text{(A3- 11)}$$

where j is an integer.

Step 4. Catchment water exchange

Groundwater exchange acts on both flow components and is calculated by:

$$F = x_2(\frac{R}{x_3})^{\frac{7}{2}} \quad \text{(A3-12)}$$

where R is the water level of the routing store, x_3 is the one day ahead maximum capacity of the routing store (mm), x_2 is the groundwater exchange coefficient (mm). x_2 represents the maximum quantity of water that can be added to or be released from each model flow component when the routing store level R equals to x_3.

Step 5. Routing store water level update

The water level of the routing store is updated by adding the output Q_9 of UH1 and F as below:

$$R = \max\{0; R + Q_9 + F\} \quad \text{(A3-13)}$$

The outflow Q_r of the routing store is calculated as:

$$Q_r = \mathrm{R}\left\{1 - [1 + (\frac{R}{x_3})^4]^{-1/4}\right\} \quad \text{(A3-14)}$$

Outflow is always less than the water level of the routing store. After outflow, the water level of the routing store is updated with:

$$R = R - Q_r \quad \text{(A3-15)}$$

Step 6. Total streamflow

The output Q_d of UH2 is also subject to the same groundwater exchange F, so the resultant flow component is calculated as:

$$Q_d = \max\{0; Q_1 + F\} \quad \text{(A3-16)}$$

The total streamflow Q is calculated as:

$$Q = Q_r + Q_d \quad \text{(A3-17)}$$

A4-1. Normalized Root-mean-square error Ratio (NRR)

The NRR measures the spread of the forecast flow ensemble. It is calculated as (Thiboult and Anctil, 2015):

$$NRR = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}[(\frac{1}{N}\sum_{n=1}^{N}\hat{y}_{t,n})-y_t]^2}}{\frac{1}{N}\{\sum_{n=1}^{N}\sqrt{\frac{1}{T}[\sum_{t=1}^{T}(\hat{y}_{t,n}-y_t)^2]}\}\sqrt{\frac{N+1}{2N}}} \quad \text{(A4-1)}$$

where T is the number of time steps of the evaluation period ($t = 1, \dots, T$). $\hat{y}_{t,n}$ and y_t are the n^{th} forecasted ensemble flow and the observed flow at time t, respectively. N is the forecast ensemble size ($n = 1, \dots, N$). A value of 1 indicates an appropriate spread. Greater and smaller values than 1 reflect too narrow and wide ensembles, respectively.

A5-1. Flow forecast hydrographs of the multi-model and single-model forecasting systems

This appendix includes, for all catchments of Table 3-1 except the 13th catchment, the first lead day flow forecast hydrographs of the multi-model and three single-model forecasting systems of the entire forecasting period. See Table 5-4 for the descriptions of the multi-model and three single-model forecasting systems.

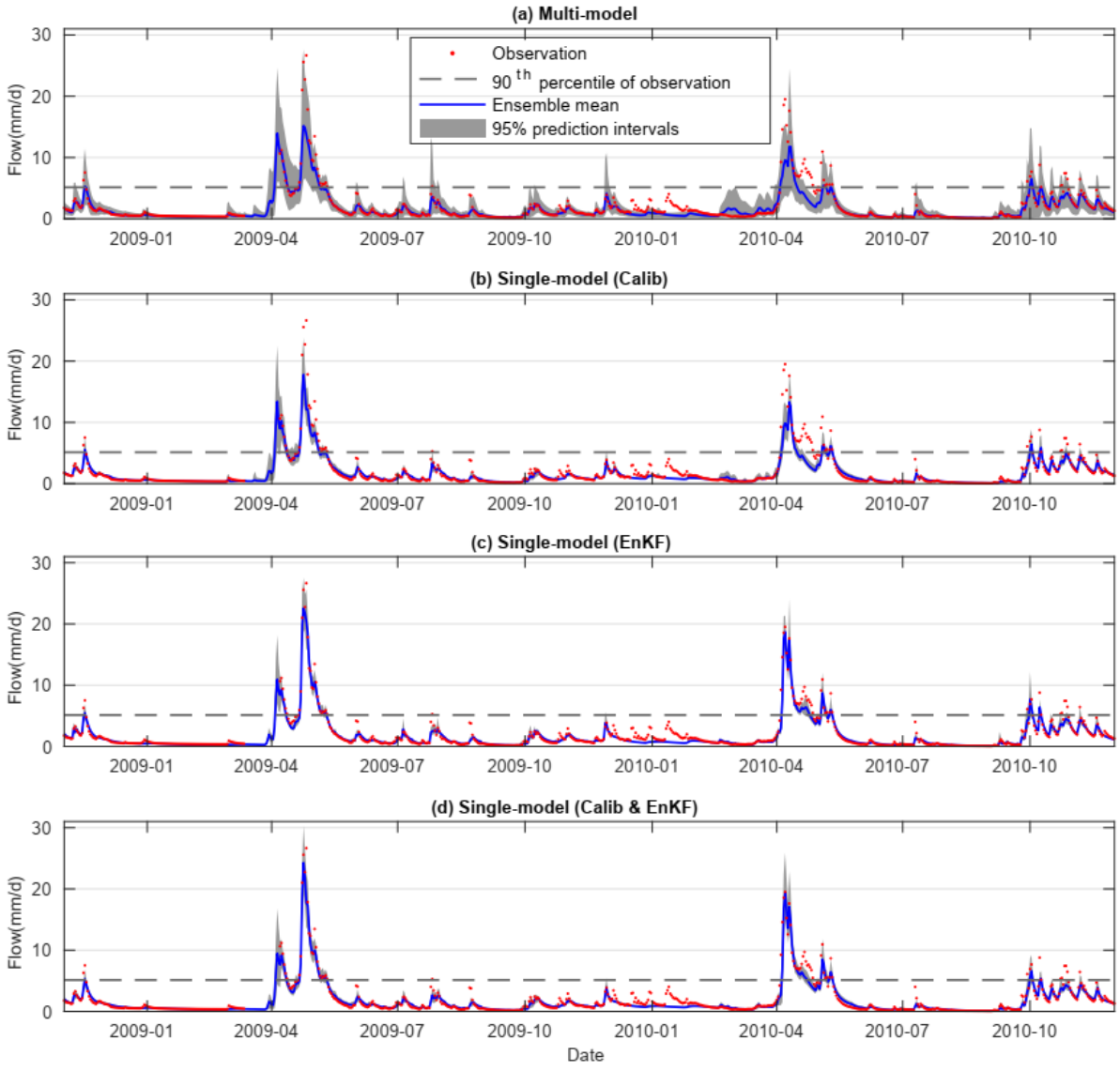

Figure A5-1. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Trois Pistoles catchment (the 1st catchment of Table 3-1) of the entire forecasting period.

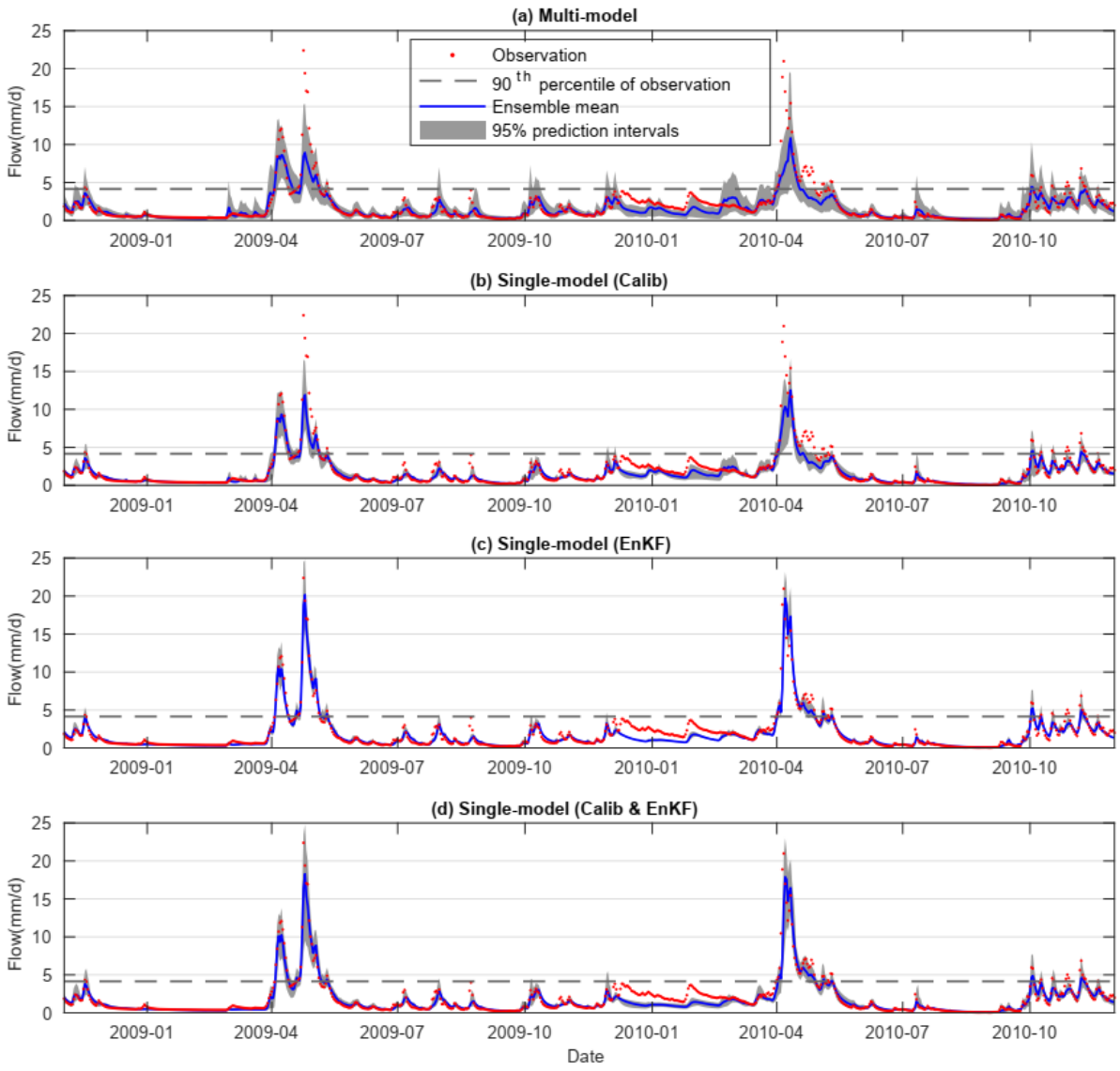

Figure A5-2. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Du Loup catchment (the 2nd catchment of Table 3-1) of the entire forecasting period.

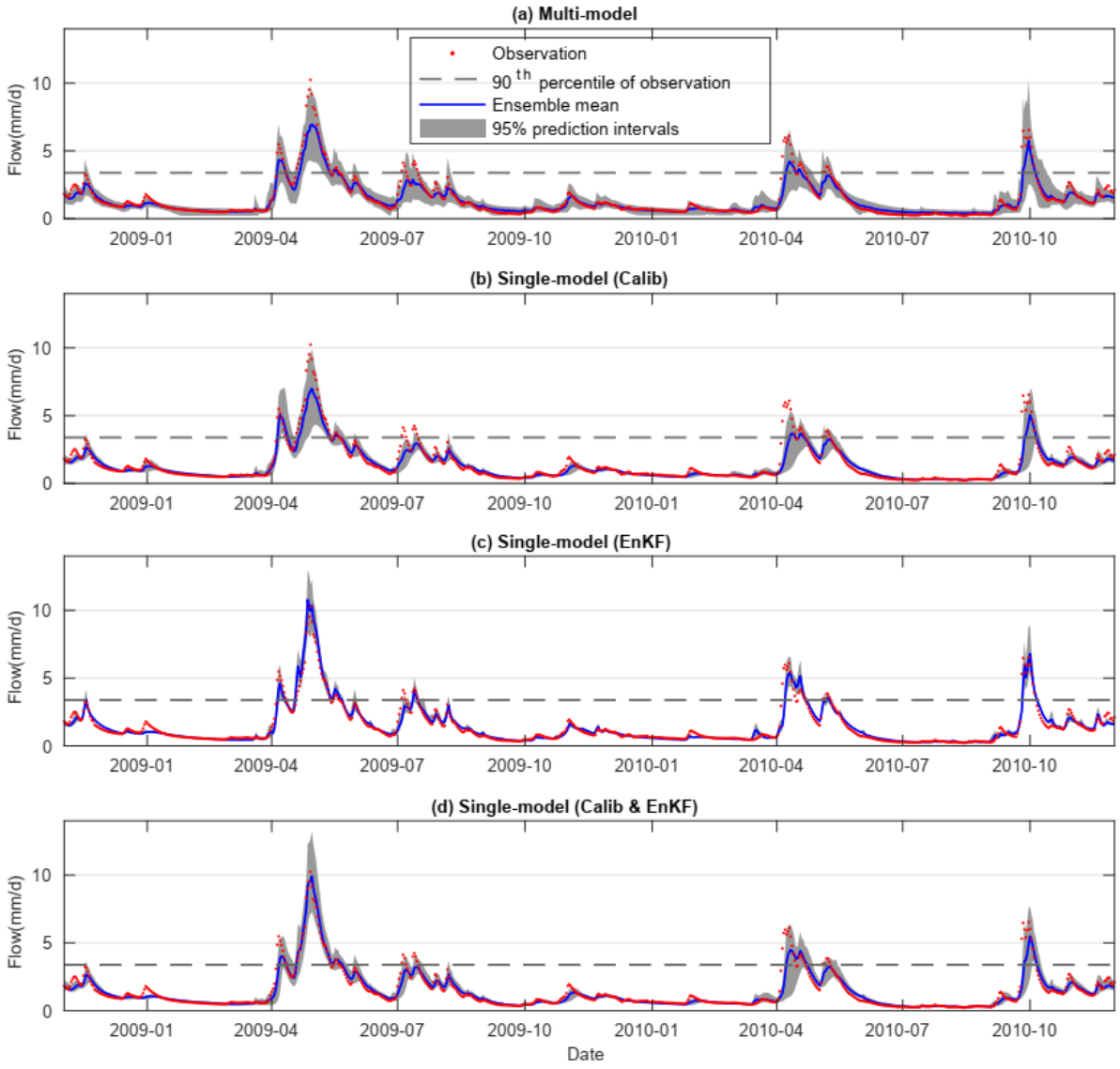

Figure A5-3. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Gatineau catchment (the 3rd catchment of Table 3-1) of the entire forecasting period.

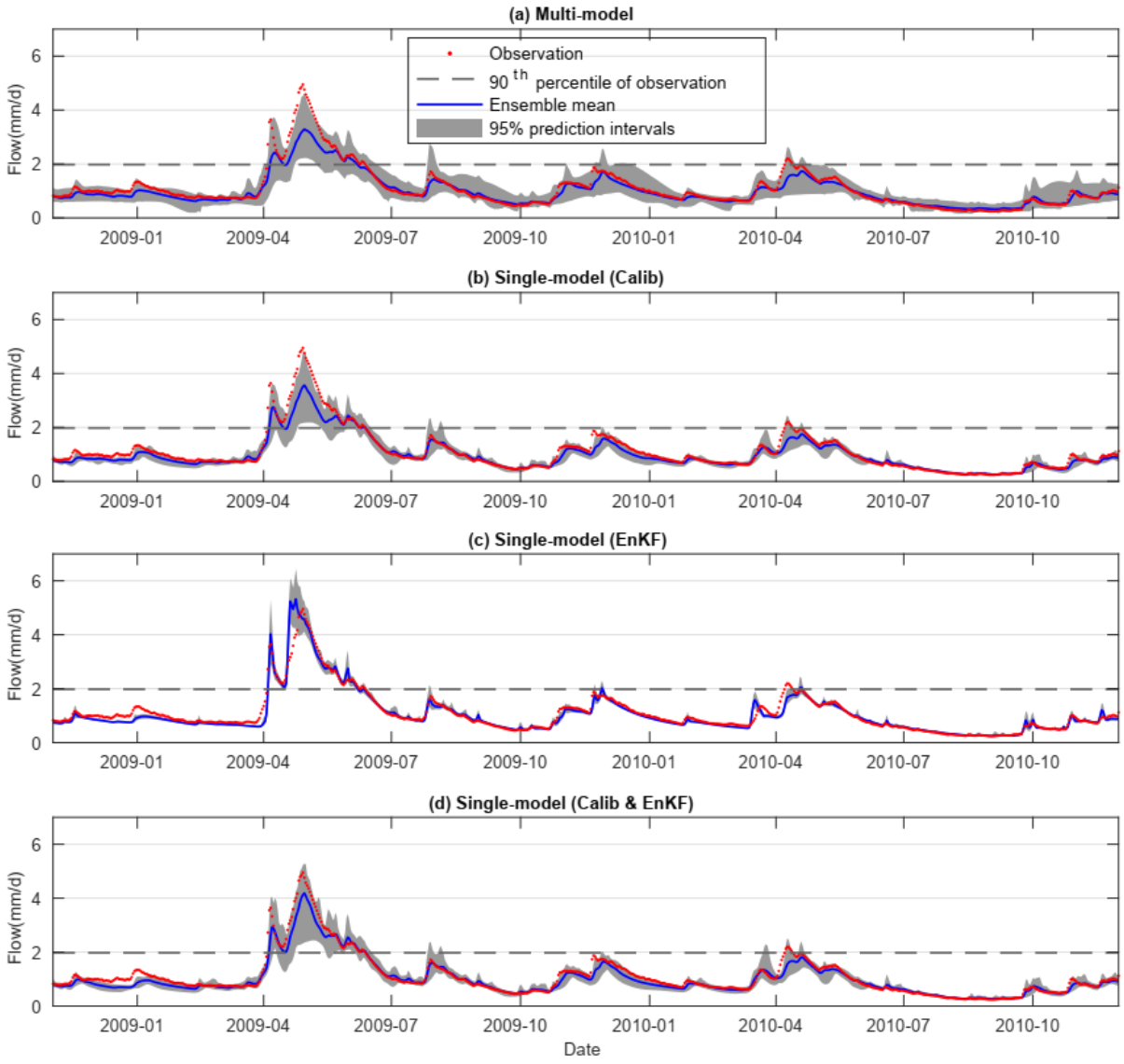

Figure A5-4. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Dumoine catchment (the 4th catchment of Table 3-1) of the entire forecasting period.

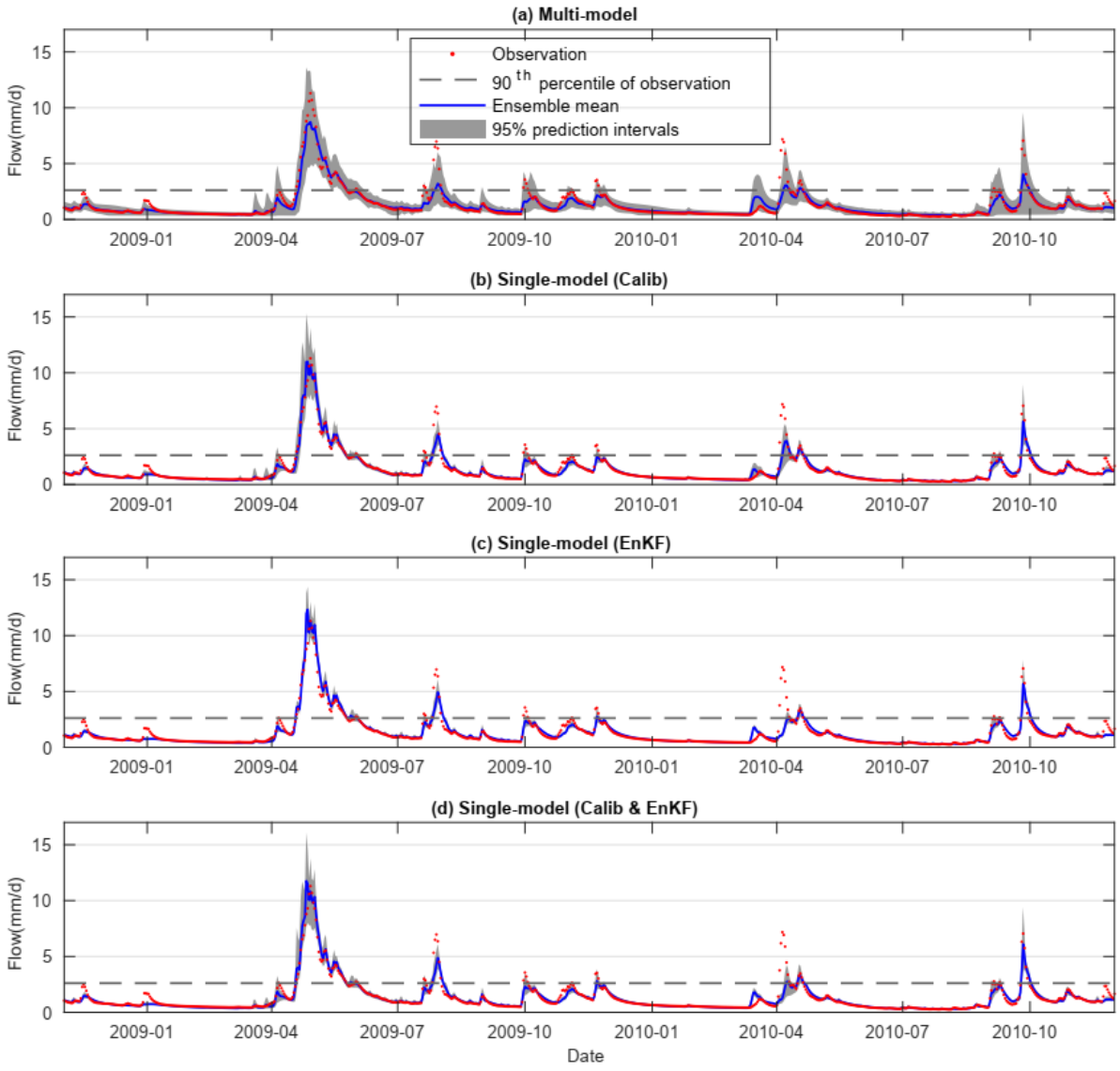

Figure A5-5. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Kinojevis catchment (the 5th catchment of Table 3-1) of the entire forecasting period.

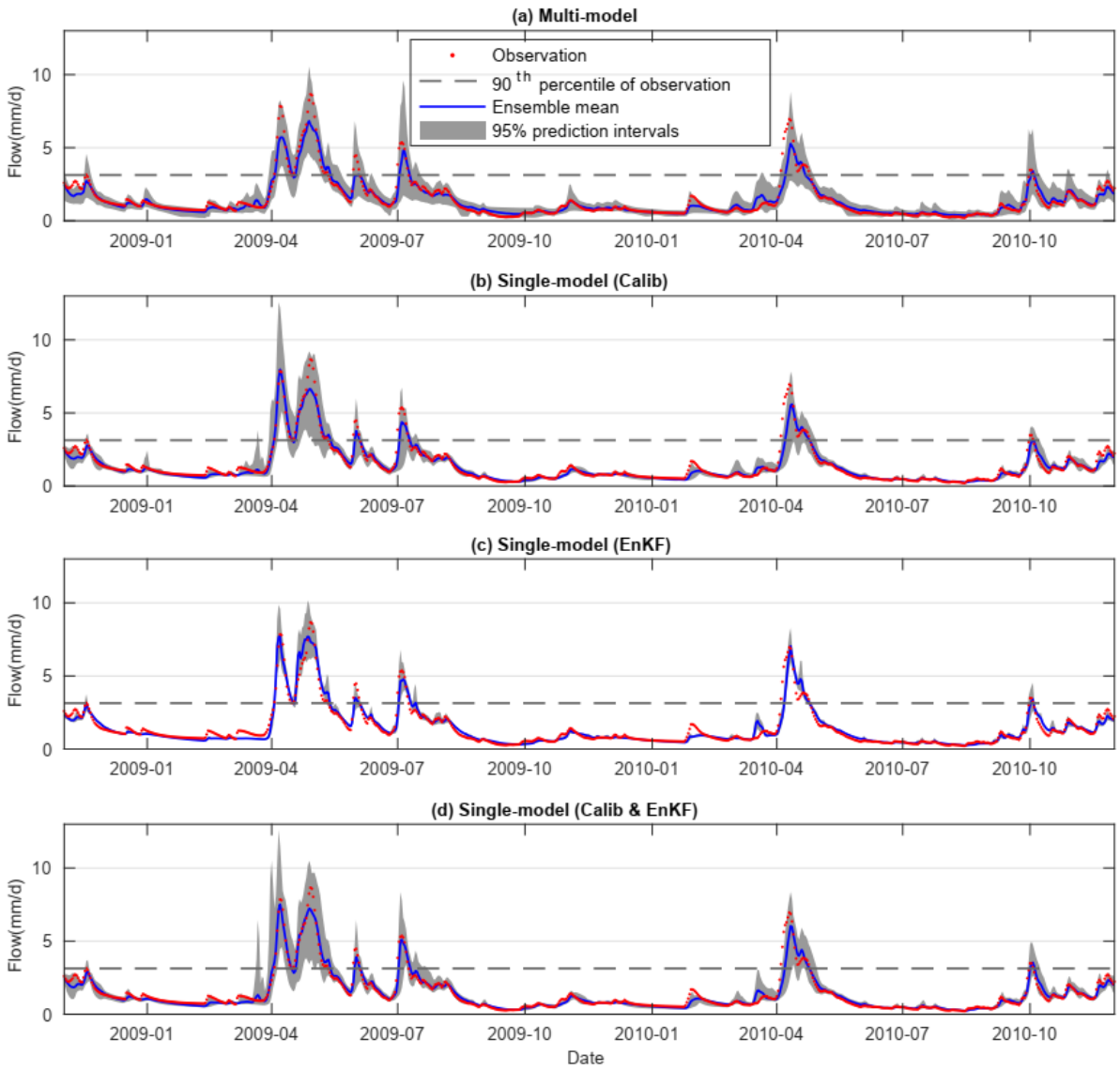

Figure A5-6. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Matawin catchment (the 6th catchment of Table 3-1) of the entire forecasting period.

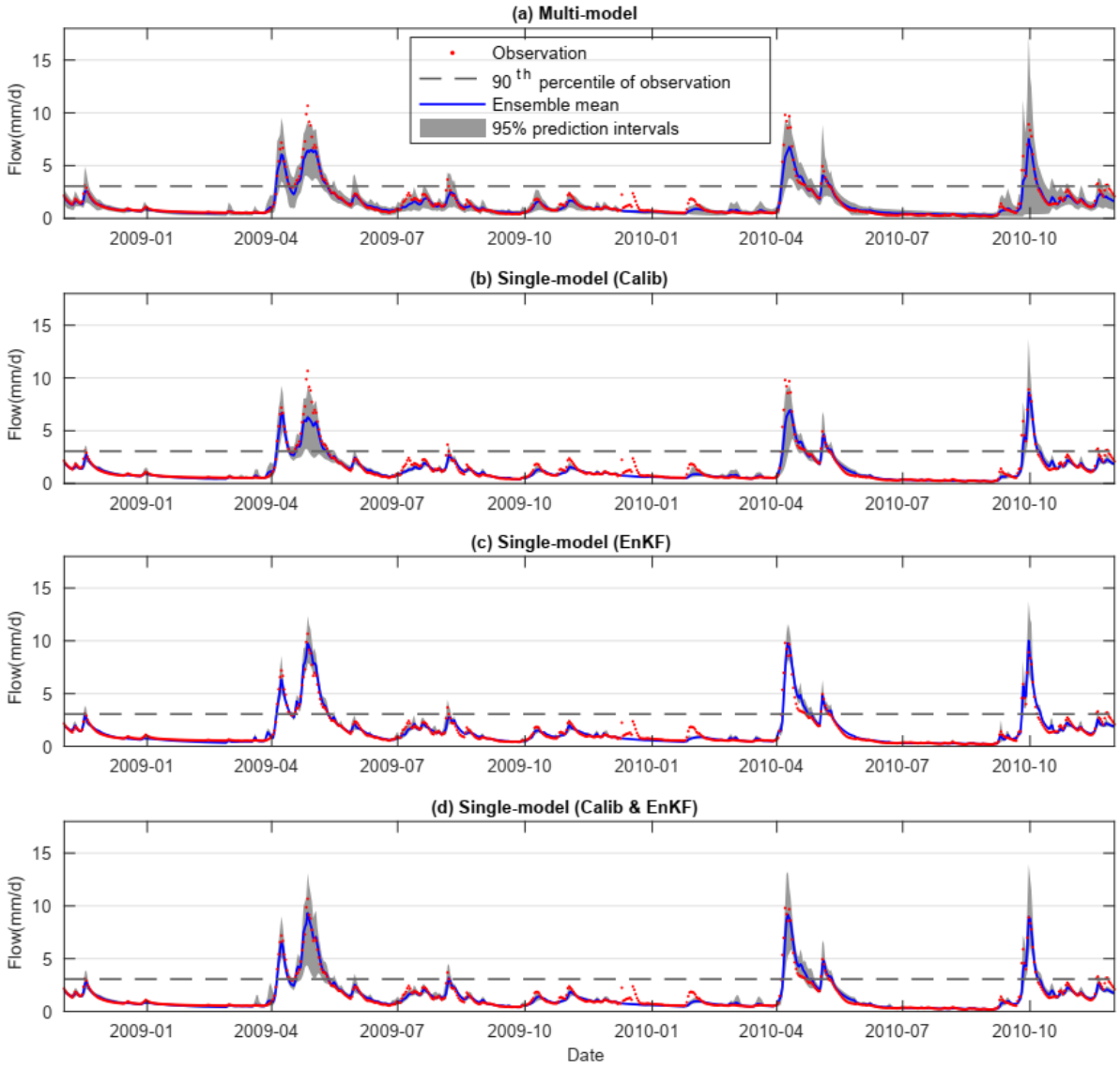

Figure A5-7. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Croche catchment (the 7th catchment of Table 3-1) of the entire forecasting period.

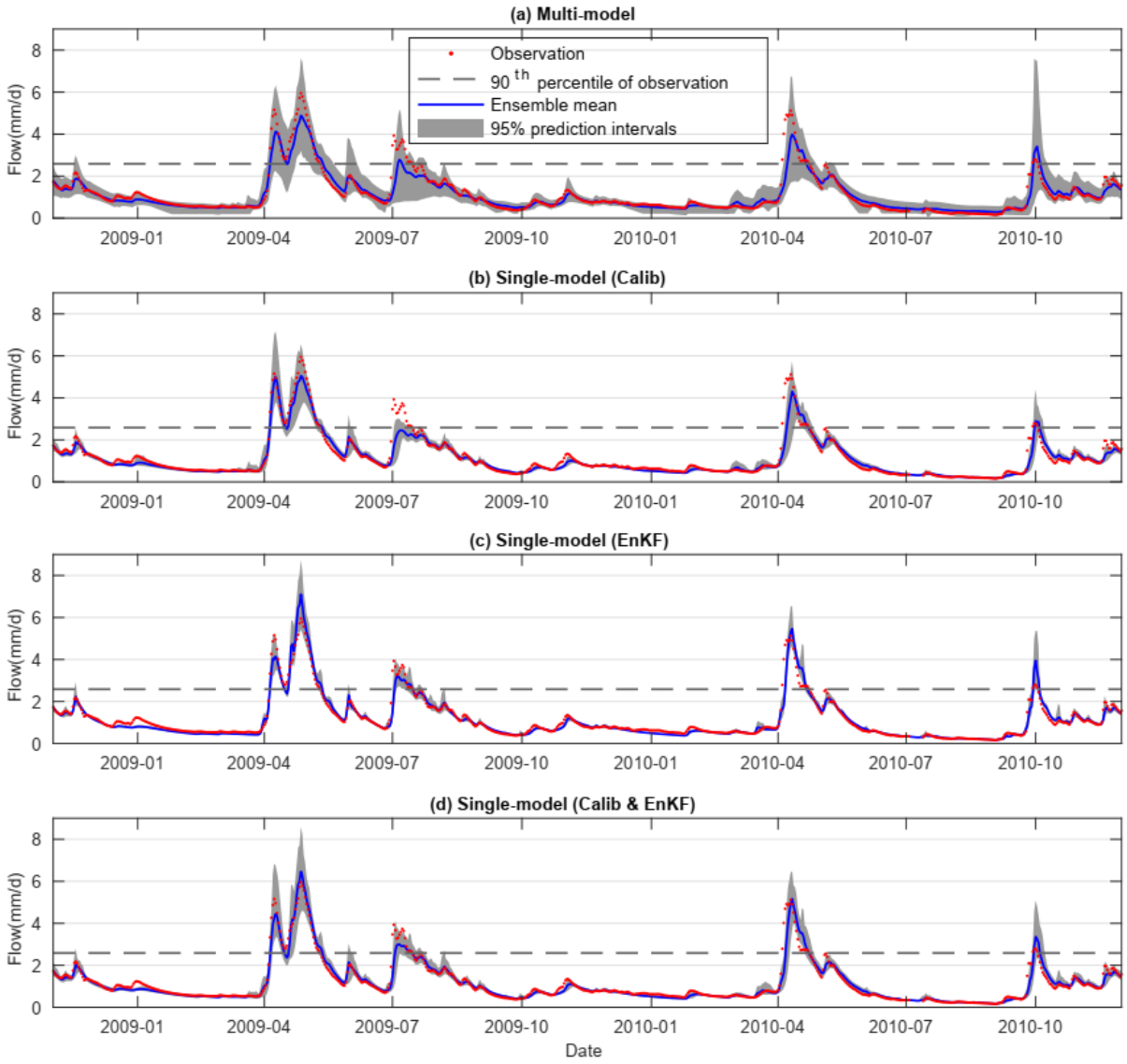

Figure A5-8. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Vermillon catchment (the 8th catchment of Table 3-1) of the entire forecasting period.

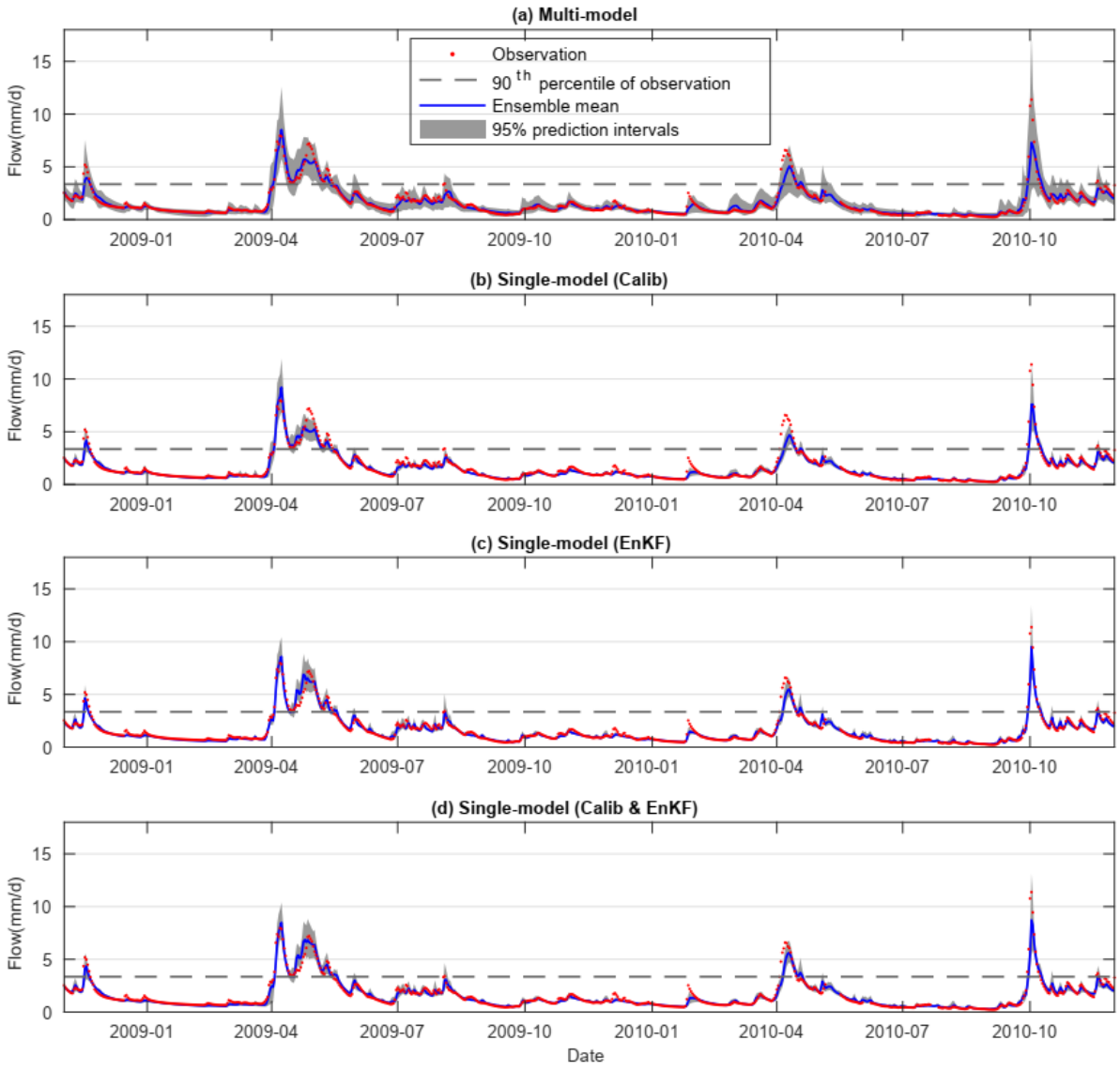

Figure A5-9. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Batiscan catchment (the 9th catchment of Table 3-1) of the entire forecasting period.

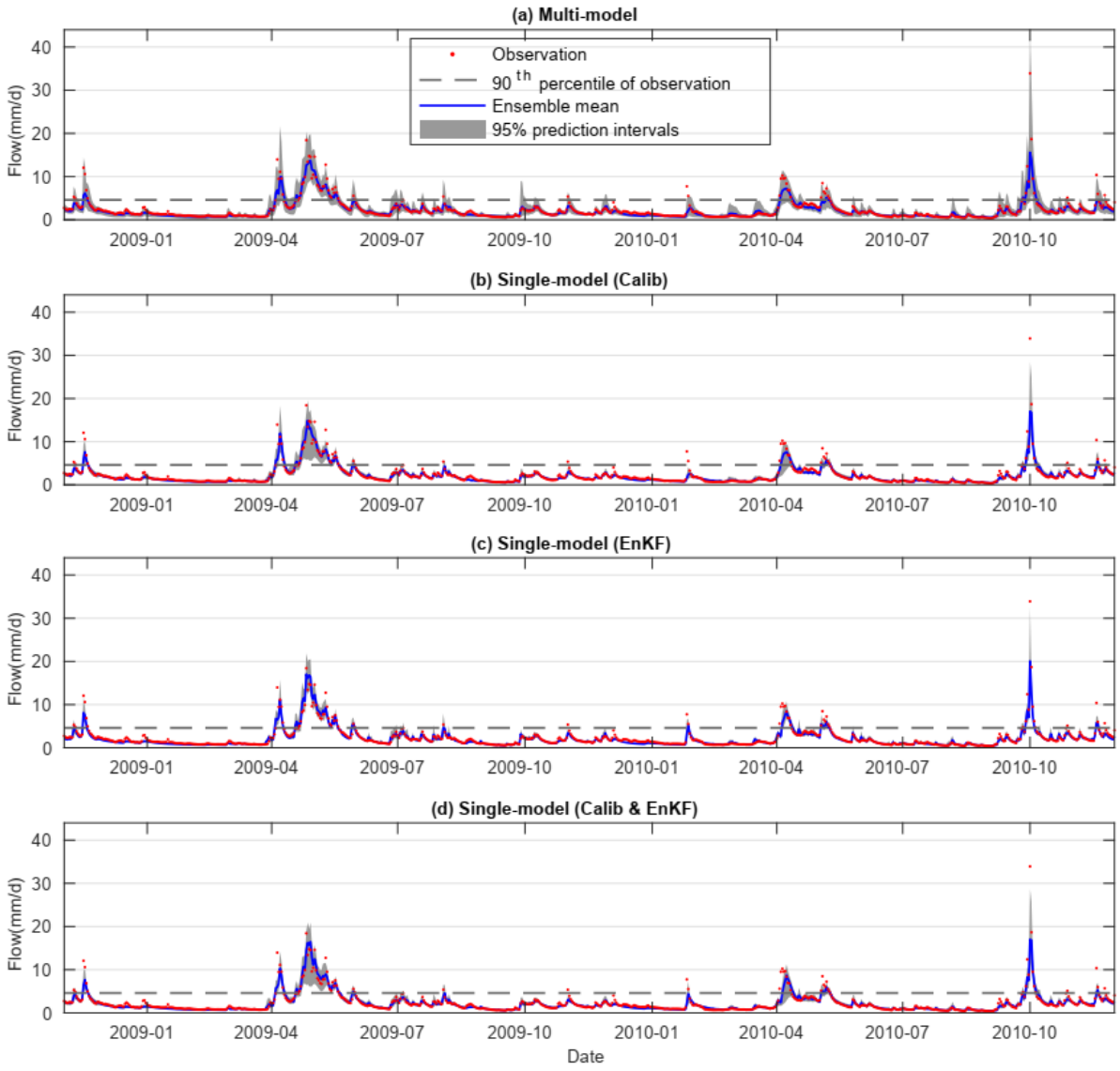

Figure A5-10. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Sainte Anne catchment (the 10th catchment of Table 3-1) of the entire forecasting period.

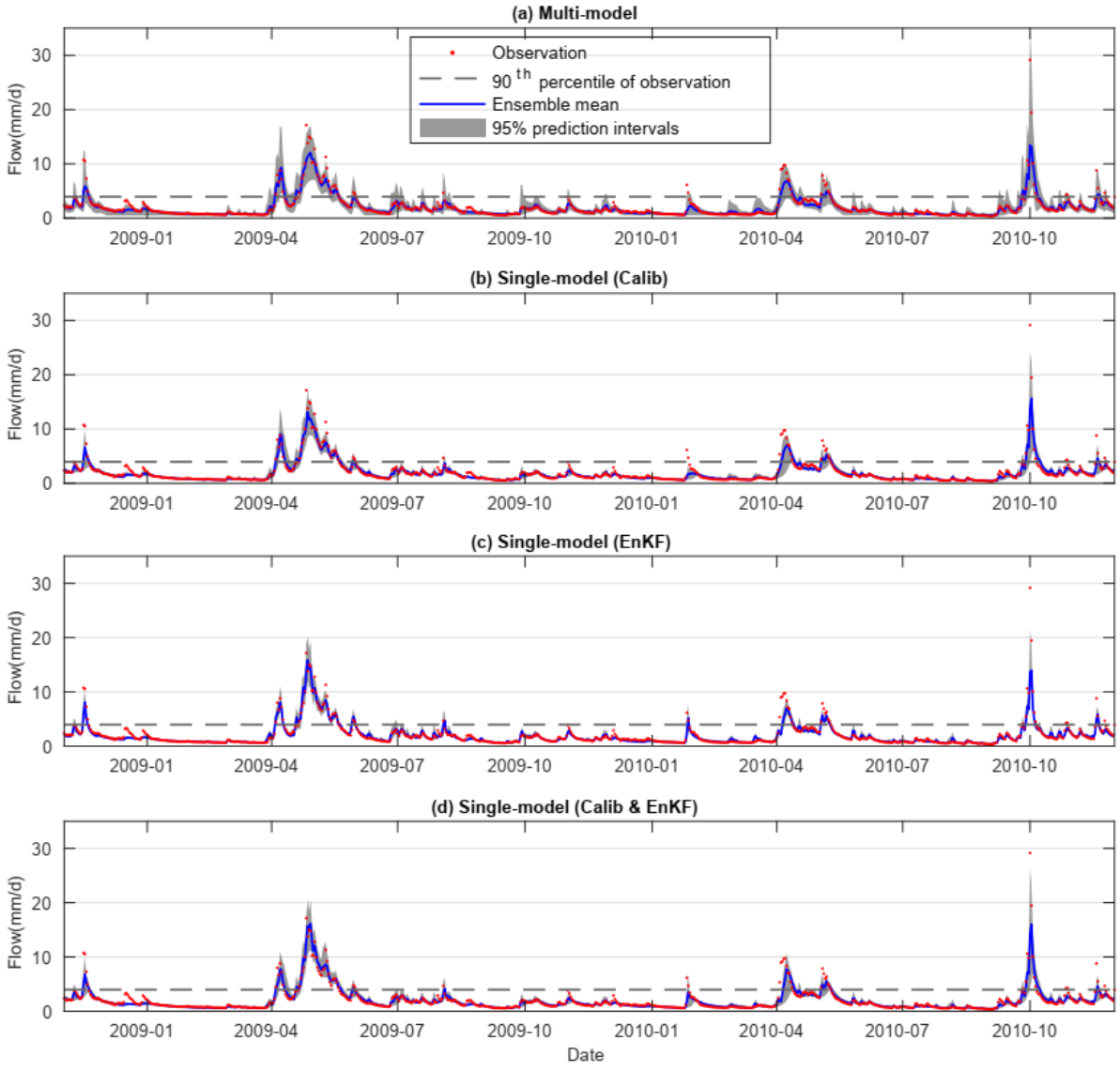

Figure A5-11. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Bras du Nord catchment (the 11th catchment of Table 3-1) of the entire forecasting period.

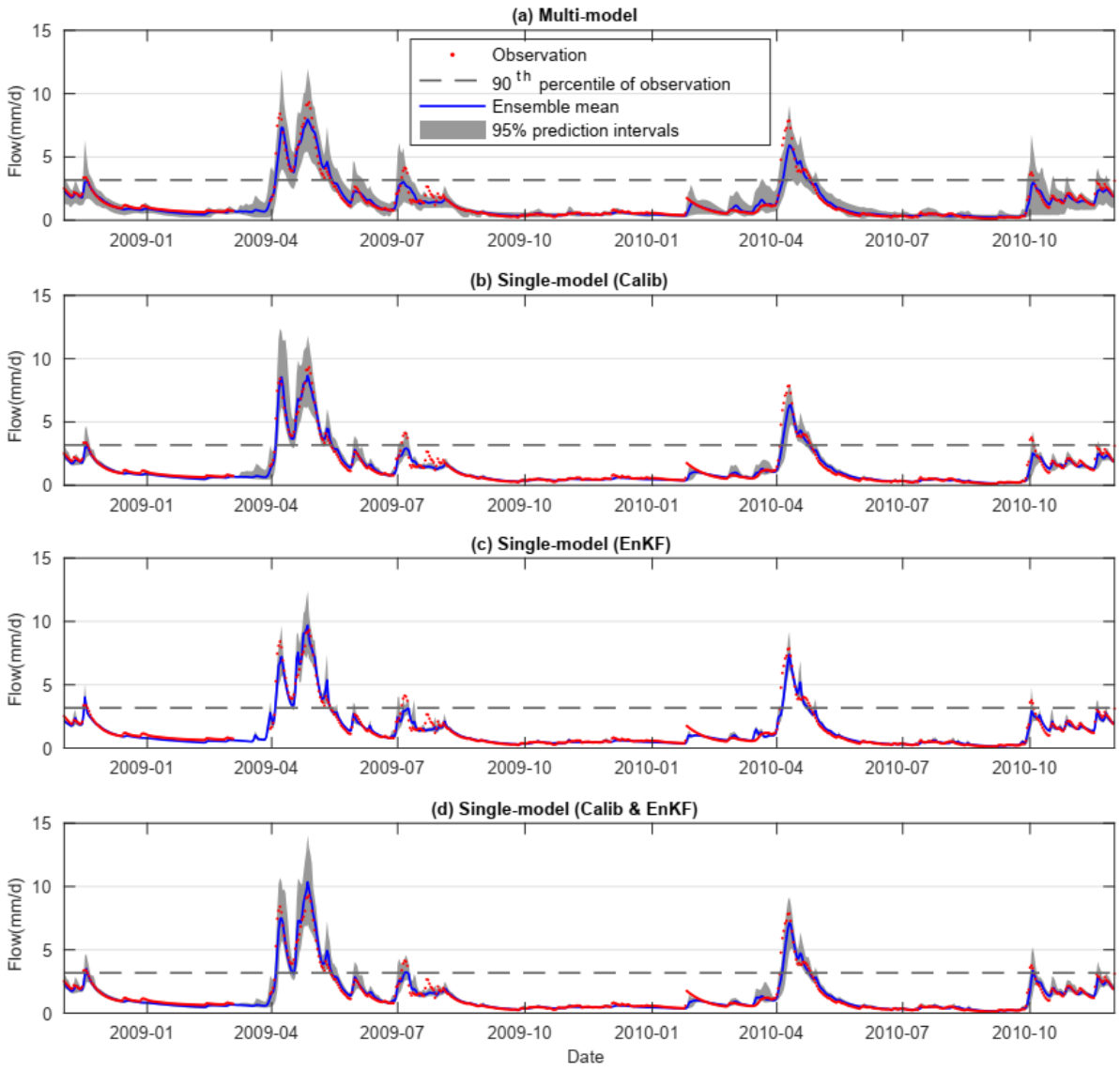

Figure A5-12. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Du Loup catchment (the 12th catchment of Table 3-1) of the entire forecasting period.

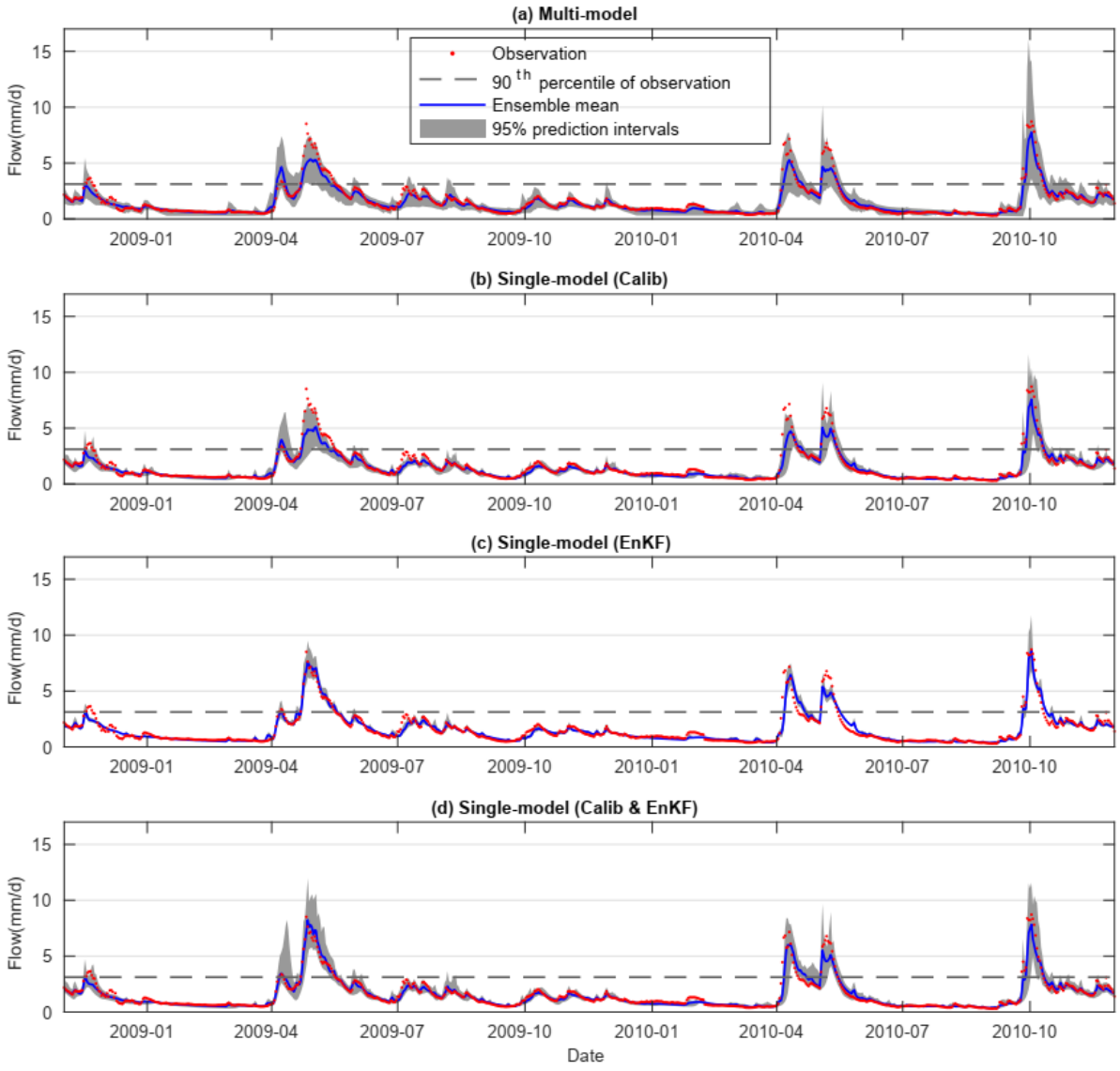

Figure A5-13. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Metabetchouane catchment (the 14th catchment of Table 3-1) of the entire forecasting period.

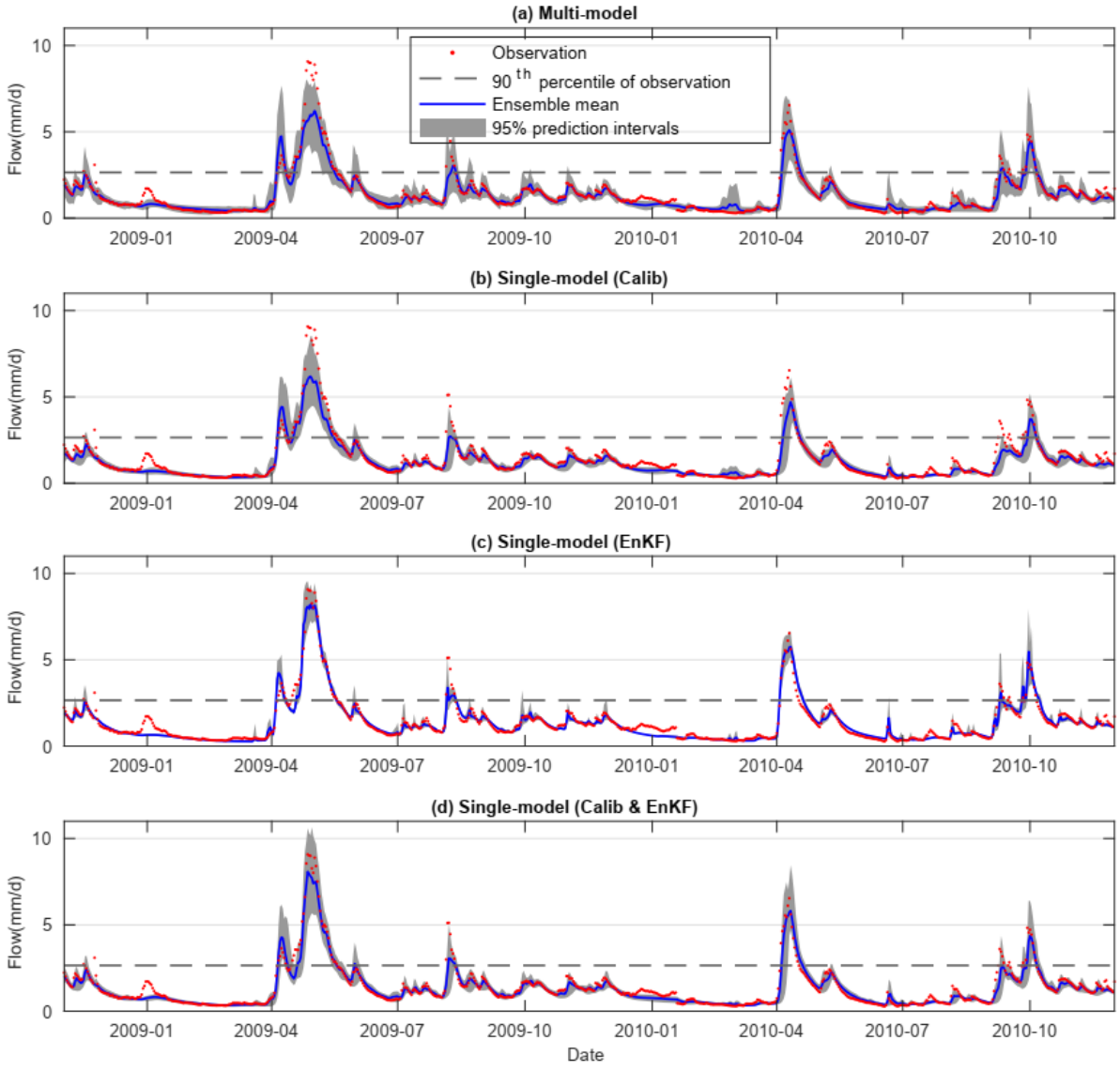

Figure A5-14. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Peribonka catchment (the 15th catchment of Table 3-1) of the entire forecasting period.

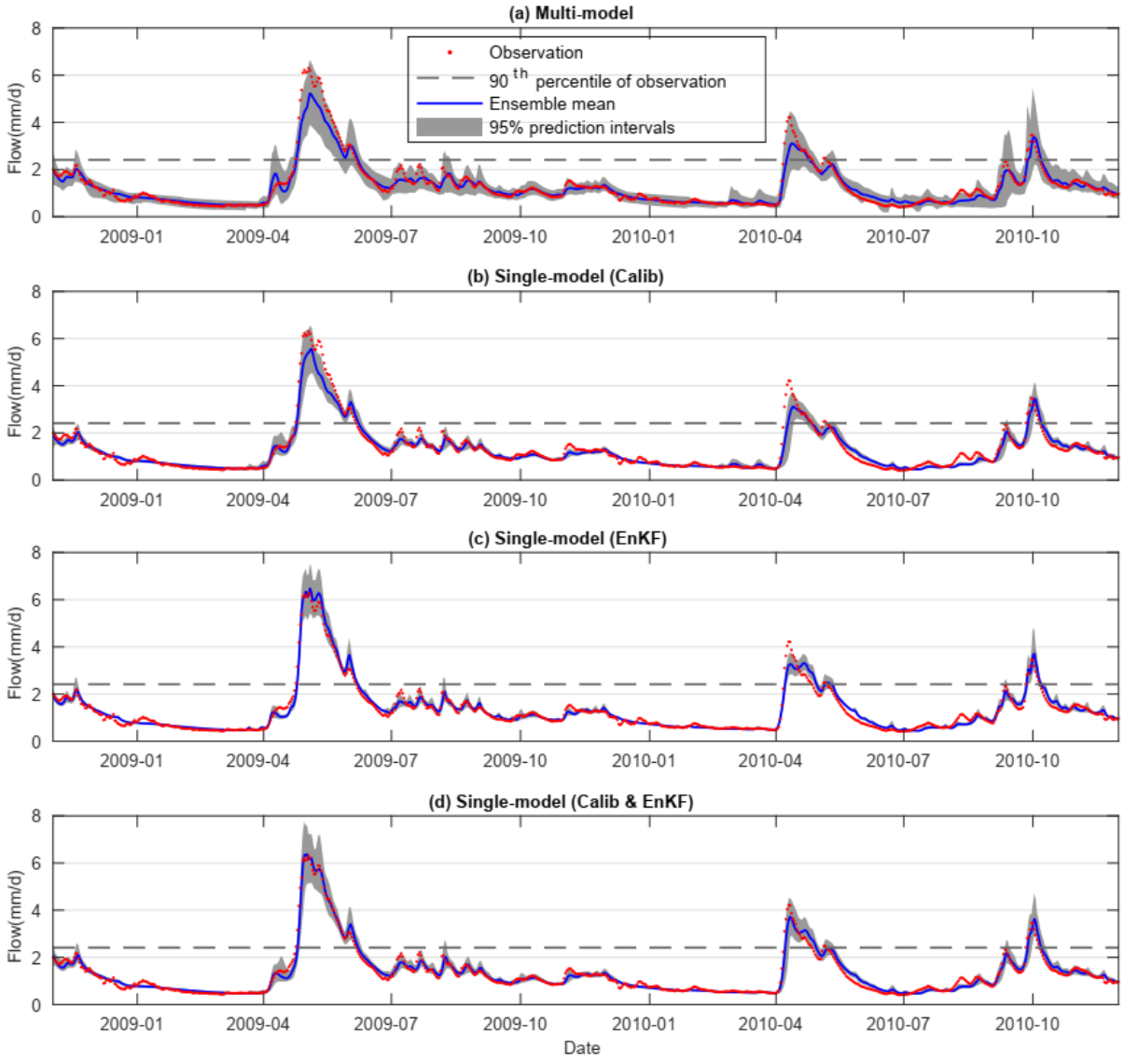

Figure A5-15. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Ashuapmushuan catchment (the 16th catchment of Table 3-1) of the entire forecasting period.

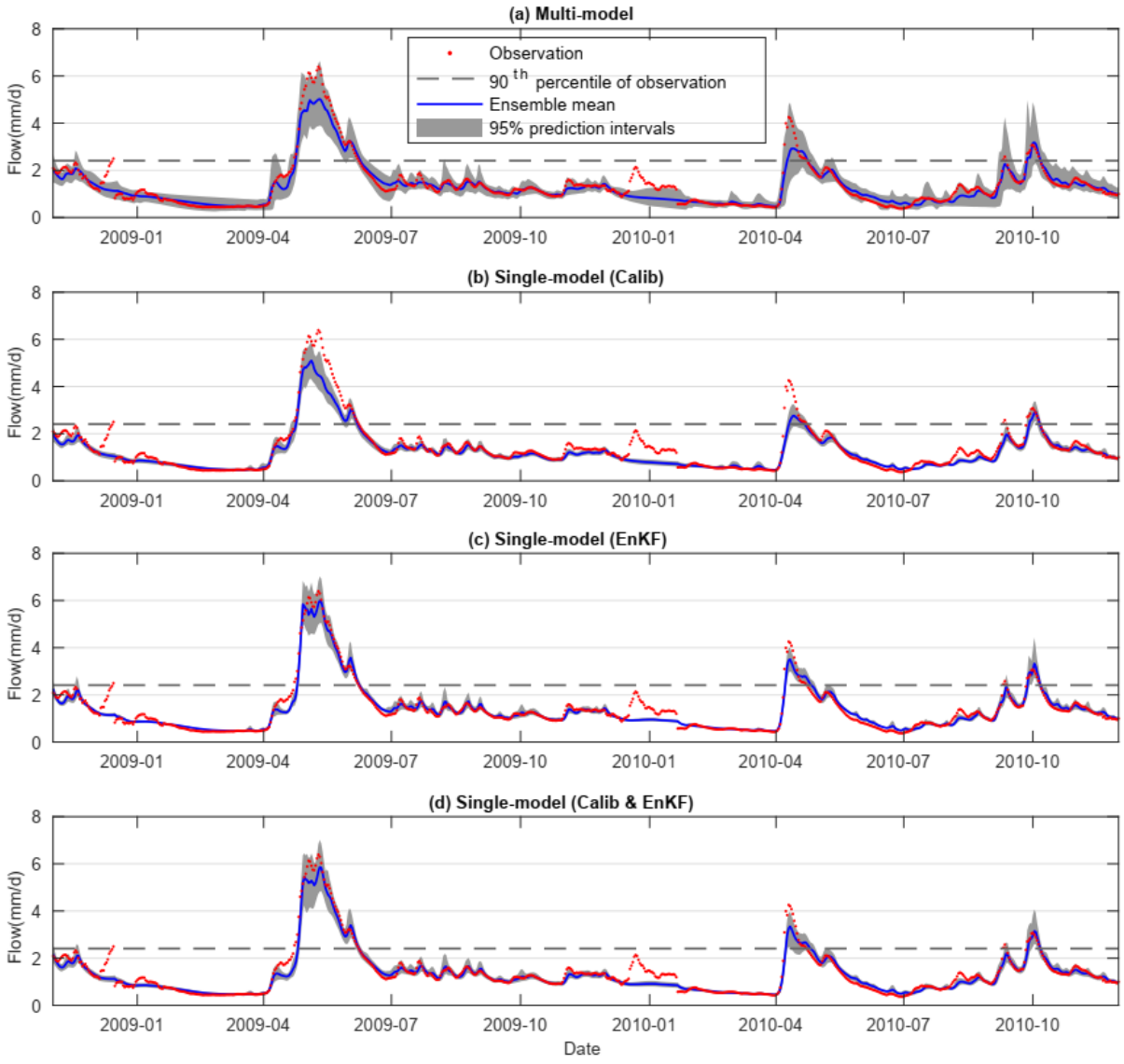

Figure A5-16. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Ashuapmushuan catchment (the 17th catchment of Table 3-1) of the entire forecasting period.

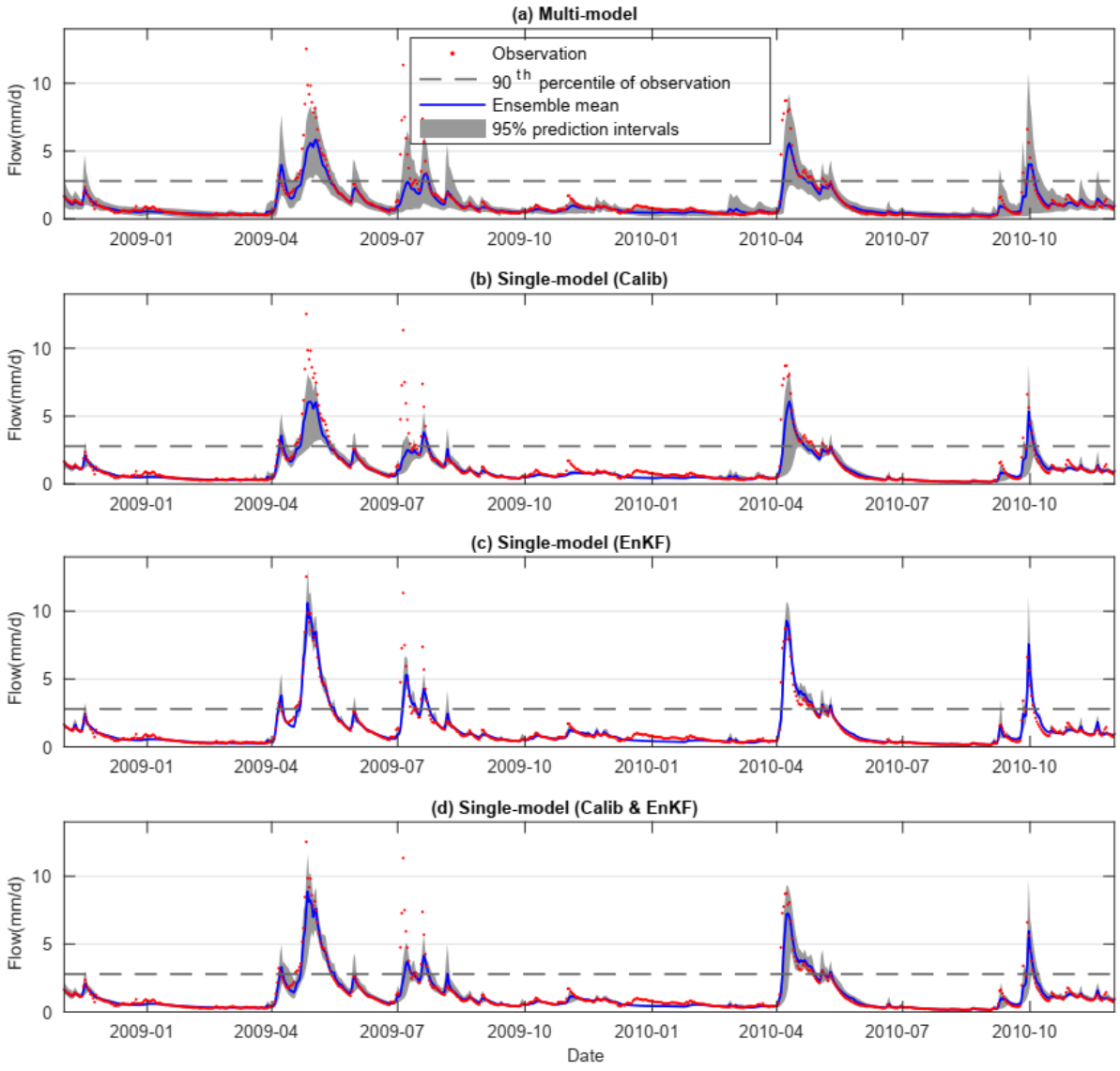

Figure A5-17. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Au Saumon catchment (the 18th catchment of Table 3-1) of the entire forecasting period.

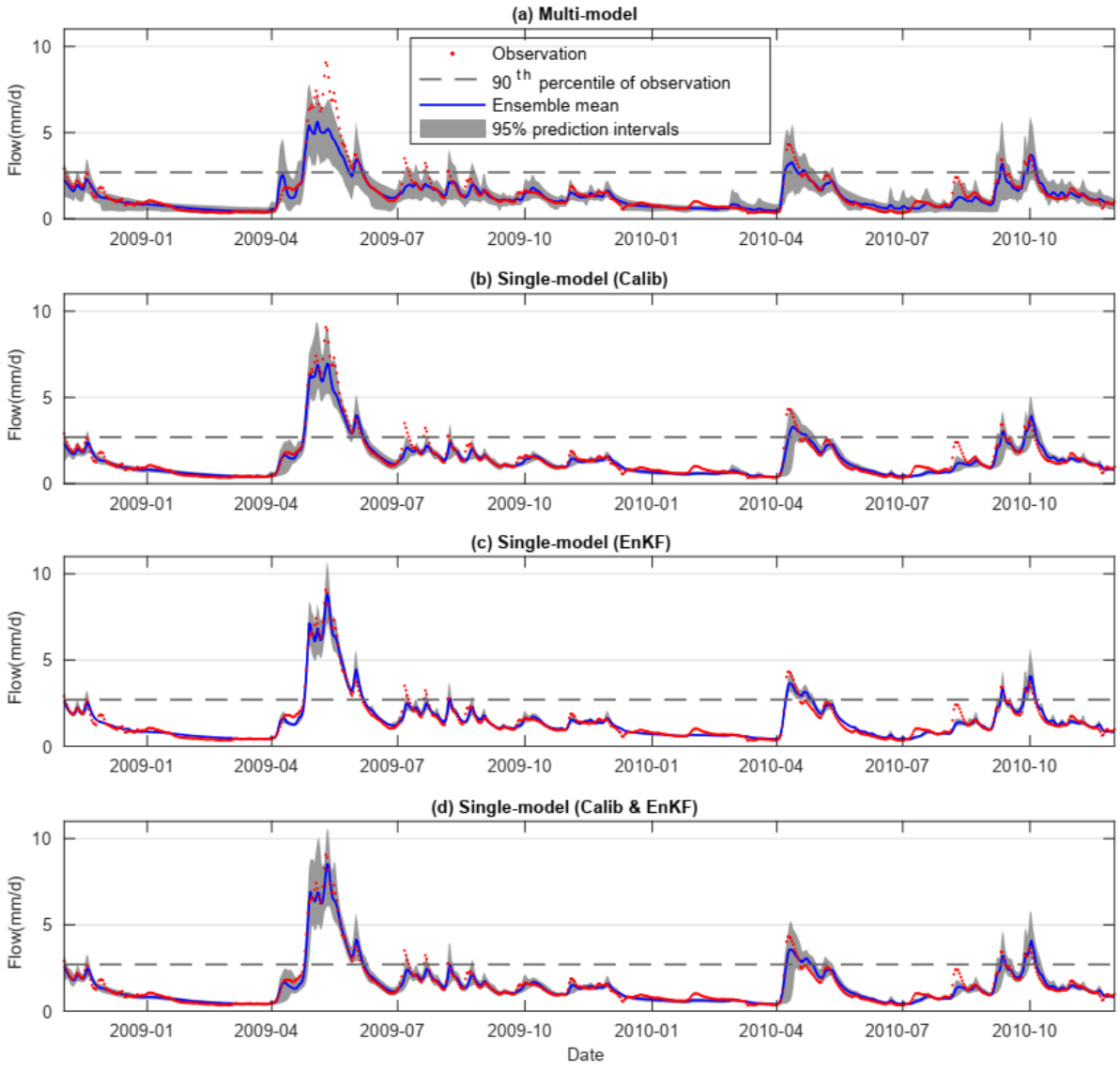

Figure A5-18. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Mistassini catchment (the 19th catchment of Table 3-1) of the entire forecasting period.

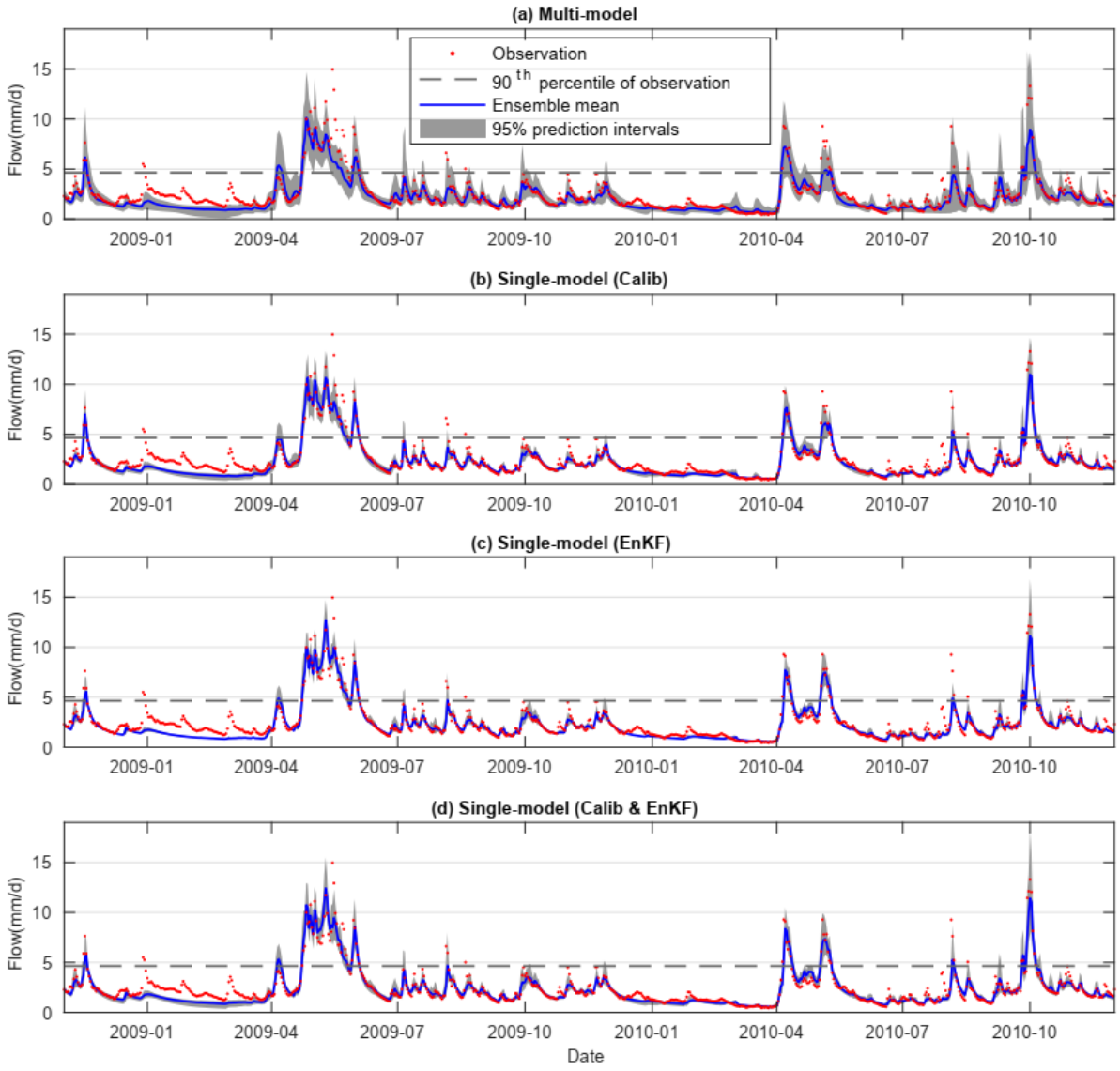

Figure A5-19. First lead day flow forecasts of the multi-model and three single-model forecasting systems for the Valin catchment (the 20th catchment of Table 3-1) of the entire forecasting period.