IDENTIFICATION OF HYDROLOGIC MODELS, INPUTS, AND CALIBRATION APPROACHES FOR ENHANCED FLOOD FORECASTING

IDENTIFICATION OF HYDROLOGIC MODELS, INPUTS, AND CALIBRATION APPROACHES FOR ENHANCED FLOOD FORECASTING

By FREZER SEID AWOL, B.Sc., M.Sc.

Faculty of Engineering

Civil Engineering

A Thesis Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

October 2019

DOCTOR OF PHILOSOPHY (2019) McMaster University

Civil Engineering Hamilton, Ontario

TITLE: Identification of Hydrologic Models, Inputs, and Calibration Approaches for Enhanced Flood Forecasting

AUTHOR: Frezer Seid Awol,

B.Sc. (Bahir Dar University)

M.Sc. (University of Stuttgart)

SUPERVISOR: Dr. Paulin Coulibaly

CO-SUPERVISOR: Dr. Ioannis Tsanis

NUMBER OF PAGES: xxvii, pp. 255

Lay Abstract

Accurate hydrological models and inputs play essential roles in creating a successful flood forecasting and early warning system. The main objective of this research is to identify adequately calibrated hydrological models and skillful weather forecast inputs to improve the accuracy of hydrological forecasting in various watershed landscapes. The key contributions include: (1) A finding that a combination of efficient optimization tools with a series of calibration steps is essential in obtaining representative parameters sets of hydrological models; (2) Simple lumped hydrological models, if used appropriately, can provide accurate and reliable hydrological forecasts in different watershed types, besides being computationally efficient; and (3) Candidate weather forecast products identified in Canada's diverse geographical regions can be used as inputs to hydrological models for improved flood forecasting. The findings from this thesis are expected to benefit hydrological forecasting centers and researchers working on model and input improvements.

Abstract

The primary goal of this research is to evaluate and identify proper calibration approaches, skillful hydrological models, and suitable weather forecast inputs to improve the accuracy and reliability of hydrological forecasting in different types of watersheds. The research started by formulating an approach that examined single- and multi-site, and single- and multi-objective optimization methods for calibrating an event-based hydrological model to improve flood prediction in a semi-urban catchment. Then it assessed whether reservoir inflow in a large complex watershed could be accurately and reliably forecasted by simple lumped, medium-level distributed, or advanced land-surface based hydrological models. Then it is followed by a comparison of multiple combinations of hydrological models and weather forecast inputs to identify the best possible model-input integration for an enhanced short-range flood forecasting in a semi-urban catchment. In the end, Numerical Weather Predictions (NWPs) with different spatial and temporal resolutions were evaluated across Canada's varied geographical environments to find candidate precipitation input products for improved flood forecasting.

Results indicated that aggregating the objective functions across multiple sites into a single objective function provided better representative parameter sets of a semi-distributed hydrological model for an enhanced peak flow simulation. Proficient lumped hydrological models with proper forecast inputs appeared to show better hydrological forecast performance than distributed and land-surface models in two distinct watersheds. For example, forcing the simple lumped model (SACSMA) with bias-corrected ensemble

inputs offered a reliable reservoir inflow forecast in a sizeable complex Prairie watershed; and a combination of the lumped model (MACHBV) with the high-resolution weather forecast input (HRDPS) provided skillful and economically viable short-term flood forecasts in a small semi-urban catchment. The comprehensive verification has identified low-resolution NWPs (GEFSv2 and GFS) over Western and Central parts of Canada and high-resolution NWPs (HRRR and HRDPS) in Southern Ontario regions that have a promising potential for forecasting the timing, intensity, and volume of floods.

Acknowledgment

I would like to express my sincere gratitude to my supervisor Prof. Paulin Coulibaly for providing me this Ph.D. opportunity and for his continuous guidance, motivation, and remarkable contributions over the past four years. His encouragement and advice help me a lot to overcome stressful periods at many stages of my study. It would have been impossible to come to this stage without his patience and guidance.

I am very grateful to my co-supervisor, Prof. Ioannis Tsanis, for his tremendous mentorship and advice in my research. The support and contributions I received from him made my study the most profitable experience. I would like to extend my heartful gratitude to the other members of my Ph.D. supervisory committee Prof. Sarah Dickson and Prof. Yiping Guo for providing me invaluable insights and suggestions that have improved my research. Many special thanks to Prof. Bryan Tolson for supporting me on DDS/PADDS optimization and modeling in general.

I would like to sincerely thank Dr. Yonas Dibike for his encouragement to commence my Ph.D. in Canada, for his invaluable support and his advice in moments of obstacles. I am thankful to Dr. Fisaha Unduche, Director at Manitoba Infrastructure, Hydrologic Forecasting Center (MI-HFC), for hosting me as an intern and providing endless supports. Many thanks to the Toronto Region Conservation Authority (TRCA) for hosting me as an intern and providing me valuable technical supports. Special thanks to the US University Corporation for Atmospheric Research (UCAR-NCAR) for providing me an internship and training opportunities. I want to thank Dr. Dilnesaw Chekol at TRCA for sharing his

precious time and co-operation. My gratitude also extends to my friend Dr. Habtamu Tolossa at MI-HFC, for his astonishing support.

I am indebted to all my colleagues at McMaster Water Resources and Hydrologic Modeling Lab (WRHML) for sharing valuable times and moments, which made our graduate life exciting and fun. I owe a big thanks to James Leach for his help in proof-reading my first paper, and for bringing knowledgeable and entertaining discussion topics in the lab.

I would like to acknowledge Environment and Climate Change Canada (ECCC), Water Survey of Canada (WSC), Canadian Surface Prediction Archive (CaSPAr), MI-HFC, TRCA, MESONET, ECMWF, NOAA National Weather Service (NWS), Computational Hydraulics International (CHI), and Prof. Altaf Arain for providing me data and models at various stages of my research. This research was made possible by the financial and official supports from the Natural Sciences and Engineering Research Council of Canada (NSERC), the School of Graduate Studies (SGS), and the Department of Civil Engineering at McMaster University. Computing resources for this research were, in part, provided by Compute Canada (www.computecanada.ca).

I am genuinely thankful to my mother and all my family members in Ethiopia for their immense supports, unconditional love, and patience. A big thank you to my friends for their supports. I am thankful to my son Eliam who made me stronger, better, and more dedicated in every single moment. My eternal gratitude to my late father, who encouraged me to strive for the most. Last and foremost, thanks to the almighty God to his countless blessings.

Table of Content

Chapter 1. Introduction..... 1

1.1. Deterministic and ensemble flood forecasting..... 1

1.2. Hydrological models and weather forecast inputs..... 3

1.3. Challenges in hydrological model calibration 7

1.4. Scope of the research 10

1.5. Research objectives and Thesis outline 12

1.6. References..... 14

Chapter 2. Event-based model calibration approaches for selecting representative distributed parameters in semi-urban watersheds..... 23

2.1. Abstract..... 24

2.2. Introduction..... 26

2.3. Study area and data 32

2.4. Methods 35

2.4.1. Model setup..... 35

2.4.2. Sensitivity analysis 41

2.4.3. Spatial and temporal parameter variability..... 43

2.4.4. Model Calibration..... 44

2.4.5. Validation..... 52

2.5. Results and discussion 53

2.5.1. Sensitivity analysis 53

2.5.2. Spatial and temporal parameter variability..... 56

2.5.3. Calibration Approaches 60

2.5.4. Validation 67
2.6. Conclusion 68
2.7. Acknowledgments 70
2.8. References 71
Chapter 3. Identification of hydrological models for enhanced ensemble reservoir inflow forecasting in a large complex Prairie watershed 78
3.1. Abstract 79
3.2. Introduction 81
3.3. Materials 84
3.3.1. Study area 84
3.3.2. Data 86
3.4. Method 88
3.4.1. Hydrological Models 89
3.4.2. Calibration and Validation 92
3.4.3. Bias correction 93
3.4.4. Hindcast simulation (model update and forecast) 94
3.4.5. Ensemble forecast verification 95
3.5. Results 99
3.5.1. Calibration and validation 99
3.5.2. Model comparison in forecast mode 102
3.6. Conclusion and Discussion 112
3.7. Author Contributions 118
3.8. Funding 118

3.9. Acknowledgments 119

3.10. Conflicts of Interest 119

Appendix A: Brief description of the calibration of models 120

A.1. SNOW17 120

A.2. SACSMA 121

A.3. MACHBV 123

A.4. VIC 124

3.11. References 128

Chapter 4. Identification of combined hydrological models and numerical weather predictions for enhanced flood forecasting in a semi-urban watershed 137

4.1. Abstract 138

4.2. Introduction 139

4.3. Study Area and Data 144

4.3.1. Study Area 144

4.3.2. Data 145

4.4. Methodology 147

4.4.1. Hydrological models 149

4.4.2. Calibration and Validation 152

4.4.3. Hindcast simulation (model update and forecast) 154

4.4.4. Hydrological and flood forecasting performance evaluation 155

4.4.5. Simple forecast averaging methods 156

4.5. Results and Discussion 158

4.5.1. Calibration and Validation 158

4.5.2. Screening weather forecast inputs 159

4.5.3. Comparison of hydrological models in forecast performance..... 163

4.5.4. Ensemble mean vs. adaptive weighted average..... 175

4.6. Conclusion 176

4.7. Acknowledgment..... 180

Appendix B: Performance evaluation metrics 181

B.1. Overall Forecast Skill and Accuracy..... 181

B.2. Model performance on peak flow magnitude and timing 182

B.3. Categorical(threshold-based) forecast verification 184

B.4. Economic Value of the model forecast 186

4.8. References..... 188

Chapter 5. Verification of Numerical Weather Predictions across Canada for Hydrologic Forecasting 197

5.1. Abstract..... 198

5.2. Introduction..... 200

5.3. Study Domains..... 205

5.4. Data..... 206

5.4.1. Forecast data 206

5.4.2. Verification data 207

5.5. Methodology..... 209

5.5.1. Traditional Grid-to-grid Verification..... 210

5.5.2. Object-based Verification (MODE)..... 211

5.5.3. Timing Error estimation..... 212

5.6. Results and Discussion 216

5.6.1. Low-Resolution Domain..... 216

5.6.2. High-Resolution Domain..... 227

5.7. Conclusion 235

5.8. References..... 239

Chapter 6. Conclusions and Recommendations..... 245

6.1. Conclusions..... 245

6.1.1. Calibration approaches for enhanced peak flow predictions..... 245

6.1.2. Hydrological model identification for large complex watersheds... 246

6.1.3. Combined hydrological model and input identification for semi-urban catchments 247

6.1.4. Identification of Numerical Weather Predictions (NWP) for enhanced flood forecasting 248

6.1.5. General conclusion and contributions..... 250

6.2. Future Work and Recommendations 251

6.3. References..... 254

List of Figures

Figure 2-1: Location of the study area in Humber River Watershed, Southern Ontario. ..34

Figure 2-2: Flowchart of proposed approach for selecting representative parameter set in event-based models.....40

Figure 2-3: Output of Regionalized Sensitivity Analysis. Figure displays the sensitivity index value of nine SWMM5 parameters for Nash-Sutcliffe Efficiency (NSE), Peak Flow Error (PE) and Volume Error (VE). Higher RSA index corresponds to higher sensitivity of parameters to the output performance. Description of parameter letter codes (x-axis) is presented in Table 2-1. 54

Figure 2-4: Cumulative Sum of the Normalized Reordered Output (CUSUNORO) used as first order sensitivity of SWMM5 parameters to three performance metrics (NSE, PE and VE). The deviation from the mean (CUSUNORO values or z(i) in Eqn. 5) is plotted against the empirical cumulative distribution of input parameters (x-axis). Higher deviation from the mean indicates higher sensitivity of parameters to corresponding performance metrics. Descriptions of parameter letter codes (for each colored lines of the plots) are presented in Table 2-1.....55

Figure 2-5: Box plots showing the spread of the lower, middle three and upper percentile values of most sensitive calibrated parameters (Imperviousness-Left and Drying Time-Right) to illustrate their variability in ten Model Sets (x-axis). Parameter values, collected from 714 sub-catchments, were ranked in ascending order. Model parameter sets represent different realization of the PCSWMM model in ten calibration events. 58

Figure 2-6: Figure showing Peak Flow variability of model parameter sets. 10 plots are constructed for 10 calibration events and each boxplot within a plot corresponds to different gauging stations. Individual boxplots are developed from 10 standardized peak flows, which are generated by ten different Model Parameter Sets in order to demonstrate the variability of different realizations of SWMM5 model. Standardized

peak flows are calculated by normalizing the deviation of the simulated peak flow from observed peak flow by the standard deviation of the simulated peak flow. Green horizontal line along the zero y-axis is computed based on observed peak flow.....59
Figure 2-7: Model Improvement (defined by Prediction Error Decrease in percentage (PED *100)) of Multi-site Simultaneous (MS-S) and Multi-site Average objective function (MS-A) calibration approaches when compared with Catchment Outlet (OU) approach at five gauging stations and ten calibration events. 63
Figure 2-8: Comparison of Taylor diagrams showing an event-by-event statistical evaluation of simulated flows from four calibration approaches (MS-S, MS-A, ME-MS, & OU) evaluated at six calibration events. The Taylor Diagrams summarized three statistical performances at five gauging stations for each event. Different colors denote respective calibration approaches while different shapes correspond to different stations (gauging sites). Perfect model sets would align themselves closer to the black arc as well as point 'OBS', which depict agreement with observations.....64
Figure 2-9: Performance ranking of 10 model parameter sets in ten calibration events. Normalized Nash-Sutcliffe Efficiency index (NSE) is used to score the performances at each gauging stations with sum over all sites and over all events displayed on the right side. Highest score corresponds to best performing model parameter set and vice versa. The heatmap shows the Normalized NSE values according to color palette displayed at the bottom side. 66
Figure 2-10: Model validation of top three model sets of MS-A approach in different events. The Taylor Skill Scores are evaluated at each of the five gauging sites for four different events. Most skillful models would have a score of 1 and the least ones have a score of 0. 67
Figure 3-1: Study area of the Upper Assiniboine River Basin 85
Figure 3-2: Illustration of the methodology adopted 89
Figure 3-3: The Bias-correction process using the Empirical Quantile Mapping method 94
Figure 3-4: Model update and forecast in the hindcast period.....95

Figure 3-5: Calibration and validation plots of SACSMA with SNOW17, MACHBV with SNOW17, and VIC with RVIC models. RMSE: Root-Mean-Square Error, and PBIAS: Percent Bias (Yapo et al., 1996) are displayed for each model to provide more information in addition to the visual inspection of the time series. 102
Figure 3-6: Mean CRPS of ensemble reservoir inflows generated with Raw (left) and Bias-corrected (right) ensemble GEFSv2 precipitation forecasts. 105
Figure 3-7: Comparison of reservoir inflow ensembles between four hydrological models using CRPS skill score (CRPSS). 105
Figure 3-8: CRPS decomposition components. The Left plot shows the reliability component, and the right plot shows the Potential CRPS component. Comparison of these attributes was made between four hydrological models..... 107
Figure 3-9: Reliability diagrams of ensemble reservoir inflows at three selected forecast lead times. Different colors show different model types. The 90% confidence intervals were shown in the reliability lines for each model. The inset histograms show the frequency of occurrence in each forecast bin. 109
Figure 3-10: Relative Operating Characteristic (ROC) curves drawn for probability thresholds exceeding 75, 80, 85, 90, and 95 percentile reservoir inflows for three days ahead forecast. The four plots are for four different hydrological models. 111
Figure 3-11: The ROC Score measured by the Area Under the ROC Curves of four hydrological models..... 112
Figure 4-1: Study Area of the Humber River Watershed 145
Figure 4-2: Methodology adopted for evaluating and selecting hydrological models and high-temporal weather forecast inputs..... 149
Figure 4-3: Screening weather forecast inputs by comparing the resulting hydrological forecast quality. Four weather forecast inputs (different colors) were fed into the five hydrological models (five boxes above) and the MAE at different lead times is estimated. 162

Figure 4-4: Taylor Diagram and Taylor Skill Score showing the statistical comparison of hydrological models at different forecast lead times using HRDPS input. The three statistical performances metrics that are displayed in Taylor Diagram are summarized by one single score in the Taylor Skill Score as shown in the right most plot (A score of one corresponds to the most skillful models). 165

Figure 4-5: Comparison of forecast performance of hydrological models using different metrics that measure peak flow magnitude, timing, overall quality, and accuracy. The Models were fed by two competing weather forecast inputs (HRDPS: Left and HRRR: Right). The normalized difference between the two inputs estimated at each forecast lead time is shown in the rightmost horizontal bar graph. 167

Figure 4-6: Flood threshold approximation: POD (Probability of Detection) versus different flow percentiles for different hydrological models. Different colors indicate outputs at different forecast leadtimes (1hr to 18hr). 169

Figure 4-7: Same as Figure 6 but with categorical forecast verification metrics or threshold-based scores. A 90 percentile of the observed streamflow is set as a flood threshold for this research. 171

Figure 4-8: Comparison of hydrological models by their forecast Economic Value (V). V is estimated using the deterministic HRDPS weather forecast input and is drawn as a function of cost-lost ratio to account for various users and forecast systems. 174

Figure 4-9: Taylor Diagram and Taylor Skill Score to show the comparison between Ensemble Mean and Adaptive weighted averaging methods. For reference, one model's (MACHBV) output is presented. 176

Figure 5-1: Location of study domains 206

Figure 5-2: Algorithm to estimate Timing Error, tei,i,t (equation 5-3). Pf and Po are forecasted and observed precipitation, respectively. 215

Figure 5-3: Sample precipitation objects identified on extreme event of 2018-06-01, forecasted 1 day earlier (LDT1) on Low resolution domain: MODE parameters: - Conv Thresh>=10mm, Conv Radi=5 grid units 218

Figure 5-4: Verification metrics for evaluating different accumulated precipitation forecasts of five NWPs in Low-Resolution Domain 219
Figure 5-5: Verification metrics for evaluating different precipitation intensity forecasts of five NWPs in Low-Resolution Domain.....220
Figure 5-6: 3-days ahead forecast skill in terms of precipitation objects features at different thresholds and grid smoothing resolution (Convolution Radii (R) (5, 10, 15 grid units) and Convolution Thresholds (T) (>=5,>=10,>=15 mm)).....223
Figure 5-7: Same as Figure 5-8 but for a week ahead forecast.....223
Figure 5-9: Timing Error analysis outputs for five NWPs located in Low-resolution domain. 225
Figure 5-10: Forecast Bias (left) and Precipitation objects count (right) of NWPs located in High-resolution domain for a precipitation threshold of 1mm accumulated in 1, 3 and 6 hours 229
Figure 5-11: Gilbert Skill (GSS) of NWPs located in High-resolution domain for a different precipitation threshold volume accumulated over 6-hours.....229
Figure 5-12: Verification metrics for evaluating different precipitation intensity forecasts of four NWPs in High-resolution Domain.....231
Figure 5-13: Timing Error analysis output for five NWPs located in High-resolution Domain 233

Figure A- 1: Soil and Land cover tiles discretization for VIC model.....126

Figure B- 1: Illustration of Series Distance (SD) method for calculating magnitude (Q) and timing error (T) for a sample event selected within a timeseries of this study.....183

List of Tables

Table 2-1: Description of SMWM5 model parameters 35
Table 2-2: Events selected for calibration and model testing 39
Table 3-1: Performance statistics of the three hydrological models from calibration and validation. The definition of the abbreviations is presented Section 3.4.2. 101
Table 4-1: Weather forecast products and details (Note: Forecasts of HRRR and RAP are available at any hour of a day. For this study four valid times are selected to compare them to the other two products) 147
Table 4-2: Metrics used for comparing the forecast performance, skill and quality of hydrological models and inputs 156
Table 4-3: List of models used and their calibration and validation result (Using Observed hourly Precipitation & Temperature data). Bold font indicates models selected for the forecast verification step. 159
Table 5-1: Numerical Weather Predictions used in the Low-resolution domain 208
Table 5-2: Numerical Weather Predictions used in the High-resolution domain 208
Table 5-3: Contingency table and associated parameters for calculating traditional verification metrics 210
Table A- 1 : SNOW17 model parameters 121
Table A- 2: SACSMA model parameters 122
Table A- 3: MACHBV model parameters 123
Table A- 4: VIC/RVIC model parameters 127
Table B- 1: Contingency table and associated parameters for calculating categorical verification metrics and economic value. Detailed definitions of symbols are presented in Section B.3 184

List of Abbreviations

AMALGAM	A Multi-Algorithm, Genetically Adaptive Multi-Objective Method
ANUSPLIN	Australian National University spline smoothing algorithm
AUC	Area Under ROC Curve
BC	Bias-Correction
BEUP	Bayesian ensemble uncertainty processer
BiasFreq	Bias Frequency
C	Cost
C/L (C/La)	Cost-loss ratio
CaPA	Canadian Precipitation Analysis
CAT	Catchment hydrological cycle Assessment Tool
CDF	Cumulative Distribution Function
CI	Confidence Intervals
CLUE-E	Conversion of Land Use and its Effect
CMC	Meteorological Service of Canada
CN	Curve Number
CONUS	Continental United States
CRHM-AHM	Cold Regions Hydrological Model - Arctic Hydrology Model
CRPS	Continuous Rank Probability Score
CRPSS	Continuous Rank Probability Skill Score
CSI	Critical Success Index
CSM	Contribution to the Sample Mean
CUSUNORO	Cumulative Sum of the Normalized Reordered Output
DDS	Dynamically Dimensioned Search
DMIP	Distributed Model Inter-comparison Project

DT	Drying Time
EC	Environment Canada
ECCC	Environment and Climate Change Canada
ECMWF	European Centre for Medium-Range Weather Forecasts
EPA	Environmental Protection Agency
EPS	ensemble prediction systems
FAO	Food and Agriculture Organization
FAR	False Alarm Rate
FBIAS	Frequency Bias
FEWS	Flood and Early Warning System
GARDENIA	Global À Réservoirs pour la simulation de DÉbits et de NIveaux Aquifères Model
GDPS	Global Deterministic Prediction System
GEFSv2	Second-generation Global Ensemble Forecast System
GEM	Global Environmental Multiscale model
GFS	Global Forecast System
GIS	Geographic Information System
GR4H	Génie Rural à 4 paramètres Heure Model
GR4J	Génie Rural à 4 paramètres Journalier Model
GRU	Group Response Unit
GSS	Gilbert Skill Score
GSSHA	Gridded Surface Subsurface Hydrologic Analysis
HBV	Hydrologiska Byråns Vattenbalansavdelning model
HEC-HMS	Hydrologic Engineering Center - Hydrologic Modeling System
HEPS	Hydrological Ensemble Prediction System
HOOPLA	HydrOlOgical Prediction Laboratory
HPC	High-Performance Computing systems

HRDPS	High-Resolution Deterministic Precipitation System
HRRR	High-Resolution Rapid Refresh
HRU	Hydrologic Response Units
HUP	Hydrologic Uncertainty Processor
HVC	Hypervolume Contribution
HYMOD	HYdrological MODel
IHACRES	Identification of unit Hydrographs And Component flows from Rainfall, Evapotranspiration, and Streamflow
IM	Imperviousness
KGE	Kling–Gupta efficiency
KINEROS2	Kinematic Runoff and Erosion Model
LDT	Lead Time
LID	Low Impact Development
MACHBV	McMaster University-Hydrologiska Byråns Vattenbalansavdelning model
MAE	Mean Absolute Error
ME-MS	Multi-Event Multi-Site
MEPS	Meteorological Ensemble Prediction System
MESH	Modelisation Environmentale Communautaire-MEC Surface and Hydrology
MET	Model Evaluation Tool
MI-HPC	Manitoba Infrastructure Hydrologic Forecasting Center
MODE	Method for Object-Based Diagnostic Evaluation
MODIS	Moderate Resolution Imaging Spectroradiometer
MOPEX	Model Parameter Estimation Experiment
MS-A	Multi-Site Average Objective Function
MS-S	Multi-Site Simultaneous Objective Function

NAM	North American Mesoscale Forecast System
NCEP	National Centers for Environmental Prediction
NOAA	National Oceanic and Atmospheric Administration
NSE	Nash-Sutcliffe Efficiency
NSGA	Non-dominated Sorted Genetic Algorithm
NVCA	Nottawasaga Valley Conservation Authority
NWP	Numerical Weather Prediction
NWS	National Weather Service
OSTRICH	Optimization Software Tool
OU	Catchment Outlet
PA-DDS	Pareto Archived Dynamically Dimensioned Search
PBIAS	Percent Bias
PCSWMM	Personal Computer Stormwater Management Model
PDM	Probability Distributed Model
PE/PFE	Peak-Flow Error
PED	Prediction Error Decrease
PFC	Peak Flow Criteria
POD	Probability of Detection
QPF	Quantitative Precipitation Forecast
RAP	Rapid Refresh
RDPS	Regional Deterministic Precipitation System
Reli	Reliability
RGA	Regionalized Sensitivity Analysis
RMSE	Root-Mean-Square-Error
ROC	Relative operating characteristics
RVIC	Routing Variable Infiltration Capacity
SACSMA	Sacramento Soil Moisture Accounting model

SD_Q	Series Distance - Magnitude Error
SD_T	Series Distance - Timing Error
SNOW17	Snow accumulation and ablation model
SPEA	Strength Pareto Evolutionary Algorithm
SWAP	Soil, Water, Atmosphere and Plant model
SWAT	Soil and Water Assessment Tool
SWMM	Storm Water Management Model
TANK	Hydrological Tank Model
THORPEX	The Observing System Research and Predictability Experiment
TIGGE	THORPEX Interactive Grand Global Ensemble
TRCA	Toronto and Region Conservation Authority
TSS	Taylor Skill Score
UBCWM	University of British Columbia Watershed Model
UTC	Coordinated Universal Time
V	Economic Value
VE	Volume Error
VIC	Variable Infiltration Capacity
WAGENINGEN	Hydrological Model type
WATFLOOD	Watershed Simulation and Flood Forecasting Model
WMO	World Meteorological Organization
WRF	Weather Research and Forecasting
WRF-ARW	Advanced Research-Weather Research and Forecasting
WRF-Hydro	Weather Research and Forecasting Hydrological and Hydraulic model

List of Symbols

a	The ratio of the standard deviations of simulated flow to observed flow
b	The ratio of the means of simulated flow to observed flow
c	Series distance connector between observed and modeled points
C	Cost of flood forecasting system/decision-maker
$CRPS_{pot}$	Potential CRPS component of the Mean Continuous Rank Probability Score (CRPS)
e^2	Mean Square Forecast Error
E_C, E_F, E_P	Expenses of the decision-maker/flood forecasting system
F_i	Forecasted river flow discharge
h	Forecast horizon (length)
L_a	Avoidable loss of flood damage
L_u	Unprotectable/unavoidable loss of flood damage
m	Number of objective functions
n_p	Number of peak flows greater than one-third of the mean peak flow observed
O_i	Observed river flow discharge
$O(p)$	A vector of multi-objective function
$o_m(p)$	The mth objective function/performance metric
$\bar{o}$	Relative frequency of occurrences
p	Optimum parameter solutions
$P(y)$	The cumulative distribution function of the ensemble river flow forecast
q_o	Peak observed river discharge/reservoir inflow time series
q_s	Peak simulated river discharge/reservoir inflow time series

Q_o	Observed river flow discharge
Q_s	Simulated river flow discharge
$Q_{p,o}$	Observed peak flow in a flow hydrograph
$Q_{p,s}$	Simulated peak flow in a flow hydrograph
r	The correlation coefficient between observed and simulated reservoir inflows
R	The correlation coefficient between observed and simulated river discharges
R_o	Maximum correlation attainable
$\overline{Reli}$	The reliability of the ensemble forecast; the mean reliability component of CRPS
S	Taylor Skill Score
SD_T	Error in peak flow timing; Series Distance – Timing error
SD_Q	Error in peak flow magnitude; Series Distance – Magnitude error
te	Timing error along the forecast horizon
t_o	Zero forecast time; current forecast time
V	The economic value of the model forecast
V_o	The volume of water under observed flow hydrograph
V_s	The volume of water under simulated flow hydrograph
w_i	Adaptive weight applied to the individual model forecast
σ_o	The variance of simulated river discharge
σ_s	The variance of observed river discharge
$1\{.\}$	A step function representing 1 for ensemble forecasts greater than the observation and 0 otherwise

Declaration of Academic Achievement

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

Chapter 2: Event-based model calibration approaches for selecting representative distributed parameters in semi-urban watersheds, by F.S. Awol, P. Coulibaly, and B.A. Tolson, Advances in Water Resources, 118, 12-27, doi: 10.1016/j.advwatres.2018.05.013, 2018. (With permission from the publisher)

Chapter 3. Identification of hydrological models for enhanced ensemble reservoir inflow forecasting in a large complex Prairie watershed, by F.S. Awol, P. Coulibaly, I. Tsanis, F. Unduche, Water, 11(11), 2201, doi: 10.3390/w11112201, 2019. (With permission from the publisher)

Chapter 4. Identification of combined hydrological models and numerical weather predictions for enhanced flood forecasting in a semi-urban watershed, by F.S. Awol, P. Coulibaly, I. Tsanis, Journal of Hydrometeorology, under review, manuscript number JHM-D-19-0174.

Chapter 5. Verification of Numerical Weather Predictions across Canada for Hydrologic Forecasting, by F.S. Awol, P. Coulibaly, I. Tsanis, Weather and Forecasting, under review, manuscript number WAF-D-19-0202.

For Chapter 2, F. S. Awol conducted the modeling and computational work with the guidance and supervision of Dr. P. Coulibaly and co-supervision of Dr. I. Tsanis. Dr. B. Tolson provided guidance on DDS and PADDS optimization algorithms. F. S. Awol wrote the manuscript, and Dr. P. Coulibaly and Dr. B. A. Tolson reviewed and edited it, the paper was published in Advances in Water Resources in 2018. For Chapter 3, F. S. Awol conducted the modeling and computational work with the guidance and supervision of Dr. P. Coulibaly and co-supervision of Dr. I. Tsanis. Dr. F. Unduche provided the operational WATFLOOD model for use as a benchmark. F. S. Awol wrote the manuscript, and Dr. P. Coulibaly, Dr. F. Unduche, and Dr. I. Tsanis reviewed and edited it, the paper was published in Water Journal in 2019. For Chapter 4, F. S. Awol conducted the modeling and computational work with the guidance and supervision of Dr. P. Coulibaly and co-supervision of Dr. I. Tsanis. F. S. Awol wrote the manuscript, and Dr. P. Coulibaly reviewed and edited it, and the paper was submitted to the Journal of Hydrometeorology in 2019. For Chapter 5, F. S. Awol conducted the verification and computational work with the guidance and supervision of Dr. P. Coulibaly and co-supervision of Dr. I. Tsanis. F. S. Awol wrote the manuscript, and Dr. P. Coulibaly reviewed and edited it, the paper was submitted to Weather and Forecasting in 2019. The work reported here was undertaken from September 2015 to October 2019.

Chapter 1. Introduction

1.1. Deterministic and ensemble flood forecasting

Recent studies indicate that the intensity of summertime convective storms, frequency of maximum hourly precipitation, and spatial expansion of heavy precipitation will increase in North America (Prein et al., 2017). The combined effect of intense and frequent rainfall events, ice jams and snow melts, anthropogenic influences, and landscape changes have contributed to locally induced floods in Canada in the past (Bonsal et al., 2019; Khandekar, 2002; Zhang et al., 2019). Whether due to natural or anthropogenic factors, flooding is becoming a chronic natural hazard, and the World Meteorological Organization (WMO) encourages a shift from the traditional structural intervention and localized approach to basin-wide integrated flood forecasting and early warning systems (FEWS) to minimize flood impact (WMO, 2011).

Operational flood forecasting systems in regional, national, continental and global scales have various capabilities depending on the domain and climatic region, hydrologic and hydraulic models, rainfall observation and forecasts, Numerical Weather Predictions (NWP), verification systems, forecast lead times and frequency, forecast style, etc. (Achleitner et al., 2012; Adams and Pagano, 2016; De Roo et al., 2003; Demargne et al., 2014; Jasper et al., 2002; Maxey et al., 2012; Pappenberger et al., 2008; Unduche et al., 2018; Zahmatkesh et al., 2019). The forecast style of flooding conveyed to the public, stakeholders or internally could be categorical (e.g., minor, moderate or major), deterministic (e.g., 100 m^3/s of river discharge), or ensemble probabilistic (e.g., 80%

chance of flooding) (Adams and Pagano, 2016). Traditional flood forecasting centers usually update a certain rainfall-runoff conversion method with experts' interpretation of meteorological outputs to issue categoric flood outlooks. In most hydrological forecasting centers, deterministic type floods are issued by forcing a single hydrological model with Quantitative Precipitation Forecasts (QPF) obtained from deterministic NWP systems. However, many are shifting from deterministic to ensemble-based probabilistic flood forecasting due to its advantages of showing the total uncertainties associated with weather forecast inputs, hydrological model parameters, and structural complexity (Cloke and Pappenberger, 2009).

Hydrological Ensemble Prediction Systems (HEPS) have improved flood risk management by offering longer forecast lead times, making advances in hazard mitigation and decision-making processes, and networking researchers with managers (Michaels, 2015). There are several ways to generate ensembles of flood forecasts. The conventional method is by forcing hydrological model(s) with Meteorological ensemble prediction systems (MEPS) informed by NWPs to conceptualize the input uncertainties (Abaza et al., 2013; Alfieri et al., 2014; Buizza et al., 2005; Calvetti and Pereira Filho, 2014; Fan et al., 2014a; Horat et al., 2018; Pietroniro et al., 2007; Thiemig et al., 2015; Zapata and Alberto, 2010; Zsótér et al., 2016). Ensembles can also be generated by several realizations of hydrological model parameter sets (e.g., using Monte Carlo simulation) in order to capture uncertainties related to hydrological model processes such as the initial state of the models and the parameterizations (Beven and Freer, 2001; Carpenter and Georgakakos, 2004; Georgakakos et al., 2004; Pappenberger et al., 2005). The third and evolving method to

generate ensembles is by using multiple hydrological models to realize the uncertainties inherited in the model structures that attempt to represent the physical world (Ajami et al., 2006; Antonetti et al., 2018; Brochero et al., 2011a, 2011b; Seiller et al., 2012, 2017; Thiboult et al., 2016, 2017; Velázquez et al., 2011). Even though improvements were made in ensemble generation approaches, some challenges exist. The under-dispersivity and bias of ensemble NWPs due to their coarse spatial resolutions have been addressed by downscaling (Gaborit et al., 2013; Renner et al., 2009) and bias-correction or post-processing methods (Bourdin et al., 2014; Crochemore et al., 2016; Cui et al., 2012; Fan and van den Dool, 2011; Jha et al., 2018). On the other hand, several post-processing methods were proposed on ensemble hydrological forecasts instead of ensemble NWPs to remove biases and uncertainties (Ajami et al., 2006; Han and Coulibaly, 2019; Hashino et al., 2007; Madadgar et al., 2014; Wood and Schaake, 2008). Concurrent with the above advancements, the effectiveness of any flood forecasting system depends on the quality, accuracy, reliability, and skill of hydrological models and weather forecast inputs.

1.2. Hydrological models and weather forecast inputs

Hydrological models or rainfall-runoff models are used to represent the natural system. Physically-based hydrological models are generally formulated using some physically measurable parameters and describe multiple components of the basin hydrologic processes with conservation of mass, momentum, and energy equations (Beven, 1993; Beven and Kirkby, 1979). On the other hand, conceptual hydrological models are designed to approximately represent the watershed system by optimizable parameters, state variables, and simplified analytical solutions to the governing equations (Nash and Sutcliffe, 1970).

In most distinctive terms, hydrological models can be classified into lumped, semi-distributed, and fully distributed models based on the spatial representation of the watershed (Corral et al., 2000; Moradkhani and Sorooshian, 2008; Sitterson et al., 2017). Lumped hydrological models have one spatially enclosed catchment upstream of the outlet gauging station. In semi-distributed models, the basin is usually discretized into several sub-catchments, grids, Hydrological Response Units (HRU) (Kalcic et al., 2015; Sanzana et al., 2013) or Group Response Units (GRU) (Kouwen et al., 1993) depending on the type of the model structure. Semi-distributed models require an embedded or external hydraulic routing component and river networks to route runoff from each sub-basin or grid cell to river nodes and the downstream catchment outlet. Fully distributed models are similar to grid-based semi-distributed models but allow for lateral transfer of water between each grid cell (Haghnegahdar et al., 2014), such as the WRF-Hydro model (Gochis et al., 2018). Hydrological models can also be divided into event-based and continuous models based on the time frame of the input and output data. Event-based hydrological models are calibrated using one or more selected events (e.g., extreme weather periods), in which each event usually spans for a few days or weeks, depending on the storm periods and catchment's response times. Event-based hydrological models are mainly used for flood forecasting purposes in urban and semi-urban areas. On the contrary, continuous hydrological models are set-up based on multiple years of historical meteorological and hydrological time series data and are usually applied for hydrological forecasting, water management, and climate change impact studies. This model accounts for continuous water and soil-moisture contents in surface and sub-surface storage zones.

The complex physical processes of some landscapes cannot be easily represented by standard hydrological models. Hydrological modeling in Canada has been a challenging topic due to the geographical and climatological variations across the large country, and the complexity of the watersheds, which include Prairies, wetlands, glaciers, permafrost, the Boreal Shield and tile drain. Some unique processes such as rain-snow partitioning, frozen ground, the interaction between potholes (wetlands), contributing and non-contributing drainage networks, orographic corrections, sublimation, and permafrost, require proper treatment in the hydrological model structures and parameters. To sufficiently account for these processes, some models were designed for specific purposes (e.g., UBCWM for mountain hydrology (Fotakis, et al. 2014), CRHM for cold regions (Pomeroy et al., 2007)) or to be adaptive in diverse landscapes (e.g., Raven (Craig et al., 2018), WATFLOOD (Kouwen 1988), or for national scale hydrological modeling (e.g., MESH (Haghnegahdar et al., 2014)).

Hrachowitz & Clark, (2017) discussed complementing modeling philosophies in hydrology between distributed and lumped models. Based on their opinion, what is significant in hydrological modeling is the way models are implemented because all models can be applied at a desired degree of detail, although models would remain, to some extent, conceptual. Most importantly, they recommended the use of diverse modeling strategies by exploiting available macroscale data with multi-scale model development. Therefore, the choice of the models to be implemented for flood forecasting, for example, shall depend on the intended purpose, the complexity and scale of the basin, and the type of weather forecast inputs available.

Once hydrological model(s) are identified for the study area, weather forecast inputs obtained from radar nowcasts, Numerical Weather Predictions (NWP), or climate change scenario predictions can be forced to the model(s) to produce hydrological predictions with varying forecast lengths. NWPs are particularly vital for cascading the inherited uncertainty from initial atmospheric conditions and providing short- to long-range flood forecasts (Pappenberger et al., 2005). Advances in hydrometeorological research have led to the development of various NWP products across the Globe. The types of NWPs depend on the scale (Global, Continental, and Regional), forecast length (Long-, Medium-, Short- and Very Short-ranges), spatial resolution (Low- and High-resolutions), and characteristics (Deterministic and Ensemble). The skill and quality of meteorological variables are influenced by the variability of NWP types, which also affect the skill and reliability of hydrological forecasts.

The chaotic nature of the atmospheric system (Lorenz, 1969) and its approximate representation by NWP systems created uncertainties in deterministic forecasts (Cuo et al., 2011). This phenomenon has led to the development of ensemble forecasts. Different ensemble NWPs vary by the methods they used to perturb initial conditions (Buizza et al., 2005).

Among the variables generated by NWPs, precipitation forecasts have been the main challenge in achieving the correct intensity, location, and timing of storms (Cuo et al., 2011). Mainly, summer precipitation forecasts by mesoscale NWP systems were considered to be difficult due to the nature of localized convective thunderstorms (Kaufmann et al., 2003). Golding, (2000), highlighted that the quality of precipitation

forecasts coupled with the catchment size and response time should be a primary requirement for flood prediction systems.

Real-time forecast data can usually be obtained directly from the providing organizations (e.g., ECCC, ECWMF, NOAA) that sometimes archive past forecasts (Bougeault et al., 2010). Verification of NWPs is typically performed using archived forecast data to examine their quality so that skillful NWPs can be identified and recommended for operational flood forecasting. Similarly, verification of hydrological forecasts generated by forcing calibrated hydrological models with archived NWP inputs has been practiced for evaluating the accuracy, reliability, and overall forecast skill using observed discharge data (Alfieri et al., 2014, Fan et al., 2014; Zsótér et al., 2016).

1.3. Challenges in hydrological model calibration

Hydrological models attempt to represent the physical processes of the natural system through non-linear mathematical formulations containing variables and model parameters. Model parameters can be physically measured or estimated through calibration. Hydrological model calibration, in simple terms, is a process of adjusting parameters to make the output variables (e.g., river discharge, reservoir inflow) as accurate as possible. The calibration process or parameter estimation by itself is an Inverse Problem, meaning that it is a process of finding causes (parameters) from a known effect (observed discharge) (Moradkhani and Sorooshian, 2008). An inverse problem or a model calibration is *ill-posed* because solutions (parameters) are non-unique and non-identifiable (Moradkhani and Sorooshian, 2008; Renard et al., 2010; Sun and Sun, 2015) which often lead to uncertainty, equifinality (Beven, 1993), and objective function surfaces with multiple local optima

(Duan and Gupta, 1992). Rigorous parameterization is a proven method to reduce the burden of calibration and validation in hydrological models by limiting the number of free parameters as few as possible (Refsgaard, 1997), which indirectly minimizes the ill-posedness problem. Model inference with Monte Carlo simulation (Moradkhani and Sorooshian, 2008) and with prior information (Renard et al., 2010) are some of the solutions recommended for the equifinality and uncertainty problems. For enhancing parameter identifiability, a sensitivity analysis is an essential step to diagnose and identify the most sensitive parameters before calibration (Shin et al., 2013). Identifiability can also be improved by an approach proposed by Shafii et al., (2017), by using flow-partitioning-based criteria as part of multi-objective optimization. Multi-objective optimization in model calibration has been advanced over the last decades. Its advantages in finding appropriate trade-offs, handling non-uniqueness, non-identifiability, and uncertainty problems, and exploring measurement-based (hard data) and information-based (soft data) criterion have been well recognized (Duan and Gupta, 1992; Efstratiadis and Koutsoyiannis, 2010; Gupta et al., 1998; Seibert and McDonnell, 2002; Tang et al., 2005).

The above literature discusses problems related to the nature of calibration by itself. The challenges in the hydrological model calibration process depend on several other factors such as the model type (lumped or distributed; event-based or continuous) and basin type (gauged or ungauged; small urban-based or large complex). In distributed hydrological models, the discretization of watersheds is critical and affects the calibration process, the computational cost, and the quality of outputs (Haghnegahdar et al., 2015). An approach proposed by Liu et al., (2016), for example, aims to assist the modeler by providing *a priori*

error metric that quantifies information losses related to routing and changes in land cover and soil type during discretization processes. To resolve calibration challenges in ungauged basins and improve river flow predictions; Bárdossy, (2007), discussed transferability of a lumped hydrological model parameters sets from donor catchments; Pokhrel and Gupta, (2010), proposed spatial regularization methods to improve performances in distributed hydrological models; and Razavi and Coulibaly, (2016), proposed a multi-model regionalization approach involving lumped models, neural network and inverse distance methods.

Some hydrological models can be used as either a lumped or semi-distributed model based on the modeler's decision and the intended purpose. The SACSMA is one of the models that has been applied as a both lumped and semi-distributed model for various forecasting applications. Kitanidis and Bras, (1980), applied SACSMA as a lumped model for analyzing uncertainties in real-time hydrological forecasting. Recently, Leach et al., 2018, used it as a lumped model for assessing the effect of near-real-time data assimilation to improve hydrological forecasting in urban basins. SACSMA has also been used as a semi-distributed model for operational hydrological forecasting purposes at the NWS River Forecasting Centers (Shamir et al., 2006). It was also adopted for exploring different calibration scenarios, and streamflow simulation approach as semi-distributed model (Ajami et al., 2004).

The conventional calibration approach, either in lumped or distributed hydrological models, is to calibrate the entire catchment parameters at the basin outlet by minimizing or maximizing single or multiple objective functions. An alternative approach to account for

interior gauging sites is a multi-site calibration method. Many forms of multi-site calibration approaches have been proposed, for example, sequential/hierarchical (Hay et al., 2006; Ozdemir et al., 2017; Singh & Bárdossy, 2015), weighted average (Asadzadeh et al., 2014; Engeland et al., 2006; Khu et al., 2008; Madsen et al., 2002), and simultaneous (Leta et al., 2017; Zhang, et al., 2010).

The aim of calibrating a continuous hydrological model is to find one set of representative model parameters over the entire calibration period by using a single-objective or multi-objective optimization algorithm. For event-based hydrological models, the calibration is generally performed in multiple independent event periods and hence results in multiple candidate parameters sets one for each event. In the application of event-based hydrological models for peak flow prediction, the question of which calibrated model parameter sets should be used can create a practical dilemma. Therefore, novel methods are needed to address the uncertainties associated with model parameterization and temporal variations of input storm events.

In general, robust calibration and validation approaches are required to identify optimum model parameters and improve hydrological and flood predictions (Kraußе et al., 2012).

1.4. Scope of the research

As indicated in earlier sections, successful flood forecasting and early warning systems depend on the quality of hydrological models and weather forecast inputs, and the characteristics of the watersheds. Without a clear methodology to identify proper parameter estimation approaches, skillful hydrological models and forecast inputs, flood forecasters

and hydrologists at large, face challenges in obtaining reliable and accurate short- and medium-range river flow forecasts in various watershed types. One of the main gaps in previous methods in the literature as well as in practical application is that hydrological model selection and development focused on calibration and validation performances based only on historical observation datasets. The forecast performances of the models should be tested with different weather forecast inputs at various forecast lead times. In addition, the forecast skill of the hydrological models should be evaluated in various types of watersheds with the appropriate weather forecast inputs. Different Numerical Weather Predictions (NWPs) have multiple ranges of spatial and temporal resolution, and the hydrological forecast quality obtained from these inputs depends on the scale and the type of the watershed. Some weather forecasts work well in large basins, and some are useful in smaller and urban catchments. The scope of this research focuses on addressing the above challenges to benefit operational flood forecasting communauty, future applications in flood and early warning systems, and research aiming at improving model development, forecast inputs and calibration methods. In this thesis,

- the necessary evaluation and verification of different calibration approaches, multiple hydrological models with diverse model structures, and various high- and low-resolution NWPs will be presented; and
- the candidate hydrological models, model parameter estimation approaches, and forecast inputs will be identified and discussed in two different watershed types.

1.5. Research objectives and Thesis outline

In order to achieve the general goal presented in Section 1.4, four independent studies were conducted. The specific objectives of each research are outlined below:

- Research objective 1 focuses on formulating and testing an appropriate calibration approach for enhancing peak flow prediction at multiple sites in semi-urban catchments using event-based hydrological models. The study tries to answer the question "in the application of event-based hydrological models, which of the calibrated model parameter sets should be used for peak flow prediction?".
- Research objective 2 aims at addressing the challenges faced by hydrologists working on flood forecasting in large and complex watersheds. Outflows from such basins often feed reservoirs. The study applies various evaluation and verification techniques to compare structurally varied hydrological models and identify the skillful and reliable ones for an improved medium-range ensemble reservoir inflow forecasting in large and complex watersheds. In general, it tries to answer the question "can medium-range reservoir inflow forecasting be accurately achieved by simple, medium level or advanced hydrological models?".
- Research objective 3 addresses the challenges in urban and semi-urban catchments because flood forecasting is often influenced by the capability of the combined hydrological models and weather forecasts to accurately predict floods. The goal is to identify a proper combination of skillful hydrological models and weather forecast inputs for an improved short-range flood forecasting in semi-urban watersheds. The research tries to answer the question "which model-input combination could be

identified for enhanced short-range flood forecasting in urban/semi-urban catchments?".

- Research objective 4 focuses on identifying candidate Numerical Weather Predictions for short- and medium-range flood forecasting in Canada's varied geographical landscape. It tries to answer the question, "which weather forecast products provide accurate forecast inputs for enhanced flood forecasting? And where?".

The thesis is organized into six chapters. This first chapter provided an introduction discussing lessons learned from past studies, the challenges, and the general background of the research. The second chapter presents an approach formulated for calibrating and validating an event-based hydrological model in a semi-urban watershed aiming at improving flood forecasting at interior and outlet gauging stations. The third chapter evaluates lumped, semi-distributed, and land-surface based hydrological models with ensemble weather forecast inputs to enhance reservoir inflow forecasting in a complex Prairie watershed. The fourth chapter investigates multiple hydrological models and several weather forecast inputs to find an appropriate combination of model and inputs for an improved short-range flood forecasting in a semi-urban catchment. The fifth chapter compares and verifies several low- and high-resolution NWPs across Canada and identifies the best candidates that could improve the prediction of the timing, intensity, and volume of floods in large and small watersheds. The sixth and final chapter summarizes the main conclusions and contributions of the thesis and provides follow up recommendations for future research.

1.6. 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. Mon. Weather Rev. 141, 3462–3476. https://doi.org/10.1175/MWR-D-12-00206.1

Achleitner, S., Schöber, J., Rinderer, M., Leonhardt, G., Schöberl, F., Kirnbauer, R., Schönlaub, H., 2012. Analyzing the operational performance of the hydrological models in an alpine flood forecasting system. J. Hydrol. 412–413, 90–100. https://doi.org/10.1016/j.jhydrol.2011.07.047

Adams, T.E., Pagano, T.C., 2016. Flood Forecasting: A global perspective. Elsevier Inc. https://doi.org/10.1016/C2014-0-01361-5

Ajami, N.K., Duan, Q., Gao, X., Sorooshian, S., 2006. Multimodel combination techniques for analysis of hydrological simulations: application to distributed model intercomparison project results. J. Hydrometeorol. 7, 755–768. https://doi.org/10.1175/JHM519.1

Ajami, N.K., Gupta, H., Wagener, T., Sorooshian, S., 2004. Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. J. Hydrol. 298, 112–135. https://doi.org/10.1016/j.jhydrol.2004.03.033

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., Salamon, P., 2014. Evaluation of ensemble streamflow predictions in Europe. J. Hydrol. 517, 913–922. https://doi.org/10.1016/j.jhydrol.2014.06.035

Antonetti, M., Horat, C., Sideris, I. V, Zappa, M., 2018. Ensemble flood forecasting considering dominant runoff processes: I. Setup and application to nested basins (Emme, Switzerland). Nat. Hazards Earth Syst. Sci. Discuss. 5194, 1–29. https://doi.org/10.5194/nhess-2018-118

Asadzadeh, M., Razavi, S., Tolson, B.A., Fay, D., 2014. Pre-emption strategies for efficient multi-objective optimization: Application to the development of Lake Superior regulation plan. Environ. Model. Softw. 54, 128–141. https://doi.org/10.1016/j.envsoft.2014.01.005

Bárdossy, A., 2007. Calibration of hydrological model parameters for ungauged catchments. Hydrol. Earth Syst. Sci. 11, 703–710. https://doi.org/10.5194/hess-11-703-2007

Beven, K., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Adv. Water Resour. 16, 41–51. https://doi.org/10.1016/0309-1708(93)90028-E

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249, 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24, 43–69. https://doi.org/10.1080/02626667909491834

Bonsal, B.R., Peters, D.L., Seglenieks, F., Rivera, A., Berg, A., 2019. Changes in freshwater availability across Canada. Canada's Chang. Clim. Rep. 261–342.

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., De Chen, H., Ebert, B., Fuentes, M., Hamill, T.M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.Y., Parsons, D., Raoult, B., Schuster, D., Dias, P.S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., Worley, S., 2010. The THORPEX interactive grand global ensemble. Bull. Am. Meteorol. Soc. 91, 1059–1072. https://doi.org/10.1175/2010BAMS2853.1

Bourdin, D.R., Nipen, T.N., Stull, R.B., 2014. Reliable probabilistic forecasts from an ensemble reservoir inflow forecasting system. Water Resour. Res. 50, 3108–3130. https://doi.org/10.1002/2014WR015462

Brochero, D., Anctil, F., Gagné, C., 2011a. Simplifying a hydrological ensemble prediction system with a backward greedy selection of members -Part 1: Optimization criteria. Hydrol. Earth Syst. Sci. 15, 3327–3341. https://doi.org/10.5194/hess-15-3327-2011

Brochero, D., Anctil, F., Gagné, C., 2011b. Simplifying a hydrological ensemble prediction system with a backward greedy selection of members - Part 2: Generalization in time and space. Hydrol. Earth Syst. Sci. 15, 3327–3341. https://doi.org/10.5194/hess-15-3327-2011

Buizza, R., Houtekamer, P.L., Pellerin, G., Toth, Z., Zhu, Y., Wei, M., 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Mon. Weather Rev. 133, 1076–1097. https://doi.org/10.1175/mwr2905.1

Calvetti, L., Pereira Filho, A.J., 2014. Ensemble hydrometeorological forecasts using WRF hourly QPF and TOPMODEL for a middle watershed. Adv. Meteorol. 2014. https://doi.org/10.1155/2014/484120

Carpenter, T.M., Georgakakos, K.P., 2004. Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model. J. Hydrol. 298, 202–221. https://doi.org/10.1016/j.jhydrol.2004.03.036

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. J. Hydrol. 375, 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Corral, C., Sempere-Torres, D., Revilla, M., Berenguer, M., 2000. A semi-distributed hydrological model using rainfall estimates by radar. Application to Mediterranean basins. Phys. Chem. Earth, Part B Hydrol. Ocean. Atmos. 25, 1133–1136. https://doi.org/10.1016/S1464-1909(00)00166-0

Craig, J.R., S. Huang, A. Khedr, S. Pearson, S. Spraakman, G. Stonebridge, C.Werstuck, & C. Zhang., 2016. Raven: user's and developer's manual. Raven Version 2.1. URL: http://www.civil.uwaterloo.ca/jrcraig/Raven/Main.html. (Accessed November 10, 2018)

Crochemore, L., Ramos, M.H., Pappenberger, F., 2016. Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. Hydrol. Earth Syst. Sci. 20, 3601–3618. https://doi.org/10.5194/hess-20-3601-2016

Cui, B., Toth, Z., Zhu, Y., Hou, D., 2012. Bias correction for global ensemble forecast. weather forecast. 27, 396–410. https://doi.org/10.1175/WAF-D-11-00011.1

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. J. Hydrometeorol. 12, 713–728. https://doi.org/10.1175/2011JHM1347.1

De Roo, A.P.J., Gouweleeuw, B., Thielen, J., Bartholmes, J., Bongioannini-Cerlini, P., Todini, E., Bates, P.D., Horritt, M., Hunter, N., Beven, K., Pappenberger, F., Heise, E., Rivin, G., Hils, M., Hollingsworth, A., Holst, B., Kwadijk, J., Reggiani, P., Dijk, M. Van, Sattler, K., Sprokkereef, E., 2003. Development of a european flood forecasting system. Int. J. River Basin Manag. 1, 49–59. https://doi.org/10.1080/15715124.2003.9635192

Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.J., Hartman, R., Herr, H.D., Fresch, M., Schaake, J., Zhu, Y., 2014. The science of NOAA's operational hydrologic ensemble forecast service. Bull. Am. Meteorol. Soc. 95, 79–98. https://doi.org/10.1175/BAMS-D-12-00081.1

Duan, Q., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res. 28, 1015–1031.

Efstratiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrol. Sci. J. 55, 58–78. https://doi.org/10.1080/02626660903526292

Engeland, K., Braud, I., Gottschalk, L., Leblois, E., 2006. Multi-objective regional modelling. J. Hydrol. 327, 339–351. https://doi.org/10.1016/j.jhydrol.2005.11.022

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. J. Hydrol. 519, 2906–2919. https://doi.org/10.1016/j.jhydrol.2014.04.038

Fan, Y., van den Dool, H., 2011. Bias correction and forecast skill of ncep gfs ensemble week-1 and week-2 precipitation, 2-m surface air temperature, and soil moisture forecasts. Weather Forecast. 26, 355–370. https://doi.org/10.1175/WAF-D-10-05028.1

Fotakis, D., Sidiropoulos, E., & Loukas, A., 2014. Integration of a hydrological model within a geographical information system: application to a forest watershed. Water, 6(3), 500–516. doi: 10.3390/w6030500

Gaborit, É., Anctil, F., Fortin, V., Pelletier, G., 2013. On the reliability of spatially disaggregated global ensemble rainfall forecasts. Hydrol. Process. 27, 45–56. https://doi.org/10.1002/hyp.9509

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

Gochis, D.J., Barlage, M., Dugger, A., Fitzgerald, K., Karsten, L., Mcallister, M., Mccreight, J., Mills, J., Rafieeinasab, A., Read, L., Sampson, K., Yates, D., Yu, W., 2018. WRF-Hydro

technical description, (version 5.0). NCAR Tech. Note 107 pp. https://doi.org/Source Code DOI:10.5065/D6J38RBJ

Golding, B.., 2000. Quantitative precipitation forecasting in the UK. J. Hydrol. 239, 286–305. https://doi.org/10.1016/S0022-1694(00)00354-1

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. Water Resour. Res. 34, 751–763. https://doi.org/10.1029/97WR03495

Haghnegahdar, A., Tolson, B.A., Craig, J.R., Paya, K.T., 2015. Assessing the performance of a semi-distributed hydrological model under various watershed discretization schemes. Hydrol. Process. 29, 4018–4031. https://doi.org/10.1002/hyp.10550

Haghnegahdar, A., Tolson, B.A., Davison, B., Seglenieks, F.R., Klyszejko, E., Soulis, E.D., Fortin, V., Matott, L.S., 2014. Calibrating environment Canada's MESH modelling system over the great lakes basin. Atmosphere-Ocean 52, 281–293. https://doi.org/10.1080/07055900.2014.939131

Han, S., Coulibaly, P., 2019. Probabilistic flood forecasting using hydrologic uncertainty processor with ensemble weather forecasts. J. Hydrometeorol. JHM-D-18-0251.1. https://doi.org/10.1175/JHM-D-18-0251.1

Hashino, T., Bradley, a. a., Schwartz, S.S., 2007. Evaluation of bias-correction methods for ensemble streamflow volume forecasts. Hydrol. Earth Syst. Sci. 11, 939–950. https://doi.org/10.5194/hess-11-939-2007

Hay, L.E., Leavesley, G.H., Clark, M.P., Markstrom, S.L., Viger, R.J., Umemoto, M., 2006. Step wise multiple objective calibration of a hydrologic model for a snowmelt dominated basin. J. Am. Water Resour. Assoc. 42, 877–890.

Horat, C., Antonetti, M., Liechti, K., Kaufmann, P., Zappa, M., 2018. Ensemble flood forecasting considering dominant runoff processes: II. Benchmark against a state-of-the-art model-chain. Nat. Hazards Earth Syst. Sci. Discuss. 1–34. https://doi.org/10.5194/nhess-2018-119

Hrachowitz, M., Clark, M.P., 2017. HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. Hydrol. Earth Syst. Sci. 21, 3953–3973. https://doi.org/10.5194/hess-21-3953-2017

Jasper, K., Gurtz, J., Lang, H., 2002. Advanced flood forecasting in Alpine watersheds by coupling meteorological observations and forecasts with a distributed hydrological model. J. Hydrol. 267, 40–52. https://doi.org/10.1016/S0022-1694(02)00138-5

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. Hydrol. Earth Syst. Sci. 22, 1957–1969. https://doi.org/10.5194/hess-22-1957-2018

Kalcic, M.M., Chaubey, I., Frankenberger, J., 2015. Defining soil and water assessment tool (SWAT) hydrologic response units (HRUs) by field boundaries. Int. J. Agric. Biol. Eng.

8, 1–12. https://doi.org/10.3965/j.ijabe.20150803.951

Kaufmann, P., Schubiger, F., Binder, P., 2003. Precipitation forecasting by a mesoscale numerical weather prediction (NWP) model: eight years of experience. Hydrol. Earth Syst. Sci. 7, 812–832. https://doi.org/10.5194/hess-7-812-2003

Khandekar, M.L., 2002. Trends and Changes in Extreme Weather Events : An assessment with focus on Alberta and Canadian Prairies, 1/927. ed, Alberta Environment. Alberta Environment, Edmonton.

Khu, S.-T., Madsen, H., di Pierro, F., 2008. Incorporating multiple observations for distributed hydrologic model calibration: An approach using a multi-objective evolutionary algorithm and clustering. Adv. Water Resour. 31, 1387–1398. https://doi.org/10.1016/j.advwatres.2008.07.011

Kitanidis, P.K., Bras, R.L., 1980. Real-time forecasting with a conceptual hydrologic model: 1. Analysis of uncertainty. Water Resour. Res. 16, 1025–1033. https://doi.org/10.1029/WR016i006p01025

Kouwen, N., Soulis, E.D., Pietroniro, A., Donald, J., Harrington, R.A., 1993. Grouped response units for distributed hydrologic modelling. J. Water Resour. Plan. Manag. 119, 289–305.

Kouwen, N., 1988. WATFLOOD: a Micro-Computer Based Flood Forecasting System Based on Real-Time Weather Radar. Can. Water Resour., J. 13, 62–77. https://doi.org/10.1007/s10228-005-0319-x

Kraußе, Cullmann, J., Saile, P., Schmitz, G.H., 2012. Robust multi-objective calibration strategies – possibilities for improving flood forecasting. Hydrol. Earth Syst. Sci 16, 3579–3606. https://doi.org/10.5194/hess-16-3579-2012

Leach, J.M., Kornelsen, K.C., Coulibaly, P., 2018. Assimilation of near-real time data products into models of an urban basin. J. Hydrol. 563, 51–64. https://doi.org/10.1016/J.JHYDROL.2018.05.064

Leta, O.T., van Griensven, A., Bauwens, W., 2017. Effect of single and multisite calibration techniques on the parameter estimation, performance, and output of a SAWT model of a spatially heterogeneous catchment. J. Hydrol. Eng. 22, 05016036. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001471

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

Lorenz, E.N., 1969. The predictability of a flow which possesses many scales of motion. Tellus 21, 289–307. https://doi.org/10.3402/tellusa.v21i3.10086

Madadgar, S., Moradkhani, H., Garen, D., 2014. Towards improved post-processing of hydrologic forecast ensembles. Hydrol. Process. 28, 104–122. https://doi.org/10.1002/hyp.9562

Madsen, H., Wilson, G., Ammentorp, H.C., 2002. Comparison of different automated strategies for calibration of rainfall-runoff models. J. Hydrol. 261, 48–59. https://doi.org/10.1016/S0022-1694(01)00619-9

Maxey, R., Cranston, M., Tavendale, A., Buchanan, P., 2012. The use of deterministic and probabilistic forecasting in countrywide flood guidance in Scotland. Hydrol. a Chang. world 01–07. https://doi.org/10.7558/bhs.2012.ns33

Michaels, S., 2015. Probabilistic forecasting and the reshaping of flood risk management. J. Nat. Resour. Policy Res. 7, 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. Hydrol. Model. Water Cycle 1–24. 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. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Ozdemir, A., Leloglu, U.M., Abbaspour, K.C., 2017. Hierarchical approach to hydrological model calibration. Environ. Earth Sci. 76, 318. https://doi.org/10.1007/s12665-017-6560-6

Pappenberger, F., Bartholmes, J., Thielen, J., Cloke, H.L., Buizza, R., de Roo, A., 2008. New dimensions in early flood warning across the globe using grand-ensemble weather predictions. Geophys. Res. Lett. 35, L10404. https://doi.org/10.1029/2008GL033837

Pappenberger, F., Beven, K.J., Hunter, N.M., Bates, P.D., Gouweleeuw, B.T., Thielen, J., Roo, A.P.J. De, 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). Hydrol. Earth Syst. Sci. Discuss. 9, 381–393. https://doi.org/10.5194/hess-9-381-2005

Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., Verseghy, D., Soulis, E.D., Caldwell, R., Evora, N., Pellerin, P., 2007. Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale. Hydrol. Earth Syst. Sci. 11, 1279–1294. https://doi.org/10.5194/hessd-3-2473-2006

Pokhrel, P., Gupta, H. V., 2010. On the use of spatial regularization strategies to improve calibration of distributed watershed models. Water Resour. Res. 46, 1–17. https://doi.org/10.1029/2009WR008066

Pomeroy, J. W., Gray, D. M., Brown, T., Hedstrom, N. R., Quinton, W. L., Granger, R. J., & Carey, S. K., 2007. The cold regions hydrological model: a platform for basing process representation and model structure on physical evidence. Hydrological Processes, 21(19), 2650–2667. doi: 10.1002/hyp.6787

Prein, A.F., Liu, C., Ikeda, K., Trier, S.B., Rasmussen, R.M., Holland, G.J., Clark, M.P., 2017.

Increased rainfall volume from future convective storms in the US. Nat. Clim. Chang. 7, 880–884. https://doi.org/10.1038/s41558-017-0007-7

Razavi, T., Coulibaly, P., 2016. Improving streamflow estimation in ungauged basins using a multi-modelling approach. Hydrol. Sci. 61, 2668–2679. https://doi.org/10.1080/02626667.2016.1154558

Refsgaard, J.C., 1997. Parameterisation, calibration and validation of distributed hydrological models. J. Hydrol. 198, 69–97. https://doi.org/10.1016/S0022-1694(96)03329-X

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resour. Res. 46, 1–22. https://doi.org/10.1029/2009WR008328

Renner, M., Werner, M.G.F., Rademacher, S., Sprokkereef, E., 2009. Verification of ensemble flow forecasts for the River Rhine. J. Hydrol. 376, 463–475. https://doi.org/10.1016/j.jhydrol.2009.07.059

Haghnegahdar, A., Tolson, B.A., Davison, B., Seglenieks, F.R., Klyszejko, E., Soulis, E.D., Fortin, V., Matott, L.S., 2014. Calibrating environment Canada's MESH modelling system over the great lakes basin. Atmosphere-Ocean 52, 281–293. https://doi.org/10.1080/07055900.2014.939131

Sanzana, P., Jankowfsky, S., Branger, F., Braud, I., Vargas, X., Hitschfeld, N., Gironás, J., 2013. Computer-assisted mesh generation based on hydrological response units for distributed hydrological modeling. Comput. Geosci. 57, 32–43. https://doi.org/10.1016/j.cageo.2013.02.006

Seibert, J., McDonnell, J.J., 2002. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. Water Resour. Res. 38, 23-1-23–14. https://doi.org/10.1029/2001wr000978

Seiller, G., Anctil, F., Perrin, C., 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrol. Earth Syst. Sci. 16, 1171–1189. https://doi.org/10.5194/hess-16-1171-2012

Seiller, G., Roy, R., Anctil, F., 2017. Influence of three common calibration metrics on the diagnosis of climate change impacts on water resources. J. Hydrol. 547. https://doi.org/10.1016/j.jhydrol.2017.02.004

Shafii, M., Basu, N., Craig, J.R., Schiff, S.L., Van Cappellen, P., 2017. A diagnostic approach to constraining flow partitioning in hydrologic models using a multiobjective optimization framework. Water Resour. Res. 53, 3279–3301. https://doi.org/10.1002/2016WR019736

Shamir, E., Carpenter, T., Fickenscher, P., 2006. Evaluation of the National Weather Service operational hydrologic model and forecasts for the American River basin. J. Hydrol. Eng. 11, 392–407. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:5(392)

Shin, M.J., Guillaume, J.H.A., Croke, B.F.W., Jakeman, A.J., 2013. Addressing ten questions about conceptual rainfall-runoff models with global sensitivity analyses in R. J. Hydrol.

https://doi.org/10.1016/j.jhydrol.2013.08.047

Shinma, T.A., Reis, L.F.R., 2014. Incorporating multi-event and multi-site data in the calibration of SWMM. Procedia Eng. 70, 75–84. https://doi.org/10.1016/j.proeng.2014.02.010

Singh, S., Bárdossy, A., 2015. Hydrological model calibration by sequential replacement of weak parameter sets using depth function. Hydrology 2, 69–92. https://doi.org/10.3390/hydrology2020069

Sitterson, J., Knightes, C., Parmar, R., Wolfe, K., Muche, M., Avant, B., 2017. An overview of rainfall-runoff model types. U.S. Environ. Prot. Agency 0–29.

Sun, N.-Z., Sun, A., 2015. Model Calibration and Parameter Estimation, Springer. https://doi.org/10.1007/978-1-4939-2323-6

Tang, Y., Reed, P., Wagener, T., 2005. How effective and efficient are multiobjective evolutionary algorithms at hydrologic model calibration? Hydrol. Earth Syst. Sci. Discuss. 2, 2465–2520. https://doi.org/hal-00298787 HAL

Thiboult, A., Anctil, F., Boucher, M.A., 2016. Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrol. Earth Syst. Sci. 20, 1809–1825. https://doi.org/10.5194/hess-20-1809-2016

Thiboult, A., Anctil, F., Ramos, M.H., 2017. How does the quantification of uncertainties affect the quality and value of flood early warning systems? J. Hydrol. 551, 365–373. https://doi.org/10.1016/j.jhydrol.2017.05.014

Thiemig, V., Bisselink, B., Pappenberger, F., Thielen, J., 2015. A pan-African medium-range ensemble flood forecast system. Hydrol. Earth Syst. Sci. 19, 3365–3385. https://doi.org/10.5194/hess-19-3365-2015

Unduche, F., Tolossa, H., Senbeta, D., Zhu, E., 2018. Evaluation of four hydrological models for operational flood forecasting in a Canadian Prairie watershed. Hydrol. Sci. J. 63, 1–17. https://doi.org/10.1080/02626667.2018.1474219

Velázquez, J.A., Anctil, F., Ramos, M.H., Perrin, C., 2011. Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. Adv. Geosci. 29, 33–42. https://doi.org/10.5194/adgeo-29-33-2011

WMO, 2011. Manual on flood forecasting and warning: WMO-No. 1072. Geneva.

Wood, A.W., Schaake, J.C., 2008. Correcting errors in streamflow forecast ensemble mean and spread. J. Hydrometeorol. 9, 132–148. https://doi.org/10.1175/2007JHM862.1

Xia, Y., Pitman, A.J., Gupta, H.V., Leplastrier, M., Henderson-Sellers, A., 2002. Calibrating a land surface model of varying complexity using multicriteria methods and the Cabauw dataset. J. Hydrometeorol. 3, 181–194. https://doi.org/10.1175/1525-7541(2002)003<0181:CALSMO>2.0.CO;2

Zahmatkesh, Z., Jha, S.K., Coulibaly, P., Stadnyk, T., 2019. An overview of river flood forecasting procedures in Canadian watersheds. Can. Water Resour. J. 44, 213–229. https://doi.org/10.1080/07011784.2019.1601598

Zapata, V., Alberto, J., 2010. Evaluation of hydrological ensemble prediction systems for operational forecasting.

Zhang, X., Flato, G., Kirchmeier-Young, M., Vincent, L., Wan, H., Wang, X., Rong, R., Fyfe, J., Li, G., Kharin, V.V., 2019. Changes in temperature and precipitation across Canada. Canada's Chang. Clim. Rep. 112–193.

Zhang, X., Srinivasan, R., Van Liew, M., 2010. On the use of multi-algorithm, genetically adaptive multi-objective method for multi-site calibration of the SWAT model. Hydrol. Process. 24, 955–969. https://doi.org/10.1002/hyp.7528

Zsótér, E., Pappenberger, F., Smith, P., Emerton, R.E., Dutra, E., Wetterhall, F., Richardson, D., Bogner, K., Balsamo, G., 2016. Building a Multimodel Flood Prediction System with the TIGGE Archive. J. Hydrometeorol. 17, 2923–2940. https://doi.org/10.1175/JHM-D-15-0130.1

Chapter 2. Event-based model calibration approaches for selecting representative distributed parameters in semi-urban watersheds

Summary of Paper 1: Awol, F.S., Coulibaly, P., Tolson, B.A. (2018). Event-based model calibration approaches for selecting representative distributed parameters in semi-urban watersheds. *Advances in Water Resources*, 118, 12-27.

In this research, an event-based calibration approach integrating multi-site, and single and multi-objective optimizations is proposed to improve peak flow prediction at interior and outlet gauging stations of a semi-urban catchment. Comparison has been performed between multi-site simultaneous (MS- S), multi-site average objective function (MS-A), multi-event multi-site (ME-MS)) and a benchmark at-catchment outlet (OU) calibration methods.

Key findings of this research study are:

- The proposed calibration and optimization formulation successfully identified representative model parameter sets.
- The two multi-site approaches (MS-S and MS-A) have better performances than multi-event ME-MS and at the catchment outlet OU approach.
- A comparison between optimized model parameter sets showed that the DDS optimization in MS-A approach improved the model performance at multiple sites.

2.1. Abstract

The objective of this study is to propose an event-based calibration approach for selecting representative semi-distributed hydrologic model parameters and to enhance peak flow prediction at multiple sites of a semi-urban catchment. The performance of three multi-site calibration approaches (multi-site simultaneous (MS-S), multi-site average objective function (MS-A), and multi-event multi-site (ME-MS)) and a benchmark at-catchment outlet (OU) calibration method, are compared in this study. Additional insightful contributions include assessing the nature of the spatio-temporal parameter variability among calibration events and developing an advanced event-based calibration approach to identify skillful model parameter-sets. This study used a SWMM5 hydrologic model in the Humber River Watershed located in Southern Ontario, Canada. For MS-S and OU calibration methods, the multi-objective calibration formulation is solved with the Pareto Archived Dynamically Dimensioned Search (PA-DDS) algorithm. For the MS-A and ME-MS methods, the single objective calibration formulation is solved with the Dynamically Dimensioned Search (DDS) algorithm.

The results indicate that the MS-A calibration approach achieved better performance than other considered methods. Comparison between optimized model parameter sets showed that the DDS optimization in MS-A approach improved the model performance at multiple sites. The spatial and temporal variability analysis indicates a presence of uncertainty on sensitive parameters and most importantly on peak flow responses in an event-based calibration process. This finding implied the need to evaluate potential model parameters sets with a series of calibration steps as proposed herein. The proposed calibration and

optimization formulation successfully identified representative model parameter set, which is more skillful than what is attainable when using simultaneous multi-site (MS-S), multi-event multi-site (MS-ME) or at basin outlet (OU) approach.

2.2. Introduction

Hydrological prediction in semi-urban watersheds requires a thorough understanding of the physical processes and the integrated response to storm events in partly urbanized and rural watersheds. In the last couple of decades, there have been research advances in understanding the urban and semi-urban hydrology with new emerging modeling tools. However, challenges remain due to the complex rainfall-runoff responses of combined urban, rural, and urbanizing areas. Such mixed responses could result in multiple peak flows, which increase prediction uncertainty (Fletcher et al., 2013). Consideration of the gradual loss of pervious surfaces in semi-urban areas within hydrological models is non-trivial because this transformation could lead to increased peak flows, and reduced flood duration and response time (Miller et al., 2014). Impervious surfaces, on the other hand, amplify irregular and periodic flows (Ackerman et al., 2005). Although the research interest grows, there are only a few guidelines mentioned in calibrating urbanizing catchments. One possible reason is due to the challenges in transferring calibrated land cover parameters between catchments (Jacobson, 2011).

Despite their limitation in setting realistic initial conditions, event-based models are conservative in nature in simulating individual flood hydrographs and peak flows and provide better flood prediction when compared to continuous hydrological models (Tramblay et al., 2012; WMO, 2011). Several event-based models have been used for urban and semi-urban catchments. For example, El-Hassan et al., 2013, compared the performances of a conceptual HEC-HMS model and physically-based distributed Gridded Surface Subsurface Hydrologic Analysis (GSSHA) model in simulating flood events of a

semi-urban watershed and showed that the latter performed better. To identify the dominant peak flow mechanisms, Kennedy et al., (2013), used the Kinematic Runoff and Erosion Model (KINEROS2) in a semi-arid urban environment, while Zhang et al., (2013), applied Dynamic Watershed Simulation Model (DWSM) in semi-urban landscape. The effect of urbanization on hydrological responses is well studied by using several models, such as Catchment hydrological cycle Assessment Tool (CAT) (Miller et al., 2014), Distributed Hydrology–Soil–Vegetation Model (DHSVM) (Cuo et al. , 2008), a coupled Conversion of Land Use and its Effect at Small regional extent (CLUE-E) and Soil and Water Assessment Tool (SWAT) (Arnold et al.,1998; Zhou et al., 2013) model. Event-based models were also used to assess their ability to reproduce past extreme, catastrophic flood events (Furl et al., 2015; Ogden et al., 2000; Sharif et al., 2013; Sharif et al., 2010).

The most widely used model for simulating extreme events in urban and semi-urban areas is the Environmental Protection Agency's Storm Water Management Model (SWMM) (Huber & Dickinson, 1988; Rossman, 2010). Gironás et al., (2010), studied the effects of various urban terrain morphologies on peak flow simulation by the SWMM model. Sun et al., 2014, compares two levels of SWMM catchment discretization (macro and micro-scale) to examine the degree of parameterizations and uncertainties using GLUE. Some advances were made on the calibration strategies of the SWMM. Krebs et al., (2013), and Zhang et al., (2013), employed Non-dominated Sorted Genetic Algorithm-II (NSGAII) and its revised version (ε-NSGAII), respectively, to optimize representative Low Impact Development (LID) scenarios in a small urbanized catchment. Herrera et al., (2006), also used NSGA-II with SWMM to analyze the trade-offs between low, medium, and high

flows. Barco et al., (2008), utilized a weighted multi-objective function and alternating starting points or constraints to optimize coupled GIS/SWMM4 model for the large urban catchment. Zaghloul et al., (2001), used Generalized Regression Neural Network to improve PCSWMM98 model simulation with inverse calibration technique, which was applied in an impervious test area.

In the application of event-based hydrological models for peak flow prediction, the question of which calibrated model parameter sets should be used can create a practical dilemma, unlike with continuous models. Despite the above efforts in improving the simulation and prediction capabilities of event-based models, novel methods are still required to address the uncertainties associated with model parameterization and temporal variations of input storm events. Robust calibration and validation approaches are required to identify optimum model parameters and improve runoff predictions (Krauße et al., 2012). Calibration procedures of hydrological models vary by their intended purpose, characteristics of the watershed, and the type and complexity of the models. The traditional approach is to calibrate the entire catchment (lumped or distributed) parameters according to model predictive performance at the basin outlet assessed via single or multiple objectives. Some authors have proposed advancing the single site calibration with a sequential/hierarchical approach (Hay et al., 2006; Ozdemir et al., 2017; Singh & Bárdossy, 2015). While the first authors sequentially calibrate a model's performance of potential evapotranspiration, water balance, and daily runoff, the second authors divided sub-basins into two hydrologic response units (HRU) and two further child HRUs based on influential parameters such as curve number and hydraulic conductivity. However, the limitation of

single site approach in improving runoff simulation at interior sites of a distributed catchment has motivated multi-site calibration methods.

One straightforward and efficient way of calibrating models to a set of distinct events would be using all calibration events in a series, yielding a unique parameter set per event, and then select the final parameter set as the one that performs best in terms of average performance across all the events (in this paper, multi-event multi-site calibration approach). However, this could lead to under- or over-estimation of flows for any arbitrary event and marks a high compromise in searching parameter sets that satisfies all events at once.

A fairly reasonable and default multi-site calibration approach to consider internal gauges is by using a weighted average of performance metrics across the gauging sites (Asadzadeh et al., 2014; Engeland et al., 2006; Haghnegahdar et al., 2014; Khu et al., 2006; Khu et al., 2008; Madsen et al., 2002; Shinma & Reis, 2014; Xia et al, 2002; Zhang et al., 2009). These studies applied continuous calibration with different types of models. Haghnegahdar et al., (2014), for example, used this approach to calibrate the Canada's Modélisation Environmentale-Surface et Hydrologie (MESH) model (Pietroniro et al., 2007) by aggregating the objective function of multiple sites into a single objective and highlighted that the method has lower computational cost than other methods involving multi-objective optimization techniques.

As an alternative to the above approach, some authors proposed multi-site simultaneous calibration approach to exclusively implement multi-objective optimization technique and generate a set of non-dominated calibration solutions (Leta et al., 2017; Zhang et al., 2010).

With this approach, objective functions at the interior sites are optimized at the same time, and the optimization result shows the tradeoffs between objective functions. Leta et al. (2017), applied a multi-site simultaneous calibration in developing SWAT Model for a heterogeneous catchment. Zhang et al., (2010), compared three optimization algorithms for multi-site simultaneous calibration of the SWAT model. The study highlighted that a multi-algorithm, genetically adaptive multi-objective method (AMALGAM) outperforms commonly used evolutionary multi-objective optimization such as Strength Pareto Evolutionary Algorithm 2 (SPEA2) and Non-dominated Sorted Genetic Algorithm II (NSGA-II). The above two studies were applied in continuous calibration approach for SWAT model. Other authors also considered multi-site step-wise/cascade (Brocca et al., 2011; Cao et al., 2006; Wang et al., 2012; Wi et al., 2015; Xue et al., 2016). Brocca et al., (2011), for example, used a distributed model with a sequential (step by step) calibration procedure to investigate its importance in flood forecasting and argued that the model improved peak flow estimation at internal sites.

To overcome the challenge of high computational cost in iterating through each sub-basin of a distributed catchment in multi-objective global search, the adaptation of tools with parsimonious characteristics is non-trivial. Asadzadeh & Tolson, (2009) developed a promising optimization tool, Pareto Archived Dynamically Dimensioned Search (PA-DDS), which is the multi-objective version of Dynamically Dimensioned Search (DDS) (Tolson and Shoemaker, 2007). PA-DDS has been compared with benchmark algorithms of NSGA-II and AMALGAM (Asadzadeh & Tolson, 2009), ε-NSGAII and AMALGAM (Asadzadeh and Tolson, 2013), and NSGAII and SPEA2 (Asadzadeh and Tolson, 2012)

and the authors concluded that PA-DDS showed improved performances with limited computational cost compared to alternative algorithms.

Behavioral parameter sets of distributed models should be identified with an efficient optimization algorithm to help overcome problems of uncertainty and over-parameterization. For example, parameters derived from the calibration process do not always give improved performances in a validation period (Beven, 1989; Beven and Freer, 2001; Brocca et al., 2011; Madsen, 2003). Mediero et al., 2011, claim that the presence of multiple acceptable parameter sets not only avoids "equifinality," but also leads to an ensemble of flood event simulations, which provide probabilities. During the calibration process, they identified the Pareto solutions and fitted a distribution function to estimate bias and confidence intervals of ensembles in the validation period.

One way of solving the problem associated with distributed catchment parameters is through the use of spatial regularization, as demonstrated by Pokhrel & Gupta, (2010). The authors used a non-linear transformation to reduce the number of parameters from Ng * Np (number of grid cells * number of parameters) to 3*Np by applying an adjustable multiplier, power term and additive constant to each prior estimated parameter value.

The above literature reviews indicate that most of multi-objective optimizations were conducted either for continuous distributed and lumped models or for an application other than flood prediction in semi-urban watersheds. The objective of this study is to develop and test different event-based calibration approaches for enhanced flood prediction in semi-urban distributed catchments. A second objective is to analyze the spatio-temporal parameter variability of calibrated parameter sets to address the uncertainty in event-based

parametrizations. The recent version of Storm Water Management Model (SMWM5) with DDS and PA-DDS optimization algorithms is used as calibration tools in this study. Section 2.3 describes the study area and data. Section 2.3 outlines the methodology including details of the model and optimization formulations, whereas the results and discussion are provided in Section 2.4. Finally, conclusion is presented in Section 2.5.

2.3. Study area and data

The research is conducted in the Humber River Watershed (Figure 2-1), which is located in Southern Ontario, Canada. The catchment area covers 911 km2, and the main Humber River drains to Lake Ontario. The distributed catchment is configured by dividing the basin into 714 sub-catchments with areas spanning between 4.3 ha (0.043 km2) and 860 ha (8.6 km^2). Humber River watershed is characterized as a semi-urban area with 54% rural, 33% urban and 13% urbanizing land covers and is administered by Toronto and Region Conservation Authority (TRCA, 2013). The hydrology and drainage patterns of the watershed are affected by its distinct topographic regions, which contain four hydrologic soil types (A, AB, B, BC, C, and D) (TRCA, 2008). The dual hydrologic soil groups AB and BC denote Sandy loam and Silt Loam soil types, respectively (NVCA, 2006).

Gauge rainfall and discharge measurements were collected from Environment Canada and Toronto and Region Conservation Authority. The temporal resolution of received data ranges from 5 to 30 minutes for rainfall data and 15 minutes to 1 hour for discharge records depending on availability. Ground-based rainfall data were used instead of gridded satellite or radar data because of unavailability of sub-hourly high-resolution temporal precipitation data in the study area. Niemi et al., (2017), also claimed that on-site gauge rainfall data

showed better runoff simulation performance than radar-based data in urbanizing catchments. In the Humber River Watershed (Figure 2-1), eleven rain gauges spatially distributed across the basin and five river flow gauging stations along the main tributaries, including one near the outlet have been used for this study. To separate the base flow from direct runoff, a simple straight-line hydrograph separation method is used (Ajmal et al., 2016; Deshmukh et al., 2013).

Significant rainfall events in spring periods are screened and selected based on criteria of (1) total rainfall amount larger than 20 mm (TRCA & AMEC, 2012), (2) spatial coverage and distribution in the watershed (rainfall amounts measured at most of the rain gauges in the watersheds), and (3) their consistency with the associated discharge measurement. As such, ten calibration events and four validation events were captured in the period spanning between 2007 and 2014 (Table 2-2).

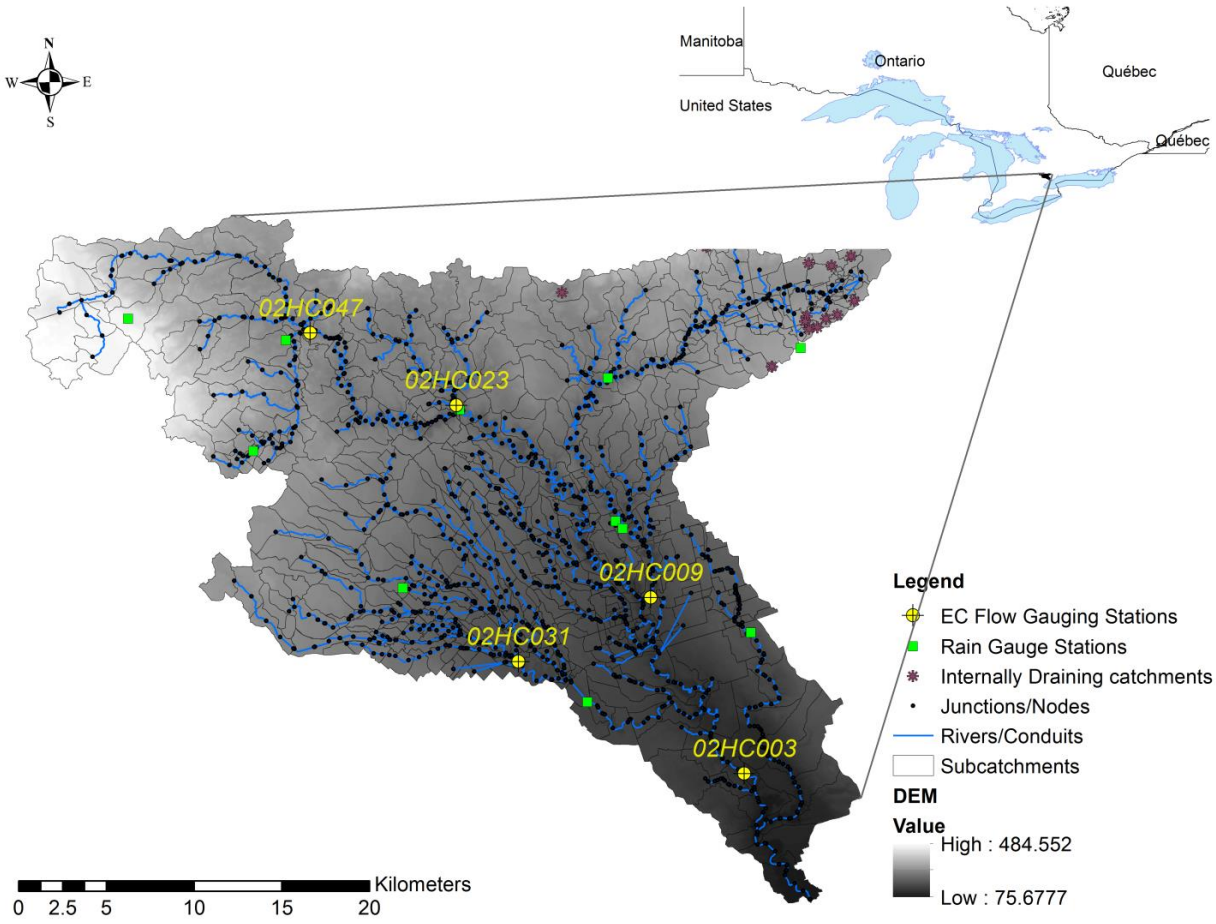

Figure 2-1: Location of the study area in Humber River Watershed, Southern Ontario.

Table 2-1: Description of SMWM5 model parameters

Parameter Codes	Description	Initial range of parameters**
IM*	Imperviousness [%]	0-99
W*	Characteristics Width of Overland flow [m]	163-124000
SP*	Depression storage in Pervious areas [mm]	1-600
CN*	Curve Number [-]	1-99
SL*	Catchment slope [%]	0.3-4.5
NI	Manning's n for overland flow in Impervious areas [-]	0.008-0.025
DT*	Drying time [days]	4-12
SM*	Depression storage in Impervious areas [mm]	0.2-5
NP	Manning's n for overland flow in Pervious areas [-]	0.08-0.4

*parameters used in calibration process
** The initial values of SWMM parameters were collected from the Toronto Region Conservation Authority (TRCA and AMEC, 2012)

2.4. Methods

2.4.1. Model setup

The Storm Water Management Model (SWMM) is a well-established event-based and continuous semi-distributed model used to simulate extreme events and peak flows in urban and semi-urban watersheds (Huber & Dickinson, 1988; Rossman, 2010). Due to the semi-urban characteristics of the study area and SWMM's wide application in operational flood forecasting (Randall et al., 2014; Robert et al., 2008), the recent version of SWMM (SWMM5) engine within PCSWMM (James et al., 2011) platform is used in this study. Curve number method and dynamic wave routing method have been used as an infiltration model and routing method respectively.

A sub-catchment in SWMM5 is represented by a non-linear reservoir model, where the conservation of mass is applied to generate overland flow (Rossman & Huber., 2015). By combining Conservation of Mass and Manning's equation, SWMM5 solves first the depth of a pond in sub-catchment (d) and then runoff at each time step using the following equations. More detailed information can be obtained from Rossman & Huber., (2015).

$$\frac{\partial d}{\partial t} = i - e - f - \alpha(d - d_s)^{5/3} \quad (2\text{-}1)$$

where, $\alpha = \frac{WS^{1/2}}{A\,n}$, in which each sub-catchment area (A) can be partitioned into pervious and impervious areas using the 'Percent Imperviousness' parameter. And the roughness (n) will be defined for each partition using the 'pervious manning's n' and 'impervious manning's n' parameters.

i = rate of rainfall + snowmelt (m/s)

e = surface evaporation rate (m/s)

i = infiltration rate (m/s)

d = ponded depth (m)

d_s= depression storage depth (m)

W = sub-catchment width (m)

S = sub-catchment slope (-)

Once d (ponded depth) is solved using equation 2-1 at each time step, the volumetric flow rate (Q in m^3/s) can be estimated by:

$$Q = \frac{WS^{1/2}}{n}(d - d_s)^{5/3} \quad (2\text{-}2)$$

Using the Curve Number method (in the current research) as an infiltration method and assuming the cumulative precipitation and infiltration at the start of the time step as P_1 and F_1 respectively, the infiltration rate (in m/s) is solved as follows (Rossman & Huber., 2015).

$$f = (F_2 - F_1)/\Delta t \quad (2\text{-}3)$$

where, $F_2 = P_2 - \frac{{P_2}^2}{P_2 + S_{max}}$

And, $S_{max} = \frac{25400}{CN} - 254$, where CN is the curve number and, S_{max} is the maximum soil moisture storage capacity (in mm).

Finally, the drying time (DT in days) is used to calculate a recovery constant (hr^{-1}), that is used to model the depletion and replenishment of the soil moisture storage capacity in the wet and dry period, respectively (Rossman & Huber., 2015).

SWMM5 consists of several physical and hydrological parameters to generate flow hydrograph, out of which nine catchment parameters (Table 2-1) are investigated to check their sensitivity against peak flow. 714 sub-catchments of Humber River watershed are assigned with unique parameter values. In Table 2-1, column three indicates the range of initial parameter values for 714 sub-catchments that are collected from previous studies and guidelines (CIVICA & TRCA, 2015; James, 2005). Event-by-event calibration and model testing are performed with simulation time steps of 15 or 30 minutes depending on input data time resolution. For defining the initial wetness of the watershed, the model was run for 1 to 2 weeks before each storm event as a 'warm up' period.

The methodology proposed in this study is summarized by a flowchart shown in Figure 2-2, which breaks down the calibration procedure into a series of phases. Phase 1 is the

model setup and calibration/validation data selection phase, which is described above. Phase 2 is the sensitivity analysis phase, the purpose of which is to find the most sensitive model parameters in semi-urban watersheds such as the Humber River Basin. Phase 3 is the spatial and temporal parameter variability assessment that aims to analyze the uncertainty associated with event-based calibration and variability of candidate parameter sets. In Phase 4, two calibrations steps are introduced. The first one compares four different types of calibration approaches and proposes ten individual candidate parameter sets obtained from the best optimization approach. The second step tests the candidate parameter sets to all calibration events and selects a certain number of parameter sets that have higher scores over the entire events and gauging sites. Phase 5 evaluates the candidate parameter set(s) in different events to refine the calibration output and select the best representative parameter set. The details and methodology associated with each of these phases are described sequentially in the following Sections (Section 2.4.2 to 2.4.5).

Table 2-2: Events selected for calibration and model testing

No.	Calibration Events	Amount of rainfall (mm)	Avg. Discharge* (mm)	Avg. Discharge* (m3/s)	No.	Validation Events	Amount of rainfall (mm	Avg. Discharge* (mm)	Avg. Discharge* (m3/s)
1	19-Aug-05	53.3	30.4	282.4	1	15-May-07	47.1	8.7	81.0
2	10-Jul-06	66.7	8.7	81.0	2	20-Oct-11	75.6	9.8	90.6
3	28-May-13	64.5	10.8	100.0	3	5-Sep-14	84.1	8.3	76.8
4	8-Jul-13	81.9	29.0	269.0	4	29-Nov-11	75.2	15.9	147.2
5	31-Jul-13	74.5	5.1	47.0					
6	27-Jul-14	29.8	7.3	67.3					
7	20-Aug-09	19.9	6.8	62.7					
8	28-Sep-10	41.4	5.4	50.3					
9	13-May-11	64.2	9.7	90.1					
10	7-May-10	37.6	9.0	83.1					

* Average discharge measured at the outlet (HC003).

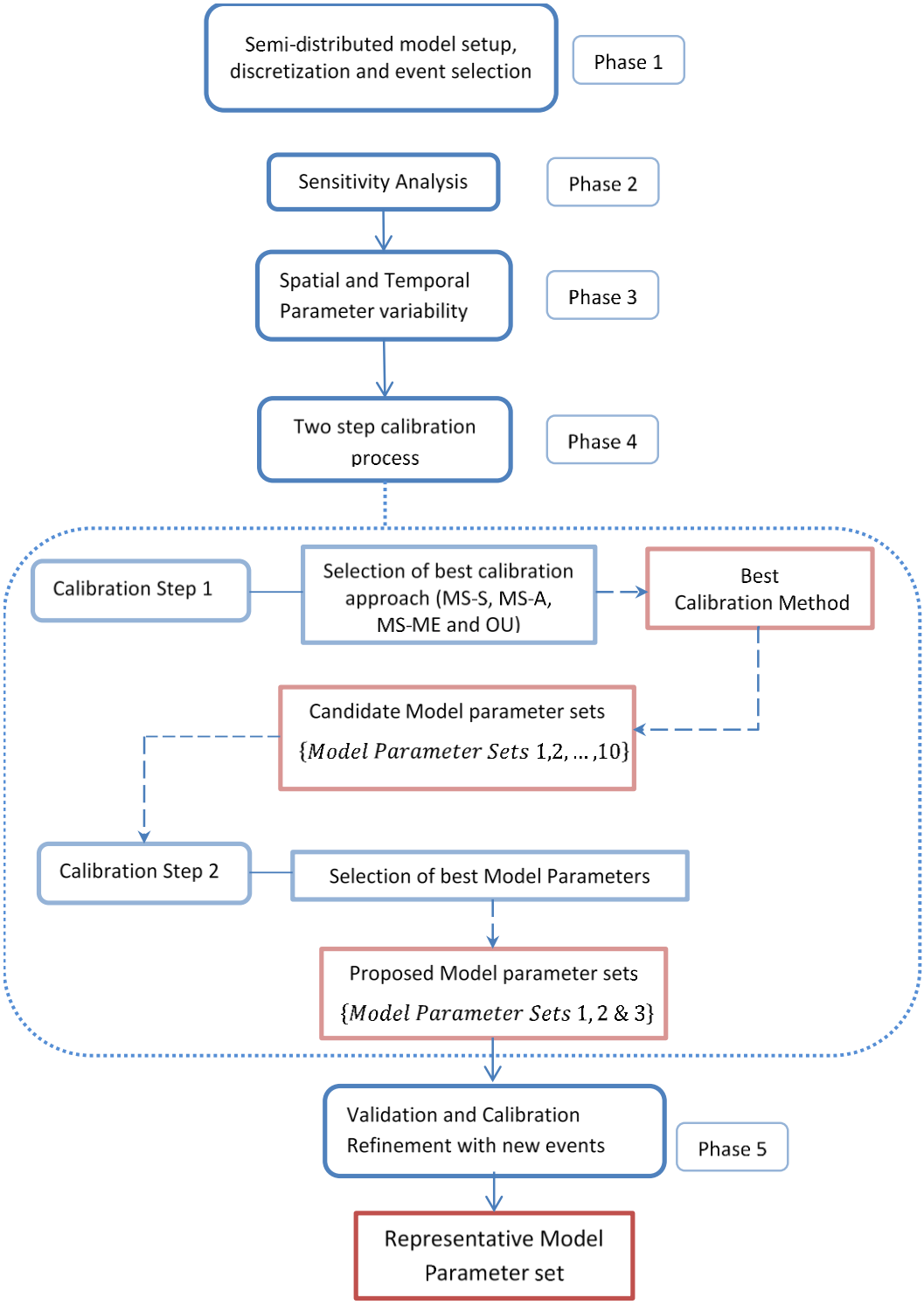

Figure 2-2: Flowchart of proposed approach for selecting representative parameter set in event-based models

2.4.2. Sensitivity analysis

The sensitivity of different versions of SWMM model parameters has been tested in different rural and urban watersheds (Barco et al., 2008; Irvine, et al., 1993). In this research, the purpose of sensitivity analysis of SWMM5 model is to identify the most sensitive parameters for the study basin. It was conducted by using two methods: Regionalized Sensitivity Analysis (RSA) (Spear and Hornberger, 1980) and Cumulative Sum of the Normalized Reordered Output (CUSUNORO) (Plischke, 2012).

Regionalized Sensitivity Analysis (RSA): Also called Generalized Sensitivity analysis or Hornberger-Spear-Young-method (Spear and Hornberger, 1980), RSA is used to identify the most sensitive parameters by distinguishing behavioral and non-behavioral parameter sets for Nash-Sutcliffe Efficiency (NSE), Peak flow Error (PE) and Volume Error (VE) model performances. 3500 parameter sets were generated by using Pareto Archived Dynamically Dimensioned Search (PA-DDS) (Asadzadeh and Tolson, 2013) optimization algorithm. The sensitivity was measured by Kolmogorov–Smirnov test statistics, which evaluates the maximum vertical distance between the curves of the cumulative distribution function of behavioral $F_n(x)$ and non-behavioral $F_{n'}(x)$ parameter sets as defined by:

$$d_{n,n'} = \sup_x |F_n(x) - F_{n'}(x)| \tag{2-4}$$

Where, $d_{n,n'}$ is the maximum vertical distance and *sup* is the supremum function. $d_{n,n'}$ (hereafter called RSA index) value ranges between 0 and 1 representing the limit between the most insensitive and sensitive parameters, respectively. Most sensitive parameters would have a higher maximum vertical distance between the curves of $F_n(x)$ and $F_{n'}(x)$.

Cumulative Sum of the Normalized Reordered Output (CUSUNORO): Initially proposed by Plischke, 2012, CUSUNORO is a graphical post-processing method to represent the first-order sensitivity index. Its principle is withdrawn from the ideas of Contribution to the Sample Mean (CSM) plot (Bolado-Lavin et al., 2009). CSM and CUSUNORO are found to be suitable for estimating the main effect, the first-order variance-based sensitivity index for cases where there is no direct access to the sampling procedure and the simulation model to map input-output relationship (Plischke, 2012).

Let π denote an arrangement of ordered values of input parameters sorted in ascending order, i.e, $x_{\pi(i)} = \{x_{i,1}, x_{i,2}, \dots, x_{i,n}\}$; hence its corresponding sorted series of outputs $y_{\pi(i)}$ can be created for all x. A scaling factor, which resembles the output variance is then created using the square root of the sum of squares $s_{yy} = \sum_{i=1}^{n}(y_{(i)} - \bar{y})$ (Plischke, 2012). Finally, the cumulative sum of normalized reordered output is defined as:

$$z(i) = \frac{1}{\sqrt{n.\, s_{yy}}} \sum_{j=1}^{i} \left(y_{\pi(i)} - \bar{y} \right) \tag{2-5}$$

The CUSUNORO values, $z(i)$, can then be plotted against the empirical cumulative distribution of input parameters x_i to visualize the sensitivity of individual parameters on the output statistics.

SWMM5 model parameters (Table 2-1) are considered as inputs, and different performance metrics were used as outputs. Input-output mapping is performed externally by using Pareto Archived Dynamically Dimensioned Search (PA-DDS) (Asadzadeh and Tolson, 2013).

2.4.3. Spatial and temporal parameter variability

The primary objective of this section is to address the variability in event-based parametrizations in a semi-urban watershed and how it can be quantified by different calibration approaches. Before starting to apply alternative and new methods of calibration formulations and optimization algorithms, we perform this exercise using a benchmark calibration approach at the catchment outlet involving limited manual and multi-objective calibration. The calibration process is described in detail, together with the other proposed approaches in Section 2.4.4.1. The outcome assists to formulate and compare alternative event-based calibration approaches in reducing the uncertainties. Different parameterizations of the SWMM5 model represent several realizations of the physical process in the event of extreme spring rainfalls. Ten individual event-based calibrations result in ten SWMM5 model parameter sets. The variability of these sets regarding the model output, as well as differences of calibrated sensitive parameters among the events, was assessed.

First, the spread of two sensitive model parameters (Imperviousness and Drying Time) in each model parameter sets were assessed by developing box plots for different percentile values. Parameter values, collected from 714 sub-catchments, were ranked in ascending order, and their percentiles were extracted accordingly. The variability of calibrated parameters in space can be observed by the degree of the spread.

Second, the uncertainty of event-based parametrization in a distributed catchment was evaluated by analyzing the peak flow response. We re-run the ten model sets for ten calibration events by regarding each model sets as an individual model and the peak flow

simulation results were extracted. The specific objective of this method is to check how variable simulated peak flows are within each model set as well as with the observation at multiple interior sites. Various boxplots were used to display standardized peak flow variability. The Standardized peak flow is calculated by normalizing the deviation of the simulated peak flows from observed peak flow by their standard deviation.

2.4.4. Model Calibration

2.4.4.1. Event-based Calibration approaches

Three multi-site event-based calibration approaches are compared with a benchmark 'At-catchment outlet' method to select potential parameters sets in Humber River basin. The calibration parameters in each of these four approaches are the same and are determined from the sensitivity analysis described above.

i) At Catchment Outlet (OU)

The conventional calibration approach of many hydrological models is to calibrate the entire catchment using a gauging station located at the basin outlet. In this approach, calibration to each of the ten events is completed independently. This calibration method is used as a benchmark to compare its results with other considered calibration approaches. Limited manual calibration is performed before using the following optimization formulation in order to get initialized solutions.

Single and multi-objective optimization techniques could be used to calibrate distributed models at basin outlets. Here, in order to find the best achievable parameter sets, multi-objective optimization with three different performance metrics (Nash-Sutcliffe Efficiency

(NSE) (Nash and Sutcliffe, 1970), Peak flow Error (PE) (Liong et al., 1995), and Volume Error (VE) (Niemi et al., 2017) are used to calibrate Humber River Watershed at HC003 gauging station. This formulation is similar to the one used by Barco et al., (2008), where they minimized a weighted objective function summing the total flow volume, peak flow rate, and instantaneous flow rate errors (each as percentage). The basic difference is that Barco et al., (2008) minimize/maximize a single weighted objective function by changing the weights depending on target flow type (e.g., peak flow or volume) whereas the approach here gives equal weight to individual objective functions and used a multi-objective PA-DDS algorithm to identify non-dominated solutions. The exercise is repeated ten times for ten calibration events with maximum iteration of 500 set for each optimization.

The multi-objective target is to maximize NSE and minimize PE and VE at station (a). i.e.

$$O_{humber} = \{o_1 = NSE_a,\ \ o_2 = VE_a,\ o_3 = PE_a\} \tag{2-6}$$

In which:

$$NSE = 1 - \frac{\sum\left(Q_{o,i} - Q_{s,i}\right)^2}{\sum\left(Q_{o,i} - \overline{Q_o}\right)^2}$$

$$VE = \frac{|V_o - V_s|}{V_o} \tag{2-7}$$

$$PE = \frac{\left|Q_{p,o} - Q_{p,s}\right|}{Q_{p,o}}$$

where, $Q_{o,i}$ & $Q_{s,i}$ are observed and simulate discharge at each time step, in cubic meter per second and $\overline{Q_o}$ is the average observed discharge; $Q_{p,o}$ & $Q_{p,s}$ are observed and simulated peak flows respectively; and V_o & V_s are the volume of water under observed and simulated

flow hydrographs respectively, in million cubic meter. NSE value ranges between –∞ and 1 with 1 indicating best performance. PE, and VE have values spanning between 0 and ∞ and better performing model sets would have values close to 0. The result of the OU calibration approach is ten parameter sets (for ten calibration events), with each set being made up of the average of non-dominated solutions corresponding to a specific flow event.

ii) Multi-Site Simultaneous multi-objective (MS-S)

Multi-objective optimization techniques have been frequently used to calibrate distributed models. A multi-objective optimization algorithm is used to find a feasible set of Pareto-optimal parameter solutions by minimizing or maximizing the objective function vector. i.e. $Min/Max\ \ \boldsymbol{O}(\boldsymbol{p}) = [o_1(p), o_2(p), o_3(p), \dots, o_m(p)]$ where the objective function vector $\boldsymbol{O}(\boldsymbol{p})$ is comprised of m objective functions or performance metrics (Zhang et al., 2010).

Multi-site simultaneous multi-objective optimization was previously considered for continuous calibration (Leta et al., 2017; Zhang et al., 2010). In the current study, it is applied for an event-based calibration process. In this calibration approach, optimization is performed independently for ten individual calibration events. For each event, the model's performance is assessed simultaneously across multiple gauging stations using Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970) performance metrics. In other words, the performance at each site in the study area is assessed by a different objective function so that performances at multiple locations are accounted for *simultaneously*. That is, for the five gauging stations in Humber River Watershed (represented by a, b, c, d, and e):

$$O_{humber} = \{o_1 = NSE_a,\ \ o_2 = NSE_b,\ o_3 = NSE_c,\ o_4 = NSE_d,\ o_5 = NSE_e\} \qquad (2\text{-}8)$$

For optimization, Pareto Archived Dynamically Dimensioned Search (PA-DDS) (Asadzadeh and Tolson, 2013) algorithm is applied to find the Pareto-optimal parameters sets. PA-DDS was used within OSTRICH (Matott, 2005) framework toolkit. The selection operation in PA-DDS of non-dominated solutions (Pareto-optimal solution) is performed using estimated Hypervolume Contribution (HVC) (Asadzadeh and Tolson, 2013). The maximum number of iterations is set as 500 and the perturbation parameter is left as the default value of 0.2. Since there are 10 calibration events, 10 PADDS optimization is performed to evaluate the objective function values of each solution.

The result of the MS-S calibration approach is multiple parameter sets or non-dominated solutions corresponding to a specific flow event. Then, equal weight is given to each objective functions $(o_1, o_2, o_3, o_4 \& o_5 \; in \; equation \; 2\text{-}8)$ to find the average of the non-dominated parameter sets and solutions for each calibration event.

iii) Multi-Site Average objective function (MS-A)

This calibration method is frequently used by several researchers to account for the interior sites of a semi- or fully distributed catchment in the calibration process by taking the weighted average of multiple objective functions. The objective functions at multiple gauging stations are aggregated into a single objective function. Then optimization is performed to maximize the aggregated single objective function.

The five sites of Humber River Watershed are evaluated by their respective Nash-Sutcliffe Efficiency index:

$$NSE = (NSEa + NSEb + NSEc + NSEd + NSEe)/5$$
$$O_{humber} = \{NSE\} \quad (2\text{-}9)$$

The single-objective function (O_{humber}) is optimized by using Dynamically Dimensioned Search (DDS) (Tolson and Shoemaker, 2007) optimization algorithm within OSTRICH framework (Matott, 2005). Similar to the MS-S approach, the MS-A DDS optimization is performed independently for 10 individual calibration events and the result is 10 candidate parameter sets. In addition, the maximum number of iterations of 500 and perturbation value of 0.2 was set.

With perfect algorithms that converge to true optimal solution/true set of non-dominated solutions, MS-A would yield one of the non-dominated solutions generated by solution of MS-S formulation. In all practical calibration situations, convergence to true optimal/Pareto-optimal set of solutions is not guaranteed and thus all results are approximate. The quality of the approximations to the true, but unknown solutions is dependent on the algorithm quality (DDS and PADDS) and is also dependent on the algorithm computational budget. PADDS and DDS computational budgets in terms of number of solutions evaluated in MS-A and MS-S are equivalent and set to 500 and replicated 10 times for 10 calibration events.

The main difference between the MS-A and the MS-S approach is on the optimization method. While MS-A is based on a single objective optimization scheme (see Equation 2-9), the MS-S approach employs multi-objective optimization function (see Equation 2-8). In the former MS-A approach, although it involves aggregating several objectives, it is based on a single objective calibration process with the help of Dynamically Dimensioned Search (DDS) method: i.e. the objective is to maximize a single NSE value which is the average NSE of all sites (including at the outlet). Conversely, MS-S approach aims to find

a feasible Pareto front by maximizing the objective function vector (rather than a single value): in which the vector comprises of NSEs at multiple sites including the outlet.

In MS-S approach, the non-dominated (Pareto-optimal) solutions are generated by finding a tradeoff between individual objective functions using Pareto Archived Dynamically Dimensioned Search (PA-DDS) algorithm. At each iteration, MS-S searches for a tradeoff of optimum parameters that simultaneously satisfies individual objective functions or simultaneously maximizes the performances of each NSEs (interior as well as outlet), whereas MS-A searches the best parameters of the whole 714 sub catchments that maximize a single NSE value (average of NSEs).

iv) Multi-event multi-site calibration (ME-MS)

This approach involves concatenating the simulated and observed discharge of separate events and treating it as a single time series. For the combined multi-event series, the performance metrics (NSE) are then computed at each gauging stations. The multi-site objective function is basically defined in a similar manner as the previous calibration approach (MS-A) (equation 2-9) and thus is also formulated as a single-objective optimization problem. One of the differences between ME-MS and the above two (MS-S and MS-A) approaches is that ME-MS is applied over all ten events, whereas the others performed event by event. The optimization was performed by DDS algorithm with maximum iteration of 500. The calibration result is one set of candidate parameter sets that are somehow appropriate for all ten flow events.

2.4.4.2. Calibration steps

In order to identify the best parameters sets across the calibration events, the results of the above four calibration approaches described in section 2.4.4.1 are processed and compared in the following two calibration steps.

Step 1:- Select best set of candidate solutions, (e.g., select best calibration approach):

Each calibration approach generates a set of candidate parameter sets. The calibration approach with better performance and score at each calibration event and gauging station is selected for the next step. This step comprises of a couple of processes. Initially, we calibrate the model to ten individual events (Table 2-2) using MS-S, MS-A, and OU approaches. At the end of each optimization or calibration approaches ten candidate parameters sets are foreseen for ten flow events. The performance of the final calibrated sets of parameters would be different for different optimization formulation. Therefore, in the next process we compared the result of these calibration approaches at each individual event. Here, since ME-MS approach is formulated by aggregating over ten calibration events, it results in one set of calibrated parameters for all events as opposed to the output of MS-S, MS-A, and OU approaches, which have ten sets of calibrated parameters. For comparison purposes, we re-apply the final calibrated parameter sets of ME-MS to ten events so that the results of four calibration approaches could be compared at individual events. In addition, comparison is also made at individual gauging stations (five sites). Finally, the best calibration approach that performed well at ten calibration events and five sites is proposed to the next calibration step. The final outcome of this step is ten calibrated parameter sets from one of the calibration approaches.

Comparison of calibration approaches is performed using model improvement scale or Prediction Error Decrease (PED) in percentage (Coulibaly, 2003) and Taylor Diagram (Taylor, 2001). The PED shows the model performance improvement of Multi-site simultaneous (MS-S), and Multi-site Average objective function (MS-A) and Multi-event multi-site (ME-MS) calibration approaches when compared to the benchmark At-Catchment Outlet (OU) approach at five gauging stations. Taylor diagram is used to precisely quantify and display the pattern similarity and statistics of different calibrated model parameter sets and the observation at multiple gauging sites. A revised normalized Taylor Diagram is constructed based on Kärnä & Baptista, 2016 by relating normalized centered root-mean-squared error with ratio of standard deviation of observed and simulated discharge and correlation coefficient through a Law of Cosines. The attributes of Taylor Diagram will be able to show the statistical proximity of individual model sets derived from two calibration approaches with the observation at five gauging stations. Details regarding Taylor Diagram can be found in Taylor, (2001).

Step 2:- From best approach candidate parameter sets, filter out poor candidates (e.g., select top three):

From the first step, ten candidate parameter sets are produced by the best calibration approach. But the performance of each candidate parameter sets in a different calibration event is not yet evaluated. In this step, we re-apply each candidate parameter set to all events and aggregate performance across the events and sites to score parameter sets. Then the most representative parameter sets are chosen based on the highest score. Normalized NSE is used to score the performance across the events and sites. Here the performance

criterion (NSE) is normalized by using the maximum and minimum values of the candidate model parameters sets at each site and event. Then the sum of the normalized NSE over the entire calibration events is estimated for each candidate model parameters sets. The top three potential model parameters sets with the highest total normalized NSE are registered and proposed for model testing and calibration refinement.

2.4.5. Validation

Validation was performed to test and refine the top three model parameter sets selected during calibration process using a data set independent of calibration period. We have selected four validation events (Table 2-2) that qualify the event selection criteria described Section 2.3. This phase is dedicated to select the most representative model parameter sets. The model testing and refinement is performed in four new events (Table 2-2). The three model sets are evaluated by using Taylor Skill Score (Taylor, 2001) to further corroborate the outcome of the previous two-step calibration processes. This score summarizes a Taylor diagram and defines a single skill score that measures the correlation coefficient and centered root-mean-squared error along with standard deviation (Taylor, 2001). It is defined as:

$$S = \frac{4(1+R)}{\left(\frac{\sigma_s}{\sigma_o} + \frac{1}{\sigma_s/\sigma_o}\right)^2 (1+R_o)} \tag{2-10}$$

Where: S indicates the Taylor Skill Score; σ_s is model variance; σ_o is observed variance; R is the model correlation coefficient, and R_o is maximum correlation attainable, here taken

as the maximum of model's correlation coefficient. The skill increases (approaches one) as σ_s and R get closer to σ_o and R_o respectively.

2.5. Results and discussion

2.5.1. Sensitivity analysis

The sensitivity analysis (Figure 2-3) indicates that Imperviousness (IM) is the most sensitive SWMM5 parameter to NSE, PE, and VE model performances in Humber River watershed. The RSA indexes show that after Imperviousness and Drying time (DT), Depression storage in Impervious areas (SM) and Pervious areas (SP) appear to be slightly sensitive to the model performances, particularly to Peak flow Error. This result is analogous to the plots of Cumulative Sum of the Normalized Reordered Output (CUSUNORO) (Figure 2-4). The CUSUNORO plots indicate that Imperviousness (IM) followed by Drying time (DT) have the largest first order contribution to NSE, VE, and PE as the departure of their cumulative sum of the normalized output from the horizontal line (y=0) is considerable. The different direction of CUSUNORO plots for NSE, VE, and PE indicates that the contribution of each parameter to the mean and variance and the output is positive if above the horizontal and negative if below the horizontal.

The results of both sensitivity analyses are reasonable for semi-urban areas like Humber River watershed, which covers about 50% pervious and 50% impervious areas. The rainfall-runoff response is governed by the percentage of imperviousness in the sub-catchments upstream of the gauging station and recovery time (drying time) of the saturated soil in pervious areas of the sub-catchments. In general, Imperviousness and Depression

storage are found to be the most sensitive parameters of SWMM model to peak flow and volume in urbanizing watersheds, which is also supported by Barco et al., (2008). For calibration, the SWMM parameters except Manning's n are considered as it has relatively less impact to NSE and Peak flow in both Impervious and Pervious areas.

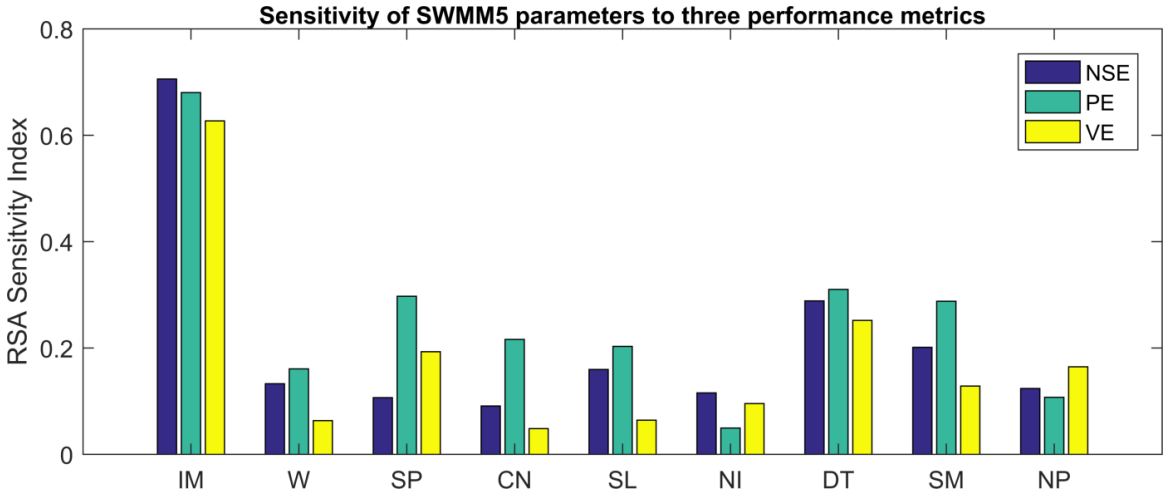

Figure 2-3: Output of Regionalized Sensitivity Analysis. Figure displays the sensitivity index value of nine SWMM5 parameters for Nash-Sutcliffe Efficiency (NSE), Peak Flow Error (PE) and Volume Error (VE). Higher RSA index corresponds to higher sensitivity of parameters to the output performance. Description of parameter letter codes (x-axis) is presented in Table 2-1.

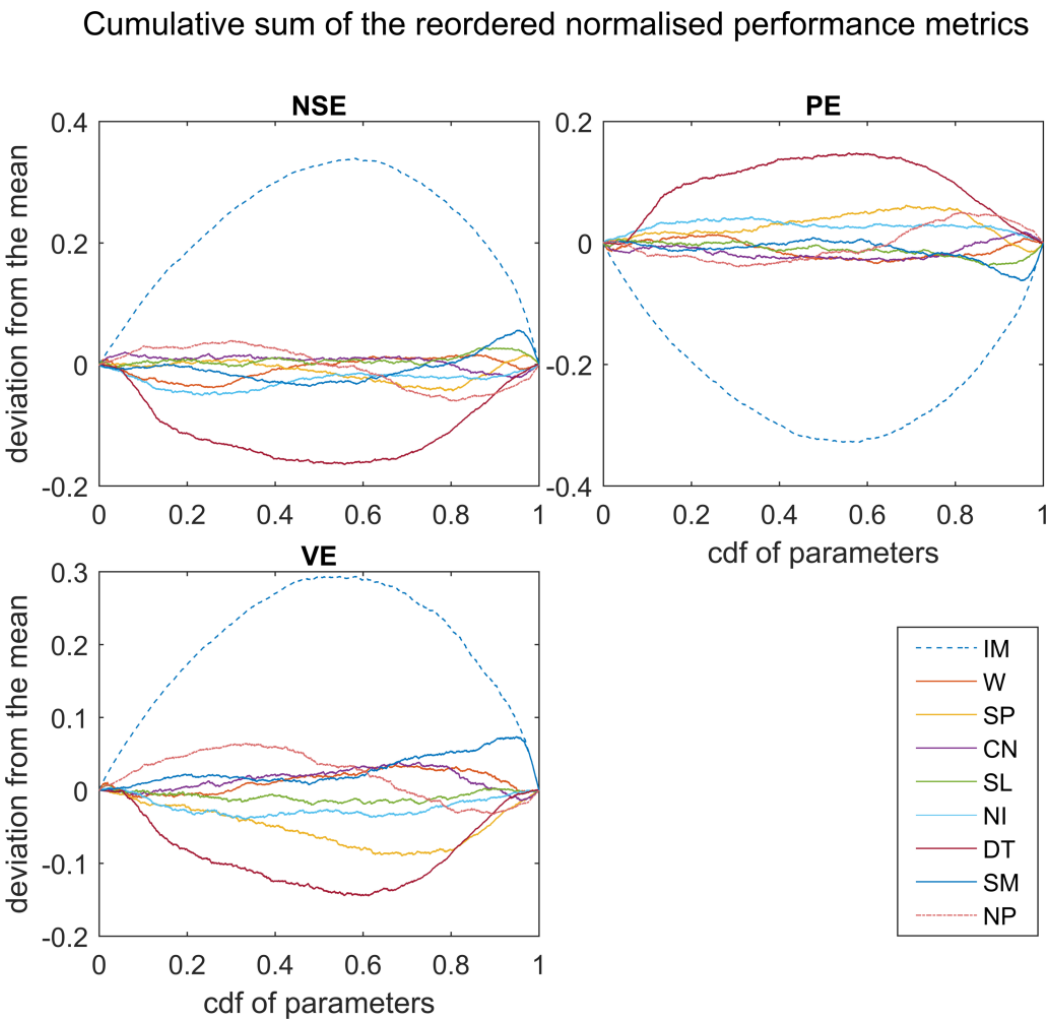

Figure 2-4: Cumulative Sum of the Normalized Reordered Output (CUSUNORO) used as first order sensitivity of SWMM5 parameters to three performance metrics (NSE, PE and VE). The deviation from the mean (CUSUNORO values or z(i) in Eqn. 5) is plotted against the empirical cumulative distribution of input parameters (x-axis). Higher deviation from the mean indicates higher sensitivity of parameters to corresponding performance metrics. Descriptions of parameter letter codes (for each colored lines of the plots) are presented in Table 2-1.

2.5.2. Spatial and temporal parameter variability

The study assessed the degree of uncertainty in event-based calibration of SWMM5 distributed model parameters sets that were obtained by an event-based calibration processes performed for ten calibration events. The parameter variability (uncertainty) was demonstrated by temporal scale (among calibration events) and spatial scale (within 714 sub-catchments). In Figure 2-5, the spatial variability of the two most sensitive parameters (Imperviousness and Drying Time) that are generated by ten calibrated parameter sets is shown. The medians and the interquartile ranges (IQR) of the box plots in higher percentile imperviousness values show variability between individual calibration events. Lower and medium percentiles values of imperviousness have relatively similar medians and IQRs among the parameter sets. In general, higher uncertainty is observed among the sub-catchments with higher imperviousness (>80% Imperviousness). This result can be reasonably expected from a semi-urban watershed where high impervious areas highly influence the rainfall-runoff response in the time of extreme events. Figure 2-5 also shows that pervious areas that have relatively faster recovery time to be in a drying state when saturated (<20% Drying Time or less than 5.5 days) shows higher variability or uncertainty. Rapid recovery time is often recognized in hydrologic soil group D such as medium and coarse sandy soils, which pertains to high rate of water transmission or infiltration (Rossman, 2010; NRCS, 2007).

Figure 2-6 shows the peak flow variability of the ten potential representative SWMM5 model parameter sets in different calibration events. The uncertainty is expressed by standardized peak flow deviation from the observation recorded at multiple gauging

stations. The degree of the deviation is quite significant in almost all events and measuring stations. The medians and associated IQRs are either above or below the green horizontal line (observation), which depicts underestimation and overestimation of peak flows by the potential model parameters sets. Outliers were also observed on many occasions. This investigation indicates the existence of high uncertainty in reproducing peak flows by the majority of model parameters sets. Within each boxplot, it can be seen that only one point (one model parameter set) matches (or close to matching) with observed peak flow, which is, in fact, the calibrated model parameter set for each event that the boxplot is constructed. The results of this variability analysis give an overview of the difficulty in selecting representative parameter sets in distributed semi-urban watersheds and the need for a robust method of calibration when dealing with event-based model parametrization.

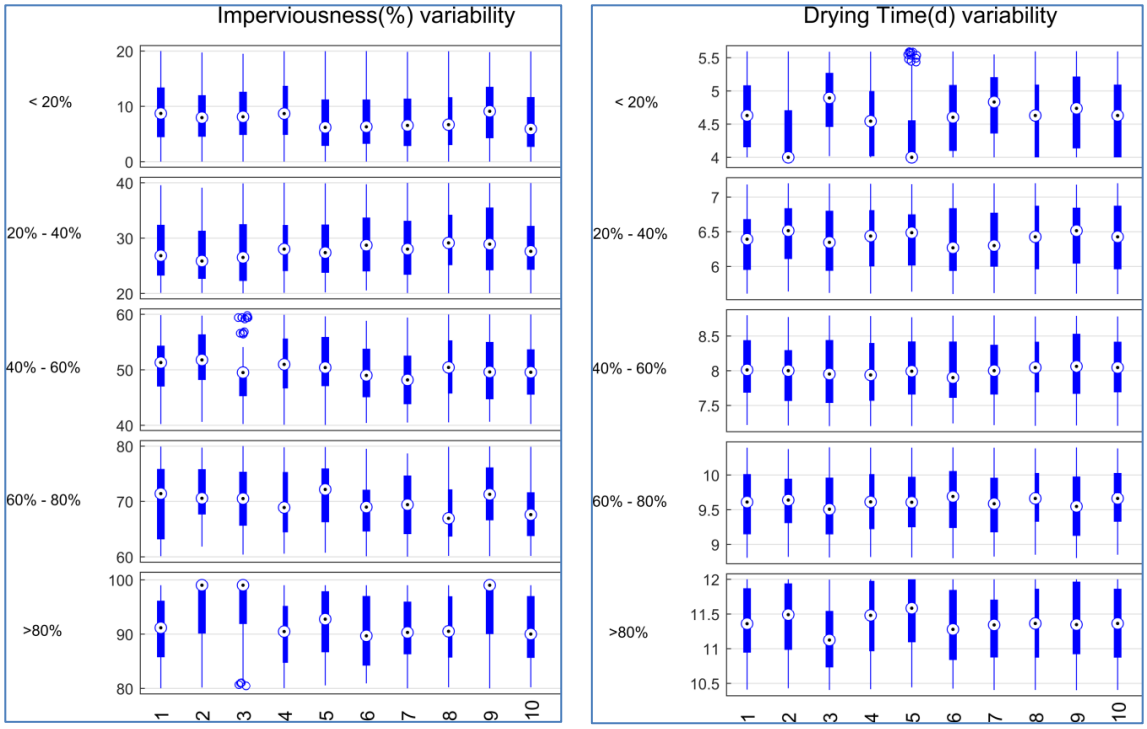

Figure 2-5: Box plots showing the spread of the lower, middle three and upper percentile values of most sensitive calibrated parameters (Imperviousness-Left and Drying Time-Right) to illustrate their variability in ten Model Sets (x-axis). Parameter values, collected from 714 sub-catchments, were ranked in ascending order. Model parameter sets represent different realization of the PCSWMM model in ten calibration events.

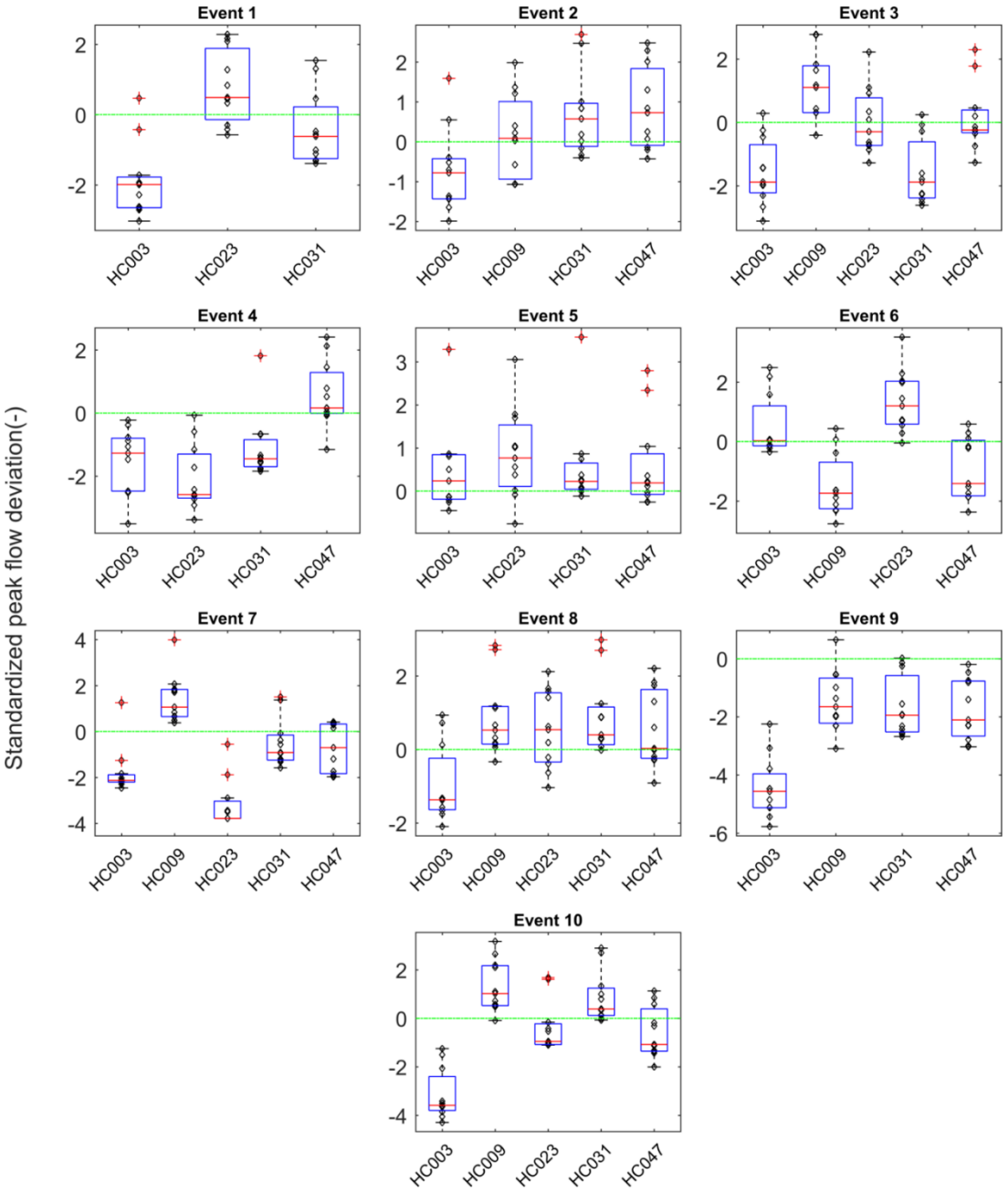

Figure 2-6: Figure showing Peak Flow variability of model parameter sets. 10 plots are constructed for 10 calibration events and each boxplot within a plot corresponds to different gauging stations. Individual boxplots are developed from 10 standardized peak flows, which are generated by ten different Model Parameter Sets in order to demonstrate the variability of different realizations of SWMM5 model. Standardized peak flows are calculated by normalizing the deviation of the simulated peak flow from observed peak flow by the standard deviation of the simulated peak flow. Green horizontal line along the zero y-axis is computed based on observed peak flow.

2.5.3. Calibration Approaches

The outputs from the four multi-objective calibration approaches presented in section 2.4.4.1 (MS-S, MS-A, ME-MS and OU) are evaluated in ten individual calibration events at five gauging stations. Their performances are compared at each calibration steps mentioned in section 2.4.4.2.

Figure 2-7 and Figure 2-8 present the comparison of calibration approaches for the first calibration step. The relative improvement of Multi-site average objective function (MS-A) and Multi-site simultaneous (MS-S) over the benchmark At-catchment outlet (OU) is quantified by the prediction (simulation) error decrease (PED) percentage. The PED (in Figure 2-7) shows the improvement of NSE of both MS-A and MS-S approaches when compared to OU at five gauging stations. Using either of the multi-site calibration approaches improves the model performance by about 28% in the interior sites when compared to the conventional at catchment outlet calibration method. Comparing the two multi-site optimization methods, aggregating the objective functions over the gauging stations (MS-A) gives a fairly better performance than calibrating the multiple sites simultaneously (MS-S). With a reference to the benchmark OU calibration, the NSE performance metric of MS-A is improved by an average of 43% as compared to MS-S where it was improved by only 29%. In fact, only 4 out of 42 calibration events and stations show slightly higher NSE performance for MS-S; out of which 3 are at the outlet. At the outlet, there are some occasions where the benchmark OU calibration shows improved performance over both MS-S and MS-A. This is a reasonable because it is generally easier to improve the performance at one location during optimizing. The calibrated parameter

sets from Multi-event multi-site (ME-MS) calibration approach is re-applied for each calibration event to evaluate and compare its result with the other methods. It is found that the performance of ME-MS is significantly lower than both multi-site optimizations as well the benchmark calibration approach. Although not shown in Figure 2-7 due to its high percentage difference to present in PED metrics with other calibration approaches, the comparison is shown in Figure 2-8.

The performance of the four calibration approaches was tested at six calibration events, and statistical comparison is shown by the Taylor Diagram in Figure 2-8. Confirming the model comparison using PED metrics in Figure 2-7, the MS-S and MS-A calibration approaches have better statistical proximity and pattern with the observation than ME-MS and OU methods. The Taylor diagrams indicate that MS-A approach has relatively more confined points towards the observation ('OBS' black dot and line) and consistently proves to be a better calibration approach than MS-S and other methods. The multi-event multi-site (ME-MS) optimization has more sparse points away from the 'OBS' proximity and produces an inconsistent performance over the calibration events.

In general, the calibrated model parameter sets generated by multi-site average objective function (MS-A) approach achieved improved model performance (NSE) and statistical measures (standard deviation, root mean squared error and correlation coefficient) during calibration step-1 and hence selected for calibration step-2.

Ten calibrated parameter sets generated by MS-A optimization approach were applied again to each of the ten calibration events and the results were extracted. Figure 2-9 demonstrates the normalized NSE performance metrics evaluated at five gauging stations.

The summation of the normalized NSE over each gauging sites and calibration events indicates that Model parameter Set 5 has the highest performance followed by Model Set 2 and 3. The result indicates that it is fairly reasonable to represent distributed semi-urban watersheds by qualifying model parameter sets generated from multiple even-based calibration process.

With the above results in mind, the DDS algorithm used by MS-A appears to converge to a better approximate true solution than the PADDS algorithm employed by MS-S approach. One of the key reasons is that MS-S result quality is summarized by precisely the objective function being optimized by MS-A. Another reason is likely that when solving the MS-S formulation, PADDS is spending substantial effort to approximate a Pareto-set in five dimensions and as such, PADDS is generating candidate solutions from much diversified parts of parameter space. In contrast, DDS is generating candidate solutions concentrated in the area of parameter space that leads to a good average objective function value.

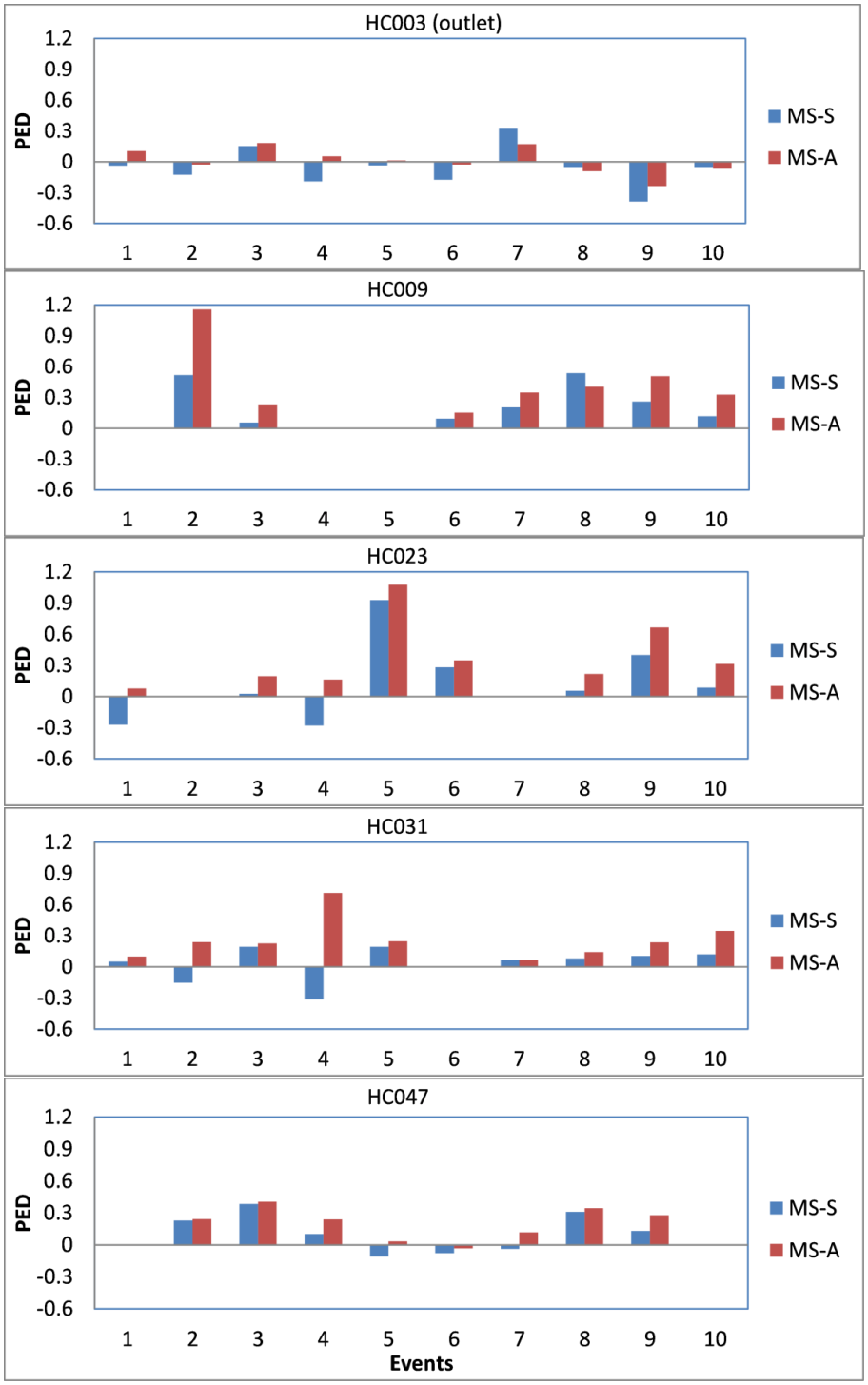

*Figure 2-7: Model Improvement (defined by Prediction Error Decrease in percentage (PED *100)) of Multi-site Simultaneous (MS-S) and Multi-site Average objective function (MS-A) calibration approaches when compared with Catchment Outlet (OU) approach at five gauging stations and ten calibration events.*

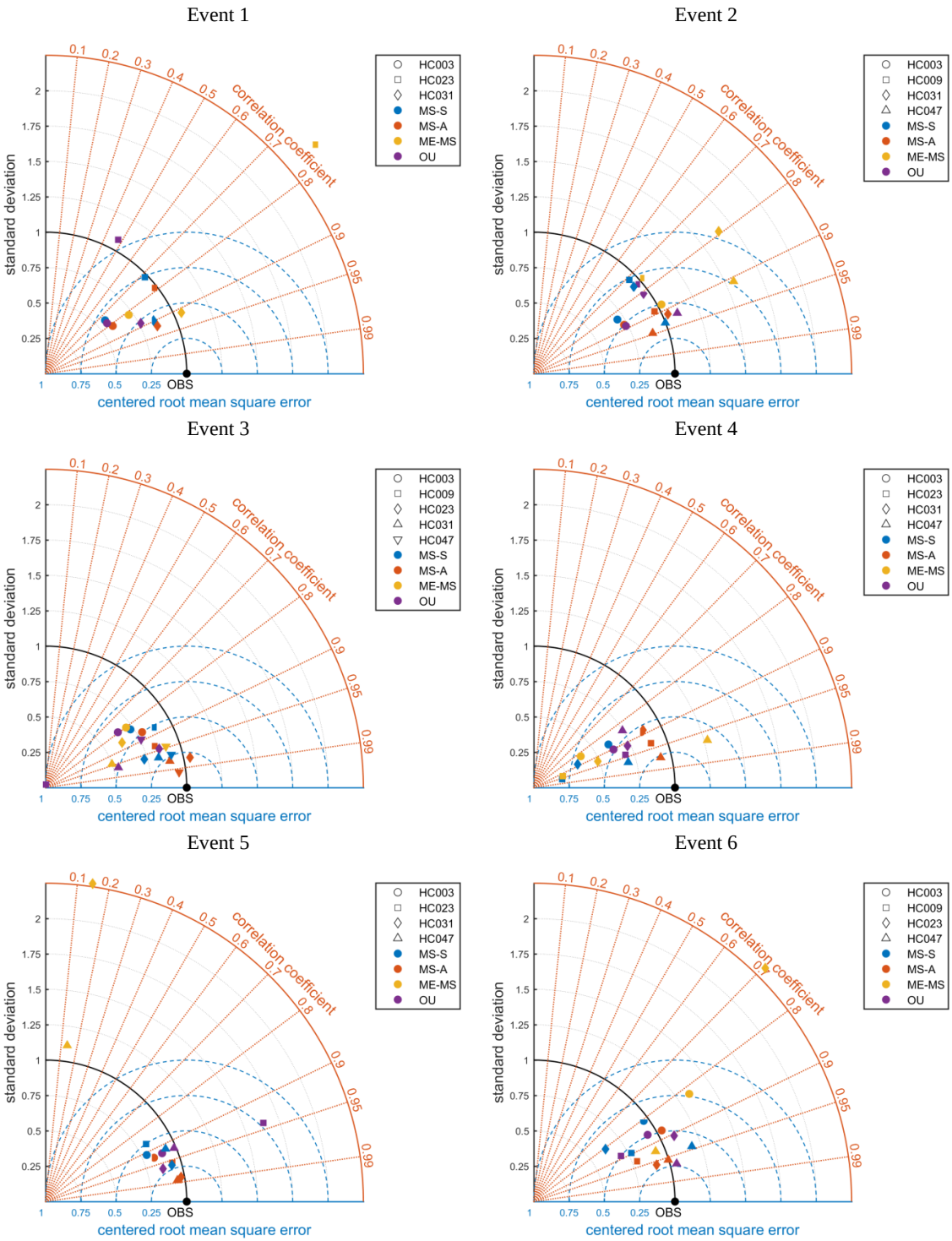

Figure 2-8: Comparison of Taylor diagrams showing an event-by-event statistical evaluation of simulated flows from four calibration approaches (MS-S, MS-A, ME-MS, & OU) evaluated at six calibration events. The Taylor Diagrams summarized three statistical performances at five gauging stations for each event. Different colors denote

respective calibration approaches while different shapes correspond to different stations (gauging sites). Perfect model sets would align themselves closer to the black arc as well as point 'OBS', which depict agreement with observations.

Calibration Events (Metric -> Normalized NSE), events 1–5

Model Set	1: HC003	1: HC023	1: HC031	2: HC003	2: HC009	2: HC031	2: HC047	3: HC003	3: HC009	3: HC023	3: HC031	3: HC047	4: HC003	4: HC023	4: HC031	4: HC047	5: HC003	5: HC023	5: HC031	5: HC047
1	1.0	1.0	1.0	1.0	0.9	0.9	0.0	1.0	0.7	0.8	0.4	0.5	0.7	0.1	0.6	0.0	1.0	1.0	0.9	0.0
2	0.7	0.8	0.9	1.0	1.0	1.0	1.0	0.9	0.7	1.0	0.2	0.4	0.4	0.8	0.3	0.9	0.9	0.9	1.0	0.9
3	0.7	0.9	0.3	0.7	0.9	0.2	0.7	0.9	1.0	1.0	1.0	1.0	0.9	0.6	1.0	0.9	0.5	1.0	0.1	1.0
4	0.4	0.2	0.1	0.4	0.2	0.0	0.9	0.6	0.1	0.5	1.0	0.7	1.0	1.0	1.0	1.0	0.5	0.3	0.0	0.8
5	0.5	1.0	0.8	0.9	0.9	1.0	0.9	0.7	0.9	0.9	0.3	0.8	0.8	0.3	0.7	1.0	0.8	1.0	1.0	0.9
6	0.4	1.0	0.8	0.9	0.5	1.0	0.4	0.7	0.4	0.7	0.0	0.6	0.0	0.0	0.1	0.5	0.6	1.0	0.9	0.1
7	0.5	0.0	0.8	0.9	0.9	0.9	0.9	0.7	0.4	0.0	0.4	0.4	0.5	0.8	0.4	0.9	0.9	0.0	1.0	0.8
8	0.3	1.0	0.7	0.8	0.9	1.0	1.0	0.7	0.0	0.9	0.0	0.0	0.1	0.2	0.0	0.9	0.2	1.0	0.9	0.8
9	0.0	0.7	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.9	0.8	0.7	1.0	0.8	0.8	0.6	0.5	0.8	0.5	0.6
10	0.3	1.0	0.7	0.7	0.9	1.0	0.8	0.7	0.5	0.9	0.1	0.5	0.1	0.3	0.0	0.9	0.0	1.0	0.9	0.4

Calibration Events (Metric -> Normalized NSE), events 6–10

Model Set	6: HC003	6: HC009	6: HC023	6: HC047	7: HC003	7: HC009	7: HC031	7: HC047	8: HC003	8: HC009	8: HC023	8: HC031	8: HC047	9: HC003	9: HC009	9: HC023	9: HC031	9: HC047	10: HC003	10: HC009	10: HC023	10: HC031	SUM	Rank
1	0.9	0.9	1.0	0.0	0.9	0.6	0.9	0.0	1.0	0.8	0.9	0.9	0.0	0.7	0.9	0.2	0.4	0.5	1.0	1.0	0.9	0.0	28	5
2	1.0	0.5	0.7	0.8	1.0	0.6	1.0	0.3	1.0	0.9	0.7	1.0	0.8	0.4	0.9	0.8	0.2	0.3	1.0	0.7	1.0	0.5	32	2
3	0.2	0.4	0.9	0.9	0.9	0.7	0.1	0.3	0.6	0.8	1.0	0.2	1.0	0.9	0.9	0.7	0.7	0.8	0.7	0.7	0.2	1.0	30	3
4	0.0	0.1	0.3	0.9	0.8	0.0	0.0	0.4	0.0	0.0	0.3	0.0	0.8	1.0	0.9	0.8	0.6	0.4	0.6	0.2	0.0	0.6	19	10
5	0.9	0.6	1.0	1.0	0.9	0.5	1.0	0.3	1.0	0.9	1.0	1.0	0.9	0.3	0.9	0.5	0.2	0.5	0.8	0.7	1.0	0.8	**33**	**1**
6	1.0	0.8	1.0	0.3	0.9	0.6	0.9	0.2	0.9	0.7	0.9	1.0	0.4	0.2	1.0	0.0	0.0	0.6	0.8	0.9	1.0	0.1	25	7
7	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	0.6	1.0	0.7	0.9	0.9	0.5	0.9	0.3	0.5	0.0	1.0	1.0	0.9	0.4	29	4
8	0.7	0.0	1.0	0.5	0.9	0.8	1.0	0.4	1.0	1.0	1.0	1.0	1.0	0.1	0.0	0.4	0.0	0.3	0.1	0.0	1.0	0.3	24	8
9	0.3	0.4	0.8	0.2	0.0	0.4	0.9	0.1	0.0	0.8	0.0	1.0	0.4	0.8	0.0	1.0	1.0	1.0	0.3	0.1	0.2	0.9	20	9
10	0.7	0.1	1.0	0.6	0.8	1.0	1.0	0.9	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.5	0.0	0.5	1.0	1.0	1.0	1.0	28	6

Normalized NSE
1 Good — 0 Bad

Figure 2-9: Performance ranking of 10 model parameter sets in ten calibration events. Normalized Nash-Sutcliffe Efficiency index (NSE) is used to score the performances at each gauging stations with sum over all sites and over all events displayed on the right side. Highest score corresponds to best performing model parameter set and vice versa. The heatmap shows the Normalized NSE values according to color palette displayed at the bottom side.

		Validation Events																					
		1					2					3					4					SUM	Rank
Metric ->		Taylor Skill Score																					
Model Sets	2	0.90	0.73	0.37	0.90	0.91	0.95	0.88	0.64	0.90	0.42	0.46	0.74	0.96	0.33	0.79	0.97	0.03	0.79	1.00	0.42	14	2
	3	1.00	0.87	0.35	0.56	0.99	0.83	0.58	0.64	0.48	0.37	0.44	0.71	0.67	0.36	0.87	0.88	0.00	0.96	0.09	0.22	12	3
	5	0.76	0.91	0.30	0.87	0.98	0.95	0.68	0.79	0.91	0.56	0.88	1.00	0.88	0.95	0.96	0.98	0.81	0.91	0.95	0.25	**16**	**1**
Stations		HC003	HC009	HC023	HC031	HC047	HC003	HC009	HC023	HC031	HC047	HC003	HC009	HC023	HC031	HC047	HC003	HC009	HC023	HC031	HC047		

Taylor Skill Score

1 Good — 0 Bad

Figure 2-10: Model validation of top three model sets of MS-A approach in different events. The Taylor Skill Scores are evaluated at each of the five gauging sites for four different events. Most skillful models would have a score of 1 and the least ones have a score of 0.

2.5.4. Validation

To verify the outcome of the above calibration processes, the top three model parameter sets (Model Set 5, 2 and 3) were evaluated at validation events because their performance from calibration step 2 are not significantly different (Summation of Normalized NSE: 30, 32 and 33 in Figure 2-9). The Taylor skill score was used to evaluate these SWMM5 model parameter sets at multiple sites and results are presented in Figure 2-10. Based on the scores, Model Set 5 appears to be more skillful than Model Set 2 and 3 as its score is close to 1 for majority of gauging stations and events. The summation of the Taylor Score over the gauges and evens (Sum=16) is the highest. Conversely Model Set 2 and 3 have lower scores because Taylor Skill Score penalizes models with little statistical pattern similarity and weak correlation with observations. In general, Taylor Skill Score is found to be a precise evaluation tool to select skillful SWMM5 model parameter sets that could represent the distributed watershed in space and time.

2.6. Conclusion

A proposed event-based calibration approach integrating multi-site and multi-objective optimizations was used to select representative SWMM5 model parameter sets in a distributed semi-urban watershed. We compared the performance of four calibration approaches in reproducing the desired spring flow responses at interior sites of Humber River Watershed. These are Multi-site simultaneous (MS-S), Multi-site average objective function (MS-A), Multi-event multi-site (ME-MS) and a benchmark At-catchment outlet (OU) calibration approaches. MS-S and OU approaches utilized PA-DDS optimization algorithm, whereas the others applied DDS algorithm.

A spatio-temporal variability of calibrated model parameter sets among different calibration events was initially assessed in anticipation of capturing the uncertainty of event-based parametrization. The results indicated that there is considerable uncertainty in calibrating highly impervious sub-catchments (>80% Imperviousness) and pervious areas with rapid recovery time (< 5.5 days of Drying Time). Another remark from the variability analysis is the presence of uncertainty in peak flow response by the model parameter sets. The uncertainty in reproducing peak flows by the majority of model parameters sets at multiple interior sites is a clear indication of a need for a robust calibration approaches in event-based distributed models.

The output from the proposed calibration approaches and steps demonstrated that multi-site average objective function (MS-A) and multi-site simultaneous (MS-S) calibration approaches showed superior performances against the Multi-event multi-site and benchmark calibration approaches. The desired flows at interior upstream sites were better

reproduced using MS-A and MS-S methods as compared to calibrating using the outlet (OU), a finding similar to Leta et al., (2017).

Most importantly, aggregating the objective functions across multiple sites into a single objective function (MS-A) outperformed the multi-site simultaneous (MS-S) approach. Individually calibrated model parameter sets from MS-A calibration approach shows significant improvement of NSE performance metrics when compared to MS-S at the majority of stations. This is also supported by Taylor diagrams, which demonstrated that the MS-A approach attained better statistical pattern and amplitude of observed hydrographs. Using MS-A method, ten parameter sets extracted from ten individual calibration events were cross-tested again at all events in the second calibration step. This step was able to identify the top three parameter sets out of ten potential model sets using their aggregated normalized NSE estimated at multiple sites. Model parameter sets 5 followed by 2 and 3 appear to outperform the rest of the model parameter sets. Validation was made at four different events to test the statistical performances using Taylor Skill Scores. And the result indicates that Model Parameter Set 5, which is calibrated using MS-A approach, is the most skillful and representative SWMM5 model parameter set in the study area.

In General, using the single objective DDS algorithm in MS-A approach to find the best average NSE of five gauging stations in the catchment area is found to be more efficient than using the multi-objective PA-DDS algorithm to find non-dominated Pareto-front of five NSE performances.

The study discovered that a combination of efficient optimization tools with a series of calibration approaches is important in finding candidate parameters sets and representing distributed catchments by event-based hydrological models. The study takes advantage of the DDS and PA-DDS algorithms to select non-dominated solutions and representative model parameter sets. Finally, the authors strongly believe that the methods and calibration approaches employed in this research can also be applied in other watersheds. An interesting result from the study is that averaging/aggregating objective functions during calibration provide better simulation output, which can be applied for any cases.

2.7. Acknowledgments

This work was supported by the Natural Science and Engineering Research Council (NSERC) Canadian FloodNet (Grant number: NETGP 451456).

2.8. References

Ackerman, D., Schiff, K.C., Weisberg, S.B., 2005. Evaluating HSPF in an arid, urbanized watershed. J. Am. Water Resour. Assoc. 41, 477–486. https://doi.org/10.1111/j.1752-1688.2005.tb03750.x

Ajmal, M., Waseem, M., Ahn, J.-H., Kim, T.-W., 2016. Runoff estimation using the NRCS slope-adjusted curve number in mountainous watersheds. J. Irrig. Drain. Eng. 142, 04016002. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000998

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assesment Part I: Model development. JAWRA J. Am. Water Resour. Assoc. 34, 73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x

Asadzadeh, M., Razavi, S., Tolson, B.A., Fay, D., 2014. Pre-emption strategies for efficient multi-objective optimization: Application to the development of Lake Superior regulation plan. Environ. Model. Softw. 54, 128–141. https://doi.org/10.1016/j.envsoft.2014.01.005

Asadzadeh, M., Tolson, B., 2013. Pareto archived dynamically dimensioned search with hypervolume-based selection for multi- objective optimization. Eng. Optim. ISSN 45, 1489–1509. https://doi.org/10.1080/0305215X.2012.748046

Asadzadeh, M., Tolson, B., 2012. Hybrid Pareto archived dynamically dimensioned search for multi-objective combinatorial optimization: application to water distribution network design. J. Hydroinformatics 14, 192–205. https://doi.org/10.2166/hydro.2011.098

Asadzadeh, M., Tolson, B.A., 2009. A New multi-objective algorithm , pareto archived DDS, in: 11th annual conference companion on genetic and evolutionary computation conference. Late Breaking Papers, Montreal, Quebec, Canada, pp. 1963–1966.

Barco, J., Wong, K.M., Stenstrom, M.K., Asce, F., 2008. Automatic calibration of the U.S. EPA SWMM model for a large urban catchment. J. Hydraul. Eng. 134, 466–474. https://doi.org/10.1061/ASCE0733-94292008134:4466

Beven, K., 1989. Changing ideas in hydrology-the case of physically-based models. J. Hydrol. Elsevier Sci. Publ. B.V 105, 157–172.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249, 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Bolado-Lavin, R., Castaings, W., Tarantola, S., 2009. Contribution to the sample mean plot for graphical and numerical sensitivity analysis. Reliab. Eng. Syst. Saf. 94, 1041–1049. https://doi.org/10.1016/j.ress.2008.11.012

Brocca, L., Melone, F., Moramarco, T., 2011. Distributed rainfall-runoff modelling for flood frequency estimation and flood forecasting. Hydrol. Process. 25, 2801–2813. https://doi.org/10.1002/hyp.8042

Cao, W., Bowden, W.B., Davie, T., Fenemor, A., 2006. Multi-variable and multi-site calibration and validation of SWAT in a large mountainous catchment with high spatial variability. Hydrol. Process. 20, 1057–1073. https://doi.org/10.1002/hyp.5933

CIVICA & TRCA, 2015. Humber River Hydrology Update, Retrieved from https://trca.ca/wp-content/uploads/2016/07/Humber-Hydrology-Update-Final Report-v19.1.pdf.

Coulibaly, P., 2003. Impact of meteorological predictions on real-time spring flow forecasting. Hydrol. Process. 17, 3791–3801. https://doi.org/10.1002/hyp.5168

Cuo, L., Lettenmaier, D.P., Mattheussen, B. V, Storck, P., Wiley, M., 2008. Hydrologic prediction for urban watersheds with the Distributed Hydrology–Soil–Vegetation Model. Hydrol. Process. 22, 4205–4213. https://doi.org/10.1002/hyp.7023

Deshmukh, D.S., Chaube, U.C., Ekube Hailu, A., Aberra Gudeta, D., Tegene Kassa, M., 2013. Estimation and comparision of curve numbers based on dynamic land use land cover change, observed rainfall-runoff data and land slope. J. Hydrol. 492, 89–101. https://doi.org/10.1016/j.jhydrol.2013.04.001

El Hassan, A.A., Sharif, H.O., Jackson, T., Chintalapudi, S., 2013. Performance of a conceptual and physically based model in simulating the response of a semi-urbanized watershed in San Antonio, Texas. Hydrol. Process. 27, 3394–3408. https://doi.org/10.1002/hyp.9443

Engeland, K., Braud, I., Gottschalk, L., Leblois, E., 2006. Multi-objective regional modelling. J. Hydrol. 327, 339–351. https://doi.org/10.1016/j.jhydrol.2005.11.022

Fletcher, T.D., Andrieu, H., Hamel, P., 2013. Understanding, management and modelling of urban hydrology and its consequences for receiving waters: A state of the art. Adv. Water Resour. 51, 261–279. https://doi.org/10.1016/j.advwatres.2012.09.001

Furl, C., Sharif, H.O., El Hassan, A., Mazari, N., Burtch, D., Mullendore, G.L., Furl, C., Sharif, H.O., Hassan, A. El, Mazari, N., Burtch, D., Mullendore, G.L., 2015. Hydrometeorological analysis of tropical storm hermine and central texas flash flooding, September 2010. J. Hydrometeorol. 16, 2311–2327. https://doi.org/10.1175/JHM-D-14-0146.1

Gironás, J., Asce, A.M., Niemann, J.D., Asce, M., Roesner, L.A., Asce, F., Rodriguez, F., Andrieu, H., 2010. Evaluation of methods for representing urban terrain in storm-water modeling. J. Hydrol. Eng. 15, 1–14. https://doi.org/10.1061/ASCÊHE.1943-5584.0000142

Haghnegahdar, A., Tolson, B.A., Davison, B., Seglenieks, F.R., Klyszejko, E., Soulis, E.D., Fortin, V., Matott, L.S., 2014. Calibrating environment canada's MESH modelling system over the great lakes basin. Atmosphere-Ocean 52, 281–293. https://doi.org/10.1080/07055900.2014.939131

Hay, L.E., Leavesley, G.H., Clark, M.P., Markstrom, S.L., Viger, R.J., Umemoto, M., 2006. Step wise multiple objective calibration of a hydrologic model for a snowmelt dominated basin. J. Am. Water Resour. Assoc. 42, 877–890.

Herrera, M., I. Heathcote, W.J. and A.B., 2006. Multi-objective calibration of SWMM for improved simulation of the hydrologic regime. J. Water Manag. Model. R225-15. https://doi.org/10.14796/JWMM.R225-15

Huber, W.C., Dickinson, R.E., 1988. Storm water management model, version 4: user's manual. Athens, GA. https://doi.org/EPA/600/3-88/001a

Irvine, K.N., Loganathan, B.G., Pratt, E.J., Sikka, H.C., 1993. Calibration of PCSWMM to estimate metals, PCBs and HCB in CSOs from an industrial sewershed. J. Water Manag. Model. R175. https://doi.org/I 0.14796/JWMM.RI75-10.

Jacobson, C.R., 2011. Identification and quantification of the hydrological impacts of imperviousness in urban catchments: A review. J. Environ. Manage. 92, 1438–1448. https://doi.org/10.1016/j.jenvman.2011.01.018

James, W., 2005. Rules for Responsible Modeling-4th Edition, 4th ed. CHI (Computational Hydraulics International), Guelph, Ontario.

James, W., Lewis A., Rossman, W., Robert C., James., 2011. User's Guide to SWMM5, 13th Edn. CHI, Guelph. ISBN: 978-0-9808853-5-4.

Kärnä, T., Baptista, A.M., 2016. Evaluation of a long-term hindcast simulation for the Columbia River estuary. Ocean Model. 99, 1–14. https://doi.org/10.1016/j.ocemod.2015.12.007

Kennedy, J.R., Goodrich, D.C., Asce, M., Unkrich, C.L., 2013. Using the KINEROS2 modeling framework to evaluate the increase in storm runoff from residential development in a semiarid environment. J. Hydrol. Eng. © 18, 698–706. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000655

Khu, S.-T., Di Pierro, F., Savi, D., Djordjevi, S., Walters, G.A., 2006. Incorporating spatial and temporal information for urban drainage model calibration: An approach using preference ordering genetic algorithm. Adv. Water Resour. 29 29, 1168–1181. https://doi.org/10.1016/j.advwatres.2005.09.009

Khu, S.-T., Madsen, H., di Pierro, F., 2008. Incorporating multiple observations for distributed hydrologic model calibration: An approach using a multi-objective evolutionary algorithm and clustering. Adv. Water Resour. 31, 1387–1398. https://doi.org/10.1016/j.advwatres.2008.07.011

Kraußе, Cullmann, J., Saile, P., Schmitz, G.H., 2012. Robust multi-objective calibration strategies – possibilities for improving flood forecasting. Hydrol. Earth Syst. Sci 16, 3579–3606. https://doi.org/10.5194/hess-16-3579-2012

Krebs, G., Kokkonen, T., Valtanen, M., Koivusalo, H., Setaïa, H., 2013. A high resolution application of a stormwater management model (SWMM) using genetic parameter optimization. Urban Water J. 10, 394–410. https://doi.org/10.1080/1573062X.2012.739631

Leta, O.T., van Griensven, A., Bauwens, W., 2017. Effect of single and multisite calibration

techniques on the parameter estimation, performance, and output of a SWAT model of a spatially heterogeneous catchment. J. Hydrol. Eng. 22, 05016036. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001471

Liong, S.-Y., Chan, W.T., ShreeRam, J., 1995. Peak-flow forecasting with genetic algorithm and SWMM. J. Hydraul. Eng. 121, 613–617. https://doi.org/10.1061/(ASCE)0733-9429(1995)121:8(613)

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv. Water Resour. 26, 205–216. https://doi.org/10.1016/S0309-1708(02)00092-1

Madsen, H., Wilson, G., Ammentorp, H.C., 2002. Comparison of different automated strategies for calibration of rainfall-runoff models. J. Hydrol. 261, 48–59. https://doi.org/10.1016/S0022-1694(01)00619-9

Matott, L.S., 2005. OSTRICH : An optimization software tool ; documentation and user ' s guide, Version 1.6. Buffalo.

Mediero, L., Garrote, L., Martín-Carrasco, F.J., 2011. Probabilistic calibration of a distributed hydrological model for flood forecasting. Hydrol. Sci. J. 56, 1129–1149. https://doi.org/10.1080/02626667.2011.610322

Miller, J.D., Kim, H., Kjeldsen, T.R., Packman, J., Grebby, S., Dearden, R., 2014. Assessing the impact of urbanization on storm runoff in a peri-urban catchment using historical change in impervious cover. J. Hydrol. 515, 59–70. https://doi.org/10.1016/j.jhydrol.2014.04.011

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

Niemi, T.J., Warsta, L., Taka, M., Hickman, B., Pulkkinen, S., Krebs, G., Moisseev, D.N., Koivusalo, H., Kokkonen, T., 2017. Applicability of open rainfall data to event-scale urban rainfall-runoff modelling. J. Hydrol. 547, 143–155. https://doi.org/10.1016/j.jhydrol.2017.01.056

NRCS, 2007. Chapter 7 hydrologic soil groups, in: national engineering handbook, Part 630 Hydrology. United States Department of Agriculture.

NVCA, 2006. Innisfil creek subwatershed plan: Appendix d - hydrologic modeling. Utopia, ON.

Ogden, F.L., Sharif, H.O., Senarath, S.U.S., Smith, J.A., Baeck, M.L., Richardson, J.R., 2000. Hydrologic analysis of the Fort Collins, Colorado, flash flood of 1997. J. Hydrol. 228, 82–100. https://doi.org/10.1016/S0022-1694(00)00146-3

Ozdemir, A., Leloglu, U.M., Abbaspour, K.C., 2017. Hierarchical approach to hydrological model calibration. Environ. Earth Sci. 76, 318. https://doi.org/10.1007/s12665-017-6560-6

Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., Verseghy, D., Soulis, E.D., Caldwell, R., Evora, N., Pellerin, P., 2007. Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale. Hydrol. Earth Syst. Sci. 11, 1279–1294. https://doi.org/10.5194/hessd-3-2473-2006

Plischke, E., 2012. An adaptive correlation ratio method using the cumulative sum of the reordered output. Reliab. Eng. Syst. Saf. 107, 149–156. https://doi.org/10.1016/j.ress.2011.12.007

Pokhrel, P., Gupta, H. V., 2010. On the use of spatial regularization strategies to improve calibration of distributed watershed models. Water Resour. Res. 46, 1–17. https://doi.org/10.1029/2009WR008066

Randall, M., James, R., James, W., Finney, K., Heralall, M., 2014. PCSWMM real time flood forecasting – Toronto, Canada. PCSWMM Real Time Flood Forecast.

Robert, W.C., Robert James, W.C., Heralall, M., James, W., 2008. Integrated web-based automated radar acquisition, processing and real-time modeling for flow/flood forecasting.

Rossman, L., Huber., W., 2015. Storm water management model reference manual: volume I-Hydrology (Revised). Washington, DC. https://doi.org/EPA/600/R-15/162A

Rossman, L.A., 2010. Storm water management model user's manual version 5.0. Cincinnati, OH. https://doi.org/EPA/600/R-05/040

Sharif, H.O., Chintalapudi, S., Hassan, A.A., Xie, H., Zeitler, J., 2013. Physically based hydrological modeling of the 2002 floods in San Antonio, Texas. J. Hydrol. Eng. 18, 228–236. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000475

Sharif, H.O., Sparks, L., Hassan, A.A., Zeitler, J., Xie, H., 2010. Application of a distributed hydrologic model to the November 17, 2004, flood of Bull Creek Watershed, Austin, Texas. J. Hydrol. Eng. 15, 651–657. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000228

Shinma, T.A., Reis, L.F.R., 2014. Incorporating multi-event and multi-site data in the calibration of SWMM. Procedia Eng. 70, 75–84. https://doi.org/10.1016/j.proeng.2014.02.010

Singh, S., Bárdossy, A., 2015. Hydrological model calibration by sequential replacement of weak parameter sets using depth function. Hydrology 2, 69–92. https://doi.org/10.3390/hydrology2020069

Spear, R.C., Hornberger, G.M., 1980. Eutrophication in peel inlet-II. Identification of critical uncertainties via generalized sensitivity analysis. Water Res. 14, 43–49. https://doi.org/10.1016/0043-1354(80)90040-8

Sun, N., Hall, M., Hong, B., Zhang, L., 2014. Impact of SWMM catchment discretization: case study in Syracuse, New York. J. Hydrol. Eng. 19, 223–234.

https://doi.org/10.1061/(ASCE)HE.1943-5584.0000777

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. 106, 7183–7192. https://doi.org/10.1029/2000JD900719

Tolson, B.A., Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resour. Res. 43, 1–16. https://doi.org/10.1029/2005WR004723

Tramblay, Y., Bouaicha, R., Brocca, L., Dorigo, W., Bouvier, C., Camici, S., Servat, E., 2012. Estimation of antecedent wetness conditions for flood modelling in northern Morocco. Hydrol. Earth Syst. Sci. 16, 4375–4386. https://doi.org/10.5194/hess-16-4375-2012

TRCA, 2013. Humber river watershed report card 2013 [WWW Document]. URL https://trca.ca/wp-content/uploads/2016/04/2173_WatershedReportCards_Humber_rev15_forWeb.pdf (accessed 4.12.17).

TRCA, 2008. Humber River Watershed Scenario Modelling and Analysis Report. Downsview, ON.

TRCA, AMEC, 2012. Hydrologic impacts of future development on flood flows and mitigation requirements in the humber river watershed-draft report submited by AMEC Environment & Infrastructure.

Wang, S., Zhang, Z., Sun, G., Strauss, P., Guo, J., Tang, Y., Yao, A., 2012. Multi-site calibration, validation, and sensitivity analysis of the MIKE SHE Model for a large watershed in northern China. Hydrol. Earth Syst. Sci 16, 4621–4632. https://doi.org/10.5194/hess-16-4621-2012

Wi, S., Yang, Y.C.E., Steinschneider, S., Khalil, A., Brown, C.M., 2015. Calibration approaches for distributed hydrologic models in poorly gaged basins: implication for streamflow projections under climate change. Hydrol. Earth Syst. Sci 19, 857–876. https://doi.org/10.5194/hess-19-857-2015

WMO, 2011. Manual on Flood Forecasting and Warning: WMO-No. 1072. Geneva.

Xia, Y., Pitman, A.J., Gupta, H.V., Leplastrier, M., Henderson-Sellers, A., 2002. Calibrating a land surface model of varying complexity using multicriteria methods and the Cabauw dataset. J. Hydrometeorol. 3, 181–194. https://doi.org/10.1175/1525-7541(2002)003<0181:CALSMO>2.0.CO;2

Xue, X., Zhang, K., Hong, Y., Gourley, J.J., Kellogg, W., Mcpherson, R.A., Wan, Z., Austin, B.N., 2016. New multisite cascading calibration approach for hydrological models: case study in the Red river basin using the VIC model. J. Hydrol. Eng. 21. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001282

Zaghloul, N.A., Abu Kiefa, M.A., 2001. Neural network solution of inverse parameters used in the sensitivity-calibration analyses of the SWMM model simulations. Adv. Eng. Softw. 32, 587–595. https://doi.org/10.1016/S0965-9978(00)00072-7

Zhang, G., Hamlett, J.M., Reed, P., Tang, Y., 2013. Multi-objective optimization of low impact development designs in an urbanizing watershed. Open J. Optim. 2, 95–108. https://doi.org/10.4236/ojop.2013.24013

Zhang, X., Srinivasan, R., Van Liew, M., 2010. On the use of multi-algorithm, genetically adaptive multi-objective method for multi-site calibration of the SWAT model. Hydrol. Process. 24, 955–969. https://doi.org/10.1002/hyp.7528

Zhang, X., Zhang, Xu, Hu, S., Liu, T., Li, G., 2013. Runoff and sediment modeling in a peri-urban artificial landscape: Case study of Olympic Forest Park in Beijing. J. Hydrol. 485, 126–138. https://doi.org/10.1016/j.jhydrol.2012.01.038

Zhou, F., Xu, Y., Chen, Y., Xu, C.-Y., Gao, Y., Du, J., 2013. Hydrological response to urbanization at different spatio-temporal scales simulated by coupling of CLUE-S and the SWAT model in the Yangtze River Delta region. J. Hydrol. 485, 113–125. https://doi.org/10.1016/j.jhydrol.2012.12.040

Chapter 3. Identification of hydrological models for enhanced ensemble reservoir inflow forecasting in a large complex Prairie watershed

Summary of Paper 2: Awol, F.S., Coulibaly, P., Tsanis, I., Unduche, F. (2019). Identification of hydrological models for enhanced ensemble reservoir inflow forecasting in a large complex Prairie watershed. *Water*, 11(11), 2201.

This research compares lumped, semi-distributed, and land-surface based models with raw and bias-corrected ensemble weather forecast inputs to identify the best model for reservoir inflow forecast in a large complex watershed.

Key findings of this research include:

- Bias-correcting precipitation forecasts for a training period of at least two years before the forecast time produced skillful ensemble hydrological forecasts.
- The lumped models forced with bias-corrected ensemble forecast inputs provided better forecast performance than distributed or land-surface models, up to a week ahead outlook.
- The benchmark distributed model was as reliable as the lumped models only up to 3 days forecast.
- Overall, the SACSMA with SNOW 17 model emerged as the best model to provide accurate and reliable medium-range forecasts in complex watersheds.

3.1. Abstract

Accurate and reliable flow forecasting in complex Canadian prairie watersheds has been one of the major challenges faced by hydrologists. In an attempt to improve the accuracy and reliability of a reservoir inflow forecast, this study investigates structurally different hydrological models along with ensemble precipitation forecasts to identify the most skillful and reliable model. The key goal is to assess whether short- and medium-range ensemble flood forecasting in large complex basins can be accurately achieved by simple conceptual lumped models (e.g., SACSMA with SNOW17 and MACHBV with SNOW17) or it requires a medium level distributed model (e.g., WATFLOOD) or an advanced macroscale land-surface based model (VIC coupled with routing module (RVIC)). Eleven (11)-member precipitation forecasts from second-generation Global Ensemble Forecast System reforecast (GEFSv2) were used as inputs. Each of the ensemble members was bias-corrected by Empirical Quantile Mapping method using the Canadian Precipitation Analysis (CaPA) as a training/verification dataset. Forecast evaluation is performed for 1-day up to 8-days forecast lead times in a 6-month hindcast period. Results indicate that bias-correcting precipitation forecasts using verifying datasets (such as CaPA) for a training period of at least two years before the forecast time, produces skillful ensemble hydrological forecasts. A comparison of models in forecast mode shows that the two lumped models (SACSMA and MACHBV) can provide better overall forecast performance than the benchmark WATFLOOD and the macroscale Variable Infiltration Capacity (VIC) model. However, for shorter lead-times, particularly up to day 3, the benchmark distributed model provides competitive reliability, as compared to the lumped models. In general, the

SACSMA model provided better forecast quality, reliability and differentiation skill than other considered models at all lead times.

3.2. Introduction

Prairie watersheds are characterized by several small depressions, potholes and wetlands, and poorly connected drainage systems that may or may not contribute to the main river system (Fang et al., 2007). They are often featured by their long winter periods, high spring snowmelt contribution to annual runoff, deep-frozen soils and rapid infiltration, intense rainfall in spring and early summer, lower soil moisture, and evaporation from summer to fall (Fang et al., 2007). Relevant methodologies were proposed to assess several aspects of the hydrological cycle such as snowpack, spring melt, soil moisture, rainfall frequency, and evaporation, in the Canadian Prairie regions (Armstrong et al., 2008; Fang et al., 2010; Hayashi and Van Der Kamp, 2000; Shook et al., 2015). The effect of climate, land use, and ecosystem change on the hydrological processes of cold and wetland regions were also studied (Eum et al., 2017; Hedstrom et al., 2001; Pattison-Williams et al., 2018). Even though some efforts were made to formulate the realistic representation of wetland processes in hydrological models (Evenson et al., 2016; Gray and Landine, 1988; Mekonnen et al., 2014; Pomeroy et al., 2007; Shook et al., 2013), challenges of hydrological forecasting and flood predictions in such complex watersheds remain at large.

Several important works have already been performed for enhancing flood prediction in several watersheds: for example, using single or multiple hydrological models (Ajami et al., 2006; Antonetti et al., 2018; Brochero et al., 2011; Seiller et al., 2017, 2012; Thiboult et al., 2016; Velázquez et al., 2011, 2010; Viney et al., 2009), or feeding ensemble numerical weather products to models (Alfieri et al., 2014; Calvetti and Pereira Filho, 2014; Fan et al., 2014b; Liechti et al., 2013; Pietroniro et al., 2007; Thiemig et al., 2010; Zsótér

et al., 2016). Velázquez et al., (2011), for example, analyzed 16 lumped hydrological models with 50-member ensemble weather inputs. They detected that the multi-model approach of a grand member ensemble provided more forecast skill and reliability than either a single model with meteorological ensembles or multiple models with the deterministic forecast at all lead times. Pietroniro et al., (2007), assessed the benefit of using Environment Canada's MESH (Modelisation Environmentale Communautaire-MEC Surface and Hydrology) model in the Great Lakes catchment with inputs from 16-member ensemble forecast variables supplied by Meteorological Service of Canada (MSC). Fan et al., (2014a), suggested the use of local or regional ensemble forecasts instead of low-resolution global ensemble inputs and data assimilation methods. In their work, they applied MGB-IHB distributed model with bias-corrected second-generation Global Ensemble Forecast System (GEFS v2) reforecast inputs and suggested that the improvements made could address the lack of spread in reservoir inflow forecasts especially in early lead times. Using the same hydrological model, Fan et al., (2015), evaluated the importance of three sets of ensemble QPFs from the TIGGE (THORPEX Interactive Grand Global Ensemble) database in larger basins that have major reservoirs and hydroelectric plants. Their verification methods confirmed that the performance of hydrological forecasts depends on the quality of each ensemble precipitation products, but they also highlighted the improved reliability and robustness of ensemble river flows obtained from the combined super ensemble inputs. Abaza et al., (2013), compared currently available Canadian meteorological forecasts and concluded that streamflow forecasting fed by Regional ensemble prediction systems (EPS) provided higher reliability

than the Global EPS followed by their deterministic counterparts, as also supported by Fan et al., (2014a). The use of multiple models with the global, regional and local ensemble and deterministic inputs has also been implemented in several operational flood forecasting centers across the globe (Achleitner et al., 2012; De Roo et al., 2003; Demargne et al., 2014; Jasper et al., 2002; Maxey et al., 2012; Florian Pappenberger et al., 2008; Unduche et al., 2018).

The main challenge in getting accurate and reliable short- and medium-range flood forecasts in large complex watersheds arises from the type of hydrological models, and the quality of weather forecast inputs applied. The choice of the models to be implemented for flood and streamflow forecasting depends on the intended purpose, the type of forecast inputs, and the complexity and scale of the study area (Hrachowitz and Clark, 2017). Given the complexity of a prairie watershed in defining wetland and non-wetland physical processes and its representation by model structures for a specific application of real-time flood forecasting, it is essential to identify the candidate hydrological model(s) from multiple diverse potential models. Once the hydrological model or group of models are identified, the skill and reliability of hydrological forecasts can be enhanced by feeding qualitative ensemble weather forecast into the models.

The limitations of previous works and the scientific challenges are that

1) only a few studies were conducted on large and complex Prairie watersheds,

2) only a lumped model or a distributed model was used independently, for hydrological forecasting study. Alternatively, in some cases, the multi-models were only a collection of lumped conceptual models,

3) identification of best hydrological model was usually based on historical meteorological or in some cases, deterministic weather forecast inputs. Evaluation and comparison of models based on raw and bias-corrected ensemble precipitation forecasts were not studied. As such, the objective of this research is designed to address these limitations and identify the best hydrological model from diverse multi-models for short- and medium-range flood forecasting in a Canadian Prairie watershed. In this study, four structurally varied hydrological models were set up in order to simulate and forecast inflows to the Shelmouth Reservoir which is located in Upper Assiniboine River Basin. A mixture of two lumped, one distributed and one macroscale land surface models were used in this research. In forecast mode, bias-corrected precipitation from second-generation Global Ensemble Forecast System (GEFS v2) reforecasts was fed into the four models in order to evaluate the reliability, skill, and overall forecast performance of the ensemble reservoir inflows.

3.3. Materials

3.3.1. Study area

The Canadian Prairies are mainly located in Saskatchewan, Manitoba, and Alberta Provinces. The research is conducted in one of the main Canadian Prairie watersheds, the Upper Assiniboine River Basin upstream of the Shelmouth Reservoir, also called Lake of the Prairie (Figure 3-1). The catchment area contributing to the reservoir inflow is approximately 18,000 km^2. While much of the basin is located in Saskatchewan, the Shelmouth Reservoir itself is located in the Province of Manitoba. Inflow into the reservoir is generated from three major upstream tributaries: the Whitesand River, the Shell Rivers

and the main stream of the Assiniboine River. This Prairie watershed which is known to have a complex hydrology is characterized by abundant potholes and wetlands, poorly interconnected streams and non-contributing areas, long and cold winter periods, deep-frozen soils and rapid infiltration, high spring snowmelt contribution to annual runoff, intense rainfall in spring and early summer, lower soil moisture and evaporation from summer to fall (Fang et al., 2007; Unduche et al., 2018). The basin's topography ranges from 250 m a.s.l. at its lowest point to 820 m a.s.l at its highest point, and its annual precipitation is approximately 460 mm (Shrestha et al., 2012). The land cover of the basin is mostly dominated by cropland, which contributes about 55-58% of the land cover (Shrestha et al., 2012).

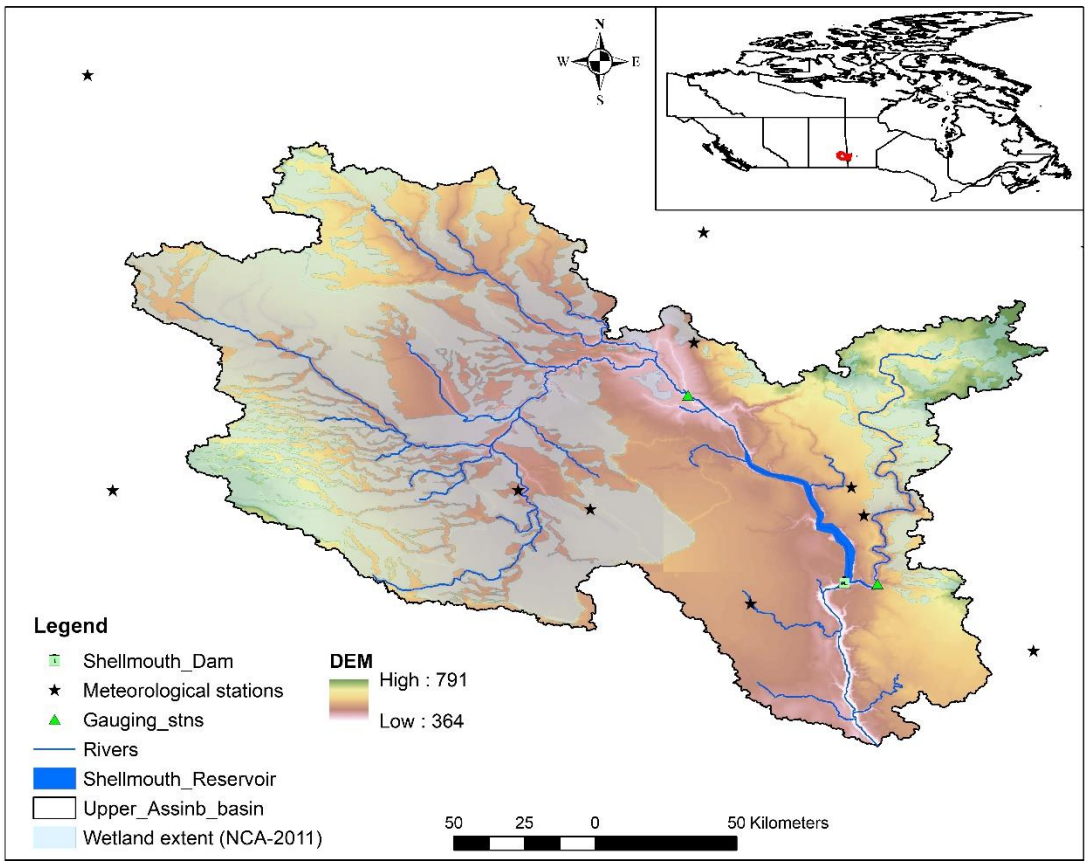

Figure 3-1: Study area of the Upper Assiniboine River Basin

3.3.2. Data

3.3.2.1. Ensemble weather forecast

An 11-member ensemble data from second-generation Global Ensemble Forecast System (GEFS v2) reforecast (Hamill et al., 2013) supplied by National Centers for Environmental Prediction (NCEP), hereafter called "GEFSv2" was used as an input to models. The GEFSv2 issues forecast once a day in 3 hourly time step up to 8 days lead time with 50 km spatial resolution and the next eight days with a lower spatial resolution. For this research, daily total precipitation forecasts from Jan 2014 to Dec 2017 were used for the input datasets; the first two years used for bias correcting the last two years. Only precipitation forecasts were used as forcing data, while other variables were taken from observation because the accuracy of flood prediction is highly impacted by precipitation forecasts than any other variables (Zsótér et al., 2016).

3.3.2.2. Observed Data

Average daily temperature and precipitation data were obtained from Environment Canada for the eleven weather gauging stations that are distributed across the catchment (Figure 3-1). These data were used as inputs to the hydrological models. The output from the hydrological models which is regarded as the simulated reservoir inflow was compared with calculated (observed) reservoir inflow in the calibration process. Detail information on the reservoir inflow estimation is provided in Section 3.3.2.3, whereas calibration and validation will be discussed in Section 3.4.2. In addition to the gauge data, precipitation data were also collected from The Canadian Precipitation Analysis (CaPA). The CaPA is

developed by statistical interpolation of a background field from short-range precipitation forecasts, and observation from radar and ground-based rainfall measurements (Mahfouf et al., 2007). The spatial and temporal resolution of CaPA is 15 km and 6 hours respectively. For this study, CaPA precipitation data is used for bias correcting global ensemble forecasts, which will be further discussed in Section 3.4.3.

3.3.2.3. Reservoir inflow

The study area is the watershed upstream of the Shelmouth Reservoir. There is no flow gauge (actual streamflow measurement) at the outlet of the watershed. At the mouth of the reservoir or the Dam section, the outflow is regulated by structural mechanisms such as releasing water through the conduits (using gates) and spillways. These releases are controlled and measured daily. Therefore, the outflow from the reservoir is a regulated outflow measured at the conduits and spillway, and due to this reason, it cannot be directly used for calibration. Instead, the reservoir inflow is implicitly considered as the outflow from the entire watershed and is used for calibrating the hydrological models. The inflow is, in this case, a collection of water from major and minor tributaries that goes into the reservoir. The estimated inflow is considered as an "unregulated" discharge observation measuring collectively the river flows coming from the tributaries.

The inflow into Shemouth Reservoir is calculated based on a simple water balance equation. Given records of daily reservoir levels, the elevation-area-storage curve of the reservoir, and the summation of outflows measured at the spillway and conduit, the water balance can be formulated by the equation 3-1. Here, losses (such as evaporation and

infiltration) within a day are assumed to be negligible, and lateral inflows are included in 'Inflow' variable.

$$Inflow - Outflow = \frac{dS}{dt} \quad (3\text{-}1)$$

Where $\frac{dS}{dt}$ is the change in storage in one-day time difference. The change in storage is obtained from the elevation-storage curve by looking at the daily average reservoir levels between the first and the second day. The reservoir inflow is calculated daily for practical application at Manitoba Hydrological Forecasting Center.

As described above, the reservoir inflow is regarded as a streamflow measurement of all the tributary rivers and streams supplying water to the reservoir. Since the inflow is not an actual flow measurement of the suppling rivers, it is prone to some degree of errors. However, the calculated reservoir inflow is believed to be the best possible method of measuring the "unregulated" watershed outflow. Also, there is uncertainty arising from the calculation method. We used a simple water balance equation to calculate the daily inflow, only accounting for the daily change in storage and the daily measured regulated reservoir outflow. The daily losses (e.g., evaporation and infiltration) are assumed to negligible. Such an assumption might contain some uncertainties. However, the uncertainty for daily water balance is not believed to be considerable, for example, comparing with monthly water balance where such losses cannot be ignored.

3.4. Method

Figure 3-2 shows the methodology adopted in this research. Ensemble weather forecasts products from GEFSv2 were collected. Each ensemble member of precipitation forecasts

was bias-corrected by the Empirical Quantile Mapping method using CaPA as a verifying data (Section 3.4.3). Four structurally various hydrological models were applied in the watershed including the benchmark model. Using raw and bias-corrected GEFSv2 ensemble inputs, the models' forecasting performances were evaluated and compared in hindcast period.

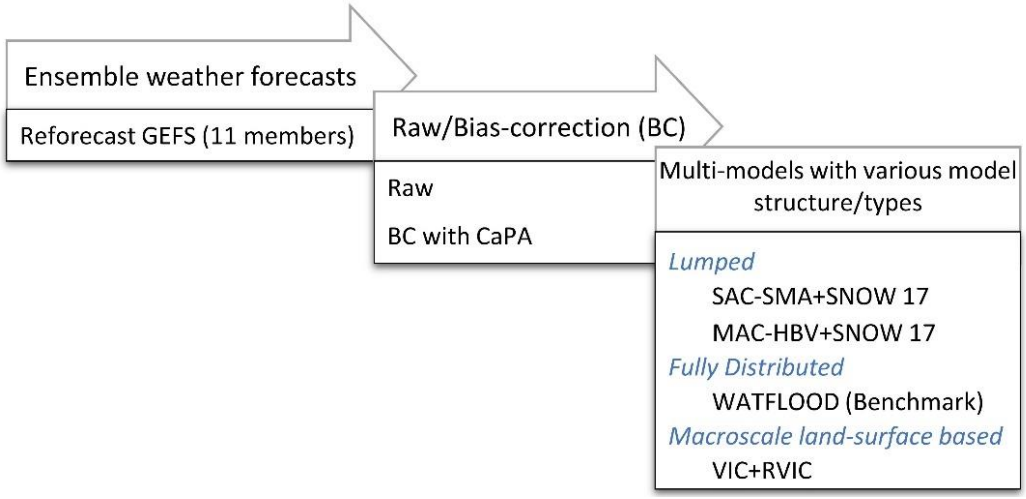

Figure 3-2: Illustration of the methodology adopted

3.4.1. Hydrological Models

Four different hydrological models with diverse model structures were applied in the study area to meet the research objectives described in Section 3.2. The models applied herein are a combination of two lumped models, one distributed and one macroscale land-surface based hydrological model. The lumped models are the Sacramento Soil Moisture Accounting (SAC-SMA) model coupled with SNOW17 (Anderson, 2006) routine and the McMaster University-Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) (Samuel

et al., 2011) model coupled with SNOW17 routine. The third model applied is the macroscale land-surface based Variable Infiltration Capacity (VIC) model (Liang et al., 1994) coupled with a routing module. VIC has been applied in nearby, similar river basins for climate change and other hydrological studies (Dibike et al., 2018; Eum et al., 2017, 2014a, 2014b). The above three models were calibrated and validated in this research.

As a benchmark, we used the distributed WATFLOOD model (Kouwen, 1988). For this study, a calibrated and operational WATFLOOD model was obtained from the Manitoba Infrastructure, Hydrological Forecasting Centre. The model has been used by the center to provide operational flood forecasting using real-time weather forecast data to issue short- and medium-range river forecasts in the Upper Assiniboine River basin and other nearby watersheds (Unduche et al., 2018). WATFLOOD is a Canadian Hydrological model specifically developed for flood forecasting and watershed simulation. The model is used as a primary routing module for the Canadian national hydrological modeling system (MESH) (Haghnegahdar et al., 2014). Newman et al., (2017), argues that a calibrated hydrological model which has a familiar practical application in local river forecasting systems has a better functional capability than reference statistical systems to test models, and employs significant water budget interactions is a suitable choice for use as a benchmark model.

For SACSMA and MACHBV models, mean areal daily temperature and precipitation time series data were created by using the Thiessen polygon method from eleven meteorological gauging stations that are distributed across the catchment. The average catchment elevation and latitude of the centroid values were used as an input to the SNOW17 model in addition

to precipitation and temperature data. These inputs were used to calibrate and validate the two lumped models.

For the VIC (version 4.2.d) model, daily gridded interpolated precipitation data was generated from 11 gauging stations, using a bilinear interpolation technique. Daily gridded minimum and maximum temperature data were provided by the Natural Resources Canada, which applied a "thin-plate smoothing splines" (ANUSPLIN) method on observations from several ground-based stations in Canada to generate long-term daily gridded data (Hopkinson et al., 2011; Hutchinson et al., 2009). ANUSPLIN has been used as forcing data for the VIC model in several studies (Dibike et al., 2018; Eum et al., 2017, 2014a, 2014b). Daily average wind speed data performed from the Global Environmental Multiscale (GEM) model (Côté et al., 1998). The grid resolution of the VIC model was about 1/8 degree. Land cover data is obtained from Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type (MCD12Q1) Version 6 data product (Friedl and Sulla-Menashe, 2015). Soil data were imported from FAO's Harmonized World Soil Database V 1.2 (FAO et al., 2009). The runoff from the land surface VIC grid cells was routed to and along the river networks using the RVIC routing module (Hamman et al., 2017) based on Lohmann et al., (1996).

For the WATFLOOD model, gridded interpolated daily precipitation and temperature data were used to set up and calibrate the model. The model was set up at a grid resolution of approximately 5 km for the Upper Assiniboine River Basin.

3.4.2. Calibration and Validation

The catchment outflow simulated by the hydrological models is considered as the reservoir inflow because the reservoir, located at the very downstream location, collects all water from tributary rivers and lateral inflows. During the calibration process, the comparison was made between the simulated and the observed daily reservoir inflow time series.

Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007) was used to optimize the calibration of SACSMA/SNOW17, MACHBV/SNOW17, and VIC/RVIC models. Calibration and Validation of the models were performed with daily timesteps from January 2005 to December 2015 with 1-year spin-up periods.

DDS has been previously compared with other optimization methods, such as shuffled complex evolution (SCE) by Tolson and Shoemaker (Tolson and Shoemaker, 2007). In their study, the dimensionality and efficiency of DDS, for example, was tested, and the authors concluded that DDS provided better results than SCE both with low- and high-dimensional problems, and is more efficient. DDS has been used to calibrate several hydrological models from simple lumped to medium level distributed models (e.g., SWAT (Aliyari et al., 2019; Ilampooranan et al., 2019), MESH (Rokaya et al., 2019), CRHM-AHM (Krogh and Pomeroy, 2019)) to very complicated land-surface based models (e.g. WRF-Hydro (Lahmers et al., 2019; Sharma et al., 2019)).

For lumped models, 10 parameters of SNOW17, 15 parameters of SACSMA, and 12 parameters of MACHBV were calibrated. The optimizing parameters are presented in Appendix A.1, A.2, and A.3 for SNOW17, SACSMA, and MACHBV, respectively. For VIC model, the total number of parameters to be optimized and calibrated is increased from

the default 13 to 53 including the wetland and routing parameters. The optimizing parameters for the VIC model are presented in Appendix A.4. Like the lumped models, the simulated reservoir inflow from VIC/RVIC model was compared with daily observed flow in the calibration process.

For all models, a single objective function obtained by a weighted average of two performance metrics was used in the DDS optimization. The performance metrics that were given equal weight are Kling–Gupta efficiency (*KGE*) (Gupta et al., 2009), and Peak Flow Criteria (*PFC*) (Coulibaly et al., 2001) as defined below.

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

$$PFC = \frac{\left(\sum_{i=1}^{n_p}\left(\left(q_{s,i} - q_{o,i}\right)^2 {q_{o,i}}^2\right)\right)^{1/4}}{\left(\sum {q_{o,i}}^2\right)^{1/2}} \tag{3-3}$$

where r is the correlation coefficient between simulated inflow and observed reservoir inflow, a and b are ratios of the standard deviation and mean of simulated inflows to the corresponding observed inflow respectively, q_s and q_o are the peak simulated and observed inflows respectively, and n_p is the number of peak flows greater than one-third of the mean peak flow observed. While *KGE* values closer to 1 indicate a better model performance, a *PFC* value closer to 0 signifies best peak flow simulation accuracy.

3.4.3. Bias correction

Each of the eleven ensemble precipitation forecasts from the GEFSv2 (Section 3.3.2.1) was bias-corrected by the Empirical Quantile Mapping method (Amengual et al., 2012). The bias-correction of ensemble forecasts were performed using the reanalysis precipitation

product of CaPA as a verifying database. Daily CaPA precipitation time series from January 2014 to December 2015 were used to bias-correct ensemble weather forecasts from January 2016 to December 2017 (Figure 3-3). That is, 4-years of daily ensemble precipitation forecasts were archived first (Jan 2014 - Dec 2017). Then CaPA data was used as a training dataset for the first 2-years of ensemble forecasts. Parameters from the quantile mapping in the training period were applied to the last 2-years of ensemble forecast time series. This step is repeated for each ensemble member to produce a bias-corrected ensemble GEFSv2 inputs.

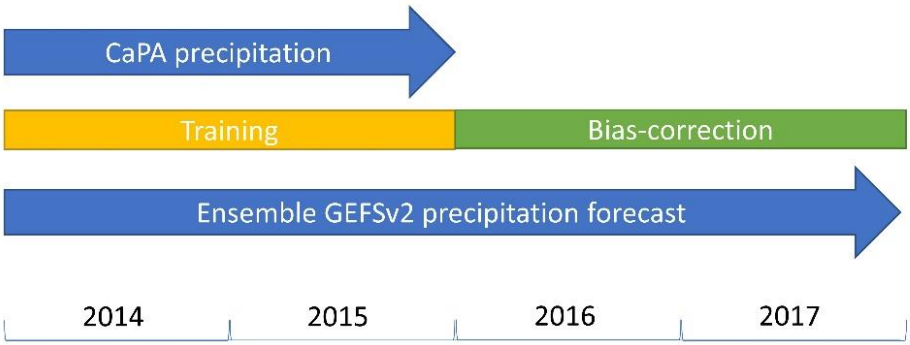

Figure 3-3: The Bias-correction process using the Empirical Quantile Mapping method

3.4.4. Hindcast simulation (model update and forecast)

Hindcast simulation is performed in order to verify the hydrological models in forecast mode. The raw and bias-corrected Reforecast GEFS ensemble datasets were fed into the four calibrated hydrological models. The focus of the study is to assess the reservoir inflow forecast accuracy and skill of the models during the high flood periods. Therefore, 2017 is selected for forecast verification, which observes frequent spring and summer floods in the area. The hindcast period was from April 2017 to September 2017. Continuous model update and forecast were performed during the hindcast period (Figure 3-4). The

hydrological models were run with observed meteorological data for at least one year before the forecast day in order to preserve and update the model's state parameters. In other words, the observed inputs were supplied to the models up to day-0. Then ensemble forecasts were fed to the models for the next eight days, and this model update, and forecast continuously every day during the entire hindcast period.

Figure 3-4: Model update and forecast in the hindcast period

3.4.5. Ensemble forecast verification

Outputs from the previous step (model update and forecast) are daily ensemble reservoir inflow forecasts from four hydrological models for 1- up to 8-day lead times. The forecast skill and reliability of each model's ensemble reservoir inflow forecasts were evaluated using various ensemble verification metrics which are outlined below.

3.4.5.1. Mean Continuous Rank Probability Score ($\overline{CRPS}$) and Skill Score ($\overline{CRPSS}$)

Mean *CRPS* measures the error of the commutative probability of the ensemble forecast. For infinite number of classes or continuous variables *CRPS* is calculated as follows (Wilks, 2006): Given cumulative distribution function of an ensemble y is P(y) and corresponding cumulative probability of observed value x with a step function 1{.}

representing 1 for ensemble values greater than observation and 0 otherwise, the CRPS and the mean *CRPS* can be computed by equation 3-4.

The mean *CRPS* can be decomposed to mean reliability ($\overline{Reli}$) and potential *CRPS* components, according to Hersbach, 2000. $\overline{Reli}$ is directly related to rank histogram but provides more information. It measures the reliability of the system by examining whether the frequency of observations that falls in any one of ranked bins is equivalent to the other bins by taking into consideration the width of the bins in which the rank histograms don't do (Hersbach, 2000). The potential *CRPS* ($CRPS_{pot}$) is the CRPS of a perfect reliable system (i.e. when $\overline{Reli}$ = 0) or for a deterministic forecast where there is no spread. $CRPS_{pot}$ is directly related to the spread of the ensembles and the presence of outliers (Hersbach, 2000). The larger the spread or, the more outliers, the larger the $CRPS_{pot}$. $\overline{CRPS}$, $\overline{Reli}$ and $CRPS_{pot}$ are negatively oriented, meaning a value of zero corresponds to a perfect ensemble forecast. Details of the derivation can be found in Hersbach, 2000.

$$CRPS = \int_{-\infty}^{\infty} (P(y) - 1\{y \geq x\})^2 \, dy$$

$$\overline{CRPS} = \frac{1}{n}\sum_{i=1}^{n} CRPS_i \tag{3-4}$$

$$\overline{CRPS} = \overline{Reli} + CRPS_{pot}$$

Continuous Rank Probability Skill Score ($\overline{CRPSS}$) is a scalar accuracy or performance measurement of the forecasting system by evaluating the mean continuous ranked probability score ($\overline{CRPS}$) of ensembles with relative to a reference forecasting system

(Bradley and Schwartz, 2011). It is positively oriented with a perfect score of 1 and is calculated by:

$$\overline{CRPSS} = 1 - \frac{\overline{CRPS}}{\overline{CRPS}_{ref}} \quad (3\text{-}5)$$

For the reference forecasting system, we used the climatological ensembles of the last twenty-four years of historical daily reservoir inflows. This is practically used by Manitoba Hydrological Forecasting Center to issue medium- and long-term ensemble forecasts at the site (Muhammad et al., 2018). Pappenberger et al., (2015), discussed the option of using climatological observations as an alternative benchmark hydrological ensemble prediction.

3.4.5.2. Reliability diagram

Reliability diagram, also called Attribute Diagram by Hsu & Murphy, (1986), is a measure of the accuracy of ensemble forecasts, which plots the observed relative frequency with respect to forecasting probability in different bins of the category (Wilks, 2006). It is a plot of forecast probability versus observed frequency, and perfect reliability is indicated by a curve lying along the diagonal of a reliability diagram (Atger, 1999).

The *CRPS* decomposition parameters of Hersbach, (2000), were used to draw reliability diagrams in this study. We apply 5% and 95% confidence intervals for the reliability diagrams using the bootstrap resampling technique to measure the conditional verification pair sample uncertainty.

3.4.5.3. Relative operating characteristics (*ROC*) and skill score (*ROC Score*)

ROC is a powerful metric to measure the probabilistic forecast occurrence of events across a range of thresholds (Mason and Graham, 2002). For each threshold, *ROC* examines the correspondence between the forecast and observation by defining the probability of detection (hit rate) and the probability of false detection (False alarm rate). ROC curve for several thresholds can then be constructed by 'Hit Rate' values as ordinate and 'False Alarm Rate' values as abscissa. A good and skillful forecast produces ROC curve above the 45 degrees diagonal but more towards the top-left position indicating high 'Hit Rate' and low 'False Alarm Rate' (Mason and Graham, 2002). *ROC* shows the discrimination skill of the ensemble forecast system (Brown et al., 2010). Discrimination skill indicates the ability of the forecasting system to categorize occurrence and non-occurrence of floods defined between user-defined probability thresholds (Brown et al., 2010; Mason and Graham, 1999).

A single scalar score can summarize the quality of *ROC* curves. *ROC* score is a function of the area under the ROC curve (AUC). (Wilks, 2006) formulates a simple equation for *ROC Score* as:

$$ROC\ Score = 2 * AUC - 1 \quad (3\text{-}6)$$

where *AUC* is the area under the curve of each Relative Operating Characteristics curves. A perfect system that has *ROC* curves close to the top left corner would have a score of 1.

3.5. Results

3.5.1. Calibration and validation

As noted in Section 3.4.2, DDS optimization was used to calibrate parameters of SACSMA with SNOW17, MACHBV with SNOW17, and VIC with RVIC models. Among the ten years chosen for calibration and validation, the recent five consecutive years (from 2011 to 2015) were used to calibrate the models, and the previous five years (2006 to 2010) were used to validate the models (Figure 3-5). The reason why we used recent data for calibration is that we want to train the models using high consecutive flood periods. Looking at the historical time series from 2006 to 2015, the recent five-years are high consecutive flood years than the previous five-years. Moreover, it is highly likely that this trend will continue past 2016 and the near future due to anticipated climate change impact in the region and other similar factors that caused the recent high consecutive flood years. Since the challenge of achieving the accurate reservoir inflow forecasting arises particularly during flood periods, and the objective of the paper focuses on improving the accuracy of flood forecasting in large complex watersheds, the hydrological models were trained/calibrated with the recent flood years.

The performance metrics of the models are summarized in Table 3-1. The KGE performance statistics indicate that the SACSMA model outperforms MACHBV followed by VIC during calibration as well as validation periods. The lumped models (SACSMA and MACHBV) appear to show better performances than the macroscale model (VIC). The Peak Flow Criteria (PFC) shows that SACSMA and MACHBV have improved and have comparable accuracy in peak flow prediction. VIC model slightly underestimates and

delays peak flows occasionally, although it maintains the hydrograph during spring and summer high inflow seasons.

The performance of the models can be seen from the simulated and observed flow hydrographs shown in Figure 3-5. Visual inspection shows that all three models comparatively capture the pattern of the observed reservoir inflow hydrographs during calibration and validation periods; although SACSMA and MACHBV models appear to reproduce the peak flows better than VIC. In addition to the visual inspection, the RMSE and PBIAS values of each model are displayed in Figure 3-5 to provide more information on the hydrographs. It can be seen that SACSMA provided better accuracy and less bias during calibration and validation, followed by MACHBV and VIC models in decreasing order of performances.

The calibration of the models was performed in a daily time step, and the optimization method (DDS algorithm) used during the calibration was the same for all models. The objective function is also the same, which is the average of *KGE* and *PFC*. However, there are differences in the number of parameters (dimensionality) among the models. Note that the lumped models were coupled with SNOW17, hence the total number of the calibrated parameters are the summation of individual models' parameters; for example, SACSMA (15) plus SNOW17 (10). As noted in Appendix A.4, the default number of VIC/RVIC model parameters was further refined to improve the calibration output and to better represent the wetland, landcover, and soil types of the basin. The parameters were refined and increased from the default 13 to 53 based on land cover classes and soil mapping units (Figure A1). A simple test has been done before refining the parameters by performing the

calibration using the default 13 parameters, and the preliminary results were much worse (*KGE* = -2.3, not shown in the results section) than after refining the parameters (*KGE* = 0.653). With the default parameters, VIC, as a macroscale land-surface based model, was not correctly estimating the water and energy balance equations in a vertical column at each grid cell and transferring water between grids and river networks by using the routing module (RVIC). After refining, the model significantly improved the water interaction in wetland areas, and different land cover and soil tiles and routed the flow to the outlet.

The message here is that an effort has been made to employ a better calibration approach with an efficient optimization algorithm for the advanced model (VIC). As discussed in Section 3.4.2, DDS is a competitive and efficient optimization tool that has been applied in several distributed and land-surface based hydrological models. Thus, it is safe to say that the conclusion (i.e. the improved performance of SACSMA and MAHBV over VIC in the calibration outputs) was not limited by the search algorithm.

Table 3-1: Performance statistics of the three hydrological models from calibration and validation. The definition of the abbreviations is presented in Section 3.4.2.

	Calibration			**Validation**		
	SACSMA	MACHBV	VIC	SACSMA	MACHBV	VIC
PFC	0.180	0.174	0.270	0.234	0.231	0.247
KGE	0.796	0.740	0.660	0.776	0.679	0.653

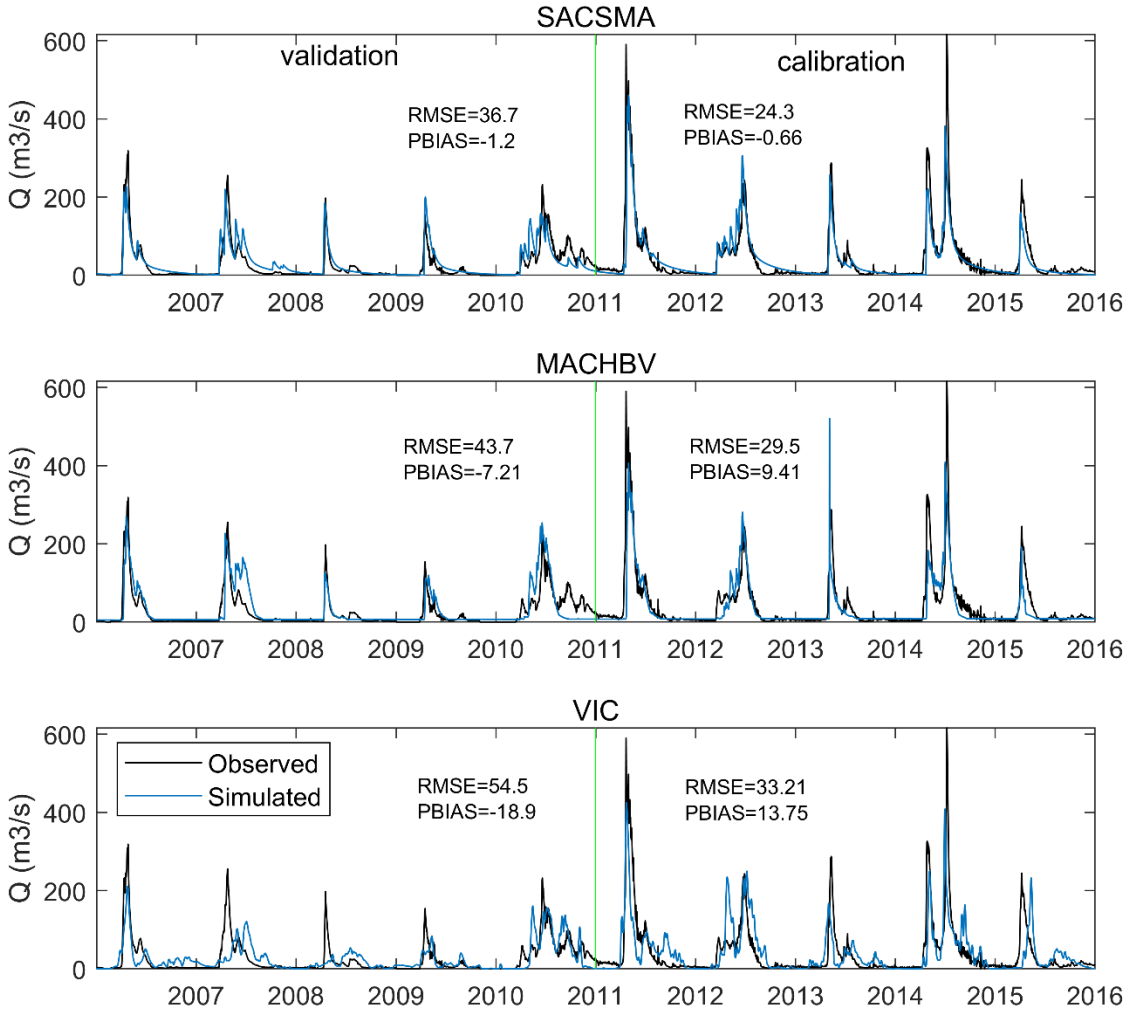

Figure 3-5: Calibration and validation plots of SACSMA with SNOW17, MACHBV with SNOW17, and VIC with RVIC models. RMSE: Root-Mean-Square Error, and PBIAS: Percent Bias (Yapo et al., 1996) are displayed for each model to provide more information in addition to the visual inspection of the time series.

3.5.2. Model comparison in forecast mode

3.5.2.1. Overall forecast quality and skill

The GEFSv2 ensemble precipitation forecasts were bias-corrected by the Empirical Quantile Mapping method using CaPA as a verifying analysis, as described in Section

3.4.3. Both raw and bias-corrected GEFSv2 inputs were fed into four hydrological models in order to (1) realize the effect of the bias-correction on the output hydrological forecasts, and (2) compare the models forecast performance pre- and post-bias correction process.

Figure 3-6 shows the mean CRPS, which measures the overall probabilistic error of the ensemble reservoir inflow forecasts generated by four hydrological models and GEFSv2 inputs. As expected, the bias-corrected GEFSv2 ensembles significantly outperform the raw GEFSv2 inputs regardless of the hydrological models used. The quality of hydrological forecasts was much improved by bias-correcting each ensemble precipitation forecast of GEFSv2 with CaPA reanalysis data. Figure 3-6 also shows a comparison between the forecast quality of the four hydrological models. For all models, the overall forecast quality declines as the lead time increases, as expected. It can be seen from the figure that the mean *CRPS* values of the SACSMA model are the lowest followed by MACHBV, WATFLOOD, and VIC in ascending order of forecast probability error. Whether using raw GEFSs or bias-corrected GEFSs as in inputs, the resultant hydrological forecast skill of the two lumped models (SACSMA and MACHBV) outperforms the benchmark distributed WATFLOOD model and the macroscale VIC model at all lead times. However, the benchmark model provides a better skill than VIC and is relatively close to the two lumped models at early lead times.

So far, the models' ensemble outputs were evaluated based on their overall forecast error. In order to add a comprehensive outlook, a reference ensemble forecasting system is used to evaluate their skills.

Figure 3-7 shows the mean CRPS skill score (CRPSS) of ensemble reservoir inflows simulated by the four different hydrological models. It can be seen from the figure that, the *CRPSS* values of the four models have a similar trend as the *CRPS* depicting that the lumped models have better forecast skill than the benchmark and macroscale models. Comparing to the reference climatological-based ensembles, SACSMA provides the best quality of ensembles at all lead times, followed by MACHBV, WATFLOOD, and VIC. The lumped models were competitive throughout the forecast horizon with minor exceptions. For the first two to three days, the skill score of the benchmark WATFLOOD is relatively close to the lumped models, but the forecast skill gradually deteriorates at later lead times.

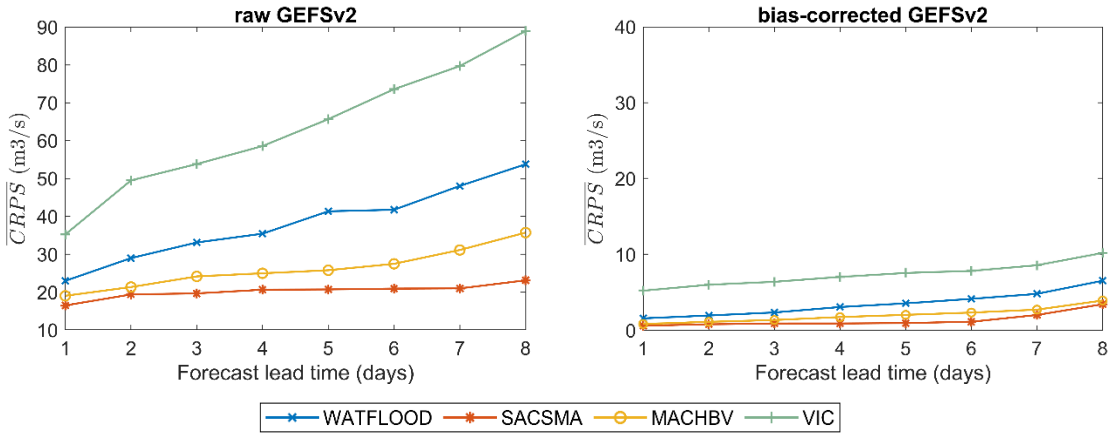

Figure 3-6: Mean CRPS of ensemble reservoir inflows generated with Raw (left) and Bias-corrected (right) ensemble GEFSv2 precipitation forecasts.

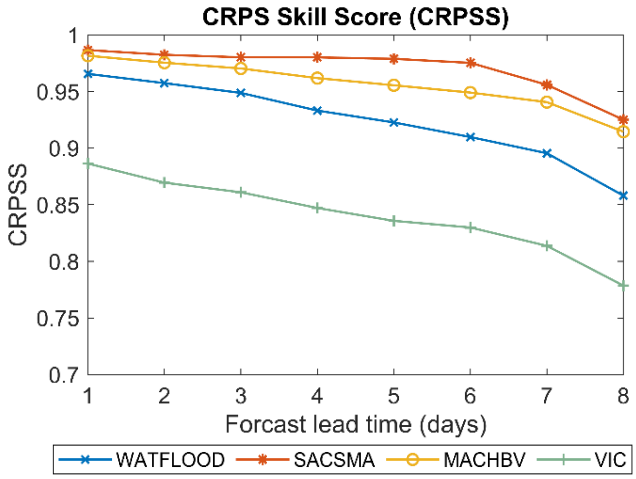

Figure 3-7: Comparison of reservoir inflow ensembles between four hydrological models using CRPS skill score (CRPSS).

3.5.2.2. Reliability

The reliability of the ensemble hydrological forecasts was evaluated by two metrics; using the reliability component of *CRPS* after decomposition of (Hersbach, 2000), and using Reliability Diagram.

Figure 3-8 shows the components of CRPS after the decomposition method of Hersbach, 2000. Here, the summation of the reliability (left) and potential CRPS (right) components of each hydrological model is the mean *CRPS*. The reliability and potential *CRPS* components follow the same trend as the mean *CRPS* and *CRPSS*. The decomposition indicates that the lumped models (SACSMA and MACHBV) were more reliable and have less spread and outliers than the benchmark WATFLOOD and macroscale VIC models. The reliability component contributes about half of the mean *CRPS*. The rest comes from the potential *CRPS*. Remarkably, it can be observed that the forecast quality of WATFLOOD during the first two or three days comes from the reliability component because this value is lower and much closer to the lumped models than the potential *CRPS* component. The overall forecast quality of SACSMA model remarkably remains the same up to lead time of day 6, as can be seen from mean *CRPS* and *CRPSS* values. This effect is mainly due to the potential *CRPS* component which remains either constant or slightly dropped as going from day 1 to day 6. SACSMA model generates ensembles that are less spread and have low number of outliers in the first six days forecast as explained by the potential *CRPS* component. The potential *CRPS* rapidly increased in all models after lead time seven, which indicates that the ensemble spread and presence of outliers start to significantly rise irrespective of the model type after a seven-day forecast. The sudden rise and decline of mean *CRPS* and *CRPSS* in most models after day seven maybe due to this effect.

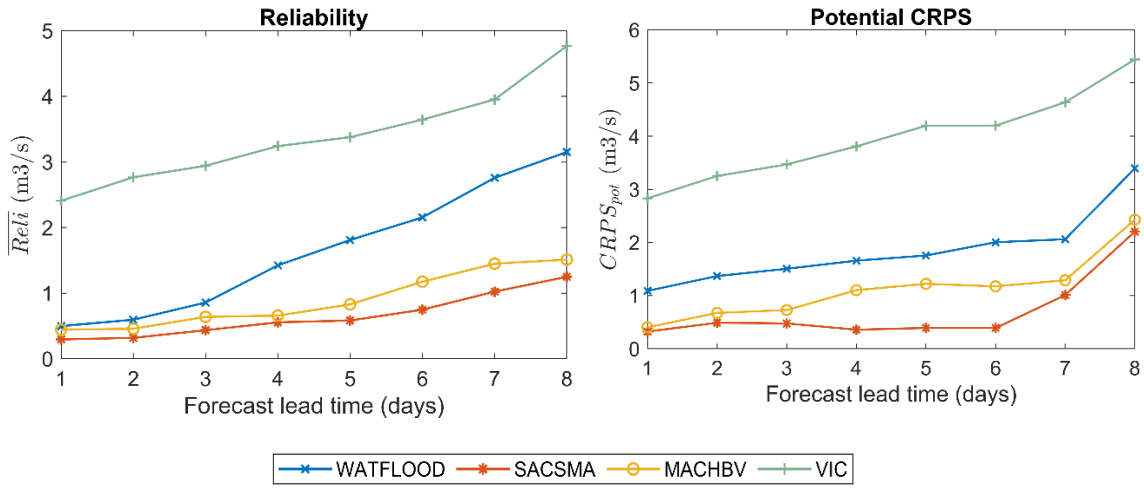

Figure 3-8: CRPS decomposition components. The left plot shows the reliability component, and the right plot shows the Potential CRPS component. Comparison of these attributes was made between four hydrological models.

Figure 3-9 shows the reliability diagrams of the hydrological models for forecast lead time of day 1, 3, and 5. For a one-day lead time forecast, the reliability curves of SACSMA, MACHBV, and WATFLOOD were all reasonably aligned along the diagonal line, which indicates that they achieve relatively more reliable forecasts. The conditional observed frequency is comparable with the forecast probability with slight exceptions in the very lower bin. For day three forecasts, this trend minimally changes, but overall, the reliability of the WATFLOOD is not significantly lower than the lumped models. For day five, the reliability curve of SACSMA is still close to the diagonal (‘perfect line’), especially on higher forecast probabilities. MACHBV is relatively reliable on day five forecast, as shown by its diagram. However, the reliability curve of WATFLOOD at day five is away from the diagonal line indicating its reliability was progressively declining after day three forecast. The reliability of VIC, although relatively moderate at day one, was reduced at day three

and five forecasts because it continuously underestimates the forecast. The 90% Confidence Intervals (CI) of the reliability diagrams showed that uncertainties in the conditional verification pair samples increased in all models as the forecast lead time increases. However, the advancement of conditional uncertainty in the forecasts in lumped models was not substantial when compared to the benchmark and macroscale models. This is because the reliability lines of SACSMA and MACHBV are within the CI bounds most of the time, and their CI's are closer to the diagonal line. Whereas, for VIC, the reliability lines are either at the lower or upper level of the CI's in all cases, and for WATFLOOD this occurs on day three lead time. This characteristic indicates that 90% of the cases, the reliability diagram attributes of VIC, and sometimes WATFLOOD did not belong to the interval where the "true" value of the attributes exists, whereas for the lumped models this does not hold.

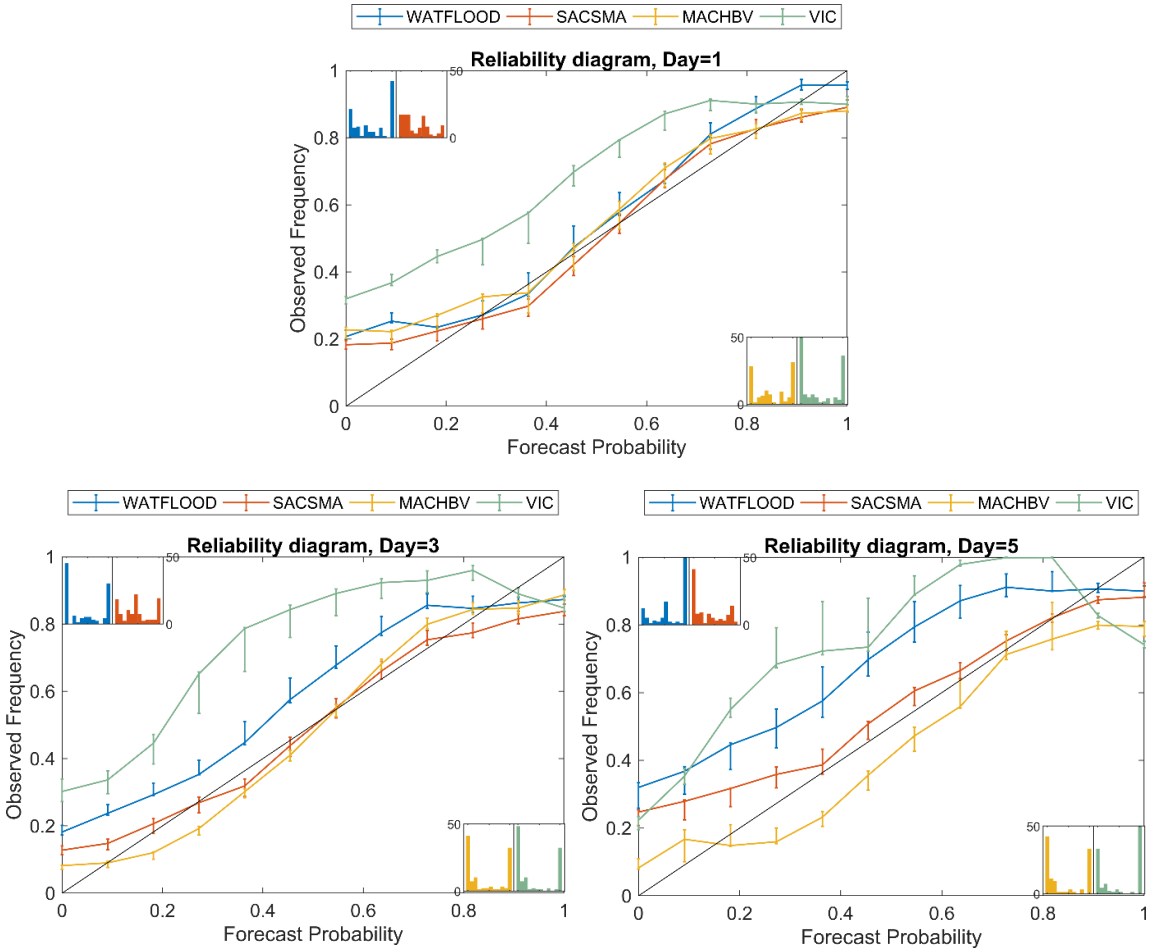

Figure 3-9: Reliability diagrams of ensemble reservoir inflows at three selected forecast lead times. Different colors show different model types. The 90% confidence intervals were shown in the reliability lines for each model. The inset histograms show the frequency of occurrence in each forecast bin.

3.5.2.3. Hit and False alarm rate distribution

As described in Section 3.4.5.3, *ROC* displays the hit-rate and false alarm rate of a forecasting system at different thresholds.

Figure 3-10 shows the *ROC* curves of the models at day 3 forecast lead time. The hit rate versus false alarm rates was drawn for varying higher probability threshold levels of reservoir inflows because the primary focus of this research is on flood forecasting. Simulated ensemble inflows exceeding 75, 80, 85, 90, and 95 percentiles of the observed reservoir inflow were taken into consideration. At day three lead time (Figure 3-10), SACSMA performed well in attaining the highest true alarm and lowest false alarm rates for all probability thresholds as compared to other models, the closest one being MACHBV. Forecasting the most extreme flood or flows exceeding 95 and 90 percentile inflows is a challenge that most models lack with different levels of forecast skill. The lumped models and subtly the benchmark are deemed sufficient to construct ensembles that have good discrimination skills to forecast up to 85 percentile reservoir inflow. Although the *ROC* curves stipulate that VIC can, in fact, reproduce 80 percentile flows up to five days ahead forecast time other probability thresholds have almost zero skills.

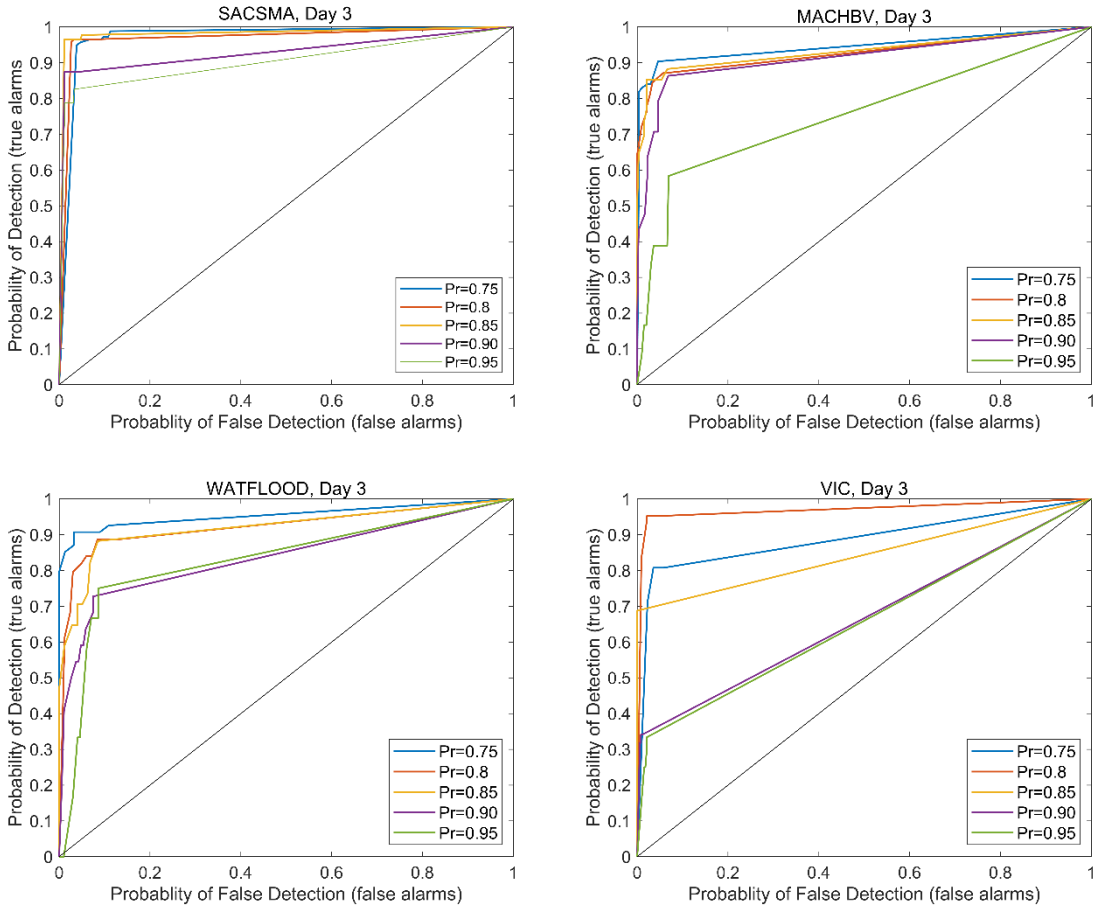

Figure 3-10: Relative Operating Characteristic (ROC) curves drawn for probability thresholds exceeding 75, 80, 85, 90, and 95 percentile reservoir inflows for three days ahead forecast. The four plots are for four different hydrological models.

In Figure 3-11, the *ROC Scores* of each model, estimated by the average of the area under the *ROC* curves for the considered probability thresholds, are shown. It summarizes the performance and discrimination skills of the models' ensembles for all forecast time horizons. The *ROC Scores* indicate that the forecast skills of WATFLOOD and VIC monotonically decrease as the lead time increases, but for the case of SACSMA and MACHBV, even though their skill unevenly decline, they have competitive and relatively

decent forecast performances. The declining *ROC Scores* indicate that as the lead time increases, the ROC curves (not shown here) progressively approach the diagonal line which is the climatological forecast or "zero skill" line (Brown et al., 2010). In general, considering all the forecast lead times and probability thresholds, SACSMA appeared to have a better discrimination skill more than the others, followed by MACHBV, WATFLOOD, and VIC in order of performance.

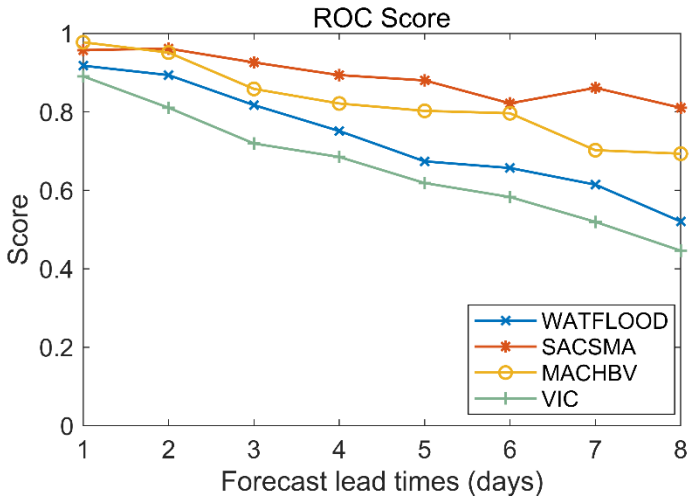

Figure 3-11: The ROC Score measured by the Area Under the ROC Curves of four hydrological models.

3.6. Conclusion and Discussion

The objective of this study was to identify hydrological models from a pool of diverse model structures that can produce better forecast skill and reliability and provide an enhanced short- and medium-range reservoir inflow forecasts in a Prairie watershed: The Upper Assiniboine River Basin. A comparison of forecast skill and reliability between the selected hydrological models was made using raw and bias-corrected ensemble

precipitation forecast products. The best model was selected from two lumped models (SACSMA with SNOW17 and MACHBV with SNOW17), a benchmark distributed model (WATFLOOD), and macroscale land-surface based model (VIC). Daily total precipitation forecasts were collected from an 11-member second-generation Global Ensemble Forecast System reforecast (GEFSv2). Each of the ensemble members was bias-corrected by Empirical Quantile Mapping method using the Canadian Precipitation Analysis (CaPA) as a training/verification dataset. Raw and bias-corrected GEFSv2 precipitation were supplied to the hydrological models to evaluate and compare the forecast skill and reliability of the ensemble inflow outputs. Forecast evaluation was performed in a 6-month hindcast period where daily ensemble reservoir inflow forecasts were issued for 1-day up to 8-days forecast lead times. SACSMA, MACHBV, and VIC models were calibrated in the study area by comparing simulated and observed inflows into Shelmouth Reservoir while WATFLOOD model, which is operationally implemented for the Provincial real-time flood forecasting was used as a benchmark.

Results indicated that simulated ensemble reservoir inflows generated by bias-corrected GEFSv2 provided significantly better forecast quality than the raw GEFSv. Even though this result is expected, two things are noticed; first, the bias-correction of each ensemble members instead of the mean or median provided a consistent and reliable ensemble inflow forecast, and second, bias-correcting forecasts using verifying datasets (such as CaPA) for a training period of at least two years before the forecast time results in an improved hydrological forecast. This method and the improved result can be beneficial for users at

operational flood forecasting centers as they would generally prefer less advanced and quick post-processing methods.

All models were supplied with bias-corrected ensemble GEFSv2, and various ensemble verification metrics were used to compare the model outputs up to eight days of forecast lead times. The overall forecast quality and skill of the models' results were evaluated by using mean CRPS and CRPSS metrics. Results indicated that the two lumped models (SACSMA and MACHBV) provided better overall forecast performance than the benchmark WATFLOOD and the macroscale VIC models. Although the lumped models (SACSMA and MACHBV) were found to be comparable, SACSMA provided enhanced forecast skill than MACHBV at all lead-times. For shorter lead-times, particularly up to day 3, WATFLOOD provided relatively competitive overall forecast quality as of the lumped models.

The CRPS decomposition by (Hersbach, 2000) was found to be vital to interpret and better analyze the overall forecast performance. This decomposition indicated that the modest forecast skill of WATFLOOD in the first 2 or 3 days came from the reliability component of CRPS. The decomposition of CRPS further indicated that the superior performance of SACSMA is due to its ability to generate ensemble inflow forecasts with less ensemble spread and low presence of outliers in the first six days of the forecast as explained by the potential CRPS component.

Reliability diagrams of the hydrological models at different lead times provided further insight into the forecast skill of the ensembles. At shorter lead times, the reliability diagrams of SACSMA, MACHBV, and WATFLOOD indicated that they all achieve

relatively reliable forecasts as the conditional observed frequency was comparable with the forecast probability. However, after day 5, the reliability of WATFLOOD deteriorated while MACHBV and SACSMA, in order of increasing performance, remain within reasonable calibration accuracy. The reliability of VIC, although relatively moderate at day one, was weak because it continuously underestimated the forecast.

In order to evaluate the discrimination skill of the ensembles, two threshold-based metrics were used to evaluate the hit-rates and false-alarm rates at different higher forecast thresholds: Relative Operating Characteristic (ROC) curve and the ROC Score measured by the area under the ROC Curves. ROC curves of the models were drawn and compared for ensemble reservoir inflows exceeding between 75 and 95 percentiles, with a 5 percent increment. For day three forecast, SACSMA and MACHBV models attained highest true alarm, and lowest false alarm rates for all probability thresholds with the former slightly outperformed the later. As the lead time increases, forecasting the most extreme flows exceeding 95 percentile inflows was a challenge for most models. However, the lumped models and moderately the benchmark were sufficiently able to generate ensemble inflows that have very good skills to forecast inflows exceeding the 85 percentiles. Overall, considering all the forecast lead times and probability thresholds, SACSMA provided better discrimination skill than the others, followed by MACHBV, WATFLOOD, and VIC in order of decreasing performance.

In general, ensemble inflow forecasts generated by the lumped models offered substantially better performances as compared to the benchmark distributed model or the macro-scale land surface models. The distributed benchmark model unequivocally provided reliability

as good as the lumped models up to three days ahead even though it deteriorates rapidly at later lead times. It is anticipated that the forecast performance of the VIC model could be improved by increasing the grid resolution of the model, which was set up at 1/8-degree horizontal resolution. Overall the SACSMA appeared to generate the most reliable and skillful ensemble reservoir forecast inflows for up to a week ahead lead times and should be considered as an alternative operational model in the study area.

The performance of different hydrological models depends on many factors such as the scale, the complexity of the basin, the spatial and temporal resolution of the input data, the structure of the models, the degree of discretization of the models, and the number of parameters to be calibrated, etc. For the models that were applied to this research, these factors are interconnected and thus affect their calibration performance jointly. The intended purpose of the hydrological models in this study is to simulate and forecast short- and medium-range reservoir inflows. Regardless of the structure and degree of discretization of the models, the objective is to obtain a time series (hydrograph) implicitly at one location, which is considered as the watershed outlet, and no interior locations or sites are needed. The way the inputs were supplied to the models depends on the type of the model (e.g., lumped, distributed) and the discretization level (e.g., spatially lumped catchment, grids, GRUs, HRUs). Hence it can be said that the calibration performance of the hydrological models was influenced jointly by the above factors.

Moreover, there are many references from the literature where lumped models outperformed various distributed or land-surface based models. The Distributed Model Inter-comparison Project (DMIP) has implemented several hydrological models at eight

basins of the River Forecasting Centers in the USA, and the results showed that lumped models (particularly SACSMA) provided better performance than distributed (such as WATFLOOD, SWAT) and land-surface based models (such as VIC, NOAH) (Reed et al., 2004). Results from the Model Parameter Estimation Experiment (MOPEX) also demonstrated a significantly improved performance of SACSMA comparing to land-surface based models including VIC and SWAP (Duan et al., 2006; Nasonova and Gusev, 2007). Maurer et al., 2010, performed a comparative study between SACSMA and VIC models, and the results revealed that the former lumped model had an evident better calibration performance over the later land-surface model.

The research is conducted to identify best performing hydrological models for improved hydrological forecasting in a specific large complex watershed of the prairie region of Canada. The Upper Assiniboine Basin is characterized as a "Prairie" watershed, which is known for its complex hydrology due to the presence of potholes. Hence, the study area is considered as one example of a complex watershed. The hydrological models have a diversified structure (lumped, distributed, and macro-scale land-surface based) and implemented to evaluate and select the model that has the best potential for simulating and predicting reservoir inflows for such a complex basin. If another kind of complex watershed with the same scale is used, it is believed that a similar conclusion would be drawn. The previous studies that provided similar conclusions were conducted in various watershed landscapes. The MOPEX project was tested in twelve watersheds that have various land cover types such as croplands, mixed forests, and natural vegetation in different altitudes (Duan et al., 2006; Nasonova and Gusev, 2007). In the DMIP study, the dominant land

cover properties of the eight basins were mainly agricultures and forests with varying topographies and soil types (Reed et al., 2004; Smith et al., 2004). The comparative study of Maurer et al., (2010), was performed in snow-dominated catchments. Overall, the same candidate model(s) would highly likely be identified to better simulate and forecast medium-range reservoir inflows in other types of complex watersheds with a similar scale and characteristics.

In general, for hydrological forecasting focusing on basin outflows and not interior sites, the study indicated that lumped models, particularly SACSMA with SNOW17, provided better performance than the distributed or land-surface models in complex watersheds. Not only the calibration but also the validation and forecast verification analysis have given the superiority in simple models. The verification of hydrological forecasts generated from bias-corrected ensemble weather forecast inputs provided enough details of the model's performances for the intended purpose.

3.7. Author Contributions

Conceptualization, F.A. and P.C.; Data curation, F.A.; Formal analysis, F.A.; Investigation, F.A.; Methodology, F.A. and P.C.; Software, F.A., P.C., and F.U.; Supervision, P.C. and I.T.; Validation, F.A. and P.C.; Resources, P.C., I.T., and F.U.; Writing—original draft preparation, F.A.; Writing—review and editing, P.C. and F.U.

3.8. Funding

This work was supported by the Natural Science and Engineering Research Council (NSERC) Canadian FloodNet (Grant number: NETGP 451456).

3.9. Acknowledgments

The authors would like to thank the Manitoba Infrastructure, Hydrologic Forecasting Center, for providing the operational WATFLOOD model along with data.

3.10. Conflicts of Interest

The authors declare no conflict of interest.

Appendix A: Brief description of the calibration of models

A.1. SNOW17

The Snow Accumulation and Ablation Model (SNOW17) model was developed by (Anderson, 2006) as part of the NWS river forecasting system. It is a conceptual model that uses a temperature index to determine energy exchange across the snow-air interface (Anderson, 2006).

Inputs to the model are:

i. The mean area observed precipitation time series obtained by Theisen Polygon method
ii. The mean area observed temperature time series obtained by Theisen Polygon method
iii. The average elevation of the catchment
iv. The latitude of the centroid of the catchment
v. The parameters that were calibrated are listed in Table A- 1.

The MATLAB version of the source code was used to set up and calibrate the model in the study area.

Outputs from SNOW17 are outflow and Snow Water Equivalent. The outflows are the summation of snowmelt and rain. The coupling mechanism of SNOW17 with SACSMA and MACHBV models is performed by forcing outflows from SNOW17 into the hydrological models.

Table A- 1 : SNOW17 model parameters

No	Parameters	Description	Unit	Ranges
1	SCF	Snowfall correction factor	–	0.4–1.6
2	MFMAX	Maximum melt factor during non-rain periods considered to occur on June 21	mm/6 h/°C	0.5–2.0
3	MFMIN	Minimum melt factor during non-rain periods considered to occur on December 21	mm/6 h/°C	0.05–0.5
4	UADJ	The average wind function during rain-on-snow periods	mm/mb/°C	0.03–0.2
5	NMF	Maximum negative melt factor	mm/6 h/°C	0.05–0.50
6	MBASE	Base temperature for non-rain melt factor above which melt typically occurs	°C	0–2.0
7	PXTEMP1	Lower Limit Temperature dividing tranistion from snow, if temp is less than or equal to pxtemp1, all precip is snow. Otherwise it is mixed linearly	°C	−2.0 to 0
7	PXTEMP2	Upper Limit Temperature dividing tranistion from snow, if temp is greater than or equal to pxtemp2, all precip is rain. Otherwise it is mixed linearly	°C	1 to 3.0
8	PLWHC	percent liquid water holding capacity of the snow pack	–	0.02–0.3
9	DAYGM	Daily melt at snow–soil interface	mm/day	0–0.3
10	TIPM	Antecedent snow temperature index	–	0.1–0.2

A.2. SACSMA

The Sacramento Soil Moisture Accounting (SAC-SMA) model has been used as a lumped conceptual model at the National Weather Service (NWS) for operational river forecasting purposes. It has also been included within the National Weather Service Hydrology Laboratory's Research Distributed Hydrologic Model (HL-RDHM) by adding several processes (Koren et al., 2004). Details description of the lumped SACSMA model can be found in (NWS, 2002).

In this research, SACSMA was implemented as a lumped continuous model in Upper Assiniboine Basin. The MATLAB version of the source code was used to set up and calibrate the model.

For calibrating the model, the following inputs are used:

i. Outflow (rain plus snowmelt) from SNOW17
ii. The catchment area of the basin
iii. Observed catchment outflow estimated by calculated reservoir inflow

The lists the parameters of the model that were calibrated by DDS optimization are presented in Table A- 2.

Table A- 2: SACSMA model parameters

No	Parameters	Description	Unit	Ranges
1	UZTWM	Upper zone tension water maximum storage	[mm]	1–150
2	UZFWM	Upper zone free water maximum storage	[mm]	1–150
3	LZTWM	Lower zone tension water maximum storage	[mm]	1–500
4	LZFPM	Lower zone free water primary maximum storage	[mm]	1–1000
5	LZFSM	Lower zone free water supplemental maximum storage	[mm]	1–1000
6	ADIMP	Additional impervious area	[-]	0.0–0.4
7	UZK	Upper zone free water lateral depletion rate	[day^{-1}]	0.1–0.5
8	LZPK	Lower zone primary free water depletion rate	[day^{-1}]	0.0001–0.025
9	LZSK	Lower zone supplemental free water depletion rate	[day^{-1}]	0.01–0.25
10	ZPERC	Maximum percolation rate	[-]	1–250
11	REXP	Exponent of the percolation equation [-]	[-]	1–5.0
12	PCTIM	Impervious fraction of the watershed area	[-]	0.0–0.1
13	PFREE	fraction percolating from upper to lower zone free water Storage	[-]	0.0–0.6

14	athorn	A constant for Thornthwaite's equation	[-]	0.1-0.3
15	Rq	Routing Coefficient	[-]	0.0–1

A.3. MACHBV

McMaster University-Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) is a modified version of the lumped conceptual HBV model edited by (Samuel et al., 2011) at McMaster University. Detail description of the model can be found in (Bergström Sten, 1978). MACHBV has been implemented in several Canadian watersheds for flood forecasting purposes (Han et al., 2019; Han and Coulibaly, 2019; Leach et al., 2018; Razavi and Coulibaly, 2017, 2016).

The model was calibrated in Upper Assiniboine River Basin in this study. The MATLAB version of the source code was used to set up and calibrate the model. The following inputs were used for calibration:

i. Outflow (rain plus snowmelt) from SNOW17
ii. The catchment area of the basin
iii. Observed catchment outflow estimated by calculated reservoir inflow

The lists the parameters of the model that were calibrated by DDS optimization are presented in Table A- 3.

Table A- 3: MACHBV model parameters

No	Parameters	Description	Unit	Ranges
1	athorn	A constant for Thornthwaite's equation	[-]	0.1-0.3
2	fc	Maximum soil box water content	[mm]	50-800
3	lp	Limit for potential evaporation	[mm/mm]	0.1*fc-0.9*fc

4	beta	A non-linear parameter controlling runoff generation	[-]	1–10
5	K0	Flow recession coefficient in an upper soil reservoir	[days	1–30
6	lsuz	A threshold value used to control response routing on an upper soil reservoir	[mm]	1-100
7	K1	Flow recession coefficient in an upper soil reservoir	[days]	2.5-100
8	cperc	A constant percolation rate parameter	[mm/day]	0.01-6
9	K2	Flow recession coefficient in a lower soil reservoir	[days]	20-1000
10	maxbas	A triangle weighting function for modelling a channel routing routine	[days]	1–20
11	rcr	Rainfall correction factor	[-]	0.5-1.5
12	a1	An exponent in relation between outflow and storage representing non-linearity of storage – discharge relationship of lower reservoir	[-]	0.5-20

A.4. VIC

The Variable Infiltration Capacity (VIC) model is a Macroscopic Land-surface distributed hydrological model. Detail description of the model can be found in (Liang et al., 1994). Since VIC computes its energy and water balance equations in a vertical column at each grid cells, an external river routing module is required to route runoff and baseflows to the edge of each grid cell throughout the river network to the catchment outflow (Lohmann et al., 1996). For this purpose, the python version of RVIC routing module (Hamman et al., 2017) is used in this research. Version 4.2.d of VIC was setup and coupled with RVIC.

Meteorological forcings to the model are:

i. Average daily gridded interpolated precipitation data from the ground network,
ii. Daily gridded minimum, and maximum temperature data from ANUSPLIN,

iii. Average daily wind speed from the Global Environmental Multiscale (GEM) model.

Physical inputs are:

i. Digital elevation model for SNOW elevation bands and flow direction computation
ii. Land cover data from Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type (MCD12Q1) Version 6 data product
iii. Soil data from FAO's Harmonized World Soil Database (HWSD) V 1.2

The grid resolution of VIC model setup was about 1/8 degree. At each grid, three elevation bands and three soil layers were used in the study area.

Some of the main processes are described below:

Snow: Rain-snow partitioning, snow accumulation, and melting are simulated at sub-grid level using temperature index method lapsed through the Elevation (SNOW) bands.

Evaporation: is simulated at each elevation band and land cover type using Penman-Monteith Approach.

For this study, the dynamic wetland module was activated to calibrate the wetland parameters because, as a Prairie watershed, the area has abundant wetland and potholes, as shown in Figure A- 1. The default parameters of the VIC/RVIC model (Table A- 4) were further refined to improve the calibration output and to better represent the physical characteristics of such a large complex basin. Wetland parameters were refined based on the vegetation type in the catchment. Each of the three major land cover classes (Figure A-1) in the area has been assigned with its own five wetland parameters. Similarly, the six soil parameters were sub-categorized into six soil groups based on the dominant and

associated soil types of the basin (Figure A- 1). After parameter refining, the total number of parameters to be calibrated was increased from the default 13 to 53, including the routing parameters.

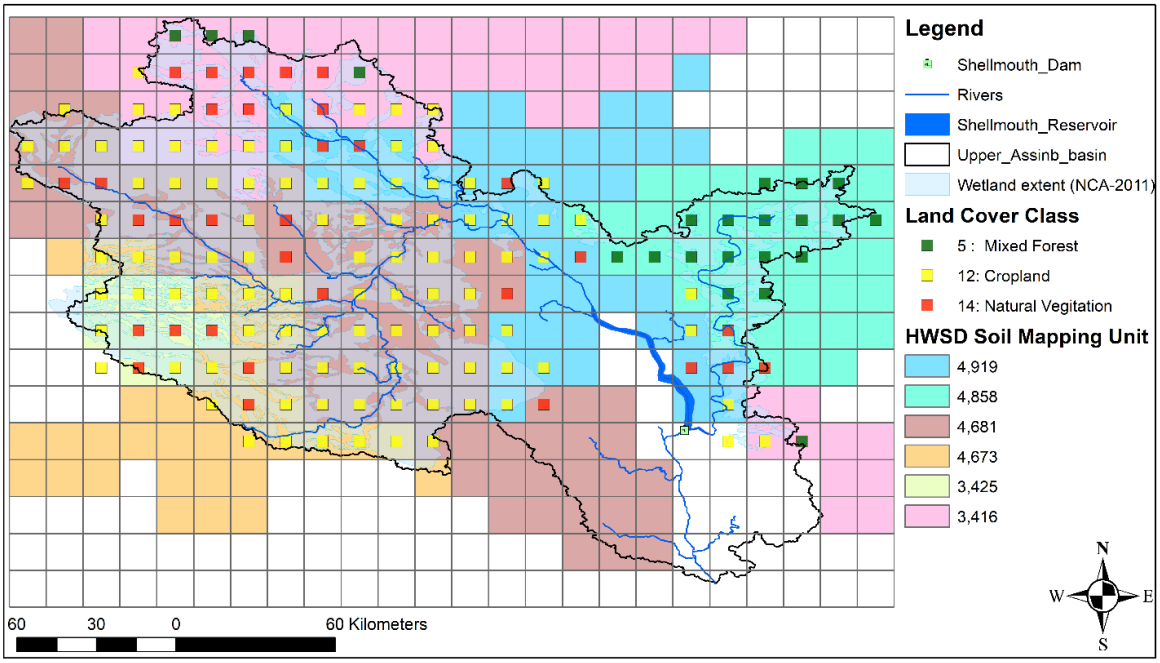

Figure A- 1: Soil and Land cover tiles discretization for VIC model

Table A- 4: VIC/RVIC model parameters

No	Notation	Range	Unit	Definition
	Soil parameters			
1	b	10^{-5} - 0.4	-	Variable infiltration curve parameter
2	Ds	10^{-3} - 1	-	Fraction of Dsmax where non-linear baseflow begins
3	Dm	0.1 - 30	mm/ day	Maximum velocity of baseflow
4	Ws	0.5-1	-	Fraction of maximum soil moisture where non-linear baseflow occurs
5	s2	0.3 - 1.5	m	Thickness of middle soil moisture layer
6	s3	0.3 - 1.5	m	Thickness of bottom soil moisture layer
	Wetland parameters			
7	bmin_depth	0.01 - 0.3	m	Lake depth below which channel outflow is 0.
8	wfrac	0.001 - 0.05	-	Width of lake outlet, as a fraction of the lake perimeter
9	depth_in	0.01 - 0.3	m	Initial lake depth
10	rpercent	0.1 - 1	-	Fraction of grid cell runoff that enters lake (instead of going directly to channel network)
11	lake_depth	0.1 - 1.5	m	Maximum allowable depth of lake
	Routing parameters			
12	Vl	0.5 - 3	m/s	Flow/Wave velocity
13	Df	200 - 4000	m^2/s	Flow diffusion

3.11. 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. Mon. Weather Rev. 141, 3462–3476. https://doi.org/10.1175/MWR-D-12-00206.1

Achleitner, S., Schöber, J., Rinderer, M., Leonhardt, G., Schöberl, F., Kirnbauer, R., Schönlaub, H., 2012. Analyzing the operational performance of the hydrological models in an alpine flood forecasting system. J. Hydrol. 412–413, 90–100. https://doi.org/10.1016/j.jhydrol.2011.07.047

Ajami, N.K., Duan, Q., Gao, X., Sorooshian, S., 2006. Multimodel combination techniques for analysis of hydrological simulations: application to distributed model intercomparison project results. J. Hydrometeorol. 7, 755–768. https://doi.org/10.1175/JHM519.1

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., Salamon, P., 2014. Evaluation of ensemble streamflow predictions in Europe. J. Hydrol. 517, 913–922. https://doi.org/10.1016/j.jhydrol.2014.06.035

Aliyari, F., Bailey, R.T., Tasdighi, A., Dozier, A., Arabi, M., Zeiler, K., 2019. Coupled SWAT-MODFLOW model for large-scale mixed agro-urban river basins. Environ. Model. Softw. 115, 200–210. https://doi.org/10.1016/j.envsoft.2019.02.014

Amengual, A., Homar, V., Romero, R., Alonso, S., Ramis, C., Amengual, A., Homar, V., Romero, R., Alonso, S., Ramis, C., 2012. A statistical adjustment of regional climate model outputs to local scales: application to Platja de Palma, Spain. J. Clim. 25, 939–957. https://doi.org/10.1175/JCLI-D-10-05024.1

Anderson, E., 2006. Snow accumulation and ablation model – SNOW-17, Nature. Silver Spring, MD. https://doi.org/10.1038/177563a0

Antonetti, M., Horat, C., Sideris, I. V, Zappa, M., 2018. Ensemble flood forecasting considering dominant runoff processes: I. Setup and application to nested basins (Emme, Switzerland). Nat. Hazards Earth Syst. Sci. Discuss. 5194, 1–29. https://doi.org/10.5194/nhess-2018-118

Armstrong, R.N., Pomeroy, J.W., Martz, L.W., 2008. Evaluation of three evaporation estimation methods in a Canadian prairie landscape Robert. Hydrol. Process. 22, 2801–2815. https://doi.org/https://doi.org/10.1002/hyp.7054

Atger, F., 1999. The Skill of ensemble prediction systems. Mon. Weather Rev. 127, 1941–1953. https://doi.org/10.1175/1520-0493(1999)127<1941:TSOEPS>2.0.CO;2

Bergström Sten, 1978. Development of a conceptual deterministic rainfall - runoff model. STOCKHOLM.

Bradley, A.A., Schwartz, S.S., 2011. Summary verification measures and their interpretation for ensemble forecasts. Mon. Weather Rev. 139, 3075–3089.

https://doi.org/10.1175/2010MWR3305.1

Brochero, D., Anctil, F., Gagné, C., 2011. Simplifying a hydrological ensemble prediction system with a backward greedy selection of members - Part 2: Generalization in time and space. Hydrol. Earth Syst. Sci. 15, 3327–3341. https://doi.org/10.5194/hess-15-3327-2011

Brown, J.D., Demargne, J., Seo, D.J., Liu, Y., 2010. The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environ. Model. Softw. 25, 854–872. https://doi.org/10.1016/j.envsoft.2010.01.009

Calvetti, L., Pereira Filho, A.J., 2014. Ensemble hydrometeorological forecasts using WRF hourly QPF and topmodel for a middle watershed. Adv. Meteorol. 2014. https://doi.org/10.1155/2014/484120

Chang, H.-L., Yang, S.-C., Yuan, H., Lin, P.-L., Liou, Y.-C., Chang, H.-L., Yang, S.-C., Yuan, H., Lin, P.-L., Liou, Y.-C., 2015. Analysis of the Relative Operating Characteristic and Economic Value Using the LAPS Ensemble Prediction System in Taiwan. Mon. Weather Rev. 143, 1833–1848. https://doi.org/10.1175/MWR-D-14-00189.1

Côté, J., Desmarais, J.G., Gravel, S., Méthot, A., Patoine, A., Roch, M., Staniforth, A., 1998. The operational CMC-MRB global environmental multiscale (GEM) model. Part II: Results. Mon. Weather Rev. 126, 1397–1418. https://doi.org/10.1175/1520-0493(1998)126<1397:TOCMGE>2.0.CO;2

Coulibaly, P., Anctil, F., Bobée, B., 2001. Multivariate Reservoir Inflow Forecasting Using Temporal Neural Networks. J. Hydrol. Eng. 6, 367–376. https://doi.org/10.1061/(ASCE)1084-0699(2001)6:5(367)

De Roo, A.P.J., Gouweleeuw, B., Thielen, J., Bartholmes, J., Bongioannini-Cerlini, P., Todini, E., Bates, P.D., Horritt, M., Hunter, N., Beven, K., Pappenberger, F., Heise, E., Rivin, G., Hils, M., Hollingsworth, A., Holst, B., Kwadijk, J., Reggiani, P., Dijk, M. Van, Sattler, K., Sprokkereef, E., 2003. Development of a European flood forecasting system. Int. J. River Basin Manag. 1, 49–59. https://doi.org/10.1080/15715124.2003.9635192

Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.J., Hartman, R., Herr, H.D., Fresch, M., Schaake, J., Zhu, Y., 2014. The science of NOAA's operational hydrologic ensemble forecast service. Bull. Am. Meteorol. Soc. 95, 79–98. https://doi.org/10.1175/BAMS-D-12-00081.1

Dibike, Y., Eum, H.-I., Prowse, T., 2018. Modelling the Athabasca watershed snow response to a changing climate. J. Hydrol. Reg. Stud. 15, 134–148. https://doi.org/10.1016/j.ejrh.2018.01.003

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., Wood, E.F., 2006. Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major

results from the second and third workshops. J. Hydrol. 320, 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031

Eum, H. Il, Dibike, Y., Prowse, T., 2017. Climate-induced alteration of hydrologic indicators in the Athabasca River Basin, Alberta, Canada. J. Hydrol. 544, 327–342. https://doi.org/10.1016/j.jhydrol.2016.11.034

Eum, H. Il, Dibike, Y., Prowse, T., 2014a. Uncertainty in modelling the hydrologic responses of a large watershed: A case study of the Athabasca River basin, Canada. Hydrol. Process. 28, 4272–4293. https://doi.org/10.1002/hyp.10230

Eum, H. Il, Dibike, Y., Prowse, T., Bonsal, B., 2014b. Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the Athabasca Watershed, Canada. Hydrol. Process. 28, 4250–4271. https://doi.org/10.1002/hyp.10236

Evenson, G.R., Golden, H.E., Lane, C.R., D'Amico, E., 2016. An improved representation of geographically isolated wetlands in a watershed-scale hydrologic model. Hydrol. Process. 30, 4168–4184. https://doi.org/10.1002/hyp.10930

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

Fan, F.M., Schwanenberg, D., Collischonn, W., Weerts, A., 2015. Verification of inflow into hydropower reservoirs using ensemble forecasts of the TIGGE database for large scale basins in Brazil. J. Hydrol. Reg. Stud. 4, 196–227. https://doi.org/10.1016/J.EJRH.2015.05.012

Fan, F.M., Schwanenberg, D., Kuwajima, J., Assis, A., 2014b. Ensemble streamflow predictions in the Três Marias basin , Brazil 16, 14191.

Fang, X., Minke, A., Pomeroy, J., Brown, T., Westbrook, C., Guo, X., Guangul, S., 2007. A Review of Canadian Prairie hydrology: principles, modelling and response to land use and drainage change, Centre for Hydrology Report #2, Version 2. Saskatoon, Saskatchewan.

Fang, X., Pomeroy, J.W., Westbrook, C.J., Guo, X., Minke, A.G., Brown, T., 2010. Prediction of snowmelt derived streamflow in a wetland dominated prairie basin. Hydrol. Earth Syst. Sci 14, 1–16. https://doi.org/10.5194/hess-14-1-2010

FAO, JRC, IIASA, ISRIC, ISS-CAS, 2009. Harmonized world soil database - Version 1.1. Rome, Italy and IIASA, Laxenburg, Austria.

Friedl, M.., Sulla-Menashe, D.., 2015. MCD12Q1 MODIS/Terra+Aqua land cover type yearly L3 Global 500m SIN Grid V006 [Data set]. https://doi.org/10.5067/MODIS/MCD12Q1.006

Gray, D.M., Landine, P.G., 1988. An energy-budget snowmelt model for the Canadian Prairies. Can. J. Earth Sci. 25, 1292–1303. https://doi.org/10.1139/e88-124

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

Haghnegahdar, A., Tolson, B.A., Davison, B., Seglenieks, F.R., Klyszejko, E., Soulis, E.D., Fortin, V., Matott, L.S., 2014. Calibrating Environment Canada's MESH modelling system over the Great Lakes basin. Atmosphere-Ocean 52, 281–293. https://doi.org/10.1080/07055900.2014.939131

Hamill, T.M., Bates, G.T., Whitaker, J.S., Murray, D.R., Fiorino, M., Galarneau, T.J., Zhu, Y., Lapenta, W., 2013. NOAA's second-generation global medium-range ensemble reforecast dataset. Bull. Am. Meteorol. Soc. 94, 1553–1565. https://doi.org/10.1175/BAMS-D-12-00014.1

Hamman, J., Nijssen, B., Clark, E., Matthews, D., Veerman, B., 2017. UW-Hydro/RVIC: RVIC 1.1.1. https://doi.org/10.5281/ZENODO.269614

Han, S., Coulibaly, P., 2019. Probabilistic flood forecasting using hydrologic uncertainty processor with ensemble weather forecasts. J. Hydrometeorol. JHM-D-18-0251.1. https://doi.org/10.1175/JHM-D-18-0251.1

Han, S., Coulibaly, P., Biondi, D., 2019. Assessing hydrologic uncertainty processor performance for flood forecasting in a semiurban watershed. J. Hydrol. Eng.

Hayashi, M., Van Der Kamp, G., 2000. Simple equations to represent the volume-area-depth relations of shallow wetlands in small topographic depressions. J. Hydrol. 237, 74–85. https://doi.org/https://doi.org/10.1016/S0022-1694(00)00300-0

Hedstrom, N.R., Granger, R.J., Pomeroy, J.W., Gray, D.M., 2001. Enhanced indicators of land use change and climate variability impacts on Prairie hydrology using the cold regions hydrological model. 58th East. Snow Conf. 262–275.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast. 15, 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hopkinson, R.F., Mckenney, D.W., Milewska, E.J., Hutchinson, M.F., Papadopol, P., Vincent, A.L.A., 2011. Impact of aligning climatological day on gridding daily maximum-minimum temperature and precipitation over Canada. J. Appl. Meteorol. Climatol. 50, 1654–1665. https://doi.org/10.1175/2011JAMC2684.1

Hrachowitz, M., Clark, M.P., 2017. HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. Hydrol. Earth Syst. Sci. 21, 3953–3973. https://doi.org/10.5194/hess-21-3953-2017

Hsu, W. ron, Murphy, A.H., 1986. The attributes diagram A geometrical framework for assessing the quality of probability forecasts. Int. J. Forecast. 2, 285–293. https://doi.org/10.1016/0169-2070(86)90048-8

Hutchinson, M.F., McKenney, D.W., Lawrence, K., Pedlar, J.H., Hopkinson, R.F., Milewska,

E., Papadopol, P., 2009. Development and testing of Canada-wide interpolated spatial models of daily minimum-maximum temperature and precipitation for 1961-2003. J. Appl. Meteorol. Climatol. 48, 725–741. https://doi.org/10.1175/2008JAMC1979.1

Ilampooranan, I., Van Meter, K.J., Basu, N.B., 2019. A race against time: modeling time lags in watershed response. Water Resour. Res. 55, 3941–3959. https://doi.org/10.1029/2018WR023815

Jasper, K., Gurtz, J., Lang, H., 2002. Advanced flood forecasting in Alpine watersheds by coupling meteorological observations and forecasts with a distributed hydrological model. J. Hydrol. 267, 40–52. https://doi.org/10.1016/S0022-1694(02)00138-5

Koren, V., Reed, S., Smith, M., Zhang, Z., Seo, D.-J., 2004. Hydrology laboratory research modeling system (HL-RMS) of the US national weather service. J. Hydrol. 291, 297–318. https://doi.org/10.1016/j.jhydrol.2003.12.039

Kouwen, N., 1988. WATFLOOD: a micro-computer based flood forecasting system based on real-time weather radar. Can. Water Resour. J. 13, 62–77. https://doi.org/10.1007/s10228-005-0319-x

Krogh, S.A., Pomeroy, J.W., 2019. Impact of future climate and vegetation on the hydrology of an Arctic headwater basin at the tundra-taiga transition. J. Hydrometeorol. 20, 197–215. https://doi.org/10.1175/JHM-D-18-0187.1

Lahmers, T.M., Gupta, H., Castro, C.L., Gochis, D.J., Yates, D., Dugger, A., Goodrich, D., Hazenberg, P. the structure of the W. hydrologic model for semiarid environments, 2019. Enhancing the structure of the WRF-hydro hydrologic model for semiarid environments. J. Hydrometeorol. 20, 691–714. https://doi.org/10.1175/JHM-D-18-0064.1

Leach, J.M., Kornelsen, K.C., Coulibaly, P., 2018. Assimilation of near-real time data products into models of an urban basin. J. Hydrol. 563, 51–64. https://doi.org/10.1016/J.JHYDROL.2018.05.064

Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J. Geophys. Res. 99, 14415. https://doi.org/10.1029/94JD00483

Liechti, K., Zappa, M., Fundel, F., Germann, U., 2013. Probabilistic evaluation of ensemble discharge nowcasts in two nested Alpine basins prone to flash floods. Hydrol. Process. 27, 5–17. https://doi.org/10.1002/hyp.9458

Lohmann, D., Nolte-Holube, R., Raschke, E., 1996. A large-scale horizontal routing model to be coupled to land surface parametrization schemes. Tellus, Ser. A Dyn. Meteorol. Oceanogr. https://doi.org/10.3402/tellusa.v48i5.12200

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

Mason, S.J., Graham, N.E., 2002. Areas beneath the relative operating characteristics (ROC)

and relative operating levels (ROL) curves: Statistical significance and interpretation. Q. J. R. Meteorol. Soc. 128, 2145–2166. https://doi.org/10.1256/003590002320603584

Mason, S.J., Graham, N.E., 1999. Conditional probabilities, relative operating characteristics, and relative operating levels. Weather Forecast. 14, 713–725. https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2

Maurer, E.P., Brekke, L.D., Pruitt, T., 2010. Contrasting lumped and distributed hydrology models for estimating climate change impacts on California watersheds. J. Am. Water Resour. Assoc. 46, 1024–1035. https://doi.org/10.1111/j.1752-1688.2010.00473.x

Maxey, R., Cranston, M., Tavendale, A., Buchanan, P., 2012. The use of deterministic and probabilistic forecasting in countrywide flood guidance in Scotland. Hydrol. a Chang. world 01–07. https://doi.org/10.7558/bhs.2012.ns33

Mekonnen, M.A., Wheater, H.S., Ireson, A.M., Spence, C., Davison, B., Pietroniro, A., 2014. Towards an improved land surface scheme for prairie landscapes. J. Hydrol. 511, 105–116. https://doi.org/10.1016/j.jhydrol.2014.01.020

Muhammad, A., Stadnyk, T., Unduche, F., Coulibaly, P., 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, 1604. https://doi.org/10.3390/w10111604

Nasonova, O.N., Gusev, Y.M., 2007. Can a land surface model simulate runoff with the same accuracy as a hydrological model?, in: IAHS-AISH Publication. pp. 258–265.

Newman, A.J., Mizukami, N., Clark, M.P., Wood, A.W., Nijssen, B., Nearing, G., Newman, A.J., Mizukami, N., Clark, M.P., Wood, A.W., Nijssen, B., Nearing, G., 2017. Benchmarking of a physically based hydrologic model. J. Hydrometeorol. 18, 2215–2225. https://doi.org/10.1175/JHM-D-16-0284.1

NWS, 2002. Conceptualization of the sacramento soil moisture accounting model introduction.

Pappenberger, F., Bartholmes, J., Thielen, J., Cloke, H.L., Buizza, R., de Roo, A., 2008. New dimensions in early flood warning across the globe using grand-ensemble weather predictions. Geophys. Res. Lett. 35, L10404. https://doi.org/10.1029/2008GL033837

Pappenberger, F., Ramos, M.H., Cloke, H.L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., Salamon, P., 2015. How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. J. Hydrol. 522, 697–713.

Pattison-Williams, J.K., Pomeroy, J.W., Badiou, P., Gabor, S., 2018. Analysis wetlands, flood control and ecosystem services in the smith creek drainage basin: A Case Study in Saskatchewan, Canada. Ecol. Econ. J. 147, 36–47. https://doi.org/10.1016/j.ecolecon.2017.12.026

Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., Verseghy, D., Soulis, E.D., Caldwell, R., Evora, N., Pellerin, P., 2007. Development of the MESH modelling

system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale. Hydrol. Earth Syst. Sci. 11, 1279–1294. https://doi.org/10.5194/hessd-3-2473-2006

Pomeroy, J.W., Gray, D.M., Brown, T., Hedstrom, N.R., Quinton, W.L., Granger, R.J., Carey, S.K., 2007. The cold regions hydrological model: A platform for basing process representation and model structure on physical evidence. Hydrol. Process. 21, 2650–2667. https://doi.org/10.1002/hyp.6787

Razavi, T., Coulibaly, P., 2017. An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds. Can. Water Resour. Journal// Rev. Can. des ressources hydriques 42, 2–20. https://doi.org/10.1080/07011784.2016.1184590

Razavi, T., Coulibaly, P., 2016. Improving streamflow estimation in ungauged basins using a multi-modelling approach. Hydrol. Sci. 61, 2668–2679. https://doi.org/10.1080/02626667.2016.1154558

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.J., 2004. Overall distributed model intercomparison project results. J. Hydrol. 298, 27–60. https://doi.org/10.1016/j.jhydrol.2004.03.031

Rokaya, P., Wheater, H., Lindenschmidt, K.E., 2019. Promoting sustainable Ice-Jam flood management along the peace river and Peace-Athabasca Delta. J. Water Resour. Plan. Manag. 145, 1–12. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001021

Samuel, J., Coulibaly, P., Metcalfe, R.A., 2011. Estimation of continuous streamflow in Ontario ungauged basins: comparison of regionalization methods. J. Hydrol. Eng. 16, 447–459. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000338

Seiller, G., Anctil, F., Perrin, C., 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrol. Earth Syst. Sci. 16, 1171–1189. https://doi.org/10.5194/hess-16-1171-2012

Seiller, G., Anctil, F., Roy, R., 2017. Design and experimentation of an empirical multistructure framework for accurate, sharp and reliable hydrological ensembles. J. Hydrol. 552, 313–340. https://doi.org/10.1016/j.jhydrol.2017.07.002

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 Resour. Res. 55, 1510–1530. https://doi.org/10.1029/2018WR023197

Shook, K., Pomeroy, J., van der Kamp, G., 2015. The transformation of frequency distributions of winter precipitation to spring streamflow probabilities in cold regions; case studies from the Canadian Prairies. J. Hydrol. 521, 394–409. https://doi.org/10.1016/j.jhydrol.2014.12.014

Shook, K., Pomeroy, J.W., Spence, C., Boychuk, L., 2013. Storage dynamics simulations in prairie wetland hydrology models: evaluation and parameterization.

https://doi.org/10.1002/hyp.9867

Shrestha, R.R., Dibike, Y.B., Prowse, T.D., 2012. Modelling of climate-induced hydrologic changes in the Lake Winnipeg watershed. J. Great Lakes Res. 38, 83–94. https://doi.org/10.1016/j.jglr.2011.02.004

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. J. Hydrol. 298, 4–26. https://doi.org/10.1016/j.jhydrol.2004.03.040

Thiboult, A., Anctil, F., Boucher, M.A., 2016. Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrol. Earth Syst. Sci. 20, 1809–1825. https://doi.org/10.5194/hess-20-1809-2016

Thiemig, V., Pappenberger, F., Thielen, J., Gadain, H., de Roo, A., Bodis, K., Del Medico, M., Muthusi, F., 2010. Ensemble flood forecasting in Africa: A feasibility study in the Juba-Shabelle river basin. Atmos. Sci. Lett. 11, 123–131. https://doi.org/10.1002/asl.266

Tolson, B.A., Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resour. Res. 43, 1–16. https://doi.org/10.1029/2005WR004723

Unduche, F., Tolossa, H., Senbeta, D., Zhu, E., 2018. Evaluation of four hydrological models for operational flood forecasting in a Canadian Prairie watershed. Hydrol. Sci. J. 63, 1–17. https://doi.org/10.1080/02626667.2018.1474219

Velázquez, J.A., Anctil, F., Perrin, C., 2010. Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments. Hydrol. Earth Syst. Sci. 14, 2303–2317. https://doi.org/10.5194/hess-14-2303-2010

Velázquez, J.A., Anctil, F., Ramos, M.H., Perrin, C., 2011. Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. Adv. Geosci. 29, 33–42. https://doi.org/10.5194/adgeo-29-33-2011

Viney, N.R., Bormann, H., Breuer, L., Bronstert, A., Croke, B.F.W., Frede, H., Gräff, T., Hubrechts, L., Huisman, J.A., Jakeman, A.J., Kite, G.W., Lanini, J., Leavesley, G., Lettenmaier, D.P., Lindström, G., Seibert, J., Sivapalan, M., Willems, P., 2009. Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions. Adv. Water Resour. 32, 147–158. https://doi.org/10.1016/j.advwatres.2008.05.006

Wilks, D., 2006. Statistical Methods in the Atmospheric Sciences, 2nd ed, International Geophysics Series. Acadamic Press.

Yapo, P.O., Gupta, H.V., Sorooshian, S., 1996. Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. J. Hydrol. 181, 23–48. https://doi.org/10.1016/0022-1694(95)02918-4

Zsótér, E., Pappenberger, F., Smith, P., Emerton, R.E., Dutra, E., Wetterhall, F., Richardson, D., Bogner, K., Balsamo, G., 2016. Building a Multimodel Flood Prediction System with the TIGGE Archive. J. Hydrometeorol. 17, 2923–2940. https://doi.org/10.1175/JHM-D-15-0130.1

Chapter 4. Identification of combined hydrological models and numerical weather predictions for enhanced flood forecasting in a semi-urban watershed

Summary of Paper 3: Awol, F.S., Coulibaly, P., and Tsanis, I. (2019). Identification of combined hydrological models and numerical weather predictions for enhanced flood forecasting in a semi-urban watershed. *Journal of Hydrometeorology*, Under Review.

In this research, twelve hydrological models (lumped and distributed) were first set-up and calibrated using historical data. Then models with an improved calibration performance were selected for forecast verification. In the verification, four medium- and high-resolution NWP inputs were used to force selected hydrological models in an attempt to identify the best combination of models and inputs to improve flood forecasting in a semi-urban catchment.

Key findings of this research include:

- The lumped model (MACHBV) combined with a high-resolution forecast input (HRDPS) provides improved accuracy, economic value, and overall skill of short-range flood forecasts (1hr-18hr) than any other model-input integration.
- Distributed models were only competent at forecasting floods in the later hours of the day (between 15hr-18hr lead times).
- There is steady persistence in flood forecasting as the top-ranking models in the very recent history will highly likely continue to perform well in the near future.

4.1. Abstract

This research aims at identifying a suitable combination of hydrological models and skillful weather predictions for enhanced short-term flood forecasting in a semi-urban watershed using several performance evaluation metrics. Twelve hydrological models were set-up and calibrated out of which five models comprised of lumped (SACSMA, MACHBV & PDM, all coupled with SNOW17) and distributed (WATFLOOD & SWMM) models were selected for forecast verification. Deterministic precipitation forecasts from High-Resolution Deterministic Precipitation System (HRDPS), High-Resolution Rapid Refresh (HRRR), North American Mesoscale Forecast System (NAM), and Rapid Refresh (RAP) were collected. Hydrological forecasts from a combination of five models and four forecast inputs were verified for 1hr to 18hr lead times in a 6-month hindcast period. Pre-screening analysis indicated that whatever the hydrological model, HRRR and HRDPS produced better hydrological forecast accuracy than NAM and RAP. A comprehensive verification revealed that MACHBV followed by SACSMA models showed better overall forecast skill and accuracy. Distributed models were only competent between 15-18h lead times. Overall, MACHBV with HRDPS has emerged as the best model-input combination. It captured the peak flow magnitude and timing, detected the flood threshold, and appeared economically viable at all forecast lead times better than any other model-input combination. Results also showed that giving adaptive weights to hydrological models based on recent performances provided enhanced combined forecasts while persistently keeping the well-performing models.

4.2. Introduction

Floods are one of the deadliest natural disasters in many regions of the world. In urban and semi-urban areas with high population density, flooding has caused the loss of many lives, damages to infrastructures, evacuations and temporary homelessness, and large insurance and disaster relief spending. In Canada, 2013 was recorded as a catastrophic year as floods affected half a million households and caused about $7 billion in damages in the City of Calgary and the Greater Toronto Area (ECCC, 2017; Sandink, 2016). Recently, the 2019 spring flooding has affected thousands of homes in parts of Quebec, Ontario, and New Brunswick, including Montreal and the capital city Ottawa (FloodList, 2019; Montreal Gazette, 2019; Statistics Canada, 2019).

A principal way of reducing flood damages in affected areas is by forecasting river flooding well ahead of time as part of an early warning system. River flow forecasting can be achieved by using historical stochastic analysis (Chen et al., 2019; Chow et al., 1983; Georgakakos, 1986; Lardet and Obled, 1994; Lindenschmidt et al., 2019), artificial neural networks (Campolo et al., 2003; Chang et al., 2014; Coulibaly et al., 2001b, 2001a, 2000; Jeong and Kim, 2005; Thirumalaiah and Deo, 1998) or hydrological and hydraulic models (Beven et al., 1984; Chen et al., 2009; Du et al., 2012; Gouweleeuw et al., 2005; Jasper et al., 2002; Muhammad et al., 2018; Vieux et al., 2004; Yucel et al., 2015). However, the latter have become a popular choice for many provincial flood forecasting centers, and national forecasting institutions (Achleitner et al., 2012; De Roo et al., 2003; Emerton et al., 2016; Hopson and Webster, 2010; Pappenberger et al., 2008; Thielen et al., 2009; Unduche et al., 2018).

Various model evaluation techniques should test the ability of hydrological models to reproduce and forecast peak flow events before application in operational flood forecasting. Researches have been conducted to enhance the prediction skill of hydrological models. Data assimilation, post-processing, and uncertainty quantification methods were some of the areas of the studies for improving the forecast skill in urban and semi-urban environments (Han et al., 2019; Han and Coulibaly, 2019; Leach et al., 2018). A diverse multi-model approach was also recommended to enhance streamflow forecasts (Hrachowitz and Clark, 2017). Leach et al., (2018), analyzed the added benefit of assimilating near real-time data (streamflow, soil moisture, and snow water equivalent) into GR4J, HYMOD, MACHBV, and SACSMA models using Ensemble Kalman Filter and suggested that the combined assimilation provides better model prediction in an urban watershed. Han et al., (2019), investigated the application of Precipitation-Dependent Hydrologic Uncertainty Processor (HUP) using HYMOD and GR4H models in a semi-urban watershed to assess hydrological uncertainty and enhance the overall quality of the deterministic forecast. Han and Coulibaly, (2019), also examine the use of Bayesian ensemble uncertainty processer (BEUP) to generate probabilistic flood forecasts from MACHBV model forced by ensemble weather forecast inputs in the same watershed. The authors conclude that BEUP post-processor can capture the main predictive uncertainties and enhance flood forecasting skills. Sharma et al., (2019), recently investigate the benefit of using diverse hydrological models to improve medium and long-term streamflow forecasts using API-C, HL-RDHM, and WRF-Hydro models and innovative post-processing methods.

Hrachowitz and Clark, (2017), suggested the implementation of diverse hydrological models at the required level of details by taking advantage of multi-scale data inputs for the intended purpose, flood forecasting, for instance. Comparison and performance evaluation of hydrological models with various calibration and validation approaches is vital for application in urban and semi-urban flood forecasting. Awol et al., 2018, for example, compares traditional and advanced calibration approaches to select appropriate SWMM model parameters in a semi-urban watershed. They highlighted the importance of using a weighted average multi-objective optimization approach using the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007) to improve flood prediction in multiple gauging sites. El Hassan et al., (2013), compared the performances of HECHMS and GSSHA models in selected historical flood events to assess models' prediction ability for a semi-urbanized watershed. Multi-model techniques have also been used to assess their added value in the operational flood forecasting context. Thiboult et al., 2017, for instance, compared several Early Warning Systems (EWS) by investigating different sources of uncertainties from the integration of multiple hydrological models and data assimilation techniques. They used the HOOPLA (HydrOlOgical Prediction Laboratory) framework (Thiboult et al., 2019), which comprises of twenty lumped conceptual hydrological models with many data assimilation scenarios. Similar tools were also used by Thiboult et al., (2016) for assessing sources of uncertainty in ensemble streamflow forecasting.

The use of meteorological forecasts from numerical weather prediction models (NWP), either deterministic or ensemble, as a driving force to hydrological models is beneficial in

providing short-term forecasts and transferring the inherited uncertainty of initial conditions (Cloke and Pappenberger, 2009). Due to fast responses to extreme events, shorter time of concentration, and flashiness of floods in urban and semi-urban watersheds, short-term weather forecast data is required for forcing hydrological models. The performance of the subsequent short-term hydrological forecast can then be evaluated for reliability, skill, and overall forecast quality. The quality of hydrological forecasts depends on the availability and skill of hourly and sub-hourly weather forecasts, particularly for urban and semi-urban catchments. Bennett et al., (2014) indicated that hydrological forecasts with hourly time steps have better accuracy than daily forecasts; and provides more information on the flood hydrographs. As important as the temporal resolution, the higher spatial resolution of weather forecasts is crucial for improved short-term flood forecasting. Abaza et al., (2013), for example, compared available regional and global ensembles, and deterministic meteorological forecasts in Canadian catchments for short-term hydrological forecasting. They set up HYDROTEL (Fortin et al., 2001) hydrological model in catchments ranging between 355 and 5820 km^2 areas. The authors found that higher resolution regional meteorological forecasts provided better hydrological forecast quality and reliability than their global counterparts. They also highlighted that deterministic forecasts were as good as the ensemble ones up to 24-hour forecast lead time and the effect of ensemble members is only recognized at later lead times.

Focusing on flash-flood prediction, Horat et al., (2018), assessed forecasting chains consisting of European-based deterministic and ensemble NWPs, radar-based real-time rainfall input, and an advanced and traditional operational setup of PREVAH hydrological

model. Their advanced hydrological model setup was based on Dominant Runoff Process (DRP) and *a priori* parameter estimation with no calibration requirement. They reported that the new forecasting chain could produce competitive and better results than the traditional approach without the need for an extensive calibration process, desirable in ungauged catchments. Their results also showed that deterministic forecasts produce better skills for short lead-times up to 24-hours as compared to ensemble forecasts.

Based on Abaza et al., (2013), and Horat et al., (2018), regional deterministic weather forecasts are deemed to be superior to ensembles for very short-term flood forecasting, particularly in small urban and semi-urban catchments that have only a few hours of response times. Hence, it is crucial to use not only high spatial resolution but also high temporal resolution NWPs with sub-hourly, hourly, or sub-daily time-steps for hydrological forecasting in a flashy urban and semi-urban catchment.

Most of the above studies applied multiple hydrological models or multiple NWPs for seasonal (monthly), medium-term (up to 15 days), and daily short-term (1 day up to a week ahead) hydrological forecasts. Alternatively, in some cases, a single hydrological model was used with sub-daily or hourly forecast lead times. Nonetheless, few studies have focused on integrated multiple hydrological models and multiple NWPs in semi-urban and urban watersheds for short-term flood forecasting purposes. As such, a comprehensive investigation of the combined multi-models and multi-inputs for short-term flood forecasting is one of the research gaps that is addressed in this study.

The primary objective of the study is to identify a proper combination of hydrological models and skillful weather forecast inputs for enhanced short-term flood forecasting in a

semi-urban watershed. The research will evaluate candidate hydrological models and weather forecast inputs using different existing forecast verification metrics based on the overall forecast skill, quality, peak flow magnitude, peak flow timing, reliability, and economic value. Besides, this study will assess the added benefit of using precipitation forecasts from four regional NWPs for forcing hydrological models. The research contributes to the verification of NWPs for short-term flood forecasting and the identification of potential hydrological models for semi-urban catchments, and thus provides findings to operational flood forecasters, and to researchers for further investigation to improve forecasting tool and products.

The paper is structured as follows. Section 4.3 will provide information on the study area, the different kinds of data used, and Section 4.4 presents the methodology and details of evaluation tools applied. In Section 4.5 the results will be presented and discussed. Finally, conclusions are provided in Section 4.6.

4.3. Study Area and Data

4.3.1. Study Area

The study area, called Humber River Watershed, is found in Ontario, Canada, and has a catchment area of about 911 km2 (Figure 4-1). The catchment is characterized as a semi-urban as it is covered by 13% urbanizing, 33% urban, and 54% rural area (TRCA, 2013). The study area is selected for this research because related studies have been conducted recently to improve the flood prediction skill (Awol et al., 2018; Han et al., 2019; Han and Coulibaly, 2019) and supplementary data were readily available.

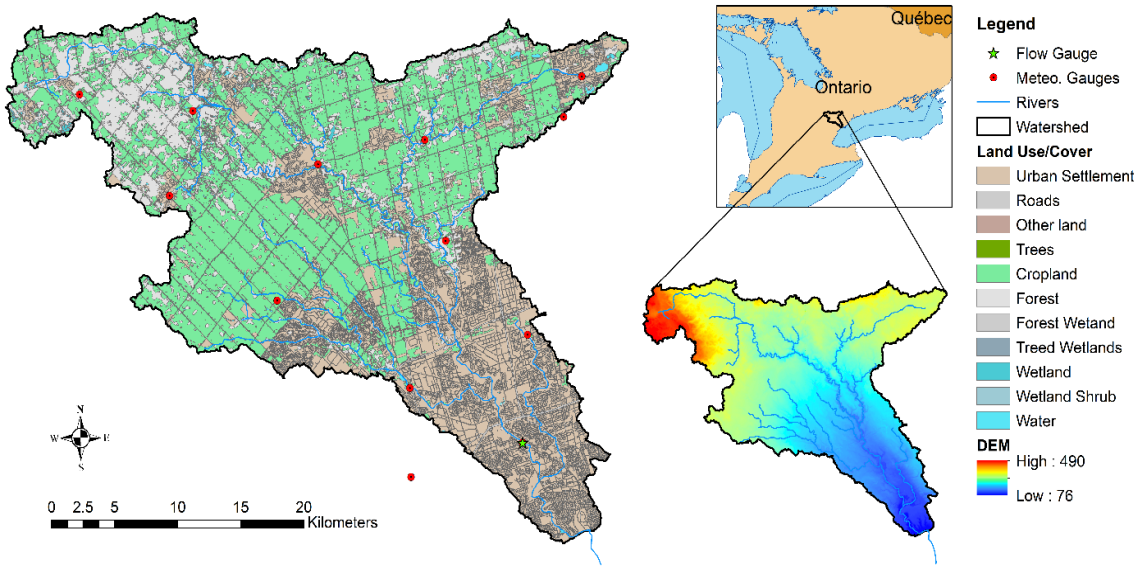

Figure 4-1: Study Area of the Humber River Watershed

4.3.2. Data

4.3.2.1. Observed meteorological and river discharge data

Historical precipitation and temperature data for the meteorological gauges, and river flow data for the outlet flow station (shown in Figure 4-1) were received from Environment Canada and Toronto Regions Conservation Authority (TRCA). Hourly data were prepared to calibrate and validate various hydrological models, which will be discussed in detail in Section 4.4.2.

4.3.2.2. Weather forecast data

For this study, regional deterministic weather forecasts inputs were chosen because studies indicated that deterministic forecasts provide better hydrological prediction skill up to 24hr forecast lead time than global and ensemble forecasts (Abaza et al., 2013; Horat et al.,

2018). Hence data were collected from two providers: Environment Canada and the NOAA. Environment Canada provided the High-Resolution Deterministic Precipitation System (HRDPS). Whereas NOAA supplied North American Mesoscale Forecast System (NAM), Rapid Refresh (RAP), and High-Resolution Rapid Refresh (HRRR) model. In terms of spatial resolution, HRDPS and HRRR (2.5 and 3 km respectively) are finer than NAM and RAP (12 and 13 km, respectively). The four deterministic products (Table 4-1) have hourly time steps but have different forecast horizons. Archives of these forecasts data were collected between June 1st, 2018 to November 30th, 2018.

Furthermore, HRRR and RAP issues forecast every hour within a day, but NAM and HRDPS are only available four times a day. For this research, the precipitation forecast variable was only used in a hindcast experiment because researches indicated that it is the most significant factor for flood forecasting (Cuo et al., 2011; Zsótér et al., 2016).

Table 4-1: Weather forecast products and details (Note: Forecasts of HRRR and RAP are available at any hour of a day. For this study four valid times are selected to compare them to the other two products)

	Valid time	Forecast lead-times
HRRR ***3km***	00Z	1h,2h,3h, …,18h
	06Z	1h,2h,3h, …,18h
	12Z	1h,2h,3h, …,18h
	18Z	1h,2h,3h, …,18h
RAP ***13km***	00Z	1h,2h,3h, …, …,21h
	06Z	1h,2h,3h, …, …,21h
	12Z	1h,2h,3h, …, …,21h
	18Z	1h,2h,3h, …, …,21h
HRDPS ***2.5km***	00Z	1h,2h,3h, …, …, …, …,48h
	06Z	1h,2h,3h, …, …, …, …,48h
	12Z	1h,2h,3h, …, …, …, …,48h
	18Z	1h,2h,3h, …, …, …, …,48h
NAM ***12km***	00Z	1h,2h,3h, …,36h,39h,41h, …,84h
	06Z	1h,2h,3h, …,36h,39h,41h, …,84h
	12Z	1h,2h,3h, …,36h,39h,41h, …,84h
	18Z	1h,2h,3h, …,36h,39h,41h, …,84h

4.4. Methodology

Figure 4-2 summarizes the methodology employed in this research. First, multiple hydrological models that are presumed to be appropriate for short-term river flow forecasting are collected and set-up in the semi-urban catchment. The hydrological models were calibrated and validated using observed hourly data. Then the best models that showed better performances in terms of overall accuracy and peak flow simulation were selected for the next step. Since the next stage involves verifying the models in forecast mode, weather forecast data are collected to feed into the screened multiple hydrological models.

The available regional deterministic weather forecast data were pre-verified using some forecast verification metrics in order to filter out those which produce poor hydrological forecast skills. This pre-verification stage is to ensure that the best deterministic products that are appropriate for semi-urban watersheds are selected before applying a full forecast verification. Comprehensive forecast verification is then performed by running the screened hydrological models using selected multiple inputs as a driving force in the hindcast period. This stage is to find the best hydrological models and input combinations that are adequate for semi-urban flood forecasting. It involves identifying competent hydrological models based on the overall forecast skill, quality, peak flow magnitude, peak flow timing, bias, reliability, and economic value aspects. Finally, the best combination of hydrological models and forecast products are selected. As a supplement, simple forecast averaging methods were applied in order to find a better forecast combination method, which will be discussed in Section 4.4.5.

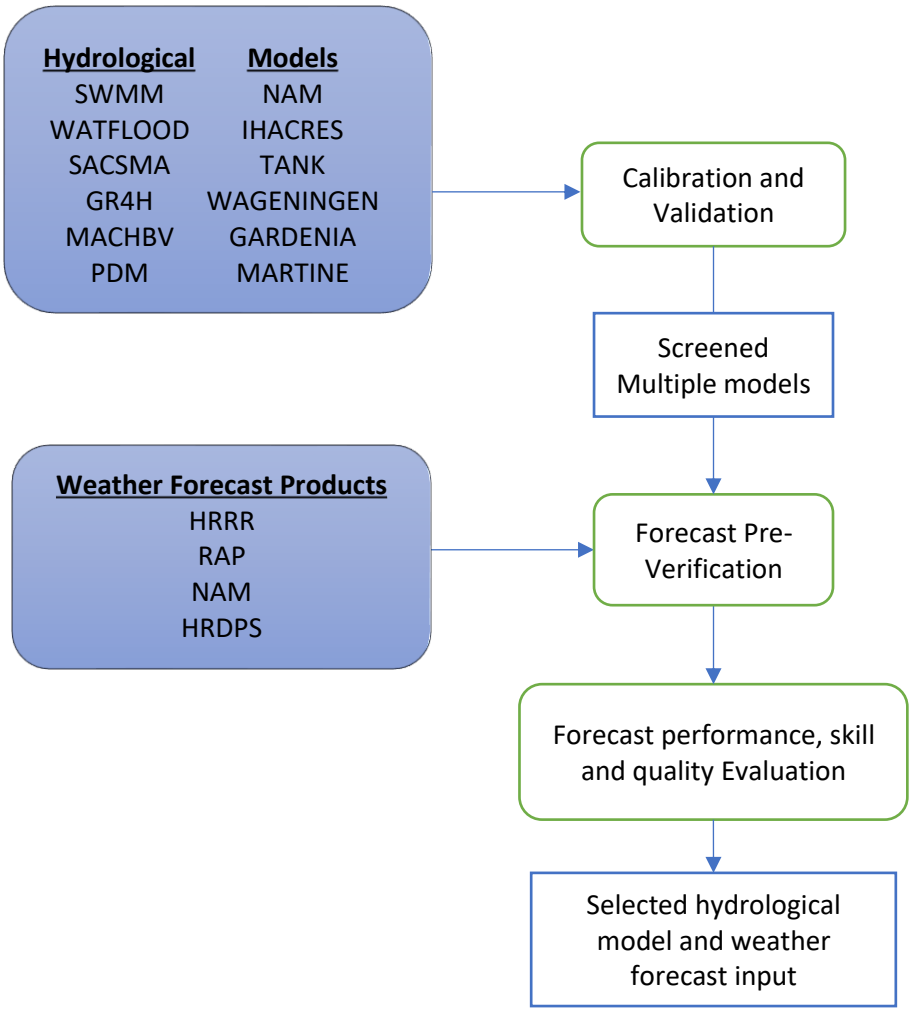

Figure 4-2: Methodology adopted for evaluating and selecting hydrological models and high-temporal weather forecast inputs.

4.4.1. Hydrological models

An attempt was made to consider potential hydrological models that are relevant for flood forecasting purposes and are locally adapted and applied for research and practical purposes. The list of models used in this study is provided in Figure 4-2 and will be discussed here.

SWMM (Storm Water Management Model) is a popular semi-distributed model developed by Environmental Protection Agency (EPA) for urban and semi-urban watersheds (Huber and Dickinson, 1988; Rossman and Huber., 2015). Previous related work has been done in the Humber River watershed using the SWMM model to investigate the best calibration approaches and select representative parameters for enhanced peak flow simulations (Awol et al., 2018). Hence, the previous model set-up is used in this study.

WATFLOOD (Kouwen, 1988) is a Canadian hydrological and routing model developed at the University of Waterloo and is specifically intended for flood forecasting, and watershed simulation which combines a conceptual Group Response Unit (GRU) distributed hydrology and a physically-based routing component (Kouwen et al., 1993). The model has been applied for flood forecasting purposes in operational and research centers (Muhammad et al., 2018; Unduche et al., 2018), and set-up was available for Humber River Watershed and the other Toronto Regions Conservation Authority (TRCA) catchments (Kouwen, 2018). For the rest of this paper, SWMM and WATFLOOD are categorized as distributed models even though the former is a semi-distributed model.

A conceptual lumped hydrological model called MACHBV (McMaster University Hydrologiska Byråns Vattenbalansavdelning) (Samuel et al., 2011) was initially developed by modifying the HBV model (Bergström Sten, 1978). The model has been applied in a similar watershed for flood forecasting studies (Han et al., 2019; Han and Coulibaly, 2019) and elsewhere tested in Canada (Razavi and Coulibaly, 2017, 2016). In this research, MACHBV is coupled with SNOW17 (Anderson, 2006) module.

The Sacramento Soil Moisture Accounting (SACSMA) model developed and used by NOAA's National Weather Service (NWS) in river forecasting centers is a conceptual, spatially lumped hydrological model that has been applied in several research projects (Burnash et al., 1973; Demargne et al., 2014; Seiller et al., 2015; Shamir et al., 2006; Velázquez et al., 2010; Vrugt et al., 2006). The model was also implemented in nearby urban and semi-urban watersheds to enhance flood forecasting (Dumedah and Coulibaly, 2013; Leach et al., 2018). In this research, SACSMA is coupled with SNOW17 module.

The rest of the hydrological models were imported from the HOOPLA (HydrOlOgical Prediction LAboratory) framework (Thiboult et al., 2019). The framework comprises of multiple lumped conceptual hydrological models, data assimilation modules and meteorological forecasting systems that allow users the option of combining different module and forecast setups for streamflow prediction studies (Thiboult et al., 2017, 2016). Out of the 20 lumped hydrological models that the authors applied in Canadian catchments, eight were selected for this study based on their overall performances from the mentioned previous studies. The models used here are GR4J (Perrin et al., 2003), NAM (Nielsen and Hansen, 1973), IHACRES (Jakeman et al., 1990), PDM (Moore and Clarke, 1981), MARTINE (Mazenc et al., 1984), GARDENIA (Thiéry, 1982), TANK (SUGAWARA, 1979) and WAGENINGEN (Warmerdam and Kole, 1997). Modifications were made to these models in order to fit into the calibration strategy employed in this research (see Section 4.4.2) in addition to coupling all the models with SNOW17. Also, the hourly GR4H model (Bennett et al., 2014) is used instead of GR4J, the original daily version, because all

other models are calibrated and simulated with hourly meteorological observed and forecast data.

4.4.2. Calibration and Validation

Apart from SWMM and WATFLOOD, all the other hydrological models described above were calibrated in the study. As mentioned in Section 4.4.2, the former two models were already calibrated in previous studies. However, the calibration of WATFLOOD was based on an unsatisfactory land cover map (Kouwen, 2019, Personal Communication). The available land cover map (Figure 4-1) was considered as a zoning map, and given how crucial impervious area is in urban and semi-urban runoff modeling, the mapping of the impervious areas as GRU's in each grid could only be guessed at. SWMM model was calibrated based on a heavily discretized 714 sub-catchments within Humber River Watershed. An advanced multi-site calibration approach was required to simulate flood events at multiple interior and outlet gauging stations (Awol et al., 2018).

The Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007) was used to calibrate all the models. Hourly historical precipitation and temperature data were used to run the models and calibrate with hourly observed river flow data. Hence, calibration and validation were performed in an hourly time step from Jan 2014 to Dec 2017 and from Jan 2010 to Dec 2013, respectively. The reason why we used recent years for calibration is that we want to train the models for relatively higher consecutive flood years (comparing to the validation period). A weighted average of three objective functions (Eqn. 4-1) was used to optimize the parameters of each model using a single-objective DDS optimization. That is, calibration was performed to maximize the average of the objective

functions. These are the Peak Flow Criteria (PFC) (Coulibaly et al., 2001a), Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970), and Kling–Gupta efficiency (KGE) (Gupta et al., 2009) which are formulated below.

$$PFC = \frac{\left(\sum_{i=1}^{n_p}\left(\left(q_{s,i} - q_{o,i}\right)^2 {q_{o,i}}^2\right)\right)^{\frac{1}{4}}}{\left(\sum {q_{o,i}}^2\right)^{\frac{1}{2}}}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}\left(Q_{o,i} - Q_{s,i}\right)^2}{\sum_{i=1}^{N}\left(Q_{o,i} - \overline{Q_o}\right)^2} \quad (4\text{-}1)$$

$$KGE = 1 - \sqrt{(r-1)^2 + (a-1)^2 + (b-1)^2}$$

Where Q_s, Q_o, q_s and q_o are simulated flow, observed flow, simulated peak flow, and observed peak flow, respectively. r, a, and b are the correlation coefficient between, the ratio of standard deviations of, and the ratio of the mean of Q_s and Q_o respectively. n_p symbolizes the number of counts where peak flows are above $1/3$ of the mean observed peak flow. NSE and KGE values closer to 1, and PFC value of 0 indicate better model performance and best peak flow simulation accuracy.

4.4.2.1. Remarks on Lumped and Distributed models

- The lumped hydrological models have one spatially enclosed catchment at the outlet gauging station. In the distributed models the study area was discretized by several sub-catchments (SWMM) or grids and Group Response Units (WATFLOOD).
- Calibration of lumped models is usually simple and requires less effort to optimize a limited number of parameters at the outlet gauging stations. Whereas in distributed models, calibration and parameter optimization often requires high computational

time and cost. The availability and quality of physical and land surface data for distributed models could influence the performance of the calibration or forecast output.

- Due to the above differences, inter-comparisons in multi-model flood forecasting could favor the performance of the forecasting skill and reliability to lumped models. However, distributed models are still needed to forecast flows at any intermediate gauged and ungauged locations within the watershed.

4.4.3. Hindcast simulation (model update and forecast)

Hydrological forecast verification was performed in the hindcast mode for the models that are selected during the calibration and validation phase. The hindcast period was chosen between June 1st, 2018 to November 30th, 2018, because frequent summer and fall season floods were observed, and archives of the four weather forecast products (Table 4-1) were available. In the hindcast mode, the selected hydrological models were forced by the precipitation forecasts while keeping the historical temperature data assuming that temperature is well forecasted and that errors in precipitation forecasts significantly affect the streamflow forecasts (Zsótér et al., 2016).

Before running with forecast data, the models are updated four times a day (at 00Z, 06Z, 12Z, and 18Z) with observed meteorological data for at least one month for the lumped models and one year for the distributed models before each forecast start time. In other words, at every forecast time (four times a day for the entire hindcast period), the models' states were continuously updated by running the models up to that time with observed meteorological data. Then forecast data were supplied to the models at each forecast time

and simulated for 18 hours forecast lead time in an hourly time step. The 18-hour forecast horizon is selected because it is the common forecast length that all the weather forecast products have, which is suitable for comparison purposes in urban and semi-urban catchments where response times are only a few hours.

4.4.4. Hydrological and flood forecasting performance evaluation

The streamflow forecast performances of the screened hydrological models (after calibration and validation) forced by the pre-screened weather forecast products were evaluated using various verification metrics. These forecast measures can be categorized into different forecast attributes based on their practical significance: the overall forecast skill, forecast accuracy/quality, ability to forecast the peak flow magnitude, ability to acquire the peak flow timing, threshold-based scores to assess reliability and bias, and the forecast economic value. Existing evaluation metrics applied in this research are provided in Table 4-2, and more details are presented in Appendix B.

Table 4-2: Metrics used for comparing the forecast performance, skill and quality of hydrological models and inputs

	Forecast attributes						
					Threshold-Based (Categorical)		
Metrics	**Overall Skill**	**Accuracy /Quality**	**Peak flow magnitude**	**Peak flow timing**	**Reliability/ Resolution/ Discrimination**	**Bias**	**Economic Value**
Taylor Skill Score (TSS)	✓						
MAE/RMSE		✓					
Magnitude Error (SD-Q)			✓				
Timing Error (SD-T)				✓			
PFC			✓				
Precipitation of Detection (POD)					✓		
False Alarm Rate (FAR)					✓		
Bias Frequency (BiasFreq)						✓	
Critical Success Index (CSI)					✓		
Economic Value (V)							✓

4.4.5. Simple forecast averaging methods

The rationale behind this task is that operational flood forecasters and users might be interested in combining or averaging the hydrological forecasts generated by multiple models or multiple weather forecasts. We tried to identify a simple forecast averaging method for the available multiple short-term streamflow forecasts provided by multiple hydrological models. Indeed, there are various advanced statistical post-processing and averaging methods in the literature (DelSole, 2007; Duan et al., 2007; Gneiting et al., 2005; Hopson and Webster, 2010; Sharma et al., 2019). However, we compare two simple methods from a practical aspect and the objectivity of this research. The first one is a simple

averaging method that gives equal weights to the model's forecast and henceforward called 'EnsMean' (mean of the ensemble).

The second one is an adaptive weighted forecast combination method (henceforward called 'AdtWeight') obtained by dynamically changing the weights throughout the hindcast period. Here, the weights are dynamically changed at every forecast time (t_o) based on the recent historical performance of each forecast. Aiolfi and Timmermann, (2006), first propose this method and often called the 'persistence forecasting method.' This method is mainly used in economic/market forecasting (Genre et al., 2013; Jordan et al., 2017; Matsypura et al., 2018).

Ranks are given to each model's forecast (i) based on the historical Mean Square Forecast Error (e^2) in a dynamic Tracking Window (t_o to ($t_o - h$)). The weights (w_i) are inversely proportional to the ranks and are found by:

$$\mathrm{w_i} = \frac{\mathrm{Rank_i^{-1}}}{\sum_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{Rank_i^{-1}}} \tag{4-2}$$

where: $Rank = f\left(S_t^i \ldots S_t^N\right)$ and $S_t^i = \left(e_{t_o,t_o-h}^{(i)}\right)^2$, N is the number of models.

The combined forecast is then estimated by multiplying each weight (w_i) to the corresponding model forecast (i) in the forecast horizon (t_o to ($t_o + h$)). Here h is taken as 18-hours, which is the length of the forecast lead time.

4.5. Results and Discussion

4.5.1. Calibration and Validation

Calibration and validation were performed using DDS optimization for the lumped hydrological models described in Section 4.4.2. The distributed models, SWMM, and WATFLOOD were previously calibrated in different studies and were directly selected for the next step. These models are also being used (included) in the local operational flood forecasting center at Toronto Regions Conservation Authority. Hence, we will discuss here the results of the eleven lumped hydrological calibration and validation performed in hourly time step.

Table 4-3 summarizes the statistical performances of the models using NSE, KGE, and PFC metrics. The results show that MACHBV, SACSMA, and PDM are the top three models that significantly outperform the other seven lumped models. NSE and KGE metrics depict that MACHBV and SACSMA, followed by PDM models, better reproduced the hourly observed streamflow data in both calibration and validation periods. Overall, the three models consistently achieved improved performance and are hence selected for the next phase, where verification using forecast data is performed.

Table 4-3: List of models used and their calibration and validation result (Using Observed hourly Precipitation & Temperature data). Bold font indicates models selected for the forecast verification step.

	Calibration			**Validation**		
Models	*NSE*	*KGE*	*PFC*	*NSE*	*KGE*	*PFC*
SACSMA	0.81	0.82	0.11	0.70	0.62	0.18
GR4H	0.48	0.63	0.15	0.43	0.55	0.20
MACHBV	0.802	0.86	0.12	0.71	0.65	0.17
NAM	0.14	0.32	0.22	0.13	0.3	0.23
PDM	0.75	0.82	0.14	0.67	0.71	0.18
IHACRES	0.65	0.76	0.12	0.54	0.68	0.19
WAGENINGEN	0.49	0.44	0.17	0.48	0.44	0.17
TANK	0.53	0.74	0.14	0.52	0.59	0.21
GARDENIA	0.56	0.81	0.14	0.48	0.65	0.19
MARTINE	0.54	0.74	0.15	0.53	0.64	0.20
SWMM	(Awol et al., 2018)					
WATFLOOD	(Kouwen, 2018)					

4.5.2. Screening weather forecast inputs

The hydrological forecasts driven by four deterministic weather forecasts products (Section 4.3.2.2) are pre-verified in this section to screen the appropriate inputs that result in better forecast accuracy in semi-urban catchments. For the rest of this paper, when comparing the forecast products, it should be noted that the comparison is based on the hydrological forecasts derived from these weather products.

Figure 4-3 presents the Mean Absolute Error (MAE) of the resultant streamflow forecasts from the weather forecast inputs using five selected hydrological models (see Section 4.5.1). MAE was computed over the 6-month verification period. It can be seen from this figure that NAM and RAP have significantly poor forecast accuracies than HRDPS and HRRR for a lead time beyond 3h, irrespective of the hydrological models used. The primary reason originates from the difference in horizontal spatial resolution. The spatial resolutions

of NAM and RAP (12 and 13km) are about 4.5 times larger than HRDPS and HRRR (2.5 and 3km). Even though the physical configuration and radar reflectivity assimilation cycle of RAP and HRRR are similar, the latter is a nested subset of the former (Benjamin et al., 2016). Besides, HRRR uses RAP as a boundary condition, whereas RAP uses the Global Forecast System (GFS), which has a coarser grid spacing of 28km (Alexander et al., 2017).

Most importantly, HRRR is preferable for forecasting rainfall events than RAP and NAM because it is a Convective-Permitting Model (CPM), which adds value for accurate prediction of convective clouds (Clark et al., 2016; Pinto et al., 2015). The results also showed that RAP achieved better forecast accuracy than NAM, particularly after 3h forecast lead time. Although no direct comparison between the two precipitation forecast products was found in literature, some studies comparing NAM and GFS (RAP's boundary condition) highlighted that the later provided skillful short-term forecast (Charles and Colle, 2009; Yan et al., 2016).

The Canadian HRDPS appears to be competitive with the NOAA's HRRR, as shown in Figure 4-3. Their differences seem insignificant in lumped models, but in distributed models, HRDPS has a slightly improved forecast accuracy than HRRR. Lumped hydrological models (MACHBV, SACSMA, and PDM) take mean spatial average precipitation forecasts as an input, whereas distributed models import the forecasts at each grid (WATFLOOD) or sub-catchment (SWMM). Due to these differences in discretization, lumped models tend to spatially average noises or uncertainties of short-term precipitation forecasts better than distributed models. This can be easily seen from Figure 4-3 that HRDPS and HRRR show almost equivalent MAEs at all lead times for lumped models but

slightly varied MAEs for distributed models. However, more verification metrics are needed to compare the two competitive weather forecast products. A more comprehensive comparison between HRDPS and HRRR and the five hydrological models are presented in the next section.

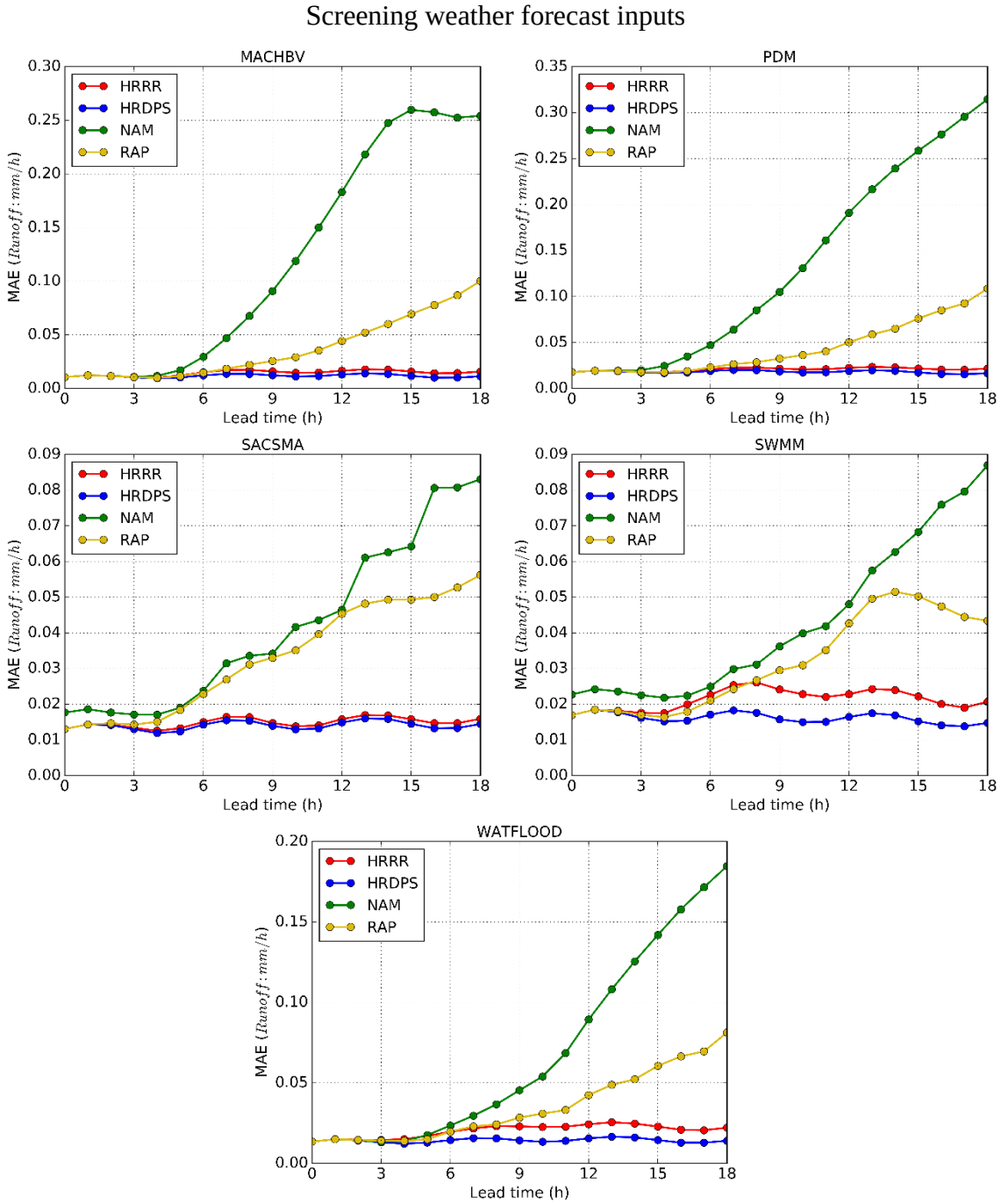

Figure 4-3: Screening weather forecast inputs by comparing the resulting hydrological forecast quality. Four weather forecast inputs (different colors) were fed into the five hydrological models (five boxes above) and the MAE at different lead times is estimated.

4.5.3. Comparison of hydrological models in forecast performance

Five hydrological models are evaluated and compared for their forecast performance using two forecast inputs. The following sub-sections provide a comparison from different forecast evaluation metrics.

4.5.3.1. Overall forecast accuracy and skill

One of the metrics that evaluate the overall forecast skill of streamflows generated by the models is Taylor Skill Score (TSS), which summarizes the Taylor Diagram. Figure 4-4 presents TSS along with the associated Taylor Diagram to illustrate the statistical performances of the hydrographs using HRDPS input. The TSS values indicated that the forecast skill of all models declines as the lead time advances from 1h to 18h, which is an expected phenomenon of deteriorating weather forecast skills with time. The result indicates that MACHBV followed by SACSMA models appears to be more skillful than the rest. Notably, the former provides a relatively good statistical pattern and correlation with the observation at all lead times compared to the other hydrological models. WATFLOOD and SWMM models show similar forecast skill throughout the forecast horizon and can be competent with the above two lumped models between 15h and 18h forecast lead time. The PDM model provides a poor forecast skill as TSS penalizes hydrographs that have low statistical similarity with the observed one. These result outlooks can also be seen from the Taylor Diagram, which shows the distribution of the models' performances in a statistical quadrant graph evaluated at each lead time (Each point corresponds to each hour of the 18h forecast lead times). Here, most of the points of

MACHBV and SACSMA are in closer proximity to the “OBS’ line and dot compared to the other models, especially the PDM model. Overall, using HRDPS input, MACHBV and SACSMA models (with the former better than the latter) are deemed to be superior for short-term streamflow forecasting in semi-urban watersheds, followed by the two distributed models (WATFLOOD and SWMM), which have relatively equivalent forecast skill.

So far, the HRDPS precipitation forecast was used as an input to the models. As HRRR and HRDPS could be competent at times (Section 4.5.2), the hydrological models were also forced with both inputs to get a comprehensive outlook of the resulting streamflow forecasts. The rest of this paper discusses the model's forecast performance using the two inputs unless otherwise stated. The top plot of Figure 4-5, shows the RMSE to present the overall forecast accuracy and quality of the streamflows generated by the models when using HRDPS and HRRR inputs and their differences. As can be seen from the RMSEs, the model’s performances using HRRR input show similar trends as using HRDPS. However, the quality of the forecast hydrographs for all models seems to deteriorate after 5h forecast lead time when using HRRR. Despite this, MACHBV and SACSMA appear to consistently provide better forecast accuracy than the other models regardless of the weather forecast inputs used.

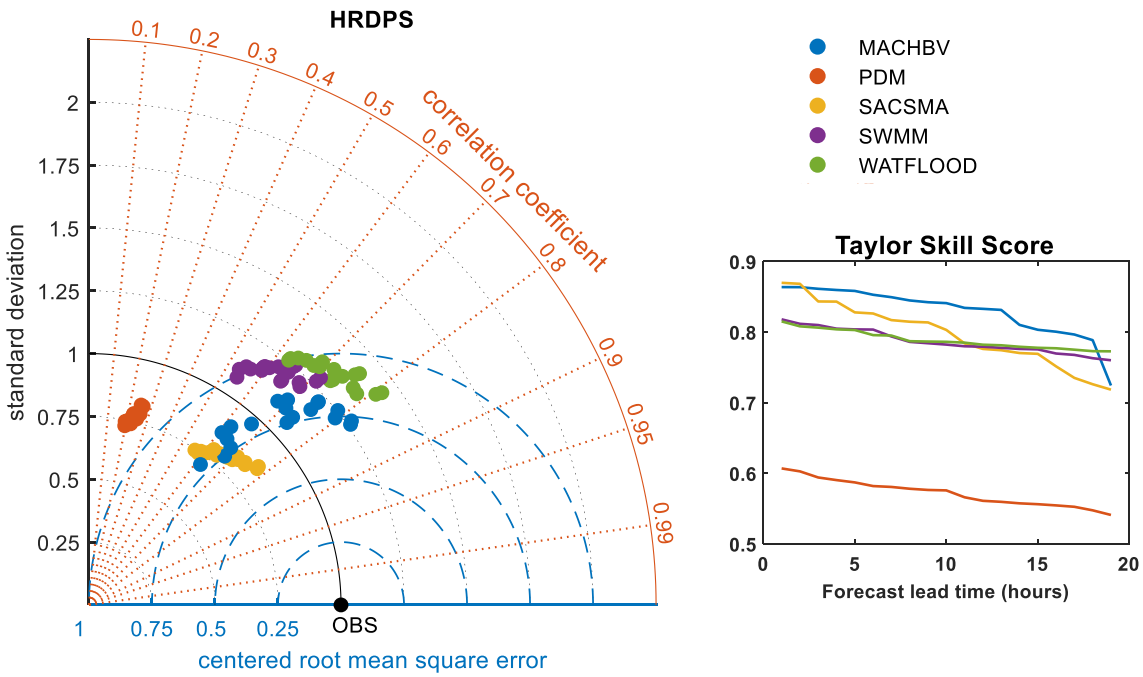

Figure 4-4: Taylor Diagram and Taylor Skill Score showing the statistical comparison of hydrological models at different forecast lead times using HRDPS input. The three statistical performance metrics that are displayed in Taylor Diagram are summarized by one single score in the Taylor Skill Score as shown in the rightmost plot (A score of one corresponds to the most skillful models).

4.5.3.2. Peak flow magnitude and timing

Figure 4-5 shows the forecast performances of the peak flow magnitude using Magnitude Error (SD_Q) and PFC metrics, and the peak flow timing using Timing Error (SD_T). The SD_Q metric shows the total vertical error, and the SD_T metric shows the total horizontal error in the rising and falling limbs of forecast hydrographs after separating multiple events from time series in the hindcast evaluation period. Applying this to the forecast time series of the models using HRDPS and HRRR inputs shows that MACHBV has the lowest magnitude error followed by SACSMA model. WATFLOOD and SWMM show relatively similar magnitude errors as SACSMA using HRDPS input, which is not the case when

using HRRR. Looking at the PFC metric, all models except PDM show similar peak flow performances across the lead time with HRDPS input. It appeared that, however, MACHBV is rather dominant followed by SACSMA (up to lead time 12h) and WATFLOOD (after lead time 12h) when forced by HRRR input. Comparing the two peak flow performance metrics, SD_Q can provide a detailed outlook of the elements of hydrographs for different forecasting systems and is a useful evaluation tool to diagnose events of a time series mimicking the hydrologist's visual inspection (Seibert et al., 2016). MACHBV forecasts provide the least peak flow timing error at all forecast lead times using both HRDPS and HRRR inputs (as can be seen by SD_T metric). SACSMA, WATFLOOD, and SWMM show relatively equivalent peak flow timing errors at all forecast lead times with HRRR input and the first 5hours with HRDPS. The timing of SACSMA's forecasts with HRDPS input was improved at latter lead times and was competitive to MACHBV.

The difference between HRDPS and HRRR is shown on the right side of each performance metric in Figure 4-5 (also in Figure 4-7). The differences in RMSE, SD_Q, and SD_T particularly indicate that HRDPS produces improved forecast quality and peak flow prediction than HRRR for all hydrological models considered. The improved quality of HRDPS might come from the capacity of this product to categorize storm dynamics using its deep convection mechanism and its relatively higher grid resolution (Milbrandt et al., 2016), which is ideal for flood forecasting in the small urbanizing catchment.

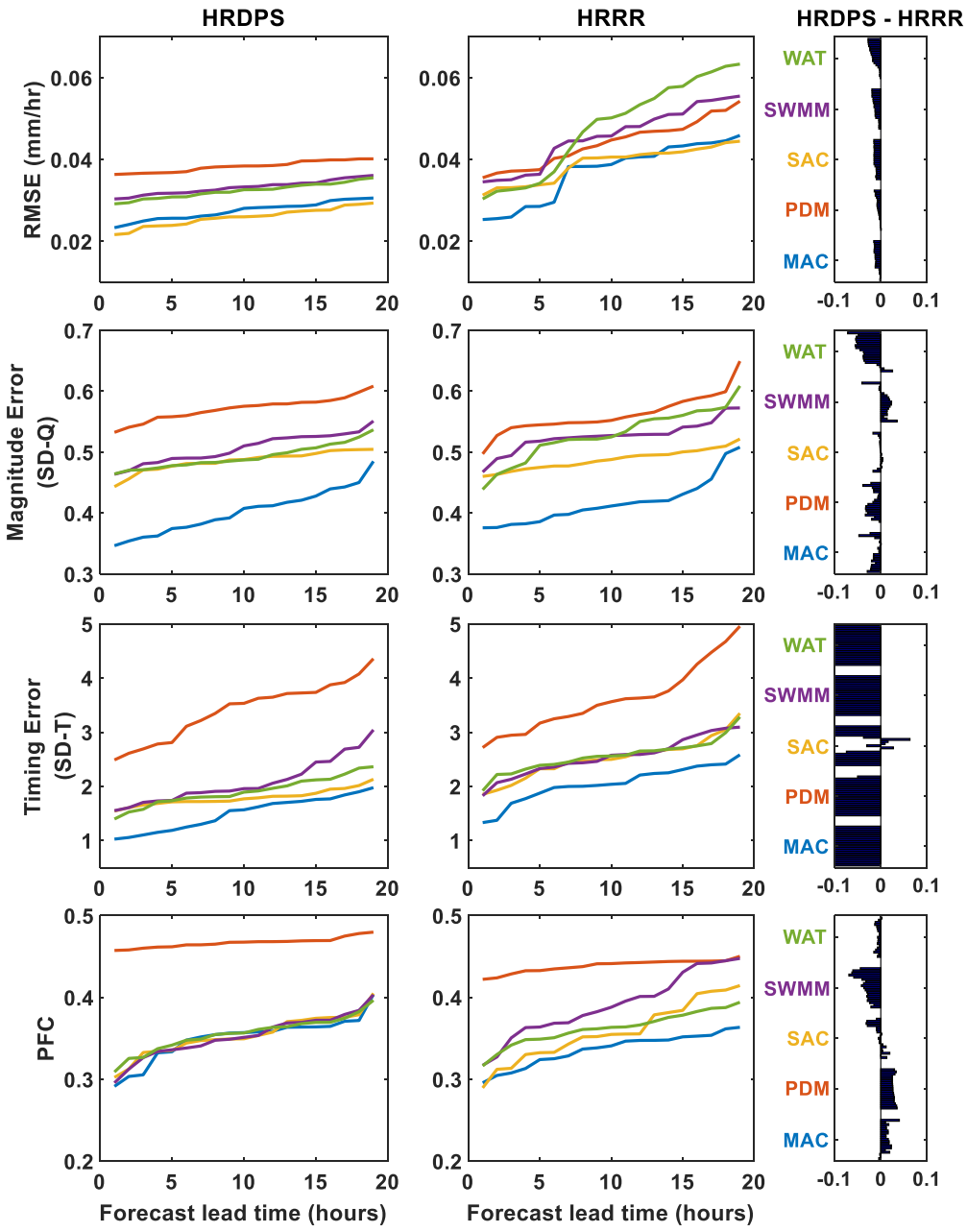

Figure 4-5: Comparison of forecast performance of hydrological models using different metrics that measure peak flow magnitude, timing, overall quality, and accuracy. The Models were fed by two competing weather forecast inputs (HRDPS: Left and HRRR: Right). The normalized difference between the two inputs estimated at each forecast lead time is shown in the rightmost horizontal bar graph.

4.5.3.3. Flood threshold approximation

The categorical forecast verification metrics and Economic value (Appendix B.3. and B.4, respectively) are formulated based on a given flood threshold. Unlike in weather forecasts where an occurrence of rain or snow can be defined easily for verification, streamflow forecasts requires some information to define flood threshold and sometimes depends on a value set by the decision-makers or operational flood forecasters. Since in this study, there is no information available for the latter, we assessed an approximate flood threshold. Based on the literature, it is most commonly set by estimating a higher percentile flow (between 80% and 99%) of historical streamflow data (Weber et al., 2006; Wu et al., 2012; Yilmaz et al., 2010). Other advanced methods are also available to define a flood threshold (Robson et al., 2017; Thielen et al., 2009b; Weeink, 2010; Wu et al., 2012). In this study, we first assumed an initial flood threshold "P" percentile flow of the historical daily observed time series between 2009 to 2018 and estimated the verification metrics. Then we check the appropriateness of this threshold by estimating the metrics multiple times for a range of thresholds above and below the initially set threshold value (P) in order to find the approximate cutoff threshold that abruptly affects the metrics.

An example is shown in Figure 4-6 for POD metrics. From this trial and error method, we have found that a 90-percentile flow is roughly a suitable flood threshold that can be used to estimate the categorical verification metrics and compare the hydrological models. This value is also within a reasonable range of the thresholds used in the above literature.

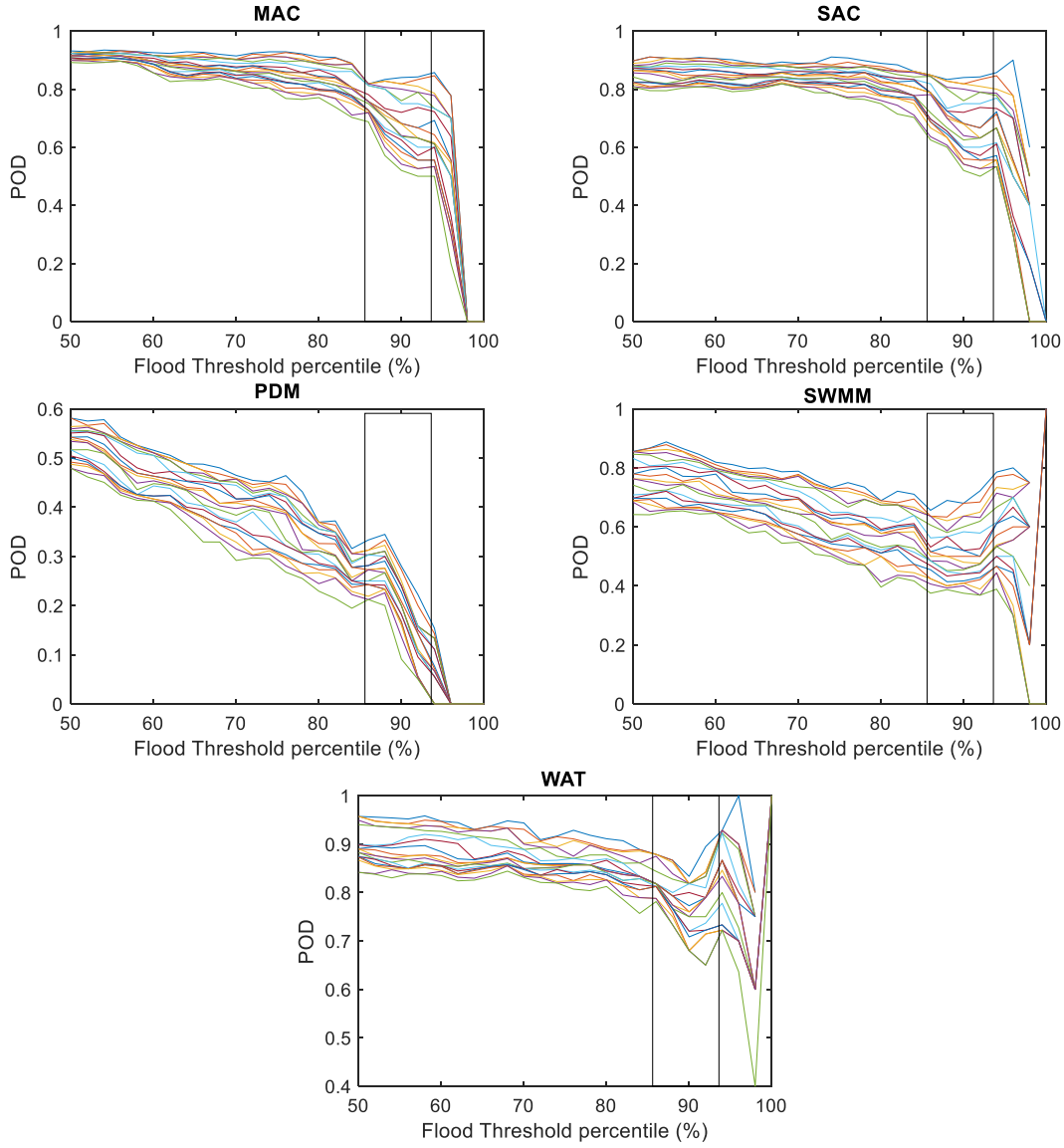

Figure 4-6: Flood threshold approximation: POD (Probability of Detection) versus different flow percentiles for different hydrological models. Different colors indicate outputs at different forecast leadtimes (1hr to 18hr).

4.5.3.4. Threshold-based scores

Figure 4-7 presents four categorical forecast verification scores for the hydrological forecasts generated by the hydrological models and weather forecast inputs: these are POD,

FAR, BiasFreq, and CSI. POD, FAR, and CSI scores indicate the reliability, discrimination, and resolution aspects of the forecast (Table 4-2). The performance of detecting the 90% flood threshold is much better experienced in MACHBV than the other hydrological models irrespective of the weather forecast product used, which can be seen from the POD. Besides, the model not only precisely detects hits but also records minimum false alarms because the FAR is significantly lower than the rest of the models for all forecast lead times. This is also supported by the CSI score, which indicates that after removing the "Correct negatives/rejections" from consideration, the ratio of the number of hits to the total number forecasted and observed floods is higher for MACHBV. SACSMA is the next well-performing model in terms of FAR at all lead times and POD and CSI scores at later lead times. WAFLOOD showed competitive performance with SACSMA at the early hours of the forecast. These trends are quite similar for both HRDPS and HRRR inputs, with the former slightly better than the later. In general, MACHBV with HRDPS input appeared to have better reliability, discrimination, and resolution skills in resolving the flood threshold at all forecast times. PDM model appeared to show poor categorical forecast verification scores.

In terms of bias, MACHBV and SACSMA appear to show similar, competitive, and increasing trend as the forecast lead time advances. The biases in the number of forecasts "yes"s over observed "yes"s revealed that MACHBV, WATFLOOD, and PDM models have over-forecasting and SACSMA and SWMM under-forecasting behaviors at all lead times. These trends are the same for HRDPS and HRRR inputs. However, the bias from

HRRR is significantly higher than HRDPS throughout the forecast lead times, regardless of the hydrological models applied.

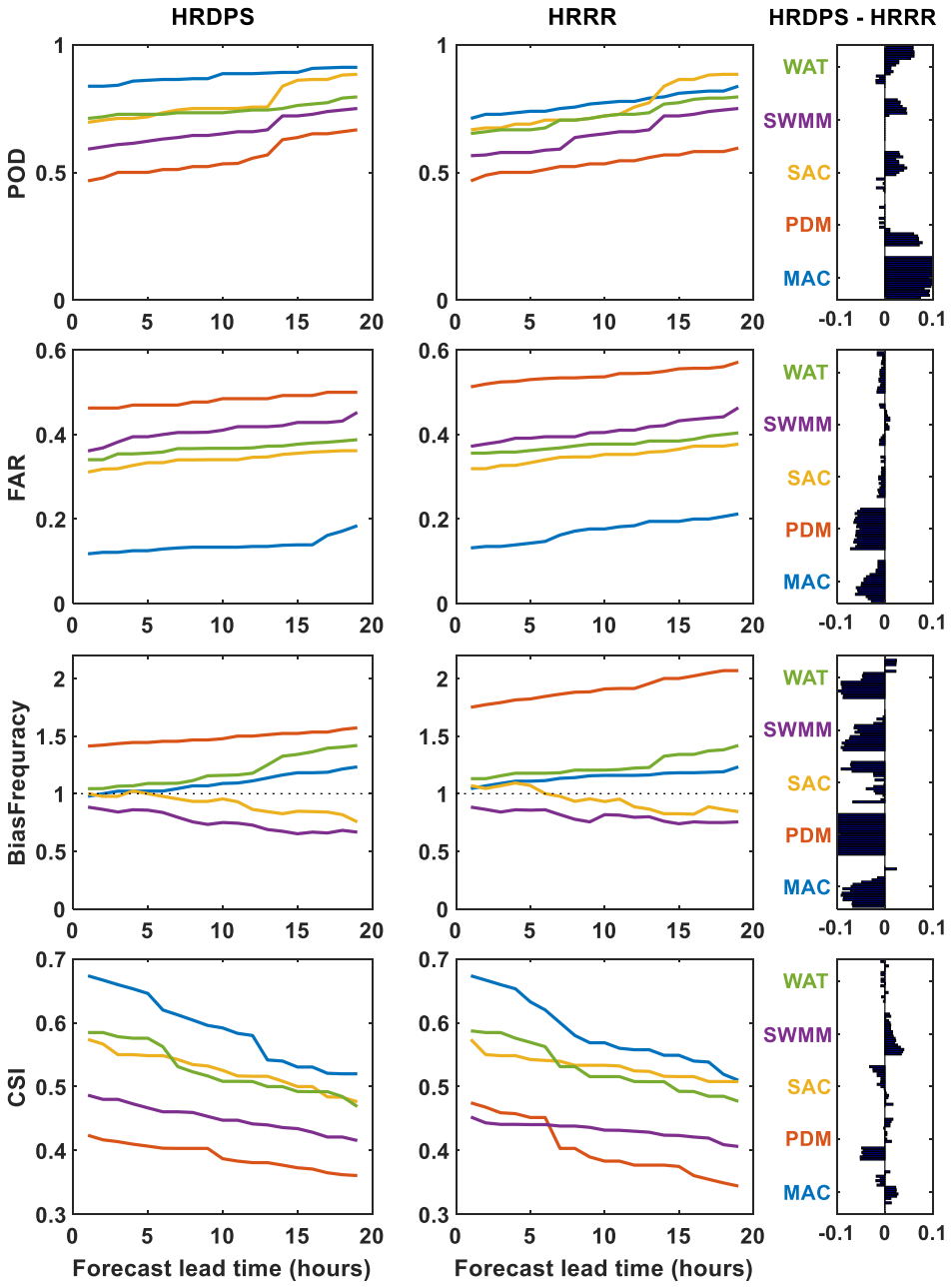

Figure 4-7: Same as Figure 6 but with categorical forecast verification metrics or threshold-based scores. A 90 percentile of the observed streamflow is set as a flood threshold for this research.

4.5.3.5. Forecast Economic Value

Using the 90-percentile flood threshold (see Section 4.5.3.3), the economic values (V) were estimated for the streamflow forecasts generated by the hydrological models forced with HRDPS input.

Figure 4-8 shows the Economic values with respect to different cost-loss ratios at different forecast lead times. The results indicate that most users or decision-makers can benefit better economic value when using MACHBV at all forecast lead times. SACSMA, WATLOOD, and SWMM follow in decreasing order of economic viability. A wide range of cost-loss ratio (C/La) with higher economic values exists for MACHBV at al lead times. This indicates that 1) several choices are available for users/decision-makers to gain better economic value, and 2) different groups of forecasting centers with a variety of cost-loss ratios can be beneficial by using the MACHBV model and HRDPS input in semi-urban watersheds. PDM is found to be a less economically viable model as the V values were low for narrow C/La values.

The optimum cost-ratio that produces maximum economic values is the same for all models, which is about 0.35. This optimum value is approximately equal to the relative frequency of occurrences (i.e., optimum C/La or maximum economic value happens when $r \approx \bar{o}$) (Richardson, 2006; Roulin, 2006). Although for all hydrological models, the optimum C/La remains the same, the maximum economic values decrease as the lead time increases (Zhu et al., 2002). This is expected because the POD and FAR, which are elements of economic value (V) (Eqn B-13), decreases and increases respectively as the lead time advances (Figure 4-7).

For medium- and long-term prediction, probabilistic and ensemble hydrological forecasts can add more economic values than deterministic forecasts (Richardson, 2006; Verkade and Werner, 2011). Although for very short-term (hourly and sub-daily) prediction, deterministic forecasts might be preferable than ensemble forecasts (Horat et al., 2018), some literature suggests that adding uncertainty and error information to deterministic forecasts could improve the economic values (Roulin, 2006; Verkade and Werner, 2011).

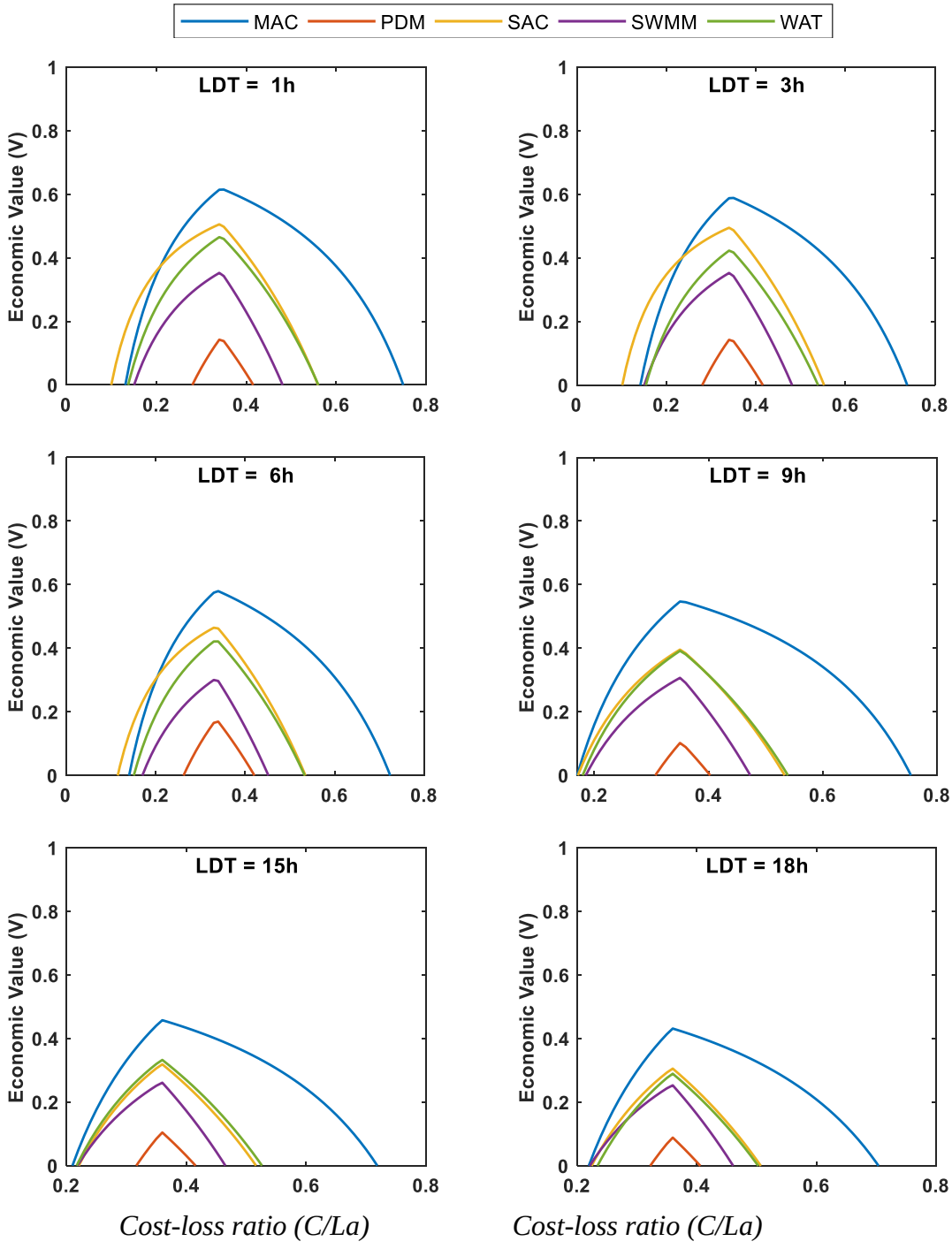

Figure 4-8: Comparison of hydrological models by their forecast Economic Value (V). V is estimated using the deterministic HRDPS weather forecast input and is drawn as a function of cost-lost ratio to account for various users and forecast systems.

4.5.4. Ensemble mean vs. adaptive weighted average

Two simple averaging methods were used to estimate the combined streamflow forecasts from all hydrological models driven by HRDPS input (See Section 4.4.5). Comparison between 'EnsMean' (mean of the ensemble) and 'AdtWeight' (adaptive weighted average) methods was performed in terms of the overall forecast accuracy and skill of their output. Figure 4-9 shows the Taylor Diagram and Taylor Skill Score (TSS) of EnsMean and AdtWeight. As a reference, MACHBV model result is presented because so far, it proved to have a superior forecast performance. The result shows that AdtWeight has better forecast skill than EnsMean and even the best model. In AdtWeight method, a dynamic tracking period of 18h is used to estimate the weights of each model forecast at every forecast start time based on their last 18h performances. Then the weights are applied to the corresponding model forecasts for the next 18h to estimate the average forecast. The result implied that dynamically changing the weights based on the recent forecast performances of the models improved the combined forecast skill as opposed to just averaging (assigning equal weights to) the models' forecasts. Furthermore, the contribution of MACHBV to the performance of the AdtWeight forecast is higher, as can be from the closer TSS values between the two. This outcome proves that there is steady persistence in hydrological forecasting as the top-ranking models in the last 18 to 24 hours will highly likely continue to perform well in the same forecast horizon.

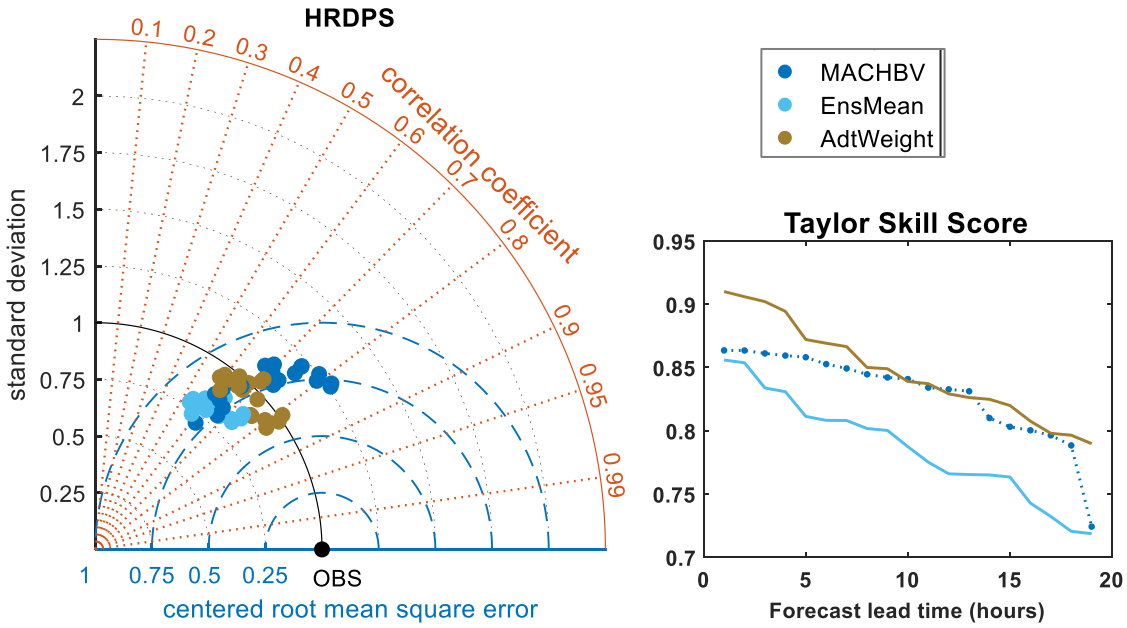

Figure 4-9: Taylor Diagram and Taylor Skill Score to show the comparison between Ensemble Mean and Adaptive weighted averaging methods. For reference, one model's (MACHBV) output is presented.

4.6. Conclusion

This study was conducted to identify the right combination of skillful hydrological models and Numerical Weather Predictions (NWPs) for enhanced short-term flood forecasting in a semi-urban watershed. Several existing verification metrics were used to evaluate and select the best models and inputs. For the research, Humber River Watershed, a semi-urban catchment located in Southern Ontario, was used as a study area. Twelve different lumped and distributed hydrological models were calibrated and validated using the DDS optimization algorithm. A total of five hydrological models were selected for further analysis based on their superior calibration performances and previous application in the

study area; three lumped models (MACHBV, SACSMA, and PDM, all coupled with SNOW17) and two semi-distributed/distributed models (SWMM and WATFLOOD).

Efforts were made to collect available high- and mid-resolution weather forecast products from local, regional, or global sources. As such, four NWP products were collected. These are the 2.5km High-Resolution Deterministic Precipitation System (HRDPS) from Environment Canada, the 3km High-Resolution Rapid Refresh (HRRR), 12km North American Mesoscale Forecast System (NAM), and 13km Rapid Refresh (RAP) from NOAA. Precipitation forecasts from these weather forecast products were used as forcing inputs to the selected hydrological models. As such, forecast verification for a different combination of five hydrological models and four weather forecast inputs was performed. Hourly hydrological forecasts were produced four times a day during a 6-month verification period and issued up to 18hr forecast lead times.

The first forecast analysis performed was to screen the weather forecast products that have better streamflow forecast performance. The pre-screening revealed that HRDPS and HRRR produce significantly better hydrological forecast accuracy than NAM and RAP. This finding implied that, for semi-urban watersheds that typically have a short time of concentrations and relatively small catchment areas, high-resolution weather products are found to be a proper precipitation forecast inputs to the hydrological models.

In order to identify appropriate flood forecasting models in a semi-urban watershed, the five hydrological models were comprehensively evaluated in forecast mode using the two fine-resolution precipitation forecast inputs. The evaluation and comparison were based on performance measures categorized into four forecast attributes: overall forecast accuracy

and skill, peak flow magnitude and timing, categorical or threshold-based scores, and economic value. In general, the MACHBV model with HRDPS input appeared to be the best model-input combination because it captured the peak flow magnitude and timing, detected the flood threshold, and appeared economically viable at all forecast lead times better than any other model-input combination.

Taylor Diagram and Taylor Skill Score results showed that forecasts from MACHBV and SACSMA (both with HRDPS input) have better overall forecast skill and accuracy than WATFLOOD and SWMM models, which showed similar skills. A similar trend is also observed with HRRR input. Overall, MACHBV with HRDPS input provided improved skills and statistical pattern proximity with the observation.

The performance of achieving the peak flow magnitude was tested by Series Distance (SD_Q) and Peak Flow Criteria (PFC) metrics. Results indicated that MACHBV's forecasts have the lowest peak flow magnitude error at all lead times regardless of the weather forecast inputs, followed by the SACSMA model. Regarding peak flow timing error, similar findings were obtained by the SD_T metric. MACHBV with HRDPS input was better at forecasting the peak flows on time while other models, to some extent, delay or postpone the peak flow times. The Series Distance metrics were found to be useful evaluation tools to diagnose and quantify vertical/horizontal errors in the rising and falling limbs of forecast hydrographs mimicking a hydrologist's visual inspection (Seibert et al., 2016).

A proper flood threshold was required to estimate categorical forecast verification metrics. In this research, we found the approximate threshold by optimizing the resulting matrices

using different ranges of thresholds. With this trial and error, we have found that roughly the 90% flow of the historical observed time series is a suitable flood threshold that can be used to estimate POD, FAR, CSI, and Bias Frequency scores. In general, the threshold-based verification showed that MACHBV with HRDPS input has better reliability, resolution, and discrimination skill in categorizing the flood threshold comparing to the other models-input integration.

The economic values of the hydrological forecasts were assessed between different models. Results showed that MACHBV with HRDPS input produced a better economic value to the various users/decision-makers followed by SACSMA, WATLOOD, and SWMM, in decreasing order of economic viability.

Comparing the lumped and distributed hydrological models in general, the lumped models except PDM outperform the latter, particularly in the first 15hr forecast lead times. However, distributed models could be competitive beyond 15hr forecast lead times. Lumped models tend to spatially average noises of short-term precipitation forecasts better than distributed models. Overall, the MACHBV model appeared to have the highest forecast skill, while PDM showed the lowest forecast skill and quality.

Comparing the two high-resolution weather forecast products, HRDPS and HRRR, the use of the former tends to be superior in generating hydrological forecasts with an improved forecast skill, quality, and peak flow prediction and timing. HRDPS also generates fewer forecast biases than HRRR regardless of the hydrological models used. The enhanced quality of HRDPS may be associated with its capacity to categorize storm dynamics in fine

grid resolutions using deep convection mechanisms (Milbrandt et al., 2016), which is ideal for flood forecasting in small urban or semi-urban catchments.

Finally, we assessed simple forecast averaging methods for users or operational flood forecasters interested in combined forecasts that can be estimated at a low computational budget. The results showed that dynamically changing weights based on each model's recent performance improved the combined forecasts better than assigning equal weights. Also, results indicated that there is steady persistence in hydrological forecasting as the top-ranking models in the very recent history will likely continue to perform well in the near future, as also suggested by Aiolfi and Timmermann, 2006.

4.7. Acknowledgment

This work was supported by the Natural Science and Engineering Research Council (NSERC) Canadian FloodNet (Grant number: NETGP 451456).

Special thanks are extended to Dr. Nickolas Kouwen, the developer of WATFLOOD, for setting up the model on Humber River Watershed.

The authors would like to thank the Toronto Region Conservation Authority (TRCA) staff for providing models, data, and support for this research.

Appendix B: Performance evaluation metrics

B.1. Overall Forecast Skill and Accuracy

Taylor Skill Score (TSS): is a single score used to summarize a Taylor Diagram (Taylor, 2001). Taylor Diagram measures correlation coefficient, centered root mean square error and standard deviation, and presents into one diagram. TSS has been used to evaluate and validate different models and approaches (Awol et al., 2018). It is formulated as:

$$TSS = \frac{4(1+R)}{\left(\frac{\sigma_s}{\sigma_o} + \frac{1}{\sigma_s/\sigma_o}\right)^2 (1+R_o)} \quad \text{(B-1)}$$

where: R, σ_s and σ_o are the correlation coefficient between variances of the forecast and observation, respectively. R_o is maximum correlation attainable, here taken as the maximum of the correlation coefficients of candidate model forecasts. TSS approaches a maximum of one when the ratio of the variances of forecast and observation is closer to unity and as R approaches R_o.

Root-Mean-Square-Error (RMSE): is the square root of the variance between forecast and observation. It is commonly used to measure the accuracy of a forecast, is scale-dependent, and is quite sensitive to large errors (Hyndman and Koehler, 2006).

Mean Absolute Error (MAE): is the mean of the absolute error between forecast and observation (Eqn. B-2). MAE is the deterministic version of the Continuous Rank Probability Score (CRPS), which is used to measure errors in ensemble forecasting

(Gneiting et al., 2005). Since deterministic forecasts are used in this study, MAE is applied to evaluate and compare the average hydrologic forecast error between the forecast outputs.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i| \quad \text{(B-2)}$$

B.2. Model performance on peak flow magnitude and timing

Peak Flow Criteria (PFC): is a metric that is used to measure the peak flow error of the forecast (Coulibaly et al., 2001a), and its formulation is given in Equation 1. PFC has been used in streamflow forecasting studies to assess a model's performance to predict peak flows (El-Shafie et al., 2009; García-Bartual, 2002; Han and Coulibaly, 2019).

Magnitude and Timing Error using Series Distance (SD) method:

First introduced by Ehret and Zehe, 2011, and later modified by Seibert et al., 2016, a Series Distance (SD) metric is an innovative way to quantify the similarity of two hydrographs mimicking the visual inspection of a hydrologist. In a Series Distance method, the correspondence between the amplitude and timing of observed and modeled hydrograph events is diagnosed by constructing a connector, as shown in an example taken from this study (Figure B- 1). A single or a combination of high flow events are first separated by rising and falling limps in hydrographs preprocessing stage. Then the source of errors in matching the amplitude and timing of modelled hydrograph with the observed hydrograph will be comprehensively quantified following a successive step such as coarse-graining, calculation of connector distances and creating a contingency table (please see Seibert et al., 2016, for more details and the code can also be found in Ehret and Seibert, 2016). The

error in amplitude corresponds to the error in peak flow magnitude (SD_Q) and is calculated by equation B-3. Similarly, the error in peak flow timing (SD_T) is estimated by equation B-4.

$$SD_Q = \frac{|Q_o(c) - Q_s(c)|}{\frac{1}{2}(Q_o(c) + Q_s(c))} \quad \text{(B-3)}$$

$$SD_T = \frac{|Q_o(t) - Q_s(t)|}{\frac{1}{2}(Q_o(t) + Q_s(t))} \quad \text{(B-4)}$$

Where: Q_o and Q_s are observed and modeled/simulated hydrographs, and c is a series distance connector between observed and modeled points.

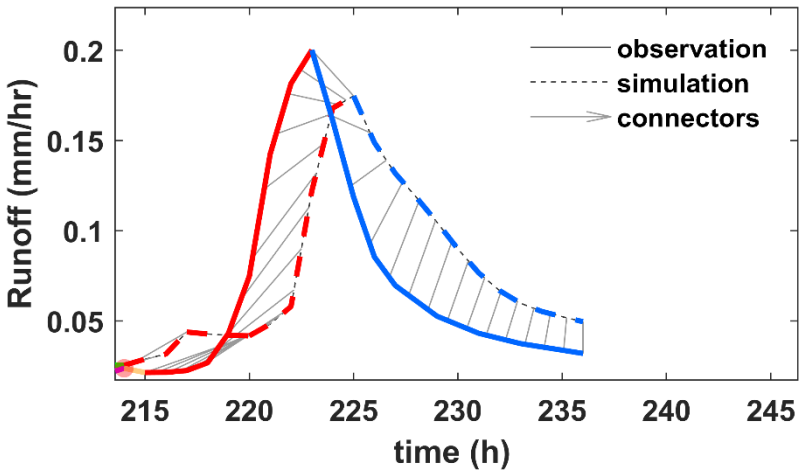

Figure B- 1: Illustration of Series Distance (SD) method for calculating magnitude (Q) and timing error (T) for a sample event selected within a timeseries of this study

Table B- 1: Contingency table and associated parameters for calculating categorical verification metrics and economic value. Detailed definitions of symbols are presented in Section B.3

		Observation				**Flood Threshold Occurred**		
		Yes	No	∑		Yes	No	aa=a/(a+c)
Forecast /action taken	Yes	a (hits)	b (false alarms)	a+b		$C+L_u$	C	bb=b/(b+d)
	No	c (misses)	d (correct negatives)	c+d		$L=L_u+L_p$	0	cc=c/(c+d)
	∑	a+c	b+d			$\bar{o}$	$1-\bar{o}$	$\bar{o}$=(a+c)/(a+b+c+d)

B.3. Categorical(threshold-based) forecast verification

Since deterministic forecasts were used to force hydrological models, the categorical verification method was used instead of a probabilistic method to measure the discrimination/reliability/resolution and bias attributes of forecast (Table 4-2). Categorical verification metrics are usually applied in weather forecasts but could be applied for hydrological forecasts by using the same principle accounting a streamflow threshold as a discrete event. A contingency table (Wilks, 2006) is then used to count frequencies of four possible cases where a certain flood threshold is equaled or exceeded by the simulated streamflow forecasts: these are “Hits = a”, “False alarms = b”, “Misses = c” and “Correct rejections = d” (Table B- 1). An approximate flood threshold is determined in this study (see Section 4.5.3.3) and used to quantify the following categorical verification used metrics. Further details on the metrics can be found in Wilks, 2006.

Probability of Detection (POD):- is the fraction of the cases when the flood threshold was forecasted by the model and was observed at the same time. A unity value of POD means the model forecasts are perfect. POD is expressed using elements of the contingency table as follows:

$$POD = \frac{a}{a + c} \tag{B-5}$$

False Alarm Rate (FAR):- is the fraction of occasions when the flood threshold was forecasted by the model but did not happen in the actual case. Zero FAR corresponds to a better forecast. It is formulated as:

$$FAR = \frac{b}{b + d} \tag{B-6}$$

Threat score or Critical Success Index (CSI): is the number of forecasts that is predicted to hit the flood threshold divided by all cases where the threshold is forecasted and/or observed (Eqn. B-7). It is an often-preferable measure because it removes the "Correct Rejections" from the total cases. In a flood forecasting system, counting several non-exceedance cases is not quite important (Wilks, 2006). A value of one corresponds to the best model forecasts.

$$CSI = \frac{a}{a + b + c} \tag{B-7}$$

Bias Frequency: measures the over-forecasting (>1) or under-forecasting (<1) behavior of the forecasting system and is given by the ratio of average forecasted cases over observed cases, as shown below (Wilks, 2006).

$$BiasFreq = \frac{a+b}{a+c} \quad \text{(B-8)}$$

B.4. Economic Value of the model forecast

The contingency table established above is used to analyze the cost-loss decision model of a forecasting system (Richardson, 2006). When the forecaster or decision-maker takes action to prevent possible damage, it has an associated cost (C) along with some unprotectable/unavoidable loss (Lu) whether the forecasted flood threshold occurred (a) or not (b) (refer Table B- 1). On the other hand, if there is no action taken and the flood threshold has occurred (c), the decision-maker will incur a total loss, which is the sum of avoidable (La) and unavoidable losses (Lu). There will be no loss in the case of "Correct rejections" (d). Now, the following expenses (E) of the decision-maker can be derived (Richardson, 2006; Verkade and Werner, 2011; Zhu et al., 2002):

- If there is only climatological information available (the relative frequency of occurrences ($\bar{o}$) is known but there is no forecast information), the baseline (climatological) expense (E_C) is to either always protect ($\bar{o}$(La+ Lu)) or never protect (C+$\bar{o}$Lu) whichever is the minimum:

$$E_C = Min[\bar{o}(L_a + L_u), C + \bar{o}L_u] \quad \text{(B-9)}$$

- For an ideal and a perfect forecasting system, whether the flood threshold occurs or not, the decision-maker will provide mitigative actions (only for those which occurs) with an average expense of:

$$E_P = \bar{o}(C + L_u) \quad \text{(B-10)}$$

- For a default forecasting system (between the above two boundary conditions), the average expense (E_F) of the decision-maker can be obtained by connecting the forecast-observation rates to the corresponding cost and loss of occurrences and non-occurrences (Table B- 1):

$$E_F = aa(C + L_u) + bb(C) + cc(L_u + L_a) \quad \text{(B-11)}$$

- The economic value (V) of the forecasting system can then be estimated as:

$$V = \frac{E_c - E_F}{E_c - E_P} \quad \text{(B-12)}$$

- Substituting and rearranging of equation B-9 to B-12 gives:

$$V = \frac{Min(r, \bar{o}) - bb(1 - \bar{o})r + aa(\bar{o})(1 - r) - \bar{o}}{Min(r, \bar{o}) - \bar{o}r} \quad \text{(B-13)}$$

Where *r* is the cost-loss ratio (C/La) (loss here is the avoidable loss), *bb* is the False Alarm Rate (FAR), and *aa* is the Probability of Detection (POD).

The maximum economic value (V=1) is obtained when the forecasting system provides a perfect and an ideal forecast (when Ef = EP); V can be negative when the system has more expenses than the baseline forecasting system (Ec), which is not desirable. The economic value (V) of a forecasting system depends on different decision-makers or users that have different cost-loss ratios (r) ranging between 0 and 1. Hence, a plot of V versus C/La provides a convenient way to find the maximum/optimum economic value of a deterministic hydrological forecast and to compare different forecasting systems (hydrological models) given the same climatological frequency and decision-maker.

4.8. 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. Mon. Weather Rev. 141, 3462–3476. https://doi.org/10.1175/MWR-D-12-00206.1

Achleitner, S., Schöber, J., Rinderer, M., Leonhardt, G., Schöberl, F., Kirnbauer, R., Schönlaub, H., 2012. Analyzing the operational performance of the hydrological models in an alpine flood forecasting system. J. Hydrol. 412–413, 90–100. https://doi.org/10.1016/j.jhydrol.2011.07.047

Aiolfi, M., Timmermann, A., 2006. Persistence in forecasting performance and conditional combination strategies. J. Econom. 135, 31–53. https://doi.org/10.1016/j.jeconom.2005.07.015

Alexander, C., Weygandt, S., Benjamin, S., Dowell, D., Hu, M., Smirnova, T., Olson, J., Kenyon, J., Grell, G., James, E., Lin, H., Ladwig, T., Brown, J., Alcott, T., Jankov, I., 2017. WRF-ARW research to operations update: The Rapid-Refresh (RAP) version 4, High-Resolution Rapid Refresh (HRRR) version 3 and convection-allowing ensemble prediction.

Anderson, E., 2006. Snow accumulation and ablation model – SNOW-17, Nature. Silver Spring, MD. https://doi.org/10.1038/177563a0

Awol, F.S., Coulibaly, P., Tolson, B.A., 2018. Event-based model calibration approaches for selecting representative distributed parameters in semi-urban watersheds. Adv. Water Resour. 118, 12–27. https://doi.org/10.1016/j.advwatres.2018.05.013

Benjamin, S.G., Weygandt, S.S., Brown, J.M., Hu, M., Alexander, C.R., Smirnova, T.G., Olson, J.B., James, E.P., Dowell, D.C., Grell, G.A., Lin, H., Peckham, S.E., Smith, T.L., Moninger, W.R., Kenyon, J.S., Manikin, G.S., Benjamin, S.G., Weygandt, S.S., Brown, J.M., Hu, M., Alexander, C.R., Smirnova, T.G., Olson, J.B., James, E.P., Dowell, D.C., Grell, G.A., Lin, H., Peckham, S.E., Smith, T.L., Moninger, W.R., Kenyon, J.S., Manikin, G.S., 2016. A North American hourly assimilation and model forecast cycle: The Rapid Refresh. Mon. Weather Rev. 144, 1669–1694. https://doi.org/10.1175/MWR-D-15-0242.1

Bennett, J.C., Robertson, D.E., Shrestha, D.L., Wang, Q.J., Enever, D., Hapuarachchi, P., Tuteja, N.K., 2014. A System for Continuous Hydrological Ensemble Forecasting (SCHEF) to lead times of 9 days. J. Hydrol. 519, 2832–2846. https://doi.org/10.1016/J.JHYDROL.2014.08.010

Bergström Sten, 1978. Development of a conceptual deterministic rainfall - runoff model. Stockholm.

Beven, K.J., Kirkby, M.J., Schofield, N., Tagg, A.F., 1984. Testing a physically-based flood forecasting model (TOPMODEL) for three U.K. catchments. J. Hydrol. 69, 119–143.

https://doi.org/10.1016/0022-1694(84)90159-8

Burnash, R., Ferral, R.L., McGuire, R.A., 1973. A generalized streamflow simulation system : conceptual modeling for digital computers. National Weather Service, NOAA, and the State of California Department of Water Resources Tech. Rep., Joint Federal–State River Forecast Center, Sacramento, CA.

Campolo, M., Soldati, A., Andreussi, P., 2003. Artificial neural network approach to flood forecasting in the River Arno, Hydrological Sciences-Journal. https://doi.org/DOI: 10.1623/hysj.48.3.381.45286

Chang, F.J., Chen, P.A., Lu, Y.R., Huang, E., Chang, K.Y., 2014. Real-time multi-step-ahead water level forecasting by recurrent neural networks for urban flood control. J. Hydrol. 517, 836–846. https://doi.org/10.1016/j.jhydrol.2014.06.013

Charles, M.E., Colle, B.A., 2009. Verification of extratropical cyclones within the NCEP operational models. part i: analysis errors and short-term NAM and GFS forecasts. Weather Forecast. 24, 1173–1190. https://doi.org/10.1175/2009WAF2222169.1

Chen, C.-S., Jhong, Y.-D., Wu, W.-Z., Chen, S.-T., 2019. Fuzzy time series for real-time flood forecasting. Stoch. Environ. Res. Risk Assess. 33, 645–656. https://doi.org/10.1007/s00477-019-01652-8

Chen, J., Hill, A.A., Urbano, L.D., 2009. A GIS-based model for urban flood inundation. J. Hydrol. 373, 184–192. https://doi.org/10.1016/J.JHYDROL.2009.04.021

Chow, K.C.A., Watt, W.E., Watts, D.G., 1983. A stochastic-dynamic model for real time flood forecasting. Water Resour. Res. 19, 746–752. https://doi.org/10.1029/WR019i003p00746

Clark, P., Roberts, N., Lean, H., Ballard, S.P., Charlton-Perez, C., 2016. Convection-permitting models: a step-change in rainfall forecasting. Meteorol. Appl. 23, 165–181. https://doi.org/10.1002/met.1538

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. J. Hydrol. 375, 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Coulibaly, P., Anctil, F., Bobée, B., 2001a. Multivariate reservoir inflow forecasting using temporal neural networks. J. Hydrol. Eng. 6, 367–376. https://doi.org/10.1061/(ASCE)1084-0699(2001)6:5(367)

Coulibaly, P., Anctil, F., Bobée, B., 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J. Hydrol. 230, 244–257. https://doi.org/10.1016/S0022-1694(00)00214-6

Coulibaly, P., Bobée, B., Anctil, F., 2001b. Improving extreme hydrologic events forecasting using a new criterion for artificial neural network selection. Hydrol. Process. 15, 1533–1536. https://doi.org/10.1002/hyp.445

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. J. Hydrometeorol. 12, 713–

728. https://doi.org/10.1175/2011JHM1347.1

De Roo, A.P.J., Gouweleeuw, B., Thielen, J., Bartholmes, J., Bongioannini-Cerlini, P., Todini, E., Bates, P.D., Horritt, M., Hunter, N., Beven, K., Pappenberger, F., Heise, E., Rivin, G., Hils, M., Hollingsworth, A., Holst, B., Kwadijk, J., Reggiani, P., Dijk, M. Van, Sattler, K., Sprokkereef, E., 2003. Development of a european flood forecasting system. Int. J. River Basin Manag. 1, 49–59. https://doi.org/10.1080/15715124.2003.9635192

DelSole, T., 2007. A Bayesian Framework for Multimodel Regression. J. Clim. 20, 2810–2826. https://doi.org/10.1175/JCLI4179.1

Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.J., Hartman, R., Herr, H.D., Fresch, M., Schaake, J., Zhu, Y., 2014. The science of NOAA's operational hydrologic ensemble forecast service. Bull. Am. Meteorol. Soc. 95, 79–98. https://doi.org/10.1175/BAMS-D-12-00081.1

Du, J., Qian, L., Rui, H., Zuo, T., Zheng, D., Xu, Y., Xu, C.-Y., 2012. Assessing the effects of urbanization on annual runoff and flood events using an integrated hydrological modeling system for Qinhuai River basin, China. J. Hydrol. 464–465, 127–139. https://doi.org/10.1016/J.JHYDROL.2012.06.057

Duan, Q., Ajami, N.K., Gao, X., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Adv. Water Resour. 30, 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Dumedah, G., Coulibaly, P., 2013. Evaluating forecasting performance for data assimilation methods: The ensemble Kalman filter, the particle filter, and the evolutionary-based assimilation. Adv. Water Resour. 60, 47–63. https://doi.org/10.1016/j.advwatres.2013.07.007

ECCC, 2017. Environment and Climate Change Canada - Weather and Meteorology - Canada's top ten weather stories of 2013 [WWW Document]. URL https://www.ec.gc.ca/meteo-weather/default.asp?lang=En&n=5BA5EAFC-1&offset=2&toc=hide (accessed 5.24.19).

Ehret, U., Seibert, S.P., 2016. The Series Distance matlab code. GIT Repos. available https//github.com/KIT-HYD/SeriesDistance. https://doi.org/10.5281/zenodo.60356

Ehret, U., Zehe, E., 2011. Series distance - An intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. Hydrol. Earth Syst. Sci. 15, 877–896. https://doi.org/10.5194/hess-15-877-2011

El-Shafie, A., Abdin, A.E., Noureldin, A., Taha, M.R., 2009. Enhancing Inflow Forecasting Model at Aswan High Dam Utilizing Radial Basis Neural Network and Upstream Monitoring Stations Measurements. Water Resour. Manag. 23, 2289–2315. https://doi.org/10.1007/s11269-008-9382-1

El Hassan, A.A., Sharif, H.O., Jackson, T., Chintalapudi, S., 2013. Performance of a conceptual and physically based model in simulating the response of a semi-urbanized watershed in

San Antonio, Texas. Hydrol. Process. 27, 3394–3408. https://doi.org/10.1002/hyp.9443

Emerton, R.E., Stephens, E.M., Pappenberger, F., Pagano, T.C., Weerts, A.H., Wood, A.W., Salamon, P., Brown, J.D., Hjerdt, N., Donnelly, C., Baugh, C.A., Cloke, H.L., 2016. Continental and global scale flood forecasting systems. Wiley Interdiscip. Rev. Water 3, n/a-n/a. https://doi.org/10.1002/wat2.1137

FloodList, 2019. Canada – thousands evacuated after rivers flood in Quebec, New Brunswick and Ontario - FloodList [WWW Document]. URL http://floodlist.com/america/canada-flood-in-quebec-new-brunswick-ontario-april-may-2019 (accessed 5.24.19).

Fortin, J.-P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J., Villeneuve, J.-P., 2001. Distributed watershed model compatible with remote sensing and GIS data. I: description of model. J. Hydrol. Eng. 6, 91–99. https://doi.org/10.1061/(ASCE)1084-0699(2001)6:2(91)

García-Bartual, R., 2002. Short term river flood forecasting with neural networks. iEMSs 2002 Int. Congr. Integr. Assesment Decis. Support 160–165.

Genre, V., Kenny, G., Meyler, A., Timmermann, A., 2013. Combining expert forecasts: Can anything beat the simple average? Int. J. Forecast. 29, 108–121. https://doi.org/10.1016/J.IJFORECAST.2012.06.004

Georgakakos, K.P., 1986. A generalized stochastic hydrometeorological model for flood and flash-flood forecasting: 1. Formulation. Water Resour. Res. 22, 2083–2095. https://doi.org/10.1029/wr022i013p02083

Gneiting, T., Raftery, A.E., Westveld, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133, 1098–1118. https://doi.org/10.1175/MWR2904.1

Gouweleeuw, B.T., Thielen, J., Franchello, G., J de Roo, A.P., Buizza, R., Buizza Flood, R., Gouweleeuw, B., De Roo, A., 2005. Flood forecasting using medium-range probabilistic weather prediction, Hydrology and Earth System Sciences.

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

Han, S., Coulibaly, P., 2019. Probabilistic flood forecasting using hydrologic uncertainty processor with ensemble weather forecasts. J. Hydrometeorol. JHM-D-18-0251.1. https://doi.org/10.1175/JHM-D-18-0251.1

Han, S., Coulibaly, P., Biondi, D., 2019. Assessing hydrologic uncertainty processor performance for flood forecasting in a semiurban watershed. J. Hydrol. Eng.

Hopson, T.M., Webster, P.J., 2010. A 1–10-Day ensemble forecasting scheme for the major river basins of Bangladesh: forecasting severe floods of 2003–07*. J. Hydrometeorol. 11, 618–641. https://doi.org/10.1175/2009JHM1006.1

Horat, C., Antonetti, M., Liechti, K., Kaufmann, P., Zappa, M., 2018. Ensemble flood forecasting considering dominant runoff processes: II. Benchmark against a state-of-the-art model-chain. Nat. Hazards Earth Syst. Sci. Discuss. 1–34. https://doi.org/10.5194/nhess-2018-119

Hrachowitz, M., Clark, M.P., 2017. HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. Hydrol. Earth Syst. Sci. 21, 3953–3973. https://doi.org/10.5194/hess-21-3953-2017

Huber, W.C., Dickinson, R.E., 1988. Storm water management model, version 4: user's manual. Athens, GA. https://doi.org/EPA/600/3-88/001a

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688. https://doi.org/10.1016/J.IJFORECAST.2006.03.001

Jakeman, A.J., Littlewood, I.G., Whitehead, P.G., 1990. Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. J. Hydrol. 117, 275–300. https://doi.org/10.1016/0022-1694(90)90097-H

Jasper, K., Gurtz, J., Lang, H., 2002. Advanced flood forecasting in Alpine watersheds by coupling meteorological observations and forecasts with a distributed hydrological model. J. Hydrol. 267, 40–52. https://doi.org/10.1016/S0022-1694(02)00138-5

Jeong, D.-I., Kim, Y.-O., 2005. Rainfall-runoff models using artificial neural networks for ensemble streamflow prediction. Hydrol. Process. 19, 3819–3835. https://doi.org/10.1002/hyp.5983

Jordan, S.J., Vivian, A., Wohar, M.E., 2017. Forecasting market returns: bagging or combining? Int. J. Forecast. 33, 102–120. https://doi.org/10.1016/J.IJFORECAST.2016.07.003

Kouwen, N., 2019. Personal communication.

Kouwen, N., 2018. WATFLOOD® / CHARM® Canadian hydrological and routing model: User Manual. Waterloo.

Kouwen, N., 1988. WATFLOOD: a micro-computer based flood forecasting system based on real-time weather radar. Can. Water Resour. J. 13, 62–77. https://doi.org/10.1007/s10228-005-0319-x

Kouwen, N., Soulis, E.D., Pietroniro, A., Donald, J., Harrington, R.A., 1993. Grouped response units for distributed hydrologic modelling By N . Kouwen , l Member , ASCE , E . D . Soulis , 2 A . Pietronirofl J . Donald , 4 captured using small subbasin elements often called hydrologic response units (HRUs) (Leavesley and Stannar. J. Water Resour. Plan. Manag. 119, 289–305.

Lardet, P., Obled, C., 1994. Real-time flood forecasting using a stochastic rainfall generator. J. Hydrol. 162, 391–408. https://doi.org/10.1016/0022-1694(94)90238-0

Leach, J.M., Kornelsen, K.C., Coulibaly, P., 2018. Assimilation of near-real time data products into models of an urban basin. J. Hydrol. 563, 51–64.

https://doi.org/10.1016/J.JHYDROL.2018.05.064

Lindenschmidt, K.-E., Rokaya, P., Das, A., Li, Z., Richard, D., 2019. A novel stochastic modelling approach for operational real-time ice-jam flood forecasting. J. Hydrol. 575, 381–394. https://doi.org/10.1016/J.JHYDROL.2019.05.048

Matsypura, D., Thompson, R., Vasnev, A.L., 2018. Optimal selection of expert forecasts with integer programming. Omega 78, 165–175. https://doi.org/10.1016/J.OMEGA.2017.06.010

Mazenc, B., Sanchez, M., Thiery, D., 1984. Analyse de l'influence de la physiographie d'un bassin versant sur les paramètres d'un modèle hydrologique global et sur les débits caractéristiques à l'exutoire. J. Hydrol. 69, 97–118. https://doi.org/10.1016/0022-1694(84)90158-6

Milbrandt, J.A., Bélair, S., Faucher, M., Vallée, M., Carrera, M.L., Glazer, A., 2016. The pan-Canadian high resolution (2.5 km) deterministic prediction system. Weather Forecast. 31, 1791–1816. https://doi.org/10.1175/WAF-D-16-0035.1

Montreal Gazette, 2019. Quebec flooding News, Articles & Images | Montreal Gazette [WWW Document]. URL https://montrealgazette.com/tag/flood (accessed 5.24.19).

Moore, R.J., Clarke, R.T., 1981. A distribution function approach to rainfall runoff modeling. Water Resour. Res. 17, 1367–1382. https://doi.org/10.1029/WR017i005p01367

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 (Switzerland) 10. 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. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nielsen, S.A., Hansen, E., 1973. Numerical simulation of the rainfall-runoff process on a daily basis, Nordic Hydrology.

Pappenberger, F., Bartholmes, J., Thielen, J., Cloke, H.L., Buizza, R., de Roo, A., 2008. New dimensions in early flood warning across the globe using grand-ensemble weather predictions. Geophys. Res. Lett. 35, 1–7. https://doi.org/10.1029/2008GL033837

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

Pinto, J.O., Grim, J.A., Steiner, M., Pinto, J.O., Grim, J.A., Steiner, M., 2015. Assessment of the high-resolution rapid refresh model's ability to predict mesoscale convective systems using object-based evaluation. Weather Forecast. 30, 892–913. https://doi.org/10.1175/WAF-D-14-00118.1

Razavi, T., Coulibaly, P., 2017. An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds. Can. Water Resour. Journal// Rev. Can. des ressources hydriques 42, 2–20. https://doi.org/10.1080/07011784.2016.1184590

Razavi, T., Coulibaly, P., 2016. Improving streamflow estimation in ungauged basins using a multi-modelling approach. Hydrol. Sci. 61, 2668–2679. https://doi.org/10.1080/02626667.2016.1154558

Richardson, D.S., 2006. Predictability and economic value. Predict. Weather Clim. 9780521848, 628–644. https://doi.org/10.1017/CBO9780511617652.026

Robson, A.J., Moore, R.J., Wells, S.C., Rudd, A., Cole, S.J., Mattingley, P.S., 2017. Understanding the performance of flood forecasting models. Environment Agency, Bristol, UK.

Rossman, L., Huber., W., 2015. Storm water management model reference manual: Volume I-Hydrology (Revised). Washington, DC. https://doi.org/EPA/600/R-15/162A

Roulin, E., 2006. Skill and relative economic value of medium-range hydrological ensemble predictions. Hydrol. Earth Syst. Sci. Discuss. 3, 1369–1406. https://doi.org/10.5194/hessd-3-1369-2006

Samuel, J., Coulibaly, P., Metcalfe, R.A., 2011. Estimation of continuous streamflow in Ontario ungauged basins: comparison of regionalization methods. J. Hydrol. Eng. 16, 447–459. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000338

Sandink, D., 2016. Urban flooding and ground-related homes in Canada: an overview. J. Flood Risk Manag. 9, 208–223. https://doi.org/10.1111/jfr3.12168

Seibert, S.P., Ehret, U., Zehe, E., 2016. Disentangling timing and amplitude errors in streamflow simulations. Hydrol. Earth Syst. Sci. 20, 3745–3763. https://doi.org/10.5194/hess-20-3745-2016

Seiller, G., Hajji, I., Anctil, F., 2015. Improving the temporal transposability of lumped hydrological models on twenty diversified U.S. watersheds. J. Hydrol. Reg. Stud. 3, 379–399. https://doi.org/10.1016/j.ejrh.2015.02.012

Shamir, E., Carpenter, T., Fickenscher, P., 2006. Evaluation of the National Weather Service operational hydrologic model and forecasts for the American River basin. J. Hydrol. Eng. 11, 392–407. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:5(392)

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 Resour. Res. 55, 1510–1530. https://doi.org/10.1029/2018WR023197

Statistics Canada, 2019. Impact of spring flooding in key areas across Canada [WWW Document]. URL https://www150.statcan.gc.ca/n1/en/daily-quotidien/190517/dq190517a-eng.pdf?st=LnI5eFBE

SUGAWARA, M., 1979. Automatic calibration of the tank model / L'étalonnage automatique d'un modèle à cisterne. Hydrol. Sci. Bull. 24, 375–388. https://doi.org/10.1080/02626667909491876

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. 106, 7183–7192. https://doi.org/10.1029/2000JD900719

Thiboult, A., Anctil, F., Boucher, M.A., 2016. Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrol. Earth Syst. Sci. 20, 1809–1825. https://doi.org/10.5194/hess-20-1809-2016

Thiboult, A., Anctil, F., Ramos, M.H., 2017. How does the quantification of uncertainties affect the quality and value of flood early warning systems? J. Hydrol. 551, 365–373. https://doi.org/10.1016/j.jhydrol.2017.05.014

Thiboult, A., Seiller, G., Anctil, F., 2019. HOOPLA v1.0.1. https://doi.org/10.5281/ZENODO.2653969

Thielen, J., Bartholmes, J., Ramos, M.-H., de Roo, A., 2009a. The European flood alert system – Part 1: Concept and development. Hydrol. Earth Syst. Sci. Discuss. 13, 125–140. https://doi.org/10.5194/hessd-5-257-2008

Thiéry, D., 1982. Utilisation d'un modèle global pour identifier sur un niveau piézométrique des influences multiples dues à diverses activités humaines., in: Exeter Symposium, 1982, Improvement of Methods of Long Term Prediction of Variations in Groundwater Resources and Regimes Due to Human Activity. p. pp--71.

Thirumalaiah, K., Deo, M.C., 1998. Real-time flood forecasting using neural networks. Comput. Civ. Infrastruct. Eng. 13, 101–111. https://doi.org/10.1111/0885-9507.00090

Tolson, B.A., Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resour. Res. 43, 1–16. https://doi.org/10.1029/2005WR004723

TRCA, 2013. Humber River Watershed Report Card 2013 [WWW Document]. URL https://trca.ca/wp-content/uploads/2016/04/2173_WatershedReportCards_Humber_rev15_forWeb.pdf (accessed 4.12.17).

Unduche, F., Tolossa, H., Senbeta, D., Zhu, E., 2018. Evaluation of four hydrological models for operational flood forecasting in a Canadian Prairie watershed. Hydrol. Sci. J. 63, 1–17. https://doi.org/10.1080/02626667.2018.1474219

Velázquez, J.A., Anctil, F., Perrin, C., 2010. Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments. Hydrol. Earth Syst. Sci. 14, 2303–2317. https://doi.org/10.5194/hess-14-2303-2010

Verkade, J.S., Werner, M.G.F., 2011. Estimating the benefits of single value and probability forecasting for flood warning. Hydrol. Earth Syst. Sci. 15, 3751–3765.

https://doi.org/10.5194/hess-15-3751-2011

Vieux, B.E., Cui, Z., Gaur, A., 2004. Evaluation of a physics-based distributed hydrologic model for flood forecasting. J. Hydrol. 298, 155–177. https://doi.org/10.1016/J.JHYDROL.2004.03.035

Vrugt, J.A., Gupta, H. V., Dekker, S.C., Sorooshian, S., Wagener, T., Bouten, W., 2006. Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model. J. Hydrol. 325, 288–307. https://doi.org/10.1016/J.JHYDROL.2005.10.041

Warmerdam, P.M.M., Kole, J., 1997. Modelling rainfall runoff processes in the Hupselse Beek Research basin, in: Ecohydrological Processes in Small Basins: Proceedings of the Strasbourg Conference, Strasbourg, 1996. pp. 155–161.

Weber, F., Perreault, L., Fortin, V., 2006. Measuring the performance of hydrological forecasts for hydropower production at BC Hydro and Hydro-Quebec. 86th AMS Annu. Meet.

Weeink, W.H. a, 2010. Thresholds for flood forecasting and warning evaluation of streamflow and ensemble thresholds. University of Twente.

Wilks, D., 2006. Statistical Methods in the Atmospheric Sciences, 2nd ed, International Geophysics Series. Acadamic Press.

Wu, H., Adler, R.F., Hong, Y., Tian, Y., Policelli, F., Wu, H., Adler, R.F., Hong, Y., Tian, Y., Policelli, F., 2012. Evaluation of global flood detection using satellite-based rainfall and a hydrologic model. J. Hydrometeorol. 13, 1268–1284. https://doi.org/10.1175/JHM-D-11-087.1

Yan, H., Gallus, W.A., Yan, H., Jr., W.A.G., 2016. An Evaluation of QPF from the WRF, NAM, and GFS models using multiple verification methods over a small domain. Weather Forecast. 31, 1363–1379. https://doi.org/10.1175/WAF-D-16-0020.1

Yilmaz, K.K., Adler, R.F., Tian, Y., Hong, Y., Pierce, H.F., 2010. Evaluation of a satellite-based global flood monitoring system. Int. J. Remote Sens. 31, 3763–3782. https://doi.org/10.1080/01431161.2010.483489

Yucel, I., Onen, A., Yilmaz, K.K., Gochis, D.J., 2015. Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall. J. Hydrol. 523, 49–66. https://doi.org/10.1016/j.jhydrol.2015.01.042

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., Mylne, K., Zhu, Y., Toth, Z., Wobus, R., Richardson, D., Mylne, K., 2002. The Economic value of ensemble-based weather forecasts. Bull. Am. Meteorol. Soc. 83, 73–83. https://doi.org/10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2

Zsótér, E., Pappenberger, F., Smith, P., Emerton, R.E., Dutra, E., Wetterhall, F., Richardson, D., Bogner, K., Balsamo, G., 2016. Building a Multimodel Flood Prediction System with the TIGGE Archive. J. Hydrometeorol. 17, 2923–2940. doi.org/10.1175/JHM-D-15-0130.1.

Chapter 5. Verification of Numerical Weather Predictions across Canada for Hydrologic Forecasting

Summary of Paper 4: Awol, F.S., Coulibaly, P., and Tsanis, I. (2019). Verification of Numerical Weather Predictions across Canada for Hydrologic Forecasting. *Weather and Forecasting*, Under Review.

This study aims at identifying skillful Numerical Weather Predictions (NWP) in Canada's varied geographic environment for enhanced short- and medium-range hydrologic forecasting application. Verification of precipitation forecast products was performed on two domains; high-resolution and low-resolution domains. Traditional grid-to-grid, emerging precipitation object-based, and timing error metrics were used to compare five NWPs at Low-resolution and four NWPs at High-resolution based on the intensity, volume, and timing aspects.

Key findings of this research include:

- In the Low-resolution domain, GEFSv2 and GFS appeared to be better candidates to supply forecast inputs to hydrological models for week ahead outlooks.
- In the High-resolution domain, HRRR and HRDPS achieved the collective aim of matching the timing, intensity, and volume of precipitation forecasts and are hence recommended for short-term flood forecasting in urban areas.
- The timing error approach was able to provide estimates of the average timing error and percentage of non-timing errors in the verification period along the forecast horizon.

5.1. Abstract

The skill of hydrological forecasting systems depends heavily on the quality of Numerical Weather Predictions (NWPs). In an attempt to identify skillful NWPs in Canada's varied geographic environment, this research performs a comprehensive verification of several low- and high-resolution precipitation forecasts from the perspective of identifying the best candidate products for enhancing short- and medium-range hydrologic forecasting. As such, five NWPs in Low-resolution and four NWPs in High-resolution domains were compared in terms of volume, intensity, and timing accuracies at different forecast lead times. In addition to existing grid-to-grid and object-based verification metrics, a new approach to estimate the timing error is proposed in this study. In the Low-resolution domain, GEFSv2 and GFS provided better forecast skills, lower biases, and good qualities of precipitation objects for various accumulated precipitation volumes and intensities. They also attained the timing of precipitation forecasts better than other products. ECMWF and GDPS not only produced higher timing errors but also contributed to a large percentage of errors attributed to non-timing aspects (e.g., magnitude). In the High-resolution domain, HRRR was the most unbiased and accurate NWP for forecasting higher precipitation intensities and a precipitation volume accumulated over multiple hours. If a particular 6hr accumulation is intended, HRDPS was superior in forecasting different precipitation volumes. Even though NAM showed lower timing errors, it resulted in a large number of grids and verification times that have magnitude related errors. For the combined objective of timing, intensity, and volume of precipitation forecasts, both HRRR and HRDPS performed well. Overall, the verification analysis identified candidate NWPs in various

geographic regions, which could be used in operational hydrology particularly for forecasting the volume, intensity, and timing of floods.

5.2. Introduction

Numerical Weather Predictions (NWPs) are essential inputs to hydrological models used in river flow and flood forecasts. Typically, Quantitative Precipitation Forecasts (QPF) from NWPs are used as a forcing input to hydrological models for simulating and forecasting discharges at different lead times in addition to cascading the inherited uncertainty from initial atmospheric conditions (Pappenberger et al., 2005).

Advance in hydrometeorological research has led to the development of various NWP products. The types of NWPs depend on the scale (Global, Continental, and Regional), forecast length (Long-, Medium-, Short- and Very Short-ranges), spatial resolution (Low- and High-resolutions), and characteristics (Deterministic and Ensemble). The skill and quality of meteorological variables are influenced by this variability of NWP types, which will also affect the forecast skill and reliability of hydrological forecasts. In order to increase the performances of NWPs, some systematic improvements have been made. For example, the deterministic Global Forecast System (GFS), which is developed by National Centers for Environmental Prediction (NCEP), has undergone significant changes through time to circumvent problems such as excessive grid-scale precipitation forecasts through vertical diffusion shallow convection scheme (Han and Pan, 2011).

The chaotic nature of the atmospheric system and its approximate representation by NWPs create uncertainties in deterministic forecasts (Cuo et al., 2011), which led to the development of ensemble forecasts that have different perturbation methods of initial conditions (Buizza et al., 2005). Among the Global ensemble products, the ensemble National Centers for Environmental Prediction (NCEP) forecasts and the ensemble

European Centre for Medium-Range Weather Forecasts (ECMWF) have relatively same spatial resolutions, provide medium-range forecasts (up to 16 days), have been operational for over a decade, are publicly archived and commonly used. Data from several ensemble NWPs across the Globe have been archived in The Observing System Research and Predictability Experiment Interactive Grand Ensemble Program (TIGGE) platform (Bougeault et al., 2010). Comparison has been made between the TIGGE ensemble NWPs. Hagedorn et al., (2012), for example, compared all TIGGE ensemble forecasts and highlighted that using the leading four ensemble NWPs (ECMWF, NCEP, ensembles forecasts issued by Meteorological Service of Canada (CMC) and UK's MetOffice) provided better performances. Buizza et al., (2005), also compared ECMWF, NCEP, and CMC and concluded that ECMWF has a better overall forecast skill followed by NCEP. The above two verifications on TIGGE were performed using variables other than precipitation.

Among the variables generated by NWPs, precipitation forecasts have been the main challenge, i.e. achieving the correct intensity, location, and timing of storms (Cuo et al., 2011). Particularly, summer precipitation forecasts by mesoscale NWPs were deemed to be difficult due to the nature of localized convective thunderstorms (Kaufmann et al., 2003). Golding, (2000), suggested that these problems have direct consequences on flood forecasting and should be dealt with critically. The author highlighted that the quality of precipitation forecast rates coupled with the catchment size and response time should be a primary requirement for flood prediction. The above challenges have contributed to the development of regional-scale high-resolution NWPs. For example, the North American

Mesoscale (NAM) and Rapid Refresh (RAP) developed by NCEP and Regional Deterministic Precipitation System (RDPS) provided by Environment Canada issue short-range deterministic forecasts at grid resolutions of approximately 12 km in North America. Even further, very short-range hourly precipitation forecasts from High-Resolution Rapid Refresh (HRRR) and High-Resolution Deterministic Precipitation System (HRDPS) were made operational at finer resolution (~2-3km) in USA and Canada.

With the increasing number of various scales of spatial and temporal resolutions of NWPs, verification of the forecasts in terms of the hydrological implications is deemed to be essential (Cuo et al., 2011). Numerous traditional and innovative verification methods were developed including grid-to-grid (Jolliffe and Stevenson, 2011; Wilks, 2006), spatial neighborhood (Ebert, 2009; Roberts and Lean, 2008), wavelet-based (Casati et al., 2004), and object- or feature-based (Davis et al., 2006a; Li et al., 2015) methods. Wolff et al., 2014, for example, compares QPFs from GFS and NAM using traditional grid-to-grid, spatial neighborhood, and object-based methods over the CONUS (Continental United States). They highlighted that more diagnostic features of precipitation forecasts such as spatial scale, coverage area, displacement, and angular orientation were resolved by using the newer verification metrics, particularly when mid- and course resolution NWPs are used. They indicated that the Method for Object-Based Diagnostic Evaluation (MODE) tool (Davis et al., 2006a, 2006b) was able to identify the lower performance of NAM forecast objects, in which the neighborhood verification method could not show or gave otherwise a higher skill score.

Yan et al., (2016), evaluated QPFs from Advanced Research-Weather Research and Forecasting (WRF-ARW), NAM, and GFS models in a small and high-resolution domain over Iowa. Their research was intended to assess the quality of very short-term precipitation forecasts (up to 12hr) that can be used as inputs to hydrological models. They found out that WRF showed better skill scores and recorded fewer errors in intensity, displacement, and areal coverage of identified precipitation systems; NAM poorly performed in the verification metrics; significant location errors were observed from late morning to afternoon in all models. One of their essential recommendations for hydrological application is to give more attention to the location errors as timing might be a factor because predicted storms were possibly shifted before or after the localized observed storms. The errors and uncertainties of precipitation forecasts are replicated in hydrological and flood forecasting (Cuo et al., 2011; Zappa et al., 2011). Ensemble QPFs has been used as an input to hydrological models to address issues related to uncertainty (Brown et al., 2012; Georgakakos and Krzysztofowicz, 2001; Han and Coulibaly, 2019; Mascaro et al., 2010; Roulin, 2006; Verkade et al., 2017).

According to Pagano et al., 2014, some of the challenges in operational river forecasting and hydrological predictions are: 1) most high-spatial resolution NWPs provide only short-range forecasts, 2) NWPs with ensemble forecasts have mostly low-spatial resolutions, and 3) precipitation forecast verifications were usually performed on large scales in which the significance to local scale flooding and hydrological studies are negligible. Given the above challenges, it is essential that available NWPs should be evaluated at various spatial scales and domains in order to identify appropriate NWPs that provides better precipitation

forecasts for flood forecasting. This is mainly because river flood forecasting depends on catchment size, catchment characteristics, time of concentration, and the spatial and temporal resolution of NWPs. Small and urban catchments that typically have shorter times of concentration require high-resolution short-range precipitation forecasts, preferably with hourly or sub-hourly time steps. On the other hand, NWPs that can issue medium-range forecasts are vital for large catchments because response times in these catchments usually take from a couple of days up to a week. As such, the quality of the precipitation forecasts should be verified at various climatological and physiological domains within the same region or continent where the NWPs are developed or issued. In previous researches, only a few studies were performed to comprehensively evaluate available NWP products at different watershed characteristics specifically intended to improve flood forecasting. Canada is a good example where watersheds with a wide range of climatological and physical characteristics exist, and numerous Regional and Global NWP products are available. It has several watersheds with diverse landscape types ranging from mountains, forests, Prairies and water bodies, and agricultures in Western and Central Provinces to semi-urban and urban in Southern Ontario (Zahmatkesh et al., 2019). The underlying research question is: Which weather forecast products provide better precipitation forecasts to enhance flood forecasting across Canada's varied watershed characteristics? Accordingly, the objective of this study is:

- ✓ To evaluate and identify better Precipitation forecasts products for
  - an improved medium-range flood forecasting in large watersheds, &
  - an improved short-range flood forecasting in small urbanizing watersheds.

- To evaluate precipitation forecasts in terms of volume, intensity, and timing error aspects that have direct practical significance in flood forecasting.

From a hydrological point of view, key contribution from this research is that candidate precipitation forecasts products can be used as input to hydrological models to evaluate the skill, reliability, and overall quality of flood forecasting.

5.3. Study Domains

To address both the diversity of the Canadian hydro-climatic regions and the various spatial and temporal resolutions of NWPs, the selected study areas contain two main domains: Low-resolution and High-resolution domains (Figure 5-1). The Low-resolution domain covers parts of Western and Central Canadian Provinces (BC, AL SK, and MN), and some parts of Northern USA where most of the watersheds are relatively large, often comprised of agricultural lands, forests, prairie and wetlands, and transboundary. The High-resolution domain covers parts of Southern Ontario where several small urban and semi-urban catchments exist. Some flood forecasting studies were conducted in watersheds located in the former domain (Muhammad et al., 2018a; Unduche et al., 2018) as well as the later domain (Awol et al., 2018; Han and Coulibaly, 2019; Leach et al., 2018). The term “Low” is assigned for the former domain to indicate the lower spatial resolution of NWPs that are used for the area to issue medium-range forecasts (daily up to a week ahead forecast with about 50km horizontal resolution). Similarly, the term “High” is given to the later domain to indicate the higher spatial and temporal resolution NWPs that are appropriate for smaller urban watersheds where the time of concentration is usually short (NWPs that issue hourly

forecasts up to 24 hours with 2-13km resolution). More detail on the type of forecast data used is found in Section 5.4.1.

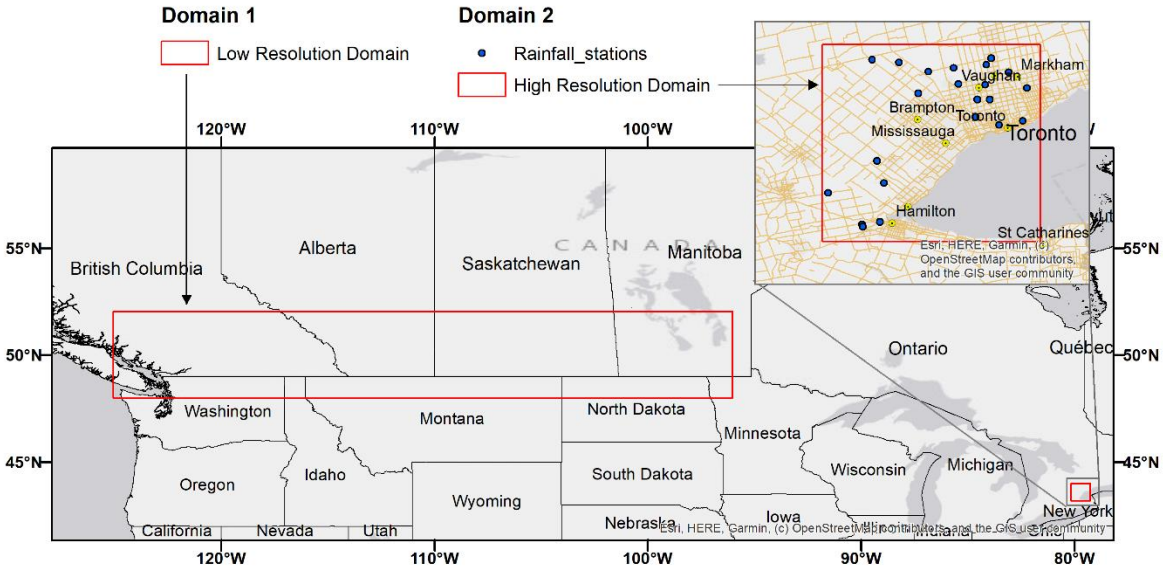

Figure 5-1: Location of study domains

5.4. Data

5.4.1. Forecast data

In the Low-resolution domain, available NWPs that can issue medium-range precipitation forecasts were collected. Table 5-1 presents the detailed features of the products. Ensemble forecasts were available from the European Centre for Medium-Range Weather Forecasts (ECMWF: 51 members), Global Ensemble Forecast System from National Centers for Environmental Prediction (NCEP: 21 members), and the second-generation Global Ensemble Forecast System version 2 reforecasts (GEFSv2: 11 members). Deterministic forecasts from the Global Deterministic Prediction System (GDPS) and Global Forecast

System (GFS) were also collected. The means of the ensemble forecasts were used in this study for comparison of performances with the deterministic ones.

For the High-resolution domain, four Regional and Continental scale NWPs that produce hourly forecasts with a relatively higher spatial resolution were collected (Table 5-2). These are the High-Resolution Deterministic Precipitation System (HRDPS), High-Resolution Rapid Refresh (HRRR), North American Mesoscale Forecast System (NAM), and Rapid Refresh (RAP).

5.4.2. Verification data

To verify the precipitation forecasts, two verification datasets were used. In the Low-resolution domain, the Canadian Precipitation Analysis (CaPA) is used. CaPA is a gridded reanalysis product estimated by statistical interpolation of observed precipitation data from radar and ground-based rain gauges as well as a background field from short-range precipitation forecasts (Mahfouf et al., 2007). CaPA has a spatial resolution of 15 km and a temporal resolution of 6 hours. For High-resolution domain, 3km-gridded hourly data is created in this study for verification of NWPs. The gridded time series is created by interpolating point precipitation data using bilinear interpolation technique. Hourly precipitation point data were collected from 18 meteorological stations spatially distributed across the High-resolution domain (Figure 5-2).

Table 5-1: Numerical Weather Predictions used in the Low-resolution domain

NWPs	type	Spatial resolution	Temporal resolution	Forecast length	Providing organization	Reference	Archive Source	Notes
ECMWF	Ensemble	~50km	6 hourly	15 days	ECMWF	(Molteni et al., 1996)	(Bougeault et al., 2010)	Publicly available version is described and used here
GDPS	Deterministic	~33km	3 hourly	10 days	Environment Canada	(Mai et al., 2019)	CaSPAr[1]	
GEFSv2	Ensemble	~50km	3 hourly	8 days	NOAA	(Hamill et al., 2013)	ESRL[2]	Forecast available from 8days to 16days at ~100km resolution
NCEP	Ensemble	~50km	6 hourly	16 days	NOAA	(Toth et al., 1993)	(Bougeault et al., 2010)	
GFS	Deterministic	~27km	3 hourly	10 days	NOAA	(Han and Pan, 2011)	(NOAA, 2015)	Forecast available from 10days to 16days at 12 hourly step

[1] https://caspar-data.ca/
[2] https://www.esrl.noaa.gov/psd/forecasts/reforecast2/download.html

Table 5-2: Numerical Weather Predictions used in the High-resolution domain

NWPs	type	Spatial resolution	Temporal resolution	Forecast length	Providing organization	Reference	Archive Source	Notes
HRDPS	Deterministic	~2.5km	hourly	48 hours	Environment Canada	(Mai et al., 2019)	CaSPAr	
HRRR	Deterministic	~3km	hourly	18 hours	NOAA	(Pinto et al., 2015)	(Blaylock et al., 2017)	
NAM	Deterministic	~12km	hourly	36 hours	NOAA	(Rogers et al., 2009)	NOMADS[1]	Forecast available from 36h to 84h in 3 hourly step
RAP	Deterministic	~13km	hourly	21 hours	NOAA	(Benjamin et al., 2016)	NOMADS	

[1] https://nomads.ncdc.noaa.gov/data/

5.5. Methodology

In order to evaluate the forecast accuracy and skill of five NWPs in the Low-resolution domain and four NWPs in High-resolution domain, different verification metrics were applied. The metrics were used to assess the volume, intensity, and timing error of precipitation forecasts. There are several spatial verification methods in the literature. In this study, the traditional grid-to-grid and evolving object-based verification methods were applied to evaluate the volume and intensity of precipitation forecasts. In addition to these metrics, a simple algorithm is developed in this study to estimate the Timing error of precipitation forecasts, which will be discussed in detail in Section 5.5.3.

The verification period for Low-resolution domain NWPs was from 2018-05-01 to 2018-11-30, in which forecasts were issued daily for 1 day up to 8 days lead times. For NWPs in High-resolution domain, the verification period was between 2018-07-01 and 2018-11-30, with forecasts issued daily from 1 hour up to 18 hours forecast lead times. For all NWPs, forecasts initialized at 00UTC every day were used. Similar initialization time was applied by Wolff et al., 2014. The verification periods and forecast horizons are chosen based on the availability of archives of all NWPs, which was aligned with the summer and fall flood periods. Also, an increasing trend of precipitation intensity and volume were detected in the past and will likely continue in these seasons for some parts of North America (Cooley and Chang, 2017).

For this research, the grid-to-grid and object-based verification metrics are implemented using the Model Evaluation Tool (MET), a verification software platform developed at the National Center for Atmospheric Research (NCAR) (Brown et al., 2009).

5.5.1. Traditional Grid-to-grid Verification

Several verification metrics have been developed on grid-to-grid comparison of observed and forecasted precipitation. The primary concept of these traditional metrics relies on developing a Contingency Table (Table 5-3) at different precipitation thresholds and forecast lead times. This table is created by using the number of hits, false alarms, correct negatives, and misses of forecast-observation (Wilks, 2006). In this study, two metrics are used to evaluate the bias and skill of forecasts: Frequency Bias (FBIAS) and Gilbert Skill Score (GSS), respectively.

Table 5-3: Contingency table and associated parameters for calculating traditional verification metrics

		Observation		
		Yes	No	∑
Forecast	Yes	a (hits)	b (false alarms)	a+b
	No	c (misses)	d (correct negatives)	c+d
	∑	a+c	b+d	

Frequency Bias (FBIAS)

Frequency bias measures the "over forecasting: >1" or "under forecasting: <1" tendency of a categorical forecast (Wilks, 2006). FBIAS is estimated by the ratio of the average forecast (forecasted 'yes's) over the average observation (observed 'yes's) at particular precipitation threshold, and in terms of the elements of Table 5-3:

$$FBIAS = \frac{a-b}{a+c} \qquad (5\text{-}1)$$

Gilbert Skill Score (GSS)

Gilbert Skill Score, also called Equitable Threat Score, measures the accuracy of forecasts by comparing the correctly predicted events with a random chance. It is estimated by equation 5-2. In this study, the bias-adjusted GSS is used based on Brill et al., (2009). GSS ranges vary from -1/3 to 1, and the value of 1 indicates a perfect forecast.

$$GSS = \frac{a + a_{random}}{a+b+c-a_{random}}$$
$$a_{random} = \frac{(a+b)(a+c)}{a+b+c+d} \qquad (5\text{-}2)$$

5.5.2. Object-based Verification (MODE)

Method for Object-based Diagnostic Evaluation (MODE) is one of the spatial verification techniques used to provide a feature-based comparison between observation and forecasts (Davis et al., 2006a). The main procedures in MODE are outlined below, and more details on the application can be obtained from Davis et al., (2006b, 2009) and Wolff et al., (2014):

- Objects (forecast and observation) are identified using two user-defined parameters (Convolution Threshold: precipitation threshold, and Convolution Radius: the number of grids used to smoothen the thresholded grid data). Various ranges of parameters are used in this study to examine the sensitivity to the identified objects' attributes.

- Different attributes are estimated for the single objects and forecast-observation pairs (e.g., Precipitation object area, object count, centroid distance, angle difference).
- Using fuzzy logic and weights assigned to object attributes, the total interest value is estimated to merge and match different single and paired objects. In this study, the recommended default weights are used for object attributes to calculate the total interest value.
- The output statistics of single, pairs, and matched objects are summarized over the verification period.

5.5.3. Timing Error estimation

The motivation to estimate the Timing Error came from identifying the possible cause of errors that short- or medium-range range precipitation forecasts inherited. Yan et al., (2016), suggested that the cause of location error in precipitation forecasts in some NWPs might be due to timing because predicted storms improperly produced before or after the localized observed storms. Having this concept, the following approach is proposed to approximately find where the shifted forecast event is located along the forecast horizon (forecast lead times). More precisely, the objective is to find how many forecast lead times the predicted precipitation event is shifted (displaced on the forecast horizon dimension). The comparison between forecasted and observed event is made on a grid-to-grid basis in the verification period. For a given precipitation threshold and at each lead time (t), the average timing error ($te_{avg,t}$) over the whole domain (all grids) and over the verification time period is given by:

$$te_{avg,t} = \frac{\sum_{j=1}^{M}\left[\sum_{i=1}^{N} te_{i,j,t}\right]}{\sum[\sum_{S} 1]} \qquad (5\text{-}3)$$

$$S = \{i: te_{t,i} \neq 0\}$$

Where N = no. of grids, M = no. of verification time periods, j = current verification time, i = current grid point, and $te_{i,j,t}$ (timing error for the current grid point, verification time and lead time) is estimated with the simple algorithm given in Figure 5-2. At each grid and verification time, the main steps to estimate the timing error at the current lead time are (refer the step numbers in Figure 5-2):

1. Check precipitation threshold (event) criteria (first the observation, then the forecast). i.e. check if $P > P_{thr}$ for both observation and forecast.
2. Find indexes of lead times that meet the criteria. If the current lead time meets the criteria, the timing error is set to zero ($te_t = 0$), and iteration will go to the next grid.
3. Find the minimum difference between current lead time and the lead times that meet the criteria. The main assumption made here is that, if timing error is found in the current lead time and more than one lead time meets the threshold criteria, the predicted event is most likely shifted (improperly produced) to the nearest lead time.

The timing error analysis was performed over 2016 grids and 226 days of verification times in the Low-resolution domain for each of the eight days forecast lead times. In the High-resolution domain, the timing error is analyzed over 1404 grids and 145 days of verification times for 1 hour up to 18 hours of forecast lead times.

In addition to the average timing error, the percentage of errors other than timing, or in other words, percentage of errors that came from magnitude error rather than timing error is given in order to give more insight. This percentage is calculated by counting the total number of grids and verification times that do not meet the threshold criteria while the threshold (event) was observed to occur on those grids and verification times.

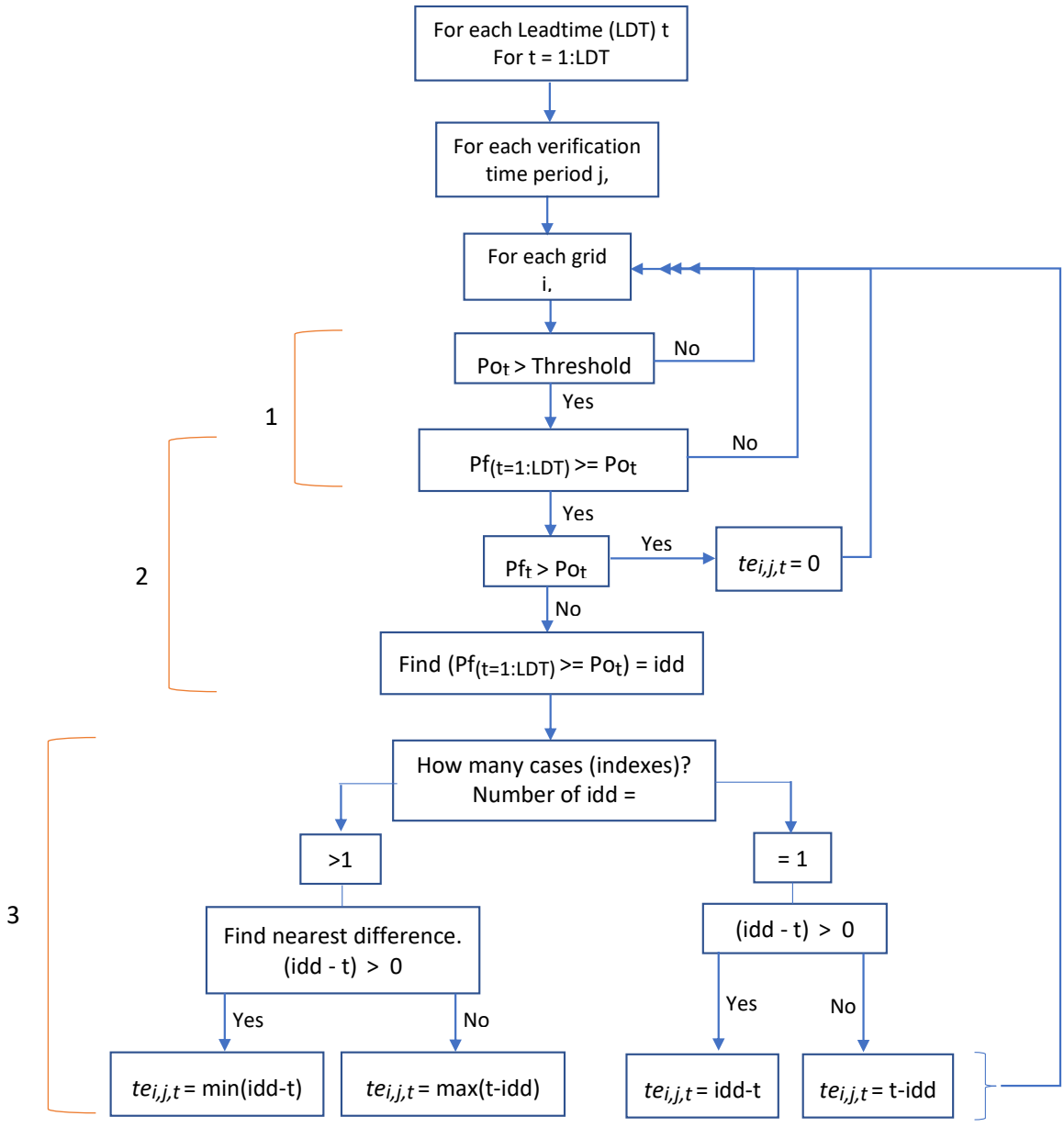

Figure 5-2: Algorithm to estimate Timing Error, $te_{i,i,t}$ (equation 5-3). Pf and Po are forecasted and observed precipitation, respectively.

5.6. Results and Discussion

5.6.1. Low-Resolution Domain

5.6.1.1. Volume error

The volume of forecasted precipitation accumulated over time for different NWPs located in the Low-resolution domain was evaluated using different verification metrics. Figure 5-3 shows sample precipitation objects identified for an event on 2018-06-01 for five NWP forecasts and CaPA observation using MODE verification. The event was forecasted by the NWPs 1 day earlier (Lead time of 1 day). In order to identify the precipitation objects, a convolution threshold of 10mm and a convolution radius of 5 grid units were used for MODE analysis. As shown in the figure, several single and paired objects were identified. It appeared that the objects identified by CaPA observation have a larger area and precipitation volume than the forecasted precipitation objects of the NWPs. The objects of GEFSv2 and NCEP were somehow matched with one of the observed objects because the area, centroid distance, and angle difference attributes were relatively better than the NWPs. The precipitation volume forecasts which were equivalent among the NWPs were significantly underestimated. Two possible reasons from this sample MODE output are: the NWPs capability in accurately predicting an extreme event was low, and the forecasted precipitation was probably shifted to the next forecast lead time(s) (presence of timing error). In general, the output indicates the need for more verification assessment for such an extreme event

Figure 5-4 presents the different verification metrics evaluated for daily accumulated volume precipitation of 10mm and 20mm exceedance thresholds. Using the traditional grid-to-grid method, the Forecast Bias and Gilbert Skill Score showed that GEFSv2 and GFS have significantly lower bias and relatively higher forecast accuracy than the other NWPs for both accumulated precipitation volumes at all lead times. The forecast bias of ECMWF, GDPS, and NCEP drastically decreased beyond two days forecast lead times regardless of the precipitation volume threshold. MODE attributes summarized over the verification period indicated that the precipitation objects identified by GEFSv2 had the lowest angular differences and centroid distance than the other NWPs, for volume above 10mm. For higher accumulated volume above 20mm, the differences between GFS, NCEP, and ECMWF could become insignificant in terms of forecast accuracy, although the former was slightly better. In general, GEFS, followed by GFS, provided an improved precipitation forecast volume, particularly at higher accumulated thresholds.

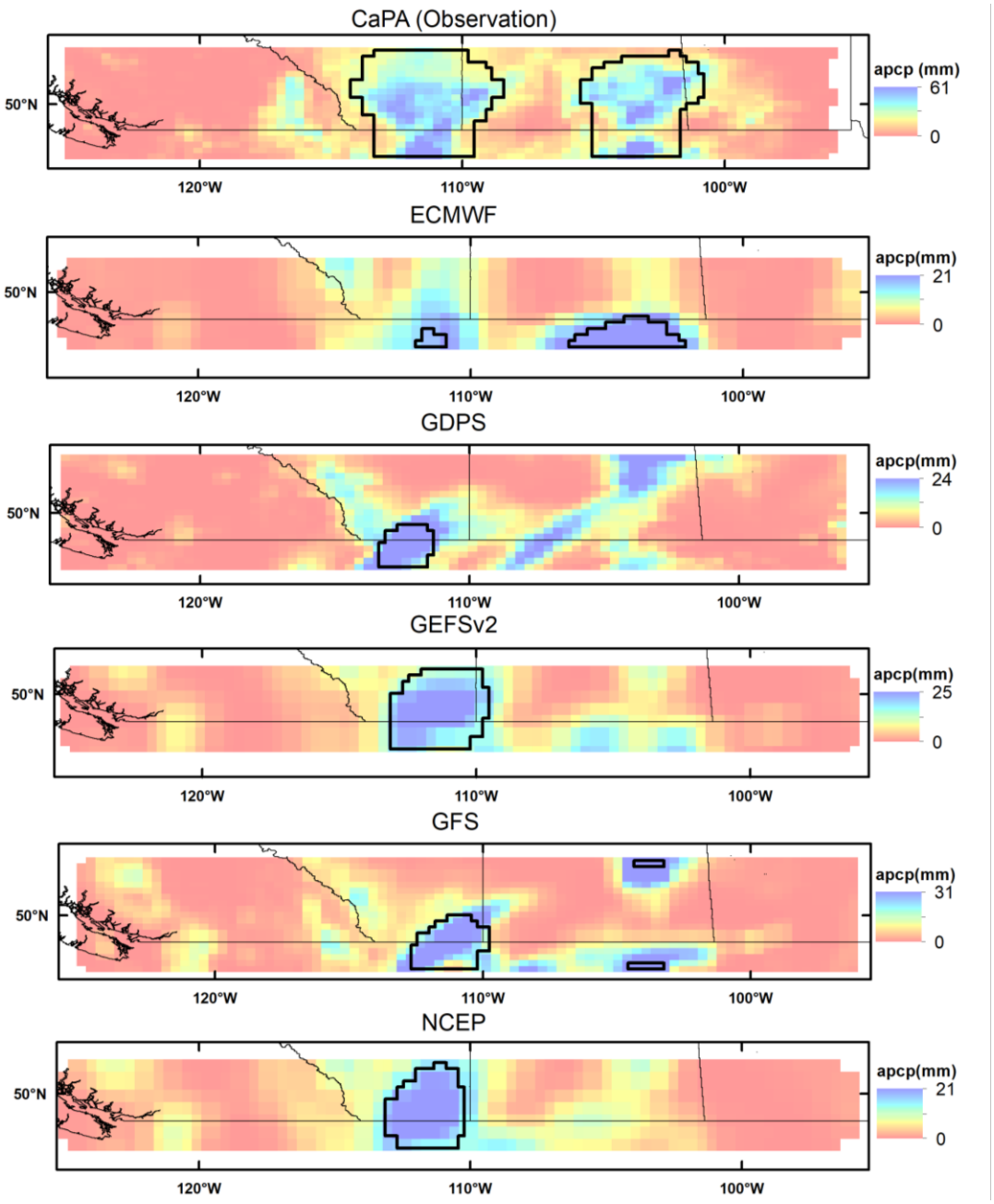

Figure 5-3: Sample precipitation objects identified on extreme event of 2018-06-01, forecasted 1 day earlier (LDT1) on Low resolution domain: MODE parameters: - Conv Thresh>=10mm, Conv Radi=5 grid units

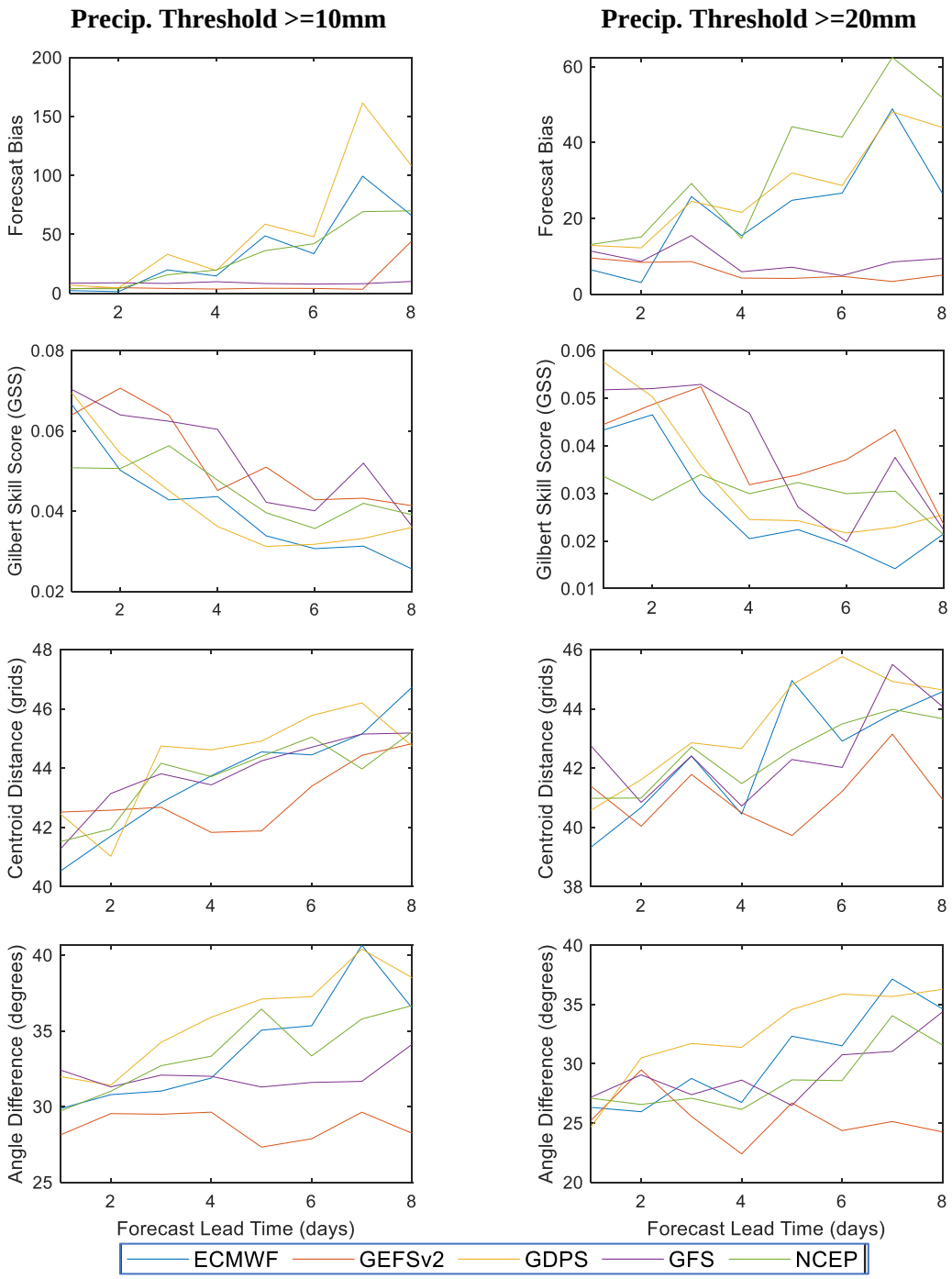

Figure 5-4: Verification metrics for evaluating different accumulated precipitation forecasts of five NWPs in Low-Resolution Domain

5.6.1.2. Intensity error

Various intensities of precipitation forecasts were evaluated in the Low-resolution domain using different methods. In many hydrological and river forecasting centers that are responsible for large catchments and rivers, short-term forecasts up to three days ahead and medium-range up to a week ahead are usually issued. The precipitation forecast intensities of different NWPs were verified on a 3-days ahead and a week ahead outlook.

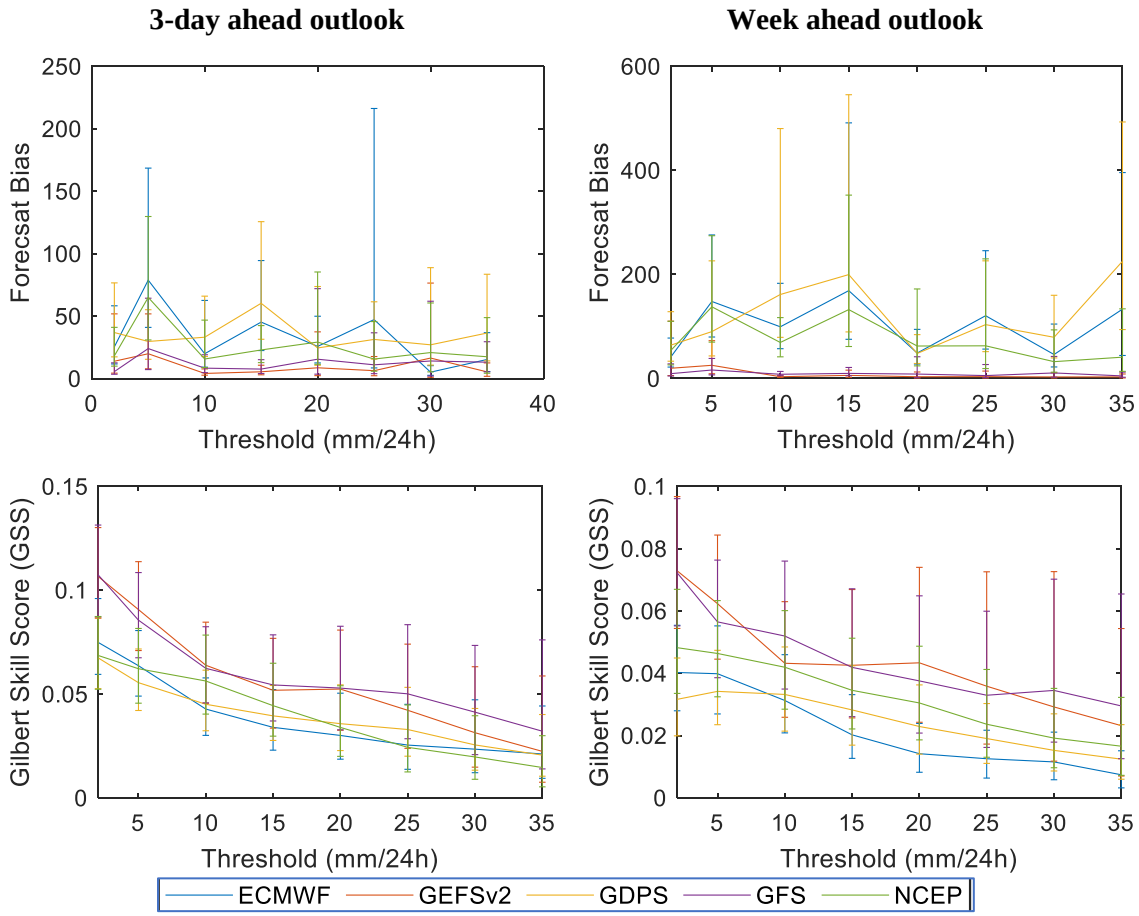

Figure 5-5: Verification metrics for evaluating different precipitation intensity forecasts of five NWPs in Low-Resolution Domain

Figure 5-5 presents the grid-to-grid verification metrics of five NWPs estimated for precipitation intensities ranging between 5 and 35 mm/day with 5mm/day increments. The forecast Bias at 3-day ahead outlook indicated that GEFSv2 and GFS produced lower biases for most precipitation intensities. Most NWPs had relatively minimal forecast biases at higher intensities above 30mm/day up to 3-day ahead forecast. For a week ahead outlook, however, ECMWF, GDPS, and NCEP significantly overestimated the forecast intensities while GEFSv2 and GFS maintained smaller forecast Biases. This result was statistically significant for GEFSv2 and GFS because the 95% confidence intervals (CI) of their Biases were narrow and closer to 1 as compared to former three NWPs, which had wider and uncertain CIs for several intensities over the forecast horizon. The forecast Skill (GSS) of GEFSv2 and GFS at all considered precipitation intensities appeared to be higher than the rest of NWPs for both 3-days and a week ahead outlook, following the similar trends observed in their forecast Bias results. As can be seen from the width of the CIs in the GSS estimates, the sampling uncertainty has increased from 3-days ahead to a week ahead forecast in most NWPs at all precipitation intensities. However, the upper, middle and lower points of the CIs in all NWP GSS estimates follow a parallel pattern showing the true GSS estimates of GEFSv2 and GFS were still higher than the rest of the NWPs. In general, GEFSv2 and GFS were more skillful than the others at 3-days and a week ahead forecasts even considering some sample uncertainties.

Figure 5-6 and Figure 5-7 present the MODE verification metrics outputs of five NWPs located in the Low-resolution domain for 3-days ahead and week ahead outlook, respectively. Aggregated Centroid Distances and Angle Differences of identified

precipitation objects were estimated at different combinations of Convolution Thresholds (5, 10, & 15 mm/day) and Convolution Radii (5, 10, 15 grid units). As the intensity increased from 5 to 15 mm/day, the quality of the identified forecasted precipitation objects decreased in all NWPs. Similarly, as the grid smoothing resolution (Convolution Radius) increased, the performances of Centroid Distance and Angle Differences decreased in all NWPs. However, the rate of quality degradation was minimum when increasing the radius from 5 to 15 grid units as compared to increasing the threshold from 5 to 15 mm/day. Reproducing the higher precipitation threshold (above 15mm/day) was a challenge for most NWPs as can be seen from the attributes of the identified precipitation objects in 3-days and a week ahead forecast (see the Centroid Distance and Angle Difference at 15mm/day in Figure 5-6 and Figure 5-7). This challenge can be associated with one of the possible reasons why the sample identified objects of an extreme case in Figure 5-3 had low qualities. For a 3-days ahead outlook (Figure 5-6), GEFSv2 produced the lowest Centroid Distances and Angle differences for a different combination of MODE parameters. Next, GFS had relatively better attributes of precipitation objects than the other three NWPs. For a week ahead outlook (Figure 5-7), both GEFSv2 and GFS were somewhat competitive in terms of the MODE verification metrics. ECMWF and GDPS were relatively weak in generating good qualities of precipitation objects at different convolution parameters.

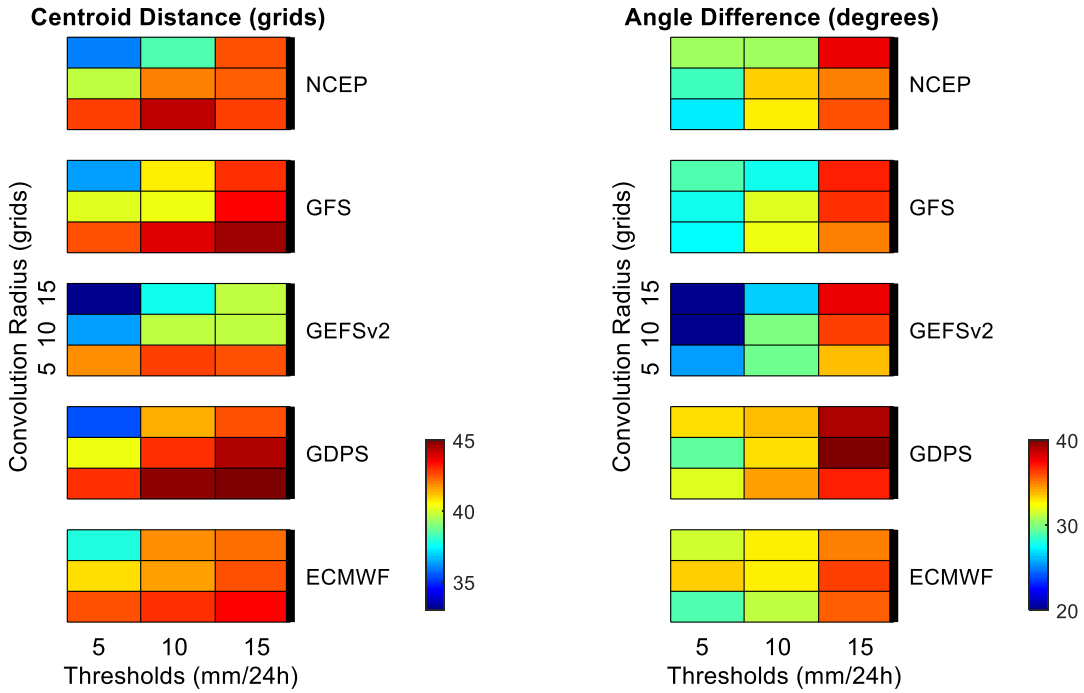

Figure 5-6: 3-days ahead forecast skill in terms of precipitation objects features at different thresholds and grid smoothing resolution (Convolution Radii (R) (5, 10, 15 grid units) and Convolution Thresholds (T) (>=5,>=10,>=15 mm))

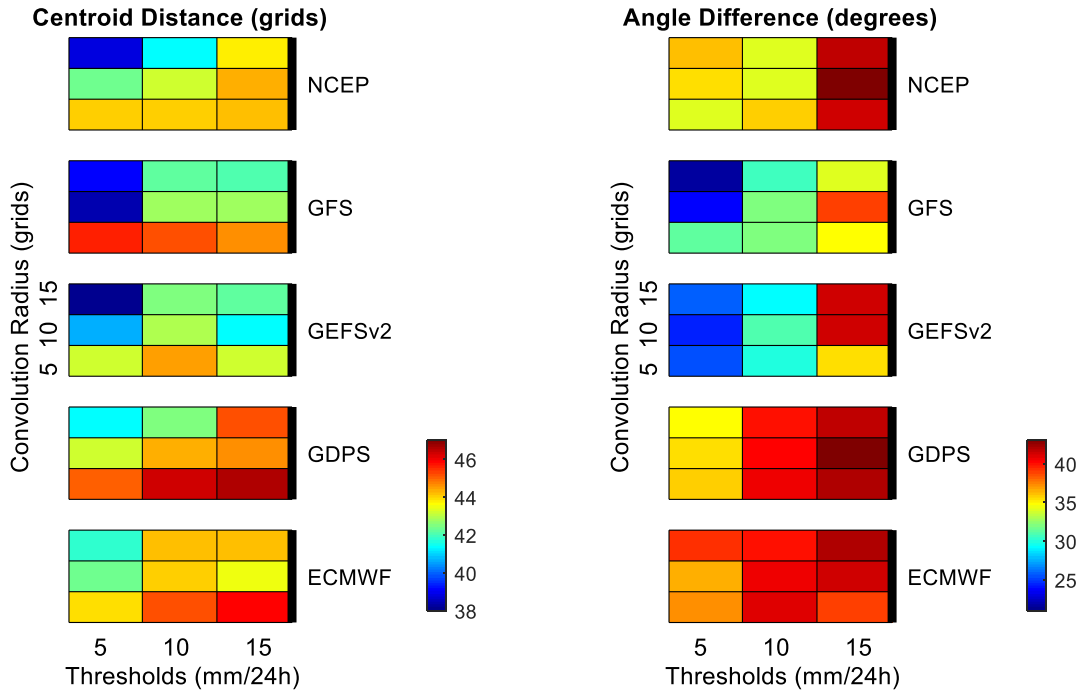

Figure 5-7: Same as Figure 5-6 but for a week ahead forecast

5.6.1.3. Timing error

Based on the approach discussed in Section 5.5.3, the average timing error of NWP forecasts was estimated for a precipitation threshold set between 10mm and 20mm for 1 day up to 8 days lead times. Figure 5-8 presents the outputs of all NWPs in the Low-resolution domain. During the first three days forecast (Figure 5-8 top-plot), GFS produced the lowest average timing error followed by GEFSv2 and NCEP, which both had similar performances regarding the timing aspect. GDPS and ECMWF were poor in achieving the timing of the forecasted event in the first five days forecast. However, both NWP appeared to be superior at 7 and 8 days forecast lead times.

The interesting finding from this analysis is that the timing errors of almost all NWPs decreased as the forecast lead time increased (see Figure 5-8 top-plot). The opposite trend is usually seen on several forecast skill metrics (such as Figure 5-4). The primary reason for the decreasing timing error along the forecast horizon is that the source of forecast error was developing from performances related to magnitude rather than timing. This can be easily seen from Figure 5-8 bottom-plot. The general trend of the Percentage of errors other than timing was increasing as the lead time increased. Meaning that as the lead time was increasing, there were more grids and verification periods evolving that did not meet the event threshold criteria either because they underestimated or overestimated the observation. ECMWF, for example, had the highest percentage of magnitude related errors in the 7^{th} and 8^{th} day even though it produced the lowest timing error in these lead times. Overall, GFS, GEFSv2, and NCEP produced lower timing errors and had a smaller

percentage of non-timing errors at all lead times, whereas errors from ECMWF's forecasts were largely attributed to both timing (first five days) and magnitude (all lead times).

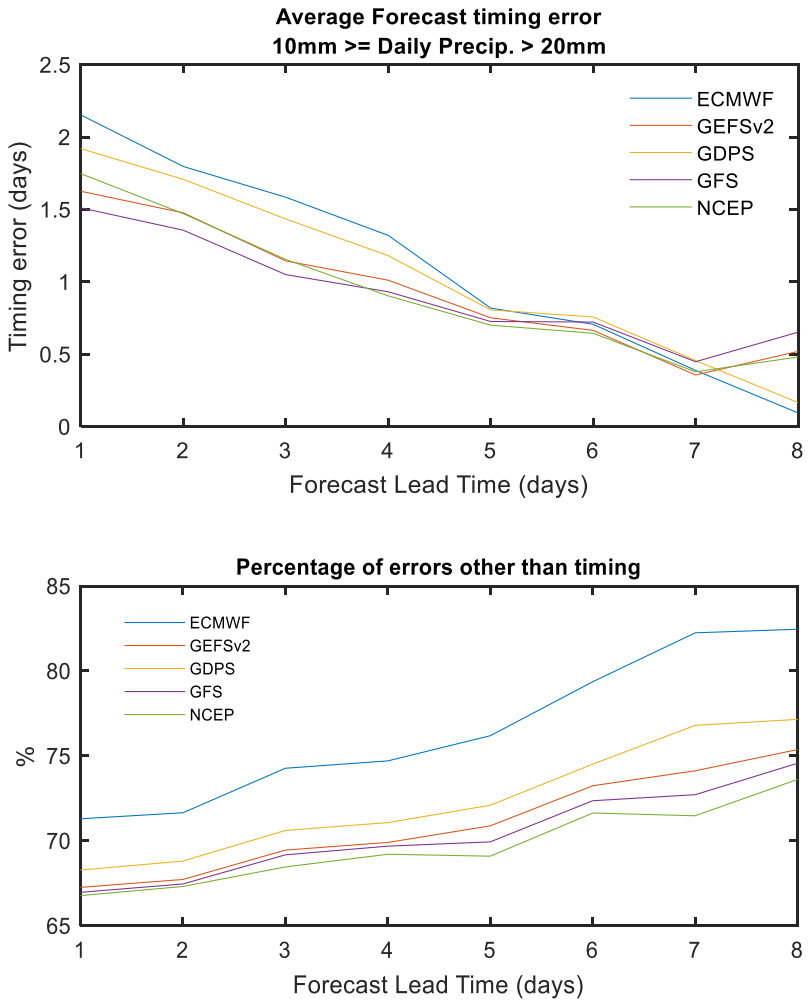

Figure 5-8: Timing Error analysis outputs for five NWPs located in Low-resolution domain.

5.6.1.4. Summary and discussion

The key messages from the operational hydrology perspective are:

- For short-range forecasts up to 3 days ahead, GEFSv2 was superior in providing unbiased, accurate and skillful forecasts in terms of the precipitation volume,

intensity, and timing. Hence, it is an excellent candidate to provide forecast inputs to hydrological models for an improved short-term (up to 3 days lead time) hydrological forecasts in relatively larger watersheds.

- For a week ahead outlook, both GEFSv2, and GFS were productive in forecasting different ranges of precipitation intensities and volumes because their forecasts have had higher skills, better accuracies, and minimum biases. They also achieved a relatively negligible timing error. If either of the two NWP forecasts are used as inputs to hydrological models, enhanced performances of hydrological forecasts could be acquired for medium-range predictions up to a week ahead.

The candidate NWPs identified in the low-resolution domain are ensemble (GEFSv2) and deterministic (GFS). The mean of the ensemble GEFSv2 was used in this research to compare performance with the other NWPs, and its results were promising, as shown earlier. The benefit of ensemble-based hydrological forecasting has been proven in the literature (e.g., Cuo et al., 2011). Therefore, in addition to the mean of the ensemble, using all the ensemble forecasts (11 members) of the GEFSv2 as inputs to hydrological models, provides added value (e.g., provides the uncertainty of the system through the spread, the quantifies the reliability, issues probabilistic forecasts, etc.) to the medium-range hydrological forecasts.

Before applying low-resolution NWPs to hydrological models, a post-processing method is recommended because the spatial scale of the hydrological model is often significantly lower than the horizontal resolution of NWPs.

5.6.2. High-Resolution Domain

5.6.2.1. Volume error

Figure 5-9 shows the forecast performances of four NWPs located in the High-resolution domain for different forecast accumulation periods. The verification was performed with a precipitation threshold of 1mm accumulated over 1hr, 3hr, and 6hrs. Forecast Bias from grid-to-grid evaluation and Precipitation object counts from MODE were estimated at different lead times. For 1-hour accumulation, RAP forecasts were too unstable because its performance highly fluctuated every hour for both traditional and object-based evaluation metrics. The reason might be due to problems related to RAP's hourly data assimilation cycle using the coarse resolution and 3-hourly forecasts of GFS as a boundary condition (Benjamin et al., 2016), but further study is needed to verify. The instability of RAP was not seen at 3hr, and 6hr accumulated forecasts but instead showed relatively biased forecasts like NAM. HRRR appeared to be the most unbiased and accurate NWP regarding the volume of precipitation forecasts at any accumulation hours, because it had the lowest forecast Bias, and its precipitation Object Counts were persistently similar to the observed counts at 1, 3, and 6hr accumulation. For HRDPS, although high in an hourly forecast, the Biases were minimal and comparable with HRRR in 3hr and 6hr accumulation.

Figure 5-10 provides the Gilbert Skill Score (GSS) of the NWPs for different amounts of precipitation volume thresholds (0.2mm, 2mm, and 5mm) accumulated over 6 hours. Results showed that HRDPS had an improved skill in forecasting 6hr accumulated volume precipitation better than the others, and this strength can be linked to its lower forecast bias

(Figure 5-9). The other NWPs were interchangeably demonstrated to have relatively moderate forecast skill, especially for higher 6hr accumulated volumes.

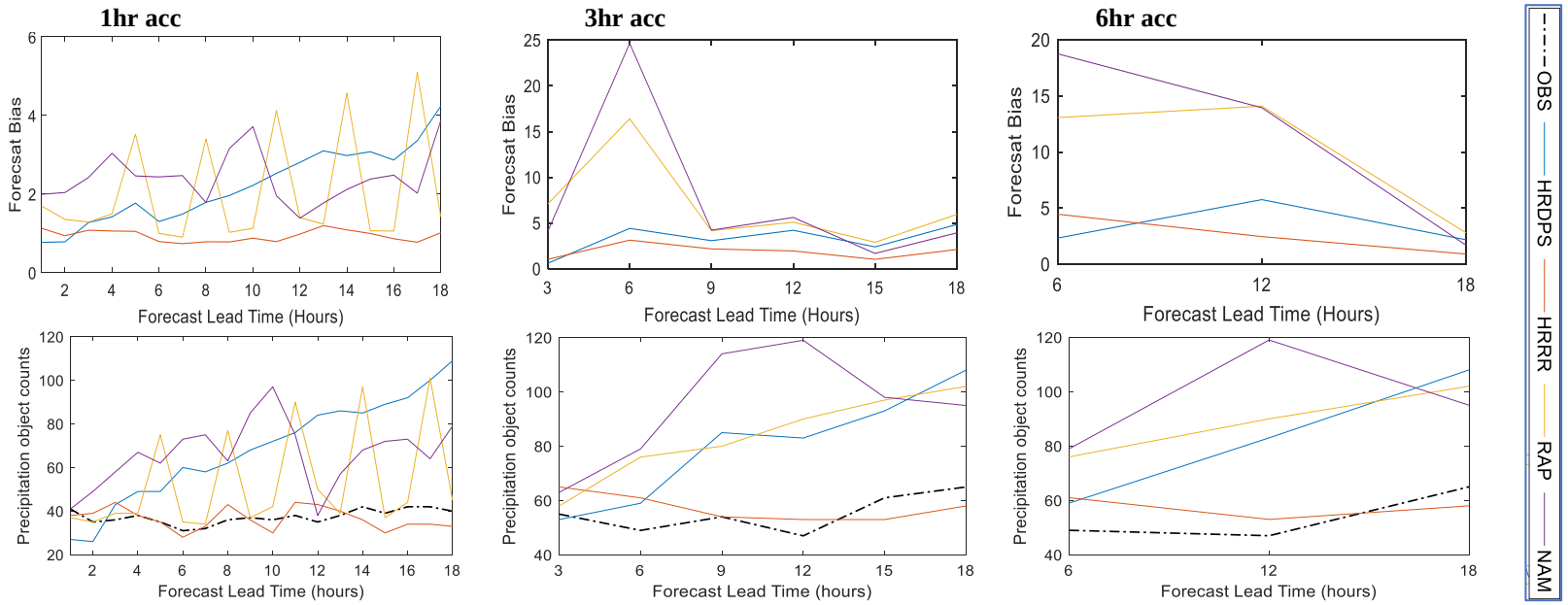

Figure 5-9: Forecast Bias (left) and Precipitation objects count (right) of NWPs located in High-resolution domain for a precipitation threshold of 1mm accumulated in 1, 3 and 6 hours

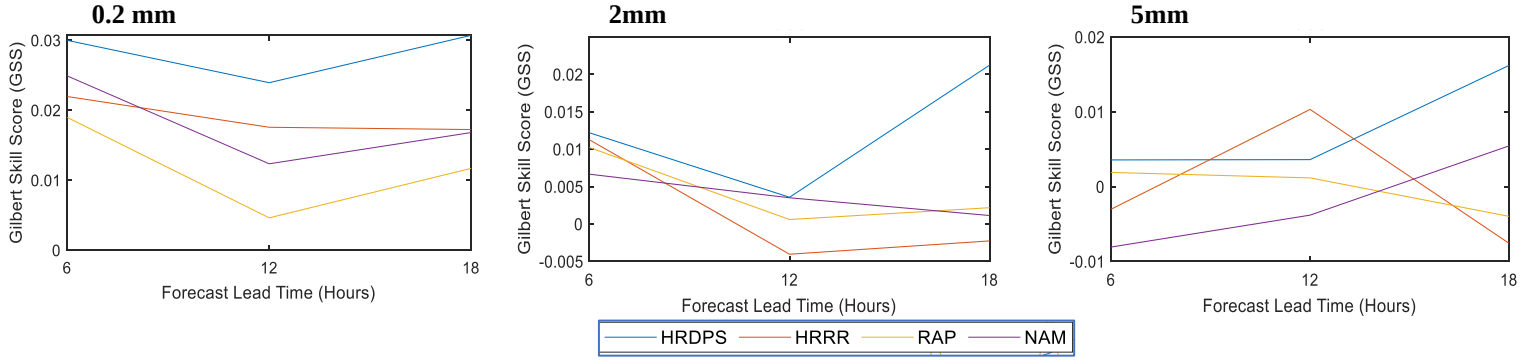

Figure 5-10: Gilbert Skill (GSS) of NWPs located in High-resolution domain for a different precipitation threshold volume accumulated over 6-hours

5.6.2.2. Intensity error

Figure 5-11 shows the Gilbert Skill Score (GSS) and Frequency Bias of four NWPs located in the High-resolution domain in terms of different precipitation intensity forecasts at 4, 6, and 12 hours lead times. Results indicated that HRRR was the most skillful NWP in forecasting higher precipitation intensities, particularly at and above 1mm/hr for all forecast lead times. The 95% confidence interval of GSS in all NWPs seemed to be similar and showed a presence of sampling uncertainties for higher precipitation intensity forecasts. Similarly, most NWPs experienced challenges related to the bias for precipitation above 1.5mm/hr because there were more uncertainties in the "true value" of the Forecast Bias along the forecast horizon. However, HRRR provided substantially lower forecast biases than the others for most precipitation intensities. This result was statistically significant because the 95% confidence interval of the Forecast Bias for HRRR was closer to 1 and had narrower width as compared to others.

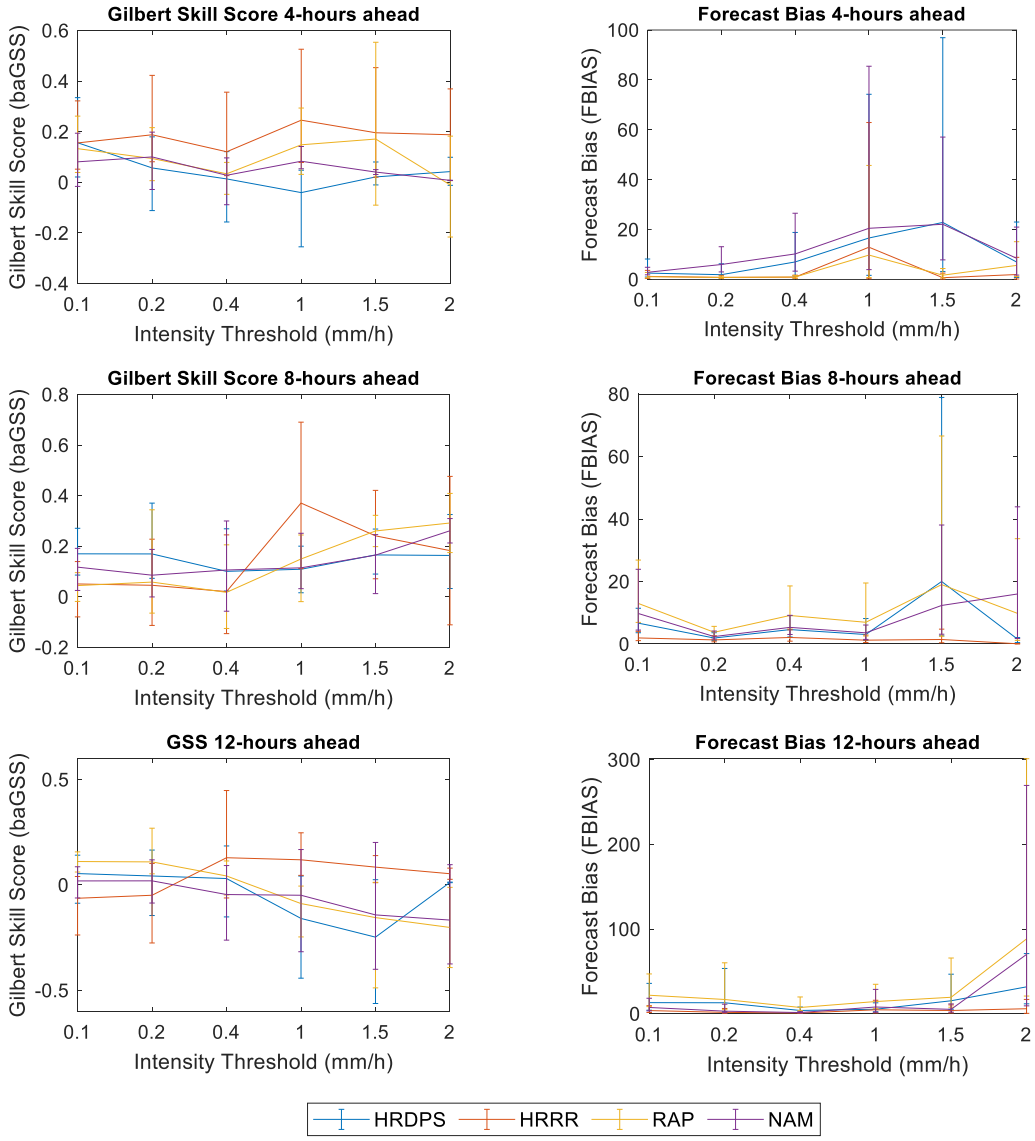

Figure 5-11: Verification metrics for evaluating different precipitation intensity forecasts of four NWPs in High-resolution Domain

5.6.2.3. Timing error

Figure 5-12 shows the Timing Error analysis results of four NWPs in the High-resolution domain. The analysis was performed for a precipitation exceedance threshold (event) of 1mm/hr. The average timing-error of all NWPs showed a decreasing trend over the forecast

horizon, which was also observed in Low-resolution domain (Section 5.6.1.3). However, the rationale that was attributed in the Low-resolution domain forecasts (increasing trend of non-timing error) could not hold here because the percentage of errors other than timing remained more or less steady for most NWPs even though it was unstable. The high temporal resolutions of NWPs (hourly) in the High-resolution domain produced fluctuating timing error outputs along the forecast lead times, which was also seen for some NWPs in other metrics (e.g., Figure 5-9).

Nevertheless, the timing error analysis produced explicable results in the High-resolution domain. The result indicated that NAM had lower average timing error up to 15 hours forecast lead times compared to the other NWPs. However, NAM also shared a significantly large percentage of errors attributed to non-timing aspects (magnitude errors) because a significant number of grids within the domain in the verification period either underestimated or overestimated the threshold criteria. HRDPS and HRRR, even though had higher timing errors than NAM, lead to a minimal number of grids and verification times that had magnitude related errors throughout the forecast lead times.

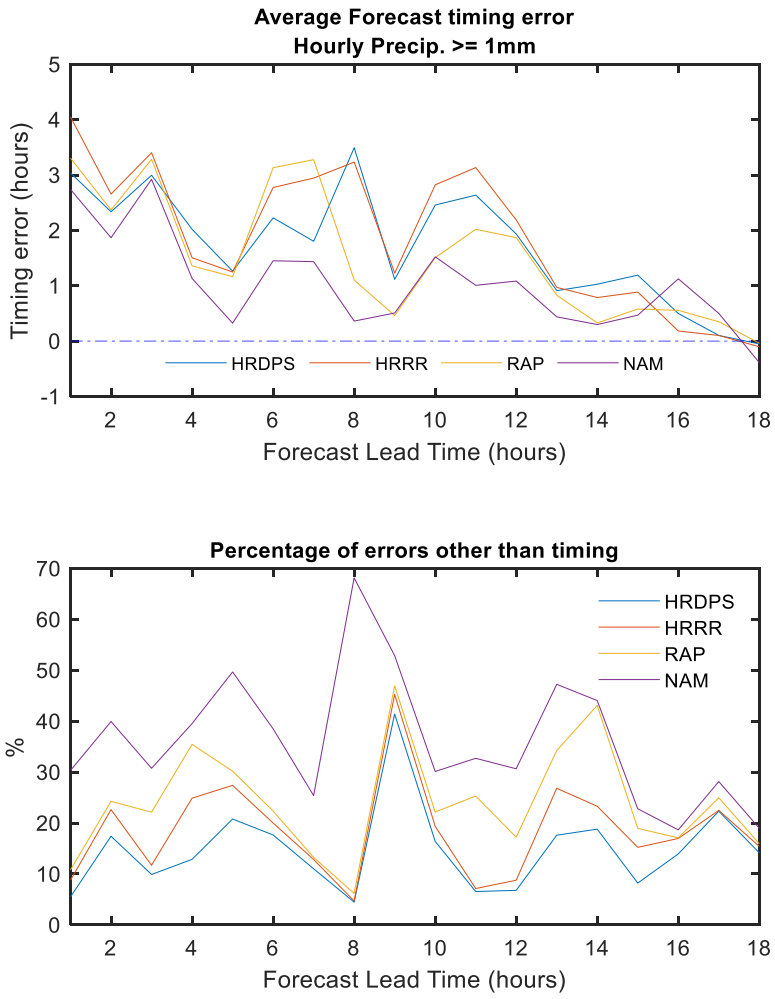

Figure 5-12: Timing Error analysis output for five NWPs located in High-resolution Domain

5.6.2.4. Summary and discussion

Based on the above results, the following summaries can be drawn from the operational hydrological forecasting perspective.

- In terms of precipitation volume, hourly forecasts were better achieved by HRRR, 3-hour cumulated forecasts were produced better by both HRRR and HRDPS, and 6-hour cumulated forecasts were recognized better by HRDPS.
- Particularly at higher precipitation intensities (above 1mm/hr), HRRR was found to be superior in achieving unbiased and skillful forecasts for lead times of 1hr up to 18hrs ahead. This finding is essential for forecasting floods produced by intense and localized storms.
- All NWPs were prone to higher timing errors in the first 3hours of the forecast, which gradually decreased as the lead time increased. NAM produced lower timing error but at the expense of having a significantly large percentage of non-timing errors within the domain grids in the verification period. On the other hand, HRDPS and HRRR had very less percentages of non-timing related errors in the domain grids over the verification period, even though they incurred higher timing error, particularly between 5hr to 13hr forecast lead times. The take-home message is that, if the particular interest is only in achieving the timing of the flood, caution should be taken on identifying the right NWPs because some candidate products (such as NAM) that have less timing error are subjected to higher magnitude or volume errors. Instead, a collective aim of the timing, intensity, and volume of floods should be sought for an enhanced short-range flood forecasting in smaller and urbanizing catchments. In this case, HRDPS and HRRR appeared to be the best candidate precipitation forecast products to force hydrological models.

High-resolution NWPs that are capable of resolving convective processes at the grid or sub-grid level are beneficial for flood forecasting, especially those induced by summertime storms (Milbrandt et al., 2016). HRRR and HRDPS have both high-spatial (2.5-3km), and high-temporal (hourly) resolutions, and are grid-level convective-permitting (Pinto et al., 2015) and sub-grid level deep-convection parameterizing (Milbrandt et al., 2016) models respectively. Hence, the use of these NWPs for very-short (1hr to 6hr) and short-range (up to 24hr) flood forecasting in urban and semi-urban catchments is paramount. As supported in literature, the capacity of NWPs to predict the combined effect of timing, intensity and amount of storms as effective as possible is a necessary condition for accurate flood forecasting because small errors can lead to significant deviations in the flood hydrograph (Jasper et al., 2002).

5.7. Conclusion

The main objective of this study was to identify Numerical Weather Predictions (NWPs) that have skillful and reliable precipitation forecasts for an improved medium- and short-range flood prediction in varied watershed characteristics of Canada. For this purpose, two domains were selected based on catchment landscape types, sizes, and the spatial and temporal scale of available NWPs. The low-resolution domain covers regions where most watersheds are vast (above thousands of km^2 areas), have agriculture, forests, and wetlands. In this domain, lower spatial resolutions NWPs that issue medium-range forecasts were used, including the deterministic GFS and GDPS, and ensemble means forecasts of ECMWF, NCEP, and GEFSv2. The high-resolution domain covers smaller urban and semi-urban catchments, which typically have shorter time of concentrations. In this domain,

higher spatial resolution NWPs that issue short-range deterministic precipitation forecasts such as HRRR, HRDPS, RAP, and NAM were used. NWPs in Low-resolution domain were verified using CaPA for up to 8 days forecast lead times in a 7-month hindcast period. Whereas NWPs in High-resolution domain were compared with gridded interpolated hourly observed data for 1hr up to 18hr forecast lead times in a 5-month hindcast period.

This study was intended to assist hydrologists in finding the best candidate precipitation forecast products in various scales so that it can be applied in hydrological models to forecast the magnitude and timing of floods. As such, the verification of different NWPs was performed on volume, intensity, and timing aspects of precipitation forecasts. In addition to the traditional grid-to-grid and emerging object-based (MODE) verification metrics, a new approach to estimate the average timing error was developed in this research. The following conclusions can be drawn from the comprehensive verification of the selected NWPs.

In the Low-resolution domain, GEFSv2 and GFS provided better skill, accuracy, and relatively unbiased forecasts for different accumulated precipitation volumes at all lead times. GEFSv2 showed good potential in identifying and matching forecast precipitation objects. Verification of precipitation intensities ranging between 5 and 35 mm/day was made at 3-days ahead and a week ahead outlooks. Results revealed that GEFSv2 and GFS maintained lower forecast biases and achieved better forecast skills at both forecast outlooks. On the other hand, ECMWF, GDPS, and NCEP significantly overestimated forecasts at different precipitation intensities, especially for a week-ahead outlook. By using MODE, the quality of identified precipitation objects was estimated for a different

combination of convolution parameters (e.g., Threshold: 5, 10, & 15 mm/day, and Radius: 5, 10, 15 grid units). The collective behavior of all NWPs was that the performances of identified objects degraded as the grid smoothing resolution and threshold intensity increased, and reproducing the higher precipitation threshold (above 15mm/day) was a challenge. Overall, GEFSv2 followed by GFS appeared to be superior in generating good qualities of precipitation objects for different MODE parameters, whereas ECMWF and GDPS produced poor forecast attributes. The timing error approach was implemented in Low-resolution domain for an event threshold criterion of 10-20mm/day. Results indicated that GFS followed by both GEFSv2 and NCEP attained better timing of the precipitation forecast up to 6 days lead time. The striking result from this analysis was that the average timing errors of almost all NWPs decreased as the forecast lead time increased, which was opposite to the trends usually observed in other forecast verification metrics. As the lead time increased, there were more grids and verification periods that could not meet the event threshold criteria because they either underestimated or overestimated the event.

Overall, GEFSv2 for 3-days ahead outlook, and both GEFSv2 and GFS for a week-ahead outlook appeared to be a proper application of the candidate NWPs in hydrological models to enhance short- and medium-range hydrological forecasting in the Low-resolution domain of Canada.

In the High-resolution domain, the volume skill was evaluated in two ways: 1mm precipitation threshold accumulated over different periods (1hr, 3hr, and 6hrs), and different precipitation volume thresholds (0.2mm, 2mm, and 5mm) accumulated over 6 hours. Verification results indicated that HRRR appeared to be the most unbiased and

accurate NWP in forecasting the precipitation at any accumulation hours. HRDPS showed competitive performance as HRRR in 3hr and 6hr accumulation. For different volume accumulations in a 6hr period, HRDPS had an improved forecast skill, which could be associated with its lower forecast bias. Regarding intensity, NWPs were evaluated for different precipitation intensities ranging from 0.1 to 2mm/hr. Results indicated that HRRR appeared to be the most skillful and unbiased NWP in forecasting higher precipitation intensities, particularly at and above 1mm/hr at all forecast lead times. The timing error analysis was performed for an event criteria of 1mm/hr precipitation exceedance. Even though NAM produced lower timing error up to 15 hours forecast lead time, it shared a considerably large percentage of errors attributed to non-timing aspects (e.g. magnitude error). HRDPS and HRRR had instead contributed to a smaller number of grids and verification times that have magnitude related errors throughout the forecast lead times. Overall, for a collective aim of the timing, intensity, and volume of floods, HRRR and HRDPS appeared to be the best candidate precipitation forecast products for an enhanced short-range flood forecasting in smaller and urbanizing catchments.

In general, the timing error approach was able to produce two essential outputs: (1) it could find the average timing error of the precipitation forecasts shifted on the forecast horizon, and; (2) it could distinguish the percentage of errors attributed to issues other than timing or the share of the number of grids and verification times that under- or over-estimated the event in the forecast lead time.

Future works are anticipated to use identified potential NWPs as inputs into hydrological models that are calibrated on selected watersheds from each domain for hydrological and flood forecast verification.

5.8. References

Awol, F.S., Coulibaly, P., Tolson, B.A., 2018. Event-based model calibration approaches for selecting representative distributed parameters in semi-urban watersheds. Adv. Water Resour. 118, 12–27. https://doi.org/10.1016/j.advwatres.2018.05.013

Benjamin, S.G., Weygandt, S.S., Brown, J.M., Hu, M., Alexander, C.R., Smirnova, T.G., Olson, J.B., James, E.P., Dowell, D.C., Grell, G.A., Lin, H., Peckham, S.E., Smith, T.L., Moninger, W.R., Kenyon, J.S., Manikin, G.S., Benjamin, S.G., Weygandt, S.S., Brown, J.M., Hu, M., Alexander, C.R., Smirnova, T.G., Olson, J.B., James, E.P., Dowell, D.C., Grell, G.A., Lin, H., Peckham, S.E., Smith, T.L., Moninger, W.R., Kenyon, J.S., Manikin, G.S., 2016. A North American hourly assimilation and model forecast cycle: The Rapid Refresh. Mon. Weather Rev. 144, 1669–1694. https://doi.org/10.1175/MWR-D-15-0242.1

Blaylock, B. K., Horel, J. D., & Liston, S. T., 2017. Cloud archiving and data mining of High-Resolution Rapid Refresh forecast model output. Computers & Geosciences, 109, 43–50. doi: 10.1016/j.cageo.2017.08.005

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., De Chen, H., Ebert, B., Fuentes, M., Hamill, T.M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.Y., Parsons, D., Raoult, B., Schuster, D., Dias, P.S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., Worley, S., 2010. The thorpex interactive grand global ensemble. Bull. Am. Meteorol. Soc. 91, 1059–1072. https://doi.org/10.1175/2010BAMS2853.1

Brill, K.F., Mesinger, F., Brill, K.F., Mesinger, F., 2009. Applying a general analytic method for assessing bias sensitivity to bias-adjusted threat and equitable threat scores. Weather Forecast. 24, 1748–1754. https://doi.org/10.1175/2009WAF2222272.1

Brown, B.G., Gotway, J.H., Bullock, R., Gilleland, E., Fowler, T., Ahijevych, D., Jensen, T., 2009. The model evaluation tools (MET): community tools for forecast evaluation, in: Preprints, 25th Conf. on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Phoenix, AZ, Amer. Meteor. Soc. A.

Brown, J.D., Seo, D.-J., Du, J., 2012. Verification of precipitation forecasts from ncep's short-range ensemble forecast (SREF) system with reference to ensemble streamflow prediction using lumped hydrologic models. J. Hydrometeorol. 13, 808–836. https://doi.org/10.1175/JHM-D-11-036.1

Buizza, R., Houtekamer, P.L., Pellerin, G., Toth, Z., Zhu, Y., Wei, M., 2005. A comparison of

the ECMWF, MSC, and NCEP global ensemble prediction systems. Mon. Weather Rev. 133, 1076–1097. https://doi.org/10.1175/mwr2905.1

Casati, B., Ross, G., Stephenson, D.B., 2004. A new intensity-scale approach for the verification of spatial precipitation forecasts. Meteorol. Appl. 11, 141–154. https://doi.org/10.1017/S1350482704001239

Cooley, A., Chang, H., 2017. Precipitation intensity trend detection using hourly and daily observations in Portland, Oregon. Climate 5. https://doi.org/10.3390/cli5010010

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. J. Hydrometeorol. 12, 713–728. https://doi.org/10.1175/2011JHM1347.1

Davis, C., Brown, B., Bullock, R., Davis, C., Brown, B., Bullock, R., 2006a. Object-based verification of precipitation forecasts. Part I: methodology and application to mesoscale rain areas. Mon. Weather Rev. 134, 1772–1784. https://doi.org/10.1175/MWR3145.1

Davis, C., Brown, B., Bullock, R., Davis, C., Brown, B., Bullock, R., 2006b. Object-based verification of precipitation forecasts. Part II: application to convective rain systems. Mon. Weather Rev. 134, 1785–1795. https://doi.org/10.1175/MWR3146.1

Davis, C.A., Brown, B.G., Bullock, R., Halley-Gotway, J., Davis, C.A., Brown, B.G., Bullock, R., Halley-Gotway, J., 2009. The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. Weather Forecast. 24, 1252–1267. https://doi.org/10.1175/2009WAF2222241.1

Ebert, E.E., 2009. Neighborhood verification: a strategy for rewarding close forecasts. Weather Forecast. 24, 1498–1510. https://doi.org/10.1175/2009WAF2222251.1

Georgakakos, K.P., Graham, N.E., Modrick, T.M., Murphy, M.J., Shamir, E., Spencer, C.R., Sperfslage, J.A., 2014. Evaluation of real-time hydrometeorological ensemble prediction on hydrologic scales in Northern California. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2014.05.032

Georgakakos, K.P., Krzysztofowicz, R., 2001. Probabilistic and ensemble forecasting. J. Hydrol. 249, 1. https://doi.org/10.1016/S0022-1694(01)00455-3

Givati, A., Gochis, D., Rummler, T., Kunstmann, H., 2016. Comparing one-way and two-way coupled hydrometeorological forecasting systems for flood forecasting in the mediterranean region. Hydrology 3, 19. https://doi.org/10.3390/hydrology3020019

Golding, B.., 2000. Quantitative precipitation forecasting in the UK. J. Hydrol. 239, 286–305. https://doi.org/10.1016/S0022-1694(00)00354-1

Hagedorn, R., Buizza, R., Hamill, T.M., Leutbecher, M., Palmer, T.N., 2012. Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. Q. J. R. Meteorol. Soc. 138, 1814–1827. https://doi.org/10.1002/qj.1895

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Jr., Y.

Zhu, and W. Lapenta., 2013. NOAA's second-generation global medium-range ensemble reforecast data set. Bull Amer. Meteor. Soc., 94, 1553-1565. doi: http://dx.doi.org/10.1175/BAMS-D-12-00014.1

Han, J., Pan, H.-L., 2011. Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. Weather Forecast. 26, 520–533. https://doi.org/10.1175/WAF-D-10-05038.1

Han, S., Coulibaly, P., 2019. Probabilistic flood forecasting using hydrologic uncertainty processor with ensemble weather forecasts. J. Hydrometeorol. JHM-D-18-0251.1. https://doi.org/10.1175/JHM-D-18-0251.1

Jasper, K., Gurtz, J., Lang, H., 2002. Advanced flood forecasting in Alpine watersheds by coupling meteorological observations and forecasts with a distributed hydrological model. J. Hydrol. 267, 40–52. https://doi.org/10.1016/S0022-1694(02)00138-5

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. Hydrol. Earth Syst. Sci. 22, 1957–1969. https://doi.org/10.5194/hess-22-1957-2018

Jolliffe, I., Stevenson, D., 2011. Forecast verification, International Journal of Forecasting. John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/9781119960003

Kaufmann, P., Schubiger, F., Binder, P., 2003. Precipitation forecasting by a mesoscale numerical weather prediction (NWP) model: eight years of experience. Hydrol. Earth Syst. Sci. 7, 812–832. https://doi.org/10.5194/hess-7-812-2003

Leach, J.M., Kornelsen, K.C., Coulibaly, P., 2018. Assimilation of near-real time data products into models of an urban basin. J. Hydrol. 563, 51–64. https://doi.org/10.1016/J.JHYDROL.2018.05.064

Li, J., Hsu, K., Aghakouchak, A., Sorooshian, & S., 2015. An object-based approach for verification of precipitation estimation. Int. J. Remote Sens. 36, 513–529. https://doi.org/10.1080/01431161.2014.999170

Mai, J., Kornelsen, K., Tolson, B., Fortin, V., Gasset, N., Bouhemhem, D., Schäfer, D., Leahy, M., Anctil, F. and Coulibaly, P. (2019). The Canadian surface prediction archive (CaSPAr): A platform to enhance environmental modeling in Canada and globally. Bulletin of the American Meteorological Society. doi: 10.1175/bams-d-19-0143.1

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

Mascaro, G., Vivoni, E.R., Deidda, R., Mascaro, G., Vivoni, E.R., Deidda, R., 2010. Implications of ensemble quantitative precipitation forecast errors on distributed streamflow forecasting. J. Hydrometeorol. 11, 69–86. https://doi.org/10.1175/2009JHM1144.1

Milbrandt, J.A., Bélair, S., Faucher, M., Vallée, M., Carrera, M.L., Glazer, A., 2016. The Pan-

canadian high resolution (2.5 km) deterministic prediction system. weather forecast. 31, 1791–1816. https://doi.org/10.1175/WAF-D-16-0035.1

Molteni, F., Buizza, R., Palmer, T. and Petroliagis, T., 1996. The ECMWF ensemble prediction system: methodology and validation. Quarterly Journal of the Royal Meteorological Society, 122(529), pp.73-119.

Muhammad, A., Stadnyk, T., Unduche, F., Coulibaly, P., 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, 1604. https://doi.org/10.3390/w10111604

NOAA., 2015. NCEP GFS 0.25 degree global forecast grids historical archive. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO. https://doi.org/10.5065/D65D8PWK, Accessed 9th July 2019

Pagano, T.C., Wood, A.W., Ramos, M.-H., Cloke, H.L., Pappenberger, F., Clark, M.P., Cranston, M., Kavetski, D., Mathevet, T., Sorooshian, S., Verkade, J.S., 2014. Challenges of operational river forecasting. J. Hydrometeorol. 140516115449007. https://doi.org/10.1175/JHM-D-13-0188.1

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). Hydrol. Earth Syst. Sci. 9, 381–393. https://doi.org/10.5194/hess-9-381-2005

Pappenberger, F., Scipal, K., Buizza, R., 2008. Hydrological aspects of meteorological verificatio. Atmos. Sci. Lett. 9, 43–52. https://doi.org/10.1002/asl.171

Pinto, J.O., Grim, J.A., Steiner, M., Pinto, J.O., Grim, J.A., Steiner, M., 2015. Assessment of the High-Resolution Rapid Refresh Model's ability to predict mesoscale convective systems using object-based evaluation. Weather Forecast. 30, 892–913. https://doi.org/10.1175/WAF-D-14-00118.1

Rebora, N., Ferraris, L., Von Hardenberg, J., Provenzale, A., 2006. Rainfall downscaling and flood forecasting: a case study in the Mediterranean area. Nat. Hazards Earth Syst. Sci. 6, 611–619.

Roberts, N.M., Lean, H.W., 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Mon. Weather Rev. 136, 78–97. https://doi.org/10.1175/2007MWR2123.1

Robertson, D.E., Shrestha, D.L., Wang, Q.J., 2013. Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. Hydrol. Earth Syst. Sci. 17, 3587–3603. https://doi.org/10.5194/hess-17-3587-2013

Rogers, E., DiMego, G., Black, T., Ek, M., Ferrier, B., Gayno, G., Janjic, Z., Lin, Y., Pyle, M., Wong, V. and Wu, W.S., 2009. The NCEP North American mesoscale modeling system: Recent changes and future plans. In Preprints, 23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction.

Roulin, E., 2006. Skill and relative economic value of medium-range hydrological ensemble predictions. Hydrol. Earth Syst. Sci. Discuss. 3, 1369–1406. https://doi.org/10.5194/hessd-3-1369-2006

Toth, Z., & Kalnay, E., 1993. Ensemble Forecasting at NMC: The generation of perturbations. Bulletin Of The American Meteorological Society, 74(12), 2317-2330. doi: 10.1175/1520-0477(1993)074<2317:efantg>2.0.co;2

Unduche, F., Tolossa, H., Senbeta, D., Zhu, E., 2018. Evaluation of four hydrological models for operational flood forecasting in a Canadian Prairie watershed. Hydrol. Sci. J. 63, 1–17. https://doi.org/10.1080/02626667.2018.1474219

Verkade, J.S., Brown, J.D., Davids, F., Reggiani, P., Weerts, A.H., 2017. Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine. J. Hydrol. 555, 257–277. https://doi.org/10.1016/j.jhydrol.2017.10.024

Wilks, D., 2006. Statistical Methods in the Atmospheric Sciences, 2nd ed, International Geophysics Series. Acadamic Press.

Wolff, J.K., Harrold, M., Fowler, T., Gotway, J.H., Nance, L., Brown, B.G., 2014. Beyond the basics: evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. Weather Forecast. 29, 1451–1472. https://doi.org/10.1175/WAF-D-13-00135.1

Yan, H., Gallus, W.A., Yan, H., Jr., W.A.G., 2016. An Evaluation of QPF from the WRF, NAM, and GFS models using multiple verification methods over a small domain. Weather Forecast. 31, 1363–1379. https://doi.org/10.1175/WAF-D-16-0020.1

Yu, W., Nakakita, E., Kim, S., Yamaguchi, K., 2016. Improving the accuracy of flood forecasting with transpositions of ensemble NWP rainfall fields considering orographic effects. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2016.05.047

Yucel, I., Onen, A., Yilmaz, K.K., Gochis, D.J., 2015. Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall. J. Hydrol. 523, 49–66. https://doi.org/10.1016/j.jhydrol.2015.01.042

Zahmatkesh, Z., Jha, S.K., Coulibaly, P., Stadnyk, T., 2019. An overview of river flood forecasting procedures in Canadian watersheds. Can. Water Resour. J. 44, 213–229. https://doi.org/10.1080/07011784.2019.1601598

Zappa, M., Jaun, S., Germann, U., Walser, A., Fundel, F., 2011. Superposition of three sources of uncertainties in operational flood forecasting chains. Atmos. Res. 100, 246–262.

https://doi.org/10.1016/j.atmosres.2010.12.005

Chapter 6. Conclusions and Recommendations

6.1. Conclusions

The research presented in this thesis focused on identifying proper calibration approaches, skillful hydrological models, and skillful weather forecast inputs to improve short- and medium-range flood forecasting in various watershed landscapes. The chapters of this thesis provided the required evaluation and verification of different calibration approaches, multiple hydrological models with diverse model structures, and various high- and low-resolution Numerical Weather Predictions, and discussed the identified potential candidate models and inputs. The accuracy and skill of candidate hydrological models and forecast input products in hydrological forecasting were evaluated in the research. The findings of this research are expected to benefit operational flood forecasting centers, future applications in flood and early warning systems, and research aiming at improving model development, forecast inputs, and calibration methods. The main conclusions of the thesis are summarized as follows:

6.1.1. Calibration approaches for enhanced peak flow predictions

- The presence of uncertainties in calibrating highly impervious sub-catchments and pervious areas with rapid recovery times as well as reproducing peak flows prompts the need for a robust calibration approach.
- An event-based calibration approach integrating multi-site, and single and multi-objective optimizations was proposed to find representative distributed model parameters in a semi-urban catchment.

- The two multi-site approaches (MS-S and MS-A) exhibit better performances than multi-event ME-MS and at the catchment outlet OU approach.
- Using the single objective DDS optimization in MS-A approach is found to be more efficient using the multi-objective PA-DDS algorithm in MS-S approach.
- The study indicated that the combination of efficient optimization tools with a series of calibration steps is essential in finding representative parameter sets.

6.1.2. Hydrological model identification for large complex watersheds

- Bias-correcting each member of an ensemble precipitation forecasts using verifying datasets (such as CaPA) for a training period of at least two years before the forecast time, produced reliable hydrological forecasts in complex watersheds.
- As indicated by the forecast reliability, accuracy and skill score measures, lumped hydrological models (SACSMA and MACHBV), provided better reservoir inflow forecast performance than the benchmark distributed (WATFLOOD) and the macroscale land-surface based (VIC) model.
- The forecast performances of lumped models are more evident for up to a week ahead forecast (1-day to 8-days). The distributed benchmark model provided reliability as good as the lumped models only up to 3-days ahead forecast.
- The calibration performances of the hydrological models were influenced jointly by several factors including the scale, the complexity of the basin, the spatial and temporal resolution of the input data, the structure of the models, the degree of discretization of the models, and the number of parameters to be calibrated.

- In general, for hydrological forecasting focusing on basin outflows and not interior sites, the study indicated that lumped models, particularly SACSMA with SNOW17, provided better forecast performance than distributed or land-surface models in complex watersheds up to a week ahead outlook.

6.1.3. Combined hydrological model and input identification for semi-urban catchments

- Based on the calibration performances of twelve hydrological models and previous application, five models comprised of lumped (SACSMA, MACHBV & PDM, all coupled with SNOW17) and distributed (WATFLOOD & SWMM) models were selected for forecast verification.
- The pre-screening of hydrological outputs from four high- and mid-resolution NWPs (HRRR, HRDPS, NAM, & RAP) showed that, for relatively small semi-urban watersheds that typically have shorter time of concentrations, high-resolution weather products (e.g., HRRR, HRDPS) were found to be proper forecast inputs to the hydrological models.
- Comprehensive evaluation based on forecast accuracy, skill, peak flow magnitude and timing, threshold-based scores, and economic value attributes has revealed that the two lumped models (MACHBV and SACSMA) were superior to the distributed models for 1hr to 18hr lead times. The distributed models were only skillful between 15hr-18hr forecast lead times.
- The best model-input combination appeared to be the MACHBV-SNOW17 model with HRDPS forecast input. This combination captured the peak flow magnitude

and timing adequately and detected the flood threshold at all forecast lead times. It has also emerged as the most economically viable model-input combination.

- Giving adaptive weights to hydrological forecasts based on recent (last 18hr-24hr) performances provided enhanced combined forecasts while persistently keeping the top-ranking models, which would highly likely continue to perform well (next 18hr-24hr).

6.1.4. Identification of Numerical Weather Predictions (NWP) for enhanced flood forecasting

- In Western and Central parts of Canada (Low-resolution domain), where the majority of watersheds are vast and transboundary, GEFSv2, followed by GFS, were found to be excellent candidates to provide precipitation forecast inputs to hydrological models for an improved short- and medium-range hydrological forecasts.
- For 3-days ahead forecast, GEFSv2, and for a week ahead outlook, both GEFSv2, and GFS were very productive in forecasting different ranges of precipitation intensities and volumes with less bias, better forecast accuracy, and minimal forecast timing error in the Low-resolution domain.
- In Southern Ontario (High-resolution domain), where abundant catchments are small urban and semi-urban, HRRR and HRDPS appeared to be the best candidates to provide precipitation forecast inputs to hydrological models for an enhanced short-range flood forecasting (1hr up to 18 hr lead times).

- If a collective aim of achieving the timing, intensity, and volume of floods is sought for, either of the two high-resolution products (HRRR and HRDPS) could offer qualitative forecast inputs in urbanizing catchments.
- Higher intensity forecasts (above 1mm/hr) and 1hr cumulated precipitation forecast volumes were more evident with HRRR, whereas 6hr cumulated forecasts were recognized better by HRDPS. Both products had shown a competent forecast potential in 3hr forecast volumes.
- The timing error approach was able to provide;
    - the average estimated forecast timing error, which, for example, showed a general decreasing trend in the Low-resolution domain as the lead time increases,
    - the percentage of non-timing related errors, which, for example, showed an increasing trend in the Low-resolution domain as the lead time increases.

6.1.5. General conclusion and contributions

The general conclusions and contributions of this research from the perspectives of the scientific research, engineering application, and universality aspects include:

- An approach was formulated for calibrating an event-based distributed hydrological model.
- Aggregating or averaging multiple performance metrics during calibration provided better simulation output, which can be applied for any cases.
- The advantage of applying lumped hydrological models in flood and hydrological forecasting was demonstrated in two complementary watersheds.
    - A lumped hydrological model (SACSMA) forced with bias-corrected ensemble forecast inputs appeared to provide reliable and skillful medium-range reservoir inflow forecasts in large complex watersheds.
    - A combination of lumped hydrological model (MACHBV) with high-resolution Numerical Weather Prediction model (HRDPS) emerged as the best model-input integration for enhanced short-range flood forecasting in semi-urban catchments.
- The same candidate model(s) would highly likely be identified to better simulate and forecast short- and medium-range hydrological forecasts in other types of watersheds with a similar scale and characteristics.
- For operational forecasters focusing on basin outflows and not interior sites, the study demonstrated that lumped models are much preferable, economically viable,

and computationally efficient than distributed or land-surface based hydrological models.

- For operational forecasters interested in averaging multiple deterministic hydrological forecasts with simple and cost-effective methods, the adaptive weighting technique can be used to provide an enhanced combined forecast while persistently keeping well-performing models.
- The research identified best candidate NWPs in two main geographic regions of Canada, which can be utilized in operational flood forecasting to predict the volume, intensity, and timing of floods.
- Modelers or hydrologists often decide the hydrological modeling time steps based on the temporal scale and quality of inputs available. The study showed the strengths of high-resolution forecast inputs for different precipitation accumulation periods. HRRR was better at 1hr, and 3hr cumulated forecasts, whereas HRDPS was better at 3hr and 6hr cumulated forecast volumes.
- A timing error estimation approach was introduced to find the average timing error of forecast variables shifted along the forecast lead times and to assess the percentage of errors attributed to non-timing aspects.

6.2. Future Work and Recommendations

Following up on the last study of this thesis (Chapter 5), a hydrological forecast verification study using the identified candidate NWPs as inputs into hydrological models that are calibrated on selected watersheds is anticipated. Essentially, an integrated or coupled meteorological and hydrological forecast verification is advantageous for application-

focused evaluations, addressing scaling and temporal effect, identifying precipitation systems contributing to floods, and exploring different post-processing methods (Cuo et al., 2011; Givati et al., 2016; Jasper et al., 2002; Pappenberger et al., 2008; Yucel et al., 2015). An ensemble-based flood forecasting system using a multi-model and multi-input approach is recommended for operational use since its advantage has been demonstrated in different studies (Cloke and Pappenberger, 2009; Demeritt et al., 2007; Pagano et al., 2014; Pappenberger et al., 2005). Furthermore, various data assimilation, Bayesian Probabilistic Forecasting, and Bayesian Model Averaging methods could be added to a coupled multi-model and multi-input hydrological forecasting system, although such multi-level integration will be computationally intensive especially from the operational river forecasting perspective. This computational burden can be minimized through a collaboration of operational forecasting centers with academic researchers who usually have access to High-performance computing (HPC) systems (e.g., Graham & Cedar HPCs of Compute Canada).

If flow conditions at interior sites of a watershed are desired, the lumped models that were identified in the thesis research (SACSMA and MACHBV) could be implemented as a semi-distributed model using multiple sub-catchments, a routing module, and a proper calibration approach (Ajami. et al., 2004).

NWPs with ensemble forecasts have mostly low-spatial resolutions. Due to this condition, the direct application of ensemble NWPs for short-range flood forecasting in urban and semi-urban watersheds has been a challenge (Pagano et al., 2014). Thus, before applying low-resolution NWPs to hydrological models, a post-processing method is recommended

because the spatial resolution of the hydrological modeling is usually significantly higher than the horizontal resolution of NWPs. The Spatio-temporal downscaling approach using RainFARM (Rebora et al., 2006), and the Rainfall Post-Processing (RPP) technique (Jha et al., 2018; Robertson et al., 2013), are some of the recommended tools that can be applied to the ensemble or deterministic NWPs to improve the reliability of flood forecasting in urban and semi-urban catchment. Separation of NWP precipitation forecasts based on topographic effect (e.g., orographic and non-orographic rainfalls) before inputting in to distributed hydrological models was also shown as an alternative method that can be used for real-time updating of flood forecasts (Yu et al., 2016).

The additional exciting research topic would be evaluating different forecast combination methods, in addition to the adaptive weighting method used in Chapter 4, to improve hydrological forecasting that is based on multi-model and multi-input approaches. Moreover, an effective way of defining flood thresholds for categorical forecast verification could be assessed, in addition to the method previously mentioned in Chapter 4.

Finally, operational forecasting centers could benefit from archiving real-time flood and hydrological forecasts for continuously evaluating the performance of the flood forecasting system.

6.3. References

Ajami, N.K., Gupta, H., Wagener, T., Sorooshian, S., 2004. Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. J. Hydrol. 298, 112–135. https://doi.org/10.1016/j.jhydrol.2004.03.033

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. J. Hydrol. 375, 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

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. J. Hydrometeorol. 12, 713–728. https://doi.org/10.1175/2011JHM1347.1

Demeritt, D., Cloke, H., Pappenberger, F., Thielen, J., Bartholmes, J., Ramos, M.H., 2007. Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. Environ. Hazards 7, 115–127. https://doi.org/10.1016/j.envhaz.2007.05.001

Georgakakos, K.P., Graham, N.E., Modrick, T.M., Murphy, M.J., Shamir, E., Spencer, C.R., Sperfslage, J.A., 2014. Evaluation of real-time hydrometeorological ensemble prediction on hydrologic scales in Northern California. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2014.05.032

Givati, A., Gochis, D., Rummler, T., Kunstmann, H., 2016. Comparing one-way and two-way coupled hydrometeorological forecasting systems for flood forecasting in the mediterranean region. Hydrology 3, 19. https://doi.org/10.3390/hydrology3020019

Jasper, K., Gurtz, J., Lang, H., 2002. Advanced flood forecasting in Alpine watersheds by coupling meteorological observations and forecasts with a distributed hydrological model. J. Hydrol. 267, 40–52. https://doi.org/10.1016/S0022-1694(02)00138-5

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. Hydrol. Earth Syst. Sci. 22, 1957–1969. https://doi.org/10.5194/hess-22-1957-2018

Pagano, T.C., Wood, A.W., Ramos, M.-H., Cloke, H.L., Pappenberger, F., Clark, M.P., Cranston, M., Kavetski, D., Mathevet, T., Sorooshian, S., Verkade, J.S., 2014. Challenges of operational river forecasting. J. Hydrometeorol. 140516115449007. https://doi.org/10.1175/JHM-D-13-0188.1

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). Hydrol. Earth Syst. Sci. 9, 381–393. https://doi.org/10.5194/hess-9-381-2005

Pappenberger, F., Scipal, K., Buizza, R., 2008. Hydrological aspects of meteorological verificatio. Atmos. Sci. Lett. 9, 43–52. https://doi.org/10.1002/asl.171

Rebora, N., Ferraris, L., Von Hardenberg, J., Provenzale, A., 2006. Rainfall downscaling and

flood forecasting: a case study in the Mediterranean area. Nat. Hazards Earth Syst. Sci. 6, 611–619.

Robertson, D.E., Shrestha, D.L., Wang, Q.J., 2013. Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. Hydrol. Earth Syst. Sci. 17, 3587–3603. https://doi.org/10.5194/hess-17-3587-2013

Yucel, I., Onen, A., Yilmaz, K.K., Gochis, D.J., 2015. Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation, and satellite-based rainfall. J. Hydrol. 523, 49–66. https://doi.org/10.1016/j.jhydrol.2015.01.042