DUAL ENTROPY MULTI-OBJECTIVE OPTIMIZATION

APPLICATION TO HYDROMETRIC NETWORK DESIGN

DUAL ENTROPY MULTI-OBJECTIVE OPTIMIZATION

APPLICATION TO HYDROMETRIC NETWORK DESIGN

By CONNOR WERSTUCK, B.A.Sc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Applied Science

McMaster University MASTER OF APPLIED SCIENCE (2016) Hamilton, Ontario

(Civil Engineering)

TITLE: Dual Entropy Multi-Objective Optimization Application to Hydrometric Network Design

AUTHOR: Connor Werstuck, B.A.Sc. (University of Waterloo)

SUPERVISOR: Dr. Paulin Coulibaly (McMaster University)

NUMBER OF PAGES: xv, 84

Abstract

Water resources managers rely on information collected by hydrometric networks without a quantitative way to assess their efficiency, and most Canadian water monitoring networks still do not meet the minimum density requirements. There is also no established way to quantify the importance of each existing station in a hydrometric network. This research examines the properties of Combined Regionalization Dual Entropy Multi-Objective Optimization (CR-DEMO), a robust network design technique which combines the merits of information theory and multi-objective optimization. Another information theory based method called transinformation (TI) which can rank the contribution of unique information from each specific hydrometric station in the network is tested for use with CR-DEMO. When used in conjunction, these methods can not only provide an objective measure of network efficiency and the relative importance of each station, but also allow the user to make recommendations to improve existing hydrometric networks across Canada. The Ottawa River Basin, a major Canadian watershed in Ontario and Quebec, was selected for analysis. Various regionalization methods which could be used in CR-DEMO such as distance weighting and a rainfall runoff model were compared in a leave one out cross validation. The effect of removing stations with regulated and unnatural flow regimes from the regionalization process is also tested. The analysis is repeated on a smaller tributary of the Ottawa River Basin, the Madawaska Watershed, to examine scale effects in TI analysis and CR-DEMO application. In this study, tests were conducted to determine whether to include stations outside of the river basin in order to provide more context to the basin boundaries. It was

found that the TI analysis complemented CR-DEMO well and it provided a detailed station ranking which was supported by CR-DEMO results. The inverse distance weighting drainage area ratio method was found to provide more accurate regionalization results compared to the rainfall-runoff model, and was thus chosen for CR-DEMO. Regionalization was shown to be more accurate when the regulated basins were omitted using leave one out cross validation. It was discovered that CR-DEMO is sensitive to scaling because some sub-basins which are relatively "well-equipped" compared to others in dire conditions may be penalized. The TI analysis was not as sensitive to scaling. Including stations outside of the Ottawa River Basin improved the information density and regionalization accuracy in the Madawaska Watershed because they provided context to sparse areas. Finally, Pareto optimal network solutions for both the Ottawa River Basin and the Madawaska Watershed were presented and analyzed. A number of optimal networks are proposed for each watershed along with "hot-spots" where new stations should be added whatever the end users' choice of network.

Acknowledgments

This research was supported jointly by Ontario Power Generation (OPG) and Natural Science and Engineering Research Council (NSERC) of Canada. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada, and by datasets from Environment Canada, Hydro-Quebec, and OPG. I am grateful to Dr. Joshua Kollat (Penn State University) who developed the ε-hBOA, and provided the source codes. I am also grateful to Dr. Jos Samuel for writing the original DEMO program used this analysis.

In addition to this I would like to thank my office mates James Leach and Dr. Jongho Keum for showing me the ropes and always being there to answer questions and to provide general support. Thank you to Dr. Kurt Kornelsen for being very helpful and knowledgeable with virtually everything hydrology related. Also thank you to my supervisor, Dr. Paulin Coulibaly, for guidance, support and help editing my submissions.

Finally, thank you to my family Dr. Geoff Werstuck, Michele Werstuck and Hayley Werstuck, and my girlfriend Julia Piccioni for your love and support along the way.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Translation	Explanation
DEMO	Dual Entropy Multi-Objective Optimization	An optimization process which uses the principles of information theory for network design.
CR-DEMO	Combined Regionalization Dual Entropy Multi-Objective Optimization.	DEMO combined with the regionalization process where synthetic data is generated at potential station locations.
DEM	Digital Elevation Model	A geographic information file containing elevation data.
EC	Environment Canada	
WMO	World Meteorological Organization	
IDW	Inverse distance weighting	A regionalization method. Discussed in the regionalization section.
IDW-DAR	Inverse distance weighting drainage area ratio	IDW normalizing by drainage area. Shown in Equations 6 and 7.
HBV	Hydrologiska Byråns Vattenbalansavdelning	A commonly used rainfall runoff model developed by Merz, R., & Blöschl, G. (2004).
HYDAT	Hydrological Database	Environment Canada's National Water Data Archive
MAC-HBV	McMaster University-Hydrologiska Byråns Vattenbalansavdelning	A modified HBV model from McMaster University by Samuel et al. (2011).
IHA	Indicators of Hydrologic Alteration	Metrics used in Monk et al. (2011) and Leach et al. (2015) to characterize streamflow.
TI	Transinformation index	The normalized transinformation value. A measure of how well a timeseries can be replicated given existing data.
NSE	Nash Sutcliffe Efficiency	A commonly used measure for the accuracy of modeled streamflow.
RR	Runoff ratio	A streamflow signature relating total rainfall to total runoff.
FDC	Slope of the flow duration curve	A streamflow signature describing the probability distribution of streamflow.
BI	Baseflow index	A streamflow signature parameter from baseflow separation.
SE	Streamflow elasticity	A streamflow signature relating the change in mean annual streamflow to mean annual precipitation.
SDR	Snow day ratio	A streamflow signature describing the fraction of precipitation events which are snow.
RLD	Rising limb density	A streamflow signature describing the fraction of days when daily average streamflow increased.
LOOCV	Leave one out cross validation	A method of measuring regionalization accuracy.

Declaration of Academic Achievement

Thesis Structure

This thesis has been assembled in accordance with the McMaster University School of Graduate Studies' "Guide for the Preparation of Master's and Doctoral Theses" as a sandwich thesis. The thesis is divided into four chapters. Chapter 1 contains background information including research objectives and a short literature review on hydrometric network design. Chapters 2 and 3 are manuscripts submitted for publication of research pertaining to the application of Dual Entropy Multi-Objective Optimization (DEMO), an information theory based method of network design. The first manuscript focuses on the regionalization technique used in the method and the presentation of transinformation analysis as a tool to be used with DEMO in hydrometric network design. The second manuscript discusses the effects of scale in the TI analysis and CR-DEMO application using the Madawaska Watershed, a sub-basin of the Ottawa River Basin. Chapter 4 details the recommendations and conclusions drawn by the research.

Author Contributions

Chapter 2

Title:	Hydrometric Network Design Using Dual Entropy Multi-Objective Optimization in the Ottawa River Basin
Authors:	Connor Werstuck, Paulin Coulibaly
Submitted to:	Hydrology Research

Contributions: Connor Werstuck obtained, organized and analyzed the hydrometric data, calibrated the necessary models and carried out the transinformation and DEMO analyses. He also interpreted the results and drew conclusions regarding the research objectives. Dr. Paulin Coulibaly provided expert guidance regarding the direction of the research and the paper, and edited Connor Werstuck's text prior to submission. This research was done using the DEMO code originally written by Dr. Jos Samuel and Dr. Joshua Kollat.

Chapter 3

Title: Assessing Scale Effects on Hydrometric Network Design Using Entropy and Multi-Objective Methods

Authors: Connor Werstuck, Paulin Coulibaly

Submitted to: Journal of the American Water Resources Association

Contributions: Connor Werstuck obtained, organized and analyzed the hydrometric data and carried out the transinformation and DEMO analyses. He also interpreted the results and drew conclusions regarding the research objectives. Dr. Paulin Coulibaly provided expert guidance regarding the direction of the research and the paper, and edited Connor Werstuck's text prior to submission. This

research was done using the DEMO code originally written by Dr. Jos Samuel and Dr. Joshua Kollat.

Copyright

The first paper entitled Hydrometric Network Design Using Dual Entropy Multi-Objective Optimization in the Ottawa River Basin has been submitted to Hydrology Research and is currently under review. This journal allows authors to freely share their publications and submissions in theses or dissertations. The second paper entitled Assessing Scale Effects on Hydrometric Network Design Using Entropy and Multi-Objective Methods is pending submission to the Journal of the American Water Resources Association.

1. Introduction

1.1 Background

Hydrometric data is the most important tool that water resources engineers have for making water allocation decisions. This data is used extensively in frequency analysis, model calibration and forecasting and infrastructure design, thus it is extremely important that it is as complete and accurate as possible. Many of Canada's most important watersheds are currently monitored using networks which are deficient (Mishra & Coulibaly, 2010). Therefore it is necessary to consider various network design and augmentation techniques (Mishra & Coulibaly, 2009) and recommend improvements to the existing Canada hydrometric monitoring infrastructure. This thesis investigates the application of Combined Regionalization Dual Entropy Multi-Objective Optimization (CR-DEMO), a network design method introduced by Samuel et al. (2013) which combines the utility of Shannon's information theory (Shannon, 1948) and a robust multi-objective optimization algorithm (Kollat et al., 2008).

There is currently no established method of ranking the information contribution from each existing station in a network. A quantitative way to do this would be extremely useful to hydrometric network operators. This research investigates transinformation (TI) analysis as a solution, as well as way to provide useful regionalized TI maps which can be used in conjunction with the CR-DEMO method. Additionally, the regionalization step of CR-DEMO can be accomplished using a number of different methods. Thus far for streamflow network design the inverse distance weighting drainage area ratio (IDW-

DAR) method has been used, but there has been no research regarding regionalization techniques and their effects on joint entropy. This thesis compares IDW-DAR to McMaster University-Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV), a rainfall-runoff model, in terms of regionalization accuracy and CR-DEMO results. The effects of including regulated sub-basins in the regionalization step are also considered. Finally, due to the fact that CR-DEMO is a relatively new technique there has been no research done regarding its properties such as scaling. This research considers both the Ottawa River Basin (ORB) and one of its sub-basins, the Madawaska Watershed, in order to investigate the scaling properties of CR-DEMO and TI analysis.

1.2 Research Objectives and Outputs

The enclosed research examines and discusses the following:

1) The utility of transinformation (TI) analysis to output TI maps and rank existing stations by information contribution for use in conjunction with CR-DEMO.
2) Which regionalization technique should be used in the operation of CR-DEMO for a daily average streamflow network in Ontario and Quebec.
3) How to deal with controlled or regulated streamflow in the area of interest when using CR-DEMO and TI analysis.
4) The effects of changing the scale in TI analysis and CR-DEMO application.
5) The justification for including stations outside of the area of interest in order to improve regionalization and transinformation for a sparsely monitored basin when using CR-DEMO and TI analysis.

In the first study, the hydrometric network of the ORB, a heavily regulated basin on the Ontario Quebec border, is analyzed. CR-DEMO and transinformation analysis are conducted and recommendations for new monitoring stations are recorded. The relationship between these two distinct information theory based analyses is discussed, and it is concluded that TI analysis is useful for comparing to and complementing CR-DEMO results. Various regionalization techniques are compared, including IDW-DAR and a rainfall-runoff model. Through a leave one out cross validation, it was found that IDW-DAR output superior regionalized data, and that omitting regulated stations improved the regionalization accuracy.

In a second study, a similar analysis was conducted on the Madawaska Watershed, a sub-basin of the ORB. The primary purpose of this study was to investigate the scaling properties of these analyses. Conclusions were also drawn about the legitimacy of including stations outside of the river basin of interest for regionalization purposes. It was found that in this particular study, CR-DEMO was sensitive to scaling due to extreme information deficit areas biasing the recommended station locations in the larger study. TI analysis was not significantly affected by scaling. It was found that including stations which were not in the ORB added information content to the regionalization process.

In addition to these outputs, a short review of the progression of information theory and multi-objective optimization in hydrometric network design is included. This review should provide some perspective on recent progress in the field and some background on the methods upon which this thesis is based.

1.3 Literature Review

Shannon's information theory (Shannon, 1948) introduced the concept of entropy as the amount of information in bits produced by a Markov process. He provided a unique entropy function which describes the amount of information contained in any known probability distribution. Lindley (1956) expanded on this research by introducing mutual information, a measure used to compare the dependence between two sets of results without knowing their prior distributions. Mutual information is another name for transinformation which is used frequently in this thesis. Jaynes (1957) added the principle of maximum entropy which states that the probability distribution describing prior data that has the maximum entropy is the least committal to missing data and makes the fewest assumptions. These contributions formed the basis of information theory which has since been applied to many different fields of science and engineering.

Prior to the introduction of entropy theory in network design, hydrometric networks were designed using a number of different methods. Fiering (1965) used nonlinear integer programming to design a network that could most efficiently estimate the annual mean streamflow at a number of known sites. Maddock (1974) estimated the information content of a station based on the reciprocal of the variance of the mean annual flow estimate. He then maximized this value and used the linear correlation between station data to determine which stations should be removed from a network due to budget constraints. Moss and Karlinger (1974) proposed a network design method that involved regression analysis between stations and considered the uncertainty in streamflow parameter estimation. Most network design was based on minimizing the error between a

network prediction and a known value or reducing uncertainty in prediction. Many of these methods are still used today such as statistical methods, spatial interpolation techniques, optimization methods, methods based on physiographic characteristics, methods driven by sampling strategy and the user survey approach (Mishra & Coulibaly, 2009).

Meanwhile, information theory was used more and more in water resources research. For example, Amorocho and Espildora (1973) conducted entropy analysis on historic time series to quantify streamflow uncertainty. They also used transinformation between simulated and measured streamflow as a model assessment tool. Sonuga (1976) used the principle of maximum entropy to find the minimally biased probability distribution of runoff given rainfall. Singh and Rajagopal (1987) discussed the application of the principle of maximum entropy as a means of deriving accurate distributions for many different processes, such as deriving frequency distributions, parameter estimation, evaluation of networks and uncertainty assessment.

Entropy was introduced into hydrologic network design by Husain (1979). In sample networks in British Columbia, Husain generated regionalized streamflow values on a grid, estimated the discrete probability distribution of each location and found the network of ten stations with the maximum joint entropy. This method was groundbreaking because it provided an objective and robust way to compare the information content of stations and networks. Husain (1989) refined his original method in a number of ways, for example, he tested other probability distribution estimators for

streamflow such as the Gamma function. Krstanovic and Singh (1992) published a two part paper which used the principle of maximum entropy and transinformation to choose which stations to remove from an existing Louisiana precipitation network. Using this methodology they produced a map of transinformation contours. Yang and Burn (1994) designed a streamflow network in Manitoba based on a measure called directional information transfer. This entropy based metric is calculated between pairs of stations and it represents the fraction of information which can be transferred between stations, which can be interpreted as dependence. This allowed for the stations to be grouped, and unimportant group members to be removed. Ozkul et al. (2000) designed a water quality monitoring network and Chen et al. (2008) designed a precipitation network both using a stepwise optimization. The station with the highest entropy was selected first, then stations were added one by one to maximize the joint entropy in the network with two, three, four stations and so on. Up until this time, information theory based network optimizations were single objective or stepwise.

Multi-objective optimization algorithms were being used in network design, but not combined with entropy theory. For example, Kollat and Reed (2007) used the epsilon-dominance non-dominated sorted genetic algorithm II (ε-NSGAII) to design long term groundwater monitoring networks based on cost, contaminant estimation error, uncertainty and contaminant mass error. Later, Kollat et al. (2008) studied the next generation epsilon-dominance hierarchical Bayesian optimization algorithm (ε-hBOA) and its application to groundwater network design and found that it was superior for network design due to its model building capability.

Alfonso et al. (2010a) began using total correlation as a network measure representing redundant information in conjunction with joint entropy. This is defined as the difference between the sum of the individual station entropies and the joint entropy of the network. Alfonso et al. (2010b) performed multi-objective optimization minimizing total correlation and maximizing joint entropy to find a set of Pareto-optimal configurations for a water level monitoring network. A similar approach was proposed by Li et al. (2012), who included joint entropy, transinformation between groups and total correlation in a weighted single objective optimization.

Samuel et al. (2013) introduced CR-DEMO, a robust network design method combining the benefits of information theory and multi-objective optimization. First, potential station locations are identified and synthetic data is regionalized to these locations. Then, the ε-hBOA algorithm is run considering the number of existing and potential stations. The algorithm outputs the Pareto front of non-dominated solutions with maximum joint entropy and minimum total correlation. This method has been used for a number of different network design applications such as streamflow networks, precipitation networks, snow depth networks and groundwater level networks. Kornelsen and Coulibaly (2015) designed a soil moisture monitoring network and showed the CR-DEMO results as hotspots on a map according to the frequency that a station appears in the Pareto front. Leach et al., (2015) added objective functions for streamflow signatures and indicators of hydrologic alteration to explicitly consider the shape of the hydrograph at different gauging sites when building a hydrometric network. This modification increases the variety of flow regimes seen in optimal networks. Finally, Keum and

Coulibaly (2016) studied the sensitivity of CR-DEMO to time series length for hydrometric networks.

2.0 Hydrometric Network Design Using Dual Entropy Multi-Objective Optimization in the Ottawa River Basin

Abstract

Water resources managers commonly rely on information collected by hydrometric networks without clear knowledge of their efficiency. Optimal water monitoring networks are still scarce especially in the Canadian context. Herein, a Dual Entropy Multi-Objective Optimization (DEMO) method uses information theory to identify locations where the addition of a hydrometric station would optimally complement the information content of an existing network. This research explores the utility of transinformation (TI) analysis, which can quantitatively measure the contribution of unique information from a hydrometric station, as a preliminary step in DEMO analysis. When used in conjunction these methods provide an objective measure of network efficiency and allow the user to make recommendations to improve existing hydrometric networks. A technique for identifying and dealing with regulated basins and their related bias on streamflow regionalization is examined to improve DEMO application. The Ottawa River Basin, a large Canadian watershed with a number of regulated hydroelectric dams, was selected for the experiment. The TI analysis approach is less costly computationally than the DEMO and it provides preliminary information which is supported by DEMO results. Regionalization was shown to be more accurate when the stations in regulated basins were omitted using leave one out cross validation. DEMO

analysis was performed with these improvements and successfully identified optimal locations for new hydrometric stations in the Ottawa River Basin.

Keywords: entropy, hydrometric network, multi-objective optimization, network design, water resources

Introduction

Hydrometric networks provide important data for water researchers and water resources decision makers. Recent research (Burn, 1998; Mishra and Coulibaly, 2009; Coulibaly et al., 2013) has shown that most Canadian watersheds do not have adequate network density as defined by the World Meteorology Organization guidelines (World Meteorological Organization, 2008). In addition, water resources engineers are frequently faced with decisions regarding priority locations to construct or maintain hydrometric stations with no standard method of quantitatively measuring the station's importance.

Information theory has been adopted as an important tool for objective hydrometric network design, for example in Husain (1989), Alfonso et al. (2013) and Mishra and Coulibaly (2014). Husain (1979) originally applied information theory to the field of hydrometric network design. In a sample British Columbia streamflow network, he used joint entropy and a stepwise optimization to maximize the information content. Another well-known example of information theory based network design is Yang and Burn (1994); in a network in Manitoba they used directional information transfer to determine which hydrometric stations produced redundant data and could be removed. Alfonso et al. (2010) introduced the use of total correlation as a design metric and included it in a multi-objective optimization. This allowed limitation of the redundant information that the network was collecting. Li et al. (2012) considered joint entropy, transinformation and total correlation as design objectives and used a weighted single objective optimization method to design a hydrometric network in Texas. This was known as the maximum information minimum redundancy method.

Dual Entropy Multi-objective Optimization (DEMO) is a new robust method of identifying locations where the addition of a hydrometric station will optimally complement the information content of the network (Samuel et al., 2013). It uses a powerful epsilon-dominance hierarchical Bayesian optimization algorithm to efficiently find the Pareto front of non-dominated network configurations. This method has been expanded to include two additional objective functions; indicators of hydrologic alteration and streamflow signatures in order to explicitly consider the spatial variability of watershed characteristics (Leach et al., 2015). A key step to this method of network design is that the entropy analysis relies on hydrometric time series at all existing and potential station locations in the watershed. In order to generate these time series, the data from existing stations must be regionalized to create synthetic data at potential station locations. The regionalization method can be decided upon by the user, however, the inverse distance weighting drainage area ratio method (IDW-DAR) has proven to be proficient at producing these time series in Ontario (Samuel et al., 2011). There has been limited research comparing the utility of each of these regionalization methods, particularly in a heavily regulated study area such as the Ottawa River Basin. Furthermore, recent studies have shown that the IDW-DAR method works well for streamflow regionalization in different Ontario watersheds (Samuel et al., 2013; Leach et al., 2015).

A transinformation analysis was performed first to provide a quantitative way of measuring the information content that each existing station adds to the dataset, as was done by Mishra and Coulibaly (2014). The McMaster University-Hydrologiska Byråns

Vattenbalansavdelning (MAC-HBV) rainfall runoff model, based on Bergström's (1976) commonly used HBV model, was used to identify hydrometric stations in sub-basins which displayed unnatural flow regimes. These stations were identified because there is a high probability that they were regulated or abnormal in some way and would negatively influence the regionalization process. A leave one out cross validation method (LOOCV) was used in order to quantify the regionalization accuracy, and to show how removing the influence of hydrometric stations with unnatural flow regimes improves the accuracy of the synthetic runoff. DEMO was then applied using the existing dataset and regionalized data at potential station locations in order to determine where additional stations should be located in order to maximize the information content of the network.

This research builds on the DEMO method used in Samuel et al. (2013) and Leach et al. (2015) by complementing the entropy analysis with transinformation analysis and introducing a method for handling regulated sub-basins. Furthermore the study results are of particular interest to water resources managers dealing with regulated basins. The results will inform decision makers in enhancing the hydrometric network of the Ottawa River Basin.

Study Area and Data

Study Area

The Ottawa River Basin is located on the Ontario-Quebec border northwest of Ottawa. It has a drainage area of 146,300 square kilometres (km^2) and an average discharge of about 1950 cubic metres per second (m^3/s). The basin contains a number of tributaries including

the Outaouais River, the Montreal River, the Kipawa River, the Madawaska River, the Gatineau River and the Lievre River (Ottawa River Regulation Planning Board, 2011). About 75% of the basin is covered in forest, more than 40% being dense mixed wood. About 6% of the basin in farmland and less than 2% of the land area in the basin has been developed. The average daily temperature in the northern part of the basin varies from about 18 °C in the summer to -15 °C in the winter. The average annual precipitation is about 840 mm, with more precipitation falling during the warmer months. Similarly, in the southern part, the average daily temperature varies between 21 °C and -10 °C, with 920 mm of average precipitation. A digital elevation model and waterbodies of the watershed are shown in Figure 1.

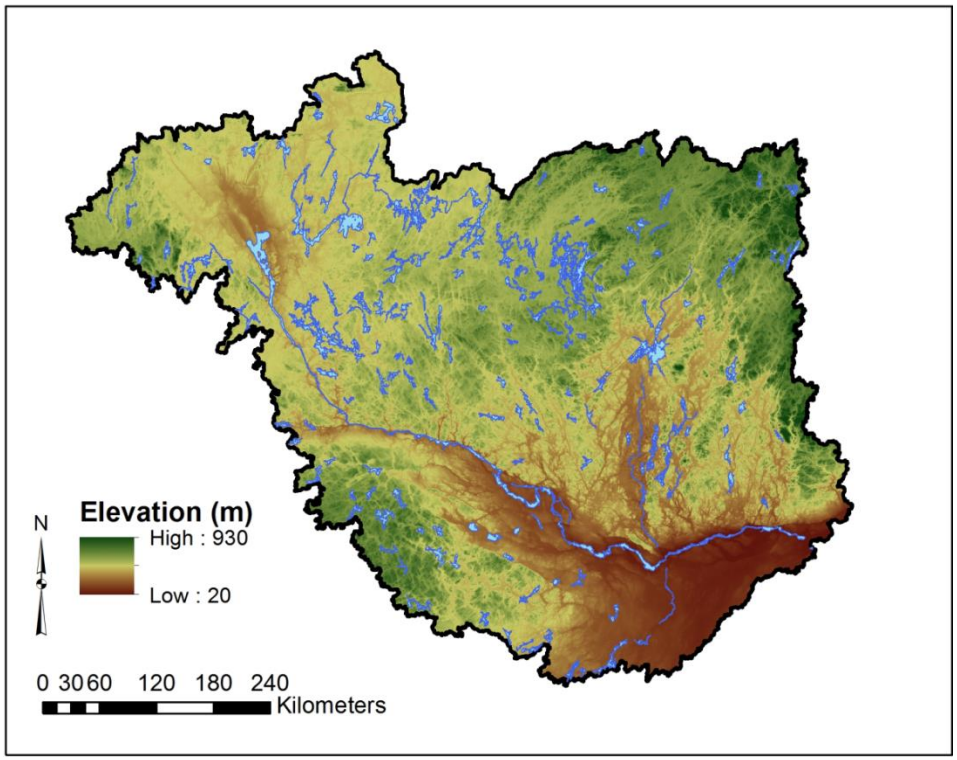

Figure 1. Ottawa River Basin Digital Elevation Model and Water Bodies

Data Preprocessing

The 50,000:1 digital elevation model and a waterbodies shape file were downloaded from Environment Canada's (EC's) Geogratis database. These were combined in ArcHydro to create the maps and to delineate sub-basins within the watershed. A total of 147 distinct sub-basins were delineated in the watershed. There are 87 flow and water level stations operated by the EC Water Survey in the Ottawa River Basin. Data for 4 additional stations in the basin was provided by Ontario Power Generation (1) and Hydro Quebec (3). The flow data for these stations was acquired using EC's HYDAT database. Only stations with at least 10 years of continuous flow data since 1985 were used. This resulted in 37 qualifying flow stations which were considered to be the existing hydrometric network. Of these 37 stations, 16 had flows which were unnatural, and this was considered in the analysis. For the purposes of this analysis, potential flow station locations were considered at the drainage point of each of the 147 sub-basins.

There are 316 EC Weather Office weather stations which have been operational in the river basin. Only stations with at least 10 years of continuous temperature and precipitation data since 1985 were selected. Qualifying stations which were within 1 km of each other were considered redundant and the one with less current data was removed. This left 82 weather stations which were used as the temperature and precipitation data network. All of the data stations mentioned above, as well as the potential station locations which were evaluated are shown in Figure 2.

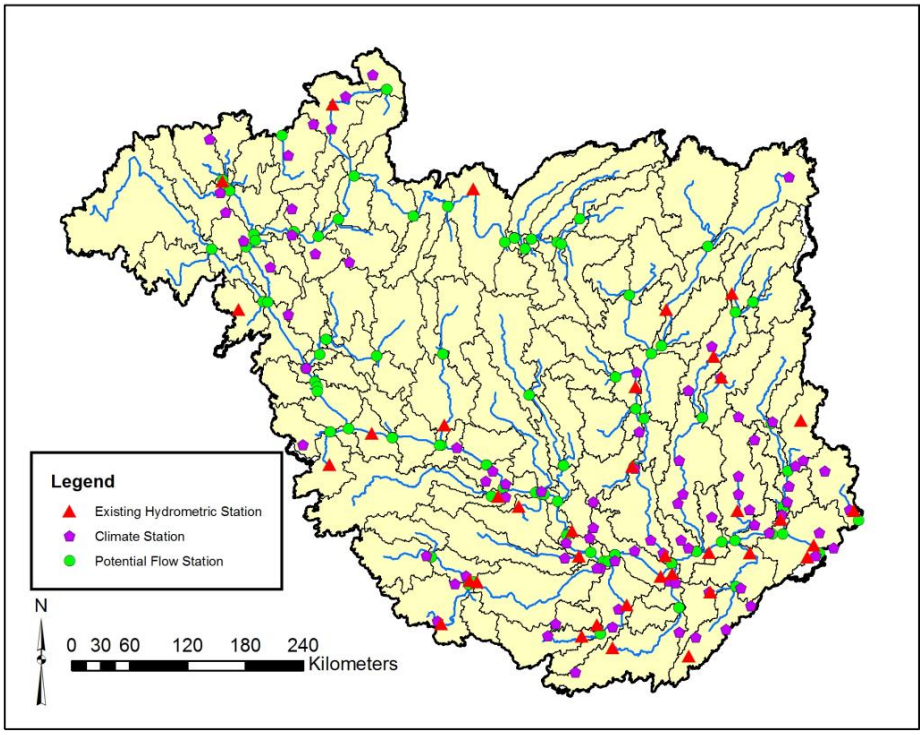

Figure 2. Ottawa River Basin Environmental Data Stations

Methodology

Methodology Overview

A flowchart of the methodology is shown in Figure 3. Each method is explained in further detail in this section.

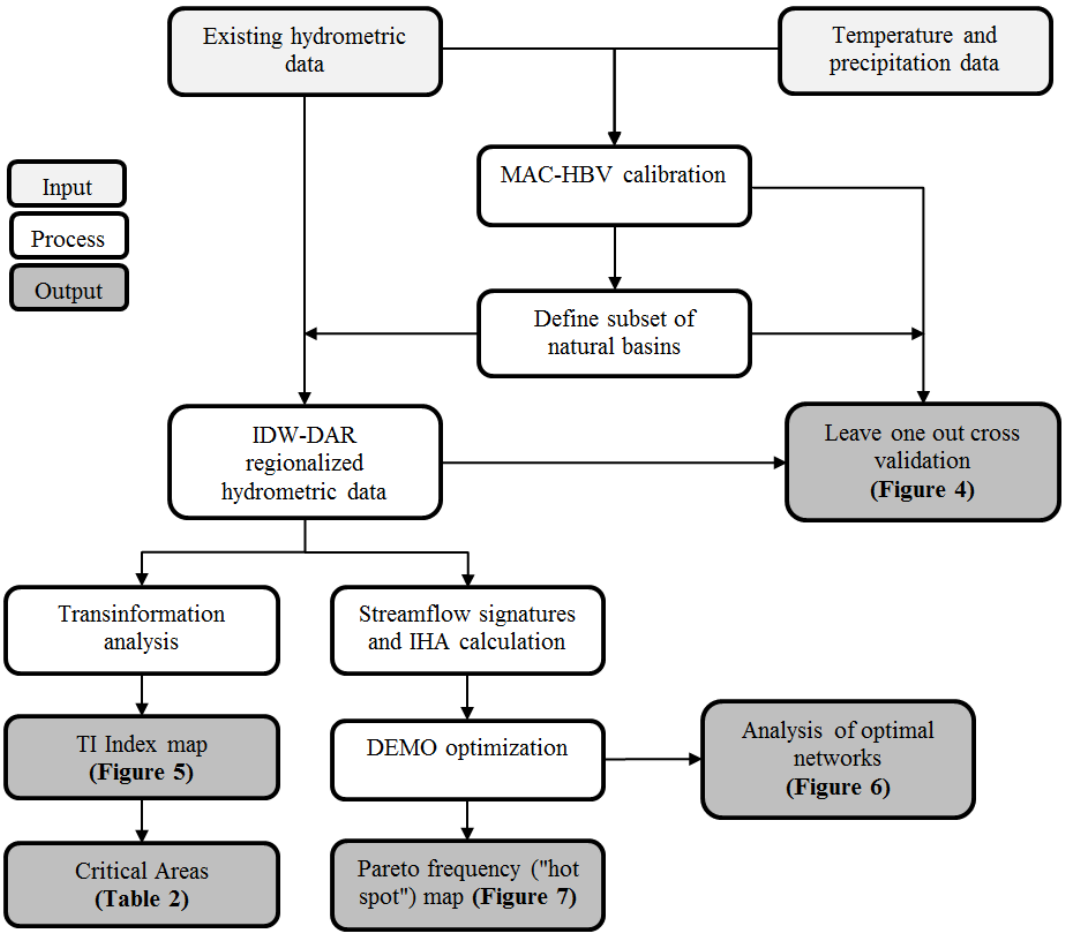

Figure 3. Methodology overview flowchart

McMaster University-Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) Optimization

MAC-HBV is a lumped conceptual rainfall-runoff model developed at McMaster University for the purpose of estimating streamflow in ungauged basins (Samuel et al., 2011). It is based on the widely used HBV model developed by (Bergström, 1976). It takes in 15 parameters, daily average temperature and daily precipitation as inputs and outputs daily average flow.

Here, the existing hydrometric stations in the basin with less than 5% missing data were each calibrated to the MAC-HBV model. Temperature and precipitation time series were generated at each location using IDW. The available data between 1995 and 2010 at each station was split evenly into two periods. A swapped optimization was performed using a particle swarm optimization algorithm; the parameters were calibrated to the first period and validated with the second, then the periods were swapped and the optimization was performed again. Similar swapped optimization procedures were used by Samuel et al. (2011) and Merz and Blöschl (2004). The parameter set which produced the highest validation Nash-Sutcliffe Efficiency (NSE) was used for each station. Parameter sets from stations which did not achieve a validation NSE of at least 0.65 were excluded from the dataset. This value was chosen in order to retain a sufficient number of stations in the dataset for analysis, while ensuring that the stations considered had high quality data. A number of the stations removed were measuring regulated flows. The purpose of this optimization was to identify any sub-basins with abnormal flow regimes which should be ignored from subsequent regionalization. A subset of 21 stations were identified which had sufficiently low missing data as well as natural flow regimes. Using only these stations meant a sparser network for the regionalization; however through the analysis it was found that it was beneficial to not include unnatural flow regimes which may bias the synthetic data.

MAC-HBV was also used to generate a second synthetic runoff dataset for comparison to the IDW- DAR regionalization method as an accuracy assessment. Samuel et al. (2011) concluded that the best method of creating runoff data using MAC-HBV is to regionalize

the parameter sets using IDW. The regionalized parameter sets were used with regionalized temperature and precipitation data to create runoff time series at the potential station locations.

Inverse Distance Weighting Drainage Area Ratio (IDW-DAR) Regionalization Method

The flow data was converted to runoff by dividing the flow by the drainage area of the sub-basin at the monitoring station. Drainage areas provided by EC were used for this calculation. Drainage areas which were unknown were estimated using ArcHydro and the digital elevation model. The IDW-DAR method can be described as follows.

$$Q_u = \sum_{i=1}^{n} w_i \left(\frac{A_u}{A_i} \right)^{\alpha} Q_i \tag{1}$$

$$w_i = \frac{\left(h_i^{-2} \right)}{\sum_{i=1}^{n} \left(h_i^{-2} \right)} \tag{2}$$

Q_u is the runoff, Q_i is the observed flow rate at each station, α is a weighting parameter set to 1 which was found to be optimal in Samuel et al. (2011), A_u is the drainage area of the output sub-basin, A_i is the drainage area of the gauged sub-basin and h_i is the distance between the centroid of the sub-basin containing the calculated flow and the gauged sub-basin. The 10 nearest neighboring stations were considered when using this method. This process was repeated to generate regionalized time series at each of the potential station locations.

The years 2001-2010 were selected for the analysis. This time period was chosen because it is fairly recent and therefore more relevant to future data than prior flow values. A 10

year time period was determined to be the recommended length for DEMO analysis of daily streamflow series (Keum and Coulibaly, 2015). The missing data were also filled using the IDW-DAR method.

Streamflow Signatures and Indicators of Hydrologic Alteration

After the flow dataset was generated for all potential station locations, streamflow signatures and IHA were calculated. These parameters contain information about different parts of the hydrograph in the sub-basin, hence it is beneficial to explicitly minimize their correlation when using them in DEMO (Leach et al., 2015). Sawicz et al. (2011) defined six streamflow signatures which could be used to detect catchment responses. The runoff ratio (RR) is the ratio of precipitation which becomes overland flow. The slope of the flow duration curve (FDC) is a descriptive measure calculated using the 33^{rd} and 66^{th} percentile of flow recorded at the location. The baseflow index (BI) is the sum of the daily percentage of flow which is considered baseflow. The streamflow elasticity describes the sensitivity of the streamflow to changes in precipitation. The snow day ratio is the ratio of days with precipitation below 2 degrees Celsius. Finally, the rising limb density is the percentage of days in which the average flow rate is higher than it was the day before. Leach et al. (2015) identified three signatures which displayed the least correlation among variables and used these three as additional DEMO inputs for each watershed. The streamflow signatures were calculated for each hydrometric station and the three signatures with the least correlation were selected. The results of this analysis are shown in Table 1. The three signatures with the

lowest absolute correlation were the SE, the SDR and the RLD, and thus these were used in the analysis.

Table 1. Streamflow signatures correlation

	RR	FDC	BI	SE	SDR	RLD	|Σ|
RR	1.00						1.70
FDC	0.31	1.00					1.16
BI	0.92	0.33	1.00				1.92
SE	0.31	-0.22	0.48	1.00			1.05
SDR	0.13	0.29	0.14	0.02	1.00		0.84
RLD	0.03	0.00	0.05	0.01	0.27	1.00	0.36

IHA were developed in order to quantify human impacts on a watershed. Monk et al. (2011) identified five key IHA parameters which represented a large portion of the variation in the sub-basins studied. The five parameters identified by Monk were used: one day maximum flow, one day minimum flow, Julian day of maximum, Julian day of minimum and number of reversals. Each of these parameters were calculated for each hydrometric station in this study. All signatures and IHA parameters were normalized to be between 0 and 1 using Equation 3 where x_i is the value of the signature being normalized at station i. This yielded a vector of streamflow signatures and a vector of IHA at each existing and potential station location. Additional objective functions were added to DEMO in order to maximize the Euclidean distance between these vectors in

order to improve the spatial variability of selected stations as shown in Leach et al. (2015).

$$x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

Information Theory

Shannon (1948) introduced a method for quantifying the amount of information contained in a given dataset. This measurement was called the data entropy. The entropy of a time series is calculated as

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) \tag{4}$$

where H is the entropy and $P(x_i)$ is the probability of event x_i. The joint entropy for N time series as would be collected from a hydrometric network is shown to be

$$H(X_1,...,X_N) = -\sum_{i_1=1}^{n1}\sum_{i_2=1}^{n2}...\sum_{i_N=1}^{nN} P(x_{1,i_1}, x_{2,i_2},..., x_{N,i_N}) \log_2 P(x_{1,i_1}, x_{2,i_2},..., x_{N,i_N}) \tag{5}$$

where x_1 through x_N represent the station locations and $x_{N,k}$ represents the k^{th} event at station N. $P(x_{1,i1}, x_{2,i2} \ldots x_{N,iN})$ is the joint probability of events i_1 through i_N at stations 1 through N. The joint entropy gives a quantitative measurement of the information content in a set of time series. Total correlation can also be calculated for a network. This is defined as

$$C(X_1,...,X_N) = \left[\sum_{i=1}^{N} H(X_i)\right] - H(X_1,...,X_N) \tag{6}$$

The total correlation value shows the amount of redundant information in the dataset. The total correlation of a time series is the multivariate extension of the bivariate transinformation (TI) which is defined as follows.

$$TI(X,Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y \mid X) \quad (7)$$

The transinformation is used in this context to compare a hydrometric time series with a synthetic version of itself generated from multiple linear regression of the rest of the dataset. The transinformation can then be regionalized using kriging to show where in the watershed there is a surplus or deficit of information.

Dual Entropy Multi-Objective Optimization (DEMO)

The regionalization method is first used to generate flow at ungauged sites which are the potential locations of additional stations. Then multi-objective optimization is used to maximize the joint entropy and minimize the total correlation in the network. This can be summarized as follows:

First, a number of potential hydrometric station locations are identified within the watershed. In this case the potential locations were at the outflow of each of the 147 sub-basins identified using ArcGIS. Using time series data from existing stations, synthetic data is generated for each of these potential locations. Once the data is regionalized to a number of potential station locations, DEMO uses an epsilon-dominance hierarchical Bayesian (ε-hBOA) optimization algorithm (Kollat et al. 2008) to determine which of these time series adds the most unique information to the dataset. Reed et al. (2013) compared various water resources optimization algorithms and found that this type of optimization algorithm was one of the most robust in water monitoring network design.

The main advantage of this algorithm is that it uses Bayesian network models to preserve the interdependencies between variables during evolution. Further information about the algorithm can be found in Kollat et al. (2008) and Leach et al. (2015).

Second, in addition to entropy and total correlation, the streamflow signatures and the indicators of hydrologic alteration (IHA) at each monitoring station were calculated. Thus four objective functions were used in optimization: the joint entropy was maximized, the total correlation was minimized and the Euclidean distance between both the IHA and streamflow signatures were maximized. This was done in order to ensure the flow regimes at the stations in the network were as diverse as possible as shown in Leach et al. (2015). The optimization output was a set of non-dominated network configurations. Additional station locations can be ranked on importance based on the frequency that they appear in the generated Pareto front.

Third, in this study, the number of additional stations was varied between 1 and 10 in order to determine the best additional station locations without limiting the search space. Given the total number of existing stations (87) and the size of the watershed, it was assumed that optimal network can be obtained by adding less than 10 new stations if optimal locations can be identified.

Transinformation (TI) Index

Transinformation is equivalent to bivariate total correlation. It is a quantitative measurement of the amount of information shared between two variables. A high transinformation value would indicate a strong dependence between two time series. Mishra and Coulibaly (2014) used transinformation to compare hydrologic time series

with synthetic time series generated using multiple linear regression of the rest of the flow data in the basin. In this study, these transinformation values were regionalized using IDW interpolation to display the areas with a redundant or deficit of information in the watershed. The transinformation values were normalized to between 0 and 1 to display the transinformation index defined by Mishra and Coulibaly (2014) using Equation 3. The period of 2001-2010 was used in this analysis.

Results

Regionalization Results

The regionalization step in DEMO is important, as the network optimization relies on the probability distribution of the regionalized data. A leave-one-out cross validation (LOOCV) was performed on the IDW-DAR and MAC-HBV time series in order to determine the accuracy of each regionalization technique and thus identify which technique produces a more reliable dataset. Time series were generated at each station location as if it were an ungauged basin. These time series were then compared to the actual runoff data at each station using NSE. Many of the stations which the MAC-HBV model could not be well calibrated to also displayed poor NSE values when modelled with IDW-DAR. It is suspected that the flow at many of these stations is regulated. The cross validation was performed twice, once with the 21 stations with sufficient data which could be modelled well with MAC-HBV only and once with all stations. The results of this analysis are displayed in Figure 4. The top whisker, top bar, middle point, bottom bar and bottom whisker show the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles

respectively. It can be seen that the IDW-DAR technique yielded the most accurate regionalization results, and that removing stations with suspect data (i.e. regulated flow) greatly improved results for both techniques. This shows why the subset of natural stations identified by this analysis was used in subsequent regionalization, while the stations with suspect data were removed.

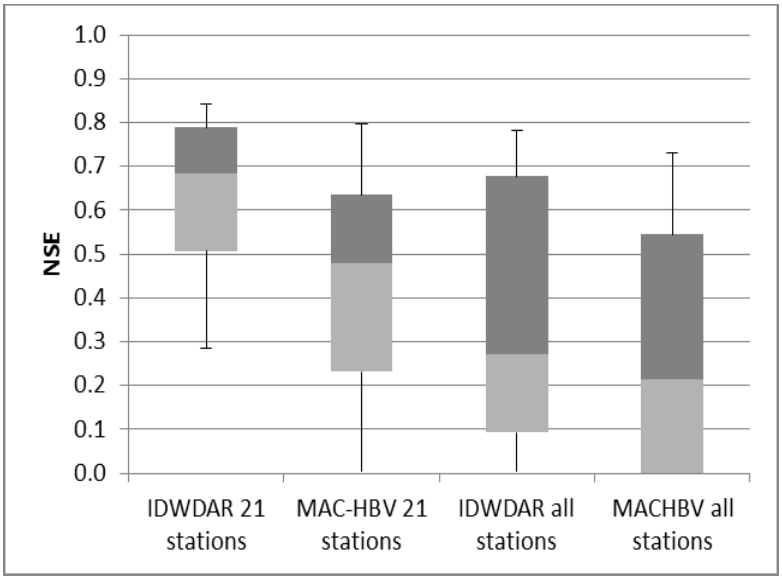

Figure 4. Leave one out cross-validation results

Transinformation Index Values

The transinformation index (TI) value was computed for each existing station. The TI value of a station is defined as the mutual information it shares with other stations within the basin. This value provides a quantitative measure of the importance of each station. Lower TI values indicate that the stations share very little common information and are hence more independent. Large TI values mean that the stations are more dependent or redundant in their information, thus there is no need to add station in these areas. Conversely, smaller TI values indicate high priority areas that should be considered for

additional stations. Here, the TI index values are interpolated across the entire watershed and classified in 4 categories. The TI index map is shown in Figure 5. The areas which registered a low TI were the Upper Ottawa River near the Dozois Reservoir and in the northeast of the watershed in the headwaters of the Gatineau River, the Liévre River and the Rouge River. These areas do not have a high station density and thus are classified as high priority areas. There are some areas in the southern region which showed low TI values, however some of these are suspected to be the results of controlled flow. The northwestern and central areas of the river basin registered high TI indexes despite low station density. This suggests that the existing stations in these areas were not set up at appropriate locations and are duplicating some information. Overall the TI index map gives a picture of critical areas where additional stations are needed in the watershed. The critical zones include the “deficit” and “highly deficit” areas which are ungauged or poorly gauged. The “average” and “above average” areas should not be a priority for adding new stations. This does not suggest that the number of stations is optimal in these areas, but rather that the existing stations are sharing a larger amount of information. This indicates that a better course of action than introducing new stations could be relocating existing ones. The Ottawa River Basin as a whole has an average TI value of 1.367 and a standard deviation of 0.351. In future research these values can be compared between Canadian watersheds to determine where to allocate monitoring network resources.

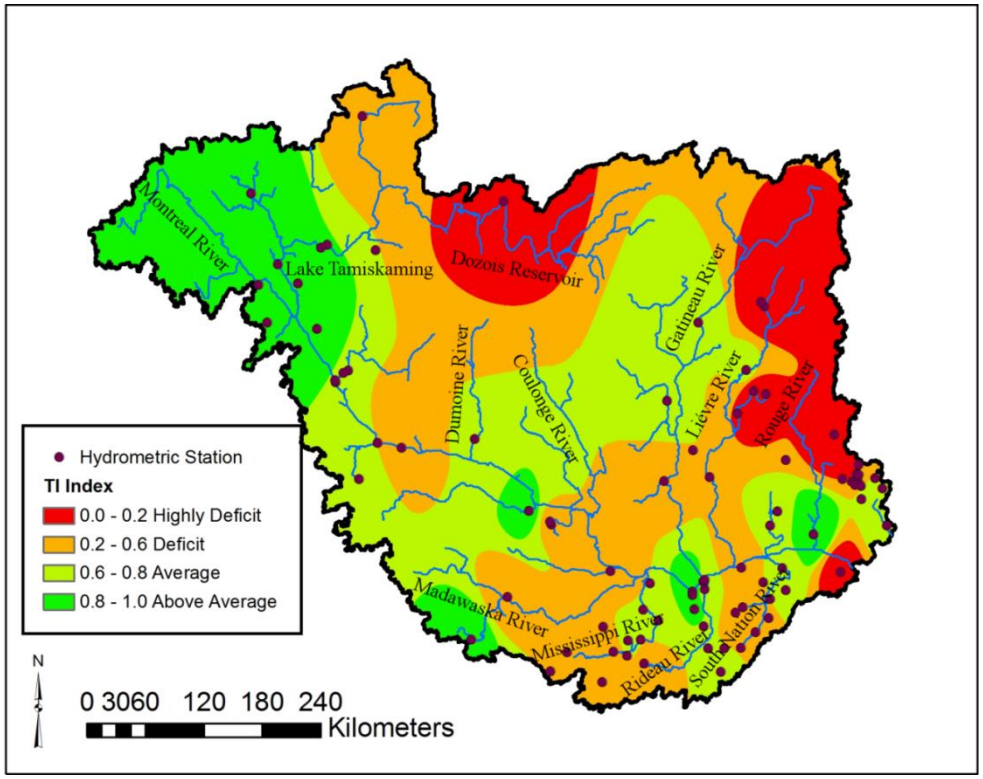

Figure 5. Map of transinformation index values of existing stations

In addition to the TI index map, the proportion of area per TI index category ("deficit", "average", etc.) is estimated for each main sub-basin. The fraction of areas classified under each TI index category is shown in Table 2. For example, it can be seen that for the Gatineau River basin (QC) and the Rideau River basin (ON), about 50% of the basin area is in deficit category; while for the Madawaska River basin (ON) and the Mississippi River basin (ON), about 46% and 63% of the basin area is in deficit category respectively. In some sub-basins (e.g. Dozois, Lievre, Rouge; see Table 2) more than 80% of basin area is in deficit category. In general, for 8 out of the 12 sub-basins, about 50% of basin area is in deficit category.

To determine the importance of each station in the network, the TI index values for specific hydrometric stations were ranked and are shown in Appendix A. The stations having lower TI values indicate they share very little common information and are important stations and are ranked higher. Stations having higher TI values are dependent and are duplicating the same information, and are thus considered of low importance and ranked lower. Note that station that did not have data dating back to the year 2001, was excluded from the analysis.

Table 2. Critical areas by major sub-basin

Prov	Basin	Area (km^2)	Ratio of Areas Under Different Categories Using TI Index			
			0.0 - 0.2 (highly deficit)	0.2 - 0.6 (deficit)	0.6 - 0.8 (average)	0.8 - 1.0 (above average)
	Ottawa River Basin	146,300	0.15	0.36	0.32	0.17
QC	Lake Tamiskaming	16,416	0.00	0.36	0.09	0.54
	Dozois Reservoir	17,641	0.43	0.40	0.16	0.00
	Gatineau River	22,567	0.18	0.32	0.49	0.01
	Liévre River	9,426	0.53	0.40	0.08	0.00
	Dumoine River	4,413	0.05	0.56	0.39	0.00
	Coulonge River	5,214	0.00	0.45	0.55	0.00
	Rouge River	5,698	0.71	0.10	0.06	0.14
ON	Montreal River	6,962	0.00	0.00	0.00	1.00
	Madawaska River	8,572	0.00	0.46	0.34	0.21
	Mississippi River	3,900	0.00	0.63	0.36	0.01
	South Nation River	3,732	0.00	0.54	0.46	0.00
	Rideau River	3,712	0.00	0.51	0.43	0.06

Results and Discussion

The results of the DEMO analysis using the IDW-DAR method are shown in Figures 6 and 7. Figure 6a shows the Pareto front generated from the multi-objective optimization. Each point in this Pareto front represents a non-dominated solution corresponding to a network configuration found by DEMO. Given that the number of additional stations was allowed to vary, the number of stations per solution is different. For example, three of these networks occurring at different points on the Pareto front are displayed in Figure 6b-d. One of the main advantages of the DEMO approach is that multiple optimal solutions are obtained, giving the decision maker more flexibility in which solution to implement.

Interestingly, the streamflow signatures and IHA Euclidean distance tend to vary inversely to the number of additional stations. Although adding a station to a certain network will always increase the Euclidean distance between these values, each network in the Pareto front is a distinct optimal solution and the distances decrease as entropy becomes the dominant objective.

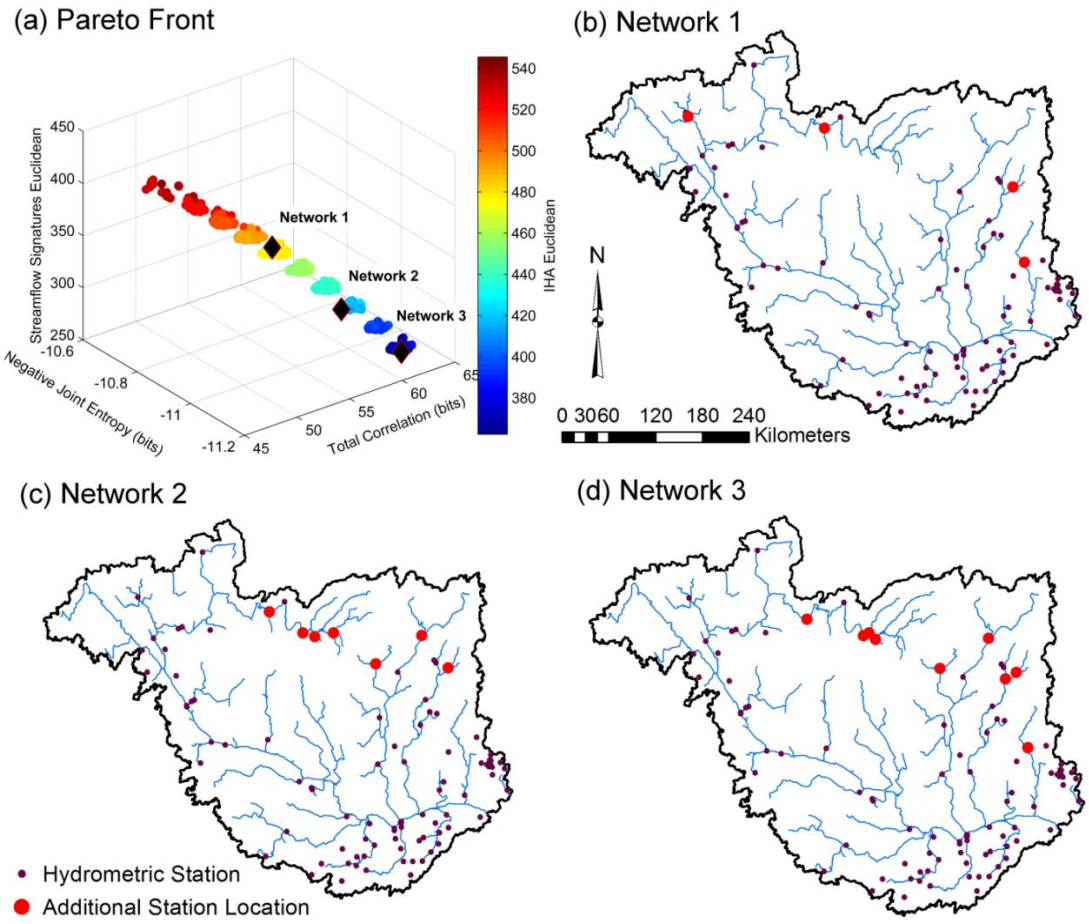

Figure 6. Examples of Pareto optimal networks

In order to analyze a large set of non-dominated Pareto front networks, the frequency at which each potential additional station appears in a Pareto optimal network was calculated. The results were spatially interpolated using IDW across the area of the basin to create critical areas or "hot-spots". These hot-spots are shown in Figure 7. The "hot-spots" shown in red indicate the areas with high probability of receiving new stations based on all the Pareto front solutions. This means that for almost all the optimum networks (or solutions) of the Pareto front, these areas ("hot-spots") were selected as locations for new stations. This information is important to guide the selection of an

optimal network by end users. A selected optimum network should have new stations at these locations. Some solutions, for example the one in Figure 7d, show several stations clustered in one location. This indicates that although this location has a high joint entropy, it also probably has a high total correlation. Thus it is Pareto optimal, but not necessarily the solution which should be chosen by the user. It is up to the user to decide on the tradeoffs between each objective function given the set of non-dominated solutions.

In this analysis, four scenarios were considered in order to ensure replicability and to study the effects which IHA and streamflow signatures have when using DEMO. All scenarios produced similar results (Figure 7). The areas around the Dozois Reservoir and the Rouge River are recommended as priority areas for new additional stations. This is consistent with the results produced by the TI Index analysis (Figure 5). It can also be seen that including IHA and signatures enhances the spatial variability of the DEMO results. This is important as this technique not only considers the joint entropy of the network, but also explicitly considers the streamflow characteristics of the station locations in order to maximize their spatial distribution. Specifically the IHA inclusion increases the hot-spot area near the Rouge River, while the signatures increase the area near the Dozois Reservoir.

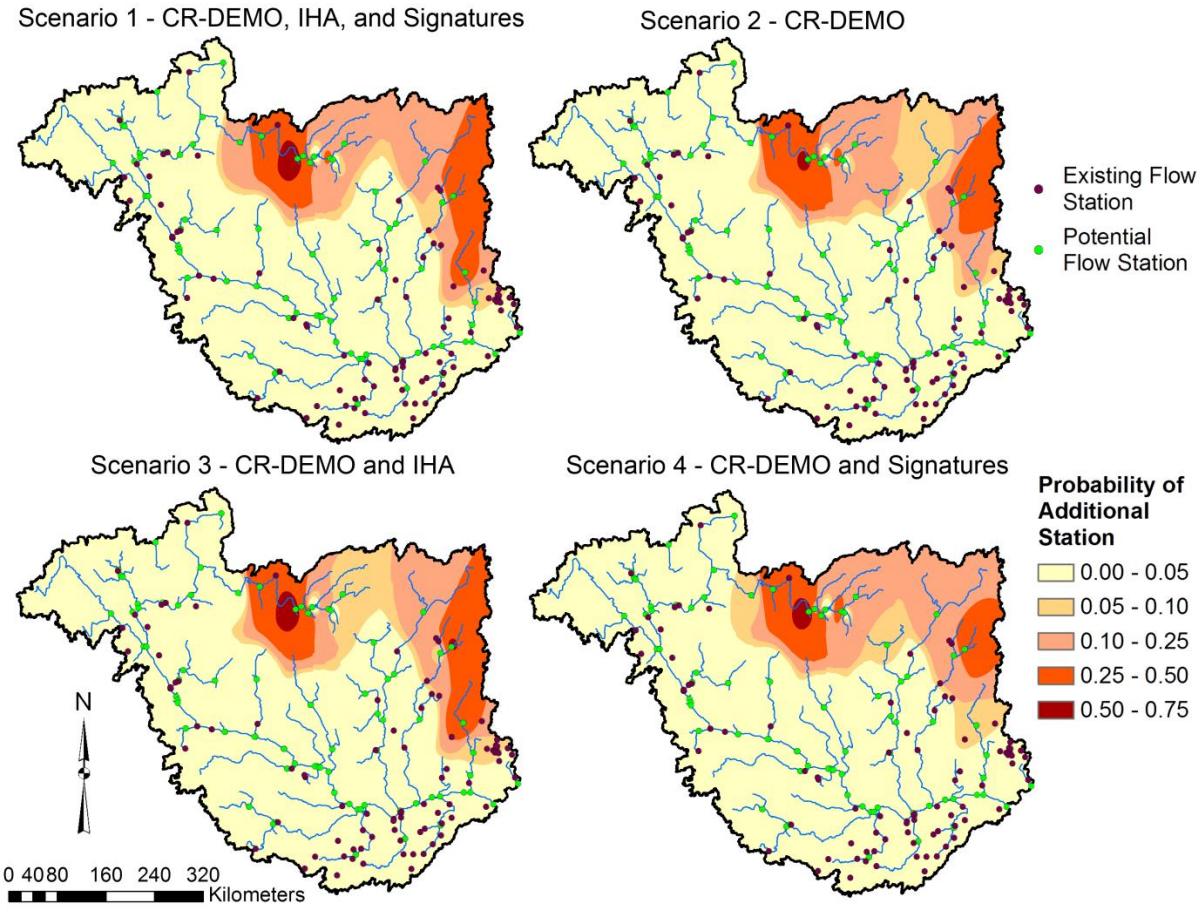

Figure 7. IDW-DAR regionalization DEMO outputs

Conclusions

Generating synthetic time series using IDW-DAR was proven to be more accurate than using MAC-HBV in a leave one out cross-validation. It was also found that the regionalization method was improved by not considering regulated basins.

The transinformation (TI) index was calculated at each station and interpolated across the Ottawa River Basin. The resulting TI index map showed deficit areas such as near the Liévre River, the Rouge River and the Dozois Reservoir where additional stations are

highly needed. Areas with redundant stations were also identified based on the TI index values. It was also shown that for 8 out of the 12 main sub-basins, about 50% of the sub-basin area is in the deficit category. This proportion exceeded 80% in some sub-basins (e.g. Dozois, Lievre, and Rouge River basin).

DEMO was used to optimize the design of the hydrometric network in the entire basin. The combination of four objective functions (joint entropy, total correlation, IHA and streamflow signatures) provided the most widely distributed hot-spots by explicitly considering both information content and runoff characteristics. The spatial distribution of the hot-spots indicated specific areas where additional stations are needed whatever the optimal network selected among the available Pareto front solutions. This information is essential for decision making in optimal network selection from DEMO solutions. Interestingly, the spatial distribution of hot-spots is consistent with the TI index map. Finally, examples of Pareto optimal networks were shown with the optimal locations for future hydrometric stations.

The main advantage of this information theory based optimization method is that it provides a quantitative measure of the overall network efficiency and the value of each individual station with the TI value. It also provides objectively optimal recommended locations in order to maximize the information content captured by the network with the DEMO results. The key limitation regarding this method of network design lies in the regionalization step. The dataset being used for optimization is limited by the accuracy of current regionalization techniques. Future research with this method could consider

further improving regionalization with new techniques or investigating the application of DEMO such as the effects of scaling.

In the future decision makers will be able to use DEMO and TI analysis to learn about and augment their monitoring networks. TI analysis is a useful tool for evaluating existing stations by ranking the information density coming from each station and in the network as a whole as shown in Table 2. This can be shown spatially to get an idea of information sparse regions as in Figure 5. From this point, the user will run DEMO analysis and find a set of non-dominated network solutions. When the frequency of Pareto-optimal selection is mapped, it should have some similarities to the TI map as shown in Figure 7. These steps will provide the user with many output configurations, samples of which are shown in Figure 6. Each solution can be optimal depending on how each objective function is weighted, so it is up to the user to choose a solution which is appropriate for their needs. Each of these tools provides useful and complementary information for evaluating and augmenting existing hydrometric networks.

Acknowledgements

This research was supported jointly by Ontario Power Generation (OPG) and Natural Science and Engineering Research Council (NSERC) of Canada. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada, and by datasets from Environment Canada, Hydro-Quebec, and OPG. The authors are grateful to Dr. Joshua Kollat (Penn State University) who developed the ε-hBOA, and provided the source codes.

References

Alfonso, L., Lobbrecht, A., & Price, R. (2010). Information theory-based approach for location of monitoring water level gauges in polders. *Water Resources Research*, *46*(3), n/a–n/a. http://doi.org/10.1029/2009WR008101

Alfonso L., He L., Lobbrecht A. & Price R. 2013 Information theory applied to evaluate the discharge monitoring network of the Magdalena River. *Journal of Hydroinformatics* 15(1), 211-228.

Bergström S. 1976 Development and application of a conceptual runoff model for Scandinavian catchments. *Series A, No. 52*, Lund Institute of Technology/Univ. of Lund, Sweden.

Burn D.H. 1997 Hydrological information for sustainable development. *Hydrological Sciences Journal* 42(4), 481-492.

Coulibaly P., Samuel J., Pietroniro A. & Harvey D. 2013 Evaluation of Canadian national hydrometric network density based on WMO 2008 standards. *Canadian Water Resources Journal* 38(2), 159-167.

Husain, T. (1979). *Shannon's Information Theory in Hydrologic Network Design and Estimation*. University of British Columbia.

Husain T. 1989 Hydrologic uncertainty measure and network design. *Journal of the American Water Resources Association* 25(3), 527-534.

Keum J., & Coulibaly P. 2015 Sensitivity of entropy method to time series length in hydrometric network design. *Journal of Hydrologic Engineering*, Submitted.

Kollat J., Reed P., & Kasprzyk J. 2008 A new epsilon-dominance hierarchical Bayesian optimization algorithm for large multi-objective monitoring network design problems. *Advances in Water Resources* 31(5), 828-845.

Leach J. M., Kornelsen K. C., Samuel J., & Coulibaly P. 2015 Hydrometric network design using streamflow signatures and indicators of hydrologic alteration. *Journal of Hydrology* 529(3), 1350-1359.

Li C., Singh V., & Mishra A. 2012 Entropy theory-based criterion for hydrometric network evaluation and design: Maximum information minimum redundancy. *Water Resources Research* 48(5), W05521.

Merz R., & Blöschl G. 2004 Regionalisation of catchment model parameters. *Journal of Hydrology (Amsterdam)* 287(1-4), 95-123.

Mishra A. K., & Coulibaly P. 2009 Developments in hydrometric network design: A review. *Reviews of Geophysics* 47(2), RG2001.

Mishra A. K., & Coulibaly P. 2014 Variability in Canadian seasonal streamflow information and its implication for hydrometric network design. *ASCE Journal of Hydrologic Engineering* 19(8), DOI:10.1061/(ASCE)HE.1943-5584.0000971.

Monk W. A., Peters D. L., Curry R., & Baird D. J. 2011 Quantifying trends in indicator hydroecological variables for regime-based groups of Canadian rivers. *Hydrological Processes* 25(19), 3086-3100.

Ottawa River Regulation Planning Board. 2011 Characteristics of the Basin. Retrieved November 11, 2015, from Ottawa River Regulation Planning Board: ottawariver.ca

Reed P. M., Hadka D., Herman J. D., Kasprzyk J. R., & Kollat J. B. 2013 Evolutionary multiobjective optimization in water resources: The past, present and future. *Advances in Water Resources* 51, 438-456.

Samuel J., Coulibaly P., & Kollat J. 2013 CRDEMO: Combined regionalization and dual entropy-multi-objective optimization for hydrometric network design. *Water Resources Research* 49(12), 8070-8089.

Samuel J., Coulibaly P., & Metcalfe R. 2011 Estimation of continuous streamflow in Ontario ungauged basins: Comparison of regionalization methods. *Journal of Hydrology* 16(5), 447-459.

Sawicz K., Wagener T., Sivapalan M., Troch P. A., & Carrillo G. 2011 Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences* 15(9), 2895-2911.

Shannon C. 1948 A mathematical theory of communication. *Bell Systems Technical Journal* 27, 379-423.

World Meteorological Organization. 2008 Guide to hydrological practices, volume I: Practices hydrology - From measurement to hydrological information. *World Meteorological Organization* 168.

Yang, Y., & Burn, D. H. (1994). An Entropy Approach to Data Collection Network Design. *Journal of Hydrology*, *157*(1-4), 307–324. http://doi.org/10.1016/0022-1694(94)90111-2

Appendix A

Table 3. TI ranking of existing stations

HYDAT Number	TI	H(X)	C(X)	Rank
02LE025	0.000	1.697	0.452	1
02LE013	0.020	2.010	0.424	2
02LC043	0.046	2.523	0.749	3
02LB032	0.056	1.927	0.521	4
02JB009	0.061	1.672	0.517	5
02KF019	0.262	1.474	0.635	6
02LB033	0.265	1.853	0.502	7
02KC009	0.380	1.058	0.579	8
02LB006	0.417	1.605	0.416	9
02LA024	0.450	1.412	0.747	10
02KD004	0.518	1.572	0.754	11
02LH032	0.522	1.555	0.769	12
02JB013	0.564	1.724	0.797	13
02KC018	0.574	1.385	0.769	14
02KA015	0.576	1.634	0.899	15
02KJ004	0.601	1.586	0.854	16
02LB007	0.618	1.716	0.397	17
02KF006	0.632	1.517	0.867	18
02LD005	0.656	2.058	0.976	19
02JE027	0.662	1.696	0.961	20
02LH033	0.696	1.740	0.986	21
02LE024	0.717	1.964	1.050	22
02LB005	0.735	1.586	0.437	23
02LC008	0.751	2.268	1.132	24
02LA004	0.778	1.513	0.861	25
02LG005	0.784	2.059	0.975	26
02KF010	0.791	1.623	0.521	27
02KB001	0.844	1.491	0.843	28
02KF005	0.865	1.433	1.073	29
02KD002	0.877	1.776	1.092	30
02JC008	0.906	1.535	0.934	31
02LC029	0.980	2.199	1.252	32
02JE028	1.000	1.537	1.086	33

3.0 Assessing Scale Effects on Hydrometric Network Design Using Entropy and Multi-Objective Methods

Abstract

In order to facilitate important water resources decisions it is important that we have access to sufficient, accurate and informative hydrometric data. Combining information theory with multi-objective optimization has led to a method of optimizing the information content provided by a hydrometric network, however, there has been little research on the effects of scale and data limitation on these methods. Herein, a Combined Regionalization Dual Entropy Multi-Objective Optimization (CR-DEMO) and a transinformation (TI) analysis were done to recommend optimal locations for additional hydrometric stations in the Madawaska Watershed in northeastern Ontario. This analysis was designed to be comparative to a similar study conducted on the Ottawa River Basin which encompasses the Madawaska Watershed to allow for an investigation of the scale effects and data limitation in this type of network design. This study concludes that transinformation analysis is not adversely affected by scaling, however, when using CR-DEMO it is very important that users carefully consider the size of the area of interest to avoid biases to the model outputs due to high variability of station information content across the area. Recommendations were made as to the ideal locations of additional stations in the Madawaska Watershed hydrometric network. In addition, this study proposes a technique for including surrounding stations when the area of interest does not have a sufficient number of existing hydrometric stations for analysis. It is shown that

even stations outside of the study watershed can provide useful information because their inclusion in the analysis increases the average transinformation in the watershed.

Introduction

Water resources allocation is an increasingly important and difficult task for water resources managers and requires representative quality data that can be obtained only with an adequate monitoring network. Decision makers rely on monitoring networks to provide them with data that is ample, accurate and informative, therefore, it is necessary that monitoring networks are designed and assessed in a transparent and robust manner. Shannon's information theory (Shannon, 1948) and multi-objective optimization algorithms (Kollat et al., 2008) have been combined to form methods of maximizing the information content of a network while minimizing redundant information (Krstanovic & Singh, 1992; Alfonso et al., 2010; Samuel et al., 2013). Combined Regionalization Dual Entropy Multi-Objective Optimization (CR-DEMO), introduced by Samuel et al. (2013) is one such method which has recently been adopted because of its flexibility and robust optimization algorithm. Recently, another information theory based method called transinformation (TI) analysis has been used in conjunction with CR-DEMO to assess the stations in an existing network and rank them based on their information contributions (Mishra & Coulibaly, 2014; Werstuck & Coulibaly, 2016). This offers another technique for measuring information density in a basin which can be compared to CR-DEMO analysis.

Due to the fact that these methods are relatively new and computationally intensive, there has been little to no research regarding the nuances of their operation, such as changing the scale of the study boundaries or operating in a watershed with too few stations for analysis. A previous TI and CR-DEMO analysis was conducted on streamflow in the

Ottawa River Basin (ORB), a large basin on the border of Ontario and Quebec (Werstuck & Coulibaly, 2016). In order to assess the scaling properties of CR-DEMO, this analysis will be repeated on the Madawaska Watershed, a sub-basin of the ORB. Like the rest of the basin, it is heavily regulated due to the operation of a number of hydroelectric generating stations. In addition, there are only 2 Environment Canada hydrometric monitoring stations operating in this 8,500 square kilometer (km^2) watershed, which does not meet the World Meteorology Organization guidelines on minimum network density (World Meteorological Organization, 2008). In fact, there are not enough stations to carry out a standard TI or CR-DEMO analysis. Due to this limitation, some modifications had to be made to the methods.

A TI analysis was conducted on the watershed by including surrounding natural flow stations in order to rank the information content provided by each station. The CR-DEMO analysis was conducted on the watershed using drainage area ratio inverse distance weighting (IDW-DAR) regionalization (Samuel et al., 2011) and including streamflow signatures and indicators of hydrologic alteration (IHA) (Leach et al., 2015) as additional objective functions. The results for each of these analyses were compared to the results of the previous ORB analysis and the effects of scale on the techniques used are discussed.

This research builds on the CR-DEMO method introduced in Samuel et al. (2013) and used in Leach et al. (2015) and Werstuck and Coulibaly (2016) by assessing the scaling properties of CR-DEMO and modifying the methods for operation on a watershed with limited information. The study results will be very useful to water resources engineers

and hydrologists who plan to use these emerging techniques for evaluating and/or augmenting their hydrometric networks.

Study Area and Data

Study Area

The Madawaska River Watershed is a sub-basin of the ORB near the Ontario-Quebec border. It is divided into the Upper Madawaska Watershed and the Lower Madawaska Watershed which comprise a total area of about 8,500 square kilometres (km^2). The 270 km long river has a daily average discharge which ranges between 24.4 cubic metres per second (m^3/s) in the late summer and 279.0 m^3/s during freshet (Ontario Power Generation, 2009)**.** The river discharges into the Ottawa River at Arnprior. More than 93% of the watershed is covered in forest, the majority of which is mixed wood. About 3% of the basin is cropland, and less than 1% has been developed (US. Geological Survey, 2010). The daily average temperature in the watershed ranges from -10 degrees celcius (°C) to 20 °C depending on the time of year. The average annual precipitation ranges from about 840 mm to 1040 mm, and is distributed evenly throughout the year (Environment Canada, 2015). There are 41 dams and hydroelectric generating stations operating on the watershed (Ontario Power Generation, 2009). A digital elevation model and the water bodies of the watershed are shown in Figure 8.

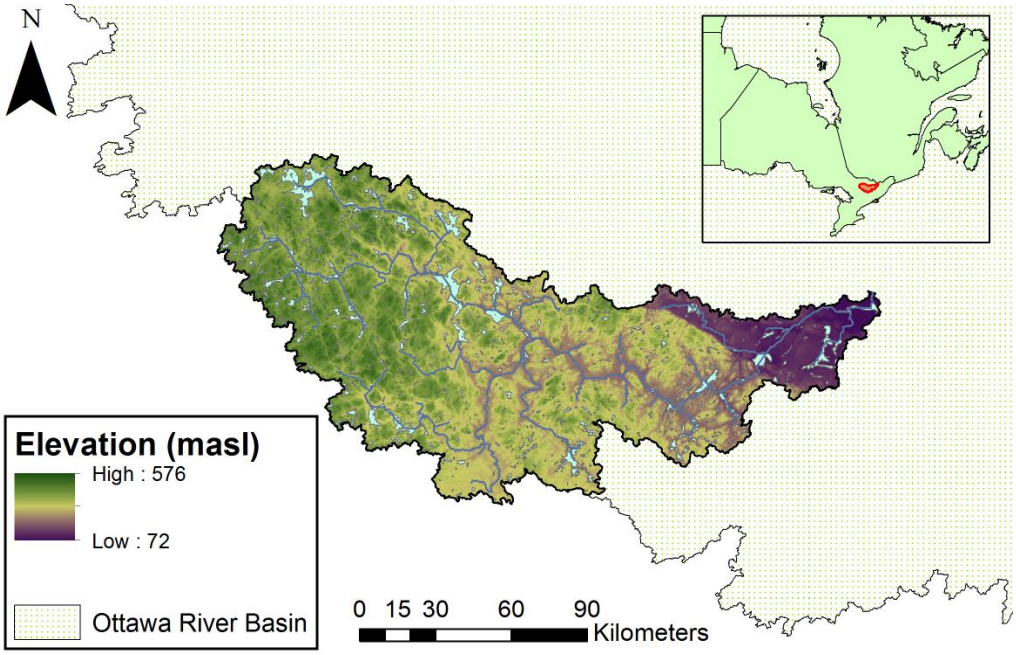

Figure 8. Madawaska Watershed digital elevation model and water bodies

Data Preprocessing

The 50,000:1 digital elevation model was downloaded from Environment Canada's (EC's) Geogratis database. This was used in ArcHydro to create the maps and to delineate sub-basins within the watershed. A total of 79 distinct subcatchments were delineated in the watershed. This is more than the 7 subcatchments which were identified in the Madawaska Watershed in the previous study (Werstuck & Coulibaly, 2016) because focusing on a smaller study area allows for finer delineation. Given the physiographic characteristics of the Madawaska Watershed, to obtain a minimum network based on the World Meteorological Organization guidelines, there should be one station for every 1875 km^2, which equals five in total (World Meteorological Organization, 2008). However, there are only two currently operating flow stations

operated by the EC Water Survey in the Madawaska Watershed. Data was acquired from nearby stations in order to compensate for this deficit in the analysis. All EC Water Survey hydrometric stations within 75 km of the watershed boundary were considered in the analysis. The runoff from the stations in the ORB should be the most similar to the actual runoff within the Madawaska Watershed, however, they are all located towards the east of the watershed. Other nearby stations which were not in the ORB were included in order to provide some context to the western region. Since these stations are located nearby, they should share many characteristics with the Madawaska Watershed which would influence their flow regimes such as soil type, land use, climate and others. In Werstuck and Coulibaly (2016), it was determined that including regulated stations was negatively effecting the regionalization process, so stations which were designated as regulated by EC were excluded from the regionalization. The flow data for these stations was acquired using EC's HYDAT database. Only stations with at least 5 years of continuous flow data between 2001 and 2010 were used.

There are 24 EC Weather Office weather stations with at least 10 years of continuous temperature and precipitation data since 1995 which are operating within 100 km of the watershed's centroid. Temperature and precipitation data from these climate stations were regionalized using inverse distance weighting (IDW) to existing and potential hydrometric station locations and used to calculate streamflow signatures. Figure 9 shows the potential station locations in the Madawaska sub basin, as well as the non-regulated hydrometric stations and the weather stations. Station 02KD004 is also displayed because

although it is regulated it is within the watershed boundary, therefore it was excluded from the regionalization process but still considered in CR-DEMO and TI analysis.

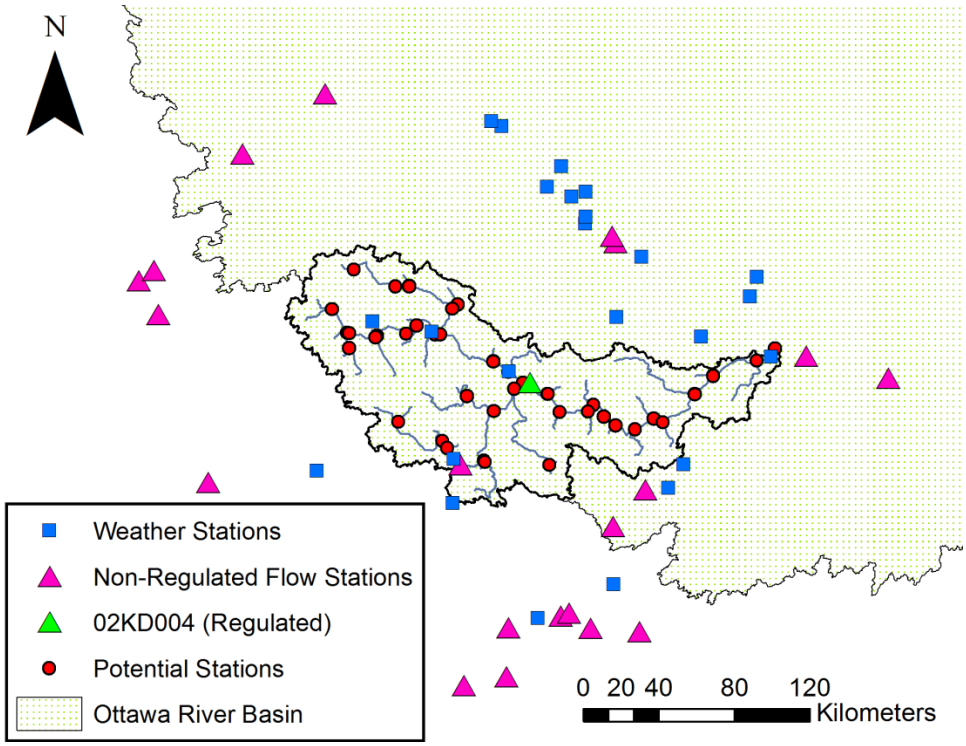

Figure 9. Madawaska Watershed environmental data stations

Methodology

Methodology Overview

A flowchart of the methodology is shown in Figure 10. Each method is explained in further detail in this section.

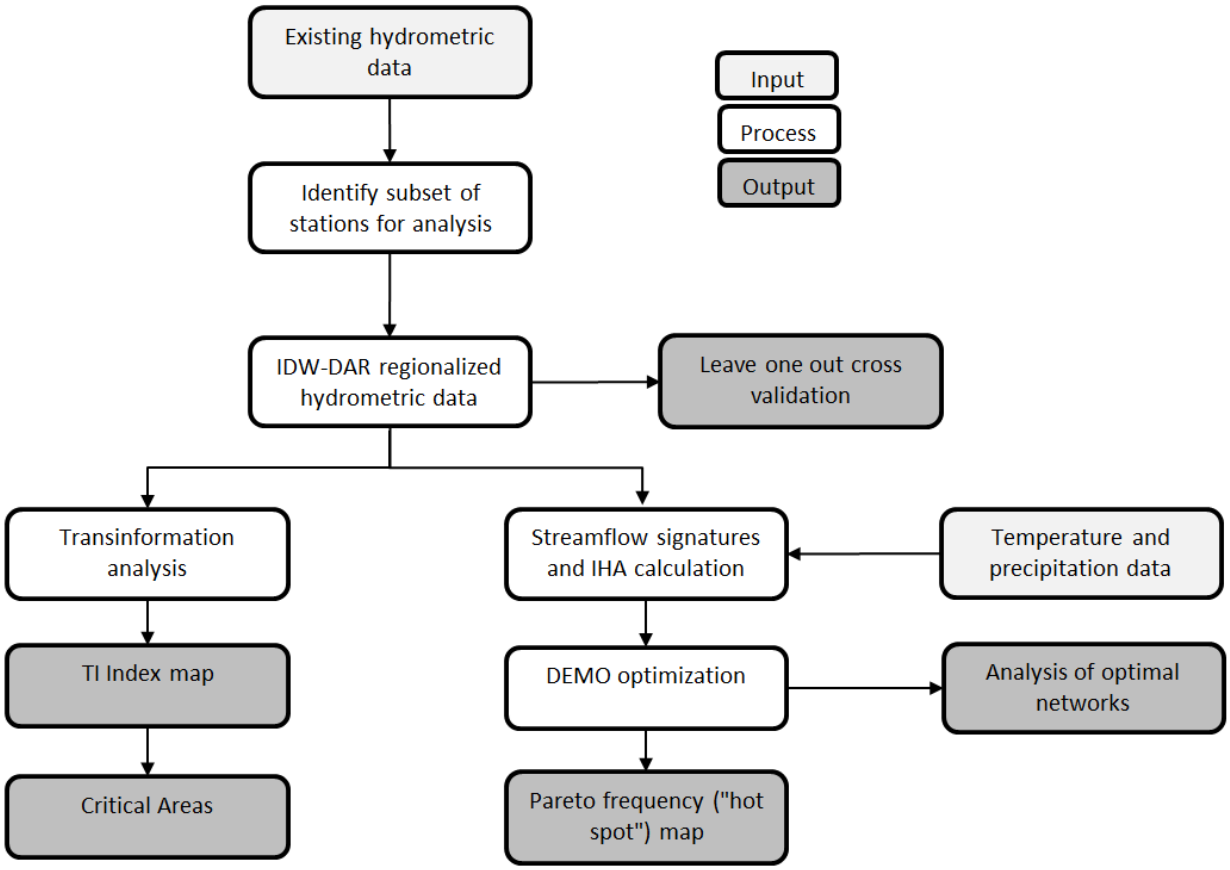

Figure 10. Madawaska Study methodology overview flowchart

Inverse Distance Weighting Drainage Area Ratio (IDW-DAR) Regionalization Method

The flow data was converted to runoff by dividing the flow by the drainage area of the sub-basin. Drainage areas provided by the EC HYDAT database were used for this calculation. The IDW-DAR method is comprised of the following equations:

$$Q_u = \sum_{i=1}^{n} w_i \left(\frac{A_u}{A_i} \right)^{\alpha} Q_i \,, \qquad w_i = \frac{\left(h_i^{-2} \right)}{\sum_{i=1}^{n} \left(h_i^{-2} \right)} \tag{1}$$

where Q_u is the runoff, Q_i is the observed flow rate at each station, α is a weighting parameter set to 1, A_u is the drainage area of the output sub-basin, A_i is the drainage area of the gauged sub-basin and h_i is the distance between the centroid of the sub-basin containing the calculated flow and the gauged sub-basin. The 10 nearest neighboring stations were considered when using this method. This process was used first to fill data at existing stations and then to generate regionalized time series at each of the potential station locations.

The years 2001-2010 were selected for the analysis. This time period was chosen in order to facilitate comparisons to Werstuck and Coulibaly (2016) where a similar analysis was done on the entire ORB. A 10 year time period was determined to be the recommended length for CR-DEMO analysis of daily streamflow series in Keum and Coulibaly (2015).

Information Theory

Shannon (1948) introduced information theory, a means of quantifying the amount of data conveyed in an informational medium. This has since been adapted to provide a quantitative means for optimal network design by using environmental data networks as seen in (Husain, 1987; Singh, 1997; Kornelsen & Coulibaly, 2015). The amount of information contained in a single time series is called the data entropy and is calculated as follows:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) \qquad (2)$$

where H is the entropy and $P(x_i)$ is the probability of event x_i. The joint entropy for N time series as would be collected from a hydrometric network is shown to be

$$H(X_1,...,X_N) = -\sum_{i_1=1}^{n1}\sum_{i_2=1}^{n2}...\sum_{i_N=1}^{nN} P(x_{1,i_1}, x_{2,i_2},..., x_{N,i_N}) \log_2 P(x_{1,i_1}, x_{2,i_2},..., x_{N,i_N}) \quad (3)$$

where x_1 through x_N represent the station locations and $x_{N,k}$ represents the k^{th} event at station N. $P(x_{1,i1}, x_{2,i2} \ldots x_{N,iN})$ is the joint probability of events i_1 through i_N at stations 1 through N. The joint entropy gives a quantitative measurement of the information content in a set of time series. Another principle of information theory is total correlation. This is defined as

$$C(X_1,...,X_N) = \left[\sum_{i=1}^{N} H(X_i)\right] - H(X_1,...,X_N) \quad (4)$$

The total correlation value shows the amount of redundant information in the dataset. The total correlation of a time series is the multivariate extension of the bivariate transinformation (TI) which is defined as follows:

$$TI(X,Y) = H(x) + H(Y) - H(X,Y) = H(Y) - H(Y \mid X) \quad (5)$$

The transinformation is used in this study to compare a hydrometric time series with a synthetic version of itself generated from multiple linear regression of the rest of the dataset. The transinformation can then be regionalized using IDW to show where in the watershed there is a surplus or deficit of information.

Transinformation (TI) Index

Transinformation is equivalent to bivariate total correlation. It is a quantitative measurement of the amount of information shared between two variables. A high transinformation value would indicate a strong dependence between two time series. Mishra and Coulibaly (2014) used transinformation to compare hydrologic data with

synthetic time series generated using multiple linear regression of the rest of the flow data in the watershed. These transinformation values were regionalized using IDW interpolation to display the areas with a redundant or deficit of information in the watershed. In this study there were only two operational hydrometric stations within the watershed, which is not enough to perform TI analysis. This analysis therefore includes stations outside of the watershed, and even some stations outside of the ORB. These stations were included in the analysis in order to provide additional information and context. Despite draining to different areas, the flow regimes at these stations should provide some information about the Madawaska Watershed because they are located nearby. The IDW regionalization was performed including these stations to obtain the TI values across the watershed.

The TI values were normalized within the watershed to between 0 and 1 to display the transinformation index defined by Mishra and Coulibaly (2014) using Equation 6. Data from the period of 2001-2010 was used in this analysis.

$$x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{6}$$

Streamflow Signatures and Indicators of Hydrologic Alteration

After the flow dataset was generated, streamflow signatures and IHA were calculated at each existing and potential station location. These parameters contain information about different parts of the hydrograph in the sub-basin, hence it is beneficial to explicitly maximize their differences when using them in CR-DEMO (Leach et al., 2015).

IHA describe the degree to which a watershed has changed over a given time period. They are commonly used to quantify the impacts of urbanization. Monk et al. (2011) identified five key IHA parameters which represented a large portion of the variation in the sub-basins studied. The five parameters identified by Monk were: one day maximum flow, one day minimum flow, Julian day of maximum, Julian day of minimum and number of reversals. These parameters were calculated for each location in this study using hydrometric data between 2001 and 2010.

Sawicz et al. (2011) defined six streamflow signatures which could be used to detect catchment responses. These were the runoff ratio (RR), the slope of the flow-duration curve (FDC), the baseflow index (BI), the streamflow elasticity (SE), the snow day ratio (SDR) and the rising limb density (RLD). Leach et al. (2015) identified three signatures which displayed the least correlation to the others and used these three as additional CR-DEMO inputs. This was done in order to reduce computational complexity while retaining most of the diversity between the signatures. Table 4 shows the sum of the intercorrelations of the streamflow signatures. The three signatures with the lowest absolute sum of intercorrelations were used in CR-DEMO. The BI, SE and RLD were identified as the most important streamflow signature parameters in this analysis.

Table 4. Madawaska Watershed streamflow signatures sum of intercorrelations

	RR	FDC	BI	SE	SDR	RLD	|Σ|
RR	1.00						2.32
FDC	-0.54	1.00					2.35
BI	-0.15	-0.49	1.00				0.87
SE	-0.62	0.37	-0.05	1.00			1.79
SDR	0.78	-0.77	0.08	-0.67	1.00		2.51
RLD	0.23	-0.18	0.11	-0.09	0.21	1.00	0.81

All signatures and IHA parameters were normalized to be between 0 and 1 using Equation 6. Additional objective functions were added to CR-DEMO in order to maximize the Euclidean distance between the vectors of streamflow signatures and the IHA at the selected station locations in order to improve the variability of flow regimes in the optimal networks.

Combined Regionalization Dual Entropy Multi-Objective Optimization (CR-DEMO)

To implement the CR-DEMO, a regionalization method is necessary to generate synthetic time series at ungauged locations where new monitoring stations may be added. After this step, a multi-objective optimization process uses the information theory principles of joint entropy and total correlation in order to identify which combination of existing and potential stations yields non-dominated solutions with maximum information content and minimum redundancy.

First, locations for potential hydrometric stations are identified in the watershed. This is done using ArcHydro and a digital elevation model to break the watershed into subcatchments. The drainage point of each subcatchment is considered to be a potential station location. Werstuck and Coulibaly (2016) determined that the IDW-DAR regionalization method was appropriate to use in this watershed to generate runoff values at each of these locations.

Second, in addition to entropy and total correlation, the streamflow signatures and IHA at each monitoring station were calculated. Thus four objective functions were used in optimization: the joint entropy was maximized, the total correlation was minimized and the Euclidean distance between both the IHA and streamflow signatures were maximized. This was done in order to ensure the information captured was as diverse as possible. The signatures and IHA explicitly increase the diversity of flow regimes chosen in natural and urbanized areas respectively.

Finally, the CR-DEMO method uses epsilon-dominance hierarchical Bayesian (ε-hBOA) optimization algorithm to solve the multi-objective network design problem and produce a Pareto front. This algorithm was determined to be efficient and robust in network design problems (Kollat et al., 2008; Reed & Kollat, 2012). Further information about the algorithm can be found in Kollat et al. (2008) and Leach et al. (2015).

The optimization output is a set of non-dominated network configurations which can be displayed in a Pareto front. Additional station locations can be ranked on importance based on the frequency that they appear in the generated Pareto front solutions. In this study, the number of additional stations was allowed to vary between 1 and 10. The

optimization was capped at 10 additional stations because solutions with too many additional stations will be computationally intensive and redundant. Solutions with fewer stations added will naturally have a lower total correlation, while solutions with more stations will have higher joint entropy. It is up to the user to determine which solution to use in practice, because depending on how the objectives are weighted, each Pareto optimal solution could be the best.

Results

Transinformation Index Results

The TI value was computed for all of the stations in the analysis. This was done by calculating the transinformation between actual runoff recorded at each station and a runoff series generated from multiple linear regression of all other stations between 2001 and 2010. The TI value gives an idea of how much mutual information is present across the watershed. This can be used as a quantitative way to assess the importance of each station. Stations existing in regions of high TI are potentially redundant, while stations existing in regions of low TI are extremely important and more stations may be necessary in these areas. This analysis is identical to the one carried out by Werstuck and Coulibaly (2016), except for the fact that stations outside of the study area were included. The TI results from the two studies are compared in Figure 11. Both studies indicate that the lower reaches of the watershed have a higher information deficit than the upper reaches.

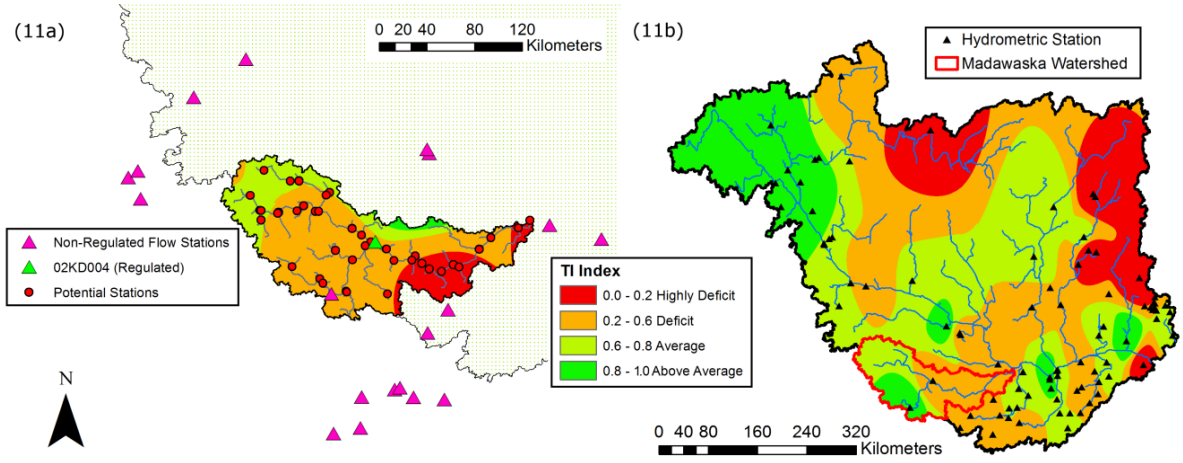

Figure 11. Map of regionalized transinformation index values

Figure 12 shows the average (bars) and standard deviation (whiskers) of the transinformation values for the stations which would result from using each dataset prior to normalization. It can clearly be seen that using only the stations in the ORB leads to sparser information in the watershed. This justifies the inclusion of stations outside of the ORB boundary because it shows that they contribute meaningful information. The dataset of natural stations narrowly registers higher transinformation values than the dataset of all stations because of the biases from regulated locations. This coupled with the fact that regulated stations lower the regionalization accuracy indicates that choosing the set of natural stations was correct. Finally, the bar to the right shows the average transinformation results of the previous ORB study. This shows that despite only having two active stations, the Madawaska Watershed has a higher information density than average in the ORB. This helps to explain why the Madawaska Watershed was not identified as a highly deficit area in the previous study.

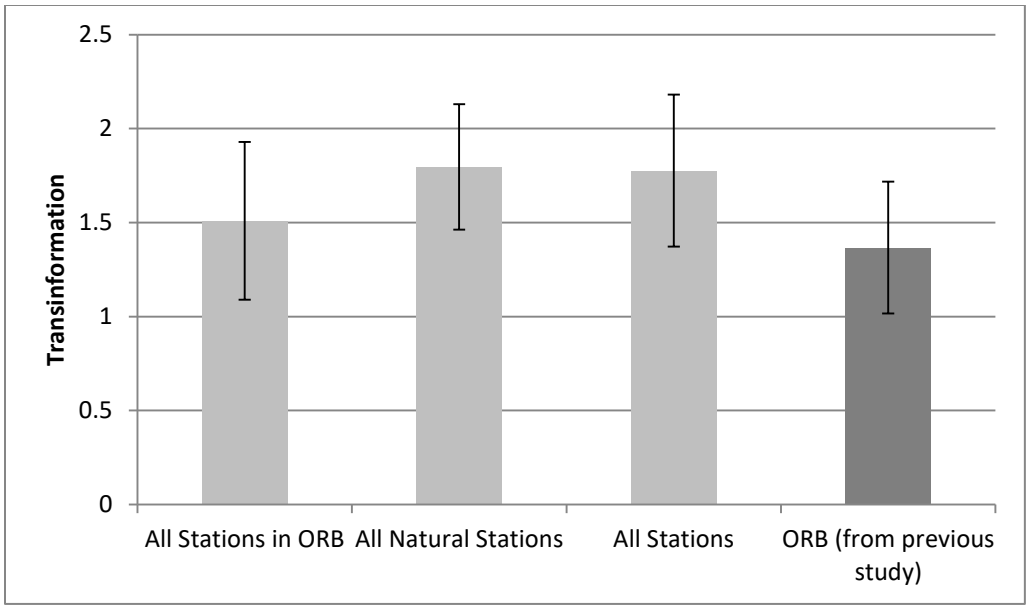

Figure 12. Transinformation values with different datasets

Table 5 shows the ratio of critical areas in the Madawaska Watershed in each analysis. When the study was scaled down to the Madawaska Watershed the deficit and highly deficit areas increase. It appears that when studied individually the Madawaska Watershed shows a larger area (about 77%) which is in a TI deficit category while in the entire ORB analysis, the deficit area for Madawaska was about 46%. This highlights the importance of selecting the appropriate basin scale for network optimization. The entropy based approach emphasizes critical areas, therefore at the ORB scale, there are more highly critical areas as compared to those of the Madawaska sub-basin. Although this could appear a limitation of the method for large basins, it could be an advantage when dealing with smaller scale watersheds.

Table 5. Madawaska Watershed critical TI areas

Basin	Area (km^2)	Ratio of Areas Under Different Categories Using TI Index			
		0.0 - 0.2 (highly deficit)	0.2 - 0.6 (deficit)	0.6 - 0.8 (average)	0.8 - 1.0 (above average)
Madawaska Watershed - Individual Study	8,572	0.14	0.63	0.21	0.02
Madawaska Watershed - ORB Study	8,572	0.00	0.46	0.34	0.21

Figure 13 shows the regionalized frequency of appearance in the Pareto front for each potential station in the CR-DEMO analysis. Figure 13a shows the results from the Madawaska Watershed and Figure 13b shows the results from the ORB study. Two major areas were identified in the lower reaches of the Madawaska Watershed: at the outflow near Arnprior and near the town of Matawatchan. An additional location on the southern tributary of Little Mississippi River near Weslemkoon Lake was also frequently selected. The selection of these downstream locations for additional stations is supported by the TI analysis. Finally, there are some upstream locations which showed some selection frequency in the set of Pareto optimal solutions. The ORB study did not identify any additional stations in the Madawaska region due to more extreme information deficits in the northeast of the ORB.

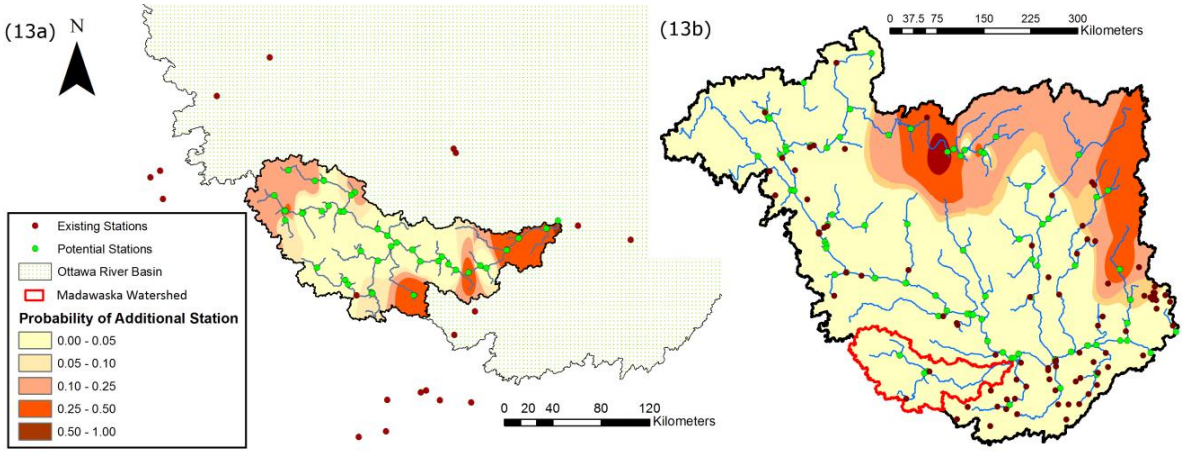

Figure 13. Madawaska Watershed CR-DEMO results using signatures and IHA

Figure 14 shows the results of the CR-DEMO analysis without including signatures and IHA as objective functions; Figure 14a in the Madawaska Watershed and Figure 14b in the ORB. These results show a higher frequency of selections in the upper reaches of the watershed than in Figure 13a. Again, the ORB result does not recommend the Madawaska Watershed as an area for additional stations.

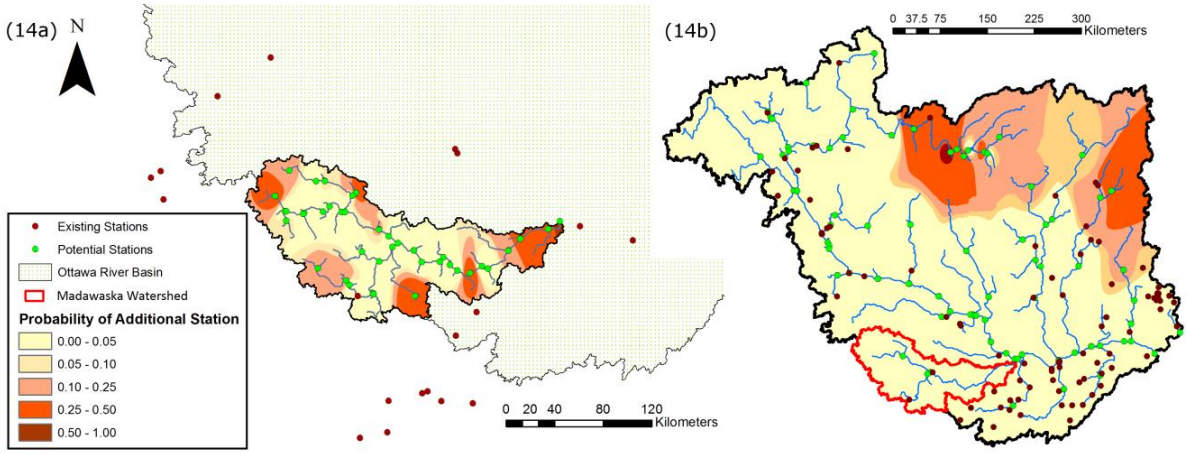

Figure 14. Madawaska Watershed CR-DEMO results without signatures and IHA

Sample network configuration solutions including signatures and IHA are shown in Figure 15. Figure 15a shows the four dimensional Pareto front of non-dominated solutions identified by CR-DEMO. Three of these solutions are identified on the Pareto front and displayed in Figures 15b through 15d. The Pareto front is naturally broken into 10 groups, each corresponding to the addition of a certain number of stations. As the number of additional stations in the solution increases, the joint entropy of the network, the total correlation of the network, the streamflow signatures Euclidean and the IHA Euclidean all increase, and the groups of solutions become larger. This output shows the flexibility of the CR-DEMO method, as each of the solutions in the Pareto front can be optimal depending on how the objectives are weighted by the user. Networks 1 to 3 are sample configurations, showing how each point on the Pareto front is a non-dominated network solution and how the number of additional stations on the Pareto front increases up and to the right (Figure 15a).

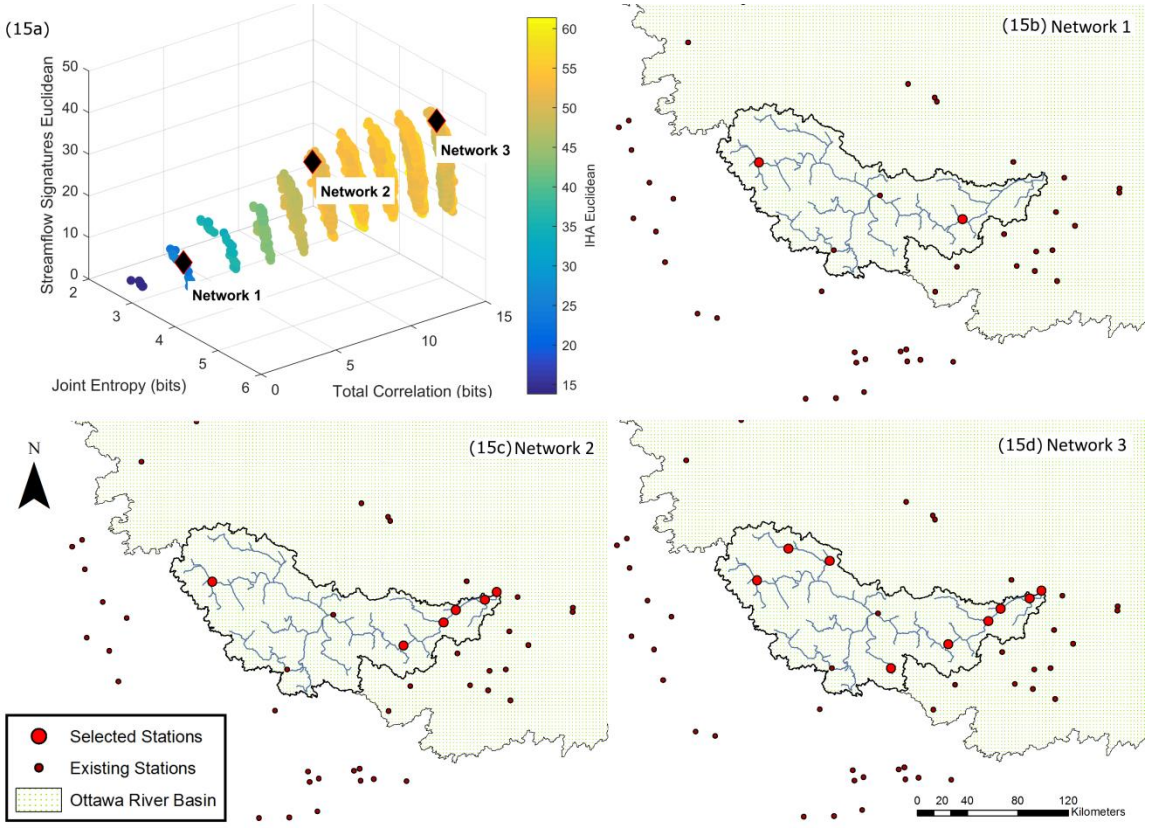

Figure 15. Madawaska Watershed Pareto front and examples of Pareto optimal networks

Figure 16 shows the two dimensional Pareto front of the CR-DEMO output. This figure clearly shows how the solutions are clustered into groups depending on the number of stations added. It also shows how the joint entropy and total correlation both increase as the number of stations increases. The joint entropy can be seen to increase the most per additional station when one to three stations are added, and increase less significantly per station past that point. This is interesting because the WMO recommendation for a minimum network in this watershed is to build three additional monitoring stations. The analysis results can be used to design an optimal network that meets the WMO

guidelines. The results also indicate that the first three stations built will have the greatest benefit to the joint entropy in the network.

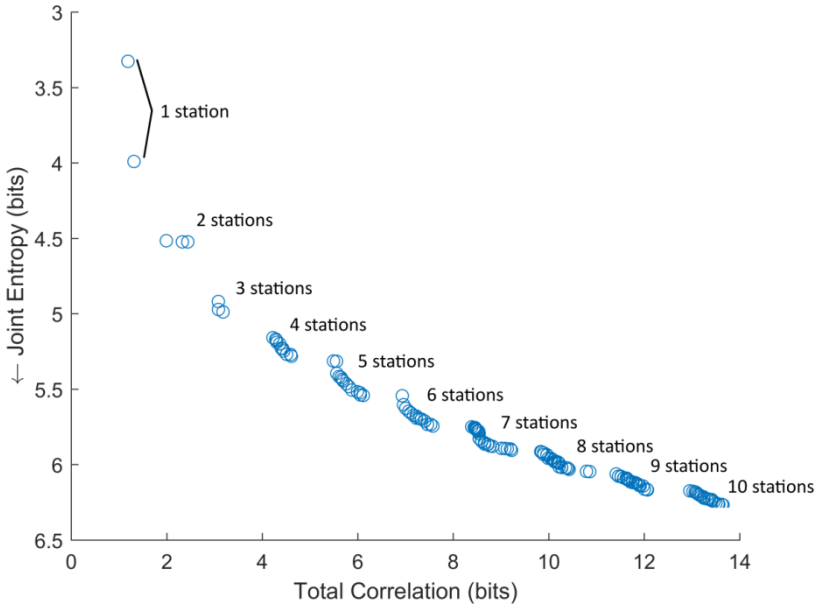

Figure 16. Two dimensional Pareto front

Conclusions

The information theory based methods of TI analysis and CR-DEMO were applied in the area of the Madawaska Watershed in order to rank the importance of the existing stations, estimate the information density in the watershed and determine optimal locations for additional stations in the hydrometric network. This study was designed to be comparable to a previous investigation by Werstuck and Coulibaly (2016) in which the same analysis was conducted on the ORB which encompasses the Madawaska Watershed. The comparison allowed conclusions to be drawn regarding the scaling properties of these

methods, as well as the feasibility of including hydrometric stations from outside of the watershed when there are too few inside of it to properly perform an analysis.

The TI analysis produced similar results in both studies. Both studies indicated that the lower reaches of the Madawaska Watershed were at a higher information deficit than its upper reaches. A comparison of the average TI values of various subsets of stations also confirmed that including nearby stations even if they were not inside of the ORB and removing regulated stations provided the most information dense dataset. The results of this analysis were reasonable and intuitive and it appears that the results of the TI analysis were not negatively affected by scaling.

CR-DEMO was conducted on the Madawaska Watershed dataset twice; once including streamflow signatures and IHA as objective functions and once just using the information theory principles of joint entropy and total correlation. Both analyses identify three key potential locations in the lower reaches of the watershed; near Arnprior, near Matawatchan and near Weslemkoon Lake. Both analyses also identified some key potential locations in the tributaries of the watershed in Algonquin Park. Interestingly, excluding streamflow signatures and IHA appears to decrease the deficit area identified by CR-DEMO. This is assumed to be because the lower reaches of the watershed have more complex streamflow signatures and IHA than the upper reaches, as reinforced by the information deficit discovered in the TI analysis. In both cases the previous study which considered the entire ORB did not identify the Madawaska Watershed as an ideal location for additional monitoring stations because of other areas in the river basin which were at a more extreme information deficit. This shows that when applying CR-DEMO

for network design, the area of interest should be carefully defined. In addition, potential station locations should only be placed in realistic locations. Otherwise, there is the potential for extreme information density or deficit outside of the area of interest biasing the results. This was the case when looking at the Madawaska Watershed results within the ORB study; despite the need for additional stations in the Madawaska Watershed, the ORB study overlooked the area in favor of locations at a greater information deficit.

The Pareto front was plotted and displayed along with sample Pareto optimal solutions. As expected, each objective increases with the addition of more stations and the Pareto front is naturally divided into groups depending on the number of additional stations. When considered in the two information theory dimensions of joint entropy and total correlation, the Pareto front can be observed as a curve. This curve shows that the sharpest increase in joint entropy occurs with the addition of the first three stations and after this threshold the joint entropy increase diminishes per station added. This is interesting because it supports the WMO minimum network density recommendation of three additional stations for the Madawaska Watershed.

It was shown that in this analysis TI analysis does not appear to be significantly affected by scaling. CR-DEMO analysis does appear to be affected by scaling and this should be considered when using this method in the future. It was also shown that nearby stations outside of the area of interest can be included in TI analysis and the regionalization step of CR-DEMO when there is insufficient data within the basin.

Acknowledgements

This research was supported jointly by Ontario Power Generation (OPG) and Natural Science and Engineering Research Council (NSERC) of Canada. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada, and by datasets from Environment Canada, Hydro-Quebec, and OPG. The authors are grateful to Dr. Joshua Kollat (Penn State University) who developed the ε-hBOA, and provided the source codes.

References

Alfonso, L., Lobbrecht, A., & Price, R. (2010a). Information theory-based approach for location of monitoring water level gauges in polders. *Water Resources Research*, *46*(3), W03528. http://doi.org/10.1029/2009WR008101

Alfonso, L., Lobbrecht, A., & Price, R. (2010b). Optimization of water level monitoring network in polder systems using information theory. *Water Resources Research*, *46*(12), W12553. http://doi.org/10.1029/2009WR008953

Amorocho, J., & Espildora, B. (1973). Entropy in the assessment of uncertainty in hydrologic systems and models. *Water Resources Research*, *9*(6), 1511–1522.

Chen, Y.-C., Wei, C., & Yeh, H.-C. (2008). Rainfall network design using kriging and entropy. *Hydrological Processes*, *22*(3), 340–346. http://doi.org/10.1002/hyp.6292

Environment Canada. (2015). Canadian Climate Normals 1981-2010 Station Data. Retrieved from http://climate.weather.gc.ca

Fiering, M. B. (1965). An optimization scheme for gaging. *Water Resources Research*, *1*(4), 463–470.

Husain, T. (1979). *Shannon's Information Theory in Hydrologic Network Design and Estimation*. University of British Columbia.

Husain, T. (1987). Hydrologic Network Design Formulation. *Canadian Water Resources Journal*, *12*(1), 44–63. http://doi.org/10.4296/cwrj1201044

Husain, T. (1989). Hydrologic Uncertainty Measure and Network Design. *Journal of the American Water Resources Association*, *25*(3), 527–534. http://doi.org/10.1111/j.1752-1688.1989.tb03088.x

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *The Physical Review*, *106*(4), 620–630.

Keum, J., & Coulibaly, P. (2016). Sensitivity of Entropy Method to Time Series Length in Hydrometric Network Design. *Journal of Hydrologic Engineering*, (Submitted).

Kollat, J. B., & Reed, P. M. (2007). A Computational Scaling Analysis of Multiobjective Evolutionary Algorithms in Long-Term Groundwater Monitoring Applications. *Advances in Water Resources*, *30*(3), 408–419. http://doi.org/10.1016/j.advwatres.2006.05.009

Kollat, J. B., Reed, P. M., & Kasprzyk, J. R. (2008). A New Epsilon-Dominance Hierarchical Bayesian Optimization Algorithm for Large Multiobjective Monitoring Network Design Problems. *Advances in Water Resources*, *31*(5), 828–845. http://doi.org/10.1016/j.advwatres.2008.01.017

Kornelsen, K. C., & Coulibaly, P. (2015a). Design of an Optimal Soil Moisture Monitoring Network Using SMOS Retrieved Soil Moisture. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(7), 3950–3959. http://doi.org/10.1109/TGRS.2014.2388451

Kornelsen, K. C., & Coulibaly, P. (2015b). Design of an Optimal Soil Moisture

Monitoring Network Using SMOS Retrieved Soil Moisture, *53*(7), 3950–3959.

Krstanovic, P. F., & Singh, V. P. (1992a). Evaluation of Rainfall Networks using Entropy: I. Theoretical Development. *Water Resources Management*, *6*(4), 279–293. http://doi.org/10.1007/BF00872281

Krstanovic, P. F., & Singh, V. P. (1992b). Evaluation of Rainfall Networks using Entropy: II. Application. *Water Resources Management*, *6*(4), 295–314. http://doi.org/10.1007/BF00872282

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

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

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

Lindley, D. V. (1956). On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005.

Maddock, T. (1974). An Optimum Reduction of Gauges to Meet Data Program

Constraints. *Hydrological Sciences Bulletin*, *19*(3), 337–345. http://doi.org/10.1080/02626667409493920

Mishra, A. K., & Coulibaly, P. (2009). Developments in Hydrometric Network Design: A Review. *Reviews of Geophysics*, *47*(2), RG2001. http://doi.org/10.1029/2007RG000243

Mishra, A. K., & Coulibaly, P. (2010). Hydrometric Network Evaluation for Canadian Watersheds. *Journal of Hydrology*, *380*(3-4), 420–437. http://doi.org/10.1016/j.jhydrol.2009.11.015

Mishra A. K., & Coulibaly P. 2014 Variability in Canadian seasonal streamflow information and its implication for hydrometric network design. *ASCE Journal of Hydrologic Engineering* 19(8), DOI:10.1061/(ASCE)HE.1943-5584.0000971.

Monk, W. A., Peters, D. L., Curry, R., & Baird, D. J. (2011). Quantifying trends in indicator hydroecological variables for regime-based groups of Canadian rivers. *Hydrological Processes*, *25*, 3086–3100.

Moss, M. E., & Karlinger, M. R. (1974). Surface Water Network Design by Regression Analysis Simulation. *Water Resources Research*, *10*(3), 427–433. http://doi.org/10.1029/WR010i003p00427

Ontario Power Generation. (2009). *Madawaska River Water Management Plan*.

Ozkul, S., Harmancioglu, N. B., & Singh, V. P. (2000). Entropy-Based Assessment of Water Quality Monitoring Networks. *Journal of Hydrologic Engineering*. Retrieved

from http://ascelibrary.org/doi/abs/10.1061/(ASCE)1084-0699(2000)5:1(90)

Reed, P. M., & Kollat, J. B. (2012). Save now, pay later? Multi-period many-objective groundwater monitoring design given systematic model errors and uncertainty. *Advances in Water Resources*, *35*, 55–68. http://doi.org/10.1016/j.advwatres.2011.10.011

Samuel, J., Coulibaly, P., & Kollat, J. B. (2013a). CRDEMO: Combined Regionalization and Dual Entropy-Multiobjective Optimization for Hydrometric Network Design. *Water Resources Research*, *49*(12), 8070–8089. http://doi.org/10.1002/2013WR014058

Samuel, J., Coulibaly, P., & Kollat, J. B. (2013b). CRDEMO: Combined regionalization and dual entropy-multiobjective optimization for hydrometric network design. *Water Resources Research*, *49*(12), 8070–8089. http://doi.org/10.1002/2013WR014058

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

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., & Carrillo, G. (2011). Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences*, *15*, 2895–2911.

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

Singh, V. P. (1997). The use of entropy in hydrology and water resources. *Hydrological Processes*, *11*(1996), 587–626.

Singh, V. P., & Rajagopal, A. K. (1987). Water for the Future: Hydrology in Perspective (Proceedings of the Rome Symposium, April 1987). In *International Associations of Hydrological Sciences* (Vol. 164).

Sonuga, J. O. (1976). Entropy Principle Applied to the Rainfall-Runoff Process. *Journal of Hydrology*, *30*, 81–94.

US. Geological Survey. (2010). 0.5 km MODIS-based Clobal Land Cover Climatology. Retrieved from http://landcover.usgs.gov/global_climatology.php

Werstuck, C., & Coulibaly, P. (2016). Hydrometric Network Design Using Dual Entropy Multi-Objective Optimization in the Ottawa River Basin. *Hydrology Research*, *Submitted*.

World Meteorological Organization. (2008). *Guide to Hydrological Practices, Volume I Hydrology – From Measurement to Hydrological Information, WMO-No. 168* (Sixth). Retrieved from http://www.hydrology.nl/images/docs/hwrp/WMO_Guide_168_Vol_I_en.pdf

Yang, Y., & Burn, D. H. (1994). An Entropy Approach to Data Collection Network

Design. *Journal of Hydrology, 157*(1-4), 307–324. http://doi.org/10.1016/0022-1694(94)90111-2

4.0 Conclusions and Recommendations

4.1 Overall Conclusions

As efficient management of water resources becomes more important, informative, high quality data collection networks become more critical. Network design using the CR-DEMO method is a valuable tool for improving the hydrometric data networks in Canadian watersheds. The research presented in this thesis investigated the use of this method in the Ottawa River Basin and one of its sub-basins, the Madawaska Watershed. This research was focused on decisions made in CR-DEMO operation such as which regionalization technique to use, how to handle stations with unnatural flow regimes and whether to include surrounding stations in the regionalization process. TI analysis was tested as a support tool to be used in conjunction with DEMO as a preliminary method of measuring the information density in the network and thus assessing the relative importance of each existing station. The scaling properties of CR-DEMO and of TI analysis were also investigated. Finally, recommendations were made for additional station locations in both the ORB as a whole and the Madawaska Watershed individually. Conclusions which were drawn from this work are as follows:

1) The IDW-DAR regionalization method performed better than MAC-HBV in a leave one out cross validation test for the basin in question. This regionalization was further improved by omitting stations which were known to be regulated.
2) TI analysis was proven to be a fast and efficient way to measure the information density in the basin. This method is less computationally intense than CR-DEMO,

and the results were consistent with the CR-DEMO results in both analyses. TI analysis output can be normalized and mapped across the watershed to spatially show the information density. It also provides a quantitative means of ranking the information contribution from each individual station, data which is very important to professionals maintaining water resources networks. It is recommended that TI analysis be used in conjunction with CR-DEMO analysis in the future.

3) It was found that CR-DEMO analysis was significantly affected by scaling. In the ORB, the northeast area was selected almost exclusively as the optimal location for additional stations. When the area of interest was scaled down to just the Madawaska Watershed, the northeast area was omitted, and thus different areas were selected for additional stations. The high information deficit in the northeast area of the ORB biased the CR-DEMO results in the Madawaska Watershed. In the future, the area of interest should be carefully considered when conducting CR-DEMO analysis, potential station locations should only be placed where a station can actually be built. It was found that scaling did not affect the TI analysis as severely. When plotted on a map, the relative location of TI dense and deficit areas in the Madawaska Watershed stayed the same.
4) It was discovered that including stations outside of the area of interest in TI analysis increased the information density of the Madawaska Watershed. This was done because there were only two operational hydrometric stations within the Madawaska watershed, so stations in the surrounding area were included to

provide context to the watershed's remote areas. In future situations when there are areas in a watershed which are remote from monitoring stations, or if there are not enough stations to perform meaningful CR-DEMO and TI analysis, monitoring stations from outside of the basin can be included in TI analysis and CR-DEMO regionalization.

5) Finally, a TI density map, a TI ranking of existing stations, a CR-DEMO hotspot map and some sample Pareto optimal configurations were provided for both the ORB and the Madawaska Watershed. These tools provide important information and recommendations for the optimal location of additional monitoring stations in these watersheds.

4.2 Contributions

The research presented in this thesis was done as part of the NSERC Strategic Project entitled: "*Decision Support System for Water Monitoring Network Evaluation and Design*". A project supported jointly by NSERC, Environment Canada, BC-Hydro, Hydro-Quebec, Ontario Power Generation and the Ontario Ministry of the Environment.

CR-DEMO has been developed in the McMaster University Water Resources and Hydrologic Modelling (WRHML) Lab led by Dr. Paulin Coulibaly. This research expanded on previous studies in network design done by Kollat et al. (2008); Samuel et al. (2013); Kornelsen and Coulibaly (2015); Keum and Coulibaly (2015) and Leach et al. (2015) in conjunction with the WRHML.

In addition, this research was submitted to Ontario Power Generation in the form of two reports detailing the recommended improvements of their hydrometric networks in the Ottawa River Basin and Madawaska Watershed.

4.3 Recommendations

The CR-DEMO method of network design holds a lot of potential as a tool for improving the distribution of Canadian hydrometric networks. Sample network configurations with recommended station locations for both the Ottawa River Basin and the Madawaska Watershed are enclosed in Chapters 2 and 3. In addition to these sample configurations, Pareto selection frequency maps, TI rankings of existing stations and TI index maps are provided for both areas. This output will be extremely beneficial to water resources decision makers operating in the ORB.

In the future, it is recommended that TI analysis be used in conjunction with CR-DEMO. This method allows the importance of existing stations to be ranked, and outputs a map of information density which supports the CR-DEMO results. The area of interest should be carefully considered when operating CR-DEMO. This research found that areas of extreme information deficit were highly recommended as additional station locations by CR-DEMO. This caused other areas such as the Madawaska Watershed, which does not have sufficient minimum network density as recommended by the WMO, to be selected less frequently. This was expected, as CR-DEMO recommends the potential station locations which optimally augment the existing network. If a station needs to be built in a specific area, the potential station locations should be limited to that area. This also

means that CR-DEMO results from a large basin cannot be scaled and applied to its individual sub-basins.

It was also discovered that the IDW-DAR regionalization method is sufficient for use with CR-DEMO in Ontario. The regionalization accuracy can be improved by omitting regulated stations from the regionalization process. If there are not enough stations in the area of interest for analysis, surrounding stations can be included in regionalization. Although the IDW-DAR method is sufficient, the regionalization step can be accomplished a number of different ways depending on user preference. When choosing a regionalization technique the most important criteria is that it can accurately estimate the shape of probability distribution at the ungauged location. As regionalization methods improve in the future this step can be modified.

References

Alfonso, L., Lobbrecht, A., & Price, R. (2010a). Information theory-based approach for location of monitoring water level gauges in polders. *Water Resources Research*, *46*(3), W03528. http://doi.org/10.1029/2009WR008101

Alfonso, L., Lobbrecht, A., & Price, R. (2010b). Optimization of water level monitoring network in polder systems using information theory. *Water Resources Research*, *46*(12), W12553. http://doi.org/10.1029/2009WR008953

Amorocho, J., & Espildora, B. (1973). Entropy in the assessment of uncertainty in hydrologic systems and models. *Water Resources Research*, *9*(6), 1511–1522.

Chen, Y.-C., Wei, C., & Yeh, H.-C. (2008). Rainfall network design using kriging and entropy. *Hydrological Processes*, *22*(3), 340–346. http://doi.org/10.1002/hyp.6292

Environment Canada. (2015). Canadian Climate Normals 1981-2010 Station Data. Retrieved from http://climate.weather.gc.ca

Fiering, M. B. (1965). An optimization scheme for gaging. *Water Resources Research*, *1*(4), 463–470.

Husain, T. (1979). *Shannon's Information Theory in Hydrologic Network Design and Estimation*. University of British Columbia.

Husain, T. (1987). Hydrologic Network Design Formulation. *Canadian Water Resources Journal*, *12*(1), 44–63. http://doi.org/10.4296/cwrj1201044

Husain, T. (1989). Hydrologic Uncertainty Measure and Network Design. *Journal of the American Water Resources Association*, *25*(3), 527–534. http://doi.org/10.1111/j.1752-

1688.1989.tb03088.x

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *The Physical Review*, *106*(4), 620–630.

Keum, J., & Coulibaly, P. (2016). Sensitivity of Entropy Method to Time Series Length in Hydrometric Network Design. *Journal of Hydrologic Engineering*, (Submitted).

Kollat, J. B., & Reed, P. M. (2007). A Computational Scaling Analysis of Multiobjective Evolutionary Algorithms in Long-Term Groundwater Monitoring Applications. *Advances in Water Resources*, *30*(3), 408–419. http://doi.org/10.1016/j.advwatres.2006.05.009

Kollat, J. B., Reed, P. M., & Kasprzyk, J. R. (2008). A New Epsilon-Dominance Hierarchical Bayesian Optimization Algorithm for Large Multiobjective Monitoring Network Design Problems. *Advances in Water Resources*, *31*(5), 828–845. http://doi.org/10.1016/j.advwatres.2008.01.017

Kornelsen, K. C., & Coulibaly, P. (2015a). Design of an Optimal Soil Moisture Monitoring Network Using SMOS Retrieved Soil Moisture. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(7), 3950–3959. http://doi.org/10.1109/TGRS.2014.2388451

Kornelsen, K. C., & Coulibaly, P. (2015b). Design of an Optimal Soil Moisture Monitoring Network Using SMOS Retrieved Soil Moisture, *53*(7), 3950–3959.

Krstanovic, P. F., & Singh, V. P. (1992a). Evaluation of Rainfall Networks using Entropy: I. Theoretical Development. *Water Resources Management*, *6*(4), 279–293. http://doi.org/10.1007/BF00872281

Krstanovic, P. F., & Singh, V. P. (1992b). Evaluation of Rainfall Networks using Entropy: II.

Application. *Water Resources Management*, *6*(4), 295–314. http://doi.org/10.1007/BF00872282

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

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

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

Lindley, D. V. (1956). On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005.

Maddock, T. (1974). An Optimum Reduction of Gauges to Meet Data Program Constraints. *Hydrological Sciences Bulletin*, *19*(3), 337–345. http://doi.org/10.1080/02626667409493920

Mishra, A. K., & Coulibaly, P. (2009). Developments in Hydrometric Network Design: A Review. *Reviews of Geophysics*, *47*(2), RG2001. http://doi.org/10.1029/2007RG000243

Mishra, A. K., & Coulibaly, P. (2010). Hydrometric Network Evaluation for Canadian Watersheds. *Journal of Hydrology*, *380*(3-4), 420–437. http://doi.org/10.1016/j.jhydrol.2009.11.015

Mishra, A. K., & Coulibaly, P. (2014). Variability in Canadian Seasonal Streamflow Information and Its Implication for Hydrometric Network Design. ASCE *Journal of Hydrologic Engineering*, *19*(8), 5014003. http://doi.org/10.1061/(ASCE)HE.1943-5584.0000971

Monk, W. A., Peters, D. L., Curry, R., & Baird, D. J. (2011). Quantifying trends in indicator hydroecological variables for regime-based groups of Canadian rivers. *Hydrological Processes*, *25*, 3086–3100.

Moss, M. E., & Karlinger, M. R. (1974). Surface Water Network Design by Regression Analysis Simulation. *Water Resources Research*, *10*(3), 427–433. http://doi.org/10.1029/WR010i003p00427

Ontario Power Generation. (2009). *Madawaska River Water Management Plan*.

Ozkul, S., Harmancioglu, N. B., & Singh, V. P. (2000). Entropy-Based Assessment of Water Quality Monitoring Networks. *Journal of Hydrologic Engineering*. Retrieved from http://ascelibrary.org/doi/abs/10.1061/(ASCE)1084-0699(2000)5:1(90)

Reed, P. M., & Kollat, J. B. (2012). Save now, pay later? Multi-period many-objective groundwater monitoring design given systematic model errors and uncertainty. *Advances in Water Resources*, *35*, 55–68. http://doi.org/10.1016/j.advwatres.2011.10.011

Samuel, J., Coulibaly, P., & Kollat, J. B. (2013a). CRDEMO: Combined Regionalization and Dual Entropy-Multiobjective Optimization for Hydrometric Network Design. *Water Resources Research*, *49*(12), 8070–8089. http://doi.org/10.1002/2013WR014058

Samuel, J., Coulibaly, P., & Kollat, J. B. (2013b). CRDEMO: Combined regionalization and dual entropy-multiobjective optimization for hydrometric network design. *Water Resources*

Research, *49*(12), 8070–8089. http://doi.org/10.1002/2013WR014058

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

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., & Carrillo, G. (2011). Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences*, *15*, 2895–2911.

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

Singh, V. P. (1997). The use of entropy in hydrology and water resources. *Hydrological Processes*, *11*(1996), 587–626.

Singh, V. P., & Rajagopal, A. K. (1987). Water for the Future: Hydrology in Perspective (Proceedings of the Rome Symposium, April 1987). In *International Associations of Hydrological Sciences* (Vol. 164).

Sonuga, J. O. (1976). Entropy Principle Applied to the Rainfall-Runoff Process. *Journal of Hydrology*, *30*, 81–94.

US. Geological Survey. (2010). 0.5 km MODIS-based Clobal Land Cover Climatology. Retrieved from http://landcover.usgs.gov/global_climatology.php

Werstuck, C., & Coulibaly, P. (2016). Hydrometric Network Design Using Dual Entropy Multi-Objective Optimization in the Ottawa River Basin. *Hydrology Research, Submitted*.

World Meteorological Organization. (2008). *Guide to Hydrological Practices, Volume I*

Hydrology – From Measurement to Hydrological Information, WMO-No. 168 (Sixth). Retrieved from http://www.hydrology.nl/images/docs/hwrp/WMO_Guide_168_Vol_I_en.pdf

Yang, Y., & Burn, D. H. (1994). An Entropy Approach to Data Collection Network Design. *Journal of Hydrology*, *157*(1-4), 307–324. http://doi.org/10.1016/0022-1694(94)90111-2