DATA ASSIMILATION FOR ENHANCED HYDROLOGIC FORECASTING

ADVANCING SEQUENTIAL DATA ASSIMILATION METHODS FOR ENHANCED HYDROLOGIC FORECASTING IN SEMI-URBAN WATERSHEDS

By JAMES MICHAEL LEACH, M.A.Sc., B.A.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: ADVANCING SEQUENTIAL DATA ASSIMILATION METHODS FOR ENHANCED HYDROLOGIC FORECASTING IN SEMI-URBAN WATERSHEDS

AUTHOR: James Michael Leach, M.A.Sc. (McMaster University), B.A.Sc. (University of Regina)

SUPERVISOR: Dr. Paulin Coulibaly

NUMBER OF PAGES: xxx, 203

Lay Abstract

The ability to accurately model hydrological systems is essential, as that allows for better planning and decision making in water resources management. The better we can forecast the hydrologic response to rain and snowmelt events, the better we can plan and manage our water resources. This includes better planning and usage of water for agricultural purposes, better planning and management of reservoirs for power generation, and better preparing for flood events. Unfortunately, hydrologic models primarily used are simplifications of the real world and are therefore imperfect. Additionally, our measurements of the physical system responses to atmospheric forcing can be prone to both systematic and random errors that need to be accounted for. To address these limitations, data assimilation can be used to improve hydrologic forecasts by optimally accounting for both model and observation uncertainties. The work in this thesis helps to further advance and improve data assimilation, with a focus on enhancing hydrologic forecasting in urban and semi-urban watersheds. The research presented herein can be used to provide better forecasts, which allow for better planning and decision making.

Abstract

Accurate hydrologic forecasting is vital for proper water resource management. Practices that are impacted by these forecasts include power generation, reservoir management, agricultural water use, and flood early warning systems. Despite these needs, the models largely used are simplifications of the real world and are therefore imperfect. The forecasters face other challenges in addition to the model uncertainty, which includes imperfect observations used for model calibration and validation, imperfect meteorological forecasts, and the ability to effectively communicate forecast results to decision-makers. Bayesian methods are commonly used to address some of these issues, and this thesis will be focused on improving methods related to recursive Bayesian estimation, more commonly known as data assimilation.

Data assimilation is a means to optimally account for the uncertainties in observations, models, and forcing data. In the literature, data assimilation for urban hydrologic and flood forecasting is rare; therefore the main areas of study in this thesis are urban and semi-urban watersheds. By providing improvements to data assimilation methods, both hydrologic and flood forecasting can be enhanced in these areas. This work explored the use of alternative data products as a type of observation that can be assimilated to improve hydrologic forecasting in an urban watershed. The impact of impervious surfaces in urban and semi-urban watersheds was also evaluated in regards to its impact on remotely sensed soil moisture assimilation. Lack of observations is another issue when it comes to data assimilation, particularly in semi- or fully-distributed models; because of this, an improved method for updating locations which do not have observations was developed which utilizes information theory's mutual information. Finally, we explored extending data assimilation into the short-term forecast by using prior knowledge of how a model will respond to forecasted forcing data.

Results from this work found that using alternative data products such as those from the Snow Data Assimilation System or the Soil Moisture and Ocean Salinity mission, can be effective at improving hydrologic forecasting in urban watersheds. They also were effective at identifying a limiting imperviousness threshold for soil moisture assimilation into urban and semi-urban watersheds. Additionally, the inclusion of mutual information between gauged and ungauged locations in a semi-distributed hydrologic model was able to provide better state updates in models. Finally, by extending data assimilation into the short-term forecast, the reliability of the forecasts could be improved substantially.

Acknowledgements

I would like to express my gratitude to my supervisor Dr. Paulin Coulibaly for his guidance, support and mentoring over the course of my doctorate studies. This achievement would not have been possible without him

I would also like to thank the members of my supervisory committee, Dr. Yipping Guo and Dr. Sarah Dickson, for their time and support.

I am also thankful to Dr. Jongho Keum and Dr. Kurt Kornelsen who were willing to share of their time and knowledge when I needed it.

Last I would like to thank my wife and family for their support and encouragement over the years.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AHPS	Advanced Hydrologic Prediction System
AMSR	Advanced Microwave Scanning Radiometer
AMSR-E	Advanced Microwave Scanning Radiometer for Earth observation science
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
ASCAT	Advanced Scatterometer
BCB	Black Creek Basin
BFS	Bayesian forecasting system
BMA	Bayesian model averaging
CDD	Canadian Disasters Database
CDF	Cumulative distribution function
CN	Curve Number
CONUS	Contiguous United States
CRPS	Mean Continuous Ranked Probability Score
$CRPS_{pot}$	potential CRPS
DA	Data assimilation
DDS	Dynamically Dimensioned Search algorithm
DEnKF	Distributed Ensemble Kalman Filter
DGG	Discrete Global Grid
DL	Direct Lookup
DRB	Don River Basin
ECCC	Environment and Climate Change Canada
EDA	Evolutionary Data Assimilation
EFAS	European Flood Awareness System
EKF	Extended Kalman Filter
EnKF	Ensemble Kalman Filter
EnKS	Ensemble Kalman Smoother
EPFM	Evolutionary Particle Filter Markov Chain Monte Carlo
ER95	95% Exceedance Ratio
ESA	European Space Agency
fdb	forecast forcing database
FEWS	Flood Early Warning System
FNN	Feedforward Neural Network
GR4J	modèle du Génie Rural à 4 paramètres Journalier
hdb	historical observation database
HEAVEN	Hybrid Ensemble and Variational Data Assimilation Framework
HRB	Humber River Basin
HRU	hydrologic response unit
HYDAT	Hydrometric Database
HYMOD	Hydrologic model

ISEA	Icosahedral Snyder Equal Area
JB	Jarque-Bera
KGE	Kling-Gupta efficiency
k-NN	k-Nearest Neighbour
KW	Kruskal-Wallis
LCD	Local Climatological Data
MAC-HBV	McMaster University - Hydrologiska Byråns Vattenbalansavdelning
MAP	Mean Areal Precipitation
masl	meters above mean seal level
MAT	Mean Areal Temperature
MCMC	Markov Chain Monte Carlo
MCP	Model Conditional Processor
MI	Mutual information
MIEnKF	Mutual information Ensemble Kalman Filter
MK	Mann-Kendall
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
MOEA	Multi-objective evolutionary algorithm
MRC	Multiplicative Random Cascade
NASA	National Aeronautics and Space Administration
NCEP-CPC	National Centers for Environmental Prediction - Climate Prediction Center
NLCD	National Land Cover Database
NOAA	National Oceanic and Atmospheric Administration
NOHRSC	National Operational Hydrologic Remote Sensing Center
NRCS	Natural Resources Conservation Service
NRR	Normalized RMSE Ratio
NSE	Nash-Sutcliffe Efficiency
NSE_{log}	NSE on log transformed data
NSERC	Natural Science and Engineering Research Council
NSE_{sqr}	NSE on square transformed data
NSM	NOHRSC Snow Model
NVE	Nash-Volume Efficiency
NWP	Numerical Weather Prediction
ODE	Ordinary Differential Equation
OL	Open-loop
pdf	probability density function
PDM	probability distributed moisture model
PF	Particle Filter
PFC	Peak Flow Criteria
PF-MCMC	Particle Filter Markov Chain Monte Carlo

ppm	Parts per million
PSAS	Physical Space Statistical Analysis System
PSO	Particle swarm optimization
RDPS	Regional Deterministic Prediction System
RFI	Radio Frequency Interference
RMSE	Root Mean Square Error
RR	Residual Resampling
SAC-SMA	Sacramento Soil Moisture Accounting
SDA	Sequential data assimilation
SIR	Sample Importance Resampling/Sequential Importance Resampling
SIRV	Sample Importance Resampling with Variable variance multiplier
SM	Soil Moisture
SMAP	Soil Moisture Active Passive
SMOS	Soil Moisture and Ocean Salinity
SNODAS	Snow Data Assimilation System
SNOW17	Snow accumulation and ablation model
SWE	Snow Water Equivalent
SWMM	Storm Water Management Model
TB1	Test basin 1
TB2	Test basin 2
TRCA	Toronto Region Conservation Authority
VAR	Variational Data Assimilation
VCA	Vertex Component Analysis
VE	Volume Error
VVM	Variable Variance Multiplier
WSC	Water Survey of Canada

List of Symbols

-, b	(superscript) prior (background) estimate
+, a	(superscript) posterior (analysis) estimate
A	sub-basin area
b	bias vector
CN, c, I_a, S,	the NRCS curve number parameters
d, g, o, u, x, y, z, Q, R, S, X, Y	random variables
D	database vector
e, f, i, q	the surface evaporation rate, infiltration rate, rainfall and snowmelt rate, and runoff rate in the SWMM model
H, h(.)	the observation operator
i	(subscript) bin number
I	available information
j	(subscript) ensemble member
K	Kalman gain
n, N, T	counters for bins, random variables, ensemble members, time steps
p, P	probability
r, r_{XY}	Pearson correlation coefficient
R_{XY}	informational coefficient of correlation
t	(subscript) the time step
T	(superscript) is the transpose operator
w, **W**	weight vector or matrix
ζ, η, ε, ν	random noise
θ	vector of parameters
Σ, R, P^b	variance or covariance
σ, ρ, β, γ, ρ, ω, μ, α, r, W, n, d_s	scalar parameters

Declaration of Academic Achievement

This thesis has been prepared in accordance with the regulations for a sandwich thesis format as a compilation of papers stipulated by the Faculty of Graduate Studies at McMaster University and as such, the publications that comprise Chapters 2, 3, 4, and 5 of this theses have been co-authored.

Chapter 2: Assimilation of near-real time data products into models of an urban basin, by J.M. Leach, K.C. Kornelsen, and P. Coulibaly, Journal of Hydrology, 563, 51-64. doi: 10.1016/j.jhydrol.2018.05.064, 2018. (With permission from publisher)

Chapter 3: The limits of soil moisture assimilation in urban watersheds, by J.M. Leach and P. Coulibaly, Journal of Hydrology, Submitted, Manuscript Number HYDROL33206.

Chapter 4: Data assimilation in ungauged basins using mutual information, by J.M. Leach and P. Coulibaly, Journal of the American Water Resources Association, Submitted, Manuscript Number JAWRA-19-0136-P.

Chapter 5: An extension of data assimilation into the short-term hydrologic forecast for improved prediction reliability, by J.M. Leach and P. Coulibaly, Advances in Water Resources, doi:10.1016/j.advwatres.2019.103443, 2019. (With permission from publisher)

All of the articles were prepared by James M. Leach. Dr. Paulin Coulibaly provided guidance for the research and aided in editing each manuscript. Dr. Kurt C. Kornelsen aided in the understanding and use of the Soil Moisture and Ocean Salinity (SMOS) mission data for Chapter 2. The work reported in this thesis was undertaken from September 2015 to October 2019.

CHAPTER 1 Introduction

1.1 Hydrologic and flood forecasting

Hydrologic forecasting can be performed using data-driven methods such as Artificial Neural Networks (ANN), statistical models such as Autoregressive Integrated Moving Average (ARIMA) and Multiple Linear Regression (MLR), and conceptual models (Hapuarachchi et al., 2011; Singh and Woolhiser, 2002). Some of these methods can be used under a lumped, semi-distributed, or full-distributed framework. Data-driven methods have been used for flow forecasting because they are simple to set up, can provide results with minimal input data, and when little is known about the underlying interactions and physical processes they can be good alternative methods (Hapuarachchi et al., 2011). However, long-term data records are required for calibration, and they do not provide predictive uncertainty. Due to the ease of use, low computational cost, and simplicity, lumped models are a popular choice for hydrologic forecasting. Semi- and fully-distributed models, however, are expected to provide more accurate results than lumped models when there is sufficient data to properly develop them (Singh and Woolhiser, 2002; Carpenter and Georgakakos, 2006). Both of these model types are not without their disadvantages. Lumped models have inherently higher uncertainty in their results, and distributed models have a much higher computational cost and data requirement (Moore et al., 2006; Singh and Woolhiser, 2002; Young, 2002).

There are several modeling approaches used for flood forecasting, each with different levels of uncertainty and forecasting lead times. Generally, the methods used to forecast river discharges with the least amount of uncertainty are those which use a purely hydraulic approach. However, these models are limited when it comes to forecasting lead times. One way to improve on these lead times is to incorporate the use of hydrological rainfall-runoff models. Several flood forecasting systems have been developed around the world. The goal of these systems is to provide real-time forecasts that help institutions and individuals make informed decisions that help minimize damages attributed to extreme weather events such as flooding. These systems include the Advanced Hydrologic Prediction System (AHPS) in the US (http://water.weather.gov/ahps/) and the European Flood Awareness System (EFAS) in Europe (https://www.efas.eu/). To help implement flood forecasting services, shell systems such as the Delft Flood Early Warning System (FEWS) have been developed. These shell systems are used to incorporate several data sources and forecasting models (Werner et al., 2013, 2004).

1.1.1 Rational

Hydrologic and flood forecasting are essential aspects of water resources management. Accurate forecasting leads to more efficient reservoir management, hydropower generation, agricultural water use, and more effective flood early warning systems. It is especially important to have accurate forecasts in urban watersheds where there is the most potential for impact on large numbers of people during flood events. In Canada, between 1900 and 2016, there have been 309 flood events recorded by the Canadian Disasters Database (CDD), which have lead to 115 fatalities and nearly 400,000 people needing to be evacuated (Public Safety Canada, 2019). Additionally, these flood events have had an estimated total cost of 9.75 billion dollars (Public Safety Canada, 2019). It is important to note that the events recorded by the CDD are not comprehensive; they only record events that have direct and significant impact to people. Figure 1-1 illustrates the frequency of various weather events between 1907 and 2016 and compares them to increasing atmospheric carbon dioxide (CO_2) levels and the increasing Canadian population. There is an increasing risk of extreme events associated with climate change (Seneviratne et al., 2012); which, along with increasing population, leads to more extreme weather and flood events that impact people. Based on this data we can see there is an increasing need for flood forecasting, which leads to need for researching ways to improve those forecasts.

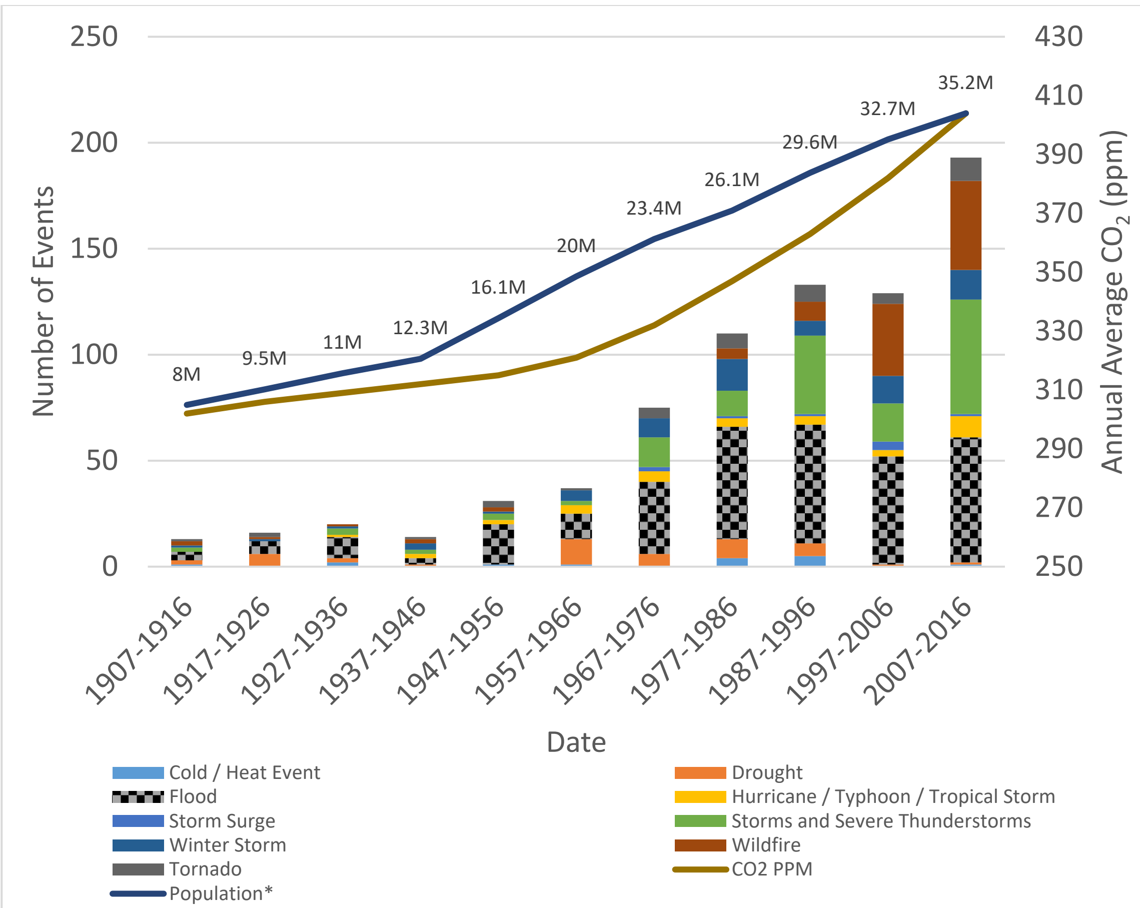

Figure 1-1: Frequency of meteorological and hydrological natural disasters in Canada from 1907 to 2016 (Public Safety Canada, 2019). Frequency is compared to the global annual average CO_2 in parts per million (ppm) (Dlugokencky et al., 2019; Institute for Atmospheric and Climate Science, 2016) and the Canadian population (Statistics Canada, 2018, 2015). Population* is shown on the same scale as the annual average CO_2.

1.1.2 Deterministic, ensemble, and probabilistic methods

There are several methods used for hydrologic forecasting, which can be broadly classified as either deterministic, ensemble, or probabilistic methods. Deterministic methods are the simplest form of forecasting. A deterministic approach will take one input and provide one output without considering any of the associated uncertainties. This can be beneficial when computation time is an issue; additionally it makes interpreting the results straightforward. However, when the inherent uncertainties are not accounted for, the forecasts can be less reliable; additionally, with the increase of computational power it can be just as efficient to use ensemble methods.

Ensemble methods provide an extension to deterministic models and can be used to quantify the various sources of uncertainty that can influence hydrological forecasts. In the most basic sense, there are three ways ensembles are used in hydrologic modeling: (1) performing the forecast multiple times with perturbed forcing data as a way to account for forcing uncertainty, (2) using multiple hydrologic models to account for uncertainty related to each model's structure, and (3) using a combination of both (1) and (2). Probabilistic methods are not independent of ensemble methods, as both can be used to quantify uncertainty better and provide more informative forecasts. These forecasts are generally superior to deterministic ones because they can provide probability associated with a forecast; though this can potentially make interpreting the results more difficult for those without a background in statistics. Bayesian methods such as Bayesian Model Averaging (BMA; Najafi and Moradkhani, 2016; Raftery et al., 2005), Sequential Data Assimilation (SDA; Liu et al., 2012; Moradkhani, 2008), and Bayesian Forecasting System (BFS;

Krzysztofowicz, 1999) are commonly used for ensemble and probabilistic hydrologic forecasting. The predictive distribution of multiple models can be combined using BMA. BFS can be used to better quantify predictive uncertainty in forecasts for deterministic or ensemble methods (Han and Coulibaly, 2019, 2017; Krzysztofowicz, 1999). Data assimilation can be used to optimally merge the various sources of uncertainty in hydrologic models such as forcing data, initial condition, parameter, model, and observation uncertainties (Moradkhani et al., 2006). The research presented in this thesis will be focused explicitly on improving SDA methods.

1.2 Data assimilation in hydrology

Data assimilation methods have been increasingly used to improve hydrologic forecasting by updating estimations as new information becomes available (Dumedah and Coulibaly, 2012; Komma et al., 2008; Li et al., 2013; Wanders et al., 2014). They do this by optimally merging the imperfect model and uncertain data in a way that reduces and quantifies the uncertainty in the system (Liu and Gupta, 2007). The most common assimilated observation in hydrology is streamflow (Abbaszadeh et al., 2018, 2019; DeChant and Moradkhani, 2012; Moradkhani et al., 2005b; Seo et al., 2003; Thiboult et al., 2015). However, other informative variables have been assimilated as well such as soil moisture (Alvarez-Garreton et al., 2014, 2015; Dumedah and Coulibaly, 2013a; Samuel et al., 2014) and snow water equivalent (Bergeron et al., 2016; Dziubanski and Franz, 2016; Franz et al., 2014; Huang et al., 2017). These observations are available from in-situ gauges or through remote sensing methods. In general, the more accurate the assimilated observation is, the closer the assimilation estimates will be to the measured values.

Likewise, if that observation is inaccurate or unavailable, the assimilation estimate will be closer to the modeled solution (Reichle, 2008). There are several sequential data assimilation methods with the most popular being the Ensemble Kalman Filter and the Particle Filter.

1.2.1 Ensemble Kalman filter

The Ensemble Kalman filter (EnKF) proposed by Evensen (1994) uses randomly generated ensemble members to estimate the probability density function (pdf) of state variables (Burgers et al., 1998; Evensen, 2003, 1994). It was developed as a nonlinear extension to the Kalman filter to address filtering problems (Liu et al., 2012). The ensemble members are generated by randomly perturbing input values as they are propagated to the next time step. These input values include model parameters, states, forcing data, and their uncertainties. The ensemble members are then input into the chosen hydrologic model to produce an ensemble of predictions. The ensemble members are updated using the Kalman gain function, which is computed using the covariance between states, parameters and forcing data as well as the residual between the simulated output and perturbed observations (Burgers et al., 1998; Evensen, 2003, 1994). Some advantages to using the EnKF are its ease of use: it does not require the use of a model in state-space form, it does not require an adjoint model, and it does not require temporally constant error covariances. However, the EnKF does assume a Gaussian distribution for model errors, and it is known to have issues when there are strong non-linear relationships between model states and observations (Clark et al., 2008; Moradkhani et al., 2005a). Optimal implementation of the EnKF can also be difficult as some models have compatibility issues with EnKF updating (Thiboult and Anctil, 2015).

There are several studies in which EnKFs are used in hydrologic modeling and forecasting (Abaza et al., 2014; Clark et al., 2008; Dumedah and Coulibaly, 2013b; Komma et al., 2008; Moradkhani et al., 2005b; Neal et al., 2007; Samuel et al., 2014; Thiboult and Anctil, 2015; Vrugt and Robinson, 2007; Wanders et al., 2014; Weerts and El Serafy, 2006). By using a multi-model approach, Thiboult and Anctil (2015) found it was easier to implement the EnKF and account for possible compatibility issues between the models and the EnKF. Samuel et al. (2014) used the SAC-SMA model and EnKF with dual state parameter estimation. They found that by using both streamflow and soil moisture observations together to update state and model parameters, streamflow forecasts would be more accurate than when using streamflow or soil moisture alone (Samuel et al., 2014). Abaza et al. (2014) used the EnKF with hydrotel for streamflow forecasting of up to 240 hours, and they compared the forecasting results of the EnKF with a manual assimilation method as well as the reference model. The results showed the EnKF had more accurate estimates than both the reference model and the manual assimilation method. Clark et al. (2008) found that by performing a log transformation on streamflow before computing error covariances improved the EnKF ability to deal with non-linear relationships between the observations and model states. Using the EnKF, it was observed that when remotely sensed soil moisture data and observed discharge was assimilated in the distributed hydrological model LISFLOOD as part of EFAS, the timing errors in the flood predictions were decreased, especially for shorter lead times (Wanders et al., 2014). Komma et al.

(2008) showed improved performance for 3 and 6-hour lead times when updating soil moisture with observed runoff in a real-time mode EnKF concept and an iterative similarity approach. Using the EnKF with a 1D hydraulic model, Neal et al. (2007) showed that updating model state and boundary conditions provided more accurate flood forecasts.

1.2.2 Particle filter

The Particle filter (PF) was developed by Gordon et al. (1993) and improved with residual resampling by Liu and Chen (1998). It approximates the state posterior pdf using prior knowledge of the state, likelihood, and observation data, with a set of weighted particles using recursive Bayesian estimation (Bengtsson et al., 2003; Snyder et al., 2008; van Leeuwen, 2009). The posterior pdf is generated from the random particles, and their associated likelihoods are determined using the residuals between the simulation and observation. Updates are usually performed indirectly on the particles (Weerts and El Serafy, 2006). PFs can be used to propagate Gaussian and non-Gaussian distributions through both linear and non-linear models unlike the standard Kalman filter which relies on the models being linear and the assumption that the distributions are Gaussian (Liu et al., 2012; Moradkhani et al., 2005a).

Originally, a large number of particles were needed to avoid the collapse of particle weights (Snyder et al., 2008), and particle weights tended to degenerate after a few iterations (Clark et al., 2008; van Leeuwen, 2009; Weerts and El Serafy, 2006). Using sequential importance resampling (SIR) and residual resampling (RR), variance reduction approaches where low weighted particles are discarded and replaced with high normalized weighted particles, the weight degeneration could be reduced (Snyder et al., 2008; van

Leeuwen, 2009). More advanced resampling methods have since addressed the issues relating to particle degeneracy. First by using Markov Chain Monte Carlo resampling (PF-MCMC) (Moradkhani et al., 2012) and Evolutionary PF-MCMC (EPFM) (Abbaszadeh et al., 2018) the number of particles and degeneracy issues could be reduced greatly, and the most recent Hybrid Ensemble and Variational Data Assimilation Framework (HEAVEN) method has now removed particle degeneracy issues entirely (Abbaszadeh et al., 2019).

1.2.3 Other data assimilation methods

There are several other data assimilation methods which have been used with hydrologic models to improve forecasting. Some of these methods include the Extended Kalman Filter (EKF; Branisavljevic et al., 2014; Karunasingha and Liong, 2018; Sun et al., 2015), Evolutionary Data Assimilation (EDA; Dumedah and Coulibaly, 2014, 2013b, 2013a), and Variational data assimilation (VAR; Alvarado-Montero et al., 2017; Lee et al., 2012, 2011; Montero et al., 2016; Seo et al., 2009, 2003).

The Extended Kalman Filter is a nonlinear extention to the Kalman filter; it uses a Taylor Series expansion of a model around a point and ignores higher moment terms as a way to linearize the models (Wishner et al., 1969). Sun et al. (2015) found that the EKF could perform well during flood rising periods and streamflow forecasts could be improved for short lead times. By using the EKF to assimilate water level data into an urban rainfall-runoff model, Branisavljevic et al. (2014) showed that water level in a retention pond could be better simulated.

Evolutionary Data Assimilation employs a flexible procedure which can be applied for both sequential and smoothing problems (Dumedah, 2012). EDA employs a multi-objective evolutionary algorithm (MOEA) combined with the cost function from a variational data assimilation approach to evolve a population of solutions through several cycles of evolution (Dumedah, 2012). MOEAs employ stochastic search algorithms that utilize the concepts of evolution and natural selection to find solutions to problems (Deb, 2001; Dumedah and Coulibaly, 2013a). The competition and natural selection in the EDA/MOEA is comparable to the Kalman gain function of the EnKF and the ensemble weights of the PF (Dumedah, 2012). EDA is characterized by the inclusion of the assimilated ensemble members into subsets of all the members that are evaluated for a time step, and that solutions are evolved both between time steps and several cycles during each time step (Dumedah, 2012). Pareto-dominance is employed to ensure that the evolved solutions are competitive under multiple evaluation conditions. The Pareto-optimal populations from each current time step are then used as the assimilated ensembles which feed into the next time steps (Dumedah, 2012). It was shown by Dumedah and Coulibaly (2013a) that EDA can estimate model state and parameterizations simultaneously for real-time forecasting, as well as improve both streamflow and soil moisture estimates.

Variational data assimilation methods are more commonly used in meteorological and oceanographic forecasting, however they have been used for hydrologic forecasting as well. They can achieve the optimal performance of Kalman filters while being more computationally efficient since they do not explicitly evaluate large error covariance matrices. Instead, variational algorithms can be used to process all data within a given assimilation interval, and take dynamic error information into account by propagating an adjoint variable (Reichle, 2008; Reichle et al., 2001). In general, variational methods produce an optimal state estimate that balances uncertainties in the model initial conditions, and measurement errors by minimizing a weighted least squares cost function. If the performance function is minimizing terms for single instants in time, the variational data assimilation method used to handle them include Optimal Interpolation, Physical Space Statistical Analysis System (PSAS), 1DVAR, and 3DVAR. If the objective function contains measurements at several different times within an assimilation interval, the assimilation method is known as 4DVAR. The 4DVAR includes dynamic features such as the propagation of the model to the exact time of the observation, and the evolution of the background error covariance within the assimilation interval (Reichle, 2008; Reichle et al., 2001). Seo et al. (2009) showed that using variational data assimilation with the Sacramento-Soil Moisture Accounting model (SAC-SMA) and the unit hydrograph model, forecast improvement was more significant for low stages and slow responding basins than it was for high stages and fast-responding basins. The more important factors limiting the performance of variational methods include large structural and/or parametric errors in soil moisture accounting and routing models and lack of flow-dependent modeling of uncertainty. Errors in hydrologic forecasting have been efficiently reduced by using variational data assimilation techniques such as 4DVAR (Bélanger and Vincent, 2005).

1.3 Research objectives

Data assimilation is an effective tool for improving hydrologic modeling. The proposed research aims to further advance sequential data assimilation methods, with a focus on urban and semi-urban watersheds. These improvements include evaluating the efficacy of assimilating data products in an urban watershed, providing justification for assimilating soil moisture into urban and semi-urban watershed models, modifying the EnKF to provide better updates in semi-distributed hydrologic models, and enhancing forecast reliability. The contributions of this research include an improved methodology to follow when using data assimilation in urban and semi-urban watersheds that can help provide more accurate forecasts. Additionally, this research provides an improved methodology for data assimilation in ungauged basins.

1.3.1 Thesis outline

This thesis consists of six chapters. Chapter 1 provides an introduction to hydrologic forecasting and data assimilation methods. It provides a general context for the research as well as outlines the research objectives for the remainder of the thesis.

Chapter 2 presents a study in which the assimilation of derived data products for soil moisture and snow water equivalent were combined with streamflow data assimilation to improve urban watershed modeling. The chapter presents a comparison of different combinations of soil moisture, snow water equivalent, and streamflow assimilation schemes for both state and dual state and parameter updating on the Don River basin in southern Ontario using multiple hydrologic models. This work was published in the Journal of Hydrology.

Chapter 3 is a follow up to work presented in Chapter 2, as it further evaluates remotely sensed soil moisture assimilation in urban and semi-urban watersheds. The chapter attempts to identify a general imperviousness threshold for watersheds, beyond which it is unproductive to assimilate soil moisture. Additionally, the chapter presents a method that can be used to quickly determine if it would be beneficial to assimilate soil moisture into a basin given some basin characteristics. This work has been submitted to the Journal of Hydrology.

Chapter 4 presents a modification to the EnKF, which incorporates the mutual information from entropy or information theory. The modification to the EnKF is meant to aid in state updating of ungauged basins in semi-distributed hydrologic models when there are non-linear dependencies present. This work has been submitted to the Journal of the American Water Resources Association.

Chapter 5 extends data assimilation into the short-term forecast by utilizing a prebuilt database of observations, states, predictions, and forcing data. The purpose of this chapter is to evaluate whether improvements to real-time forecasts can be made given the knowledge of how well, historically, the forecast performed. This work was published in Advances in Water Resources journal.

Chapter 6 is a summary of the main conclusions as well as recommendations for future research.

1.4 References

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

Abbaszadeh, P., Moradkhani, H., Daescu, D.N., 2019. The Quest for Model Uncertainty Quantification: A Hybrid Ensemble and Variational Data Assimilation Framework. Water Resour. Res. 55, 2407–2431. https://doi.org/10.1029/2018WR023629

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

Alvarado-Montero, R., Schwanenberg, D., Krahe, P., Helmke, P., Klein, B., 2017. Multi-parametric variational data assimilation for hydrological forecasting. Adv. Water Resour. 110, 182–192. https://doi.org/10.1016/j.advwatres.2017.09.026

Alvarez-Garreton, C., Ryu, D., Western, A.W., Crow, W.T., Robertson, D.E., 2014. The impacts of assimilating satellite soil moisture into a rainfall–runoff model in a semi-arid catchment. J. Hydrol. 519, 2763–2774. https://doi.org/10.1016/j.jhydrol.2014.07.041

Alvarez-Garreton, C., Ryu, D., Western, A.W., Su, C.H., Crow, W.T., Robertson, D.E., Leahy, C., 2015. Improving operational flood ensemble prediction by the assimilation of satellite soil moisture: Comparison between lumped and semi-distributed schemes. Hydrol. Earth Syst. Sci. 19, 1659–1676. https://doi.org/10.5194/hess-19-1659-2015

Bélanger, E., Vincent, A., 2005. Data assimilation (4D-VAR) to forecast flood in shallow-waters with sediment erosion. J. Hydrol. 300, 114–125. https://doi.org/10.1016/j.jhydrol.2004.06.009

Bengtsson, T., Snyder, C., Nychka, D., 2003. Toward a nonlinear ensemble filter for high-dimensional systems. J. Geophys. Res. Atmos. 108, n/a-n/a. https://doi.org/10.1029/2002JD002900

Bergeron, J.M., Trudel, M., Leconte, R., 2016. Combined assimilation of streamflow and snow water equivalent for mid-term ensemble streamflow forecasts in snow-dominated regions. Hydrol. Earth Syst. Sci. Discuss. 1–34. https://doi.org/10.5194/hess-2016-166

Branisavljevic, N., Hutton, C., Kapelan, Z., Vamvakeridou-Lyroudia, L., Savić, D.A., 2014. Real-Time Runoff Prediction Based on Data Assimilation And Model Bias Reduction, in: 11th International Conference on Hydroinformatics. New York, NY.

Burgers, G., Jan van Leeuwen, P., Evensen, G., 1998. Analysis Scheme in the Ensemble Kalman Filter. Mon. Weather Rev. 126, 1719–1724. https://doi.org/10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2

Carpenter, T.M., Georgakakos, K.P., 2006. Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. J. Hydrol. 329, 174–185. https://doi.org/10.1016/j.jhydrol.2006.02.013

Clark, M.P., Rupp, D.E., Woods, R.A., Zheng, X., Ibbitt, R.P., Slater, A.G., Schmidt, J., Uddstrom, M.J., 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Adv. Water Resour. 31, 1309–1324. https://doi.org/10.1016/j.advwatres.2008.06.005

Deb, K., 2001. Multi-objective optimization using evolutionary algorithms. Wiley, Chichester, New York, New York.

DeChant, C.M., Moradkhani, H., 2012. Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting. Water Resour. Res. 48, n/a-n/a. https://doi.org/10.1029/2011WR011011

Dlugokencky, E., Tans, P., NOAA/ESRL, 2019. Trends in Atmospheric Carbon [WWW Document]. URL www.esrl.noaa.gov/gmd/ccgg/trends/

Dumedah, G., 2012. Formulation of the Evolutionary-Based Data Assimilation, and its Implementation in Hydrological Forecasting. Water Resour. Manag. 26, 3853–3870. https://doi.org/10.1007/s11269-012-0107-0

Dumedah, G., Coulibaly, P., 2014. Integration of an evolutionary algorithm into the ensemble Kalman filter and the particle filter for hydrologic data assimilation. J. Hydroinformatics 16, 74–94. https://doi.org/10.2166/hydro.2013.088

Dumedah, G., Coulibaly, P., 2013a. 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

Dumedah, G., Coulibaly, P., 2013b. Evolutionary assimilation of streamflow in distributed hydrologic modeling using in-situ soil moisture data. Adv. Water Resour. 53, 231–241. https://doi.org/10.1016/j.advwatres.2012.07.012

Dumedah, G., Coulibaly, P., 2012. Integration of evolutionary based assimilation into Kalman-type methods for streamflow simulations in ungauged watersheds. J. Hydrol. 475, 428–440. https://doi.org/10.1016/j.jhydrol.2012.10.033

Dziubanski, D.J., Franz, K.J., 2016. Assimilation of AMSR-E snow water equivalent data in a spatially-lumped snow model. J. Hydrol. 540, 26–39. https://doi.org/10.1016/j.jhydrol.2016.05.046

Evensen, G., 2003. The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367. https://doi.org/10.1007/s10236-003-0036-9

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143. https://doi.org/10.1029/94JC00572

Franz, K.J., Hogue, T.S., Barik, M., He, M., 2014. Assessment of SWE data assimilation for ensemble streamflow predictions. J. Hydrol. 519, 2737–2746. https://doi.org/10.1016/j.jhydrol.2014.07.008

Gordon, N.J., Salmond, D.J., Smith, A.F.M., 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proc. F Radar Signal Process. 140, 107. https://doi.org/10.1049/ip-f-2.1993.0015

Han, S., Coulibaly, P., 2019. Probabilistic Flood Forecasting Using Hydrologic Uncertainty Processor with Ensemble Weather Forecasts. J. Hydrometeorol. 20, 1379–1398. https://doi.org/10.1175/JHM-D-18-0251.1

Han, S., Coulibaly, P., 2017. Bayesian flood forecasting methods: A review. J. Hydrol. 551, 340–351. https://doi.org/10.1016/j.jhydrol.2017.06.004

Hapuarachchi, H.A.P., Wang, Q.J., Pagano, T.C., 2011. A review of advances in flash flood forecasting. Hydrol. Process. 25, 2771–2784. https://doi.org/10.1002/hyp.8040

Huang, C., Newman, A.J., Clark, M.P., Wood, A.W., Zheng, X., 2017. Evaluation of snow data assimilation using the ensemble Kalman filter for seasonal streamflow prediction in the western United States. Hydrol. Earth Syst. Sci. 21, 635–650. https://doi.org/10.5194/hess-21-635-2017

Institute for Atmospheric and Climate Science, 2016. Mole fraction of carbon dioxide in air [WWW Document]. URL ftp://data.iac.ethz.ch/CMIP6/input4MIPs/UoM/GHGConc/CMIP/yr/atmos/UoM-CMIP-1-1-0/GHGConc/gr3-GMNHSH/v20160701/ (accessed 9.20.10).

Karunasingha, D.S.K., Liong, S.-Y., 2018. Enhancement of chaotic hydrological time series prediction with real-time noise reduction using Extended Kalman Filter. J. Hydrol. 565, 737–746. https://doi.org/10.1016/j.jhydrol.2018.08.044

Komma, J., Blöschl, G., Reszler, C., 2008. Soil moisture updating by Ensemble Kalman Filtering in real-time flood forecasting. J. Hydrol. 357, 228–242. https://doi.org/10.1016/j.jhydrol.2008.05.020

Krzysztofowicz, R., 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resour. Res. 35, 2739–2750. https://doi.org/10.1029/1999WR900099

Lee, H., Seo, D.-J., Koren, V., 2011. Assimilation of streamflow and in situ soil moisture data into operational distributed hydrologic models: Effects of uncertainties in the data and initial model soil moisture states. Adv. Water Resour. 34, 1597–1615. https://doi.org/10.1016/j.advwatres.2011.08.012

Lee, H., Seo, D.-J., Liu, Y., Koren, V., McKee, P., Corby, R., 2012. Variational assimilation of streamflow into operational distributed hydrologic models: effect of spatiotemporal scale of adjustment. Hydrol. Earth Syst. Sci. 16, 2233–2251. https://doi.org/10.5194/hess-16-2233-2012

Li, Y., Ryu, D., Western, A.W., Wang, Q.J., 2013. Assimilation of stream discharge for flood forecasting: The benefits of accounting for routing time lags. Water Resour. Res. 49, 1887–1900. https://doi.org/10.1002/wrcr.20169

Liu, J.S., Chen, R., 1998. Sequential Monte Carlo Methods for Dynamic Systems. J. Am. Stat. Assoc. 93, 1032–1044. https://doi.org/10.1080/01621459.1998.10473765

Liu, Y., Gupta, H. V., 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resour. Res. 43, n/a-n/a. https://doi.org/10.1029/2006WR005756

Liu, Y., Weerts, A.H., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., van Dijk, A.I.J.M., van Velzen, N., He, M., Lee, H., Noh, S.J., Rakovec, O., Restrepo, P., 2012. Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. Hydrol. Earth Syst. Sci. 16, 3863–3887. https://doi.org/10.5194/hess-16-3863-2012

Montero, R.A., Schwanenberg, D., Krahe, P., Lisniak, D., Sensoy, A., Sorman, A.A., Akkol, B., 2016. Moving horizon estimation for assimilating H-SAF remote sensing data into the HBV hydrological model. Adv. Water Resour. 92, 248–257. https://doi.org/10.1016/j.advwatres.2016.04.011

Moore, R.J., Cole, S.J., Bell, V. a., Jones, D. a., 2006. Issues in flood forecasting: ungauged basins, extreme floods and uncertainty. Front. Flood Res. 305, 103–122.

Moradkhani, H., 2008. Hydrologic remote sensing and land surface data assimilation. Sensors 8, 2986–3004. https://doi.org/10.3390/s8052986

Moradkhani, H., DeChant, C.M., Sorooshian, S., 2012. Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method. Water Resour. Res. 48. https://doi.org/10.1029/2012WR012144

Moradkhani, H., Hsu, K.-L., Gupta, H., Sorooshian, S., 2005a. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. Water Resour. Res. 41, W05012. https://doi.org/10.1029/2004WR003604

Moradkhani, H., Hsu, K., Hong, Y., Sorooshian, S., 2006. Investigating the impact of remotely sensed precipitation and hydrologic model uncertainties on the ensemble streamflow forecasting. Geophys. Res. Lett. 33, L12401. https://doi.org/10.1029/2006GL026855

Moradkhani, H., Sorooshian, S., Gupta, H. V., Houser, P.R., 2005b. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour. 28, 135–147. https://doi.org/10.1016/j.advwatres.2004.09.002

Najafi, M.R., Moradkhani, H., 2016. Ensemble Combination of Seasonal Streamflow Forecasts. J. Hydrol. Eng. 21, 04015043. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001250

Neal, J.C., Atkinson, P.M., Hutton, C.W., 2007. Flood inundation model updating using an ensemble Kalman filter and spatially distributed measurements. J. Hydrol. 336, 401–415. https://doi.org/10.1016/j.jhydrol.2007.01.012

Public Safety Canada, 2019. The Canadian Disaster Databse [WWW Document]. URL https://www.publicsafety.gc.ca/cnt/rsrcs/cndn-dsstr-dtbs/index-en.aspx

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Mon. Weather Rev. 133, 1155–1174. https://doi.org/10.1175/MWR2906.1

Reichle, R.H., 2008. Data assimilation methods in the Earth sciences. Adv. Water Resour. 31, 1411–1418. https://doi.org/10.1016/j.advwatres.2008.01.001

Reichle, R.H., Entekhabi, D., McLaughlin, D.B., 2001. Downscaling of radio brightness measurements for soil moisture estimation: A four-dimensional variational data assimilation approach. Water Resour. Res. 37, 2353–2364. https://doi.org/10.1029/2001WR000475

Samuel, J., Coulibaly, P., Dumedah, G., Moradkhani, H., 2014. Assessing model state and forecasts variation in hydrologic data assimilation. J. Hydrol. 513, 127–141. https://doi.org/10.1016/j.jhydrol.2014.03.048

Seneviratne, S.I., Nicholls, N., Easterling, D., Goodess, C.M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., Zhang, X., 2012. Changes in climate extremes and their impacts on the natural physical environment., in: Field, C.B., Barros, V., Stocker, T.F., Qin, D., Dokken, D.J., Ebi, K.L., Mastrandrea, M.D., Mach, K.J., Plattner, G.-K., Allen, S.K., Tignor, M., Midgley, P.M. (Eds.), Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC). Cambridge University Press, Cambridge, UK, and New York, NY, USA, pp. 109–230.

Seo, D.-J., Cajina, L., Corby, R., Howieson, T., 2009. Automatic state updating for operational streamflow forecasting via variational data assimilation. J. Hydrol. 367, 255–275. https://doi.org/10.1016/j.jhydrol.2009.01.019

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

Singh, V.P., Woolhiser, D.A., 2002. Mathematical Modeling of Watershed Hydrology. J. Hydrol. Eng. 7, 270–292. https://doi.org/10.1061/(ASCE)1084-0699(2002)7:4(270)

Snyder, C., Bengtsson, T., Bickel, P., Anderson, J., 2008. Obstacles to High-Dimensional Particle Filtering. Mon. Weather Rev. 136, 4629–4640. https://doi.org/10.1175/2008MWR2529.1

Statistics Canada, 2018. Population and Dwelling Count Highlight Tables, 2016 Census [WWW Document]. URL https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Comprehensive.cfm

Statistics Canada, 2015. Estimated population of Canada, 1605 to present [WWW Document]. URL https://www150.statcan.gc.ca/n1/pub/98-187-x/4151287-eng.htm

Sun, L., Nistor, I., Seidou, O., 2015. Streamflow data assimilation in SWAT model using Extended Kalman Filter. J. Hydrol. 531, 671–684. https://doi.org/10.1016/j.jhydrol.2015.10.060

Thiboult, A., Anctil, F., 2015. On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments. J. Hydrol. 529, 1147–1160. https://doi.org/10.1016/j.jhydrol.2015.09.036

Thiboult, A., Anctil, F., Boucher, M. -a., 2015. Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrol. Earth Syst. Sci. Discuss. 12, 7179–7223. https://doi.org/10.5194/hessd-12-7179-2015

van Leeuwen, P.J., 2009. Particle Filtering in Geophysical Systems. Mon. Weather Rev. 137, 4089–4114. https://doi.org/10.1175/2009MWR2835.1

Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43, n/a-n/a. https://doi.org/10.1029/2005WR004838

Wanders, N., Karssenberg, D., de Roo, A., de Jong, S.M., Bierkens, M.F.P., 2014. The suitability of remotely sensed soil moisture for improving operational flood forecasting. Hydrol. Earth Syst. Sci. 18, 2343–2357. https://doi.org/10.5194/hess-18-2343-2014

Weerts, A.H., El Serafy, G.Y.H., 2006. Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. Water Resour. Res. 42, n/a-n/a. https://doi.org/10.1029/2005WR004093

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., Heynert, K., 2013. The Delft-FEWS flow forecasting system. Environ. Model. Softw. 40, 65–77. https://doi.org/10.1016/j.envsoft.2012.07.010

Werner, M., van Dijk, M., Schellekens, J., 2004. DELFT-FEWS: an open shell flood forecasting system, in: S., L., Phoon, K., Babovic, V. (Eds.), 6th International Conference on Hydroinformatics. World Scientific Publishing Company, Singapore, pp. 1205–1212.

Wishner, R.P., Tabaczynski, J.A., Athans, M., 1969. A comparison of three non-linear filters. Automatica 5, 487–496. https://doi.org/10.1016/0005-1098(69)90110-1

Young, P.C., 2002. Advances in real-time flood forecasting. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 360, 1433–1450. https://doi.org/10.1098/rsta.2002.1008

CHAPTER 2 Assimilation of near-real time data products into models of an urban basin

Summary of Paper 1: Leach, J.M., Kornelsen, K.C., and Coulibaly, P. (2018). Assimilation of near-real time data products into models of an urban basin, Journal of Hydrology, 563, 51-64, doi: 10.1016/j.jhydrol.2018.05.064

Summary:

This research sets the focus for the rest of the thesis to be on the application and improvement of data assimilation in urban watersheds. This work used the Ensemble Kalman Filter (EnKF) to assimilation SNODAS Snow Water Equivalent and SMOS L2 Soil Moisture products into models of an urban basin to improve hydrologic forecasting.

The results of this research demonstrated:

- Simple models such as HyMod and GR4J can be better suited to modelling flashy urban watersheds compared to more complex models like SAC-SMA or MAC-HBV.
- Was able to combined assimilation of streamflow observations with SMOS soil moisture and/or SNODAS snow water equivalent data products into urban models using the Ensemble Kalman Filter for improved performance.
- Spectral unmixing can be used to better evaluate an urban basin's level of development for determining if it could benefit from soil moisture assimilation.
- Soil moisture assimilation can benefit urban hydrologic modelling and forecasting, and the soil moisture assimilation scheme was able to provide the best forecast results.

2.1 Abstract

The goal of this study was to determine if assimilating a combination of various derived data products can help circumvent some of the difficulties associated with urban watershed modeling. Combinations of the SNODAS (Snow Data Assimilation System) snow water equivalent data, the SMOS (Soil Moisture and Ocean Salinity) L2 soil moisture, and streamflow observations were used for the data assimilation schemes. Combinations of these observation data sets were assimilated into lumped conceptual rainfall-runoff models of the highly-urbanized Don River basin (in southern Ontario) to determine if assimilation of geophysical variables will have a significant impact on simulations and forecasting in an urbanized watershed. The Ensemble Kalman Filter (EnKF) data assimilation method was used for these analyses, with various rainfall-runoff models that include GR4J, HyMod, MAC-HBV, and SAC-SMA models. The best data assimilation scheme for hydrologic modeling involved using a combination of streamflow, soil moisture, and snow water equivalent while performing both state and parameter updating. These results suggest that using a combination of soil moisture and snow water equivalent from the SMOS and SNODAS data products can improve simulations and ensemble forecasts in an urban basin.

Keywords: Data assimilation, urban watershed, ensemble forecasting, hydrologic modeling, soil moisture, snow water equivalent

2.2 Introduction

Urbanization is an increasing global trend which can have impacts on the hydrology of a watershed. These impacts include an increase in impervious surfaces, reduced infiltration, lower baseflow, increased runoff, and more flashy-ness in the hydrograph, all of which contribute to the difficulty in simulating an urbanized basin (McPherson and Schneider, 1974). Urban hydrology is important to understand and model due to the impacts it has on the often dense local population. Therefore, it is important to determine simple and easy methods which can be used to overcome the difficulties and improve rainfall-runoff modeling in urban areas. Previous studies have assessed the use of data assimilation to improve urban basin modeling by integrating one observation type such as water level or discharge (Branisavljevic et al., 2014; Hutton et al., 2014). This paper will explore the improvements data assimilation, which integrates multiple observation types, can have on urban basin modeling with various conceptual rainfall-runoff models.

Both the soil moisture and snow water equivalent (SWE) play important roles in the hydrology of a watershed and have been shown to improve streamflow estimations when assimilated into a hydrologic model (Huang et al., 2017; Moradkhani, 2008; Samuel et al., 2014). Better quantification of soil moisture in a watershed leads to a more accurate estimation of the rainfall quantity that becomes runoff or infiltration. In northern and/or snow-dominated basins snowmelt can be a large contributor to runoff, therefore better estimates of snowmelt, in the form of snow water equivalent, can lead to better estimation of streamflow (Moradkhani, 2008). These observations are available through several data products which can provide informative variables at near real-time frequency, can be used for data assimilation, and are of interest to hydrologists.

Possible data products include the European Space Agency's (ESA) Soil Moisture and Ocean Salinity (SMOS) data (Rodríguez-Fernández et al., 2017), the National Aeronautics and Space Administration's (NASA) Soil Moisture Active Passive (SMAP) data, and the National Operational Hydrologic Remote Sensing Center's (NOHRSC) Snow Data Assimilation System (SNODAS) data to name a few (Entekhabi et al., 2008; Kerr et al., 2010; National Operations Hydrologic Remote Sensing Center, 2004). The NOHRSC has provided daily gridded estimates of snow parameters such as SWE through the SNODAS program since 2004 (National Operations Hydrologic Remote Sensing Center, 2004). This paper will explore the use of the SMOS L2 soil moisture and the SNODAS snow water equivalent data products for assimilation into conceptual rainfall-runoff models to determine if they can be used to improve hydrologic modeling in an urban basin. Data assimilation will be used to help merge these datasets with the hydrologic models while also accounting for the uncertainty in both the models and the data products (Liu et al., 2012; Reichle, 2008).

The Ensemble Kalman Filter (EnKF) proposed by Evensen (1994) was used for data assimilation in these analyses. There are several examples in which the EnKF has been used for hydrologic modeling and forecasting (Abaza et al., 2014; Alvarez-Garreton et al., 2015; Clark et al., 2008; Crow and Ryu, 2009; Dumedah and Coulibaly, 2013; Komma et al., 2008; Massari et al., 2015; Moradkhani et al., 2005; Neal et al., 2007; Samuel et al., 2014; Thiboult and Anctil, 2015; Vrugt and Robinson, 2007; Wanders et al., 2014; Weerts and El Serafy, 2006). Moradkhani et al., (2005) proposed dual state parameter updating using the EnKF, and Samuel et al. (2014) found that using both streamflow and soil moisture observations together to update state and model parameters provided more accurate forecasts. Snow data assimilation, which includes assimilation of SWE, has also been examined in previous studies and has been shown to improve hydrologic simulations and forecasts (Bergeron et al., 2016; Dziubanski and Franz, 2016; Huang et al., 2017; Liu et al., 2012; Moradkhani, 2008). Building on these previous findings, this study will assess some combinations of streamflow, SWE, soil moisture in a dual state parameter updating scheme with different hydrologic models to enhance streamflow forecast in urban watershed.

2.3 Study area and data

The study area being focused on in this paper is the Don River basin (DRB) in Toronto, Ontario, Canada (Figure 2-1). The DRB is managed by the Toronto Region Conservation Authority (TRCA). It contains several sub-catchments, the largest of them being the Upper East Don, German Mills Creek, Lower East Don, Upper West Don, Lower West Don, Taylor-Massey Creek, and the Lower Don River. The DRB is approximately 350 km^2 and is a mostly urban watershed being roughly 80% developed, with the remaining area being split between crops, and pasture, forest, and wetland (Natural Resources Canada, 2009). This area has an average daily temperature of 9.4 °C (the average daily minimum and maximum temperatures are 5.9 °C to 12.9 °C respectively) and an average annual precipitation of 831.3 mm/year based on the 1981-2010 Canadian Climate Normals (Environment and Climate Change Canada, 2017). Major soils in the DRB include sandy loam, loam, clay loam, and clay (Ontario Ministry of Agriculture, 2015), and its elevation ranges from 75 to 330 meters above mean sea level (masl).

Daily precipitation, temperature, and snow depth data sets were obtained from Environment and Climate Change Canada (ECCC) weather stations, and evapotranspiration was estimated using the Penman-Monteith equation (Monteith, 1965).

Three observation data sets were used for data assimilation, they are daily streamflow data from Environment Canada's hydrometric database (HYDAT), daily SWE from SNODAS (National Operations Hydrologic Remote Sensing Center, 2004), and daily soil moisture (SM) from the ESA's SMOS satellite (Kerr et al., 2010).

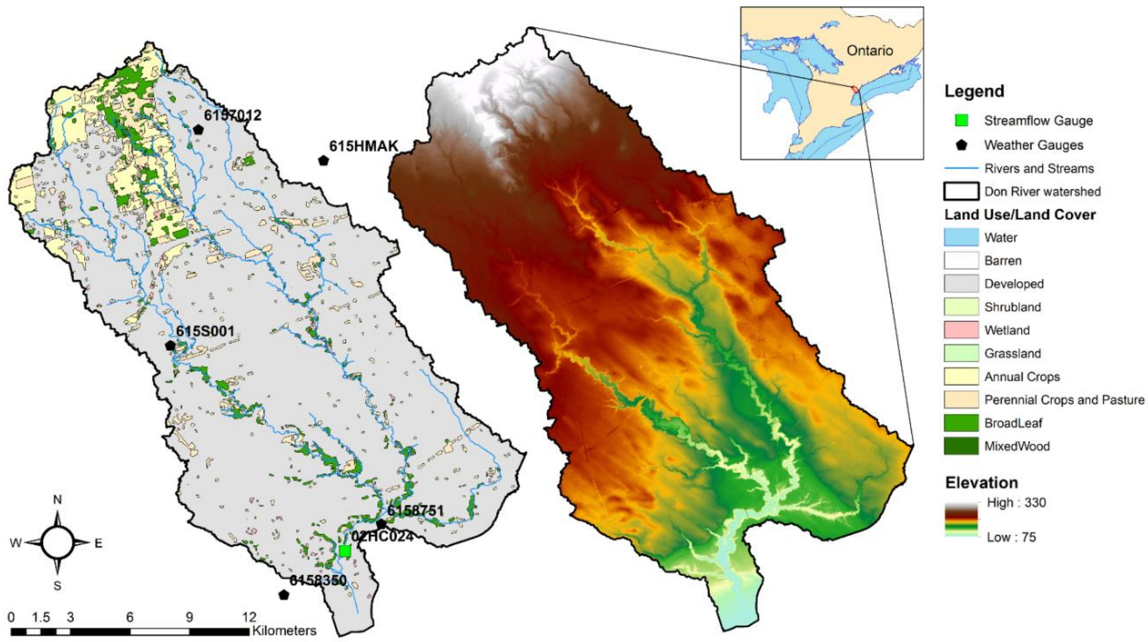

Fig. 2-1. Land use and land cover (left) (Natural Resources Canada, 2009), and topography (right) for the Don River basin in Ontario.

2.4 Methodology

2.4.1 Data processing

2.4.1.1 SNODAS Snow water equivalent data

One of this study's goals is to assess the assimilation of snow data as SWE into models of an urban basin. Several SNODAS data products are available including SWE and snow depth. The SNODAS SWE data product will be the source for the data being assimilated. The SNODAS products are developed as follows. First, data from the Rapid Refresh numerical weather prediction model (Rapid Update Cycle numerical weather prediction model for dates before May 1, 2012) are downscaled from 13 to 1 km^2. Next, these data are used to drive the NOHRSC Snow Model (NSM) at a resolution of 1 km^2. Finally, available remote sensing, radar, and ground station snow observations are assimilated into the model using a Newtonian nudging technique to produce a best estimate of near real-time snow conditions (Carroll et al., 2006; Clow et al., 2012). To determine the validity of using the SNODAS data, the ECCC snow on ground (snow depth) data available at gauges within and near the DRB was used to validate and bias correct the SNODAS snow depth and SWE data. This was done since SWE and snow depth are related and there are no SWE measurements available from ECCC.

A cumulative distribution function (CDF) matching bias correction method which uses polynomial fitting (Drusch et al., 2005; Kornelsen and Coulibaly, 2015) was used to correct the SNODAS snow depth such that it's CDF matched the ECCC snow depth. A consequence of this method is the ability to apply the polynomial to other data sets in order to implement a similar bias correction. Since no actual ECCC SWE data set is available in the study area, the bias correction used to correct SNODAS snow depth was applied to the SNODAS SWE data set to correct potential bias it may have. This assumes that, since snow depth and SWE are related, they would contain a similar relative bias when compared with the ECCC observations. For comparison only, a pseudo SWE data set was generated using a quick estimate from the ECCC snow depth values to compare with the bias corrected SNODAS SWE. This quick estimate was based on Environment Canada (2013) and Dubé (2003) where it is shown that using a 10:1 ratio, although not exact, for converting snow depth to SWE can provide a good estimate. To evaluate the correction method, the Bias, the Root Mean Square Error (RMSE), and the Nash-Sutcliffe Efficiency (NSE) were used. They are defined as follows:

$$Bias = \bar{Y} - \bar{X} \tag{2-1}$$

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i)^2\right)^{\frac{1}{2}} \tag{2-2}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \tag{2-3}$$

where X_i is the observed data (ECCC data) at time i, $\bar{X}$ is the mean of the observed data, Y_i is the simulated data (SNODAS data) at time i, $\bar{Y}$ is the mean of the simulated data, and N is the sample size. The RMSE was used to measure the fit between the observed and simulated data (lower values being preferable). The bias shows the consistent difference between the data and is better as the absolute bias approaches zero. NSE can be used to show how accurately the simulated and observed values match each other and ranges from -∞ to 1, with 1 being a perfect match.

2.4.1.2 SMOS L2 Soil moisture data

The SMOS satellite was launched in November 2009, it has a revisit time for both its ascending (6 am) and descending (6 pm) passes every three days at the equator (Kerr et al., 2012, 2010). Soil moisture data from the SMOS Soil Moisture Level 2 User Data Products, which also contains retrieved parameters such as optical thickness and surface temperature, was used for this study (Kerr et al., 2010). This dataset is retrieved using the SMOS satellite's L-Band Microwave radiometer which operates in the 1.4 GHz band range (Kerr et al., 2012). In the SMOS L2 data product, soil moisture is retrieved on a 15 km

Discrete Global Grid (DGG), however, the resolution of the instrument on the satellite is 43 km around each DGG center (Kerr et al., 2012, 2010). Therefore, there is overlap in the sensing area between DGGs. Within the DRB there are two SMOS DGG points, DGG206279 and DGG206792, that have retrieved soil moisture data available every three days. Because of the inherent overlap and to minimize uncertainty the mean of the two DRB DGGs was calculated as the basin average soil moisture when two retrievals were available. To use the soil moisture values at these points, the data was first filtered based on their probability of radio frequency interference (RFI) as well as their data quality index values. This filtering was done for both the ascending overpass, which is retrieved in the mornings, and descending overpass, which is retrieved during the evenings, soil moisture data. Retrievals with RFI probability and data quality index values greater than 0.1 were removed (Kornelsen et al., 2016). Assimilation of soil moisture only occurred during the non-freezing months of May to October to avoid the impact of retrievals of frozen soils. Figure 2-2 illustrates the similarity of the SMOS soil moisture data from each DGG point within the DRB. Using the two sample Kolmogorov-Smirnov test, the distribution of the ascending and descending retrievals at each DGG were compared and shown to be from the same distribution ($\alpha = 0.01$). From this comparison it was determined that a combination of the data from each DGG could be used to obtain one ascending and one descending soil moisture data set for the DRB which could both be used for assimilation.

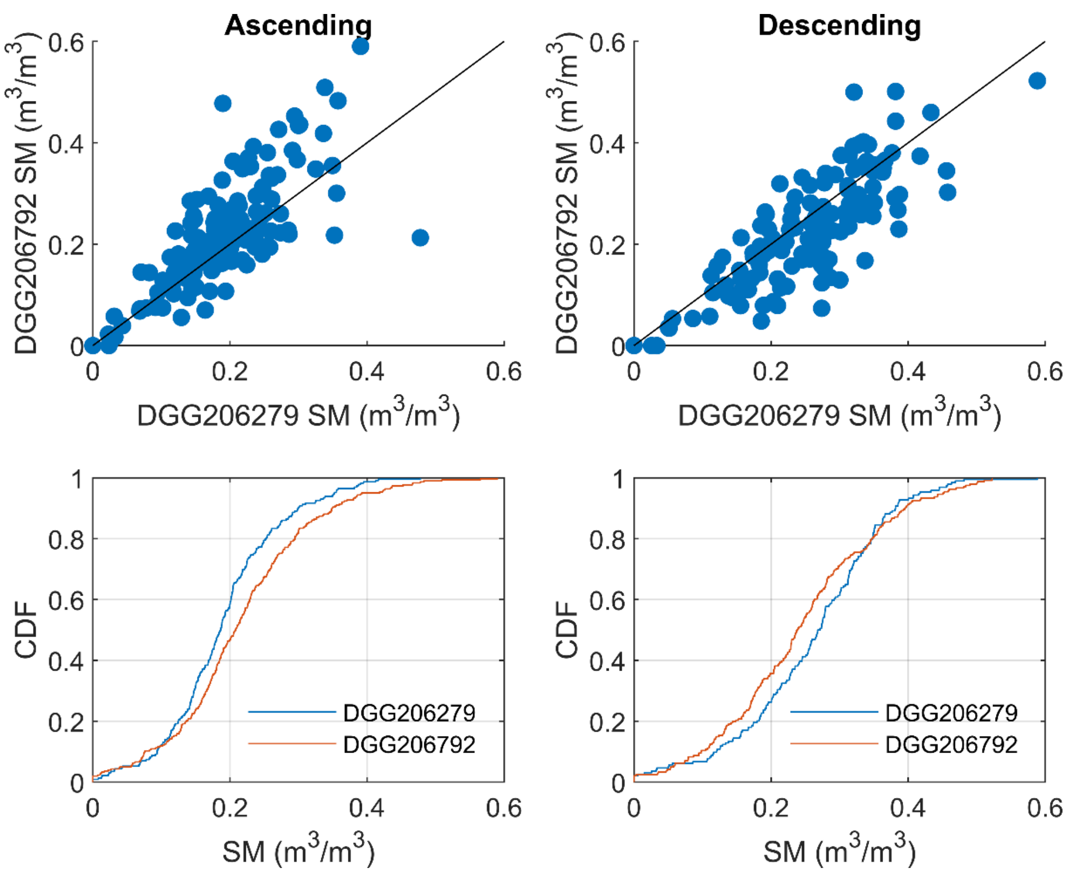

Fig. 2-2. Comparison of SMOS L2 soil moisture at each DGG point within the DRB for 2011-01-01 to 2015-12-31 data for both the ascending and descending passes.

The significance of the role of soil moisture in rainfall runoff modeling can often be under-valued in urban areas due to the influence of impervious surfaces. This presumption may also hold for the DRB as a highly-urbanized basin (Figure 2-1). To explore the validity of this assumption and justify the use of assimilation of SMOS soil moisture, spectral unmixing was used to determine the pervious areas of the watershed.

2.4.1.3 Spectral unmixing

Spectral unmixing is a method that can determine the contents of an image pixel if that pixel contains more than one material (Nascimento and Dias, 2005). For this research, it will be used to identify the pervious land cover in the DRB as part of the justification for assimilating soil moisture data in a highly developed urban basin (Figure 2-1). Using Google's Earth Engine (Google Earth Engine Team, 2015), a mosaic of Landsat 8 surface reflectance images over the DRB was created to help filter out cloud cover and other atmospheric interferences, the mosaic was then spectrally unmixed to find the relative proportion of pervious versus impervious regions. Each pixel is presumed to contain some mixed proportion of relevant land cover types (pervious, impervious and water). By comparing the spectral characteristics of the mixed pixel to within scene pure pixels, referred to as endmembers, the relative proportion of each land cover type with a Landsat pixel can be determined (Small, 2002, 2001). The vertex component analysis (VCA) method was chosen to identify spectral endmembers since it is quick and performs as well as manually determining the endmembers (Nascimento and Dias, 2005). VCA exploits the fact that the endmembers used to identify substances are the vertices of a simplex, and it is based on the assumption that pure pixels exist in the data. The VCA algorithm iteratively finds a preset number of purest endmembers which correspond to the most abundant land covers. In this case, VCA was used to determine three endmembers of the Landsat 8 mosaic using bands 1 - 7 (Ultra Blue, Blue, Green, Red, Near Infrared, Shortwave Infrared 1, Shortwave Infrared 2) which were then spectrally unmixed into a 3-band image that could be used to determine the land cover fraction of each pixel (water, vegetation, urban).

2.4.2 Hydrologic models

Four lumped hydrologic models were used in this study to model the DRB. They are the GR4J (modèle du Génie Rural à 4 paramètres Journalier), HYMOD (HYdrologic MODel), MAC-HBV (McMaster University - Hydrologiska Byråns

Vattenbalansavdelning), and SAC-SMA (Sacramento Soil Moisture Accounting) models. These models were calibrated using the particle swarm optimization (PSO) (Kennedy and Spears, 1998), which has been used to calibrate rainfall-runoff models previously (Chau, 2006; Gill et al., 2006; Li et al., 2009), to determine optimal parameter sets for the open loop simulations used for comparison with the data assimilation results. Each model was calibrated using data from 2001-01-01 to 2010-12-31 and validated against 2011-01-01 to 2013-12-31 values. Using the PSO, the optimal parameter sets were found by maximizing the Nash-Volume Error (NVE) performance metric from Samuel et al., (2012):

$$NVE = 0.5NSE - 0.1|VE| + 0.25NSE_{log} + 0.25NSE_{sqr} \tag{2-4}$$

where NSE is the Nash-Sutcliffe efficiency, VE is the volume error ($VE = Bias/\bar{X}$), NSE_{log} is the NSE calculated using the log streamflow values (for low flows), and NSE_{sqr} is the NSE found using the squared streamflow values (for high flows).

The GR4J is an empirical hydrologic model that has four parameters and runs on a daily scale (Perrin et al., 2003). In this study, GR4J was modified to include the degree day snow routine described in Samuel et al. (2011) so that it would contain a SWE state which could be updated during assimilation; the modified model will henceforth be referred to as GR4J-SR. Inputs for the GR4J-SR model include precipitation, evapotranspiration, temperature, and the parameter set optimized by the PSO. The model parameters and states are shown in Table 2-1.

Table 2-1. Description of GR4J-SR model parameters and state variables updated through data assimilation.

Parameter	Description	Range	Units	Calibrated Value
x1	Capacity of the production soil store	1 – 1500	mm	1499.997
x2	Water exchange coefficient	-10 – 5	mm	-0.319
x3	Capacity of the routing store	1 – 500	mm	18.507
x4	Time parameter for unit hydrographs	0.5 – 4	days	1.244
tr	Rainfall threshold temperature	0 – 2.5	°C	0.252
scf	Snow correction factor	0.4 – 1.6	-	1.052
ddf	Degree day factor	0 – 5	mm/day/°C	4.194
rcr	Rainfall correction factor	0.5 – 1.5	-	1.473
	State variable			
S	Production Store	-	mm	-
R	Routing store	-	mm	-
swe	Snow water equivalent	-	mm	-

The hydrologic model HYMOD, developed by Boyle (2001), is a simple conceptual rainfall-runoff model. As with the GR4J model, the HYMOD model used for these analyses was modified to include the snow routine from Samuel et al. (2011); the modified model will be referred to as HYMOD-SR. Inputs for the HYMOD-SR model include precipitation, evapotranspiration, temperature, and the parameter set optimized by the PSO. The model parameters and states are shown in Table 2-2.

Table 2-2. Description of HyMod-SR model parameters and state variables updated through data assimilation.

Parameter	Description	Range	Units	Calibrated Value
Alpha	Factor distributing the runoff between the quick and slow reservoirs	0.1 – 0.99	-	0.332
B_{exp}	Degree of spatial variability of soil moisture capacity	0.1 – 3	-	0.664
C_{max}	Maximum storage capacity	1 – 1000	mm	263.219
Rs	Residence time of slow flow reservoir	0.01 – 0.99	day	0.370
Rq	Residence times of quick flow reservoirs	0.001 – 0.1	day	0.009
tr	Rainfall threshold temperature	0 – 2.5	°C	2.126
scf	Snow correction factor	0.4 – 1.6	-	0.773
ddf	Degree day factor	0 – 5	mm/day/°C	3.634
rcr	Rainfall correction factor	0.5 – 1.5	-	1.212
	State variable			
S	Watershed storage	-	mm	-
qfr1	Quick flow reservoir 1	-	mm	-
qfr2	Quick flow reservoir 2	-	mm	-
qfr3	Quick flow reservoir 3	-	mm	-
sfr	Slow flow reservoir	-	mm	-
swe	Snow water equivalent	-	mm	-

The MAC-HBV model is a lumped conceptual rainfall-runoff model developed by Samuel et al. (Samuel et al., 2012, 2011), based on the HBV model (Bergsröm, 1976), for estimating streamflow in ungauged Ontario basins. The MAC-HBV incorporates a nonlinear response function, a routing routine, a degree day snow routine that is used to determine the SWE from the forcing data, a soil moisture routine used to show the change in soil moisture storage in the catchment, and uses a nonlinear storage-discharge relationship in the soil layers (Samuel et al., 2012, 2011). Both the SWE and soil moisture states from these routines will be updated using the chosen data assimilation method. Inputs into the MAC-HBV model include precipitation, temperature, and the calibrated parameter set from the PSO. The parameters and state variables used are summarized in Table 2-3.

The SAC-SMA is a lumped conceptual rainfall-runoff model with several parameters and model states (Table 2-4). There are five storages within the model used to represent the water accumulation in the catchment (Burnash, 1995; Burnash et al., 1973; Koren et al., 2004). The upper soil layer consists of both upper zone water storage contents, while the lower soil layer consists of the three lower zone water storage contents. The sum of the upper and lower soil layers was considered as the soil moisture for use in data assimilation (Samuel et al., 2014). The degree day snow routine within the model is used to determine the SWE from the forcing data (Samuel et al., 2014). The SAC-SMA model has been applied in several studies and is extensively used for operational streamflow forecasting (Samuel et al., 2014; Vrugt et al., 2006b, 2006a; Vrugt and Robinson, 2007).

Table 2-3. Description of MAC-HBV model parameters and state variables updated through data assimilation.

Parameter	Description	Range	Units	Calibrated Value
tr	Rain threshold	0 – 2.5	°C	1.194
scf	Snow correction factor	0.4 – 1.6	-	0.677
ddf	Degree-day factor	0 – 5	mm/day/°C	3.914
athorn	Coefficient of a simplified version of Thornthwaite formula to calculate potential evapotranspiration	0.1 – 0.3	-	0.160
fc	Maximum soil moisture storage	50 – 800	mm	69.740
flp	Fraction of maximum soil moisture value above which actual evapotranspiration reaches evapotranspiration potential	0.1 – 0.9	-	0.693
beta	Nonlinear function parameter represents the relative contribution to runoff from rain or snowmelt	0 – 10	-	0.792
k0	Storage coefficient 0	1 – 30	day	2.889
lsuz	Threshold value	1 – 100	mm	1.000
k1	Storage coefficient 1	30 – 100	day	66.079
cperc	Constant percolation parameter	0.01 – 6	mm/day	2.261
k2	Storage coefficient 2	100 – 500	day	263.227
maxbas	Runoff distribution parameter	1 – 20	day	1.597
rcr	Rainfall correction factor	0.5 – 1.5	-	1.092
alpha	Non-linearity coefficient	0.5 – 1.25	-	1.099
State variable				
swe	Snow water equivalent	-	mm	-
ssm	Soil moisture storage	-	mm	-
suz	Upper zone storage	-	mm	-
slz	Lower zone storage	-	mm	-
qg	Routed runoff	-	mm/day	-

2.4.3 Data Assimilation using the Ensemble Kalman Filter

The EnKF was used for both state and dual state parameter estimating (Moradkhani et al., 2005; Samuel et al., 2014). For this study, eight data assimilation schemes were used for each hydrologic model (Table 2-5). State updating was performed using streamflow, soil moisture, snow water equivalent, and combined soil moisture-snow water equivalent observations. Dual state-parameter updating was also performed using streamflow observations to update the parameter values in each case. Samuel et al., (2014) previously showed that using streamflow to update the parameter values and soil moisture to update state values allowed for more accurate estimates of streamflow and soil moisture since the combination allows for the model to better adjust over time.

The SM and SWE values were used to update their related states and parameters while the streamflow observations were used to update all (or remaining) states and parameters, this is illustrated in Figure 2-3. The related states and parameters updated for each model using SNODAS SWE, based on notation from Tables 2-1 to 2-4, are *swe*, *tr*, *scf*, *ddf*, and *rcr*. The related states and parameters updated using SMOS SM, based on notation from Tables 2-1 to 2-4, are *S* and *x1* (GR4J-SR); *S*, B_{exp}, and C_{max} (HyMod-SR); *ssm*, *fc*, and *flp* (MAC-HBV); and *uztwc*, *uzfwc*, *lztwc*, *lzfpc*, *lzfsc*, *uztwm*, *uzfwm*, *lztwm*, *lzfpm*, and *lzfsm* (SAC-SMA).

For the assimilation schemes using SNODAS SWE, the updates to state values were performed for the months of November – April, if data was available. When updating using

SMOS soil moisture the ascending pass data was prioritized over the descending pass such that the descending retrieval was

Table 2-4. Description of SAC-SMA model parameters and state variables updated through data assimilation.

Parameter	Description	Range	Units	Calibrated Value
uztwm	Upper-zone tension water maximum storage	1 - 150	mm	66.858
uzfwm	Upper-zone free water maximum storage	1 - 150	mm	69.711
uzk	Upper-zone free water lateral depletion rate	0.1 - 0.5	1/day	0.356
pctim	Impervious fraction of the watershed area	0 - 0.9	-	0.160
adimp	Additional impervious area	0 - 0.4	-	0.119
zperc	Maximum percolation rate	1 - 250	-	73.071
rexp	Exponent of the percolation equation	1 - 5	-	3.659
lztwm	Lower-zone tension water maximum storage	1 - 500	mm	406.084
lzfsm	Lower-zone free water supplemental maximum storage	1 - 1000	mm	66.081
lzfpm	Lower-zone free water primary maximum storage	1 - 1000	mm	654.933
lzsk	Lower-zone supplemental free water depletion rate	0.01 - 0.25	1/day	0.178
lzpk	Lower-zone primary free water depletion rate	0.0001 - 0.025	1/day	0.004
pfree	Fraction percolating from upper to lower-zone free water storage	0 - 0.6	-	0.596
rq	Residence time parameters of quick flow	0 - 0.99		0.758
ddf	Degree day factor	0 - 5	mm/day/°C	3.932
scf	Snowfall correction factor	0.4 - 1.6	-	0.843
tr	Upper threshold temperature, to distinguish between rainfall, snowfall and a mix of rain and snow	0 - 2.5	°C	1.164
athorn	A constant for thornthwaite's equation	0.1 - 0.3	-	0.199
rcr	Rainfall correction factor	0.5 - 1.5	-	1.175
	State variable			
uhg1	Linear reservoir to route upper-zone channel inflow 1	-	mm	-
uhg2	Linear reservoir to route upper-zone channel inflow 2	-	mm	-
uhg3	Linear reservoir to route upper-zone channel inflow 3	-	mm	-
swe	Snow water equivalent	-	mm	-
uztwc	Upper-zone tension water storage content	-	mm	-
uzfwc	Upper-zone free water storage content	-	mm	-
lztwc	Lower-zone tension water storage content	-	mm	-
lzfpc	Lower-zone free primary water storage content	-	mm	-
lzfsc	Lower-zone free secondary water storage content	-	mm	-
adimc	Additional impervious area content linked to stream network	-	mm	-

assimilated only if there was no soil moisture data available from the ascending retrieval. Since the models are run at a daily timestep the ascending overpass (6 am retrieval) is assumed to provide a better representation of the antecedent soil moisture conditions of the day, additionally the ascending has been shown to perform better than the descending pass (Jackson et al., 2012; Kornelsen et al., 2016). Since data assimilation is used to help remove random error in the model outputs, the SMOS soil moisture data was bias corrected using CDF matching to each of the models' soil moisture states (ascending and descending separately). CDF matching was chosen based on the results of Kornelsen and Coulibaly (2015). This adjustment was made so that the observations better fit the model to help avoid systematic errors, since unbiased errors are an important assumption for data assimilation methods (Reichle, 2008).

The EnKF, developed by Evensen (2003, 1994), uses a Monte Carlo approach to estimate the posterior distributions of the model output by using an ensemble estimate of their priors. Due to uncertainty related to initial conditions, especially when using data assimilation for both parameter and state estimating, several simulations can be run to determine the performance of the data assimilation method. The optimal ensemble size was determined for each model based on this fact (See Section 3.3.1).

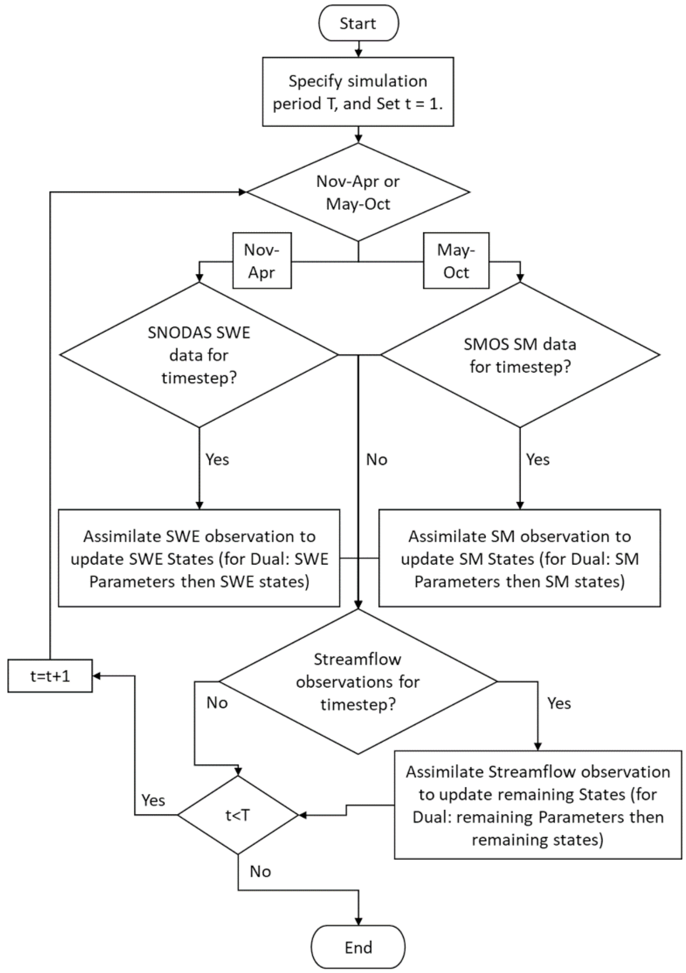

Fig. 2-3. Flowchart illustrating updating procedure used for data assimilation schemes.

The EnKF was formulated as follows (Moradkhani et al., 2005; Samuel et al., 2014):

$$u_t^i = u_t + \zeta_t^i \tag{2-5}$$

where u_t^i is the perturbed forcing data at time *t* for ensemble member *i*, u_t is the unperturbed forcing data, and ζ_t^i is the noise added to the forcing data to generate *i* ensemble members. The precipitation was perturbed using lognormally distributed noise, $\zeta_t^i{\sim}logN(0,\Sigma_t^u)$, whereas the temperature and evaporation was perturbed using normally distributed noise, $\zeta_t^i{\sim}N(0,\Sigma_t^u)$. The variance of the noise, Σ_t^u, is used to influence the ensemble spread to better account for uncertainty, it is calculated by taking the square of the product of the forcing data u_t and a proportionality factor γ. This proportionality term is considered a hyper-parameter which can be adjusted to improve results and ensure adequate ensemble spread (Moradkhani et al., 2005; Thiboult and Anctil, 2015).

The state variables being updated in each model are, in general, calculated as follows:

$$\mathrm{x}_{t+1}^{i-} = f\left(x_t^{i+}, u_t^i, \theta\right) \tag{2-6}$$

where x_{t+1}^{i-} is the non-updated vector of state variables in the model at time *t+1* for ensemble member *i*, *f(.)* is the operator within the models that propagate the state variables, x_t^{i+} is the updated vector of state variables at time *t* for ensemble *i*, and θ is the vector of parameters (θ_t^{i+} for dual state parameter updating). For the dual state parameter updating assimilation scheme, the parameter set θ_t^{i+} at *t = 1* is generated from a uniform distribution for each ensemble member and each parameter based on the ranges in Tables 2-1 to 2-4.

Table 2-5. Description and naming convention of each data assimilation scheme used in this study.

DA Scheme Number	Description of Assimilation scheme
DAS1	State updating using streamflow observations
DAS2	State updating using SMOS L2 soil moisture data
DAS3	State updating using SNODAS SWE data
DAS4	State updating using SMOS L2 soil moisture and SNODAS SWE data sets
DAS5	Dual state and parameter updating using streamflow observations.
DAS6	Dual state and parameter updating in which SMOS L2 soil moisture dataset was used to update related parameter and state variables when available, otherwise streamflow is used.
DAS7	Dual state and parameter updating in which SNODAS SWE dataset was used to update related parameter and state variables when available, otherwise streamflow is used.
DAS8	Dual state and parameter updating in which SMOS L2 soil moisture data and SNODAS SWE are used to update related parameter and state variables when available, otherwise streamflow is used.

The hydrologic model outputs are then calculated as follows:

$$\hat{y}_{t+1}^{i} = h\left(x_{t+1}^{i-}, \theta\right) + \nu_{t+1}^{i}, \nu_{t+1}^{\mathrm{i}} \sim N(0, \Sigma_{\mathrm{t+1}}^{\mathrm{m}}) \tag{2-7}$$

where $\hat{y}_{t+1}^{i}$ is the simulated model output at time *t+1* for ensemble member *i*, *h(.)* is the hydrologic model, ν_{t+1}^{i} is normally distributed noise at time *t* for ensemble member *i*, and Σ_{t+1}^{m} is the variance of the noise found by taking the square of the product of the model output and its hyper-parameter ω. This noise is used to represent uncertainty related to the model structure.

To perform the update, an ensemble of observations is generated at each time step by perturbing measured data as follows:

$$y_{t+1}^{i} = y_{t+1} + \eta_{t+1}^{i}, \eta_{t+1}^{i} \sim N\left(0, \Sigma_{t+1}^{y}\right) \tag{2-8}$$

where y_{t+1}^{i} is the perturbed observation data used for updating the state (or parameter) vector for *i* ensemble members at time *t+1*, y_{t+1} is the observation at time *t+1*, η_{t+1}^{i} is the normally distributed noise with variance Σ_{t+1}^{y} used for perturbing the observations into *i* ensemble members, and the variance of the noise is found by taking the square of the product of the observation data with the hyper-parameter ρ.

The Kalman gain used in the EnKF is calculated as follows:

$$K_{t+1} = \Sigma_{t+1}^{xy}\left(\Sigma_{t+1}^{yy} + \Sigma_{t+1}^{y}\right)^{-1} \tag{2-9}$$

where K_{t+1} is the Kalman gain at time *t+1*, Σ_{t+1}^{xy} is the cross-covariance of the state variable ensembles with the prediction ensemble (streamflow, soil moisture, or snow water equivalent), and Σ_{t+1}^{yy} is the error covariance of matrix of the streamflow (or soil moisture or snow water equivalent) prediction ensemble.

After calculating equations 2-5 to 2-9, the updated state (or parameter) vector is then found by:

$$x_{t+1}^{i+} = x_{t+1}^{i-} + K_{t+1}\left(y_{t+1}^{i} - \hat{y}_{t+1}^{i}\right) \tag{2-10}$$

where x_{t+1}^{i+} is the updated state (or parameter) vector for time *t+1* and ensemble member *i*. After the states (or parameters) are updated, realism is checked to ensure values are not out of their allowed ranges. This is repeated every time step until the simulation has been completed.

2.4.3.1 Determining hyper-parameter values and optimal ensemble sizes

The hyper-parameters are used to control the ensemble spread caused by perturbation of forcing data (γ), observations (ρ), and simulated model results (ω) (Moradkhani et al., 2005; Thiboult and Anctil, 2015). The Normalized RMSE Ratio (NRR) was used to help determine the optimal ensemble size and hyper-parameter values such that they would not cause too much or too little spread (Murphy, 1988), and has been used in previous studies (Alvarez-Garreton et al., 2014; Moradkhani et al., 2005; Thiboult and Anctil, 2015). The NRR is defined as follows:

$$NRR = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\left[\frac{1}{N}\sum_{i=1}^{N}\hat{y}_t^i\right] - y_t\right)^2}}{\frac{1}{N}\left\{\sum_{i=1}^{N}\sqrt{\frac{1}{T}\left[\sum_{t=1}^{T}\left(\hat{y}_t^i - y_t\right)^2\right]}\right\}\sqrt{\frac{N+1}{2N}}} \tag{2-11}$$

where T is the length of the time series, N is the number of ensemble members y_t is the observation at time *t*, and $\hat{y}_t^i$ is the simulated value at time *t* for ensemble *i*. The ideal value for NRR is 1, while if NRR < 1 or NRR > 1 there is too much or too little spread respectively.

Using the DAS5 assimilation scheme, hyper-parameter values for the forcing data, streamflow observations, and models were generated from the uniform distribution U(0,0.5) and tested against ensemble sizes ranging from 25 to 500. The combinations which provided the best NSE value along with an NRR value closest to 1 were chosen for each model. Additionally, the hyper-parameter used to perturb the SMOS L2 soil moisture and the SNODAS SWE were determined based on uncertainty analysis performed in previous studies (Al Bitar et al., 2012; Clow et al., 2012; Kerr et al., 2012; Kornelsen and Coulibaly, 2015; Zhang and Yang, 2016). Finally, a pairwise comparison of the ensemble means was used to confirm that at the chosen ensemble size, based on NRR, the mean NSE was not significantly different from subsequent increases in ensemble size.

2.4.4 Evaluation of assimilation schemes

To determine the performance of each data assimilation scheme and model combination for the Don River basin, the ensemble means of the hydrologic assimilation experiments were evaluated using the Kling-Gupta efficiency (KGE) performance metric (Gupta et al., 2009):

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

where r is the linear correlation coefficient between the simulated and observed runoff values, $\alpha = \sigma_s/\sigma_o$ is the measure of relative variability in the simulated and observed values, and $\beta = \mu_s/\mu_o$ is the ratio between the simulated and observed values and is used to represent bias (Gupta et al., 2009). The KGE has a range from $-\infty$ to 1, with 1 being the optimal value. This metric uses the same components as the NSE; when these components are near their optimal values the KGE will be near its optimal value, which is not always the case for the NSE (Gupta et al., 2009).

Additionally, the ensemble forecast performance of each assimilation scheme and model combination was also examined. In this case, the mean Continuous Ranked Probability Score (CRPS) performance metric, which assesses accuracy and resolution, was used to evaluate the one to fourteen-day ensemble forecast performances. The CPRS is formulated as (Matheson and Winkler, 1976; Unger, 1985):

$$CRPS(F, x) = \int_{-\infty}^{\infty} (F(y) - \mathbf{1}\{y \geq x\})^2 dy \qquad (2\text{-}13)$$

where *F(y)* is the cumulative distribution function of the forecast distribution (ensemble forecasts), *y* is the predicted variable (simulated runoff), *x* is used to verify the distribution (observed runoff), and $\mathbf{1}\{y \geq x\}$ is the Heaviside step function that provides a value of 1 if the predicted value is larger than the observed and 0 otherwise. A perfect forecast is indicated with a CPRS value of 0, there is no upper limit to the value.

Using these performance metrics the potential improvements of each assimilation scheme will be easily quantified which will help identify the best performing assimilation scheme, hydrologic model, and combination of the two for performing hydrologic simulations and forecasts in an urban basin.

2.5 Results and Discussions

2.5.1 Data processing results

2.5.1.1 SNODAS

The raw and bias corrected SNODAS data sets were compared with ECCC data at various weather gauges as well as a Thiessen polygon weighted basin average value for the period of 2011-01-01 to 2015-12-31. Over this period, the SNODAS data showed systematically higher values than the observations, shown in Tables 2-6 and 2-7, which is likely attributed to SNODAS not being ideally calibrated for small urban basins coupled with forcing data issues due to downscaling 13 km to 1 km grids. This bias can be seen in the mean annual maximum values for snow depths being 435 mm and 299 mm and the average snow cover days being 110 days and 88 days for SNODAS and ECCC data respectively.

Table 2-6. Bias, RMSE, and NSE of raw and bias corrected SNODAS Snow depth values at grid point nearest to ECCC weather station.

Station	Raw Snow Depth			Bias Corrected Snow Depth		
	Bias (mm)	RMSE (mm)	NSE	Bias (mm)	RMSE (mm)	NSE
615HMAK	24.14	69.27	0.17	0.24	40.52	0.72
6157012	27.99	77.52	0.39	0.46	49.97	0.75
6158350	45.15	88.25	-0.24	0.00	41.62	0.73
615S001	21.84	67.02	0.19	0.38	38.71	0.73
6158751	12.94	48.27	0.55	0.00	32.87	0.79
Basin Average	19.16	53.50	0.39	0.55	27.02	0.85

Table 2-7. Bias, RMSE, and NSE of raw and bias corrected SNODAS SWE values at grid point nearest to ECCC weather station.

Station	Raw SWE			Bias Corrected SWE		
	Bias (mm)	RMSE (mm)	NSE	Bias (mm)	RMSE (mm)	NSE
615HMAK	7.71	19.14	-5.30	-1.02	5.33	0.51
6157012	9.77	22.50	-4.16	0.46	6.01	0.63
6158350	14.51	27.06	-10.62	-1.57	6.19	0.39
615S001	7.10	18.62	-5.26	-0.37	4.94	0.56
6158751	5.71	15.88	-3.85	0.70	4.95	0.53
Basin Average	6.48	16.79	-5.02	-0.13	3.72	0.72

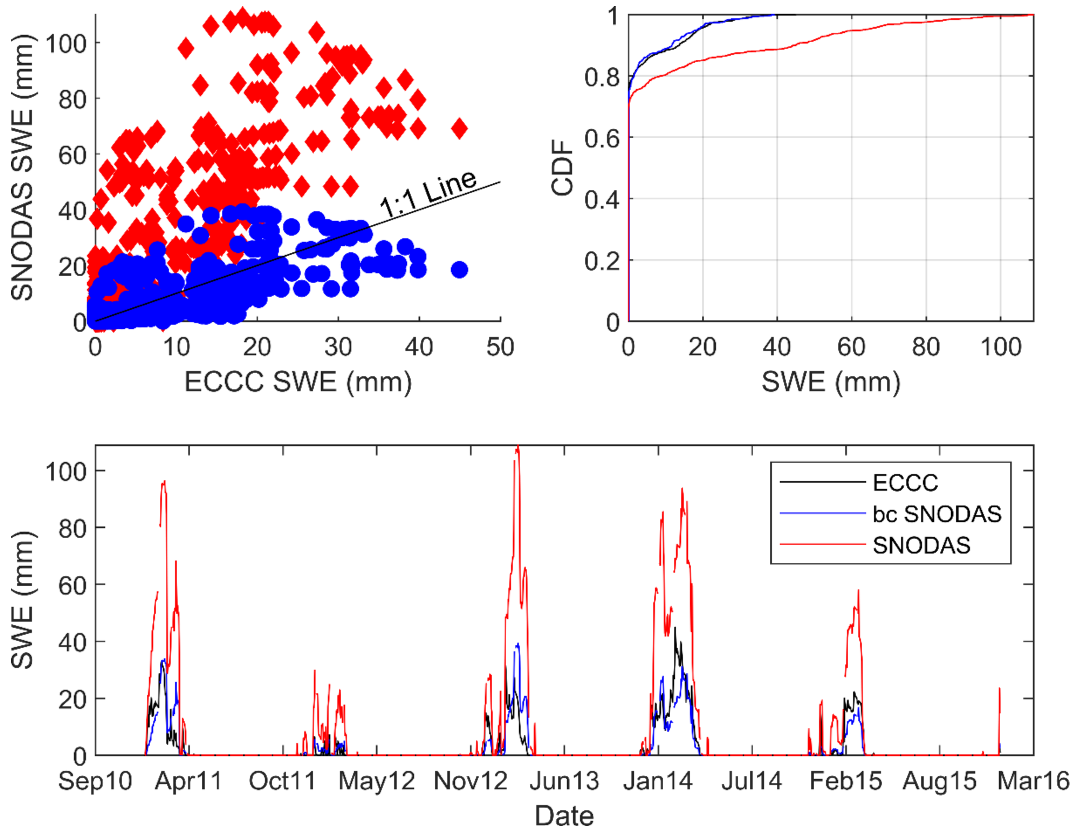

Fig. 2-4. Comparisons of raw (red) and bias corrected (blue) SNODAS basin average SWE data to ECCC SWE estimates for 2011-01-01 to 2015-12-31. The correlation between the SNODAS SWE and ECCC SWE estimate is illustrated (top left). The CDF plot shows how well the distributions are matched before and after bias correction (top right). A timeseries comparison of the ECCC SWE estimate with the raw and bias corrected SNODAS SWE (bottom).

The basin average estimate of SWE is illustrated in Figure 2-4, where there still exists a slight bias to the SNODAS SWE product. Despite this, the corrected SNODAS SWE values can be used to provide a good estimate for the basin average SWE.

2.5.1.2 VCA Spectral unmixing

The DRB is recognized as a highly developed urban basin (Figure 2-1), in which the assimilation of soil moisture would often be dismissed. To demonstrate the potential value of soil moisture assimilation in such a basin, a better estimate of pervious land cover was produced by spectrally unmixing a Landsat 8 surface reflectance image mosaic of the DRB. A temporal mosaic of images was used to better remove cloud cover and using data from the period 2013-06-01 to 2015-09-30. In the unmixed image, it was assumed that vegetation cover was equivalent to pervious areas. This is shown by vegetation fraction in Figure 2-5. Taking the averages of each pixel, the urban, vegetation, and water coverages are shown to be 54%, 44%, and 2% respectively. Based on these results there is a significant portion of the study area that can be considered pervious surface and it should be acceptable to use soil moisture assimilation in the DRB.

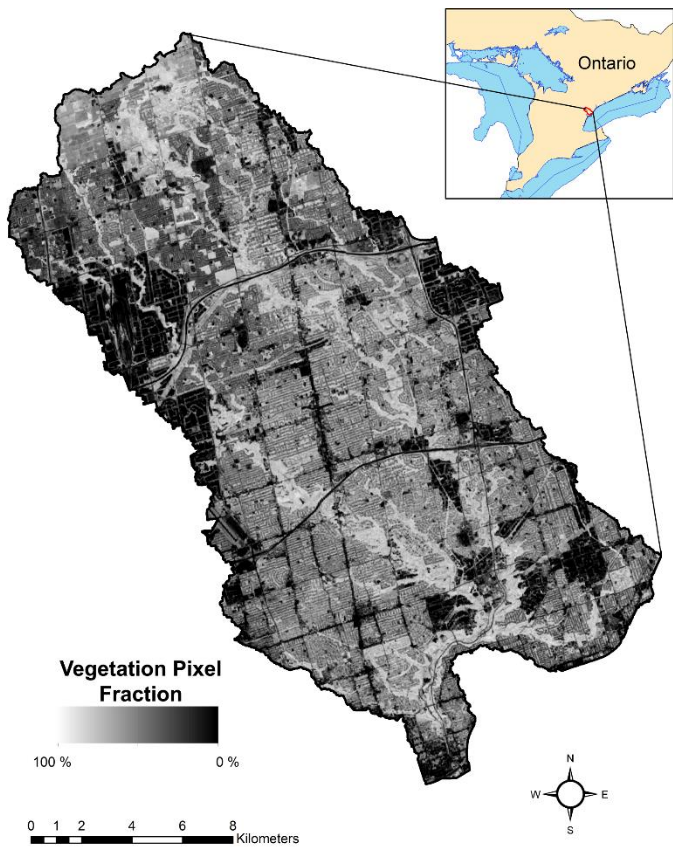

Fig. 2-5. 'Vegetation' band of spectrally unmixed Landsat 8 image mosaic illustrating the fraction of each pixel in the Don River basin containing pervious surface (vegetation).

2.5.2 Model calibration and validation

The resulting performance metrics for both the calibration (2001 – 2010) and validation (2011 – 2013) periods are summarized in Table 2-8. Of note, the performance during calibration varies between each model much more than during the validation period, in which the GR4J-SR performs noticeably better than the other three models. During the validation, however, the GR4J-SR, HyMod-SR, and SAC-SMA have relatively similar performances with the SAC-SMA being slightly better. In both cases, the MAC-HBV model does not perform as well. These results are likely due to the differing model structures. The GR4J-SR and HyMod-SR are both simple and easily adaptable models therefore applying and using them in an urban basin is fairly easy. The SAC-SMA model can handle impervious areas due to the wide variety of parameters within the model making it adaptable for use in an urban basin as well. However, the MAC-HBV model does not have any parameters specifically for urban areas within the basin, which may have led to its poorer performance. These validation results will be used as an open loop comparison to the data assimilation schemes being tested to help illustrate the improvements the assimilation has on the hydrologic simulations.

Table 2-8. Summary of the calibration and validation results for the GR4J-SR, HyMOD-SR, MAC-HBV, and SAC-SMA hydrologic models in the DRB.

	CALIBRATION (2001-2010)			VALIDATION (2011-2013)		
Hydrologic Model	NVE	KGE	NSE	NVE	KGE	NSE
GR4J-SR	0.66	0.79	0.68	0.58	0.78	0.61
HyMod-SR	0.57	0.74	0.64	0.58	0.71	0.60
MAC-HBV	0.43	0.63	0.40	0.31	0.56	0.32
SAC-SMA	0.54	0.75	0.60	0.60	0.72	0.63

2.5.3 Optimal hyper parameters and ensemble size for each hydrologic model

The ensemble size test results are illustrated in Figure 2-6. These results show that increasing the ensemble size increases the NSE performance of the models' ensemble mean and decreases the amount that the performance can vary. This test showed that for the HyMod-SR the optimal ensemble size was 200, for the MAC-HBV and SACSMA models the optimal ensemble size was 250, and the optimal ensemble size for GR4J-SR was 325. Additionally, Figure 2-6 shows that the MAC-HBV and SAC-SMA model performances have higher variability than the other two models, this is due to them having more parameters to update.

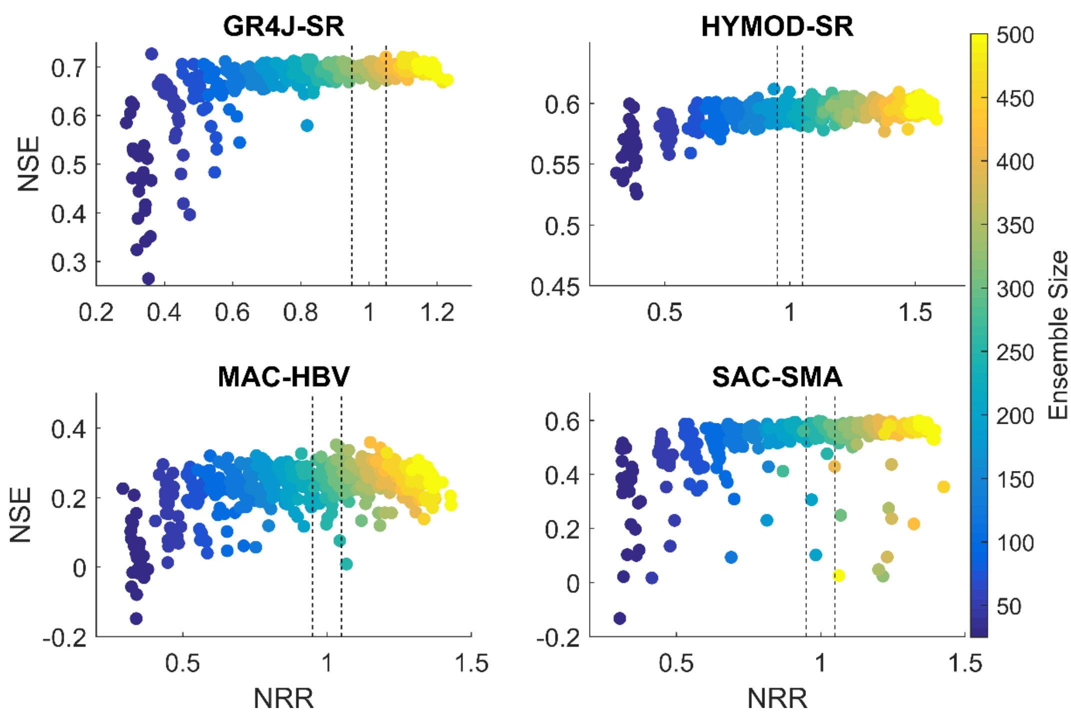

Fig. 2-6. NSE and NRR performance for each model at different ensemble sizes under the DAS5 scheme. For each ensemble size, the model was run 25 times. The dashed line represents the NRR considered acceptable for this study.

To ensure the best results from the assimilation schemes tuning of the hyper-parameters was done to better quantify the uncertainty in the observations and models

(Moradkhani et al., 2005; Thiboult and Anctil, 2015). Summarized in Table 2-9 are the hyper-parameter values used for each model.

Table 2-9. Hyper-parameter set used for data assimilation with the EnKF for each hydrologic model.

Model	γ_{Precip}	γ_{Temp}	γ_{ET}	ρ_{Q}	ρ_{SM}	ρ_{SWE}	ω
GR4J-SR	0.28	0.01	0.25	0.42	0.15	0.25	0.27
HyMod-SR	0.24	0.08	0.27	0.21	0.15	0.25	0.30
MAC-HBV	0.45	0.01	-	0.16	0.15	0.25	0.22
SAC-SMA	0.32	0.07	-	0.24	0.15	0.25	0.47

2.5.4 Hydrologic model performance using different data assimilation schemes

The main focus of this study is to determine the effects of assimilating soil moisture, snow water equivalent, and streamflow, under various data assimilation schemes, can have on hydrologic simulation and forecasting in an urban basin. To assess the effect assimilating these observations have on simulating the DRB two performance metrics were examined, the KGE and the NSE. The NSE being the traditional measure of hydrologic model performance, and the KGE which is relatively newer and uses the same components as NSE to partition performance into contributing components (Gupta et al., 2009). Due to the metrics similarities, the assimilation schemes being compared are considered acceptable if they perform better than the calibrated model for only one of the metrics.

From the results shown in Figure 2-7, it is apparent that models perform the best under the assimilation schemes DAS1-4 based on the improvements seen compared to the open loop simulation. When comparing the average relative improvements over the open loop simulation for both KGE and NSE, DAS2 performed the best. The GR4J-SR model had the largest improvement over the OL seen from DAS6 which had an increase in NSE from 0.6 to 0.69, with similar increase in performance seen from the other dual state parameter estimating schemes. This agrees with the results of Samuel et al. (2014), which stated that the combination helped to account for variations through time in the model parameters. HyMod-SR showed minor improvement in KGE for DAS3 and DAS4, MAC-HBV showed minor KGE improvement in DAS1-4, and SAC-SMA showed minor improvements in KGE for DAS2-8. These results indicate that the only model, of the four tested, to significantly benefit from data assimilation when simulating the runoff of an urban basin is GR4J-SR. Additionally, these results suggest that the improvement from state updates is smoothed out the longer the simulation which can be seen in the forecast results. Finally, the results show that data assimilation which updates both states and parameters has the potential to reduce the performance of a model (MAC-HBV DAS5-8), although this may be due to MAC-HBV lacking a parameter which explicitly considers urban areas.

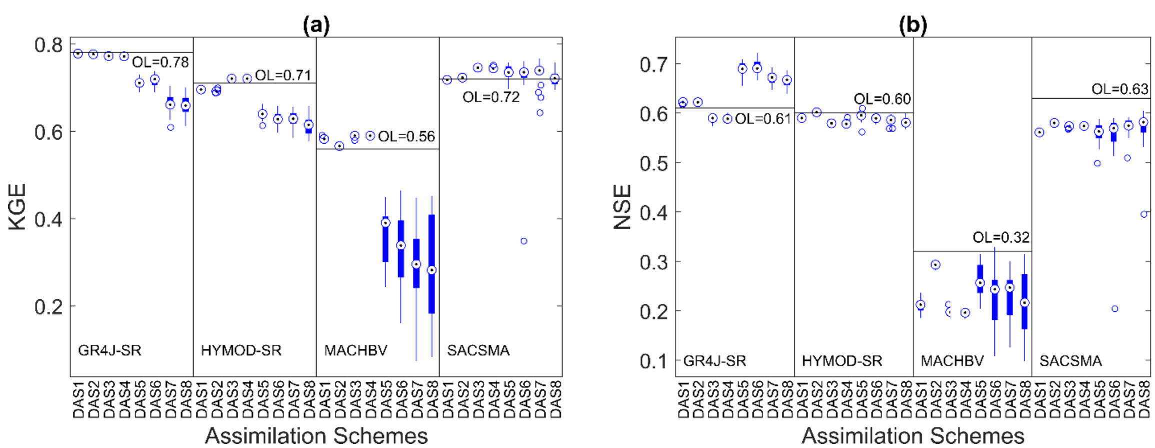

Fig. 2-7. Results of 20 simulations for each data assimilation scheme and model. This figure illustrates the spread of (a) KGE values and (b) NSE values in relation to the open loop (OL) runs for the 2011-01-01 to 2013-12-31 time period.

Illustrated in Figure 2-8 are the 2013-01-01 to 2013-12-31 simulated ensemble means for each hydrologic model and the four best performing data assimilation schemes.

Here it is apparent that the ensemble means do not always capture the peak flows despite the model's performances. However, when looking at the ensemble simulation results from DAS2 for example (Figure 2-9), the extreme ensemble values can capture the majority of those peak flows. Also apparent in Figure 2-9 is that MAC-HBV has some issues capturing low flows within the DRB, which is likely due to it being initially developed to simulate natural catchments, and its inability to account for impervious areas.

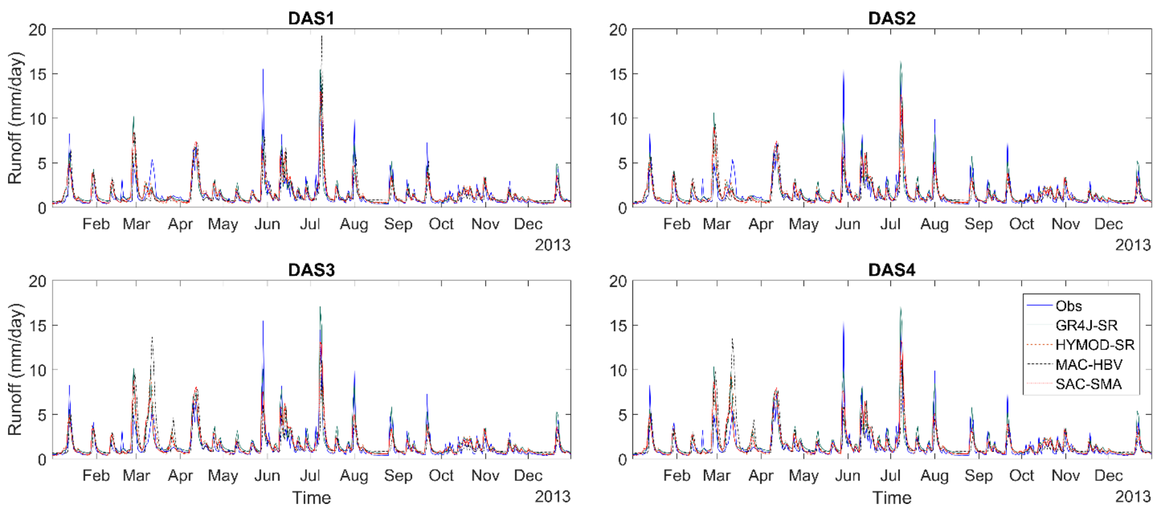

Fig. 2-8. Each model's simulation results, for January – December 2013, under the DAS1, DAS2, DAS3, and DAS4 assimilation schemes comparing ensemble means to observed streamflow.

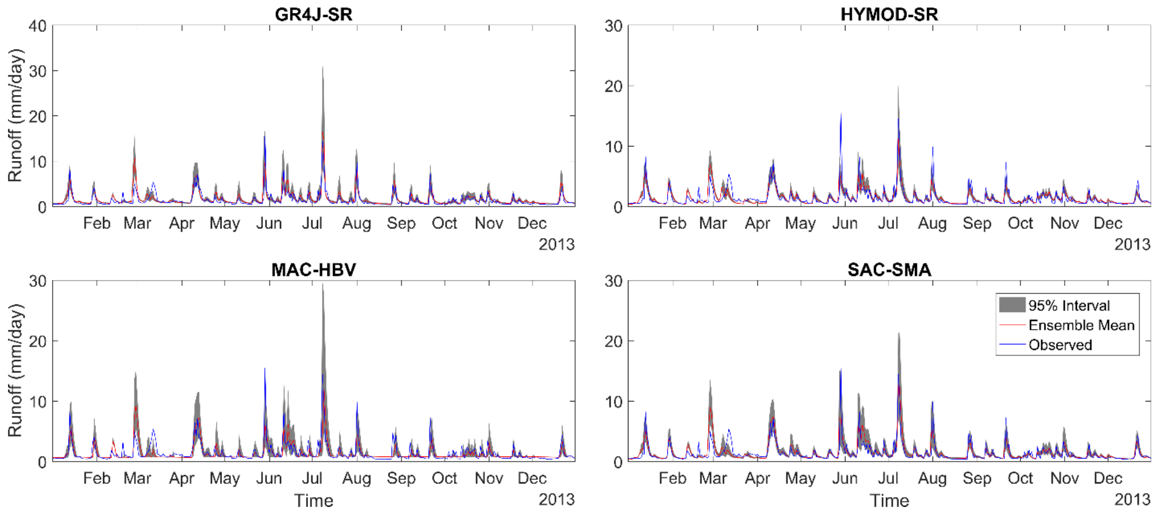

Fig. 2-9. The ensemble results, for January – December 2013, illustrating the 95% interval of each ensemble simulation for the DAS2 assimilation scheme.

2.5.5 Performance of ensemble forecasts

Ensemble forecasts were made using 'perfect' weather forecasts (meteorological observations) and the updated states and parameter values for each model and assimilation scheme. The perfect forecast was chosen to help reduce uncertainty in the forcing data. These forecasts were made for lead times of up to 14 days over the 2011-01-01 to 2013-12-31 period. To evaluate the performance of these ensemble forecasts the CRPS was used, the results for each model and assimilation scheme are shown in Figure 2-10. For the HyMod-SR, MAC-HBV, and SAC-SMA models, the best performances for each assimilation scheme are generally shown for the 1-day ahead forecast. The GR4J-SR, however, performed its best for the 2-day ahead forecast which is likely due to the model's structure, more specifically how it utilizes unit hydrographs for routing. As the forecast lead times increase the mean CRPS values begin to degrade and converge towards the OL values for each model and assimilation scheme, with the exception of DAS8 for SAC-SMA which shows consistent performance over the forecast horizon. The performance of the state updating assimilation schemes show as good or better performance for each model over all forecast horizons, which agrees with the literature in that more accurate states produce better forecast results (Moradkhani, 2008; Reichle, 2008). However, the performance of the dual state parameter assimilation schemes varies depending on the model, which is most apparent with the HyMod-SR and MAC-HBV forecast results. Where, for short term forecasts, updates to states and parameters (DAS5-8) provide the HyMod-SR model with better ensemble forecasting skill which degraded slower when compared to only state updating (DAS1-4), while for MAC-HBV DAS5-8 provides much worse performance which is also seen in Figure 2-10. These differences could be attributed to structural differences in the models such as the simple fast and slow storages in HyMod and GR4J which can easily adapt to urban basins as well as SAC-SMAs explicit consideration of urban areas, while MAC-HBV lacks the ability to account for the quicker runoff from more than half the basin.

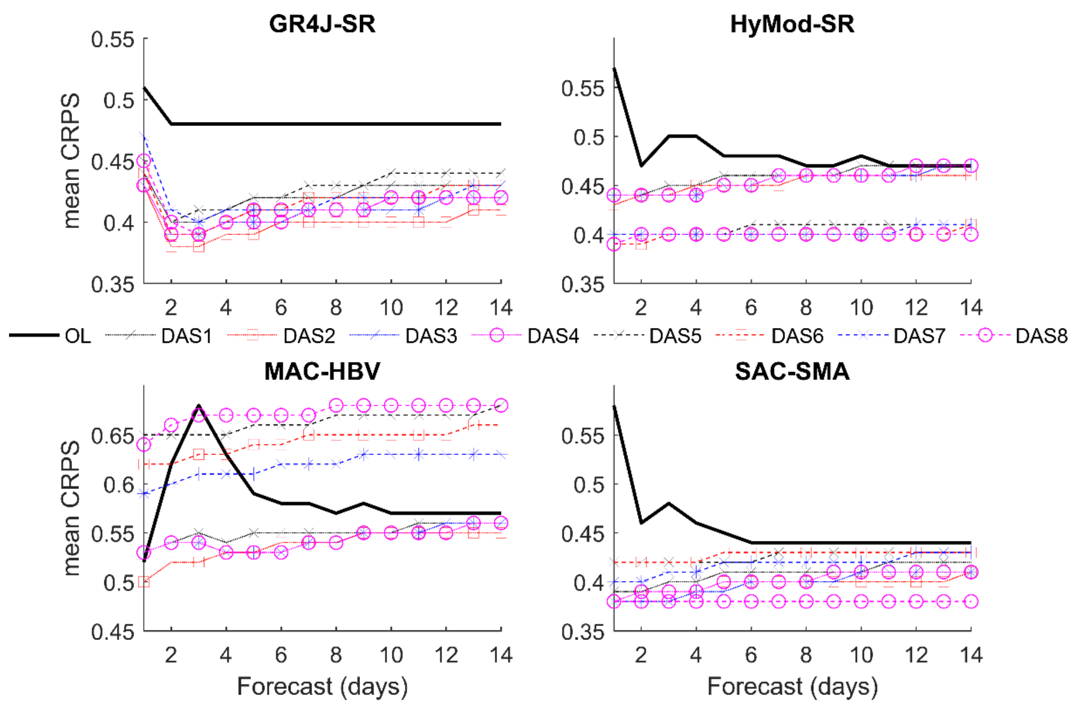

Fig. 2-10. Mean CRPS values for each assimilation scheme and hydrologic model for the 1 to 14 day running forecast.

Among the four models, the top performing assimilation schemes for ensemble forecasting were DAS2, DAS3, and DAS4 when compared to the OL, which suggests that assimilating the SMOS L2 SM and/or the SNODAS SWE data can improve the short-term forecasts for an urban model. Additionally, the best overall performance was seen from the SAC-SMA model under DAS8, however, the GR4J-SR, HyMod-SR, and SAC-SMA models had similar forecasting performances on average. The DAS2 assimilation scheme showed the most improvement in ensemble forecast performance over all forecast horizons compared to the OL. This is likely caused by the size and climate of the basin, the watershed response is quick, and the snow cover time is short, allowing for antecedent soil moisture estimation to be the most meaningful variable of the three for runoff prediction. When comparing the forecast performance from each assimilation scheme and model, by ranking them from lowest to highest CRPS, the best to worst performing for ensemble forecasting are DAS8, DAS2, DAS3, DAS7, DAS4, DAS6, DAS5, and DAS1. This shows that the forecasts benefit from the additional information provided from soil moisture and SWE.

Ultimately, these results show that assimilation of the SMOS L2 soil moisture and the SNODAS SWE data products can improve hydrologic modeling and forecasting in the Don River basin. This indicates that assimilating soil moisture and snow water equivalent could potentially improve hydrologic models of urban basins. Additionally, when moving forward into semi-distributed or distributed models, these gridded data products would likely be more useful as there will be less chance of losing information due to aggregation of data when computing the basin average.

2.6 Conclusions

This study examined the performance of four lumped conceptual rainfall-runoff model's ability to simulate the streamflow in the urban Don River Basin under various data assimilation schemes. As with previous studies, assimilation of streamflow can improve the urban models performance (e.g. Branisavljevic et al., 2014; Hutton et al., 2014, and others), additionally this study showed that further improvements to model performance could be made through assimilating soil moisture and/or SWE, although some models benefitted more than others. The results showed that the simpler GR4J-SR model had the most improved hydrologic simulations, based on NSE, from data assimilation under the dual state and parameter updating assimilation schemes, with the best performing being that of DAS6. For flow forecasting, SAC-SMA performed the best followed by HyMod-SR, however, the GR4J-SR model was comparable for 2-days forecasts or longer. DAS2-

4 performed the best for the 1-day forecast, while the DAS2, DAS3, and DAS8 provided the best overall performance when comparing all assimilation schemes for every forecast horizon.

In general, the results show that assimilating the SNODAS SWE and SMOS L2 SM data products can provide some improvement to different aspects of hydrologic simulation and forecasting, which is apparent through the use of multiple performance metrics, even in urban basins such as the DRB. However, further improvements could likely be made to the results by using a different snow models such as SNOW17 which has been used in previous studies when assimilating SWE data (DeChant and Moradkhani, 2011; Dziubanski and Franz, 2016; Huang et al., 2017), or a more comprehensive analysis for error quantification since the EnKF is known to perform better when the uncertainties associated with the model and observations are better quantified (Huang et al., 2017; Moradkhani et al., 2005). Additionally, potential improvements could be made through identifying and addressing problematic variables which could arise from updating with different types of observational data. Finally, the analysis performed here were for one basin on the daily time scale with lumped models; further analyses are needed to determine if these results are valid at finer spatial and temporal scales as well as other basins.

2.7 Acknowledgements

This research was supported by the Natural Science and Engineering Research Council (NSERC) of Canada, grant NSERC Canadian FloodNet (NETGP-451456). The authors are grateful to the European Space Agency for providing SMOS L2 data products, the National Operational Hydrologic Remote Sensing Center for providing SNODAS data products, and for data available from Environment and Climate Change Canada. Finally, the authors would like to thank the anonymous reviewers who helped improve the quality of this manuscript through their thoughtful comments and suggestions.

2.8 References

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

Al Bitar, A., Leroux, D., Kerr, Y.H., Merlin, O., Richaume, P., Sahoo, A., Wood, E.F., 2012. Evaluation of SMOS Soil Moisture Products Over Continental U.S. Using the SCAN/SNOTEL Network. IEEE Trans. Geosci. Remote Sens. 50, 1572–1586. doi:10.1109/TGRS.2012.2186581

Alvarez-Garreton, C., Ryu, D., Western, A.W., Crow, W.T., Robertson, D.E., 2014. The impacts of assimilating satellite soil moisture into a rainfall–runoff model in a semi-arid catchment. J. Hydrol. 519, 2763–2774. doi:10.1016/j.jhydrol.2014.07.041

Alvarez-Garreton, C., Ryu, D., Western, A.W., Su, C.H., Crow, W.T., Robertson, D.E., Leahy, C., 2015. Improving operational flood ensemble prediction by the assimilation of satellite soil moisture: Comparison between lumped and semi-distributed schemes. Hydrol. Earth Syst. Sci. 19, 1659–1676. doi:10.5194/hess-19-1659-2015

Bergeron, J.M., Trudel, M., Leconte, R., 2016. Combined assimilation of streamflow and snow water equivalent for mid-term ensemble streamflow forecasts in snow-dominated regions. Hydrol. Earth Syst. Sci. Discuss. 1–34. doi:10.5194/hess-2016-166

Bergsröm, S., 1976. Development and Application of a conceptual runoff model fro Scandinavian catchment. Department of Water Resources Engineering, Lund Institute of Technology, University of Lund.

Boyle, D.P., 2001. Multicriteria calibration of hydrologic models. University of Arizona.

Branisavljevic, N., Hutton, C., Kapelan, Z., Vamvakeridou-Lyroudia, L., Savić, D.A., 2014. Real-Time Runoff Prediction Based on Data Assimilation And Model Bias Reduction, in: 11th International Conference on Hydroinformatics. New York, NY.

Burnash, R.J.C., 1995. The NWS River Forecast System - Catchment Model, in: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications.

Burnash, R.J.C., Ferral, R.L., McGuire, R.A., 1973. A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers. U.S. Department of Commerce, National Weather Service, and State of California Department of Water Resources.

Carroll, T.R., Cline, D.W., Olheiser, C., Rost, A., Nilsson, A., Fall, G., Bovitz, C., Li, L., 2006. NOAA's national snow analyses., in: 74th Annual Meeting of the Western Snow Conference. pp. 31–43.

Chau, K.W., 2006. Particle swarm optimization training algorithm for ANNs in stage prediction of Shing Mun River. J. Hydrol. 329, 363–367. doi:10.1016/j.jhydrol.2006.02.025

Clark, M.P., Rupp, D.E., Woods, R.A., Zheng, X., Ibbitt, R.P., Slater, A.G., Schmidt, J., Uddstrom, M.J., 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Adv. Water Resour. 31, 1309–1324. doi:10.1016/j.advwatres.2008.06.005

Clow, D.W., Nanus, L., Verdin, K.L., Schmidt, J., 2012. Evaluation of SNODAS snow depth and snow water equivalent estimates for the Colorado Rocky Mountains, USA. Hydrol. Process. 26, 2583–2591. doi:10.1002/hyp.9385

Crow, W.T., Ryu, D., 2009. A new data assimilation approach for improving runoff prediction using remotely-sensed soil moisture retrievals. Hydrol. Earth Syst. Sci. 13, 1–16. doi:10.5194/hess-13-1-2009

DeChant, C.M., Moradkhani, H., 2011. Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation. Hydrol. Earth Syst. Sci. 15, 3399–3410. doi:10.5194/hess-15-3399-2011

Drusch, M., Wood, E.F., Gao, H., 2005. Observation operators for the direct assimilation of TRMM microwave imager retrieved soil moisture. Geophys. Res. Lett. 32, L15403. doi:10.1029/2005GL023623

Dubé, I., 2003. From mm to cm... Study of snow/liquid water ratios in Quebec. QC, Quebec, Canada.

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. doi:10.1016/j.advwatres.2013.07.007

Dziubanski, D.J., Franz, K.J., 2016. Assimilation of AMSR-E snow water equivalent data in a spatially-lumped snow model. J. Hydrol. 540, 26–39. doi:10.1016/j.jhydrol.2016.05.046

Entekhabi, D., Njoku, E., O'Neill, P., Spencer, M., Jackson, T., Entin, J., Im, E., Kellogg, K., 2008. The Soil Moisture Active/Passive Mission (SMAP), in: IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium. IEEE, p. III-1-III-4. doi:10.1109/IGARSS.2008.4779267

Environment and Climate Change Canada, 2017. Canadian Climate Normals [WWW Document]. URL http://climate.weather.gc.ca/climate_normals/ (accessed 10.1.17).

Environment Canada, 2013. Environment and Climate Change Canada - Weather and Meteorology - Weather Tools - FAQ [WWW Document]. URL https://ec.gc.ca/meteo-weather/default.asp?lang=En&n=108C6C74-1#wsA0744309 (accessed 5.17.17).

Evensen, G., 2003. The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367. doi:10.1007/s10236-003-0036-9

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143. doi:10.1029/94JC00572

Gill, M.K., Kaheil, Y.H., Khalil, A., McKee, M., Bastidas, L., 2006. Multiobjective particle swarm optimization for parameter estimation in hydrology. Water Resour. Res. 42, n/a-n/a. doi:10.1029/2005WR004528

Google Earth Engine Team, 2015. Google Earth Engine: A planetary-scale geospatial analysis platform. [WWW Document]. URL https://earthengine.google.com

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. doi:10.1016/j.jhydrol.2009.08.003

Huang, C., Newman, A.J., Clark, M.P., Wood, A.W., Zheng, X., 2017. Evaluation of snow data assimilation using the ensemble Kalman filter for seasonal streamflow prediction in the western United States. Hydrol. Earth Syst. Sci. 21, 635–650. doi:10.5194/hess-21-635-2017

Hutton, C.J., Kapelan, Z., Vamvakeridou-Lyroudia, L., Savić, D., 2014. Real-time data assimilation in urban rainfall-runoff models. Procedia Eng. 70, 843–852. doi:10.1016/j.proeng.2014.02.092

Jackson, T.J., Bindlish, R., Cosh, M.H., Zhao, T., Starks, P.J., Bosch, D.D., Seyfried, M., Moran, M.S., Goodrich, D.C., Kerr, Y.H., Leroux, D., 2012. Validation of Soil Moisture and Ocean Salinity (SMOS) Soil Moisture Over Watershed Networks in the U.S. IEEE Trans. Geosci. Remote Sens. 50, 1530–1543. doi:10.1109/TGRS.2011.2168533

Kennedy, J., Spears, W.M., 1998. Matching algorithms to problems: an experimental test of the particle swarm and some genetic algorithms on the multimodal problem generator, in: 1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360). IEEE, pp. 78–83. doi:10.1109/ICEC.1998.699326

Kerr, Y.H., Waldteufel, P., Richaume, P., Wigneron, J.P., Ferrazzoli, P., Mahmoodi, A., Al Bitar, A., Cabot, F., Gruhier, C., Juglea, S.E., Leroux, D., Mialon, A., Delwart, S., 2012. The SMOS Soil Moisture Retrieval Algorithm. IEEE Trans. Geosci. Remote Sens. 50, 1384–1403. doi:10.1109/TGRS.2012.2184548

Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.-J., Font, J., Reul, N., Gruhier, C., Juglea, S.E., Drinkwater, M.R., Hahne, A., Martín-Neira, M., Mecklenburg, S., 2010. The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle. Proc. IEEE 98, 666–687. doi:10.1109/JPROC.2010.2043032

Komma, J., Blöschl, G., Reszler, C., 2008. Soil moisture updating by Ensemble Kalman Filtering in real-time flood forecasting. J. Hydrol. 357, 228–242. doi:10.1016/j.jhydrol.2008.05.020

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. doi:10.1016/j.jhydrol.2003.12.039

Kornelsen, K.C., Coulibaly, P., 2015. Reducing multiplicative bias of satellite soil moisture retrievals. Remote Sens. Environ. 165, 109–122. doi:10.1016/j.rse.2015.04.031

Kornelsen, K.C., Davison, B., Coulibaly, P., 2016. Application of SMOS Soil Moisture and Brightness Temperature at High Resolution With a Bias Correction Operator. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9, 1590–1605. doi:10.1109/JSTARS.2015.2474266

Li, H., Zhang, Y., Chiew, F.H.S., Xu, S., 2009. Predicting runoff in ungauged catchments by using Xinanjiang model with MODIS leaf area index. J. Hydrol. 370, 155–162. doi:10.1016/j.jhydrol.2009.03.003

Liu, Y., Weerts, A.H., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., van Dijk, A.I.J.M., van Velzen, N., He, M., Lee, H., Noh, S.J., Rakovec, O., Restrepo, P., 2012. Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. Hydrol. Earth Syst. Sci. 16, 3863–3887. doi:10.5194/hess-16-3863-2012

Massari, C., Brocca, L., Tarpanelli, A., Moramarco, T., 2015. Data Assimilation of Satellite Soil Moisture into Rainfall-Runoff Modelling: A Complex Recipe? Remote Sens. 7, 11403–11433. doi:10.3390/rs70911403

Matheson, J.E., Winkler, R.L., 1976. Scoring Rules for Continuous Probability Distributions. Manage. Sci. 22, 1087–1096. doi:10.1287/mnsc.22.10.1087

McPherson, M.B., Schneider, W.J., 1974. Problems in modeling urban watersheds. Water Resour. Res. 10, 434–440. doi:10.1029/WR010i003p00434

Monteith, J.L., 1965. Evaporation and environment. Symp. Soc. Exp. Biol. 19, 205–34.

Moradkhani, H., 2008. Hydrologic remote sensing and land surface data assimilation. Sensors 8, 2986–3004. doi:10.3390/s8052986

Moradkhani, H., Sorooshian, S., Gupta, H. V., Houser, P.R., 2005. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour. 28, 135–147. doi:10.1016/j.advwatres.2004.09.002

Murphy, J.M., 1988. The impact of ensemble forecasts on predictability. Q. J. R. Meteorol. Soc. 114, 463–493. doi:10.1002/qj.49711448010

Nascimento, J.M.P., Dias, J.M.B., 2005. Vertex component analysis: a fast algorithm to unmix hyperspectral data. IEEE Trans. Geosci. Remote Sens. 43, 898–910. doi:10.1109/TGRS.2005.844293

National Operations Hydrologic Remote Sensing Center, 2004. Snow Data Assimilation System (SNODAS) Data Products at NSIDC, Version 1 [January 2011 - December 2015]. Boulder, Colorada, USA. NSIDC: National Snow and Ice Data Center. doi:http://dx.doi.org/10.7265/N5TB14TC

Neal, J.C., Atkinson, P.M., Hutton, C.W., 2007. Flood inundation model updating using an ensemble Kalman filter and spatially distributed measurements. J. Hydrol. 336, 401–415. doi:10.1016/j.jhydrol.2007.01.012

Ontario Ministry of Agriculture, 2015. Soil survey complex.

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. 279, 275–289. doi:10.1016/S0022-1694(03)00225-7

Reichle, R.H., 2008. Data assimilation methods in the Earth sciences. Adv. Water Resour. 31, 1411–1418. doi:10.1016/j.advwatres.2008.01.001

Rodríguez-Fernández, N.J., Muñoz Sabater, J., Richaume, P., De Rosnay, P., Kerr, Y.H., Albergel, C., Drusch, M., Mecklenburg, S., 2017. SMOS near-real-time soil moisture product: Processor overview and first validation results. Hydrol. Earth Syst. Sci. 21, 5201–5216. doi:10.5194/hess-21-5201-2017

Samuel, J., Coulibaly, P., Dumedah, G., Moradkhani, H., 2014. Assessing model state and forecasts variation in hydrologic data assimilation. J. Hydrol. 513, 127–141. doi:10.1016/j.jhydrol.2014.03.048

Samuel, J., Coulibaly, P., Metcalfe, R. a., 2012. Identification of rainfall-runoff model for improved baseflow estimation in ungauged basins. Hydrol. Process. 26, 356–366. doi:10.1002/hyp.8133

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. doi:10.1061/(ASCE)HE.1943-5584.0000338

Small, C., 2002. Multitemporal analysis of urban reflectance. Remote Sens. Environ. 81, 427–442. doi:10.1016/S0034-4257(02)00019-6

Small, C., 2001. Estimation of urban vegetation abundance by spectral mixture analysis. Int. J. Remote Sens. 22, 1305–1334. doi:10.1080/01431160151144369

Thiboult, A., Anctil, F., 2015. On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments. J. Hydrol. 529, 1147–1160. doi:10.1016/j.jhydrol.2015.09.036

Unger, D.A., 1985. A method to estimate the continuous ranked probability score, in: Ninth Conference on Probability and Statistices in Atmospheric Sciences. American Meteorological Society, Virginia Beach, VA, pp. 206–213.

Vrugt, J.A., Gupta, H. V., Dekker, S.C., Sorooshian, S., Wagener, T., Bouten, W., 2006a. Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model. J. Hydrol. 325, 288–307. doi:10.1016/j.jhydrol.2005.10.041

Vrugt, J.A., Gupta, H. V., Nualláin, B., Bouten, W., 2006b. Real-Time Data Assimilation for Operational Ensemble Streamflow Forecasting. J. Hydrometeorol. 7, 548–565. doi:10.1175/JHM504.1

Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43, n/a-n/a. doi:10.1029/2005WR004838

Wanders, N., Karssenberg, D., de Roo, A., de Jong, S.M., Bierkens, M.F.P., 2014. The suitability of remotely sensed soil moisture for improving operational flood forecasting. Hydrol. Earth Syst. Sci. 18, 2343–2357. doi:10.5194/hess-18-2343-2014

Weerts, A.H., El Serafy, G.Y.H., 2006. Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. Water Resour. Res. 42, n/a-n/a. doi:10.1029/2005WR004093

Zhang, Y.-F., Yang, Z.-L., 2016. Estimating uncertainties in the newly developed multi-source land snow data assimilation system. J. Geophys. Res. Atmos. 121, 8254–8268. doi:10.1002/2015JD024248

CHAPTER 3 The limits of soil moisture assimilation in urban watersheds

Summary of Paper 2: Leach, J.M. and Coulibaly, P. (submitted) The limits of soil moisture assimilation in urban watersheds, Journal of Hydrology, Manuscript Number: HYDROL33206

Summary:

This research was a follow-up to Chapter 2 with the goal of further evaluating remotely sensed soil moisture assimilation in urban watersheds. From this work, a general imperviousness threshold was able to be determined for assimilating soil moisture.

The key findings of this research include:

- Assimilation of remotely sensed soil moisture has limiting impervious threshold, beyond which the assimilation may negatively impact model performance
- A method was developed to quickly identify imperviousness threshold for urban basins using some basin characteristics
- A general imperviousness threshold range for soil moisture assimilation was also able to be determined based on results from both synthetic and real world experiments.

3.1 Abstract

Data assimilation is well suited for merging observed and simulated values to improve hydrologic forecasts. Typically, observation data that is assimilated into hydrologic models can be obtained from in situ gauges, radar, and remote sensing. In an urban environment, the benefits of soil moisture assimilation are not clearly defined. This work aims to identify the feasible imperviousness range at which it is advantageous to assimilate remotely sensed soil moisture to improve hydrologic forecasting in an urban watershed. A synthetic experiment was set up to simulate the retrieval of soil moisture onto a spatial grid and assimilate it into a hydrologic model. Sub-catchments were set up such that they would represent areas within that retrieval grid with varying levels of development. Multiple rainfall events were simulated with forecasts of up to 12 hours for each sub-catchment. The results of these simulations indicate that when areal average soil moisture is assimilated into an urban sub-catchment model which has a level of impervious that exceeds a threshold value, there is a decrease in model forecast performance, indicating that soil moisture assimilation is no longer beneficial. A quick way to determine the imperviousness threshold was then derived using a modified NRCS-CN method which matches the results of the synthetic experiments. This methodology was then further tested using real-world urban watersheds and shown to be a valid approach which can be used to quickly determine whether soil moisture assimilation would be beneficial for the watershed of interest.

Author keywords: Soil moisture, urban, data assimilation, ensemble Kalman filter

3.2 Introduction

In hydrologic modeling, data assimilation methods are used to assimilate observation data, obtained from in-situ gauges or remote sensing methods, into models to improve state and parameter estimates for better forecasts (Moradkhani, 2008; Moradkhani et al., 2005a, 2005b; Reichle, 2008). There are several data assimilation methods which have been used with hydrologic models to improve forecasts, some common ones being the Extended Kalman Filter (Sun et al., 2015; Wishner et al., 1969), the Ensemble Kalman Filter (EnKF) (Evensen, 1994; Thiboult and Anctil, 2015; Vrugt et al., 2006), and the Particle Filter (DeChant and Moradkhani, 2014; Gordon et al., 1993; Moradkhani et al., 2005a). Each assimilation method has been shown to improve forecasts by integrating available observations into the model while also accounting for uncertainty in those observations and the model.

Remote sensing techniques can provide informative variables for hydrologic modeling such as snow depth, snow water equivalent, and soil moisture (Andreadis and Lettenmaier, 2006; Moradkhani, 2008). Remote sensing soil moisture data products are available from satellite missions such as the Advanced Microwave Scanning Radiometer for Earth observation science (AMSR-E) (Njoku et al., 2003), the Advanced Scatterometer (ASCAT) (Bartalis et al., 2007), the Soil Moisture Active Passive (SMAP) (Entekhabi et al., 2008), and the Soil Moisture and Ocean Salinity (SMOS) (Kerr et al., 2010). Remotely sensed soil moisture data sets, for example, have been used in many data assimilation studies, and have been shown to improve hydrologic forecasts (Alvarez-Garreton et al., 2015; Dziubanski and Franz, 2016; Leach et al., 2018; Moradkhani, 2008). However, these previous studies have focused on assimilation of soil moisture into larger rural basins with little urban development and have rarely been used to improve urban hydrologic forecasting (Alvarez-Garreton et al., 2015; Leach et al., 2018; Lee et al., 2011; Leroux et al., 2016; Massari et al., 2015; Samuel et al., 2014).

Soil moisture plays an important role in the hydrologic cycle by controlling the partition between runoff and infiltration (Kornelsen and Coulibaly, 2014; Moradkhani, 2008). Antecedent soil moisture can directly influence surface runoff, and by properly accounting for the antecedent soil moisture conditions rainfall-runoff modeling can be improved (Nishat et al., 2010). However, antecedent soil moisture has also been shown to have less influence in urban watersheds for small precipitation events when runoff is dominated by the impervious area's contribution (Boyd et al., 1993; Miller and Viessman, 1972), and large events where the rainfall volume dominates the runoff (Nishat et al., 2010). These studies then indicate that there is a range of storm events which could be better simulated through better quantification of the soil moisture in an urban watershed.

The role of soil moisture in rainfall-runoff modeling is often under-valued in urban areas due to the influence of impervious surfaces. The lack of studies which investigate the assimilation of soil moisture within urban catchments can likely be attributed to the level of impervious surfaces within those watersheds. However, even in highly urbanized watersheds, soil moisture assimilation has been shown to improve hydrologic forecasts (Leach et al., 2018). This is possible since watersheds considered highly-urbanized still have pervious areas in which antecedent soil moisture conditions have influence. These areas can easily be revealed through image analysis, such as spectrally unmixing multi-band Earth images like those available from the Landsat missions.

The work presented here will attempt to identify a potential limiting threshold for the level of development/imperviousness which will aid in determining whether it is useful to assimilate soil moisture data when modeling an urban watershed. For the analysis herein, the EnKF data assimilation method was chosen due to its relative simplicity and ease of use. A synthetic experiment was used to represent the assimilation of remotely sensed soil moisture into similar watersheds with various levels of development. This synthetic experiment aided in identifying the limiting imperviousness threshold and in developing an efficient method for identifying the threshold using available information about the watershed. The experiment is meant to emulate urban catchments, which can vary widely in their levels of development. The goal of this work is to determine the impact that the percent of impervious surface in a watershed has on soil moisture assimilation and at what impervious percentage threshold the impact occurs. This threshold identification method will then be applied to selected real-world watersheds to aid in validating its accuracy and generalizability.

3.3 Methodology

3.3.1 Hydrologic Models

3.3.1.1 Storm Water Management Model

The Storm Water Management Model (SWMM) was used for this study. SWMM is an established semi-distributed model for event-based and continuous simulation in semi-urban and urban watersheds. Within SWMM, sub-catchments are represented by a nonlinear reservoir model, and runoff is solved for through conservation of mass and the Manning equation.

$$\frac{\partial d}{\partial t} = i - e - f - q \tag{3-1}$$

$$q = \frac{1.49WS^{1/2}}{An}(d - d_s)^{5/3} \tag{3-2}$$

where *d* is the depth, *t* is timestep, *i* is the rate of rainfall and snowmelt, *e* is the surface evaporation rate, *f* is the infiltration rate, *q* is the runoff rate, *W* is the characteristic width, *S* is the average slope, *A* is the surface area of the sub-catchment, *n* is the surface roughness, and d_s is the depression storage. SWMM sub-catchments are also partitioned into sub-areas of pervious surface or impervious surface (with or without depression storage), which can also be subject to flow re-routing in which some fraction of runoff from one sub-area gets routed through the other. There are also multiple infiltration methods available for use within SWMM (Green-Ampt was used for this study), an optional two-layer groundwater component (unsaturated upper zone and saturated lower zone), and an optional three-layer snowmelt component (plowable snow, pervious area snow, and impervious area snow). The interested reader can find more information on SWMM from Rossman and Huber (2016).

3.3.1.2 NRCS Curve Number

The Natural Resources Conservation Service (NRCS) Curve Number (CN) method is an empirical model that relates precipitation volume to direct runoff. It can be defined through the following equations (NRCS 2004):

$$S = \frac{25400}{CN} - 254 \tag{3-3}$$

$$I_a = cS \tag{3-4}$$

$$Q = \begin{cases} \dfrac{(R - I_a)^2}{R - I_a + S}, & R > I_a \\ 0, & R \leq I_a \end{cases} \tag{3-5}$$

where S is the maximum potential soil moisture storage (mm), I_a is the initial abstraction (mm), c is the ratio of I_a to S and has been shown to range from 0.05 to 0.2 (for this study 0.1 was used) (Lim et al., 2006), CN is the NRCS curve number which is an empirical parameter related to soil type and land use, R is the rainfall (and snowmelt) volume (mm), and Q is the direct runoff (mm).

The CN model was set up to be analogous to how sub-areas are partitioned in SWMM sub-catchments (Fig. 3-2b). Instead of having one model with a CN being the weighted average of the impervious and pervious areas, two NRCS-CN models were created where the CN for the impervious area was 95 and the CN for the pervious area was 50. This set up was chosen so that the internal routing from the impervious to the pervious area could be accounted for. The direct runoff from the pervious sub-area would then be a function of the routed runoff from the impervious sub-area, and the total runoff that leaves the watershed is the sum of the direct runoff from the pervious sub-area and the non-routed direct runoff from the impervious sub-area.

3.3.2 The Ensemble Kalman Filter

Since it is a conventional data assimilation method, and due to its relative ease of implementation, the Ensemble Kalman Filter (EnKF) (Evensen, 2003, 1994) was used for these analyses. The general formulations of the EnKF for state updating are as follows (Evensen, 2003; Moradkhani et al., 2005b):

$$P_t^b = \frac{\left(\widehat{\boldsymbol{X}}_t^b - \overline{\widehat{\boldsymbol{X}}}_t^b\right)\left(\widehat{\boldsymbol{X}}_t^b - \overline{\widehat{\boldsymbol{X}}}_t^b\right)^T}{N-1} \tag{3-6}$$

$$\boldsymbol{q}_t = H_t\left(\widehat{\boldsymbol{X}}_t^b - \overline{\widehat{\boldsymbol{X}}}_t^b\right) = \left(\widehat{\boldsymbol{Y}}_t - \overline{\widehat{\boldsymbol{Y}}}_t\right) \tag{3-7}$$

$$P_t^b H_t^T = \frac{\left(\widehat{\boldsymbol{X}}_t^b - \overline{\widehat{\boldsymbol{X}}}_t^b\right)}{N-1}\boldsymbol{q}_t^T = \Sigma_t^{\hat{X}\hat{Y}} = Cov\left(\widehat{\boldsymbol{X}}_t, \widehat{\boldsymbol{Y}}_t\right) \tag{3-8}$$

$$H_t P_t^b H_t^T = \frac{\boldsymbol{q}_t \boldsymbol{q}_t^T}{N-1} = \Sigma_t^{\hat{Y}\hat{Y}} = Cov\left(\widehat{\boldsymbol{Y}}_t, \widehat{\boldsymbol{Y}}_t\right) \tag{3-9}$$

$$K_t = P_t^b H_t^T\left(H_t P_t^b H_t^T + R_t\right)^{-1} = \Sigma_t^{\hat{X}\hat{Y}}\left(\Sigma_t^{\hat{Y}\hat{Y}} + R_t\right)^{-1} \tag{3-10}$$

$$\hat{x}_{j,t}^a = \hat{x}_{j,t}^b + K_t\left(y_t + \varepsilon_{j,t} - \hat{y}_{j,t}\right) \tag{3-11}$$

where N is the number of Ensemble Members, j (subscript) is the ensemble member, T (superscript) is the transpose operator, H is the observation operator, t (subscript) the time step, $\hat{x}$ is the state estimate, $\widehat{\boldsymbol{X}}$ is the state vector, K is the Kalman gain, b (superscript) is the background estimate, a (superscript) is the analysis estimate, P^b is the background covariance matrix, R is the observation uncertainty, y is the observation being assimilated, $\hat{y}$ is the simulated observation, $\widehat{\boldsymbol{Y}}$ is the simulated observation vector, and ε is zero-mean random noise. These equations are repeated every time step that observations are available for assimilation until the final time step has been reached. For these analyses, the assimilated observation was the areal average soil moisture, and it was used to update the SWMM's states whenever observations were available.

3.3.3 Experiment design

Two experiments were set up to aid in determining a quick method for determining the limiting imperviousness threshold of a sub-catchment at which soil moisture assimilation would be no longer beneficial. The first experiment used synthetic watersheds in SWMM to evaluate the method under ideal conditions and then a surrogate model was created using the NRCS-CN method which could be quickly used to identify the imperviousness threshold using simple watershed characteristics. The second experiment involved modeling real watersheds using SWMM, assimilating SMOS L2 soil moisture into SWMM using the EnKF, and then checking if the surrogate model threshold value would be consistent with them as well.

3.3.3.1 Synthetic experiment

Two synthetic test basins of different sizes were developed for these analyses. Test basin 1 (TB1) has an area similar to a small town, whereas test basin 2 (TB2) was set to have the area of an average neighborhood. Each test basin had 20 near-identical sub-catchments within them, which differed only in their impervious percentage (Fig. 3-2a.). The test basins were also set up to have different aquifer characteristics, with TB1 and TB2 having those of silt-loam and sandy-clay-loam, respectively. The different synthetic basins were used to help show that the method applies to multiple cases. The parameter values used for the synthetic urban sub-catchments, and their descriptions, are listed in Table 3-1. The chosen parameter values are also similar to existing urban watersheds (Liu et al., 2013).

Table 3-1. SWMM parameters used for the synthetic urban sub-catchments. Parameter descriptions adapted from Rossman and Huber (2016). The infiltration and aquifer parameters were obtained from Table A.2 in Rossman and Huber (2016). Other parameter values were chosen as they were similar to real-world urban watershed parameters (Liu et al., 2013).

Parameter	Description	TB1	TB2	Unit
	Sub-catchments			
area	Sub-catchment area	1600	70	ha
%Imperv	Impervious percentage of sub-catchment	5 - 100	5 - 100	%
Width	Characteristic width of sub-catchment	500	200	m
Slope	Sub-catchment slope	0.5	1	%
	Sub-areas			
Nimp	Manning's n for impervious area	0.01	0.01	-
Nperv	Manning's n for pervious area	0.1	0.1	-
Simp	Depression storage for impervious area	0.05	1	mm
Sperv	Depression storage for pervious area	0.05	2	mm
%Zero	Percent of impervious area without depression storage	25	85	%
%Routed	percent of runoff routed from impervious to pervious area	25 - 75	25 - 75	%
	Infiltration - Green-Ampt			
Psi	Soil capillary suction	166.878	169.926	mm
Ksat	Saturated hydraulic conductivity	6.604	6.604	mm/hr
IMD	Initial soil moisture deficit (porosity - field capacity)	0.2	0.217	-
	Aquifers			
Por	Porosity	0.501	0.398	-
WP	Wilting point	0.135	0.136	-
FC	Field capacity	0.2	0.244	-
Ks	Aquifer saturated hydraulic conductivity	7.5	1.524	mm/hr
Kslp	Slope of the logarithm of hydraulic conductivity vs. moisture deficit	27	15	-
Tslp	Slope of soil tension vs. moisture content	15	15	mm
ETu	Fraction of total evaporation available in unsaturated zone	0.4	0.2	-
ETs	Maximum depth evapotranspiration can occur	10	1.5	m
Seep	Deep groundwater seepage rate	0.04	0.01	mm/hr
Umc	Unsaturated zone moisture content at start of simulation	0.3	0.244	-
	Groundwater			
A1	Groundwater flow coefficient	0.08	0.1	-
B1	Groundwater flow exponent	5	1	-
A2	Surface water flow coefficient	0.05	0.1	-
B2	Surface water flow exponent	6	1	-
A3	GW-SW interaction coefficient	0.1	0	-

The synthetic experiment flowchart is illustrated in Fig. 3-1. The precipitation data set used to force the models contains eight years of hourly data and was from a rain gauge located in Toronto, Ontario. This precipitation data set was chosen since it could provide real-world precipitation depths, event durations, and inter-event times. By forcing the models with hourly precipitation, synthetic observations of runoff and soil moisture were generated for each sub-catchment in each test basin. This method was repeated with 25, 50, and 75 percent of impervious runoff being routed internally to the pervious area first before contributing to total runoff.

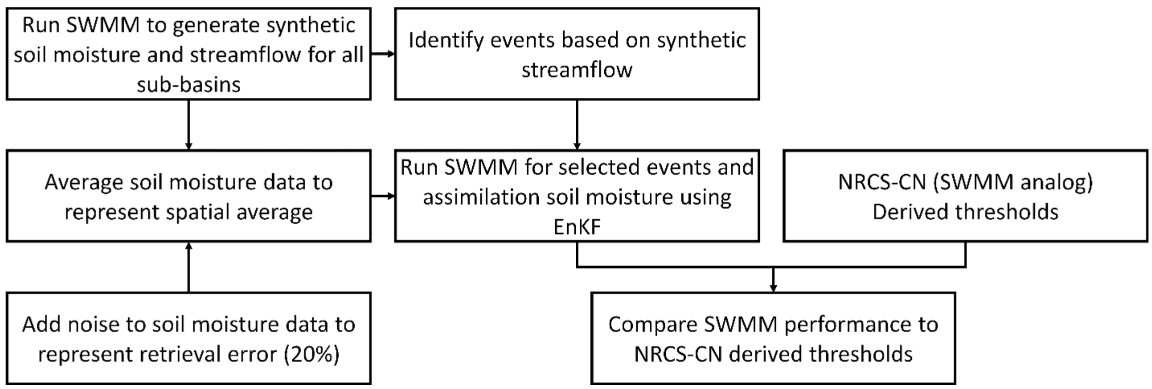

Figure 3-1. Synthetic experiment flowchart for TB1 and TB2. This process was repeated for each test basin when using a percent routed value of 25, 50, and 75 percent.

Taking the average of the synthetic soil moisture observations that were generated for each sub-catchment, the areal average soil moisture was calculated for each test basin. The areal average soil moisture values were then perturbed to represent observation error as follows:

$$y_t = \tilde{y}_t + \varepsilon_t, \varepsilon_t \sim N(0, \Sigma) \quad (3\text{-}12)$$

where y_t is the synthetic areal average soil moisture with observation error, $\tilde{y}_t$ is the synthetic areal average soil moisture observations, and ε_t represents the normally distributed random noise with zero mean and Σ variance added to represent measurement error. The variance of the random noise was equivalent to 20% observation error. In this case, only random measurement error was used since systematic errors should be removed from observations before they are assimilated (Reichle, 2008).

The generated soil moisture was meant to represent retrieved soil moisture values similar to those from ASCAT, SMAP, or SMOS. Data retrieved through remote sensing methods are mapped to regular Discrete Global Grids (DGG) (Bartalis et al., 2007; Entekhabi et al., 2008; Kerr et al., 2010; Wagner et al., 1999). One of these is the Icosahedral Snyder Equal Area (ISEA) grid. For example, the SMOS data retrievals are mapped onto the ISEA4H9 grid which is a 15-km hexagonal DGG. Illustrated in Fig. 3-2 is an example of a hexagonal DGG, as well as a simple example of how the imperviousness of a watershed could vary within a grid cell. The synthetic areal average soil moisture generated for these analyses are meant to emulate this type of retrieved data, since it has been shown that the SMOS data can be directly used on the 15-km DGG despite the spatial discrepancy between the grid and the instrument's resolution (Al Bitar et al., 2012; Dumedah et al., 2014; Kerr et al., 2010).

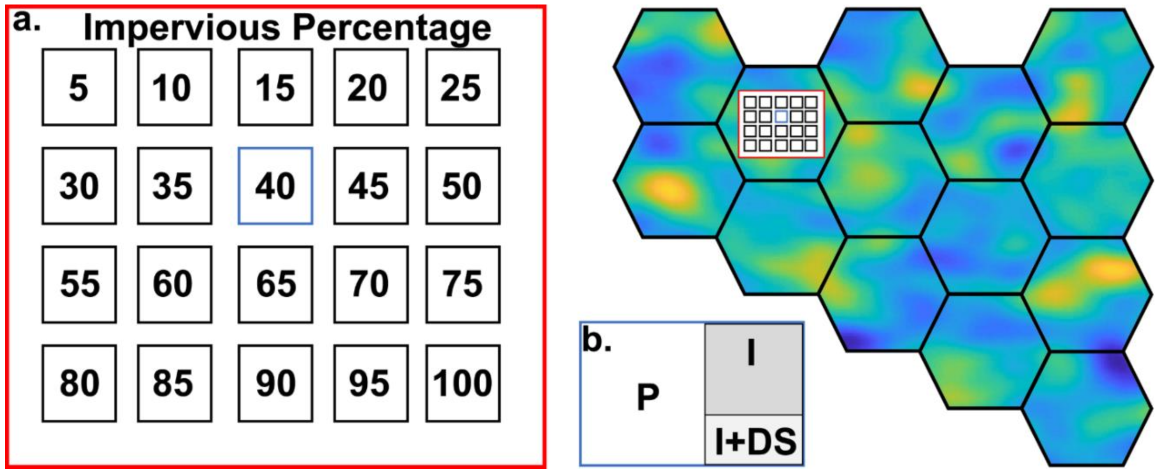

Figure 3-2. Synthetic example in which the areal average soil moisture is assumed to be retrieved onto hexagonal grids like those of SMOS (Dumedah et al., 2014). For this test, the sub-catchments within the hexagonal grid (a) are assumed to have different imperviousness percentages. Each sub-catchment is discretized (b) into Pervious (P), Impervious (I), and Impervious with Depression Storage (I+DS) sub-areas within SWMM (Rossman and Huber, 2016).

To set up the assimilation experiment, precipitation events were first identified for each runoff routing scheme. These events were chosen by identifying high flow events that occurred, ignoring the first year of generated synthetic observations. This was done so that the model states would be stabilized, and each storm event's synthetic runoff observations would have had at least one year of spin up. Doing this provided 58-67 precipitation events, depending on the basin and routing scheme, to account for uncertainty due to different initial conditions. A data assimilation run was set up for each event that modeled a seven-day window, centered around each peak, in which the synthetic areal average soil moisture observations were assimilated at each timestep. After each assimilation step, the SWMM was run 12 hours ahead to determine the influence that the soil moisture assimilation had on the short-term forecast within the urban sub-catchment. Additionally, this forecast horizon allowed for determining how long the soil moisture assimilation would influence the predicted runoff.

The synthetic data assimilation experiment was used to explore the benefit of assimilating remotely sensed areal average soil moisture data for use in event forecasting. The experiment was meant to simulate the areal average soil moisture that would be retrieved over some urban area; this is illustrated in Fig. 3-2. This average soil moisture was assimilated into the hydrologic models which were set up in such a way as to represent multiple sub-catchments within a retrieval grid. Each sub-catchment contains a different level of development, which is represented as the impervious percentage of the basin. The results of the SWMM data assimilation runs were then compared with the NRCS-CN model to provide further insight into the results and to develop a surrogate model which could quickly identify the imperviousness threshold values using sub-catchment characteristics.

3.3.3.2 Real-world experiment

To confirm that the synthetic experiments results and threshold method are consistent with real-world data, the method was then tested using several urban watersheds. The urban watersheds used for testing were identified based on the classifications in the Dudley et al. (2019) dataset. Hourly streamflow data for each watershed were obtained using the R-Studio package *"dataRetrieval"* (De Cicco et al., 2018). The NCEP-CPC Stage 4 Precipitation dataset, which contains 4-km gridded hourly data generated from gauge and radar data, was used for precipitation forcing (Cooperative Distributed Interactive Atmospheric Catalog System et al., 2000), hourly temperature was obtained from the Local Climatological Data (LCD) Dataset through National Oceanic and Atmospheric Administration (NOAA), and soil moisture observations for each watershed were the SMOS L2 Soil moisture. Imperviousness percentage for each watershed was determined using the National Land Cover Database (NLCD) 2016 Urban Imperviousness dataset (Yang et al., 2018); the NLCD 2016 database provides impervious surface percent for the CONUS on 30x30 meter pixels. Yang et al. (2018) reported that the overall agreement of land use pixel identification ranged from 70 to mid 80% in the Eastern United States. More specifically, they reported that the pixels which identify developed areas, which were used to determine the imperviousness of a pixel, were mislabeled 18 to 50% of the time. This uncertainty will be accounted for when calculating the imperviousness threshold for each watershed.

The forcing datasets were preprocessed and filtered based on their level of missing data. If more than 5% of the precipitation and temperature were missing for a grid or gauge it was omitted from the analysis, and if less than 5% of the data was missing it was infilled using simple linear regression. The infilled data was then converted to mean areal precipitation and temperature data, MAP and MAT respectively, using the Thiessen polygon method. The SMOS soil moisture data were filtered based on probability of radio frequency interference and data quality index values; when those values exceeded 0.1, the corresponding data was removed (Kornelsen et al., 2016). After filtering the soil moisture data, there were 43 remaining watersheds with available soil moisture; their locations are shown in Fig. 3-3. These watersheds range in size from 11.55 to 198.91 square kilometers, their level of development ranges from 10 to 100 percent (Dudley et al., 2019), and their level of imperviousness ranges from 1 to 55 percent (Yang et al., 2018); each watershed was modeled using the SWMM.

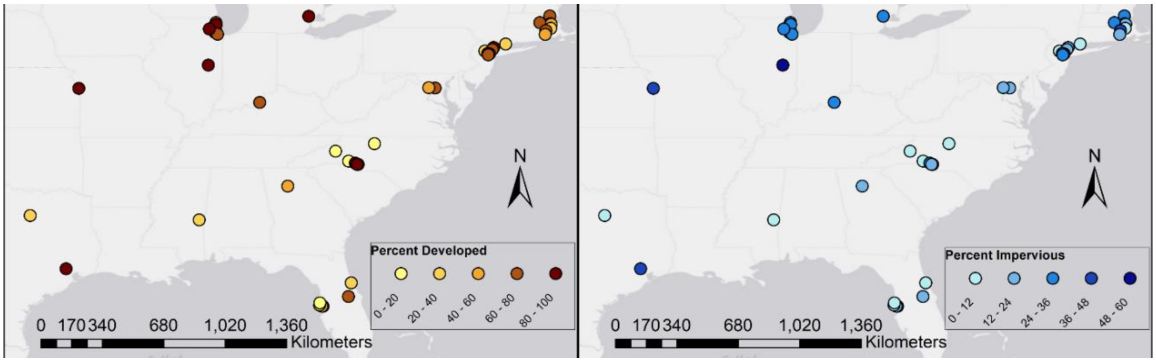

Figure 3-3. Urban watershed locations illustrating their level of development (Dudley et al., 2019) (left) and their level of imperviousness (Yang et al., 2018) (right). Notice that the impervious percentage is less than the percent developed and that the relationship between percent impervious and percent developed is non-linear (a high level of development does not indicate a high level of imperviousness).

Each of the real-world urban watersheds was modeled in a lumped fashion within the SWMM. Additionally, unlike the synthetic SWMM watershed setups, the snowmelt routine was included for the real watersheds since they were calibrated by running the model continuously. The parameters which were calibrated for each urban watershed are listed in Table 3-2. The Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007) was used to calibrate the selected urban watersheds with a calibration period of 2009-01-01 to 2011-12-31 (2008 was used for spin-up), and a validation period of 2012-01-01 to 2014-12-31. The Nash Volume Efficiency (NVE) cost function was used (Samuel et al., 2012) with modified weights placing more emphasis on peak flows:

$$NVE = 0.3NSE - 0.1|VE| + 0.2NSE_{log} + 0.5NSE_{sqr} \quad (3\text{-}13)$$

where *NSE* is the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970), *VE* is the volume error, NSE_{log} is the NSE calculated using log-transformed streamflow, and NSE_{sqr} is the NSE calculated using squared streamflow.

Table 3-2. The SWMM parameters were calibrated for each of the real urban watersheds using the provided parameter ranges. Parameter descriptions and ranges adapted from Rossman and Huber (2016).

Parameter	Description	Range	Unit
Snowmelt			
Stemp	Air temperature at which precipitation falls as snow	-3 to 3	C
ATIwt	Antecedent temperature index weight	0 to 1	-
RNM	Negative melt ratio	0 to 1	-
Sub-catchments			
Width	Characteristic width of sub-catchment	100 to 100000	m
Slope	Sub-catchment slope	0.1 to 10	%
Sub-areas			
Nimp	Manning's n for impervious area	0.011 to 0.15	-
Nperv	Manning's n for pervious area	0.05 to 0.8	-
Simp	Depression storage for impervious area	1.27 to 2.54	mm
Sperv	Depression storage for pervious area	2.54 to 3.302	mm
%Zero	Percent of impervious area without depression storage	0 to 100	%
%Routed	percent of runoff routed from impervious to pervious area	0 to 100	%
Infiltration - Green-Ampt			
Psi	Soil capillary suction	49.022 to 320.04	mm
Ksat	Saturated hydraulic conductivity	0.254 to 120.396	mm/hr
IMD	Initial soil moisture deficit (porosity - field capacity)	0.37 to 0.5	-
Aquifers			
Por	Porosity	0.398 to 0.501	-
WP	Wilting point	0.024 to 0.265	-
FC	Field capacity	0.062 to 0.378	-
Ks	Aquifer saturated hydraulic conductivity	0.254 to 120.396	mm/hr
Kslp	Slope of the logarithm of hydraulic conductivity vs. moisture deficit	5 to 100	-
Tslp	Slope of soil tension vs. moisture content	5 to 100	mm
ETu	Fraction of total evaporation available in unsaturated zone	0 to 1	-
ETs	Maximum depth evapotranspiration can occur	0 to 5	m
Seep	Deep groundwater seepage rate	0 to 8	mm/hr
Ebot	Elevation of the bottom of the aquifer	*	m
Egw	Groundwater table elevation at the start of the simulation	*	m
Umc	Unsaturated zone moisture content at start of simulation	0.37 to 0.5	-
Groundwater			
A1	Groundwater flow coefficient	0.00005 to 0.001	-
B1	Groundwater flow exponent	0 to 4	-

*Note: Ebot and Egw range from 0 to their mean elevation above sea level in meters, this changes for each watershed.

Once the watersheds were calibrated, the data assimilation experiment was set up to assimilate SMOS L2 Soil moisture into each SWMM model. The assimilation was performed during the same time period as the model validation for comparison. The NRCS-

CN surrogate model was then set up for each real watershed, and the results were evaluated to determine if the fast threshold identification method is consistent with real data.

3.3.4 Performance metrics

To determine the impact soil moisture assimilation had on the short term forecast, the following performance metrics were used: the Kling-Gupta efficiency (KGE) (Gupta et al., 2009), Bias, Root Mean Square Error (RMSE), and Mean Continuous Ranked Probability Score (CRPS) (Matheson and Winkler, 1976). Better performance is indicated when Bias, CRPS, and RMSE approach zero and when KGE approaches unity. The metrics are defined as follows:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{3-14}$$

$$Bias = \mathrm{E}[\hat{y} - y] \tag{3-15}$$

$$RMSE = \left(\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2\right)^{1/2} \tag{3-16}$$

$$CRPS(F, y) = \frac{1}{N}\sum_{t=1}^{N}\int_{-\infty}^{\infty}(F(\hat{y}_t) - \mathbf{1}\{\hat{y}_t \geq y_t\})^2 d\hat{y} \tag{3-17}$$

where r is the linear correlation coefficient between the simulated and synthetic observed runoff values, α is the measure of relative variability in the simulated and synthetic observed values, β is the ratio between the simulated and observed values, y_t is the observed runoff at time *t*, $\hat{y}_t$ is the simulated streamflow at time *t*, *N* is the number of time steps in the event, $F(\hat{y}_t)$ is the cumulative distribution function of the forecast, and $\mathbf{1}\{\hat{y}_t \geq y_t\}$ is the Heaviside step function that provides a value of 1 if the predicted value is larger than the observed and 0 otherwise.

3.4 Results and Discussion

3.4.1 Synthetic urban watershed results

The results of the synthetic experiments, where 50% of the impervious flow was routed, are illustrated in Fig. 3-4 and 3-5. From the figures, the KGE, Bias, RMSE, and CRPS are all shown to have similar performances. These results show how the model performance changes with imperviousness and forecast horizon after the states are updated with areal average soil moisture. Fig. 3-4 and 3-5 allows for visualizing how far the update's influence propagates forward in the forecast. From these results, it appears that there is an impervious threshold beyond which assimilating areal average soil moisture is not beneficial and may negatively impact the model's performance. The results also show that when the basin is almost entirely impervious, the effects on the model's performance caused by the assimilation are negligible within two hours, while at lower imperviousness percentages the effects can last several hours longer. These results are caused by the update increasing the water in the sub-areas, particularly the very high impervious subareas (95-99%) by an amount larger than they can store, so it instead runs off immediately causing a large, but limited, effect in the forecast.

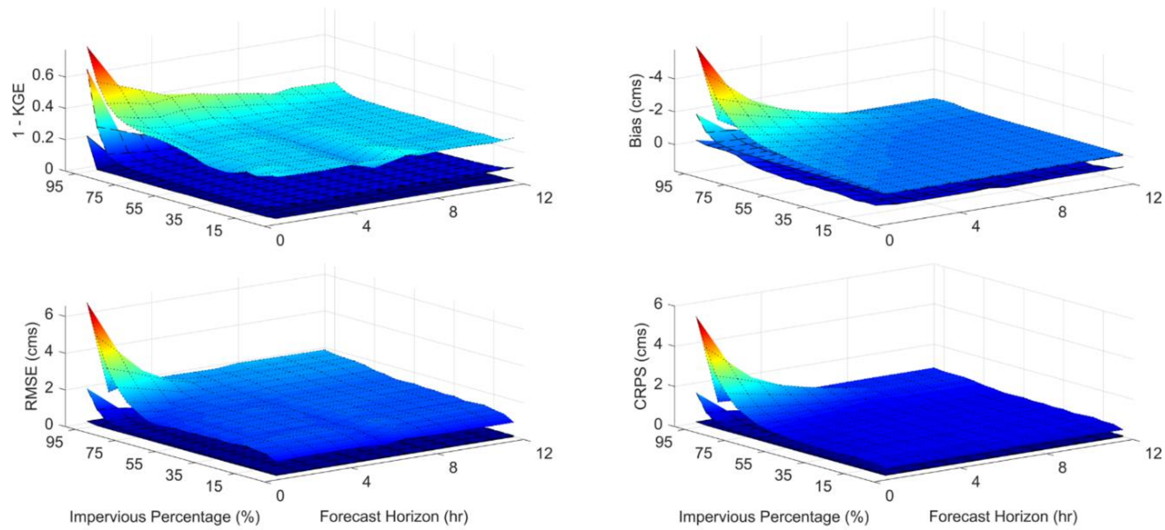

Figure 3-4. Surface plots of 1-KGE, Bias, RMSE, and CRPS for the large synthetic basin (TB1) results with 50 percent of impervious runoff routed to the pervious area. The surfaces which represent the minimum (dotted line), average (dashed line), and maximum (solid line) performance values for the 65 events modeled at various impervious percentages and forecast horizons. These plots represent the region in which the results for all the experiment runs are located within.

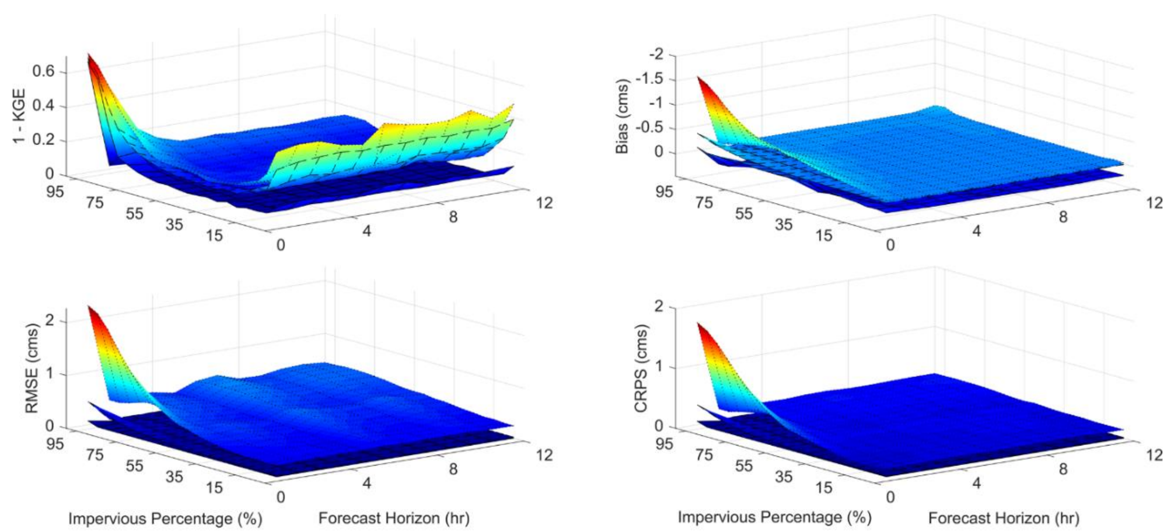

Figure 3-5. Surface plots of 1-KGE, Bias, RMSE, and CRPS for the small synthetic basin (TB2) results with 50 percent of impervious runoff routed to the pervious area. The surfaces which represent the minimum (dotted line), average (dashed line), and maximum (solid line) performance values for the 61 events modeled at various impervious percentages and forecast horizons. These plots represent the region in which the results for all the experiment runs are located within.

To further analyze the results several statistical tests were performed, the first being the Jarque-Bera (JB) test for normality. The JB test showed that the distribution of performances from the events was not normally distributed, indicating that nonparametric tests should be used for analysis. The Mann-Kendall (MK) trend test was utilized to determine if the decrease in performance was statistically significant for each forecast horizon. The MK test results in Table 3-3 show that there was a significantly decreasing trend in performance as imperviousness increased, on average, for all the performance metrics, which diminished as the forecast horizon increased. This indicated that there was a decrease in performance caused when the areal average soil moisture was assimilated into basins with higher imperviousness. Finally, the Kruskal-Wallis (KW) test was used to determine at what impervious percent a significant decrease in performance was seen. The KW indicated that there was a significant decrease in performance when the imperviousness was greater than 90 percent.

Table 3-3. TB1 P-values for Mann-Kendall test for trend absence, at a significance level of α=0.05, for each forecast horizon out to 12 hours. P-values less than 0.05 indicate a significantly decreasing trend is present for each performance metric as the imperviousness level of the watershed increases. The results of TB2 are omitted since the trend analysis shows similar results.

	Forecast Horizon											
Performance Metric	1 hr	2 hr	3 hr	4 hr	5 hr	6 hr	7 hr	8 hr	9 hr	10 hr	11 hr	12 hr
KGE												
max	0.26	0.97	0.72	0.18	0.35	0.92	0.46	0.28	0.18	1.00	0.58	0.82
average	0.00	0.01	0.01	0.06	0.07	0.18	0.28	0.28	0.32	0.54	0.92	1.00
min	0.00	0.01	0.01	0.00	0.00	0.00	0.31	0.02	0.01	0.04	0.06	0.18
Bias												
max	0.00	0.00	0.00	0.06	0.97	0.04	0.02	0.00	0.00	0.00	0.00	0.00
average	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.42	0.23	0.16
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RMSE												
max	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
average	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CRPS												
max	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
average	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

From visual inspection of the results illustrated in Fig. 3-4 and 3-5, the impact of the soil moisture assimilation was noticeable when the imperviousness is greater than 75 percent. However, the KW test showed that the soil moisture assimilation significantly impacted the performance when the imperviousness was greater than 90 percent. When considering these results, it is important also to note that synthetic experiments tend to have higher performances. Therefore this threshold should be lower than the point at which the KW test shows significant declination in performance. The visual inspection of the results then led to the suggested general threshold value of 75 percent watershed imperviousness when 50 percent of impervious sub-area runoff is rerouted to the pervious sub-area, beyond which it is not beneficial to assimilate soil moisture in an urban model.

The previously discussed analysis focused on the results from when 50 percent of impervious sub-area runoff was first routed to the pervious sub-area before contributing to total runoff of the watershed. Similar results are shown when that routing percentage is 25 and 75. However, the thresholds are shifted to higher (85%) and lower (65%) imperviousness values, respectively. To determine the cause of the impervious threshold value and why it is shifted when the routing percentage changes, a simple example was set up using the NRCS-CN method. The ratio of runoff from the impervious fraction (Q_{imp}) of a sub-catchment to total runoff (Q_{tot}) of a watershed was determined for the different levels of connectivity of impervious to pervious sub-areas within the sub-catchments and for different rainfall volumes. The threshold values were then found to correlate well to where the maximum of the Q_{imp}/Q_{tot} ratio was as a function of imperviousness and percent routed, these thresholds are illustrated with vertical lines in Fig. 3-6. Beyond the threshold value the performances of the modeled events begin to degrade, which is shown by the interquartile range of KGE performance beginning to decline in performance beyond the thresholds. These thresholds exist because the effects that soil moisture from the pervious sub-area have on the total runoff are small compared to the runoff being routed to the pervious sub-area from the impervious sub-area of the sub-catchment. The larger this connection between pervious and impervious areas is, the lower the impervious percent of the basin has to be before the influence of soil moisture is no longer meaningful.

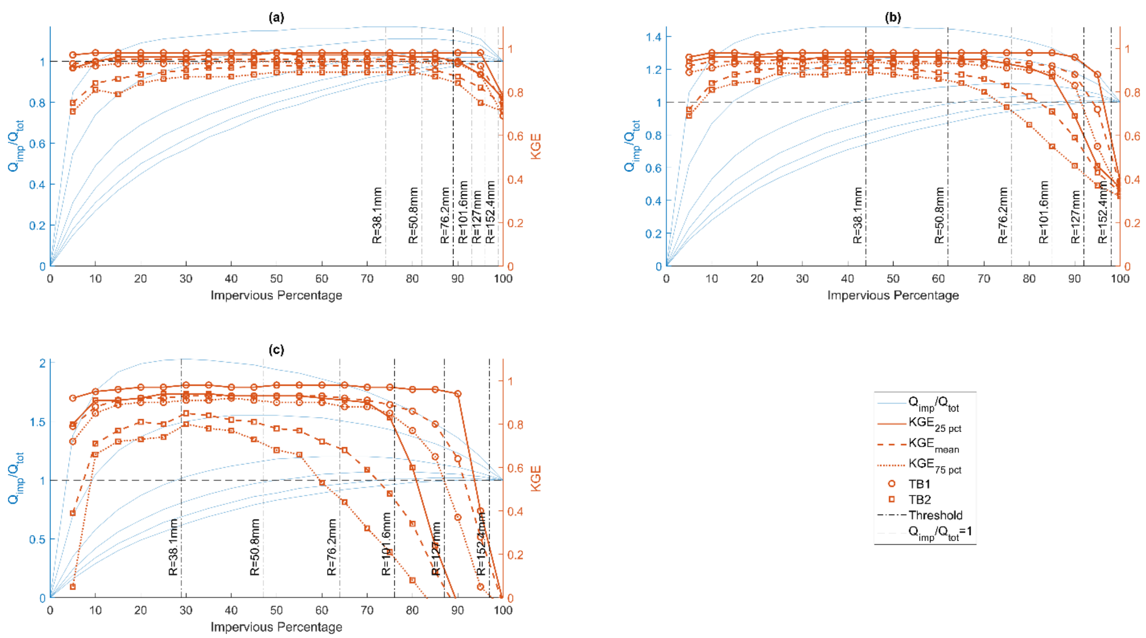

Figure 3-6. Test basin 1 (circle) and 2 (square) KGE performances from 1-hr forecast and the ratio of runoff from impervious sub-area to total runoff from the sub-catchment for (a) 25% connected watershed, (b) 50% connected watershed, and (c) 75% connected watershed. The runoff generated from the impervious portions of the sub-catchment (I and I+DS in Fig. 3-2b.) is denoted as Q_{imp}, and the total runoff from the sub-catchment is denoted as Q_{tot}. Impervious threshold values are shown to be a function of both precipitation volume (R) and basin connectivity.

3.4.2 Real urban watershed results

The calibration and validation results for the real urban watersheds are summarized in Fig. 3-7. From Fig. 3-7, we see that the performance during the calibration period is similar to that of the validation period for most of the watersheds. Additionally, we see that many of the urban watersheds were modeled by SWMM reasonably well. However, only watersheds which obtained a minimum KGE value of 0.6 were chosen for the data assimilation experiment. The justification for this KGE value was that the threshold method requires the *%Routed* parameter to be very well-calibrated.

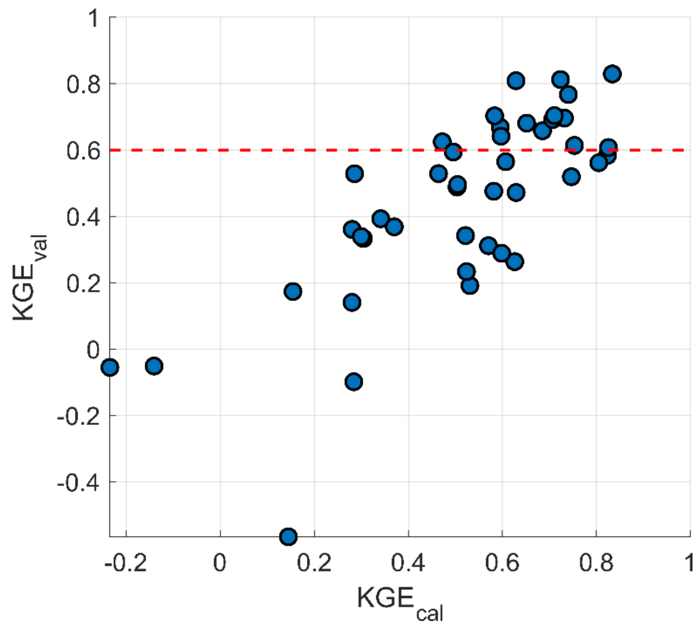

Figure 3-7. KGE Performances during Calibration (2009-01-01 to 2011-12-31) and Validation (2012-01-01 to 2014-12-31) periods for selected urban watersheds modeled in SWMM. The dashed line indicates the minimum acceptable KGE during the validation period.

Unlike for the calibration and validation periods, where the performance metrics were calculated from a continuous time series, the data assimilation results were evaluated based on the performance of the 12 hours immediately following each soil moisture assimilation. This was done so that the real-world watersheds would be evaluated similarly to the synthetic watersheds, and to observe how the soil moisture assimilation would influence the short term forecast. Illustrated in Fig. 3-8 are the predicted threshold ranges which are similar to where the Q_{imp}/Q_{tot} graphs in Fig. 3-6 reach their maximum values as a function of precipitation, which rarely exceeds 75 mm, and *%Routed* (*%Routed* is described in Table 3-2). The percent impervious estimate was also included in this figure for comparison to the derived threshold range. To account for the uncertainty in the percent impervious estimate, the imperviousness was assumed to be up to 25 percent larger (or smaller) for each watershed. This assumption was then used to determine whether there would be a negative impact when assimilating remotely sensed soil moisture into the urban watershed model.

Evaluation of the data assimilation results to determine if there was a significant impact or not was performed as follows. First, the difference between the open-loop model and the data assimilation results was taken for each forecast step out to 12 hours. Next, the MK test for significant trend was used to determine if there was a significantly increasing trend in those residuals, which would indicate that immediately after assimilation there was a performance drop that diminished over time. The predicted and actual impacts are illustrated in the confusion matrix in Fig. 3-8. From the confusion matrix we see that the overall accuracy of the method is 87% and Type I errors occur 18% of the time; there are no Type II errors.

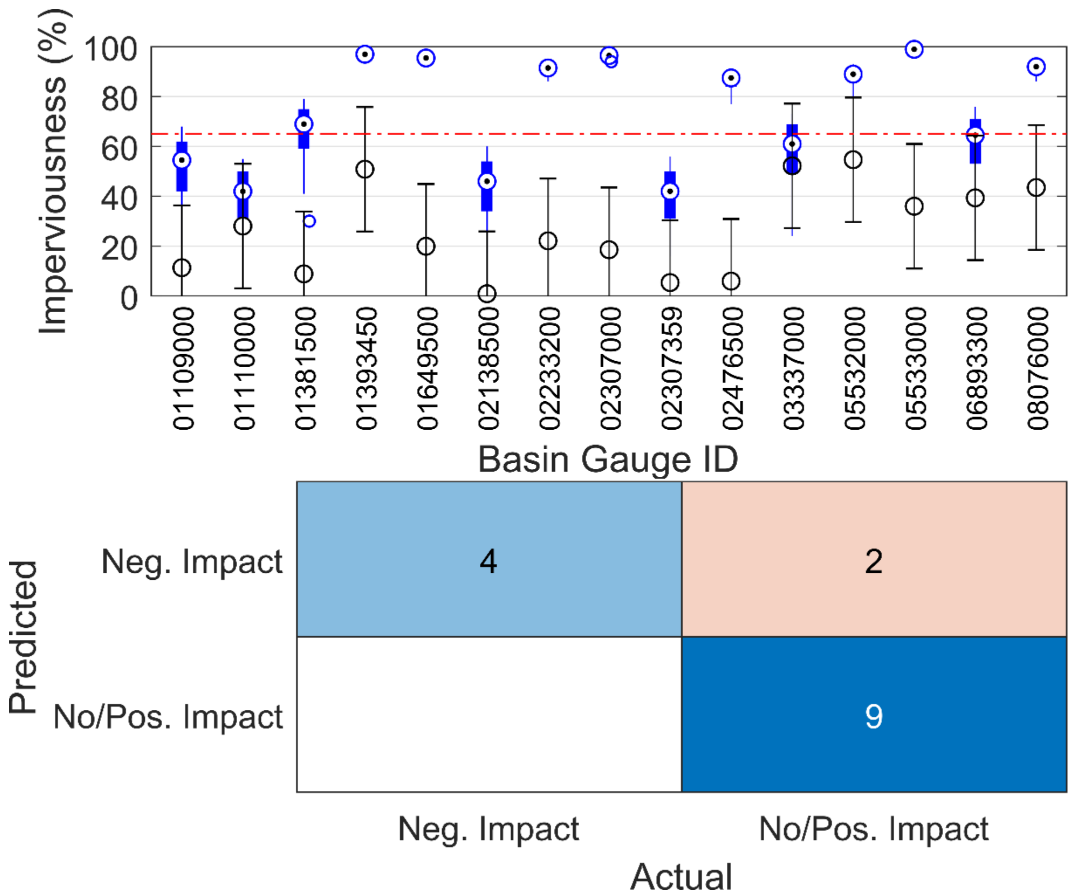

Figure 3-8. Summary of derived imperviousness threshold (top) and the impact of soil moisture assimilation in real-world urban watersheds (bottom). The boxplots represent the derived impervious threshold range which corresponds to the vertical lines in Fig. 3-6 and is a function of the calibrated percent routed parameter and precipitation. The circles with error bars represent the imperviousness estimate for each basin with uncertainty. The red dashed line indicates a general imperviousness threshold based on impact and expected precipitation values. The bottom plot illustrates the confusion matrix comparing whether there was a negative impact or not from assimilating soil moisture and what the predicted impact was using the NRCS-CN surrogate method.

In general, the surrogate model accurately predicted all the watersheds which would be negatively impacted from assimilating remotely sensed soil moisture, while also having a false discovery rate of 33%. However, it should be noted that the imperviousness threshold range is dependent on the *%Routed* parameter, which needed to be calibrated and therefore may not be the actual value for the watershed due to equifinality. There is also the assumed parameter value of $c=0.1$ in equation 3-4, which, if increased or decreased, will cause the threshold to shift down or up, respectively. Finally, there may be some additional uncertainties that influence the results based on the structural error of the SWMM, or due choosing to model each watershed as lumped instead of semi-distributed.

These results suggest that using the surrogate NRCS-CN method is a quick way to determine if assimilating remotely sensed soil moisture into an urban model will be beneficial. Additionally, they show that there is a limiting imperviousness threshold which should be considered before using assimilating remotely sensed soil moisture data. For the real-world urban watersheds, the general threshold is suggested to be about 65 percent imperviousness; this is due to the expected precipitation event volumes which occur. In the synthetic experiment the suggested threshold was about 75 percent imperviousness. Therefore depending on precipitation and actual basin characteristics, it can be expected that the threshold of imperviousness could be in the range of 65% to 75%.

3.5 Conclusions

From the modeling results presented here for the synthetic basin experiment, an imperviousness threshold was identified which shows some limits to the usefulness of soil moisture assimilation in urban watersheds. This threshold was shown to be mainly a function of the impervious to pervious connectivity within a watershed, the basin's imperviousness, and the precipitation volume. We found that for the synthetic case the assimilation of soil moisture had little impact on the model performance at the lower impervious levels, which was to be expected since the difference between the simulated and synthetic soil moistures should be negligible. It was also shown that the assimilation of areal average soil moisture into a model of an urban watershed that is more impervious than the threshold allows for, could cause the model's performance to deteriorate. For a real-world case, it is expected that this degradation will be more pronounced as it is known that synthetic cases tend to have better performances than real-world cases.

Additionally, we found that the threshold method is reasonably consistent when using real-world data. We conclude that, in general, it is feasible to assimilate areal average soil moisture data into a model of an urban watershed as long as the imperviousness of the watershed is below the threshold value of 65% impervious. This threshold can then be refined further on a case-by-case basis. If the impervious percentage of the watershed exceeds the threshold, the modeler will need to adjust their assimilation strategy for soil moisture data, as it may begin to degrade model performance otherwise.

3.6 Acknowledgments

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

3.7 References

Al Bitar, A., Leroux, D., Kerr, Y.H., Merlin, O., Richaume, P., Sahoo, A., Wood, E.F., 2012. Evaluation of SMOS Soil Moisture Products Over Continental U.S. Using the SCAN/SNOTEL Network. IEEE Trans. Geosci. Remote Sens. 50, 1572–1586. https://doi.org/10.1109/TGRS.2012.2186581

Alvarez-Garreton, C., Ryu, D., Western, A.W., Su, C.H., Crow, W.T., Robertson, D.E., Leahy, C., 2015. Improving operational flood ensemble prediction by the assimilation of satellite soil moisture: Comparison between lumped and semi-distributed schemes. Hydrol. Earth Syst. Sci. 19, 1659–1676. https://doi.org/10.5194/hess-19-1659-2015

Andreadis, K.M., Lettenmaier, D.P., 2006. Assimilating remotely sensed snow observations into a macroscale hydrology model. Adv. Water Resour. 29, 872–886. https://doi.org/10.1016/j.advwatres.2005.08.004

Bartalis, Z., Wagner, W., Naeimi, V., Hasenauer, S., Scipal, K., Bonekamp, H., Figa, J., Anderson, C., 2007. Initial soil moisture retrievals from the METOP-A Advanced Scatterometer (ASCAT). Geophys. Res. Lett. 34, L20401. https://doi.org/10.1029/2007GL031088

Boyd, M.J., Bufill, M.C., Knee, R.M., 1993. Pervious and impervious runoff in urban catchments. Hydrol. Sci. J. 38, 463–478. https://doi.org/10.1080/02626669309492699

Cooperative Distributed Interactive Atmospheric Catalog System, Earth Observing Laboratory, National Center for Atmospheric Research, University Corporation for Atmospheric Research, Climate Prediction Center, National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce, 2000. NCEP/CPC Four Kilometer Precipitation Set, Gauge and Radar. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO. https://doi.org/10.5065/D69Z93M3

De Cicco, L.A., Hirsch, R.M., Lorenz, D., Watkins, W., 2018. dataRetrieval: R packages for discovering and retrieving water data available from Federal hydrologic web services. https://doi.org/10.5066/P9X4L3GE

DeChant, C.M., Moradkhani, H., 2014. Toward a reliable prediction of seasonal forecast uncertainty: Addressing model and initial condition uncertainty with ensemble data assimilation and Sequential Bayesian Combination. J. Hydrol. 519, 2967–2977. https://doi.org/10.1016/j.jhydrol.2014.05.045

Dudley, R.W., Hirsch, R.M., Archfield, S.A., Blum, A.G., Renard, B., 2019. Low-streamflow trends and basin characteristics for 2,482 U.S. Geological Survey streamgages in the conterminous U.S. (ver. 2.0, February 2019): U.S. Geological Survey data release. https://doi.org/10.5066/P9LO24MG

Dumedah, G., Walker, J.P., Rudiger, C., 2014. Can SMOS Data be Used Directly on the 15-km Discrete Global Grid? IEEE Trans. Geosci. Remote Sens. 52, 2538–2544. https://doi.org/10.1109/TGRS.2013.2262501

Dziubanski, D.J., Franz, K.J., 2016. Assimilation of AMSR-E snow water equivalent data in a spatially-lumped snow model. J. Hydrol. 540, 26–39. https://doi.org/10.1016/j.jhydrol.2016.05.046

Entekhabi, D., Njoku, E., O'Neill, P., Spencer, M., Jackson, T., Entin, J., Im, E., Kellogg, K., 2008. The Soil Moisture Active/Passive Mission (SMAP), in: IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. III-1-III–4. https://doi.org/10.1109/IGARSS.2008.4779267

Evensen, G., 2003. The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367. https://doi.org/10.1007/s10236-003-0036-9

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143. https://doi.org/10.1029/94JC00572

Gordon, N.J., Salmond, D.J., Smith, A.F.M., 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proc. F Radar Signal Process. 140, 107. https://doi.org/10.1049/ip-f-2.1993.0015

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

Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.-J., Font, J., Reul, N., Gruhier, C., Juglea, S.E., Drinkwater, M.R., Hahne, A., Martín-Neira, M., Mecklenburg, S., 2010. The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle. Proc. IEEE 98, 666–687. https://doi.org/10.1109/JPROC.2010.2043032

Kornelsen, K.C., Coulibaly, P., 2014. Root-zone soil moisture estimation using data-driven methods. Water Resour. Res. 50, 2946–2962. https://doi.org/10.1002/2013WR014127

Kornelsen, K.C., Davison, B., Coulibaly, P., 2016. Application of SMOS Soil Moisture and Brightness Temperature at High Resolution With a Bias Correction Operator. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9, 1590–1605. https://doi.org/10.1109/JSTARS.2015.2474266

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

Lee, H., Seo, D.-J., Koren, V., 2011. Assimilation of streamflow and in situ soil moisture data into operational distributed hydrologic models: Effects of uncertainties in the data and initial model soil moisture states. Adv. Water Resour. 34, 1597–1615. https://doi.org/10.1016/j.advwatres.2011.08.012

Leroux, D.J., Pellarin, T., Vischel, T., Cohard, J.-M., Gascon, T., Gibon, F., Mialon, A., Galle, S., Peugeot, C., 2016. Assimilation of SMOS soil moisture into a distributed hydrological model and impacts on the water cycle variables over the Ouémé catchment in Benin. Hydrol. Earth Syst. Sci. Discuss. 1–26. https://doi.org/10.5194/hess-2015-548

Lim, K.J., Engel, B.A., Muthukrishnan, S., Harbor, J., 2006. EFFECTS OF INITIAL ABSTRACTION AND URBANIZATION ON ESTIMATED RUNOFF USING CN TECHNOLOGY. J. Am. Water Resour. Assoc. 42, 629–643. https://doi.org/10.1111/j.1752-1688.2006.tb04481.x

Liu, G., Schwartz, F.W., Kim, Y., 2013. Complex baseflow in urban streams: an example from central Ohio, USA. Environ. Earth Sci. 70, 3005–3014. https://doi.org/10.1007/s12665-013-2358-3

Massari, C., Brocca, L., Tarpanelli, A., Moramarco, T., 2015. Data Assimilation of Satellite Soil Moisture into Rainfall-Runoff Modelling: A Complex Recipe? Remote Sens. 7, 11403–11433. https://doi.org/10.3390/rs70911403

Matheson, J.E., Winkler, R.L., 1976. Scoring Rules for Continuous Probability Distributions. Manage. Sci. 22, 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087

Miller, C.R., Viessman, W., 1972. Runoff volumes from small urban watersheds. Water Resour. Res. 8, 429–434. https://doi.org/10.1029/WR008i002p00429

Moradkhani, H., 2008. Hydrologic remote sensing and land surface data assimilation. Sensors 8, 2986–3004. https://doi.org/10.3390/s8052986

Moradkhani, H., Hsu, K.-L., Gupta, H., Sorooshian, S., 2005a. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. Water Resour. Res. 41, W05012. https://doi.org/10.1029/2004WR003604

Moradkhani, H., Sorooshian, S., Gupta, H. V., Houser, P.R., 2005b. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour. 28, 135–147. https://doi.org/10.1016/j.advwatres.2004.09.002

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

Natural Resources Conservation Service (NRCS), 2004. Chapter 10 Estimation of Direct Runoff from Storm Rainfall, in: Part 630 Hydrology, National Engineering Handbook. U.S. Department of Agriculture, Washington, D.C.

Nishat, S., Guo, Y., Baetz, B.W., 2010. Antecedent soil moisture conditions of different soil types in South-western Ontario, Canada. Hydrol. Process. 24, 2417–2424. https://doi.org/10.1002/hyp.7647

Njoku, E.G., Jackson, T.J., Lakshmi, V., Chan, T.K., Nghiem, S.V., 2003. Soil moisture retrieval from AMSR-E. IEEE Trans. Geosci. Remote Sens. 41, 215–229. https://doi.org/10.1109/TGRS.2002.808243

Reichle, R.H., 2008. Data assimilation methods in the Earth sciences. Adv. Water Resour. 31, 1411–1418. https://doi.org/10.1016/j.advwatres.2008.01.001

Rossman, L.A., Huber, W.C., 2016. Storm Water Management Model Reference Manual Volume I – Hydrology. U.S. Environ. Prot. Agency 231. https://doi.org/EPA/600/R-15/162A

Samuel, J., Coulibaly, P., Dumedah, G., Moradkhani, H., 2014. Assessing model state and forecasts variation in hydrologic data assimilation. J. Hydrol. 513, 127–141. https://doi.org/10.1016/j.jhydrol.2014.03.048

Samuel, J., Coulibaly, P., Metcalfe, R. a., 2012. Identification of rainfall-runoff model for improved baseflow estimation in ungauged basins. Hydrol. Process. 26, 356–366. https://doi.org/10.1002/hyp.8133

Sun, L., Nistor, I., Seidou, O., 2015. Streamflow data assimilation in SWAT model using Extended Kalman Filter. J. Hydrol. 531, 671–684. https://doi.org/10.1016/j.jhydrol.2015.10.060

Thiboult, A., Anctil, F., 2015. On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments. J. Hydrol. 529, 1147–1160. https://doi.org/10.1016/j.jhydrol.2015.09.036

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

Vrugt, J.A., Gupta, H. V., Nualláin, B., Bouten, W., 2006. Real-Time Data Assimilation for Operational Ensemble Streamflow Forecasting. J. Hydrometeorol. 7, 548–565. https://doi.org/10.1175/JHM504.1

Wagner, W., Lemoine, G., Rott, H., 1999. A Method for Estimating Soil Moisture from ERS Scatterometer and Soil Data. Remote Sens. Environ. 70, 191–207. https://doi.org/10.1016/S0034-4257(99)00036-X

Wishner, R.P., Tabaczynski, J.A., Athans, M., 1969. A comparison of three non-linear filters. Automatica 5, 487–496. https://doi.org/10.1016/0005-1098(69)90110-1

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S.M., Case, A., Costello, C., Dewitz, J., Fry, J., Funk, M., Granneman, B., Liknes, G.C., Rigge, M., Xian, G., 2018. A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. ISPRS J. Photogramm. Remote Sens. 146, 108–123. https://doi.org/10.1016/j.isprsjprs.2018.09.006

CHAPTER 4 Data assimilation in ungauged basins using mutual information

Summary of Paper 3: Leach, J.M. and Coulibaly, P. (submitted) Data assimilation in ungauged basins using mutual information, Journal of the American Water Resources Association, Manuscript Number: JAWRA-19-0136-P

Summary:

The goal of this research was to improve data assimilation with semi-distributed models. Modifying the Kalman gain with mutual information, an information entropy measure, we showed that a better update to model states could be made in sub-basins which do not have gauges.

The results of this research demonstrate:

- The Ensemble Kalman Filter (EnKF) could be modified to include the mutual information entropy measure, termed the MIEnKF.
- The EnKF was compared to the MIEnKF using the Lorenz 63 model, a common testbed model. This comparison showed that by using the MIEnKF to update unmeasured states with measured ones (analogous to using observations from a gauged sub-basin to update states in an ungauged one), the performance could be improved.
- After showing its validity for a simple case, the MIEnKF was then tested using the HyMod and Raven models. These results showed that it could be used to update ungauged sub-basin states more effectively than the EnKF.

4.1 Abstract

The primary goal for this work was to improve the ability of the Ensemble Kalman Filter (EnKF) to update states and parameters of a semi-distributed hydrologic model when streamflow observations are not available for all sub-basins. The developed method utilizes an entropy term, namely Mutual Information (MI) between the gauged and ungauged locations, to improve model states and parameter updating using EnKF. The proposed Mutual Information EnKF (MIEnKF) was first tested using Lorenz 63 model for theoretical understanding, then on a small semi-distributed HyMod model, and finally using a larger semi-distributed model built using the Raven hydrologic model. The Don and Humber River basins in Southern Ontario were chosen as study areas; they are urban and semi-urban watersheds, respectively. Overall, the results show that incorporating the MI into the EnKF for updating model states in ungauged sub-basins can improve streamflow prediction in those ungauged basins. However, when attempting to use the MIEnKF to also update model parameters in ungauged sub-basins, the method was not as effective because of the differences in parameter distributions. Finally, the proposed MIEnKF method was shown to be applicable in different models and is scalable from a low to a high number of ungauged sub-basins.

(**KEYWORDS:** data assimilation; ensemble Kalman filter; hydrologic model; entropy; mutual information; streamflow)

4.2 Introduction

As computation power increases, it is becoming more and more viable to use higher resolution models for hydrologic forecasting. In general, it is expected that as the spatial resolution of a model increases and the watershed characteristics are better represented, the forecast can be improved (Singh and Woolhiser 2002; Carpenter and Georgakakos 2006). However, increasing the model's spatial resolution can be limited by data availability (Singh and Woolhiser 2002). As the model becomes more discretized, it would be ideal to have the number of streamflow gauges match the number of sub-basins in the model. In practice, however, this is not always the case. For example, when using the hydrologic response unit (HRU) method, there may be many more discretized areas than there are gauges. Additionally, as the drainage area becomes more discretized, the more difficult it becomes to calibrate the several sub-basins by using few or one streamflow gauge (Awol, Coulibaly, and Tolson 2018). At first glance, it seems like a simple solution to this gauging problem would be to increase the number of gauges. However, there is a trend in which the number of gauged basins is actually on the decline (Mishra and Coulibaly 2010, 2009), which is an important issue to hydrologists and decision-makers. The need arises then for regionalization methods, which can use the information from existing gauges to estimate ungauged basins (Razavi and Coulibaly 2013). These estimations, however, come with unavoidable uncertainty.

Sequential data assimilation (SDA) is a recursive application of Bayes theorem which uses new observations as they become available to update the prior estimate for a more accurate posterior estimate (Liu et al. 2012). In the case of hydrologic modeling, observations such as streamflow are used to update the states and parameters of a model to improve hydrological simulations and forecasts (Moradkhani et al. 2005; Samuel et al. 2014). It does this by accounting for the uncertainties in the model, forcing data, and observations to provide an optimal solution (Liu et al. 2012; Reichle 2008). For the simplest cases of hydrologic models, lumped models, using SDA methods to update model states and parameters is straightforward because the basin is considered homogenous and the observations are directly relatable to the model states and parameters. This becomes more difficult as the complexity of the model increases, specifically when the model is semi or fully distributed. The reason for this is due to the number of states and parameters requiring updating becoming disproportionately larger than the available observations, and those observations being more likely related to a further downstream sub-basin.

Streamflow observations remain the most common observation used for updating hydrologic model states, despite there being several distributed data products available, such as those from the Soil Moisture Ocean Salinity (SMOS), Soil Moisture Active Passive (SMAP), and Advanced Microwave Scanning Radiometer (AMSR) missions (Leach, Kornelsen, and Coulibaly 2018; Abbaszadeh, Moradkhani, and Yan 2018; Pathiraja et al. 2016; Abaza et al. 2015; Thiboult and Anctil 2015; Dumedah and Coulibaly 2012; Clark et al. 2008; Moradkhani et al. 2005). Since it is infeasible to have a gauge for every river, stream, and creek in a watershed, updating ungauged locations continues to be a challenge in hydrologic data assimilation due to the lack of available observations. Although it is possible to use observations from nearby gauges to update these basins, not knowing the underlying spatial covariance/correlations between them makes updating the ungauged locations difficult (Clark et al. 2008). Despite this, data assimilation has been used to update locations where observations are not available (Clark et al. 2008; Rakovec et al. 2012; Xie and Zhang 2010), in some cases by using a localization method which gives more weight to updates closer to where observations are available (Rasmussen et al. 2015).

Information theory can be used to describe the amount of information a random variable contains (Shannon 1948). The information content between two or more random variables can be quantified; this value is known as mutual information. Mutual information is a measure of the linear and non-linear dependence between random variables and is independent of the original units of measurement. Put another way, the mutual information between variables X and Y, if we know Y, is the information known about $X|Y$. This is analogous to how the states and parameters of a hydrologic model can be updated by assimilating streamflow. A recent application of information theory with data assimilation was Nearing et al. (2018), in which the efficiency of the Ensemble Kalman Filter (EnKF) update was evaluated. They found for their case study that the EnKF used a small portion of the information content of remote sensing retrievals to provide updates.

Currently, uses of information theory entropy with data assimilation are limited. However, due to the relationship of mutual information with correlation and covariance, it is reasonable to assume that entropy could be used to help in performing updates. The main goal of this paper is to determine if accounting for the mutual information, a measure of shared information between random variables, can provide a more effective update of states and parameters. This work explores the use of mutual information as a stand-in for the unknown spatial correlation/covariance between gauged and ungauged sub-basins, and if doing so allows for an improved update at the ungauged locations. A modified EnKF was developed which incorporates the mutual information between the ungauged and gauged basins, denoted as the Mutual Information EnKF (MIEnKF), to test this. Several iterations of the MIEnKF were tested to determine which produced the best updates of states and parameters in ungauged locations. These results were tested against the results of the EnKF and the implementation of the EnKF from Clark et al. (2008).

4.3 Study area and data

4.3.1 Study area

The study area focused on in this paper is the Don River basin (DRB) and Humber River basin (HRB) in Toronto, Ontario, Canada (Figure 4-1); the Toronto Region Conservation Authority (TRCA) manages both basins. These two basins contain several sub-catchments, the largest of them being the East Don, West Don, Taylor-Massey Creek, East Humber River, West Humber River, and Black Creek. The DRB is 316 km^2 and is a mostly urban watershed being roughly 93% developed, with the remaining area being split between crops, forest, and wetland (Agriculture and Agri-Food Canada 2015). The HRB is 892 km^2 and is a semi-urban watershed being roughly 37% developed, primarily in the downstream region, with the remaining area being split between crops (44%), forest/wetland (18%), and water/other (1%) (Agriculture and Agri-Food Canada 2015). These basins have an average daily temperature of 8.0 °C, the average daily minimum and maximum temperatures being 3.4 °C to 12.5 °C respectively, and average annual precipitation of 841.1 mm/year based on the 1981-2010 Canadian Climate Normals (Environment and Climate Change Canada 2017). Major soils in the basins include sandy loam, loam, clay loam, and clay (Government of Canada 2018). The DRB's elevation ranges from 75 to 330 meters above mean sea level (masl), and the HRB's elevation ranges from 85 to 490 masl (Natural Resources Canada 2015).

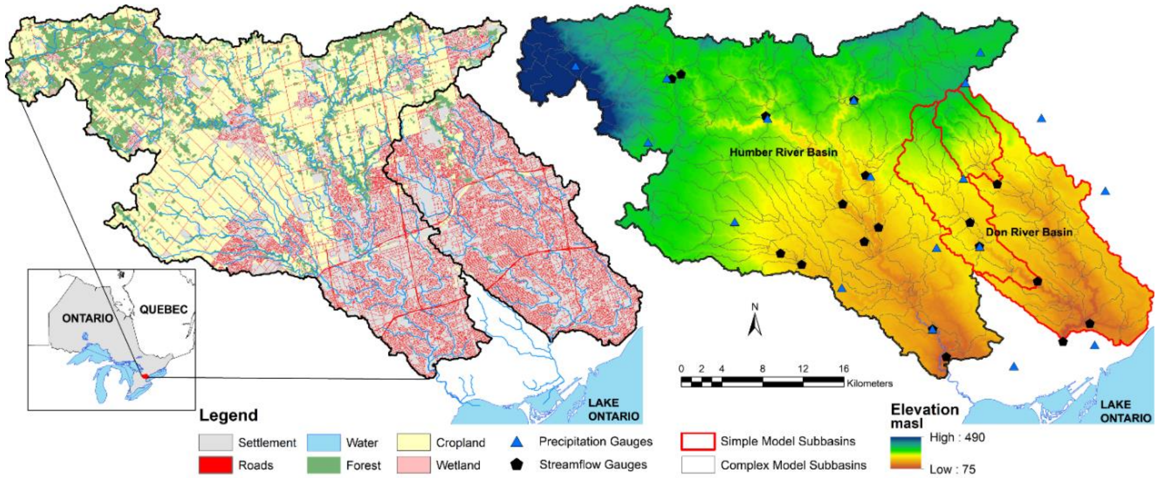

Figure 4-1: The Don and Humber River basins illustrating land-use and land cover (left) and elevation (right). Also shown are the streamflow and precipitation gauge locations (right), the sub-basins used in the HyMod Model (right - red boundaries), and the sub-basins in the Raven model (right - grey boundaries).

4.3.2 Data preprocessing

Hourly streamflow was provided by the TRCA and by Environment and Climate Change Canada (ECCC)/Water Survey of Canada (WSC). Hourly precipitation data was provided by the TRCA, and hourly temperature data was provided by the TRCA and ECCC. Missing temperature data was infilled using ordinary kriging. Missing precipitation data was infilled using a combination of ordinary kriging and disaggregated, daily to hourly, ECCC precipitation data. The disaggregation was performed using Multiplicative Random Cascade-based disaggregation (Olsson 1995, 1998; Ganguli and Coulibaly 2017). Potential evapotranspiration was estimated using the Penman-Monteith equation (Monteith 1965).

4.4 Methodology

4.4.1 Numerical Models

The research presented here will compare variants of the EnKF which have been developed to better update states from ungauged (or unmeasured locations) using data from another location that is gauged (or measured). We provide a conceptual method for modifying the Kalman gain and empirically test it with three models, the Lorenz 63 model (Lorenz 1963), the HyMod model (Boyle 2001), and the Raven model (Craig and the Raven Development Team 2018). The Lorenz 63 model is meant to be a simple case for comparing the methods. The HyMod model is meant to be more complex than Lorenz, but simpler than Raven. The Raven model is meant to be the most complex case, as well as to show that the method is scalable to largely discretized watershed models and is valid with different hydrologic models and watersheds (urban and semi-urban).

4.4.1.1 Lorenz 63 Model.

The Lorenz system is meant to be a simplified atmospheric convection model and is a common testbed for data assimilation methods. The Lorenz 63 model is a chaotic system which is both dynamic and non-linear; it is represented using three ordinary differential equations (ODEs; Lorenz 1963):

$$\frac{dx}{dt} = \sigma(y_t - x_t) + \varepsilon_t^x \tag{4-1}$$

$$\frac{dy}{dy} = x_t(\rho - z_t) - y_t + \varepsilon_t^y \tag{4-2}$$

$$\frac{dz}{dt} = x_t y_t - \beta z_t + \varepsilon_t^z \tag{4-3}$$

where *x* is the rate of convection, *y* is the horizontal temperature variation, *z* is the vertical temperature variation, σ is the Prandtl number, ρ is the normalized Rayleigh number, β is a non-dimensional wavenumber, and ε is Gaussian random noise added to represent model error. The fourth-order Runge-Kutta method using an iteration step of $\Delta t=0.01$ was used to solve this system of ODEs.

To evaluate the data assimilation techniques, the Lorenz 63 model was first run without random noise being added. The results of this run were saved as the true value to be used for evaluation. When performing the assimilation experiment with the Lorenz 63 model, a vector of observations was created by adding Gaussian random noise to the vector of true values to simulate measurement error.

4.4.1.2 HyMod Model.

The HyMod model (Boyle 2001) was used to set up a simple distributed model in the DRB with three sub-basins corresponding to the ECCC-WSC streamflow gauges within the basin (02HC056, 02HC005, and 02HC024). HyMod is a simple conceptual rainfall-runoff model and for this research was modified to include a degree-day snow routine and Muskingum-Cunge routing. The calibrated parameters, as well as the model states, are listed in Table 4-1. The HyMod model was used to test the data assimilation methods on a simpler semi-distributed hydrologic model before using them on more complex semi-distributed hydrologic models.

Table 4-1: Description of HyMod model parameters and state variables updated through data assimilation.

Parameter	Description	Range	Units	Calibrated Value 02HC056	02HC005	02HC024
Alpha	Factor distributing the runoff between the quick and slow reservoirs	0.1 – 0.99	-	0.438	0.209	0.268
B_{exp}	Degree of spatial variability of soil moisture capacity	0.1 – 3	-	0.497	2.992	0.569
C_{max}	Maximum storage capacity	1 – 1000	mm	170.877	139.466	568.734
Rs	Residence time of slow flow reservoir	0.01 – 0.99	day	0.417	0.210	0.423
Rq	Residence times of quick flow reservoirs	0.001 – 0.1	day	0.007	0.020	0.011
tr	Rainfall threshold temperature	0 – 2.5	°C	2.395	1.345	0.062
scf	Snow correction factor	0.4 – 1.6	-	0.664	1.346	0.768
ddf	Degree day factor	0 – 5	mm day-1 °C-1	2.788	4.986	4.949
rcr	Rainfall correction factor	0.5 – 1.5	-	0.737	1.499	1.312
K	Storage coefficient	0 – 1	day	0.407	0.304	-
X	Weighting factor	0 – 0.5	-	0.126	0.494	-
		State variables				
S	Watershed storage	-	mm	-	-	-
qfr1	Quick flow reservoir 1	-	mm	-	-	-
qfr2	Quick flow reservoir 2	-	mm	-	-	-
qfr3	Quick flow reservoir 3	-	mm	-	-	-
sfr	Slow flow reservoir	-	mm	-	-	-
swe	Snow water equivalent	-	mm	-	-	-

4.4.1.3 Raven Model.

More discretized models for both the DRB and the HRB were built using the Raven modeling framework (Craig and the Raven Development Team 2018). The Raven models are set up to be semi-distributed with multiple soil layers, and to use iterative hydrologic routing for channel routing. The calibrated parameters, as well as the states that were updated, are shown in Table 4-2. The DRB and HRB were discretized into several HRUs using land use-land cover, soil, and elevation data; each HRU was then used as sub-basin within the model framework. In total, by using the HRU method for discretizing a watershed, 58 sub-basins were identified for the DRB Raven model, and 149 sub-basins were identified for the in HRB Raven model.

Table 4-2: Description of calibrated Raven model parameters and state variables updated through data assimilation.

Parameter	Description	Range	Units
Global Parameters			
Rainsnow Delta	The range of temperature where mixed precipitation occurs	0.0 - 2.0	°C
Rainsnow Temperature	Center of the range of temperature where mixed precipitation occurs	-1.0 - 1.0	°C
Precipitation Lapse Rate	Precipitation lapse rate	0 - 5.0	mm/d/km
Adiabatic Lapse Rate	Adiabatic lapse rate	4.0 - 12.0	°C/km
Irreducible snow saturation	The maximum liquid water content of snow (fraction)	0 - 1.0	-
Soil Parameters for each Soil Class			
Beta	Infiltration exponent	0 - 10.0	-
Baseflow coefficient	Linear baseflow storage/routing coefficient	0.001 - 10000.0	1/d
Baseflow N	Baseflow exponent	1.0 - 10.0	-
Max Perc Rate	Percolation rate	1.0 - 20.0	mm/d
Max Interflow Rate	Max interflow rate	0.01 - 1000.0	mm/d
Land Use Parameters for each Land Use Class			
Impermeability	The fraction of the surface that is impermeable	0 - 1.0	-
Forest Cover	The fraction of land covered by a canopy	0 - 1.0	-
Forest Sparseness	Canopy sparseness	0 - 0.99	-
Refreeze factor	Maximum refreeze factor	0 - 5.0	mm/d/°C
Depth max	Maximum depression storage	0 - 20.0	mm
Vegetation Parameters for each Vegetation Class			
Max HT	Maximum vegetation height	0 - 30.0	m
Max LAI	Maximum leaf area index	0 - 6.0	-
Max Leaf Cond	Maximum leaf conductance	0 - 14.0	mm/s
Max Capacity	Maximum rain canopy capacity	0 - 5.0	mm
Rain icept frac	Rain throughfall fraction	0 - 0.2	-
Max snow capacity	Maximum snow canopy capacity	0 - 5.0	mm
Snow icept frac	Snow throughfall fraction	0 - 0.2	-
Seasonal Canopy LAI	Relative leaf area index. Found for each month (fraction).	0 - 1.0	-
State variables			
Surface water	Streams, rivers, rivulets routed to basin outlet through in-catchment routing	-	mm
Atmospheric water	Receiving water	-	mm
Atmospheric Precipitation	Providing water	-	mm
Ponded Water	Water waiting to infiltrate or runoff	-	mm
Soil Layer Storage	Water stored in soil layers	-	mm
Canopy	Liquid water on the canopy	-	mm
Canopy Snow	Snow on canopy	-	mm
Snow	Frozen snow depth	-	mm
Snow liquid	Liquid snow cover	-	mm
Depression Storage	Depression/surface storage	-	mm

4.4.1.4 Model Calibration.

The hydrologic models were calibrated using the Dynamically Dimensioned Search algorithm (DDS), a heuristic global search algorithm (Tolson and Shoemaker 2007). The Nash-Volume Error (NVE) (Samuel, Coulibaly, and Metcalfe 2012) was used as the cost function for calibration. The NVE was modified by using different weights for its components, as follows:

$$NVE = w_1 NSE - 0.1|VE| + w_2 NSE_{log} + w_3 NSE_{sqr} \tag{4-4}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \tag{4-5}$$

$$VE = (\bar{Y} - \bar{X})/\bar{X} \tag{4-6}$$

where *NSE* is the Nash-Sutcliffe efficiency (Nash and Sutcliffe 1970) with *X* being observed values and *Y* being simulated values, *VE* is the volume error, NSE_{log} is the NSE calculated using the log streamflow values, NSE_{sqr} is the NSE found using the squared streamflow values, and $\boldsymbol{w}=\{w_1, w_2, w_3\}$ is the corresponding weights for each NSE controlling how much weight is applied to the mean (w_1), low (w_2), and high (w_3) flows.

4.4.2 Information theory

The amount of information, or uncertainty, a random variable contains is its marginal entropy (Shannon 1948). Marginal entropy is analogous to variance and is defined mathematically as:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{4-7}$$

where *H(X)* is the marginal entropy of a random variable *X*, *n* is the number of bins being used to represent the random variable, and p_i is the probability of the value being in bin *i*. The data quantization method used for has been shown to impact the results of the entropy calculation (Keum and Coulibaly 2017); for simplicity, however, Sturges' method was used (Sturges 1926). The marginal entropy can be expanded to two or more variables as follows (Shannon 1948):

$$H(X_1, X_2, \dots, X_N) = -\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \dots \sum_{i_N=1}^{n_N} p\left(x_{1,i_1}, x_{2,i_2}, \dots, x_{N,i_N}\right) \log\left(p\left(x_{1,i_1}, x_{2,i_2}, \dots, x_{N,i_N}\right)\right) \tag{4-8}$$

where $H(X_1,X_2,\dots,X_N)$ is the joint entropy of *N* random variables, n_1, $n_2, \dots,$ n_N are the number of bins being used to represent the corresponding *N* random variables, and $p(x_{1,i_1}, x_{2,i_2}, \dots, x_{N,i_N})$ is the joint probability of the *N* variables.

The information content that is found in one random variable but not another in a system is defined as the conditional entropy (Shannon 1948; V. P. Singh 1997):

$$H(X|Y) = H(X,Y) - H(Y) \leq H(X) \tag{4-9}$$

where *H(X|Y)* is the conditional entropy between two random variables and *H(X,Y)* is the joint entropy. Knowing the marginal and conditional entropies between random variables the mutual information can be determined. Mutual information is the measure of the shared information, both linear and non-linear dependencies, between two or more random variables and it is analogous to covariance. The two-variable form of mutual information, also known as transinformation, is defined as follows (Shannon 1948; V. P. Singh 1997):

$$T(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = T(Y,X) \tag{4-10}$$

where *T(X,Y)* is the mutual information between two variables or two groups of variables. A generalized measure of dependence between continuous random variables was developed by Linfoot (1957) called the informational coefficient of correlation. This measure can aid in showing the relationship between mutual information and covariance:

$$R_{XY} = \sqrt{1 - \exp\big(-2T(X,Y)\big)} \tag{4-11}$$

where R_{XY} is the informational coefficient of correlation and $T(X,Y)$ is the mutual information (transinformation) between random variables X and Y. Note that for discrete random variables, the modification by Lu (2011) denoted as the L-measure should be used. The informational coefficient of correlation becomes the Pearson correlation coefficient when the two random variables are linearly correlated, continuous, and have a joint bivariate normal distribution. The Pearson correlation coefficient can be calculated as follows:

$$r_{XY} = \frac{\Sigma^{XY}}{\sqrt{\Sigma^{XX}\Sigma^{YY}}} \tag{4-12}$$

where r_{XY} is the Pearson correlation coefficient, Σ^{XY} is the covariance between random variables X and Y, Σ^{XX} is the variance of X, and Σ^{YY} is the variance of Y.

4.4.3 Data assimilation techniques

4.4.3.1 The Ensemble Kalman Filter.

The EnKF, developed by Evensen (1994, 2003), is a technique used for sequential data assimilation. The EnKF uses a Monte Carlo approach for prior and posterior distribution estimation and can be used to update both states and parameters of a model as new observations become available (Moradkhani et al. 2005; Samuel et al. 2014). Ensemble size and hyper-parameter selection are important considerations that can influence the performance of the EnKF (Moradkhani et al. 2005; Thiboult and Anctil 2015), however these are beyond the scope of the work in this manuscript and have instead been chosen based on previous work done in the study area (Leach, Kornelsen, and Coulibaly 2018).

The following steps are taken when performing the update (Moradkhani et al. 2005; Samuel et al. 2014). First the states are propagated forward in time:

$$x_{j,t}^{-} = f\left(x_{j,t-1}^{+}, u_{j,t}, \theta\right) \tag{4-13}$$

where $x_{j,t}^{-}$ is the prior estimate of state variables in the model at time *t* for ensemble member *j*, *f(.)* is the operator within the models that propagate the state variables, $x_{j,t-1}^{+}$ is the posterior estimate of state variables at time *t-1* for ensemble *j*, $u_{j,t}$ is the perturbed forcing data at time *t* for ensemble member *j*, and θ is the vector of parameters. When updating states and parameters, θ is replaced with $\theta_{j,t-1}^{+}$ following the dual updating method of Moradkhani et al. (2005).

The streamflow is then predicted using the prior estimated states as follows:

$$\hat{y}_{j,t} = h\left(x_{j,t}^{-}, \theta\right) + \nu_{j,t} \tag{4-14}$$

where $\hat{y}_{j,t}$ is the simulated streamflow at time *t* for ensemble member *j*, *h(.)* is the hydrologic model and $\nu_{j,t}$ is normally distributed noise at time *t* for ensemble member *j* used to represent the model's structural uncertainty.

Streamflow observations at time *t* are then perturbed to generate an ensemble to represent the uncertainty related to the observations:

$$y_{j,t} = y_t + \eta_{j,t} \tag{4-15}$$

where $y_{j,t}$ is the perturbed streamflow used to update the model states (or parameters) at time *t* for ensemble j, y_t is the observation at time *t*, $\eta_{j,t}$ is the normally distributed noise used for perturbing the streamflow.

The Kalman gain, which determines how much confidence should be put in the new observation, is then calculated as follows:

$$K_t = \Sigma_t^{X\hat{Y}}\left(\Sigma_t^{\hat{Y}\hat{Y}} + \Sigma_t^{YY}\right)^{-1} \tag{4-16}$$

where K_t is the Kalman gain for time *t*, $\Sigma_t^{X\hat{Y}}$ is the background covariance at time *t*, $\Sigma_t^{\hat{Y}\hat{Y}}$ is the variance of the predicted streamflow at time *t*, and Σ_t^{YY} is the variance of the observed streamflow a time *t*.

The updated state (or parameter) can then be found as follows:

$$x_{j,t}^{+} = x_{j,t}^{-} + K_t\left(y_{j,t} - \hat{y}_{j,t}\right) \tag{4-17}$$

where $x_{j,t}^{+}$ is the updated posterior state for ensemble *j* and time *t* and is the optimal estimate given the uncertainties in the model and observations. Equations 4-13 to 4-17 are repeated until the simulation is completed.

4.4.3.2 Distributed Ensemble Kalman Filter.

Data assimilation using the EnKF with semi-distributed hydrologic models has been done previously (Clark et al. 2008; Xie and Zhang 2010). The method described in Clark et al. (2008) was used for distributed data assimilation as a comparison, denoted as DEnKF. The procedure is similar to the EnKF update strategy with some exceptions highlighted in the following.

When observations are available from one or more gauge a $nobs \times 1$ observation vector, $\boldsymbol{Y_t}$, and an $nobs \times nobs$ observation error matrix, $\boldsymbol{\Sigma}_t^{YY} = \rho \times diag(\boldsymbol{Y}_t)$, are created. Where $\boldsymbol{Y}_t = \{y_{t,1}, \ldots, y_{t,nobs}\}$ is the vector of streamflow observations made at time t, $diag(.)$ is the vector diagonalization function, ρ is the observation hyper-parameter (coefficient of variation), and $nobs$ is the number of gauged basins with observations at timestep t. The streamflow observations are then perturbed into a $nobs \times N$ matrix using normally distributed noise. Next, the $nobs \times nobs$ model error covariance matrix, $\boldsymbol{\Sigma}_t^{\hat{Y}\hat{Y}} = cov(\widehat{\boldsymbol{Y}}_t)$, is calculated in which $\widehat{\boldsymbol{Y}}_t$ represents the $nobs \times N$ matrix of ensemble streamflow predictions for time t at the gauged locations.

The next step in this implementation is to cycle through each unobserved model state (or parameter) and update it using data from observed locations, and this is done as follows. A $(nobs + 1) \times N$ matrix is created of the prior state estimates, $\boldsymbol{X}_{tk}^-$, in which the extra row is from a location without observations for time *t* and state *k*. Then the $(nobs + 1) \times nobs$ background covariance matrix is found, $\boldsymbol{\Sigma}_{tk}^{X\hat{Y}} = cov(\boldsymbol{X}_{tk}^-, \widehat{\boldsymbol{Y}}_t)$, and the $(nobs + 1) \times nobs$ Kalman gain can be determined, $\boldsymbol{K_{tk}} = \boldsymbol{\Sigma}_{tk}^{X\hat{Y}} (\boldsymbol{\Sigma}_t^{\hat{Y}\hat{Y}} + \boldsymbol{\Sigma}_t^{YY})^{-1}$. Each ensemble member *j* for each state *k* is then updated using the following:

$$\boldsymbol{X}_{tk}^+ = \boldsymbol{X}_{tk}^- + \boldsymbol{K}_{tk}(\boldsymbol{Y}_t - \widehat{\boldsymbol{Y}}_t) \tag{4-18}$$

where $\boldsymbol{X}_{tk}^+$ is the updated state (or parameter) vector in which the $(nobs + 1)$ entry is the updated vector of state (or parameter) *k* for the ungauged basin at time *t*, $\boldsymbol{X}_{tk}^-$ is the background state (or parameter) vector for state *k* being updated at time *t*, $\boldsymbol{K}_{tk}$ is the Kalman gain used to update state *k* at time *t*, $\boldsymbol{Y}_t$ is the vector of observed streamflows at time *t*, $\widehat{\boldsymbol{Y}}_t$

is the vector of predicted streamflow for the observed basins for at time *t*. This is repeated for each of the state (or parameters) listed in Tables 1 and 2 until the final timestep.

4.4.3.3 Mutual Information Ensemble Kalman Filter.

One of the assumptions used in the derivation of the Kalman gain is that there is no correlation between model and observation error (Evensen 1994). Without that assumption, the Kalman gain and the analysis error covariance can be shown to be:

$$K_t = \left(\Sigma_t^{X\hat{Y}} + \Sigma_t^{XY}\right)\left(\Sigma_t^{\hat{Y}\hat{Y}} + \Sigma_t^{\hat{Y}Y} + \left(\Sigma_t^{\hat{Y}Y}\right)^T + \Sigma_t^{YY}\right)^{-1} \tag{4-19}$$

$$\Sigma_t^{XX+} = \Sigma_t^{XX-} - K_t\left(\Sigma_t^{\hat{Y}X-} + \Sigma_t^{XY-}\right) \tag{4-20}$$

where K_t is the Kalman gain, $\Sigma_t^{X\hat{Y}}$ is the covariance of the states and the predictand, Σ_t^{XY} is the covariance of the states and observations, $\Sigma_t^{\hat{Y}Y}$ is the covariance of the predictand and observations, + (superscript) indicates posterior estimate, and - (superscript) indicates prior estimate. It should be expected that ungauged sub-basins in a semi-distributed model would have a non-zero correlation between the model and observation error (states at ungauged location and observation at gauged location). This is because those ungauged sub-basins were calibrated using observations from the gauged sub-basins; though this relationship may be non-linear/non-Gaussian. Because of this potential non-linearity, mutual information is proposed to be used in place of the cross-covariance between ungauged sub-basin states and gauged sub-basin observations. From some empirical testing, the modified Kalman gain used for the MIEnKF was found to be:

$$K_t = \left(\Sigma_t^{X\hat{Y}} + T(X,Y)/H(X,Y)\right)\left(\Sigma_t^{\hat{Y}\hat{Y}} + T(\hat{Y},Y)/H\left(\hat{Y},Y\right) + \Sigma_t^{YY}\right)^{-1} \tag{4-21}$$

where *T(,)* is the mutual information, *H(,)* is the joint entropy, *X* is the matrix containing states (or parameters) from both gauged and ungauged basins, *T* (superscript) is the transpose operator, *Y* is the matrix containing observed streamflow from one or more gauged locations, and $\hat{Y}$ is the matrix containing predicted streamflow at one or more gauged locations.

4.4.4 Method comparison and evaluation

The Lorenz model was used to evaluate the performance of the proposed method against previously developed EnKF implementations. To do this, multiple tests were run which evaluated the performance of the Lorenz system with the EnKF using the standard Kalman gain (Eq. 4-16), the Kalman gain considering correlation between model and measurement error (Eq. 4-19), the DEnKF implementation, and the modified Kalman gain using mutual information (Eq. 4-21). To set up an analogous test for updating ungauged basins using observations from gauged basins, the Lorenz states were used to update each other (x update y and z, y update x and z, z update x and y). These tests were run 100 times and then evaluated using the Root Mean Square Error (RMSE). The RMSE is:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(Y - \bar{\hat{Y}}\right)^2} \tag{4-22}$$

where Y is the measured (observed) value, $\bar{\hat{Y}}$ is the ensemble mean of the prediction, t is the model step, and T is the total number of steps. The RMSE values range from zero to infinity, with zero being a perfect fit.

To test the performance of each data assimilation method on the hydrologic models, several leave one/multiple out cross-validations were run and compared using the DRB HyMod model. These leave one/multiple out comparisons allowed for evaluating how well the proposed method could be used to update one or multiple ungauged locations efficiently. Each test was run 20 times using an ensemble size of 200, and the average performance of them was compared. To test the scalability of the MIEnKF method to more discretized models, it was also evaluated using the Raven models of DRB and HRB. For the tests run using the HyMod model, it was effective to update every timestep that observations were available; however, due to the high computational cost when using a distributed model, an optimal updating strategy was determined for Raven. The performance of Raven was tested against different ensemble sizes, model time steps, and update frequencies on the DRB Raven model to determine the best update strategy. The best performing options chosen for the DRB were then applied to the HRB. The evaluation using the Raven model also allows for determining if the method is useful for other models and at different time steps (since HyMod was run as daily and Raven run as hourly).

For the Raven model, the ensemble means from each data assimilation scheme were evaluated using the Peak Flow Criteria (PFC) (Coulibaly, Anctil, and Bobée 2001), and Kling-Gupta Efficiency (KGE) (Gupta et al. 2009). The PFC was chosen to highlight how well the methods could improve peak flow prediction, since the peak flows are more pronounced at the hourly timestep, and the KGE was chosen to show how well, in general the methods could improve the modeling of the hydrographs. Both metrics are defined as follows:

$$PFC = \frac{\left(\sum_{p=1}^{T_p}\left(Q_p - \bar{\bar{Q}}_p\right)^2 Q_p^2\right)^{1/4}}{\left(\sum_{p=1}^{T_p} Q_p^2\right)^{1/2}} \tag{4-23}$$

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\sigma_{\bar{\hat{Y}}}/\sigma_Y - 1\right)^2 + \left(\mu_{\bar{\hat{Y}}}/\mu_Y - 1\right)^2} \tag{4-24}$$

where Q_p is the observed runoff peak, $\bar{\bar{Q}}_p$ is the ensemble mean of the simulated runoff peak, T_p is the number of observed peaks above a threshold, and $p \in \{1, \dots, T_p\}$. The threshold value for peak flows was found by identifying the peaks, then taking one-third of the mean of those peaks. The KGE is found using the Pearson correlation coefficient, r, the relative variability, $\sigma_{\bar{\hat{Y}}}/\sigma_Y$, and the mean ratio, $\mu_{\bar{\hat{Y}}}/\mu_Y$, between the ensemble mean of the simulated runoff and the observed runoff. Since an essential consideration in urban and flood-prone environments is the peak flow values, the PFC will be used to evaluate how well the data assimilation methods can improve the peak flow simulation in the ungauged basins. The KGE will be used to evaluate the general performance of the data assimilation methods, specifically how effective they are at improving performance in ungauged basins. The PFC values can range from zero to infinity with zero being optimal, and the KGE can range from negative infinity to one with one being optimal.

4.5. Results and discussions

4.5.1 Comparison of data assimilation methods in a chaotic system

The Lorenz 63 model performance using the data assimilation methods are summarized in Table 4-3. From these results, we can see that when assuming there is a linear correlation between model and measurement error using the EnKF, the performance is worse. Additionally, these results show that using observations of one state to update another can reduce performance. This reduction in performance can, however, be dampened when considering model and measurement error to be correlated. From these results, we also see that the MIEnKF can better update variables when their direct observations are not available. The MIEnKF can also perform comparably well to the EnKF when all observations are available, and in general outperforms the EnKF when observations are not available; both in cases when the measurement and model error are considered linearly correlated or not. This suggests that replacing the cross-covariance with the entropy terms allows for better transfer of information from measured to unmeasured states. Although using the Lorenz 63 model is not exactly the same as when considering gauged and ungauged locations in a watershed, these results can be considered somewhat analogous to that kind of scenario.

Table 4-3: Mean and standard deviation of RMSE for 100 runs of Lorenz 63 model with the EnKF and the MIEnKF. Considered cases where model and measurement error are correlated or not (C or noC), and considering when measurements are available or not (noUG meaning all are available, XY meaning that XY are unavailable, etc.). DEnKF values are (nearly) identical to the EnKF ones for the Lorenz 63 model, so they were omitted.

Test options	$\mu_X \pm \sigma_X$	$\mu_Y \pm \sigma_Y$	$\mu_Z \pm \sigma_Z$
EnKF_noC_noUG	0.58 ± 0.03	0.81 ± 0.05	0.87 ± 0.04
EnKF_noC_XY	12.70 ± 0.07	12.40 ± 0.06	6.77 ± 0.08
EnKF_noC_XZ	5.95 ± 0.18	11.22 ± 0.21	16.59 ± 0.23
EnKF_noC_YZ	7.57 ± 0.20	13.44 ± 0.25	16.93 ± 0.25
EnKF_C_noUG	0.63 ± 0.03	0.87 ± 0.05	0.91 ± 0.04
EnKF_C_XY	17.68 ± 0.24	15.29 ± 0.20	9.89 ± 0.20
EnKF_C_XZ	4.40 ± 0.19	9.66 ± 0.27	18.98 ± 0.22
EnKF_C_YZ	5.99 ± 0.19	11.84 ± 0.26	19.16 ± 0.19
MIEnKF_C_noUG	0.64 ± 0.03	0.86 ± 0.05	0.91 ± 0.04
MIEnKF_C_XY	0.66 ± 0.03	0.87 ± 0.05	0.93 ± 0.04
MIEnKF_C_XZ	4.09 ± 8.78	1.38 ± 0.24	21.15 ± 17.85
MIEnKF_C_YZ	0.87 ± 0.03	3.07 ± 5.03	19.80 ± 19.80

Note: Extreme outliers removed (RMSE>100).

A comparison of the EnKF and the MIEnKF, when using the measurement of the Z state to update the X and Y states, can be seen in Figure 4-2. Here we see that the standard EnKF is not able to correct for the error added to the Lorenz 63 model when using measured states to update unmeasured states. This is not the case for the MIEnKF; it can model the system much better when information is missing. We also see from the analysis error covariance plot in Figure 4-2d that the MIEnKF variance generally stays lower than the EnKF with correlation considered, with occasional spikes. This implies that including the entropy measures is useful in updating systems when there are non-linear dependencies present.

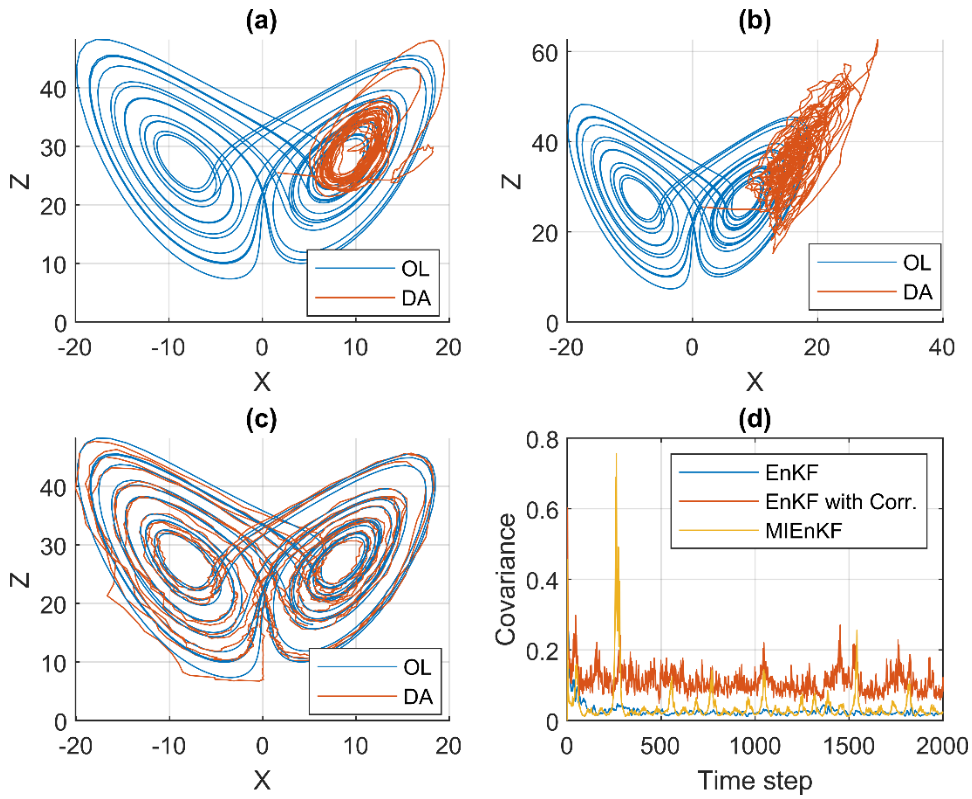

Figure 4-2: Comparison of EnKF and MIEnKF results when the measurement of Z is used to update X and Y with the open-loop (OL) results in blue, and the data assimilation (DA) results in orange (for a, b, and c). (a) is the results of the EnKF with correlation of model and measurement error assumed to be zero, (b) is the EnKF with non-zero correlation, (c) is the MIEnKF results, and (d) is a comparison of the analysis error covariance of X.

4.5.2 Simple hydrological model

4.5.2.1 Calibration and Validation of HyMod model.

The HyMod model calibration and validation results are provided in Table 4-4. HyMod was calibrated using observations from 2007-01-01 to 2012-12-31 and validated using observations from 2013-01-01 to 2015-12-31. The upstream sub-basins, gauges 02HC056 and 02HC005, were calibrated first in the case of the semi-distributed HyMod model. All three sub-basins have similar performances in the calibration and validation periods. However, gauge 02HC005 (basin 2) performs significantly poorer. This lower performance in basin 2 is likely due to the G. Ross Lord Dam which is upstream of the gauge since its operations are not adequately modeled by HyMod.

Table 4-4: Calibration and validation results for the semi-distributed HyMod model setup for the Don River basin using daily streamflow. The optimization function used was the NVE with the values used to control weighting on medium, low, and high flows being *w={0.3,0.2,0.5}*.

		Calibration period 2007-2012			Validation Period 2013-2015		
Sub-basin	Gauge	NVE	VE	NSE	NVE	VE	NSE
1	02HC056	0.54	0.01	0.60	0.55	0.03	0.55
2	02HC005	0.43	0.18	0.46	0.39	0.19	0.43
3	02HC024*	0.62	0.01	0.66	0.56	0.05	0.60

Note: * Indicates most downstream gauge of the semi-distributed model.

4.5.2.2 Data Assimilation with HyMod.

Several implementations of the MIEnKF were evaluated to identify the best way to incorporate the mutual information between gauged and ungauged locations for improving state and parameter updating in ungauged locations. From those tests, the modified Kalman gain shown in Eq. 4-21 proved to be the best. The EnKF, DEnKF, and MIEnKF were then all evaluated using multiple assimilation schemes. These assimilation schemes included all combinations of the three daily streamflow gauges being used or not used for updating. Since the focus of these analyses is on the ability to update ungauged locations more effectively, the performance of each 'ungauged' basin was compared for each assimilation scheme, shown in Tables 4-5 and 4-6. From these results, it is apparent that the selected MIEnKF methods perform better than both the EnKF and DEnKF methods at updating the states of ungauged basins. The improvement to state updating by incorporating mutual information in the EnKF can likely be attributed to the similarity in the ensemble distributions of the states and observations. More specifically, the cumulative distribution of the ensemble for each state variable from an ungauged sub-basins is similar to that of the cumulative distribution of the gauged sub-basins' streamflow (Figure 4-3a). For mutual information there is more information shared between random variables when their distributions are similar; thus the more information is available for the update. However, the ensemble distributions for each parameter are not very similar between sub-basins, which leads to little or no added information (Figure 4-3b).

Table 4-5: NVE for DA methods using observations from sub-basins 1 and 2 for updating HyMod. The *italicized* performances are those from the 'ungauged' basin during the run.

	Sub-basin	Gauge	EnKF	DEnKF	MIEnKF
State	1	02HC056	0.43	0.28	0.59
	2	02HC005	0.35	0.24	0.38
	3	02HC024	0.43	0.41	0.51
Dual	1	02HC056	0.45	-0.27	0.60
	2	02HC005	0.31	-0.23	0.03
	3	02HC024	0.40	0.18	0.37

Table 4-6: NVE for DA methods using observations from sub-basin 3 for updating HyMod. The *italicized* performances are those from the 'ungauged' basins during the run.

	Sub-basin	Gauge	EnKF	DEnKF	MIEnKF
State	1	02HC056	0.43	0.45	0.55
	2	02HC005	0.34	0.34	0.38
	3	02HC024	0.43	0.43	0.52
Dual	1	02HC056	-1.89	-8.69	-1.82
	2	02HC005	-0.34	-0.29	-0.34
	3	02HC024	0.48	0.43	0.47

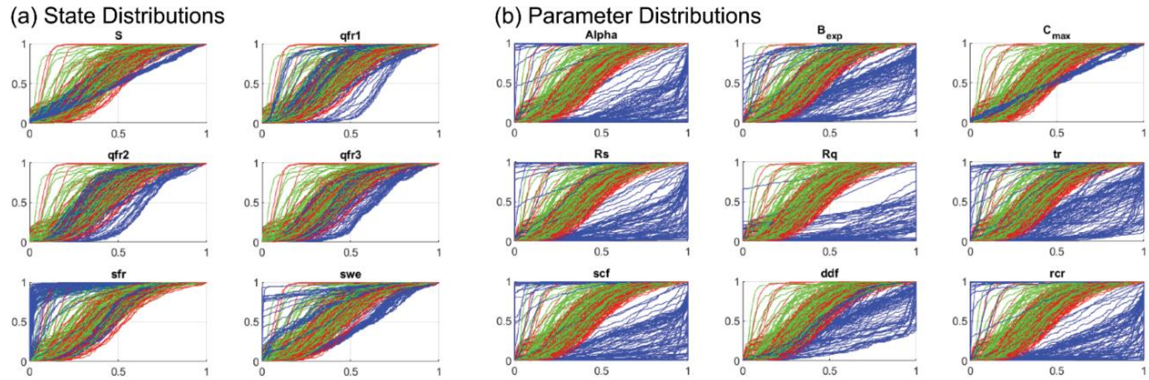

Figure 4-3: Comparison of the normalized ensemble streamflow cumulative distribution functions (CDFs) from 100 timesteps with (a) the normalized ensemble CDFs of states listed in Table 4-1 and (b) the normalized ensemble CDFs of parameters listed in Table 4-1. The green lines are the ensemble streamflow CDFs from sub-basin 1, the red lines are the ensemble streamflow CDFs from sub-basin 2, and the blue are the ensemble CDFs for either states or parameters. Values were normalized for easier comparison since mutual information depends on distribution shape, not magnitude.

Additionally, selected model runs of the MIEnKF hydrographs are illustrated in Figure 4-4. These results are meant to show how well the ensemble mean can match the observations despite the data assimilation update being made using observations from a different sub-basin. Here Figure 4-4a shows the results from when the states from sub-basin 1 are updated using streamflow from sub-basin 3, Figure 4-4b shows the results from when sub-basin 2 states are updated using streamflow from sub-basin 3, and Figure 4-4c shows the results from sub-basin 3 states are updated using streamflow from both sub-basins 1 and 2. We concluded that the ensemble results can reasonably match the observed values when updating using observations from other sub-basins. These results agree with those of the Lorenz63 model and strengthen the justification for incorporating mutual information into the EnKF to better update ungauged locations. To further test the MIEnKF in its applicability to different models, as well as its scalability, it will also be used with Raven models of both the DRB and the HRB.

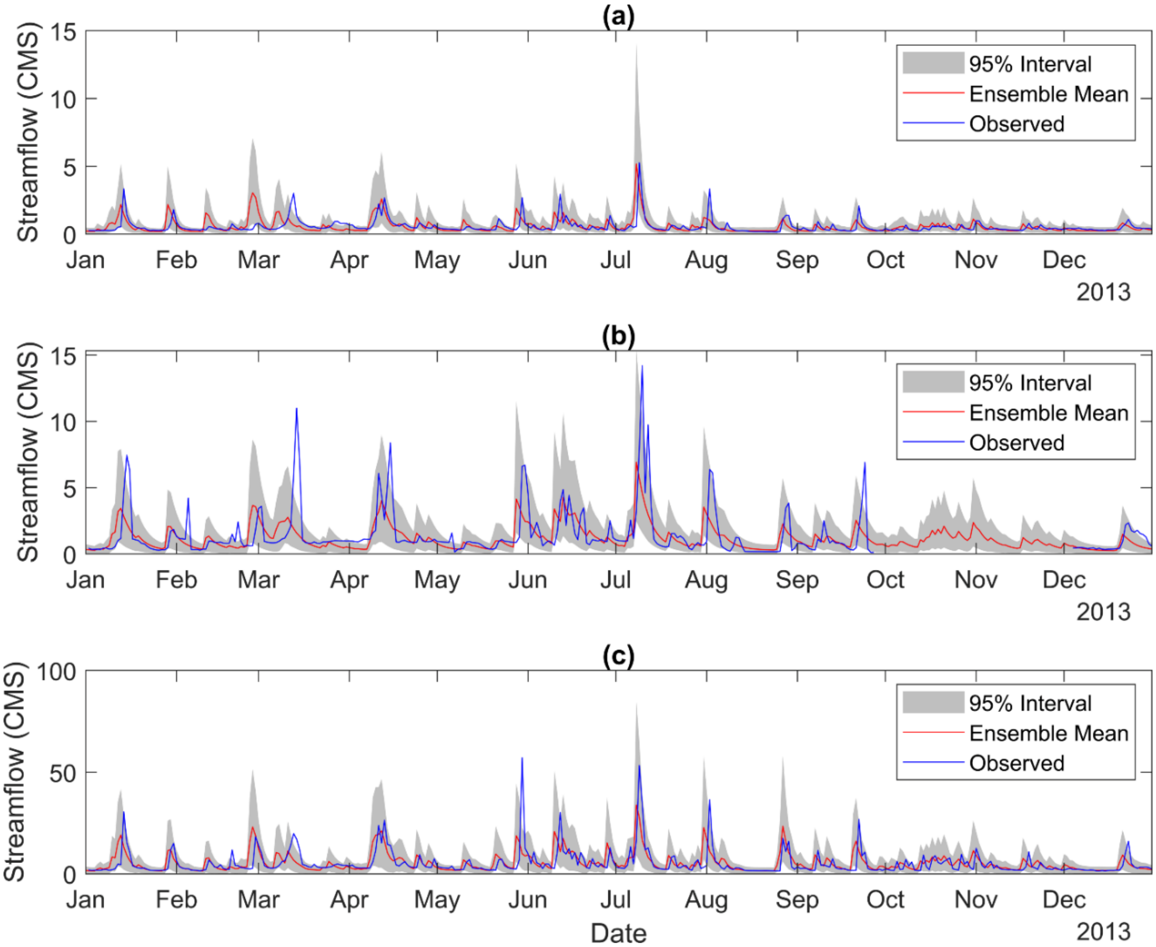

Figure 4-4: Comparison of ensemble mean and 95% confidence interval of the MIEnKF method against the observed streamflow hydrographs for (a) sub-basin 1 states being updated using observations from sub-basin 3, (b) sub-basin 2 states being updated using observations from sub-basin 3, and (c) sub-basin 3 states being updated using observations from sub-basins 1 and 2.

4.5.3 Complex hydrological model

4.5.3.1 Calibration and validation of Raven Model.

The Raven model calibration and validation results are provided in Tables 4-7 and 4-8. Raven was calibrated with observations from 2009-01-01 to 2012-12-31 and validated from 2013-01-01 to 2015-12-31. The differences in calibration periods between the HyMod and Raven models were simply due to the availability of hourly and daily streamflow observations.

Table 4-7: Calibration and validation results for the semi-distributed semi-physically based Raven model of the Don River basin using hourly streamflow. Gauges that start with '02HC' are ECCC gauges, and those that start with 'HY' are TRCA gauges; some gauges may be used by both agencies.

		Calibration period 2009-2012			Validation Period 2013-2015		
Sub-basin	Gauge	NVE	VE	NSE	NVE	VE	NSE
12	02HC056	0.36	0.53	0.41	0.09	0.59	0.19
30	HY017	0.34	0.64	0.47	0.21	0.68	0.40
46	HY018 (02HC005)	0.33	0.09	0.43	-2.27	0.19	-0.91
56	HY062	0.22	0.32	0.38	-1.21	0.30	-0.07
58	HY019 (02HC024)*	0.45	0.27	0.59	0.03	0.25	0.14

Note: * Indicates most downstream gauge of the semi-distributed model.

Table 4-8: Calibration and validation results for the semi-distributed semi-physically based Raven model of the Humber River basin using hourly streamflow. Gauges that start with '02HC' are ECCC gauges, and those that start with 'HY' are TRCA gauges; some gauges may be used by both agencies.

		Calibration period 2009-2012			Validation Period 2013-2015		
Sub-basin	Gauge	NVE	VE	NSE	NVE	VE	NSE
32	02HC047	-0.16	0.67	-0.31	-0.30	0.65	-0.56
57	02HC032	0.01	0.66	0.05	0.18	0.61	-0.16
59	02HC051	-	-	-	-0.44	0.72	-0.59
69	02HC023	0.08	0.64	0.07	-0.09	0.63	-0.11
86	HY054	0.11	0.55	0.26	0.02	0.69	0.05
91	02HC025	0.01	0.61	-0.07	-0.08	0.60	-0.20
104	02HC009	0.11	0.62	0.14	0.04	0.62	0.02
111	HY053	-0.08	0.79	0.01	-1.40	0.50	0.08
119	HY035	0.16	0.65	0.25	-0.41	0.61	-0.04
127	02HC031	0.18	0.59	0.32	0.15	0.55	0.21
145	02HC003*	0.19	0.58	0.28	0.34	0.52	0.51
148	02HC027*	0.43	0.51	0.62	0.37	0.49	0.54

Note: * Indicates the two most downstream gauges of the semi-distributed model.

Both the DRB and HRB Raven models were calibrated using hourly streamflow data. The NVE objective function weights were chosen so that peak and mean flows would be prioritized ($w=\{0.5,0,0.5\}$). This weighting of the NVE was chosen since including low flows provided worse performance in both the calibration and validation period. The HRB Raven model is shown to have similar performance for both the calibration and validation periods, while the DRB Raven model generally performs worse during the validation period. Additionally, the DRB Raven model is also shown to have poorer performance than the DRB HyMod model. The reason for this difference is due to the average daily streamflow being fairly consistent between calibration and validation periods while the hourly streamflow is much higher during the validation period, as illustrated in Figure 5. In general, the model validation results for the DRB are considered to be acceptable when considering these points.

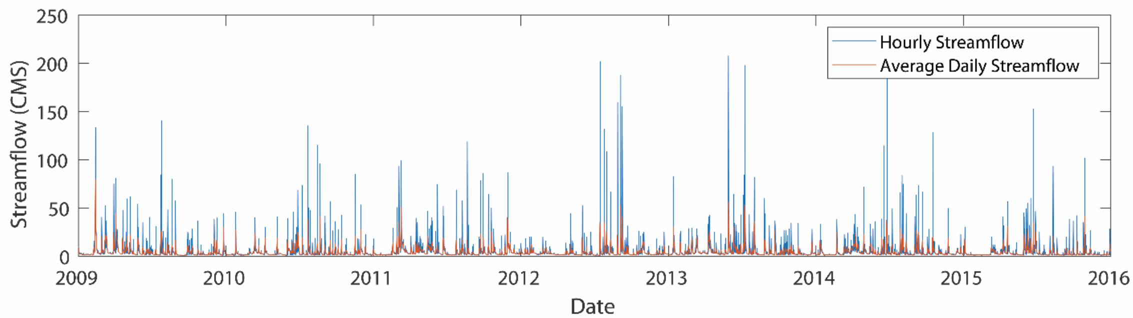

Figure 4-5: Comparison of hourly and average daily streamflow values for gauge 02HC024(HY019) in the DRB.

4.5.3.2 Optimal Timestep and Update Strategy for Raven Models.

Distributed and semi-distributed models generally have larger computational time requirements compared with those of lumped models; this computation time is further increased as the temporal resolution increases. When these factors are considered in an ensemble data assimilation framework, the computation time of assimilation experiments can quickly increase. It is desirable to determine an optimal updating strategy because of this fact. Due to the setup being used for the semi-distributed HyMod model, three sub-basins and daily timesteps, it was feasible to update whenever observations were available. For the semi-distributed Raven models, an optimal update strategy had to be determined.

In order to determine the best data assimilation update strategy, different update frequencies, model time steps, and ensemble sizes were compared. The results of the different ensemble sizes showed little improvement in performance compared to the increased computation time. However, the update frequency and model timestep did affect the model performance, as illustrated in Figure 4-6. From these results, it seems that the optimal updating strategy for the Raven models should be to perform the update every 24 hours while the model should run with 1-hour time steps. This update strategy was then used for both the DRB and the HRB Raven models.

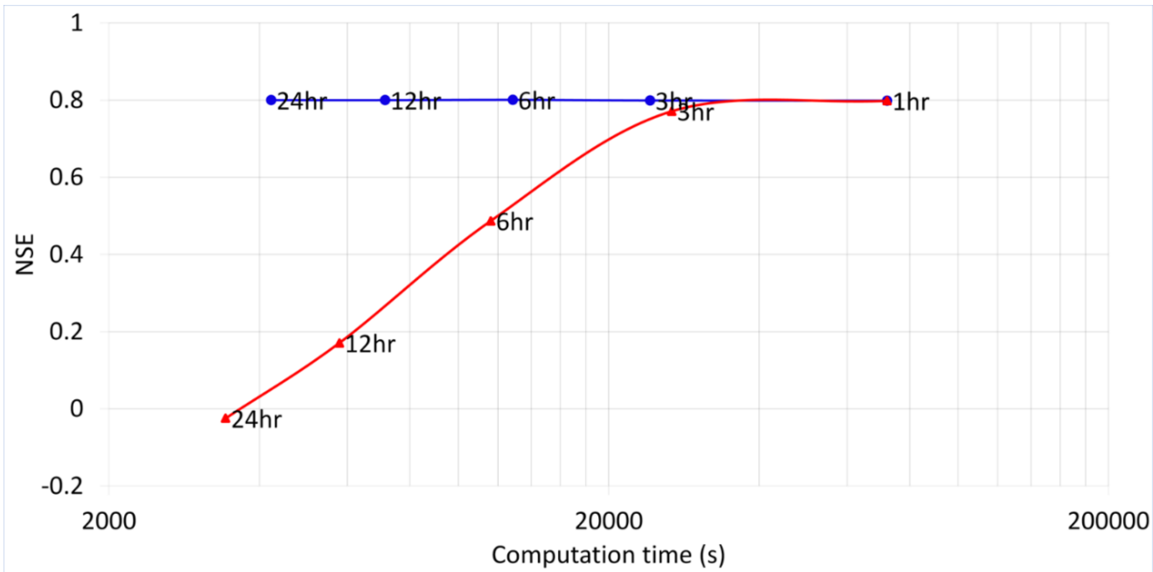

Figure 4-6: Comparison of update frequency and model timesteps to computation time for the DRB Raven model; run on a 2.1GHz 12-Core Processor. The point labels indicate the update frequency, the blue line shows the model performance with 1-hour time steps, and the red line shows the model performance when the timestep is equal to the update frequency.

4.5.3.3 Data Assimilation with Raven Models.

To further evaluate the chosen MIEnKF implementation, it was used to update the Raven models of the DRB and HRB. This was done to show the applicability of the method, mainly that it is scalable to largely discretized models, it can be used for both urban and semi-urban watersheds, and that it can be used with multiple models. Just as with the HyMod tests, leave 1-to-N gauges out tests were performed, and the performances of the EnKF, DEnKF, and MIEnKF methods were compared. Shown in Table 4-9 is a summary of the DRB and HRB Raven models' performances, specifically the percentage of ungauged sub-basins that have improved performance compared to the open-loop. From these results, it seems like the methods have similar results to those of the semi-distributed HyMod model.

Additionally, none of the methods perform particularly well when updating parameters and states. Focusing on the state updating results for the DRB and directly comparing the assimilation methods, the MIEnKF provides the better KGE performance for 52% of the ungauged sub-basins, while the EnKF and DEnKF provide better performance in 23% and 25% of the ungauged sub-basins respectively. The PFC provides similar results with the MIEnKF providing better performance for 69% of the sub-basins.

Table 4-9: Percentage of ungauged sub-basins, updated using data from gauged basins, which have the same or better performance than the open-loop run. Percentages are found from the leave 1-to-N gauges out tests. For example, the DRB has 30 possible scenarios in which one or more gauged basins can be considered ungauged, of those tests, there would be 75 'ungauged' basins, and if all 75 has better performance than the validation run, it would be indicated as 100%.

		State Updating			Dual Updating		
	Metric	EnKF	DEnKF	MIEnKF	EnKF	DEnKF	MIEnKF
DRB	KGE	40%	41%	57%	23%	31%	20%
	PFC	20%	20%	20%	20%	20%	1%
HRB	KGE	82%	79%	93%	13%	15%	0%
	PFC	18%	18%	18%	27%	17%	19%

Although it is not the focus of the paper, Table 4-10 provides the KGE performances when all gauges are used for updating in the DRB Raven model. Here, results suggest that the MIEnKF method can improve model performance of gauged basins as well when using it to update model states. However, as with the ungauged sub-basins, the MIEnKF method performs poorly at updating the model parameters.

Table 4-10: KGE performance for state and dual updating with assimilation methods in the DRB Raven model. Bolded values indicate the best performing state or dual assimilation method for each gauge if it also performs better than the OL validation.

Method		HC056	HY017	HY018	HY062	HY019
Open Loop	Validation	0.24	0.09	0.20	0.45	0.51
	EnKF	0.24	0.18	0.00	0.41	0.69
State	DEnKF	0.10	0.15	0.05	0.42	0.66
	MIEnKF	0.40	0.24	-0.09	0.41	0.77
	EnKF	-0.30	0.46	-1.12	-0.12	0.50
Dual	DEnKF	-0.70	0.14	-0.41	0.12	0.76
	MIEnKF	-1.81	0.26	-0.98	-1.07	0.29

The results indicate that for state updating, the MIEnKF is the superior method and that it performs as well or better than the open-loop more often than either the EnKF or DEnKF for both the KGE and PFC performance criteria (Table 4-9). Additionally, these results indicate that for parameter updating, all methods perform poorly. The difference in performance between the state and the parameter updating is likely due to parameters being more spatially variable in these sub-basins. More specifically, when updating only states, the values are easily related between the gauged and ungauged sub-basins even when using data from a rural location to update an urban one (Figure 4-3a). However, when attempting to use data from an urban sub-basin to update the parameters of a rural sub-basin or vice versa, the values are not well related, and therefore the performance is not improved by any of the data assimilation methods (Figure 4-3b).

This work suggests that the MIEnKF method can use observation information from gauged sub-basins to update the states of ungauged sub-basins better than either EnKF or DEnKF, and can potentially act as a pseudo-replacement for the unknown spatial covariances between the locations. Therefore, when using data assimilation to update states in an ungauged location, the MIEnKF will provide improved updates over the original EnKF implementation.

4.6 Conclusions

The goal of this work was to utilize the mutual information between gauged and ungauged locations to provide an improved update for states and parameters when using data assimilation. To test this, a method was developed which utilized the entropy metric of mutual information, called the MIEnKF. The MIEnKF was first tested using the Lorenz 63 model, and then on a small semi-distributed HyMod model. The results of the Lorenz 63 model with MIEnKF implied that it could better update the system when non-linear dependencies are present. It was also shown to be able to update the states of an ungauged sub-basin better when compared to the EnKF or DEnKF. Next, the scalability of the method was evaluated by testing it on larger semi-distributed models built using the Raven hydrologic modeling framework. Similar performance was shown with the larger Raven model implementations as those from the HyMod model. Additionally, this showed that the MIEnKF method was usable in different hydrologic models as well as being valid for both urban and semi-urban watersheds. Finally, it was shown that the proposed method was particularly good at updating the states of ungauged basins, while it was not good at updating parameters, especially when land use was significantly different.

Future work will be aimed at optimally transferring information about basin parameters from gauged to ungauged sub-basins, which would be useful for both data assimilation and optimization of semi-distributed and distributed hydrologic models.

Additionally, future work will be used to solidify better the theory behind the improvements gained by using MIEnKF.

4.7 Acknowledgments

This research was supported by the Natural Science and Engineering Research Council (NSERC) of Canada, grant NSERC Canadian FloodNet (NETGP-451456). The authors are grateful to the Toronto Region Conservation Authority (TRCA), Environment and Climate Change Canada (ECCC), and Water Survey of Canada (WSC) for providing the data. We would also like to thank Dr. James Craig for his development and support of Raven.

4.8 Literature Cited

Abaza, Mabrouk, François Anctil, Vincent Fortin, and Richard Turcotte. 2015. "Exploration of Sequential Streamflow Assimilation in Snow Dominated Watersheds." *Advances in Water Resources* 80 (June). Elsevier Ltd: 79–89. doi:10.1016/j.advwatres.2015.03.011.

Abbaszadeh, Peyman, Hamid Moradkhani, and Hongxiang Yan. 2018. "Enhancing Hydrologic Data Assimilation by Evolutionary Particle Filter and Markov Chain Monte Carlo." *Advances in Water Resources* 111 (April 2017). Elsevier: 192–204. doi:10.1016/j.advwatres.2017.11.011.

Agriculture and Agri-Food Canada. 2015. "Land Use 1990, 2000 and 2010 (LU1990, LU2000, LU2010)." https://open.canada.ca/data/en/dataset/18e3ef1a-497c-40c6-8326-aac1a34a0dec.

Awol, Frezer Seid, Paulin Coulibaly, and Bryan A. Tolson. 2018. "Event-Based Model Calibration Approaches for Selecting Representative Distributed Parameters in Semi-Urban Watersheds." *Advances in Water Resources* 118 (November 2017). Elsevier Ltd: 12–27. doi:10.1016/j.advwatres.2018.05.013.

Boyle, Douglas Patrick. 2001. "Multicriteria Calibration of Hydrologic Models." University of Arizona. http://hdl.handle.net/10150/290657.

Carpenter, Theresa M., and Konstantine P. Georgakakos. 2006. "Intercomparison of Lumped versus Distributed Hydrologic Model Ensemble Simulations on Operational Forecast Scales." *Journal of Hydrology* 329 (1–2): 174–85. doi:10.1016/j.jhydrol.2006.02.013.

Clark, Martyn P., David E. Rupp, Ross A. Woods, Xiaogu Zheng, Richard P. Ibbitt, Andrew G. Slater, Jochen Schmidt, and Michael J. Uddstrom. 2008. "Hydrological Data Assimilation with the Ensemble Kalman Filter: Use of Streamflow Observations to Update States in a Distributed Hydrological Model." *Advances in Water Resources* 31 (10): 1309–24. doi:10.1016/j.advwatres.2008.06.005.

Coulibaly, Paulin, François Anctil, and Bernard Bobée. 2001. "Multivariate Reservoir Inflow Forecasting Using Temporal Neural Networks." *Journal of Hydrologic Engineering* 6 (5): 367–76. doi:10.1061/(ASCE)1084-0699(2001)6:5(367).

Craig, James R., and the Raven Development Team. 2018. "Raven User's and Developer's Manual (Version 2.8)." http://raven.uwaterloo.ca/.

Dumedah, Gift, and Paulin Coulibaly. 2012. "Integration of Evolutionary Based Assimilation into Kalman-Type Methods for Streamflow Simulations in Ungauged

Watersheds." *Journal of Hydrology* 475 (December). Elsevier B.V.: 428–40. doi:10.1016/j.jhydrol.2012.10.033.

Environment and Climate Change Canada. 2017. "Canadian Climate Normals." http://climate.weather.gc.ca/climate_normals/.

Evensen, Geir. 1994. "Sequential Data Assimilation with a Nonlinear Quasi-Geostrophic Model Using Monte Carlo Methods to Forecast Error Statistics." *Journal of Geophysical Research* 99 (C5): 10143. doi:10.1029/94JC00572.

———. 2003. "The Ensemble Kalman Filter: Theoretical Formulation and Practical Implementation." *Ocean Dynamics* 53 (4): 343–67. doi:10.1007/s10236-003-0036-9.

Ganguli, Poulomi, and Paulin Coulibaly. 2017. "Does Nonstationarity in Rainfall Require Nonstationary Intensity–Duration–Frequency Curves?" *Hydrology and Earth System Sciences* 21 (12): 6461–83. doi:10.5194/hess-21-6461-2017.

Government of Canada. 2018. "Canadian Soil Information Service." http://sis.agr.gc.ca/cansis/index.html.

Gupta, Hoshin V., Harald Kling, Koray K. Yilmaz, and Guillermo F. Martinez. 2009. "Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modelling." *Journal of Hydrology* 377 (1–2). Elsevier B.V.: 80–91. doi:10.1016/j.jhydrol.2009.08.003.

Keum, Jongho, and Paulin Coulibaly. 2017. "Information Theory-Based Decision Support System for Integrated Design of Multivariable Hydrometric Networks." *Water Resources Research* 53 (7): 6239–59. doi:10.1002/2016WR019981.

Leach, James M., Kurt C. Kornelsen, and Paulin Coulibaly. 2018. “Assimilation of Near-Real Time Data Products into Models of an Urban Basin.” *Journal of Hydrology* 563 (August). Elsevier: 51–64. doi:10.1016/j.jhydrol.2018.05.064.

Linfoot, E.H. 1957. “An Informational Measure of Correlation.” *Information and Control* 1 (1): 85–89. doi:10.1016/S0019-9958(57)90116-X.

Liu, Y., A. H. Weerts, M. Clark, H.-J. Hendricks Franssen, S. Kumar, H. Moradkhani, D.-J. Seo, et al. 2012. “Advancing Data Assimilation in Operational Hydrologic Forecasting: Progresses, Challenges, and Emerging Opportunities.” *Hydrology and Earth System Sciences* 16 (10): 3863–87. doi:10.5194/hess-16-3863-2012.

Lorenz, Edward N. 1963. “Deterministic Nonperiodic Flow.” *Journal of the Atmospheric Sciences* 20 (2): 130–41. doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Lu, Shan. 2011. “Measuring Dependence via Mutual Information (MSc Thesis).” Kingston, Ontario: Queen’s University. http://www.mast.queensu.ca/~web/Papers/shan-lu-msc11.pdf.

Mishra, Ashok K., and Paulin Coulibaly. 2009. “Developments in Hydrometric Network Design: A Review.” *Reviews of Geophysics* 47 (2): RG2001. doi:10.1029/2007RG000243.

———. 2010. “Hydrometric Network Evaluation for Canadian Watersheds.” *Journal of Hydrology* 380 (3–4): 420–37. doi:10.1016/j.jhydrol.2009.11.015.

Monteith, J. L. 1965. “Evaporation and Environment.” *Symposia of the Society for Experimental Biology* 19: 205–34. http://www.ncbi.nlm.nih.gov/pubmed/5321565.

Moradkhani, Hamid, Soroosh Sorooshian, Hoshin V. Gupta, and Paul R. Houser. 2005. “Dual State–Parameter Estimation of Hydrological Models Using Ensemble Kalman Filter.” *Advances in Water Resources* 28 (2): 135–47. doi:10.1016/j.advwatres.2004.09.002.

Nash, J.E., and J.V. Sutcliffe. 1970. “River Flow Forecasting through Conceptual Models Part I — A Discussion of Principles.” *Journal of Hydrology* 10 (3): 282–90. doi:10.1016/0022-1694(70)90255-6.

Natural Resources Canada. 2015. “Canadian Digital Elevation Model (CDEM).” https://open.canada.ca/data/en/dataset/7f245e4d-76c2-4caa-951a-45d1d2051333.

Nearing, Grey, Soni Yatheendradas, Wade Crow, Xiwu Zhan, Jicheng Liu, and Fan Chen. 2018. “The Efficiency of Data Assimilation.” *Water Resources Research* 54 (9): 6374–92. doi:10.1029/2017WR020991.

Olsson, J. 1995. “Limits and Characteristics of the Multifractal Behaviour of a High-Resolution Rainfall Time Series.” *Nonlinear Processes in Geophysics* 2 (1): 23–29. doi:10.5194/npg-2-23-1995.

———. 1998. “Evaluation of a Scaling Cascade Model for Temporal Rain- Fall Disaggregation.” *Hydrology and Earth System Sciences* 2 (1): 19–30. doi:10.5194/hess-2-19-1998.

Pathiraja, S., L. Marshall, A. Sharma, and H. Moradkhani. 2016. “Hydrologic Modeling in Dynamic Catchments: A Data Assimilation Approach.” *Water Resources Research* 52 (5): 3350–72. doi:10.1002/2015WR017192.

Rakovec, O., A. H. Weerts, P. Hazenberg, P. J. J. F. Torfs, and R. Uijlenhoet. 2012. "State Updating of a Distributed Hydrological Model with Ensemble Kalman Filtering: Effects of Updating Frequency and Observation Network Density on Forecast Accuracy." *Hydrology and Earth System Sciences* 16 (9): 3435–49. doi:10.5194/hess-16-3435-2012.

Rasmussen, J., H. Madsen, K. H. Jensen, and J. C. Refsgaard. 2015. "Data Assimilation in Integrated Hydrological Modeling Using Ensemble Kalman Filtering: Evaluating the Effect of Ensemble Size and Localization on Filter Performance." *Hydrology and Earth System Sciences* 19 (7): 2999–3013. doi:10.5194/hess-19-2999-2015.

Razavi, Tara, and Paulin Coulibaly. 2013. "Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods." *Journal of Hydrologic Engineering* 18 (8): 958–75. doi:10.1061/(ASCE)HE.1943-5584.0000690.

Reichle, Rolf H. 2008. "Data Assimilation Methods in the Earth Sciences." *Advances in Water Resources* 31 (11). Elsevier Ltd: 1411–18. doi:10.1016/j.advwatres.2008.01.001.

Samuel, Jos, Paulin Coulibaly, Gift Dumedah, and Hamid Moradkhani. 2014. "Assessing Model State and Forecasts Variation in Hydrologic Data Assimilation." *Journal of Hydrology* 513 (May). Elsevier B.V.: 127–41. doi:10.1016/j.jhydrol.2014.03.048.

Samuel, Jos, Paulin Coulibaly, and Robert a. Metcalfe. 2012. "Identification of Rainfall-Runoff Model for Improved Baseflow Estimation in Ungauged Basins." *Hydrological Processes* 26 (3): 356–66. doi:10.1002/hyp.8133.

Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *The Bell System Technical Journal* 27 (3–4): 379–423, 623–56.

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

Singh, Vijay P., and David A. Woolhiser. 2002. “Mathematical Modeling of Watershed Hydrology.” *Journal of Hydrologic Engineering* 7 (4): 270–92. doi:10.1061/(ASCE)1084-0699(2002)7:4(270).

Sturges, Herbert A. 1926. “The Choice of a Class Interval.” *Journal of the American Statistical Association* 21 (153): 65–66. doi:10.1080/01621459.1926.10502161.

Thiboult, A., and F. Anctil. 2015. “On the Difficulty to Optimally Implement the Ensemble Kalman Filter: An Experiment Based on Many Hydrological Models and Catchments.” *Journal of Hydrology* 529 (October). Elsevier B.V.: 1147–60. doi:10.1016/j.jhydrol.2015.09.036.

Tolson, Bryan A., and Christine A. Shoemaker. 2007. “Dynamically Dimensioned Search Algorithm for Computationally Efficient Watershed Model Calibration.” *Water Resources Research* 43 (1): 1–16. doi:10.1029/2005WR004723.

Xie, Xianhong, and Dongxiao Zhang. 2010. “Data Assimilation for Distributed Hydrological Catchment Modeling via Ensemble Kalman Filter.” *Advances in Water Resources* 33 (6). Elsevier Ltd: 678–90. doi:10.1016/j.advwatres.2010.03.012.

CHAPTER 5 An extension of data assimilation into the short-term hydrologic forecast for improved prediction reliability

Summary of Paper 4: Leach, J.M. and Coulibaly, P. (2019). An extension of data assimilation into the short-term hydrologic forecast for improved prediction reliability, Advances in Water Resources, doi:10.1016/j.advwatres.2019.103443

Summary:

This research developed and evaluated a methodology which allows for the extension of data assimilation into the forecast. By building a database which contains the prior knowledge of how a watershed modelled with a particular hydrologic model responds to forcing data, pseudo-observations could then be pulled from it and assimilated during the forecast period. This method was tested using the Ensemble Kalman Filter (EnKF) and the Particle Filter (PF) on two highly urbanized watersheds. The results suggest that this can improve the forecast reliability.

The results of this research demonstrate:

- Using the pre-built database which stores prior information about how a watershed model responds to inputs, as well as the corresponding historical observations, the capability of data assimilation can be extended into the forecast period.
- This database allows for pseudo-observations to be assimilated into a model when real observations are not available. These pseudo-observations are true observations from the past which have been looked-up based on the current model states, forcing data, and prediction, or they are artificially generated observations from a pre-trained neural network.
- By extending data assimilation into the short-term forecast using this lookup method, the reliability of forecasts can be improved.

5.1 Abstract

Typically, when using data assimilation to improve hydrologic forecasting, observations are assimilated up to the start of the forecast. This is done to provide more accurate state and parameter estimates which, in turn, allows for a better forecast. We propose an extension to the traditional data assimilation approach which allows for assimilation to continue into the forecast to further improve the forecast's performance and reliability. This method was tested on two small, highly urbanized basins in southern Ontario, Canada; the Don River and Black Creek basins. Using a database of forcing data, model states, predicted streamflow, and streamflow observations, a lookup function was used to provide an observation during the forecast which can be assimilated. This allows for an indirect way to assimilate the numerical weather prediction forcing data. This approach can help in addressing prediction uncertainty, since an ensemble of previous observations can be pulled from the database which correspond to the forecast probability density function given previous information. The results show that extending data assimilation into the forecast can improve forecast performance in these urban basins, and it was shown that the forecast reliability could be improved by up to 78 percent.

Keywords: Data Assimilation; Ensemble Kalman Filter; Particle Filter; Short-term forecast; nearest neighbour; artificial neural network

5.2 Introduction

With our changing climate, extreme weather events, especially heavy rainfalls, are becoming more common which leads increasing need for improved flood forecasting methods. Historically, deterministic methods were used to predict floods with a chosen model being forced with a deterministic Numerical Weather Prediction (NWP) product; but we know that this method is not ideal. Advances over the last two decades have led to ensemble and probabilistic approaches to better improve flood forecasting as well as to quantify the associated uncertainties related to the forecast (Liu et al., 2012). These more advanced forecasting and uncertainty quantification methods include sequential data assimilation (SDA) (Liu et al., 2012; Moradkhani et al., 2005), Bayesian forecasting system (BFS) (Biondi and De Luca, 2012; Han and Coulibaly, 2019; Krzysztofowicz, 1999), Bayesian model averaging (BMA) (Najafi and Moradkhani, 2016; Raftery et al., 2005), and model conditional processor (MCP) (Todini, 2008). With improved methods we are able to more accurately predict floods sooner, leading to a faster response and a reduced associated cost.

Sequential data assimilation is a popular method used in hydrologic modeling to improve hydrologic and flood forecasting. As an application of Bayes Theorem, SDA can incorporate the uncertainty from both the model and observations to update boundary conditions and improve forecasts. There are several SDA methods which have been used for hydrologic forecasting such as the Extended Kalman Filter (EKF) (Sun et al., 2015), Ensemble Kalman Filter (EnKF) (Leach et al., 2018; Moradkhani et al., 2005; Samuel et al., 2014; Thiboult et al., 2015; Vrugt and Robinson, 2007), Ensemble Kalman Smoother (EnKS) (Crow and Ryu, 2009; Li et al., 2013), Evolutionary Data Assimilation (EDA) (Dumedah and Coulibaly, 2013), the Particle Filter (PF) (DeChant and Moradkhani, 2014; Moradkhani et al., 2005; Parrish et al., 2012; Yan and Moradkhani, 2016), and most recently the Evolutionary PF Hybrid methods (Abbaszadeh et al., 2019, 2018; Ju et al., 2019; Zhu et al., 2018); with the EnKF and PF being popular due to them being relatively simple to implement. Additionally, various types of observations have been successfully assimilated into hydrologic models, using these SDA techniques, such as streamflow (Abaza et al., 2017; Abbaszadeh et al., 2018; Yan and Moradkhani, 2016), soil moisture (Ju et al., 2019; Leach et al., 2018; Meng et al., 2017; Yan and Moradkhani, 2016), and snow water equivalent (Dziubanski and Franz, 2016; Huang et al., 2017; Leach et al., 2018; Smyth et al., 2019); these data sets are obtained from in situ gauges or remote sensing techniques. Through assimilating these observations, both models states and parameters can be updated to improve hydrologic simulation and forecasting (Leach et al., 2018; Moradkhani et al., 2005; Samuel et al., 2014).

One downside to hydrologic DA and modeling in general, is limited observational data due to many locations being ungauged. This causes issues for SDA with distributed models, as it has been shown that SDA in distributed modeling is better when more observations are available from multiple locations (Clark et al., 2008; Rakovec et al., 2012; Xie and Zhang, 2010). This data limitation also impacts forecasts, since the ability to update is limited by observations being available to assimilate at the start of a forecast. If a real-time gauge, or a gauge in general, does not exist, then there may not be observations available at the beginning of the forecast to be used by SDA techniques to update the states and parameters. Due to the fact that these SDA methods are ensemble forecasting methods, they are limited in that they can address NWP uncertainty and merge model and observation uncertainty, but are not guaranteed to address prediction uncertainty (Biondi and Todini, 2018).

In this study, we aim to address the limitation of SDA during the forecast mode, specifically the lack of available observations to assimilate (since there are none). To do this, we first build a database for our selected watershed. The database contains results from a chosen hydrologic model, such as predicted streamflows, predicted states, the forcings used to drive the model, and the corresponding observed streamflows. This database essentially stores how a specific watershed responds to forcings given some initial conditions, and it is used to generate a probability density function (pdf) of pseudo-observations ($\tilde{y}$), conditional on available information (I), denoted as $f\{\tilde{y}|I\}$ (Biondi and Todini, 2018). This pdf is then assimilated back into the model at each forecast step with the goal of improving the forecast's performance and reliability. This method could potentially also be extended to ungauged basins through the use of regionalization techniques to generate a database for any basin, although that is not explored within this work.

The proposed methodology will allow for data assimilation to be extended into the forecast period by using our prior knowledge of how the model responds to different inputs. In doing so, we can better account for the prediction uncertainty with SDA and provide better state estimates at each forecast horizon. Additionally, during a dual state and parameter updating scheme, this method will allow for the model parameters to be calibrated to the chosen weather forecast product. Real-time calibration to the weather forecast is beneficial since models are traditionally calibrated using historical gauge data, however, model parameter distributions can vary depending on what forcing data products are used to calibrate the model (Kornelsen and Coulibaly, 2016).

5.3 Study Areas and Data

The study areas being focused on are the Black Creek and Don River basins in Toronto, Ontario, Canada (Figure 5-1). These watersheds are managed by the Toronto Region Conservation Authority (TRCA). Both the Black Creek basin (BCB) and the Don River basin (DRB) are highly urbanized, being roughly 97% and 93% developed, respectively (Agriculture and Agri-Food Canada, 2015). These basins have an average daily temperature of 8.0 °C, an average daily minimum and maximum temperatures of 3.4 °C to 12.5 °C respectively, and an average annual precipitation of 841.1 mm/year based on the 1981-2010 Canadian Climate Normals (Environment and Climate Change Canada, 2017).

Hourly data was provided by the TRCA, Environment and Climate Change Canada (ECCC), and Water Survey of Canada (WSC) (20080101 to 20171231). Gaps in the precipitation data were infilled using a combination of disaggregated daily to hourly ECCC precipitation data and ordinary kriging. The disaggregation was performed using multiplicative random cascade (MRC)-based disaggregation (Ganguli and Coulibaly, 2017; Olsson, 1998, 1995). Ordinary kriging was also used to infill gaps in the temperature data. Forecast forcing data was obtained from the Regional Deterministic Prediction System (RDPS). The RDPS provides four 48-hour forecasts per day, in 3-hour increments, on a 10 km grid. The 00-hour forecast was used for this study, with 3-hour precipitation and temperature data from 20150401 to 20171231.

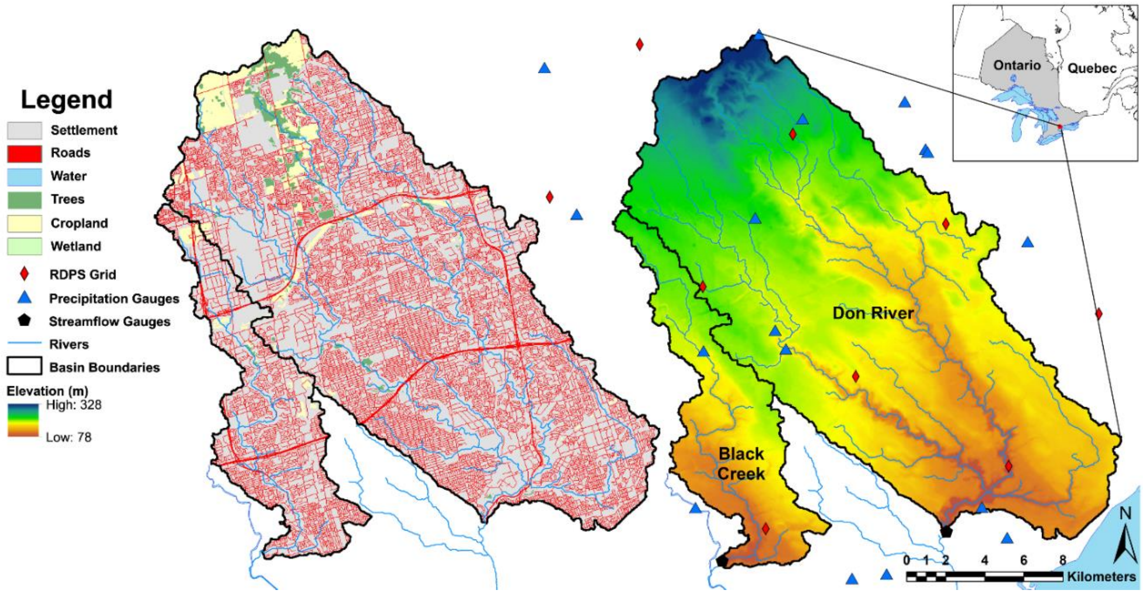

Figure 5-1: Land-use/land-cover (left) and elevation (right) for Don River and Black Creek in Ontario, Canada. Locations of streamflow gauges, precipitation gauges, and RDPS grid points are also illustrated.

5.4 Methodology

5.4.1 Hydrologic Model

HyMod is a lumped hydrologic model based on the probability distributed moisture model (PDM) (Moore, 1985). For this study HyMod was modified to include the degree-day snow routine from Samuel et al. (2011) and the simplified Thornthwaite evapotranspiration method (Samuel et al., 2011). The model was calibrated using a 3-hour time interval, this timestep was chosen to match the RDPS dataset’s forecast timestep. Each watershed model was calibrated using the Dynamically Dimensioned Search (DDS) Algorithm (Tolson and Shoemaker, 2007), which has one algorithm parameter, *r*, used to control the neighborhood size for the random search. Using DDS, the calibration period was 20090101 to 20121231 and the validation period was 20130101 to 20151231, with the period of 20080101 to 20081231 used for spin-up. The Nash Volume Efficiency (NVE) from Samuel et al. (2012), with modified weights, was the chosen objective function. The NVE is defined as follows:

$$NVE = 0.3NSE - 0.1|VE| + 0.2NSE_{log} + 0.5NSE_{sqr} \quad (5\text{-}1)$$

$$NSE = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - E[\boldsymbol{y}])^2} \quad (5\text{-}2)$$

$$VE = \frac{E[\boldsymbol{y} - \hat{\boldsymbol{y}}]}{E[\boldsymbol{y}]} \quad (5\text{-}3)$$

where *NSE* is the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970), NSE_{log} is the NSE calculated using log-transformed runoff (low flow), NSE_{sqr} is the NSE calculated using the squared runoff (high flow), *VE* is the volume error, y_t is the observed runoff at time *t*, $\hat{y}_t$ is the simulated runoff at time *t*, and $E[.]$ is the expectation function.

Descriptions for the HyMod model parameters and states are listed in Table 5-1. The calibrated parameter sets for both the BCB and DRB are also provided in Table 5-1 along with their parameter ranges and associated units.

Table 5-1: HyMod parameters and state variables, descriptions, parameter ranges, as well as the calibrated parameter sets for each watershed.

Parameter	Description	Range	Units	Calibrated Value	
				DRB	BCB
Alpha	Factor distributing the runoff between the quick and slow reservoirs	0.1 – 0.99	-	0.492	0.407
B_{exp}	Degree of spatial variability of soil moisture capacity	0.1 – 3	-	1.497	2.963
C_{max}	Maximum storage capacity	1 – 1000	mm	36.400	58.403
Rs	Residence time of slow flow reservoir	0.01 – 0.99	day	0.182	0.345
Rq	Residence times of quick flow reservoirs	0.001 – 0.1	day	0.004	0.002
tr	Rainfall threshold temperature	0 – 2.5	°C	1.827	2.187
scf	Snow correction factor	0.4 – 1.6	-	0.420	0.528
ddf	Degree day factor	0 – 5	mm/day/°C	0.221	3.845
rcr	Rainfall correction factor	0.5 – 1.5	-	1.457	1.482
athorn	Potential evapotranspiration coefficient for simplified Thornthwaite equation	0.1 – 0.3	-	0.131	0.106
	State variables				
S	Watershed storage	-	mm	-	-
qfr1	Quick flow reservoir 1	-	mm	-	-
qfr2	Quick flow reservoir 2	-	mm	-	-
qfr3	Quick flow reservoir 3	-	mm	-	-
sfr	Slow flow reservoir	-	mm	-	-
swe	Snow water equivalent	-	mm	-	-

5.4.2 Data Assimilation

For hydrologic forecasting, the benefit of SDA is that it can take new information at each time step and update our prior estimate of the states (and parameters), to provide a more accurate posterior which leads to a better forecast. The posterior distribution of states at time *t* can be found using the recursive Bayes' Theorem:

$$P(x_t|y_{1:t}) = P(x_t|y_t, y_{1:t-1}) = \frac{P(y_t|x_t)P(x_t|y_{1:t-1})}{\int P(y_t|x_t)P(x_t|y_{1:t-1})dx_t} \tag{5-4}$$

where $P(x_t|y_{1:t})$ is the posterior distribution, $P(x_t|y_{1:t-1})$ is the prior distribution, $P(y_t|x_t)$ is the likelihood for time t, and $\int P(y_t|x_t)P(x_t|y_{1:t-1})dx_t$ is the normalization factor. Since the analytic solution is generally intractable, a numerical approximation can be made using Monte Carlo methods.

5.4.2.1 Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) is a Monte Carlo based SDA method that can be used to optimally combine model and observation uncertainties to provide an improved forecast (Evensen, 2003, 1994). The EnKF is a commonly used data assimilation method which can be used to update model states and parameters; its update procedure is shown as follows (Moradkhani et al., 2005):

$$\widehat{\boldsymbol{X}}_t^- = f\left(\widehat{\boldsymbol{X}}_{t-1}^+, \boldsymbol{\theta}, \boldsymbol{u}_t\right) + \boldsymbol{\zeta}_t \tag{5-5}$$

$$\widehat{\boldsymbol{y}}_t = h(\widehat{\boldsymbol{X}}_t^-, \boldsymbol{\theta}) + \boldsymbol{\nu}_t \tag{5-6}$$

$$\boldsymbol{K}_t = \frac{\mathrm{cov}(\widehat{\boldsymbol{X}}_t^-, \widehat{\boldsymbol{y}}_t)}{\mathrm{var}(\widehat{\boldsymbol{y}}_t) + R} \tag{5-7}$$

$$\widehat{\boldsymbol{X}}_t^+ = \widehat{\boldsymbol{X}}_t^- + \boldsymbol{K}_t(\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t) \tag{5-8}$$

where $\widehat{\boldsymbol{X}}_t$ is the predicted ensemble of states for time *t*, $\boldsymbol{\theta}$ is the vector of model parameters, $\boldsymbol{u}_t$ is the ensemble of perturbed forcing data at time *t*, $f(.)$ is the state propagation functions which moves them forward in time, and $\boldsymbol{\zeta}_t$ is the random noise which represents process error. The superscripts – and + represent the prior and posterior updated states, respectively. The hydrologic model is represented by $h(.)$, $\widehat{\boldsymbol{y}}_t$ is the predicted runoff at time *t*, and $\boldsymbol{\nu}_t$ is noise which represents model uncertainty. The Kalman gain, $\boldsymbol{K}_t$, at time *t* is found by dividing the cross-covariance of predicted states and runoff by the variance of the innovation (the difference of the observed and predicted runoff). The observed runoff $\boldsymbol{y}_t$ is perturbed by random noise $\mathrm{N}(0, R)$ to generate an ensemble of observations, with *R* being the prior known observation uncertainty. The difference of the observed and predicted runoff is then multiplied by the Kalman gain and added to the prior state estimate to create the posterior state estimate. This process, Eq. 5-5 to 5-8, is repeated until there are no more observations, or the spin-up period is complete.

5.4.2.2 Particle Filter

Another common SDA method is the Particle Filter (PF). As with the EnKF, the PF can be used to update model states and parameters (Moradkhani, 2008; Moradkhani et al., 2005). There are many variants of the particle filter that utilize different particle resampling methods. Four resampling methods were selected for this analysis, they were the Residual Resampling (RR) method (Liu and Chen, 1998), two versions of the Sample Importance Resampling method (SIR and SIRV) (Smith and Gelfand, 1992), and the Markov Chain Monte Carlo (MCMC) method (Leisenring and Moradkhani, 2012; Moradkhani et al., 2012; Yan et al., 2015; Yan and Moradkhani, 2016). The SIR and RR methods add constant noise to the particles at every timestep in order to add particle diversity and combat degeneracy (Liu and Chen, 1998), while the SIRV and MCMC methods add variable noise, using the variable variance multiplier (VVM), to combat particle degeneracy (Leisenring and Moradkhani, 2012; Moradkhani et al., 2012). The general formulation of the PF, in conjunction with Eq. 5-5 and 5-6, is as follows (Gordon et al., 1993; Moradkhani et al., 2005):

$$L\left(y_t \middle| \boldsymbol{x}_{i,t}^{-}\right) = \frac{1}{\sqrt{2\pi R}} \exp\left(-\frac{\left(y_t + \varepsilon_{i,t} - \hat{y}_{i,t}\right)^2}{2R}\right) \tag{5-9}$$

$$P\left(y_t \middle| \boldsymbol{x}_{i,t}^{-}\right) = \frac{L\left(y_t \middle| \boldsymbol{x}_{i,t}^{-}\right)}{\sum_{i=1}^{N} L\left(y_t \middle| \boldsymbol{x}_{i,t}^{-}\right)} \tag{5-10}$$

$$w_{i,t}^{+} = \frac{w_{i,t}^{-} P\left(y_t \middle| \boldsymbol{x}_{i,t}^{-}\right)}{\sum_{i=1}^{N} w_{i,t}^{-} P\left(y_t \middle| \boldsymbol{x}_{i,t}^{-}\right)} \tag{5-11}$$

$$P(\boldsymbol{X}_t^{+} | y_{1:t}) \approx \sum_{i=1}^{N} w_{i,t}^{+} \delta\left(\boldsymbol{X}_t - \boldsymbol{x}_{i,t}\right) \tag{5-12}$$

where *N* is the number of particles, $w_{i,t}^{-}$ is the prior weight (set to 1/N) for particle *i* at time *t*, $w_{i,t}^{+}$ is the posterior weight for particle *i* at time *t*, $\delta(.)$ is the Dirac delta function, $\varepsilon_{i,t} \sim N(0, R)$ is random noise which represents observation error, and $L(.)$ is the Gaussian likelihood function. The posterior estimate can then be resampled to eliminate low weight particles using the chosen resampling method.

There are many resampling methods that can be used with the PF, including the RR, SIR, and MCMC resampling methods. The RR method for reweighting residual particles (particles with a weight less than *1/N*) is:

$$w_{i,t}^{*+} = \frac{N w_{i,t}^{+} - \lfloor N w_{i,t}^{+} \rfloor}{N - \sum_{i=1}^{N} \lfloor N w_{i,t}^{+} \rfloor} \tag{5-13}$$

where $w_{i,t}^{*+}$ are the weights of the N^* residual particles ($N^* = N - \sum_{i=1}^{N} \lfloor N w_{i,t}^{+} \rfloor$), that will be used to construct an empirical cumulative distribution function (CDF). This CDF is then resampled from to increase the number of particles from N^* to *N*, after which the particle weights are set to *1/N*. When using SIR, like RR, a CDF of the particle weights is constructed. From this CDF, *N* particles are resampled proportionally to their weights with higher weighted particles being resampled more often (when their probability is higher than uniform probability). The MCMC resampling method is an extension of the SIR method, it uses a metropolis acceptance ratio, α, to decide if an update should be accepted or rejected.

$$\alpha = \min\left(1, \frac{P\left(\boldsymbol{X}_t^p \middle| y_{1:t}\right)}{P(\boldsymbol{X}_t^+ | y_{1:t})}\right) \tag{5-14}$$

The proposed joint probability distribution of the updated particles is $P\left(\boldsymbol{X}_t^p \middle| y_{1:t}\right)$, and the particles are updated when the acceptance ratio exceeds a threshold value generated from a uniform random variable $u \sim U(0,1)$.

5.4.3 Pseudo-observation lookup routine

To provide pseudo-observations during the forecast period a database was built that contains past information about the watershed. By using this database, data assimilation can be extended into the forecast period. This database contains the HyMod state values (Table 5-1), the precipitation and temperature data used to force the model and generate those state values, and the predicted streamflow associated with those states and forcings for each timestep during the time period allocated for building the database. Additionally, the database contains the corresponding observed streamflow values, and the prediction error values for each timestep in the database. To evaluate the influence forcing data uncertainty may have on the lookup methods, two databases were built for this study using different forcing datasets.

The first database was built using historical precipitation and temperature gauge data to run HyMod, denoted as the historical observation database (*hdb*). The *hdb* was built using gauge data from 20090101 to 20161231, and results from each timesteps were archived. The second database was created using archived RDPS precipitation and temperature, denoted as the forecast forcing database (*fdb*). Since the available archive of RDPS forecasts is much shorter than that of the gauges, the period of 20150401 to 20161231 was used to build the *fdb*.

Three lookup methods were evaluated and used to pull pseudo-observations from the pre-built databases. The lookup methods are *k*-Nearest Neighbour (*k*-NN), direct lookup (DL), and a feedforward neural network (FNN). Each method was evaluated based on how they affected the performance of the forecasts. The term pseudo-observation was used to describe the values pulled from the database because they are true observations that correspond to past vectors (of states, forcings, and predicted streamflow) which most closely resembles the forecasted vector (of states, forcings, and predicted streamflow). The DL method directly provides the closest values from the database as a pseudo-observation, the *k*-NN method provides the average of the *k* closest values as a pseudo-observation, and the FNN generates a pseudo-observation which corresponds to the forecasted vector. The general flowchart illustrating how the method works, and where the lookup routine fits into the forecast, is provided in Figure 5-2.

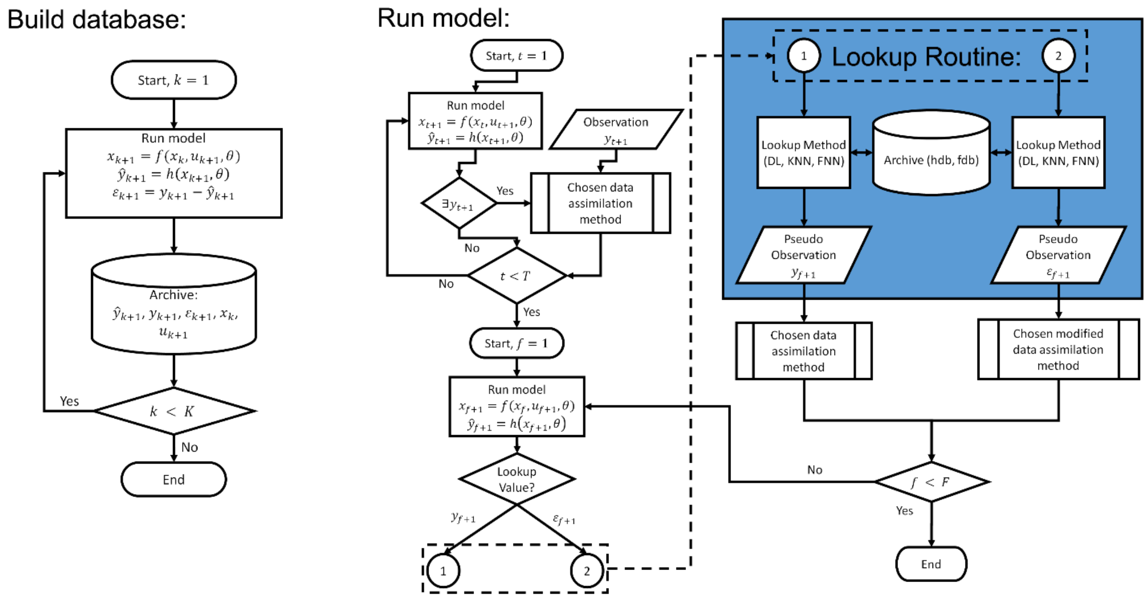

Figure 5-2: Flowchart illustrating database setup, spin-up data assimilation, and data assimilation in the forecast.

5.4.3.1 Direct lookup and k-Nearest Neighbour

The *k*-NN method is a simple machine learning algorithm which can be used for regression or classification (Altman, 1992). When using it for regression, the *k* nearest neighbours to the sample point of interest are averaged to provide a predicted value. In the case of this analysis, the sample point is a vector containing the initial model states, forcing data, and predicted streamflow for a timestep, which is then compared to the points in a database, either the *hdb* or *fdb*, and a corresponding streamflow observation, or prediction error, is returned. This observation is the mean of the *k* closest points to the sample point. The direct lookup method is the special case of the *k*-NN algorithm when $k = 1$. It is possible to use different distance and weighting methods for k-NN, however for this work the Euclidean distance with equal weights was used. The general formulation for the *k*-NN method is as follows:

$$\begin{cases} c(j) = \underset{i}{\operatorname{argmin}} \| \boldsymbol{x}_t - \boldsymbol{\mathcal{D}}_i \| , \qquad for\ j = 1 \\ c(j) = \underset{i}{\operatorname{argmin}} \| \boldsymbol{x}_t - \boldsymbol{\mathcal{D}}_i \| , \qquad i \neq c(j-1), \qquad for\ j = 2, \ldots, k \end{cases} \tag{5-15}$$

$$\tilde{y}_t = \frac{1}{k} \sum_{j=1}^{k} \mathscr{y}_{c(j)} , \qquad \mathscr{y}_{c(j)} \in \boldsymbol{\mathcal{D}} \tag{5-16}$$

where $\boldsymbol{x}_t$ is the vector of states, forcing data, and predicted streamflow at time *t*, $\boldsymbol{\mathcal{D}}_i$ is a database vector entry *i* that is of the same structure as $\boldsymbol{x}_t$, $\|.\|$ is the Euclidean distance function (2-norm), and $c(j)$ is the database index value of the nearest $j = 1, \ldots, k$ neighbours. The pseudo streamflow observation for time *t*, $\tilde{y}_t$, is then found by taking the average of the archived streamflow observation, $\mathscr{y}_{c(j)}$. The prediction error (or innovation) can be retrieved from the database in a similar manner.

5.4.3.2 Feedforward Neural Network

To generate comparable results to those of the DL and *k*-NN lookup methods, a FNN was trained on each database and used to generate pseudo-observations which could be assimilated during the forecast period. More specifically, the FNN was trained to reproduce the past true observation which corresponds to a past vector (of states, forcings, and predicted streamflow), so that if given a forecasted vector the FNN would generate a pseudo-observation for it. The specific type of FNN used as a lookup routine was the multilayer perceptron (MLP). A simple formulation of the MLP is shown as follows (Hagan et al., 1997):

$$\tilde{y}_t = f(\boldsymbol{W}_2\, \mathrm{g}(\boldsymbol{W}_1 \boldsymbol{x}_t + \boldsymbol{b}_1) + \boldsymbol{b}_2) \tag{5-17}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the input layer and hidden layer neurons weight matrices, respectively; $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are the input layer and hidden layer bias vectors, respectively; $f(.)$

is the linear activation function for the output neuron, and $g(.)$ is the non-linear activation function (hyperbolic tangent) for the hidden neurons.

The MLP was pre-trained separately for the *hdb*, and the *fdb* using the Levenberg-Marquardt optimization method with Mean squared normalized error as the objective function. For training purposes, the databases were randomly split into 70% training data used for fitting, 15% validation data used for unbiased evaluation while tuning hyper-parameters, and 15% testing data used for an unbiased evaluation of the final model. The network architecture included an input layer with nine neurons, one hidden layer with twenty neurons, and the output layer with one neuron. The input layer neurons were the initial model states, forcing data, and predicted streamflow for a timestep, and the output neuron returned either the streamflow observation or prediction error (innovation) for that timestep.

5.4.3.3 Modification to Kalman Gain and Likelihood functions

When the prediction error (innovation) is retrieved from the databases, the data assimilation formulations need to be modified. This is because the prediction error is not independent from the predicted streamflow, therefore the variance of $\boldsymbol{Y}_t = \boldsymbol{\varepsilon}_t + \widehat{\boldsymbol{Y}}_t$ is $\mathrm{var}(\boldsymbol{Y}_t) = \mathrm{var}(\widehat{\boldsymbol{Y}}_t) + \mathrm{var}(\boldsymbol{\varepsilon}_t) + 2 \times \mathrm{cov}(\boldsymbol{\varepsilon}_t, \widehat{\boldsymbol{Y}}_t)$. Since $R \approx \mathrm{var}(\boldsymbol{Y}_t)$, the Kalman gain in Eq. 5-7 when using the prediction error becomes:

$$\boldsymbol{K}_t = \frac{\mathrm{cov}(\widehat{\boldsymbol{X}}_t^-, \widehat{\boldsymbol{Y}}_t)}{2 \times \mathrm{var}(\widehat{\boldsymbol{Y}}_t) + \mathrm{var}(\boldsymbol{\varepsilon}_t) + 2 \times \mathrm{cov}(\boldsymbol{\varepsilon}_t, \widehat{\boldsymbol{Y}}_t)} \tag{5-18}$$

A similar modification can be made to the likelihood function used for the PF in Eq. 5-9, such that it becomes:

$$L\left(Y_t \middle| X_{i,t}^-\right) = \frac{\exp\left(-\frac{\left(\varepsilon_{i,t}\right)^2}{2 \times \left(\mathrm{var}\left(\widehat{\boldsymbol{Y}}_t\right) + \mathrm{var}(\boldsymbol{\varepsilon}_t) + 2 \times \mathrm{cov}\left(\boldsymbol{\varepsilon}_t, \widehat{\boldsymbol{Y}}_t\right)\right)}\right)}{\sqrt{2\pi \times \left(\mathrm{var}\left(\widehat{\boldsymbol{Y}}_t\right) + \mathrm{var}(\boldsymbol{\varepsilon}_t) + 2 \times \mathrm{cov}\left(\boldsymbol{\varepsilon}_t, \widehat{\boldsymbol{Y}}_t\right)\right)}} \tag{5-19}$$

The assimilation of the prediction error was motivated by the idea that there may be a large difference in the innovation during the forecast when the pseudo-streamflow observation is assimilated compared to what it was historically. In doing this test, the impact that difference has could be better evaluated.

5.4.4 Evaluation methods

To evaluate the performance of the data assimilation methods when running the model using historical gauge forcing, the Kling-Gupta Efficiency (KGE) (Gupta et al., 2009), the Root Mean Square Error (RMSE), and the Peak Flow Criteria (PFC) (Coulibaly et al., 2001) were used. These criteria were chosen to provide a baseline of performance for each data assimilation method before the extension into the forecast. These metrics are defined as follows:

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\sigma_{\bar{\hat{Y}}}/\sigma_Y - 1\right)^2 + \left(\mu_{\bar{\hat{Y}}}/\mu_Y - 1\right)^2} \tag{5-20}$$

$$RMSE = \left(\frac{1}{T}\sum_{t=1}^{T}\left(\bar{\hat{Y}}_t - Y_t\right)^2\right)^{\frac{1}{2}} \tag{5-21}$$

$$PFC = \frac{\left(\sum_{p=1}^{T_p}\left(Q_p - \bar{\hat{Q}}_p\right)^2 Q_p^2\right)^{1/4}}{\left(\sum_{p=1}^{T_p} Q_p^2\right)^{1/2}} \tag{5-22}$$

The KGE is found using the linear correlation coefficient, r, the relative variability, $\sigma_{\bar{\hat{Y}}}/\sigma_Y$, and the mean ratio, $\mu_{\bar{\hat{Y}}}/\mu_Y$, between the ensemble mean of the simulated runoff and the observed runoff. Where Y_t is the observed runoff, $\bar{\hat{Y}}_t$ is the ensemble mean of the simulated runoff, *t* is the timestep, *T* is the number of timesteps, Q_p is the observed runoff peak, $\bar{\hat{Q}}_p$ is the ensemble mean of the simulated runoff peak, T_p is the number of observed peaks above a threshold, and $p \in \{1, \ldots, T_p\}$. The threshold value for peak flows was found by identifying the peaks, then taking one-third of the mean of those peaks. Since an essential consideration in urban and flood-prone environments is the peak flow values, the PFC will be used to evaluate how well the data assimilation methods can improve the peak flow simulation in the ungauged basins. The RMSE and KGE will be used to evaluate the general performance of the data assimilation methods, specifically how effective they are at improving performance in ungauged basins. The RMSE and PFC values can range from zero to infinity with zero being optimal, and the KGE can range from negative infinity to one, with one being optimal.

The ensemble forecast performance of each assimilation scheme and model combination was evaluated using the mean Continuous Ranked Probability Score (CRPS) performance metric. The CRPS can be used to assess accuracy and resolution of the 48-hour ensemble forecast performances. The mean CPRS is formulated as follows (Matheson and Winkler, 1976; Unger, 1985):

$$\overline{CRPS}(F, y) = \frac{1}{N}\sum_{t=1}^{N}\int_{-\infty}^{\infty}(F(\hat{y}_t) - \mathbf{1}\{\hat{y}_t \geq y_t\})^2 d\hat{y} \tag{5-23}$$

where $F(\hat{y}_t)$ is the cumulative distribution function of the ensemble forecasts, $\hat{y}_t$ is the predicted runoff, y_t is the observed runoff, and $\mathbf{1}\{\hat{y}_t \geq y_t\}$ is the Heaviside step function that provides a value of 1 if the predicted value is larger than the observed and 0 otherwise. A CPRS value of 0 indicates a perfect forecast.

To determine the reliability of the forecasts, the mean CRPS was decomposed according to Hersbach (2000), such that $\overline{CRPS} = \overline{Reli} + CRPS_{pot}$. This decomposition is analogous to that of the Brier score decomposition which can provide reliability, resolution, and uncertainty (Hersbach, 2000; Murphy, 1973). Reliability of a forecast is a measure of its statistical accuracy and whether or not the forecasting system has correct statistical properties (Hersbach, 2000; Murphy, 1973). More specifically, the forecast can be considered reliable if it is not over or under dispersed and is unbiased. In this way, the reliability also has a connection to the rank histogram, which can be used to evaluate the spread and bias of an ensemble. The reliability and potential CRPS are calculated as follows:

$$\overline{Reli} = \sum_{i=0}^{N} \bar{g}_i(\bar{o}_i - p_i)^2 \tag{5-24}$$

$$CRPS_{pot} = \sum_{i=0}^{N} \bar{g}_i\bar{o}_i(1 - \bar{o}_i) \tag{5-25}$$

where $\bar{g}_i$ is the average width of bin *i*, $\bar{o}_i$ is the average frequency that the observation is less than the middle of bin *i*, p_i is the fraction *i/N*, and *N* is the ensemble size. There are *N+1* bins and each is defined by the distance between consecutive ensemble members when *0<i<N*, while for *i=0,N* the bins are defined by the distance between the observation and

outlier members. The $CRPS_{pot}$ represents what the $\overline{CRPS}$ could be if the forecast was perfectly reliable, and $\overline{Reli}$ is an indication of how reliable the forecast is. For both metrics the optimal value is 0. The interested reader can see Hersbach (2000) for a more detailed description of the decomposition.

The Reliability performance measure from Renard et al. (2010) and the 95% Exceedance Ratio (ER95) from Moradkhani et al. (2006) were also used for evaluating the performance of the forecasts. These metrics evaluate the overall reliability of the predictive distribution and the spread of the ensemble, respectively. They are calculated as follows:

$$Reliability = 1 - \frac{2}{T}\sum_{t=1}^{T}\left|P\left(\hat{Y}_t \leq y_t\right) - P_t^U\right| \tag{5-26}$$

$$ER95 = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_{97.5\%,t} < y_t \ or \ \hat{y}_{2.5\%,t} > y_t\right) \times 100\% \tag{5-27}$$

where $P\left(\hat{Y}_t \leq y_t\right)$ is the (sorted) probability that the ensemble prediction $\hat{Y}_t$ will be less than the observation y_t, T is the total number of time steps, P_t^U is the theoretical cumulative probability (cumulative uniform distribution), $\hat{y}_{97.5\%,t}$ is the 97.5 percentile of the predicted streamflow ensemble, and $\hat{y}_{2.5\%,t}$ is the 2.5 percentile of the predicted streamflow ensemble. The Reliability measure can range from 0 to 1 with 1 being perfectly reliable, and the ER95 ranges from 0 to 100% with a perfect ensemble having a value of 5%.

As part of assessing the impact of extending the assimilation into the forecast, four cases were evaluated: (1) for both the spin up and forecast period, the model was run without data assimilation; (2) for the spin-up period, data assimilation was used, and during the forecast it was just the model; (3) data assimilation was used in both the spin up and forecast periods; and (4) the model was used during the spin-up period and data assimilation was used in the forecast. Additionally, for all four cases during the spin-up period the models were forced with historical gauge data, and during the forecast period the models were forced with the RDPS forecasted forcings.

5.5 Results and discussions

5.5.1 Model calibration, validation, and standard data assimilation results

Using the DDS optimization, multiple calibrations for each basin were performed, with each run consisting of 50000 iterations with an *r* of 0.2, then another 50000 iterations, starting from the previous' endpoint, with an *r* of 0.1. The best HyMod model calibration (20090101 to 20121231) and validation (20130101 to 20151231) results, for both basins, are summarized in Table 5-2. From these results, we see that both basins perform reasonably well during the calibration period. However, larger runoff events occurred during the validation period in both basins, which can explain why they perform slightly worse during that period. Additionally, the Black Creek basin model performs better than the Don River basin model, which is likely due to the controlled reservoir in the Don River basin, making it more difficult to model.

Table 5-2: Calibration and Validation performance of 3-hour HyMod model for each watershed

Basin	Calibration			Validation		
	NSE	NVE	VE	NSE	NVE	VE
Don River	0.56	0.45	0.06	0.42	0.32	0.07
Black Creek	0.66	0.54	0.03	0.55	0.37	0.03

The models were also run using each data assimilation during the validation period so that the performance of each method could be compared to the open loop case. The results of these runs are summarized in Table 5-3.; From this table, it can been seen that the data assimilation with only state updating performs best for the 3-hour ahead forecast using the historical gauge data. This was likely due to the models being well calibrated using the historical gauge data, so re-calibrating using the dual state and parameter updating strategies was not beneficial. The PF methods with Sequential Importance Resampling and Residual Resampling performed the best for these state updating strategies, however the EnKF was comparable. Note that when using the dual state and parameter updating strategy, the EnKF method outperformed the other data assimilation methods.

Table 5-3: Comparison of performances of data assimilation methods to open loop validation period (20130101 to 20151231). Note that these runs provide only the 3-hour ahead forecasts using historical gauge data as forcing.

DA Method	Update	Don River			Black Creek		
		KGE	RMSE (mm/3hr)	PFC	KGE	RMSE (mm/3hr)	PFC
Open loop	-	0.66	0.20	0.41	0.60	0.41	0.61
EnKF	State	0.72	0.15	0.35	0.77	0.22	0.39
	Dual	0.54	0.15	0.39	0.67	0.21	0.42
PF-SIR	State	0.75	0.14	0.33	0.86	0.16	0.31
	Dual	0.28	0.17	0.40	0.46	0.23	0.46
PF-RR	State	0.75	0.14	0.33	0.85	0.16	0.33
	Dual	0.14	0.19	0.43	0.06	0.28	0.46
PF-SIRV	Dual	0.40	0.16	0.38	0.58	0.19	0.37
PF-MCMC	Dual	0.51	0.16	0.38	0.57	0.20	0.41

5.5.2 Standard data assimilation performance using the RDPS forecasts

The data assimilation methods were then evaluated based on their forecast performance with the RDPS forcing data. Illustrated in Figure 5-3 are the mean CRPS performances for both the Don River and Black Creek basin HyMod models using each assimilation method out to the 48-hour forecast. As a side note, a continuous archive of Regional Deterministic Prediction System (RDPS) was not available therefore only 302 of the RDPS forecasts were used from 2017 (forecasted at time 00), and the results in Figure 5-3 are the average of them. From the forecast results it appears that the data assimilation methods with only state updating performed nearly identically; but they were not the best performing methods as before (forced with historical gauge data for a 3-hour forecast). When forcing the HyMod model with RDPS it is instead better to use the dual state and parameter updating strategy with data assimilation to achieve the best performance; this is consistent with the literature (Leach et al., 2018; Moradkhani et al., 2005; Samuel et al., 2014; Yan et al., 2015).

Specifically, it is shown that the EnKF performs well for the first 6 hours for the Don River basin and the first 18 hours for the Black Creek basin. The particle filter with SIR, SIRV, and MCMC resampling outperforms the EnKF beyond those forecast horizons. Additionally, the particle filter with RR does not perform well for either basin when forcing with RDPS, and it will therefore be omitted from further analyses.

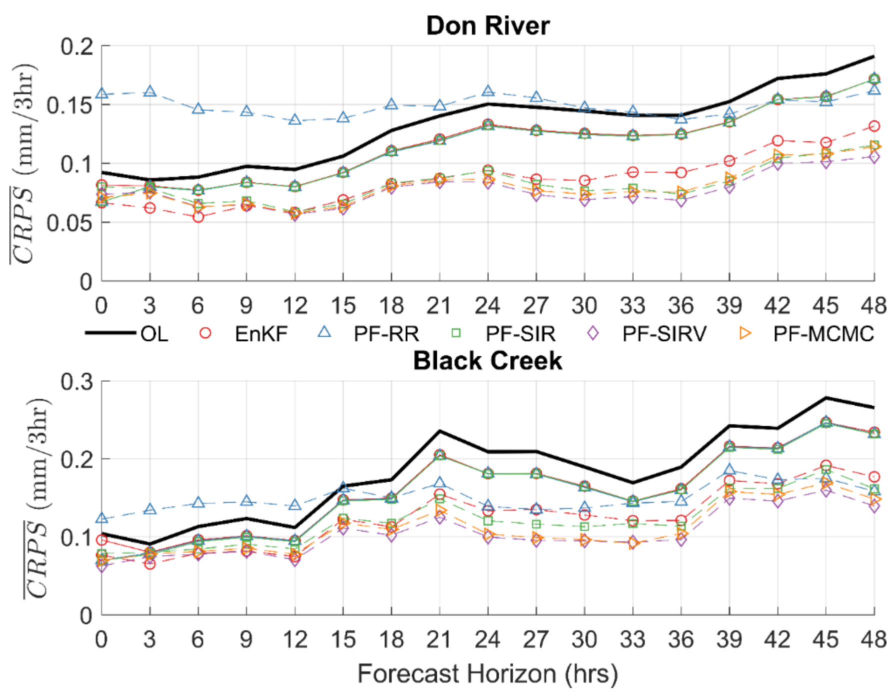

Figure 5-3: Standard data assimilation ensemble forecast performances illustrating how the mean CRPS values for each assimilation method for each watershed changes out to the 48-hour forecast. Solid lines indicate state updating and dashed lines indicate dual state and parameter updating. The solid black line indicates the open loop (OL) performance with no data assimilation.

5.5.3 Evaluation of data assimilation in the forecast

To evaluate the performance of extending data assimilation into the forecast, several scenarios were evaluated and compared. These scenarios included different combinations of lookup methods, lookup values, along with which database was being used. These combinations meant that twelve scenarios were tested for each assimilation method under Cases 3 and 4 (listed in section 5.4.4), the forecast results of which are provided in Figures 5-4 and 5-5 for the Don River basin and Black Creek basin, respectively. These results suggest that, in general, extending data assimilation into the forecast will provide better performance than the open loop model (Case 1). Additionally, many of the scenarios also outperform the traditional data assimilation (Case 2). However, how much of an improvement varies from method to method.

From Figure 5-4, the assimilation methods with state updating, EnKF and PF-SIR, have better performance under Cases 3 and 4. When updating both states and parameters in the forecast, the performances of the EnKF-D, PF-SIRV, and PF-MCMC can be further improved beyond that of Case 2, but only when data assimilation is performed during both the spin up and forecast. Both the PF-SIRV and PF-MCMC methods provide very similar performance here, as was also shown in Moradkhani et al. (2012), with the PF-SIRV providing slightly better performance in the Don River basin under Case 3 assimilating archived streamflow predictions provided using the direct lookup method and the *fdb*. These results suggest that even if observations are not directly available for a basin leading up to the forecast, but there were observations some time in the past, data assimilation could still be used to improve the forecast performance as long as a similar event occurred previously.

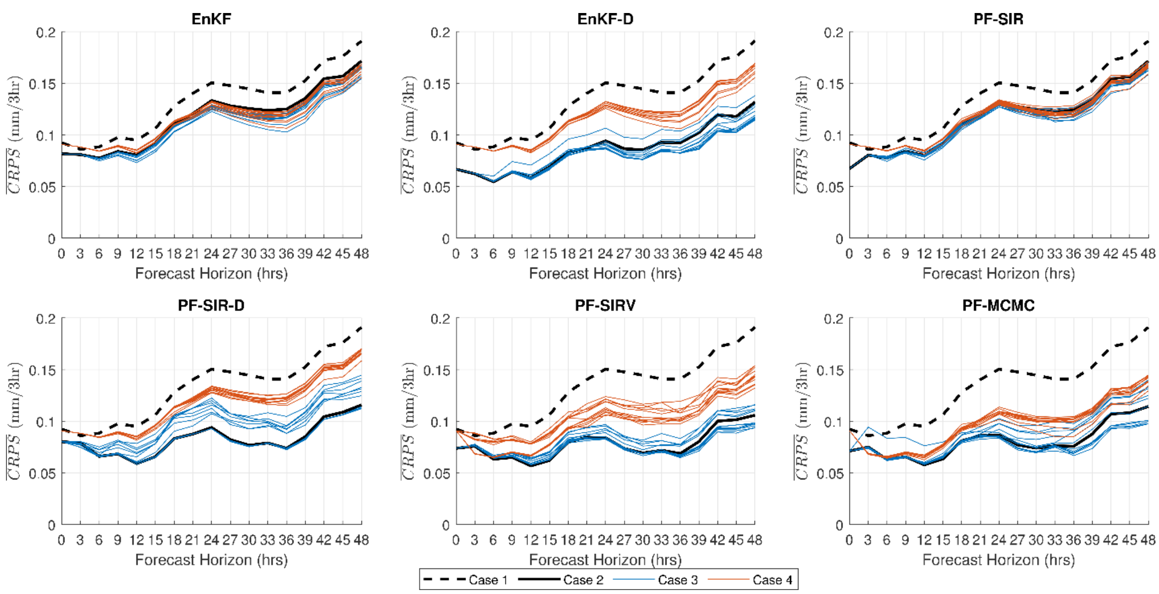

Figure 5-4: Don River basin ensemble forecast performance for the various data assimilation methods and tested cases out to the 48-hour forecast. The dashed black line is the open loop (Case 1), the solid black line is the standard data assimilation framework (Case 2), the blue lines illustrate the various tested scenarios with data assimilation in spin up and forecast (Case 3), and the orange lines illustrate the various tested scenarios with data assimilation only in the forecast (Case 4).

Similar performance is seen for the Black Creek basin model as with the Don River basin model with both the state updating methods being further improved for both Case 3 and 4, and the dual state and parameter updating methods being further improved for Case 3. However, there are some scenarios that also improve the dual updating for Case 4 in the Black Creek basin, and the best case was the same as that of the Don River basin (PF-SIRV). We suspect that the ability for this proposed method to work in these basins is due to them being small urban basins with low memory and short time of concentrations. This makes the basins respond in a similar manner when similar boundary conditions exist, leading to the database lookup method working well. If the basins were larger, it is likely that multiple time lags would need to be considered which would add more complexity to the lookup routine; although it would be possible to do.

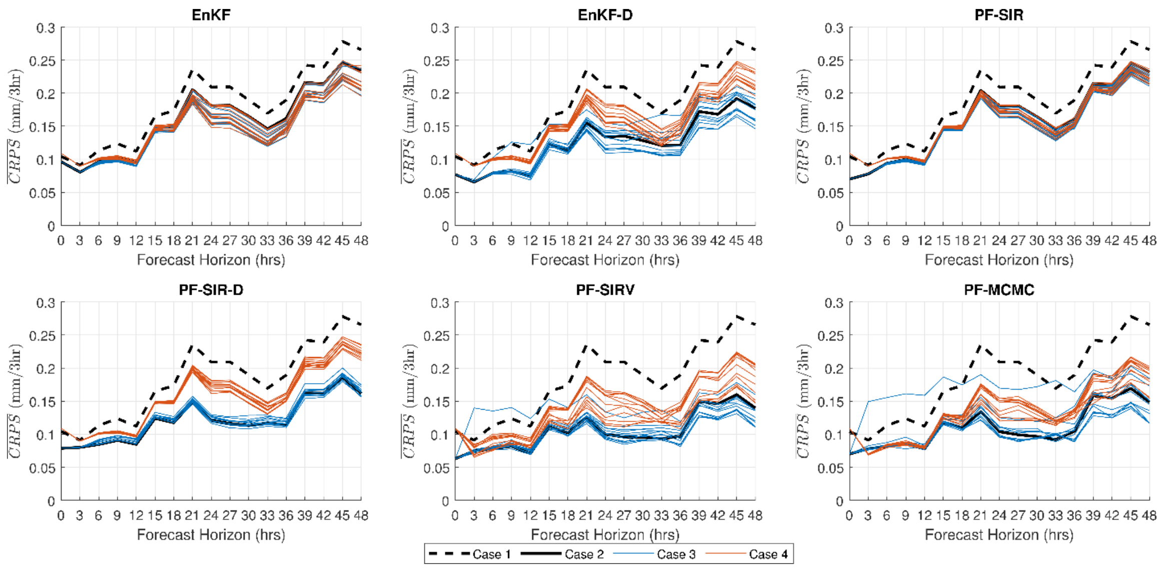

Figure 5-5: Black Creek basin ensemble forecast performance for the various data assimilation methods and tested cases out to the 48-hour forecast. The dashed black line is the open loop (Case 1), the solid black line is the standard data assimilation framework (Case 2), the blue lines illustrate the various tested scenarios with data assimilation in spin up and forecast (Case 3), and the orange lines illustrate the various tested scenarios with data assimilation only in the forecast (Case 4).

To better illustrate the difference between cases, a random event was selected from the Don River basin and its two-day modeled window was shown in Figure 5-6. The results presented here are from the best performing case and scenario options that were previously identified. From Figure 5-6 each case can perform reasonably well, however Case 3 is able to provide better results further into the forecast; this agrees with the results in Figure 5-4.

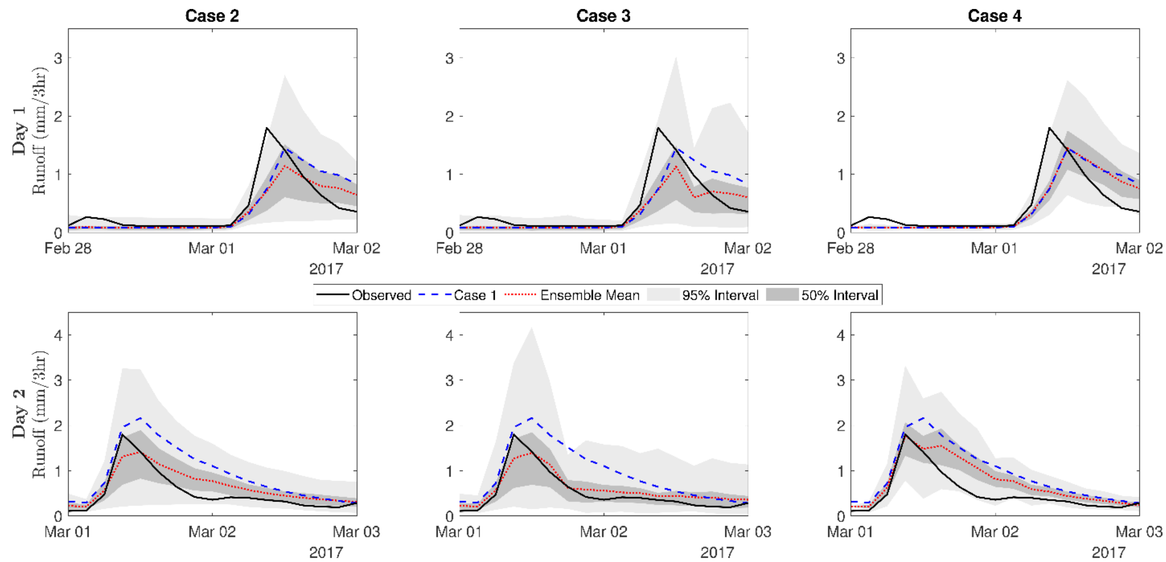

Figure 5-6: Hydrographs for selected event in the Don River basin illustrating ensemble forecast performance of the different cases with PF-SIRV updating states and parameters. The assimilated observation in Cases 3 and 4 was the streamflow value found using the direct lookup method. The RDPS forecasted forcings database (fdb) was used.

To evaluate the various options used for Cases 3 and 4, the results of the scenario options have been summarized in Figures 5-7 and 5-8 for the Don River and Black Creek basins, respectively. From these plots we can make a few conclusions about the best scenario options to use. The first is that PF-SIRV or PF-MCMC with dual updating during the forecast period should be used (Case 3). The results from using each database suggest that the *fdb* is slightly better to use, although it would likely be beneficial to use whichever database is larger, since there will be more archived events available. It is also slightly better to pull the archived streamflow instead of the archived prediction error from the database with either of the lookup methods; further optimization may change this, however. Since this manuscript is presenting a proof of concept for extending data assimilation into the forecast, there is likely some optimization that could be made to further improve the lookup routine and functions. Suggested improvements include more neighbours or a different distance metric for the *k*-NN and more hidden layers and/or neurons for the FNN.

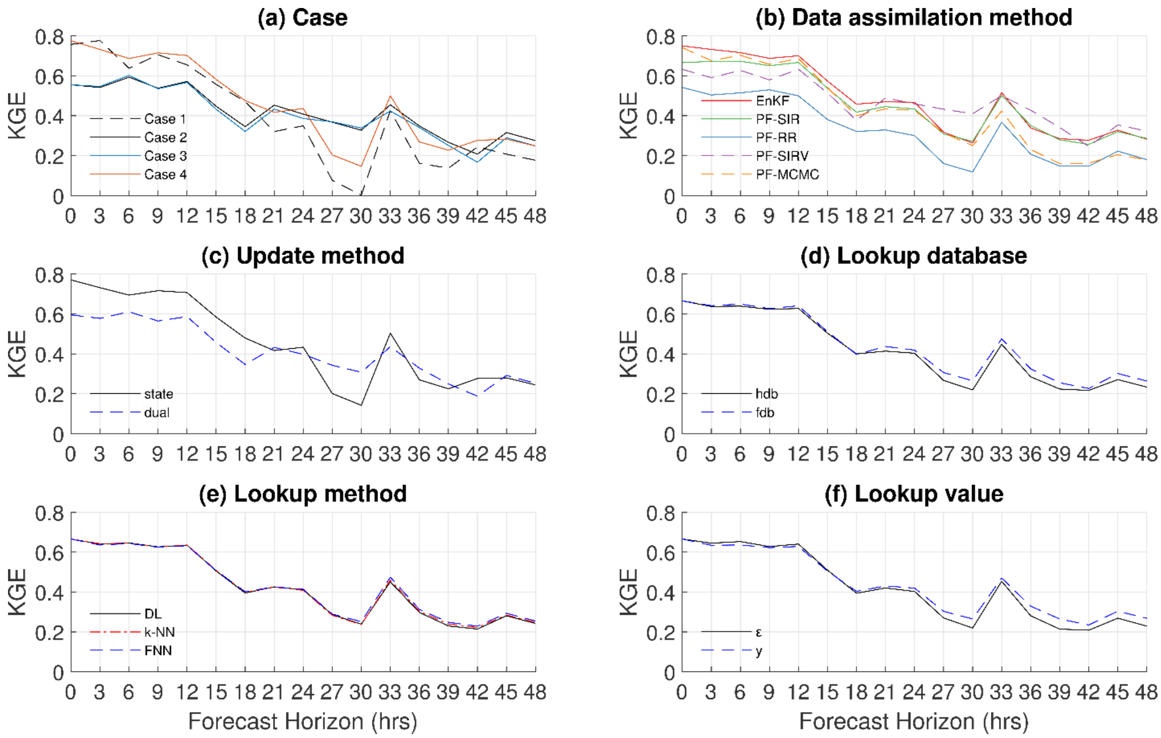

Figure 5-7: Average performance for KGE in the Don River basin for (a) all cases, (b) each data assimilation method, (c) update method, (d) lookup database, (e) lookup method, (f) and lookup value, out to the 48-hour forecast.

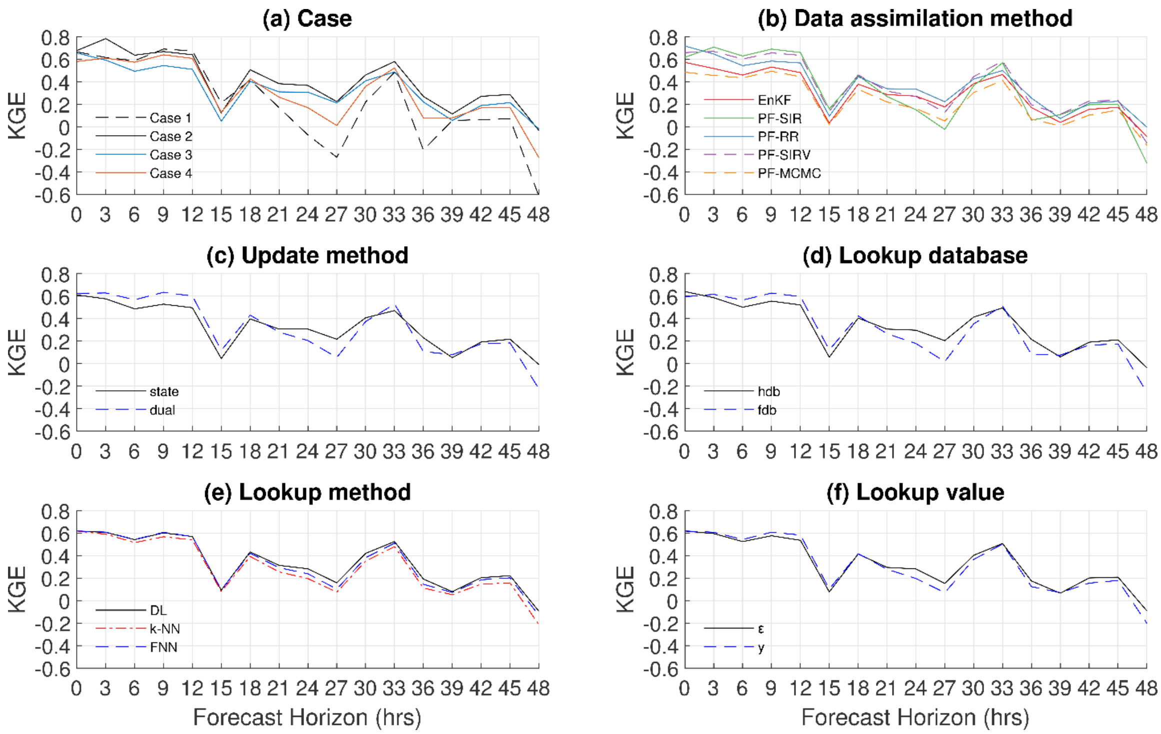

Figure 5-8: Average performance for KGE in the Black Creek basin for (a) all cases, (b) each data assimilation method, (c) update method, (d) lookup database, (e) lookup method, (f) and lookup value, out to the 48-hour forecast.

5.5.4 Effect on model forecast reliability

To evaluate the impact extending data assimilation into the forecast has on forecast reliability, the CRPS was decomposed based on Hersbach (2000) to get the potential CRPS and reliability components. The relative changes to these performance metrics are illustrated in Figure 5-9 for the using PF-SIRV data assimilation method. Here the difference between Case 2 and 3 and Case 2 and 4 are taken so that a positive percent difference indicates an improvement for Case 3 or 4. From these results we see for the Don River basin model that Case 3 scenarios improve the forecast reliability and potential CRPS, increasing forecast reliability by up to 70 percent. However, the Case 4 results in general are only able to improve the potential CRPS results. The Black Creek basin model has similar performance to that of the Don River model, with forecast reliability of Case 3 also improving by up to 78 percent. Unlike the Don River model, however, the Black Creek model also has improved forecast reliability with Case 4. With the potential CRPS for both cases instead having poorer performance.

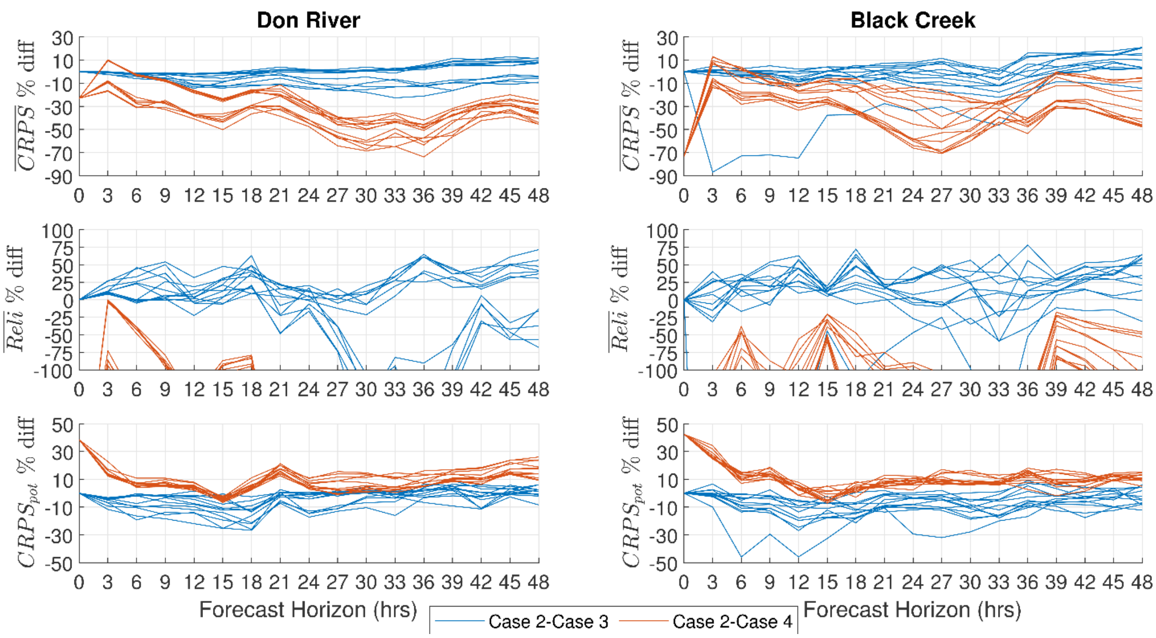

Figure 5-9: Percent difference of CRPS and its decomposed values, CRPS potential and Reliability, for each watershed when comparing cases using the PF-SIRV data assimilation method. The blue indicates the difference between Case 2 and Case 3 for the various tested scenarios with data assimilation in spin up and forecast. The orange indicates the difference between Case 2 and Case 4 for the various tested scenarios with data assimilation only in the forecast. Positive values indicate improvements over Case 2.

Additional probabilistic performance metrics were also evaluated to determine the improvement of Case 3 and Case 4 over Case 2. Those metrics were the 95% Exceedance Ratio (Moradkhani et al., 2006) and Reliability (Renard et al., 2010). The percent difference of each metric was calculated between Case 3 and 2 and Case 4 and 2, so that improvements over the traditional data assimilation (Case 2) could be determined. From Figure 5-10 it is apparent that extending data assimilation into the forecast can both improve the spread of the ensemble and improve overall reliability, particularly under Case 3 where data assimilation is used during both the spin-up period and is extended into the forecast. The Renard Reliability can be improved by up to 67% on the Don River and 28% on Black Creek for Case 3, and the ensemble spread (ER95) can be improved by up to 97% on the Don River and 98% on Black Creek. These results generally agree with those of the decomposed CRPS reliability.

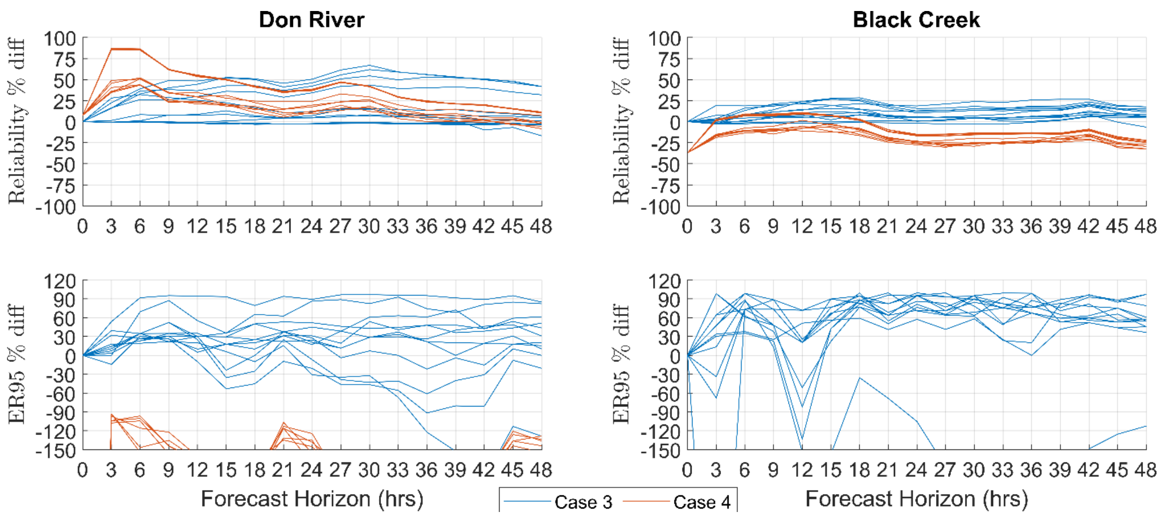

Figure 5-10: Percent difference in Reliability (Renard et al., 2010) and ER95 (Moradkhani et al., 2006) for Don River and Black Creek using PF-SIRV data assimilation method. The blue indicates the percent improvement in Case 3 over Case 2 for the various tested scenarios with data assimilation in spin up and forecast. The orange indicates the percent improvement in Case 4 over Case 2 for the various tested scenarios with data assimilation only in the forecast. Positive values indicate improvements over Case 2.

These results lead to the conclusion that forecast performance can be improved if data assimilation is extended into the forecast by using pseudo-observations obtained from a historical database of the basin model. This improvement is enhanced when observations are available during the forecast spin-up period for assimilation, although they are not required by the method. The proposed method suggests that, if a database can be built which contains the information of how a basin will respond to precipitation and other forcings, data assimilation can be performed using pseudo-observations in the absence of real observation data to improve forecast reliability. Additionally, we suspect that the larger the period of record used to build the database, the larger the improvement in forecast performance should be; as it will be more likely that a historical event occurred which matches the event being forecasted.

5.6 Conclusions

The results of this work provide a proof of concept for the extension of data assimilation into the forecast. We found that when assimilating pseudo-observations pulled from a prebuilt database, which contains information on how a specific watershed modeled using a specific model responds to inputs, the short-term forecast can be improved. Specifically, we showed that for small, highly urbanized basins with short times of concentration, this method works well at improving forecast reliability. The largest improvements were shown when data assimilation was used during both the spin-up period to the forecast as well as into the forecast (Case 3). However, we also showed that forecast performance could be improved with just data assimilation in the forecast (Case 4). Although the improvements shown for Case 4 are not as significant as Case 3, they indicate that assimilating with pseudo-observations in the forecast is still better than the open-loop results (Case 1).

A drawback to extending data assimilation into the forecast is that the method is largely dependent on the ability to build a database. For best results, it is suggested that a large period of record should exist for the watershed of interest. Additionally, since the database is watershed and model specific, if the modelers were interested in multi-modeling, they would need to build a database for each model. Finally, as the model becomes more complex, such as a fully distributed model, there would be an increased computational cost associated with the database and lookup function. There is potential to improve the developed method, however, through further optimization of the lookup method, or expanding the database lookup over a span of several timesteps; which may be needed on larger, more complex basins.

5.7 Acknowledgements

This research was supported by the Natural Science and Engineering Research Council (NSERC) of Canada, grant NSERC Canadian FloodNet (NETGP-451456). The authors are grateful to the Toronto Region Conservation Authority (TRCA), Environment and Climate Change Canada (ECCC), and Water Survey of Canada (WSC) for providing the data. Finally, the authors would like to thank the anonymous reviewers for their thoughtful comments and suggestions which helped to improve the manuscript.

5.8 References

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

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

Agriculture and Agri-Food Canada, 2015. Land Use 1990, 2000 and 2010 (LU1990, LU2000, LU2010) [WWW Document]. URL https://open.canada.ca/data/en/dataset/18e3ef1a-497c-40c6-8326-aac1a34a0dec (accessed 12.1.17).

Altman, N.S., 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am. Stat. 46, 175–185. https://doi.org/10.1080/00031305.1992.10475879

Biondi, D., De Luca, D.L., 2012. A Bayesian approach for real-time flood forecasting. Phys. Chem. Earth, Parts A/B/C 42–44, 91–97. https://doi.org/10.1016/j.pce.2011.04.004

Biondi, D., Todini, E., 2018. Comparing Hydrological Postprocessors Including Ensemble Predictions Into Full Predictive Probability Distribution of Streamflow. Water Resour. Res. 54, 9860–9882. https://doi.org/10.1029/2017WR022432

Clark, M.P., Rupp, D.E., Woods, R.A., Zheng, X., Ibbitt, R.P., Slater, A.G., Schmidt, J., Uddstrom, M.J., 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Adv. Water Resour. 31, 1309–1324. https://doi.org/10.1016/j.advwatres.2008.06.005

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)

Crow, W.T., Ryu, D., 2009. A new data assimilation approach for improving runoff prediction using remotely-sensed soil moisture retrievals. Hydrol. Earth Syst. Sci. 13, 1–16. https://doi.org/10.5194/hess-13-1-2009

DeChant, C.M., Moradkhani, H., 2014. Toward a reliable prediction of seasonal forecast uncertainty: Addressing model and initial condition uncertainty with ensemble data assimilation and Sequential Bayesian Combination. J. Hydrol. 519, 2967–2977. https://doi.org/10.1016/j.jhydrol.2014.05.045

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

Dziubanski, D.J., Franz, K.J., 2016. Assimilation of AMSR-E snow water equivalent data in a spatially-lumped snow model. J. Hydrol. 540, 26–39. https://doi.org/10.1016/j.jhydrol.2016.05.046

Environment and Climate Change Canada, 2017. Canadian Climate Normals [WWW Document]. URL http://climate.weather.gc.ca/climate_normals/ (accessed 10.1.17).

Evensen, G., 2003. The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367. https://doi.org/10.1007/s10236-003-0036-9

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143. https://doi.org/10.1029/94JC00572

Ganguli, P., Coulibaly, P., 2017. Does nonstationarity in rainfall require nonstationary intensity–duration–frequency curves? Hydrol. Earth Syst. Sci. 21, 6461–6483. https://doi.org/10.5194/hess-21-6461-2017

Gordon, N.J., Salmond, D.J., Smith, A.F.M., 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proc. F Radar Signal Process. 140, 107. https://doi.org/10.1049/ip-f-2.1993.0015

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

Hagan, M.T., Demuth, H.B., Beale, M., 1997. Neural Network Design. PWS Publishing Company, Boston, MA.

Han, S., Coulibaly, P., 2019. Probabilistic Flood Forecasting Using Hydrologic Uncertainty Processor with Ensemble Weather Forecasts. J. Hydrometeorol. 20, 1379–1398. https://doi.org/10.1175/JHM-D-18-0251.1

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

Huang, C., Newman, A.J., Clark, M.P., Wood, A.W., Zheng, X., 2017. Evaluation of snow data assimilation using the ensemble Kalman filter for seasonal streamflow prediction in the western United States. Hydrol. Earth Syst. Sci. 21, 635–650. https://doi.org/10.5194/hess-21-635-2017

Kornelsen, K.C., Coulibaly, P., 2016. Investigating the Relationship Between Precipitation and Model Parameter Distributions During Calibration: Initial Results from 72 Canadian Basins., in: 2016 HEPEX Workshop - Ensemble for Better Hydrological Forecasts. Québec, Quebec, Canada.

Krzysztofowicz, R., 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resour. Res. 35, 2739–2750. https://doi.org/10.1029/1999WR900099

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

Leisenring, M., Moradkhani, H., 2012. Analyzing the uncertainty of suspended sediment load prediction using sequential data assimilation. J. Hydrol. 468–469, 268–282. https://doi.org/10.1016/j.jhydrol.2012.08.049

Li, Y., Ryu, D., Western, A.W., Wang, Q.J., 2013. Assimilation of stream discharge for flood forecasting: The benefits of accounting for routing time lags. Water Resour. Res. 49, 1887–1900. https://doi.org/10.1002/wrcr.20169

Liu, J.S., Chen, R., 1998. Sequential Monte Carlo Methods for Dynamic Systems. J. Am. Stat. Assoc. 93, 1032–1044. https://doi.org/10.1080/01621459.1998.10473765

Liu, Y., Weerts, A.H., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., van Dijk, A.I.J.M., van Velzen, N., He, M., Lee, H., Noh, S.J., Rakovec, O., Restrepo, P., 2012. Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. Hydrol. Earth Syst. Sci. 16, 3863–3887. https://doi.org/10.5194/hess-16-3863-2012

Matheson, J.E., Winkler, R.L., 1976. Scoring Rules for Continuous Probability Distributions. Manage. Sci. 22, 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087

Meng, S., Xie, X., Liang, S., 2017. Assimilation of soil moisture and streamflow observations to improve flood forecasting with considering runoff routing lags. J. Hydrol. 550, 568–579. https://doi.org/10.1016/j.jhydrol.2017.05.024

Moore, R.J., 1985. The probability-distributed principle and runoff production at point and basin scales. Hydrol. Sci. J. 30, 273–297. https://doi.org/10.1080/02626668509490989

Moradkhani, H., 2008. Hydrologic remote sensing and land surface data assimilation. Sensors 8, 2986–3004. https://doi.org/10.3390/s8052986

Moradkhani, H., DeChant, C.M., Sorooshian, S., 2012. Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method. Water Resour. Res. 48. https://doi.org/10.1029/2012WR012144

Moradkhani, H., Hsu, K., Hong, Y., Sorooshian, S., 2006. Investigating the impact of remotely sensed precipitation and hydrologic model uncertainties on the ensemble streamflow forecasting. Geophys. Res. Lett. 33, L12401. https://doi.org/10.1029/2006GL026855

Moradkhani, H., Sorooshian, S., Gupta, H. V., Houser, P.R., 2005. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour. 28, 135–147. https://doi.org/10.1016/j.advwatres.2004.09.002

Murphy, A.H., 1973. A New Vector Partition of the Probability Score. J. Appl. Meteorol. 12, 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

Najafi, M.R., Moradkhani, H., 2016. Ensemble Combination of Seasonal Streamflow Forecasts. J. Hydrol. Eng. 21, 04015043. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001250

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

Olsson, J., 1998. Evaluation of a scaling cascade model for temporal rain- fall disaggregation. Hydrol. Earth Syst. Sci. 2, 19–30. https://doi.org/10.5194/hess-2-19-1998

Olsson, J., 1995. Limits and characteristics of the multifractal behaviour of a high-resolution rainfall time series. Nonlinear Process. Geophys. 2, 23–29. https://doi.org/10.5194/npg-2-23-1995

Parrish, M.A., Moradkhani, H., DeChant, C.M., 2012. Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation. Water Resour. Res. 48, n/a-n/a. https://doi.org/10.1029/2011WR011116

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Mon. Weather Rev. 133, 1155–1174. https://doi.org/10.1175/MWR2906.1

Rakovec, O., Weerts, A.H., Hazenberg, P., Torfs, P.J.J.F., Uijlenhoet, R., 2012. State updating of a distributed hydrological model with Ensemble Kalman Filtering: effects of updating frequency and observation network density on forecast accuracy. Hydrol. Earth Syst. Sci. 16, 3435–3449. https://doi.org/10.5194/hess-16-3435-2012

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

Samuel, J., Coulibaly, P., Dumedah, G., Moradkhani, H., 2014. Assessing model state and forecasts variation in hydrologic data assimilation. J. Hydrol. 513, 127–141. https://doi.org/10.1016/j.jhydrol.2014.03.048

Samuel, J., Coulibaly, P., Metcalfe, R. a., 2012. Identification of rainfall-runoff model for improved baseflow estimation in ungauged basins. Hydrol. Process. 26, 356–366. https://doi.org/10.1002/hyp.8133

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

Smith, A.F.M., Gelfand, A.E., 1992. Bayesian Statistics without Tears: A Sampling–Resampling Perspective. Am. Stat. 46, 84–88. https://doi.org/10.1080/00031305.1992.10475856

Sun, L., Nistor, I., Seidou, O., 2015. Streamflow data assimilation in SWAT model using Extended Kalman Filter. J. Hydrol. 531, 671–684. https://doi.org/10.1016/j.jhydrol.2015.10.060

Thiboult, A., Anctil, F., Boucher, M. -a., 2015. Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrol. Earth Syst. Sci. Discuss. 12, 7179–7223. https://doi.org/10.5194/hessd-12-7179-2015

Todini, E., 2008. A model conditional processor to assess predictive uncertainty in flood forecasting. Int. J. River Basin Manag. 6, 123–137. https://doi.org/10.1080/15715124.2008.9635342

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

Unger, D.A., 1985. A method to estimate the continuous ranked probability score, in: Ninth Conference on Probability and Statistices in Atmospheric Sciences. American Meteorological Society, Virginia Beach, VA, pp. 206–213.

Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43, n/a-n/a. https://doi.org/10.1029/2005WR004838

Xie, X., Zhang, D., 2010. Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter. Adv. Water Resour. 33, 678–690. https://doi.org/10.1016/j.advwatres.2010.03.012

Yan, H., DeChant, C.M., Moradkhani, H., 2015. Improving soil moisture profile prediction with the particle filter-Markov chain Monte Carlo method. IEEE Trans. Geosci. Remote Sens. 53, 6134–6147. https://doi.org/10.1109/TGRS.2015.2432067

Yan, H., Moradkhani, H., 2016. Combined assimilation of streamflow and satellite soil moisture with the particle filter and geostatistical modeling. Adv. Water Resour. 94, 364–378. https://doi.org/10.1016/j.advwatres.2016.06.002

CHAPTER 6 Conclusions and Recommendations

6.1 Conclusions

This thesis presented advanced methods that can be used to improve data assimilation techniques for hydrologic modeling and forecasting. First, it was shown that assimilating data products using the Ensemble Kalman Filter (EnKF; Evensen 1994) can improve hydrologic modeling and forecasting in an urban watershed. Next, remotely sensed soil moisture assimilation was further evaluated in regards to its limits for assimilation into urban and semi-urban watersheds. Thirdly, the Mutual information Ensemble Kalman filter (MIEnKF) was developed as a way to better update ungauged basin states when using observations from a gauged basin in a semi-distributed hydrologic model. Finally, forecast reliability was improved by extending data assimilation into the short-term forecast through the use of a pre-built database; with the database containing information on how the selected watershed responds to initial conditions and forcing data.

The main conclusions of the thesis are summarized as follows.

6.1.1 Using data products to improve forecasting in an urban basin

- Simple conceptual models like GR4J (Perrin et al., 2003) and HyMod (Boyle, 2001) can be used to model urban watersheds with similar or better performance as more complex models like SAC-SMA (Burnash, 1995; Burnash et al., 1973).
- Assimilating data products like SMOS L2 soil moisture (Kerr et al., 2012, 2010) and SNODAS snow water equivalent (National Operations Hydrologic Remote Sensing Center, 2004) can be beneficial for hydrologic forecasting in urban watersheds.

- A combination of streamflow and soil moisture assimilation can provide the best hydrologic forecast performance in the Don River basin.

6.1.2 Limits for assimilating soil moisture in urban basins

- Through testing of synthetic urban catchments, it was determined that an imperviousness threshold exists, beyond which it is not beneficial to assimilate remotely sensed soil moisture data. This threshold is a function of the rainfall volume, internal basin routing, and the basin's imperviousness.
- A surrogate model was developed using the NRCS-CN (Natural Resources Conservation Service (NRCS), 2004), which can be used to quickly determine the threshold value for a watershed using information about the watershed.
- The method was then validated by modeling several real-world urban and semi-urban watersheds. These real-world experiments involved assimilating SMOS soil moisture data and determining whether there was a negative impact on model performance. Based on the results of both the synthetic and real-world tests, a general limiting imperviousness range was identified to be between 65 and 75 percent.
- The results showed that the method was reasonably accurate in identifying the imperviousness threshold for each real-world watershed; with an accuracy of 87 percent in determining if there will or will not be a negative impact.

6.1.3 Combining information theory and data assimilation

- As a way to account for non-linear dependencies, data assimilation and information theory were combined; specifically the EnKF (Evensen, 1994) and mutual information (Keum et al., 2017; Shannon, 1948; Singh, 1997) and the new method was termed the MIEnKF. The purpose of this combination is to better aid in data assimilation for ungauged basins in semi-distributed hydrologic models.

- The Lorenz 63 model (Lorenz, 1963) was used as a testbed to compare the EnKF and MIEnKF. The results of these tests showed that in the traditional case of data assimilation, that being when a measured variable is used to update the predicted value for that variable, the methods provided similar performance. However, when a measured variable was used to update another variable in which it had non-linear dependencies, the MIEnKF was able to provide more useful updates.
- The MIEnKF was then tested using a simple semi-distributed hydrologic model and a more complex semi-distributed hydrologic model and again found similar results. These results show that the method is valid for different models, is scalable to more discretized models, is usable for both urban and semi-urban basins, and works at daily or hourly timescales.

6.1.4 Using past information to improve the reliability of forecasts

- It is possible to store the past information on how a hydrologic model for a specific watershed responded to input forcing data given the model's initial conditions in a database. This database can be used to generate a probability density function (pdf) of observations that is conditional on previously known information which can then be assimilated back into a hydrologic model.
- To extend data assimilation into the short-term forecast, given some meteorological forecast and the initial conditions for a timestep, a pdf can be generated using the

pre-built database, which can then be assimilated during the forecast. This method allows for assimilation to occur when observations are not available.

- Through testing with the EnKF and various particle filter implementations (Gordon et al., 1993; Leisenring and Moradkhani, 2012; Moradkhani et al., 2012, 2005), this extension of data assimilation into the short-term forecast was shown to improve the ensemble spread and enhance the forecast reliability.

6.2 Recommendations for Future Research

The research presented in this thesis is aimed towards advancing data assimilation methodology in urban and semi-urban watersheds. The results demonstrate that improvements can be made to existing methods which can aid in improving hydrologic and flood forecasting. However, these methods could still be improved upon, and there are several recommendations for future work which can aid in this.

One recommendation is to test the impact of assimilating data products into semi- and fully-distributed hydrologic models of urban basins to determine if the same improvements to forecasting are present. Additionally, using a more advanced snowmelt routine when assimilating SNODAS snow water equivalent could provide improved results over those of the degree-day snowmelt routine. Finally, these tests should be run on hourly or sub-hourly scales in the urban watersheds to better account for the time of concentration as well as to show validity for flood forecasting.

It is recommended that the theoretical basis for the MIEnKF method be further developed. Additionally, further improvements to this method could be made to better allow for updating parameters of ungauged basins.

Finally, it is recommended that the database method used for extending data assimilation into the forecast be further evaluated using the EPFM (Abbaszadeh et al., 2018) and the HEAVEN (Abbaszadeh et al., 2019) data assimilation methods. It should also be evaluated using larger watersheds in which the impact of the time of concentration will likely require modifications to the database. Further work should also be done to test this method on both semi- and fully-distributed hydrologic models.

6.3 References

Abbaszadeh, P., Moradkhani, H., Daescu, D.N., 2019. The Quest for Model Uncertainty Quantification: A Hybrid Ensemble and Variational Data Assimilation Framework. Water Resour. Res. 55, 2407–2431. https://doi.org/10.1029/2018WR023629

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

Boyle, D.P., 2001. Multicriteria calibration of hydrologic models. University of Arizona.

Burnash, R.J.C., 1995. The NWS River Forecast System - Catchment Model, in: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications.

Burnash, R.J.C., Ferral, R.L., McGuire, R.A., 1973. A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers. U.S. Department of Commerce, National Weather Service, and State of California Department of Water Resources.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143. https://doi.org/10.1029/94JC00572

Gordon, N.J., Salmond, D.J., Smith, A.F.M., 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proc. F Radar Signal Process. 140, 107. https://doi.org/10.1049/ip-f-2.1993.0015

Kerr, Y.H., Waldteufel, P., Richaume, P., Wigneron, J.P., Ferrazzoli, P., Mahmoodi, A., Al Bitar, A., Cabot, F., Gruhier, C., Juglea, S.E., Leroux, D., Mialon, A., Delwart, S., 2012. The SMOS Soil Moisture Retrieval Algorithm. IEEE Trans. Geosci. Remote Sens. 50, 1384–1403. https://doi.org/10.1109/TGRS.2012.2184548

Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.-J., Font, J., Reul, N., Gruhier, C., Juglea, S.E., Drinkwater, M.R., Hahne, A., Martín-Neira, M., Mecklenburg, S., 2010. The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle. Proc. IEEE 98, 666–687. https://doi.org/10.1109/JPROC.2010.2043032

Keum, J., Kornelsen, K., Leach, J., Coulibaly, P., 2017. Entropy Applications to Water Monitoring Network Design: A Review. Entropy 19, 613. https://doi.org/10.3390/e19110613

Leisenring, M., Moradkhani, H., 2012. Analyzing the uncertainty of suspended sediment load prediction using sequential data assimilation. J. Hydrol. 468–469, 268–282. https://doi.org/10.1016/j.jhydrol.2012.08.049

Lorenz, E.N., 1963. Deterministic Nonperiodic Flow. J. Atmos. Sci. 20, 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

Moradkhani, H., DeChant, C.M., Sorooshian, S., 2012. Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method. Water Resour. Res. 48. https://doi.org/10.1029/2012WR012144

Moradkhani, H., Hsu, K.-L., Gupta, H., Sorooshian, S., 2005. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. Water Resour. Res. 41, W05012. https://doi.org/10.1029/2004WR003604

National Operations Hydrologic Remote Sensing Center, 2004. Snow Data Assimilation System (SNODAS) Data Products at NSIDC, Version 1 [January 2011 - December 2015]. Boulder, Colorada, USA. NSIDC: National Snow and Ice Data Center. https://doi.org/http://dx.doi.org/10.7265/N5TB14TC

Natural Resources Conservation Service (NRCS), 2004. Chapter 10 Estimation of Direct Runoff from Storm Rainfall, in: Part 630 Hydrology, National Engineering Handbook. U.S. Department of Agriculture, Washington, D.C.

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

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423, 623–656.

Singh, V.P., 1997. The use of entropy in hydrology and water resources. Hydrol. Process. 11, 587–626. https://doi.org/10.1002/(SICI)1099-1085(199705)11:6<587::AID-HYP479>3.0.CO;2-P