Hydrometric network design using dual entropy multi-objective optimization in the Ottawa River Basin

Connor Werstuck and Paulin Coulibaly

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. 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 also examined. The Ottawa River Basin, a large Canadian watershed with a number of regulated hydroelectric dams, was selected for the experiment. The TI analysis approach provides preliminary information which is supported by DEMO results. Regionalization was shown to be more accurate when the regulated basin stations 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.

Key words | entropy, hydrometric network, multi-objective optimization, network design, water resources

Connor Werstuck
Paulin Coulibaly (corresponding author)
Department of Civil Engineering and School of Geography and Earth Sciences,
McMaster University,
1280 Main Street West,
Hamilton,
Ontario,
Canada,
L8S 4L8
E-mail: *couliba@mcmaster.ca*

INTRODUCTION

Hydrometric networks provide important data for water researchers and water resources decision-makers. Recent research (Burn 1997; Mishra & 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 stations' 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 & 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 & 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

doi: 10.2166/nh.2016.344

the network was collecting. Li *et al.* (2012) considered joint entropy, transinformation (TI) 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. Xu *et al.* (2015) considered both mutual information and accuracy metrics such as Nash–Sutcliffe efficiency (NSE) in a multi-objective optimization process. This method was used to investigate the effects of network size on the data quality. In addition to information theory, tools such as cluster analysis, feature selection, artificial neural networks, probability constraints, Box–Hill discrimination and spatiotemporal variogram models are used in network design today (Xu *et al.* 2013; Asghari & Nasseri 2014; Shafiei *et al.* 2014; Wu *et al.* 2015; Chang *et al.* 2016; Pham & Tsai 2016).

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 (IHA) 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). 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 TI 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 & 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 (LOOCV) method 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 TI 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 1,950 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 is 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.

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. In total, 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 four additional stations in the basin were 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.

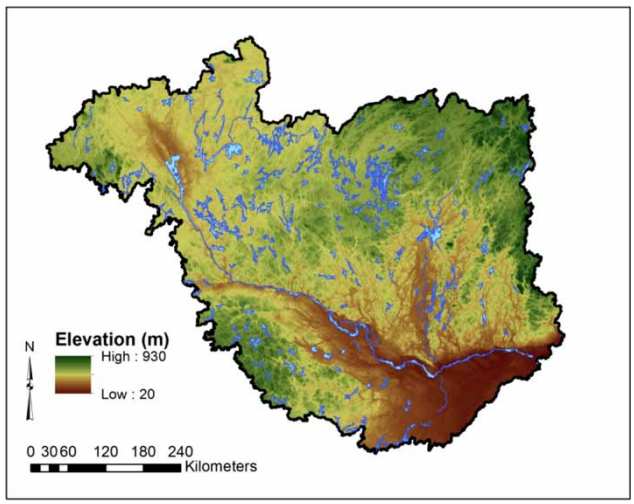

Figure 1 | Ottawa River Basin digital elevation model and water bodies.

METHODOLOGY

Methodology overview

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

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 & Blöschl (2004). The parameter set which produced the highest validation 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

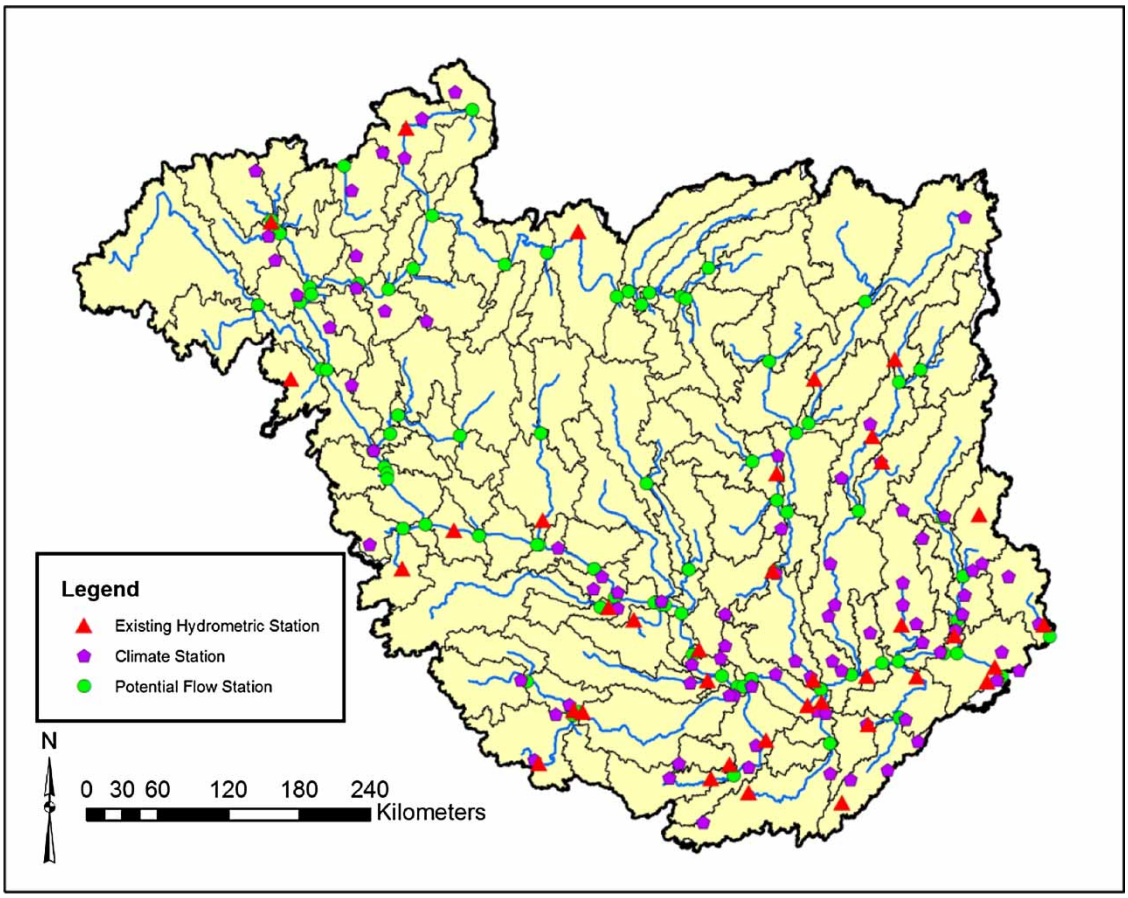

Figure 2 | Ottawa River Basin environmental data stations.

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.

IDW-DAR regionalization method

The flow data were 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:

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

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

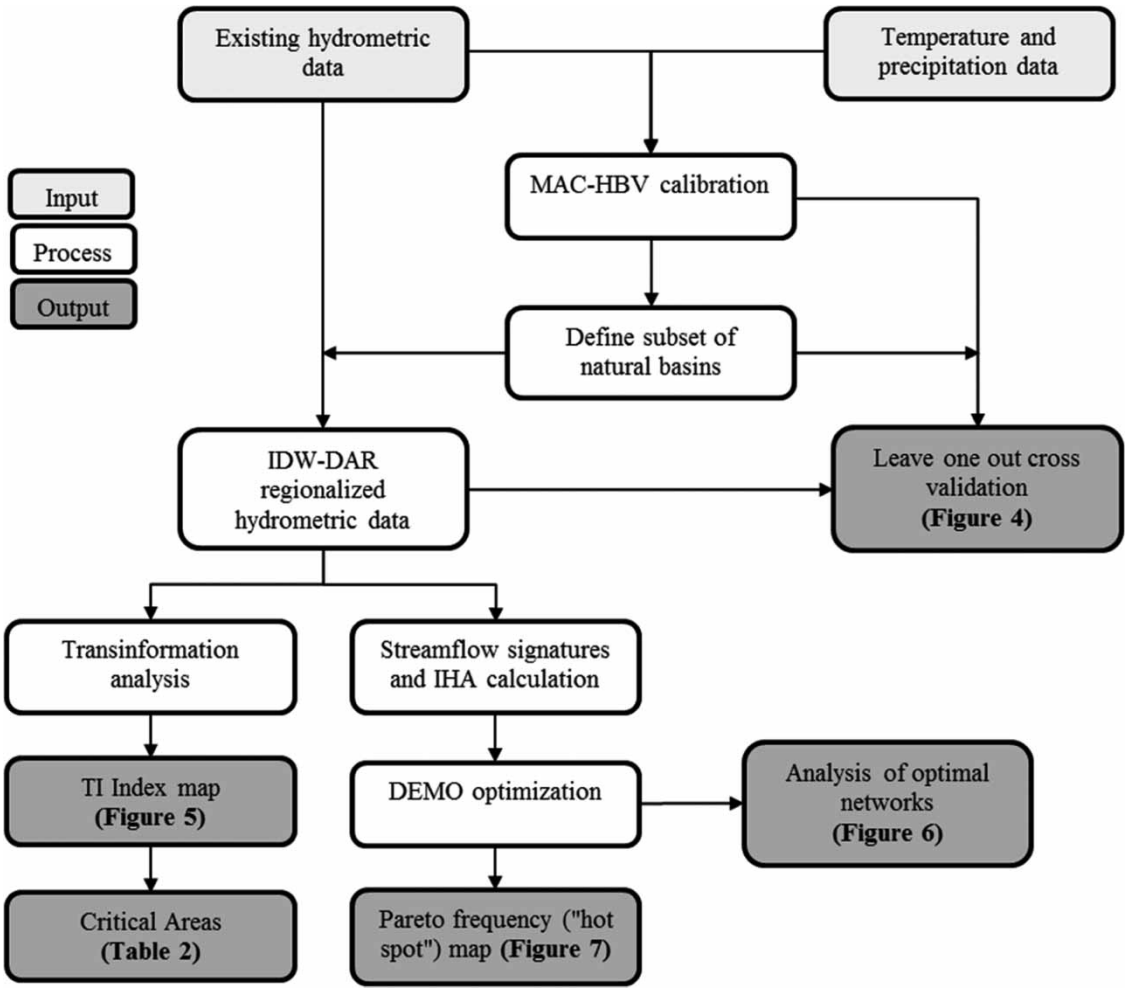

Figure 3 | Methodology overview flowchart.

where 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 & Coulibaly in press). The missing data were also filled using the IDW-DAR method.

Streamflow signatures and IHA

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 33rd and 66th 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 °C. 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 streamflow elasticity (SE), snow day ratio (SDR) and rising limb density (RLD), and thus these were used in the analysis.

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 proportion 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 was 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).

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

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

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:

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

where x_1 through x_N represent the station locations and $x_{N,k}$ represents the datapoint at station N at timestep k. 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:

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

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 TI which is defined as follows:

$$TI(X, Y) = H(x) + H(Y) - H(X, Y) = H(Y) - H(Y|X) \tag{7}$$

The TI 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 TI can then be regionalized using kriging to show where in the watershed there is a surplus or deficit of information.

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 are generated for each of these potential locations. Once the data are 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 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 distances 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 one and ten 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 ten new stations if optimal locations can be identified.

TI index

TI is equivalent to bivariate total correlation. It is a quantitative measurement of the amount of information shared between two variables. A high TI value would indicate a strong dependence between two time series. Mishra & Coulibaly (2014) used TI 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 TI values were regionalized using IDW interpolation to display the areas with a redundant or deficit of information in the watershed. The TI values were normalized to between 0 and 1 to display the TI index defined by Mishra & 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 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. Certain stations were represented particularly poorly by MAC-HBV. These same stations tended to produce low NSE values when modelled using 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 10th, 25th, 50th, 75th and 90th 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.

TI index values

The TI index 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 stations 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 four 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.

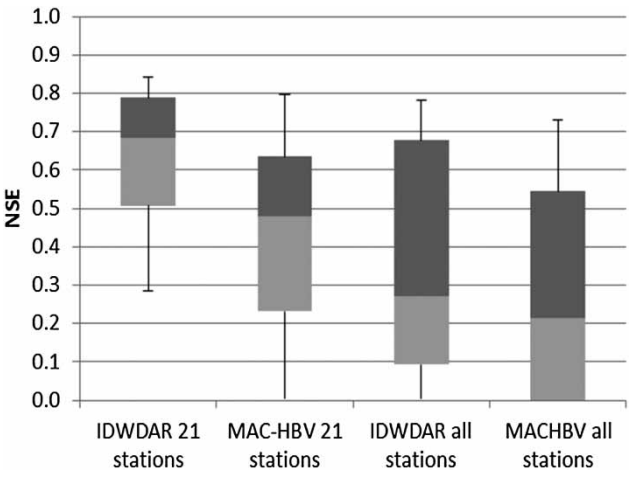

Figure 4 | Leave one out cross-validation results.

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 the 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 the deficit category, respectively. In some sub-basins (e.g., Dozois, Lievre, Rouge; see Table 2) more than 80% of basin area is in the deficit category. In general, for eight out of the 12 sub-basins, about 50% of basin area is in the 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 the Appendix (available

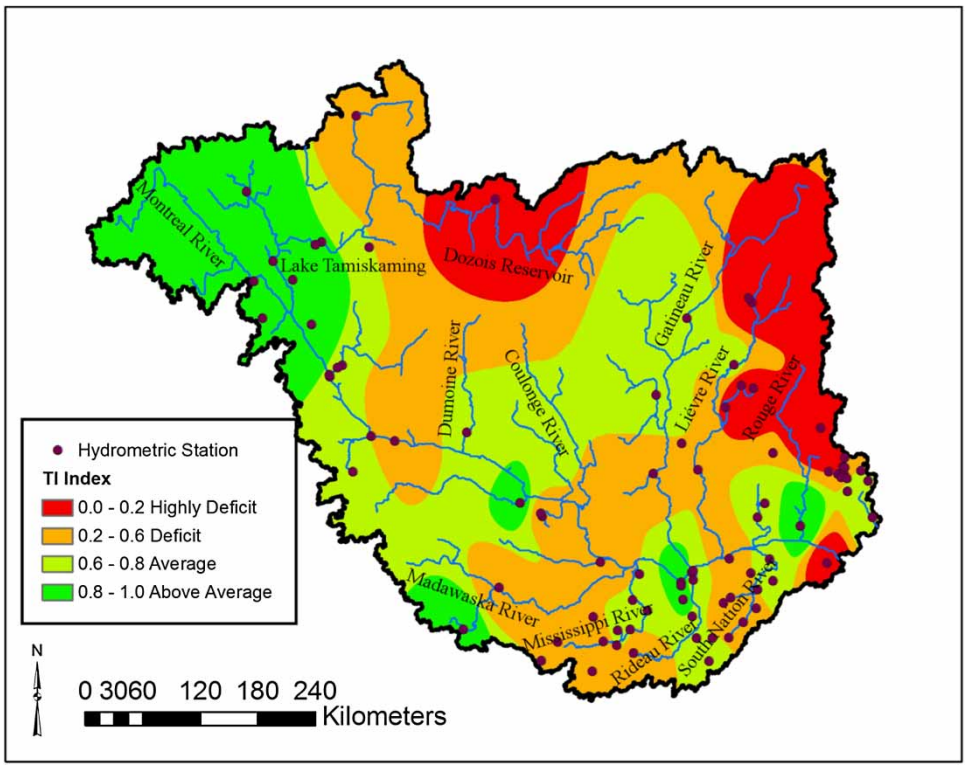

Figure 5 | Map of TI index values of existing stations.

Table 2 | Critical areas by major sub-basin

			Ratio of areas under different categories using TI index			
Prov	**Basin**	**Area (km^2)**	**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

with the online version of this paper). 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 stations that did not have data dating back to the year 2001 were excluded from the analysis.

RESULTS AND DISCUSSION

The results of the DEMO analysis using the IDW-DAR method are shown in Figures 6 and 7. Figure 6(a) 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 6(b)–6(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.

In order to analyse 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 Figure 7 indicate the areas with high probability of receiving new stations based on all the Pareto front solutions. This means that for almost all the

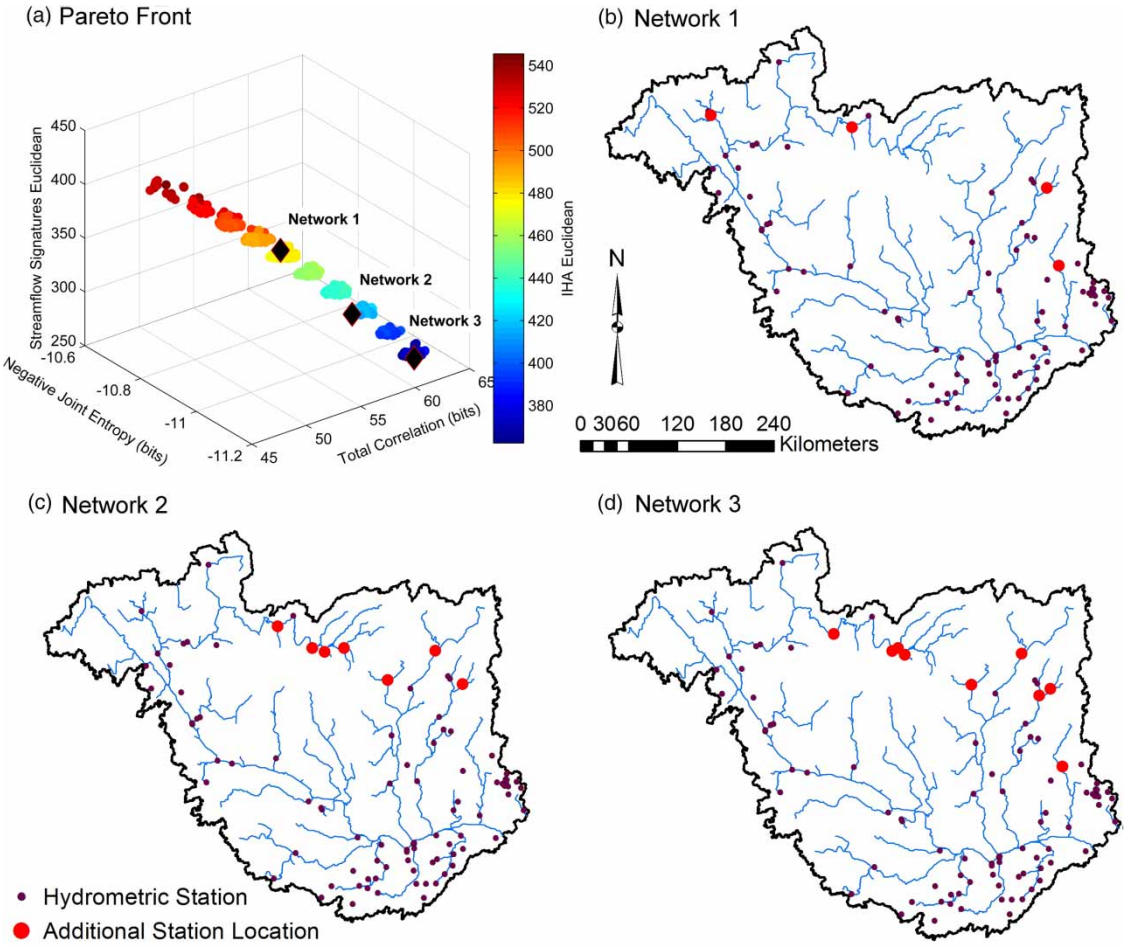

Figure 6 | Examples of Pareto optimal networks.

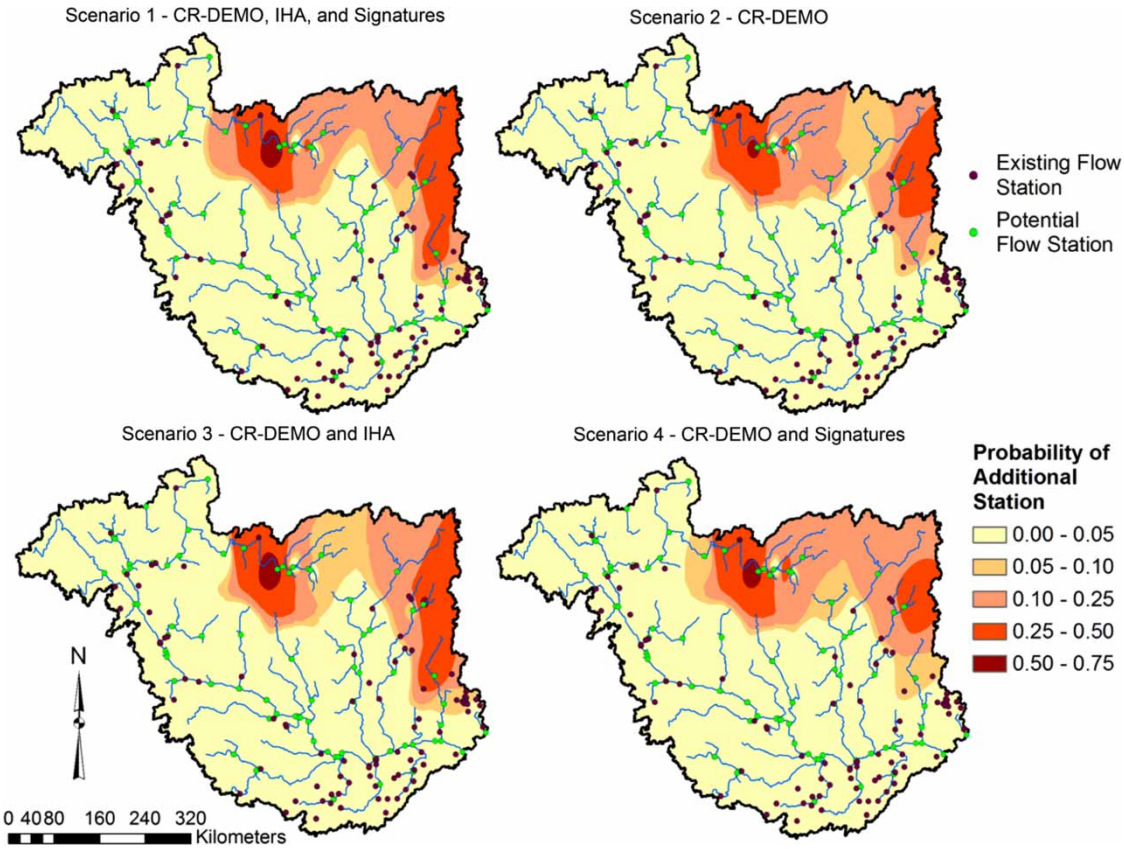

Figure 7 | IDW-DAR regionalization DEMO outputs.

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 7(d), 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.

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 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 eight 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 advantages of this information theory-based optimization method include that: (i) it provides a quantitative measure of the overall network efficiency and the value of each individual station with the TI value; (ii) it also provides objectively optimal locations in order to maximize the information content captured by the network; and (iii) it offers a set of optimal solutions (or networks) to users to select from based on their specific needs and priorities. 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 sensitivity of DEMO to the choice of potential stations or 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. The authors acknowledge two anonymous reviewers for their comments that helped to improve the manuscript.

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), W03528. 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.

Asghari, K. & Nasseri, M. 2014 Spatial rainfall prediction using optimal features selection approaches. *Hydrology Research* **46** (3), 343–355. http://doi.org/10.2166/nh.2014.178.

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

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

Chang, C., Wu, S., Hsu, C., Shen, J. & Lien, H. 2016 An evaluation framework for identifying the optimal raingauge network based on spatiotemporal variation in quantitative precipitation estimation. *Hydrology Research*. http://doi.org/10.2166/nh.2016.169 (in press).

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. PhD Thesis, University of British Columbia, Canada.

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

Keum, J. & Coulibaly, P. Sensitivity of entropy method to time series length in hydrometric network design. *Journal of Hydrologic Engineering* doi:10.1061/(ASCE)HE.1943-5584.0001508 (in press).

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* **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. *Journal of Hydrology* **19** (8), 05014003.

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.

Pham, H. V. & Tsai, F. T.-C. 2016 Optimal observation network design for conceptual model discrimination and uncertainty reduction. *Water Resources Research* **52**, 1245–1264. http://doi.org/10.1002/2015WR017474.

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. & Metcalfe, R. 2011 Estimation of continuous streamflow in Ontario ungauged basins: comparison of regionalization methods. *Journal of Hydrology* **16** (5), 447–459.

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.

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.

Shafiei, M., Ghahraman, B., Saghafian, B., Pande, S., Gharari, S. & Davary, K. 2014 Assessment of rain-gauge networks using a probabilistic GIS based approach. *Hydrology Research* **45** (4), 551–562. http://doi.org/10.2166/nh.2013.042.

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*, 6th edn. WMO-NO, 168. World Meteorological Organization Commission for Hydrology, Geneva.

Wu, S., Lien, H.-C., Hsu, C.-T. & Shen, J.-C. 2015 Modeling probabilistic radar rainfall estimation at ungauged locations based on spatiotemporal errors which correspond to gauged data. *Hydrology Research* **46** (1), 39–59. http://doi.org/10.2166/nh.2013.197.

Xu, H., Xu, C., Chen, H., Zhang, Z. & Li, L. 2013 Assessing the influence of rain gauge density and distribution on hydrological model performance in a humid region of China. *Journal of Hydrology* **505**, 1–12. http://doi.org/10.1016/j.jhydrol.2013.09.004.

Xu, H., Xu, C., Roar, N., Xu, Y., Zhou, B. & Chen, H. 2015 Entropy theory based multi-criteria resampling of rain gauge networks for hydrological modelling – a case study of humid area in southern China. *Journal of Hydrology* **525**, 138–151. http://doi.org/10.1016/j.jhydrol.2015.03.034.

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.

First received 13 May 2016; accepted in revised form 11 September 2016. Available online 5 December 2016