Probabilistic Ensemble-based Streamflow Forecasting Framework

PROBABILISTIC ENSEMBLE-BASED STREAMFLOW FORECASTING FRAMEWORK

By PEDRAM DARBANDSARI, M.Sc., B.Sc.

Faculty of Engineering

Civil Engineering

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

DOCTOR OF PHILOSOPHY

DOCTOR OF PHILOSOPHY (2021) McMaster University

Civil Engineering Hamilton, Canada

TITLE: Probabilistic Ensemble-based Streamflow Forecasting Framework

AUTHOR: Pedram Darbandsari

M.Sc. (University of Tehran)

B.Sc. (University of Tehran)

SUPERVISOR: Dr. Paulin Coulibaly

NUMBER OF PAGES: xxxi, 310

Lay Abstract

Possessing a reliable streamflow forecasting framework is of special importance in various fields of operational water resources management, non-structural flood mitigation in particular. Accurate and reliable streamflow forecasts lead to the best possible in-advanced flood control decisions which can significantly reduce its consequent loss of lives and properties. The main objective of this research is to develop an enhanced ensemble-based probabilistic streamflow forecasting approach through proper quantification of predictive uncertainty using an ensemble of streamflow forecasts. The key contributions are: (1) implementing multiple diverse forecasts with full coverage of future possibilities in the Bayesian ensemble-based forecasting method to produce more accurate and reliable forecasts; and (2) developing an ensemble-based Bayesian post-processing approach to enhance the hydrologic uncertainty quantification by taking the advantages of multiple forecasts and initial flow observation. The findings of this study are expected to benefit streamflow forecasting, flood control and mitigation, and water resources management and planning.

Abstract

Streamflow forecasting is a fundamental component of various water resources management systems, ranging from flood control and mitigation to long-term planning of irrigation and hydropower systems. In the context of floods, a probabilistic forecasting system is required for proper and effective decision-making. Therefore, the primary goal of this research is the development of an advanced ensemble-based streamflow forecasting framework to better quantify the predictive uncertainty and generate enhanced probabilistic forecasts. This research started by comprehensively evaluating the performances of various lumped conceptual models in data-poor watersheds and comparing various Bayesian Model Averaging (BMA) modifications for probabilistic streamflow simulation. Then, using the concept of BMA, two novel probabilistic post-processing approaches were developed to enhance streamflow forecasting performance. The combination of the entropy theory and the BMA method leads to an entropy-based Bayesian Model Averaging (En-BMA) approach for enhanced probabilistic streamflow and precipitation forecasting. Also, the integration of the Hydrologic Uncertainty Processor (HUP) and the BMA methods is proposed for probabilistic post-processing of multi-model streamflow forecasts.

Results indicated that the MACHBV and GR4J models are highly competent in simulating hydrological processes within data-scarce watersheds, however, the presence of the lower skill hydrologic models is still beneficial for ensemble-based streamflow forecasting. The comprehensive verification of the BMA approach in terms of streamflow predictions has identified the merits of implementing some of the previously recommended modifications

and showed the importance of possessing a mutually exclusive and collectively exhaustive ensemble. By targeting the remaining limitation of the BMA approach, the proposed En-BMA method can improve probabilistic streamflow forecasting, especially under high flow conditions. Also, the proposed HUP-BMA approach has taken advantage of both HUP and BMA methods to better quantify the hydrologic uncertainty. Moreover, the applicability of the modified En-BMA as a more robust post-processing approach for precipitation forecasting, compared to BMA, has been demonstrated.

Acknowledgements

First and foremost, I wish to express my deepest gratitude to my esteemed supervisor Prof. Paulin Coulibaly whose consistent and relentless supervision, guidance, support, and motivation have been invaluable throughout this study. His persistent encouragement and helpful advice enabled me to overcome difficulties over the past four years. Achieving this milestone in my life would have not been possible without his unparalleled help and guidance. I would like to extend my sincere thanks to the other Ph.D. supervisory committee members, Prof. Zoe Li and Prof. Yiping Guo for their constructive and insightful suggestions that have improved my research.

I would like to thank all my colleagues at McMaster Water Resources and Hydrologic Modeling Lab (WRHML) for their unwavering knowledge- and time-sharing. Their kind help and support have made my graduate life at McMaster University a wonderful time. I also specially thank Dr. James Leach for proofreading my third and fourth papers and Dr. Shasha Han for providing the source code for the HUP method.

I acknowledge the Ministry of Natural Resources and Forestry (MNRF), Environment and Climate Change Canada (ECCC), Water Survey of Canada (WSC), National Oceanic and Atmospheric Administration (NOAA), and Canadian Surface Prediction Archive (CaSPAr) for study data. I also gratefully acknowledge the financial and official supports of the Natural Science and Engineering Research Council of Canada (NSERC), the Department of Civil Engineering, and the school of Graduate Studies at McMaster University.

Last and Foremost, I owe a big thanks to all my family members whose support was a milestone in the completion of my Ph.D. degree. I am genuinely grateful to my wife Yasamin, for her unconditional love and tremendous encouragement. My deepest thank you goes to my parents for their immense emotional and financial support, dedication, and love over my whole life.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADA	Average Deviation Amplitude
ANN	Artificial Neural Network
ATIMR	Antecedent Temperature Index (ATI) Melt Rate
B, BW	Bandwidth
BE	Big East River watershed
BFS	Bayesian Forecasting System
BL	Black River watershed
BMA	Bayesian Model Averaging
CAFFEWS	Canadian Adaptive Flood Forecasting and Early Warning System
CaPA	Canadian Precipitation Analysis
CasPAr	Canadian Surface Prediction Archive
CDF	Cumulative Distribution Functions
CMC	Canadian Meteorological Centre
CR	Containing Ratio
CRPS	Continuous Ranked Probability Score
DD, DDM	Degree-Day snowmelt method
DDS	Dynamically Dimensioned Search algorithm
EC,ECCC	Environment and Climate Change Canada
EM	Expectation-Maximization algorithm
En-BMA	Entropy-based Bayesian Model Averaging
ENQT	Empirical Normal Quantile Transformation
EPS	Ensemble Prediction System
ESP	Ensemble Streamflow Prediction
F	Flow
GDPS	Global Deterministic Prediction System
GEFS	Global Ensemble Forecasting System Reforecast Project Version 2
GEPS	Global Ensemble Prediction System
GFS	Global Forecast System
GR4J	Génie Rural à 4 Paramètres Journaliers

HEC-HMS	Hydrologic Engineering Center's Hydrologic Modeling System
HRDPS	High-resolution Regional Deterministic Prediction System
HUP	Hydrologic Uncertainty Processor
HUP-BMA	HUP-embedded BMA
KGE	Kling Gupta Efficiency
MACHBV, MAC-HBV	McMaster University Hydrologiska Byrans Vattenbalansavdelning
MAE	Mean Absolute Error
MECE	Mutually Exclusive and Collectively Exhaustive
M-EnBMA	Modified Entropy-based BMA
M-M	Multi-Model
M-MI	Multi-Model Multi-Input
M-MIP	Multi-Model Multi-Input Multi-Parameter
M-MP	Multi-Model Multi-Parameter
MNVE	Modified Nash Volume Error
MSE	Modified Shapiro-Wilk statistic
NOAA	National Oceanic and Atmospheric Administration
NQT	Normal Quantile Transform
NSE	Nash Sutcliffe Efficiency
NSERC	Natural Science and Engineering Research Council of Canada
NVE	Nash Volume error
NWSRFS	National Weather Service River Forecast System
P	Precipitation
PCC	Pearson Correlation Coefficient
PDF	Probability Distribution Functions
PE	Peak Error
PET	Potential Evapotranspiration
PUP	Precipitation Uncertainty Processor
PWRMSE	Peak Weighted Root Mean Square Error
Q-Q	Quantile-Quantile
RDPS	Regional Deterministic Prediction System
REPS	Regional Ensemble Prediction System

RMSE	Root Mean Square Error
RSA	Regional Sensitivity Analysis
S17	Snow17 model
SACSMA, SAC-SMA	Sacramento soil moisture accounting
SHE	European hydrological system
SMARG	The Modified Version of the Soil Moisture and Accounting Routing
VE	Volume Error

List of Symbols

ADIMP	Additional impervious area (SACSMA model)
alpha1	An exponent in relation between outflow and storage (MACHBV model)
athorn	A constant for Thornthwaite's equation
beta	A non-linear parameter controlling runoff generation (MACHBV model)
C	Decay coefficient of soil evaporation (SMARG model)
C(.)	Total Correlation function
cperc	Constant Percolation rate parameter (MACHBV model)
DDF	Degree-day factor (DDM model)
E(.)	Expected value function
eCDF(.)	Empirical Cumulative Distribution Function
F, f	Forecasted streamflow/precipitation time series
f'	The cubic root of forecasted precipitation time series
fc	Maximum for soil water content (MACHBV model)
f_n	The likelihood of model river discharge at lead time n
f_{Qn}	The likelihood of model river discharge in normal space at lead time n
G	The groundwater runoff coefficient (SMARG model)
g_n	Prior density of actual river discharge at lead time n
g_{Qn}	Prior density of actual river discharge in normal space at lead time n
H	Direct runoff factor (SMARG model)
H(.)	The marginal entropy function The Heaviside function
I[.]	General indicator function
K	Number of ensemble member
k0	Flow recession coefficient in an upper soil reservoir (MACHBV model)
k1	Flow recession coefficient in an upper soil reservoir (MACHBV model)
k2	Flow recession coefficient in a lower soil reservoir (MACHBV model)
Kg	Number of time step for groundwater routing (SMARG model)
L(.)	Log-likelihood function
lp/fc	Limit for potential evapotranspiration (MACHBV model)
lsuz	A threshold value used to control response routing on an upper soil reservoir (MACHBV model)

LZFPM	Lower-zone free water primary maximum storage (SACSMA model)
LZFSM	Lower-zone free water supplemental maximum (SACSMA model)
LZPK	Lower-zone primary free water lateral depletion rate (SACSMA model)
LZSK	Lower-zone supplemental free water lateral depletion rate (SACSMA model)
LZTWM	Lower-zone tension water maximum storage (SACSMA model)
M	The ensemble of hydrologic models (vector)
maxbas	Parameter of a triangle weighting function (MACHBV model)
mfmax	Maximum melt factor (SNOW17 model)
mfmin	Minimum melt factor (SNOW17 model)
N	The length of time series Number of variables The length of forecasting horizon The Nash Cascade model parameter (SMARG model)
N(.),Q(.)	The Normal distribution function
NK	Number of time step for surface runoff routing (SMARG model)
nmf	Maximum negative melt factor (SNOW17 model)
N_{Qin}, N_{Oin}	Number of data falls within the 95% confidence bound
NSE90	NSE based on flows more than 90 percentile
NSEL, NSE_{log}	NSE based on logarithmic transformed data
NSES, NSE_{sqrt}	NSE based on squared transformed data
PCTIM	Impervious fraction of the watershed area (SACSMA model)
P_f	Deterministic precipitation forecast time series
PFREE	Fraction percolating from upper to lower zone free water storage (SACSMA model)
plwhc	The water holding capacity of the snow pack (SNOW17 model)
q(.)	The standard normal density function
$Q^{-1}(.)$	Inverse of normal standard distribution function
q_l, q^l, f_l	The lower bound of the 95% confidence interval
Q_O, O	Observed streamflow/precipitation time series
Q_{O90}	Observed streamflow more than 90 percentile time series
Q_{Omax}	Annual maximum observed peak flow time series
Q_S	Simulated streamflow time series
Q_{S90}	Simulated streamflow more than 90 percentile time series

Q_{Smax}	Annual maximum simulated peak flow time series
q_u, q^u f_u	The upper bound of the 95% confidence interval time series
RCF	Rain correction factor (model)
REXP	Exponent of the percolation equation (SACSMA model)
RIVA	Riparian vegetation area (SACSMA model)
Rq	Routing coefficient (SACSMA model)
r_{Qn}	Transition density of actual river discharge at lead time n in normal space
RSERV	Fraction of lower zone free water not transferable to tension water (SACSMA model)
SCF	Snow correction factor (DDM model)
S_F	The optimal subset of ensemble of forecasts matrix (En-BMA method)
SIDE	Ratio of deep recharge to channel base flow (SACSMA model)
T	Conversion parameter for calculating potential evaporation (SMARG model) Air temperature (SNOW17 and DDM model)
T(.)	Transinformation (mutual information) function
T_f	Deterministic temperature forecast time series
T_m	Melting temperature (DDM model) Coefficient of transferability
T_r	Upper threshold temperatures (DDM model)
T_S	Lower threshold temperatures (DDM model)
uadj	The average wind function (SNOW17 model)
UZFWM	Upper-zone free water maximum storage (SACSMA model)
UZK	Upper-zone free water lateral depletion rate (SACSMA model)
UZTWM	Upper-zone tension water maximum storage (SACSMA model)
v	Precipitation indicator
w	The weight vector or matrix (BMA method)
x1	Maximum capacity of the production store (GR4J model)
x2	The groundwater exchange coefficient (GR4J model)
x3	Maximum capacity of routing store (GR4J model)
x4	The unit hydrograph time base (GR4J model)
X_n	Actual river discharge at lead time n in normal space (random variable)
x_n	Actual river discharge at lead time n in normal space (a realization)
$\hat{X}_n$	Model river discharge at lead time n in normal space (random variable)
$\hat{x}_n$	Model river discharge at lead time n in normal space (a realization)

Y	Infiltration capacity (SMARG model)
Y_n	Actual river discharge at lead time n (random variable)
y_n	Actual river discharge at lead time n (a realization)
$\hat{Y}_n$	Model river discharge at lead time n (random variable)
$\hat{y}_n$	Model river discharge at lead time n (a realization)
Y^f, F	Ensemble of forecasts/simulations matrix
Z	The total depth of all soil layers (SMARG model)
$ZPERC$	Maximum percolation rate (SACSMA model)
α	the discrepancy between Q-Q plot and the bisector line
β	The stopping threshold parameter (En-BMA method)
$\bar{\Lambda}_n$	Marginal distributions of model river discharge at lead time n
Γ_n	Marginal distributions of actual river discharge at lead time n
δ_i	The seconder predictor variable in logistic regression
θ	Vector of parameters
κ_n	Expected density of model river discharge conditioned on initial observation
λ	The Box–Cox coefficient
μ	The mean of the gamma distribution
σ,τ	Standard deviation
Φ_n	Posterior distribution of actual river discharge at lead time n
φ_{Qn}	Posterior density of actual river discharge in normal space at lead time n

Declaration of Academic Achievement

This thesis was prepared in a sandwich style following the regulations provided by the School of Graduate Studies at McMaster University. It includes the published and submitted papers listed below:

Chapter 2: Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: Case study of Northern Ontario, Canada, by P. Darbandsari and P.Coulibaly, Journal of Hydrology: Regional Studies, 31, 100730, doi: 10.1016/j.ejrh.2020.100730, 2020 (With permission from the publisher)

Chapter 3: Inter-comparison of different Bayesian model averaging modifications in streamflow simulation, by P. Darbandsari and P. Coulibaly, Water, 11(8), 1707, doi: 10.3390/w11081707, 2019 (With permission from the publisher)

Chapter 4: Introducing entropy-based Bayesian model averaging for streamflow forecast, by P. Darbandsari and P. Coulibaly, Journal of Hydrology, 11(8), 1707, doi: 10.1016/j.jhydrol.2020.125577, 2020 (With permission from the publisher)

Chapter 5: An Integration of Hydrologic Uncertainty Processor and Bayesian Model Averaging for Streamflow Forecasting, by P. Darbandsari and P. Coulibaly, Water Resources Research, under review, manuscript number 2020WR029433.

Chapter 6: Assessing Entropy-based Bayesian Model Averaging Method for Probabilistic Precipitation Forecasting, by P. Darbandsari and P. Coulibaly, Journal of Hydrometeorology, under review, manuscript number JHM-D-21-0086.

For Chapters 2 and 3, P. Darbandsari conducted the modeling and computational works under the guidance and supervision of Prof. P. Coulibaly. P. Darbandsari wrote both manuscripts, and Prof. P. Coulibaly reviewed and edited them. Chapter 2 was published in Journal of Hydrology, Regional Studies in 2020, and Chapter 3 was published in Water journal in 2020. Chapters 4 and 5 were conceptualized by Prof. P. Coulibaly and P. Darbandsari. P. Darbandsari conducted the programming, formal analysis, and computational works under the direct guidance and supervision of Prof. P. Coulibaly. P. Darbandsari wrote both papers and Prof. P. Coulibaly reviewed and edited them. Chapter 4 was published in Journal of Hydrology in 2020, and Chapter 5 was submitted to the Water Resources Research journal in 2020. For Chapter 6, P. Darbandsari conducted the modeling and formal analysis with the guidance of Prof. P. Coulibaly. P. Darbandsari wrote the manuscript, and Prof. P. Coulibaly reviewed and edited it. The paper was submitted to the Journal of Hydrometeorology in 2021. The work reported here was undertaken from January 2017 to May 2021.

Chapter 1. Introduction

1.1 Background

Accurate and reliable streamflow forecasting is receiving particular importance in various fields of water resources management, flood control and mitigation in particular. Compared with other natural disasters, flood is the most common natural hazard in Canada leading to catastrophic environmental and socio-economic damages (Caballero & Rahman, 2014; Thistlethwaite et al., 2018). Dramatically increasing frequency of extreme events in the recent decades, mostly caused by climate change, has brought more attention to flood mitigation measures (Han & Coulibaly, 2017; Reggiani et al., 2009). In general, these measures are categorized into two groups: structural and non-structural. The structural measures tried to reduce the negative flood effects by changing the characteristics of the landscapes, such as constructing flood-control reservoirs and diversions (Heidari, 2009; Meyer et al., 2012), while the non-structural interventions are more sustainable, less expensive, and the only effective way for protecting life and property from floods in many flood-prone regions (Barbetta et al., 2017; Kundzewicz, 2002). One of the most effective non-structural flood mitigation measures is the application of a reliable flood forecasting system, where hydrologic and hydraulic models are used for flood predictions. Although some regional streamflow prediction frameworks have been developed and used in Canada (Zahmatkesh et al., 2019), there is currently no nationwide flood forecasting and early warning system.

One of the integral components of any operational streamflow prediction framework is rainfall-runoff models (or hydrologic models) which are used as simplified characterizations of different hydrologic processes (such as snowmelt, infiltration, evapotranspiration, runoff, etc.). Thanks to the development of various types of hydrologic models in the last century, flood forecasting has improved significantly. Classifying rainfall-runoff models based on their mathematical representations leads to three main groups. In the category of conceptual models, such as Sacramento soil moisture accounting (SAC-SMA) (Burnash et al., 1973), McMaster University Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) (Samuel et al., 2011), and Hydrologic Engineering Center's Hydrologic Modeling System (HEC-HMS) (Scharffenberg, 2016), interconnected conceptual elements are used for representing different hydrologic components. These models are popular for flood forecasting due to their low computational cost and simplicity. The low required input data makes the conceptual models the best option for data-scare regions (Anshuman et al., 2019; Tegegne et al., 2017). On the other hand, by focusing on physical characteristics of the hydrologic processes in time and space, the physically-based distributed models, such as the European hydrological system (SHE) (Abbott et al., 1986a, 1986b), have been developed. Reliable practical application of the physically-based models requires a large amount of data for the proper estimation of the spatially distributed parameters representing the physical properties of the watershed (Todini, 2011; Young, 2002). Moreover, the black-box models (also known as data-driven models), ranging from simple linear models to Artificial Neural Network (ANN), analyze the relationships between inputs (e.g. temperature and precipitation) and the output of interest (e.g.

streamflow or water level), without consideration of the catchment physical processes. The lack of physical interpretation and the strong dependence on the calibration data are the main concerns of the data-driven models (Shrestha, 2009; Todini, 2011). Although much progress has been made to improve the capability of different hydrologic models, there are still lots of simplifications in their structures. Also, each model has parameters that cannot be perfectly estimated due to errors in historical data. So, no hydrological model can provide error-free streamflow prediction in all conditions (Chen et al., 2013).

Streamflow forecasting is subject to various sources of uncertainty. Besides the inaccurate future meteorological forecasts, the imperfection of the hydrologic models, the uncertain parameters estimation, and the unknown initial conditions are the other important sources of uncertainties in the case of streamflow prediction (Moradkhani & Sorooshian, 2008). Decision-making based on a single model deterministic forecast, which only provides a point estimation of the future value without taking into account the inherent uncertainties, is very risky and can lead to irreversible economic and social damages (Liu et al., 2018). So, generating probabilistic forecasts by quantifying and reducing the predictive uncertainty is one of the most important parts of any operational flood forecasting framework (Biondi & Todini, 2018). Predictive uncertainty is defined as the conditional distribution of future unknown values based on the information provided by the forecasting model(s) (Todini, 2011). Using an ensemble of streamflow predictions (ESP) is one of the most widely used approaches for quantifying the predictive uncertainty (Madadgar & Moradkhani, 2014; Michaels, 2015). As a conventional approach, the ESP was generated by forcing a hydrologic model with multiple meteorological forecasts (Abaza et al., 2013;

Baran et al., 2019; Fan et al., 2014; Qu et al., 2017) which reflects the uncertainties associated with forcing inputs. As an alternative and evolving way, multiple hydrologic models can be used for generating ESP to capture the model structural uncertainties (Jiang et al., 2018; Seiller et al., 2017; Todini, 2008; Xu et al., 2019). Also, using various realizations of the model's parameters or assimilating initial states could be used for constructing ESP for conceptualizing uncertainties associated with measurement errors and parameters estimation process (Dong et al., 2013; Pappenberger et al., 2005; Parrish et al., 2012).

Although, using multiple forecasts, compared to the deterministic one, provides more information about the future event, it is still required to estimate the correct and reliable predictive uncertainty for sound and proper decision making (Biondi & Todini, 2018; J. Liu & Xie, 2014; Reggiani & Weerts, 2008). Therefore, a statistical post-processing approach, which is used to reduce and quantify the predictive uncertainty, is a crucial component of any operational forecasting system. A recent review of the various post-processing approaches can be found in Han and Coulibaly (2017) and Li et al., (2017). In general, the post-processing methods tried to use the full capability of all available information to characterize the predictive uncertainty and generate reliable probabilistic forecasts. Some of these methods, such as the well-known Hydrologic Uncertainty Processor (HUP) (Krzysztofowicz & Kelly, 2000), use single deterministic forecasts for generating predictive results (e.g. Krzysztofowicz & Herr, 2001; Krzysztofowicz & Kelly, 2000; Liu et al., 2018; Montanari & Grossi, 2008), while some have been extended to take

the advantage of considering ensemble streamflow forecasts (e.g. Han and Coulibaly 2019; Khajehei & Moradkhani, 2017; Raftery et al., 2005; Reggiani et al., 2009; Seo et al., 2006).

Among the latter methods, the Bayesian Model Averaging (BMA) approach (Hoeting et al., 1999; Raftery et al., 1997, 2005) is the most common and widely used ensemble-based post-processing method for both meteorological (e.g. Cane et al., 2013; Ji et al., 2019; J. Liu & Xie, 2014; Ma et al., 2018; Vrugt et al., 2006) and hydrological (Duan et al., 2007; Jiang et al., 2018; Liang et al., 2013; Sharma et al., 2019) forecasts. By using the weighted average of the conditional distribution of the predictand based on each ensemble forecast member, BMA estimates the forecast predictive uncertainty conditioned on the whole ensemble. Although different modifications have been proposed to enhance the BMA capability of quantifying predictive uncertainty in the case of dealing with different hydrometeorological variables (e.g. Soughter et al. (2007), Fraley et al. (2010), and Yang et al. (2012) for precipitation and Yan and Moradkhani (2016), Madadgar and Moradkhani, (2014), Vrugt and Robinson (2007), and He et al. (2018) for streamflow), there are still some inherent assumptions and limitations in the BMA structure which requires further research to develop a promising probabilistic flood forecasting framework. Besides the quality of individual members of the forecast ensemble, the characteristics of the whole ensemble play an important role in the reliable performance of the BMA method. Based on the law of total probability, as the main assumption of the BMA approach, the two properties of capturing the whole future variability as well as possessing mutually independent members of the ensemble are required for generating reliable BMA based probabilistic forecasts (Madadgar & Moradkhani, 2014; Refsgaard et al., 2012; Sharma et

al., 2019). Furthermore, the existing limitations in different methods motivate some research to integrate various techniques for devising a more reliable forecasting approach. In the context of Bayesian Model Averaging, Sharma et al. (2019), for instance, combines the Quantile Regression and BMA methods respectively for bias-correcting the ensemble of daily streamflow forecasts and probabilistically merging them. Another example is the employment of the data assimilation technique (i.e. Particle Filter) within the BMA structure to better quantify the predictive uncertainty (Parrish et al., 2012). Also, Ajami et al. (2007), Yen et al. (2014), and Jiang et al. (2018) developed BMA based methods for multi-source uncertainty analysis in an integrated manner, however, they have not been used for operational streamflow forecasting.

1.2 Research Objectives and Thesis Outline

Using the concept of the Bayesian Model Averaging (BMA) approach and aiming at enhancing the assessment of predictive uncertainty through utilizing an ensemble of streamflow forecasts, this study focuses on developing a reliable ensemble-based probabilistic streamflow forecasting/simulation framework. To achieve the primary objective, the following secondary objectives have been accomplished, which leads to five journal papers presented in Chapters 2 to 6 of the thesis:

- A literature review on the Bayesian Model Averaging (BMA) concepts and their applications for streamflow forecasting.
- Developing, investigating, and selecting the appropriate hydrologic models.

- Examining the current variants of the BMA method for streamflow forecasting/simulation.
- Proposing Entropy-based BMA method for enhanced probabilistic ensemble streamflow forecasting.
- Developing multi-model Bayesian post-processor for probabilistic streamflow forecasting through integrating the Hydrologic Uncertainty Processor (HUP) and the Bayesian Model Averaging (BMA) approaches.
- Modifying and assessing the Entropy-based BMA method for post-processing precipitation forecasts.

The thesis is organized into six chapters. After presenting an overview of the background, scope, and objectives of the research in Chapter 1, Chapter 2 comprehensively compares the performances of different conceptual hydrologic models for streamflow simulation in snow-dominated data-poor watersheds and evaluates the reliability of the archived Canadian Precipitation Analysis (CaPA) as an alternative forcing input of rainfall-runoff models in the case of sparse ground-based meteorological measurements. Chapter 3 thoroughly assesses the effects of various previously recommended modifications of the Bayesian Model Averaging approach on the quality of the final BMA-derived probabilistic streamflow predictions. In Chapter 4, by using the concepts of the Entropy theory, the new Entropy-based Bayesian Model Averaging (En-BMA) technique has been developed in order to generate more accurate and reliable streamflow forecasts. Chapter 5 presents a new ensemble-based Bayesian post-processing approach where a combination of Hydrologic Uncertainty Processor and Bayesian Model Averaging methods is used for

better quantifying hydrologic uncertainty and providing more reliable streamflow forecasts. In chapter 6, by taking a step forward, a modified version of En-BMA approach was proposed and evaluated for post-processing an ensemble of precipitation forecasts.

1.3 References

Abaza, M., Anctil, F., Fortin, V., & Turcotte, R. (2013). A Comparison of the Canadian Global and Regional Meteorological Ensemble Prediction Systems for Short-Term Hydrological Forecasting. *Monthly Weather Review*, *141*(10), 3462–3476. https://doi.org/10.1175/MWR-D-12-00206.1

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E., & Rasmussen, J. (1986a). An introduction to the European Hydrological System — Systeme Hydrologique Europeen, "SHE", 1: History and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, *87*(1), 45–59. https://doi.org/10.1016/0022-1694(86)90114-9

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E., & Rasmussen, J. (1986b). An introduction to the European Hydrological System — Systeme Hydrologique Europeen, "SHE", 2: Structure of a physically-based, distributed modelling system. *Journal of Hydrology*, *87*(1), 61–77. https://doi.org/10.1016/0022-1694(86)90115-0

Ajami, N. K., Duan, Q., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004745

Anshuman, A., Kunnath-Poovakka, A., & Eldho, T. I. (2019). Towards the use of conceptual models for water resource assessment in Indian tropical watersheds under monsoon-driven climatic conditions. *Environmental Earth Sciences*, *78*(9), 282. https://doi.org/10.1007/s12665-019-8281-5

Baran, S., Hemri, S., & Ayari, M. E. (2019). Statistical Postprocessing of Water Level Forecasts Using Bayesian Model Averaging With Doubly Truncated Normal Components. *Water Resources Research*, *55*(5), 3997–4013. https://doi.org/10.1029/2018WR024028

Barbetta, S., Coccia, G., Moramarco, T., Brocca, L., & Todini, E. (2017). The multi temporal/multi-model approach to predictive uncertainty assessment in real-time flood forecasting. *Journal of Hydrology*, *551*, 555–576. https://doi.org/10.1016/j.jhydrol.2017.06.030

Biondi, D., & Todini, E. (2018). Comparing Hydrological Postprocessors Including Ensemble Predictions Into Full Predictive Probability Distribution of Streamflow. *Water Resources Research*, *54*(12), 9860–9882. https://doi.org/10.1029/2017WR022432

Burnash, R. J. C., Ferral, R. L., & McGuire, R. A. (1973). *A generalized streamflow simulation system: Conceptual modeling for digital computers*. Joint Federal-State River Forecast Center, United States National Weather Service.

Caballero, W. L., & Rahman, A. (2014). Development of regionalized joint probability approach to flood estimation: A case study for Eastern New South Wales, Australia. *Hydrological Processes*, *28*(13), 4001–4010. https://doi.org/10.1002/hyp.9919

Cane, D., Ghigo, S., Rabuffetti, D., & Milelli, M. (2013). Real-time flood forecasting coupling different postprocessing techniques of precipitation forecast ensembles with a distributed hydrological model. The case study of may 2008 flood in western Piemonte, Italy. *Natural Hazards and Earth System Sciences*, *13*(2), 211–220. https://doi.org/10.5194/nhess-13-211-2013

Chen, X., Yang, T., Wang, X., Xu, C.-Y., & Yu, Z. (2013). Uncertainty Intercomparison of Different Hydrological Models in Simulating Extreme Flows. *Water Resources Management*, *27*(5), 1393–1409. https://doi.org/10.1007/s11269-012-0244-5

Dong, L., Xiong, L., & Zheng, Y. (2013). Uncertainty analysis of coupling multiple hydrologic models and multiple objective functions in Han River, China. *Water Science and Technology*, *68*(3), 506–513. https://doi.org/10.2166/wst.2013.255

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Fan, F. M., Collischonn, W., Meller, A., & Botelho, L. C. M. (2014). Ensemble streamflow forecasting experiments in a tropical basin: The São Francisco river case study. *Journal of Hydrology*, *519*, 2906–2919. https://doi.org/10.1016/j.jhydrol.2014.04.038

Fraley, C., Raftery, A. E., & Gneiting, T. (2010). Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging. *Monthly Weather Review*, *138*(1), 190–202. https://doi.org/10.1175/2009MWR3046.1

Han, S., & Coulibaly, P. (2017). Bayesian flood forecasting methods: A review. *Journal of Hydrology*, *551*, 340–351. https://doi.org/10.1016/j.jhydrol.2017.06.004

He, S., Guo, S., Liu, Z., Yin, J., Chen, K., & Wu, X. (2018). Uncertainty analysis of hydrological multi-model ensembles based on CBP-BMA method. *Hydrology Research*, *49*(5), 1636–1651. https://doi.org/10.2166/nh.2018.160

Heidari, A. (2009). Structural master plan of flood mitigation measures. *Natural Hazards and Earth System Sciences*, *9*(1), 61–75. https://doi.org/10.5194/nhess-9-61-2009

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, *14*(4), 382–401. JSTOR.

Ji, L., Zhi, X., Zhu, S., & Fraedrich, K. (2019). Probabilistic Precipitation Forecasting over East Asia Using Bayesian Model Averaging. *Weather and Forecasting*, *34*(2), 377–392. https://doi.org/10.1175/WAF-D-18-0093.1

Jiang, S., Ren, L., Xu, C.-Y., Liu, S., Yuan, F., & Yang, X. (2018). Quantifying multi-source uncertainties in multi-model predictions using the Bayesian model averaging scheme. *Hydrology Research*, *49*(3), 954–970. https://doi.org/10.2166/nh.2017.272

Khajehei, S., & Moradkhani, H. (2017). Towards an improved ensemble precipitation forecast: A probabilistic post-processing approach. *Journal of Hydrology*, *546*, 476–489. https://doi.org/10.1016/j.jhydrol.2017.01.026

Krzysztofowicz, R., & Herr, H. D. (2001). Hydrologic uncertainty processor for probabilistic river stage forecasting: Precipitation-dependent model. *Journal of Hydrology*, *249*(1), 46–68. https://doi.org/10.1016/S0022-1694(01)00412-7

Krzysztofowicz, R., & Kelly, K. S. (2000). Hydrologic uncertainty processor for probabilistic river stage forecasting. *Water Resources Research*, *36*(11), 3265–3277. https://doi.org/10.1029/2000WR900108

Kundzewicz, Z. W. (2002). Non-structural Flood Protection and Sustainability. *Water International*, *27*(1), 3–13. https://doi.org/10.1080/02508060208686972

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., & Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *WIREs Water*, *4*(6), e1246. https://doi.org/10.1002/wat2.1246

Liang, Z., Wang, D., Guo, Y., Zhang, Y., & Dai, R. (2013). Application of Bayesian Model Averaging Approach to Multimodel Ensemble Hydrologic Forecasting. *Journal of Hydrologic Engineering*, *18*(11), 1426–1436. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000493

Liu, J., & Xie, Z. (2014). BMA Probabilistic Quantitative Precipitation Forecasting over the Huaihe Basin Using TIGGE Multimodel Ensemble Forecasts. *Monthly Weather Review*, *142*(4), 1542–1555. https://doi.org/10.1175/MWR-D-13-00031.1

Liu, Z., Guo, S., Xiong, L., & Xu, C.-Y. (2018). Hydrological uncertainty processor based on a copula function. *Hydrological Sciences Journal*, *63*(1), 74–86. https://doi.org/10.1080/02626667.2017.1410278

Ma, Y., Hong, Y., Chen, Y., Yang, Y., Tang, G., Yao, Y., Long, D., Li, C., Han, Z., & Liu, R. (2018). Performance of Optimally Merged Multisatellite Precipitation Products Using the Dynamic Bayesian Model Averaging Scheme Over the Tibetan Plateau. *Journal of Geophysical Research: Atmospheres*, *123*(2), 814–834. https://doi.org/10.1002/2017JD026648

Madadgar, S., & Moradkhani, H. (2014). Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resources Research*, *50*(12), 9586–9603. https://doi.org/10.1002/2014WR015965

Meyer, V., Priest, S., & Kuhlicke, C. (2012). Economic evaluation of structural and non-structural flood risk management measures: Examples from the Mulde River. *Natural Hazards: Journal of the International Society for the Prevention and Mitigation of Natural Hazards*, *62*(2), 301–324.

Michaels, S. (2015). Probabilistic forecasting and the reshaping of flood risk management. *Journal of Natural Resources Policy Research*, *7*(1), 41–51. https://doi.org/10.1080/19390459.2014.970800

Montanari, A., & Grossi, G. (2008). Estimating the uncertainty of hydrological forecasts: A statistical approach. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2008WR006897

Moradkhani, H., & Sorooshian, S. (2008). General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis. In *Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrological Models* (pp. 1–24). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77843-1_1

Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J., & de Roo, A. P. J. (2005). Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrology and Earth System Sciences*, *9*(4), 381–393. https://doi.org/10.5194/hess-9-381-2005

Parrish, M. A., Moradkhani, H., & DeChant, C. M. (2012). Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation: TOWARD REDUCTION OF MODEL UNCERTAINTY. *Water Resources Research*, *48*(3). https://doi.org/10.1029/2011WR011116

Qu, B., Zhang, X., Pappenberger, F., Zhang, T., & Fang, Y. (2017). Multi-Model Grand Ensemble Hydrologic Forecasting in the Fu River Basin Using Bayesian Model Averaging. *Water*, *9*(2), 74. https://doi.org/10.3390/w9020074

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, *92*(437), 179–191. https://doi.org/10.1080/01621459.1997.10473615

Refsgaard, J. C., Christensen, S., Sonnenborg, T. O., Seifert, D., Højberg, A. L., & Troldborg, L. (2012). Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Advances in Water Resources*, *36*, 36–50. https://doi.org/10.1016/j.advwatres.2011.04.006

Reggiani, P., Renner, M., Weerts, A. H., & Gelder, P. A. H. J. M. van. (2009). Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. *Water Resources Research*, *45*(2). https://doi.org/10.1029/2007WR006758

Reggiani, P., & Weerts, A. H. (2008). A Bayesian approach to decision-making under uncertainty: An application to real-time forecasting in the river Rhine. *Journal of Hydrology*, *356*(1), 56–69. https://doi.org/10.1016/j.jhydrol.2008.03.027

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2011). Estimation of Continuous Streamflow in Ontario Ungauged Basins: Comparison of Regionalization Methods. *Journal of*

Hydrologic Engineering, *16*(5), 447–459. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000338

Scharffenberg, W. (2016). *HEC-HMS User's Manual, Version 4.2*. U.S. Army Corps of Engineers Institute for Water Resources Hydrologic Engineering Center (CEIWR-HEC).

Seiller, G., Roy, R., & Anctil, F. (2017). Influence of three common calibration metrics on the diagnosis of climate change impacts on water resources. *Journal of Hydrology*, *547*, 280–295. https://doi.org/10.1016/j.jhydrol.2017.02.004

Seo, D.-J., Herr, H. D., & Schaake, J. C. (2006). A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrology and Earth System Sciences Discussions*, *3*(4), 1987–2035. https://doi.org/10.5194/hessd-3-1987-2006

Sharma, S., Siddique, R., Reed, S., Ahnert, P., & Mejia, A. (2019). Hydrological Model Diversity Enhances Streamflow Forecast Skill at Short- to Medium-Range Timescales. *Water Resources Research*, *55*(2), 1510–1530. https://doi.org/10.1029/2018WR023197

Shrestha, D. L. (2009). *Uncertainty analysis in rainfall-runoff modelling - application of machine learning techniques: UNESCO-IHE PhD thesis*. [PhD. thesis, IHE Delft Institute for Water Education]. https://www.cabdirect.org/cabdirect/abstract/20123116250

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., & Fraley, C. (2007). Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, *135*(9), 3209–3220. https://doi.org/10.1175/MWR3441.1

Tegegne, G., Park, D. K., & Kim, Y.-O. (2017). Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. *Journal of Hydrology: Regional Studies*, *14*, 49–66. https://doi.org/10.1016/j.ejrh.2017.10.002

Thistlethwaite, J., Minano, A., Blake, J. A., Henstra, D., & Scott, D. (2018). Application of re/insurance models to estimate increases in flood risk due to climate change. *Geoenvironmental Disasters*, *5*(1), 8. https://doi.org/10.1186/s40677-018-0101-9

Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, *6*(2), 123–137. https://doi.org/10.1080/15715124.2008.9635342

Todini, E. (2011). History and perspectives of hydrological catchment modelling. *Hydrology Research*, *42*(2–3), 73–85. https://doi.org/10.2166/nh.2011.096

Vrugt, J. A., Gupta, H. V., Nualláin, B., & Bouten, W. (2006). Real-Time Data Assimilation for Operational Ensemble Streamflow Forecasting. *Journal of Hydrometeorology*, *7*(3), 548–565. https://doi.org/10.1175/JHM504.1

Vrugt, J. A., & Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004838

Xu, J., Anctil, F., & Boucher, M.-A. (2019). Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems. *Journal of Hydrology*, *578*, 124002. https://doi.org/10.1016/j.jhydrol.2019.124002

Yan, H., & Moradkhani, H. (2016). Toward more robust extreme flood prediction by Bayesian hierarchical and multimodeling. *Natural Hazards*, *81*(1), 203–225. https://doi.org/10.1007/s11069-015-2070-6

Yang, C., Yan, Z., & Shao, Y. (2012). Probabilistic precipitation forecasting based on ensemble output using generalized additive models and Bayesian model averaging. *Acta Meteorologica Sinica*, *26*(1), 1–12. https://doi.org/10.1007/s13351-012-0101-8

Yen, H., Wang, X., Fontane, D. G., Harmel, R. D., & Arabi, M. (2014). A framework for propagation of uncertainty contributed by parameterization, input data, model structure, and calibration/validation data in watershed modeling. *Environmental Modelling & Software*, *54*, 211–221. https://doi.org/10.1016/j.envsoft.2014.01.004

Young, P. C. (2002). Advances in real-time flood forecasting. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *360*(1796), 1433–1450. https://doi.org/10.1098/rsta.2002.1008

Zahmatkesh, Z., Jha, S. K., Coulibaly, P., & Stadnyk, T. (2019). An overview of river flood forecasting procedures in Canadian watersheds. *Canadian Water Resources Journal / Revue Canadienne Des Ressources Hydriques*, *44*(3), 213–229. https://doi.org/10.1080/07011784.2019.1601598

Chapter 2. Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: Case study of Northern Ontario, Canada

Summary of Paper 1: Darbandsari, P., & Coulibaly, P. (2020). Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: Case study of Northern Ontario, Canada. Journal of Hydrology: Regional Studies, 31, 100730.

By considering the effects of calibration process and multiple precipitation input scenarios on the models' performance, the main goal of this research is it to comprehensively evaluate and compare various conceptual rainfall-runoff models with distinct structures for daily stream flow prediction in snow-dominated watersheds with low data availability. Also, the implementation of the archive Canadian Precipitation Analysis (CaPA) as an alternative forcing input of the hydrologic models in Northern Ontario, Canada is evaluated.

Key findings of this research are as follows:

- The necessity of considering the effects of calibration process in any model comparison study is revealed.
- The MACHBV hydrologic model shows the most consistent results in daily streamflow simulations, and the GR4J and SACSMA models possess competitive performances.
- The GR4J model outperformed the other six models for high flow prediction.

- The HEC-HMS based models possess the relatively lower performances for daily streamflow simulation.
- The aggregated daily archived CaPA precipitation data is a reliable alternative for hydrologic modelling in data-poor watersheds.
- The effectiveness of using a more complex snowmelt estimation routine depends on the structure of the conceptual rainfall-runoff model.

2.1 Abstract

Study Region: Big East River and Black River watersheds in Northern Ontario, Canada as snow-dominated, data-poor case studies.

Study focus: In this study, seven lumped conceptual models were thoroughly compared in order to determine the best performing model for reproducing different components of the hydrograph, including low and high flows in data-poor catchments. All models were calibrated using five various objective functions for reducing the effects of calibration process on models' performance. Additionally, the effects of precipitation, an important factor, particularly in data-scarce regions, were assessed by comparing two precipitation input scenarios: (1) low-density ground-based gauge data, and (2) the Canadian Precipitation Analysis (CaPA) data. The final goal of this study was to compare the effects of using either the Degree-Day or SNOW17 snowmelt estimation methods on the accuracy of streamflow simulation.

New hydrological insights: The results indicate that, in general, MACHBV is the best performing model at simulating daily streamflow in a data-poor watershed, and both

SACSMA and GR4J can provide competitive results. Additionally, MACHBV and GR4J are superior to the other conceptual models regarding high flow simulation. Moreover, it was found that incorporating the more complex SNOW17 snowmelt estimation method did not always enhance the performance of the hydrologic models. Finally, the results also confirmed the reliability of the CaPA data as an alternative forcing precipitation in the case of low data availability.

Keywords: Model inter-comparison; Simple conceptual models, Data scarce regions; Canadian Precipitation Analysis (CaPA); Snowmelt estimation; Canada

2.2 Introduction

Possessing reliable hydrological models is an important issue for operational hydrology and water resources management (Donnelly-Makowecki & Moore, 1999; Razavi & Coulibaly, 2017), and this is a unique challenge in data-scarce regions (Adjei et al., 2015). Various types of rainfall-runoff models, from lumped empirical to fully distributed physically-based ones, have been designed with different mathematical representations of hydrological processes (Beven, 2011; Lü et al., 2013; Moradkhani & Sorooshian, 2008). Empirical or data-driven models are based on mathematical equations not specifically related to the physical processes of the watershed. Although these models have some advantages, such as having higher performance efficiency, they are only valid within the boundaries of the given data (Shrestha & Solomatine, 2009). In contrast, physically-based distributed models are better at representing spatial variability when characterizing the water cycle processes, and can produce more reliable results (Moradkhani & Sorooshian, 2008; Smith et al., 2004). Having parameters with physical interpretation and spatial

variability is the main advantage of physically-based distributed models, however, their proper estimation requires more computational cost and huge amount of data (Shrestha, 2009), this poses a problem when the area of interest has low data availability (Gan et al., 2006; Grayson et al., 2002; Tegegne, Park, & Kim, 2017; Young, 2002). Conceptual lumped models are another group of hydrologic models, which can provide a desirable alternative to empirical and physically-based distributed models. These models are commonly based on several interconnected conceptual elements representing different hydrologic components. There are various conceptual hydrologic models with different structures and processes, and the popularity of these models is due to their simplicity and low computational costs.

Inter-comparison of various models is one of the most convenient approaches for assessing the influence of model structure and aiding in the selection of the best performing model (Breuer et al., 2009; Garavaglia et al., 2017). Also, through the comparison a multi-model ensemble can be generated which can then be used for quantifying model structural uncertainty (Seiller et al., 2012). Various studies have been conducted using model inter-comparison experiments in the field of streamflow simulation (e.g. Chiew et al., 1993; Das et al., 2008; Gan et al., 1997; Koch et al., 2016; Michaud & Sorooshian, 1994; Shi et al., 2011; Suliman et al., 2015; Te Linde et al., 2008; Tegegne et al., 2017a; Vansteenkiste et al., 2014; Zhang et al., 2016); however, few of them focused on regions with low data availability. Refsgaard and Knudsen (1996) compared a physically-based distributed, a lumped conceptual, and a semi-distributed hydrologic models for three data-scarce regions in Zimbabwe. Their results showed that although using the distributed model provided

reliable results, it did not outperformed simpler ones in term of streamflow simulation at the outlet. The study by Tegegne et al. (2017b) found that, through the inter-comparison of lumped conceptual models with a physically-based semi-distributed model in data poor catchments, use of a more complex model could not be justified. Anshuman et al. (2019) proposed considering conceptual over physically-based models in the case of facing watersheds with low data availability. In addition, comparing a semi-distributed and a lumped model by Srivastava et al. (2020) again showed the superiority of the lumped model for hydrological modeling in data-limited basins. Although it can be argued that simple lumped conceptual hydrologic models could be the best choice for modeling rainfall-runoff process in data-scarce watersheds, the need of comprehensive comparison of the performance of various lumped conceptual model structures is strongly felt on regions with limited data availability.

Furthermore, the applicability of a conceptual hydrologic model is highly related to how well its parameters are estimated (Sorooshian et al., 1993). Although using a multi-objective calibration procedure provides valuable information about the parameter equifinality and uncertainty, there are different studies ranging from flood forecasting (e.g. Han et al., 2019; Reggiani et al., 2009; Wijayarathne & Coulibaly, 2020) to climate change assessment (e.g. Ashofteh et al., 2017; Li et al., 2014), where one optimal parameter set is utilized for streamflow simulation or forecasting. Various performance statistics have been developed to evaluate the performance of hydrological models and each of them can be considered as an objective function for estimating parameters (Lü et al., 2013; Wöhling et al., 2013). Therefore, the selected objective function may affect the performance of the

calibrated rainfall-runoff model. Although reducing these effects on the model inter-comparison process seems important for achieving more reliable results (Gan et al., 1997; Ouermi et al., 2019), it has not received much attention in previous studies.

Reliable and accurate historical forcing data has a considerable effect on the calibration process and the corresponding model performance (Te Linde et al., 2008). Temperature can be accurately estimated using the low-density measurements; however, accounting for the spatial and temporal variability of precipitation is challenging in data-poor watersheds (Price et al., 2014). Consequently, assessing the potential of utilizing other sources of precipitation data (e.g., Satellite or Radar-based data) as input into a rainfall-runoff model seems necessary in the case of limited ground-based observation stations. This has recently motivated research, in data-scarce regions, to evaluate the influence of using other sources of precipitation (e.g. Climate Forecast System Reanalysis data (Dile & Srinivasan, 2014; Fuka et al., 2014), Tropical Rainfall Measuring Mission precipitation analysis (Adjei et al., 2015; Collischonn et al., 2008; Worqlul et al., 2017), and North American Regional Reanalysis (Choi et al., 2009)) for models calibration. Their general finding is that the precipitation products provide valuable information for data-scarce regions while their evaluation at local scale is required due to regional variability of their quality (Lakew et al., 2020; Sirisena et al., 2018). In addition, by removing possible random and systematic errors, the application of a bias correction method can enhance the applicability of precipitation products at regional scale (Habib et al., 2014).

Canada specific, a Canadian Precipitation Analysis (CaPA) data is a gridded precipitation product, which is generated using various sources of precipitation, such as observations,

radar data, and forecasts (Mahfouf et al., 2007). There are few studies evaluating CaPA in terms of hydrologic performance. Eum et al. (2014) applied CaPA into the VIC model for streamflow simulation of Canadian mountainous catchments. In the case study by Gaborit et al. (2017), a comparison was made between CaPA and ground-based observations in terms of the accuracy of runoff predictions using two lumped models. Also, Boluwade et al. (2018), assessed the reliability of CaPA as forcing of Watflood hydrologic model. In previous studies, the accuracy of various precipitation products were evaluated, while none have utilized them in a model inter-comparison experiment for reducing the effects of forcing input on model performance. Also, assessing precipitation products utilizing multiple hydrologic models lead to more robust results than their proxy evaluation using streamflow data and an auxiliary model (Fortin et al., 2018).

The main objective of this study is to evaluate various lumped conceptual models with different structures for continuous daily streamflow prediction in order to propose the most suitable one for operational hydrology in watersheds with low data availability. By focusing on two data-poor watersheds in Northern Ontario, Canada, the performance of seven different lumped conceptual models are thoroughly compared using different statistics focusing on low and high flow conditions. In the proposed inter-comparison framework, we utilized an ensemble of calibrated parameter sets for each hydrologic model, derived from implementing different objective functions, which reduces the effects of the calibration process on models' performance. Also, besides the low-density gauge measurements, we used CaPA precipitation data as an alternative forcing input in the model comparison process. This helps us reach more robust conclusions about the direct ability

of different hydrologic model structures for streamflow simulation. In addition, we have the opportunity to evaluate the reliability of using CaPA data as an alternative input forcing precipitation in the case of low available historical meteorological data in Northern Ontario. Moreover, given that the regions of interest are snow-dominated watersheds, where the snowmelt freshet is the main cause of floods, using a proper snowmelt estimation routine in model structure seems necessary for accurate streamflow simulation. Therefore, as a side objective of this study, a comparison is made between two popular temperature-index methods, the simple Degree-day (Samuel et al., 2011) and the more complex SNOW17 (Anderson, 2006, 1973), to evaluate their potential for improving stream flow estimation in snow-dominated basins with low data availability. It is of note that there are lots of studies comparing temperature index with complex energy balance snowmelt models (Bowling et al., 2003; Debele et al., 2010; Essery et al., 2013; Troin et al., 2015), while the comparison between different conceptual methods, where the only required inputs are temperature and precipitation, has not received much attention (Agnihotri & Coulibaly, 2020).

2.3 Methods

2.3.1 Study Area and data description

The study regions are the Big East River and the Black River watersheds, which have areas of 620 and 1522 km2, respectively. Both watersheds are located in the Muskoka/Bracebridge areas of Northern Ontario, Canada (Figure 2-1). The terrain elevation of the Big East River watershed ranges from 293 to 564 meter above sea level (masl) while the Black River changes from 221 to 421 masl. There are no major urban

areas in either watersheds and, with the exception of the southern part of the Black River used for agriculture, the dominant land cover of two basins are mixed forest vegetation. There are only Six Environment and Climate Change Canada (EC) meteorological stations with more than 10 years of reliable data and all are located near, but not within, the aforementioned watersheds boundaries (Figure 2-1). Each watershed has one hydrometric station, located at their outlets, confirming the status of low data availability. Based on all available historical data, the long-term daily mean air temperature of the regions is around 5°C and the warmest and coldest months are February and July, respectively. The temperature is near or below freezing point from November to March, indicating a need for modeling snow storage and snowmelt processes. The highest discharge for both watersheds occurs during spring, indicating the runoff is snowmelt dominated (Table 2-1).

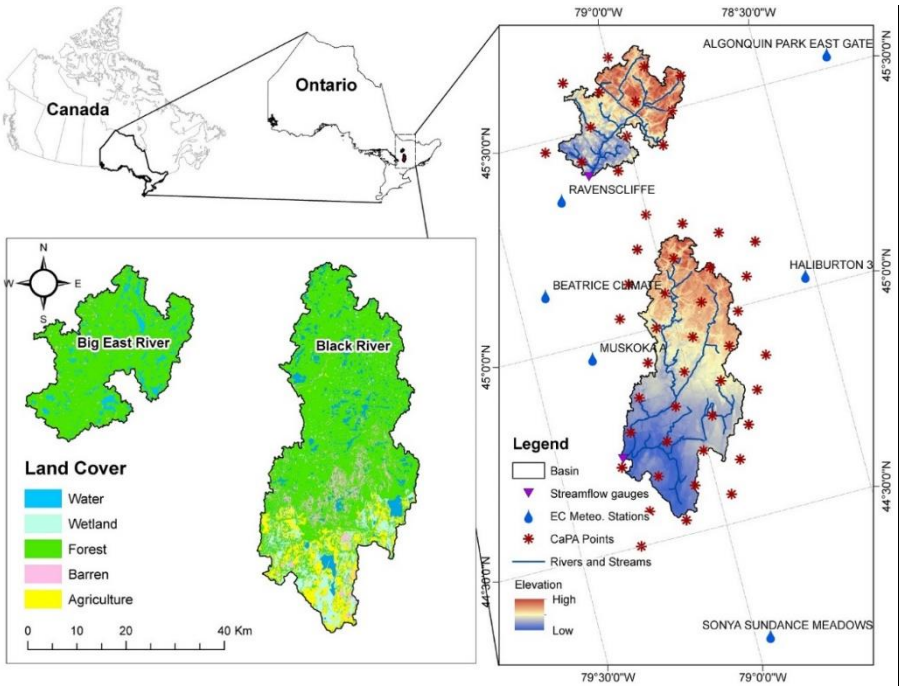


Figure 2-1 The study areas: Big East River and Black River watersheds (modified after Darbandsari & Coulibaly, 2019)

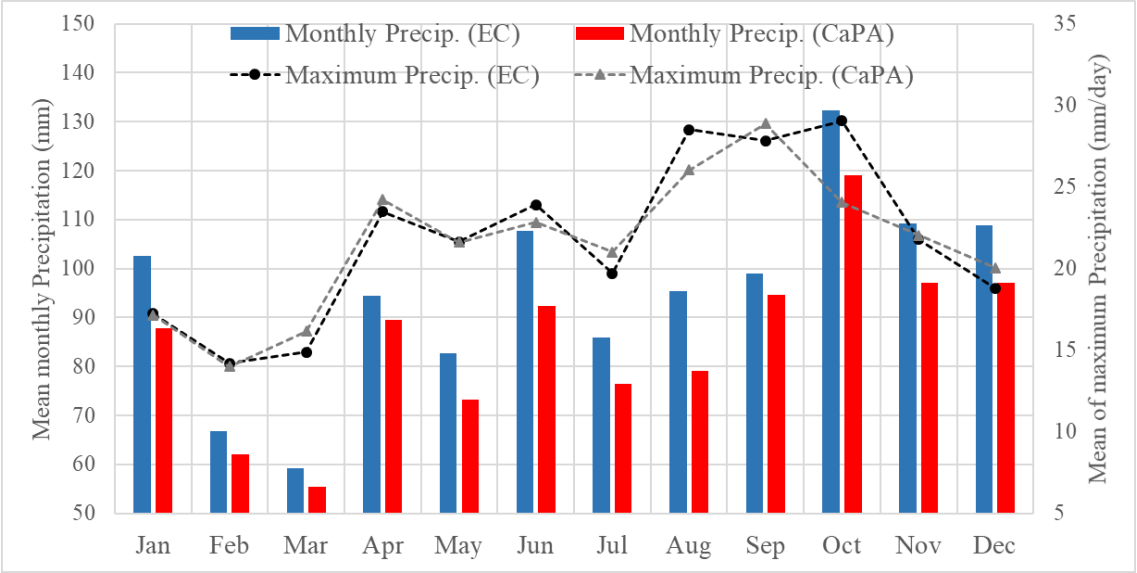
Table 2-1 The details of all utilized datasets and the climate characteristics of different measurements based on available historical data from 2006 to 2015

	Stations	Type	latitude	longitude	Range (years)	Precipitation or Flow (mm/month)			
						Winter	Spring	Summer	Fall
ECCC stations[4]	Beatrice	P[1]	45.14	-79.4	2006-2015	108	89	81	97
	Algonquin	P	45.53	-78.27	2006-2015	90	73	72	85
	Haliburton	P	45.03	-78.5	2006-2015	104	85	78	104
	Muskoka	P	44.97	-79.3	2006-2015	98	55	72	87
	Sonya	P	44.21	-78.95	2006-2015	78	63	67	87
	Ravenscliffe	P	45.35	-79.27	2006-2015	128	106	81	102
CaPA points		P	-	-	2006-2015	[77,120][2]	[70,110]	[81,111]	[96,139]
Outlet (BE)[5]		F[1]	45.39	-79.16	2006-2015	53	92	27	43
Outlet (BL)[5]		F	44.71	-79.28	2006-2015	54	80	15	28
Average monthly Temperature (°C)						[-12,-5][3]	[-5,12]	[16,19]	[0,14]

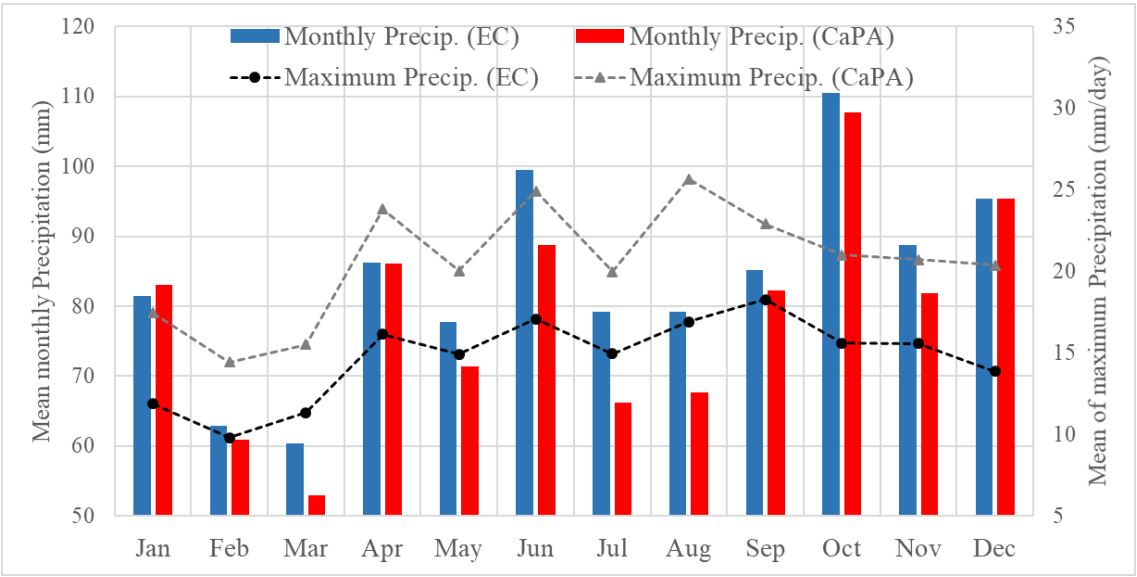
[1] P and F are the abbreviations of precipitation and flow, respectively.
[2] The range of historical seasonal precipitation value derived from all CaPA points shown in Figure 2-1.
[3] The changes of historical average monthly temperature in each season
[4] The meteorological stations are operated by Environment and Climate Change Canada.
[5] The hydrometric stations are operated by the Water Survey of Canada.

Apart from EC meteorological stations, the archive of Canadian Precipitation Analysis (CaPA), produced by the Meteorological Service of Canada (Mahfouf et al., 2007), is another source of precipitation time series available for both study regions. CaPA is a near real-time gridded precipitation analysis based on the combination of observation and climate model data with a spatial and temporal resolution of 15 km and 6 hours, respectively (Lespinas et al., 2015). CaPA points (i.e., center of each CaPA grid), located inside or near both watersheds, are illustrated in Figure 2-1. CaPA provides better spatial coverage of both watersheds so considering it as an alternative forcing precipitation input may enhance the performance of hydrologic models in the case of limited data

measurements. Consequently, two different precipitation input scenarios will be evaluated. In the first scenario, the daily precipitation comes from interpolation of the available EC meteorological stations to the center of the watersheds using the Inverse Distance Weighting (American Society of Civil Engineers, 1996) method. The second scenario involves using the Thiessen polygon method (Thiessen, 1911) to generate mean areal precipitation with the daily aggregated CaPA data. The primary comparison of two scenarios shows that in general, CaPA data underestimates the precipitation amount in comparison with EC (Figure 2-2). However, CaPA, specifically in Black River watershed, proposes more intense rainfall events. In this study, both aforementioned scenarios are separately used to calibrate the models' parameters and the reliability of using CaPA data is evaluated based on the performance of the calibrated models.



(a) Big East River watershed



(**b**) Black River watershed

Figure 2-2 The comparison of mean areal precipitation of (a) Big East River and (b) Black River watersheds derived from EC and CaPA data

2.3.2 Rainfall-runoff Models

The seven structurally different conceptual hydrologic models compared in this study are: (1) the Sacramento soil moisture accounting (SACSMA) (Burnash et al., 1973); (2) the McMaster University Hydrologiska Byrans Vattenbalansavdelning (MACHBV) (Samuel et al., 2011); (3) Génie Rural à 4 Paramètres Journaliers (GR4J) (Perrin et al., 2003); (4) the modified version of the Soil moisture and accounting routing (SMARG) (Liang, 1992); and three different implementations of the Hydrologic Engineering Center's Hydrologic Modeling System (HEC-HMS) software (Scharffenberg, 2016). These models are chosen mainly based on their structural diversity and performances in previous studies. SACSMA is widely used for operational flood forecasting in United States (e.g. Day, 1985; Seo et al., 2003; Vrugt et al., 2006) and it was proven to perform well in Canadian catchments (Agnihotri & Coulibaly, 2020; Wijayarathne & Coulibaly, 2020). MACHBV is the modified version of HBV (Bergström, 1976) which is specifically developed for enhancing streamflow estimation of ungauged basins (Samuel et al., 2012, 2011). SMARG is another well-known conceptual model with variable number of soil storage, which is proven to provide better performance than its original version in humid and semi-humid regions. (Tan & O'Connor, 1996). GR4J is a parsimonious hydrologic model with two conceptual storages, and was successfully applied in Canadian cold regions (Gaborit et al., 2017; Martel et al., 2020; Seiller et al., 2012). Moreover, HEC-HMS is a widely used platform all over the world, which provides several different methods for developing structurally different models to simulate rainfall-runoff process (Gyawali & Watkins, 2013; Teng et al., 2018). The main different characteristics of the structures of these seven models are

summarized in Table 2-2. These models possess distinct complexities and their number of parameters varies based on their corresponding number of hydrologic processes and descriptions. A brief explanation of each model is provided in the following sections.

It is worthy of note that in all SACSMA, MACHBV, SMARG, and GR4J, the daily potential evapotranspiration (PET) was calculated using the simplified Thornthwaite equation (Samuel et al., 2011; Thornthwaite, 1948) by multiplying mean daily temperature by athorn as the only parameter being determined through calibration process (Table 2-3). In addition, the simple Degree-day (DD) snowmelt routine was added to these models for representing changes in the snowpack as well as discriminating snow and rainfall. A brief explanation of the DD method is presented in Section 2.3.3.

2.3.2.1 Sacramento Soil Moisture Accounting (SACSMA)

The SACSMA is a well-known conceptual lumped hydrologic model, which is used by the National Weather Service River Forecast System (NWSRFS) for flood forecasting. In this model, the surface of the basin is divided into pervious and impervious areas. The soil profile of the pervious portion is partitioned into the thin upper and thicker lower zones. A total of five state variable reservoirs are used to determine the accumulation of "tension" and "free" water storages, representing the water bound and not bound to the soil particles, respectively (Caldwell et al., 2015; Razavi & Coulibaly, 2017). Moreover, the Nash cascade method is implemented as the routing approach in this model. As can be seen in Table 2-3, the model possesses 17 parameters that must be specified by the user or through an automatic calibration.

Table 2-2 Characteristics of the model structure of seven different conceptual hydrologic models used in this study

Model	Conceptual storage	Type of flows	Routing	Evapotranspiration
SACSMA	•Upper soil tension water •Upper soil free water •Lower soil tension water •Lower soil primary free water •Lower soil supplemental free water	•Direct flow •Surface flow •Interflow •Baseflow	•Direct, Surface and Interflow: Cascade of three linear reservoirs	Occurred from: •Upper soil tension water •Upper soil free water •Lower soil tension water
MACHBV	•Soil moisture layer •Upper soil reservoir •Lower soil reservoir	•Upper soil flow •Lower soil flow	•Non-linear Equilateral triangular weighting	Occurred from: •Soil moisture layer
SMARG	•Multiple soil layers •Groundwater storage	•Direct flow •Surface flow •Interflow •Baseflow	•Direct, Surface and Interflow: cascade of multiple reservoirs (Nash model) •Baseflow: Single linear reservoir	Occurred from: •Total precipitation •All soil layers
GR4J	•Production soil storage •Routing soil storage	•Fast flow •Slow flow	•Fast flow: Unit hydrograph and non-linear routing storage •Slow flow: Unit Hydrograph	Occurred from: •Total precipitation •Production store
HEC1	•Canopy storage •Soil Storage	•Direct flow •Surface flow •Baseflow	•Direct and Surface flow: Clark unit hydrograph •Baseflow: exponential recession model	Occurred from: •Canopy storage
HEC2	•Canopy storage •Surface Storage •Upper zone storage •Tension zone storage	•Direct flow •Surface flow •Baseflow	•Direct and Surface flow: Clark unit hydrograph •Baseflow: exponential recession model	Occurred from: •Canopy storage •Upper zone storage •Tension zone storage
HEC3	•Canopy storage •Surface Storage •Upper zone storage •Tension zone storage •GW upper layer (GW1) •GW lower layer (GW2)	•Direct flow •Surface flow •Baseflow GW1 •Baseflow GW2	•Direct and Surface flow: Clark unit hydrograph •Two separate single linear reservoir for both baseflow components	Occurred from: •Canopy storage •Upper zone storage •Tension zone storage

2.3.2.2 McMaster University Hydrologiska Byrans Vattenbalansavdelning (MACHBV)

The MACHBV is a nonlinear variant of the conceptual HBV model (Bergström, 1976). In this model, a soil moisture routine accounts for fluctuations of the soil moisture storage of the basin. A response function, comprising upper and lower soil reservoirs, is used for

estimating the amount of runoff based on the recharge from soil moisture routine. Finally, considering the modified routing routine proposed by Samuel et al. (2012), where the nonlinear storage-discharge relationship is considered in the lower layer of deep soil, the final streamflow is obtained. These processes are controlled through 10 parameters, shown in Table 2-3. More detailed description of the MACHBV model can be found in Samuel et al. (2011, 2012).

2.3.2.3 The modified version of Soil moisture and accounting routing (SMARG)

The SMARG is a 10-parameter lumped conceptual rainfall-runoff model following the structure of its earliest version (i.e., SMAR; (O'Connell et al., 1970)). In this modification, a single reservoir groundwater component is added for considering the effects of groundwater on total estimated discharge (Liang, 1992). Therefore, it is more reliable in humid regions where the groundwater component contributes significantly in generating runoff (Tan & O'Connor, 1996). This model uses a nonlinear water balance routine for simulating runoff generation process by visualizing the watershed as a stack of horizontal soil storage layers. Then, the generated surface runoff is transferred through Nash cascade of equal linear reservoirs model (Nash, 1957) while the previously mentioned single linear reservoir is used for routing the groundwater discharge. The brief descriptions of the model parameters and their initial ranges are provided in Table 2-3.

2.3.2.4 Génie Rural à 4 Paramètres Journaliers (GR4J)

The GR4J is a daily conceptual rainfall-runoff model, developed based on the GR3J model (Edijanto et al., 1999), with only four parameters needing to be calibrated (Table 2-3). This

model consists of three subsequent steps. First, net precipitation and evapotranspiration is calculated. Afterward, a portion of net rainfall goes to the one parameter ($x1$) production store where the actual evapotranspiration and percolation are determined. Then, the remaining portion of net precipitation and percolation are used for determining discharge. A unit hydrograph with time base of $x2$ and a one-parameter ($x3$) non-linear routing store transfer 90 percent of the available water as slow flow; while the other ten percent is considered as a fast flow and routed by a unit hydrograph with time base of $2 \times x2$. Finally, after applying the one parameter ($x4$) groundwater exchange component, the total discharge is computed by adding the two aforementioned routed flows (Perrin et al., 2003).

Table 2-3 Parameters of the SACSMA, MACHBV, SMARG, and GR4J models and their initial and optimized ranges

Parameter	Description	Unit	Initial Range	Optimized Range**
SACSMA				
UZTWM	Upper-zone tension water maximum storage	mm	1-150	13 - 145
UZFWM	Upper-zone free water maximum storage	mm	1-150	4 - 150
LZTWM	Lower-zone tension water maximum storage	mm	1-500	2 - 400
LZFPM	Lower-zone free water primary maximum storage	mm	1-1000	235 - 979
LZFSM	Lower-zone free water supplemental maximum	mm	1-1000	147 - 940
ADIMP	Additional impervious area	-	0-0.4	0 - 0.32
UZK	Upper-zone free water lateral depletion rate	day-1	0.1-0.5	0.14 - 0.5
LZPK	Lower-zone primary free water lateral depletion rate	day-1	0.0001-0.025	0.01 - 0.02
LZSK	Lower-zone supplemental free water lateral depletion rate	day-1	0.01-0.25	0.04 - 0.19
ZPERC	Maximum percolation rate	-	1-250	4 - 234
REXP	Exponent of the percolation equation	-	0.01-6	1 - 5.9
PCTIM	Impervious fraction of the watershed area	-	0-0.1	0 - 0.05
PFREE	Fraction percolating from upper to lower zone free water storage	days	0-0.6	0 - 0.57
Rq	Routing coefficient	-	0-0.99	0.2 - 0.55
RIVA*	Riparian vegetation area	-	0	-
SIDE*	Ratio of deep recharge to channel base flow	-	0	-
RSERV*	Fraction of lower zone free water not transferable to tension water	-	0.3	-

MACHBV				
fc	Maximum for soil water content	mm	50-800	54 - 346
lp/fc	Limit for PET to determine actual ET	mm/mm	0.1-0.9	0.3 - 0.9
lsuz	A threshold value used to control response routing on an upper soil reservoir	mm	1-100	6 - 90
cperc	Constant Percolation rate parameter	mm/day	0.01-6	0.15 - 5.92
beta	A non-linear parameter controlling runoff generation	-	0-10	0.9 - 10
k0	Flow recession coefficient in an upper soil reservoir	days	1-30	2 - 30
k1	Flow recession coefficient in an upper soil reservoir	days	2.5-100	6 - 37
k2	Flow recession coefficient in a lower soil reservoir	days	20-1000	346 - 992
alpha1	An exponent in relation between outflow and storage	-	0.5-20	1.1 - 12
maxbas	Parameter of a triangle weighting function for modeling a routing routine	days	1-20	1.3 - 4.4
SMARG				
T	Conversion parameter for calculating potential evaporation	-	0-1	0.47 - 0.98
C	Decay coefficient of soil evaporation	-	0-1	0.1 - 1
Z	The total depth of all soil layers	mm	0.01-500	72 - 495
H	Direct runoff factor	-	0-1	0.05 - 0.74
Y	Infiltration capacity	mm/day	0-200	24 - 152
N	the Nash Cascade model parameter (number of resorvoirs)	-	1-20	1 - 18
NK	number of time step for surface runoff routing	days	0-200	4 - 145
G	The groundwater runoff coefficient	-	0-1	0.63 - 1
Kg	number of time step for groundwater routing	days	0-200	5 - 51
m*	Memory length of the routing response function	days	100	-
GR4J				
x1	Maximum capacity of the production store	mm	1-1500	99 - 512
x2	The groundwater exchange coefficient	mm/day	-10-5	-6.7 - 2.3
x3	Maximum capacity of routing store	mm	1-500	146 - 439
x4	The unit hydrograph time base	days	0.5-4	2.2 - 3.6
Simplified Thornwaite's PET formula				
athorn	A constant for Thornthwaite's equation (PET)	-	0.1-0.3	0.15 - 0.3

*Predefined fixed values are used for these parameters based on previous studies

** Optimized ranges are based on the ensemble of calibrated parameter sets of each model for both watersheds with different objective functions and input scenarios(the outlier values are removed)

2.3.2.5 Lumped HEC-HMS based models

The HEC-HMS software, developed by the US Army Corps of Engineers, is designed for both continuous and event-based simulation of the rainfall-runoff process of dendritic

watersheds (Scharffenberg, 2016). It is a reliable platform for developing variants of lumped and semi-distributed hydrologic models by allowing the user to choose from an assortment of methods which can model different components of hydrologic cycles (i.e., surface, canopy, loss, transform, baseflow and routing processes). The general structure of any continuous hydrologic model developed based on the HEC-HMS framework is illustrated in Figure 2-3. As can be seen, the meteorological component, including the snow-rainfall discrimination and snowmelt module, estimates the excess water that may contribute to runoff generation. Then, the basin component, consisting of the conceptual simulations of different physical phenomena, determines the streamflow value at the outlet of the watershed. In this study, three lumped, structurally different hydrologic models were developed using different combinations of the available baseflow and loss methods in the HEC-HMS platform. The first HEC-HMS model, called HEC1 hereafter, uses the Deficit and Constant loss method and the Recession baseflow method, while the other two models (i.e., HEC2 and HEC3) use the soil moisture accounting loss method with the Recession or Linear Reservoir baseflow methods, respectively. The parameters of the utilized approaches, presented in Table 2-4, show the total number of 7, 15, and 17 parameters for HEC1, HEC2, and HEC3 models, respectively (excluding the parameters of snowmelt routine).

It is worth mentioning that the only snowmelt method available in the HEC-HMS software is the temperature index that is an extension of the degree-day (DD) approach. However, unlike the DD approach where a constant snowmelt rate is used for each degree above a base temperature, the melting rate in HEC-HMS is determined as a function of an

antecedent temperature index (ATI) (Gyawali & Watkins, 2013; Razmkhah et al., 2016). As can be seen in Table 2-4, due to the importance of ATI melt rate (ATIMR) function on the accuracy of the snowmelt process, 22 different scenarios are defined for the ATIMR curve proposed by USACE (Khalida et al., 2014; U.S. Army Corps of Engineers, 1991) and the best one is determined through the calibration process. Additionally, all three HEC-HMS based models use the monthly average PET option in their meteorology models; the monthly average was calculated using the Hargreaves equation (Hargreaves & Samani, 1985), which is proved to be one of the most promising temperature-based PET estimation method in cold regions (Almorox et al., 2015).

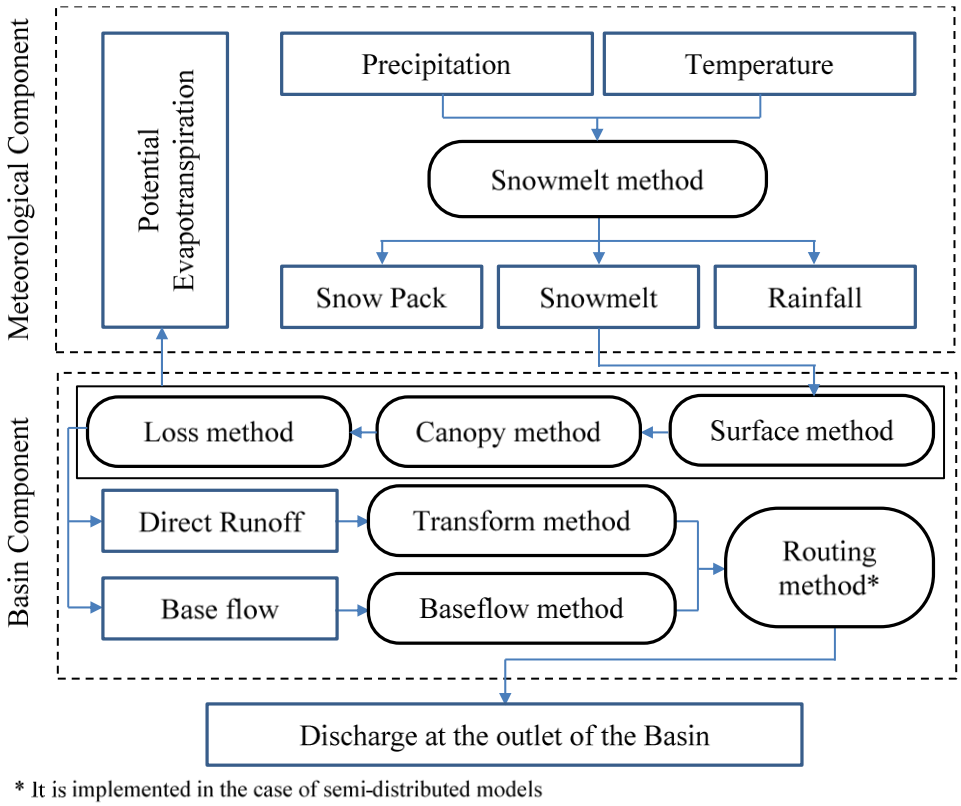


Figure 2-3 The general structure of HEC-HMS based hydrologic models (Feldman, 2000; Scharffenberg, 2016)

Table 2-4 The parameters of various HEC-HMS hydrologic process and their acceptable ranges

Process	Method	Description	Unit	Range
Canopy	Simple	The maximum storage capacity of canopy	mm	0-1500
		PET pan coefficient	-	0-10
Surface	Simple	The maximum storage capacity of surface	mm	0-1500
Loss	Deficit & Constant	The maximum storage capacity of the soil	mm	0-500
		Percolation rate	mm/hr	0.1-5
	SMA	Maximum infiltration from surface to the soil	mm/hr	0-500
		Storage capacity of the soil top layer	mm	0-1500
		A part of soil storage not affecting by gravity (Tension storage)	mm	0-1500
		Percolation rate from soil to GW1 layer	mm/hr	0-500
		Storage capacity of GW1 layer	mm	0-1500
		Percolation rate from GW1 to GW2 layer	mm/hr	0.01-500
		Lag time determining lateral outflow from GW1	hr	0.01-10000
		Available storage in the GW2 layer	mm	0.01-1500
		Deep percolation	mm/hr	0.01-500
		Lag time determining lateral outflow from GW2	hr	0.01-10000
Transform	Clark	Time of concentration	hr	0-1000
		Storage coefficient accounting for storage effects	hr	0.01-1000
Baseflow	Linear Reservoir	Time coefficient for linear reservoir in GW1 layer	hr	0 - 10000
		The number of reservoir used for routing	#	1-100
		Time coefficient for linear reservoir in GW2 layer	hr	0 - 10000
		The number of reservoir used for routing	#	1-100
	Recession	Recession constant	-	0-1
		Ratio of flow to peak flow for resetting base flow	-	0-1
Snowmelt	Temperature Index	Temperature for discriminating between snow and rainfall	C	-2-3
		Melting threshold temperature	C	-2.5-2
		Melt rate in the wet rain condition	mm.C/day	0 - 10
		Limit for Discriminating between dry rain and wet rain	mm/day	0 - 200
		Coefficient for updating the antecedent meltrate index	-	0.9-0.9995
		Relationship between meltrate and ATI (ATIMR function)	-	22 scenarios
		Threshold of rainfall caused rapid change in snow temperature	mm/day	0-20
		Coefficient for updating the antecedent cold content index	-	0.4 - 1
		The maximum liquid water capacity in the snowpack	%	3 - 10
		Melt rate caused by ground heat	mm/day	0 - 2
		Changing temperature in different elevation	C/1000m	-5

2.3.3 Snowmelt routines

As previously mentioned, due to the significant effects of the snowmelt process on stream flows in both Big East River and Black River watersheds, two different temperature-index models, requiring only temperature and precipitation data, are selected. Both methods possess the same snow-rainfall discrimination procedure where the Upper (T_r) and Lower (T_s) threshold temperatures are used as two calibrating parameters to distinguish between snow and rainfall (Samuel et al., 2011). In addition, in both models the amount of rain and snow are simply modified by multiplying by rain correction (RCF) and snow correction (SCF) factors, respectively. For snowmelt estimation, the Degree-Day (DD) method relies on a linear relationship between snowmelt and air temperature (T). If T is less than the melting temperature threshold (T_m), the melt rate is calculated by multiplying difference between T and T_m by degree-day factor (DDF). The SNOW17 approach, however, considers some of the physical processes involved in snowmelt (e.g., energy exchange between air and snow, the effects of rain on snow, the snowpack heat storage and deficit), without needing additional input data (Agnihotri & Coulibaly, 2020; Anderson, 2006). During the non-rain period, the same concept as DD is used for melt rate calculation whereas the melt factor seasonally changes based on two parameters, maximum ($mfmax$) and minimum ($mfmin$) melt factors. Additionally, the melt during rain is determined by a simplified empirical energy balance equation (Shamir & Georgakakos, 2006). Other parameters which should be calibrated for SNOW17 include the average wind function ($uadj$), antecedent snow temperature index ($tipm$), maximum negative melt factor (nmf),

and the water holding capacity of the snow pack ($plwhc$). Detailed descriptions of both snow routing approaches can be found in the above-cited references.

2.3.4 Optimization and Evaluation Processes

The main purpose of the calibration of a rainfall-runoff model is to select the best parameter values by minimizing the difference between observations and model streamflow predictions (Chiew et al., 1993). However, the choice of a proper objective function from a bunch of well-known performance statistics is not a straightforward task. In this study, for decreasing the influence of choosing objective function on models' inter-comparison, five criteria, including Nash Sutcliffe Efficiency (NSE), Nash Volume error (NVE), Kling Gupta Efficiency (KGE), Modified Nash Volume Error (MNVE), and Peak Weighted Root Mean Square Error (PWRMSE), were selected and considered as different single objective functions in order to find the best parameter set for each of the rainfall-runoff models (Table 2-5). While the first three metrics are formulated to accurately simulate medium flows, the latter two ones focus on providing more accurate high flow simulation.

Table 2-5 Performance statistics used as objective functions and evaluation criteria

Criteria	Mathematical Formulation	Range
Objective functions		
Nash Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970)	$1-\frac{\sum_{i=1}^{N}\left(Q_{s_t}-Q_{o_i}\right)^2}{\sum_{i=1}^{N}\left(Q_{o_i}-\overline{Q_o}\right)^2}$ (1)	-∞ to **1**(2)
Nash Volume Error (NVE) (Samuel et al., 2011)	$0.5NSE-0.1VE+0.25NSEL+0.25NSES$	-∞ to **1**
Kling Gupta Efficiency (KGE) (Gupta et al., 2009)	$1-\sqrt{(r-1)^2+(a-1)^2+(b-1)^2}$ (3)	-∞ to **1**
Modified NVE ($MNVE$) (Darbandsari & Coulibaly, 2019)	$NSES-0.1VE$	-∞ to **1**
Peak-weighted root mean square error ($PWRMSE$) (Cunderlik & Simonovic, 2004)	$\left(\frac{1}{N}\left(\sum_{i=1}^{N}\left(Q_{s_i}-Q_{o_i}\right)^2\times\frac{Q_{o_i}+\overline{Q_o}}{2\overline{Q_o}}\right)\right)^{\frac{1}{2}}$	**0** to +∞
Evaluation Criteria		
Volume Error (VE) (Samuel et al., 2011)	$\frac{\left\lvert\sum_{i=1}^{N}\left(Q_{s_i}-Q_{o_i}\right)\right\rvert}{\sum_{i=1}^{N}Q_{o_i}}$	**0** to +∞
NSE based on squared transformed data ($NSES$) (Razavi & Coulibaly, 2017)	$1-\frac{\sum_{i=1}^{N}\left(\log\left(Q_{s_i}\right)-\log\left(Q_{o_i}\right)\right)^2}{\sum_{i=1}^{N}\left(\log\left(Q_{o_i}\right)-\overline{\log(Q_o)}\right)^2}$	-∞ to **1**
NSE based on logarithmic transformed data ($NSEL$) (Razavi & Coulibaly, 2017)	$1-\frac{\sum_{i=1}^{N}\left(Q_{s_i}^2-Q_{o_i}^2\right)^2}{\sum_{i=1}^{N}\left(Q_{o_i}^2-\overline{Q_o^2}\right)^2}$	-∞ to **1**
Peak Error (PE) (Das et al., 2008)	$\frac{\lvert\overline{Q_{s_{max}}}-\overline{Q_{o_{max}}}\rvert}{\overline{Q_{o_{max}}}}$ (4)	**0** to +∞
Coefficient of Transferability (T_m) (Das et al., 2008)	$\max(NSE_{Cal}-NSE_{val},0)$	**0** to +∞
NSE based on flows more than 90 percentile ($NSE90$)	$1-\frac{\sum_{i=1}^{N}\left(Q_{s90_i}-Q_{o90_i}\right)^2}{\sum_{i=1}^{N}\left(Q_{o90_i}-\overline{Q_{o90}}\right)^2}$ (5)	-∞ to **1**

(1) Q_{s_i} and Q_{o_i} respectively represent the observed and simulated flows for day i.

(2) Bold and underlined value indicates perfect performance.

(3) r: linear correlation coefficient between Q_o and Q_s, a: standard deviation of Q_s over the standard deviation of Q_o, b: the mean of Q_s over the mean of Q_o

(4) $Q_{s_{max}}$ and $Q_{o_{max}}$ show the mean annual observed and simulated peak flows, respectively.

(5) Observed and simulated flows more than 90 percentile are denoted by Q_{s90_i} and Q_{o90_i}, respectively.

All models are calibrated for the period from 2006 to 2011, with the year 2006 considered as warm-up period. The three-year period (2012 to 2015) was used for the validation. The dynamically dimensioned search (DDS) algorithm (Tolson & Shoemaker, 2007) was used to calibrate the parameters for each model in this study. DDS is a heuristic global single-solution based search algorithm that was developed to calibrate complex hydrological models with a large number of parameters (Arsenault et al., 2014). The main distinguishing feature of this method is the transition from global to local search by dynamically rescaling the dimension of the search space. The Ostrich calibration toolkit (Matott, 2005) was used to run the DDS optimization for each model.

It is of note that although there are two available automatic calibration algorithms in HEC-HMS software, they perform poorly in finding the optimized parameter sets, especially when there are a relatively high number of parameters to be calibrated (Cunderlik & Simonovic, 2004). From the literature, with the exception of Dariane et al. (2016) who developed an automatic calibration for the HEC-HMS program based on genetic algorithm, there are no studies that calibrate all parameters of a continuous HEC-HMS model with snowmelt routine using auto-optimization methods. In this paper, Ostrich, Matlab, HEC-HMS, and HEC-DSSVue are linked together in order to apply the DDS optimization algorithm to calibrate HEC-HMS models (Figure 2-4).

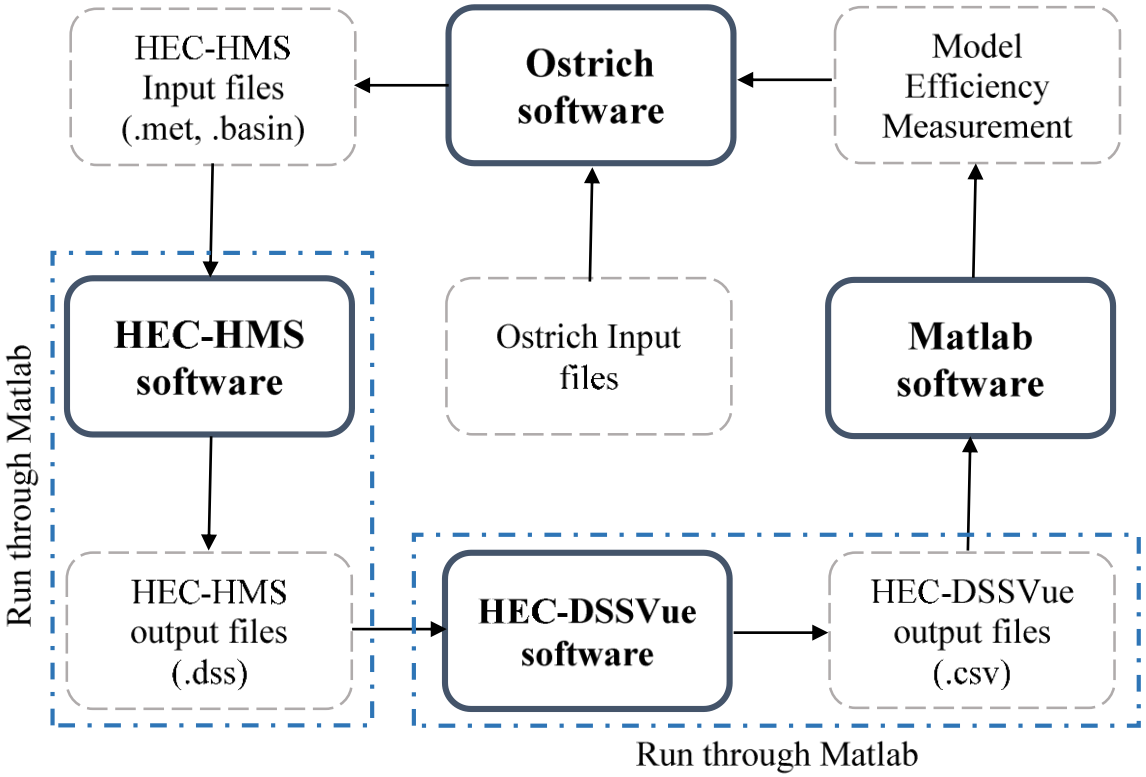


Figure 2-4 The general structure of HEC-HMS based hydrologic models (Feldman, 2000; Scharffenberg, 2016)

Moreover, in this study, seven different evaluation metrics are used for assessing the performance of the calibrated models (Table 2-5). Apart from standard NSE, we utilized NSE calculated based on logarithmic ($NSEL$) and squared ($NSES$) transformed stream flows for reflecting the accuracy of low and high flows, respectively. Volume Error (VE) assesses the long-term performance of the model simulation and Peak Error (PE) evaluates the models' ability in capturing peak flows. Also, the transferability of a model in time is illustrated using the coefficient of transferability (T_M) where the lower values show better model parameters' transferability from calibration to validation period. Moreover, for more specific evaluation of different models' ability in reproducing high flows, NSE is

calculated and compared using the 90 percentile streamflow values ($NSE90$). We also utilize representative hydrographs and scatterplots, as graphical tools, for visually assessing the performance of different models.

2.4 Results

2.4.1 Objective functions evaluation

The combination of seven hydrologic models, two different precipitation input scenarios, and five various objective functions leads to 70 different calibrated models for each watershed. Evaluating the effect of considering different objective functions is the prime step before comparing various hydrologic models. Therefore, a comparison is made between the performance of the calibrated models using the five performance criteria in Table 2-5 as an objective function for both the Big East River and Black River watersheds in the validation period (Figure 2-5). The first thing that stands out from the comparison results in both basins is that neither of the objective functions has complete superiority in providing the most promising parameters set of all models. Based on the results, it is shown that although the $PWRMSE$ and $MNVE$ statistics give more weights to high flow values, using them as an objective function does not significantly improve the performance of the calibrated models regarding high flows. Additionally, the models calibrated with those metrics (*PWRMSE* and *MNVE*) are the worst at modeling low flows, leading to large variation in $NSEL$. Moreover, as expected, calibrated models, utilizing NSE as an objective function, perform well according to NSE based performance criteria (i.e., NSE, $NSES$, and $NSEL$). However, there are some concerns about their performance regarding

volume and peak error related measurements, especially in Big East River watershed where both VE and PE uncertainty of NSE based calibrated models are relatively large. In addition, the comparison of $NSEL$ measurements shows that using NVE for calibrating models outperform the other ones in low flow simulation in both watersheds. The relative good performance of the KGE based calibrated models is noticeable in Big East River while in Black River watershed, they are not the best ones. By looking at all different performance criteria simultaneously, it can be seen that using KGE and NVE provide good and reliable performance when considering different aspects of the hydrographs (i.e. low, high, and peak flows). However, it is impossible to determine one objective function as the best one that can be used for calibrating all hydrologic models in both watersheds.

For better clarifying the effects of choosing objective function on model results, Table 2-6 exemplifies the validation performances of MACHBV and SMARG hydrologic models being calibrated using NVE and KGE as objective functions. In Big East River, using NVE provides better parameter estimation of MACHBV while KGE performs better for SMARG. The opposite is true in Black River watershed where KGE and NVE are better objective functions for MACHBV and SMARG, respectively. The use of NVE leads to the superiority of MACHBV in Big East River while SMARG perform better if both models are calibrated with KGE. The same issue arise in Black River where comparing KGE calibrated models shows the advantages of MACHBV over SMARG, however, the same performances are achieved in the case of using NVE.

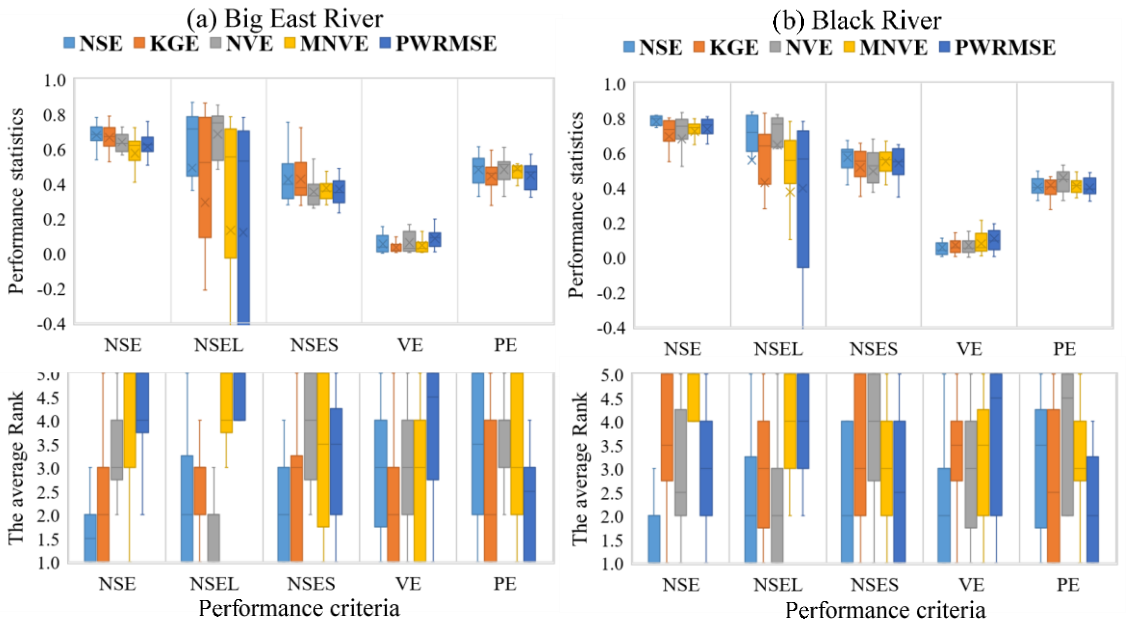


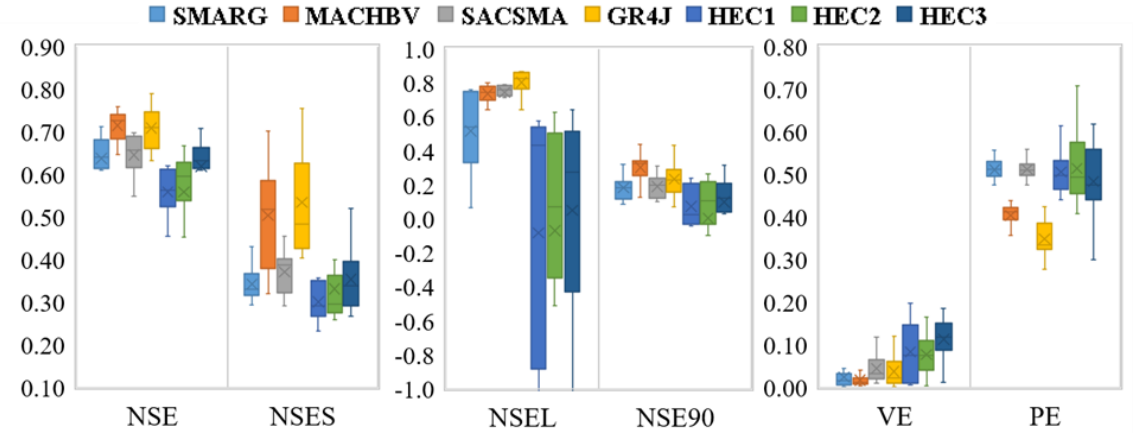
Figure 2-5 Box Plots of different performance criteria and their corresponding ranks for simulated daily stream flows by implementing five various objective functions during the validation period, derived from seven hydrologic models using two different input scenarios for (a) Big East River and (b) Black River watersheds

Table 2-6 The validation performances of MACHBV and SMARG hydrologic models being calibrated by using NVE and KGE as objective functions and ground based measurements as forcing precipitation input

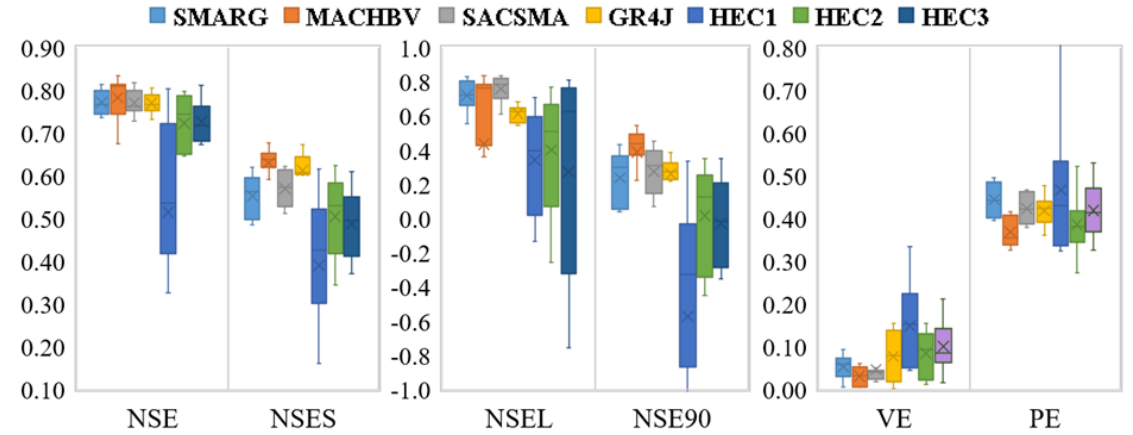
Criteria	Big East River				Black River			
	NVE		KGE		NVE		KGE	
	MACHBV	SMARG	MACHBV	SMARG	MACHBV	SMARG	MACHBV	SMARG
NSE	**0.81**	0.74	0.67	**0.75**	0.65	**0.68**	**0.73**	0.65
NSEL	0.79	0.79	0.36	**0.83**	**0.79**	0.76	**0.73**	0.54
NSES	**0.63**	0.50	**0.54**	0.52	0.33	**0.35**	**0.53**	0.31
VE	**0.01**	0.07	0.06	**0.05**	0.01	0.01	0.02	**0.01**
PE	**0.40**	0.48	**0.42**	0.45	**0.44**	0.51	**0.42**	0.47

2.4.2 Model Comparison

The main goal of this study is to find the most promising conceptual hydrologic model in data-poor watersheds by reducing the effects of the calibration process on models' performance. Therefore, two types of inter-comparison are made between the seven aforementioned hydrologic models. First, we compared all calibrated models with each other. Consequently, the twenty calibrated sets of parameters for each hydrologic model are considered to create the box plots of the evaluation performance metrics for the validation period (Figure 2-6). The most obvious conclusion derived from Figure 2-6 is that, in general, the HEC-HMS based models perform worst in comparison to the other hydrologic models. This relatively poor performance is more significant when the low flows are the main concerns of the simulation. Moreover, regarding high flows, the MACHBV shows the best performance while the high capability of the GR4J model in high flows simulation cannot be ignored, specifically in the Big East River watershed where the PE criteria of the GR4J models is the best one. By looking at all six performance statistics, the MACHBV is the most consistent hydrologic model for both watersheds and can be considered the best rated one. The performance of both the SACSMA and GR4J models also provide reliable and valuable results, whereas the HEC based models do not.



(a) Big East River watershed



(b) Black River watershed

Figure 2-6 Box plots of different performance statistics for simulated daily stream flows by using different hydrologic models, derived from sets of calibrated parameters based on five objective functions and two various input scenarios, for the (a) Big East River and (b) Black River watersheds during the validation period

The second type of model inter-comparison is between the best-calibrated set of parameters for each model using the EC precipitation input scenario. We chose the best parameter set by comparing all the aforementioned performance statistics focusing on the validation period. The set of parameters with the best rank and rational condition (i.e., the values of different performance measurements are controlled manually and parameters' set

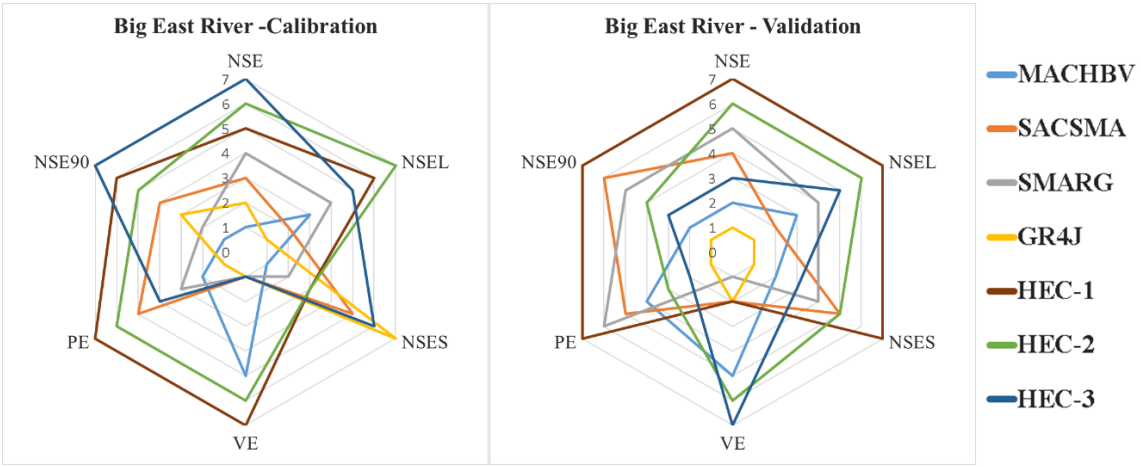
providing unreasonable predictions are ignored) was determined as the best selected one. The evaluated performance statistics of the final selected models in both calibration and validation periods and the relative comparison between them are presented in Table 2-7 and Figure 2-7, respectively. What stands out in the results for the Black River watershed is that, in general, MACHBV is the best-rated model based on most of the criteria in both calibration and validation periods. However, the results are more complex for the Big East River watershed, where the GR4J model is the best performing one during the validation period while MACHBV, SACSMA and GR4J models perform competitively based on the calibration period results. Moreover, regarding high flow-based criteria, MACHBV and GR4J respectively provide the best results in Black River and Big East River watersheds. In addition, as concluded in the previous comparison, although HEC2 and HEC3, compared with HEC1, relatively lead to more reliable results, the three HEC based models do not perform as well as the other four conceptual models in either watersheds.

Table 2-7 The performance statistics of the best-calibrated models in both watersheds

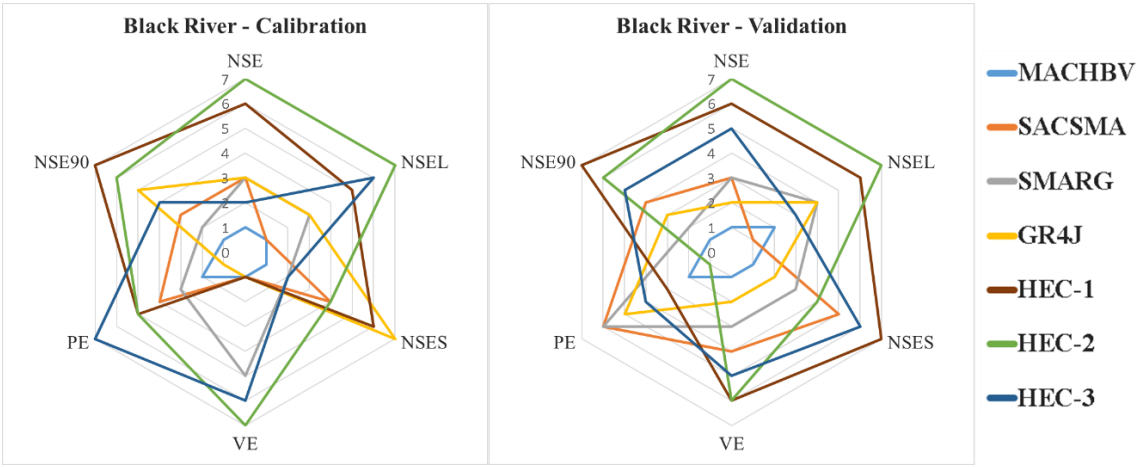
Basin	Model	Objective function[1]	Calibration						Validation					
			NSE	NSEL	NSES	VE	PE	NSE90	NSE	NSEL	NSES	VE	PE	NSE90
Black River Watershed	MACHBV	3	**0.90**[2]	**0.85**	**0.90**	0.00	0.14	**0.77**	**0.81**	0.79	**0.63**	**0.01**	0.40	**0.44**
	SACSMA	3	0.88	0.85	0.85	0.00	0.19	0.67	0.76	**0.82**	0.53	0.04	0.46	0.18
	SMARG	5	0.88	0.81	0.86	0.04	0.16	0.71	0.76	0.68	0.55	0.03	0.46	0.27
	GR4J	1	0.88	0.81	0.82	0.00	**0.11**	0.56	0.79	0.68	0.60	0.02	0.45	0.26
	HEC-1	4	0.84	0.57	0.84	0.00	0.20	0.54	0.72	0.55	0.50	0.15	0.42	-0.07
	HEC-2	4	0.80	0.39	0.85	0.11	0.20	0.55	0.65*	0.10	0.54	0.15	**0.38**	0.01
	HEC-3	1	0.89	0.49	0.86	0.06	0.23	0.57	0.75	0.78	0.51	0.11	0.44	0.05
Big East River Watershed	MACHBV	2	**0.84**	0.76	**0.81**	0.01	0.05	**0.65**	0.73	0.73	0.53	0.02	0.42	0.33
	SACSMA	2	0.77	0.79	0.63	0.00	0.27	0.45	0.66	0.79	0.35	0.01	0.47	0.15
	SMARG	4	0.74	0.66	0.72	0.00	0.17	0.54	0.63	0.54	0.37	0.00	0.48	0.20
	GR4J	2	0.82	**0.81**	0.51	0.00	**0.03**	0.49	**0.79**	**0.86**	**0.72**	0.01	**0.28**	**0.43**
	HEC-1	4	0.68	0.35	0.70	0.04	0.31	0.23	0.55	0.20	0.28	0.01	0.50	0.00
	HEC-2	5	0.67	0.31	0.70	0.03	0.28	0.35	0.62	0.21	0.35	0.05	0.41	0.21
	HEC-3	2	0.60	0.55	0.57	0.00	0.23	0.06	0.71	0.40	0.52	0.09	0.30	0.31

1 The objective functions lead to the best parameter sets of each model (1 = NSE; 2 = KGE; 3 = NVE; 4 = $MNVE$; 5 = $PWRMSE$)

2 The best and the worst values are bolded and underlined, respectively.



(a) Big East River watershed



(b) Black River watershed

Figure 2-7 The rank of different hydrologic models based on various performance statistics in the validation and calibration periods for (a) Big East River and (b) Black River watersheds

For a better comparison of different model performance in simulating high flows, scatter plots of the simulated and observed daily peak flows (peak flows greater than 75 percentile) are illustrated in Figure 2-8. Additionally, for qualitative inspection, Figure 2-9 presents a representative portion of the simulated and observed hydrographs based on various hydrologic models with their best parameter sets (presented in Table 2-7). As was

concluded previously, MACHBV and GR4J show better performance in comparison to the other models regarding high flows in both watersheds. However, it is worthy of note that even the best hydrologic models possess a clear tendency to underestimate peak flows especially in the Black River Watershed. This underestimation can likely be attributed to both the models' structures as well as the quality of forcing input data. The poor estimation of mean areal precipitation, derived from low-density meteorological stations, can lead to systematic under/overestimation of stream flows (Collischonn et al., 2008; Tegegne et al., 2017b).

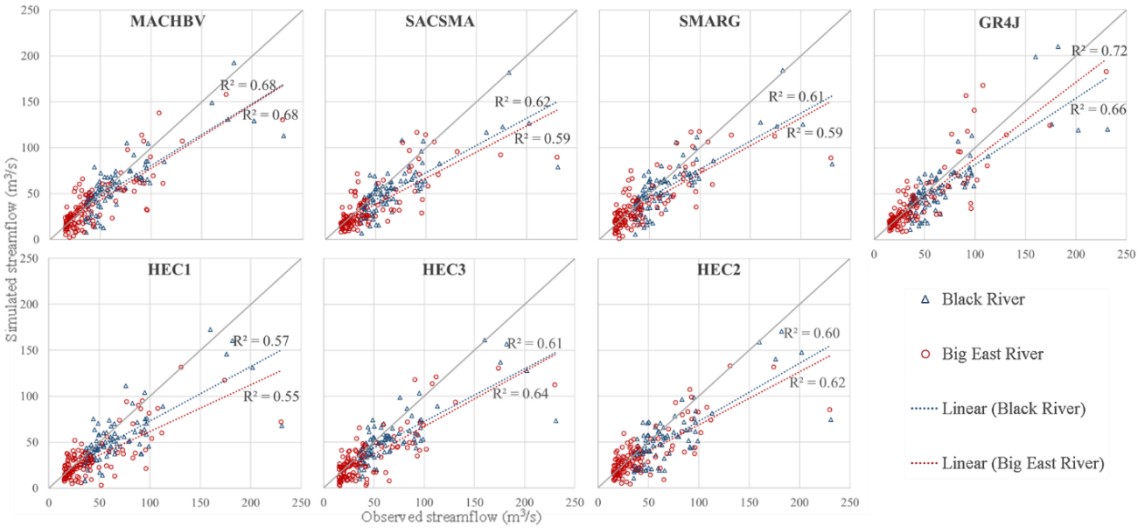
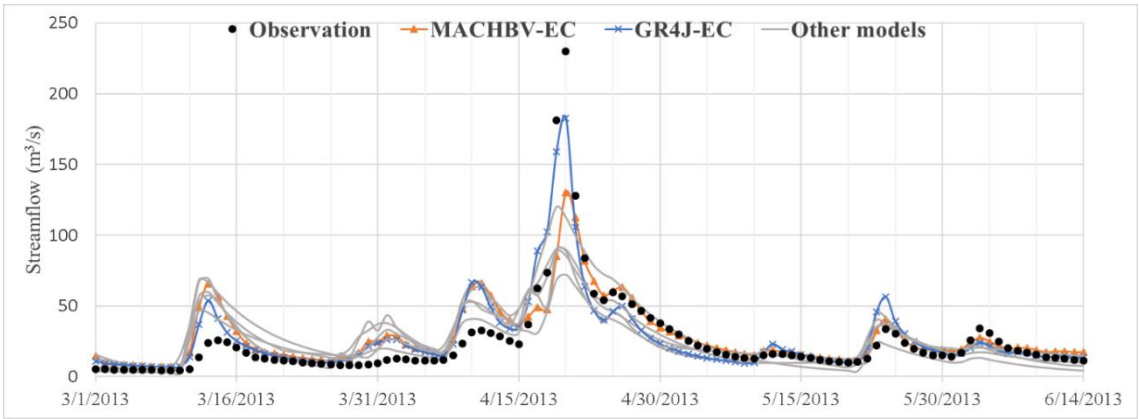
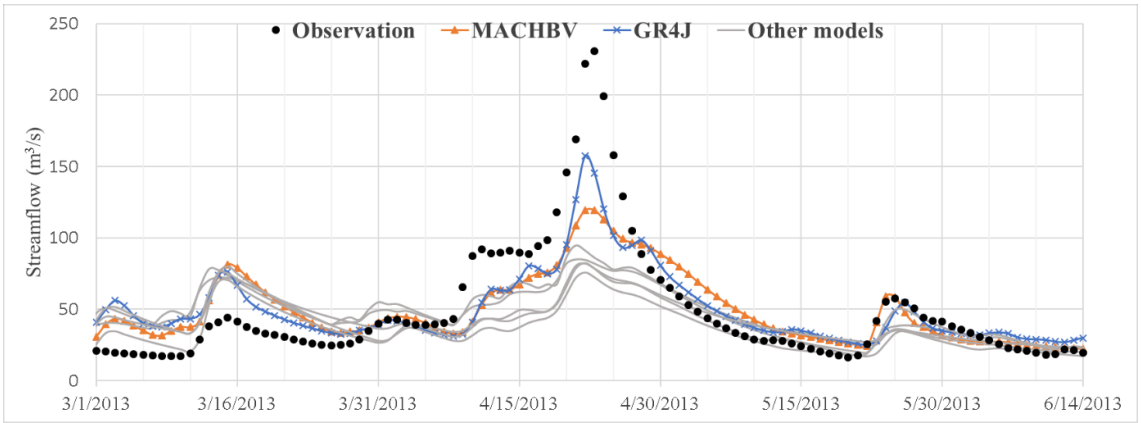


Figure 2-8 The scatter plots depicting the simulated and observed daily peak flows, greater than the 75 percentile, for the whole period (i.e. calibration and validation periods) and their fitted regression lines for both Big East River and Black River watersheds



(a) Big East River watershed



(b) Black River watershed

Figure 2-9 Observed and simulated daily runoff discharges at the outlet of (a) Black River watershed and (b) Big East River Watershed for a representative portion of the validation period

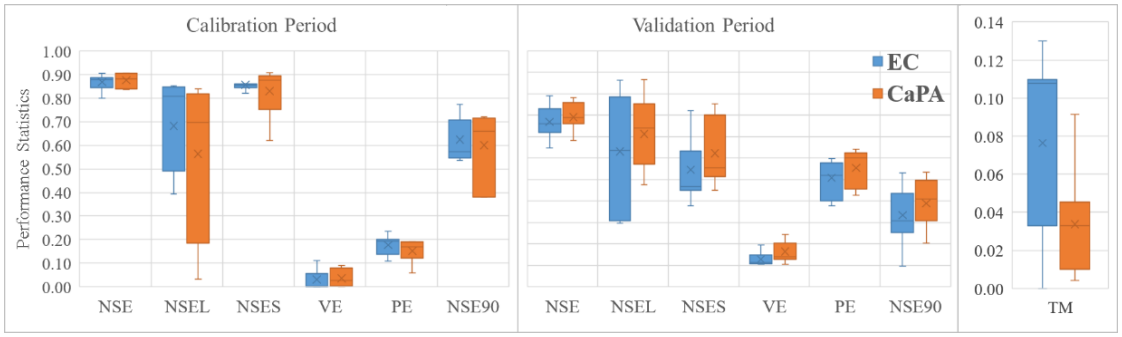
2.4.3 The effect of forcing precipitation input (CaPA versus Ground-based stations)

In the case of low-density meteorological measurements, utilizing other sources of data as forcing inputs into hydrologic models is necessary. Here, in order to evaluate the reliability of Canadian Precipitation analysis (CaPA) data, a comparison is made between the simulation results, derived from the best optimal parameter set of all hydrologic models using the two predefined input scenarios. The results, as shown in Figure 2-10, indicate

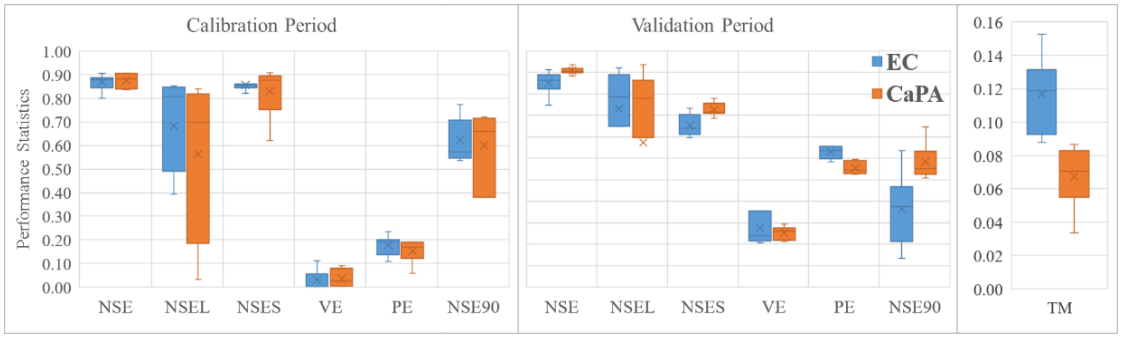
that in general, the CaPA based calibrated models performed approximately similar to the calibrated models based on EC data. However, the average transferability statistic implies that using CaPA data, in comparison to EC, provides more consistent performance for the models during the calibration and validation periods. Also, by focusing on high flows, it can be concluded that CaPA based calibrated models performed better specifically in the Black River watershed where both $NSE90$ and PE criteria of CaPA based calibrated models possess a noticeable superiority over EC based ones.

Moreover, the effects of using different precipitation scenarios on model performances are separately evaluated for various hydrologic models, using the percentage of improvement in low, medium, and high flows. These percentages are respectively defined as the percent increase in $NSEL$, NSE, and $NSES$ criteria when CaPA is used as forcing input in comparison to the EC scenario (Table 2-8). As can be seen, the effects of using CaPA as an alternative forcing precipitation input is not consistent with different hydrologic model. Regarding high flows, the results shows the advantage of using CaPA for almost all models in both Big East River and Black River watersheds, where respectively the average of 18 and 12 percentage of performance improvement occurs based on $NSES$ criterion. However, the effects of implementing CaPA on low flow performance of different models do not follow the same trend. In Big East River watershed, using CaPA leads to lower $NSEL$ of MACHBV, SACSMA, and SMARG while the low flow performance improvement of HEC-HMS based models is significantly high. On the other hand, in Black River watershed, except HEC2 and HEC3, where implementing EC and CaPA precipitation scenarios respectively results in better low flow simulation, other models' performances regarding

low flows does not significantly changes when different input scenarios are used. Besides HEC2 with 24% *NSE* based improvement, the general performances of different models (i.e. using NSE measurement) do not face significant changes in the case of using CaPA as forcing input.



(a) Big East River watershed



(b) Black River watershed

Figure 2-10 Box plots of different performance measurements for (a) Big East River and (b) Black River watersheds using best-estimated parameter set of all hydrologic models. These are derived from two input precipitation scenarios: Environment Canada (EC) and Canadian Precipitation Analysis (CaPA)

Table 2-8 The model performance statistics of different best-calibrated models by implementing EC and CaPA precipitation scenarios

Basin	Model	Obj. Func.[1]		NSE			NSEL			NSES		
		EC	CaPA	EC	CaPA	% I[2]	EC	CaPA	I (%)	EC	CaPA	% I
Big East River	MACHBV	3	3	0.73	0.76	4%[2]	0.73	0.67	-8%	0.53	0.60	13%
	SACSMA	3	1	0.66	0.70	6%	0.79	0.75	-4%	0.35	0.45	29%
	SMARG	5	5	0.63	0.69	9%	0.54	0.38	-30%	0.37	0.43	17%
	GR4J	1	2	0.79	0.78	-1%	0.86	0.87	0%	0.72	0.75	5%
	HEC1	4	1	0.55	0.58	6%	0.20	0.34	72%	0.28	0.35	26%
	HEC2	4	4	0.62	0.67	8%	0.21	0.32	55%	0.35	0.56	59%
	HEC3	1	1	0.71	0.66	-7%	0.40	0.64	59%	0.52	0.41	-21%
Black River	MACHBV	2	5	0.81	0.83	3%	0.79	0.76	-3%	0.63	0.68	7%
	SACSMA	2	1	0.76	0.82	7%	0.82	0.84	2%	0.53	0.62	15%
	SMARG	4	2	0.76	0.80	5%	0.68	0.72	6%	0.55	0.62	12%
	GR4J	2	1	0.79	0.78	-1%	0.68	0.68	-1%	0.60	0.66	8%
	HEC1	4	3	0.72	0.80	11%	0.55	0.57	3%	0.50	0.62	24%
	HEC2	5	1	0.65	0.80	24%	0.10	0.30	181%	0.54	0.58	8%
	HEC3	2	1	0.75	0.74	-2%	0.78	0.61	-22%	0.51	0.54	5%

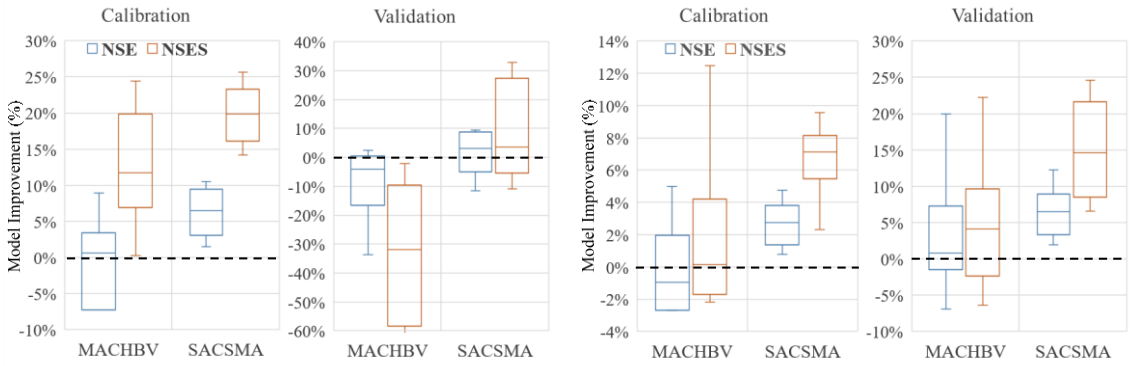
1 The objective functions lead to the best parameter sets of each model (1 = NSE; 2 = KGE; 3 = NVE; 4 = $MNVE$; 5 = $PWRMSE$).
2 I = The percentage of Improvement. Its positive values are underlined.

2.4.4 Evaluation of Snowmelt Estimation Methods: Degree-Day and SNOW17 models

For assessing the effects of implementing a more complex snowmelt routine, a comparison is made between the MACHBV and SACSMA hydrologic models in conjunction with Degree-Day (DD) and SNOW17 snowmelt estimation approaches. Four model structures, stemming from the combination of two aforementioned hydrologic models and snow modules, are calibrated using two predefined input scenarios and five aforementioned objective functions in both watersheds. In order to facilitate the evaluation, the percentages of model improvement in medium and high flows were used. The former one is defined as a percent increase in NSE when the SNOW17 model is implemented in comparison to DD

method while the latter one is based on $NSES$ criteria for focusing on high flows, specifically. Positive values of model improvement shows the advantage of using the SNOW17 method.

Figure 2-11 presents the model improvements in both calibration and validation periods using all calibrated models for each watershed. For the SACSMA, in both watersheds during both periods, the median is more than zero indicating enhancement of the model performance in the case of adapting the SNOW17 model. However, the positive effect of using the SNOW17 model, combined with the MACHBV model structure, cannot be concluded. For instance, although the model improvements during the calibration period in the Big East River watershed are almost positive, especially regarding high flows, opposite results are obtained in the validation period, where the high flow model improvement median is around -30 percent. Moreover, in order to complete the assessment, a comparison has been made between the best optimal parameter set of each model structure (Table 2-9). In line with the previous comparison, the results indicate that coupling the SACSMA and SNOW17 provides better results especially in high flows where the maximum of approximately 25% improvement occurs. However, the DD approach seems to be more appropriate to be used in conjunction with the MACHBV model structure especially in the Big East River watershed. In addition, the performance of the models regarding the whole hydrographs (i.e., model improvement based on NSE) varies between ~-2% and ~8% depicting no specific enhancement of general models' performance in the case of implementing more complex snowmelt models.



(a) Big East River (b) Black River

Figure 2-11 Box plots of the MACHBV and the SACSMA model improvements in (a) Big East River and (b) Black River watersheds using all estimated parameters of the models. The positive value of model improvement reveals the positive effect of utilizing the SNOW-17 method, while the negative value shows the advantage of the DD approach

Table 2-9 The model performance statistics of the best-calibrated SACSMA and MACHBV in conjunction with DD and SNOW-17 snowmelt methods in both watersheds

Basin	Model	Input	Snow model	Calibration			Validation			Model Improvement (%)			
										Calibration		Validation	
				NSE	NSES	PE	NSE	NSES	PE	NSE	NSES	NSE	NSES
Black River Watershed	MACHBV	EC	DD	0.90	0.90	0.14	0.81	0.63	0.40	0.3%	2.6%	1.5%	2.2%
			S17*	0.91	0.92	0.11	0.82	0.65	0.38				
		CaPA	DD	0.90	0.91	0.12	0.83	0.68	0.33	-0.3%	0.7%	-0.1%	-0.5%
			S17	0.90	0.91	0.11	0.83	0.67	0.31				
	SACSMA	EC	DD	0.88	0.85	0.19	0.76	0.53	0.46	3.3%	7.8%	7.9%	21.0%
			S17	0.91	0.92	0.11	0.82	0.65	0.38				
		CaPA	DD	0.90	0.89	0.18	0.82	0.62	0.39	-0.4%	2.1%	1.8%	9.4%
			S17	0.90	0.91	0.11	0.83	0.67	0.31				
Big East River Watersheds	MACHBV	EC	DD	0.84	0.81	0.05	0.73	0.53	0.42	-1.7%	3.6%	-3.5%	-32.8%
			S17	0.82	0.84	0.12	0.70	0.36	0.33				
		CaPA	DD	0.80	0.79	0.12	0.76	0.70	0.36	1.9%	5.7%	-7.1%	-31.1%
			S17	0.81	0.84	0.09	0.70	0.48	0.35				
	SACSMA	EC	DD	0.77	0.63	0.27	0.66	0.35	0.47	9.1%	28.5%	4.8%	12.6%
			S17	0.84	0.82	0.19	0.69	0.40	0.47				
		CaPA	DD	0.74	0.64	0.34	0.70	0.45	0.50	6.3%	20.4%	2.0%	16.6%
			S17	0.79	0.77	0.22	0.71	0.53	0.45				

* S17 is the abbreviation of SNOW17.

2.5 Discussion

Assessing the effects of different objective functions on model performances indicate that in general, using KGE and NVE as an objective function can provide more reliable estimation of different models' parameters. NVE is a combination of NSE and VE (Table 2-5) and its best value gives both the lowest difference between computed and observed flows and small volume error, (Lindström, 1997; Samuel et al., 2011). KGE also consider three measures (correlation, bias, and variability) simultaneously and provide more consistent results than NSE especially in basins where the variability of the observed flow is high (Buzacott et al., 2019; Gupta et al., 2009), such as Big East River watershed where the streamflow coefficient of variation is 1.21 compared with 1.05 in Black River. However, the two aforementioned criteria do not always lead to the best optimal parameter sets for different hydrologic models in both watersheds and will affect the performance of various models in different manners (Figure 2-5 and Table 2-6). This proves the necessity of considering multiple objective functions in model inter-comparison process in order to find more robust and comprehensive conclusions.

Inter-comparison of different conceptual hydrologic models suggests that MACHBV is the most consistent model providing reliable low, medium, and high flow estimation in both basins. This conclusion is in line with the original purpose of developing MACHBV, which was to simulate stream flows of ungauged watersheds in Ontario (Samuel et al., 2011, 2012). Also, the parsimonious GR4J model, with the lowest complexity, possess competitive performances in both watersheds, especially regarding high flow simulation (Figure 2-8 and 2-9). Besides the proven capability of GR4J model

structure in daily streamflow simulation (Wijayarathne & Coulibaly, 2020), this advantage may be due to the remarkable ability of GR4J to compensate the problem of having poor precipitation input. The parameters of GR4J model (i.e. $X1$, $X2$, and $X3$) can be relatively distorted through the calibration process in order to provide good results even with limited input data (Andréassian et al., 2001; Drogue & Khediri, 2016; Simonneaux et al., 2008). This ability is also demonstrated where along with MACHBV, GR4J possesses the lowest changes of performance in comparison to other models when CaPA is used as another source of precipitation (Table 2-8). On the other hand, compared with other conceptual models, the HEC-HMS based ones have relatively poor performances in both watersheds, which is more significant in low flow simulation. With HEC1 and HEC2, the possible reason of their poor performance, which is more pronounced than HEC3, is the low capability of the recession method in accurately estimating base flow. In addition, this may be due to the use of the fixed monthly estimated PET for HEC-HMS based models, compared with other models where the daily PET is determined during the calibration process. Therefore, evaluating the effects of using an external PET estimation model linked, and calibrated with HEC-HMS, on the accuracy of streamflow simulation is recommended.

Evaluating the effects of CaPA as another source of precipitation for both watersheds indicate that the effect of using different input scenarios is not similar for different hydrologic models and considering it in model comparison process is required for possessing robust conclusions. As previously stated, changing rainfall has the least effect on MACHBV and GR4J models (Table 2-8), proving their suitability for study regions with low data availability. Also, the reliability of CaPA as another source of precipitation

for both watersheds is confirmed by comparing the performance of different hydrologic models. CaPA based calibrated models yield better high flow estimation than using sparse ground-based measurements for model calibration (Figure 2-10 and Table 2-8). This is more obvious in Black River watershed and can be justifiable by the fact that EC based input scenario underestimates severe daily rainfall events in the Black River watershed (Figure 2-2). In line with previous studies, the results of this study show that using reliable spatially distributed data can provides more accurate mean areal precipitation estimates and affect the performances of the lumped hydrologic models in data-poor regions (Collischonn et al., 2008; Martel et al., 2020). However, more comprehensive evaluation of data with high spatial resolution and the effects of its spatial heterogeneity on hydrologic model performances required the application of a distributed or semi-distributed hydrologic models (Mazzoleni et al., 2019).

Snowmelt module is an important part of any hydrologic model in snow-dominated watersheds. Therefore, this study compared Degree-Day (DD) and more complex SNOW17 methods, relying on temperature and precipitation as the only inputs, in conjunction with two hydrologic models (i.e. MACHBV and SACSMA). In general, in line with Agnihotri and Coulibaly (2020), the results indicate the competitive performance of DD in both watersheds, which may be related to the land use characteristics of the regions, which are forested. Regarding high flows, the SNOW17 performs better than DD in Black River watershed while in Big East River, this superiority is less noticeable and diminishes from calibration to validation periods. This may be attributed to the steeper topography of the Big East River watershed and not dividing it into different elevation zones (Agnihotri

& Coulibaly, 2020; Anderson, 2006). Another possible reason for this outcome is the lack of long-term historical data in data-scarce regions. The higher degrees of freedom of SNOW17 method, compared with DDM, can lead to parameter overfitting when the length of the calibration period is not long enough due to low data availability.

2.6 Summary and Conclusions

Inter-comparison of various conceptual hydrologic models for continuous daily streamflow simulation in watersheds with low data availability is the main goal of this study. Consequently, the performance of seven lumped conceptual rainfall-runoff models with different structures (i.e., SACSMA, MACHBV, SMARG, GR4J, and three HEC-HMS based models) were compared in two data-poor and snow-dominated watersheds, Big East River and Black River, located in Northern Ontario, Canada. All models were calibrated using five different criteria (i.e. NSE, KGE, NVE, $MNVE$ and $PWRMSE$) and two different input scenarios in order to relax the influence of calibration process on the models' results.

The comparison results suggest that although the SACSMA and GR4J hydrologic models possess competitive performances, MACHBV shows the best results in simulating the daily stream flows for both watersheds. Also, the GR4J model shows the highest accuracy for high flow prediction in both watersheds. The results also indicate that the HEC-HMS based models provide lower performance, especially for low flows. From inter-comparing the different structures of the HEC-HMS models, using the soil moisture accounting and linear reservoir approaches are preferred to Deficit and Constant loss and recession methods for continuous daily streamflow simulation. Moreover, alongside model

comparison, the evaluation of the effects of using different objective functions shows that KGE and NVE are the most consistent criteria leading to reliable parameter estimation with reasonable performance regarding different parts of the hydrograph, while comparing different model structures require considering the effects of objective function selection. Furthermore, the comparison between gauged and CaPA based calibrated models in both watersheds indicates the high potential of CaPA data as a good alternative in the case of low data availability. CaPA not only provides the same level of performance in general but also leads to better results than ground-based data regarding high flows. In addition, due to the importance of accurate snowmelt estimation in snow-dominated watersheds, we compared the performance of the MACHBV and SACSMA hydrologic models in conjunction with Degree-Day method (DD) and more complex SNOW17 snowmelt estimation methods in both watersheds. In general, incorporation of SNOW17 does not significantly improve the performance of either hydrologic models. By focusing on high flows, however, the results show that implementing SNOW17 with SACSMA is consistently superior, while the DD method can perform comparably well with MACHBV.

In general, this study reveals that besides considering the effects of calibration process, utilizing different precipitation input scenarios can lead to more robust conclusion of model comparison process in data-poor watersheds. The findings of this study suggest that MACHBV and GR4J are the most robust lumped conceptual rainfall-runoff models, reacting well to poor mean areal rainfall estimation in data-scarce watersheds and performing well regarding different aspects of the hydrographs, while the SACSMA also reliably simulates streamflow in both watersheds. In addition, this study confirms the high

potential of the archived aggregated daily CaPA data to be considered as a hydrologic forcing in data-poor watersheds in Northern Ontario. Moreover, another implication of the results is that implementing the more complex SNOW17 model for snowmelt estimation in watersheds with low data availability does not always provide more reliable results, and its effectiveness depends on the hydrologic model structure.

Limitations of this study are as follows:

- The study was designed for assessing the general performance of models regarding all aspects of hydrographs (i.e. low, medium, and high flows), simultaneously. However, inter-comparing various models being calibrated regarding different particular class of flows, separately, would be advisable.
- Although the hydrologic responses of the two considered watersheds are not quite similar, the findings of this study remain applicable within the same topographic and climatologic conditions. So, we recommend further application of the proposed model inter-comparison for different types of watersheds (i.e., mountainous, semi-arid, semi-urban) with low data availability for providing more comprehensive conclusions.
- There are other conceptual models with different structures that are worth to be investigated in future studies.

The reliability of Canadian Precipitation Analysis in estimating mean areal precipitation as an input of lumped models in data-scarce regions were revealed in this study, however, further studies need to be carried out to comprehensively assessed the accuracy of spatial

heterogeneity of CaPA data in Northern Ontario, using a distributed or semi-distributed hydrologic models.

2.7 Acknowledgements

This research was supported by the Natural Science and Engineering Research Council (NSERC) of Canada, grant NSERC Canadian FloodNet [NETGP-451456]. The authors would like to thank Dr. James Leach for his support with reviewing and editing of the manuscript. We acknowledge the Ontario Ministry of Natural Resources and Forestry, Surface Water Monitoring Center for providing some of the data.

2.8 References

Adjei, K. A., Ren, L., Appiah-Adjei, E. K., & Odai, S. N. (2015). Application of satellite-derived rainfall for hydrological modelling in the data-scarce Black Volta trans-boundary basin. *Hydrology Research*, *46*(5), 777–791. https://doi.org/10.2166/nh.2014.111

Agnihotri, J., & Coulibaly, P. (2020). Evaluation of Snowmelt Estimation Techniques for Enhanced Spring Peak Flow Prediction. *Water*, *12*(5), 1290. https://doi.org/10.3390/w12051290

Almorox, J., Quej, V. H., & Martí, P. (2015). Global performance ranking of temperature-based approaches for evapotranspiration estimation considering Köppen climate classes. *Journal of Hydrology*, *528*, 514–522. https://doi.org/10.1016/j.jhydrol.2015.06.057

American Society of Civil Engineers. (1996). *Hydrology handbook*. ASCE.

Anderson, E. A. (1973). *National Weather Service river forecast system: Snow accumulation and ablation model*. U.S. DEPARTMENT OF COMMERCE: National Oceanic and Atmospheric Administration, National Weather Service.

Anderson, E. A. (2006). *Snow Accumulation and Ablation Model – SNOW-17*. Natl. Ocean. Atmospheric Adm. Natl. Weather Serv. Silver Springs MD. https://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/part2/_pdf/22snow17.pdf

Andréassian, V., Perrin, C., Michel, C., Usart-Sanchez, I., & Lavabre, J. (2001). Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models. *Journal of Hydrology*, *250*(1), 206–223. https://doi.org/10.1016/S0022-1694(01)00437-1

Anshuman, A., Kunnath-Poovakka, A., & Eldho, T. I. (2019). Towards the use of conceptual models for water resource assessment in Indian tropical watersheds under monsoon-driven climatic conditions. *Environmental Earth Sciences*, *78*(9), 282. https://doi.org/10.1007/s12665-019-8281-5

Arsenault, R., Poulin, A., Côté, P., & Brissette, F. (2014). Comparison of Stochastic Optimization Algorithms in Hydrological Model Calibration. *Journal of Hydrologic Engineering*, *19*(7), 1374–1384. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000938

Ashofteh, P.-S., Rajaee, T., & Golfam, P. (2017). Assessment of Water Resources Development Projects under Conditions of Climate Change Using Efficiency Indexes (EIs). *Water Resources Management*, *31*(12), 3723–3744. https://doi.org/10.1007/s11269-017-1701-y

Bergström, S. (1976). Development and Application of a Conceptual Runoff Model for Scandinavian Catchments. *Lund,Sweden: Lund Institute of Technology/Univ. of Lund, A*, *52*. https://www.researchgate.net/publication/255274162_Development_and_Application_of_a_Conceptual_Runoff_Model_for_Scandinavian_Catchments

Beven, K. J. (2011). *Rainfall-Runoff Modelling: The Primer*. John Wiley & Sons.

Boluwade, A., Zhao, K.-Y., Stadnyk, T. A., & Rasmussen, P. (2018). Towards validation of the Canadian precipitation analysis (CaPA) for hydrologic modeling applications

in the Canadian Prairies. *Journal of Hydrology*, *556*, 1244–1255. https://doi.org/10.1016/j.jhydrol.2017.05.059

Bowling, L. C., Lettenmaier, D. P., Nijssen, B., Graham, L. P., Clark, D. B., El Maayar, M., Essery, R., Goers, S., Gusev, Y. M., Habets, F., van den Hurk, B., Jin, J., Kahan, D., Lohmann, D., Ma, X., Mahanama, S., Mocko, D., Nasonova, O., Niu, G.-Y., … Yang, Z.-L. (2003). Simulation of high-latitude hydrological processes in the Torne–Kalix basin: PILPS Phase 2(e): 1: Experiment description and summary intercomparisons. *Global and Planetary Change*, *38*(1), 1–30. https://doi.org/10.1016/S0921-8181(03)00003-1

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., & Viney, N. R. (2009). Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use. *Advances in Water Resources*, *32*(2), 129–146. https://doi.org/10.1016/j.advwatres.2008.10.003

Burnash, R. J. C., Ferral, R. L., & McGuire, R. A. (1973). *A generalized streamflow simulation system: Conceptual modeling for digital computers*. Joint Federal-State River Forecast Center, United States National Weather Service.

Buzacott, A. J. V., Tran, B., van Ogtrop, F. F., & Vervoort, R. W. (2019). Conceptual Models and Calibration Performance—Investigating Catchment Bias. *Water*, *11*(11), 2424. https://doi.org/10.3390/w11112424

Caldwell, P. V., Kennen, J. G., Sun, G., Kiang, J. E., Butcher, J. B., Eddy, M. C., Hay, L. E., LaFontaine, J. H., Hain, E. F., Nelson, S. A. C., & McNulty, S. G. (2015). A comparison of hydrologic models for ecological flows and water availability. *Ecohydrology*, *8*(8), 1525–1546. https://doi.org/10.1002/eco.1602

Chiew, F. H. S., Stewardson, M. J., & McMahon, T. A. (1993). Comparison of six rainfall-runoff modelling approaches. *Journal of Hydrology*, *147*(1), 1–36. https://doi.org/10.1016/0022-1694(93)90073-I

Choi, W., Kim, S. J., Rasmussen, P. F., & Moore, A. R. (2009). Use of the North American Regional Reanalysis for Hydrological Modelling in Manitoba. *Canadian Water Resources Journal / Revue Canadienne Des Ressources Hydriques*, *34*(1), 17–36. https://doi.org/10.4296/cwrj3401017

Collischonn, B., Collischonn, W., & Tucci, C. E. M. (2008). Daily hydrological modeling in the Amazon basin using TRMM rainfall estimates. *Journal of Hydrology*, *360*(1), 207–216. https://doi.org/10.1016/j.jhydrol.2008.07.032

Cunderlik, J., & Simonovic, S. (2004). *Calibration, Verification and Sensitivity Analysis of the HEC-HMS Hydrologic Model*. Department of Civil and Environmental Engineering, The University of Western Ontario. https://ir.lib.uwo.ca/wrrr/11

Darbandsari, P., & Coulibaly, P. (2019). Inter-Comparison of Different Bayesian Model Averaging Modifications in Streamflow Simulation. *Water*, *11*(8), 1707. https://doi.org/10.3390/w11081707

Dariane, A. B., Javadianzadeh, M. M., & James, L. D. (2016). Developing an Efficient Auto-Calibration Algorithm for HEC-HMS Program. *Water Resources Management*, *30*(6), 1923–1937. https://doi.org/10.1007/s11269-016-1260-7

Das, T., Bárdossy, A., Zehe, E., & He, Y. (2008). Comparison of conceptual model performance using different representations of spatial variability. *Journal of Hydrology*, *356*(1), 106–118. https://doi.org/10.1016/j.jhydrol.2008.04.008

Day, G. N. (1985). Extended Streamflow Forecasting Using NWSRFS. *Journal of Water Resources Planning and Management*, *111*(2), 157–170. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)

Debele, B., Srinivasan, R., & Gosain, A. K. (2010). Comparison of Process-Based and Temperature-Index Snowmelt Modeling in SWAT. *Water Resources Management*, *24*(6), 1065–1088. https://doi.org/10.1007/s11269-009-9486-2

Dile, Y. T., & Srinivasan, R. (2014). Evaluation of CFSR climate data for hydrologic prediction in data-scarce watersheds: An application in the Blue Nile River Basin.

JAWRA Journal of the American Water Resources Association, *50*(5), 1226–1241. https://doi.org/10.1111/jawr.12182

Donnelly-Makowecki, L. M., & Moore, R. D. (1999). Hierarchical testing of three rainfall–runoff models in small forested catchments. *Journal of Hydrology*, *219*(3), 136–152. https://doi.org/10.1016/S0022-1694(99)00056-6

Drogue, G., & Khediri, W. B. (2016). Catchment model regionalization approach based on spatial proximity: Does a neighbor catchment-based rainfall input strengthen the method? *Journal of Hydrology: Regional Studies*, *8*, 26–42. https://doi.org/10.1016/j.ejrh.2016.07.002

Edijanto, Nascimento, N. D. O., Yang, X., Makhlouf, Z., & Michel, C. (1999). GR3J: A daily watershed model with three free parameters. *Hydrological Sciences Journal*, *44*(2), 263–277. https://doi.org/10.1080/02626669909492221

Essery, R., Morin, S., Lejeune, Y., & B Ménard, C. (2013). A comparison of 1701 snow models using observations from an alpine site. *Advances in Water Resources*, *55*, 131–148. https://doi.org/10.1016/j.advwatres.2012.07.013

Eum, H.-I., Dibike, Y., Prowse, T., & Bonsal, B. (2014). Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the Athabasca Watershed, Canada. *Hydrological Processes*, *28*(14), 4250–4271. https://doi.org/10.1002/hyp.10236

Feldman, A. D. (2000). *Hydrologic Modeling System HEC-HMS Technical Reference Manual*. U.S. Army Corps of Engineers Hydrologic Engineering Center.

Fortin, V., Roy, G., Stadnyk, T., Koenig, K., Gasset, N., & Mahidjiba, A. (2018). Ten Years of Science Based on the Canadian Precipitation Analysis: A CaPA System Overview and Literature Review. *Atmosphere-Ocean*, *56*(3), 178–196. https://doi.org/10.1080/07055900.2018.1474728

Fuka, D. R., Walter, M. T., MacAlister, C., Degaetano, A. T., Steenhuis, T. S., & Easton, Z. M. (2014). Using the Climate Forecast System Reanalysis as weather input data

for watershed models. *Hydrological Processes*, *28*(22), 5613–5623. https://doi.org/10.1002/hyp.10073

Gaborit, É., Fortin, V., Tolson, B., Fry, L., Hunter, T., & Gronewold, A. D. (2017). Great Lakes Runoff Inter-comparison Project, phase 2: Lake Ontario (GRIP-O). *Journal of Great Lakes Research*, *43*(2), 217–227. https://doi.org/10.1016/j.jglr.2016.10.004

Gan, T. Y., Dlamini, E. M., & Biftu, G. F. (1997). Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. *Journal of Hydrology*, *192*(1), 81–103. https://doi.org/10.1016/S0022-1694(96)03114-9

Gan, T. Y., Gusev, Y., Burges, S. J., & Nasonova, O. (2006). *Performance comparison of a complex physics- based land surface model and a conceptual, lumped-parameter hydrological model at the basin-scale*. 196–207.

Garavaglia, F., Lay, M. L., Gottardi, F., Garçon, R., Gailhard, J., Paquet, E., & Mathevet, T. (2017). Impact of model structure on flow simulation and hydrological realism: From a lumped to a semi-distributed approach. *Hydrology and Earth System Sciences*, *21*(8), 3937–3952. https://doi.org/10.5194/hess-21-3937-2017

Grayson, R. B., Blöschl, G., Western, A. W., & McMahon, T. A. (2002). Advances in the use of observed spatial patterns of catchment hydrological response. *Advances in Water Resources*, *25*(8), 1313–1334. https://doi.org/10.1016/S0309-1708(02)00060-X

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

Gyawali, R., & Watkins, D. W. (2013). Continuous Hydrologic Modeling of Snow-Affected Watersheds in the Great Lakes Basin Using HEC-HMS. *Journal of Hydrologic Engineering*, *18*(1), 29–39. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000591

Habib, E., Haile, A. T., Sazib, N., Zhang, Y., & Rientjes, T. (2014). Effect of Bias Correction of Satellite-Rainfall Estimates on Runoff Simulations at the Source of the Upper Blue Nile. *Remote Sensing*, *6*(7), 6688–6708. https://doi.org/10.3390/rs6076688

Han, S., Coulibaly, P., & Biondi, D. (2019). Assessing Hydrologic Uncertainty Processor Performance for Flood Forecasting in a Semiurban Watershed. *Journal of Hydrologic Engineering*, *24*(9), 05019025. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001828

Hargreaves, G. H., & Samani, Z. A. (1985). Reference Crop Evapotranspiration from Temperature. *Applied Engineering in Agriculture*, *1*(2), 96–99.

Khalida, F., Scharffenberg, W. A., & Bombardelli, F. A. (2014). Assessment of the Melt Rate Function in a Temperature Index Snow Model Using Observed Data. *Journal of Hydrologic Engineering*, *19*(7), 1275–1282. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000925

Koch, J., Cornelissen, T., Fang, Z., Bogena, H., Diekkrüger, B., Kollet, S., & Stisen, S. (2016). Inter-comparison of three distributed hydrological models with respect to seasonal variability of soil moisture patterns at a small forested catchment. *Journal of Hydrology*, *533*, 234–249. https://doi.org/10.1016/j.jhydrol.2015.12.002

Lakew, H. B., Moges, S. A., & Asfaw, D. H. (2020). Hydrological performance evaluation of multiple satellite precipitation products in the upper Blue Nile basin, Ethiopia. *Journal of Hydrology: Regional Studies*, *27*, 100664. https://doi.org/10.1016/j.ejrh.2020.100664

Lespinas, F., Fortin, V., Roy, G., Rasmussen, P., & Stadnyk, T. (2015). Performance Evaluation of the Canadian Precipitation Analysis (CaPA). *Journal of Hydrometeorology*, *16*(5), 2045–2064. https://doi.org/10.1175/JHM-D-14-0191.1

Li, F., Xu, Z., Liu, W., & Zhang, Y. (2014). The impact of climate change on runoff in the Yarlung Tsangpo River basin in the Tibetan Plateau. *Stochastic Environmental*

Research and Risk Assessment, *28*(3), 517–526. https://doi.org/10.1007/s00477-013-0769-z

Liang, G. C. (1992). *A note on the revised SMAR model*. Memorandum to the River Flow Forecasting Workshop Group, Department of Engineering Hydrology, University College Galway. https://ci.nii.ac.jp/naid/10018251872/

Lindström, G. (1997). A Simple Automatic Calibration Routine for the HBV Model. *Hydrology Research*, *28*(3), 153–168. https://doi.org/10.2166/nh.1997.0009

Lü, H., Hou, T., Horton, R., Zhu, Y., Chen, X., Jia, Y., Wang, W., & Fu, X. (2013). The streamflow estimation using the Xinanjiang rainfall runoff model and dual state-parameter estimation method. *Journal of Hydrology*, *480*, 102–114. https://doi.org/10.1016/j.jhydrol.2012.12.011

Mahfouf, J.-F., Brasnett, B., & Gagnon, S. (2007). A Canadian precipitation analysis (CaPA) project: Description and preliminary results. *Atmosphere-Ocean*, *45*(1), 1–17. https://doi.org/10.3137/ao.v450101

Martel, J.-L., Brissette, F., & Poulin, A. (2020). Impact of the spatial density of weather stations on the performance of distributed and lumped hydrological models. *Canadian Water Resources Journal / Revue Canadienne Des Ressources Hydriques*, *45*(2), 158–171. https://doi.org/10.1080/07011784.2020.1729241

Matott, L. S. (2005). *OSTRICH: An optimization software tool: Documentation and users guide*. University at Buffalo.

Mazzoleni, M., Brandimarte, L., & Amaranto, A. (2019). Evaluating precipitation datasets for large-scale distributed hydrological modelling. *Journal of Hydrology*, *578*, 124076. https://doi.org/10.1016/j.jhydrol.2019.124076

Michaud, J., & Sorooshian, S. (1994). Comparison of simple versus complex distributed runoff models on a midsized semiarid watershed. *Water Resources Research*, *30*(3), 593–605. https://doi.org/10.1029/93WR03218

Moradkhani, H., & Sorooshian, S. (2008). General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis. In *Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrological Models* (pp. 1–24). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77843-1_1

Nash, J. E. (1957). The form of the instantaneous unit hydrograph. *Comptes Rendus et Rapports Assemblee Generale de Toronto*, *3*, 114–121. http://nora.nerc.ac.uk/id/eprint/508550/

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

O'Connell, P. E., Nash, J. E., & Farrell, J. P. (1970). River flow forecasting through conceptual models part II - The Brosna catchment at Ferbane. *Journal of Hydrology*, *10*(4), 317–329. https://doi.org/10.1016/0022-1694(70)90221-0

Ouermi, K. S., Paturel, J.-E., Adounpke, J., Lawin, A. E., Goula, B. T. A., & Amoussou, E. (2019). Comparison of hydrological models for use in climate change studies: A test on 241 catchments in West and Central Africa. *Comptes Rendus Geoscience*, *351*(7), 477–486. https://doi.org/10.1016/j.crte.2019.08.001

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*(1), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Price, K., Purucker, S. T., Kraemer, S. R., Babendreier, J. E., & Knightes, C. D. (2014). Comparison of radar and gauge precipitation data in watershed models across varying spatial and temporal scales. *Hydrological Processes*, *28*(9), 3505–3520. https://doi.org/10.1002/hyp.9890

Razavi, T., & Coulibaly, P. (2017). An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged

watersheds. *Canadian Water Resources Journal / Revue Canadienne Des Ressources Hydriques*, *42*(1), 2–20. https://doi.org/10.1080/07011784.2016.1184590

Razmkhah, H., Saghafian, B., Ali, A.-M. A., & Radmanesh, F. (2016). Rainfall-runoff modeling considering soil moisture accounting algorithm, case study: Karoon III River basin. *Water Resources*, *43*(4), 699–710. https://doi.org/10.1134/S0097807816040072

Refsgaard, J. C., & Knudsen, J. (1996). Operational Validation and Intercomparison of Different Types of Hydrological Models. *Water Resources Research*, *32*(7), 2189–2202. https://doi.org/10.1029/96WR00896

Reggiani, P., Renner, M., Weerts, A. H., & Gelder, P. A. H. J. M. van. (2009). Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. *Water Resources Research*, *45*(2). https://doi.org/10.1029/2007WR006758

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2011). Estimation of Continuous Streamflow in Ontario Ungauged Basins: Comparison of Regionalization Methods. *Journal of Hydrologic Engineering*, *16*(5), 447–459. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000338

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2012). Identification of rainfall–runoff model for improved baseflow estimation in ungauged basins. *Hydrological Processes*, *26*(3), 356–366. https://doi.org/10.1002/hyp.8133

Scharffenberg, W. (2016). *HEC-HMS User's Manual, Version 4.2*. U.S. Army Corps of Engineers Institute for Water Resources Hydrologic Engineering Center (CEIWR-HEC).

Seiller, G., Anctil, F., & Perrin, C. (2012). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrology and Earth System Sciences*, *16*(4), 1171–1189. https://doi.org/10.5194/hess-1116-1171-2012

Seo, D.-J., Koren, V., & Cajina, N. (2003). Real-Time Variational Assimilation of Hydrologic and Hydrometeorological Data into Operational Hydrologic Forecasting. *Journal of Hydrometeorology*, *4*(3), 627–641. https://doi.org/10.1175/1525-7541(2003)004<0627:RVAOHA>2.0.CO;2

Shamir, E., & Georgakakos, K. P. (2006). Distributed snow accumulation and ablation modeling in the American River basin. *Advances in Water Resources*, *29*(4), 558–570. https://doi.org/10.1016/j.advwatres.2005.06.010

Shi, P., Chen, C., Srinivasan, R., Zhang, X., Cai, T., Fang, X., Qu, S., Chen, X., & Li, Q. (2011). Evaluating the SWAT Model for Hydrological Modeling in the Xixian Watershed and a Comparison with the XAJ Model. *Water Resources Management*, *25*(10), 2595–2612. https://doi.org/10.1007/s11269-011-9828-8

Shrestha, D. L. (2009). *Uncertainty analysis in rainfall-runoff modelling - application of machine learning techniques: UNESCO-IHE PhD thesis.* [PhD. thesis, IHE Delft Institute for Water Education]. https://www.cabdirect.org/cabdirect/abstract/20123116250

Shrestha, D. L., & Solomatine, D. P. (2009). Assessing uncertainty in rainfall-runoff models: Application of data-driven models. In *Flood Risk Management: Research and Practice* (pp. 1563–1573). CRC Press.

Simonneaux, V., Hanich, L., Boulet, G., & Thomas, S. (2008). *Modelling runoff in the Rheraya Catchment (High Atlas, Morocco) using the simple daily model GR4J. Trends over the last decades.* 13th IWRA World Water Congress, Montpellier, France.

Sirisena, T. A. J. G., Maskey, S., Ranasinghe, R., & Babel, M. S. (2018). Effects of different precipitation inputs on streamflow simulation in the Irrawaddy River Basin, Myanmar. *Journal of Hydrology: Regional Studies*, *19*, 265–278. https://doi.org/10.1016/j.ejrh.2018.10.005

Smith, M. B., Seo, D.-J., Koren, V. I., Reed, S. M., Zhang, Z., Duan, Q., Moreda, F., & Cong, S. (2004). The distributed model intercomparison project (DMIP): Motivation

and experiment design. *Journal of Hydrology*, *298*(1), 4–26. https://doi.org/10.1016/j.jhydrol.2004.03.040

Sorooshian, S., Duan, Q., & Gupta, V. K. (1993). Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting Model. *Water Resources Research*, *29*(4), 1185–1194. https://doi.org/10.1029/92WR02617

Srivastava, A., Deb, P., & Kumari, N. (2020). Multi-Model Approach to Assess the Dynamics of Hydrologic Components in a Tropical Ecosystem. *Water Resources Management*, *34*(1), 327–341. https://doi.org/10.1007/s11269-019-02452-z

Suliman, A. H. A., Jajarmizadeh, M., Harun, S., & Mat Darus, I. Z. (2015). Comparison of Semi-Distributed, GIS-Based Hydrological Models for the Prediction of Streamflow in a Large Catchment. *Water Resources Management*, *29*(9), 3095–3110. https://doi.org/10.1007/s11269-015-0984-0

Tan, B. Q., & O'Connor, K. M. (1996). Application of an empirical infiltration equation in the SMAR conceptual model. *Journal of Hydrology*, *185*(1), 275–295. https://doi.org/10.1016/0022-1694(95)02993-1

Te Linde, A. H., Aerts, J. C. J. H., Hurkmans, R. T. W. L., & Eberle, M. (2008). Comparing model performance of two rainfall-runoff models in the Rhine basin using different atmospheric forcing data sets. *Hydrology and Earth System Sciences*, *12*(3), 943–957. https://doi.org/10.5194/hess-12-943-2008

Tegegne, G., Park, D. K., Kim, Y., & Kim, Y.-O. (2017). Selecting hydrologic modelling approaches for water resource assessment in the Yongdam watershed. *Journal of Hydrology (New Zealand)*, *56*(2), 155.

Tegegne, G., Park, D. K., & Kim, Y.-O. (2017). Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. *Journal of Hydrology: Regional Studies*, *14*, 49–66. https://doi.org/10.1016/j.ejrh.2017.10.002

Teng, F., Huang, W., & Ginis, I. (2018). Hydrological modeling of storm runoff and snowmelt in Taunton River Basin by applications of HEC-HMS and PRMS models. *Natural Hazards, 91*(1), 179–199. https://doi.org/10.1007/s11069-017-3121-y

Thiessen, A. H. (1911). Precipitation averages for large areas. *Monthly Weather Review*, *39*(7), 1082–1089. https://doi.org/10.1175/1520-0493(1911)39<1082b:PAFLA>2.0.CO;2

Thornthwaite, C. W. (1948). An Approach toward a Rational Classification of Climate. *Geographical Review*, *38*(1), 55–94. JSTOR. https://doi.org/10.2307/210739

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Troin, M., Arsenault, R., & Brissette, F. (2015). Performance and Uncertainty Evaluation of Snow Models on Snowmelt Flow Simulations over a Nordic Catchment (Mistassibi, Canada). *Hydrology*, *2*(4), 289–317. https://doi.org/10.3390/hydrology2040289

U.S. Army Corps of Engineers. (1991). *SSARR model, streamflow synthesis and reservoir regulation*. US Army Corps of Engineers.

Vansteenkiste, T., Tavakoli, M., Van Steenbergen, N., De Smedt, F., Batelaan, O., Pereira, F., & Willems, P. (2014). Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation. *Journal of Hydrology*, *511*, 335–349. https://doi.org/10.1016/j.jhydrol.2014.01.050

Vrugt, J. A., Gupta, H. V., Nualláin, B., & Bouten, W. (2006). Real-Time Data Assimilation for Operational Ensemble Streamflow Forecasting. *Journal of Hydrometeorology*, *7*(3), 548–565. https://doi.org/10.1175/JHM504.1

Wijayarathne, D. B., & Coulibaly, P. (2020). Identification of hydrological models for operational flood forecasting in St. John's, Newfoundland, Canada. *Journal of Hydrology: Regional Studies*, *27*, 100646. https://doi.org/10.1016/j.ejrh.2019.100646

Wöhling, T., Samaniego, L., & Kumar, R. (2013). Evaluating multiple performance criteria to calibrate the distributed hydrological model of the upper Neckar catchment. *Environmental Earth Sciences*, *69*(2), 453–468. https://doi.org/10.1007/s12665-013-2306-2

Worqlul, A. W., Yen, H., Collick, A. S., Tilahun, S. A., Langan, S., & Steenhuis, T. S. (2017). Evaluation of CFSR, TMPA 3B42 and ground-based rainfall data as input for hydrological models, in data-scarce regions: The upper Blue Nile Basin, Ethiopia. *CATENA*, *152*, 242–251. https://doi.org/10.1016/j.catena.2017.01.019

Young, P. C. (2002). Advances in real-time flood forecasting. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *360*(1796), 1433–1450. https://doi.org/10.1098/rsta.2002.1008

Zhang, L., Xin, J., He, C., Zhang, B., Zhang, X., Li, J., Zhao, C., Tian, J., & DeMarchi, C. (2016). Comparison of SWAT and DLBRM for Hydrological Modeling of a Mountainous Watershed in Arid Northwest China. *Journal of Hydrologic Engineering*, *21*(5), 04016007. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001313

Chapter 3. Inter-Comparison of Different Bayesian Model Averaging Modifications in Streamflow Simulation

Summary of Paper 2: Darbandsari, P., & Coulibaly, P. (2019). Inter-comparison of different Bayesian model averaging modifications in streamflow simulation. Water, 11(8), 1707.

In the context of streamflow predictions, this research work aims at evaluating the effects of various previously recommended Bayesian Model Averaging (BMA) modifications, including the implementation of different data transformation approaches, various distribution types, heteroscedastic variance, and different BMA parameter estimation methods on the reliability and accuracy of BMA predictive results.

Key findings of this research include:

- The contributions of different members of the ensemble in the BMA final results are not always in accordance with their individual performances, which shows the significant importance of establishing an ensemble with independent members, capturing the whole observational variability.
- The expectation-maximization algorithm is a robust optimization method for reliably estimating the original BMA parameters.
- The application of the non-constant (i.e. heteroscedastic) variance enhances the capability of the BMA method for quantifying predictive uncertainty, especially for high streamflow values.

- Applying the data transformation method, in general, leads to more reliable predictive results while it reduces the sharpness of the probabilistic high flow streamflow predictions.
- The effects of employing more representative distribution types in the BMA formulation are marginal.
- The combination of data transformation approach and non-constant variance yields under confident results with large width of confidence interval bounds in high flows.

3.1 Abstract

Bayesian model averaging (BMA) is a popular method using the advantages of forecast ensemble to enhance the reliability and accuracy of predictions. The inherent assumptions of the classical BMA has led to different variants. However, there is not a comprehensive examination of how these solutions improve the original BMA in the context of streamflow simulation. In this study, a scenario-based analysis was conducted for assessment of various modifications and how they affect BMA results. The evaluated modifications included using various streamflow ensembles, data transformation procedures, distribution types, standard deviation forms, and optimization methods. We applied the proposed analysis in two data-poor watersheds located in northern Ontario, Canada. The results indicate that using more representative distribution types do not significantly improve BMA-derived results, while the positive effect of implementing non-constant variance on BMA probabilistic performance cannot be ignored. Also, higher reliability was obtained by applying a data transformation procedure; however, it can reduce the results' sharpness significantly. Moreover, although considering many streamflow simulations as ensemble

members does not always enhance BMA results, using different forcing precipitation scenarios besides multi-models led to better BMA-based probabilistic simulations in data-poor watersheds. Also, the reliability of the expectation-maximization algorithm in estimating BMA parameters was confirmed.

Keywords: Bayesian model averaging; multi-model ensemble hydrologic simulation; uncertainty analysis; Canada

3.2 Introduction

Different types of hydrologic models, varying from empirical and conceptual to fully distributed physically based models, have been developed in order to increase the accuracy of hydrological forecasts. However, none of these models describe all aspects of hydrological processes sufficiently and without avoiding errors. Therefore, it remains difficult to choose one of them as superior in all conditions (Chen et al., 2013; Z. Liu et al., 2016).

Different uncertainties in rainfall-runoff modeling, arising mostly from parameters, inputs, and the structure of the model (Moradkhani & Sorooshian, 2008; Shrestha, 2009), need to be quantified reliably and accurately as possible. This can be done by generating a streamflow ensemble system (Madadgar & Moradkhani, 2014; Michaels, 2015; Seo et al., 2006). Although using streamflow ensemble based on multi-input and multi-parameter sets can enhance the uncertainty quantification process, it cannot address the uncertainty within a single hydrologic model structure (i.e., model structural uncertainty) (Georgakakos et al., 2004; Vrugt & Robinson, 2007). Consequently, in recent years, some multi-model

approaches have been developed in order to find more reliable results by combining multiple model forecasts.

The model averaging approaches can be divided into two main groups. The first one includes methods leading to a one-point deterministic result by using the weighted average of the deterministic model forecasts or simulations, such as simple model averaging, Granger–Ramanathan averaging (Granger & Ramanathan, 1984), and artificial neural network (ANN) methods (Shamseldin et al., 1997; Shamseldin & O'Connor, 1999). The second group contains combination techniques like Bayesian model averaging (BMA) (Hoeting et al., 1999; Raftery, 1993; Raftery et al., 1997, 2005) which quantify the predictive uncertainty and provide probabilistic results. In the BMA method, individual models are weighted using their likelihood measures and probabilistic results are generated by combining the probability distribution of various individual forecasts. It has been shown that BMA is one of the most promising multi-model combination approaches in producing more reliable and accurate results in comparison to the other methods (Arsenault et al., 2015; Raftery et al., 2005; Viallefont et al., 2001).

There are many different fields, from medicine to management, where the BMA method is applied (Tian et al., 2014). Bayesian model averaging has been largely used in meteorology (Liu & Xie, 2014; Ma et al., 2018; Raftery et al., 2005; Sloughter et al., 2007; Sun et al., 2018). In recent years, the BMA approach has been applied in various water resources and hydrologic studies ranging from groundwater modeling (Neuman, 2003; Rojas et al., 2008; Zeng et al., 2016) to flood frequency analysis (Yan & Moradkhani, 2016). Moreover, various studies have successfully applied the BMA method in the field of hydrological

modelling (Ajami et al., 2007; Dong et al., 2013; Duan et al., 2007; Huo et al., 2019; Liang et al., 2013; Najafi & Moradkhani, 2016; Qu et al., 2017; Yen et al., 2014).

There are some potential issues and limitations for the standard Bayesian model averaging approach. One of the main assumptions of the classic BMA methodology is estimation of forecast posterior probability distribution by a Gaussian function. It has been raised that this assumption leads to inappropriate results in the case of non-normal data, such as streamflow or precipitation where skewed distributions (e.g., gamma) are more representative. This has motivated some research to relax this assumption by considering different types of distributions (Sloughter et al., 2007; Vrugt & Robinson, 2007) or applying a data transformation procedure in order to generate approximately normal data (Duan et al., 2007; Z. Liang et al., 2013; Qu et al., 2017; Todini, 2008; Yan & Moradkhani, 2016). Additionally, in the original BMA, a single constant variance for conditional probability distribution functions (PDFs) is implemented. This seems to be unsuitable for streamflow data where the larger errors are expected regarding high flows. Consequently, some studies proposed considering heteroscedastic (non-constant) variance changing monotonically with the flow level in order to enhance the predictive performance of the BMA model (Vrugt, 2016; Vrugt & Robinson, 2007). Although a significant number of studies tried to reduce the effect of the aforementioned assumptions, none have comprehensively assessed the sensitivity of BMA methodology in applying various aforementioned modifications and how they affect BMA final probabilistic results.

Moreover, in the original BMA method, the expectation-maximization (EM) algorithm (McLachlan & Krishnan, 2008) was proposed to find the optimal values of BMA

parameters. However, it is argued that the EM algorithm is not always able to find the global solution properly and the final solution is sensitive to the initial values (Duan et al., 2007; Raftery et al., 2005; Sloughter et al., 2007; Vrugt & Robinson, 2007). As a result, some studies have proposed replacing the EM algorithm with other global optimization techniques for possessing more reliable solutions solutions (Ebtehaj et al., 2010; Vrugt et al., 2008; Vrugt & Robinson, 2007), while no studies have assessed how the accuracy and reliability of the BMA results are influenced by this modification.

Furthermore, the streamflow ensemble for BMA application can be derived in various ways, such as utilizing different hydrologic models (Duan et al., 2007; Zhang et al., 2009), considering various forcing inputs scenarios (Liang et al., 2013; Neto et al., 2018; Strauch et al., 2012), or using different parameter sets of each hydrologic model (Dong et al., 2013). It has been claimed that a high number of members in the ensemble does not always increase the potential ability of the BMA method (Madadgar & Moradkhani, 2014; Neuman, 2003). However, there is no thorough evaluation of how an ensemble generated from different sources can affect the performance of the BMA method.

Although some studies have proposed more complicated BMA-based methods (i.e., GLUE-BMA (Rojas et al., 2008), BMA-PF (Parrish et al., 2012, Cop-BMA (Madadgar & Moradkhani, 2014), and CBP-BMA (He et al., 2018)), there are still many studies being done using the original BMA approach based on the aforementioned modifications. Consequently, the need of a comprehensive assessment of the different BMA variants is strongly felt. This study aims to fill this gap by closely evaluating how the various previously recommended modifications affect the accuracy and reliability of the BMA-

generated probabilistic results. The conclusions are expected to contribute toward the improvement of the knowledge of different BMA variants dealing with streamflow simulations and forecasting and provide practical and useful recommendations about the effectiveness of various modifications. The organization of this paper is as follows: Section 3.3 elaborates on all materials and methods used in this study, including the study areas and data, the standard BMA method and its various components, the proposed BMA scenario-based analysis, the different hydrologic models, and the evaluation performance statistics. In Section 3.4, the inter-comparison results of the proposed BMA modifications are presented and discussed, and, finally, a summary and conclusion section are provided.

3.3 Materials and Methods

3.3.1 Study Area and Data

The Big East River (620 km^2) and the Black River (1522 km^2) watersheds, located in the northern part of Ontario, Canada, are chosen for the implementation of the proposed BMA scenario-based analysis (Figure 3-1). Both basins are mostly forested regions and their landscapes are moderately sloped with mean elevations of 450 and 300 meters above sea level for the Big East River and Black River watersheds, respectively. The historical daily streamflow data at the outlet of both watersheds (the only hydrometric station of each watershed) illustrate that high flows mostly occur in April when the snowmelt process plays an important role. Moreover, as can be seen from Figure 3-1, the only six available Environment Canada (EC) meteorological stations with reliable and sufficient historical data are located outside the boundaries of both watersheds. This represents an actual condition of watersheds with limited data availability. Analysis of the precipitation and

temperature time-series of these six stations approximately shows the annual mean precipitation and the daily average temperature of 1050 mm and 5 °C, respectively. Moreover, the winter and summer average temperature are −9 °C and 18 °C, respectively, showing that all four seasons are defined clearly in both study areas (Figure 3-2).

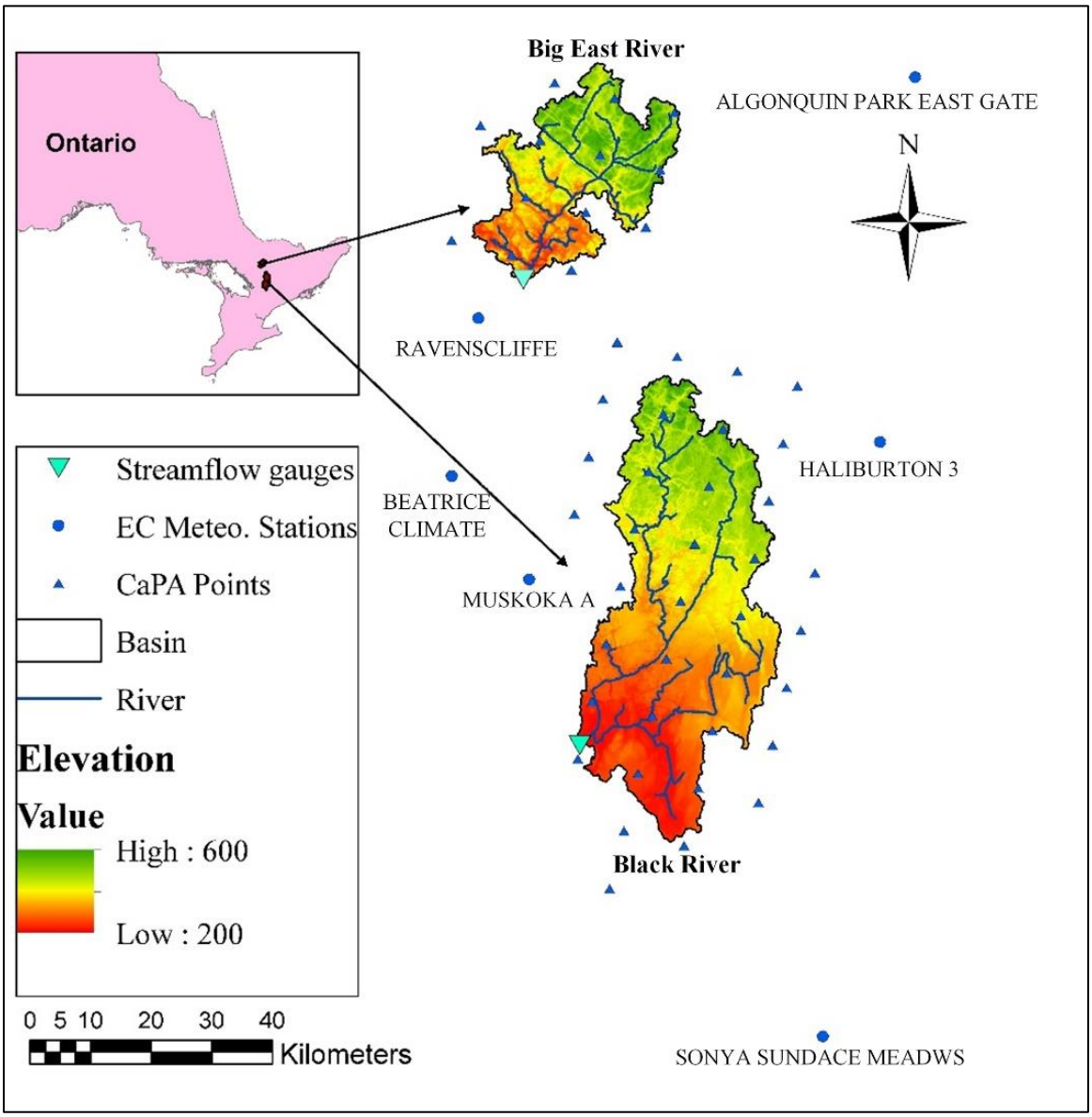


Figure 3-1 Location map of the Big East River and Black River watersheds

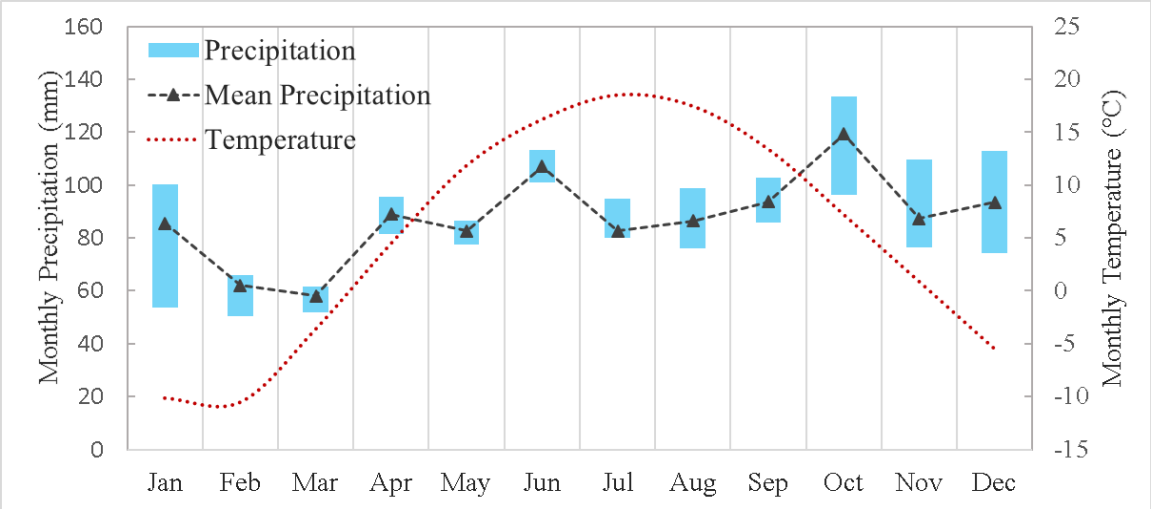


Figure 3-2 The box-plot and average of monthly precipitation and the mean monthly temperature for the observation period (2006–2015) based on data from six available meteorological stations

Besides the ground-based precipitation data, the archive of the daily aggregated form of the Canadian Precipitation Analysis (CaPA) was used as an alternative precipitation forcing input for hydrologic modeling of both watersheds. The CaPA is a gridded precipitation product with a spatial resolution of 15 km produced by the Meteorological Service of Canada based on the combination of various data sources, such as radar data, climate model data, and observations (Lespinas et al., 2015). It was shown that the archived CaPA is a potential reliable source of precipitation for data-scarce regions (Boluwade et al., 2018). In order to initially assess the precipitation variability of each basin using different datasets, primary analysis was performed. Two mean areal precipitation time-series for each watershed were derived from interpolated EC ground-based data using an inverse distance weighting method (American Society of Civil Engineers, 1996) and the CaPA data by applying a Thiessen polygon approach (Thiessen, 1911). As can be seen

from Figure 3-3, although CaPA provided more intense rainfalls specifically in the Black River watershed, it underestimated the amount of precipitation compared with the EC data in both watersheds. Moreover, the calculated daily correlation coefficients between EC- and CaPA-derived datasets (0.83 and 0.87 for the Big East River and Black River watersheds, respectively) show evidence of a linear relationship. However, by focusing on intense rainfall events (precipitation > 10 mm/day), the correlation coefficients were dramatically decreased to 0.42 and 0.48 for the Big East River and Black River watersheds, respectively. Therefore, there are remarkable differences between two datasets, especially at intense rainfall events, suggesting a significant amount of input uncertainty in poor-data watersheds. So, the authors used CaPA as a second forcing data for hydrologic models, which can help obtain a better quantification of the predictive uncertainty in the rainfall-runoff process using a Bayesian model averaging approach.

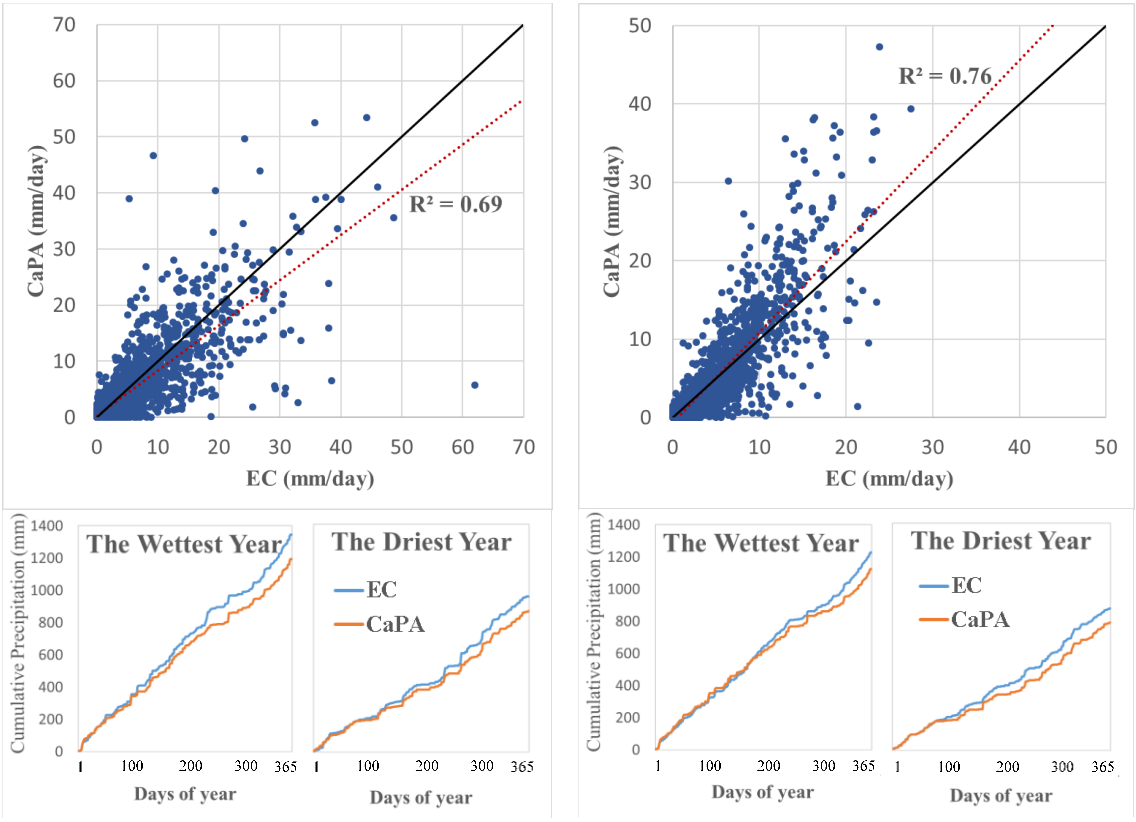


*Figure 3-3 The scatter plots of the mean areal interpolated Environment Canada (EC) and Canadian Precipitation Analysis (CaPA) data and their corresponding cumulative precipitation of the driest and wettest years during the period 2006–2015 for both the (**a**) Big East River and (**b**) Black River watersheds*

3.3.2 Standard Bayesian Model Averaging Technique

Bayesian model averaging is a statistical method for estimating probabilistic prediction based on various competing forecasts, possessing more reliability and accuracy than initial ensemble predictions. In this approach, the weighted averages of the individual forecasts' probability distribution functions (PDFs) are used for generating the posterior distribution of forecasting variables. It was claimed through different studies that the higher weights

are considered for better performing predictions in the training period (Duan et al., 2007; He et al., 2018; Liang et al., 2013; Vrugt et al., 2008; Yen et al., 2014).

Consider y as a quantity which is going to be forecasted (i.e., predictand) and, therefore, $Y = (y_1, y_2, \dots, y_T)$ denotes the training period of observation with data length T. Having K different models (i.e., $M = (M_1, M_2, \dots, M_K)$) results in $Y^f = (Y^{M_1}, Y^{M_2}, \dots, Y^{M_K})$, the ensemble of model predictions for the aforementioned training period, where $Y^{M_i} = (y_1^{M_i}, y_2^{M_i}, \dots, y_T^{M_i})$. Based on the law of total probability and the assumption about the independence of different model forecasts, the PDF of the predictand conditioned on the models over the given training period can be formulated as follows (Raftery et al., 1997):

$$P(y|Y^{M_1}, Y^{M_2}, \dots, Y^{M_K}, Y) = \sum_{i=1}^{k} P(y|\, Y^{M_i}, Y) \times P(\, Y^{M_i}|Y) \qquad (3\text{-}1)$$

where $P(y|Y^{M_i}, Y)$ is the posterior distribution of y given the prediction of model M_i and observed data Y, which simply can be considered as the forecast PDF of y based on model M_i. Moreover, $P(Y^{M_i} \,|\, Y)$ is the posterior probability or the likelihood of the model's M_i prediction being correct over the training period. Due to the assumption of models' independency, the posterior probabilities of models should sum to unity, $\sum_{i=1}^{K} P(\, Y^{M_i}|Y) = 1$, and, consequently, they can be considered as weights (i.e., $w_i = P(\, Y^{M_i} \,|\, Y)$ is the weight of model i). Furthermore, in the BMA approach, it is assumed that the model forecasts are unbiased, meaning that the expected value of the difference between observation and each model forecast should be equal to zero (i.e., $E(Y - Y^{M_i}) = 0$ for $i \in [1, K]$). So, before BMA implementation, a bias-correction method should be used in order

to create an unbiased ensemble of predictions. Although there are several bias-correction methods which all can be used for this aim, a linear-regression technique is utilized in the original BMA (Raftery et al., 2005). The bias-corrected results, $F^{M_i} = a_i \times Y^{M_i} + b_i$ (where a_i and b_i are the coefficients of the linear regression model), are replaced with the original model forecasts (Y^{M_i}). Therefore, the BMA predictive model (Equation 3-1) can be rewritten as follows:

$$P(y|Y^{M_1}, Y^{M_2}, \dots, Y^{M_K}, Y) = \sum_{i=1}^{k} w_i \times P(y|\, F^{M_i}, Y) \qquad (3\text{-}2)$$

On the other hand, in the original BMA method, it is assumed that the aforementioned posterior probability (i.e., $P(y|F^{M_i}, Y)$) follows the normal (Gaussian) distribution, $g(y|F^{M_i}, \sigma_i^2)$, with mean F^{M_i} and variance σ_i^2, reflecting the uncertainty within the individual model i. As explained in the introduction, some studies discussed that this assumption is a poor choice for a non-Gaussian forecast variable like streamflow. Therefore, they proposed implementing more representative distribution types (e.g., gamma distribution) or applying data transformation procedures (e.g., the Box–Cox transformation method (Box & Cox, 1964)) for transforming data from their original space to a Gaussian space. It is worth mentioning that in the case of applying a data transformation procedure, the reverting process has to be able to apply in order to revert back to the original variable space.

Finally, based on Equation 3-2 and considering the Gaussian distribution, the BMA predictive mean and its associated variance can be determined using the two following

equations (Raftery et al., 1997, 2005). The mean value is the weighted average of individual predictions, and the BMA variance consists of (1) between-model variance, reflecting the spread of the ensemble, and (2) within-model variance that represents the uncertainty regarding each model having the best forecast.

$$E(y|Y^{M_1}, Y^{M_2}, \dots, Y^{M_K}, Y) = \sum_{i=1}^{k} w_i \times F^{M_i} = \sum_{i=1}^{k} w_i \times (a_i \times Y^{M_i} + b_i) \qquad (3\text{-}3)$$

$$Var(y|Y^{M_1}, Y^{M_2}, \dots, Y^{M_K}, Y) = \sum_{i=1}^{k} w_i \left(F^{M_i} - \sum_{n=1}^{k} w_n \times F^{M_n} \right)^2 + \sum_{i=1}^{k} w_i \, \sigma_i^2$$

$$= \sum_{i=1}^{k} w_i \left((a_i \times Y^{M_i} + b_i) - \sum_{n=1}^{k} w_n \times (a_n \times Y^{M_n} + b_n) \right)^2 \qquad (3\text{-}4)$$

$$+ \sum_{i=1}^{k} w_i \, \sigma_i^2$$

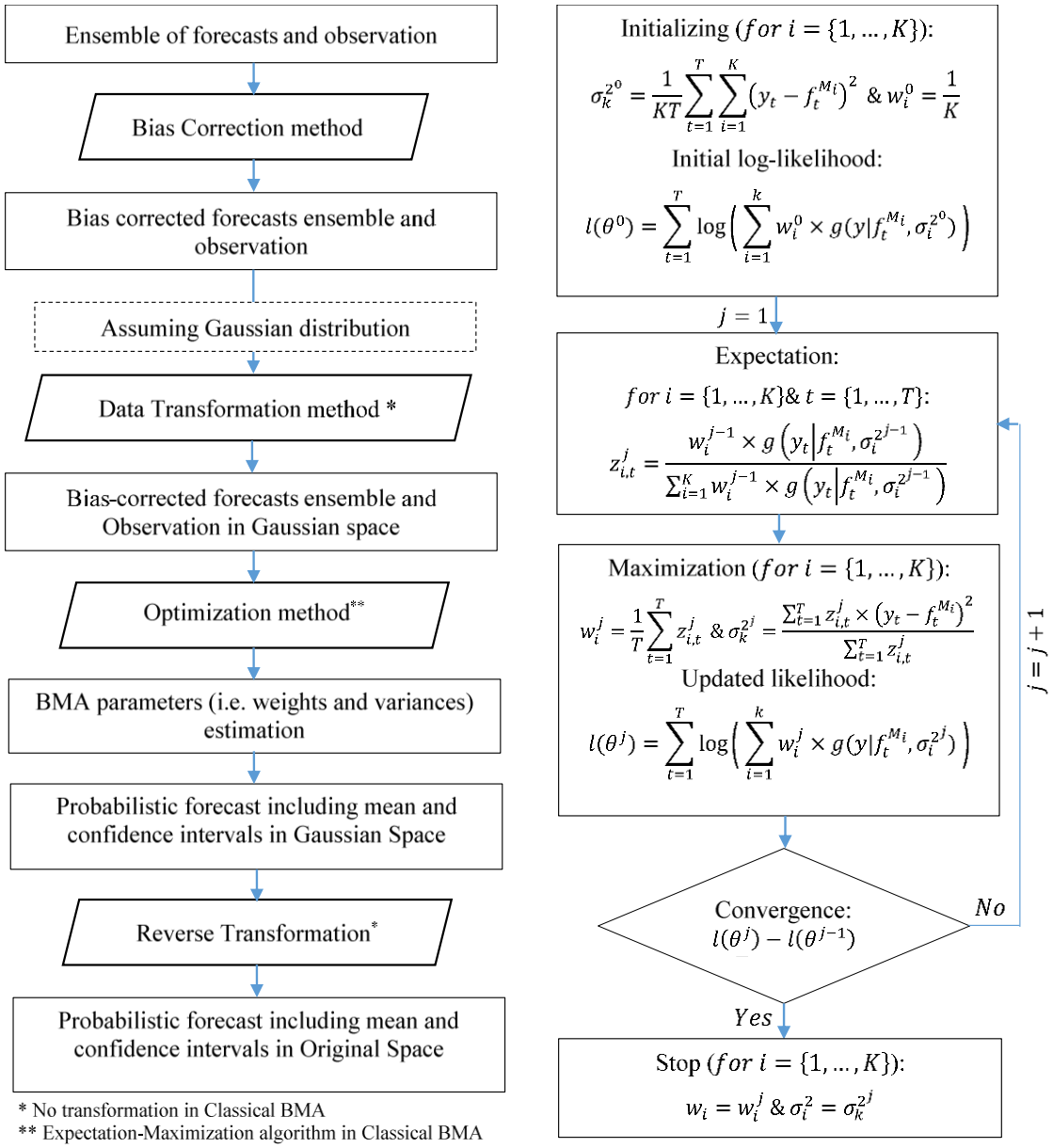
Successful implementation of the BMA method relies on the proper estimation of the parameters including weights (w_i) and variances (σ_i^2) of each individual prediction ($i = 1, \dots k$). Following Raftery et al. (2005), in the standard BMA, the EM algorithm is utilized in order to maximize the log-likelihood function of the parameter vector ($\theta = \{w_i, \sigma_i^2, i = 1,2,\dots,K\}$) being approximated as follows:

$$L(\theta) = Log\big(P(y|Y^{M_1}, Y^{M_2}, \dots, Y^{M_K}, Y)\big) = Log\left(\sum_{i=1}^{k} w_i \times g(y|F^{M_i}, \sigma_i^2)\right) \qquad (3\text{-}5)$$

Given that there is no analytical solution for maximizing the summation of the aforementioned function over the training period, an iterative procedure such as the EM

algorithm was used. In this procedure, the optimization problem was set by introducing a latent variable (Z_k). Apart initialization, this algorithm included an (1) expectation step, where the latent variable was calculated based on the current values of parameters, and a (2) maximization step, where the parameters were estimated according to the determined value of the latent variable (Figure 3-4b). It is worthy of note that, although the EM algorithm is computationally efficient, it is argued that using other optimization methods can lead to more robust estimation of the parameters.

According to the above equations, the flowchart of the classical BMA implementation is depicted in Figure 3-4a. As previously stated, some studies have been done in order to improve the reliability of the standard BMA approach by modifying some parts of the BMA structure. However, no comprehensive evaluation has been completed in order to clarify the effects of these modifications.



(a) **(b)**

*Figure 3-4 The flowcharts for (**a**) standard Bayesian model averaging (BMA) and (**b**) the step-by-step procedure of the expectation-maximization (EM) algorithm*

3.3.3 BMA Scenario-Based Analysis

In order to achieve the main goal of this research, we designed a BMA scenario-based analysis (Table 3-1) to see how the predictive streamflow simulation of the BMA approach was affected by modifying or changing some steps of the original BMA procedure. Implementation of the proposed evaluation allowed to assess how the accuracy and reliability of the BMA probabilistic results are sensitive to considering (1) different streamflow ensemble scenarios; (2) various data transformation methods; (3) more representative distribution types; (4) different standard deviation definitions; and (5) different optimization methods for parameter estimation. These scenarios are chosen in a way that cover most of the aforementioned modifications proposed by previous studies (explained in Section 3.2). Therefore, the effects of each modification or the combinations of modifications on BMA results can be assessed completely through the proposed analysis. The following paragraphs present a brief description of all aforementioned modification sections.

Table 3-1 The BMA scenario-based analysis

Streamflow Ensemble	Data Transformation Method	Distribution Type	Standard Deviation Type	Optimization Method
Multi-Model (M-M[1])	No Transformation (T0)	Normal (C1)	Common Constant (V1)	Expectation-Maximization Algorithm (EM)
Multi-Model Multi-Input (M-MI)	Box–Cox Type 1 (T1)	Gamma (C2)	Individual Constant (V2)	
Multi-Model Multi-Parameter (M-MP)	Box–Cox Type 2 (T2)	Log-Normal (C3)	Common Non-Constant (V3)	Dynamically Dimensioned Search (DDS)
Multi-Model Multi-Input Multi-Parameter (M-MIP)	Logarithmic Transform (T3)	Weibull (C4)	Individual Non-Constant (V4)	
	Empirical Normal Quantile Transform (T4)		Common Non-Constant + Constant Value (V5)	
			Individual Non-Constant + Constant Value (V6)	

1 The ID of each scenario is presented in the parentheses

3.3.3.1 Streamflow Ensemble

As mentioned before, the ensemble can stem from different sources. Apart from considering different hydrologic models, various forcing precipitation inputs, as well as different reliable parameter sets of each rainfall-runoff model, can be considered for generating an ensemble of streamflow simulations. In this study, four different scenarios were determined to see how the BMA performance would change by considering a different number of ensemble members coming from various sources. In the first scenario, which was named "Multi-Model", the ensemble was only based on different hydrologic models. In the two other scenarios (i.e., Multi-Model Multi-Input and Multi-Model Multi-Parameter), besides multiple hydrologic models, different precipitation datasets and

various parameter sets were respectively utilized. Moreover, the last scenario was defined using all aforementioned sources (i.e., Multi-Model Multi-Input Multi-Parameter).

3.3.3.2 Data Transformation Methods

Four different data transformation procedures were assessed in the case of assuming normal function for the posterior distributions. The Box–Cox transformation method is a family of power transformations, and one of the common approaches is formulated as follows (Box & Cox, 1964):

$$Z' = \begin{cases} \dfrac{Z-1}{\lambda} & \lambda \neq 0 \\ \log(Z) & \lambda = 0 \end{cases} \qquad (3\text{-}6)$$

Z and Z' are the original and transformed data, respectively. λ is the Box–Cox coefficient and its common optimum value will be estimated using (1) observation data (i.e., Type 1) or (2) observation and simulations data (i.e., Type 2) by maximizing the log-likelihood function. Moreover, in the logarithmic transformation method, the daily streamflow data are transformed using natural logarithm in order to make them approximately follow the normal distribution. Another data transformation method evaluated in this study was the Empirical Normal Quantile Transformation (ENQT) procedure (Krzysztofowicz, 1997). In this approach, the transformed data were calculated using the following equation, where Q^{-1} is the inverse of the standard normal distribution and the empirical cumulative distribution of each value is denoted by $eCDF(Z)$.

$$Z' = Q^{-1}\big(eCDF(Z)\big) \qquad (3\text{-}7)$$

It is of note that, instead of the empirical distribution, the generalized Pareto distribution is fitted to extrapolate the upper tail of the sample in the case of having a value which falls outside the range of the calibration data.

3.3.3.3 Distribution Types

Apart from using normal distribution, which is the main assumption of the original BMA method, the log-normal, gamma, and Weibull distributions are implemented as the conditional probability distribution function $P(y|F^{M_i}, Y)$ in Equation 3-2. These distributions are more representative for highly skewed data such as daily stream flows and may lead to better results.

3.3.3.4 Standard Deviation Types

In this study, following Vrugt (2016), six various standard deviation parameterizations of the forecast distributions were assessed. The terms "common" and "individual" are used when all members of the ensembles have the same and distinct standard deviations, respectively. The other two terms illustrate if the standard deviations are dependent on the magnitude of the streamflow data ("non-constant") or not ("constant"). Moreover, the last two types are defined by adding constant value in order to make the standard deviation be more than zero in all cases. The equations of all aforementioned standard deviation types and their corresponding number of parameters are presented in Table 3-2. In these equations, $\sigma_{i,j}$ and $Q_{i,j}$, respectively, denote the standard deviation and the daily discharge of the ith simulated streamflow at time-step j. Also, K is the total number of members in the ensemble.

Table 3-2 The definitions and formulations of different standard deviation parameterizations

Standard Deviation Type	Formulation	BMA parameters
Common Constant (V1[1])	$\sigma_i = \sigma$	$\theta = \{w_i, \sigma\} \;\; i \in [1, K]$
Individual Constant (V2)	$\sigma_i = \{\sigma_1, \sigma_2, \dots, \sigma_K\}$	$\theta = \{w_i, \sigma_i\} \;\; i \in [1, K]$
Common Non-Constant (V3)	$\sigma_{i,j} = c \times Q_{i,j}$	$\theta = \{w_i, c\} \;\; i \in [1, K]$
Individual Non-Constant (V4)	$\sigma_{i,j} = c_i \times Q_{i,j}$	$\theta = \{w_i, c_i\} \;\; i \in [1, K]$
Common Non-Constant Type 2 (V5)	$\sigma_{i,j} = c \times Q_{i,j} + d$	$\theta = \{w_i, c, d\} \;\; i \in [1, K]$
Individual Non-Constant Type 2 (V6)	$\sigma_{i,j} = c_i \times Q_{i,j} + d_i$	$\theta = \{w_i, c_i, d_i\} \;\; i \in [1, K]$

[1] The ID of each type is presented in the parentheses.

3.3.3.5 Optimization Methods

Given the criticism of the EM algorithm regarding its ability to achieve the global optimum estimation and its lack of flexibility in applying to the various aforementioned modifications, the dynamically dimensioned search (DDS) method (Tolson & Shoemaker, 2007) was used as the alternative optimization technique for estimating the BMA parameters. Dynamically dimensioned search is a single global optimization method which finds the optimal solution by dynamically rescaling the search space dimension. Similar to the EM algorithm, the log-likelihood of the BMA parameter vector is considered as the objective function in the DDS optimization approach. Correspondingly, the DDS parameter estimations can be utilized as benchmarks for evaluating the application of the EM algorithm.

3.3.4 Hydrological Models

Using different hydrologic models for generating an ensemble of competing simulated stream flows is the main basis of the BMA approach (Vrugt & Robinson, 2007). As listed

in Table 3-3, the seven different rainfall-runoff models implemented in this study are SAC-SMA, MAC-HBV, SMARG, GR4J, and three HEC-HMS (Scharffenberg, 2016) based models. There are different methods available for each part of the hydrologic cycle in the HEC-HMS platform. In this study, we used the rational combination of loss (i.e., deficit and constant, and soil moisture accounting) and baseflow (i.e., recession and linear reservoir) methods for generating the HEC-HMS-based models with different structures. In the HEC-HMS type 1 and 2, the recession baseflow method is implemented with the deficit and constant and soil moisture accounting loss approaches, respectively, while HEC-HMS type 3 is developed using the combination of the soil moisture accounting and linear reservoir methods.

All of the aforementioned models are lumped conceptual ones, which have been shown to provide comparable or even better performance in comparison to the more complex models (e.g., distributed models) in data-poor watersheds (Anshuman et al., 2019; Refsgaard & Knudsen, 1996; Tegegne et al., 2017). Moreover, by adding the simplified Thornwaite formula (Samuel et al., 2011; Thornthwaite, 1948) to the first four models and feeding HEC-HMS models the average monthly potential evapotranspiration calculated using Hargreaves equation (Hargreaves & Samani, 1985), the only inputs to all models are the mean areal daily precipitation and temperature. Also, streamflow estimation at the outlet of the watershed is the only output of these models. It is worth mentioning that due to the importance of the snow accumulation and melt process in cold regions, three different snowmelt modules are implemented with different hydrologic models. The available temperature-index method in the HEC-HMS software (Scharffenberg, 2016) was used for

the three aforementioned HEC-HMS-based models. The simple degree-day snowmelt module (DDM) (Samuel et al., 2011) was added to the SMARG and GR4J models, while the SACSMA and MACHBV models were combined with the more complex SNOW17 snowmelt estimation method (Anderson, 1973, 2006) for snow–rainfall discrimination and quantifying snowpack changes over the simulation period.

On the one hand, in the DDM approach, the snowmelt is calculated using a linear relationship between snowmelt and air temperature, where a constant melt rate factor is considered. However, the antecedent temperature index is used for melt-rate determination in the HEC-HMS snowmelt approach (Gyawali & Watkins, 2013). On the other hand, the SNOW17 is a process-based temperature-index method that considers different physical processes in the snowmelt procedure such as energy exchange between air and snow, heat storage and deficit of the snowpack, liquid water storage, etc. Also, upper and lower preset temperature thresholds are used for distinguishing between rainfall and snowfall in both the DDM and SNOW17 models (Agnihotri, 2018). For a more detailed description of all snow routines, the readers are referred to the aforementioned citations.

Table 3-3 Hydrologic models used in this study

Model ID	Full Name	Reference	Number of Parameters
SAC-SMA	Sacramento Soil Moisture Accounting	Burnash et al. (1973)	19
MAC-HBV	McMaster University Hydrologiska Byrans Vattenbalansavdelning	Samuel et al. (2012)	15
SMARG	Modified Soil Moisture Accounting and Routing	Tan and O'Connor. (1996)	14
GR4J	Génie Rural à 4 Paramètres Journaliers	Edijatno et al. (1999)	9
HEC-HMS1	Hydrologic Engineering Center's Hydrologic Modeling System-Type 1	USACE-HEC (Scharffenberg, 2016)	17
HEC-HMS2	Hydrologic Engineering Center's Hydrologic Modeling System-Type 2	USACE-HEC (Scharffenberg, 2016)	25
HEC-HMS3	Hydrologic Engineering Center's Hydrologic Modeling System-Type 3	USACE-HEC (Scharffenberg, 2016)	27

Furthermore, five different objective functions, including Nash–Sutcliffe efficiency (NSE) [68], Kling–Gupta efficiency (KGE) (Gupta et al., 2009), Nash volume error (NVE) (Samuel et al., 2011), peak-weighted root mean square error (PWRMSE) (Cunderlik & Simonovic, 2004), and modified Nash volume error (MNVE) were used through the dynamically dimensioned search (DDS) algorithm for finding the optimized parameter sets of each individual model. The latter objective function was defined in order to greatly focus on high flows by using the NSE based on square of discharge ($NSES$):

$$MNVE = NSES - 0.1VE \tag{3-8}$$

where volume error (VE) is:

$$VE = \frac{\left|\sum_{i=1}^{N}\left(Q_{s_i} - Q_{O_i}\right)\right|}{\sum_{i=1}^{N} Q_{O_i}} \tag{3-9}$$

and NSE based on square of discharge ($NSES$) is calculated as follows:

$$NSES = 1 - \frac{\sum_{i=1}^{N}\left(Q_{s_i}^2 - Q_{O_i}^2\right)^2}{\sum_{i=1}^{N}\left(Q_{O_i}^2 - \overline{Q_O^2}\right)^2} \quad (3\text{-}10)$$

In the above equations, Q_{O_i} and Q_{S_i} are the observed and simulated streamflow, respectively, while N is the data length. The years 2006 to 2011 were considered the calibration period and the validation was carried out for the 2012–2015 (4 years) period. It is of note that the best performing parameter set of each individual model, determined based on validation results, is utilized for generating multi-model and multi-model multi-input ensemble scenarios. For a detailed description of the aforementioned hydrologic models and objective functions, the readers are referred to the cited references.

3.3.5 Performance Evaluation Metrics

Five model evaluation statistics are used for comparing the accuracy, reliability, and sharpness of the results of different BMA variants. The accuracy is defined as the error between deterministic simulations and their corresponding observations. In this study, besides the well-known Nash–Sutcliffe efficiency criteria, NSE being calculated according to squared ($NSES$; Equation 3-10) and logarithmic ($NSEL$; Equation 3-11) transformed streamflow data, were the two other deterministic performance criteria being, respectively, focused on the accuracy of the high- and low-flow simulations.

$$NSEL = 1 - \frac{\sum_{i=1}^{N}\left(Ln\left(Q_{s_i}\right) - Ln\left(Q_{O_i}\right)\right)^2}{\sum_{i=1}^{N}\left(Ln\left(Q_{O_i}\right) - \overline{Ln(Q_o)}\right)^2} \quad (3\text{-}11)$$

Q_{O_i} is the observed variable and Q_{S_i} represents the simulated variable which is considered to be the expected value of the BMA predictive simulation. Also, N is the length of the dataset. All NSE-based criteria vary between $-\infty$ and 1 with the best value of 1.

Furthermore, two other probabilistic performance measurements proposed by Xiong et al. (2009) were adopted for quantitative evaluation of the BMA probabilistic results. The containing ratio (CR) is defined as the percentage of the observed data which falls within the 95% confidence interval, and the average bandwidth (B) is the average width of the corresponding bound. The former measures the reliability while the latter is used for quantifying the sharpness of the results. Given two forecasts with the same CR (i.e., same reliability), the one with a smaller B shows a greater precision.

$$CR = \frac{NQ_{in}}{N} \times 100\% \qquad \textit{(3-12)}$$

$$B = \frac{1}{N}\sum_{t=1}^{N}\left(q_u(t) - q_l(t)\right) \qquad \textit{(3-13)}$$

In the above equations, the number of observations being contained in the 95% confidence interval is denoted by NQ_{in} $q_u(t)$ and $q_u(t)$, respectively, show the upper and lower boundaries of the 95% confidence interval at time-step t. In addition, for evaluating the probabilistic performance of different BMA variants regarding high flows, we calculated the two aforementioned probabilistic indices using the streamflow values of more than 90 percentiles (denoted by $CR90$ and $B90$ for the containing ratio and the average bandwidth, respectively).

3.4 Results and Discussion

3.4.1 Choosing the Best Ensemble Scenario

One of the vague points of the BMA approach in the literature is the optimal number of members of the ensemble and how they should be generated. The prime step before employing any BMA variants is constructing the most reliable ensemble, which provides the best results. Therefore, as the first section of the proposed analysis, the four aforementioned scenarios of different streamflow simulation ensembles were used in the original BMA for both the Big East River and Black River watersheds, and a comparison was made among their results (Table 3-4). Given the two different input scenarios and five various parameter sets for each hydrologic model, there were 7, 14, 35, and 70 simulated stream flows for the Multi-Model (M-M), Multi-Model Multi-Input (M-MI), Multi-Model Multi-Parameter (M-MP), and Multi-Model Multi-Input Multi-Parameter (M-MIP) ensemble scenarios, respectively.

Table 3-4 Validation statistics of the BMA model using four ensemble scenarios in both watersheds

Criteria	Big East River Watershed				Black River Watershed			
	M-MIP	**M-MP**	**M-MI**	**M-M**	**M-MIP**	**M-MP**	**M-MI**	**M-M**
NSE[1]	0.76	0.74	0.79	0.77	0.82	0.81	0.84	0.81
$NSES$[1]	0.45	0.42	0.54	0.49	0.57	0.55	0.62	0.56
$NSEL$[1]	0.84	0.84	0.82	0.83	0.79	0.80	0.78	0.77
CR[1]	0.95	0.94	0.96	0.96	0.92	0.90	0.91	0.88
B[1]	17	18	19	23	27	28	24	27
$CR90$[1]	0.72	0.64	0.73	0.68	0.62	0.46	0.62	0.49
$B90$[1]	39	32	38	34	55	48	41	36

[1] NSE: Nash Sutcliffe efficiency; $NSES$: NSE based on squared transformed streamflow; $NSEL$: NSE based on logarithmic transformed streamflow; CR: containing ratio; B: average bandwidth; $CR90$: containing ratio based on stream flows more than 90 percentile; $B90$: average bandwidth based on stream flows more than 90 percentile

If the BMA performance based on the Multi-Model (M-M) ensemble scenario is considered as the benchmark, there was no significant improvement when the performance statistics focusing on the whole and low discharges were considered. However, by focusing on the high flow-based criteria, the results show that considering the forcing precipitation as another source of uncertainty besides hydrologic models enhanced both the deterministic and probabilistic BMA results. This improvement was more significant in the Black River watershed, where the accuracy and reliability of the BMA using the M-MI scenario increased by about 10 and 25 percent based on the $NSES$ and $CR90$ criteria, respectively. It is worth mentioning that, all seven additional members of the streamflow simulations (generated by considering CaPA as forcing inputs of each individual model) being used in M-MI compared to M-M, possessed lower individual deterministic predictive skills than existing models in both ensemble scenarios.

Moreover, surprisingly, although the Multi-Model Multi-Parameter ensemble scenario included all members being utilized in the benchmark scenario, the overall performances of the BMA method implementing them slightly deteriorated in both watersheds. This may be due to the main initial assumption of the BMA methodology, where the law of total probability needs not only collectively exhaustive but also independent members of the ensemble. Furthermore, using 70 members in a streamflow ensemble (constructed by considering all aforementioned sources) enhanced the probabilistic performance of the BMA, specifically in high flows, while its performance was not as reliable and sharp as in the case where the M-MI scenario was applied.

Altogether, it can be concluded that the M-MI ensemble scenario was the most appropriate one, providing better probabilistic and deterministic results. Accordingly, for the rest of the application of the proposed analysis, the Multi-Model Multi-Input ensemble scenario, including 14 members of streamflow simulations, was implemented for both watersheds. As a result, 48 probabilistic streamflow simulations were generated considering the combination of the different modifications, including distribution, standard deviation, and data transformation methods (Table 3-1). The parameters for all 48 BMA variants were calibrated using the DDS optimization method for the period from 2006 to 2011, considering one year as a warm-up period, and the years 2012 to 2015 were considered for validation.

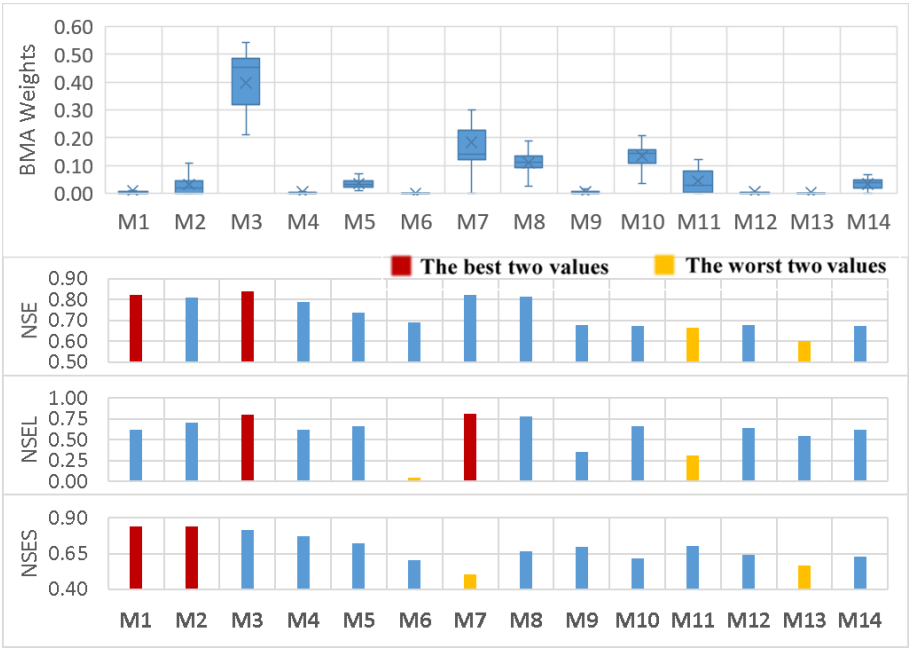
3.4.2 BMA Weights versus Models' Performance Statistics

In the first place, besides assessing the effects of various modifications, a comparison was made between the BMA weights of different members of the ensemble and the performance of the corresponding models during the calibration period for both the Big East River and Black River watersheds (Figure 3-5).

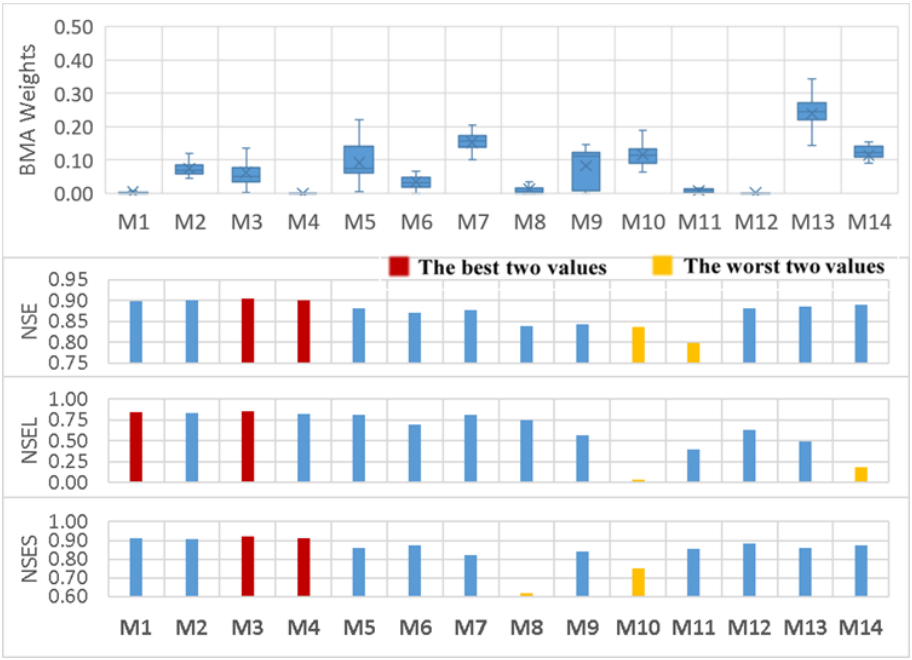
Interestingly, it can be seen that the distributions of the weights amongst different members do not properly agree with the previous belief, where the weights reflect the models' performance. For instance, in the Big East River watershed, although M1 was one of the most promising simulations comparing different performance statistics, its weights were not predominant compared to other BMA variants. In addition, in the Black River watershed, M10 had relatively high weights, while its performance was not good in comparison to the other models. Similarly, the first four members of the ensemble (i.e., M1

to M4) possessed the most reliable deterministic results, although they received relatively low weights.

Moreover, closer inspection of the graphs (in Figure 3-5) shows that low flows played an important role in the determination of the BMA weights, specifically in the Big East River watershed where the specified weights relatively fit with the $NSEL$ performance statistics. This may be justifiable by the fact that more than 90 percent of the daily streamflow observations were less than 25 m^3/s while this fraction was around 60 for the Black River watershed (Figure 3-6).



(**a**)



(**b**)

*Figure 3-5 The boxplots of the calibrated BMA weights stem from different BMA modifications in comparison with the different performance criteria of each individual daily streamflow simulation for (**a**) the Big East River and (**b**) Black River watersheds during the calibration period*

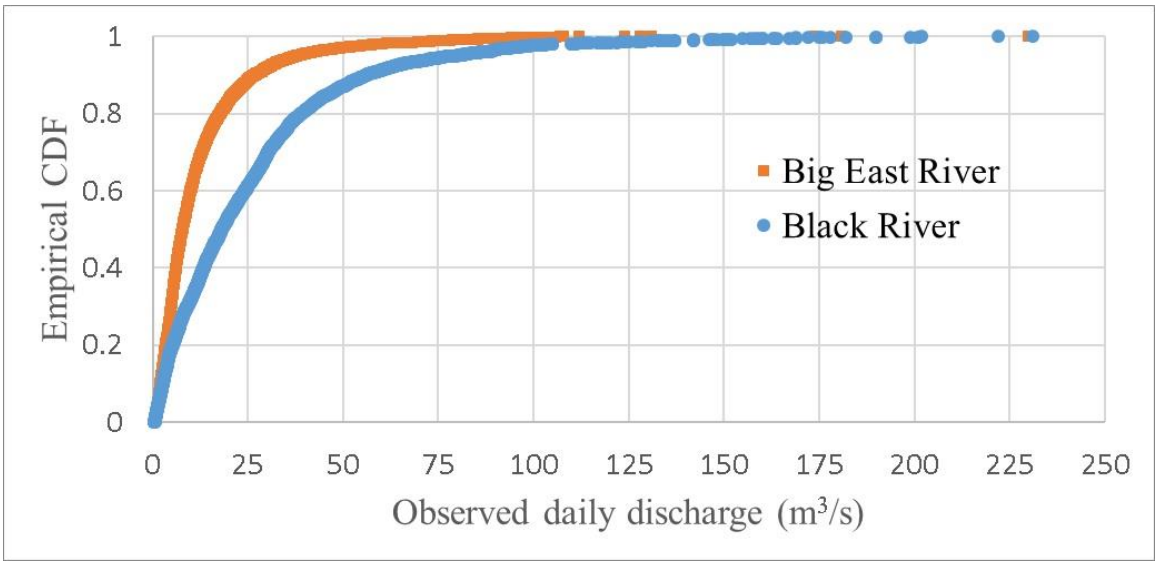


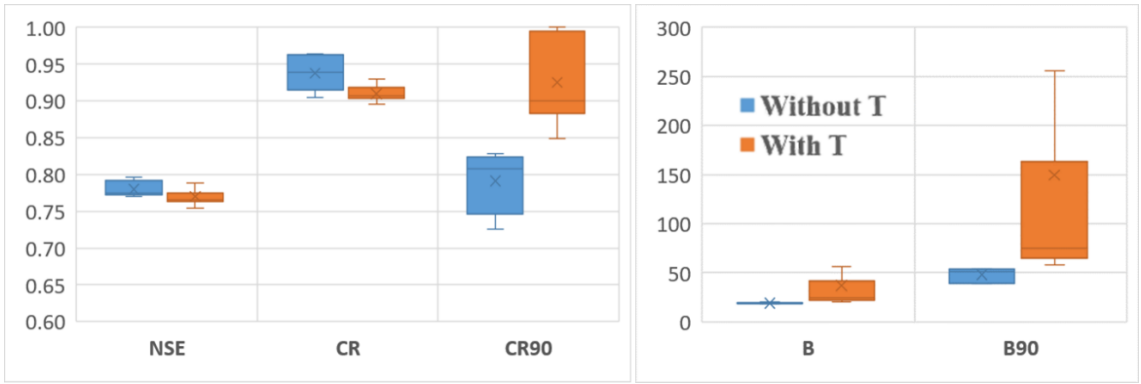
Figure 3-6 Empirical cumulative probability distribution of the daily streamflow observations at the outlet of the Big East River and Black River watersheds

3.4.3 The Effects of Different Modifications

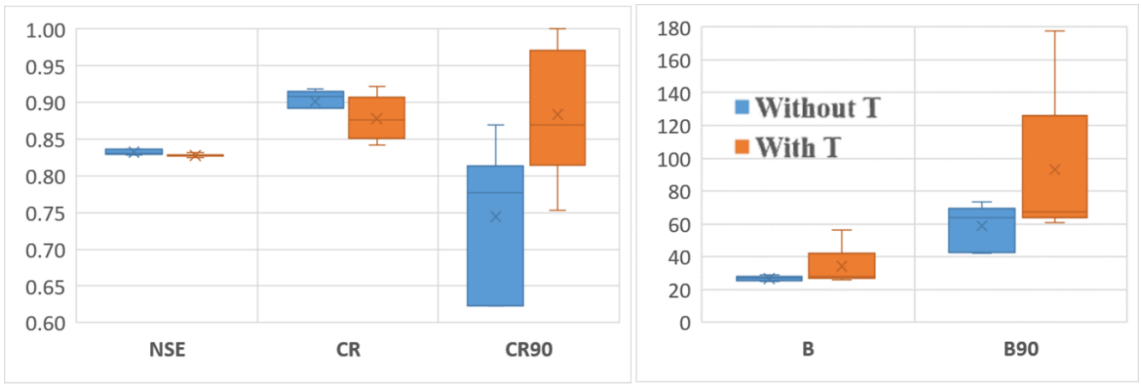
The evaluations of various BMA modifications (i.e., different distribution and standard deviation types, and data transformation methods) will be provided in this section. As discussed previously, one recommended solution in order to enhance the performance of the original BMA approach is using data transformation procedures for generating approximately normally distributed data. Figure 3-7 compares the accuracy and reliability of the BMA variants with and without application of data transformation procedures. It can be recognized that, in general, the BMA deterministic performance did not change significantly by applying data transformation methods. On the other hand, although the data transformation caused a remarkable enhancement of the BMA's reliability in high flows, the sharpness of the results was largely reduced.

Further analysis (Figure 3-8) shows that the influence of applying data transformation modification on the BMA performance is highly related to the types of standard deviation being implemented in the procedure. In the case of considering common and individual non-constant variance types (i.e., V3 and V4, respectively), implementation of a data transformation method leads to under confident and negatively biased probabilistic results. It is much more recognizable in high flows where the containing ratios of the 95% confidence interval are around one, while their corresponding bandwidths increase largely. However, for other types of standard deviations where a constant value can play an important role, the reliability of the high flows' simulation is partly improved without a drastic drop in their sharpness.

Moreover, Table 3-5 represents the performance criteria of different BMA variants, being developed using normal distribution and variance types V5 and V4, to compare different data transformation procedures. Based on the results, the only data transformation procedure providing acceptable probabilistic results with the use of heteroscedastic standard deviation without a constant value (i.e., V3 and V4) was the empirical normal quantile transform (i.e., T4) method. However, in general, by looking at the BMA variants based on variance type V5, as a representative of the other standard deviation forms, none of the methods appeared superior to the others, indicating that changing the data transformation approaches had little impact on BMA model performance.



(**a**)

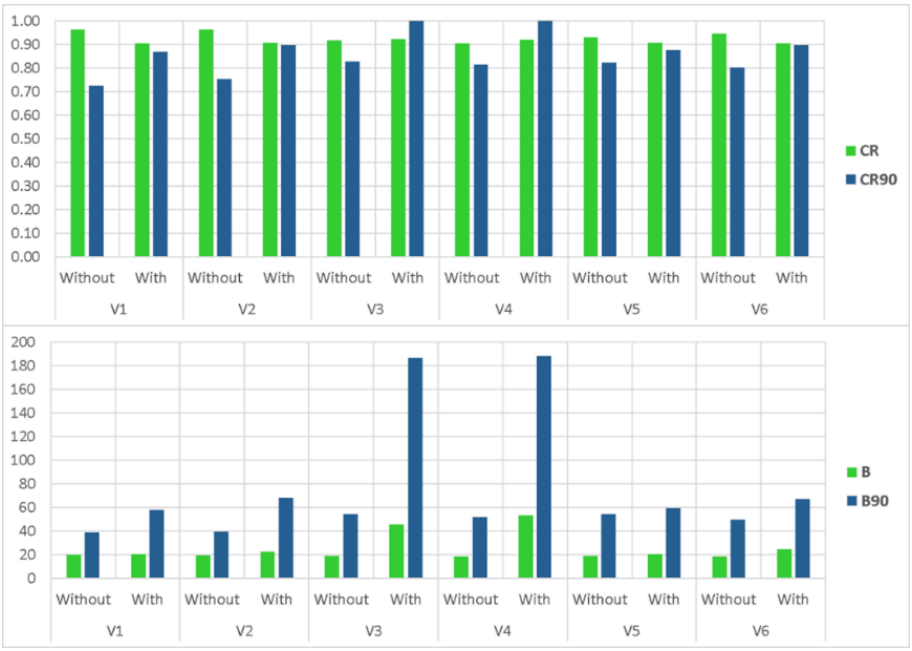


(**b**)

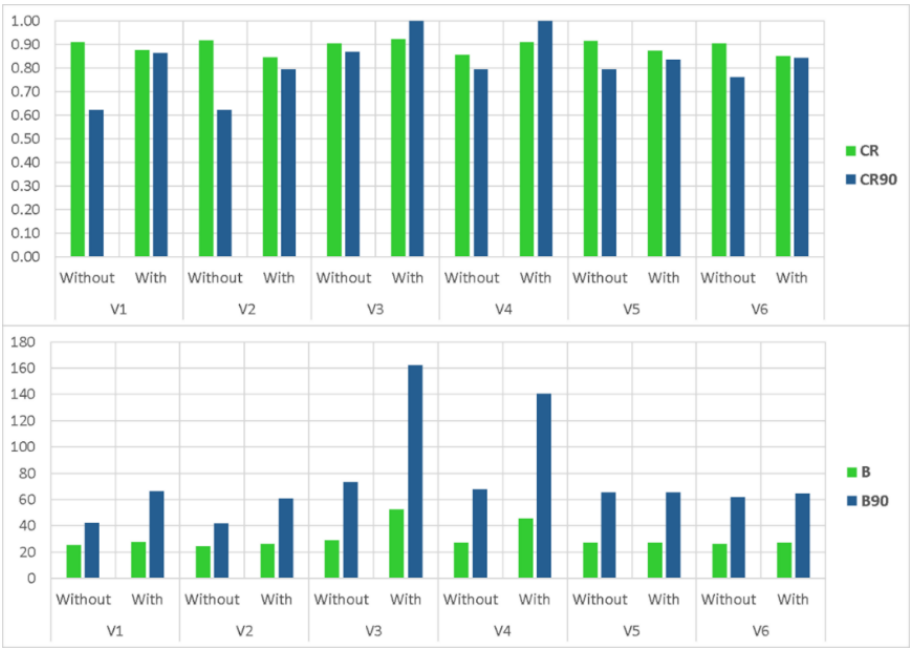
*Figure 3-7 The boxplots of the different evaluation metrics for the BMA streamflow simulations by implementation (With T) or non-implementation of data transformation (without T) methods being derived from considering normal distribution and different proposed standard deviation types for the (**a**) Big East River and (**b**) Black River watersheds during the validation period*

Table 3-5 Probabilistic evaluation criteria of different BMA variants based on different data transformation methods for both watersheds in the validation period

Basin	Criteria	BMA Variant							
		C1V5T1	**C1V5T2**	**C1V5T3**	**C1V5T4**	**C1V4T1**	**C1V4T2**	**C1V4T3**	**C1V4T4**
BE	CR	0.91	0.90	0.91	0.90	0.92	0.93	0.92	0.91
	B	25	22	21	24	127	73	53	30
	$CR90$	0.90	0.88	0.88	0.89	1.00	1.00	1.00	0.98
	$B90$	82	65	60	65	720	364	188	87
BL	CR	0.87	0.88	0.87	0.86	0.91	0.91	0.91	0.88
	B	27	27	29	27	46	46	52	30
	$CR90$	0.84	0.80	0.92	0.85	0.99	1.00	0.99	0.88
	$B90$	66	64	73	64	143	141	170	76



(**a**)

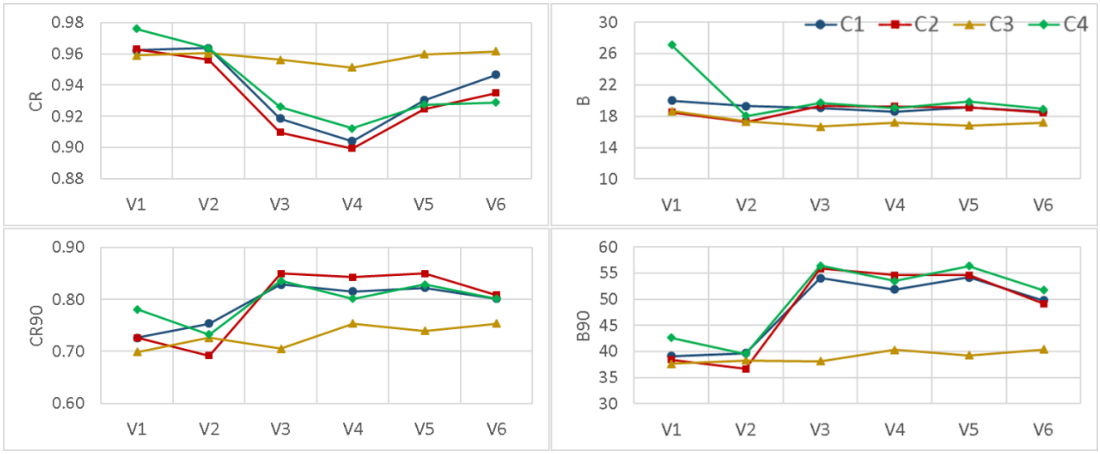


(**b**)

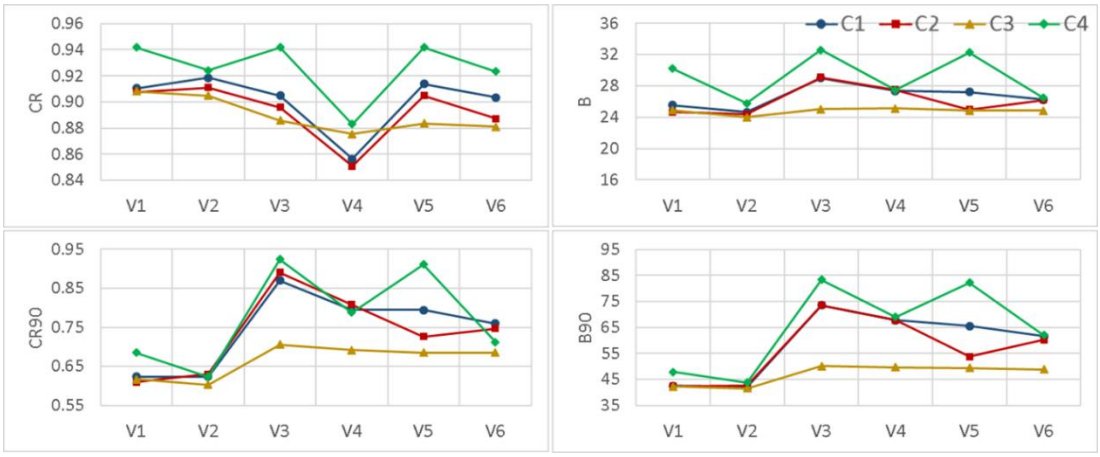
*Figure 3-8 The comparison of different performance statistics for various BMA modifications generated by considering different standard deviation types and non-implementation ("Without") and implementation ("With") of their corresponding best data transformation method for the validation period in the (**a**) Big East River and (**b**) Black River watersheds*

Besides using data transformation procedures, the two other BMA modifications evaluated in this study were considering other distribution types and implementing various standard deviation forms (Figure 3-9). The comparison between the applications of four different distribution functions proposed in the scenario-based analysis shows that, in general, the implementation of the log-normal distribution (i.e., C3) enhances the reliability and sharpness of the BMA results simultaneously. However, it underestimates when considering high flows, which is not appropriate in most operational hydrologic fields such as flood forecasting. As can be seen from the figure, in the case of using a common constant standard deviation type (i.e., V1), even though the coverage of the 95% confidence interval slightly increased by applying the Weibull distribution, the model lost its sharpness by leading to a higher bandwidth in both watersheds. Moreover, by assessing the effects of using different standard deviation types, it is apparent that considering "non-constant" types leads to more reliable results especially for high flows. However, using "individual" variance types does not affect the BMA performance in comparison to their corresponding "common" ones.

Taken together, these results suggest that changing the distribution type of the BMA posterior probability from normal to more representative ones does not enhance the BMA probabilistic performance, significantly. However, implementation of "non-constant" standard deviation types improved the BMA predictive results specifically regarding high flows.



(**a**)



(**b**)

*Figure 3-9 Comparison of the probabilistic performance of the BMA models being modified using different distribution and variance types for the validation period in the (**a**) Big East River and (**b**) Black River watersheds*

3.4.4 Expectation-Maximization Algorithm versus Dynamically Dimensioned Search Method

The EM algorithm was implemented in the classical BMA method, which is criticized for not being able to reach global optimum estimations. Here, as a part of the evaluation, six different BMA variants were calibrated using the EM algorithm, and a comparison was

made with the corresponding DDS-based calibrated models. The results, as shown in Figure 3-10, indicate that the differences among estimated BMA weights using EM and DDS methods were negligible, and both methods led to the approximately similar optimal solution.

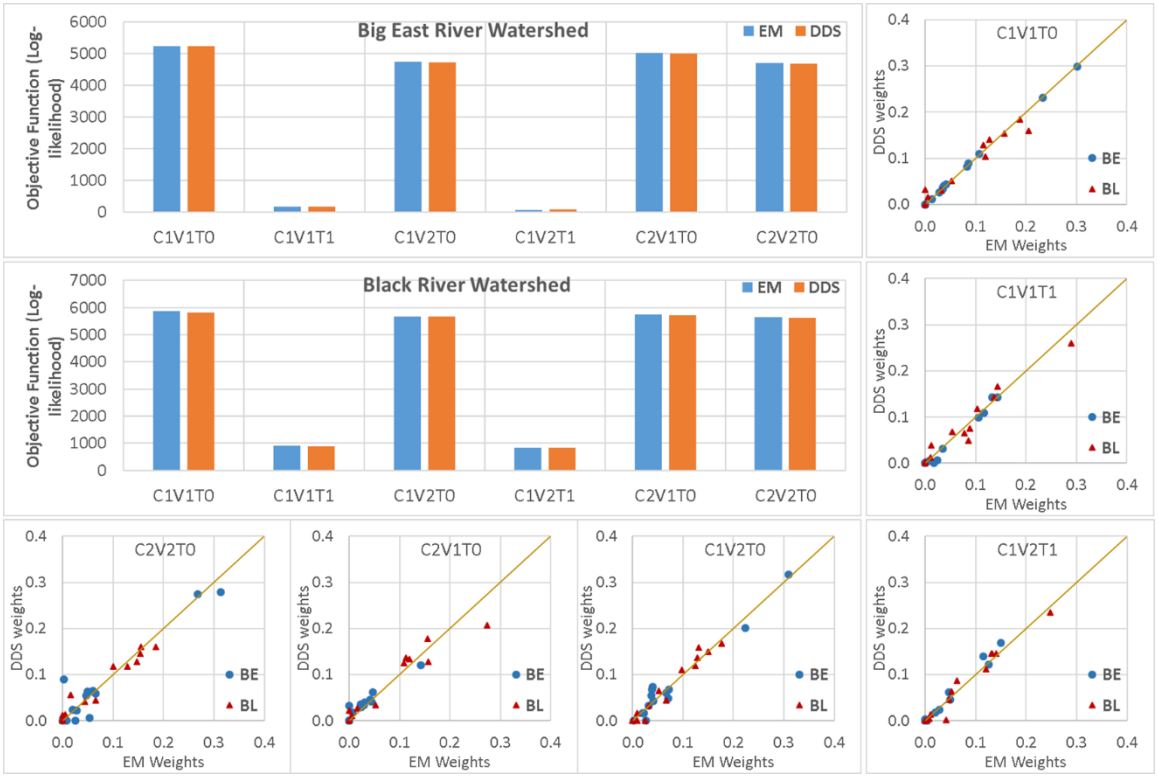


Figure 3-10 A comparison of the log-likelihood and weights of the calibrated BMA models using dynamically dimensioned search (DDS) and expectation-maximization (EM) algorithms as the optimization process

To specify the logic behind these results, the authors applied the regional sensitivity analysis (RSA) method (Hornberger & Spear, 1981) to original BMA with "common" (Figure 3-11) and "individual" (Figure 3-12) constant standard deviation types (i.e., C1V1T0 and C1V2T0 BMA variants, respectively). In this method, the Monte Carlo

simulation technique is used for generating various parameter sample sets, and then, the samples are divided into two behavioral and non-behavioral ones based on a predefined threshold. So, qualitative comparison of the empirical cumulative distribution functions (CDFs) of the behavioral and non-behavioral parameter sets illustrate the most sensitive parameter(s). The RSA results for both the Big East River and Black River watersheds reveal that the objective function is significantly sensitive to standard deviation values, while the models' weights can be considered non-sensitive parameters.

Therefore, the variation of the log-likelihood function is evaluated by changing the most sensitive parameters (standard deviations) between their lower and upper bounds while the other parameters are constant and equal to their nominal values (i.e., the calibrated values). The results, illustrated in Figure 3-13, show that in all evaluated cases, the negative log-likelihood, which is the objective function for both optimization processes, is a convex function so that a local optimization method such as the EM algorithm can lead to global optimal estimation of parameters. Consequently, although the EM algorithm is considered a local optimization method, it can estimate the original BMA parameters like other global optimization techniques. It is of note that the original EM method can only be applied for the constant variance types and it requires modifications if other distribution or standard deviation types need to be incorporated. However, DDS or any other global optimization techniques can be used by different BMA modifications without any difficulty.

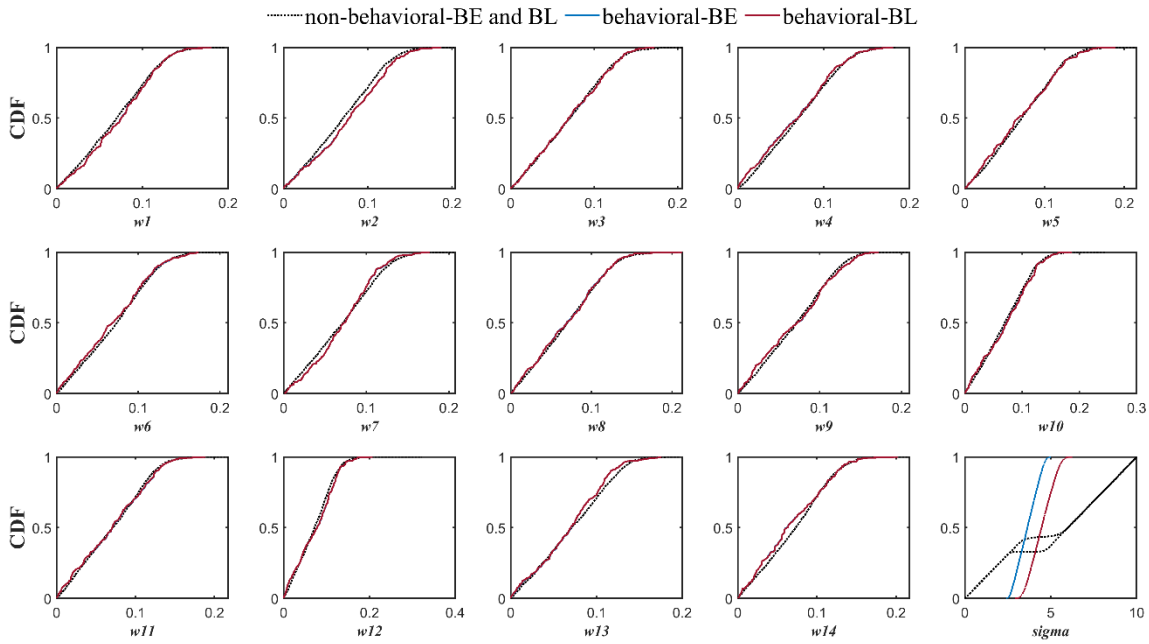


Figure 3-11 The regional sensitivity analysis (RSA) plots for the parameters of the C1V1T0 BMA variant for both the Big East River and Black River watersheds

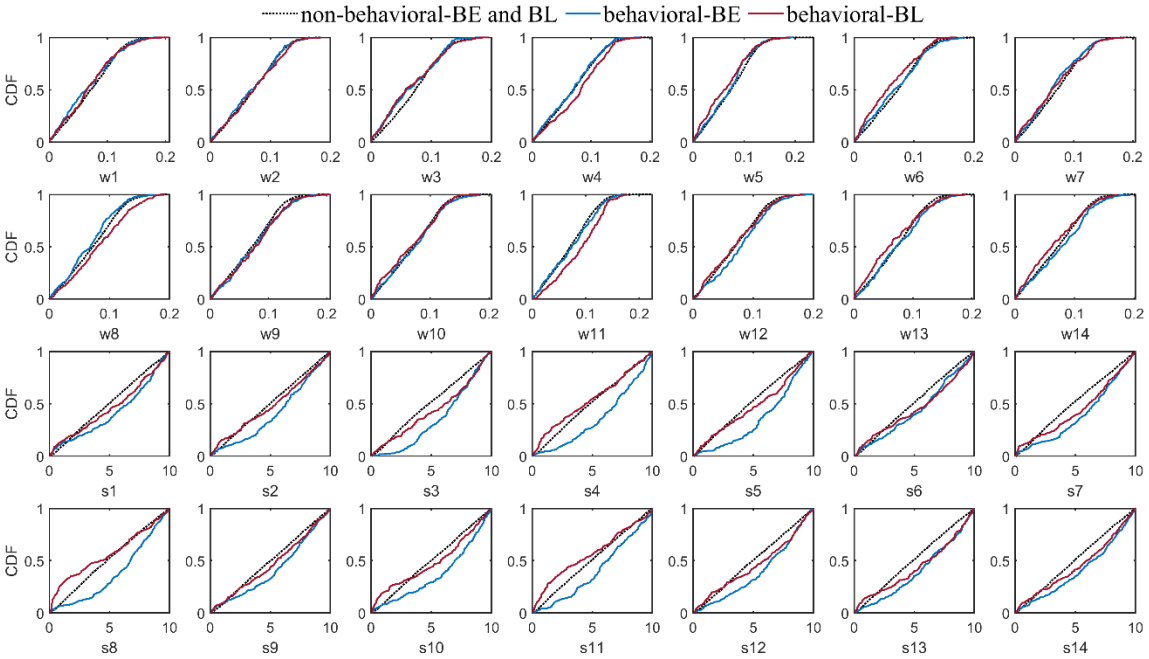
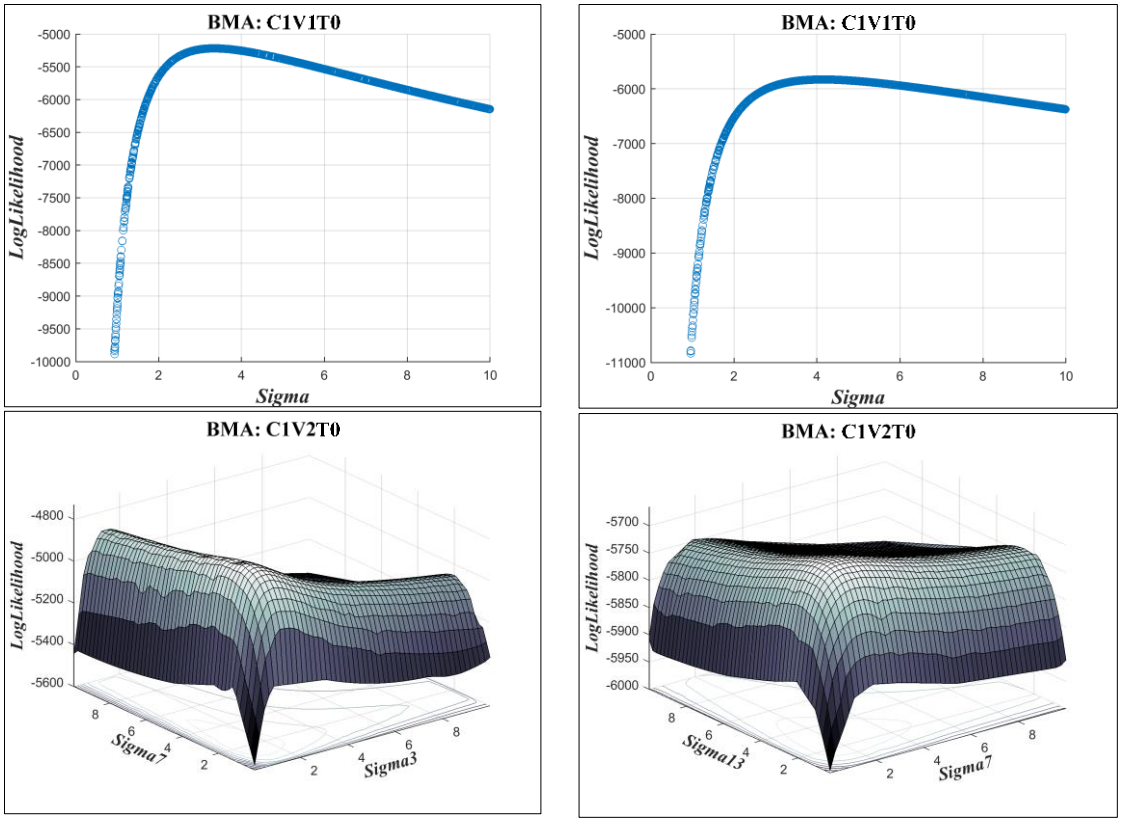


Figure 3-12 The RSA plots for the parameters of the C1V2T0 BMA variant for both the Big East River and Black River watersheds



(**a**) (**b**)

*Figure 3-13 The changes of the objective function regarding the most sensitive parameter(s) for the C1V1T0 and C1V2T0 BMA variants in both the (**a**) Big East River and (**b**) Black River watersheds*

Finally, in order to complete the evaluation and find the most promising types of BMA modifications, the best combinations were selected for each distribution type and their performances during the validation period were compared with each other (Table 3-6). Additionally, for qualitative inspection of the best models, Figure 3-14 illustrates the mean and the 95% predictive bounds of the BMA streamflow simulations for a representative portion of the validation period. What stands out in Table 3-6 is that the standard deviation types in all the best-selected BMA models were the non-constant ones, and most of them

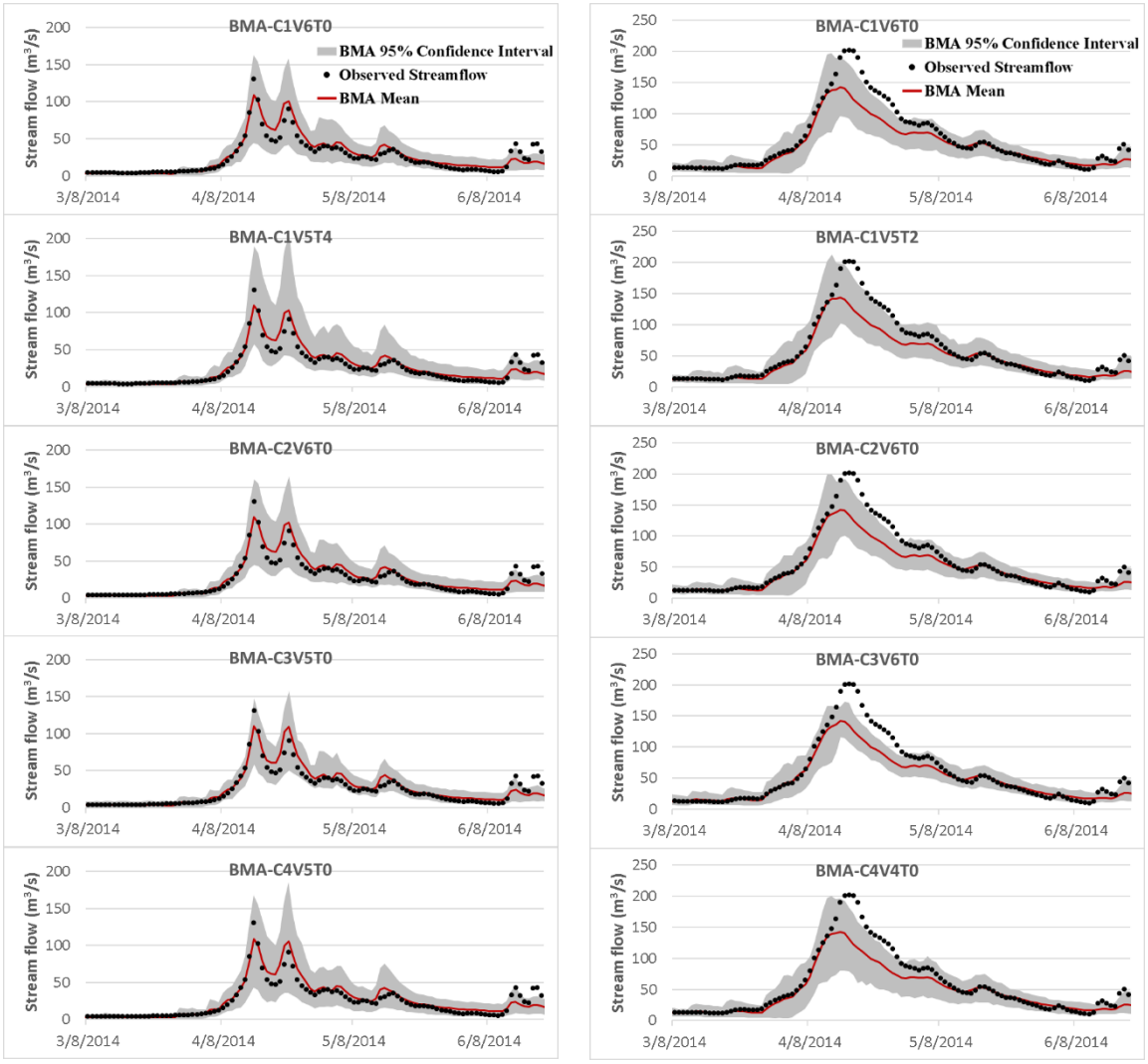
were the heteroscedastic variance with a constant value (i.e., V5 and V6). Moreover, as expected based on the previous comparison, although the best BMA modification with data transformation procedure provided higher reliability, the sharpness of the results partially deteriorated in high flows in both watersheds. Also, it can be seen that the best BMA model using the log-normal distribution type underestimated high flows significantly, while its other performance statistics showed almost the same predictive performance in comparison to the other best models. It is worthy of note that there was no significant difference among the accuracy of the various best-selected BMA variants.

Table 3-6 The comparison of the performances of the best-selected BMA types for both the Big East River and Black River watersheds during the validation period

	Criteria	NSE	NSES	NSEL	CR	B	CR90	B90
Big East River	C1V6T0	0.77	0.49	0.81	0.95	19	0.80	50
	C1V5T4	0.77	0.49	0.82	0.91	21	0.88	60
	C2V6T0	0.77	0.49	0.82	0.93	18	0.81	49
	C3V5T0	0.78	0.54	0.83	0.96	17	0.74	40
	C4V5T0	0.77	0.51	0.82	0.93	20	0.83	56
Black River	C1V6T0	0.83	0.60	0.80	0.90	26	0.76	61
	C1V5T2	0.83	0.59	0.80	0.87	27	0.84	66
	C2V6T0	0.83	0.61	0.80	0.89	26	0.75	60
	C3V6T0	0.83	0.61	0.79	0.89	25	0.71	50
	C4V4T0	0.83	0.59	0.80	0.88	27	0.79	69

Furthermore, as it was concluded beforehand, there was not a significant difference among the predictive performances of the different BMA variants utilizing various distribution types. However, the implementation of the gamma distribution type seemed to provide more balanced and consistent results in comparison to the other ones in this case. It is of

note that even by comparing the most promising models, which possessed approximately similar performances, the calibrated weights showed some changes confirming that there were no specific BMA weight combinations that led to the best results (Figure 3-15).



(**a**) (**b**)

*Figure 3-14 Time-series of the mean and 95% predictive bounds of daily streamflow derived from the best-selected BMA models for a representative portion of the validation period for both the (**a**) Big East River and (**b**) Black River watersheds*

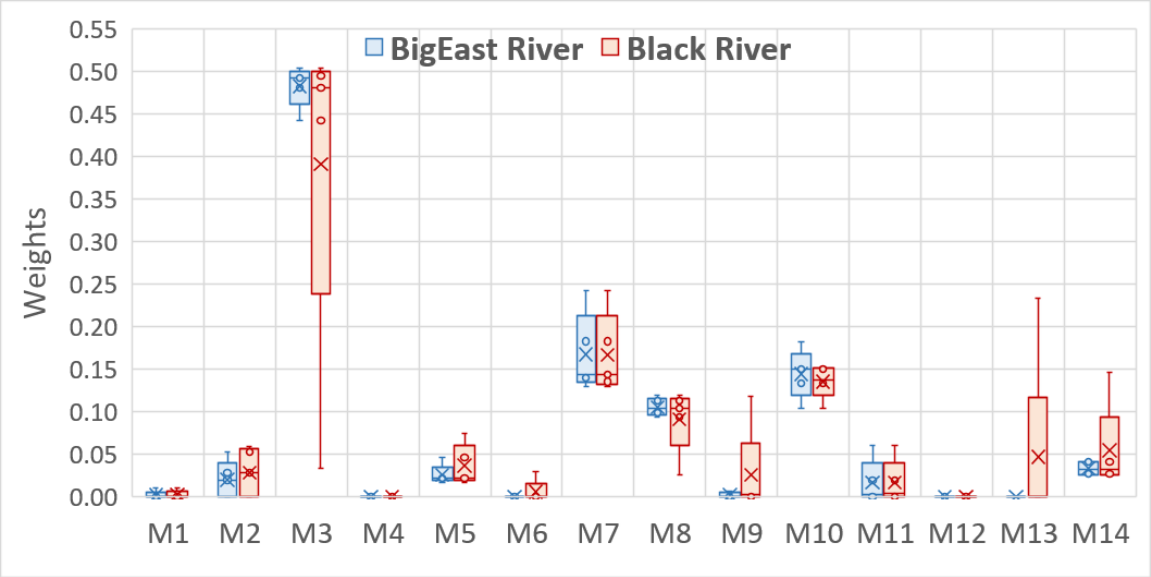


Figure 3-15 Scatter plots of different models' weights derived from the best-selected BMA variants

3.5 Summary and Conclusions

This study provides the first assessment of the previously proposed modifications for the original BMA methodology and documents how they affect the probabilistic and deterministic performance of the BMA-derived results for daily streamflow simulation. A scenario-based analysis was designed where the application of four diverse streamflow ensemble scenarios, different data transformation procedures, various distribution types, six different types of standard deviation, and two optimization algorithms were assessed thoroughly.

The summary of the obtained results from applying the proposed evaluation into two data-poor watersheds is as follows:

1. Comparing different ensemble scenarios indicated that, besides using multi-models, considering various forcing precipitation scenarios in generating members of an ensemble leads to better probabilistic and deterministic results in data scarce regions, where the estimation of mean areal precipitation always comes with noticeable errors. However, not only using a multi-model multi-parameter scenario did not provide better results, it also slightly reduced the reliability of the BMA simulations.
2. In contrast to earlier findings, however, the results showed that the BMA weights were not completely in accordance with individual model performance. There were some highly weighted hydrologic models with relatively lower performance in comparison to the others in both watersheds. In addition, various BMA modifications led to different combinations of weights and all had almost the same predictive power.
3. Applying data transformation generally yielded an improvement in the reliability of the BMA results. However, except for the empirical normal quantile approach, using other data transformation methods concurrent with implementing non-constant standard deviation without a constant parameter dramatically deteriorated the sharpness of the results, specifically in high flows.
4. Incorporation of the more representative distribution types did not show a particular superiority over the classic BMA method, where the posterior predictive distributions were assumed to be Gaussian. However, implementing non-constant standard deviations enhanced the predictive capability of the BMA model,

especially for high flows that are often of particular attention in operational hydrology.

5. The expectation-maximization algorithm provided almost the same results as the dynamically dimensioned search (DSS) method, which showed its ability to estimate BMA parameters well enough. However, the only drawback was that it could not easily be applied for all BMA variants when the distribution or standard deviation types were changed.

In general, the findings of this study suggest that the simulation skill of individual members are less important than how the whole ensemble captures the variability of the observation without overlapping. In other words, using ensemble members with diverse simulation skills can enhance the quality of the BMA results, while simply increasing the number of members in the ensemble does not always lead to better results. Although possessing high-performance models is necessary for obtaining reliable results, there is some information that is only provided by the relatively lower performing models and, consequently, considering them as members of the ensemble can enhance the BMA's predictive performance. The notable BMA weights of some of these models are another convincing justification for this conclusion. In addition, it was shown that in regions where the network of meteorological stations was sparse, using other sources of precipitation data, such as archived radar- or satellite-based products as inputs into the hydrologic models, can lead to a more exhaustive streamflow ensemble that enhances the BMA's performance.

Moreover, another implication of these results is that the most effective BMA modification in the positive direction (i.e., enhancing the predictive performance) is the implementation

of non-constant standard deviation. Increasing the variance of errors in line with flow level seems to be more realistic and enhances the reliability of the BMA results significantly for high flows (an average of 20% improvement in the reliability of high-flow simulations in both the Big East River and Black River watersheds over the whole period). However, considering the more representative distribution types does not highly affect the BMA-derived probabilistic and deterministic results. Moreover, although using data transformation procedures enhanced the reliability of the results, even more than applying non-constant variance, it can lead to a notable wide confidence interval width in high flows. Therefore, much more attention must be paid to the sharpness of the high-flow probabilistic simulation in the case of implementing data transformation. Furthermore, the results showed the robustness of the EM algorithm for estimating the original BMA parameters, while it was not easily applicable to all BMA modifications. Thus, applying a global optimization method is recommended in the case of using various BMA variants.

Although the two watersheds in this study share approximately the same land use and climatology, their hydrologic responses are not quite similar and lead to two different empirical CDFs of streamflow data. Therefore, it can be said that the aforementioned conclusions about the effects of different modifications on BMA results can be considered as useful recommendations in future studies. However, in order to provide more comprehensive conclusions, it is worth applying the proposed BMA modifications analysis in watersheds with very different topography and climatology (e.g., mountainous or coastal areas and tropical or semi-arid regions) in future studies. Furthermore, although possessing mutually exclusive and collectively exhaustive ensemble members is one of the main

assumptions of the BMA method, no studies have tried to overcome this issue. Although this study assessed the effects of various ensemble scenarios on BMA performance and provided fresh insight into the importance of establishing an ensemble with the aforementioned properties, there has not been a specific method about how these members should be generated and selected. Consequently, further studies need to be carried out to establish new ideas for solving this remaining challenge.

3.6 Author Contributions

Conceptualization, P.D. and P.C.; Data curation, P.D.; Formal analysis, P.D.; Investigation, P.D.; Methodology, P.D. and P.C.; Software, P.D. and P.C.; Supervision, P.C.; Validation, P.D. and P.C.; Visualization, P.C.; Writing—original draft, P.D.; Writing—review and editing, P.C.

3.7 Funding

This work was funded by the Natural Sciences and Engineering Research Council of Canada under Canadian FloodNet project (Grant number: NETGP 451456).

3.8 Acknowledgments

We would like to thank the Ministry of Natural Resources and Forestry, Surface Water Monitoring Center for providing some of the study data and Tara Razavi for providing Matlab source codes for the MACHBV, SACSMA, and Snow17 models.

3.9 Conflicts of Interest

The authors declare no conflict of interest.

3.10 References

Agnihotri, J. (2018). *Evaluation of snowmelt estimation techniques for enhanced spring peak flow prediction* [MS Thesis, McMaster University]. https://macsphere.mcmaster.ca/handle/11375/24099

Ajami, N. K., Duan, Q., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004745

American Society of Civil Engineers. (1996). *Hydrology handbook*. ASCE.

Anderson, E. A. (1973). *National Weather Service river forecast system: Snow accumulation and ablation model*. U.S. DEPARTMENT OF COMMERCE: National Oceanic and Atmospheric Administration, National Weather Service.

Anderson, E. A. (2006). *Snow Accumulation and Ablation Model – SNOW-17*. Natl. Ocean. Atmospheric Adm. Natl. Weather Serv. Silver Springs MD. https://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/part2/_pdf/22snow17.pdf

Anshuman, A., Kunnath-Poovakka, A., & Eldho, T. I. (2019). Towards the use of conceptual models for water resource assessment in Indian tropical watersheds under monsoon-driven climatic conditions. *Environmental Earth Sciences*, *78*(9), 282. https://doi.org/10.1007/s12665-019-8281-5

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., & Martel, J.-L. (2015). A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. *Journal of Hydrology*, *529*, 754–767. https://doi.org/10.1016/j.jhydrol.2015.09.001

Boluwade, A., Zhao, K.-Y., Stadnyk, T. A., & Rasmussen, P. (2018). Towards validation of the Canadian precipitation analysis (CaPA) for hydrologic modeling applications in the Canadian Prairies. *Journal of Hydrology*, *556*, 1244–1255. https://doi.org/10.1016/j.jhydrol.2017.05.059

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252. JSTOR.

Burnash, R. J. C., Ferral, R. L., & McGuire, R. A. (1973). *A generalized streamflow simulation system: Conceptual modeling for digital computers*. Joint Federal-State River Forecast Center, United States National Weather Service.

Chen, X., Yang, T., Wang, X., Xu, C.-Y., & Yu, Z. (2013). Uncertainty Intercomparison of Different Hydrological Models in Simulating Extreme Flows. *Water Resources Management*, *27*(5), 1393–1409. https://doi.org/10.1007/s11269-012-0244-5

Cunderlik, J., & Simonovic, S. (2004). *Calibration, Verification and Sensitivity Analysis of the HEC-HMS Hydrologic Model*. Department of Civil and Environmental Engineering, The University of Western Ontario. https://ir.lib.uwo.ca/wrrr/11

Dong, L., Xiong, L., & Zheng, Y. (2013). Uncertainty analysis of coupling multiple hydrologic models and multiple objective functions in Han River, China. *Water Science and Technology*, *68*(3), 506–513. https://doi.org/10.2166/wst.2013.255

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Ebtehaj, M., Moradkhani, H., & Gupta, H. V. (2010). Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling: HYDROLOGIC PARAMETER ESTIMATION. *Water Resources Research*, *46*(7). https://doi.org/10.1029/2009WR007981

Edijanto, Nascimento, N. D. O., Yang, X., Makhlouf, Z., & Michel, C. (1999). GR3J: A daily watershed model with three free parameters. *Hydrological Sciences Journal*, *44*(2), 263–277. https://doi.org/10.1080/02626669909492221

Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J., & Butts, M. B. (2004). Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology*, *298*(1), 222–241. https://doi.org/10.1016/j.jhydrol.2004.03.037

Granger, C. W., & Ramanathan, R. (1984). Improved Methods of Combining Forecasts: ABSTRACT. *Journal of Forecasting (Pre-1986); Chichester*, *3*(2), 197–204.

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

Gyawali Rabi & Watkins David W. (2013). Continuous Hydrologic Modeling of Snow-Affected Watersheds in the Great Lakes Basin Using HEC-HMS. *Journal of Hydrologic Engineering*, *18*(1), 29–39. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000591

Hargreaves, G. H., & Samani, Z. A. (1985). Reference Crop Evapotranspiration from Temperature. *Applied Engineering in Agriculture*, *1*(2), 96–99.

He, S., Guo, S., Liu, Z., Yin, J., Chen, K., & Wu, X. (2018). Uncertainty analysis of hydrological multi-model ensembles based on CBP-BMA method. *Hydrology Research*, *49*(5), 1636–1651. https://doi.org/10.2166/nh.2018.160

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, *14*(4), 382–401. JSTOR.

Hornberger, G. M., & Spear, R. C. (1981). Approach to the preliminary analysis of environmental systems. *Journal of Environmental Management*, *12*(1), 7–18.

Huo, W., Li, Z., Wang, J., Yao, C., Zhang, K., & Huang, Y. (2019). Multiple hydrological models comparison and an improved Bayesian model averaging approach for ensemble prediction over semi-humid regions. *Stochastic Environmental Research and Risk Assessment*, *33*(1), 217–238. https://doi.org/10.1007/s00477-018-1600-7

Krzysztofowicz, R. (1997). Transformation and normalization of variates with specified distributions. *Journal of Hydrology*, *197*(1), 286–292. https://doi.org/10.1016/S0022-1694(96)03276-3

Lespinas, F., Fortin, V., Roy, G., Rasmussen, P., & Stadnyk, T. (2015). Performance Evaluation of the Canadian Precipitation Analysis (CaPA). *Journal of Hydrometeorology*, *16*(5), 2045–2064. https://doi.org/10.1175/JHM-D-14-0191.1

Liang, G. C. (1992). *A note on the revised SMAR model*. Memorandum to the River Flow Forecasting Workshop Group, Department of Engineering Hydrology, University College Galway. https://ci.nii.ac.jp/naid/10018251872/

Liang, Z., Wang, D., Guo, Y., Zhang, Y., & Dai, R. (2013). Application of Bayesian Model Averaging Approach to Multimodel Ensemble Hydrologic Forecasting. *Journal of Hydrologic Engineering*, *18*(11), 1426–1436. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000493

Liu, J., & Xie, Z. (2014). BMA Probabilistic Quantitative Precipitation Forecasting over the Huaihe Basin Using TIGGE Multimodel Ensemble Forecasts. *Monthly Weather Review*, *142*(4), 1542–1555. https://doi.org/10.1175/MWR-D-13-00031.1

Liu, Z., Guo, S., Zhang, H., Liu, D., & Yang, G. (2016). Comparative Study of Three Updating Procedures for Real-Time Flood Forecasting. *Water Resources Management*, *30*(7), 2111–2126. https://doi.org/10.1007/s11269-016-1275-0

Ma, Y., Hong, Y., Chen, Y., Yang, Y., Tang, G., Yao, Y., Long, D., Li, C., Han, Z., & Liu, R. (2018). Performance of Optimally Merged Multisatellite Precipitation Products Using the Dynamic Bayesian Model Averaging Scheme Over the Tibetan Plateau. *Journal of Geophysical Research: Atmospheres*, *123*(2), 814–834. https://doi.org/10.1002/2017JD026648

Madadgar, S., & Moradkhani, H. (2014). Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resources Research*, *50*(12), 9586–9603. https://doi.org/10.1002/2014WR015965

McLachlan, G., & Krishnan, T. (2008). *The EM Algorithm and Extensions* (2 edition). Wiley-Interscience.

Meira Neto, A., Oliveira, P. T. S., Rodrigues, D. B., & Wendland, E. (2018). Improving Streamflow Prediction Using Uncertainty Analysis and Bayesian Model Averaging. *Journal of Hydrologic Engineering*, *23*(5), 05018004. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001639

Michaels, S. (2015). Probabilistic forecasting and the reshaping of flood risk management. *Journal of Natural Resources Policy Research*, *7*(1), 41–51. https://doi.org/10.1080/19390459.2014.970800

Moradkhani, H., & Sorooshian, S. (2008). General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis. In S. Sorooshian, K.-L. Hsu, E. Coppola, B. Tomassetti, M. Verdecchia, & G. Visconti (Eds.), *Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrological Models* (Vol. 63, pp. 1–24). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77843-1_1

Najafi, M. R., & Moradkhani, H. (2016). Ensemble Combination of Seasonal Streamflow Forecasts. *Journal of Hydrologic Engineering*, *21*(1), 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. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Neuman, S. P. (2003). Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment*, *17*(5), 291–305. https://doi.org/10.1007/s00477-003-0151-7

Parrish, M. A., Moradkhani, H., & DeChant, C. M. (2012). Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation: TOWARD REDUCTION OF MODEL UNCERTAINTY. *Water Resources Research*, *48*(3). https://doi.org/10.1029/2011WR011116

Qu, B., Zhang, X., Pappenberger, F., Zhang, T., & Fang, Y. (2017). Multi-Model Grand Ensemble Hydrologic Forecasting in the Fu River Basin Using Bayesian Model Averaging. *Water*, *9*(2), 74. https://doi.org/10.3390/w9020074

Raftery, A. E. (1993). Bayesian Model Selection in Structural Equation Models. In *Testing Structural Equation Models* (Vol. 154, pp. 163–180). SAGE.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, *92*(437), 179–191. https://doi.org/10.1080/01621459.1997.10473615

Refsgaard, J. C., & Knudsen, J. (1996). Operational Validation and Intercomparison of Different Types of Hydrological Models. *Water Resources Research*, *32*(7), 2189–2202. https://doi.org/10.1029/96WR00896

Rojas, R., Feyen, L., & Dassargues, A. (2008). Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2008WR006908

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2011). Estimation of Continuous Streamflow in Ontario Ungauged Basins: Comparison of Regionalization Methods. *Journal of Hydrologic Engineering*, *16*(5), 447–459. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000338

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2012). Identification of rainfall–runoff model for improved baseflow estimation in ungauged basins. *Hydrological Processes*, *26*(3), 356–366. https://doi.org/10.1002/hyp.8133

Scharffenberg, W. (2016). *HEC-HMS User's Manual, Version 4.2*. U.S. Army Corps of Engineers Institute for Water Resources Hydrologic Engineering Center (CEIWR-HEC).

Seo, D.-J., Herr, H. D., & Schaake, J. C. (2006). A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrology and Earth System Sciences Discussions*, *3*(4), 1987–2035. https://doi.org/10.5194/hessd-3-1987-2006

Shamseldin, ASAAD Y., & O'Connor, K. M. (1999). A real-time combination method for the outputs of different rainfall-runoff models. *Hydrological Sciences Journal*, *44*(6), 895–912. https://doi.org/10.1080/02626669909492288

Shamseldin, Asaad Y., O'Connor, K. M., & Liang, G. C. (1997). Methods for combining the outputs of different rainfall–runoff models. *Journal of Hydrology*, *197*(1), 203–229. https://doi.org/10.1016/S0022-1694(96)03259-3

Shrestha, D. L. (2009). *Uncertainty analysis in rainfall-runoff modelling - application of machine learning techniques: UNESCO-IHE PhD thesis.* [PhD. thesis, IHE Delft Institute for Water Education]. https://www.cabdirect.org/cabdirect/abstract/20123116250

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., & Fraley, C. (2007). Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, *135*(9), 3209–3220. https://doi.org/10.1175/MWR3441.1

Strauch, M., Bernhofer, C., Koide, S., Volk, M., Lorz, C., & Makeschin, F. (2012). Using precipitation data ensemble for uncertainty analysis in SWAT streamflow simulation. *Journal of Hydrology*, *414–415*, 413–424. https://doi.org/10.1016/j.jhydrol.2011.11.014

Sun, R., Yuan, H., & Yang, Y. (2018). Using multiple satellite-gauge merged precipitation products ensemble for hydrologic uncertainty analysis over the Huaihe River basin. *Journal of Hydrology*, *566*, 406–420. https://doi.org/10.1016/j.jhydrol.2018.09.024

Tegegne, G., Park, D. K., & Kim, Y.-O. (2017). Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. *Journal of Hydrology: Regional Studies*, *14*, 49–66. https://doi.org/10.1016/j.ejrh.2017.10.002

Thiessen, A. H. (1911). Precipitation averages for large areas. *Monthly Weather Review*, *39*(7), 1082–1089. https://doi.org/10.1175/1520-0493(1911)39<1082b:PAFLA>2.0.CO;2

Thornthwaite, C. W. (1948). An Approach toward a Rational Classification of Climate. *Geographical Review*, *38*(1), 55–94. JSTOR. https://doi.org/10.2307/210739

Tian, Y., Booij, M. J., & Xu, Y.-P. (2014). Uncertainty in high and low flows due to model structure and parameter errors. *Stochastic Environmental Research and Risk Assessment*, *28*(2), 319–332. https://doi.org/10.1007/s00477-013-0751-9

Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, *6*(2), 123–137. https://doi.org/10.1080/15715124.2008.9635342

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Viallefont, V., Raftery, A. E., & Richardson, S. (2001). Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine*, *20*(21), 3215–3230. https://doi.org/10.1002/sim.976

Vrugt, J. A. (2016). *MODELAVG: A MATLAB Toolbox for Postprocessing of Model Ensembles*. http://faculty.sites.uci.edu/jasper/files/2016/04/manual_Model_averaging.pdf

Vrugt, J. A., Diks, C. G. H., & Clark, M. P. (2008). Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling. *Environmental Fluid Mechanics*, *8*(5), 579–595. https://doi.org/10.1007/s10652-008-9106-3

Vrugt, J. A., & Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004838

Xiong, L., Wan, M., Wei, X., & O'Connor, K. M. (2009). Indices for assessing the prediction bounds of hydrological models and application bysection generalised likelihood uncertainty estimation / Indices pour évaluer les bornes de prévision de modèles hydrologiques et mise en œuvre pour une estimation d'incertitude par vraisemblance généralisée. *Hydrological Sciences Journal*, *54*(5), 852–871. https://doi.org/10.1623/hysj.54.5.852

Yan, H., & Moradkhani, H. (2016). Toward more robust extreme flood prediction by Bayesian hierarchical and multimodeling. *Natural Hazards*, *81*(1), 203–225. https://doi.org/10.1007/s11069-015-2070-6

Yen, H., Wang, X., Fontane, D. G., Harmel, R. D., & Arabi, M. (2014). A framework for propagation of uncertainty contributed by parameterization, input data, model structure, and calibration/validation data in watershed modeling. *Environmental Modelling & Software*, *54*, 211–221. https://doi.org/10.1016/j.envsoft.2014.01.004

Zeng, X., Wu, J., Wang, D., Zhu, X., & Long, Y. (2016). Assessing Bayesian model averaging uncertainty of groundwater modeling based on information entropy method. *Journal of Hydrology*, *538*, 689–704. https://doi.org/10.1016/j.jhydrol.2016.04.038

Zhang, X., Srinivasan, R., & Bosch, D. (2009). Calibration and uncertainty analysis of the SWAT model using Genetic Algorithms and Bayesian Model Averaging. *Journal of Hydrology*, *374*(3), 307–317. https://doi.org/10.1016/j.jhydrol.2009.06.023

Chapter 4. Introducing entropy-based Bayesian model averaging for streamflow forecast

Summary of Paper 3: Darbandsari, P., & Coulibaly, P. (2020). Introducing entropy-based Bayesian model averaging for streamflow forecast. Journal of Hydrology, 591, 125577.

In this study, a new ensemble-based probabilistic post-processing framework is proposed where an entropy-based selection procedure is implemented to generate an ensemble of forecasts with mutually exclusive and collectively exhaustive characteristics prior to the Bayesian Model Averaging (BMA) application. Comparison has been performed between the traditional BMA and the proposed approach (En-BMA) for probabilistic daily streamflow forecasting.

Key findings of this research work are:

- Higher information can be provided by generating an ensemble of streamflow forecasts using various hydrologic models being calibrated by different objective functions.
- The proposed entropy-based selection procedure can select the subset of forecasts with high information content and low mutual dependency which are the vital requirements for reliable performance of the BMA method.
- The proposed En-BMA post-processing approach, compared to the BMA method, provides more reliable and accurate high flow forecasts.

- The superiority of the proposed En-BMA method over BMA presents in all lead-times while it is more noticeable for shorter ones.

4.1 Abstract

Bayesian Model Averaging (BMA) is a well-known statistical post-processing approach for probabilistically merging individual forecasts. In BMA, the posterior distribution of the predictand variable is determined by implementing the law of total probability. Therefore, possessing an ensemble of independent members (mutually exclusive) with the highest information content about observation variability (collectively exhaustive) is the main inherent assumption of the original BMA method. Mutually exclusive and collectively exhaustive are two contradictory criteria. Although constructing an ensemble of members that fully satisfied these two properties is practically impossible, providing a balance between them is a key requirement for enhancing the BMA performance. Through coupling BMA with Shannon entropy of information theory, this study proposes an entropy-based selection procedure to construct an ensemble of streamflow forecasts by better addressing the aforementioned contradictory criteria prior to performing the BMA. We investigate the effects of using ensembles with the aforementioned properties by comparing the results of original BMA with the proposed entropy-based BMA (En-BMA) for short- to medium-range daily streamflow forecasts in two different watersheds. The results indicate that the En-BMA leads to better results particularly for high flow predictions. Both probabilistic and deterministic high flow forecasts are more accurate and reliable when using the En-BMA approach. However, for the average flow forecasts, there are no clear differences in the general performance of both methods. The improvements observed are more

pronounced for shorter lead-times and less pronounced, but still present, for longer lead times.

Keywords: Bayesian model averaging; Streamflow forecasting; Uncertainty; Entropy; Information theory

4.2 Introduction

Reliable streamflow prediction is an essential task for various water management issues, from flood forecasting and reservoir operation to recreational activities. Various sources of uncertainties associated with forcing inputs, initial conditions, model parameters, and model structures affect the reliability of hydrological forecasts (Moradkhani & Sorooshian, 2008; Shrestha, 2009; Xu et al., 2019). Generating an ensemble streamflow prediction (ESP) is one of the most common approaches for quantifying different uncertainties (Madadgar & Moradkhani, 2014; Michaels, 2015; Seo et al., 2006). It is shown that an ESP is more skillful and functional than deterministic systems for operational purposes (Boucher et al., 2011; Xu et al., 2019). Besides using various meteorological forcing inputs and perturbing initial states of the model, ESP can be created by utilizing multiple hydrologic models in order to quantify the model structural uncertainty and prevent statistical bias of the prediction (Darbandsari & Coulibaly, 2019; Parrish et al., 2012).

Various statistical and post-processing tools have been developed for optimally merging the individual members of the ESP. Simple averaging (DelSole, 2007) and Granger–Ramanathan averaging (Granger & Ramanathan, 1984) are the simplest ones providing one-point deterministic results. However, some more complex methods, such as Bayesian

Model Averaging (BMA; (Hoeting et al., 1999; Raftery, 1993; Raftery et al., 1997, 2005)), generate probabilistic forecasts by quantifying predictive uncertainty. BMA is a statistical procedure using the weighted average of the probability distribution function (PDF) of different individual forecasts for generating predictive forecast distributions. In comparison to the other multi-model combination methods, the higher capability of the BMA approach in producing more accurate and reliable forecasts has been shown by various studies (Arsenault et al., 2015; Viallefont et al., 2001).

Exploring the application of the BMA approach in the field of streamflow predictions/simulations has led to different variants by relaxing some inherent assumptions of the original BMA. Besides proposing some minor modifications, such as implementing more representative distribution types (Vrugt & Robinson, 2007) or applying data transformation (Duan et al., 2007; Liang et al., 2013; Qu et al., 2017), several more complicated BMA based post-processing methods have been developed, such as combining BMA and Data Assimilation (Parrish et al., 2012; Rings et al., 2012), Copulas and BMA (Madadgar & Moradkhani, 2014), and Copula Bayesian Processors with BMA (He et al., 2018). All BMA variants listed attempt to relax the Gaussian assumption of the posterior distributions, while there are some fundamental limitations of the original BMA method which remain.

One of the primary principles of the standard BMA formulations is the law of total probability. This principle leads to the assumption of possessing mutually exclusive (i.e. independent), as well as collectively exhaustive (i.e. capturing observation variability) members of the ensemble. In other words, having an ensemble of members with the least

shared information and higher capability of covering possible futures, is a potential requirement for reliable performance of the BMA approach. It has been shown that selecting independent forecast members enhance the reliability of the BMA results (Sharma et al., 2019); however, capturing the variability of the observation by using an ensemble is not possible except by having collectively exhaustive members (Madadgar & Moradkhani, 2014). Simply constructing a large ensemble of members can provide more information about observation and relatively assure the latter requirement; while, it may limit the former one, as the larger number of members can lead to higher redundant information within the ensemble (Refsgaard et al., 2012). Given the contradiction between the two criteria, it is impossible to possess a mutually exclusive and collectively exhaustive ensemble simultaneously. Therefore, constructing an ESP by providing a balance between the two criteria is necessary and can positively affect the performance of the BMA method.

The information theory, also known as Shannon entropy, was first introduced by Shannon (1948) and has become very popular in several scientific fields. The definition of the entropy term in the context of information theory is a measure of uncertainty in a random variable; and based on the fact that the amount of uncertainty will be reduced if more information is available, entropy corresponds to the amount of information contained in a data set (Keum & Coulibaly, 2017a). There are various applications of information theory in solving different water-related issues (Mishra & Coulibaly, 2009; Singh, 1997). One common application of entropy in water resources has been to aid in the design of hydrometric monitoring networks (Alfonso et al., 2010, 2013; Keum et al., 2019; Keum &

Coulibaly, 2017b; Leach et al., 2015; Li et al., 2012), where gathering high-quality information with minimal redundancy is the main objective.

This study seeks to establish a new entropy-based selection procedure using the proven capability of the information theory concept to provide information with minimal redundancy. The new method will integrate entropy with BMA in order to overcome the remaining challenge of possessing mutually exclusive and collectively exhaustive (MECE) ensemble. In the proposed Entropy-based BMA (En-BMA) framework, before estimating BMA parameters, three entropy measures (joint entropy, total correlation, and transinformation) are utilized to narrow down the streamflow forecasts for constructing ensemble with the MECE properties. The applicability and efficiency of the proposed En-BMA approach have been assessed in two different watersheds for short- to medium-range (up to 7 days) daily streamflow predictions. By providing a balance between two conflicting criteria, the results show the superiority of the En-BMA in providing better probabilistic and deterministic high flow forecasts over the standard BMA approach.

The structure of the paper is as follows. Section 4.3 introduces the underlying concepts of our new En-BMA method. The experimental setup, including brief explanations of the study areas, hydrologic models, and different verification metrics, are presented in Section 4.4. Section 4.5 provides the comparison results between original BMA and proposed En-BMA methods, and finally, the summary and conclusions are drawn in Section 4.6.

4.3 Methodology

4.3.1 Definition of Entropy terms

As defined in information theory, entropy is a measure of the amount of information required to describe a random variable. In other words, it is the amount of uncertainty represented by the probability distribution of a random variable. The basis of the Shannon entropy is that the information gained from an event with occurrence probability p is $\log(1/p)$. This stems from the fact that the anticipatory uncertainty from an event varies inversely with its probability. Also, the logarithmic function is the only transition that can be used in order to make sure that the information gained by the joint occurrence of two independent events is equal to the sum of the information from each one:

$$\log\left(\frac{1}{p(x_1)p(x_2)}\right) = \log\left(\frac{1}{p(x_1)}\right) + \log\left(\frac{1}{p(x_2)}\right) = -\log\big(p(x_1)\big) - \log\big(p(x_2)\big) \quad \textit{(4-1)}$$

Consequently, in the case of a discrete variable (X) with K outcomes with probabilities ($p_1, p_2, \dots, p_K$), the average information of X is denoted by:

$$H(X) = E\left(\log\left(\frac{1}{p_1 p_2 \dots p_k}\right)\right) = -\sum_{i=1}^{K} p_i \log(p_i) \quad \textit{(4-2)}$$

where $E(.)$ is the expectation function and $H(X)$ is the marginal entropy of a single variable X in bits, because the base of the logarithm is assumed to be equal to 2. Therefore, marginal entropy is the amount of information gained by knowing a single variable and it varies between zero, for a deterministic case, and $logN$, for the most uncertain cases

(uniformly distributed variables). It is of note that for continuous variables, such as streamflow data, a finite number of discrete data intervals must be chosen.

A similar procedure is used for calculating the total information content in more than two variables (e.g., N variables), which is known as joint (multivariate) entropy ($H(X_1, X_2, \dots, X_N)$):

$$H(X_1, X_2, \dots, X_N) = \sum_{i_1=1}^{K_1} \sum_{i_2=1}^{K_2} \dots \sum_{i_N=1}^{K_N} p\left(x_{1,i_1}, x_{2,i_2}, \dots x_{N,i_N}\right) \log\left(p\left(x_{1,i_1}, x_{2,i_2}, \dots x_{N,i_N}\right)\right) \quad (4\text{-}3)$$

where $p\left(x_{1,i_1}, x_{2,i_2}, \dots x_{N,i_N}\right)$ is the joint probability of all variables and K_j ($j \in [1, N]$) is the number of values or class intervals for variable x_j in the case of discrete or continuous variables, respectively. If there are independent variables, multivariate entropy is equal to the summation of their marginal entropies and its maximum value will occur in the case of independent, equally likely variables.

In general, marginal and joint entropies are related as follows:

$$H(X_1, X_2, \dots, X_N) = \sum_{i=1}^{N} H(X_i) - C(X_1, X_2, \dots, X_N) \quad (4\text{-}4)$$

where $C(X_1, X_2, \dots, X_N)$ is the total correlation which estimates the amount of duplicated information in multiple variables. It is of note that increasing the number of variables could potentially lead to larger total correlation (Figure 4-1). If the number of variables is reduced to two, Equation 4-4 can be rewritten as follows:

$$H(X_1, X_2) = H(X_1) + H(X_2) - T(X_1, X_2) \quad (4\text{-}5)$$

where $T(X_1, X_2)$ is the amount of mutual information (or transinformation) between variables X_1 and X_2. In other words, it shows the information content of one variable that is contained in another. Transinformation is symmetric (i.e., $T(X_1, X_2) = T(X_2, X_1)$) and will be equal to zero when two variables are statistically independent. The larger value of transinformation depicts the higher dependence between the variables and the maximum value occurs in the case of functionally dependent variables. Therefore, transinformation varies in the range of zero to $min\big(H(X_1), H(X_2)\big)$.

For more clarification, Figure 4-1 illustrates the schematic of the various aforementioned entropy terms for a case of three variables, where the circle sizes indicate the amount of marginal entropy. As can be seen, transinformation is only meaningful for two variables (or group of variables) and is not equal to total correlation when we possess three or more data sets. Moreover, the total correlation is the sum of all order duplications in the system. Based on the definition of the previously mentioned entropy terms, it can be concluded that these concepts can be used in order to relax some of the remaining assumptions and hypotheses of the BMA method (i.e., independent and mutually exclusive predictions).

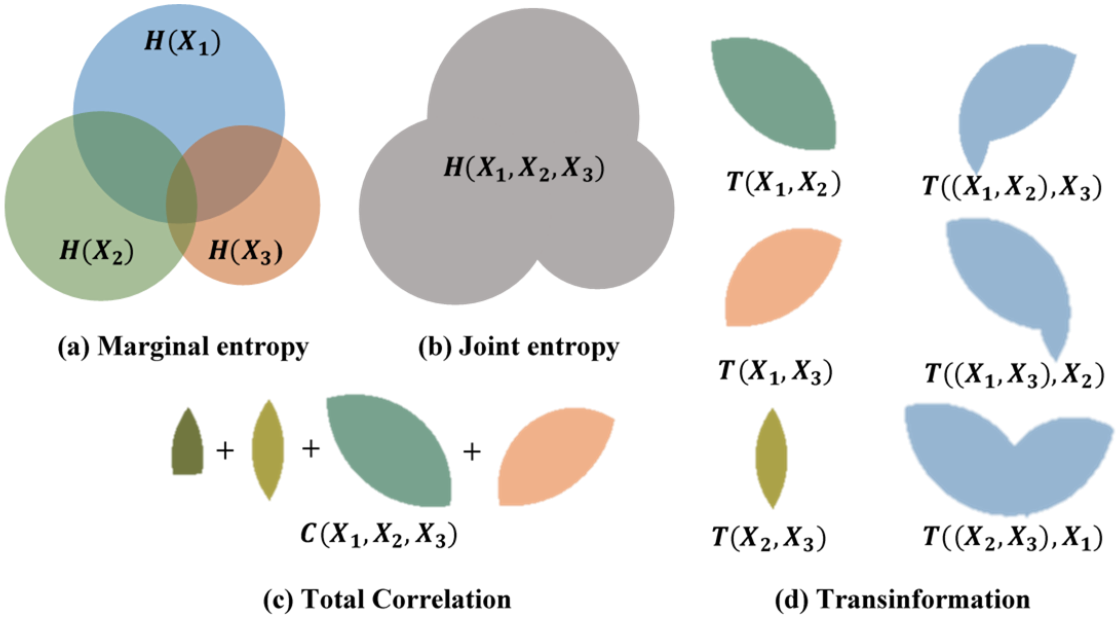


Figure 4-1 The schematics of (a) marginal entropy, (b) joint entropy, (c) total correlation, and (d) transinformation

4.3.2 Bayesian Model Averaging with Moving Window

BMA is a probabilistic post-processor where the conditional PDFs of various forecasts are combined in order to generate more skillful predictions. BMA was first introduced for statistical linear regression applications (Hoeting et al., 1999; Kass & Raftery, 1995). Raftery et al. (Raftery et al., 2005) extended the application of BMA to dynamic models. Given the detailed description of the BMA approach in the literature (Darbandsari & Coulibaly, 2019; Raftery et al., 2005), we briefly explained the BMA basic concepts for the sake of completeness.

Consider the quantity Δ to be the predictand (i.e., forecasted variable) based on the training data D and the ensemble of independent predictions $= \{F_1, F_2, \dots, F_k\}$. Based on the law of total probability, the BMA predictive PDF of Δ can be represented as.

$$P(\Delta|F, D) = \sum_{i=1}^{k} P(\Delta|F_i, D) \times P(F_i|D) \qquad (4\text{-}6)$$

where $P(F_i|D)$ is the posterior probability of the forecast F_i being correct given the observational data. This term can be viewed as a weight (w_i) reflecting how well the ensemble member i fits the observations in the training period. Moreover, $P(\Delta|F_i, D)$ is the PDF of the predictand Δ conditional on the forecast F_i and observed data D. In the standard BMA approach, this posterior probability is assumed to follow the Gaussian distribution centered at the forecast value with standard deviation σ_i.

In order to estimate the parameters of the BMA approach, (i.e., weights and variances of each individual forecast), the Expectation-Maximization (EM) algorithm, proposed by Raftery et al. (2005), is applied for maximizing the log-likelihood function of the parameter vector (θ):

$$L(\theta) = Log\big(P(\Delta|F_1, F_2, \dots, F_K, D)\big) = Log\left(\sum_{i=1}^{K} w_i \times g(y|F_i, \sigma_i^2)\right) \qquad (4\text{-}7)$$

M is an iterative approach, including expectation and maximization steps, where a latent variable is used for searching the optimal values of parameters. Although obtaining a global optimum solution is not guaranteed, it has been shown that EM is as reliable and efficient as more complex global optimization techniques (Darbandsari & Coulibaly, 2019; Vrugt

et al., 2008). For a more detailed description of the EM algorithm, the reader is referred to the above-cited references.

Using a fixed training set of data leads to a static estimation of the BMA parameters, which does not change with respect to the hydrologic regime. However, updating the parameters when new observations are available may provide more reliable results. Therefore, following Raftery et al. (2005), the moving window approach is implemented where the shorter window of simulation-observation pairs surrounding each forecast is used as the recursive training period for calculating BMA parameters. By capturing the time-dependent relative performance of various members of the ensemble, the BMA with moving window leads to better probabilistic forecasts (Parrish et al., 2012; Vrugt & Robinson, 2007).

4.3.3 Entropy-based Bayesian Model Averaging method

As previously stated, establishing a balanced ensemble of forecasts with mutually exclusive (independent) and collectively exhaustive (capturing the observation variability) members can potentially lead to more reliable BMA derived predictive forecasts. Here, by using three aforementioned entropy terms, we introduce an easy-to-implement selection procedure, through which the generated ensemble can (1) possess minimum redundancy and (2) assure the highest overall information.

Figure 4-2 represents the proposed Entropy-based selection algorithm for optimally choosing the subset of forecast members with minimum redundancy and maximum information for BMA application. As can be seen, the method has a nested loop structure.

The goal of the inner loop is finding less independent forecasts when the number of selected members is fixed. Therefore, the total correlation between selected members is considered as an objective function, and its values for potentially selected subsets (S) are compared in order to find the optimal one (S_F). Different subsets are initialized by iteratively removing one member of the candidate forecasts and finally, the subset with the lowest dependence within its members is the selected one. In other words, the finally omitted member in each loop possesses the most duplicate information in common with other members, which leads to the highest redundancy of the ensemble.

In the outer loop, the stopping criteria are defined in order to provide collectively exhaustive ensemble. Therefore, we introduce two entropy-based ratios. The first is the ratio of joint entropy of the selected optimal members, derived from the inner loop, to that of all the candidate members ($F = \{F_1, F_2, .., F_K\}$). This ratio shows how much of the information contained by the whole ensemble is covered using the selected subset. Although it illustrates the exhaustiveness of the selected members, it is not a proper term for representing the amount of information about capturing the variability in the observations. Therefore, the second criterion is defined as transinformation between the final selected subset and observations over transinformation between all candidate members and observations. This provides a better estimation of the maximum information. However, both ratios must be used together to assure collectively exhaustive criterion in both calibration and forecasting periods. Figure 4-3 exemplifies the application of the proposed selection procedure for one forecast by illustrating how different entropy terms change in both inner and outer loops. As previously stated, the inner loop of the procedure

determines the best subset of ensemble with minimum total correlation and the outer loop continues until both stopping criteria are fulfilled. The steady decrease of total correlation in the outer loop shows selection of less dependent members while the imposed stopping threshold parameter (β), ensures possessing an ensemble with sufficient information content.

The integration of the proposed Entropy-based selection algorithm with BMA using the moving window scheme is presented in Figure 4-4. For each forecast, the moving window with length N is considered as the training period for implementing the selection procedure and estimating BMA parameters. The results of the training phase are used during the forecasting mode, where at first, the ensemble of streamflow forecasts up to T days ahead are generated using the selected members and deterministic precipitation (P_f) and temperature (T_f) forecasts, and then, the BMA approach with estimated parameters is utilized as a post-processor for probabilistically merging the generated forecasts. It is of note that, in this study, we used observed precipitation and temperature as perfect deterministic meteorological forecasts. By allowing the selected members to vary for each forecast, we hope the proposed En-BMA is able to provide better deterministic and probabilistic results in comparison to the original BMA where the same ensemble is implemented.

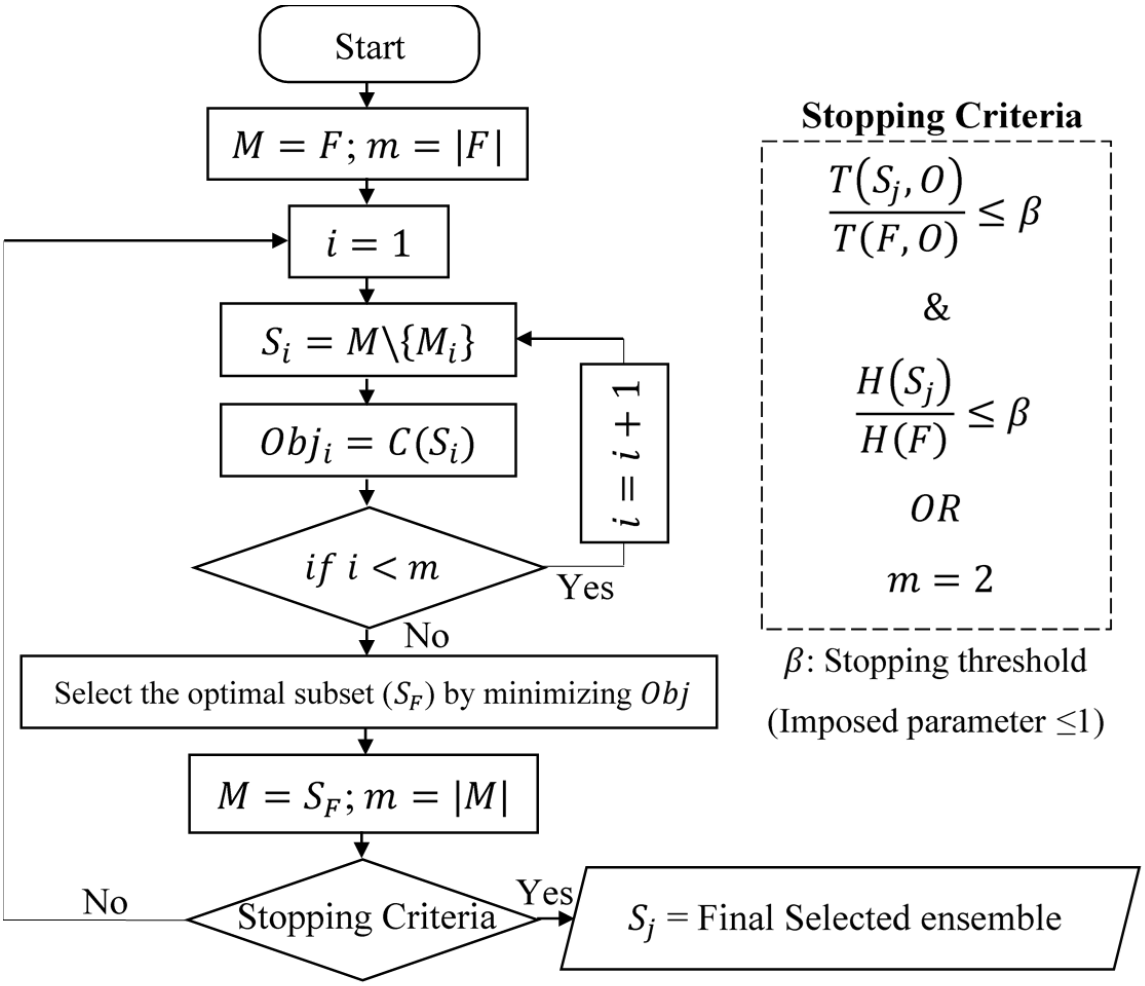


Figure 4-2 The greedy algorithm of the Entropy-based selection procedure. $F = \{F_1, F_2, .., F_K\}$ is the set of all candidate forecast members. m shows the number of members of the ensemble. S_i is a candidate ensemble subset after removing member i. $H(.)$, $C(.)$ and $T(.)$ respectively are the functions of joint entropy, total correlation and transinformation

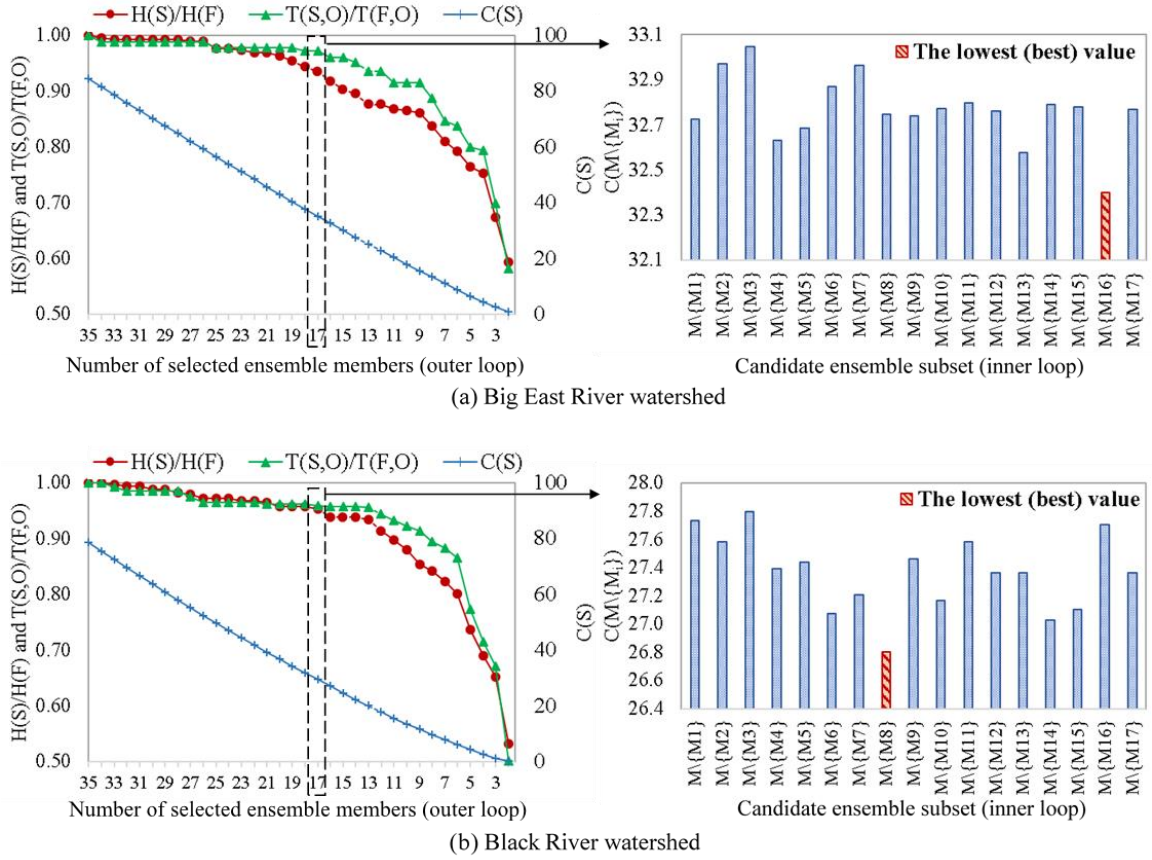


Figure 4-3 An example of the application of the proposed selection procedure in both (a) Big East River and (b) Black River watersheds. F and S are the ensembles considering all and selected members, respectively, and O is the observation. $H(.)$, $T(.)$, and $C(.)$ respectively show the functions of joint entropy, transinformation and total correlation in bits

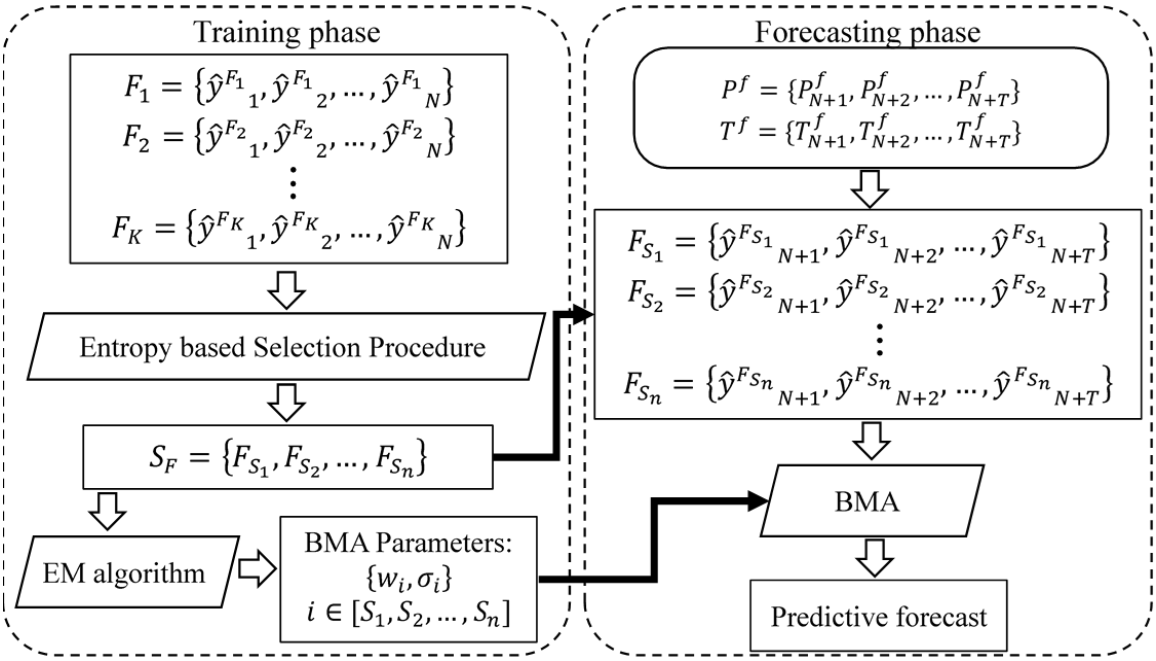


Figure 4-4 The main flowchart of the Entropy-based Bayesian Model Averaging (En-BMA) with Moving window scheme. N: moving window length; T: forecast lead-time; K: total number of candidate members; S_n: the number of selected members; P^f and T^f are inputs for the selected models

4.4 Experimental Setup

4.4.1 Study Area

The proposed En-BMA approach is applied to the Big East River (BE) and Black River (BL) basins, located in northern Ontario, Canada (Figure 4-5). BE is a mostly forested area of 620 km^2 while BL, with a drainage area of 1522 km^2, is covered with the combination of agricultural lands and natural forests. Both watersheds are moderately sloped with altitudes approximately changing from 200 to 400 and 200 to 600 meters above sea level for BE and BL, respectively. From the six available meteorological stations, located around both watersheds, the mean annual precipitation ranges between 887 to 1249 mm. In

addition, the average temperature variations from -10.5°C in January to 18.5°C in July depict the occurrence of all four seasons in both watersheds (Table 4-1). Moreover, in April, when the temperature rises to above the freezing point and snowmelt begins, the highest amount of discharge at the outlet of both watersheds can be seen. This depicts the importance of considering the snowmelt routine in the hydrological modeling of both watersheds. It is noted that the low-density ground-based stations shows the status of data-poor watersheds where the conceptual models are the most appropriate ones for rainfall-runoff modeling (Anshuman et al., 2019; Tegegne et al., 2017).

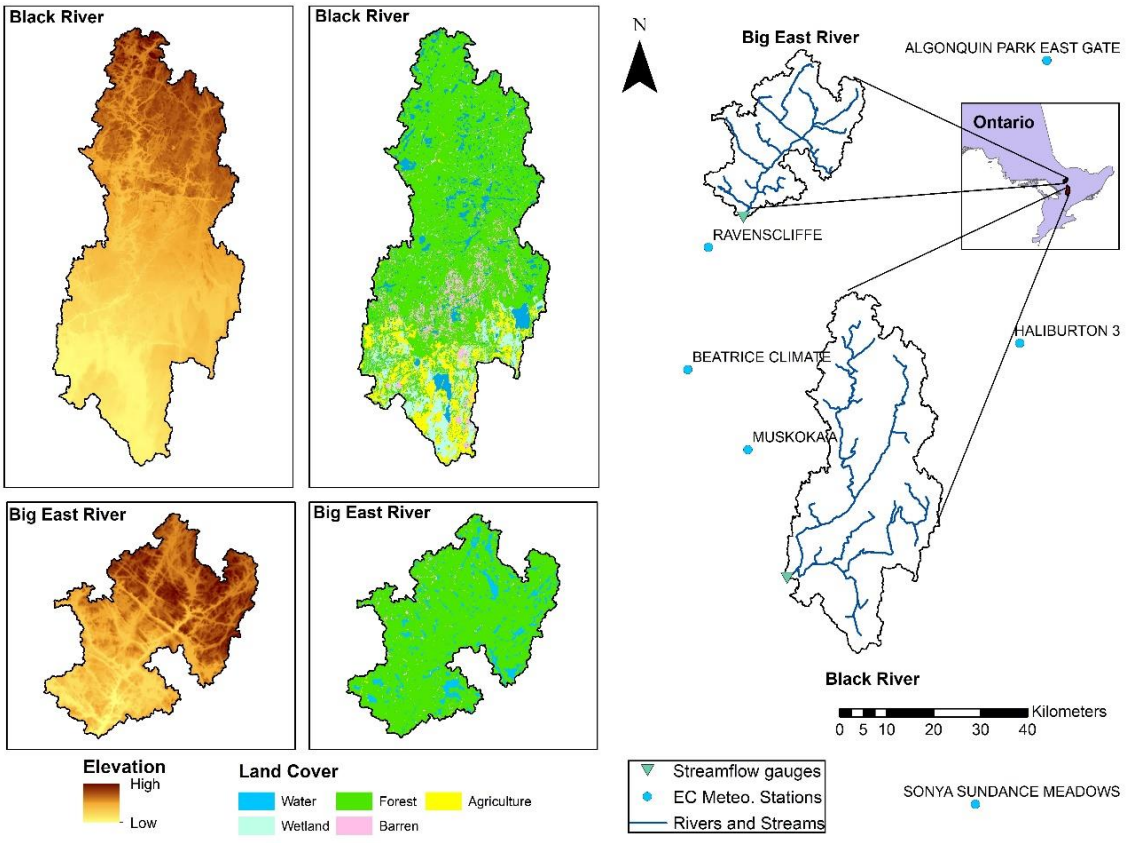


Figure 4-5 The study areas: Big East River and Black River watersheds

Table 4-1 The climate characteristics of both basins using all available meteorological and hydrometric data

Month	Precipitation (mm)			Mean daily temperature (°C)			Discharge (mm)	
	Mean	Max	Min	Mean	Max	Min	BE outlet	BL outlet
January	81	119	50	-10.2	-9.5	-11.4	55	62
February	59	75	40	-10.5	-10.0	-11.6	33	35
March	57	63	49	-3.6	-3.2	-4.5	59	64
April	87	96	70	4.5	5.1	3.8	152	125
May	81	87	69	11.9	12.3	11.3	64	51
June	108	122	98	16.2	16.7	15.6	39	22
July	85	97	73	18.5	19.3	17.9	23	15
August	88	105	74	17.5	18.2	16.7	19	8
September	94	109	82	13.4	14.0	12.4	21	9
October	117	151	86	7.2	7.7	6.2	41	24
November	92	125	66	0.8	1.2	-0.1	67	52
December	94	122	72	-5.5	-4.9	-7.0	71	65

4.4.2 Ensemble streamflow forecasts

In this study, as presented in Table 4-2, seven different lumped conceptual rainfall-runoff models, employed for generating an ensemble of streamflow forecasts, are SACSMA, MACHBV, SMARG, GR4J and three different configurations of HEC-HMS model. These models possess unique structural complexities with varying numbers of parameters. Daily precipitation and temperature are the only inputs to the chosen models, with different methods used for calculating potential evapotranspiration (PET) depending on the model (Table 4-2). Moreover, as stated in the previous section, snowmelt is one of the most important hydrologic processes in our study areas. Therefore, for adding more diversity, three different snowmelt modules are implemented with various models (Table 4-2). In the simple degree-day method (DDM; Samuel et al., 2011) which uses five parameters, a linear

relationship between temperature and the amount of snowmelt is considered. The 10-parameter snowmelt routine of the HEC-HMS models (Scharffenberg, 2016) uses the antecedent temperature index for calculating the melt rate. Snow17 (Anderson, 1973, 2006), including 11 parameters, is a more complex temperature index approach where some of the snowmelt related physical processes are considered.

Table 4-2 Rainfall-runoff models implemented in this study

Model	Reference	Number of parameters	PET method	Snowmelt routine
SACSMA	Burnash et al. (1973)	14	Thornwaite (Samuel et al., 2011)	Snow17
MACHBV	Samuel et al. (2011)	10	Thornwaite	Snow17
SMARG	Tan and O'Connor (1996)	9	Thornwaite	DDM
GR4J	Perrin et al. (2003)	4	Thornwaite	DDM
HECHMS1*	Scharffenberg (2016)	7	Hargreaves (Hargreaves and Samani, 1985)	HECHMS
HECHMS2*	Scharffenberg (2016)	15	Hargreaves	HECHMS
HECHMS3*	Scharffenberg (2016)	17	Hargreaves	HECHMS

* HECHMS1: recession + deficit and constant approaches; HECHMS2: recession + soil moisture accounting approaches; HECHMS3: linear reservoir + soil moisture accounting approaches

By considering 6 years of historical data (i.e., years 2006-2011) as the calibration period, we use the dynamically dimensioned search algorithm (Tolson & Shoemaker, 2007) for estimating models' parameters using five different objective functions. Kling-Gupta efficiency (Gupta et al., 2009), Nash-Sutcliffe efficiency (Nash & Sutcliffe, 1970), and Nash volume error (Samuel et al., 2011) focus on medium flows in different ways, while modified Nash volume error (MNVE; Darbandsari and Coulibaly, 2019) and peak weighted root mean square error (PWRMSE; Cunderlik and Simonovic, 2004) are aimed to simulate high flows more accurate. As shown in Equation 4-8, in $MNVE$, the combination of volume

error (VE) and NSE based on squared transformed streamflow (NSE_{sqrt}, reflecting the accuracy of high flows) is used:

$$MNVE = NSE_{sqrt} - 0.1 \times VE \quad (4\text{-}8)$$

Also, $PWRMSE$ is formulated as follows:

$$PWRMSE = \left(\frac{1}{N} \left(\sum_{i=1}^{N} (f_i - O_i)^2 \times \frac{O_i + \bar{O}}{2\bar{O}} \right) \right)^{\frac{1}{2}} \quad (4\text{-}9)$$

where N is the data length and f_i, O_i, and $\bar{O}$ respectively are the simulated, observed, and the mean of observed flows. As can be seen, $PWRMSE$ gives higher weights to errors near the peak flows. Utilizing multiple objective functions can provide better BMA predictions in different flow ranges (Dong et al., 2013). Combining the seven hydrologic models and five objective functions leads to a set of 35 calibrated models, this set can then be used to generate an ensemble of streamflow forecasts.

4.4.3 Performance measures

The accuracy, reliability and sharpness are the most important aspects of any predictive forecast need to be evaluated. In this study, we used six different evaluation metrics and some visual graphical tools to cover all of the aforementioned properties. In terms of the accuracy, three deterministic-based measures, including Nash Sutcliffe Efficiency (NSE), Volume Error (VE), and the root mean square error ($RMSE$) are employed:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(f_i - O_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2} \quad (4\text{-}10)$$

$$VE = \frac{\left|\sum_{i=1}^{N}(f_i - O_i)\right|}{\sum_{i=1}^{N} O_i} \quad (4\text{-}11)$$

$$RMSE = \left[\frac{1}{N}\left(\sum_{i=1}^{N}(f_i - O_i)^2\right)\right]^{1/2} \quad (4\text{-}12)$$

where f_i and O_i, are respectively the forecast and observed variable. $\bar{O}$ is the observation mean and N is the dataset length. NSE varies between -∞ and 1 with the best value of 1, while VE and $RMSE$ possess a range of [0, ∞] when lower values show better model performance.

Moreover, the mean continuous ranked probability score (CRPS; Hersbach, 2000) evaluates the accuracy of the results in a probabilistic way by comparing the cumulative distribution of forecasts ($P_i^f(x)$) and observations ($P_i^O(x)$), determined by using Heaviside function ($H(x - O_i)$):

$$CRPS = \frac{1}{N}\sum_{i=1}^{N}\int_{x=-\infty}^{x=+\infty}\left(P_i^f(x) - P_i^O(x)\right)^2 d_x \quad (4\text{-}13)$$

$$P_i^O(x) = H(x - O_i) = \begin{cases} 0 & x < O_i \\ 1 & x \geq O_i \end{cases} \quad (4\text{-}14)$$

The range for $CRPS$ is 0 to ∞ where smaller values indicate better performance.

The Containing ratio ($CR95$) and the average Bandwidth ($B95$) of the 95% confidence interval (Xiong et al., 2009) are the two other performance statistics, being used to assess

the reliability and sharpness of the probabilistic forecasts, respectively. $CR95$ is the ratio of the observations, covered by the 95% prediction bound. It ranges between 0 and 1 with an optimal value of 0.95. $B95$ determines the average width of the aforementioned interval and it is negatively oriented, with lower values indicating better forecasts.

$$CR95 = \frac{N_{O_{in}}}{N} \tag{4-15}$$

$$B95 = \frac{1}{N}\sum_{t=i}^{N}\left(f_u(i) - f_l(i)\right) \tag{4-16}$$

In the above equations, $N_{O_{in}}$ is the number of observations that have fallen within the 95% bound and the upper and lower boundaries of the corresponding bound is denoted by $f_u(i)$ and $f_l(i)$, respectively. Simultaneously evaluating these two criteria is vital for providing precise conclusions. For instance, a forecast with a good $CR95$ may still be underconfident by providing high $B95$, indicating an overestimation of the uncertainty bound.

Apart from previously presented verification metrics, we adopted the predictive quantile-quantile plot (Q-Q plot; Laio and Tamea, 2007) as additional visual statistical verification of the forecast reliability, where the comparison is made between forecast and the cumulative uniform distributions. The more the Q-Q plot follows the bisector line, the more reliable the forecast is. Therefore, another reliability measure (α) can also be calculated from a Q-Q plot, which represents the discrepancy between Q-Q plot and the bisector line (Renard et al., 2010):

$$\alpha = 1 - 2 \times \left[\frac{1}{N} \sum_{i=1}^{N} \left| P_i^f(O_i) - U(O_t) \right| \right] \quad (4\text{-}17)$$

$P_i^f(O_i)$ and $U(O_t)$ respectively determine the non-exceedance probability of the observed value using forecast and uniform cumulative distributions. $\alpha = 1$ shows the perfect reliability while its worst value is zero. Moreover, the representative hydrographs are another tool for assessing both deterministic and probabilistic forecasts visually. Furthermore, for more specific evaluation and comparison of different methods' performance regarding high flows, all aforementioned performance measures are also determined based on the high flow data (90^{th} percentile of streamflow).

4.5 Results and Discussion

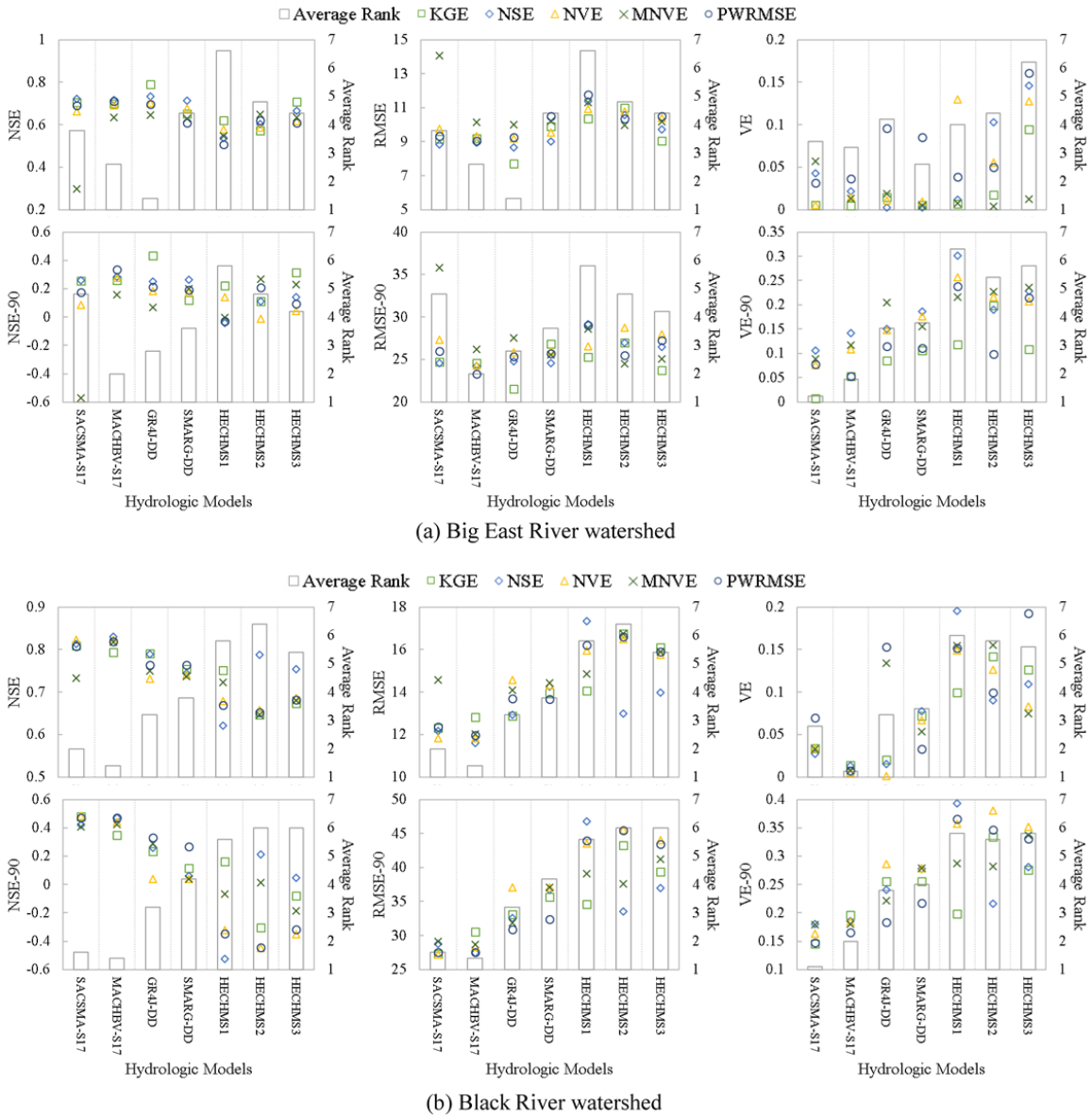
4.5.1 Rainfall-Runoff models calibration

As previously stated, for each hydrologic model, five optimized parameter sets are obtained by considering different criteria (i.e. KGE, NSE, NVE, $MNVE$, and $PWRMSE$) as an objective function in the optimization process using the calibration period from 2006 to 2011. A comparison between the performances of different models, calibrated using different objective functions, in the three-years validation period (2012-2015) in terms of NSE, $RMSE$, VE, and their corresponding values derived from flows more than 90 percentile (i.e. NSE-90, $RMSE$-90, and VE-90) can be found in Figure 4-6. In the Big East River watershed, the results, in general, indicate the superiority of GR4J model in simulating daily streamflows, however, by focusing on high flows, MACHBV shows the most reliable performance. In the Black River, on the other hand, MACHBV is the most

consistent model by providing the best results based on almost all performance metrics while SACSMA performs competitively regarding high flow simulation. The fact that MACHBV was initially developed for streamflow simulation in Ontarian's ungauged basins (Samuel et al., 2011, 2012), can justify the robust performance of MACHBV in both watersheds, which are considered as data-scarce regions with low-density ground based measurements (Darbandsari & Coulibaly, 2019).

Morever, comparing the use of different objective functions shows that implementing KGE as an objective function consistently lead to relatively better performance for most of the hydrologic models in the Big East River watershed, while its application in the Black River is not among the best ones. NSE based calibrated models, compared with NVE, provide better results in terms of NSE and $RMSE$ criteria in both watersheds, however, their performances are worst regarding volume error (VE) metric for most of the models. By combining NSE and VE metrics, NVE criterion provides a balance among volume error and difference between streamflow simulations and their corresponding observations (Lindström, 1997; Samuel et al., 2011). Furthermore, the results shows that using objective functions focusing on high flows (i.e. $PWRMS$ and $MNVE$) does not always lead to better calibrated models regarding high flows. Overall, the main conclusion that stands out from comparing the different objective functions, is that it is practically impossible to select one criteria which gives the best optimal parameters sets for all hydrologic models based on different performance measurements in both watersheds. Therefore, besides helping in finding the best optimal parameter sets for each hydrologic model, using different objective

functions provides larger number of ensemble members with higher potential capability in capturing future flow possibilities.



(a) Big East River watershed

(b) Black River watershed

Figure 4-6 The performance evaluation of various calibrated hydrologic models for the validation period (years 2012-2015) in (a) Big East River and (b) Black River watersheds

4.5.2 Multi-model versus Multi-model Multi-objective ensemble scenarios

Prior to the application of the proposed En-BMA method, we evaluate the effects of using multiple objective functions for generating ensemble members to be merged by BMA with different moving window lengths. Considering two ensemble scenarios of multi-model (E-7) and multi-model with multiple parameter sets generated using different objective functions (E-35), Figure 4-7 illustrates the performance statistics of the 1-day ahead BMA forecasts as a function of the number of days of the moving window for both Big East River and Black River watersheds. The most influenced properties of the forecasts, changing based on the window length, are the reliability ($CR95$) and sharpness ($B95$) of the results. Although the shorter training period leads to smaller uncertainty bounds, it highly reduces the reliability of the forecasts in both watersheds and both ensemble scenarios. These results is qualitatively in line with previous studies (Raftery et al., 2005; Vrugt & Robinson, 2007) showing that increasing moving window length enhances the reliability of the forecasts while it reduces the sharpness. Other performance measures focusing on the accuracy of the results show negligible changes regarding the moving window length. Therefore, given that a similar trend has been seen for other forecast horizons (up to 7 days), we select a length of 100 days where the reliability of the results appears to reach stability, beyond which the sharpness of the results deteriorates without significant improvement of the containing ratio.

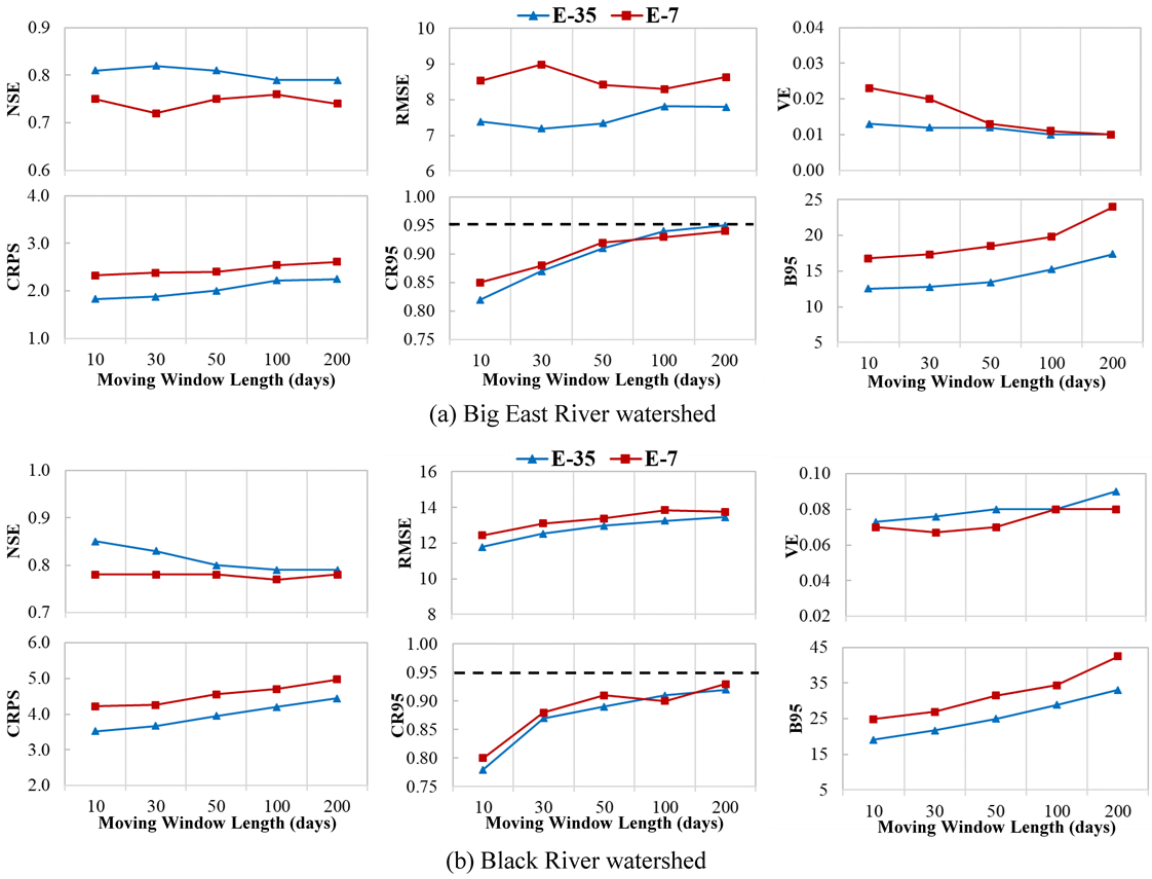


Figure 4-7 The comparison of different performance statistics of BMA 1-day ahead forecasts using two different ensemble scenarios and different mowing window length in (a) Big East River and (b) Black River watersheds. E-7 and E-35 respectively show the multi-models and multi-models multi-objectives ensemble scenarios

Moreover, although the superiority of scenario E-35 over E-7 can be seen from Figure 4-7, a comprehensive comparison of both ensemble scenarios has been made using BMA with a 100-day moving window for 1-, 3-, 5-, and 7-day ahead forecasts. It is of note that N-day ahead forecasts mean the times series of forecasts for lead-time *N* independently and not as aggregated of 1- to N-days ahead. Therefore, all criteria, being calculated based on N-day ahead time series of forecasts, only indicate the performance at that particular lead-time. The results, as shown in Figure 4-8, indicate that using multiple models with

multiple objective functions consistently enhance the probabilistic performance of the BMA method for different lead-times in both watersheds. These improvements are also apparent in the deterministic performance statistics regarding high flows, especially in the Black River watershed. Consequently, it can be concluded that constructing an ensemble using multiple models with multiple parameter sets based on different objective functions leads to better BMA results. Accordingly, for the En-BMA application, the multi-model multi-objective ensemble scenario, including 35 members of streamflow predictions, was implemented for both watersheds.

Figure 4-8 The percent improvement of different performance statistics in both Big East River and Black River watersheds. The positive value of percent improvement shows the positive effect of utilizing E-35 in comparison to E-7

4.5.3 The effects of the stopping threshold value on En-BMA

The main parameter of the proposed En-BMA approach is the stopping threshold value (β), which needs to be determined beforehand. Therefore, the sensitivity analysis was carried out for different β values (i.e., 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99) to assess how selecting this threshold can affect the streamflow forecasts. The first noticeable effect of the stopping threshold is on the number of selected members (Figure 4-9); by increasing β a higher numbers of members are selected for BMA application. Larger stopping threshold value focuses more on increasing information content of the whole ensemble rather than independence of its members, while choosing smaller β leads to lower redundant information without paying that much attention to the ensemble exhaustiveness. In addition, different performance measurements for the 1-day lead time forecasts are presented in Table 4-3. What stands out in this table is that in general, a higher threshold leads to sharper forecasts for both watersheds, while the deterministic performance slightly deteriorates especially in high flows. Therefore, in this study, the threshold value of 0.95 was chosen for both watersheds, which provides a balance between the different performance statistics. These results indicate the same value of having mutually exclusive as well as collectively exhaustive ensemble (Madadgar & Moradkhani, 2014; Refsgaard et al., 2012), and indicate the importance of selecting a proper threshold value prior to the En-BMA application.

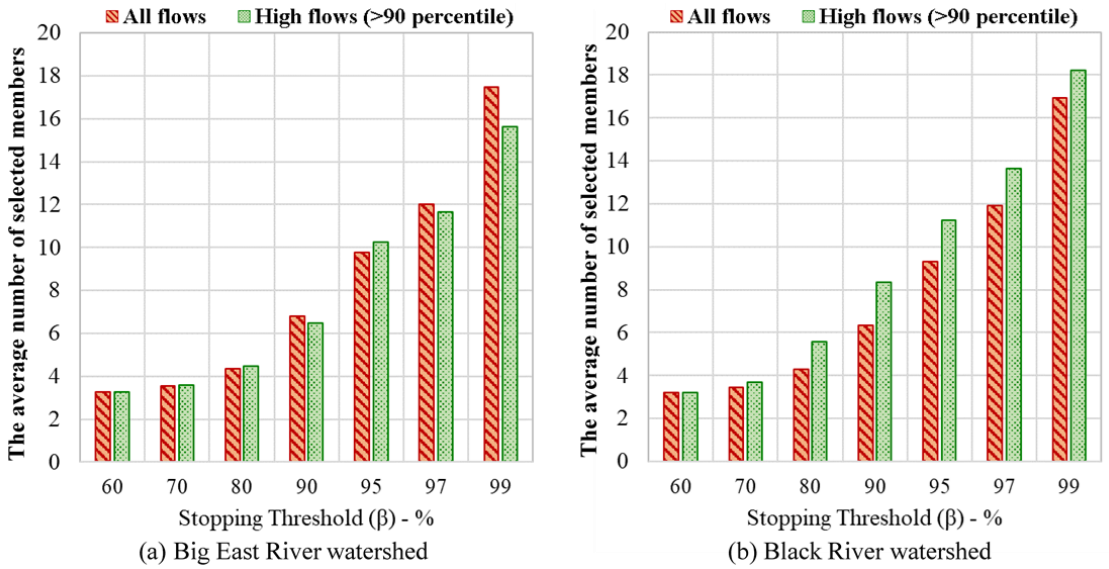
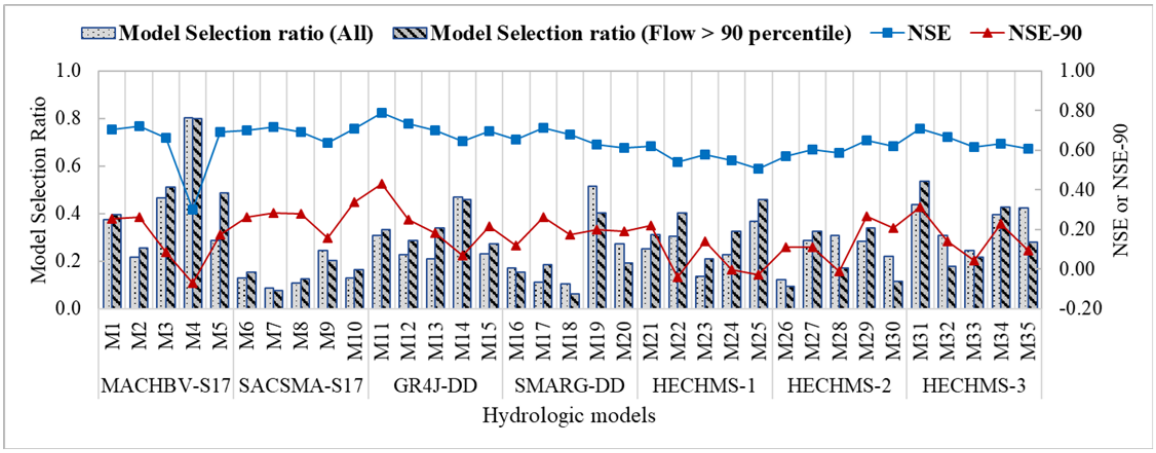


Figure 4-9 The average number of selected members and using different stopping threshold in (a) Big East River and (b) Black River watersheds

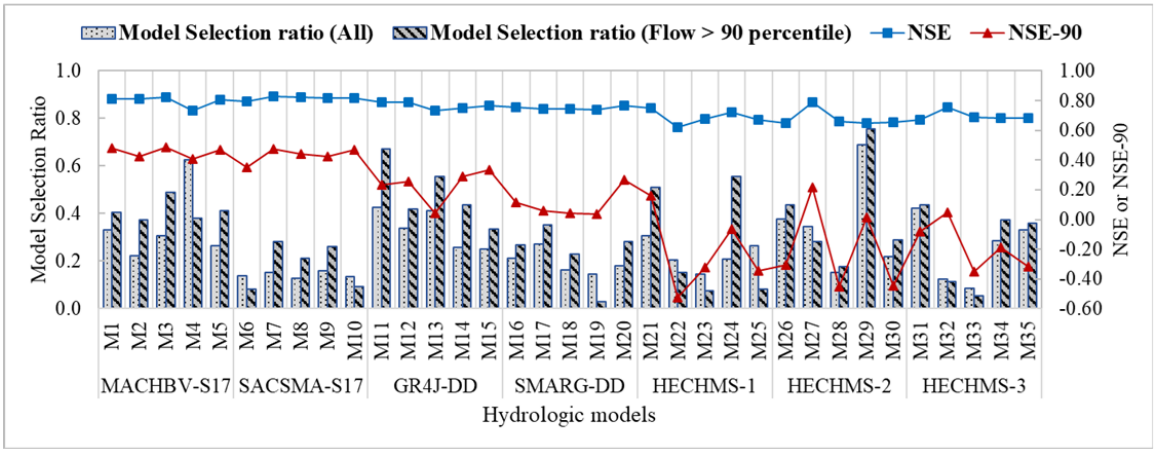
Table 4-3 Different performance statistics of 1-day ahead forecasts derived from the proposed En-BMA method with different stopping threshold values

Basin	β	All flows						Flows more than 90 percentile					
		NSE	**RMSE**	**VE**	**CRPS**	**CR95**	**B95**	**NSE**	**RMSE**	**VE**	**CRPS**	**CR95**	**B95**
Big East River	**0.6**	0.79	7.8	0.04	2.7	0.95	26	0.71	17.7	0.06	8.3	0.74	37
	0.7	0.8	7.7	0.04	2.6	0.95	25	0.71	17.6	0.06	8.3	0.75	36
	0.8	0.79	7.7	0.03	2.6	0.95	24	0.68	18.6	0.06	8.5	0.76	36
	0.9	0.8	7.6	0.03	2.5	0.94	22	0.67	18.8	0.04	8.6	0.78	35
	0.95	0.79	7.7	0.013	2.3	0.94	17	0.64	19.6	0.03	8.9	0.78	33
	0.97	0.77	8.1	0.04	2.5	0.95	17	0.58	21.2	0.01	9.5	0.77	33
	0.99	0.76	8.3	0.03	2.5	0.94	17	0.56	21.6	0.03	10.2	0.74	32
Black River	**0.6**	0.81	12.4	0.02	4.8	0.92	43	0.46	31.2	0.23	16.0	0.73	56
	0.7	0.81	12.4	0.02	4.7	0.92	42	0.47	31.0	0.22	15.9	0.74	55
	0.8	0.82	12.4	0.02	4.7	0.93	41	0.49	30.5	0.21	15.9	0.72	52
	0.9	0.82	12.4	0.03	4.8	0.92	37	0.47	31.0	0.21	15.5	0.70	49
	0.95	0.81	12.6	0.04	4.5	0.92	34	0.43	32.2	0.22	14.6	0.73	48
	0.97	0.81	12.6	0.04	4.4	0.91	33	0.43	32.3	0.22	14.7	0.72	47
	0.99	0.81	12.7	0.04	4.3	0.9	32	0.41	32.8	0.23	14.7	0.7	47

Besides assessing the effects of the stopping threshold, the contribution of each member into the forecasts is illustrated in Figure 4-10. A simple comparison between the performance of each individual model during the independent validation period (i.e., years 2012-2015) and the frequency of their selection for forecasting application shows that the selection ratios are not completely in accordance with models' performance, and even some relatively lower performing members have been frequently selected based on the proposed procedure. This is justifiable by the fact the entropy terms used in the proposed selection procedure, evaluate the information content of the whole ensemble rather than focusing on individual members. This expresses that besides high-performance models, considering some members with unique information is necessary for possessing mutually exclusive and collectively exhaustive ensemble (Darbandsari & Coulibaly, 2019).



(a) Big East River watershed



(b) Black River watershed

Figure 4-10 The contribution of each member into the forecasts and the NSE performance statistic of each member for the whole validation period based on all flows and flows more than 90 percentile in (a) Big East River and (b) Black River watersheds

4.5.4 En-BMA versus BMA

BMA and the proposed En-BMA method with a stopping threshold of 0.95 are employed to forecast the streamflow up to seven days ahead within the validation period in both Big East River and Black River watersheds. Figure 4-11 and Figure 4-12 compare the accuracy, reliability, and sharpness of the BMA and En-BMA methods using six different

performance statistics (presented in Section 4-4). It can be recognized that, in general, when all flows are considered, there is a small loss of deterministic performance in the Big East River watershed by applying the proposed entropy-based approach, while in the Black River, all deterministic measures (i.e. NSE, $RMSE$, and VE) show marginal advantage of En-BMA compared to BMA for all lead-times. However, by focusing on high flows (Figure 4-12), the superiority of the En-BMA over BMA in both watersheds is shown based on all performance statistics. These improvements exist in all forecasting horizons, but more so during shorter lead times (e.g., the 1-day ahead NSE improvement of 8% and 65% in comparison to 4% and 32% for 7-day ahead forecasts in Big East River and Black River watersheds, respectively).

In terms of probabilistic forecasts, almost the same conclusions can be derived. The general performance statistics based on all data (i.e., Figue 4-11) show that applying En-BMA may slightly deteriorate the probabilistic performance of the results; however, improvements are notable in both watersheds for the high flows (Figure 4-12), especially for shorter forecast horizons. For instance, the containing ratios regarding high flows (CR95-90) at 1-day ahead forecasts improve more than 10% in both watersheds with the same or less corresponding bandwidth. This improvement extended to the longer lead times in Black River watersheds. However, in the Big East River, although the high flow forecasts reliability of the En-BMA approach is better, sharpness was deteriorated.

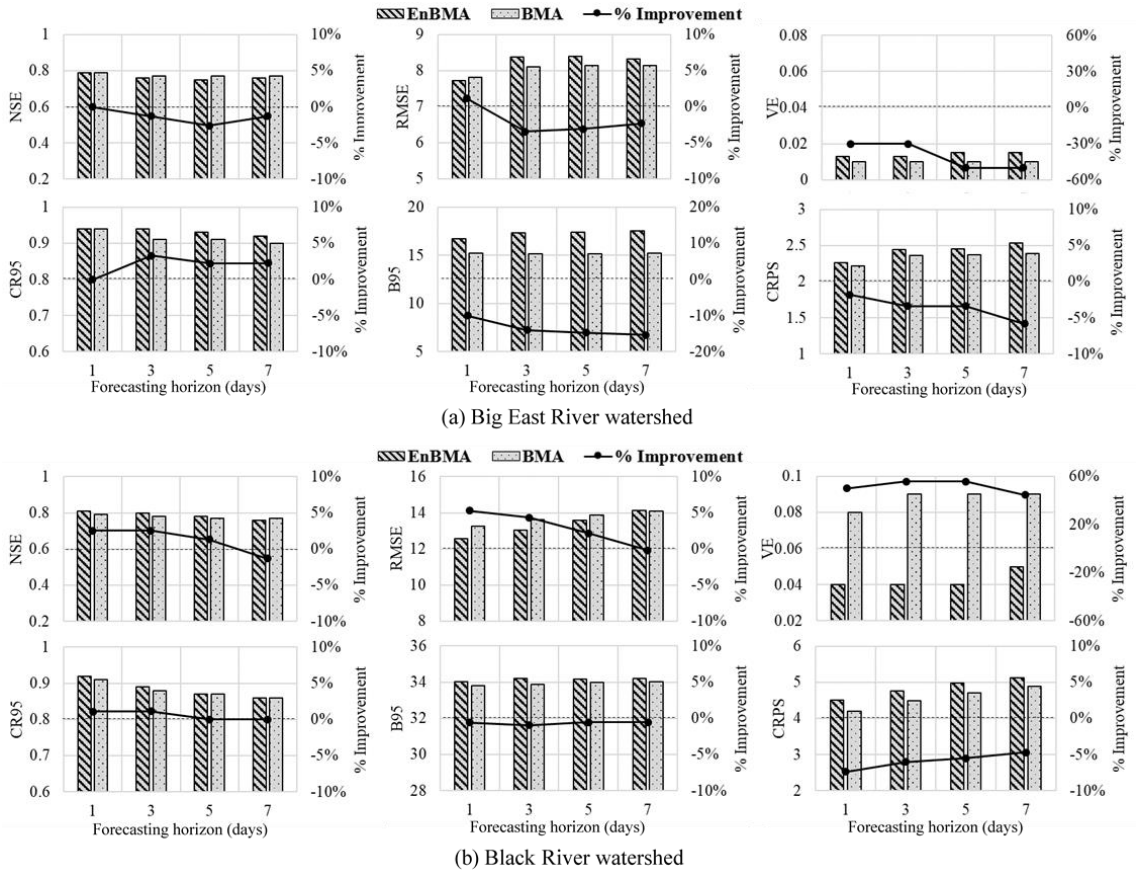


Figure 4-11 Comparison of different performance metrics for 1 to 7 days-ahead streamflow forecasting derived from BMA and En-BMA methods in (a) Big East River and (b) Black River watersheds. % improvement is defined as the percentage increase when using En-BMA instead of BMA, with positive values indicating it was advantageous to use En-BMA

As stated previously, the quantile-quantile plot and its corresponding reliability measure (α) are also used for comparing the performance of En-BMA and BMA methods. When looking at the results derived from all streamflow data (Figure 4-13), there is a negligible difference between the performances of both approaches in both watersheds. Although En-BMA leads to better results in the Big East River watershed, it slightly deteriorates in the Black River watershed. However, in term of high flow forecasting (Figure 4-14), the higher

reliability of the En-BMA results as compared with BMA is noticeable. The percent improvement shown for the α values when using the En-BMA approach for 1-day ahead forecasts was approximately 13% and 18% in the Big East River and the Black River watersheds, respectively. This superiority was shown to decrease with increasing lead-times; however, a positive percent improvement was still found for all lead-times. It is worthy of note that in both the Q-Q plot and the volume error values (Figure 4-12 and Figure 4-14, respectively), both BMA and En-BMA methods underestimate high flows in the Black River watershed. These underestimations were seen for all calibrated hydrologic models used in this study. This may be due to the limitation of the precipitation data in the study area. The mean areal forcing precipitation data may not be representative of the actual precipitation patterns. Therefore affecting both the En-BMA and BMA results. However, it does not affect the comparing process.

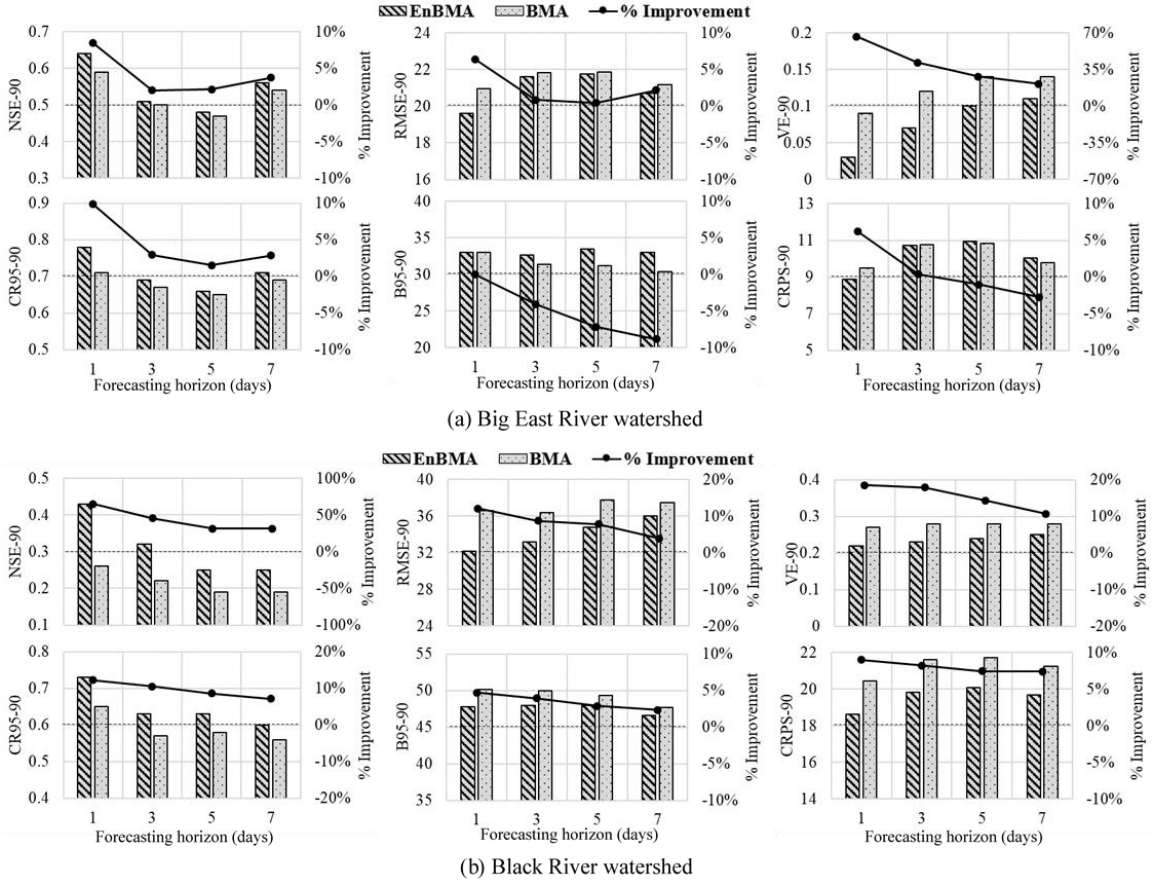


Figure 4-12 Comparison of different performance metrics for 1 to 7 days-ahead high flow forecasting derived from BMA and En-BMA methods in (a) Big East River and (b) Black River watersheds. % improvement is defined as the percentage increase when using En-BMA instead of BMA, with positive values indicating it was advantageous to use En-BMA

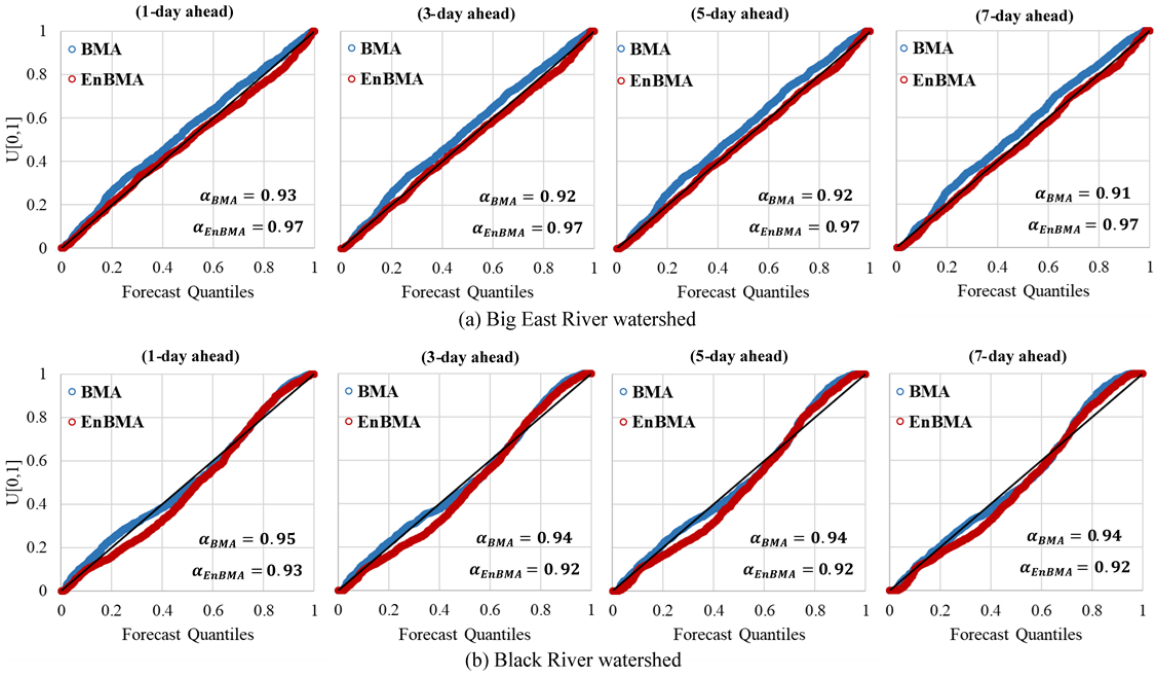


Figure 4-13 Comparison of the predictive Q-Q plot of different lead times (1-day to 7-day) derived from BMA and En-BMA results for (a) Big East River and (b) Black River watersheds

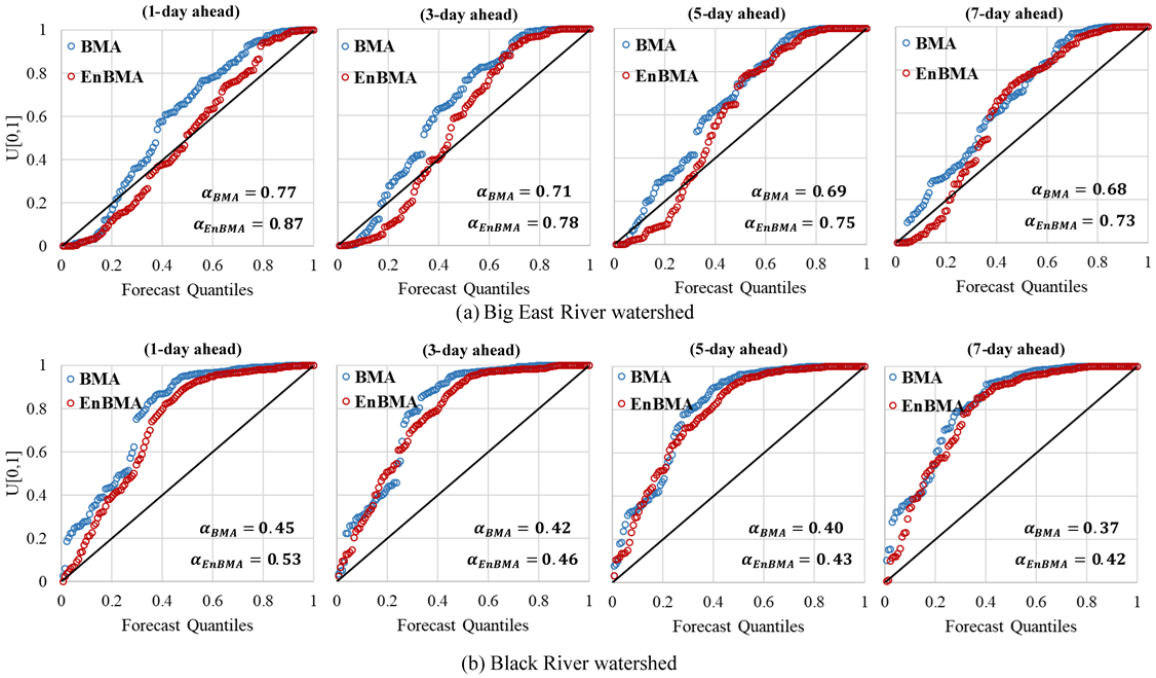


Figure 4-14 Comparison of the predictive Q-Q plot of different lead times (1-day to 7-day) derived from BMA and En-BMA high flow results for (a) Big East River and (b) Black River watersheds

Finally, in order to complete the comparison, Figure 4-15 illustrates a representative portion of hydrographs including observed, BMA, and En-BMA derived mean and 95% prediction uncertainty for different forecasting horizons (i.e., 1, 3, 5, and 7 days). In line with previous conclusions, the plots of different lead times show that in both watersheds, En-BMA outperforms BMA in terms of both probabilistic and deterministic performance regarding high flow predictions. These outperforming results are more noticeable for shorter lead times (1-day and 3-day ahead forecasts). It is worth mentioning that, in general, both En-BMA and BMA results have almost the same accuracy and reliability regarding low flows in both watersheds. This is due to the fact that the temporal variability of streamflow forecasts based on different ensemble members is marginal in low flows and

narrowing down the ensemble member using En-BMA method does not significantly affects the results. On the other hand, however, it can be seen that implementing En-BMA leads to less sharp low flow forecasts in comparison to the original BMA, especially in base flows after a rainfall event (Figure 4-15). This is justifiable by the fact that the moving windows sorounding these days, used in the selection procedure, include high flow events which leads to the selected ensemble with larger variability.

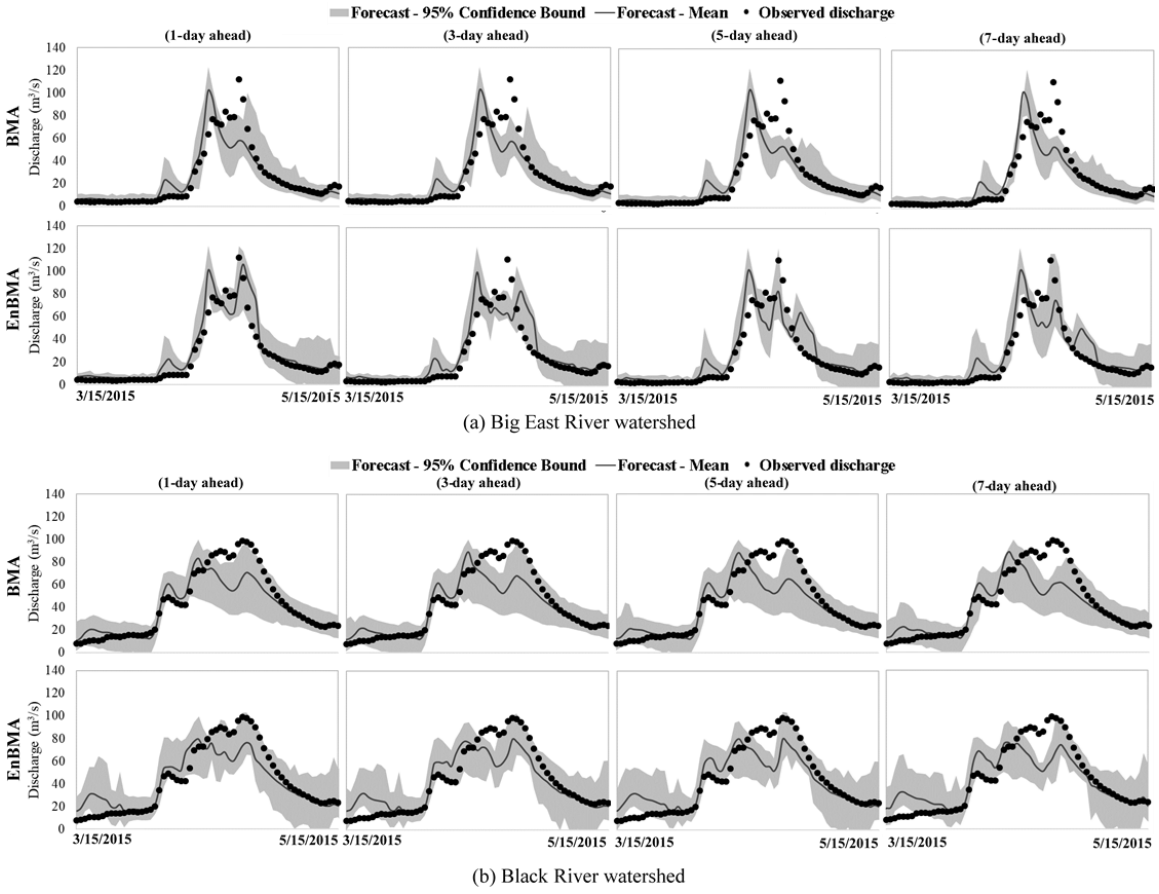


Figure 4-15 Time-series of the mean and 95% predictive bounds derived from En-BMA and BMA forecasts of various lead times compared with observations from a representative period in (a) Big East River and (b) Black River watersheds

4.6 Summary and Conclusion

The multi-model ensemble prediction system is a well-known approach to quantify and reduce model structural uncertainty. Among various post-processing methods, Bayesian Model Averaging (BMA) is one of the most reliable statistical tools for generating predictive forecasts by relatively merging individual ones. In BMA, the law of total probability is used for estimating the predictive distribution of the forecast variable as a weighted average of the PDF of individual forecasts. Therefore, having mutually exclusive and collectively exhaustive members of the ensemble is a fundamental need in order to reach more reliable results. However, these two requirements are in conflict with each other, so providing a balance between them seems necessary for possessing better BMA based predictive forecasts. Given the mentioned challenge, in order to narrow down the streamflow forecasts for meeting the two contrasting criteria, this study developed a novel entropy-based selection method to be employed prior to the BMA. Since information theory measures have shown success in different hydrometric network design applications, where the same competing objectives are considered, we utilized three entropy terms (i.e., total correlation, joint entropy and transinformation) for generating an independent and exhaustive ensemble. In the proposed structure, minimizing total correlation assures the minimum redundancy between selected members while joint entropy of members and transinformation between members and observation lead to an ensemble with higher information.

We compared the application of the BMA and the proposed En-BMA methods for generating probabilistic streamflow forecasts at short- to medium range lead times (1 to 7-

day ahead forecasts) in two data-poor watersheds, located in Ontario, Canada. Seven conceptual lumped hydrologic models with different structures and five different objective functions were used to create an ensemble of 35 streamflow forecasts for each watershed. We used six different evaluation metrics, the Q-Q plot, and representative hydrographs for comprehensively comparing BMA and En-BMA results regarding the whole time series and the high flows, separately. The summary of the most important obtained results in both watersheds is as follows.

- The simulation results, comparing the calibrated models using different objective functions, as well as the comparison of BMA f with two different ensemble scenarios, indicate that using multiplie objective functions for calibrating various hydrologic models leads to an ensemble of members with higher diversity, and can enhance the BMA performance. This conclusion is in line with previous studies suggesting the use of diverse ensemble members in conjunction with BMA (Dong et al., 2013; Parrish et al., 2012; Sharma et al., 2019).
- Evaluating the proposed entropy based selection procedure illustrates the same importance of having collectively exhaustive ensemble as mutually exclusive members. Besides independency, the number of ensemble members should be large enough in order to have enough information about all future possibilities, otherwise, the BMA application may be unreliable and leads to overestimation of predicitive uncertainties (Madadgar & Moradkhani, 2014; Refsgaard et al., 2012).
- Comparing the application of BMA and the proposed En-BMA methods in both watersheds shows no significant difference between BMA and En-BMA methods

when the whole forecast time series is considered. However, in term of high flow forecasts, En-BMA provided better deterministic and probabilistic results. Based on NSE and $RMSE$ scores for high flows, the accuracy of the forecasts enhanced significantly after implementing En-BMA in both study watersheds. Also, the Q-Q plots and the containing ratio measurements indicate higher reliability of the En-BMA derived probabilistic forecasts for high flows, as compared to BMA, without losing its sharpness; this is more apparent at the shorter lead times.

In general, besides confirming the merits of using multiple models with multiple objectives over only considering multi-models for generating an ensemble of streamflow forecasts, the results suggest that the proposed En-BMA method outperforms the traditional BMA in both deterministic and probabilistic ways, through constructing a mutually exclusive and collectively exhaustive ensemble of streamflow forecasts, especially for high flows which are of particular interest in operational hydrology. The findings of this study call for further studies on employing other entropy measures for generating proper streamflow ensemble for BMA applications. In addition, as the proposed entropy based selection procedure is not restricted to specific types of variables, apart from streamflow forecasts, further studies could employ the proposed method for other variables, such as precipitation and temperature, with different time intervals (e.g. hourly).

Although, both basins, used in this study, have similar climatologic conditions, their hydrologic responses are quite different and yield two distinct probability distributions of streamflow data, which is the most effective characteristics in the calculation of entropy terms and the BMA application. This difference suggests that the findings of this research

can be easily generalized to other future studies. Such future applications of the proposed En-BMA method should cover diverse watersheds with different climatology, land cover, and topography. Furthermore, it is noteworthy that the BMA approach and therefore the En-BMA, estimates predictive distribution by using the information derived from an ensemble of multi-model streamflow forecasts, while, there are some other valuable information, such as the known initial and boundary conditions, that can be used for reducing hydrologic uncertainty. This study evaluates the positive direct effects of possessing mutually exclusive and collectively exhaustive ensemble on BMA results, however, for operational purposes (such as flood forecasting), explicitly deciphering the initial condition uncertainty by implementing an updating procedure (e.g. data assimilation methods) in conjunction with the proposed En-BMA approach is recommended.

4.7 Acknowledgments

unding: This work was supported by the Natural Science and Engineering Research Council (NSERC) of Canada, grant NSERC Canadian FloodNet (NETGP-451456). The authors thank the Ministry of Natural Resources and Forestry, Surface Water Monitoring Center for providing some data of the study regions. The authors acknowledge Dr. James Leach for his help with reviewing and editing of the manuscript.

4.8 References

Alfonso, L., He, L., Lobbrecht, A., & Price, R. (2013). Information theory applied to evaluate the discharge monitoring network of the Magdalena River. *Journal of Hydroinformatics*, *15*(1), 211–228. https://doi.org/10.2166/hydro.2012.066

Alfonso, L., Lobbrecht, A., & Price, R. (2010). Optimization of water level monitoring network in polder systems using information theory. *Water Resources Research*, *46*(12). https://doi.org/10.1029/2009WR008953

Anderson, E. A. (1973). *National Weather Service river forecast system: Snow accumulation and ablation model*. U.S. DEPARTMENT OF COMMERCE: National Oceanic and Atmospheric Administration, National Weather Service.

Anderson, E. A. (2006). *Snow Accumulation and Ablation Model – SNOW-17*. Natl. Ocean. Atmospheric Adm. Natl. Weather Serv. Silver Springs MD. https://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/part2/_pdf/22snow17.pdf

Anshuman, A., Kunnath-Poovakka, A., & Eldho, T. I. (2019). Towards the use of conceptual models for water resource assessment in Indian tropical watersheds under monsoon-driven climatic conditions. *Environmental Earth Sciences*, *78*(9), 282. https://doi.org/10.1007/s12665-019-8281-5

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., & Martel, J.-L. (2015). A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. *Journal of Hydrology*, *529*, 754–767. https://doi.org/10.1016/j.jhydrol.2015.09.001

Boucher, M.-A., Anctil, F., Perreault, L., & Tremblay, D. (2011). A comparison between ensemble and deterministic hydrological forecasts in an operational context. *Advances in Geosciences*, *29*, 85–94. https://doi.org/10.5194/adgeo-29-85-2011

Cunderlik, J., & Simonovic, S. (2004). *Calibration, Verification and Sensitivity Analysis of the HEC-HMS Hydrologic Model*. Department of Civil and Environmental Engineering, The University of Western Ontario. https://ir.lib.uwo.ca/wrrr/11

Darbandsari, P., & Coulibaly, P. (2019). Inter-Comparison of Different Bayesian Model Averaging Modifications in Streamflow Simulation. *Water*, *11*(8), 1707. https://doi.org/10.3390/w11081707

DelSole, T. (2007). A Bayesian Framework for Multimodel Regression. *Journal of Climate*, *20*(12), 2810–2826. https://doi.org/10.1175/JCLI4179.1

Dong, L., Xiong, L., & Zheng, Y. (2013). Uncertainty analysis of coupling multiple hydrologic models and multiple objective functions in Han River, China. *Water Science and Technology*, *68*(3), 506–513. https://doi.org/10.2166/wst.2013.255

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Granger, C. W., & Ramanathan, R. (1984). Improved Methods of Combining Forecasts: ABSTRACT. *Journal of Forecasting (Pre-1986); Chichester*, *3*(2), 197–204.

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

He, S., Guo, S., Liu, Z., Yin, J., Chen, K., & Wu, X. (2018). Uncertainty analysis of hydrological multi-model ensembles based on CBP-BMA method. *Hydrology Research*, *49*(5), 1636–1651. https://doi.org/10.2166/nh.2018.160

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science, 14*(4), 382–401. JSTOR.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Keum, J., Awol, F. S., Ursulak, J., & Coulibaly, P. (2019). Introducing the Ensemble-Based Dual Entropy and Multiobjective Optimization for Hydrometric Network Design Problems: EnDEMO. *Entropy*, *21*(10), 947. https://doi.org/10.3390/e21100947

Keum, J., & Coulibaly, P. (2017a). Information theory-based decision support system for integrated design of multivariable hydrometric networks. *Water Resources Research*, *53*(7), 6239–6259. https://doi.org/10.1002/2016WR019981

Keum, J., & Coulibaly, P. (2017b). Sensitivity of Entropy Method to Time Series Length in Hydrometric Network Design. *Journal of Hydrologic Engineering*, *22*(7), 04017009. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001508

Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences, 11*, 1267–1277.

Leach, J. M., Kornelsen, K. C., Samuel, J., & Coulibaly, P. (2015). Hydrometric network design using streamflow signatures and indicators of hydrologic alteration. *Journal of Hydrology*, *529*, 1350–1359. https://doi.org/10.1016/j.jhydrol.2015.08.048

Li, C., Singh, V. P., & Mishra, A. K. (2012). Entropy theory-based criterion for hydrometric network evaluation and design: Maximum information minimum redundancy. *Water Resources Research*, *48*(5). https://doi.org/10.1029/2011WR011251

Liang, Z., Wang, D., Guo, Y., Zhang, Y., & Dai, R. (2013). Application of Bayesian Model Averaging Approach to Multimodel Ensemble Hydrologic Forecasting. *Journal of*

Hydrologic Engineering, *18*(11), 1426–1436. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000493

Lindström, G. (1997). A Simple Automatic Calibration Routine for the HBV Model. *Hydrology Research*, *28*(3), 153–168. https://doi.org/10.2166/nh.1997.0009

Madadgar, S., & Moradkhani, H. (2014). Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resources Research*, *50*(12), 9586–9603. https://doi.org/10.1002/2014WR015965

Michaels, S. (2015). Probabilistic forecasting and the reshaping of flood risk management. *Journal of Natural Resources Policy Research*, *7*(1), 41–51. https://doi.org/10.1080/19390459.2014.970800

Mishra, A. K., & Coulibaly, P. (2009). Developments in hydrometric network design: A review. *Reviews of Geophysics*, *47*(2). https://doi.org/10.1029/2007RG000243

Moradkhani, H., & Sorooshian, S. (2008). General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis. In *Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrological Models* (pp. 1–24). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77843-1_1

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Parrish, M. A., Moradkhani, H., & DeChant, C. M. (2012). Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation: TOWARD REDUCTION OF MODEL UNCERTAINTY. *Water Resources Research*, *48*(3). https://doi.org/10.1029/2011WR011116

Qu, B., Zhang, X., Pappenberger, F., Zhang, T., & Fang, Y. (2017). Multi-Model Grand Ensemble Hydrologic Forecasting in the Fu River Basin Using Bayesian Model Averaging. *Water*, *9*(2), 74. https://doi.org/10.3390/w9020074

Raftery, A. E. (1993). Bayesian Model Selection in Structural Equation Models. In *Testing Structural Equation Models* (Vol. 154, pp. 163–180). SAGE.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, *92*(437), 179–191. https://doi.org/10.1080/01621459.1997.10473615

Refsgaard, J. C., Christensen, S., Sonnenborg, T. O., Seifert, D., Højberg, A. L., & Troldborg, L. (2012). Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Advances in Water Resources*, *36*, 36–50. https://doi.org/10.1016/j.advwatres.2011.04.006

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2009WR008328

Rings, J., Vrugt, J. A., Schoups, G., Huisman, J. A., & Vereecken, H. (2012). Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments. *Water Resources Research*, *48*(5). https://doi.org/10.1029/2011WR011607

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2011). Estimation of Continuous Streamflow in Ontario Ungauged Basins: Comparison of Regionalization Methods. *Journal of*

Hydrologic Engineering, *16*(5), 447–459. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000338

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2012). Identification of rainfall–runoff model for improved baseflow estimation in ungauged basins. *Hydrological Processes*, *26*(3), 356–366. https://doi.org/10.1002/hyp.8133

Scharffenberg, W. (2016). *HEC-HMS User's Manual, Version 4.2*. U.S. Army Corps of Engineers Institute for Water Resources Hydrologic Engineering Center (CEIWR-HEC).

Seo, D.-J., Herr, H. D., & Schaake, J. C. (2006). A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrology and Earth System Sciences Discussions*, *3*(4), 1987–2035. https://doi.org/10.5194/hessd-3-1987-2006

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sharma, S., Siddique, R., Reed, S., Ahnert, P., & Mejia, A. (2019). Hydrological Model Diversity Enhances Streamflow Forecast Skill at Short- to Medium-Range Timescales. *Water Resources Research*, *55*(2), 1510–1530. https://doi.org/10.1029/2018WR023197

Shrestha, D. L. (2009). *Uncertainty analysis in rainfall-runoff modelling - application of machine learning techniques: UNESCO-IHE PhD thesis*. [PhD. thesis, IHE Delft Institute for Water Education]. https://www.cabdirect.org/cabdirect/abstract/20123116250

Singh, V. P. (1997). The Use of Entropy in Hydrology and Water Resources. *Hydrological Processes*, *11*(6), 587–626. https://doi.org/10.1002/(SICI)1099-1085(199705)11:6<587::AID-HYP479>3.0.CO;2-P

Tegegne, G., Park, D. K., & Kim, Y.-O. (2017). Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. *Journal of Hydrology: Regional Studies*, *14*, 49–66. https://doi.org/10.1016/j.ejrh.2017.10.002

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Viallefont, V., Raftery, A. E., & Richardson, S. (2001). Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine*, *20*(21), 3215–3230. https://doi.org/10.1002/sim.976

Vrugt, J. A., Diks, C. G. H., & Clark, M. P. (2008). Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling. *Environmental Fluid Mechanics*, *8*(5), 579–595. https://doi.org/10.1007/s10652-008-9106-3

Vrugt, J. A., & Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004838

Xiong, L., Wan, M., Wei, X., & O'Connor, K. M. (2009). Indices for assessing the prediction bounds of hydrological models and application by generalised likelihood uncertainty estimation / Indices pour évaluer les bornes de prévision de modèles hydrologiques et mise en œuvre pour une estimation d'incertitude par vraisemblance généralisée. *Hydrological Sciences Journal*, *54*(5), 852–871. https://doi.org/10.1623/hysj.54.5.852

Xu, J., Anctil, F., & Boucher, M.-A. (2019). Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems. *Journal of Hydrology*, *578*, 124002. https://doi.org/10.1016/j.jhydrol.2019.124002

Chapter 5. HUP-BMA: An Integration of Hydrologic Uncertainty Processor and Bayesian Model Averaging for Streamflow Forecasting

Summary of Paper 4: Darbandsari, P., & Coulibaly, P. (2021). HUP-BMA: An Integration of Hydrologic Uncertainty Processor and Bayesian Model Averaging for Streamflow Forecasting. Water Resources Research, under review.

In this research work, after evaluating the effects of implementing different deterministic forecasts within the Hydrologic Uncertainty Processor (HUP) method, a new ensemble-based Bayesian post-processing (HUP-BMA) approach is proposed where the Bayesian Model Averaging concept is used to enhance the uncertainty quantification by combining the predictive distributions derived from HUP with different hydrologic models.

Key findings of this study include:

- For short lead times and low flow values, the HUP method can compensate the low quality of the used deterministic forecasts by generating more accurate and reliable probabilistic results.
- The HUP performance is noticeably affected by the performance of the deterministic forecasts in longer lead times and higher flow magnitudes.
- The proposed HUP-BMA method, compared to HUP, takes the advantage of multiple deterministic forecasts for better quantifying hydrologic uncertainty and generating more accurate and reliable probabilistic results.

- The modified HUP-BMA unconditioned on initial flow values leads to better probabilistic forecasts for longer lead times when the dependence between the actual and initial flow values is low.

5.1 Abstract

Uncertainty quantification and providing probabilistic streamflow forecasts are of particular interest for water resource management. The Hydrologic Uncertainty Processor (HUP) is a well-known Bayesian approach used to quantify hydrologic uncertainty based on observations and deterministic forecasts. This uncertainty quantification is model-specific; however, utilizing information from multiple hydrologic models should be advantageous and should lead to better probabilistic forecasts. Using seven, structurally different, conceptual models, this study firstly aims at evaluating the effects of implementing different hydrologic models on HUP performance. Secondly, using the concepts of the Bayesian Model Averaging (BMA) approach, a multi-model HUP-based Bayesian post-processor (HUP-BMA) is proposed where the combination of posterior distributions derived from HUP with different hydrologic models are used to better quantify the hydrologic uncertainty. All post-processing approaches are applied for medium-range daily streamflow forecasting (1 to 14 days ahead) in two watersheds located in Ontario, Canada. The results indicate that that the HUP forecasts for short lead-times are negligibly affected by implementing different hydrologic models, while with increasing lead-time and flow magnitude, they significantly depend on the quality of the deterministic forecast. Moreover, the superiority of the proposed HUP-BMA method over HUP is demonstrated based on various verification metrics in both watersheds. Additionally, HUP-

BMA outperformed the original BMA in quantifying hydrologic uncertainty for short lead-times. However, by increasing lead-time, considering the effects of initial observed flow on HUP-BMA formulation may be not beneficial. So, its modified version unconditioned on initial observations is preferred.

Keywords: Uncertainty, Streamflow forecasting, Bayesian Model Averaging, Hydrologic Uncertainty Processor

5.2 Introduction

Probabilistic streamflow forecasting is of increasing interest in various fields of water resources management from real-time flood forecasting to long-term management of water systems. Accurate and reliable short- to medium-range streamflow forecasts, with lead-times ranging from hours to days, can play an important role in flood control, mitigation, and early warning systems (Bravo et al., 2009; Thiemig et al., 2015). Unlike deterministic forecasts, which provide a point estimation of the river flow, probabilistic forecasts try to quantitatively assess the inherent uncertainties associated with the streamflow predictions and provide a predictive uncertainty distribution, which is required for reliable and informed decision making (Biondi & Todini, 2018; Liu et al., 2018; Reggiani & Weerts, 2008; Todini, 2008). Predictive uncertainty is defined as the posterior probability distribution of future events conditioned on all available information at the time of forecast (Todini, 2011). There are various sources of uncertainties within streamflow forecasting which can be categorized into two main groups (Krzysztofowicz & Kelly, 2000; Seo et al., 2006): (1) input (meteorological forcing) uncertainty, and (2) hydrologic uncertainty. Apart from the unknown future meteorological variables, other sources of uncertainties, such as

errors in observational measurements, structure and parameter values of the hydrologic model, and initial conditions (Ajami et al., 2007; Madadgar & Moradkhani, 2014; Montanari et al., 2009), are aggregated as hydrologic uncertainty and the significance of their quantification is dependent on factors such as the forecasting horizon (Biondi & Todini, 2018).

In recent years, various post-processors have been developed for quantifying and reducing the uncertainty of hydrological forecasts, which are comprehensively reviewed in Li et al. (2017) and Han and Coulibaly (2017). Among these approaches, the Bayesian Forecasting System (BFS; Krzysztofowicz, 1999) appears a reliable and robust probabilistic forecasting framework, which can explicitly address input and hydrologic uncertainties using the precipitation uncertainty processor (PUP; Kelly & Krzysztofowicz, 2000) and the hydrologic uncertainty processor (HUP; Krzysztofowicz & Kelly, 2000), respectively. Using Bayes theorem, the HUP explicitly quantifies the hydrologic uncertainty by providing a posterior distribution for any deterministic forecast, derived from a hydrologic model, based on the assumption of possessing a perfect precipitation forecast (Han et al., 2019; Liu et al., 2018).

There are several studies that evaluated the application of the HUP for hydrologic uncertainty estimation (Biondi et al., 2010; Krzysztofowicz & Herr, 2001; Krzysztofowicz & Kelly, 2000; Liu et al., 2016, 2018; Reggiani et al., 2009; Reggiani & Weerts, 2008), some of which also improved the HUP procedure. Krzysztofowicz and Herr (2001), for instance, proposed the precipitation-dependent version of HUP, which is shown to be more efficient. Using the HUP concept, Reggiani et al. (2009) developed the Bayesian Ensemble

Uncertainty Processor which implicitly quantifies input uncertainty by aggregating the HUP-based posterior distributions of various streamflow forecasts stemming from an ensemble of precipitation forecasts. Also, the copula-based HUP, proposed by Liu et al. (2018), used the advantage characteristics of copula functions to develop the prior density and the likelihood function without transforming data into Gaussian space, which is required in the HUP method. Their results show that the proposed modified approach is as reliable as HUP in terms of probabilistic streamflow forecast. Although different studies investigated HUP from various aspects, very few studies evaluate the effects of using different hydrologic models on HUP performance. Recently, by comparing the use of HUP with two lumped hydrologic models, Han et al. (2019) show that the quality of a deterministic model is an important factor in HUP in order to produce reliable and accurate probabilistic forecasts.

On the other hand, some multi-model post-processing approaches are combining multiple model forecasts for generating more reliable results. Multi-models hydrological predictions, compared to the single deterministic one, provide more information about the unknown future events and can better reflect the uncertainties associated with streamflow forecasting, however, the statistical post-processing approach is still required to produce accurate and reliable forecasts (Li et al., 2017; Muhammad et al., 2018; Reggiani & Weerts, 2008). There are various deterministic model averaging techniques (e.g. Bates-Granger averaging (Bates & Granger, 1969), Granger–Ramanathan averaging (Granger & Ramanathan, 1984), etc.), which provide one point-estimation of the predictand; however, some post-processing methods have been developed to treat multi-model streamflow forecasts in a

probabilistic way by quantifying predictive uncertainty induced by the imperfection of models' structures. Bayesian Model averaging (BMA) (Raftery, 1993; Raftery et al., 2005) is one the most well-known statistical multi-model post-processing approaches which has been widely and successfully applied in streamflow simulation and forecasting studies (e.g. Darbandsari & Coulibaly, 2019; Duan et al., 2007; Huo et al., 2019; Madadgar & Moradkhani, 2014; Parrish et al., 2012; Sharma et al., 2019; Vrugt & Robinson, 2007).

In BMA, the predictive posterior distribution is quantified as a weighted average of the conditional probability distributions of individual forecasts, which are assumed to follow the Gaussian distribution. For highly skewed variables like streamflow, the conditional probability distribution with the aforementioned assumption is a poor choice. So, various studies proposed different BMA modifications, such as implementing other distribution types (Vrugt & Robinson, 2007), and applying data transformation procedure (Liang et al., 2013; Qu et al., 2017), to more complex Copula-embedded BMA (Madadgar & Moradkhani, 2014) where any assumption about the distribution shape is relaxed using the properties of copula functions. Moreover, addressing the law of total probability as another inherent assumption of the BMA approach, Darbandsari and Coulibaly (2020b) recently proposed the Entropy-based BMA method where an ensemble with mutually exclusive and collectively exhaustive properties is constructed prior to the BMA application, leading to better probabilistic high flow forecasts. Also, using the integration of BMA with other techniques, some other multi-model methods have been proposed for uncertainty analysis and quantifications (e.g. Ajami et al., 2007; Parrish et al., 2012; Poeter & Hill, 2007; Rojas et al., 2008; Sharma et al., 2019; Yen et al., 2014). Although BMA and its variants are

among the most reliable approaches for estimating predictive uncertainty based on an ensemble of forecasts, they do not consider the effects of initial conditions explicitly, and the use of external updating procedure is required for better estimation of hydrologic uncertainty for operational streamflow forecasting (Darbandsari & Coulibaly, 2020b; Todini, 2008; Xu et al., 2019).

The main objective of this study is to evaluate the benefits of using multiple deterministic forecasts within the HUP procedure for better quantifying hydrologic uncertainty. Besides evaluating the effects of using different hydrologic models on HUP performance, an extension of the Bayesian post-processor is proposed by integrating the HUP and BMA approaches (called HUP-BMA hereafter) which can incorporate multi-model ensemble streamflow forecasts. Two different watersheds are used as case studies to assess the applicability and efficiency of the proposed HUP-BMA method for short- to medium-range daily streamflow forecast (1- to 14- days ahead) using different deterministic and probabilistic performance criteria. Compared with HUP, which requires a single deterministic forecast, the HUP-BMA approach takes the advantage of an ensemble of individual predictions to better quantify and reduce the hydrologic uncertainty and enhance the accuracy and reliability of streamflow forecasts. Similar to the HUP, the parameters of the proposed approach can be calibrated offline and the method can be easily implemented for operational use (Han et al., 2019; Krzysztofowicz & Herr, 2001). Moreover, using the advantages of the HUP method, this post-processor could be an alternative for BMA in operational streamflow forecasting by deciphering the initial condition uncertainty, which

leads to the better quantification of the hydrologic uncertainty associated with short-term streamflow forecasts.

The remainder of the paper is as follows. In section 5.3, the underlying concepts of the applied methodologies (e.g. HUP and BMA), an overview of the proposed HUP-BMA, and a detailed explanation of various evaluation metrics for assessing the performance of forecasts are given. Section 5.4 presents the experimental setup, including brief descriptions of the case studies and data, and the employed rainfall-runoff models. Section 5.5 discussed the results, and the summary and conclusions are presented in Section 5.6.

5.3 Methods

5.3.1 Hydrologic Uncertainty Processor

Hydrologic uncertainty processor (HUP), firstly introduced by Krzysztofowicz and Kelly (2000), is a Bayesian method for quantifying the hydrologic uncertainty conditioned on initial observation and a deterministic prediction, based on the assumption that the precipitation uncertainty is zero. In other words, HUP aims to estimate hydrologic uncertainty using real-time observations and a deterministic forecast from a hydrologic model. Given the detailed information about the HUP method in the literature (Krzysztofowicz, 2002; Krzysztofowicz & Kelly, 2000), a brief explanation of its basic concepts is provided for the sake of completeness.

Let Y_0 be the observed river discharge at the initial date of forecast (i.e. $n = 0$) and $Y = (Y_1, Y_2, \dots, Y_N)$ is the vector of the actual river discharge at forecasting times 1 to N. Similarly, $\hat{Y} = (\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_N)$, denoted as the model river discharge, is the vector of

estimates of $Y_n: n = (1,2, \dots, N)$ derived from the output of a deterministic hydrologic model based on the perfect precipitation forecast. The realizations of the above mentioned random variables Y_0, Y_n, and $\hat{Y}_n$ are respectively presented by y_0, y_n, and $\hat{y}_n$. Using Bayes theorem, the HUP procedure tries to quantify hydrologic uncertainty by supplying posterior densities of the actual river discharge at lead-time n (y_n), as the quantity to be forecasted, conditioned on $\hat{Y}_n = \hat{y}_n$, and $Y_0 = y_0$ (Krzysztofowicz, 1999):

$$\varphi_n(y_n|\hat{y}_n, y_0) = \frac{f_n(\hat{y}_n|y_n, y_0) g_n(y_n|y_0)}{\kappa_n(\hat{y}_n|y_0)} \quad (5\text{-}1)$$

where $g_n(y_n|y_0)$ is the prior uncertainty of the actual river discharge at lead-time n given the observation $Y_0 = y_0$, and $f_n(\hat{y}_n|y_n, y_0)$ is the likelihood of model river discharge. The expected density of the model river discharge conditional on the observed initial discharge ($\kappa_n(\hat{y}_n|y_0)$) can be determined as a function of the prior density and the likelihood function, using the law of total probability:

$$\kappa_n(\hat{y}_n|y_0) = \int_{-\infty}^{+\infty} f_n(\hat{y}_n|y_n, y_0) g_n(y_n|y_0) dy_n \quad (5\text{-}2)$$

Through the HUP process, the following main steps have been taken to estimate the aforementioned posterior density:

Normal Quantile Transform

HUP is a meta-Gaussian model where the families of conditional densities are assumed to follow the Gaussian distribution after transforming data into a Normal space (Biondi et al., 2010; Kelly & Krzysztofowicz, 2000; Krzysztofowicz & Herr, 2001). Therefore, the

Normal Quantile Transform (NQT) is used as a primary step of the HUP method for converting Y_n and $\hat{Y}_n$ into the Normal space (random variates X_n and $\hat{X}_n$ respectively) using the following equations:

$$X_n = Q^{-1}\left(\Gamma_n(Y_n)\right) \quad \forall\, n \in \{0,1,2,\dots,N\} \tag{5-3}$$

$$\hat{X}_n = Q^{-1}\left(\bar{\Lambda}_n\left(\hat{Y}_n\right)\right) \quad \forall\, n \in \{1,2,\dots,N\} \tag{5-4}$$

where $Q^{-1}(.)$ is the inverse of standard normal distribution and Γ_n and $\bar{\Lambda}_n$ are the marginal distributions of the actual and model river discharges at lead-time n, respectively. NQT is one of the most general transformation approach (Krzysztofowicz, 1997), which has been applied successively in analyzing streamflow forecast uncertainty (Kelly & Krzysztofowicz, 1997; Liang et al., 2013; Montanari & Brath, 2004; Reggiani et al., 2009). This conversion makes the HUP formulation reliable for variables with any types of marginal distributions, and heteroscedastic and nonlinear dependence structure, which are necessary features for streamflow forecasting (Krzysztofowicz & Herr, 2001). The lower case letters x_n and $\hat{x}_n$, indicate the experimental values (realizations) of the transformed variates X_n and $\hat{X}_n$, respectively.

Prior density and likelihood function in the transformed space

As can be seen from Equation 5-1, proper estimations of the prior density and the likelihood function are the key requirements for the Bayes theorem application. In the HUP, by assuming the strictly stationary lag-one Markovian process, the stochastic dependence

structure between each two consecutive actual river discharges in the transformed space (i.e. X_n and X_{n-1}) is governed by:

$$X_n = c_n \times X_{n-1} + \varepsilon_n \quad \forall\, n \in \{1,2,\dots,N\} \tag{5-5}$$

where c_n is the parameter and ε_n is an independent normally distributed variate with mean zero and variance $1 - c_n^2$. Based on the stationary assumption and successive application of the aforementioned lag-one process, the transition density ($r_{Q_n}(x_n|x_{n-1})$) and consequently the prior density ($g_{Q_n}(x_n|x_0)$) in the Gaussian space is estimated as follows:

$$r_{Q_n}(x_n|x_{n-1}) = \frac{1}{(1-c_n^2)^{0.5}}\, q\left(\frac{x_n - c_n x_{n-1}}{(1-c_n^2)^{0.5}}\right) \tag{5-6}$$

$$g_{Q_n}(x_n|x_0) = \frac{1}{t_n^{0.5}}\, q\left(\frac{x_n - C_n x_0}{t_n^{0.5}}\right) \tag{5-7}$$

Subscript Q denotes a density in the space of transformed variates. $q(.)$ is the standard normal density function, and C_n and t_n are dependent parameters calculated by (Krzysztofowicz & Herr, 2001):

$$C_n = \prod_{i=1}^{n} c_i \tag{5-8}$$

$$t_n = 1 - C_n^2 \tag{5-9}$$

For estimating the likelihood function, the linear regression is used for characterizing the dependence structure between the transformed actual (X_n) and model ($\hat{X}_n$) river discharges:

$$\hat{X}_n = a_n \times X_n + d_n \times X_0 + b_n + \theta_n \quad \forall\, n \in \{1,2,\dots,N\} \tag{5-10}$$

where, θ_n is an independent normally distributed variate with mean zero and variance σ_n^2, and a_n, d_n, and b_n are the regression parameters. Therefore, the likelihood function of the predictand x_n in Normal space is determined as follows:

$$f_{Q_n}(\hat{x}_n|x_n, x_0) = \frac{1}{\sigma_n} q\left(\frac{\hat{x}_n - (a_n x_n + d_n x_0 + b_n)}{\sigma_n}\right) \tag{5-11}$$

Posterior density in transformed space

After estimating the prior density (Equation 5-7) and the likelihood function (Equation 5-11), which are both normal-linear, the theory of conjugate families of distribution (DeGroot, 2005) is utilized to derive the closed-form expression of the posterior density in the transformed space through Bayes theorem (Krzysztofowicz & Kelly, 2000):

$$\varphi_{Q_n}(x_n|\hat{x}_n, x_0) = \frac{1}{T_n} q\left(\frac{x_n - (A_n \hat{x}_n + D_n x_0 + B_n)}{T_n}\right) \tag{5-12}$$

where A_n, B_n, D_n, and T_n are the parameters which are determined as follows:

$$A_n = \frac{a_n t_n^2}{a_n^2 t_n^2 + \sigma_n^2} \tag{5-13}$$

$$B_n = \frac{-a_n b_n t_n^2}{a_n^2 t_n^2 + \sigma_n^2} \tag{5-14}$$

$$D_n = \frac{C_n \sigma_n^2 - a_n d_n t_n^2}{a_n^2 t_n^2 + \sigma_n^2} \tag{5-15}$$

$$T_n = \left(\frac{t_n^2 \sigma_n^2}{a_n^2 t_n^2 + \sigma_n^2}\right)^{0.5} \quad (5\text{-}16)$$

Transform back to the original space

Using the Jacobian of transformation, the estimated posterior density in the Gaussian space ($\varphi_{Q_n}(x_n|\hat{x}_n, x_0)$), is converted back to the original space for determining the meta-Gaussian posterior distribution ($\Phi_n(y_n|\hat{y}_n, y_0)$), which becomes (Krzysztofowicz & Kelly, 2000):

$$\Phi_n(y_n|\hat{y}_n, y_0) = Q\left(\frac{Q^{-1}\left(\Gamma_n(y_n)\right) - A_n Q^{-1}\left(\bar{\Lambda}_n(\hat{y}_n)\right) - D_n Q^{-1}\left(\Gamma_n(y_0)\right) - B_n}{T_n}\right) \quad (5\text{-}17)$$

Same as Equations 5-3 and 5-4, Γ_n and $\bar{\Lambda}_n$ are the marginal distributions for the actual (y_n) and the model ($\hat{y}_n$) river discharge variates at forecasting time n.

Altogether, using a joint sample of realizations $\{(y_0, y_1, \dots, y_N; \hat{y}_1, \dots, \hat{y}_N)\}$, which is formed based on historically observed discharges and the output of the deterministic hydrologic model, the estimation procedure of the HUP method for forecasting time $n = (1,2, \dots, N)$ includes: (1) estimating the marginal prior distributions (i.e. Γ_n, and $\bar{\Lambda}_n$), (2) transforming training data into the Gaussian space using the NQT approach (Equations 5-3 and 5-4), (3) estimating the prior densities and the likelihood functions parameters in the transformed space (i.e. c_n, a_n, b_n, d_n and σ_n in Equations 5-5 to 5-11), and (4) determining the parameters of the posterior densities using Equations 5-13 to 5-16 (i.e. A_n,

B_n, D_n, T_n). These estimated parameters can then be utilized in the forecasting mode for probabilistic streamflow forecasting up to N days ahead.

The precipitation-dependent HUP, proposed by Krzysztofowicz and Herr (Krzysztofowicz & Herr, 2001), is a two-branch procedure based on the non-occurrence ($v = 0$) and occurrence ($v = 1$) of precipitation. Before the HUP estimation procedure, the training period is divided into two groups of data with and without the occurrence of precipitation in their initial date. Two sets of HUP parameters are then separately estimated for each branch (i.e. $\Gamma_{n,v}, \bar{\Lambda}_{n,v}, c_{n,v}$, $a_{n,v}$, $b_{n,v}$, $d_{n,v}$, $A_{n,v}$, $B_{n,v}$, $D_{n,v}$, $T_{n,v}$ $\forall\, v \in \{0.1\}$). The probabilistic forecasts have been generated using the first branch ($v = 0$) if there is no precipitation while the second branch ($v = 1$) is utilized for streamflow forecasting in the case of precipitation occurrence. By better capturing the model structural uncertainty and explicitly considering the effects of transition between the recession and the rising limbs of the hydrographs, the precipitation-dependent HUP has better predictive capabilities than the independent one (Biondi et al., 2010; Krzysztofowicz & Herr, 2001). Therefore, the precipitation-dependent HUP, abbreviated as HUP hereafter, is used in this study.

5.3.2 Multi-model Bayesian processor (HUP-BMA)

As previously stated, the HUP procedure employs a deterministic hydrologic model to quantify the hydrologic uncertainty and generate the posterior distribution. However, there are various structurally different hydrologic models, and using all information derived from multi-model ensemble predictions can enhance the reliability and accuracy of the probabilistic forecasts (Ajami et al., 2007; Dong et al., 2013; Jiang et al., 2018). Here, by

implementing the concept of the Bayesian Model Averaging (BMA) method (Raftery, 1993), the HUP-derived posterior distributions based on multiple hydrologic models are integrated in order to estimate the final predictive distribution conditioned on the initially observed flow and multi-model streamflow forecasts ensemble.

Bayesian Model Averaging (BMA) is one of the most well-known multi-model post-processing approaches where the combination of forecast conditional densities, derived from different models, are weighted for generating the final posterior distribution. Following the law of total probability, the BMA predictive distribution of a forecasted variable at lead-time n (y_n) conditioned on the ensemble of K different multi-model forecasts ($M_n = (\hat{y}_n^1, \hat{y}_n^2, \dots, \hat{y}_n^K)$) is defined as follows (Raftery et al., 2005):

$$P(y_n|M_n) = \sum_{i=1}^{K} w_i \times P\left(y_n\middle|\hat{y}_n^i\right) \qquad (5\text{-}18)$$

where, w_i are the BMA weights need to be estimated, showing how well the forecast $\hat{y}_n^i$ fits the observation in the calibration period. $P\left(y_n\middle|\hat{y}_n^i\right)$ is the forecast (or conditional) probability distribution of the predictand y_n given the ensemble member $\hat{y}_n^i$. In the original BMA, this conditional density (i.e. $P\left(y_n\middle|\hat{y}_n^i\right)$) is assumed to be approximately normally distributed with mean $\mu_n^i = \rho_{1,n}^i + \rho_{2,n}^i \hat{y}_n^i$ and variance ${\tau^2}_n^i$. $\rho_{1,n}^i$, $\rho_{2,n}^i$ are the bias correction coefficients, which are estimated by a simple linear regression of y_n on $\hat{y}_n^i$ in the calibration period. Therefore, Equation 5-18 can be rewritten as follows:

$$P(y_n|M_n) = \sum_{i=1}^{K} w_i \times N(y_n|\mu_n^i, {\tau^2}_n^i) \qquad (5\text{-}19)$$

The parameters of the standard BMA approach, including weights (w_i) and variances (${\tau^2}_n^i$) of each member $i \in \{1,2,\dots,K\}$, are estimated through the Expectation-Maximization (McLachlan & Krishnan, 2008) iterative algorithm where a two-step procedure is used for searching the optimal parameter values by maximizing the log-likelihood function (Figure 5-1a). The assumption of the Gaussian conditional probability distribution (i.e. $P(y_n|\hat{y}_n^i) \sim N(y_n|\mu_n^i, {\tau^2}_n^i)$) in the original development of the BMA method might not be a proper choice for river discharge. Therefore, using more representative distribution (e.g. Gamma) (Vrugt & Robinson, 2007) or applying a data transformation procedure (Duan et al., 2007; Qu et al., 2017; Roy et al., 2017; Sharma et al., 2019; Todini, 2008) is recommended in order to achieve more reliable results.

As previously mentioned, using the basic concept of BMA, this study tried to merge an ensemble of posterior distributions derived from the application of the HUP (Equation 5-17) in conjunction with K different models (i.e. $\Phi_n^i(y_n|\hat{y}_n^i, y_0)\ \forall\ i \in \{1,2,\dots,K\}$) in order to generate reliable and accurate predictive forecasts. Therefore, the parametric conditional normal distributions ($N(y_n|\mu_n^i, {\tau^2}_n^i)\ \forall\ i \in \{1,2,\dots,K\}$) in Equation 5-19 are replaced by pre-estimated HUP derived posterior probabilities in Equation 5-17, and the final PDF conditioned on all forecast members ($\Phi_n(y_n|M_n, y_0)$; $M_n = (\hat{y}_n^1, \hat{y}_n^2, \dots, \hat{y}_n^K)$) is determined by:

$$\Phi_n(y_n|M_n, y_0) = \sum_{i=1}^{K} w_i \times \Phi_n^i\left(y_n\middle|\hat{y}_n^i, y_0\right) \quad (5\text{-}20)$$

As can be seen, in the proposed HUP-embedded-BMA, called HUP-BMA hereafter, the weights are the only BMA parameters that need to be estimated. So, the modified EM algorithm, proposed by Madadgar and Moradkhani (2014), is used where in contrast with standard EM, the pre-specified HUP-based posterior probabilities remain the same for all the iterations (Figure 5-1b). Compared with BMA, the proposed HUP-BMA method does not need the external applications of the linear bias correction and data transformation because these processes have been already embedded in the HUP procedure. Additionally, compared with BMA, the HUP-BMA quantifies the initial condition uncertainty by explicitly implementing the initial state knowledge (i.e. actual river discharge at time zero).

Figure 5-2 illustrates the calibration process of the proposed HUP-BMA approach. As can be seen, for each forecast lead-time n, after dividing data into two groups based on the occurrence ($v = 1$) and non-occurrence ($v = 0$) of precipitation, the HUP calibration procedure is done for each member of the streamflow forecast ensemble. The estimated HUP parameters are then utilized for calculating the posterior distributions based on different members over the whole calibration period. Finally using the estimated HUP derived densities, the modified EM algorithm is employed for determining the BMA parameters ($w_{i,n,v}$). It is worthy of note that two sets of weights are determined for each branch of data with and without precipitation occurrence. Finally, the offline calibrated HUP-BMA method can be executed online for streamflow forecasting based on multiple deterministic forecasts derived from various hydrologic models. Using all the information

obtained from different models, the proposed HUP-BMA procedure can better estimate the hydrologic uncertainty and provide more accurate and reliable forecasts.

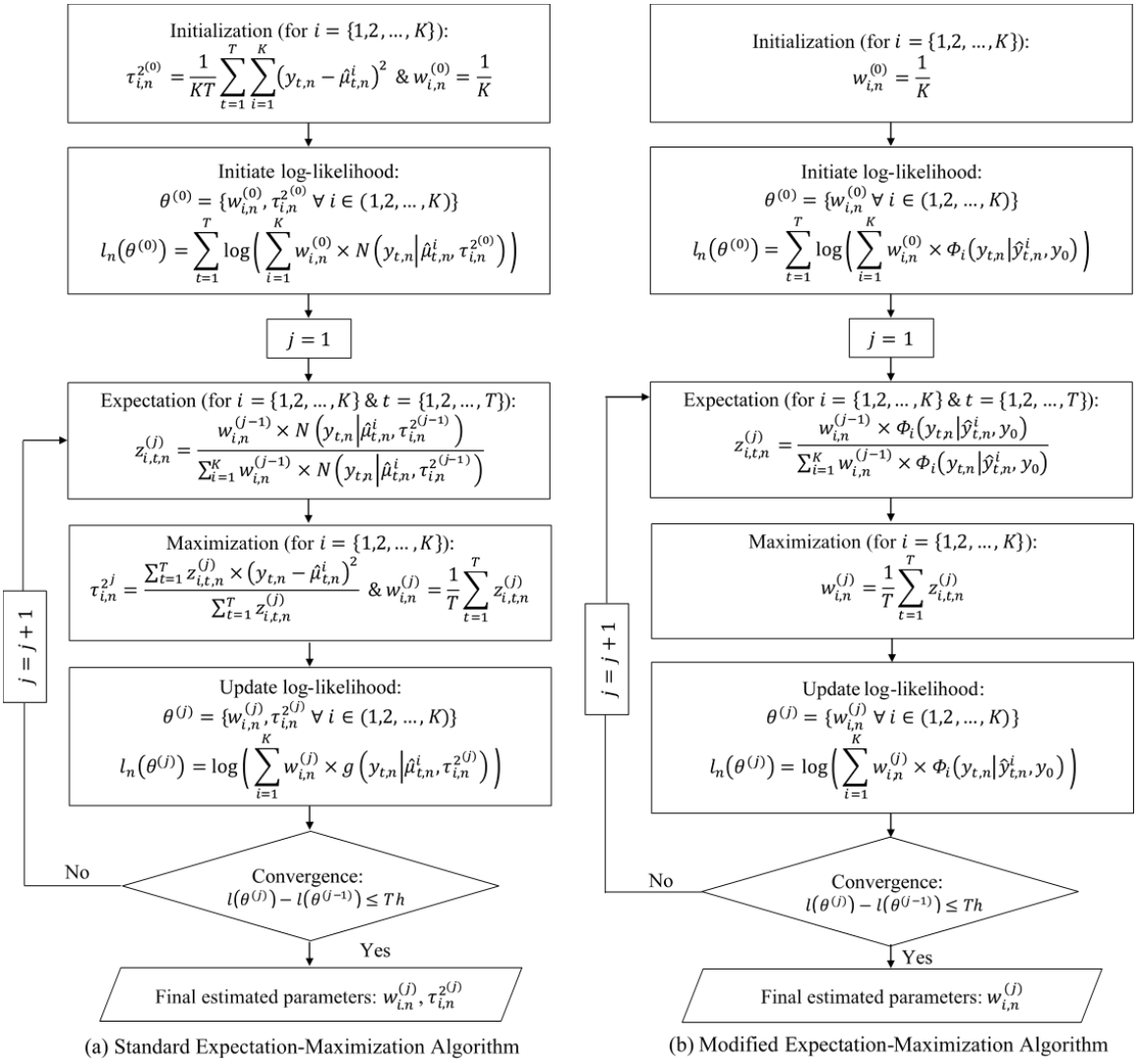


Figure 5-1 The step-by-step procedure of (a) the standard Expectation-Maximization (EM) and (b) the modified EM algorithms at forecasting time n. z is a latent variable, K is the number of ensemble members, T is the length of the calibration period, and Th is the pre-specified tolerance level

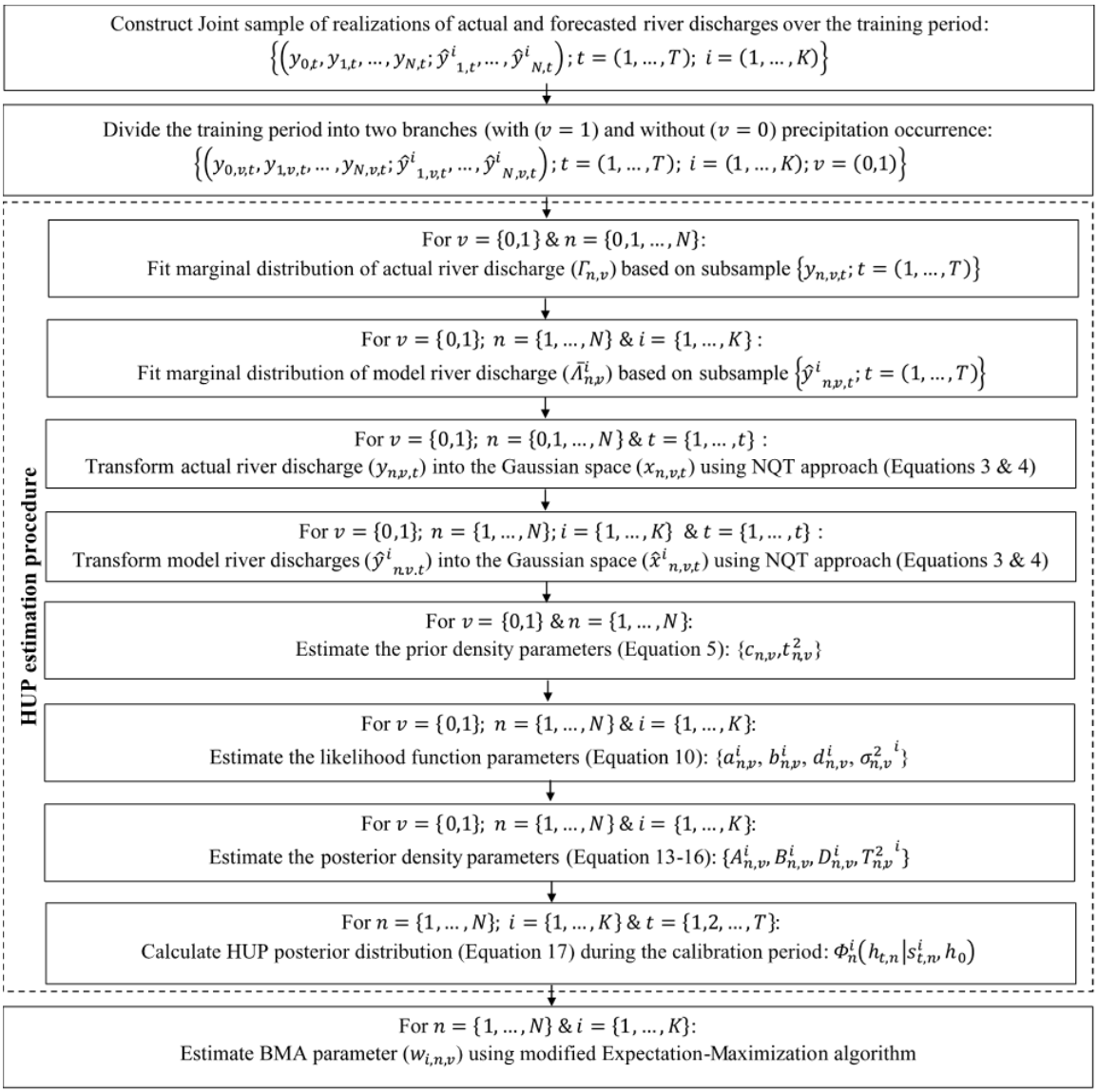


Figure 5-2 The flow chart of the proposed HUP-BMA calibration process. T is the length of the calibration period, K is the number of forecasts ensemble members, and N is the length of the forecasting horizons

5.3.3 Performance evaluation metrics

In this study, we utilize eight different criteria in order to assess the forecast performance of different post-processing approaches in terms of accuracy, reliability, and sharpness as

the three main important aspects of any probabilistic predictions (Elshall et al., 2018). Accuracy is defined as the level of agreement between forecasts and their corresponding observations. Considering the mean value of predictive distribution as the deterministic forecast, the well-known Nash Sutcliffe Efficiency (NSE) and the Mean Absolute Error (MAE) criteria are calculated to deterministically evaluate the accuracy of the forecasts:

$$NSE = 1 - \frac{\sum_{t=1}^{N}(F_t - O_t)^2}{\sum_{t=1}^{N}(O_t - \bar{O})^2} \quad (5\text{-}21)$$

$$MAE = \frac{\sum_{t=1}^{N}(|F_t - O_t|)}{N} \quad (5\text{-}22)$$

In the above equation, N is the length of time series, O_t and F_t respectively are the observed and forecasted variables, and $\bar{O}$ shows the mean of the observed flows. NSE, possessing a range of $(-\infty, 1]$, reflects how well the forecast represents the observed data by considering the observation mean as the benchmark. The higher NSE values correspond to the better predictions while its negative values occur when the residuals of the forecast are larger than observation variance (Nash & Sutcliffe, 1970; Strauch et al., 2012). Although NSE gives more weights to lager errors (Krause et al., 2005; Seiller et al., 2012), MAE is the absolute criteria which provides a more balanced error measurements for assessing the average performance (Kisi & Cimen, 2011; Willmott & Matsuura, 2005). It varies between 0 and +∞ with the best value of 0. Also, NSE of log-transformed ($NSEL$) and squared-transformed ($NSES$) streamflows are used as two other deterministic metrics emphasizing on the accuracy of the lower and higher forecasted flow values, respectively (Darbandsari & Coulibaly, 2020a; Razavi & Coulibaly, 2017).

Moreover, two probabilistic-based measures, the mean continuous ranked probability score ($CRPS$) (Hersbach, 2000) and the average deviation amplitude ($ADA95$) (Xiong et al., 2009) are used to evaluate the accuracy of the predictive forecasts:

$$CRPS = \frac{1}{N}\sum_{t=1}^{N}\int_{x=-\infty}^{x=+\infty}\left(P_t^F(x) - P_t^O(x)\right)^2 d_x \ , P_t^O(x) = H(x - O_t) \qquad (5\text{-}23)$$

$$ADA95 = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{1}{2}\left(q_t^u + q_t^l\right) - O_t\right| \qquad (5\text{-}24)$$

where, $P_t^F(.)$ and $P_t^O(.)$ represent the probability distributions of the forecasted and observed flows, respectively, and $H(x - O_t)$ is the Heaviside function, being equal to zero if $x < O_t$, and 1 otherwise. q_t^l and q_t^u are the lower and upper boundaries of the 95 percent prediction bound. The $CRPS$ is the average of the squared error of the forecast cumulative probability distributions compared to the observation, and $ADA95$ calculates the discrepancy between the middle point of the confidence bound and observations. Both criteria have a negative orientation, in which the better forecasts receive lower values.

Reliability and sharpness are the other two important aspects of any probabilistic forecast. In this study, the containing ratio ($CR95$), defined as the ratio of observations enveloped by 95% confidence interval, and the average bandwidth ($BW95$), representing the average width of the corresponding bound, are adopted to respectively evaluate the reliability and sharpness of the predictive forecast (Xiong et al., 2009):

$$CR95 = \frac{N_{O_{in}}}{N} \qquad (5\text{-}25)$$

$$BW95 = \frac{1}{N}\sum_{t=1}^{N}\left(q_t^u - q_t^l\right) \quad (5\text{-}26)$$

$N_{O_{in}}$ is the number of observations covered by the 95% confidence bound, and all other variables are defined similar to the previous equations. Simultaneous evaluation of these two metrics is necessary for providing the right conclusions. For $CR95$, varies between 0 and 1, the value closer to 95% is preferable, however, it can lead to overestimation of the uncertainty with a large prediction bound (i.e. large $BW95$). On the other hand, the forecasts with a narrow confidence interval (i.e. smaller $BW95$) can be overconfident if the ratio of covering is low (Parrish et al., 2012).

5.4 Experimental Setup

5.4.1 Study area and data

We carried out the aforementioned post-processing approaches on two hydrologic basins located in Northern Ontario, Canada: (1) Big East River, and (2) Black River watersheds. Apart from the southern part of Black River, which is used for agricultural purposes, the dominant area of both watersheds is mixed forest vegetation. There are no meteorological stations within the boundaries of either basin (Figure 5-3). These two watersheds are poorly gauged. Low-density ground-based measurements can be used for estimating temperature while capturing the temporal and spatial variability of precipitation required more reliable data (Price et al., 2014). Therefore, the archived daily aggregated Canadian Precipitation Analysis (CaPA) data (Mahfouf et al., 2007) was used to estimate the mean areal precipitation of both watersheds. CaPA, with a temporal and spatial resolution of 6 hours

and 15 km, respectively, is a precipitation product generated based on the combination of observations and climate model data (Lespinas et al., 2015), which has been demonstrated to have reliable performance as precipitation forcing of hydrologic models in Canadian catchments (Boluwade et al., 2018; Darbandsari & Coulibaly, 2020a; Eum et al., 2014).

The geophysical and climate characteristics of both basins are summarized in Table 5-1. As can be seen, temperature changes indicate the presence of all four seasons in the study areas. Moreover, the seasonal precipitation amount in spring is less than in the other seasons; however, the flow is the highest. This shows the significant impact of the snowmelt process on the hydrological characteristics of both regions. Furthermore, although the climatologic conditions of both basins are almost the same, they possess quite distinct hydrologic responses leading to the outlet streamflow values with very different statistical characteristics (Figure 5-4).

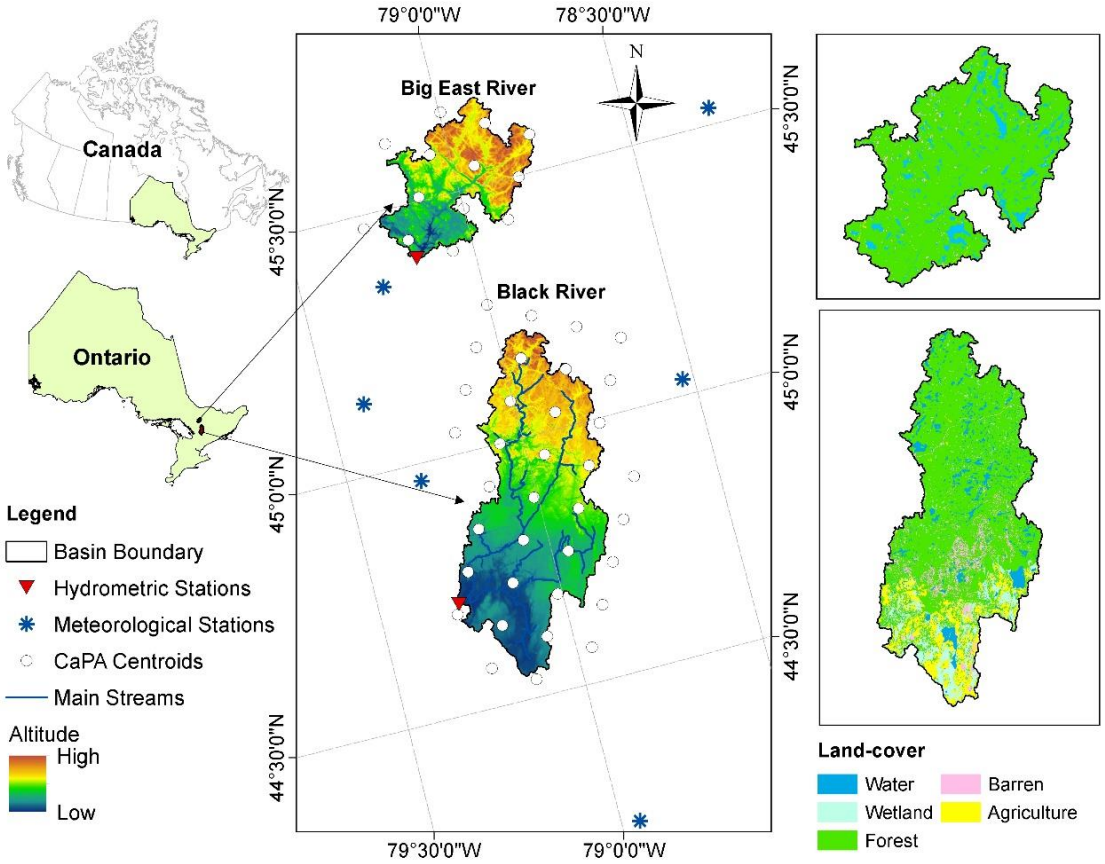


Figure 5-3 Location map of the Big East River and Black River watersheds

Table 5-1 Geophysical and climatic characteristics of the Big East River and Black River watersheds

Characteristics	**Basins**							
	Big East River				**Black River**			
Area (km2)	620				1522			
Elevation range (m.a.s.l)	[290-570]				[220-420]			
Average Slope (m/km)	10.2				4.9			
Data period	2006-2015				2006-2015			
Long-term seasonal statistics	**Spring**	**Summer**	**Fall**	**Winter**	**Spring**	**Summer**	**Fall**	**Winter**
Average precipitation (mm/season)	218	248	310	247	210	223	272	239
Average daily temperature (C)	3.5	16.7	6.2	-10.1	4.2	17.1	7.1	-8.5
Average outlet flow (mm/season)	275	82	129	159	240	45	85	162
Average of daily flow (m3/s)	21.4	6.4	10.2	12.7	46.1	8.7	16.5	31.7

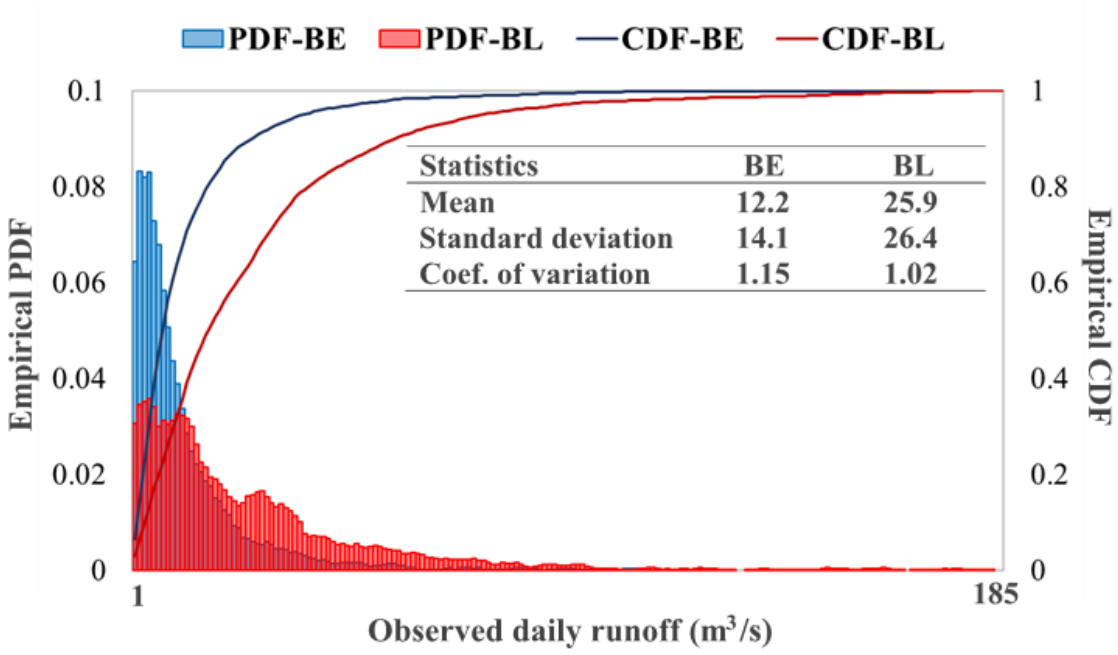


Figure 5-4 The empirical PDF and CDF of the daily streamflow observation at the outlet of the Big East River (BE) and Black River (BL) watersheds and their corresponding statistical measures

5.4.2 Hydrological models

With no climatic and hydrometric stations within the basins and lack of some geographical information, such as soil data, both watersheds can be categorized as data-scarce regions where conceptual models are more suitable for simulating the rainfall-runoff process (Refsgaard & Knudsen, 1996; Srivastava et al., 2020; Tegegne et al., 2017). In this study, we used seven structurally different conceptual hydrologic models with various parameterizations for streamflow forecasting in both basins (Table 5-2). The 4-parameter parsimonious GR4J model, with the least complexity among the others, relies on two conceptual storages and the unit hydrograph concept for simulating the whole rainfall-runoff process (Perrin et al., 2003), and it was proven to perform well in regions with low

data availability (Anshuman et al., 2019; Darbandsari & Coulibaly, 2020a). SMARG, the modified version of the SMAR model (O'Connell et al., 1970) with superior ability in semi-humid and humid regions (Tan & O'Connor, 1996), uses the variable number of soil layers for generating surface and groundwater runoffs, being respectively transferred using the Nash cascade of variable linear reservoirs and a single linear reservoir. MACHBV, the nonlinear version of HBV (Bergström, 1976), originally developed for better simulation of stream flows of ungauged basins in Canada (Samuel et al., 2011, 2012), includes a single soil moisture storage, and a two-layer response function for estimating the runoff value and the non-linear Equilateral triangular weighting function for flow routing. In the well-known SACSMA model, widely used for flood forecasting (Dong-Jun Seo et al., 2003; Vrugt et al., 2006; Wijayarathne & Coulibaly, 2020), five soil storage layers and the Nash cascade of three linear reservoirs are utilized for simulating the hydrologic processes of the basin (Razavi & Coulibaly, 2017). Moreover, various available conceptual approaches for modeling different components of rainfall-runoff cycles make the HEC-HMS software a reliable platform for generating structurally different models (Scharffenberg, 2016; Teng et al., 2018). Therefore, here, the first HEC-HMS based model (i.e. HEC1) combines the Deficit and Constant loss method and the Recession baseflow method while the conjunction of the soil moisture accounting with the Recession and Linear Reservoir baseflow methods are used for developing the other two configurations (i.e. HEC2 and HEC3, respectively).

Moreover, as can be seen in Table 5-2, two potential evapotranspiration estimation methods are used with different models, so the daily mean areal precipitation and

temperature are the only inputs of the models. Also, due to the importance of the snowmelt process in both study regions, three different snowmelt routines, being employed in conjunction with different models, are the simple degree-day (Samuel et al., 2011), the HEC-HMS snowmelt routine (Scharffenberg, 2016), and Snow-17 (Anderson, 2006) with five, ten, and eleven user-specified parameters, respectively. Different complexities of the snowmelt methods lead to more diverse hydrological models. Further descriptions of all hydrologic models as well as the utilized snow routing approaches can be found in the above-cited references.

Table 5-2 The main characteristics of the hydrologic models implemented in this study

Model	Number of Conceptual Storages	Number of optimized parameters	Snowmelt estimation routine	PET estimation method	Reference
SMARG	Variable	9	Degree-Day	1*	Tan and O'Connor (1996)
GR4J	2	4	Degree-Day	1	Perrin et al., (2003)
MACHBV	3	10	Snow-17	1	Samuel et al. (2011)
SACSMA	5	14	Snow-17	1	Burnash et al. (1973)
HEC1**	2	7	HEC-HMS	2*	Scharffenberg (2016)
HEC2**	4	15	HEC-HMS	2	Scharffenberg (2016)
HEC3**	6	17	HEC-HMS	2	Scharffenberg (2016)

* 1: Simplified Thornwaite formula (Samuel et al., 2011); 2: Hargreaves formula (Hargreaves & Samani, 1985)

** Three different HEC-HMS configurations

The dynamically dimensioned search optimization algorithm (Tolson & Shoemaker, 2007) is used for automatically estimating the models' parameters based on the 6-year calibration period (2006-2011). Three different performance evaluation metrics, including Nash Sutcliffe Efficiency (NSE), Kling Gupta Efficiency (KGE) (Gupta et al., 2009) and Nash Volume Error (NVE) (Samuel et al., 2011) are used as various single objective functions

in the calibration process, leading to three different optimized sets of parameters for each hydrologic model at each watershed. Finally, the best parameter set of each hydrologic model is selected by visually comparing the performance of three potential parameter sets using the whole observed and simulated hydrographs as well as the Mean Absolute Error (MAE; Equation 5-22) and the Nash Sutcliff Efficiency (NSE; Equation 5-21) as the overall performance measurements. This helps us achieve more robust conclusions about the potential capability of each rainfall-runoff model by avoiding the possible systematic errors or over-parameterization issues which may be caused by the automatic optimization process (Gan et al., 1997; Ouermi et al., 2019; Wöhling et al., 2013).

Similar to NSE (Equation 5-21), both KGE and NVE criteria are positively oriented with the best value of 1, which are formulated as follows:

$$KGE = 1 - \sqrt{(r-1)^2 + (a-1)^2 + (b-1)^2} \qquad (5\text{-}27)$$

$$NVE = 0.5 \times NSE + 0.25 \times NSES + 0.25 \times NSEL - 0.1 \times VE \qquad (5\text{-}28)$$

where r is the linear correlation coefficient between observed (O_t) and simulated (F_t) flows and a and b respectively denote the ratios of the standard deviation and mean of the simulated flows to the corresponding statistics of the observations. Also, the Volume Error (VE) is defined as:

$$VE = \frac{|\sum_{t=1}^{N} F_t - O_t|}{\sum_{t=1}^{N} O_t} \qquad (5\text{-}27)$$

where N is the length of the calibration period.

5.5 Results and Discussion

5.5.1 Rainfall-runoff models calibration

The best optimal parameter set for each hydrologic model was selected after manually comparing the three estimated sets derived from using various objective functions. Using the final calibrated parameters, Table 5-3 presents the performance of different hydrologic models for the streamflow simulation at the outlet of the Big East River and the Black River watersheds in both calibration and validation periods. In addition, for visual inspection, Figure 5-5 illustrates the hydrographs of the observed and simulated streamflows for the year 2013, as a representative portion of the validation period, in both watersheds. In the Big East River watershed, although MACHBV performance is as accurate as GR4J in the calibration period, all criteria over the validation period (Table 5-3) show the notable superiority of GR4J over the other hydrologic models regarding different aspects of hydrographs (i.e. low and high flows). The comparison of the representative hydrographs (Figure 5-5) also shows the relatively better performance of the GR4J model in capturing high flows. This may be due to the demonstrated ability of the GR4J model to reliably cope with the problem of having poor mean areal precipitation estimates, as the main issue in data-scarce regions. Despite its relative simplicity, the GR4J parameters are sufficiently flexible to compensate for the over and underestimation of the mean areal precipitation (see Andréassian et al., 2001; Drogue & Khediri, 2016). Moreover, besides GR4J, the MACHBV and the SACSMA models performed competitively regarding low flow simulation in the validation period while the other four models (i.e.

SMARG, HEC1, HEC2, and HEC3) tend to over- or under- predict the low flows (Table 5-3 and Figure 5-5).

In the Black River watershed, there is no significant difference between the general performances of different hydrologic models over the validation period (NSE values range from 0.78 to 0.83 compared to 0.58 to 0.78 in the Big East River watershed). However, by focusing on low and high flows separately, the discrepancies between the model performances become more apparent. The $NSEL$ values in both calibration and validation periods suggest that MACHBV and SACSMA are capable of simulating low flows more accurately in the Black River watershed. Regarding high flows, however, performances in calibration and validation are not the same. Both MACHBV and SACSMA provide the highest performance in capturing high flows during the calibration period ($NSES = 0.91$), while, over the validation period, GR4J performs as well as the SAC-SMA model in high flow simulation. Similar to the Big East River watershed, the GR4J can better capture the peak flows (Figure 5-5) in the Black River watershed.

Overall, the GR4J model provides the most reliable results in the Big East River watershed for both calibration and validation periods, however, there are still some streamflow events that are better estimated by relatively lower performing models (Figure 5-5). In the Black River watershed, on the other hand, although MACHBV and SACSMA performed slightly better in the validation period, it is hard to select one of the models as the most robust one regarding different aspects of the hydrograph. Altogether, as it is visually recognizable from Figure 5-5, using an ensemble of multi-model streamflow predictions, compared to

the individual ones, can provide more information about the observation using the properties of various hydrologic models.

Temporal parameter transferability of the different models is compared using the coefficient of transferability (Das et al., 2008), which is calculated as a difference between the NSE values in the calibration and validation periods (i.e. $T = max(NSE_{Cal} - NSE_{Val}, 0)$) where the lower value is better. This criterion shows that the models' performances deteriorate from calibration to validation period and these losses of performance do not occur in the same way in different models and different regions (Table 5-3). For instance, the HEC1 model possesses the highest loss of performance in the Big East River watershed, making it the poorest model in the validation period, while, in Black River, it is among the models with the lowest performance alteration between calibration and validation periods. Therefore, for possessing more robust results and conclusions, the validation period is only used for the application of the aforementioned post-processing approaches (i.e. HUP, BMA, and HUP-BMA) by dividing it into calibration (2012-2014) and forecasting (2014-2015) phases.

Table 5-3 The performances of different calibrated hydrologic models for both calibration (2006-2011) and validation (2012-2015) periods in the Big East River and Black River watersheds

Basin	Period	Criteria	SMARG	GR4J	HEC1	HEC2	HEC3	MACHBV	SACSMA
Big East River	Obj. Func.[1]		KGE	NSE	NVE	NSE	NSE	NVE	KGE
	Calibration	NSE	0.69	0.81	0.67	0.68	0.68	0.81	0.79
		NSEL	0.04	0.78	0.67	0.64	0.62	0.71	0.62
		NSES	0.60	0.67	0.62	0.64	0.63	0.84	0.77
		MAE	5.0	3.5	4.9	4.7	4.6	3.6	4.0
	Validation	NSE	0.69	0.78	0.58	0.67	0.66	0.70	0.71
		NSEL	0.38	0.87	0.54	0.47	0.64	0.78	0.78
		NSES	0.43	0.75	0.35	0.56	0.41	0.48	0.53
		MAE	5.1	3.7	6.4	5.7	4.7	4.7	4.6
	Transferability (T)		0	0.03	0.12	0.01	0.02	0.11	0.08
Black River	Obj. Func.		NSE	KGE	NSE	NSE	NSE	NVE	NVE
	Calibration	NSE	0.87	0.84	0.84	0.88	0.89	0.90	0.90
		NSEL	0.70	0.75	0.03	0.63	0.18	0.83	0.83
		NSES	0.87	0.62	0.75	0.88	0.88	0.91	0.91
		MAE	6.6	6.9	7.3	6.5	6.2	5.8	6.1
	Validation	NSE	0.80	0.78	0.80	0.80	0.81	0.80	0.83
		NSEL	0.72	0.68	0.57	0.49	0.03	0.83	0.81
		NSES	0.62	0.66	0.62	0.58	0.61	0.63	0.67
		MAE	8.0	8.8	7.7	7.6	7.2	7.3	7.0
	Transferability (T)		0.07	0.06	0.04	0.08	0.08	0.10	0.07

1 The objective function, which provides the best optimal parameter set of each rainfall-runoff model.

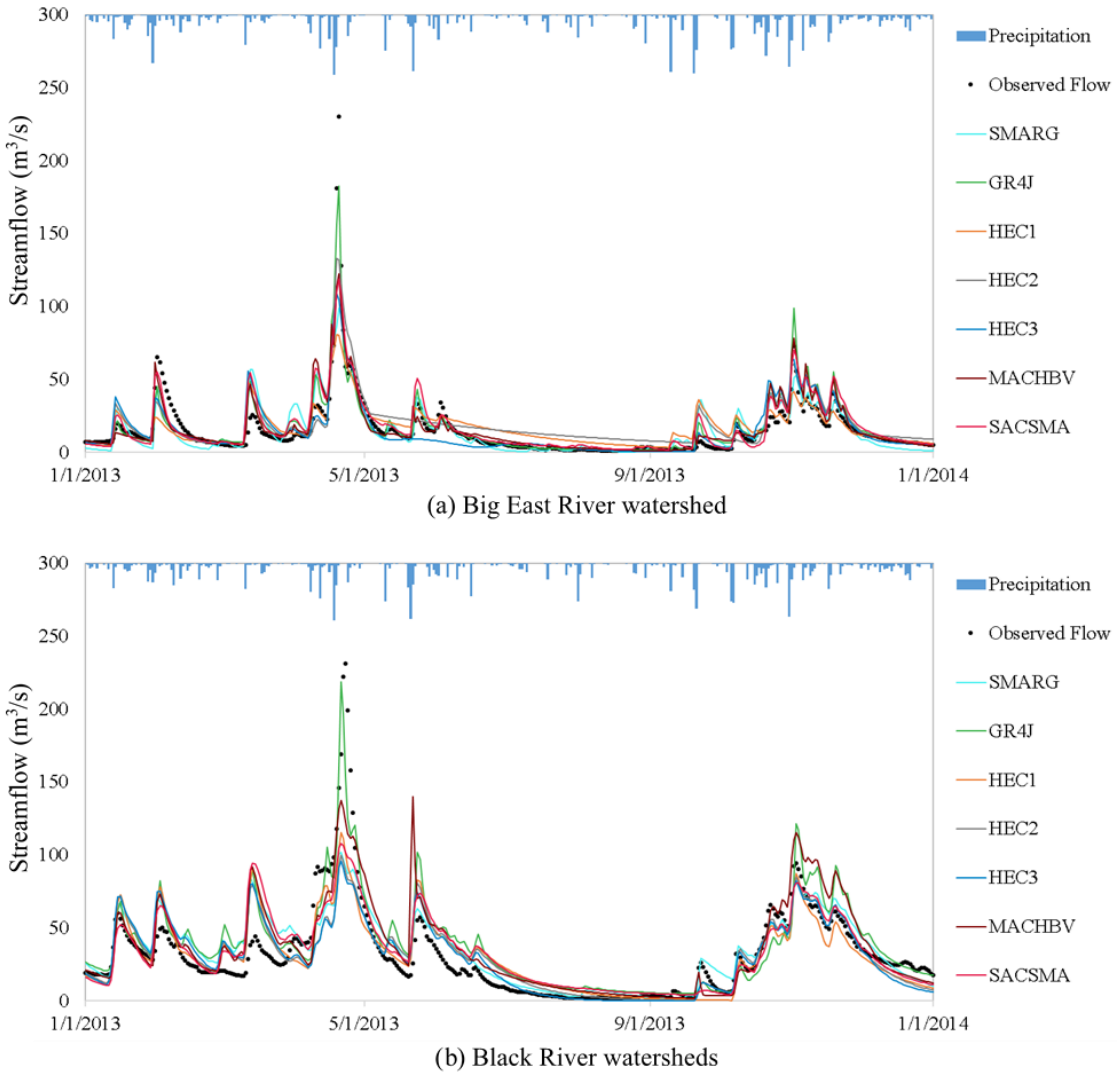


Figure 5-5 Observed and simulated hydrographs of the daily streamflow derived from different hydrologic models for the year 2013 of the validation period in (a) Big East River and (b) Black River watersheds

5.5.2 Calibration of the HUP and HUP-BMA methods

As previously stated, the time-period of 2012 to 2014 was used in order to calibrate different post-processing approaches. As the first step of estimating HUP parameters, after dividing the data based on the precipitation indicator (v), the marginal distributions for

actual (y_{nv}) and forecasted river discharges of different hydrologic models ($\hat{y}_{nv}^{i}$) for each lead-time n, need to be separately estimated. For this purpose, the modified version of the Shapiro-Wilk statistic (Ashkar & Aucoin, 2012) is utilized for testing the goodness of fit of ten various distributions, and the most suitable distribution function is the one with the largest MSW value. MSW test is shown to be a powerful approach for selecting the best distributions in the case of possessing non-Gaussian data with small sizes (Ashkar & Aucoin, 2012; Han et al., 2019). As an example, the best-selected distributions for each sample of data (i.e. actual y_{nv} and forecasted $\hat{y}_{nv}^{i}$ streamflow values) at lead-time $n = 1$ and their corresponding MSW statistics are presented in Table 5-4. It is of note that, due to the presence of no gaps in data and using observation as the perfect precipitation forecasts, the statistics of each variates do not change significantly as a function of lead-time. So, the aforementioned estimation procedure leads to the same best marginal distribution functions with slightly different parameters for different lead-times $n = (2,3, \dots, N)$.

Table 5-4 Sample statistics and their selected prior marginal distributions for lead-time $n = 1$

Basin	Variate	$V = 0$				$V = 1$			
		Mean	s.d[1]	Distribution	MSW[2]	Mean	s.d.	Distribution	MSW
Big East	h_1	11.7	14.3	BS[3]	0.984	14.7	18.7	LL	0.996
	s_1^{SMAR}	12.3	14.6	Weibull	0.995	14.5	16.2	GP	0.994
	s_1^{GR4J}	12.0	12.0	BS	0.974	15.6	17.7	LN	0.994
	s_1^{HEC1}	15.6	13.9	Weibull	0.970	16.5	14.4	BS	0.979
	s_1^{HEC2}	15.4	15.2	GEV	0.990	15.6	15.2	GEV	0.993
	s_1^{HEC3}	11.1	13.8	LN	0.984	13.8	15.2	Kernell	0.984
	s_1^{MAC}	12.4	14.8	Kernel	0.962	15.1	18.0	LN	0.987
	s_1^{SAC}	13.3	15.1	BS	0.992	16.2	16.8	BS	0.996
Big East	h_1	23.6	32.4	Weibull	0.976	29.1	28.4	Gamma	0.984
	s_1^{SMAR}	27.1	24.5	Gamma	0.986	31.0	23.9	Gamma	0.996
	s_1^{GR4J}	26.9	26.1	BS	0.988	30.8	24.8	GEV	0.989
	s_1^{HEC1}	23.9	24.3	GP	0.962	27.3	22.9	GEV	0.981
	s_1^{HEC2}	24.1	24.0	GP	0.964	28.0	23.6	GEV	0.980
	s_1^{HEC3}	22.3	24.7	Kernell	0.937	26.7	24.4	Kernel	0.974
	s_1^{MAC}	25.5	26.5	GP	0.986	29.9	26.9	Gamma	0.988
	s_1^{SAC}	25.9	25.3	Gamma	0.983	29.0	24.8	Gamma	0.994

1 s.d is the abbreviation of standard deviation

2 MSW = Modified Shapiro-Wilk statistic with the perfect value of 1.

3 BS=Binaum-Saunders; LL=Log-Logistic; GP=Generalized Pareto; LN=Log-Normal; GEV=Generalized Extreme Value

After estimating the regression parameters of the prior density (i.e. Equation 5-5) and the likelihood function (i.e. Equation 5-10) in the Gaussian space, the parameters of the HUP posterior distributions are determined using Equations 5-13 to 5-16 for each deterministic forecast based on different hydrologic models (Figure 5-6). The increasing trends in A and decreasing trends in D for both branches of data (i.e. $v = \{0,1\}$) and all hydrologic models, which are more pronounced in the Big East River watershed, show that the initial river discharge (h_0) is less informative and the forecasts are more affected by s_n when lead-time

increases. Parameter *B* is also estimated to be very close to zero in all cases for both watersheds. This is the outcome of the zero estimation of the intercept parameter of the likelihood functions (*b* in Equation 5-10), which is the obvious consequence of transforming data into the standard normal space where the expected value for the intercept is always zero; so, this parameter can be ignored in the HUP formulation in future studies (Vrugt et al., 2008).

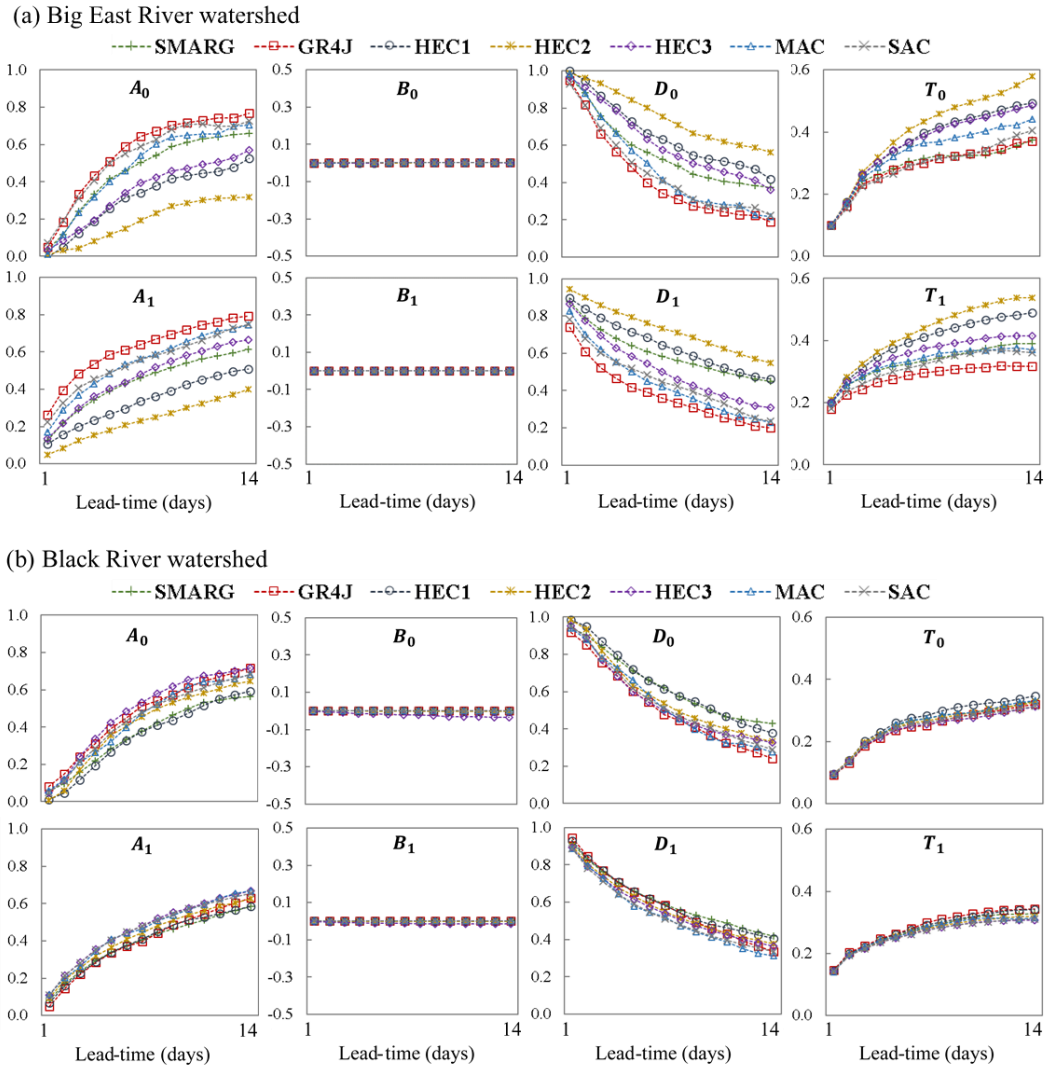


Figure 5-6 The precipitation-dependent HUP posterior distribution parameters with different hydrologic models in both (a) Big East River and (b) Black River watersheds

Using the HUP derived posterior distributions from different hydrologic models in the calibration period (years 2012-2014), the BMA parameters (i.e. weights) were separately estimated for each branch of data based on the modified expectation-maximization algorithm. As can be seen in Figure 5-7, in the Big East River watershed, GR4J possesses the highest weights in almost all cases, especially when precipitation occurs. This can be justifiable by the fact that in the Big East River watershed, the GR4J model noticeably outperformed all other hydrologic models, which is more pronounced regarding high flows (Table 5-3). Besides HEC1 and HEC2, the weights of MACHBV are always zero in Big East River, while it shows relatively good potential in forecasting streamflow compared to the other models. This result may stem from the inherent assumption of the BMA formulation, which is about possessing independent members with the ability to capture observational variability (Darbandsari & Coulibaly, 2019; Lu et al., 2013; Madadgar & Moradkhani, 2014; Refsgaard et al., 2012). A high-performance model may receive a low weight due to its similar error structure with another model in the calibration period, while a member with relatively lower forecasting skill may assign a higher weight by providing unique information to the ensemble.

In the Black River watershed, on the other hand, there is not a specific hydrologic model receiving the highest weights, which is due to the fact that the general performances of all models are relatively good in this basin (Table 5-3). In the case of precipitation occurrence ($v = 1$), MACHBV and HEC3 gained relatively larger weights compared to the other models for almost all the lead-times, while the weights of SACSMA, as one of the best performing model in the Black River watershed, are not significant, again showing the

importance of possessing exclusive information. If no precipitation occurs ($v = 0$), the GR4J weights are the largest one in the shorter lead-times (i.e. $n = \{1,2,3\}$), however, by increasing the lead-time the HEC3 model became the dominant one. Although HEC3 has relatively lower performance, especially regarding low flows (Table 5-3) in the Black River watershed, its high weights stem from the reliable performances of the HUP method in conjunction with HEC3, which will be shown in section 5.5.3. It is worth mentioning that, as will be discussed in the following sections, there are not many differences between the HUP results from different hydrologic models for shorter lead-times, which makes it less important how the BMA weights are distributed among different models.

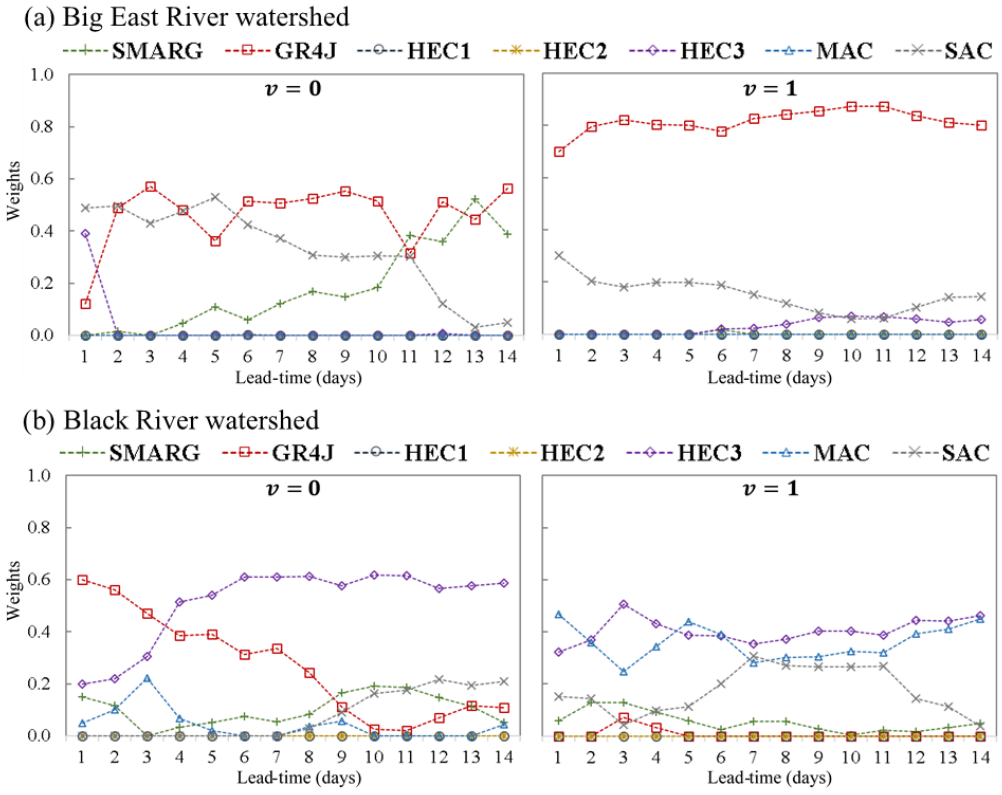


Figure 5-7 The determined HUP-BMA weights of different hydrologic models in (a) Big East River and (b) Black River watersheds (v is the indicator of the precipitation occurrence)

5.5.3 HUP-BMA versus HUP

Both calibrated HUP-BMA and HUP in conjunction with different hydrologic models were employed for probabilistically forecasting daily stream flows up to 14 days ahead during the one-year (2015) verification period in both Big East River and Black River watersheds. Under the assumption of perfect precipitation forecasts, the forecasting here is essentially hindcasting, which focuses on estimating the hydrologic uncertainty (Liu et al., 2018; Han et al., 2019). Using the performance criteria, proposed in Section 5.3.3, Figures 5-8 and 5-9 compared the accuracy of the streamflow forecasts derived from different post-processing approaches deterministically and probabilistically, respectively. Also, the reliability and sharpness of the predictive forecasts are simultaneously compared in Figure 5-10 using the containing ratio ($CR95$) and the average bandwidth ($BW95$) criteria. Additionally, for qualitative comparison, Figure 5-11 shows a representative portion of the observations and probabilistic forecasts derived from the HUP-BMA method and the HUP method in conjunction with SMAR, GR4J, and HEC3 rainfall-runoff models.

In general, the results show that the increasing hydrologic uncertainty for longer lead-times causes the deterioration of the accuracy of the forecasts derived from all post-processing approaches, which is expected; however, the relative deterioration between sequential lead-time steps decreases as lead-time increases (Han et al., 2019). Considering HUP-BMA as an example, the relative differences of MAE and $CRPS$ statistics between lead-times 1 and 2 respectively are 67% and 70% for Big East River and 66% and 53% for Black River, while the same differences between lead-times 10 and 11 are around 2% for both metrics in both watersheds. Additionally, the reliability of the forecasts (evaluated by $CR95$

measurements) do not follow any specific trend with increasing lead-time, while increasing $BW95$ as a function of n shows that all approaches lead to worse probabilistic forecasts with less sharpness in longer lead-times (Figure 5-10).

General comparison of the HUP results, using all performance metrics, shows that the differences between the HUP-derived forecasts based on different hydrologic models are not noticeable in short lead-times. This result, which is confirmed by the sample 1-day ahead forecasted hydrographs illustrated in Figure 5-11, may be due to the fact that the HUP predictions for short forecasting horizons are mostly influenced by the observed actual river discharge (y_0) rather than the forecast derived from the hydrologic model ($\hat{y}_n$). However, by increasing lead-time, the influence of modeled river flow becomes more pronounced, and consequently, selecting a hydrologic model clearly affects the HUP performance (Figures 5-8 to 5-11).

In the Big East River watershed, HUP with the GR4J model has the same forecasts accuracy as the HUP-BMA method (Figure 5-8, 5-9, and 5-11), with similar sharpness and less reliability (Figure 5-10), this is not surprising as the GR4J model was already performing the best (Table 5-3). On the other hand, in the Black River watershed, HEC3 was not the best performing model, however, in comparison with other hydrologic models, its application in conjunction with HUP leads to the most reliable and accurate results for the different lead-times (Figure 5-8 to 5-11). It is worthy of note that although HEC3 performed relatively poor regarding low flows (Table 5-3), the HUP-HEC3 possesses the highest $NSEL$ values, compared with other HUP-based results. Also, as can be seen in Figure 5-11, the HUP method leads to almost similar base flow forecasts regardless of the

implemented hydrologic models. In line with Han et al., (2019), this result confirms the capability of the HUP structure in generating reliable low flow forecasts while the modeled deterministic results are not relatively good. On the other hand, however, the implemented hydrologic model noticeably affects the performances of the HUP method regarding high flow predictions (Figure 5-11). These effects are more pronounced for longer lead-times when the initial condition of flow is less effective. Besides the quality of the hydrologic models, the inefficiency of conditioning the prior density and the likelihood function to the initial river discharge, as another reason for the poor performances of HUP-based post-processing methods for high flow forecasting at longer lead-times in the Big East River watershed, will be discussed in the following section. Altogether, in both watersheds, the HUP-BMA method always performed as well as or better than the best HUP and hydrologic model combination regarding the accuracy, reliability, and sharpness of the probabilistic streamflow forecasts. This demonstrates the robust ability of HUP-BMA to use all beneficial information derived from HUP based on different models for better quantifying hydrologic uncertainty and producing enhanced probabilistic streamflow forecasts.

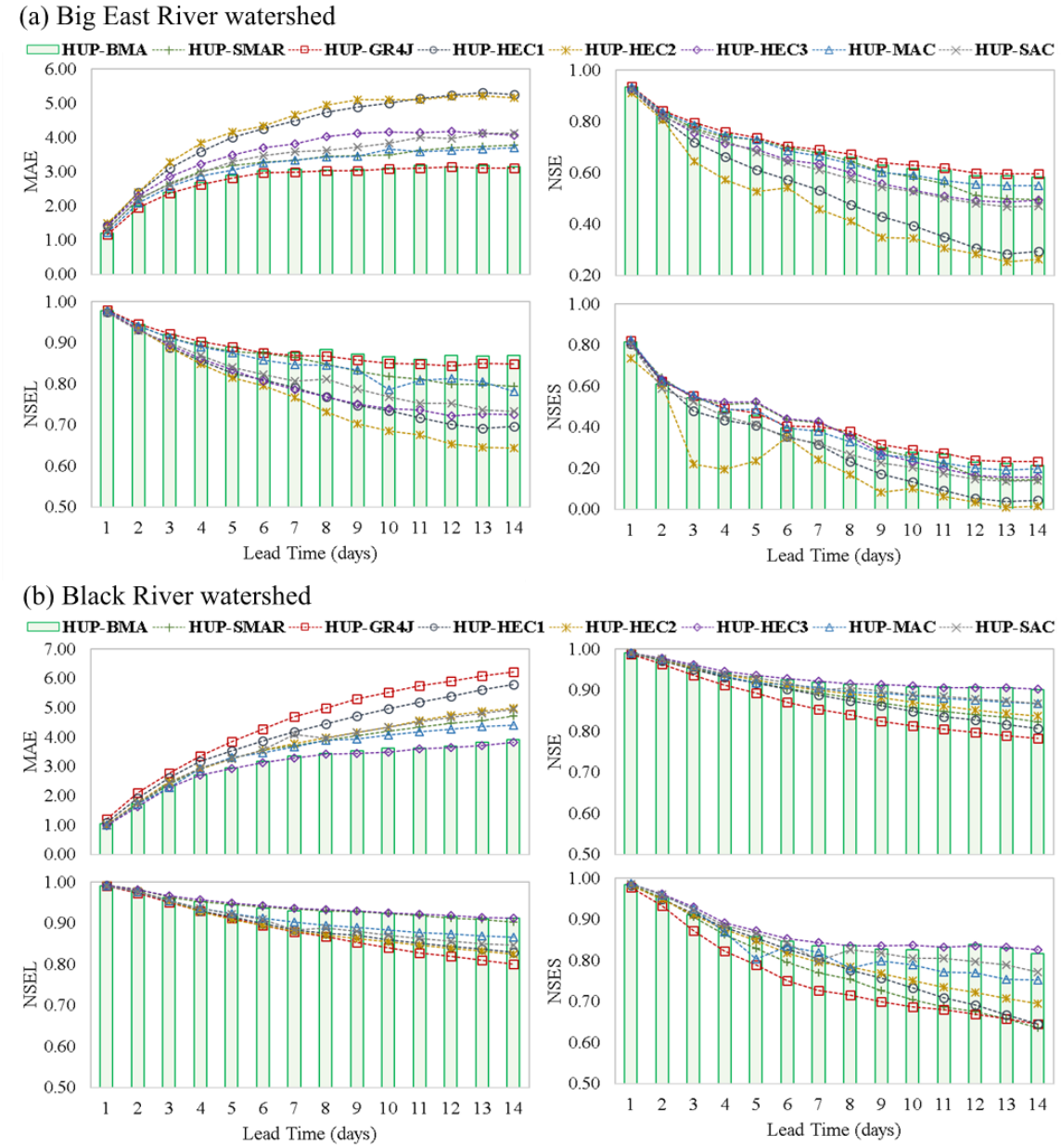


Figure 5-8 Deterministic performances of the proposed HUP-BMA compared with HUP with different hydrologic models using various criteria (i.e. MAE, NSE, NSEL, NSES) for short- to mid-range streamflow forecasts (1 to 14 days ahead) in (a) Big East River and (b) Black River watersheds

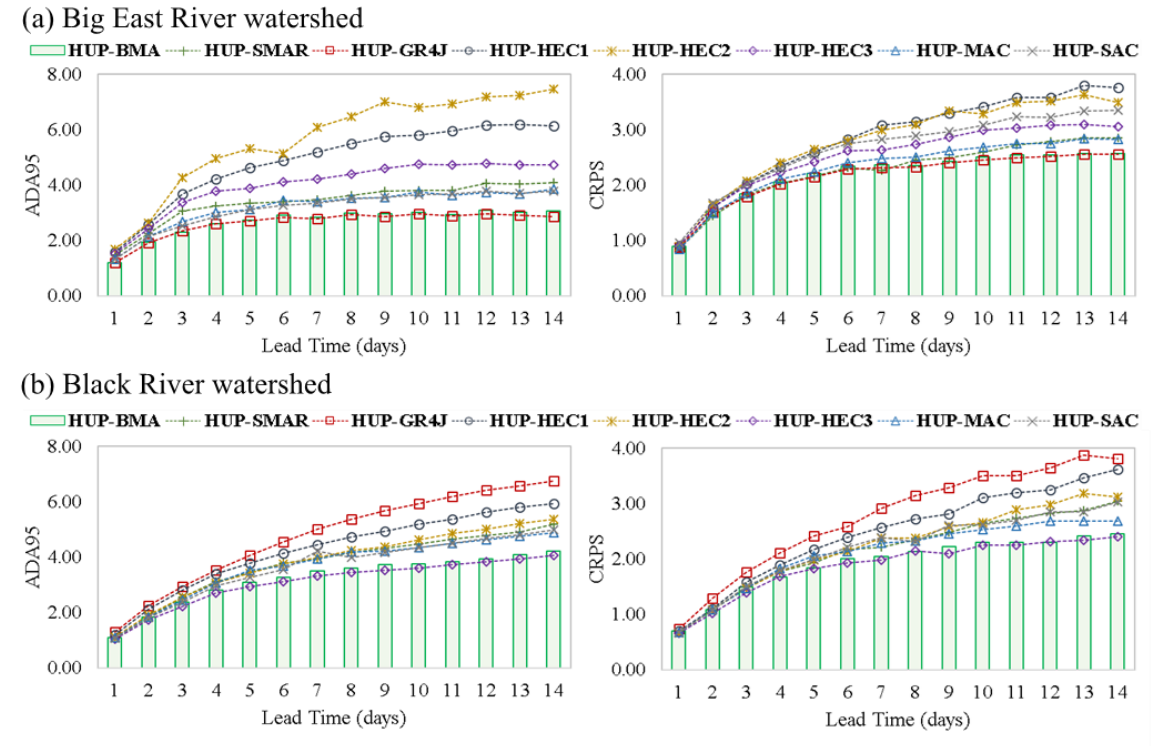


Figure 5-9 Comparison of the accuracy of probabilistic forecasts derived from the proposed HUP-BMA and HUP based on different hydrologic models using two performance metrics (i.e. ADA95 and CRPS) for 1to 14 days ahead in (a) Big East River and (b) Black River watersheds

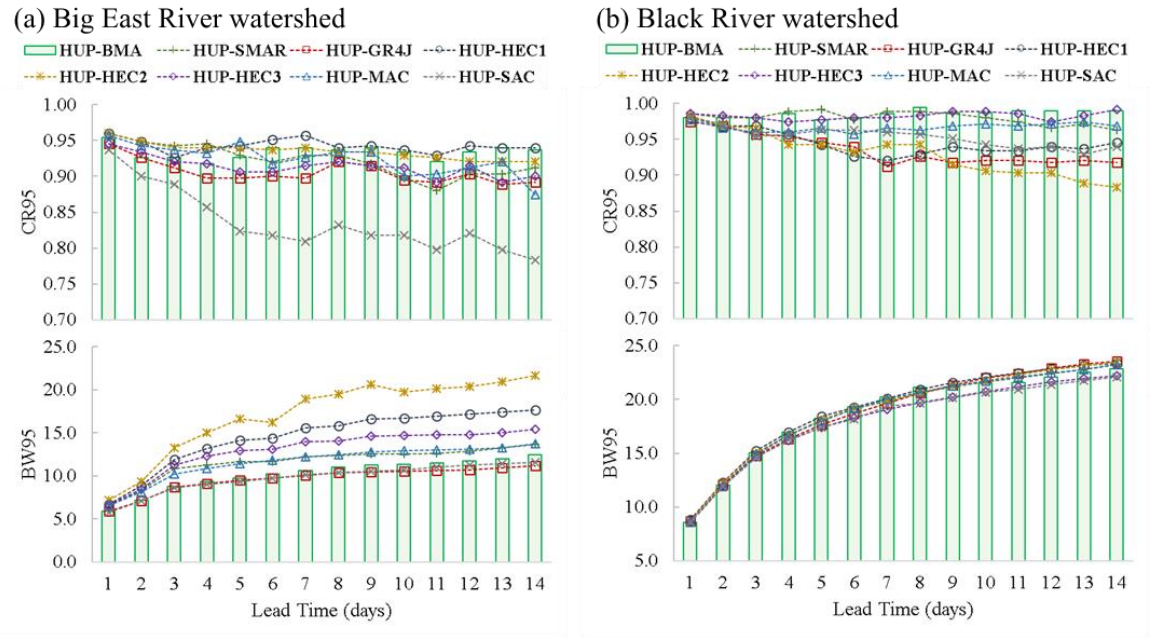


Figure 5-10 Comparing the reliability (CR95) and sharpness (BW95) of the proposed HUP-BMA and HUP with different hydrologic models for 1- to 14- days ahead probabilistic streamflow forecasts in (a) Big East River and (b) Black River watersheds

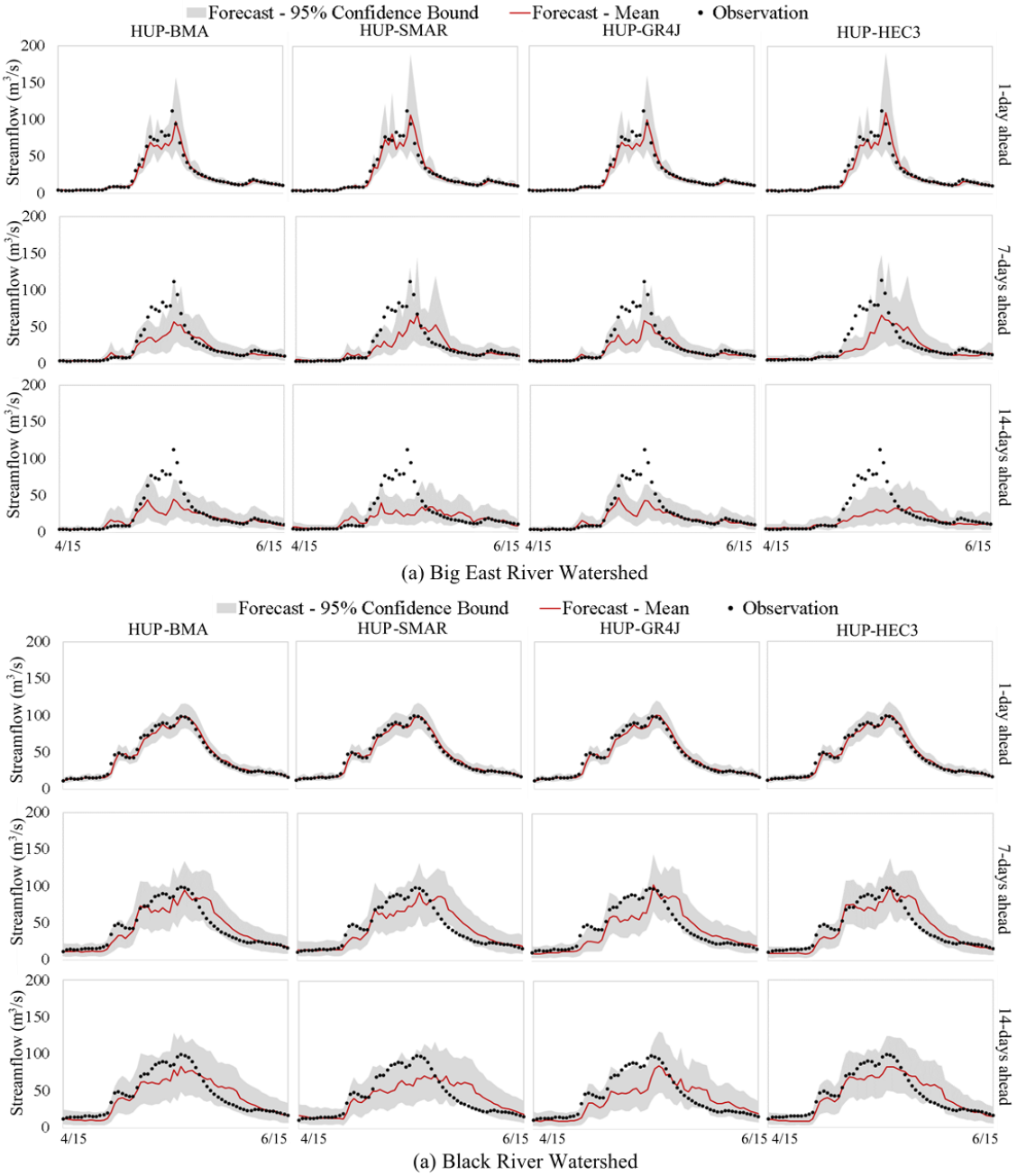


Figure 5-11 Time-series of the mean and 95% confidence bounds of 1-, 7-, and 14-days ahead streamflow forecasts derived from HUP-BMA, and HUP in conjunction with, SMAR, GR4J, and HEC3 hydrologic models, compared with the observation from a representative part of the verification period in (a) Big East River and (b)Black River watersheds

5.5.4 The modified HUP-BMA unconditioned on initial observation

For assessing the effects of using the initial condition on the proposed HUP-BMA formulation, the performance of the HUP-BMA method is compared with the BMA approach, which is not explicitly taking into account the initial observation. Figure 5-12 illustrates the percentage of performance improvement, defined as the percent enhancement in different evaluation metrics when HUP-BMA is used as a post-processing approach, compared to BMA, in both Big East River and Black River watersheds. In general, except for the $CR95$ criterion, the percent improvements based on all performance metrics show a decreasing trend by increasing lead-time in both watersheds, confirming the previously aforementioned fact that the performance of HUP-BMA deteriorates for longer lead-times. In the Big East River watershed, comparing the accuracy of the forecasts, using all deterministic (i.e. NSE, $NSEL$, $NSES$, and MAE) and probabilistic (i.e. $CRPS$ and $ADA95$) measures shows that although HUP-BMA outperformed BMA in shorter lead-times, the proposed HUP-BMA provides relatively worse results for n larger than 6. In terms of reliability and sharpness, HUP-BMA provides sharper results with the same reliability for short lead-times, however, increasing lead-time leads to the same probabilistic performance of BMA and HUP-BMA. In the Black River watershed, the same trend as Big East River is observed where the differences between the performances of HUP-BMA and BMA are decreased by increasing lead-times in both deterministic and probabilistic manners, however, the advantage of HUP-BMA over BMA still presents for longer lead-times.

The observed performance deterioration of HUP-BMA, compared with BMA, by increasing lead-time probably comes from the main structural difference between these two approaches, which is taking into account the initially observed discharge in the HUP-BMA formulation, and as previously stated, this consideration does not positively affect the forecasts for longer lead-times. Comparing the correlation coefficient of the initial river discharge (i.e. y_0) with the actual river discharge at lead-time n (i.e. $y_n;\ n = \{1, \dots, N\}$) confirms that increasing lead-time causes lower dependence between these two variables. This deterioration is more noticeable in the Big East River watershed where the correlation coefficient is less than 0.5 for n larger than six. However, the larger area and the slighter slope of the Black River watershed (Table 5-1), leads to a higher correlation among flows (with the lowest value of 0.51 for $n = 14$). This result can justify the relatively worse performances of HUP-BMA, compared with BMA, in Big East River for streamflow forecasting at longer lead-times.

To solve this issue in the Big East River watershed, inspired from Reggiani et al. (2009), the modified HUP-BMA is proposed based on the unconditioned HUP procedure where both prior density and likelihood function are no longer conditioned on y_0 for $n > 6$ (i.e. parameters c_n in Equation 5-5 and d_n in Equation 5-10 are zero for both branches (i.e. $v = \{0,1\}$) at $n > 6$). As expected, a comparison between the performances of HUP-BMA and the modified HUP-BMA in the Big East River watershed for 7 to 14 days ahead (Figure 5-13) shows that the accuracy of streamflow forecasts significantly enhance by using the modified HUP-BMA, especially regarding high flows in longer lead-times (e.g. percent improvement based on NSE and $NSES$ at $n = 14$ are more than 20 and 60 percent,

respectively). Regarding reliability and sharpness, although the modified version leads to forecasts with slightly higher reliability (better CR95), its higher average bandwidth shows the lower sharpness in the case of implementing unconditioned HUP. Moreover, comparisons between the modified HUP-BMA and the BMA method using different performance metrics (Figure 5-13) indicate the same overall behaviors of both methods. Except for $ADA95$, indicating the superiority of the modified HUP-BMA method over BMA, and $CRPS$, which shows the slight advantage of using BMA, all other measurements indicate negligible difference between the performances of the two methods.

Finally, as a visual inspection, Figure 5-14 illustrates a representative part of the observed and 14-day ahead forecasted streamflow hydrographs, derived from the HUP-BMA, the modified HUP-BMA, and the BMA methods. In line with previous conclusions, the plots indicate the superiority of the modified HUP-BMA method in producing better probabilistic and deterministic high flow forecasts over the original HUP-BMA. Figure 5-14 is also indicated the almost same performance of the modified HUP-BMA and the BMA approaches in quantifying hydrologic uncertainty. Although, there are still some differences between the formulations of the modified HUP-BMA and the BMA methods, the comparable performance of both approaches is expected as the overall structure of both methods is somewhat similar after removing the effects of the initial condition in the modified HUP-BMA approach. It is worthy of note that the lead-time when the modified version of the HUP-BMA is beneficial is a function of the watershed characteristics (e.g. the characteristics of the hydrographs) and should be separately assessed and specified for each study.

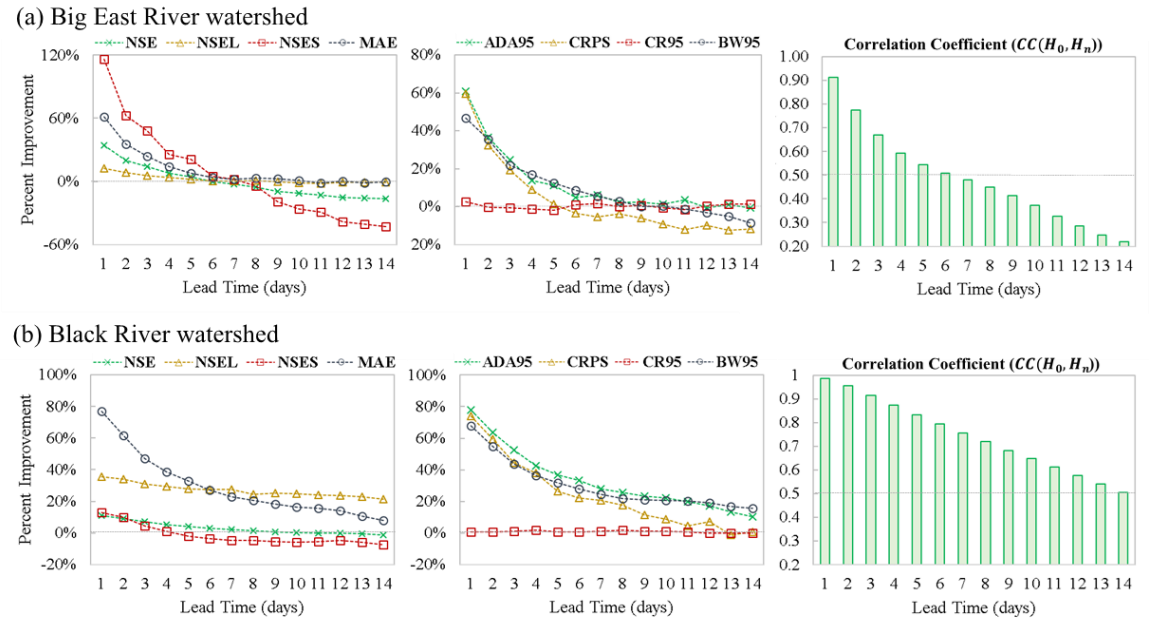


Figure 5-12 Comparing the performance of HUP-BMA and BMA using the percent improvement of different criteria and the correlation between the actual flow H_0 and H_n $\forall$ $n = \{1,2,\dots,14\}$ in (a) Big East River and (b) Black River watersheds. The percent improvement is defined as the percentage of improvement when using HUP-BMA instead of BMA, with positive values indicating the advantage of using HUP-BMA

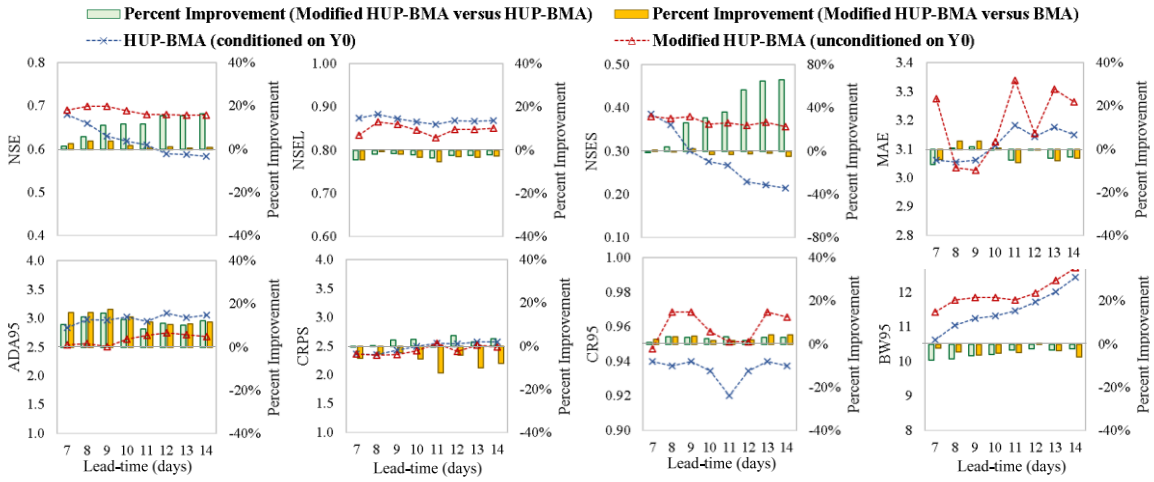


Figure 5-13 Comparison of different performance metrics for 7 to 14 days-ahead streamflow forecasting derived from HUP-BMA, modified HUP-BMA, and BMA in Big East River watershed. Percent improvement is defined as the percentage of improvement when using modified HUP-BMA instead of HUP-BMA or BMA, with positive values indicating the advantages of using modified HUP-BMA

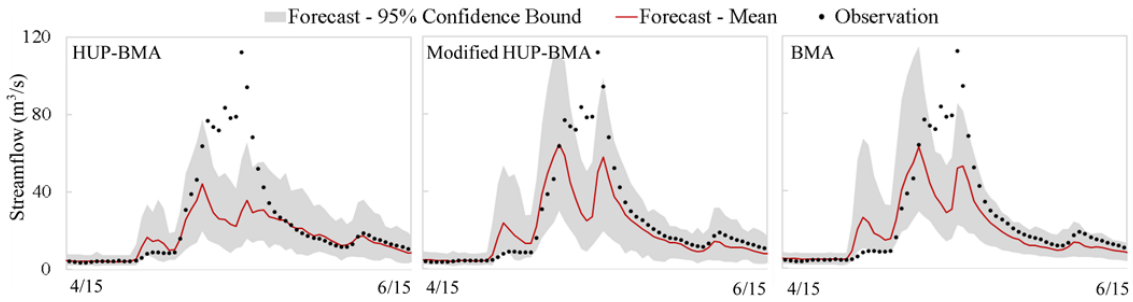


Figure 5-14 Time-series of the mean and 95% confidence bounds of 14-days ahead streamflow forecasts derived from HUP-BMA, modified HUP-BMA, and BMA, compared with observations, from a representative portion of the verification period in the Big East River watershed

5.6 Summary and Conclusion

Considering hydrologic uncertainty is one of the main steps for generating reliable probabilistic streamflow forecasts. The Hydrologic Uncertainty Processor (HUP) is a well-known approach for quantifying hydrologic uncertainty by providing a posterior distribution conditioned on the initial observation and a deterministic model forecast. However, various structurally different hydrologic models can be used in conjunction with the HUP and significantly affect its capability of providing reliable and accurate probabilistic forecasts. Considering seven conceptual hydrologic models with different structures, this study assessed the effects of implementing different models combined with the HUP on streamflow forecasting. Also, a new multi-model Bayesian processor (HUP-BMA) is proposed by using the concept of the Bayesian Model Averaging (BMA) approach, where the HUP-derived posterior distributions based on various models are combined in order to generate the final predictive forecasts. All post-processing approaches are utilized for daily streamflow forecasting up to 14 days ahead in two study regions in

Northern Ontario, Canada, and their performances are compared using various deterministic and probabilistic verification metrics. The summary of the most important conclusions based on the detailed analysis of the results are as follows:

- In general, although the HUP was proven as a robust approach for quantifying hydrologic uncertainty, its performance is influenced by the hydrologic model selected and used for generating the deterministic forecast. To some extent, the HUP method can compensate for the low quality of the deterministic forecast, however, this capability will reduce by increasing lead-time and flow values, so the predictive results are more affected by selecting the most promising hydrologic model for longer lead-times and high flows.
- Comparing the application of HUP-BMA and the HUP based on various hydrologic models in both watersheds indicated that the proposed HUP-BMA method compensates for the dependence of the HUP results on the quality of the hydrologic model. Using the advantages of multiple models, HUP-BMA always provides better or similar deterministic and probabilistic forecasts, compared with the HUP combined with different models.
- For shorter lead-times, considering the effects of initially observed discharge in HUP-BMA formulation leads to a better quantification of hydrologic uncertainty, compared with the original BMA, where the only information used is a multi-model ensemble of deterministic forecasts. However, by increasing lead-time, this superiority is reduced and conditioning the formulations on initial river discharge becomes less beneficial and can lead to deteriorating predictions. So, considering

unconditioned prior density and likelihood function in the HUP as well as in the HUP-BMA formulation for longer lead-times is showing as a potential approach for better quantifying hydrologic uncertainty and producing the most accurate probabilistic forecasts.

In general, the findings of this study suggest that while retaining all the advantages of the HUP method, the proposed HUP-BMA approach addresses the effects of selecting a single hydrologic model by using the information derived from a multi-model ensemble of streamflow forecasts to better quantify hydrologic uncertainty. Also, by explicitly considering the effects of initial observations, the HUP-BMA method, compared with BMA, better estimates the hydrologic uncertainty in short-range streamflow forecasts. However, for longer lead-times, the unconditioned revision of HUP-BMA formulation may lead to more accurate results. Although the hydrologic responses of both poorly-gauged watersheds are quite different, evaluation of the proposed multi-model Bayesian processor in regions with different climatologic and topographic conditions needs to be carried out in future studies. Moreover, additional verification of the proposed method for case studies with longer time-series of data and higher temporal resolution is advisable for further research to better understanding the processor. The proposed HUP-BMA method includes a segregated calibration of the HUP parameters and weights while developing and evaluating an integrated calibration structure, where all HUP-BMA parameters are estimated simultaneously, could lead to better results and worth being investigated in the future studies. Moreover, this study was designed to relatively assess the performances of the proposed HUP-BMA for estimating hydrologic uncertainty, however, its evaluation in

conjunction with the probabilistic precipitation forecasts (or ensemble of precipitation forecasts) for quantifying total predictive uncertainty requires further studies. In addition, providing criteria as a function of characteristics of the observed hydrograph, for systematically determining the specific lead-time after which the unconditioned version of the processor is more beneficial, requires further studies.

5.7 Acknowledgments

We acknowledge that this work was supported by the Natural Science and Engineering Research Council (NSERC) of Canada, grant NSERC Canadian FloodNet (NETGP-451456). The authors would like to thank James Leach for his support with reviewing and editing the manuscript, and Shasha Han for providing Matlab source code for the HUP processor. Historical daily streamflow observation at the outlet of both watersheds can be obtained from Water Survey of Canada (https://wateroffice.ec.gc.ca/mainmenu/historical_data_index_e.html). Daily temperature data for the selected stations is provided by Environment and Climate Change Canada (https://climate.weather.gc.ca/historical_data/search_historic_data_e.html). The Canadian Precipitation Analysis (CaPA) data can be obtained from the Canadian Surface Prediction Archive (CasPAr) website (https://caspar-data.ca/).

5.8 References

Ajami, N. K., Duan, Q., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004745

Anderson, E. A. (2006). *Snow Accumulation and Ablation Model – SNOW-17*. Natl. Ocean. Atmospheric Adm. Natl. Weather Serv. Silver Springs MD. https://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/part2/_pdf/22snow17.pdf

Andréassian, V., Perrin, C., Michel, C., Usart-Sanchez, I., & Lavabre, J. (2001). Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models. *Journal of Hydrology*, *250*(1), 206–223. https://doi.org/10.1016/S0022-1694(01)00437-1

Anshuman, A., Kunnath-Poovakka, A., & Eldho, T. I. (2019). Towards the use of conceptual models for water resource assessment in Indian tropical watersheds under monsoon-driven climatic conditions. *Environmental Earth Sciences*, *78*(9), 282. https://doi.org/10.1007/s12665-019-8281-5

Ashkar, F., & Aucoin, F. (2012). Choice between competitive pairs of frequency models for use in hydrology: A review and some new results. *Hydrological Sciences Journal*, *57*(6), 1092–1106. https://doi.org/10.1080/02626667.2012.701746

Bates, J. M., & Granger, C. W. J. (1969). The Combination of Forecasts. *OR*, *20*(4), 451–468. JSTOR. https://doi.org/10.2307/3008764

Bergström, S. (1976). Development and Application of a Conceptual Runoff Model for Scandinavian Catchments. *Lund,Sweden: Lund Institute of Technology/Univ. of Lund, A*, *52*. https://www.researchgate.net/publication/255274162_Development_and_Application_of_a_Conceptual_Runoff_Model_for_Scandinavian_Catchments

Biondi, D., & Todini, E. (2018). Comparing Hydrological Postprocessors Including Ensemble Predictions Into Full Predictive Probability Distribution of Streamflow. *Water Resources Research*, *54*(12), 9860–9882. https://doi.org/10.1029/2017WR022432

Biondi, D., Versace, P., & Sirangelo, B. (2010). Uncertainty assessment through a precipitation dependent hydrologic uncertainty processor: An application to a small

catchment in southern Italy. *Journal of Hydrology*, *386*(1), 38–54. https://doi.org/10.1016/j.jhydrol.2010.03.004

Boluwade, A., Zhao, K.-Y., Stadnyk, T. A., & Rasmussen, P. (2018). Towards validation of the Canadian precipitation analysis (CaPA) for hydrologic modeling applications in the Canadian Prairies. *Journal of Hydrology*, *556*, 1244–1255. https://doi.org/10.1016/j.jhydrol.2017.05.059

Bravo, J. M., Paz, A. R., Collischonn, W., Uvo, C. B., Pedrollo, O. C., & Chou, S. C. (2009). Incorporating Forecasts of Rainfall in Two Hydrologic Models Used for Medium-Range Streamflow Forecasting. *Journal of Hydrologic Engineering*, 14(5), 435–445. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000014

Burnash, R. J. C., Ferral, R. L., & McGuire, R. A. (1973). *A generalized streamflow simulation system: Conceptual modeling for digital computers*. Joint Federal-State River Forecast Center, United States National Weather Service.

Darbandsari, P., & Coulibaly, P. (2019). Inter-Comparison of Different Bayesian Model Averaging Modifications in Streamflow Simulation. *Water*, *11*(8), 1707. https://doi.org/10.3390/w11081707

Darbandsari, P., & Coulibaly, P. (2020a). Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: Case study of Northern Ontario, Canada. *Journal of Hydrology: Regional Studies*, *31*, 100730. https://doi.org/10.1016/j.ejrh.2020.100730

Darbandsari, P., & Coulibaly, P. (2020b). Introducing entropy-based Bayesian model averaging for streamflow forecast. *Journal of Hydrology*, *591*, 125577. https://doi.org/10.1016/j.jhydrol.2020.125577

Das, T., Bárdossy, A., Zehe, E., & He, Y. (2008). Comparison of conceptual model performance using different representations of spatial variability. *Journal of Hydrology*, *356*(1), 106–118. https://doi.org/10.1016/j.jhydrol.2008.04.008

DeGroot, M. H. (2005). *Optimal Statistical Decisions*. John Wiley & Sons.

Dong, L., Xiong, L., & Zheng, Y. (2013). Uncertainty analysis of coupling multiple hydrologic models and multiple objective functions in Han River, China. *Water Science and Technology*, *68*(3), 506–513. https://doi.org/10.2166/wst.2013.255

Drogue, G., & Khediri, W. B. (2016). Catchment model regionalization approach based on spatial proximity: Does a neighbor catchment-based rainfall input strengthen the method? *Journal of Hydrology: Regional Studies*, *8*, 26–42. https://doi.org/10.1016/j.ejrh.2016.07.002

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Elshall, A. S., Ye, M., Pei, Y., Zhang, F., Niu, G.-Y., & Barron-Gafford, G. A. (2018). Relative model score: A scoring rule for evaluating ensemble simulations with application to microbial soil respiration modeling. *Stochastic Environmental Research and Risk Assessment*, *32*(10), 2809–2819. https://doi.org/10.1007/s00477-018-1592-3

Eum, H.-I., Dibike, Y., Prowse, T., & Bonsal, B. (2014). Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the Athabasca Watershed, Canada. *Hydrological Processes*, *28*(14), 4250–4271. https://doi.org/10.1002/hyp.10236

Gan, T. Y., Dlamini, E. M., & Biftu, G. F. (1997). Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. *Journal of Hydrology*, 192(1), 81–103. https://doi.org/10.1016/S0022-1694(96)03114-9

Granger, C. W., & Ramanathan, R. (1984). Improved Methods of Combining Forecasts: ABSTRACT. *Journal of Forecasting (Pre-1986); Chichester*, *3*(2), 197–204.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving

hydrological modelling. *Journal of Hydrology*, *377*(1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Han, S., & Coulibaly, P. (2017). Bayesian flood forecasting methods: A review. *Journal of Hydrology*, *551*, 340–351. https://doi.org/10.1016/j.jhydrol.2017.06.004

Han, S., Coulibaly, P., & Biondi, D. (2019). Assessing Hydrologic Uncertainty Processor Performance for Flood Forecasting in a Semiurban Watershed. *Journal of Hydrologic Engineering*, *24*(9), 05019025. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001828

Hargreaves, G. H., & Samani, Z. A. (1985). Reference Crop Evapotranspiration from Temperature. *Applied Engineering in Agriculture*, *1*(2), 96–99.

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Huo, W., Li, Z., Wang, J., Yao, C., Zhang, K., & Huang, Y. (2019). Multiple hydrological models comparison and an improved Bayesian model averaging approach for ensemble prediction over semi-humid regions. *Stochastic Environmental Research and Risk Assessment*, *33*(1), 217–238. https://doi.org/10.1007/s00477-018-1600-7

Jiang, S., Ren, L., Xu, C.-Y., Liu, S., Yuan, F., & Yang, X. (2018). Quantifying multi-source uncertainties in multi-model predictions using the Bayesian model averaging scheme. *Hydrology Research*, *49*(3), 954–970. https://doi.org/10.2166/nh.2017.272

Kelly, K. S., & Krzysztofowicz, R. (1997). A bivariate meta-Gaussian density for use in hydrology. *Stochastic Hydrology and Hydraulics*, *11*(1), 17–31. https://doi.org/10.1007/BF02428423

Kelly, K. S., & Krzysztofowicz, R. (2000). Precipitation uncertainty processor for probabilistic river stage forecasting. *Water Resources Research*, *36*(9), 2643–2653. https://doi.org/10.1029/2000WR900061

Kisi, O., & Cimen, M. (2011). A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *Journal of Hydrology*, *399*(1), 132–140. https://doi.org/10.1016/j.jhydrol.2010.12.041

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

Krzysztofowicz, R. (1997). Transformation and normalization of variates with specified distributions. *Journal of Hydrology*, *197*(1), 286–292. https://doi.org/10.1016/S0022-1694(96)03276-3

Krzysztofowicz, R. (1999). Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, *35*(9), 2739–2750. https://doi.org/10.1029/1999WR900099

Krzysztofowicz, R. (2002). Bayesian system for probabilistic river stage forecasting. *Journal of Hydrology*, *268*(1), 16–40. https://doi.org/10.1016/S0022-1694(02)00106-3

Krzysztofowicz, R., & Herr, H. D. (2001). Hydrologic uncertainty processor for probabilistic river stage forecasting: Precipitation-dependent model. *Journal of Hydrology*, *249*(1), 46–68. https://doi.org/10.1016/S0022-1694(01)00412-7

Krzysztofowicz, R., & Kelly, K. S. (2000). Hydrologic uncertainty processor for probabilistic river stage forecasting. *Water Resources Research*, *36*(11), 3265–3277. https://doi.org/10.1029/2000WR900108

Lespinas, F., Fortin, V., Roy, G., Rasmussen, P., & Stadnyk, T. (2015). Performance Evaluation of the Canadian Precipitation Analysis (CaPA). *Journal of Hydrometeorology*, *16*(5), 2045–2064. https://doi.org/10.1175/JHM-D-14-0191.1

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., & Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *WIREs Water*, *4*(6), e1246. https://doi.org/10.1002/wat2.1246

Liang, Z., Wang, D., Guo, Y., Zhang, Y., & Dai, R. (2013). Application of Bayesian Model Averaging Approach to Multimodel Ensemble Hydrologic Forecasting. *Journal of Hydrologic Engineering*, *18*(11), 1426–1436. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000493

Liu, Z., Guo, S., Xiong, L., & Xu, C.-Y. (2018). Hydrological uncertainty processor based on a copula function. *Hydrological Sciences Journal*, *63*(1), 74–86. https://doi.org/10.1080/02626667.2017.1410278

Liu, Z., Guo, S., Zhang, H., Liu, D., & Yang, G. (2016). Comparative Study of Three Updating Procedures for Real-Time Flood Forecasting. *Water Resources Management*, *30*(7), 2111–2126. https://doi.org/10.1007/s11269-016-1275-0

Lu, D., Ye, M., Meyer, P. D., Curtis, G. P., Shi, X., Niu, X.-F., & Yabusaki, S. B. (2013). Effects of error covariance structure on estimation of model averaging weights and predictive performance. *Water Resources Research*, 49(9), 6029–6047. https://doi.org/10.1002/wrcr.20441

Madadgar, S., & Moradkhani, H. (2014). Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resources Research*, *50*(12), 9586–9603. https://doi.org/10.1002/2014WR015965

Mahfouf, J.-F., Brasnett, B., & Gagnon, S. (2007). A Canadian precipitation analysis (CaPA) project: Description and preliminary results. *Atmosphere-Ocean*, *45*(1), 1–17. https://doi.org/10.3137/ao.v450101

McLachlan, G., & Krishnan, T. (2008). *The EM Algorithm and Extensions* (2 edition). Wiley-Interscience.

Montanari, A., & Brath, A. (2004). A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, *40*(1). https://doi.org/10.1029/2003WR002540

Montanari, A., Shoemaker, C. A., & Giesen, N. van de. (2009). Introduction to special section on Uncertainty Assessment in Surface and Subsurface Hydrology: An

overview of issues and challenges. *Water Resources Research*, *45*(12). https://doi.org/10.1029/2009WR008471

Muhammad, A., Stadnyk, T. A., Unduche, F., & Coulibaly, P. (2018). Multi-Model Approaches for Improving Seasonal Ensemble Streamflow Prediction Scheme with Various Statistical Post-Processing Techniques in the Canadian Prairie Region. *Water*, *10*(11), 1604. https://doi.org/10.3390/w10111604

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

O'Connell, P. E., Nash, J. E., & Farrell, J. P. (1970). River flow forecasting through conceptual models part II - The Brosna catchment at Ferbane. *Journal of Hydrology*, *10*(4), 317–329. https://doi.org/10.1016/0022-1694(70)90221-0

Ouermi, K. S., Paturel, J.-E., Adounpke, J., Lawin, A. E., Goula, B. T. A., & Amoussou, E. (2019). Comparison of hydrological models for use in climate change studies: A test on 241 catchments in West and Central Africa. *Comptes Rendus Geoscience*, *351*(7), 477–486. https://doi.org/10.1016/j.crte.2019.08.001

Parrish, M. A., Moradkhani, H., & DeChant, C. M. (2012). Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation: TOWARD REDUCTION OF MODEL UNCERTAINTY. *Water Resources Research*, *48*(3). https://doi.org/10.1029/2011WR011116

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*(1), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Poeter, E. P., & Hill, M. C. (2007). *MMA, A Computer Code for Multi-Model Analysis* (TM 6-E3). United States Geological Survey - Nevada, Henderson, Nevada. https://doi.org/10.2172/920086

Price, K., Purucker, S. T., Kraemer, S. R., Babendreier, J. E., & Knightes, C. D. (2014). Comparison of radar and gauge precipitation data in watershed models across varying spatial and temporal scales. *Hydrological Processes*, *28*(9), 3505–3520. https://doi.org/10.1002/hyp.9890

Qu, B., Zhang, X., Pappenberger, F., Zhang, T., & Fang, Y. (2017). Multi-Model Grand Ensemble Hydrologic Forecasting in the Fu River Basin Using Bayesian Model Averaging. *Water*, *9*(2), 74. https://doi.org/10.3390/w9020074

Raftery, A. E. (1993). Bayesian Model Selection in Structural Equation Models. In *Testing Structural Equation Models* (Vol. 154, pp. 163–180). SAGE.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Razavi, T., & Coulibaly, P. (2017). An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds. *Canadian Water Resources Journal / Revue Canadienne Des Ressources Hydriques*, *42*(1), 2–20. https://doi.org/10.1080/07011784.2016.1184590

Refsgaard, J. C., Christensen, S., Sonnenborg, T. O., Seifert, D., Højberg, A. L., & Troldborg, L. (2012). Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Advances in Water Resources*, *36*, 36–50. https://doi.org/10.1016/j.advwatres.2011.04.006

Refsgaard, J. C., & Knudsen, J. (1996). Operational Validation and Intercomparison of Different Types of Hydrological Models. *Water Resources Research*, *32*(7), 2189–2202. https://doi.org/10.1029/96WR00896

Reggiani, P., Renner, M., Weerts, A. H., & Gelder, P. A. H. J. M. van. (2009). Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. *Water Resources Research*, *45*(2). https://doi.org/10.1029/2007WR006758

Reggiani, P., & Weerts, A. H. (2008). A Bayesian approach to decision-making under uncertainty: An application to real-time forecasting in the river Rhine. *Journal of Hydrology*, *356*(1), 56–69. https://doi.org/10.1016/j.jhydrol.2008.03.027

Rojas, R., Feyen, L., & Dassargues, A. (2008). Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2008WR006908

Roy, T., Serrat-Capdevila, A., Gupta, H., & Valdes, J. (2017). A platform for probabilistic Multimodel and Multiproduct Streamflow Forecasting. *Water Resources Research*, *53*(1), 376–399. https://doi.org/10.1002/2016WR019752

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2011). Estimation of Continuous Streamflow in Ontario Ungauged Basins: Comparison of Regionalization Methods. *Journal of Hydrologic Engineering*, *16*(5), 447–459. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000338

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2012). Identification of rainfall–runoff model for improved baseflow estimation in ungauged basins. *Hydrological Processes*, *26*(3), 356–366. https://doi.org/10.1002/hyp.8133

Scharffenberg, W. (2016). *HEC-HMS User's Manual, Version 4.2*. U.S. Army Corps of Engineers Institute for Water Resources Hydrologic Engineering Center (CEIWR-HEC).

Seiller, G., Anctil, F., & Perrin, C. (2012). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrology and Earth System Sciences*, *16*(4), 1171–1189. https://doi.org/10.5194/hess-1116-1171-2012

Seo, D.-J., Herr, H. D., & Schaake, J. C. (2006). A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrology and Earth System Sciences Discussions*, *3*(4), 1987–2035. https://doi.org/10.5194/hessd-3-1987-2006

Seo, D.-J., Koren, V., & Cajina, N. (2003). Real-Time Variational Assimilation of Hydrologic and Hydrometeorological Data into Operational Hydrologic Forecasting. *Journal of Hydrometeorology*, *4*(3), 627–641. https://doi.org/10.1175/1525-7541(2003)004<0627:RVAOHA>2.0.CO;2

Sharma, S., Siddique, R., Reed, S., Ahnert, P., & Mejia, A. (2019). Hydrological Model Diversity Enhances Streamflow Forecast Skill at Short- to Medium-Range Timescales. *Water Resources Research*, *55*(2), 1510–1530. https://doi.org/10.1029/2018WR023197

Srivastava, A., Deb, P., & Kumari, N. (2020). Multi-Model Approach to Assess the Dynamics of Hydrologic Components in a Tropical Ecosystem. *Water Resources Management*, *34*(1), 327–341. https://doi.org/10.1007/s11269-019-02452-z

Strauch, M., Bernhofer, C., Koide, S., Volk, M., Lorz, C., & Makeschin, F. (2012). Using precipitation data ensemble for uncertainty analysis in SWAT streamflow simulation. *Journal of Hydrology*, *414–415*, 413–424. https://doi.org/10.1016/j.jhydrol.2011.11.014

Tan, B. Q., & O'Connor, K. M. (1996). Application of an empirical infiltration equation in the SMAR conceptual model. *Journal of Hydrology*, *185*(1), 275–295. https://doi.org/10.1016/0022-1694(95)02993-1

Tegegne, G., Park, D. K., & Kim, Y.-O. (2017). Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. *Journal of Hydrology: Regional Studies*, *14*, 49–66. https://doi.org/10.1016/j.ejrh.2017.10.002

Teng, F., Huang, W., & Ginis, I. (2018). Hydrological modeling of storm runoff and snowmelt in Taunton River Basin by applications of HEC-HMS and PRMS models. *Natural Hazards*, *91*(1), 179–199. https://doi.org/10.1007/s11069-017-3121-y

Thiemig, V., Bisselink, B., Pappenberger, F., & Thielen, J. (2015). A pan-African medium-range ensemble flood forecast system. *Hydrology and Earth System Sciences*, *19*(8), 3365–3385. https://doi.org/10.5194/hess-19-3365-2015

Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, *6*(2), 123–137. https://doi.org/10.1080/15715124.2008.9635342

Todini, E. (2011). History and perspectives of hydrological catchment modelling. *Hydrology Research*, *42*(2–3), 73–85. https://doi.org/10.2166/nh.2011.096

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Vrugt, J. A., Diks, C. G. H., & Clark, M. P. (2008). Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling. *Environmental Fluid Mechanics*, *8*(5), 579–595. https://doi.org/10.1007/s10652-008-9106-3

Vrugt, J. A., Gupta, H. V., Nualláin, B., & Bouten, W. (2006). Real-Time Data Assimilation for Operational Ensemble Streamflow Forecasting. *Journal of Hydrometeorology*, *7*(3), 548–565. https://doi.org/10.1175/JHM504.1

Vrugt, J. A., & Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004838

Wijayarathne, D. B., & Coulibaly, P. (2020). Identification of hydrological models for operational flood forecasting in St. John's, Newfoundland, Canada. *Journal of Hydrology: Regional Studies*, *27*, 100646. https://doi.org/10.1016/j.ejrh.2019.100646

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079

Wöhling, T., Samaniego, L., & Kumar, R. (2013). Evaluating multiple performance criteria to calibrate the distributed hydrological model of the upper Neckar catchment. *Environmental Earth Sciences*, *69*(2), 453–468. https://doi.org/10.1007/s12665-013-2306-2

Xiong, L., Wan, M., Wei, X., & O'Connor, K. M. (2009). Indices for assessing the prediction bounds of hydrological models and application by generalised likelihood uncertainty estimation / Indices pour évaluer les bornes de prévision de modèles hydrologiques et mise en œuvre pour une estimation d'incertitude par vraisemblance généralisée. *Hydrological Sciences Journal*, *54*(5), 852–871. https://doi.org/10.1623/hysj.54.5.852

Xu, J., Anctil, F., & Boucher, M.-A. (2019). Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems. *Journal of Hydrology*, 578, 124002. https://doi.org/10.1016/j.jhydrol.2019.124002

Yen, H., Wang, X., Fontane, D. G., Harmel, R. D., & Arabi, M. (2014). A framework for propagation of uncertainty contributed by parameterization, input data, model structure, and calibration/validation data in watershed modeling. *Environmental Modelling & Software*, 54, 211–221. https://doi.org/10.1016/j.envsoft.2014.01.004

Chapter 6. Assessing Entropy-based Bayesian Model Averaging Method for Probabilistic Precipitation Forecasting

Summary of Paper 4: Darbandsari, P., & Coulibaly, P. (2021). The Application of the (Modified) Entropy-based Bayesian Model 1 Averaging Method for Probabilistic Precipitation Forecasting. *Journal of Hydrometeorology*, under review.

The main goal of this research work is to investigate the applicability of a variant of the Entropy-based Bayesian Model Averaging (En-BMA) approach for precipitation forecasting. Some modifications are proposed to enhance the En-BMA method for post-processing ensemble of precipitation forecasts. After verifications of seven different individual forecasts, comparison has been made between the sub-daily probabilistic precipitation forecasts derived from the modified En-BMA and the widely used traditional BMA methods.

Key findings of this research work include:

- Among different precipitation forecasts, the Regional Ensemble Prediction System (REPS) appeared to be the most robust one for the Northern Ontario regions, while none of them can be selected as the most accurate one in all lead times and locations.
- Implementing the proposed modifications enhances the performances of the En-BMA in the case of precipitation forecasting.

- Considering the whole time series of forecasts, both modified En-BMA and BMA methods shows competitive performances with the former outperforming the latter on extreme or large precipitation events.

6.1 Abstract

Bayesian Model Averaging (BMA) is a popular ensemble-based post-processing approach where the weighted average of the individual members is used to generate predictive forecasts. As the BMA formulation is based on the law of total probability, possessing the ensemble of forecasts with mutually exclusive and collectively exhaustive properties is one of the main BMA inherent assumptions. Trying to meet these requirements led to the entropy-based BMA (En-BMA) approach. En-BMA uses the entropy-based selection procedure to construct an ensemble of forecasts with the aforementioned characteristics before the BMA implementation. This study aims at investigating the potential of the En-BMA approach for post-processing precipitation forecasts. Some modifications are proposed to make the method more suitable for precipitation forecasting. Considering the 6-hour accumulated precipitation forecasts with lead times of 6 to 24 hours from seven different models, we evaluate the effects of the proposed modifications and comprehensively compared the probabilistic forecasts, derived from the BMA and the modified En-BMA methods in two different watersheds. The results, in general, indicate the advantage of implementing the proposed modifications in the En-BMA structure for possessing more accurate precipitation forecasts. Moreover, the superiority of the modified En-BMA method over BMA in generating predictive precipitation forecasts is demonstrated based on different performance criteria in both watersheds and all forecasting

horizons. These outperforming results of the En-BMA are more pronounced for large precipitation values which are particularly important for hydrologic forecasting.

6.2 Introduction

Accurate and reliable precipitation forecasting is fundamental for various operational water resources management tasks, flood control and mitigation in particular (Cuo et al., 2011; Shrestha et al., 2013; Steenbergen & Willems, 2014). The inherent uncertainties associated with precipitation forecasts, which mostly stem from the initial conditions and model structures (Jha et al., 2018; Taillardat et al., 2016), make it difficult to incorporate deterministic forecasts without uncertainty quantification into practical applications. This limitation shows the importance of generating reliable probabilistic precipitation forecasts that meet the needs of users. One of the most common approaches for quantifying different uncertainties and generating probabilistic forecasts is ensemble forecasting (Han & Coulibaly, 2020; Ji et al., 2019; Yang et al., 2012). Constructing ensemble prediction systems (EPS) can effectively enhance uncertainty estimation (Ma et al., 2018; Robertson et al., 2013). Besides an ensemble of forecasts from a single model with different initial conditions, EPS can be generated using multiple forecasting models, which leads to a better quantification of predictive uncertainty (Liu & Xie, 2014; Saedi et al., 2020; Xu et al., 2019; Yang et al., 2012).

Deriving reliable probabilistic forecasts from EPS requires the application of a post-processing approach, which utilizes the full capability of the ensemble for quantifying predictive uncertainty (Liu & Xie, 2014; Scheuerer & Hamill, 2015). Bayesian Model Averaging (BMA) is one of the most widely used ensemble post-processing approach,

which was first proposed for merging multiple statistical models (Hoeting et al., 1999; Kass & Raftery, 1995) and then, it was extended for dynamical models and forecast ensembles (Raftery et al., 2005). By using the full information of the forecasts' ensemble, BMA generates reliable and sharp predictive distribution through weighted averaging of the posterior distributions conditioned on different individual members. The BMA weights are determined by the conditional probability of each member given observation, which represents the forecasting skills of the member in the training period.

Given that the original BMA uses the Gaussian function to estimate the posterior probabilities (Raftery et al., 2005), it is not reliable for precipitation where the predictive distribution is not normal (high possibility of being zero and highly skewed for non-zero values; (Sloughter et al., 2007; Yang et al., 2012)). In order to relax the aforementioned assumption, the BMA method was modified for skewed variables, such as precipitation, through proper modeling of the distribution. Sloughter et al. (2007) developed BMA for precipitation using a two-stage strategy, where the predictive distributions of each forecast are modeled by a mixture of a point mass at zero and a gamma distribution for positive values. The capability of the proposed BMA variant for generating reliable probabilistic precipitation forecasts was shown by various studies (Aminyavari & Saghafian, 2019; Fraley et al., 2010; Ji et al., 2019; Liu & Xie, 2014; Saedi et al., 2020; Vogel et al., 2018). Besides, Yang et al. (2012) used the Tweedie distribution, which can simultaneously model the probability of precipitation and its amount. Moreover, stratifying precipitation forecasts using threshold values (Ji et al., 2019) or ensemble spread (Zhu et al., 2015) was recommended for possessing more reliable results especially for heavy events.

Although some studies tried to extend the BMA applicability for precipitation forecasting by properly modeling the predictive distribution functions, there are some more inherent assumptions in the BMA scheme that require some attention. BMA is based on the law of total probability (Raftery et al., 2005), and consequently, for reliable performance of the BMA, an ensemble with independent members (exclusiveness) and high coverage of the observation variability (exhaustiveness) is required (Darbandsari & Coulibaly, 2020a; Refsgaard et al., 2012). In other words, an ensemble of forecasts with mutually exclusive and collectively exhaustive properties can lead to better BMA results. However, these two properties are in contradiction with each other. Simply increasing the number of members can relatively assure the latter property (exhaustiveness), while this may contravene the exclusiveness requirement by increasing the redundant information within the ensemble (Madadgar & Moradkhani, 2014). Therefore, generating a balanced EPS with the two aforementioned properties seems necessary in any BMA applications. Recently, Darbandsari and Coulibaly (2020a) proposed an entropy-based Bayesian model averaging (En-BMA) approach to relax the assumption of possessing a mutually exclusive and collectively exhaustive ensemble. Prior to the BMA application, the optimal subset of members is selected using three different entropy terms by simultaneously minimizing the redundant information between members while keeping the overall information amounts at the highest level. Their study shows the superiority of their proposed approach over BMA for daily streamflow forecasting.

So far, no studies have assessed the En-BMA approach for probabilistic precipitation forecasting. The main objective of this work is to propose a modified version of the

entropy-based BMA method for post-processing ensemble of precipitation forecasts. The multi-model EPS that incorporates the precipitation forecasts of seven different numerical models are used to evaluate the applicability of the modified En-BMA for sub-daily precipitation forecasting in two basins located in Ontario, Canada. Besides assessing the performance of individual forecasts and showing the importance of using an ensemble system, the relative performance of the modified En-BMA, compared with BMA, shows the advantages of possessing ensemble precipitation forecasts with mutually exclusive and collectively exhaustive properties for generating more accurate ensemble-based probabilistic results. The remainder of the paper is organized as follows. Section 6.3 presents the underlying concepts of the BMA and the En-BMA post-processing approaches. In section 6.4, we briefly describe the study areas and data. Section 6.5 discusses the results, including the comprehensive comparison of BMA and En-BMA, and a summary and conclusions are presented in Section 6.6.

6.3 Methodology

6.3.1 Bayesian Model Averaging (BMA) for precipitation forecast

Bayesian Model Averaging (BMA; (Hoeting et al., 1999; Raftery et al., 2005)) is a statistical post-processing approach where a weighted combination of the predictive distribution functions from different competing individual forecasts is used for generating more reliable probabilistic results. In the original BMA, the posterior distribution of forecasted variable y given K different forecasts ensemble ($F = \{f_1, f_2, \dots, f_K\}$) is formulated using the law of total probability (Raftery et al., 2005):

$$P(y|f_1, f_2, \dots, f_k, Y) = \sum_{i=1}^{k} w_i \times P(y|f_i, Y) \quad (6\text{-}1)$$

where w_i is the BMA weight, which presents the forecasting skill of the corresponding ensemble member (f_i) over the training period Y, and $P(y|f_i, Y)$ is the posterior distribution of y conditioned on individual forecast i. The aforementioned posterior distributions are assumed to follow a Gaussian distribution in the original BMA (Raftery et al., 2005), while various studies show that this is a poor choice for precipitation forecasts with a large number of zero values and highly skewed distributions for positive dates. Sloughter et al. (Sloughter et al., 2007) proposed one of the most well-known BMA variants for precipitation where the main modification includes replacing the Gaussian distribution with a mixture of a point mass at zero and the gamma distribution, and transforming data using cube root. So, the main equation of the BMA can be rewritten as follows (Sloughter et al., 2007):

$$\begin{aligned} P(y'|f_1', f_2', \dots, f_K', Y) \\ = \sum_{i=1}^{k} w_i \\ \times (P(y = 0|f_i') \times \mathrm{I}[\mathrm{y} = 0] + P(y > 0|f_i') \times g_i(y'|f_i') \\ \times \mathrm{I}[\mathrm{y} > 0]) \end{aligned} \quad (6\text{-}2)$$

y' and f_K' respectively shows the cube root of observations and forecasts and $\mathrm{I}[.]$ is the general indicator function which will be unity if the condition inside the bracket holds.

$P(y = 0|f_i')$ and $P(y > 0|f_i')$ respectively represents the probability of no precipitation and probability of precipitation given the forecast member f_i' which is estimated using logistic regression (Sloughter et al., 2007):

$$logitP(y = 0|f_i') = \log\frac{P(y = 0|f_i')}{P(y > 0|f_i')} = a_{0i} + a_{1i} \times f_i' + a_{2i} \times \delta_i \quad (6\text{-}3)$$

$$\delta_i = \begin{cases} 1 \; if \; f_i = 0 \\ 0 \; if \; f_i \neq 0 \end{cases} \quad (6\text{-}4)$$

δ_i is the second predictor variable (Equation 6-4), which is considered to enhance the logistic regression performance (Sloughter et al., 2007). Also a_{0i}, a_{1i}, and a_{2i} are the member specific parameters that need to be estimated by logistic regression based on the training data.

In the case of occurring precipitation, the posterior distribution of the cubic root of precipitation conditioned on each ensemble member i is modeled using the gamma distribution ($g_i(y'|f_i')$) with the following mean (μ_i) and variance (σ_i^2) (Sloughter et al., 2007):

$$\mu_i = b_{0i} + b_{1i}f_i' \quad (6\text{-}5)$$

$$\sigma_i^2 = c_0 + c_1 f_i \quad (6\text{-}6)$$

b_{0i} and b_{1i} are the parameters which need to be estimated separately for each ensemble member i using a simple linear regression between the cube root of non-zero observations (y') as predicant and the cubic root of the corresponding forecasts (f_i') as the predictor over the training period. c_0 and c_1 are the variance parameters, which are used to capture the

heteroscedastic characteristics of the standard deviations as a function of the predictor values (Sloughter et al., 2007; Vrugt & Robinson, 2007).

Weights (w_i) and variance parameters (c_0 and c_1) are determined using the modified expectation-maximization (EM) algorithm by maximizing the log-likelihood function ($L(w_i, c_0, c_1)$):

$$L(w_i, c_0, c_1) = Log\left(\sum_{i=1}^{K} P(y'|f_1', f_2', \dots, f_K', Y)\right) \quad (6\text{-}7)$$

EM is an iterative algorithm (Figure 6-1). After initialization, the latent variable (z) is calculated using the current values of the parameters in the expectation step. In the maximization step, the updated value of z is used to calculate the weights while the variance parameters are estimated numerically by maximizing the objective function (Equation 6-7) using the updated weights. The EM algorithm is not a global optimization method, and its sensitivity to the initial parameter values could lead to local maxima (Sloughter et al., 2007; Vrugt et al., 2008), so it can be replaced with a global optimization technique for possessing more robust estimation. In this study, the application of the dynamically dimensioned search (DDS) optimization method (Tolson & Shoemaker, 2007) as an alternative for the modified EM algorithm is evaluated. The global optimal solution in the DDS approach is determined by dynamically rescaling the dimension of the search space.

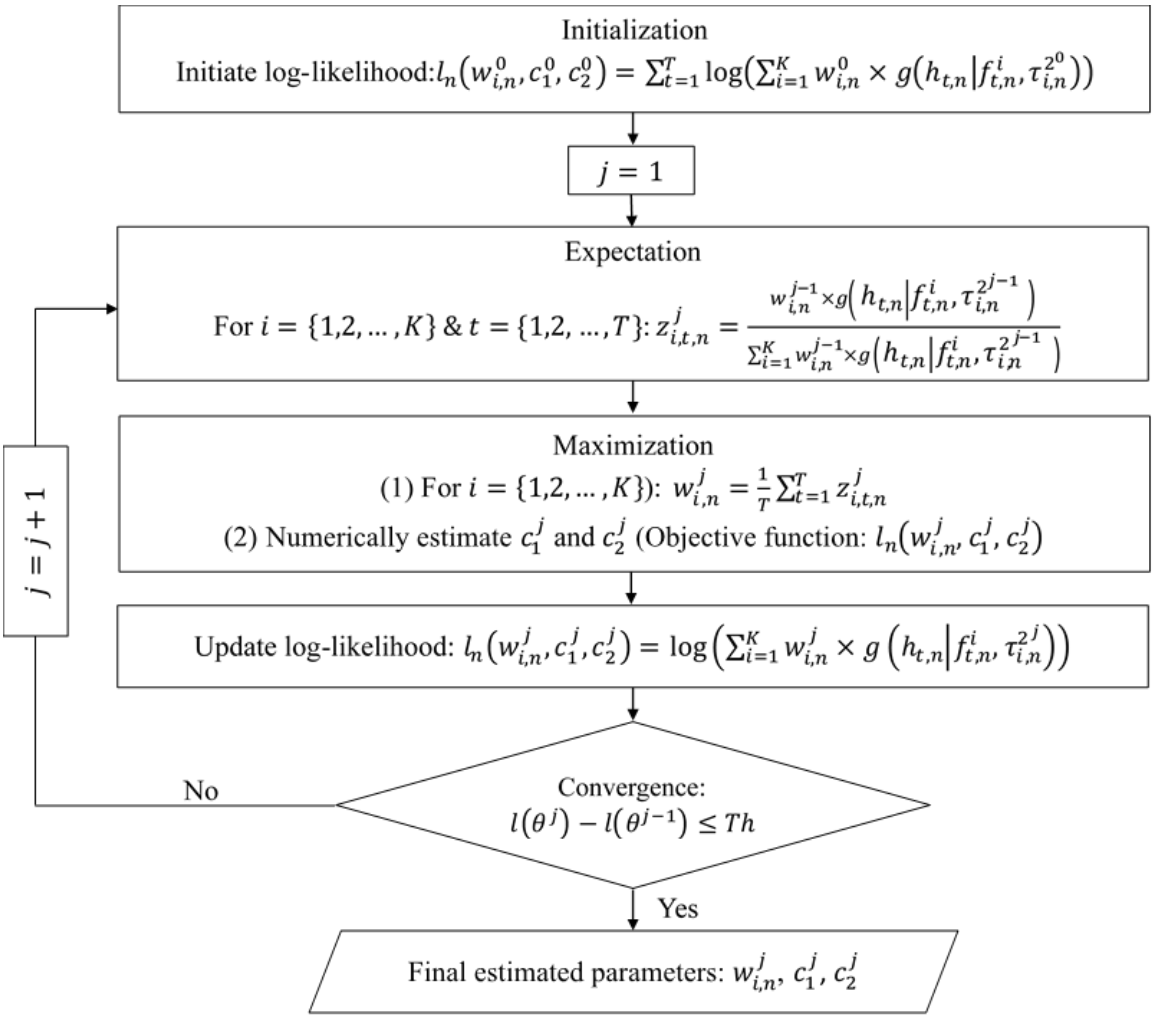


Figure 6-1 The modified Expectation-Maximization algorithm (after Sloughter et al., 2007)

Altogether, the proper estimation of the BMA parameters includes three main steps: using the training data, (1) the linear regression is used for estimating b_{0i} and b_{1i} parameters, (2) the a_{0i}, a_{1i}, and a_{2i} parameters are estimated using logistic regression, and (3) weights (w_i) and variance parameters (c_1 and c_2) are determined based on the EM algorithm. It is worthy of note that the moving window scheme is used for defining the training period. The sliding window of observation-forecast pairs before the initial date is taken as the recursive

training period. Therefore, selecting the optimal window length is the initial step before the BMA implementation.

6.3.2 Entropy-based BMA for precipitation forecast

As previously mentioned, possessing a mutually exclusive and collectively exhaustive ensemble of forecasts is a prerequisite for proper implementation of the BMA postprocessor. In the Entropy-based Bayesian Model Averaging (En-BMA) approach, which is initially proposed for streamflow forecasting (Darbandsari & Coulibaly, 2020a), an optimal subset of forecasts ensemble with lower dependency and higher information content was selected before applying the BMA method. Since detailed descriptions of the En-BMA concepts are provided in Darbandsari and Coulibaly (2020a), a brief overview of this approach and the proposed modifications to make it more suitable for precipitation forecasting are presented here.

The En-BMA method focuses on constructing the optimal ensemble of forecasts for BMA application using different entropy terms. Without any prior assumption about the statistical characteristics of the data sets, entropy provides a measure of the corresponding information content included in the data (Leach et al., 2015; Mishra & Coulibaly, 2009; Singh, 1997). Based on the Shannon entropy of information theory (Shannon, 1948), the marginal entropy shows the amount of information retained by a single variable ($H(X)$), while in the case of more than two variables (e.g. $X_1, X_2, \dots, X_N$), the term, joint entropy ($H(X_1, X_2, \dots, X_N)$), is defined as a measure of the overall information content gained from knowing all variables. The highest possible joint entropy value of multiple variables will

be equal to the sum of their marginal entropies in the case of possessing independent variables (Keum et al., 2019; Li et al., 2012). The relationship between marginal entropy and joint entropy leads to the definition of the total correlation ($C(X_1, X_2, \dots, X_N)$), which is a measure of the redundant information in multiple variables (Alfonso et al., 2013; Keum & Coulibaly, 2017):

$$C(X_1, X_2, \dots, X_N) = \sum_{i=1}^{N} H(X_i) - H(X_1, X_2, \dots, X_N) \quad (6\text{-}8)$$

In the case of possessing two variables (or groups of variables), the total correlation is transformed to the transinformation ($T(X_1, X_2)$) which represents the amount of dependence between two variables (or groups of variables):

$$T(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2) \quad (6\text{-}9)$$

The transinformation changes between zero, for fully independent variables, and $\min\big(H(X_1), H(X_2)\big)$, in the case of functionally dependent ones (Darbandsari & Coulibaly, 2020b).

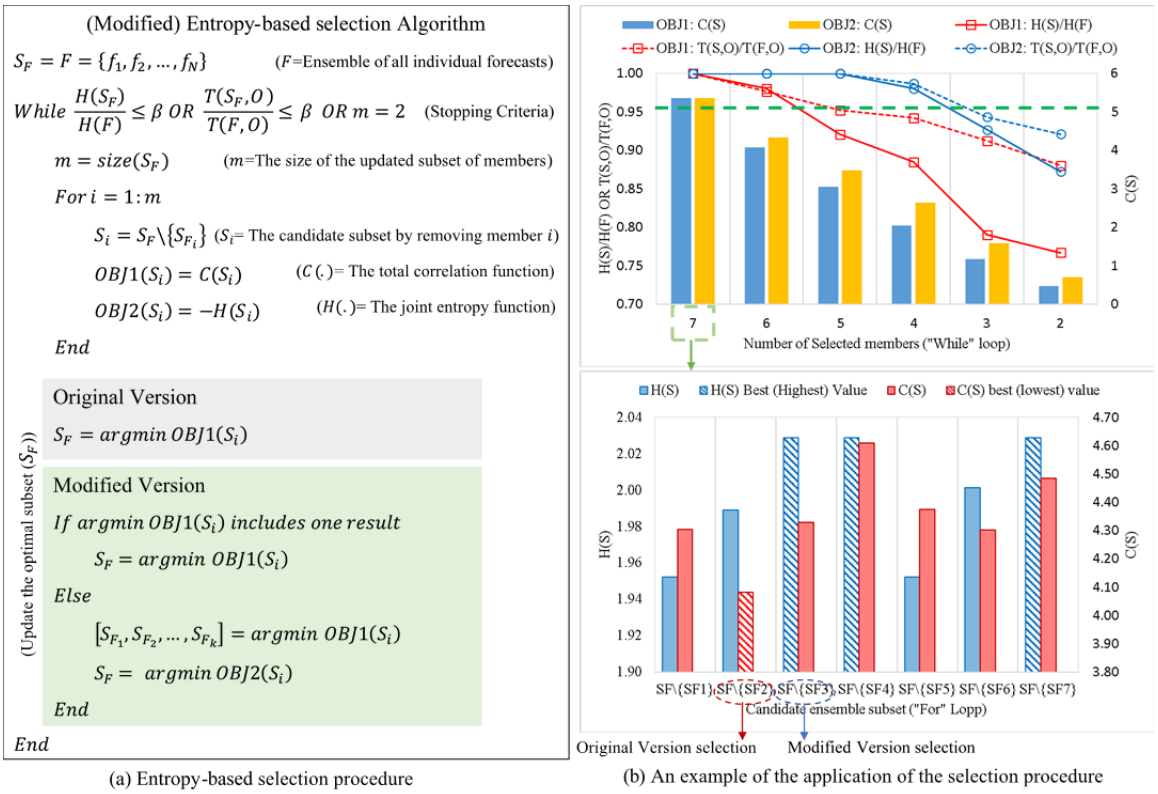
Considering individual forecasts and observation data sets as different variables, the aforementioned entropy terms can be used to generate an ensemble of forecasts with maximum information and minimum redundancy. For achieving this goal, the entropy-based selection algorithm with a nested loop structure is developed (Darbandsari & Coulibaly, 2020b). In this algorithm, two stopping criteria (the joint entropy of the selected subset over the joint entropy of all forecasts, and the ratio of the transinformation between selected members (as a group of variables) and observations to that of all candidates and

observations) are used to prevent the loss of information (Figure 6-2a). Also, in the original procedure, the total correlation is implemented as an objective function of the inner loop to select the optimal ensemble with the lowest redundant information among subsets with the same number of members.

In this study, an alternative objective function based on joint entropy is also evaluated. At first glance, it seems that defining the cost function of the inner loop based on the joint entropy will not lead to the lowest shared information. However, using joint entropy before the total correlation can prevent information loss resulting from removing members. So, with the same amount of information content, the final selected subset using the modified objective function possesses a lower number of members with a lower total correlation value, compared to the original version. As a representative example, shown in Figure 6-2b, the total correlation of the four-member optimal subset, derived from the modified selection procedure, is around 60 percent lower than the corresponding value of the final selected subset (including six members) from the original method, while the overall information content of both optimal subsets is at the same level.

Moreover, another modification that is applied in this study is about where the entropy-based selection procedure will be implemented. In the original En-BMA method, the selection procedure has been used prior to the BMA application (Figure 6-3a). As shown in Figure 6-3b, here we proposed an alternative framework, where the first step of the BMA procedure for precipitation, a linear regression, has been applied before narrowing down the ensemble members. In other words, the variables, being used in the selection procedure are the non-zero cubic root transformed and linear regressed forecasts and observation. As

the BMA posterior distribution for precipitation forecasts is defined based on the aforementioned data (Equation 6-2), their characteristics as an ensemble play an important role in the BMA performance, and considering them in the selection algorithm could lead to better results.



(a) Entropy-based selection procedure

(b) An example of the application of the selection procedure

Figure 6-2 The (modified) entropy-based selection procedure: (a) the Pseudo Code and (b) examples of their applications

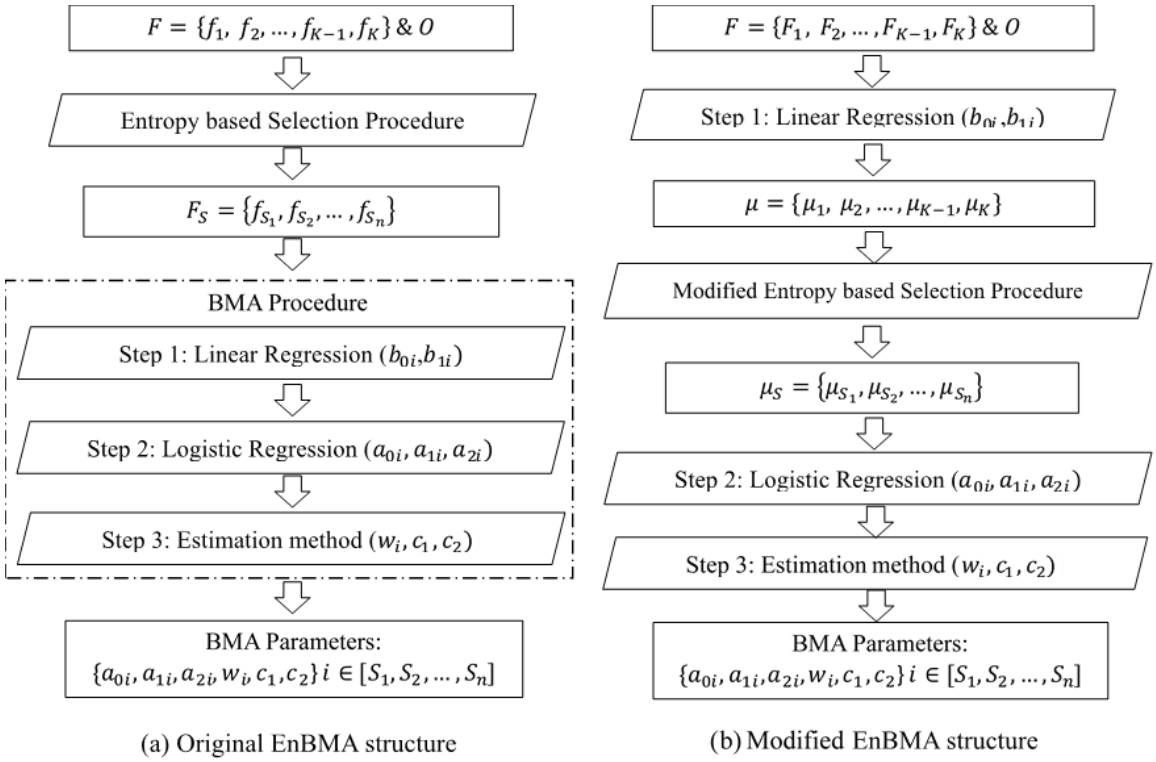


Figure 6-3 The structure of (a) the original and (b) the modified entropy-based BMA methods

6.3.3 Verification metrics

In this study, different evaluation metrics, including Mean Absolute Error (MAE), Root Mean Squared Error ($RMSE$), Pearson Correlation Coefficient (PCC), and Continuous Ranked Probability Score ($CRPS$) are used for the verification analysis of various post-processing approaches. MAE, $RMSE$, and PCC are the deterministic measures which are formulated as follows:

$$MAE = \frac{1}{N}\sum_{t=1}^{N}(|f_t - O_t|) \qquad (6\text{-}10)$$

$$RMSE = \left(\frac{1}{N} \left(\sum_{i=1}^{N} (f_t - O_t)^2 \right) \right)^{\frac{1}{2}} \quad (6\text{-}11)$$

$$PCC = \frac{\left[\sum_{t=1}^{N} (O_t - \bar{O})(f_t - \bar{f}) \right]}{\sqrt{\sum_{t=1}^{N} (O_t - \bar{O})^2 \sum_{t=1}^{N} (f_t - \bar{f})^2}} \quad (6\text{-}12)$$

where O_t and f_t are the observation and forecast (the mean of the predictive distributions of probabilistic forecasts), and $\bar{O}$ and $\bar{f}$ are respectively the observation and forecast mean over the verification period. $RMSE$ reflects the closeness between observation and forecast by giving more weights to large values (Coulibaly et al., 2005), while MAE is a more balanced criterion, assessing the average model performance using the difference between forecast and observation (Willmott & Matsuura, 2005). Both MAE and $RMSE$ varies between 0 and +∞ with the best value of 0. PCC, possessing a range of [-1,1], shows the linear dependency between forecasts and observed value (Verkade et al., 2013). The better forecasts possess higher PCC values while its negative values reflects the inverse correlation.

$CRPS$ is a probabilistic-based criterion, which is defined as the squared error of the probability distributions of the forecast (P_t^f) compared to the observed one (P_t^O) (Hersbach, 2000):

$$CRPS = \frac{1}{N} \sum_{t=1}^{N} \int_{-\infty}^{+\infty} \left(P_t^f(x) - P_t^O(x) \right)^2 d_x \quad (6\text{-}13)$$

$$CDF_t^O(x) = H(x - O_t) \quad (6\text{-}14)$$

$H(x - O_t)$ in Equation 6-14 is the Heaviside function, being equal to 1 if $x > O_t$, and 0 otherwise. $CRPS$, with a range of 0 to +∞, is negatively oriented where smaller values correspond to the better predictions. Besides the previously stated performance criteria, we use the reliability plot (Laio & Tamea, 2007) as a graphical tool for assessing the statistical reliability of the probabilistic forecasts. Also, its corresponding metric (α), changing between 0 and 1 with the best value of 1, reflects the reliability of the forecasts by calculating the difference between the reliability plot (forecasts cumulative probability distribution) and the bisector line (cumulative uniform distribution):

$$\alpha = 1 - 2 \times \left(\frac{1}{N} \sum_{t=1}^{N} P_t^f(O_i) - U(O_t) \right) \quad (6\text{-}15)$$

To facilitate the comparison of different post-processing approaches, the percentage of performance improvement, called percent improvement hereafter, is used. This term is defined based on different performance metrics as percent improvements of their values in the case of using one method, compared with another.

6.4 Study areas and data

We assessed the applications of the BMA and the (Modified) En-BMA for the post-processing ensemble of mean-areal sub-daily precipitation forecasts in the Big East River and the Black River watersheds, with the catchment areas of approximately 600 and 1500 km^2, respectively. As can be seen in Figure 6-4, both watersheds are located in Northern Ontario, Canada, and can be considered as poorly-gauged basins as there are no

meteorological stations within their boundaries. So, the Canadian Precipitation Analysis (CaPA) data are used as a reference dataset for verifying precipitation forecasts in both regions. CaPA is a 6-hourly precipitation product with a spatial resolution of 15 km, which is generated based on the combination of various precipitation sources (Lespinas et al., 2015; Mahfouf et al., 2007). Its reliability as an alternative to weather stations in Canadian data-poor catchments was shown by various studies (Darbandsari & Coulibaly, 2020b; Eum et al., 2014).

In this study, seven different numerical weather prediction systems, employed for generating an ensemble of precipitation forecasts, are Global Deterministic Prediction System (GDPS), Global Ensemble Prediction System (GEPS), Global Forecast System (GFS), Global Ensemble Forecasting System Reforecast Project Version 2 (GEFS), Regional Deterministic Prediction System (RDPS), Regional Ensemble Prediction System (REPS), and High-resolution Regional Deterministic Prediction System (HRDPS). Table 6-1 represents the detailed features of the aforementioned products. The seven-member 6-hr accumulated precipitation forecasts ensemble with lead times of 6 to 24 hours, issued at 0000 UTC, are used in this study. The verification period is from 2019/07/03 to 2020/08/31. We use the means of the single model ensemble forecasts (i.e. GEPS, REPS, and GEFS) for comparing their performances with deterministic models (i.e. GDPS, RDPS, GFS) and constructing the multi-model ensemble. It is of note that the spatially averaged predictions and verification data over each basin, which is determined using the Thiessen polygon approach (Thiessen, 1911), are considered as the mean areal precipitation data and be used for evaluating different post-processing methods.

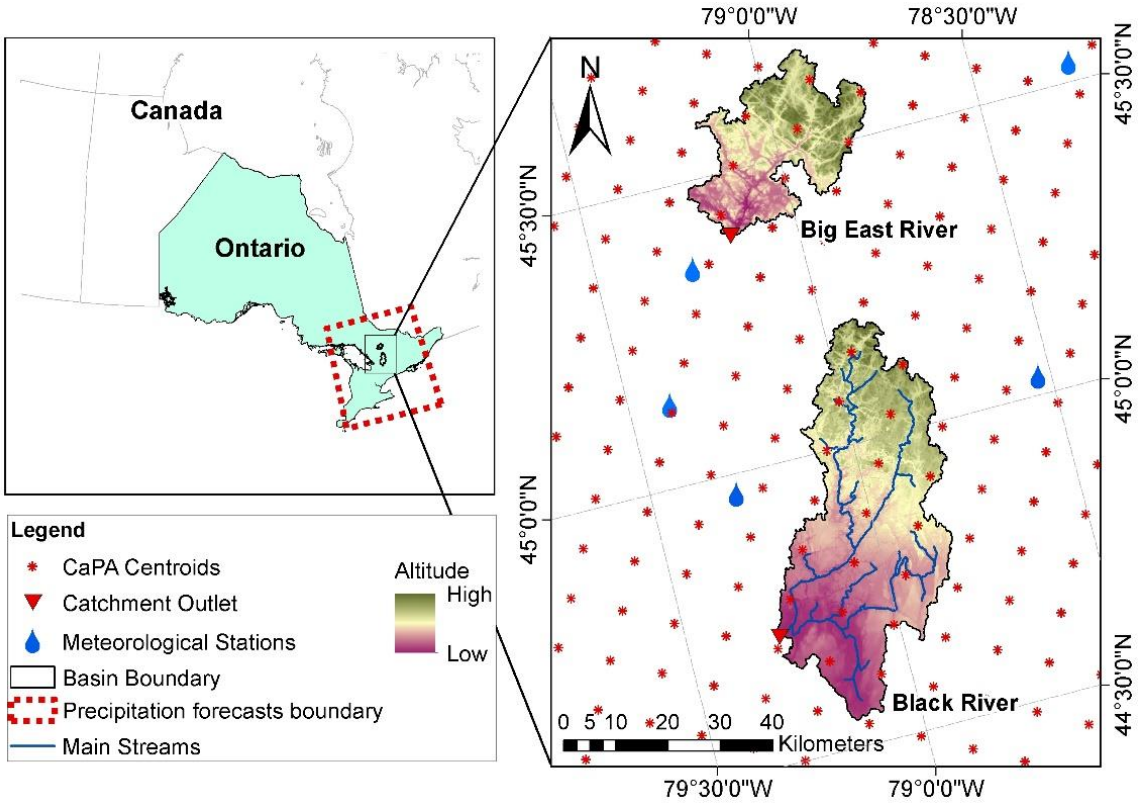


Figure 6-4 The location maps of the Big East River and Black River watersheds

Table 6-1 The detailed descriptions of the numerical weather prediction models used in this study

NWP	Resolution		Base Time (UTC)	Forecast length (hr)	Availability	Organization
	Spatial	Temporal				
GDPS	~25 km	1 hourly	00, 12	0 – 240	From 2019/07/03	CMC[1]
GEPS	~50 km	1 hourly	00, 12	0 – 384	From 2018/09/19	CMC
GFS	~27 km	3 hourly	00, 06, 12, 18	0 – 384	From 2015/01/15	NOAA[2]
GEFS	~50 km	3 hourly	00	0 – 192	From 2000/01/01	NOAA
REPS	~15 km	1 hourly	00, 12	0 – 72	From 2019/07/03	CMC
RDPS	~10 km	1 hourly	00, 06, 12, 18	0 – 84	From 2015/01/01	CMC
HRDPS	~2.5 km	1 hourly	00, 06, 12, 18	0 – 48	From 2017/05/22	CMC

1 Canadian Meteorological Centre

2 National Oceanic and Atmospheric Administration

6.5 Results and Discussions

6.5.1 Individual model performance

Before evaluating the post-processing approaches, a primary comparison was made between the performances of different individual forecasts in terms of MAE, PCC, and $RMSE$ (Figure 6-5). In general, the results in both basins indicate that REPS leads to the most consistent precipitation forecasts based on most of the criteria at different forecasting horizons. This is in line with Abaza et al. (Abaza et al., 2013), showing the REPS potential for precipitation forecasting in Canadian catchments. Also, GEPS performs competitively in both watersheds. Regarding MAE and $RMSE$, the GEFS model provides accurate results in both basins, however, the PCC criterion in the Big East River watershed shows the relatively poor performance of GEFS compared to the other forecasts. Moreover, the HRDPS forecasts are relatively accurate for lead-times 6 and 12; however, its performance is among the worst ones for longer lead-times (i.e. 18 and 24), especially in the Big East River watershed. Although the performances of GFS in the Big East River watershed are relatively reliable, it possesses the lowest skill for precipitation forecasts in Black River, especially for 6-, 12-, and 18-hours ahead forecasts.

Altogether, the main conclusion, standing out from comparing various individual models, is that neither of the models has complete superiority to provide the most promising precipitation forecasts in both basins and for different lead-times. Although in general, the REPS model can be considered as the good performing one, it does not always provide the best results based on different verification metrics. Therefore, selecting the best model is

practically impossible. This shows the importance of using an ensemble of multi-model forecasts instead of relying on individual ones (Liu & Xie, 2014; Qu et al., 2017).

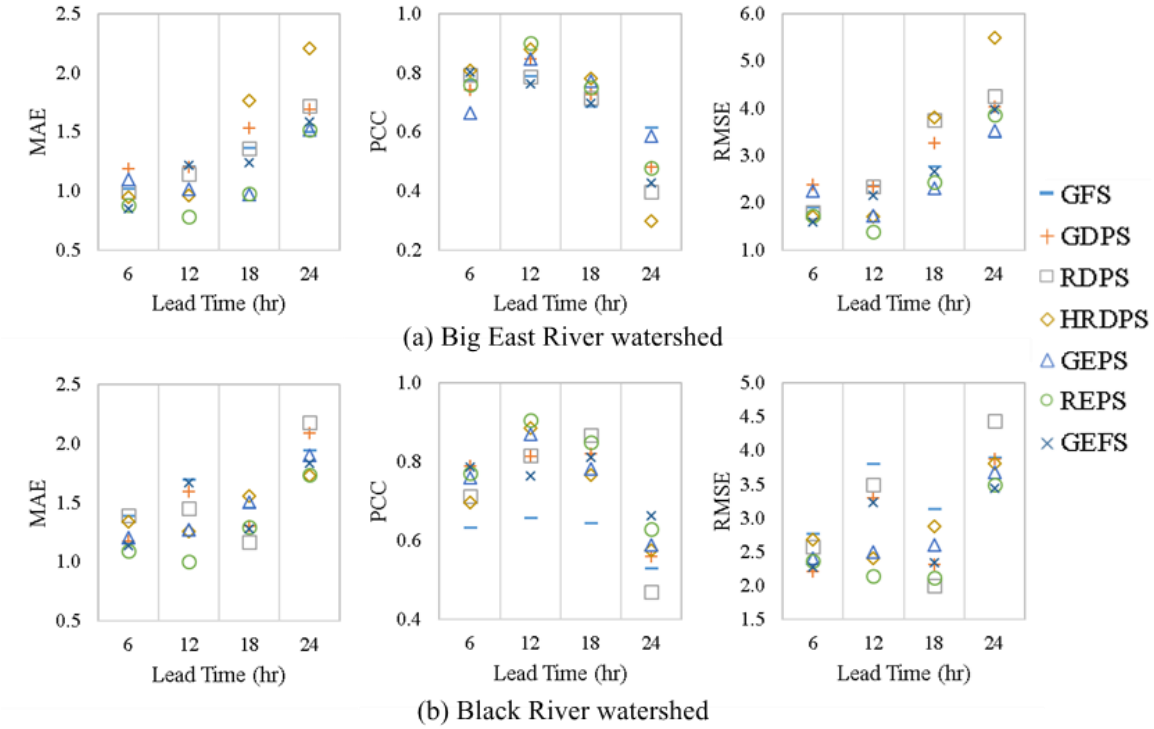


Figure 6-5 Comparison of different performance measurements for 6 to 24 hours-ahead forecasts derived from different forecasting models in (a) Big East River and (b) Black River watersheds

6.5.2 BMA evaluation

Following Raftery et al.(Raftery et al., 2005), a temporal moving window, including sample data from N previous days, is used for estimating BMA parameters. So, determining the optimal window length (N) is the primary and important step prior to the applications of the BMA (and En-BMA). As the optimal length of moving window varies based on areas and data sets, it should be specifically determined for each study (Liu &

Xie, 2014; Sloughter et al., 2007). Here, by using different evaluation metrics, we compare the performance of BMA as a function of the moving window length (Figure 6-6). In line with previous studies (Raftery et al., 2005; Schmeits & Kok, 2010; Sloughter et al., 2007; Vrugt & Robinson, 2007), the results in both Big East River and Black River watersheds show that increasing the length of training period leads to better BMA results in terms of different performance metrics. However, these trends of improvements decrease with increasing moving window length.

Moreover, as previously mentioned, the BMA formulation, proposed for precipitation forecasting is based on the linear regression between observation and forecasts in non-zero precipitation dates. Therefore, a sufficient number of non-zero precipitation dates is required for the proper estimation of the regression parameters. As expected, by increasing the length of the moving window, the number of dates with precipitation occurrence increases (Figure 6-6), which can be one of the main reasons for better BMA performance with higher N values in the case of forecasting precipitation. It is worthy of note that although larger moving window length provides higher information, it could be impractical in the real world, due to the limited length of available time-series (Xu et al., 2019). In this study, we select a 100-day moving window for the applications of both BMA and En-BMA methods.

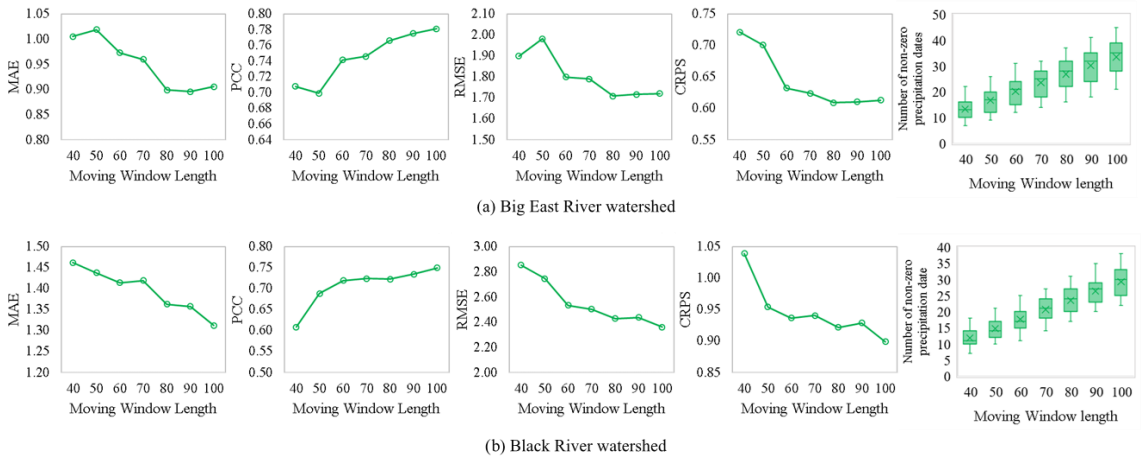


Figure 6-6 The performance statistics of the BMA 6-hour ahead forecasts and the number of non-zero precipitation dates as a function of moving window length in (a) Big East River and (b) Black River watersheds

Besides the length of the training period, how well the BMA parameters are estimated can also have noticeable effects on the results. The modified EM algorithm is proposed for estimating BMA parameters, while it is argued that this method may have some difficulty in finding the global optimal estimations. Here, in order to evaluate the capability of the modified EM method in estimating BMA parameters, a dynamically dimensioned search (DDS) global optimization method is considered as the benchmark, and a comparison is made between BMA models calibrated with both approaches. The results in both basins, as shown in Figure 6-7, indicate that both methods lead to approximately similar parameters and objective function values. Therefore, the modified EM algorithm is sufficiently reliable for estimating BMA parameters. This conclusion is in line with Darbandsari and Coulibaly (2019) where an in-depth analysis in the case of streamflow forecasting shows that the log-likelihood is a convex function of the most sensitive BMA

parameters, leading to the reliable performance of the local optimization technique, such as the EM algorithm.

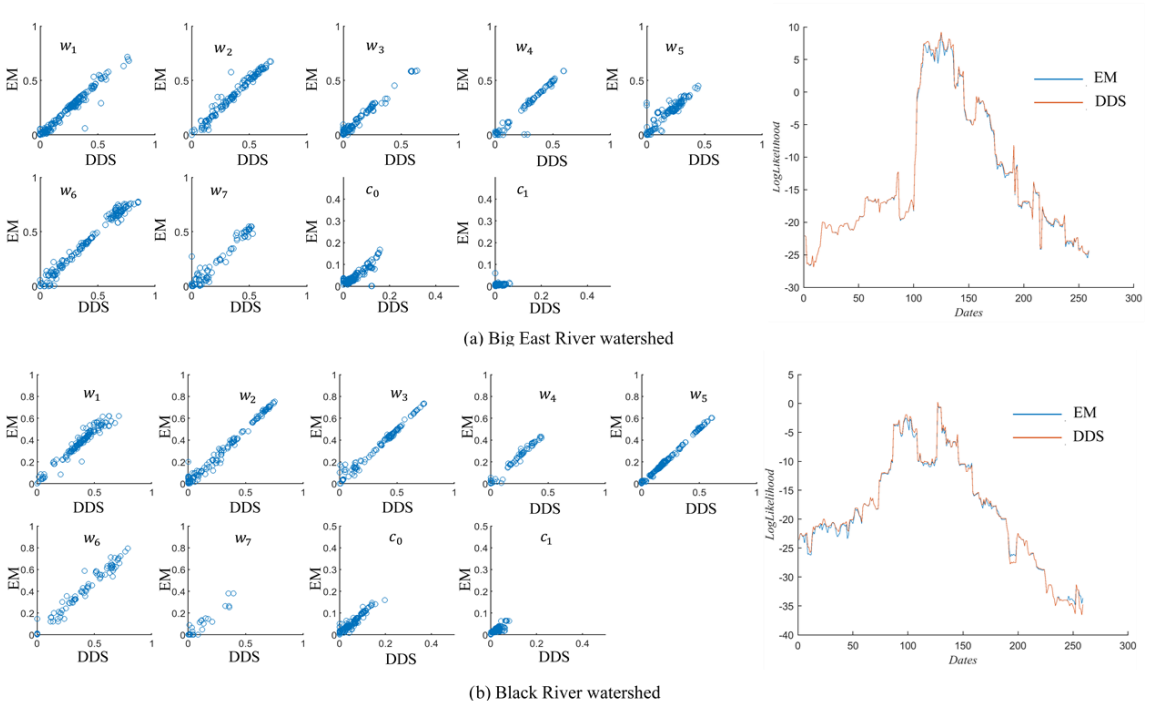


Figure 6-7 A comparison of the BMA parameters and the objective function (loglikelihood) values derived from the modified expectation-maximization (EM) algorithm and the dynamically dimensioned search (DDS) method in (a) Big East River and (b) Black River watersheds

The contribution of each forecast member into the BMA predictive results is determined by its corresponding estimated weights. Figure 6-8 simply compared the average weights of different models with their performances in both basins at different lead-times. What stands out in this figure is that the weights are not completely following the performance of individual members. There are some cases where members with relatively lower performance, possess higher weights and vice versa. For example, HRDPS, which was selected as the worst model in the Big East River watershed at lead-time 24 (Section 6.5.1),

has relatively high weights compared with the other members. On the other hand, GEFS possesses the most accurate 6-hour ahead forecasts in Black River; however, lower BMA weights are allocated to it. These results show that apart from the forecasting skills of individual models, the diversity of the ensemble members also plays an important role in the BMA application (Darbandsari & Coulibaly, 2020b; Sharma et al., 2019).

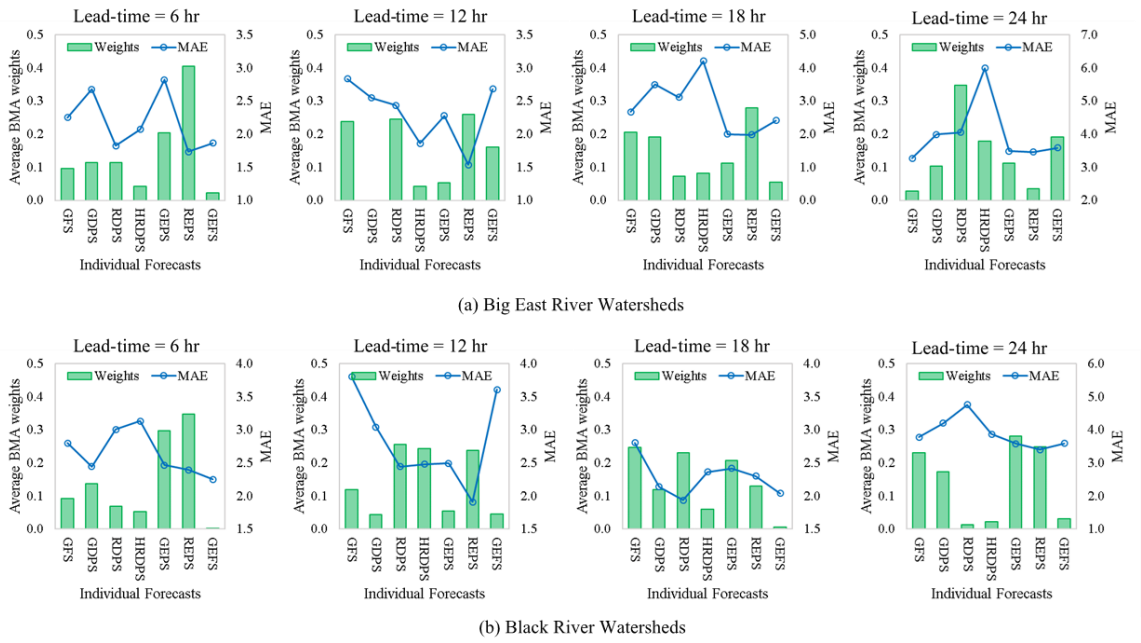


Figure 6-8 he average BMA weights and the MAE performance statistics of each member at different forecasting horizons in (a) Big East River and (b) Black River watersheds

6.5.3 En-BMA evaluation

In both En-BMA and its modified version, the only parameter that needs to be specified is the stopping threshold value (β). This parameter is used as a criterion that is not allowed to be violated, and implicitly shows the maximum amount of information that we want to keep in the system. We evaluate the effects of choosing different threshold values as the

first step of the implementation of the En-BMA and the Modified En-BMA (M-EnBMA) methods for forecasting precipitation. As can be seen from Figure 6-9, the average number of selected members is noticeably affected by changing the stopping threshold value. As expected, increasing β leads to a higher number of selected members; however, this increase does not follow the same trend using both En-BMA and M-EnBMA methods. The M-EnBMA approach has a lower number of selected members, compared to En-BMA, in the case of using the same threshold value. This is justifiable by the fact of using joint entropy as the objective function in the inner loop of the modified selection procedure (Figure 6-3), which helps to keep information at the highest possible level by narrowing the ensemble down.

Also, using MAE and $CRPS$ measurements, the performances of both post-processing approaches in producing 6-hour ahead forecasts have been evaluated as a function of β (Figure 6-9). The En-BMA predictions in both basins indicate that using threshold values less than 0.9 leads to unreliable precipitation forecasts and the values of 0.9 and 0.95 provide the best results. On the other hand, in the case of applying M-EnBMA, the improving trend of the performances continues by increasing β. The $\beta = 0.99$ leads to the selection of around 4 members (on average) with the best performances in both basins. So, for the rest of this study, the threshold values of 0.95 and 0.99 were respectively selected for En-BMA and M-EnBMA applications in both watersheds.

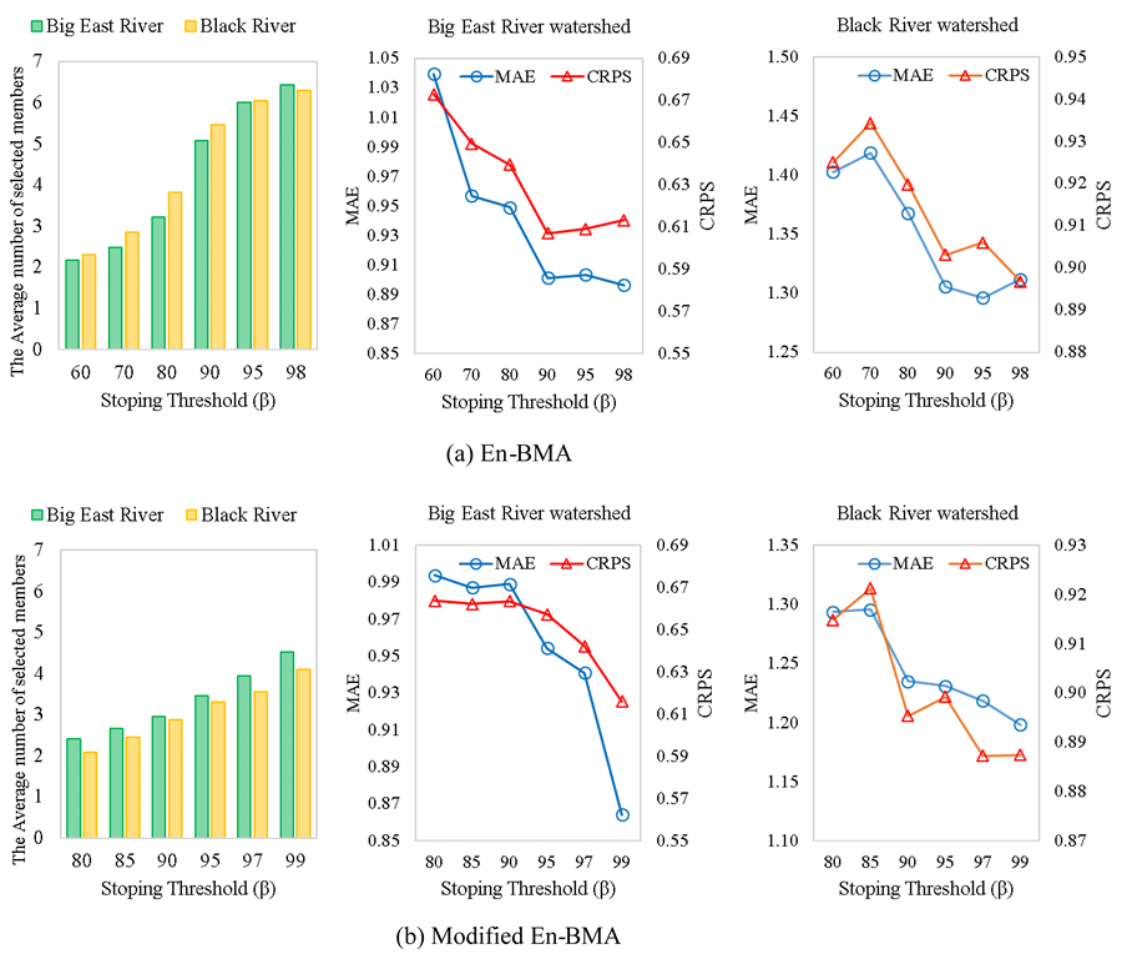


Figure 6-9 The effects of stopping threshold values on the average number of selected members and the performances of the 6-hour ahead forecasts, derived from (a) the En-BMA and (b) the modified En-BMA post-processing methods in both Big East River and Black River watersheds

Prior to the weights, how frequently an individual member is selected shows its participation in generating predictive forecasts in the case of using entropy-based approaches. The M-EnBMA selection ratios of various members in both basins at different lead-times, as shown in Figure 6-10, are not always in accordance with their corresponding individual performance (Figure 6-5). REPS, as the most promising model in both watersheds, does not possess the predominant ratios of selection. As another example in

the Big East River watershed, HRDPS contribution in the final selected ensemble is relatively high, while its performances at lead-time 18 and 24 hours are among the worst ones. In conclusion, these results indicate that even the low-skill models could have some unique information, and their presence in the ensemble is required to meet the collectively exhaustive and mutually exclusive properties. Also, the distribution of the average En-BMA weights amongst the selected members does not properly agree with their corresponding selection ratios (Figure 6-10), while their comparison with the BMA weights shows a positive relationship between them. This may be due to the fact that the same structure as the original BMA is utilized in the proposed entropy-based methods.

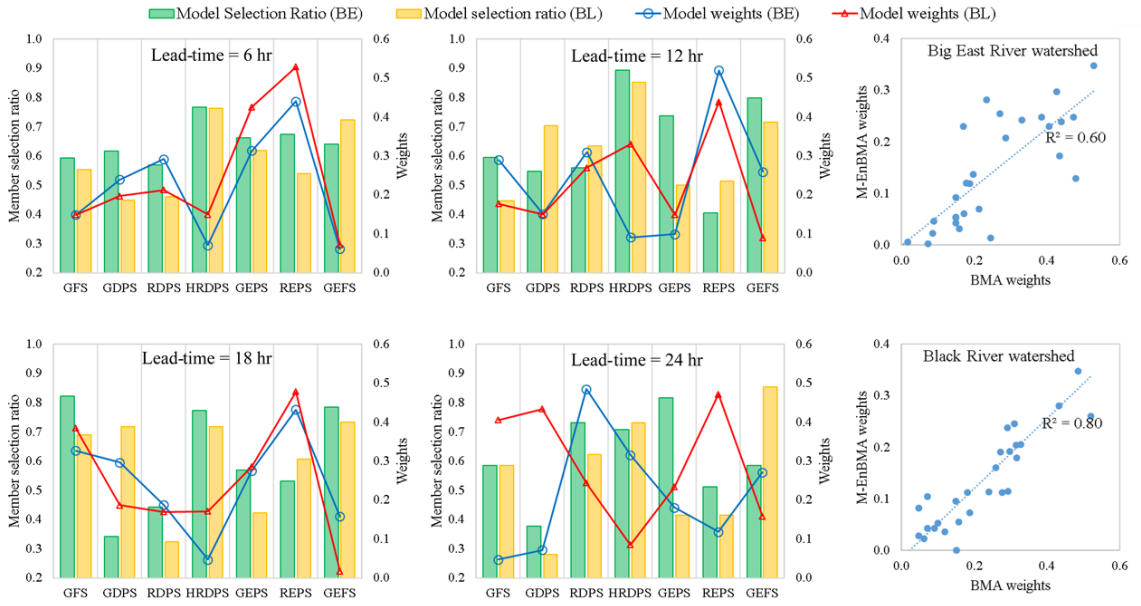


Figure 6-10 The average weights and selection ratio of each member in the case of applying the modified En-BMA (M-EnBMA) approach and a comparison between the M-EnBMA and BMA weights in both Big East River and Black River watersheds at different forecasting horizons (6 to 24 hours ahead).

For evaluating the proposed modifications, we made a comparison between the En-BMA and the M-EnBMA methods using the percent improvements based on different metrics where positive values show the advantages of applying modifications (Table 6-2). The results in both watersheds indicate the superiority of the modified version over the original one based on all performance metrics. This advantage exists in all forecasting horizons by providing 2 to 15 percent performance improvement in general. No relationship can be found between improvements and the forecasting horizons. In the Big East River watershed; the highest improvement can be seen in the 12 hr ahead forecasts, where MAE, $RMSE$, and $CRPS$ indicate more than 10 percent of improvement. However, in the Black River watershed, the largest difference between En-BMA and M-EnBMA occurs at lead-time 18 (MAE, PCC, and RMSE improve more than 10 percent). Altogether, it is concluded that using the proposed modifications in the En-BMA structure leads to better precipitation forecasting. So, the modified En-BMA method is considered for the rest of this paper to be compared with the BMA approach.

Table 6-2 The percentage of improvements derived from using the modified En-BMA instead of BMA based on different performance metrics in both Big East River and Black River watersheds at different forecasting horizons (6 to 24 hours ahead)

Basin	**Big East River**				**Black River**			
Lead-time (hr)	**Percent Improvement (%)[1]**				**Percent Improvement (%)**			
	MAE	**PCC**	**RMSE**	**CRPS**	**MAE**	**PCC**	**RMSE**	**CRPS**
6	4.1	2.0	2.0	3.5	8.2	2.4	3.6	2.3
12	12.8	7.7	11.8	11.5	5.3	2.5	2.6	4.7
18	6.6	4.9	4.9	2.9	12.3	14.5	14.0	8.4
24	2.4	6.4	2.2	2.8	2.4	3.6	3.6	7.0

1 The positive values of percent improvement indicate the advantage of using M-EnBMA over En-BMA.

6.5.4 Modified En-BMA versus BMA

BMA and the proposed modified En-BMA, called M-EnBMA hereafter, are implemented to post-process the ensemble of 6-hourly accumulated precipitation forecasts up to 24 hours ahead within the verification period (2019/07/03 to 2020/08/31). Using four different criteria, presented in Section 6.3.3, Figure 6-11 compared the performance of both methods in the Big East River and the Black River watersheds. The main conclusion that stands out from the comparison in both watersheds and different lead-times is that in general, all performance statistics show the superiority of M-EnBMA over BMA. The percentages of performance improvement based on various metrics are always positive showing the advantage of using M-EnBMA compared with BMA for post-processing precipitation forecasts.

These enhancements are not constant as a function of different performance metrics. The $CRPS$ criterion, which assesses the accuracy of the probabilistic forecasts, shows an average of 5% consistent enhancements in all forecasting horizons at both basins. However, the percent improvements based on the other three deterministic-based measurements (i.e. MAE, PCC, and $RMSE$), vary as a function of forecasting horizons. In the Big East River watershed, the greatest superiority of M-EnBMA over BMA occurs for 18 hours ahead forecasts (the MAE, PCC, and $RMSE$ improvements are respectively 8%, 9%, and 8%) while, both BMA and M-EnBMA methods leads to almost similar results at lead-time 24. In Black River, on the other hand, the lowest difference between BMA and M-EnBMA

methods can be seen at 18 hours ahead forecasts where almost the same results are derived from both approaches.

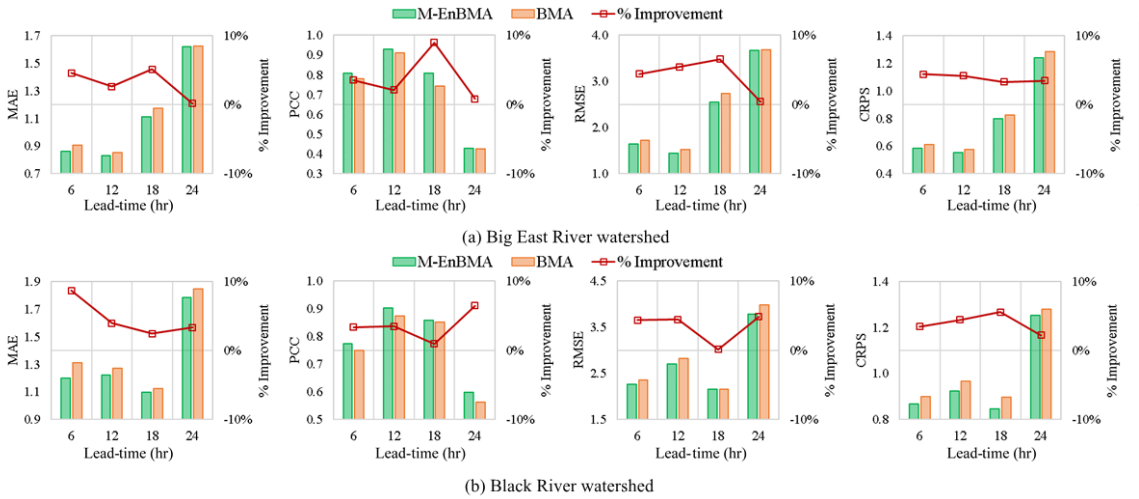


Figure 6-11 Comparison of different performance measurements for 6 to 24 hours-ahead forecasts derived from the BMA and the modified En-BMA (M-EnBMA) methods in (a) Big East River and (b) Black River watersheds. The positive value of % improvement shows the advantage of using M-EnBMA instead of BMA

Also, to specifically evaluate the performances of both methods in reproducing high precipitation values, MAE is calculated and compared using precipitation values more than 90 percentile (MAE_{90}) and values more than 5 mm (MAE_5). These comparisons in general, as presented in Table 6-3, again show that M-EnBMA is better in terms of forecasting large values. These enhancements are more pronounced compared to the ones calculated using the whole data. In the case of 24 hours ahead forecasts in the Big East River and 18 hours ahead forecasts in the Black River watershed, as both methods possess the same overall performance based on the whole verification period (Figure 6-11), the MAE_{90} and MAE_5 improvements are also marginal. However, in most of the other forecasting horizons in

both basins, the highest improvements can be seen based on MAE_{90} and MAE_{5}, compared to MAE. For instance, at 12-hour ahead forecasts, MAE shows around five percent of improvement in both basins, while the MAE_{90} and MAE_{5} improvements respectively are 6% and 12% in the Big East River and 11% and 10% in the Black River watershed.

Table 6-3 The performance statistics of the BMA and the modified En-BMA (M-EnBMA) focusing high precipitation values in both Big East River and Black River watersheds at different forecasting horizons (6 to 24 hours ahead)

Basin	Lead-time (hr)	Method	MAE_{90} [1]	Improvement (%)[2]	$MAE_{<5}$ [1]	Improvement (%)[2]
BE[3]	6	BMA	1.93	10.3	3.26	15.2
		M-EnBMA	1.73		3.84	
	12	BMA	1.79	5.1	3.40	11.7
		M-EnBMA	1.70		3.01	
	18	BMA	2.83	10.3	4.95	7.5
		M-EnBMA	2.54		4.58	
	24	BMA	4.21	1.4	6.14	3.0
		M-EnBMA	4.15		5.95	
BL[3]	6	BMA	2.23	4.3	4.26	6.9
		M-EnBMA	2.13		3.97	
	12	BMA	2.72	11.1	5.64	9.9
		M-EnBMA	2.41		5.08	
	18	BMA	2.08	2	3.65	2.7
		M-EnBMA	2.04		3.80	
	24	BMA	4.17	2.7	6.16	38.3
		M-EnBMA	4.06		3.8	

1 MAE_{90} and $MAE_{>5}$ are the MAE calculated respectively based on values more than 90 percentile and 5 mm
2 The positive values of percent improvement indicate the advantage of using M-EnBMA over BMA.
3 BE and BL are the abbreviations of Big East River and Black River watersheds, respectively.

Moreover, Figures 6-12 and 6-13 evaluate the statistical reliability of both BMA and M-EnBMA methods using the reliability diagram and its corresponding reliability

measurements (α) respectively in the Big East River and the Black River watersheds. Comparing the BMA and the M-EnBMA results based on all precipitation data indicates the same reliability of both approaches in both watersheds and all forecasting horizons. However, by focusing on large values, the superiority of M-EnBMA over BMA in terms of generating reliable probabilistic forecasts is noticeable. Using precipitation data more than 90 percentile illustrates the slight enhancement of the reliability of the probabilistic forecasts in the case of applying M-EnBMA, compared to BMA (around 5% improvement in α values in most of the lead-times in both basins). Also, in the Big East River watershed, except for 24 hours ahead forecasts, where the difference between the reliability of BMA and M-EnBMA is negligible, the percent improvement of α for forecasting large precipitation values (more than 5 mm) are always more than 10% (Figure 6-12). The same conclusion derived from the Black River watershed where the implementation of M-EnBMA leads to 21%, 27%, 15%, and 18% improvement in the reliability of 6-, 12-, 18- and 24-ahead forecasts of more than 5 mm precipitation values (Figure 6-13). It is worthy of note that all predictive reliability plots fell above the uniform line, which indicates that both BMA and M-EnBMA possess negative biases by under-predicting the precipitation values at different lead-times. These under-estimations are more pronounced by focusing on larger values (Figure 6-13), which stem from the presence of the negative biases in all individual precipitation forecasts in both basins.

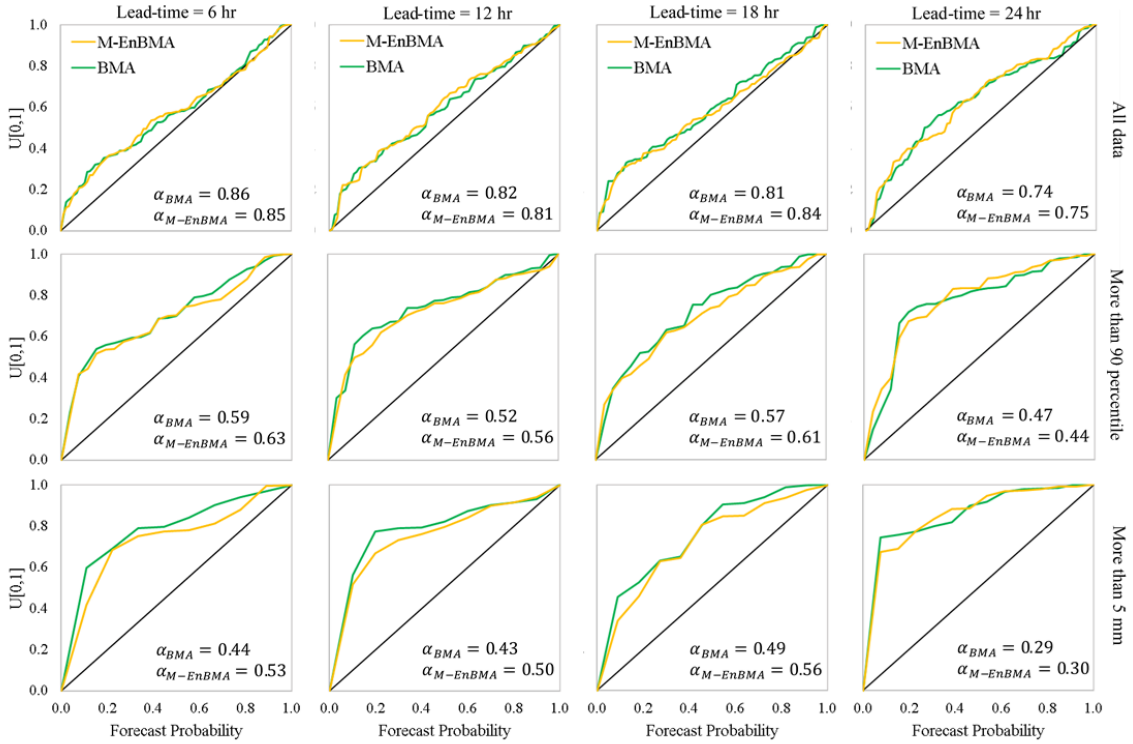


Figure 6-12 The reliability plot and their corresponding α values at different forecasting horizons (6 to 24 hours ahead), derived from both BMA and modified M-EnBMA (M-EnBMA) results regarding the whole time-series, values more than 90 percentile, and values more than 5 mm in the Big East River watershed

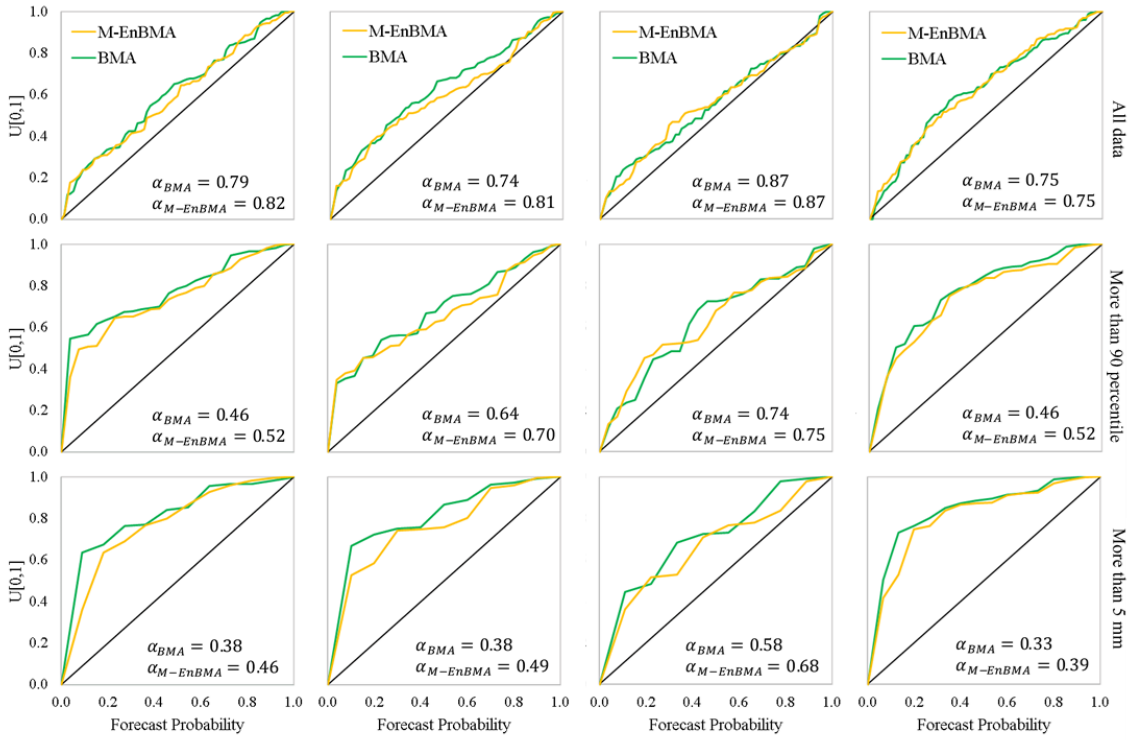


Figure 6-13 The reliability plot and their corresponding α values at different forecasting horizons (6 to 24 hours ahead), derived from both BMA and modified En-BMA (M-EnBMA) results regarding the whole time-series, values more than 90 percentile, and values more than 5 mm in the Black River watershed

6.6 Summary and Conclusion

Bayesian Model Averaging (BMA) is one of the most common post-processing approaches in hydro-meteorological studies, which uses the weighted average of predictive distributions based on individual members to produce probabilistic forecasts. As the BMA formulation is based on the law of total probability, it requires independent members (mutually exclusive) with a high capability of capturing the future variability (collectively exhaustive) for producing more accurate forecasts. In this study, the modified version of the entropy-based BMA (En-BMA) method is proposed for precipitation forecasting where

a modified entropy-based selection procedure is used within the BMA structure in order to select the optimal subset of forecasts by keeping the information of the ensemble in the highest possible level with the lowest redundancy. Considering seven different 6-hourly accumulated precipitation forecasts up to 24 hours ahead in two study areas, in Ontario, Canada, this study assessed the effects of the proposed modifications and provide a comprehensive comparison between the modified En-BMA and BMA in generating predictive precipitation forecasts.

From the initial comparison of individual forecasts, we found that although REPS can be considered as the most robust precipitation forecasts in both study regions, it is not possible to select one model as the best one for all forecasting horizons and locations, confirming the advantage of possessing an ensemble of multi-model forecasts. Moreover, the results indicate that applying the proposed modifications in the En-BMA structure tended to improve the accuracy of the forecasts in all forecasting horizons and both watersheds, compared to the original En-BMA. The modified entropy-based selection procedure constructs the optimal subset of the ensemble with a lower number of members (i.e. lower redundant information), compared to the original one, while the information content is at the same level. This provides a better balance between the mutually exclusive and collectively exhaustive properties and consequently leads to better results. Lastly, we compared the forecasts from the modified En-BMA and the BMA methods. Considering the whole forecast time series, we found that the modified En-BMA provides slightly more accurate precipitation forecasts while in terms of reliability, both methods possess the same performance. However, by focusing on large precipitation values, which are receiving

particular attention in hydrology, there is a significant advantage of implementing the modified En-BMA method over the BMA approach for generating more reliable and accurate probabilistic precipitation forecasts.

As the structure of the proposed entropy-based selection procedure is not limited to variables with any specific characteristics, the findings of this research can be generalized to other future studies where the application of the BMA with point-mass and a gamma distribution is reliable. Covering watersheds with various climatologic conditions and forecasts with shorter (e.g. hourly) or longer (e.g. daily) temporal resolution is recommended in future applications of the proposed En-BMA method.

6.7 Acknowledgment

This study was funded by the Natural Science and Engineering Research Council (NSERC) of Canada, grant NSERC Canadian Floodnet (NETGP-451456).

6.8 Data Availability Statement

The archive of the Canadian Precipitation Analysis (CaPA), Regional Ensemble Prediction System (REPS), Regional Deterministic Prediction System (RDPS), Global Ensemble Precipitation System (GEPS), Global Deterministic Prediction System (GDPS), and High-resolution Deterministic Prediction System (HRDPS) can be obtained from The Canadian Surface Prediction Archive (CaSPAr) (Mai et al., 2019). The archive of the Global Ensemble Forecasting System Reforecast Version 2 (GEFS) and the Global Forecast System (GFS) archived forecasts can be respectively obtained from National Oceanic and

Atmospheric Administration, Physical Science Laboratory (https://psl.noaa.gov/) and National Centre for Atmospheric Research (https://rda.ucar.edu/) websites.

6.9 References

Abaza, M., Anctil, F., Fortin, V., & Turcotte, R. (2013). A Comparison of the Canadian Global and Regional Meteorological Ensemble Prediction Systems for Short-Term Hydrological Forecasting. *Monthly Weather Review*, *141*(10), 3462–3476. https://doi.org/10.1175/MWR-D-12-00206.1

Alfonso, L., He, L., Lobbrecht, A., & Price, R. (2013). Information theory applied to evaluate the discharge monitoring network of the Magdalena River. *Journal of Hydroinformatics*, *15*(1), 211–228. https://doi.org/10.2166/hydro.2012.066

Aminyavari, S., & Saghafian, B. (2019). Probabilistic streamflow forecast based on spatial post-processing of TIGGE precipitation forecasts. *Stochastic Environmental Research and Risk Assessment*, *33*(11), 1939–1950. https://doi.org/10.1007/s00477-019-01737-4

Coulibaly, P., Haché, M., Fortin, V., & Bobée, B. (2005). Improving Daily Reservoir Inflow Forecasts with Model Combination. *Journal of Hydrologic Engineering*, *10*(2), 91–99. https://doi.org/10.1061/(ASCE)1084-0699(2005)10:2(91)

Cuo, L., Pagano, T. C., & Wang, Q. J. (2011). A Review of Quantitative Precipitation Forecasts and Their Use in Short- to Medium-Range Streamflow Forecasting. *Journal of Hydrometeorology*, *12*(5), 713–728. https://doi.org/10.1175/2011JHM1347.1

Darbandsari, P., & Coulibaly, P. (2019). Inter-Comparison of Different Bayesian Model Averaging Modifications in Streamflow Simulation. *Water*, *11*(8), 1707. https://doi.org/10.3390/w11081707

Darbandsari, P., & Coulibaly, P. (2020a). Introducing entropy-based Bayesian model averaging for streamflow forecast. *Journal of Hydrology*, *591*, 125577. https://doi.org/10.1016/j.jhydrol.2020.125577

Darbandsari, P., & Coulibaly, P. (2020b). Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: Case study of Northern Ontario, Canada. *Journal of Hydrology: Regional Studies*, *31*, 100730. https://doi.org/10.1016/j.ejrh.2020.100730

Eum, H.-I., Dibike, Y., Prowse, T., & Bonsal, B. (2014). Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the Athabasca Watershed, Canada. *Hydrological Processes*, *28*(14), 4250–4271. https://doi.org/10.1002/hyp.10236

Fraley, C., Raftery, A. E., & Gneiting, T. (2010). Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging. *Monthly Weather Review*, *138*(1), 190–202. https://doi.org/10.1175/2009MWR3046.1

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, *14*(4), 382–401. JSTOR.

Jha, S. K., Shrestha, D. L., Stadnyk, T. A., & Coulibaly, P. (2018). Evaluation of ensemble precipitation forecasts generated through post-processing in a Canadian catchment. *Hydrology and Earth System Sciences*, *22*(3), 1957–1969. https://doi.org/10.5194/hess-22-1957-2018

Ji, L., Zhi, X., Zhu, S., & Fraedrich, K. (2019). Probabilistic Precipitation Forecasting over East Asia Using Bayesian Model Averaging. *Weather and Forecasting*, *34*(2), 377–392. https://doi.org/10.1175/WAF-D-18-0093.1

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Keum, J., Awol, F. S., Ursulak, J., & Coulibaly, P. (2019). Introducing the Ensemble-Based Dual Entropy and Multiobjective Optimization for Hydrometric Network Design Problems: EnDEMO. *Entropy*, *21*(10), 947. https://doi.org/10.3390/e21100947

Keum, J., & Coulibaly, P. (2017). Information theory-based decision support system for integrated design of multivariable hydrometric networks. *Water Resources Research*, *53*(7), 6239–6259. https://doi.org/10.1002/2016WR019981

Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, *11*, 1267–1277.

Leach, J. M., Kornelsen, K. C., Samuel, J., & Coulibaly, P. (2015). Hydrometric network design using streamflow signatures and indicators of hydrologic alteration. *Journal of Hydrology*, *529*, 1350–1359. https://doi.org/10.1016/j.jhydrol.2015.08.048

Lespinas, F., Fortin, V., Roy, G., Rasmussen, P., & Stadnyk, T. (2015). Performance Evaluation of the Canadian Precipitation Analysis (CaPA). *Journal of Hydrometeorology*, *16*(5), 2045–2064. https://doi.org/10.1175/JHM-D-14-0191.1

Li, C., Singh, V. P., & Mishra, A. K. (2012). Entropy theory-based criterion for hydrometric network evaluation and design: Maximum information minimum redundancy. *Water Resources Research*, *48*(5). https://doi.org/10.1029/2011WR011251

Liu, J., & Xie, Z. (2014). BMA Probabilistic Quantitative Precipitation Forecasting over the Huaihe Basin Using TIGGE Multimodel Ensemble Forecasts. *Monthly Weather Review*, *142*(4), 1542–1555. https://doi.org/10.1175/MWR-D-13-00031.1

Ma, Y., Hong, Y., Chen, Y., Yang, Y., Tang, G., Yao, Y., Long, D., Li, C., Han, Z., & Liu, R. (2018). Performance of Optimally Merged Multisatellite Precipitation Products Using the Dynamic Bayesian Model Averaging Scheme Over the Tibetan Plateau. *Journal of Geophysical Research: Atmospheres*, *123*(2), 814–834. https://doi.org/10.1002/2017JD026648

Madadgar, S., & Moradkhani, H. (2014). Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resources Research*, *50*(12), 9586–9603. https://doi.org/10.1002/2014WR015965

Mahfouf, J.-F., Brasnett, B., & Gagnon, S. (2007). A Canadian precipitation analysis (CaPA) project: Description and preliminary results. *Atmosphere-Ocean*, *45*(1), 1–17. https://doi.org/10.3137/ao.v450101

Mishra, A. K., & Coulibaly, P. (2009). Developments in hydrometric network design: A review. *Reviews of Geophysics*, *47*(2). https://doi.org/10.1029/2007RG000243

Qu, B., Zhang, X., Pappenberger, F., Zhang, T., & Fang, Y. (2017). Multi-Model Grand Ensemble Hydrologic Forecasting in the Fu River Basin Using Bayesian Model Averaging. *Water*, *9*(2), 74. https://doi.org/10.3390/w9020074

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Refsgaard, J. C., Christensen, S., Sonnenborg, T. O., Seifert, D., Højberg, A. L., & Troldborg, L. (2012). Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Advances in Water Resources*, *36*, 36–50. https://doi.org/10.1016/j.advwatres.2011.04.006

Robertson, D. E., Shrestha, D. L., & Wang, Q. J. (2013). Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. *Hydrology and Earth System Sciences*, *17*(9), 3587–3603. https://doi.org/10.5194/hess-17-3587-2013

Saedi, A., Saghafian, B., Moazami, S., & Aminyavari, S. (2020). Performance evaluation of sub-daily ensemble precipitation forecasts. *Meteorological Applications*, *27*(1), e1872. https://doi.org/10.1002/met.1872

Scheuerer, M., & Hamill, T. M. (2015). Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions. *Monthly Weather Review*, *143*(11), 4578–4596. https://doi.org/10.1175/MWR-D-15-0061.1

Schmeits, M. J., & Kok, K. J. (2010). A Comparison between Raw Ensemble Output, (Modified) Bayesian Model Averaging, and Extended Logistic Regression Using ECMWF Ensemble Precipitation Reforecasts. *Monthly Weather Review*, *138*(11), 4199–4211. https://doi.org/10.1175/2010MWR3285.1

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sharma, S., Siddique, R., Reed, S., Ahnert, P., & Mejia, A. (2019). Hydrological Model Diversity Enhances Streamflow Forecast Skill at Short- to Medium-Range Timescales. *Water Resources Research*, *55*(2), 1510–1530. https://doi.org/10.1029/2018WR023197

Shrestha, D. L., Robertson, D. E., Wang, Q. J., Pagano, T. C., & Hapuarachchi, H. a. P. (2013). Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose. *Hydrology and Earth System Sciences*, *17*(5), 1913–1931. https://doi.org/10.5194/hess-17-1913-2013

Singh, V. P. (1997). The Use of Entropy in Hydrology and Water Resources. *Hydrological Processes*, *11*(6), 587–626. https://doi.org/10.1002/(SICI)1099-1085(199705)11:6<587::AID-HYP479>3.0.CO;2-P

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., & Fraley, C. (2007). Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, *135*(9), 3209–3220. https://doi.org/10.1175/MWR3441.1

Steenbergen, N. V., & Willems, P. (2014). *Quantification Of Rainfall Forecast Uncertainty And Its Impact On Flood Forecasting*. 9.

Taillardat, M., Mestre, O., Zamo, M., & Naveau, P. (2016). Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics. *Monthly Weather Review*, *144*(6), 2375–2393. https://doi.org/10.1175/MWR-D-15-0260.1

Thiessen, A. H. (1911). Precipitation averages for large areas. *Monthly Weather Review*, *39*(7), 1082–1089. https://doi.org/10.1175/1520-0493(1911)39<1082b:PAFLA>2.0.CO;2

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Verkade, J. S., Brown, J. D., Reggiani, P., & Weerts, A. H. (2013). Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, *501*, 73–91. https://doi.org/10.1016/j.jhydrol.2013.07.039

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., & Gneiting, T. (2018). Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall over Northern Tropical Africa. *Weather and Forecasting*, *33*(2), 369–388. https://doi.org/10.1175/WAF-D-17-0127.1

Vrugt, J. A., Diks, C. G. H., & Clark, M. P. (2008). Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling. *Environmental Fluid Mechanics*, *8*(5), 579–595. https://doi.org/10.1007/s10652-008-9106-3

Vrugt, J. A., & Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004838

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079

Xu, J., Anctil, F., & Boucher, M.-A. (2019). Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems. *Journal of Hydrology*, *578*, 124002. https://doi.org/10.1016/j.jhydrol.2019.124002

Yang, C., Yan, Z., & Shao, Y. (2012). Probabilistic precipitation forecasting based on ensemble output using generalized additive models and Bayesian model averaging. *Acta Meteorologica Sinica*, *26*(1), 1–12. https://doi.org/10.1007/s13351-012-0101-8

Zhu, J., Kong, F., Ran, L., & Lei, H. (2015). Bayesian Model Averaging with Stratified Sampling for Probabilistic Quantitative Precipitation Forecasting in Northern China during Summer 2010. *Monthly Weather Review*, *143*(9), 3628–3641. https://doi.org/10.1175/MWR-D-14-00301.1

Chapter 7. Conclusions and Recommendations

7.1 Conclusions

This thesis was mainly focused on the development of a novel probabilistic ensemble streamflow forecasting framework to reliably quantify and reduce predictive uncertainty by using the full potential of the forecasts ensemble. The study areas in this research cover the Big East River and Black River watersheds as snow-dominated data-poor catchments situated in Northern Ontario, Canada. First, after developing and evaluating various conceptual hydrologic models and examining different variants of the Bayesian Model Averaging (BMA) method, the promising Entropy-based BMA (En-BMA) post-processing approach was proposed for enhanced probabilistic streamflow forecasting where entropy theory concepts are used for relaxing the remaining limitations of the BMA method. Then, by taking advantage of both BMA and the Hydrologic Uncertainty Processor (HUP) methods, an ensemble-based Bayesian post-processor was developed for better quantifying the predictive uncertainty in the context of streamflow forecasting. Last, the proposed En-BMA method was modified to be used as a post-processor for enhancing ensemble precipitation forecasts. The outcomes of this study are expected to benefit the hydrology community at large, the operational streamflow forecasting centers, and could also be integrated into any flood forecasting and early warning systems. The main findings of the five thesis chapters are summarized as follows:

7.1.1 Conceptual hydrologic models in data-scare regions

- Using various objective functions for calibrating parameters and estimating the mean areal precipitation in different ways possess heterogeneous effects on the performances of various conceptual hydrologic models. So, considering these effects in any model inter-comparison process is required and leads to more consistent findings and conclusions.
- Comparing five different objective functions shows the advantages of using Nash Volume Error (NVE) and Kling Gupta Efficiency (KGE) for calibrating models' parameters used for continuous daily streamflow simulation.
- Among seven structurally different conceptual models, MACHBV provides the most robust streamflow predictions by providing relatively accurate results regarding low, medium, and high flows in data-poor watersheds. Focusing on high flows indicates the competitive performances of the GR4J model.
- HEC-HMS based models lead to the relatively worst results. The poor performances of three HEC-HMS based models, especially regarding low flows, may stem from the poor estimation of the base flow and the use of pre-specified monthly PET in their structures.
- Incorporation of the snowmelt method in hydrologic modeling of snow-dominated watersheds is a necessary task. However, using a more complex snow routine does not always lead to better streamflow predictions. The advantages of using more complex snowmelt methods depend on the structure of the hydrologic model, which shows the importance of comparing different snowmelt approaches regardless of

their complexity, to evaluate their compatibility with the selected rainfall-runoff model.

- The reliability and suitability of the CaPA precipitation data to be considered as forcing inputs of the hydrologic models in regions with sparse ground-based measurements was confirmed in Northern Ontario, Canada. The calibrated models using CaPA data lead to better streamflow simulation compared to the ones being calibrated based on the ground-based observations.

7.1.2 Bayesian Model Averaging method for streamflow simulation/forecasting

- Bayesian Model Averaging is a well-known probabilistic ensemble-based post-processing approach taking the advantages of multiple forecasts for reliably quantifying the predictive uncertainty and generating probabilistic results.
- As the main input of the BMA method, the ensemble of streamflow simulations has a direct impact on the accuracy and reliability of the BMA derived predictive results. Besides the simulation skill of different models, the diversity of the ensemble members and the capability of the ensemble, as a whole, to capture the observational variability are the important features to enhance the results. In the context of streamflow simulation in data-scarce regions, using multiple inputs and multiple models for constructing members of an ensemble leads to better BMA results.
- Among various modifications being proposed to enhance the BMA method dealing with streamflow data, implementation of the heteroscedastic variance improves the BMA predictive performance. Also, incorporating data transformation procedure,

in general, improves the reliability while there are some concerns about the sharpness of the BMA predictive results, especially regarding high flows. Simultaneous application of non-constant variance and data transformation modifications is not recommended as it significantly deteriorates the sharpness of the probabilistic results.

- The expectation-maximization algorithm is an efficient method for reliably estimating BMA parameters and there is no need for using a more complex global optimization approach.

7.1.3 Entropy-based BMA for probabilistic streamflow forecasting

- An Entropy-based Bayesian Model Averaging (En-BMA) approach, integrating entropy theory concepts and BMA, is proposed to provide enhanced probabilistic streamflow forecasts.
- The combination of multiple objective functions and multiple hydrologic models generates a more diverse ensemble of streamflow forecasts with a higher capability of capturing future possibilities and yields more reliable and accurate BMA probabilistic results.
- By keeping the information content while reducing the redundancy within the ensemble, the proposed entropy-based selection procedure narrows down the streamflow forecasts for providing a balance between the mutually exclusive and collectively exhaustive properties which are the inherent assumptions of the BMA method.

- Although both BMA and En-BMA methods generate comparable streamflow forecasts in general, different performance evaluation metrics show the superiority of the proposed En-BMA method over BMA, for providing better probabilistic and deterministic high flow forecasts. The advantages of En-BMA can be seen in all forecasting horizons, while it is less obvious as lead time increases.
- The proposed entropy-based selection method is a non-parametric procedure without any assumptions about the distributions of the variable. So, it can be integrated with any variant of the BMA method to be suitably used for other types of variables (e.g. precipitation, soil moisture, and snowmelt).

7.1.4 HUP-embedded BMA for streamflow probabilistic forecasting

- HUP is a statistical post-processing approach that relied on a deterministic streamflow forecast for quantifying predictive uncertainty and providing probabilistic results. The effects of the forecasting skills of the deterministic hydrologic model on the HUP predictive performance are negligible in terms of low flow forecasting. However, by focusing on high flows, using a better performing rainfall-runoff model yields significantly better HUP-derived probabilistic forecasts, especially in longer lead times.
- Making the most of the respective strengths of both HUP and BMA methods, the HUP embedded BMA (HUP-BMA) was proposed to provide more reliable and accurate probabilistic streamflow forecasts by implicitly taking into account the effects of initial conditions and taking the advantages of considering multiple forecasts.

- The superiority of the proposed HUP-BMA method over both HUP and BMA approaches for short-range streamflow forecasts is shown by various deterministic and probabilistic performance metrics. However, increasing lead-time and reducing the dependence of the actual flow with the initial observation lead to the deterioration of the HUP-BMA performance. In the latter case, removing the dependence of the HUP-BMA formulation on initial flow values is beneficial and improves the ability of the method to generate more reliable and accurate probabilistic forecasts.

7.1.5 Modified Entropy-based BMA for probabilistic precipitation forecasting

- The modified entropy-based selection procedure is proposed to better meet the mutually exclusive and collectively exhaustive requirements of the BMA method. The implementation of the proposed selection procedure within the structure of the BMA variant for precipitation leads to the modified version of the Entropy-based BMA method for probabilistic precipitation forecasting.
- The proposed modifications are required to make the En-BMA method suitable for reliable and accurate ensemble-based probabilistic precipitation forecasting. As indicated by various performance measurements, the modified En-BMA provides more accurate post-processed precipitation forecasts in all lead times, compared to the original En-BMA.
- Various deterministic evaluation metrics based on the whole forecasting period show slight improvements of the results in the case of applying the modified En-BMA compared with BMA, while both methods produce comparable probabilistic

forecasts in term of reliability scores. By focusing on high precipitation values, however, the advantages of implementing the modified En-BMA over BMA are noticeable in terms of generating both deterministic and probabilistic sub-daily precipitation forecasts.

7.1.6 General Conclusions

The general conclusions and contributions of this thesis are as follows:

- MACHBV and GR4J are the most robust rainfall-runoff models for simulating the hydrological processes of watersheds with low data availability in northern Ontario watersheds.
- The archive of CaPA data is a reliable source of precipitation in Northern Ontario, Canada, which can be used as an alternative forcing inputs of rainfall-runoff models in regions with sparse ground-based measurements.
- The proposed entropy-based BMA post-processing approach yields enhanced probabilistic high flow forecasting by relaxing the assumption of having a mutually exclusive and collectively exhaustive ensemble of forecasts.
- The modified entropy-based BMA method is proposed for generating enhanced probabilistic precipitation forecasts, which shows large forecast improvement on high precipitation values.
- Short-term streamflow forecasting using the proposed HUP-embedded BMA post-processor is more reliable and accurate than both HUP and BMA methods.

- Formulating HUP-BMA without considering the effects of initial values is identified to produce better probabilistic medium-range streamflow forecasts especially in watersheds with a short time of concentration.

7.2 Recommendations for Future Research

This thesis is a primary step in the process of developing a reliable probabilistic streamflow forecasting framework to be used in the Canadian Adaptive Flood Forecasting and Early Warning System (CAFFEWS). Therefore, great efforts are still required to achieve this objective. One of the primary sources of uncertainty in hydrologic modeling is the forcing inputs, precipitation in particular. Apart from the unknown future precipitation, there are errors in observation data that can significantly affect the quality of the streamflow simulation/forecasting process. This effect is more noticeable in regions without a dense ground-based meteorological network (Sirisena et al., 2018; Tegegne et al., 2017; Worqlul et al., 2017). Therefore, besides CaPA, which is identified as a reliable alternative source of precipitation, there are some other ways of deriving precipitation data, such as the radar-based and satellite-based precipitation estimates, that should be deeply evaluated as forcing inputs of hydrologic models in data-poor watersheds in Canada.

The proposed entropy-based selection procedure is the initial step of using the entropy concept for overcoming the mutually exclusive and collectively exhaustive assumptions of the BMA approach. Further improvements can be achieved by revising the proposed selection structure or implementing other entropy measures, such as Net Information (Markus et al., 2003). As another interesting topic, following up on the last study of this

thesis (Chapter 6), the segregated structure of the proposed entropy-based selection procedure encourage future studies to investigate the combination of the proposed selection method with various BMA variant, such as Copula-embedded BMA (Madadgar & Moradkhani, 2014).

As shown in Chapters 3 and 4 of this thesis, the proposed En-BMA and HUP-BMA methods produce reliable post-processed streamflow forecasts, however, a more comprehensive evaluation has yet to be performed in various types of watersheds, in terms of size, climatologic conditions, and geography. The additional interesting research topic would be the integration of the proposed En-BMA and HUP-BMA methods to take advantage of both systems for generating enhanced ensemble-based probabilistic streamflow forecasts. The proper combination of these two methods can reduce the two important limitations of the BMA approach by considering a mutually exclusive and collectively exhaustive ensemble of forecasts in conjunction with the effects of initial flow values and could lead to a better quantification of predictive uncertainty associated with streamflow forecasting. Also, it is worth to examine the implementation of a reliable bias-correction method prior to the application of either HUP-BMA and En-BMA methods as recent studies showed the advantages of using bias-corrected forecasts in ensemble-based post-processing approaches (e.g. Han & Coulibaly, 2019; Sharma et al., 2019).

Moreover, the unknown future meteorological forcing input of hydrological models is one of the main sources of uncertainty within streamflow forecasting and it is required to be reduced and quantified in any operational flood forecasting framework. Generating an ensemble of streamflow forecasts using multiple precipitation products and multiple

hydrologic models is one way of accounting for the aforementioned issue; and has recently received particular interests in the context of Bayesian Model Averaging (e.g. Awol et al., 2021; Roy et al., 2017; Sharma et al., 2019; Xu et al., 2019). Therefore, as another topic of considerable interest, the application of the proposed ensemble-based Bayesian post-processing methods (i.e. En-BMA and HUP-BMA) in conjunction with multi-input multi-model streamflow forecasts as a platform for probabilistic streamflow forecasting could be evaluated. Lastly, an advanced Bayesian ensemble probabilistic streamflow forecasting framework could be developed by taking the advantage of probabilistic post-processing of both precipitation and streamflow forecasts. The appropriate implementation of the well-post-processed probabilistic precipitation forecasts in a multi-model streamflow forecasting method can greatly reduce and quantify the predictive uncertainty and lead to enhanced streamflow forecast. The proposed En-BMA method would be a good choice for developing the aforementioned streamflow forecasting framework as it is a reliable approach in the context of both streamflow and precipitation data.

7.3 References

Awol, F. S., Coulibaly, P., & Tsanis, I. (2021). Identification of Combined Hydrological Models and Numerical Weather Predictions for Enhanced Flood Forecasting in a Semiurban Watershed. *Journal of Hydrologic Engineering*, *26*(1), 04020057. https://doi.org/10.1061/(ASCE)HE.1943-5584.0002018

Han, S., & Coulibaly, P. (2019). Probabilistic Flood Forecasting Using Hydrologic Uncertainty Processor with Ensemble Weather Forecasts. *Journal of Hydrometeorology*, *20*(7), 1379–1398. https://doi.org/10.1175/JHM-D-18-0251.1

Madadgar, S., & Moradkhani, H. (2014). Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resources Research*, *50*(12), 9586–9603. https://doi.org/10.1002/2014WR015965

Markus, M., Vernon Knapp, H., & Tasker, G. D. (2003). Entropy and generalized least square methods in assessment of the regional value of streamgages. *Journal of Hydrology*, *283*(1), 107–121. https://doi.org/10.1016/S0022-1694(03)00244-0

Roy, T., Serrat-Capdevila, A., Gupta, H., & Valdes, J. (2017). A platform for probabilistic Multimodel and Multiproduct Streamflow Forecasting. *Water Resources Research*, *53*(1), 376–399. https://doi.org/10.1002/2016WR019752

Sharma, S., Siddique, R., Reed, S., Ahnert, P., & Mejia, A. (2019). Hydrological Model Diversity Enhances Streamflow Forecast Skill at Short- to Medium-Range Timescales. *Water Resources Research*, *55*(2), 1510–1530. https://doi.org/10.1029/2018WR023197

Sirisena, T. A. J. G., Maskey, S., Ranasinghe, R., & Babel, M. S. (2018). Effects of different precipitation inputs on streamflow simulation in the Irrawaddy River Basin, Myanmar. *Journal of Hydrology: Regional Studies*, *19*, 265–278. https://doi.org/10.1016/j.ejrh.2018.10.005

Tegegne, G., Park, D. K., & Kim, Y.-O. (2017). Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. *Journal of Hydrology: Regional Studies*, *14*, 49–66. https://doi.org/10.1016/j.ejrh.2017.10.002

Worqlul, A. W., Yen, H., Collick, A. S., Tilahun, S. A., Langan, S., & Steenhuis, T. S. (2017). Evaluation of CFSR, TMPA 3B42 and ground-based rainfall data as input for hydrological models, in data-scarce regions: The upper Blue Nile Basin, Ethiopia. *CATENA*, *152*, 242–251. https://doi.org/10.1016/j.catena.2017.01.019

Xu, J., Anctil, F., & Boucher, M.-A. (2019). Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems. *Journal of Hydrology*, *578*, 124002. https://doi.org/10.1016/j.jhydrol.2019.124002