New Brunswick Hydrometric Network Analysis and Rationalization

Jasmin Boisvert[1]

Nassir El-Jabi[1*]

Salah-Eddine El Adlouni[2]

Daniel Caissie[3]

Alida Nadège Thiombiano[4]

1. Department of Civil Engineering, Université de Moncton, Moncton, N.B., Canada.
2. Department of Mathematics and Statistics, Université de Moncton, Moncton, N.B., Canada.
3. Fisheries and Oceans Canada, Moncton, N.B., Canada.
4. Centre Eau–Terre-Environnement, Institut National de la Recherche Scientifique, Quebec Canada

*Corresponding author: Nassir El-Jabi, Department of Civil Engineering, Université de Moncton, Moncton, N.B., E1A 3E9, Canada,

Telephone: (506) 858-4296; Fax: (506) 858-4082; E-mail: Nassir.el-jabi@umoncton.ca

Word Count : 8397

Abstract

The availability of hydrometric data, as well as its spatial distribution, is important for water resources management. An overly dense network or an under developed network can cause inaccurate hydrological regional estimates. This study's objective is to propose a methodology for rationalizing a network, specifically the New Brunswick Hydrometric Network. A hierarchical clustering analysis allowed dividing the province into two regions (North and South), based on latitude and high flow timing. These groups were subsequently split separately into three homogeneous subgroups, based on the generalized extreme value (GEV) distribution shape parameter of each station for annual maximum flow series. An entropy method was then applied to compute the amount of information shared between stations, ranking each station's importance. A station with a lot of shared information is redundant (less important), whereas one with little shared information is unique (very important). The entropy method appears to be a useful decisional tool in a network rationalization.

Keywords : hydrometric network; GEV; clustering analysis; entropy ranking; New Brunswick

Résumé

La disponibilité des données hydrométriques ainsi que la distribution spatiale des stations hydrométriques sont d'une grande importance pour la gestion des ressources en eau. La couverture spatiale est souvent très faible, ce qui peut causer des simulations hydrologiques inexactes. L'objectif de cette étude est de proposer une méthodologie pour la rationalisation du réseau hydrométrique du Nouveau-Brunswick. L'approche proposée combine une méthode de classification basée sur le comportement des extrêmes hydrologiques et une mesure de

l'information conjointe produite par l'ensemble des stations disponibles dans le réseau. Une approche de classification hiérarchique a permis de diviser la province en deux secteurs dits homogènes (Nord et Sud) en fonction de la latitude et de l'occurrence des débits maxima annuels. Chacun de ces groupes a été divisé en trois sous-groupes homogènes, selon la valeur du paramètre de forme de la distribution GEV des débits maxima annuels de chaque station. Une méthode basée sur l'entropie a permis le classement des stations en fonction de leur importance dans leur groupe respectif (Nord ou Sud), en calculant la quantité d'information conjointe entre les stations. Ainsi, une station qui comporte beaucoup d'information commune avec d'autres stations est considérée redondante, et donc moins importante. Une station avec très peu d'information partagée est considérée unique, et donc très importante. Le classement des stations par ordre d'importance peut être un outil décisionnel utile. Les stations ont été ordonnées selon leur importance, et la méthode d'entropie se présente comme un outil décisionnel utile dans la rationalisation du réseau.

1. Introduction

The importance of hydrometric gauging station networks for surface water monitoring is well established, given the usefulness of collected hydrometric data for decision making related to water resources management around the world (Hannah et al. 2011). However, the density of these networks is still being impacted by the shift of social and economic priorities of governments, like that observed in Canada (Burn 1997; Coulibaly et al. 2013; Mishra and Coulibaly 2009). In fact, Pilon et al. (1996) showed that, through the 1990s, data collection from Canadian National Hydrometric Network (CNHN) declined mainly due to financial pressure that impacted the budget of relevant agencies. More recently, Coulibaly et al. (2013) noticed that only 12% of the Canadian terrestrial area, the majority of which is in the southern portion of the country, is covered by hydrometric networks that meet the minimum standards according to the World Meteorological Organization (WMO) physiographic guidelines. Moreover, 49% of the Canadian terrestrial area is gauged by a sparse network and the remaining 39% is ungauged (Coulibaly et al. 2013). Although the negative implications of this may not be immediately apparent, many water resource decisions, project designs and project management rely on information gained by hydrometric gauging stations. In other words, short-comings in a gauging network can lead to greater hydrological uncertainty, which can lead to inefficient project design and resource management, which in turn can have diverse consequences. For example, uncertainty can lead to over-designing, which adds unnecessary extra project costs. In addition, under-designing is also a possibility, which could lead to project failure and extra costs as well. Poor resource management can also impact the population as well as the environment. Although reducing the amount of gauging stations within a network is not ideal according to WMO guidelines, financial and budget restraints may make it necessary. Therefore it seems an evaluation of the network must be undertaken in order to properly analyze options for station reduction or displacement to minimize information loss, thus rationalizing the network. The required assessment must define and integrate appropriate criteria for each region for the network to be properly updated. It is in this context that the present study aims to propose a rationalization of the hydrometric gauging network of New Brunswick (NB). This will be accomplished using a hierarchical clustering analysis and the generalized extreme value (GEV) distribution shape parameter analysis as preliminary evaluation tools of hydroclimatic behaviour and homogeny between gauging stations, and subsequently, with the entropy concept to quantify the importance of each station regarding information content. However, in order to have a more complete rationalization process, data managers and users should be consulted for their input, as other criteria (e.g., quality of rating curves, size of the drainage basin, etc.) may also be of importance in a final decision.

Mishra and Coulibaly (2009) provided a review of common methodologies developed to address hydrometric network design or redesign in response to growing management and financial challenges for governments and data users. Using the entropy concept, Mishra and Coulibaly (2010) provided an evaluation of hydrometric network density and the worth of each station, in major watersheds across Ontario, Quebec, Alberta, New Brunswick and Northwest Territories. Their study highlighted the generally deficient status of hydrometric networks, mainly over the northern part of Ontario and Alberta, as well as in the Northwestern regions of Canada. The entropy concept, derived from Shannon information theory (Shannon 1948), assesses the information content of each gauging station of a given network in relation to all other stations of that network. It was adapted to suit hydrological concerns by Hussain (1987; 1989). Its applications showed its usefulness for optimal hydrometric network design in many studies (e.g., Alfonso et al. 2013; Li et al. 2012; Mishra and Coulibaly 2010; Singh 1997; Yeh et al. 2011). Nevertheless, multivariate analysis methods such as clustering analysis also remain useful statistical tools in the hydrometric network rationalization process. These methods are commonly used to identify homogeny in a dataset, and potentially form groups of similar individuals (in this case hydrometric gauging stations), which is an important step for network rationalization and optimization (Daigle et al. 2011; Khalil and Ouarda 2009). For example, Khalil et al. (2011) used clustering analysis to extract different sub-hydrological units in order to better perform their assessment and redesign of the water quality monitoring network. In hydrology, three parameter distributions are recommended, especially the Generalized Extreme Value (GEV) distribution. Indeed, its statistical properties indicate how the flexibility of this three parameter class of distributions can capture skew and fat tails (El Adlouni et al., 2010).

In studies characterizing natural flow regimes, environmental flows and floods in New Brunswick (Aucoin et al. 2011; El-Jabi et al. 2015), it was found that the GEV distribution was an appropriate distribution to model the annual maximum and minimum flows at most of the gauging stations in New Brunswick. For instance, the Anderson-Darling test showed slightly better performances for the GEV than for the 3 parameter lognormal distribution. Unlike the normal distribution that arises from the use of the central limit theorem on sample averages, the extreme value distribution arises from the theorem of Fisher and Tippet (1928) on extreme values or maxima in sample data. The class of GEV distributions is very flexible and its shape parameter controls the size of the tails corresponding to three special cases (Gumbel, Fréchet and Weibull). Therefore, the GEV shape parameter (kappa) is a good indicator of the distribution of the extreme high and low flow events, and thus could be useful in differentiating between different hydrological regimes. Consequentially, it seems that the shape parameter fitted to the annual maximum flow series (GEVkapMax) and the shape parameter fitted to the annual minimum flow series (GEVkapMin) are good for characterising flows in the province. Since the annual maximum flows were particularly well modeled by the GEV distribution, and the maximum flows are generally of more interest, the GEVkapMax was deemed as an appropriate metric to analyze and potentially be used for identifying homogenous groups of datasets.

Network optimization cannot be accomplished by solely using these purely statistical approaches, as other factors must be taken into account. For example, a gauging station linked to a hydroelectric facility may not be statistically important in a network, but would most likely be important from a resources management perspective. Data user needs and perception must be integrated in any analysis of a network. It has been recommended and integrated in previous studies (e.g., Burn 1997; Coulibaly et al. 2013; Davar and Brimley 1990). Environment Canada and New Brunswick Department of Municipal Affairs and Environment (1988) investigated accuracy requirements identified by users in order to define a minimum and target networks. They considered mean, low and high flows in this approach, which consisted of developing regional equations for each of the three categories. They initially identified 16 homogenous regions in the province, considering that there should be a small, medium, and large gauged basin in each homogenous area. This implied that 48 stations, plus an additional 6 for larger regions (total of 54 stations), was identified as a minimum network. They also identified a target network, this time considering that 10 stations were necessary per region in order to properly define regional regression equations. However, they also refined the initial 16 homogenous regions into 7 regions. This implied that 70 stations (plus an additional 7 for variations in size) were suggested as the target (total of 77 stations). They concluded that it was important to coordinate hydrographic gauging with meteorological gauging, that more gauging was necessary for smaller catchments, and that the central part of the province lacked gauging stations. They also evaluated the hydrometric network using an audit approach, through which a ranked prioritization of stations was provided based on the hydrometric, socio-economic and environmental worth of each station according to data user perceptions. They also considered site characteristics, economic activity, federal and provincial commitments, special needs, as well as a station's regional and operational users in their audit approach. Davar and Brimley (1990) used a similar approach to identifying a minimum and target network as Environment Canada and New Brunswick Department of Municipal Affairs and Environment (1988), but their audit was slightly different. The existing stations and proposed new stations were evaluated using an audit approach, based on site characteristics, client needs (regional hydrology and operational), and regional water resource importance. They created different scenarios that had different impacts and values (based on audit points) in function of different costs (adding, removing, or maintaining the amount of gauging stations in the network). Overall, their recommendations included: reallocating resources to meet the minimum network; create a committee for ongoing planning and analysis, as well as communication with the user community; emphasize the importance of regional hydrology; coordinate with other related data gathering, such as water quality and atmospheric data. The size of the New Brunswick Hydrometric Network has been subject to change over the years. The network's major expansion occurred in the 1960's, with a peak size of 75 stations achieved in 1978-1979. The network maintained a size of between 69 and 72 stations from 1980-1993. In 1994, the number of stations was reduced to 56, a reduction of 22% in the number of stations. By the year 2000, only 46 stations were left active in New Brunswick a further reduction of 14% (i.e., a reduction of 36% compared to the 1980-1993 period). This was contradictory to the studies done by Environment Canada and New Brunswick

Department of Municipal Affairs and Environment (1988) and Davar and Brimley (1990), which advocated an increase in network size. However, financial pressure required the reduction in gauging station numbers.

Due to the fact that network expansion and reduction was mostly done considering site specific needs, as opposed to the network as a whole, rationalization of the network is still relevant in New Brunswick. This study aims to provide a methodology for such a rationalization, and in the case of further budget reductions, a supporting tool for management and decision making.

2. Methodology

It should be noted that for all the methods used in this study, the specific discharge (discharge per unit area; m^3/s per km^2) was used. This was done to remove drainage area as an overwhelmingly dominant variable when it comes to explaining flow rates. In other words, the drainage area was removed as a variable in order to better compare larger basins with smaller basins in terms of flow.

2.1 Hierarchical clustering analysis

The first approach used in this study was the hierarchical clustering analysis. The objective of this analysis was to divide the province into similar hydrological groups. This was done using a clustering analysis. Thereafter, the GEV analysis further refined these hydrological groupings, and the entropy analysis was carried out in this framework for finer assessment. Rationalization and optimization assessment of the network has been shown to be better conducted with the division of a network into climatic regions (Burn and Goulter 1991; Khalil et al. 2011).

The attributes from which similarities will be defined need to be specified for clustering analysis (Burn and Goulter 1991). Once this is done, clusters are formed by grouping similar observations together in such a way that variance is minimized within a cluster and maximized between clusters (Khalil and Ouarda 2009). The division of the complete network into clusters was done using hierarchical agglomerative clustering (based on Euclidean distance), accomplished using R software toolbox (R Core Team 2015). In this type of clustering, each individual station is initially considered as being its own cluster. Afterwards, an iterative process is used in which only the two most similar clusters (least Euclidean distance between two clusters of all possible combinations) are joined together to form one new cluster per iteration. This is repeated until a single cluster remains, containing all the individuals. In this study, two attributes were used for the clustering analysis: latitude of each station, and high flow timing. The latter was computed as the 30-day period with the highest mean flow (moving average). The two attributes (latitude and timing) were chosen with the purpose of dividing the province based on climate. The high flow timing is typically dependent on temperature, due to snowmelt. The northern part of the province is typically cooler than the southern part. As such, using latitude and high flow timing, it is expected that the province will be divided into clusters based on both geographical location and hydro-climatological processes.

2.2 GEV shape parameter analysis

Following the clustering analysis, each group will be characterized by the GEV shape parameter, fitted to the maximum annual data, of each station. The objective of this analysis is to further subdivide the climatic regions into smaller homogenous groups of similar data. The GEV probability density function is given by Equation 1.

$$(1) \qquad f(x) = \frac{1}{\alpha}[1 - \frac{\kappa}{\alpha}(x-u)]^{\frac{1}{\kappa}-1} \exp\{-[1-\frac{\kappa}{\alpha}(x-u)]^{\frac{1}{\kappa}}\}$$

where x is an observation of the random variable in this case the specific discharge, κ is the shape parameter, α is the scale parameter, and u is a location parameter. In addition, the following restriction applies: $x < u + \alpha/\kappa$ if $\kappa > 0$; $x > u + \alpha/\kappa$ if $\kappa < 0$. The shape parameter, as suggested by its name, represents the shape of the right tail or the left tail of the distribution. This means that depending on the parameter, the distribution can be symmetrical ($\kappa = 0$), skewed with a heavy left tail ($\kappa > 0$), or skewed with a heavy right tail ($\kappa < 0$). The GEV shape parameter (kappa) has three statistically significant categories. These will be used to further subdivide the groups classified by the clustering analysis into smaller subgroups. The first category, where kappa is between]-0.33; +0.33[, has a finite mean, variance and coefficient of skewness. The second category, where kappa belongs to the interval]-0.5; -0.33] or [+0.33; +0.5[, has an infinite coefficient of skewness. The third category, where kappa is between]-∞; -0.5] or [0.5; ∞[, is for datasets with an infinite variance as well as an infinite skewness coefficient. It should be noted that a negative GEV shape parameter (kappa) value produces a positive skewness (heavy right side of the distribution), which is most common for hydrological maxima.

2.3 Entropy analysis

Once the hydrological similarity assessment based on the existing hydrometric network is carried out using the clustering and GEV shape parameter analyses, the entropy method is used to evaluate the worth of each station in the network. The objective of the entropy concept analysis is to quantify the information contained in the random variable (specific discharge) measured at the different gauging stations. This is important since it provides an objective criterion to describe each station. However, it is actually the measure of trans-information that is of particular interest in this study. The measure of trans-information, a function of marginal entropy and joint entropy, indicates if the same information is measured by multiple stations (redundancy), or if the information measured by a station is unique (optimal). This gives an idea of the relative importance of each station, given the principles of information maximization (Hussain 1987; 1989; Singh 1997; Mishra and Coulibaly 2010). This allows for better decision making when it comes to choosing if a station should be removed, displaced, or continued. For example, a station that provides similar information to the network as other stations is highly redundant and can be removed without significant loss of information. In contrast, a station whose information is unique is highly valuable to the network, and should not be removed. It should be noted that a limitation of this method is the fact that the data from each stations has to be in the same time period (of at least 20 years), and the whole period must be covered. Therefore, the time period where the greatest amount of stations has concurrent measurements/data is required.

A station malfunctioning for a few days or even months is not uncommon in a network. Therefore, it is important before proceeding to the entropy calculations to deal with missing data. To complete the data time series, a correlation matrix between stations with missing values and stations without missing data can be constructed (Mishra and Coulibaly 2010), using a linear regression analysis (Ouarda et al. 1996).

The trans-information (or mutual information) $T\left(X,Y\right)$ which is of interest, is described in Equation 2 as the information about a predicted variable transferred by the knowledge of a predictor (Mishra and Coulibaly 2010) as follows:

$$T(X,Y)=H(X)+H(Y)-H(X,Y) \tag{2}$$

where $T(X,Y)$ is the trans-information; $H(X)$ and $H(Y)$ are the discrete form of entropy of the continuous random variables X and Y. $H(X)$ was formulated by Shannon (1948) and later updated by Hussain (1987; 1989) for use with hydrological time series data and given by:

$$H(X)=-\sum_{k=1}^{K}p(x_k)\log\left[p(x_k)\right] \tag{3}$$

$H(Y)$ is given by the same equation as $H(X)$, but substituting *k* for *l*. This information coefficient only gives a measure of information from the concerned random variable; hence the importance of joint entropy between the interested variables (flow time series), as described by Equation 3 as $H(X,Y)$ for the bivariate case. This allows the measurement of the overall information retained by random variables (Li et al. 2012). The logical extension can be made for the multivariate case.

$$H(X,Y)=-\sum_{k=1}^{K}\sum_{l=1}^{L}p(x_k,\mathrm{y}_l)\log\left[p(x_k,\mathrm{y}_l)\right] \tag{4}$$

In the above equations, x_k is an outcome corresponding to k; $p(x_k)$ is the probability of x_k and is based on the empirical frequency of the variable *X*; y_1 is an outcome corresponding to *l*; $p(\mathrm{y}_1)$ is the probability of y_1 and is based on the empirical frequency of the variable *Y*; $p(x_k,\mathrm{y}_1)$ is the joint probability of an outcome corresponding to k for X and *l* for Y. K and L are the finite number of class intervals (as divided by the points x_k and y_1) for the corresponding variables with the general assumption that $K = L$*;* In the case where the entropy concept is being applied to a hydrometric gauged network, the variable X becomes $Z(i)$; the actual quantity of information contained at station i. The variable Y becomes $\hat{Z}$ the quantity of information at station *i*, but this time derived from the linear regression demonstrated in Equation 5.

$$\hat{Z} = a(i) + b(i) * G(i) \tag{5}$$

In this equation, $G(i)$ is a matrix of data from all other stations, $a(i)$ and $b(i)$ are the parameters of the regression between station i and all other stations, assuming a linear relation between stations is deemed appropriate. The trans-information becomes T(Z, $\hat{Z}$) (Burn 1997; Mishra and Coulibaly 2010). The data used for all these computations is the annual series of maximum monthly specific discharge.

Since the entropy analysis is performed over a 20 year window, each station has a data series of 20 points, each one representing the average specific discharge for the month with the highest average specific discharge of that year.

Once the trans-information has been evaluated for each station, it can be used to rank station in order of importance (Li et al. 2012; Yeh et al. 2011). Stations with smaller trans-information values are the most important stations, since they contain little redundant information, and thus get ranked the highest (1 being the most important).

3. Case Study: New Brunswick Hydrometric Network

The hydrometric gauging station network being analyzed by this study is the New Brunswick Hydrometric Network (NBHN). There are also a few gauging stations located in Québec and in Maine (U.S.) that can be considered relevant to New Brunswick, since the watersheds of some rivers located in New Brunswick are partially located outside the province. The current network, as identified by Environment Canada, contains 67 stations. Of these 67 stations, 46 are active and 21 are discontinued.

The first measurements taken in the province were in 1918. The major expansion of the network occurred in the late 1960's, continuing in the early 1970's. This was caused by an increased demand for data for water supply, fisheries, and flood forecasting (Davar et al. 1990). Many stations were originally established to suit specific needs, often short-term. After their objectives were completed, these stations were kept in service. This method of network expansion was considered acceptable at the time (Davar et al. 1990). Although this method did in fact create an expanded network, it is not necessarily the most effective method. Since new stations were added in locations for a specific purpose (i.e. a single project), little consideration was given to the network as a whole. This implies that new stations may have been placed in similar locations to existing stations, causing redundancy in the information measured. An objective of the analysis and optimization of the network carried out by this study is to identify this redundancy in information.

4. Results and Discussion

4.1 Hierarchical clustering analysis

Two clusters were formed in the hierarchical clustering analysis, based on high flow timing. A dendrogram was obtained using the hierarchical clustering technique (Figure 1) where station number identifies each site. The two major groups formed by the clustering analysis can be seen on this figure (identified in red). All 67 stations identified by Environment Canada were used in this analysis.

Each horizontal bar connecting two stations (or groups) corresponds to the maximum difference in timing of the stations within the two connected groups. For example, the stations 01AP004 and 01BU002 (3rd and 4th from the top), are connected by a horizontal line positioned at a value close to 0, implying they have very similar high flow timing and latitude. Furthermore, station 01AR006 is connected to the previously mentioned group of two stations by a line positioned at a value close to 1, indicating a difference in Euclidean distance (timing and latitude) between 01AR006 and the other two stations of close to 1, which is also a small distance. It should be noted that the method used for clustering was the complete linkage method. This means that the distance between clusters was calculated as the maximum possible Euclidean distance between a pair of stations, one from each cluster. This is important when selecting which two clusters to join together in an iteration, since other methods could use criterion such as the minimum distance (single linkage), the average distance (mean linkage), or other criterion, possibly yielding different results. The complete linkage method does not perform as well when there are many outliers in the population being analyzed. Since all of the stations are in the same geographic area, there should be few outliers in terms of latitude and high flow timing among the stations analyzed. Therefore, the complete linkage method was deemed appropriate for this study.

Since the groups are mostly positioned in a north-south fashion, the two groups were named North Group (NG) and South Group (SG). These results are consistent with previous studies where stations were divided in a north and south group, when dealing with mean annual flow regimes, e.g., Environment Canada and New Brunswick Department of Municipal Affairs and

Environment 1988. As such, in the present study the North Group (NG) and South Group (SG) will be analyzed separately in the analyses that follow, the GEV shape parameter and entropy analyses. It should be noted that the results of the clustering analysis were slightly modified for the final classification into the two groups (NG and SG). Notably, stations 01BV007, 01BU004, 01AL003, and 01AL002 had flow timings similar to the North Group, despite being more southern stations (Figure 2). These stations were analysed part of the South Group, as they were a significant distance from the north, and typically surrounded by southern stations. Similar reasoning was applied to station 01BO003, which was clustered in the south, but located in the north (subsequently analyzed as part of the North Group). Stations 01AG002 and 01AG003 could have easily been part of the North or South Group, as they are very close to the perceived divisional north-south line (see Figure 2); however, they were identified part the of South Group in the analysis and kept within this group. Of the 67 stations used for the clustering, 31 were placed in NG and 36 in SG. Table 1 contains the results of the clustering analysis (NG or SG) as well as the results of the GEV shape parameter analysis (see section 4.2 below).

4.2 GEV shape parameter analysis

Applying the GEV shape parameter analysis allowed dividing the North Group and South Group each into three respective subgroups. The first subgroups (NG1 and SG1), have kappa values between]-0.33; +0.33[. The second category (NG2 and SG2), have kappa values between]-0.5; -0.33] or [+0.33; +0.5[. The third category (NG3 and SG3), have kappa values between]-∞; -0.5] or [0.5; ∞[. Table 1 lists the six groups and the stations within each group. It should be noted that of the 67 stations used in the clustering analysis, 01AD004 (NG) and 01BV007 (SG) were removed from the GEV shape parameter analysis, given the poor quality of data (short record length and interpolated data). Table 1 show that most stations behaved accordingly to the category 1 (i.e., a finite mean, variance and coefficient of skewness) followed by category 2 (i.e., an infinite coefficient of skewness). This is interesting because the values of the GEV shape parameter in these both categories are the most probable in hydrology and moreover they allow avoiding unfeasible estimations (Martins and Stedinger 2000).The category 3 formed the least amount of stations (i.e., an infinite variance as well as an infinite skewness coefficient) with only 5 stations in the North Group and 2 stations in the South Group.

4.3 Entropy analysis

The annual maximum specific discharge was used for the entropy analysis. The window chosen for the analysis was 1976-1995. This is the period of time with the maximum amount of data among stations, i.e., at least 20 years of record, no significant gaps in data and a concurrent time period. Of the 65 stations used for the GEV shape parameter analysis, only 53 (23 in NG, 30 in SG) respected the above conditions. As such, these remaining 53 stations were used in the entropy analysis. For the stations with acceptable gaps in data (up to 25% missing data accepted), the individual station with complete data that showed the maximum correlation with a station having missing data was used to fill the data. This was done for 8 stations in the North group, filling in anywhere from 1 to 5 years of data (average of 3 years). This was also done for 7 stations in the South group, all of which were for 2 years.

The results of the entropy computation are presented in Tables 2a and 2b for the North Group and South Group, respectively. They are constituted by the marginal entropies H(Z) and H(G), the joint entropy H(Z,G), the trans-information T(Z, $\hat{Z}$) and the rank R values. It is important to remind that Z and G are respectively the quantity of information at individual station and that from the matrix of all others stations excluded the one of interest. Additionally, $\hat{Z}$ is the quantity of information resulting from the linear regression between Z and G. The rank of the stations is simply the order associated to the sorted values of T(Z, $\hat{Z}$), so that to the lowest values correspond the smallest rank which are equivalent to the most important stations (Mishra and Coulibaly 2010). It is important to note that stations 01BL001, 01AK001, and 01AP002 are considered to be the most important stations, given that their values of H(G), H(Z,G) and T(Z, $\hat{Z}$) are zero (Table 2a and 2b). This implies that the information measured by these stations is unique, and consequently very important.

Table 3 show the ranking of the stations divided into their respective groups based on the GEV parameter. It is important to remember that removing the majority or entirety of a group is not advisable, since each group has some statistical importance. It would be preferable to remove a few of the least important stations per group (especially within a large group), as opposed to several from the same group, even if the stations from a single group are ranked lower by the entropy analysis. Figure 2 shows the positions of all the stations of the network and their ranks (in bracket), including information on if the station is current in operation or discontinued.

4.4 Stations excluded from the entropy analysis

Of the 67 stations initially identified as being part of the New Brunswick network of hydrometric gauging stations, only 53 were analyzed by the entropy method. The remaining 14 stations must also be dealt with by other means. These stations are listed in Table 4. Many of the stations listed in Table 4 are already inactive (discontinued). No reasoning or analysis will be applied to these stations, since it is assumed that they will not likely be reactivated. This leaves six stations that need further consideration. Stations 01AD004 and 01BV004 have long record lengths (46 and 52 years respectively) and therefore should most likely remain part of the network, since such a long record length is not common in the province. Station 01AF009 is part of NG3, which is a small group, and is the only member of this group in the northwest of New Brunswick. It may be wise to keep 01AF009, particularly if other stations of this group are already being removed. Station 01BJ012, represents a small drainage basin within the North Group, has a reasonably long record length (29 years), but is located near 01BJ003, 01BJ004 (inactive) and 01BJ007. Therefore one of these stations could potentially be removed (note that 01BJ007 has a lower ranking than 01BJ003 within the NG1; Table 3). Station 01BP002 has a small drainage area (28.7 km^2) and a reasonably long record length (24 years). It is also close to the center of the province; where there seems to be a lack of gauging stations (Figure 2). This station could be either removed or kept from the network depending on the importance of this station in terms of location, size and length of records. Whether or not nearby stations are being removed should also be considered before deciding if 01BP002 should be kept or removed. Very similar reasoning and conclusions can be applied to station 01BU009.

4.4 Other considerations

It is also important to take into account information about each station's worth using, for example, expert knowledge in order to make advised choices of an optimal network design (Hannah et al. 2011). For example, a statistically insignificant station according to the entropy analysis could in fact be very important because of its use in conjunction with a hydroelectric dam or a water supply. Similar elements to this example can be helpful through consultations with data users and managers, in order to properly design a rationalized hydrometric network for New Brunswick. A brief analysis was carried out on the groupings to determine if there were patterns regarding drainage area, specific discharge, and record length of each station to see if, for example, the majority of smaller basins or larger basins were contained in a single grouping.

No such patterns seemed to exist, and it seems that the groups each contain a broad range of drainage areas, specific discharges, and record lengths. A more detailed analysis could be undertaken for the sake of completeness.

Consideration should also be given to reactivating some of the more important stations that have already been discontinued. This can be accomplished by removing a higher quantity of less important stations than what is necessary, allowing some of those stations to be removed or displaced to a better location, especially if the new station would contribute to a better network and a better spatial coverage. As such, it would be recommended when choosing which stations to removed or displaced that a separate evaluation be done using existing regional regression equations. In fact, the question becomes, does the removal of a particular station of group of station significantly impact the regional hydrological equations (mean annual flow, high and low flows)? An analysis of these regression equations should be done to see how they would be affected if a few selected stations were to be removed from the computation. This can give additional insight as to whether or not a station should be removed or kept.

5. Conclusions

Water management requires an optimal hydrometric network, as shown by the growing interest for hydrometric network evaluation and rationalization, in order to address challenges ahead in monitoring and data collection network stations. The present study provides a contribution to support decision makers, like data users and monitoring networks managers, in the process of selecting optimal representative stations for New Brunswick hydrometric network. The proposed methodology is flexible and can be applied to other case studies. The present study proceeded by first dividing New Brunswick into two groups, using clustering analysis based on high flow timing and latitude. This had the effect of creating a north-south division. However, this division was not a perfect divide of north and south stations, where some northern stations had high flow timings similar to southern stations, and vice-versa. The GEV shape parameter (maximum annual flow series) was then used to split each group into three sub-groups based on specific characteristics of the distribution (e.g., tail). The purpose of these divisions was to avoid suggesting the complete or majority removal of stations from a single homogenous group, since removing a few stations of each group would be preferable. Finally, an entropy analysis was done to quantify the amount of information that was redundant at each station, thereby quantifying the importance of each station, based on its measurement of unique information. This allowed the ranking of stations in order of importance, which in turn allowed the prioritization of stations. This prioritization can thereafter be used to determine the removal or displacement of stations that would allow for a more optimal network. Some reasoning and analysis was done regarding the stations that did not meet the criteria for entropy analysis to better judge whether or not they were important.

6. Recommendations

The present study showed difference among stations within each group and subgroup. It is not recommended to remove the majority or entirety of stations within a subgroup. This is particularly the case for NG3 and SG3 as they are the subgroups with the least amount of stations. It is instead preferable to remove some stations from each subgroup, as opposed to many from one subgroup.

Reactivating some of the more important stations that have been deactivated should be considered. These stations contributed unique information to the network, and so would be useful to have active. An analysis of regression equations should also be undertaken as an additional insight to how the network would react to certain stations being removed.

Acknowledgements

This study was funded by the New Brunswick Environmental Trust Fund. The authors remain thankful to Mr. Darryl Pupek and Dr. Don Fox from the New Brunswick Department of Environment and Local Government for their support.

References

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

Aucoin, F., Caissie, D., El-Jabi, N., and Turkkan, N. 2011. Flood frequency analyses for New Brunswick rivers. Can. Tech. Rep. Fish. Aquat. Sci. 2920: xi + 77p.

Burn, D.H. 1997. Hydrological information for sustainable development. Hydrological Sciences Journal **42**(4): 481-492. doi: 10.1080/02626669709492048

Burn, D.H., and Goulter, I.C. 1991. An approach to the rationalization of streamflow data collection networks. Journal of Hydrology **122**: 71-91.

Coulibaly, P., Samuel, J., Pietroniro, A., and Harvey, D. 2013. Evaluation of Canadian national hydrometric network density based on WMO 2008 standards. Canadian Water Resources Journal **38**(2): 159-167. doi: 10.1080/07011784.2013.787181

Daigle, A., St-Hilaire, A., Beveridge, D., Caissie, D., and Benyahya, L. 2011. Multivariate analysis of the low-flow regimes in eastern Canadian rivers. Hydrological Sciences Journal **56**(1): 51-67. doi: 10.1080/02626667.2010.535002

Davar, Z.K., and Brimley, W.A. 1990. Hydrometric network evaluation: audit approach. Journal of Water Resources Planning and Management **116**(1): 134-146.

El Adlouni, S., Chebana, F., and Bobée, B. (2010). Generalized Extreme Value vs. Halphen System: An exploratory study. Journal of Hydrologic Engineering **15**(2): 79-89.

El-Jabi, N., Turkkan, N., and Caissie, D. 2015. Characterisation of Natural Flow Regimes and Environmental Flows Evaluation in New Brunswick. New Brunswick Environmental Trust Fund.

Environment Canada and New Brunswick Dept. of Municipal Affairs and Envir. 1988. New Brunswick hydrometric network evaluation. Prepared by the Water Resources Planning Branch of the Province of New Brunswick and Inland Waters Directorate, Dartmouth, Nova Scotia, Canada.

Fisher, R.A., and Tippett, L.H.C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society **24**(2): 180-190. doi:10.1017/S0305004100015681.

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Kerstin, S., and Tallaksen, L. M. 2011. Large-scale river flow archives: importance, current status and future needs. Hydrol. Process. **25**: 1191-1200.

Husain, T. 1989. Hydrologic uncertainty measure and network design. Water Resources Bulletin **25**(3): 527-534.

Hussain, T. 1987. Hydrologic network design formulation. Canadian Water Resources Journal **12**(1): 44-63. doi: 10.4296/cwrj1201044

Khalil, B., and Ouarda, T.B.M.J. 2009. Statistical approaches used to assess and redesign surface water-quality-monitoring networks. J.Environ.Monit **11**: 1915-1929.

Khalil, B., Ouarda, T.B.M.J., and St-Hilaire, A. 2011. A statistical approach for the assessment and redesign of the Nile Delta drainage system water-quality-monitoring locations. J. Environ. Monit. **13**: 2190-2205.

Li, C., Singh, V.P., and Mishra, A.K. 2012. Entropy theory-based criterion for hydrometric network evaluation and design: maximum information minimum redundancy. Water Resources Research **48**(5), W05521. doi: 10.1029/2011WR011251.

Martins, E.S., and Stedinger, J.R. (2000) Generalized maximum likelihood generalized extreme value quantile estimators for hydrological data. Water Resources Research **36**(3): 737-744. doi: 10.1029/1999WR900330

Mishra, A.K., and Coulibaly, P. 2009. Developments in hydrometric network design: a review. Reviews of Geophysics **47** (RG2001): 1-24. doi: 10.1029/2007RG000243.

Mishra, A.K., and Coulibaly, P. 2010. Hydrometric network evaluation for Canadian watersheds. Journal of hydrology **380**: 420-437.

Ouarda, T.B.M.J., Rasmussen, P.F., and Bobée, B. 1996. Rationalisation du réseau hydrométrique de la province de Québec pour le suivi des changements climatiques. INRS-Eau. **476**: 1-80.

Pilon, P.J., Day, T.J., Yuzyk, T.R., and Hale, R.A. 1996. Challenges facing surface water monitoring in Canada. Canadian Water Resources Journal **21**: 157-164. doi: 10.4296/cwrj210157

R Development Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available from http://www.R-project.org [accessed April 2016].

Shannon, C.E. 1948. A mathematical theory of communication. Bell Syst. Tech. J. **27**: 379-423, 623-656.

Singh, V.P. 1997. The use of entropy in hydrology and water resources. Hydrological Processes **11**: 587-626.

Yeh, H.C., Chen, Y.C., Wei, C., and Chen, R.H. 2011. Entropy and kriging approach to rainfall network design. Paddy Water Environ **9**:343-355. doi: 10.1007/sl0333-010-0247-x

Van Groenwoud, H.V. 1984. The climatic regions of New Brunswick : a multivariate analysis of meteorological data. Can. J. For. Res. **14**: 389-394.

Table 1. Division of the North and South Groups into subgroups based on the GEVkapMax parameter

Table 2a. Entropy values and ranking of each station (North Group)

Table 2b. Entropy values and ranking of each station (South Group).

Table 3. Entropy values and ranking of each station per subgroup

Table 4. Stations excluded from the Entropy analysis

Figure 1. Hierarchical clustering of NB gauged hydrometric stations.

Figure 2. Map of gauging stations, as well as their group and rank. Names of inactive stations are shown in gray.

Table 1. Division of the North and South Groups into subgroups based on the GEVkapMax parameter

NG1	NG2	NG3	SG1	SG2	SG3
Kap ϵ	Kap ϵ	Kap < -0.5	Kap ϵ	Kap ϵ	Kap < -0.5
]-0.33 ; +0.33[	]-0.5 ; -0.33[		]-0.33 ; +0.33[	]-0.5 ; -0.33[	
01AD002	01AF003	01AF009	01AG003	01AG002	01AN002
01AD003	01BE001	01AH005	01AJ003	01AJ004	01AR008
01AE001	01BJ001	01BJ004	01AJ010	01AJ011	
01AF002	01BJ010	01BL001	01AK001	01AK005	
01AF007	01BK003	01BR001	01AK006	01AK008	
01AF010	01BK004		01AK007	01AR011	
01AH002	01BL002		01AL002	01BU002	
01BC001	01BL003		01AL003		
01BJ003	01BO002		01AL004		
01BJ007	01BO003		01AM001		
01BJ012			01AN001		
01BO001			01AP002		
01BP001			01AP004		
01BP002			01AP006		
01BQ001			01AQ001		
			01AQ002		
			01AR004		
			01AR005		
			01AR006		
			01BS001		
			01BU003		
			01BU004		
			01BU009		
			01BV004		
			01BV005		
			01BV006		

Table 2a. Entropy values and ranking of each station (North Group)

Station	$H(Z)$	$H(G)$	$H(Z,G)$	$T(Z,\widehat{Z})$	R
01BL001	1.5694	-	-	-	0*
01BO002	1.8449	1.8744	2.7499	0.9694	1
01AF007	2.2071	2.0100	3.1765	1.0406	2
01BQ001	2.0100	2.0428	2.9876	1.0652	3
01BO001	1.8744	2.1644	2.9142	1.1245	4
01AF003	2.0100	2.0673	2.9253	1.1520	5
01BL003	2.0681	1.9416	2.8233	1.1865	6
01BL002	2.2071	2.0681	3.0681	1.2071	7
01AD003	1.7926	2.2071	2.7499	1.2499	8
01BO003	2.0681	2.0428	2.7876	1.3233	9
01BR001	2.0428	2.0681	2.7876	1.3233	9
01AH002	2.1266	2.2253	3.0058	1.3462	10
01BJ003	1.9171	2.2071	2.7681	1.3561	11
01BE001	1.8623	2.1233	2.6253	1.3602	12
01BJ001	2.1266	2.2499	2.9926	1.3839	13
01AH005	2.1744	2.1499	2.9303	1.3939	14
01BJ010	2.1499	2.2253	2.9765	1.3987	15
01BC001	1.9233	1.9416	2.3876	1.4773	16
01BJ007	1.9623	2.0058	2.4855	1.4826	17
01BP001	2.1499	2.2071	2.8520	1.5050	18
01AE001	2.1478	2.2071	2.7876	1.5673	19
01AD002	2.2253	2.1050	2.7520	1.5784	20
01AF002	2.1744	2.2681	2.8520	1.5905	21

*A rank of 0 means the station's information is unique, and thus very important.

Table 2b. Entropy values and ranking of each station (South Group).

Stations	$H(Z)$	$H(G)$	$H(Z,G)$	$T(Z,\hat{Z})$	R
01AK001	2.1449	-	-	-	0*
01AP002	2.1449	-	-	-	0
01AG003	1.8253	1.9876	2.9876	0.8253	1
01AL004	2.1121	2.0855	3.1142	1.0834	2
01AR005	1.9050	2.2071	3.0058	1.1063	3
01AM001	1.6989	2.0694	2.6549	1.1134	4
01AR011	2.1744	1.9623	3.0142	1.1224	5
01AR004	1.9623	2.1744	3.0142	1.1224	6
01AG002	2.2253	2.1478	3.1681	1.2050	7
01BU002	2.2071	2.1499	3.1142	1.2427	8
01AR006	2.0428	2.1499	2.9303	1.2623	9
01BV006	2.0428	2.0855	2.8499	1.2784	10
01BU003	2.0673	2.1449	2.8926	1.3196	11
01BS001	2.1121	2.2499	3.0303	1.3316	12
01AL002	1.8253	2.0100	2.4926	1.3428	13
01AK005	1.9303	1.9303	2.5071	1.3536	14
01AP004	1.8478	2.1926	2.6765	1.3639	15
01AK007	2.1926	2.1644	2.9303	1.4266	16
01AQ002	2.0681	2.0694	2.6926	1.4449	17
01AP006	2.2171	2.0549	2.8171	1.4549	18
01AJ004	2.0303	2.0058	2.4694	1.5668	19
01AN001	2.1926	2.1050	2.7253	1.5723	20
01AJ003	1.9876	2.0794	2.4926	1.5744	21
01AK008	2.1499	2.1303	2.7058	1.5744	22
01AL003	2.0694	2.1499	2.5897	1.6295	23
01AJ010	2.2071	2.2071	2.5765	1.8377	25
01AJ011	2.1644	2.2499	2.6499	1.7644	24
01AQ001	2.1171	2.1499	2.4142	1.8527	26
01AN002	2.2071	2.2253	2.5338	1.8987	27
01AK006	2.1233	2.0681	2.2855	1.9058	28

*A rank of 0 means the station's information is unique, and thus very important.

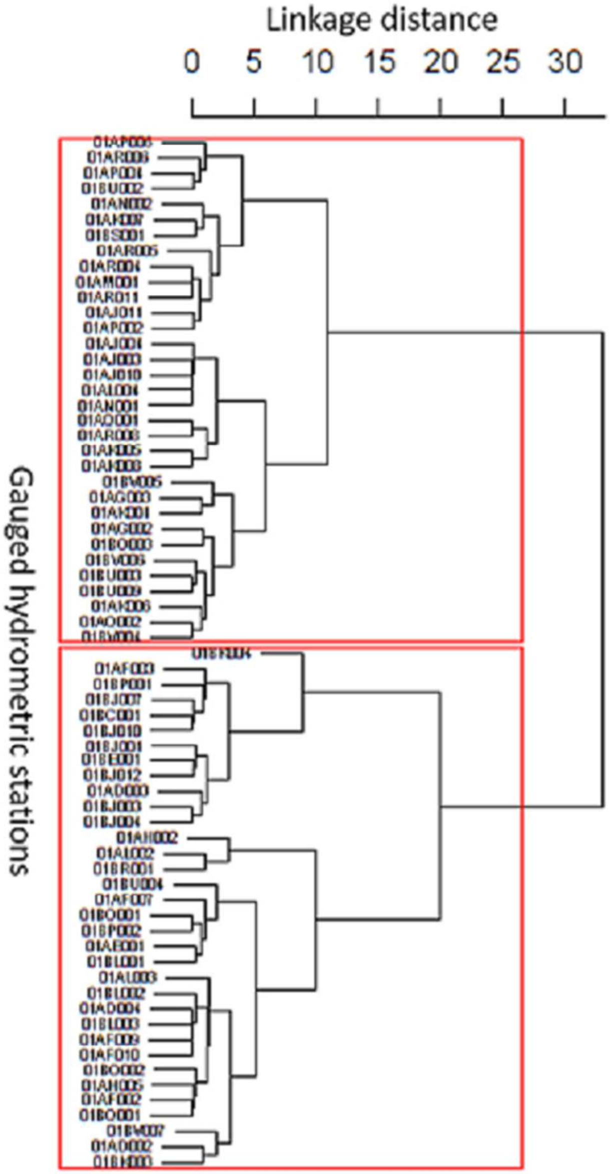
Table 3. Entropy values and ranking of each station per subgroup

NG1 (Rank)	NG2 (Rank)	NG3 (Rank)	SG1 (Rank)	SG2 (Rank)	SG3 (Rank)
01AF007 (2)	01BO002 (1)	01BL001 **(0)**	01AK001 **(0)**	01AR011 (5)	01AN002 (27)
01BQ001 (3)	01AF003 (5)	01BR001 (9)	01AP002 **(0)**	01AG002 (7)	
01BO001 (4)	01BL003 (6)	01AH005 (14)	01AG003 (1)	01BU002 (8)	
01AD003 (8)	01BL002 (7)		01AL004 (2)	01AK005 (14)	
01AH002 (10)	01BO003 (9)		01AR005 (3)	01AJ004 (19)	
01BJ003 (11)	01BE001 (12)		01AM001 (4)	01AK008 (22)	
01BC001 (16)	01BJ001 (13)		01AR004 (6)	01AJ011 (24)	
01BJ007 (17)	01BJ010 (15)		01AR006 (9)		
01BP001 (18)			01BV006 (10)		
01AE001 (19)			01BU003 (11)		
01AD002 (20)			01BS001 (12)		
01AF002 (21)			01AL002 (13)		
			01AP004 (15)		
			01AK007 (16)		
			01AQ002 (17)		
			01AP006 (18)		
			01AN001 (20)		
			01AJ003 (21)		
			01AL003 (23)		
			01AJ010 (25)		
			01AQ001 (26)		
			01AK006 (28)		
01AF010 **(UR)***	01BK003 **(UR)**	01AF009 **(UR)**	01BU004 **(UR)**		01AR008 **(UR)**
01BJ012 **(UR)**	01BK004 **(UR)**	01BJ004 **(UR)**	01BU009 **(UR)**		
01B0P2 **(UR)**			01BV004 **(UR)**		
			01BV005 **(UR)**		

***UR** indicates that the station was excluded from the entropy analysis

Table 4. Stations excluded from the Entropy analysis

Station N°	Station Name	Active	Record length (years)	Drainage Area (Km^2)	Mean Annual Flow (m^3/s)
01AD004	SAINT JOHN RIVER AT EDMONSTON	Yes	46	15500	200.41
01AF009	IROQUOIS RIVER AT MOULIN MORNEAULT	Yes	21	182	4.09
01AF010	GREEN RIVER AT DEUXIEME SAULT	No	16	1030	28.65
01AR008	BOCABEC RIVER ABOVE TIDE	No	14	43	1.10
01BJ004	EEL RIVER NEAR EEL RIVER CROSSING	No	17	88.6	2.08
01BJ012	EEL RIVER NEAR DUNDEE	Yes	29	43.2	0.94
01BK003	NEPISIGUIT RIVER AT NEPISIGUIT FALLS	No	31	1840	33.92
01BK004	NEPISIGUIT RIVER NEAR PABINEAU FALLS	No	18	2090	45.09
01BP002	CATAMARAN BROOK AT REPAP ROAD BRIDGE	Yes	24	28.7	0.64
01BU004	PALMERS CREEK NEAR DORCHESTER	No	20	34.2	0.92
01BU009	HOLMES BROOK SITE NO.9 NEAR PETITCODIAC	Yes	17	6.2	0.12
01BV004	BLACK RIVER AT GARNET SETTLEMENT	Yes	52	40.4	1.32
01BV005	RATCLIFFE BROOK BELOW OTTER LAKE	No	12	29.3	0.99
01BV007	UPPER SALMON RIVER AT ALMA	No	13	181	7.28



127x236mm (96 x 96 DPI)

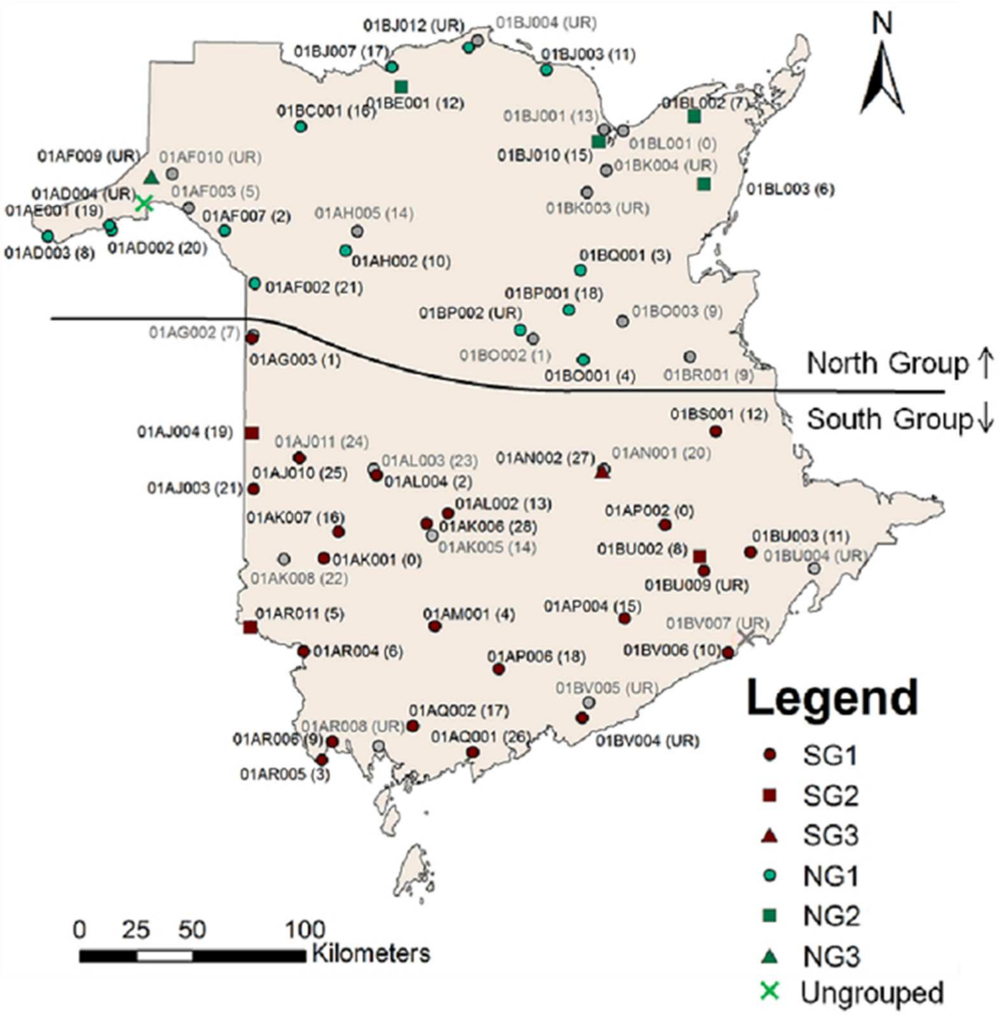


Figure 2. Map of gauging stations, as well as their group and rank. Names of inactive stations are shown in gray.

182x186mm (96 x 96 DPI)