Introduction of the GAM model for regional low-flow frequency analysis at ungauged basins and comparison with commonly used approaches

T. B.M.J. Ouarda[1*], C. Charron[1], Y. Hundecha[2], A. St-Hilaire[1], F. Chebana[1]

[1]Canada Research Chair in Statistical Hydro-Climatology, INRS-ETE, 490 de la Couronne, Quebec, QC, G1K9A9, Canada

[2]Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

*Corresponding author:

Email: taha.ouarda@ete.inrs.ca

Tel: +1 418 654 3842

August 2018

Abstract

Generalized Additive Models (GAMs) are introduced in this study for the regional estimation of low-flow characteristics at ungauged basins and compared to other approaches commonly used for this purpose. GAMs provide more flexibility in the shape of the relationships between the response and explanatory variables in comparison to classical models such as multiple linear regression (MLR). Homogeneous regions are defined here using the methods of hierarchical cluster analysis, canonical correlation analysis and region of influence. GAMs and MLR are then used within the delineated regions and also for the whole study area. In addition, a spatial interpolation method is also tested. The different models are applied for the regional estimation of summer and winter low-flow quantiles at stations in Quebec, Canada. Results show that for a given regional delineation method, GAMs provide improved performances compared to MLR.

Keywords: Low flows; Regional estimation; Canonical correlation analysis; Region of influence; Hierarchical cluster analysis; Generalized additive models.

1. Introduction

Assessment of low-flow characteristics is traditionally performed using different approaches including flow duration curves, frequency analysis of extreme low-flow events and continuous low-flow intervals, baseflow separation and characterization of streamflow recessions (Smakhtin, 2001). Knowledge of the magnitude and frequency of low flows for streams is important for water-supply planning and design, waste-load allocation, reservoir storage design, and maintenance of quantity and quality of water for irrigation, recreation, and wildlife conservation (Smakhtin, 2001). The frequency analysis of extreme low flows consists in fitting appropriate probability distributions to the annual minimum flow (Lawal and Watt, 1996; Nathan and McMahon, 1990; Ouarda et al., 2008b; Russell, 1992) defined as the annual minimum of daily or monthly discharges or averages of consecutive flows over a certain number of days (Zalants, 1991). The most used low-flow statistics in hydrology are the quantiles $Q_{d,T}$ of the minimum mean discharge over *d* days corresponding to a return period of *T* years. These low-flow quantiles are operationally related to the concept of environmental flows, which are flow regimes designed to maintain a river in some agreed ecological condition (Acreman, 2005; Smakhtin and Eriyagama, 2008).

The reliability of the estimates of the desired low-flow characteristics, however, depends on the amount of available streamflow data from which the at-site estimates are obtained. In practice, it is often the case that many streams are poorly monitored, do not have enough record to enable estimation of the required low flows, or are simply ungauged. To circumvent this problem, various approaches have been attempted, which enable estimation of low-flow characteristics at ungauged basins. A comprehensive review of methods of low-flow estimation at ungauged sites has been presented by Smakhtin (2001). Statistical regionalization methods have been among the most widely used schemes over the last decades to estimate low-flow characteristics at ungauged or poorly gauged locations using data from gauged sites (Charron and Ouarda, 2015; Durrans and Tomic, 1996; Gustard et al., 1997; Holmes et al., 2005; Laaha and Blöschl, 2006; Rees et al., 2006; Requena et al., 2018; Tsakiris et al., 2011).

In practice, regionalization of low-flow characteristics is generally carried out with one of two commonly used approaches. The first approach consists of estimating low-flow characteristics from a set of explanatory variables using a regression model calibrated with at-site estimates of low-flow characteristics at gauged stations (Fennessey and Vogel, 1990; Vogel and Kroll, 1990). The second approach is based on the assumption that the low-flow distribution functions at all sites within a region considered to be homogeneous are the same when standardized by a site specific index flow (Dalrymple, 1960). The parameters of the regional low-flow distribution function are generally estimated from the corresponding parameters of the local low-flow distribution functions obtained at each gauged site within the region. Regional estimation of the required low-flow quantile is then performed by rescaling the quantile value estimated from the regional distribution by the index flow.

In both regionalization approaches, the identification of sites that constitute a homogeneous region is usually carried out. Different approaches can be implemented to achieve this. It would be logical to group sites based on similarity of certain statistical properties of their flow records. This, however, would only be possible if all the sites were properly gauged. In order to allow estimation at ungauged sites, therefore, other methods that do not require analysis of flow records are used. In the absence of detailed information on catchment characteristics, sites may be grouped based on their geographic proximity (Smakhtin, 2001). However, geographic proximity does not guarantee the similarity of catchments and this does not necessarily lead to the grouping of hydrologically similar sites. Indeed, the hydrological response of a catchment is a function of a set of physiographic and meteorological attributes of the catchment which are often not continuous in space. Alternatively, such attributes can be employed as surrogates of the hydrological behaviour to define homogeneous regions.

Several methodologies for grouping sites into homogeneous regions were developed in the past for the regionalization of flood flows (Acreman and Sinclair, 1986; Burn, 1990; Hosking and Wallis, 1993; Ouarda, 2016). Homogeneous regions have been defined as geographically contiguous regions, geographically non-contiguous regions, or as hydrological neighbourhoods. For the delineation of geographically non-contiguous regions, clustering methods such as hierarchical cluster analysis (HCA) are often used. HCA identifies sites that are identical with one another based on the distance between sites within the physiographic-meteorological space. The HCA method groups sites into fixed regions, which are exclusive of one another. On the other hand, neighbourhood approaches identify hydrologically similar sites for each target site separately. That means, every site can have a unique set of stations within its neighbourhood. Obviously, this does not necessarily lead to homogeneous regions that are exclusive of one another as in the case of HCA. This, consequently, might lead to having a large number of stations in the neighbourhood of each target site depending on the criteria employed for region delineation. The neighbourhood approach can be based on the region of influence (ROI) principle (Burn, 1990) or on the use of canonical correlation analysis (CCA) (Ouarda et al., 2001). In a comparison study dealing with regional flood frequency analysis approaches, Ouarda et al. (2008a) indicated that the neighbourhood approach for the delineation of groups of hydrologically homogeneous basins is superior to the fixed set of regions approaches. This kind of comparison, although well established for floods, has not been carried out for regional low-flow frequency analysis methods.

The spatial interpolation (SI) approach is based on the assumption that there is a continuous and gradual spatial variation of flow characteristics. Based on this assumption, an areal mapping of the flow characteristics is produced by interpolating the values at gauged sites to estimate the values at unsampled locations. Interpolation techniques, such as regression or kriging, were used for flow regionalization by a number of authors (Daviau et al., 2000; Eaton et al., 2002; Huang and Yang, 1998). In order to avoid the scaling effect due to the differences in the sizes of the contributing drainage areas at the observation sites, the map is produced using specific flows (flows standardized by the size of the contributing area). Since flow characteristics estimated at any gauged location in a region are assumed to be representative of the whole catchment upstream of the gauge, the calculated flow values are usually assigned to the centroids of gauged catchments (Smakhtin, 2001). The SI method does not take any of the physiographic and meteorological attributes of a catchment into consideration and the information for the regional estimation of the flow characteristics is acquired based only on geographic proximity. This proximity, however, does not always guarantee similarity in the hydrological response of catchments (Ouarda et al., 2001). Nevertheless, the approach can be useful in the absence of detailed catchment physiographic and meteorological information.

Multiple linear regression (MLR), generally used in the regionalization of hydrological extreme variables, assumes a linear relation between the response variable and the explanatory variables. However, this assumption is not always met. To account for the presence of potential non-linearities, alternative methods such as artificial neural networks (ANNs) or Generalized Additive Models (GAMs) have been proposed. The use of ANNs for prediction and forecasting in the fields of environmental and water resources modelling has become increasingly popular since the early 1990s (Maier et al., 2010; Wu et al., 2014). ANNs were applied for the regionalization of flood flows in Shu and Ouarda (2007), and low flows in Ouarda and Shu

(2009). The use of GAMs has been gaining rapid popularity in a number of fields such as public health (Bayentin et al., 2010; Leitte et al., 2009; Vieira et al., 2009), renewable energy (Ouarda et al., 2016), environmental studies (Wen et al., 2011; Wood and Augustin, 2002) and hydrology (Rahman et al., 2018). Chebana et al. (2014) introduced GAMs for the regionalization of flood flows. Nonlinear models were proven in a number of studies to be superior to the traditional regression linear model for the estimation of hydrological extreme variables (Durocher et al., 2015, 2016a, 2016b; Ouali et al., 2016a, 2016b, 2017; Wazneh et al., 2013, 2016).

The aim of the present work is to extend the application of the most recent methods used in regional flood frequency analysis to the analysis of low-flow characteristics and compare their performances in terms of reproducing at-site estimates. It is proposed here to introduce GAMs to the regional estimation of low-flow characteristics and compare their performances with the MLR approach frequently used in regionalization studies. The method of index flow is not considered here based on the fact that it obtained equivalent performances to MLR in previous studies (Ouarda et al., 2001). GAMs and MLR are used in conjunction with the methods HCA, ROI and CCA for the delineation of homogeneous regions. GAMs and MLR are also applied on the whole study area without the delineation of homogeneous regions. This is justified by the fact that in Chebana et al. (2014), GAMs, in conjunction with the neighbourhood approach, did not provide a significant gain in performance compared to the linear approach. A SI method using splines is also applied in the present study. The regional models are applied to a group of catchments in the province of Quebec (Canada) and performances are compared.

The paper is organized as follows: A brief theoretical overview of the regionalization approaches that are considered in this research is presented in the next section. The case study is presented in Section 3. The methodology is presented in Section 4 and the results of the intercomparison are illustrated in Section 5. Finally, the conclusions are presented in Section 6.

2. Theoretical background

2.1. Delineation of homogeneous regions

2.1.1. Hierarchical cluster analysis (HCA)

HCA is a collection of statistical methods which identify groups of samples that behave similarly or show similar characteristics. The first step in HCA is the establishment of the similarity between each pair of stations in the dataset. This is done by computing the distance between stations in the space defined by a group of selected physiographic-meteorological variables using a distance function. The selected catchment attributes are chosen from those that exhibit a relationship with the flow characteristics and for which the values are available for all sites in the network (Burn, 1989). Then, stations are grouped into a binary hierarchical cluster tree. In HCA, each station is initially assigned to its own singleton cluster by using a linkage function which is based on the distance information generated in the first step. The analysis then proceeds iteratively, at each stage joining the two most similar clusters into a new one, until there is only one overall cluster. To represent the results of a cluster analysis, a dendrogram (tree diagram) is used. Cluster formation is followed by a procedure for determining groupings of clusters to create hydrologically homogeneous regions. This step can be carried out either by detecting natural groupings in the hierarchical tree or simply by cutting off the tree at a point which may be determined by the targeted number of clusters.

The application of HCA to the delineation of homogeneous regions is hence not automatic, as the user must intervene at each step to select among a number of choices. In the first step, the user must select the most relevant physiographic and/or meteorological variables that will be used in the computation of the distances between stations. A variety of distances, such as the Euclidean distance, Mahalanobis distance or City-block distance may be employed at this stage. The choice of the linkage function (nearest neighbour, furthest neighbour, Ward's method, etc.) also has a significant impact on how the clusters are formed. Finally, the choice of the cut-off distance on the hierarchical tree must reflect the objective pursued by the user, e.g. finding the optimal number of clusters. For a more thorough description of the various aspects of the HCA technique, the reader is referred to textbooks such as Rencher and Christensen (2012).

2.1.2. Canonical correlation analysis (CCA)

Canonical correlation analysis (CCA) consists in reducing two groups of variables into pairs of canonical variables, which are linear combinations of the variables in each group and are established in such a way that the correlations between the pairs are maximized. There are, in general, as many canonical pairs (p) as the minimum number of variables in either of the two groups. The analysis is usually performed on the standardized data and the canonical variables are also standardized such that they have a unit variance. In the context of identifying the hydrological neighbourhood corresponding to a given basin for the regionalization of low flows, the variables constituting the first group are defined as a set of low-flow characteristics, which are generally established as low flows associated with different occurrence probabilities. Those constituting the second group can be defined based on a set of physiographic and/or meteorological characteristics of the drainage basins.

The identification of the hydrological neighbourhood of a basin using CCA is performed based on the sampling theory of the canonical variables and the corresponding canonical correlations. Let **W** and **V** be *p*-dimensional vectors of the canonical variables corresponding to the hydrological and the physiographic-meteorological variables respectively, $(\lambda_1,\ldots,\lambda_p)$ a sequence of the corresponding canonical correlation coefficients, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1,\ldots,\lambda_p)$. If **W** and **V** are jointly *p*-normally distributed, the conditional distribution of **W** given **V** is approximately *p*-normal:

$$\left(\mathbf{W} \mid \mathbf{V} = \mathbf{v}_0\right) \approx N_p\left(\mathbf{\Lambda}\mathbf{v}_0, \mathbf{I}_p - \mathbf{\Lambda}^2\right), \qquad (1)$$

where $\mathbf{I}_p$ is a $p\times p$ identity matrix, and $\mathbf{v_0}$ denotes the corresponding values of the canonical physiographic variables for the target basin. Eq. (1) implies that **W** would be scattered around a mean position $\mathbf{\Lambda}\mathbf{v}_0$ with a conditional probability density function given by:

$$f\left(\mathbf{W} \mid \mathbf{V} = \mathbf{v}_0\right) = \left(2\pi\right)^{-p/2}\left|\mathbf{I}_p - \mathbf{\Lambda}^2\right|^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{W} - \mathbf{\Lambda}\mathbf{v}_0\right)'\left(\mathbf{I}_p - \mathbf{\Lambda}^2\right)^{-1}\left(\mathbf{W} - \mathbf{\Lambda}\mathbf{v}_0\right)\right], \qquad (2)$$

where $\left(\mathbf{W} - \mathbf{\Lambda}\mathbf{v}_0\right)'$ denotes the transpose of the matrix $\left(\mathbf{W} - \mathbf{\Lambda}\mathbf{v}_0\right)$. The Mahalanobis distance given by the quadratic form of the conditional distribution, $D^2 = \left(\mathbf{W} - \mathbf{\Lambda}\mathbf{v}_0\right)'\left(\mathbf{I}_p - \mathbf{\Lambda}^2\right)^{-1}\left(\mathbf{W} - \mathbf{\Lambda}\mathbf{v}_0\right)$, can be used to define a homogeneous neighbourhood for the target basin as the region in the canonical space **W** where the realizations **w** of **W** for which $\mathbf{V} = \mathbf{v}_0$ would be found.

The $100(1-\alpha)\%$ confidence level neighbourhood is therefore defined as the set of basins having location vectors **W** in the hydrological canonical space such that:

$$\left(\mathbf{W}-\mathbf{\Lambda v}_0\right)'\left(\mathbf{I}_p-\mathbf{\Lambda}^2\right)^{-1}\left(\mathbf{W}-\mathbf{\Lambda v}_0\right)\le \boldsymbol{\chi}^2_{\alpha,p}\,, \tag{3}$$

where $\boldsymbol{\chi}^2_{\alpha,p}$ is such that, for an observed Mahalanobis distance χ^2, $\mathrm{P}\left(\boldsymbol{\chi}^2 \le \boldsymbol{\chi}^2_{\alpha,p}\right)=1-\alpha$. Eq. (3) describes the interior of an ellipsoidal region in the canonical space **W**. Detailed description of the theoretical background as well as application of the CCA methodology for the identification of hydrological neighbourhoods is presented in Ouarda et al. (2000).

2.1.3. Region of influence (ROI)

Similar to the CCA approach, the ROI method is also based on the identification of homogeneous neighbourhoods for each target site and was first proposed by Acreman (1987). Later, Burn (1990) adopted it for the regionalization of flood flows and named it the “region of influence” method. ROI was used for the estimation of low-flow statistics in Holmes et al. (2002, 2005). In this method, each station is considered the centre of its own region formed by stations with similar flow characteristics. The identification of a ROI for a given station is based on a Euclidean distance in a multidimensional space defined by a set of statistical measures of the hydrological attributes of a site as well as the physiographic and meteorological attributes of the contributing basin. For ungauged sites, only physiographic and meteorological catchment attributes are used to define the space. The ROI for a station constitutes all stations within a certain critical distance from the target site. A similar concept is implemented in this work for the regionalization of low-flow characteristics.

To avoid the possible bias that might result due to the inconsistency of the scales of the different attributes, the Euclidean distance D_{ij} between stations i and j is computed using the standardized values of the hydrological and physiographic-meteorological attributes as:

$$D_{ij} = \left(\sum_{k=1}^{K} \left(C_k^i - C_k^j \right)^2 \right)^{1/2}, \tag{4}$$

where C_k^i and C_k^j are the standardized values of attribute k for stations i and j respectively, and K is the number of attributes used to define the Euclidean space. The attributes used to define the space are selected based on the knowledge of their relevance to low-flow characteristics of the contributing basin. Once they are selected, the stations to be included into the ROI for a given target station are selected as those within a certain threshold distance δ_i:

$$\mathrm{ROI}_i = \left\{ k : D_{ik} \le \delta_i \right\}. \tag{5}$$

The value of δ_i is fixed in such a way that there is a good compromise between the number of stations in the neighbourhood and the hydrological homogeneity of the selected stations. δ_i has a specific value for a given site and is a function of a set of physical conditions pertaining to the site. More details concerning the method and the definition of the thresholds are given in Ouarda (2016).

2.2. Regional estimation methods

2.2.1. Multiple linear regression (MLR)

The method of MLR allows to obtain a regional estimate of the low flow by establishing a direct relationship between the hydrological variables (low-flow quantiles) and the physiographic-meteorological explanatory variables. Topographic parameters such as relief of the catchment (Vogel and Kroll, 1990, 1992), which is defined as the difference between the elevations of the summit of the catchment and that of the gauging station, are among the physiographic variables widely used for the estimation of low-flow quantiles. Additionally, geological parameters such as the proportions of gravel and silt also have a significant influence on low flows (Dingman and Lawlor, 1995). Among the meteorological variables, mean annual precipitation is the most widely used variable (Chang and Boyer, 1977). Other parameters, such as the 10-year return period value of the maximum temperature over seven consecutive days, have also been implemented (Chang and Boyer, 1977).

The MLR method is applied on a group of catchments which are similar in terms of the statistical properties of their hydrological responses (Hosking and Wallis, 1993). It is often assumed that the relationship between the explanatory variables and the *T*-year return period *d*-day minimum flow has the following form:

$$Q_{d,T} = \theta_0 \exp(\varepsilon) \prod_{i=1}^{p} X_i^{\theta_i} , \tag{6}$$

where θ_i is a model coefficient associated with the explanatory variable X_i (θ_0 is the ordinate at the origin), p is the number of explanatory variables used in the model and ε is the multiplicative error of the model. This error can also be additive and in that case, the relationship becomes:

$$Q_{d,T} = \theta_0 \prod_{i=1}^{p} X_i^{\theta_i} + \varepsilon . \tag{7}$$

A logarithmic transformation is generally applied to linearize the relation in Eq. (6):

$$\log Q_{d,T} = \log\theta_0 + \sum_{i=1}^{p}\theta_i \log X_i + \varepsilon . \tag{8}$$

The coefficients θ_i of the model are generally estimated using the ordinary least squares approach (Thomas and Benson, 1970), the weighted least squares method (Tasker, 1980) or the generalized least squares method (Kroll and Stedinger, 1998; Stedinger and Tasker, 1985).

2.2.2. Spatial interpolation (SI)

Interpolation of low flows is generally performed at grids (regular or irregular) across the study region using techniques such as 1) linear interpolation, where low flows are assumed to vary linearly between adjacent observations, and 2) averaging technique, where the mean of low flows of all stations contained within the grid cell is used as estimator, either as a simple average or area-weighted average (Arnell, 1995). An interpolation method widely used in earth sciences is the minimum curvature method (Smith and Wessel, 1990). This method consists in fitting a twice differentiable surface through the observations. Physically, it can be interpreted as stretching and deforming an elastic plate so that it fits all the observations. This might, however, result in large oscillations and unrealistic inflection points in the fitted surface. To avoid this, Smith and Wessel (1990) introduced a tension term in the flexibility equation that leads to minimization of the oscillations and the inflection points. Formally, the fitted surface is the solution of Eq. (9):

$$(1-\rho)\nabla^4 H + \rho\nabla^2 H = 0, \tag{9}$$

where H is the low flow standardized by the drainage area, ∇^4 and ∇^2 are the biharmonic and Laplace operators respectively, and $\rho \in [0,1]$ is the tension term. Eq. (9) is solved under the constraint that the observed values are honoured at the observation locations. $\rho = 0$ leads to undesirable oscillations of the surface and $\rho = 1$ yields a harmonic surface. Johnston and Merrifield (2000) suggested a value of $\rho = 0.25$ for the interpolation at regular grids of geographic coordinates from irregularly spaced stations.

2.2.3. Generalized additive models (GAMs)

GAMs, introduced by Hastie and Tibshirani (1986), extend the generalized linear models (GLMs) by replacing the linear predictor by a set of smooth functions of the explanatory variables. GLMs are themselves a generalization of MLR in which the response variable Y can follow any distribution of the exponential family and the link function g transforms Y to a scale where the model is linear. For a response variable Y, GAMs can be expressed by:

$$g(\mathrm{E}(Y \mid \mathbf{X})) = \alpha + \sum_{j=1}^{p} f_j(X_j) \quad , \tag{10}$$

where f_j is the smooth function of the j-th explanatory variable X_j, $\mathbf{X}$ is a matrix whose columns correspond to a set of p explanatory variables, α is an intercept and $g(.)$ is a monotonic link function. With the smooth functions, GAMs are more flexible than GLMs by allowing a non-linear relation between the response variable and each of the explanatory variables.

The smooth function f_j can be defined by a linear combination of q basis functions $b_{ji}(x)$:

$$f_j(x) = \sum_{i=1}^{q} \beta_{ji} b_{ji}(x), \tag{11}$$

where β_{ji} are smoothing coefficients. The smooth function in GAMs is often estimated by a spline defined by a curve composed of piecewise polynomial functions, joined together at points called knots. A number of spline types have been proposed in the literature: cubic splines, P-splines, B-splines, etc. In a regression spline, the number of knots is considerably reduced. For such spline, the position of the knots needs then to be chosen. However, with penalized splines, the exact location and the number of the knots are not as important as the smoothing parameters which control the smoothness of the spline.

The natural cubic spline interpolates each data value. To avoid the problem of overfitting, GAMs are usually optimized by maximizing the penalized log-likelihood:

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{p}\lambda_j \boldsymbol{\beta}'\mathbf{S}_j\boldsymbol{\beta} \quad , \tag{12}$$

where $\boldsymbol{\beta}$ is a matrix of smoothing coefficients, $\boldsymbol{\beta}'$ is the transpose of $\boldsymbol{\beta}$, $l(\boldsymbol{\beta})$ is the log-likelihood function, λ_j is the smoothing parameter of the *j*-th smooth function f_j, and $\mathbf{S}_j$ is a matrix of known coefficients (Wood, 2008). The parameter λ_j controls the degree of smoothness of the smooth function. With values ranging from 0 to 1, 0 corresponds to the un-penalized case and 1 to the completely smoothed case. The optimum value of λ_j is a right balance between the fitting objective and smoothness. The function $l_p(.)$ is maximized for $\boldsymbol{\lambda}$, a given vector of smoothing parameters, by the penalized iteratively reweighted least squares method (P-IRLS; Wood, 2004). $\boldsymbol{\lambda}$ is found iteratively according to a criterion such as the generalized cross validation (GCV; Wahba, 1985), unbiased risk estimator (UBRE; Craven and Wahba, 1978) or maximum likelihood (ML).

3. Case study

The proposed approaches are applied to the hydrometric station network of southern Quebec (Canada). The hydrological and physiographic-meteorological variables used in the present study come from a low-flow frequency analysis study by Charron and Ouarda (2015). In the present study, we analyse separately the summer and winter low-flow quantiles $Q_{d,T}$ corresponding to return periods of T = 2 and 10 years for a duration of d = 7 days, and to a return period of T = 5 years for a duration of d = 30 days. These indices are the most widely used in Canada for the analysis of water supply systems during droughts and for the study of the waste assimilative capacity of streams (Ouarda et al., 2008b). Data from 190 hydrometric stations managed by the Ministry of Environment of Quebec (MENV) were used (Data are available at https://www.cehq.gouv.qc.ca/hydrometrie/historique_donnees/default.asp). The database does not include any nested catchments. Only stations that meet the following three criteria were retained: First, the gauged river should have a flow regime that is natural. Secondly, the station should have a historical record period of at least 10 years. Finally, the historical data at the station should meet the basic assumptions of independence and stationarity. The non-parametric test of Wald and Wolfowitz (1943) was used to test the independence of the d-day low-flow series, and the non-parametric Kendall test (Kendall, 1975) was used to test the stationarity of the d-day low-flow series.

Finally, 134 and 135 stations were retained for the analysis of $Q_{30,T}$ for the summer and winter seasons, respectively. Similarly, 129 and 133 stations were retained for the analysis of $Q_{7,T}$ for the summer and winter seasons, respectively. Fig. 1 shows the location of the gauging stations that were retained for any dry season and any low-flow duration. The diameters of the circles are proportional to the basin areas which vary between 0.69 and 96,600 km^2 with a median value of 1548 km^2. The stations cover a large area in the southern half of the province of Quebec. The largest catchments are located towards the northern part of the study area. The average flow record size is 32 years of data. Winter mean temperatures for the study area vary between -10 °C in the south and -21 °C in the north. Summer mean temperatures vary between 20 °C in the south and 12 °C in the north. The typical annual hydrograph in the area is characterized by an important spring flood caused by snow melt, followed by a summer dry season. Important rainstorms usually cause another flood season in the fall, followed by a winter dry season caused by the lack of liquid precipitation and during which the soil is often frozen. Note that low-flow data at a number of these stations were analysed in several previous studies for the detection of non-stationarities and for the multivariate characterization of low-flow descriptors (Ehsanzadeh et al., 2011; Khaliq et al., 2008; Lee et al., 2013, 2017).

A local low-flow frequency analysis was carried out at each station of the database in order to estimate at-site low-flow quantiles $Q_{d,T}$ corresponding to the various return periods T and durations d. Low-flow d-day series were fitted with the following statistical distributions (Rao and Hamed, 2000): the Generalized Extreme Value distribution (GEV), Gumbel (EV1), Weibull (W2), two- and three-parameter Lognormal (LN2 and LN3 respectively), Gamma (G), Person type III (P3), Log-Pearson type III (LP3) and Generalized Pareto (GP) distributions. The distribution that best fits the data at each station is then selected based on the Bayesian information criterion (BIC; Schwarz, 1978) to allow for appropriate local estimation of low-flow quantiles. Fig. 2 illustrates the frequency with which the various distributions were selected for the winter and summer 7-day low flows. Descriptive characteristics of the obtained quantiles are summarized in Table 1.

A set of physiographic and meteorological variables for each catchment of the study area are available and come from Charron and Ouarda (2015). The characteristics of the selected stations are provided in the supplementary Table S1. Table 1 lists all the variables as well as their descriptive statistics. Catchment delineation for the hydrometric stations of this study was performed in the ESRI ArcGIS environment using the ESRI Arc Hydro Tools available at resources.arcgis.com/en/communities/hydro. Arc Hydro Tools include functionalities for catchment delineation from Digital Elevation Models (DEMs). The DEM used in this study is Canada 3D available from Natural Resources Canada at http://ftp.geogratis.gc.ca/pub/nrcan_rncan/elevation/canada3d/. Catchment rasters obtained were after converted to polygon features which were used to compute the spatial averages of the physiographic and meteorological variables in this study.

The catchment area (AREA), the latitude (LAT) and longitude (LONG) of the catchment centroid were computed directly from the catchment polygon. The average slope of the catchment (MSLP) was computed from the DEM. The variables related to the land coverage, mean curve number (MCN), percentage of forest cover (PFOR) and percentage of lakes (PLAKE), were computed from digital maps of Quebec (Maps are available from Natural Resources Canada at http://open.canada.ca/en/open-maps). MCN consists of an area-weighted average of the curve number (CN) values in the catchment. The major factors that determine CN are the hydrological soil group, cover type, treatment, hydrological condition, and antecedent runoff condition (USDA, 1986). Its values range from 0 to 100 with a lower value representing the most pervious soil and a higher value representing the most impervious soil. Fig. 3 shows the distribution of the values of CN within the study area.

The five meteorological variables, mean total annual precipitation (PTMA), average summer/fall liquid precipitation (PLMS), average degree-days below 0 °C (DDBZ), average degree-days above 13 °C (DDH13) and average number of days where mean temperature exceeds 27 °C (NDH27), were computed through a spatial interpolation of the meteorological data of the MENV. Universal kriging (Isaaks and Srivastava, 1989) was implemented for the spatial interpolation. Using the geographic location of every meteorological station, an interpolation of meteorological contour lines was performed for the whole province. The meteorological stations which were selected had at least 15 years of data and the earliest starting year is 1940.

4. Methodology

4.1. Regional models

The methods presented in Section 2 for the delineation of homogeneous regions are used in conjunction with the methods MLR and GAMs for the transfer of hydrological information. These regional models are denoted by HCA+MLR, ROI+MLR, CCA+MLR, HCA+GAM, ROI+GAM and CCA+GAM. As indicated in Section 1, other tested models are obtained by applying MLR and GAMs to the whole dataset without delineation of homogeneous regions. These models are denoted respectively by ALL+MLR and ALL+GAM. In this study, the R package *mgcv* (Wood, 2006) is used to estimate the GAMs parameters. Cubic regression splines are considered as smooth functions and the GCV score is used to optimize λ. The knots in smooth functions are placed at a number of quantiles of the distribution of the unique values x of a given explanatory variable.

For each regional model, different physiographic-meteorological attributes are used for the summer and winter seasons. A backward stepwise regression method, applied to all stations, is used to select the optimal explanatory variables to be used with the methods MLR and GAMs. This stepwise method is presented in the next section. To apply the delineation methods, variables considered to be the most relevant in terms of explaining the low-flow processes need to be selected. In this study, the variables selected for MLR with the stepwise regression method constitute the physiographic-meteorological variables used in each of the delineation methods. The same homogeneous regions obtained for a given delineation method are used in conjunction with either MLR or GAMs (i.e. the same regions are used for HCA+MLR and HCA+GAM, for ROI+MLR and ROI+GAM, and for CCA+MLR and CCA+GAM).

The SI method is also applied to the study area using the minimum curvature method presented in Section 2.2.2. In that case, only variables LAT and LONG are used for interpolation of specific quantiles and thus no selection of variables is required. The spatial interpolation performed in this study was carried out with the Generic Mapping Tools (Wessel et al., 2013). Once the map is produced, the low flow at an ungauged basin is estimated by multiplying the contour value corresponding to the location of its centroid by its drainage area. The contour value corresponding to the basin centroid is computed using the nearest neighbour approach from the grid values.

With the standard methods used to define the threshold in ROI and CCA, the size of homogeneous regions can vary considerably from one region to another. For instance, for a given fixed threshold, stations located on the edge of the cloud of points defined by the canonical space for CCA or the Euclidian space for ROI will have fewer stations within their neighbourhood, while stations located near the center of the cloud of points will have more stations within their neighbourhoods (Leclerc and Ouarda, 2007). Given that the sample size is essential for the reliability of the estimates obtained by MLR and GAMs, it was decided that for each target station, the size of the region is increased until a selected optimal size is reached. It was decided to fix the size of each region to three times the number of parameters to estimate in GAMs, which has more parameters to estimate than the MLR model. The number of parameters to estimate in GAMs depends on the number of predictors in the model and the number of knots in the smooth functions.

4.2. Stepwise regression

To select the optimal explanatory variables, the backward stepwise method is used (Marra and Wood, 2011). In this approach, the regression method (MLR or GAMs) is initially applied with a model including all the explanatory variables. At each step, the variable with the highest *p*-value, for the null hypothesis that the parameter (for MLR) or the smooth term (for GAMs) is zero, is removed. The procedure ends when the *p*-values of all the remaining variables are below a given threshold (5%). For the aim of simplicity, the explanatory variables obtained with the stepwise regression procedure applied to $Q_{7,2}$ are used as the explanatory variables to estimate the other quantiles. Quantile $Q_{7,2}$ is used as the quantile of reference because, having the smallest return period, it can be considered the most reliable quantile.

4.3. Validation

A leave-one-out cross-validation technique (Jackknife method) was employed to evaluate the performance of the regional estimates of the low-flow quantiles. The at-site estimate of the quantile value of interest at a given station is temporarily removed from the sample and a new value is estimated from the regression relationship established using data from the remaining stations within the homogeneous region. This process is repeated for the entire set of gauged sites. The estimated quantiles are then compared with the at-site quantile estimates computed from the observed values. The following five indices are used to evaluate the performances: the Nash criterion (NASH), the root mean squared error (RMSE), the relative root mean squared error (rRMSE), the mean bias (BIAS), and the relative mean bias (rBIAS). These performance indices are frequently used for the assessment of low flows (see Ouarda and Shu, 2009).

5. Results

In this section, results of the selection of the physiographic and meteorological variables included in the MLR and GAMs are first presented. Then, results related to the delineation methods and the SI method are discussed. Finally, a comparison of the different regionalization models is presented.

5.1. Selection of the physiographic and meteorological variables for MLR

Pearson correlation coefficients between the various explanatory variables and low-flow quantiles are presented in Table 2. These results suggest that the catchment area (AREA) is a particularly important variable and explains most of the variance of low-flow quantiles. Other important variables are PLAKE, mean annual total and liquid precipitation (PTMA and PLMS), number of days where the temperature is higher than 27 °C (NDH27), degree-days below 0 °C and higher than 13 °C (DDBZ and DDH13), and latitude (LAT). The log-linear regression model in Eq. (8) is considered for the estimation of the low-flow quantiles. Following the application of the backward stepwise procedure with MLR, the models for the summer season are defined by:

$$\log\left(\tilde{Q}_{30,5}\right) = -31.69 + 1.07\log(\text{AREA}) + 1.94\log(\text{DDBZ}) - 0.62\log(\text{MCN}) \\ + 2.07\log(\text{PTMA}) - 0.17\log(\text{NDH27}) + 0.05\log(\text{PLAKE}), \tag{13}$$

$$\log\left(\tilde{Q}_{7,2}\right) = -25.93 + 1.05\log(\text{AREA}) + 1.78\log(\text{DDBZ}) - 0.76\log(\text{MCN}) \\ + 1.50\log(\text{PTMA}) - 0.15\log(\text{NDH27}) + 0.08\log(\text{PLAKE}), \tag{14}$$

$$\log\left(\tilde{Q}_{7,10}\right) = -32.26 + 1.09\log(\text{AREA}) + 2.13\log(\text{DDBZ}) - 0.80\log(\text{MCN}) \\ + 1.97\log(\text{PTMA}) - 0.19\log(\text{NDH27}) + 0.04\log(\text{PLAKE}), \tag{15}$$

and the models for the winter season are defined by:

$$\log\left(\tilde{Q}_{30,5}\right) = -9.40 + 0.98\log(\text{AREA}) + 0.14\log(\text{PLAKE}) \\ + 0.79\log(\text{PLMS}) - 0.28\log(\text{MCN}), \tag{16}$$

$$\log\left(\tilde{Q}_{7,2}\right) = -9.02 + 0.97\log(\text{AREA}) + 0.15\log(\text{PLAKE}) \\ + 0.81\log(\text{PLMS}) - 0.36\log(\text{MCN}), \tag{17}$$

$$\log\left(\tilde{Q}_{7,10}\right) = -9.63 + 1.00\log(\text{AREA}) + 0.17\log(\text{PLAKE}) \\ + 0.92\log(\text{PLMS}) - 0.54\log(\text{MCN}), \tag{18}$$

where the predictors in Eqs. (13)-(18) are ordered from the most to the least significant. The stepwise procedure allows a selection of variables that minimizes the correlations between the explanatory variables. The AREA is the most important variable and variables AREA, MCN and PLAKE are important for both seasons. Mean annual total precipitation PTMA and mean annual liquid precipitation PLMS are selected for the summer and winter season respectively. Two temperature-related variables are selected for summer low flows (degree-days below 0 °C DDBZ and number of days higher than 27 °C NDH27) while no temperature variables are selected for winter low flows.

5.2. Selection of the physiographic and meteorological variables for GAMs

A different selection of variables is expected with GAMs because predictors presenting a non-linear relationship with the explained variable were disadvantaged with MLR over those presenting a linear relationship. The logarithmic transformation of the response variables was necessary in order to meet the assumption of constant variance of the residuals. It was also found that applying the logarithmic transformation to the variable AREA improves considerably the performances. Following the application of the backward stepwise procedure with GAMs, and given that a large number of variables would also require a large number of stations in the neighbourhoods, the optimal number of variables during summer was identified to be 6. The model used for the summer season within the models HCA+GAM, ROI+GAM and CCA+GAM is then defined by:

$$\begin{aligned}\log\left(Q_{d,T}\right) = \alpha + f_1\left(\log \text{AREA}\right) + f_2(\text{DDH13}) + f_3(\text{MCN})\\ + f_4(\text{PLMS}) + f_5(\text{PLAKE}) + f_6(\text{DDBZ}) + \varepsilon\end{aligned} \quad (19)$$

Following the application of the backward stepwise procedure with GAMs, the model for the winter season is defined by:

$$\begin{aligned}\log\left(Q_{d,T}\right) = \alpha + f_1\left(\log \text{AREA}\right) + f_2(\text{PLAKE}) + f_3(\text{PLMS})\\ + f_4(\text{MCN}) + f_5(\text{DDBZ}) + \varepsilon\end{aligned} \quad (20)$$

Variables AREA, PLAKE, MCN, mean annual liquid precipitation PLMS and degree-days below 0 °C DDBZ are important for both seasons. In addition, with GAMs, degree-days higher than 13 °C DDH13 is included for summer low flows.

The smooth functions obtained for $\log(Q_{7,10})$ for the summer and winter seasons are presented in Figs. 4 and 5 respectively. Smooth functions allow interpreting the influence of each variable without the effect of the others. It can be observed that log(AREA) is perfectly linear with $\log(Q_{7,10})$ for both seasons with narrow confidence intervals and small residuals. Some variables present important non-linear behaviours (e.g. MCN for both seasons, degree-days below 0 °C DDBZ for summer, and mean annual liquid precipitation PLMS and PLAKE for winter) while others are linear (e.g. degree-days higher than 13 °C DDH13 and PLAKE for summer, and degree-days below 0 °C DDBZ for winter). The slopes of the smooth functions of PLAKE are positive. This is explained by the fact that lakes sustain the streamflow during dry periods. The slopes of the smoothing functions of MCN are negative, reflecting the fact that more impervious (pervious) soil retains (releases) more water during dry seasons. The smooth functions of mean annual liquid precipitation PLMS for both seasons are increasing because precipitation recharges groundwater. The negative slope and the positive slope of the smoothing functions of degree-days higher that 13 °C DDH13 and degree-days below 0 °C DDBZ, respectively, for summer low flows indicate that the colder the region is, the higher the low flow will be during summer. A possible explanation is that temperature influences snow melt during spring and for colder regions, the release of water from snow melt is delayed, resulting then in higher low flows during the summer season. In the case of winter low flows, the slope of the smooth function of degree-days below 0 °C DDBZ is negative because colder temperatures increase the length of the dry season leading to a decrease in low flows. Note that these previous conclusions cannot be made only on the basis of the correlation coefficients in Table 2. For instance, the positive coefficient of correlation for PLAKE is in agreement with the positive slope of the smooth function of PLAKE. However, in the case of the precipitation-related variables, correlations are negative while the slopes of the smooth functions are positive, and in the case of degree-days below 0 °C DDBZ for winter, correlations are positive while the slope of the smooth functions are negative. Thus, conclusions drawn from Pearson's correlations differ from those obtained from GAMs. Because of their additive nature, GAMs allow to interpret the impact of a given explanatory variable on the response variable independently of the other explanatory variables. These results demonstrate that relationships based only on correlations can be misleading.

5.3. Delineation of regions with HCA, ROI and CCA

For the application of the HCA method, the standardized Euclidean distance measure based on the catchment descriptors selected for each season was employed to determine the similarity between stations. Clustering was performed using Ward's algorithm (Ward, 1963), which is based on minimizing the sum of the square of the distances between each site within a given cluster and the centroid of the cluster to ensure maximum similarity of the elements of the cluster (group). Fig. 6 shows the dendrogram obtained after application of this algorithm for the summer season. The choice of the cut-off distance has a significant impact on the number of stations in the regions and on the performances. The distance should not be too short to avoid very small regions in which case the regression would be impossible or would lead to weak performances. With this method, the number of stations in each region could be very different. In the present case, the cut-off distance is selected to provide three regions for both seasons. The regions include 61, 33 and 42 stations for summer and 76, 30 and 30 stations for winter respectively.

Considering that 6 and 5 variables, respectively, were used for the summer and winter low flows and that 5 knots were considered in the smooth functions, the optimal neighbourhood size for the ROI and CCA methods was fixed at 75 and 63 stations for the summer and the winter season, respectively. CCA requires the normality of the hydrological and physiographic-meteorological variables. Some variables were hence transformed to achieve normality. As one can see in Table 1, some of the physiographic and meteorological variables show clear asymmetry. Thus, a logarithmic transformation was applied to the low-flow quantiles, AREA and DDBZ. For PLAKE, a square root transformation was found to be more appropriate. Fig. 7 illustrates the hydrological and physiographic-meteorological canonical spaces for both seasons. No consistent clusters of stations are visible in the canonical hydrological spaces, indicating that the delineation of fixed regions may not be the most appropriate approach. This confirms that the neighbourhood approach adopted in the present study is more appropriate.

5.4. Method of spatial interpolation (SI)

The studied quantiles at each station were standardized by the area of the drainage basin corresponding to the station. The obtained values of specific quantiles were estimated at a regular grid of 2’ longitude × 2’ latitude using the minimum curvature method discussed in Section 2.2.2. Fig. 8 shows the contour maps of specific quantiles of $Q_{7,2}$ for low flows during the summer and winter seasons. The map for the summer season displays generally a vertical gradient of specific quantiles with a positive trend towards the north. This indicates an increase in the specific quantiles from warmer to colder regions. The same relation of summer low flows with temperature was observed previously in Section 5.2 with the smooth functions. For the winter season, no similar vertical gradient is visible and the distribution of specific quantiles is more homogeneous through the study area. This indicates a weaker influence of the temperature on winter low flows which was also observed in Sections 5.1 and 5.2.

5.5. Comparison of regional models

A comparison of the performances obtained with the different regional models is carried out in this subsection. The performance indices obtained from the cross-validation analysis for summer and winter low-flow quantile estimates are presented in Tables 3 and 4, respectively. The indices associated with relative errors (rBIAS and rRMSE) provide a different set of information than the indices associated with absolute errors (NASH, BIAS, RMSE) since the latter ones end up giving an overly large weight to a few extremely large basins. This is especially the case for the present database since basin areas range from less than one km^2 to almost 100,000 km^2. Plots of regional estimates versus at-site values for the summer and winter low-flow quantiles $Q_{7,10}$ are presented in Figs. 9 and 10, respectively. Plots of the relative residuals for summer and winter low-flow quantiles $Q_{7,10}$ are presented in Figs. 11 and 12, respectively. It can be noticed in these later figures that the highest relative errors are obtained for catchments with small to moderate areas and which have thus more weights in the indices of relative errors.

The cross-validation results indicate that, according to NASH, better fits are obtained for summer low flows than for winter low flows. This may be explained by the facts that more significant variables were included in the regional models for summer low flows and that the correlations presented in Table 2 are for most cases higher for the summer quantiles. On the other hand, higher rBIAS and rRMSE values are obtained for summer low flows. Among the models using MLR, the ROI+MLR model provides generally the best performances for both seasons regardless of the absolute or relative error indices. Methods using the neighbourhood approach in conjunction with MLR (CCA+MLR and ROI+MLR) provide generally better performances than the method using the fixed regions approach (HCA+MLR). The difference in relative error between the two approaches can be significant, as for instance rRMSE is 58% with HCA+MLR for the summer quantile $Q_{7,10}$ while it is 45% with ROI+MLR.

The application of GAMs without the delineation of regions (ALL+GAM) leads to an improvement of the absolute error indices in comparison to the models that use MLR. With respect to the relative error indices, performances of ALL+GAM are rather similar to those obtained with ALL+MLR, HCA+MLR and CCA+MLR, but not as good as those obtained with ROI+MLR. When GAMs are used in conjunction with HCA or ROI, significant improvements are obtained compared to ALL+GAM. The delineation method that benefits the most from the introduction of GAMs is HCA, where the performances obtained are comparable or better than those of ROI+GAM. In this regard, HCA+GAM is the best model with respect to RMSE and rRMSE for the winter low flows. For a given delineation method as well as for the information transfer methods applied to the whole study area, better performances are generally obtained with the model using GAMs instead of the one using MLR. Overall best results are obtained with ROI+GAM and HCA+GAM for both seasons, as these two combinations usually lead to best performance indices for both absolute and relative cases.

Results also indicate that SI obtained good performances with respect to the absolute error indices. However, poor results are obtained for the summer season with respect to the relative error indices. Good performances for the summer season with respect to the absolute error indices can be attributed to the fact that the biggest basin is much better estimated with SI than with the other methods as it can be noticed in Fig. 9. These general poor performances are somewhat expected considering the spatial discontinuity in catchment physiographic and meteorological characteristics. However, the method has the advantage of allowing the estimation at ungauged basins in cases where other catchment characteristics are not available.

6. Conclusions

In this study, GAMs were introduced for the estimation of low-flow quantiles. Comparison with other methods commonly employed for the regionalization of low flows was also carried out. In all, nine regionalization models were compared. For six of them, MLR and GAMs were applied within homogeneous regions using three different methods for the delineation of homogeneous regions: hierarchical clustering analysis of the sites based on their relative proximity within the physiographic-meteorological space, the region of influence approach based on the proximity of the target site with the other sites within the physiographic-meteorological space, and canonical correlation analysis of a group of low-flow characteristics and a group of physiographic and meteorological attributes of the sites. Within each delineated region, either MLR or GAMs were used for the transfer of hydrological information. For two other models, MLR and GAMs were applied to all stations of the study area without the delineation of homogeneous regions. Finally, in the last model, a technique of spatial interpolation was applied over the specific low flows of the study area.

The models were applied to a large selection of catchments in the province of Quebec. The dataset on which the proposed methods were applied represents a challenge because it includes a wide range of catchment sizes, including basins smaller than one km^2 to others as large as 100,000 km^2. Additionally, most of the quantiles are concentrated around rather low values.

GAMs allow to relate the hydrological variables to the explanatory variables through non-linear functions, while the commonly used MLR assumes a linear relationship between the response variable and the explanatory variables. However, hydrological processes are complex in nature and the assumption of linearity is not always met. In order to improve the estimates, GAMs were introduced here for the estimation of low-flow characteristics. The main advantage of GAMs is that they provide explicit expressions of the functions between the response variable and each of the explanatory variables.

A stepwise regression method was applied to the study dataset to select the optimal variables to be included in the regional models. It was observed that some variables included in GAMs present important non-linear behaviours. A leave-one-out cross-validation technique was implemented to evaluate the performance of each of the approaches. GAMs applied to the whole set of stations without homogeneous regions were found to lead to a good performance with respect to the absolute error indices, while with respect to the relative error indices, this model was found to be comparable to the approaches using MLR. When the homogeneous regions approach was used in conjunction with GAMs, better performances were obtained compared to the approach where GAMs are applied to the whole study area. These results prove that it is best practice to delineate homogeneous regions before applying GAMs. Performances were also improved when GAMs instead of MLR were used with the homogeneous regions approach. In general, GAMs with the HCA and ROI approaches provide the best overall results. These results indicate that it is relevant to use GAMs for the regional estimation of low-flow characteristics. The results of this study show that the use of GAMs instead of the linear model improves significantly the performances. GAMs can be easily applied with available software tools. The delineation of homogeneous regions represents an additional effort but results in significant improvements.

Another approach implemented here is based on the spatial interpolation of low-flow characteristics from gauged sites to estimate the values at ungauged sites. While geographic proximity of catchments by itself is not a good indicator of hydrological similarity between catchments, the spatial interpolation method, which is based on the estimation of the low-flow characteristics from the geographic pattern of the low flows is also found to produce acceptable results. This is, indeed, a desirable outcome in that it signifies the usefulness of such an approach in the absence of more informative descriptors for the regionalization of low flows.

Future work should focus on the extension of the Regional Streamflow Estimation Based Frequency Analysis (RSBFA) approach to the low-flow case. This approach was recently developed by Requena et al. (2017) and is based on the prior estimation of daily streamflows at the target ungauged site (Shu and Ouarda, 2012). Future research should also explore the impact of adopting the RSBFA on the combination of local and regional low-flow information when the target site is partially gauged, and compare the results to more complex statistical models such as the Bayesian model proposed by Seidou et al. (2006). The extension of the regional models compared in the present study to the multivariate case is also of interest.

Acknowledgements

The support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Ministry of Environment of Quebec is gratefully acknowledged. The authors are grateful to the Editor-in-Chief, Dr. Dan Ames, and to three anonymous reviewers for their comments which helped improve the quality of the manuscript.

REFERENCES

Acreman, M.C., 1987. Regional flood frequency analysis in the U.K.: Recent research-new ideas. Report of the Institute of Hydrology. Wallingford, UK.

Acreman, M., 2005. Linking science and decision-making: features and experience from environmental river flow setting. Environ. Model. Soft. 20(2), 99-109. https://doi.org/10.1016/j.envsoft.2003.08.019.

Acreman, M.C., Sinclair, C.D., 1986. Classification of drainage basins according to their physical characteristics: An application for flood frequency analysis in Scotland. J. Hydrol. 84, 365-380. https://doi.org/10.1016/0022-1694(86)90134-4.

Arnell, N.W., 1995. Grid mapping of river discharge. J. Hydrol. 167, 39–56. https://doi.org/10.1016/0022-1694(94)02626-M.

Bayentin, L., El Adlouni, S., Ouarda, T., Gosselin, P., Doyon, B., Chebana, F., 2010. Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989-2006 in Quebec, Canada. Int. J. Health Geogr. 9, 5. https://doi.org/10.1186/1476-072X-9-5.

Burn, D.H., 1989. Cluster analysis as applied to regional flood frequency. J. Water Resour. Plan. Manag. 115(5), 567-582. https://doi.org/10.1061/(ASCE)0733-9496(1989)115:5(567).

Burn, D.H., 1990. Evaluation of regional flood frequency analysis with a Region of Influence approach. Water Resour. Res. 26(10), 2257-2265. https://doi.org/10.1029/WR026i010p02257.

Chang, M., Boyer, D.G., 1977. Estimates of low flows using watershed and climatic parameters. Water Resour. Res. 13(6), 997-1001. https://doi.org/10.1029/WR013i006p00997.

Charron, C., Ouarda, T.B.M.J., 2015. Regional low-flow frequency analysis with a recession parameter from a non-linear reservoir model. J. Hydrol. 524, 468-475. https://doi.org/10.1016/j.jhydrol.2015.03.005.

Chebana, F., Charron, C., Ouarda, T.B.M.J., Martel, B., 2014. Regional frequency analysis at ungauged sites with the generalized additive model. J. Hydrometeorol. 15(6), 2418-2428. https://doi.org/10.1175/jhm-d-14-0060.1.

Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions. Numer. Math. 31(4), 377-403. https://doi.org/10.1007/bf01404567.

Dalrymple, T., 1960. Flood frequency analyses, manual of hydrology: Part 3. U.S. Geol. Surv. Water Supply Pap. 1543-A.

Daviau, J.L., Adamowski, K., Patry, G.G., 2000. Regional flood frequency analysis using GIS, L-moment and geostatistical methods. Hydrol. Process. 14(15), 2731-2753. https://doi.org/10.1002/1099-1085(20001030)14:15<2731::AID-HYP89>3.0.CO;2-U.

Dingman, S.L., Lawlor, S.C. 1995. Estimating low flow quantiles from drainage basin characteristics in New Hampshire and Vermont. J. Am. Water Resour. Assoc. 31(2), 243-256. https://doi.org/10.1111/j.1752-1688.1995.tb03377.x.

Durocher, M., Chebana, F., Ouarda, T.B.M.J., 2015. A nonlinear approach to regional flood frequency analysis using projection pursuit regression. J. Hydrometeorol. 16(4), 1561-1574. https://doi.org/10.1175/jhm-d-14-0227.1.

Durocher, M., Chebana, F., Ouarda, T.B.M.J., 2016a. Delineation of homogenous regions using hydrological variables predicted by projection pursuit regression. Hydrol. Earth Syst. Sci. 20(12), 4717-4729. https://doi.org/10.5194/hess-20-4717-2016.

Durocher, M., Chebana, F., Ouarda, T.B.M.J., 2016b. On the prediction of extreme flood quantiles at ungauged locations with spatial copula. J. Hydrol. 533, 523-532. https://doi.org/10.1016/j.jhydrol.2015.12.029.

Durrans, S.R., Tomic, S., 1996. Regionalization of low-flow frequency estimates: an Alabama case study. J. Am. Water Resour. Assoc. 32(1), 23–37. https://doi.org/10.1111/j.1752-1688.1996.tb03431.x.

Eaton, B., Church, M., Ham, D., 2002. Scaling and regionalization of flood flows in British Columbia, Canada. Hydrol. Process. 16(16), 3245-3263. https://doi.org/10.1002/hyp.1100.

Ehsanzadeh, E., Ouarda, T.B.M.J., Saley, H.M., 2011. A simultaneous analysis of gradual and abrupt changes in Canadian low streamflows. Hydrol. Process. 25(5), 727-739. https://doi.org/10.1002/hyp.7861.

Fennessey, N., Vogel, R.M., 1990. Regional flow-duration curves at ungaged sites in Massachusetts. J. Water Resour. Plan. Manag. 116, 530-549. https://doi.org/10.1061/(ASCE)0733-9496(1990)116:4(530).

Gustard, A., Rees, H.G., Croker, K.M., Dixon, J.M., 1997. Using regional hydrology for assessing European water resources. In: FRIEND'97 — Regional Hydrology: Concepts and Models for Sustainable Water Resource Management. IAHS Publ. No. 246, Postojna, Slovenia, pp. 107–115.

Hastie, T., Tibshirani, R., 1986. Generalized Additive Models. Stat. Sci. 1(3), 297-310.

Holmes, M.G.R., Young, A.R., Goodwin, T.H., Grew, R., 2005. A catchment-based water resource decision-support tool for the United Kingdom. Environ. Model. Softw. 20(2), 197-202. https://doi.org/10.1016/j.envsoft.2003.04.001.

Holmes, M.G.R., Young, A.R., Gustard, A., Grew, R., 2002. A region of influence approach to predicting flow duration curves within ungauged catchments. Hydrol. Earth Syst. Sci. 6(4), 721-731. https://doi.org/10.5194/hess-6-721-2002.

Hosking, J.R.M., Wallis, J.R., 1993. Some statistics useful in regional frequency analysis. Water Resour. Res. 29(2), 271-281. https://doi.org/10.1029/92wr01980.

Huang, W.-C., Yang, F.-T., 1998. Streamflow estimation using kriging. Water Resour. Res. 34(6), 1599-1608. https://doi.org/10.1029/98WR00555.

Isaaks, E., Srivastava, R., 1989. An introduction to applied geostatistics. Oxford Univ. Press, New York.

Johnston, T.M.S., and M.A. Merrifield, 2000. Interannual geostrophic current anomalies in the near-equatorial western Pacific. J. Phys. Oceanogr. 30(1), 3-14. https://doi.org/10.1175/1520-0485(2000)030<0003:Igcait>2.0.Co;2.

Kendall, MG., 1975. Rank correlation methods. Griffin, London.

Khaliq, M.N., Ouarda, T.B.M.J., Gachon, P., Sushama, L., 2008. Temporal evolution of low-flow regimes in Canadian rivers. Water Resour. Res. 44(8), W08436. https://doi.org/10.1029/2007wr006132.

Kroll, C.N., Stedinger, J.R., 1998. Regional hydrologic analysis: Ordinary and generalized least squares revisited. Water Resour. Res. 34(1), 121-128. https://doi.org/10.1029/97WR02685.

Laaha, G., Blöschl, G., 2006. A comparison of low flow regionalisation methods - catchment grouping. J. Hydrol., 323, 193-214. https://doi.org/10.1016/j.jhydrol.2005.09.001.

Lawal, S.A., Watt, W.E., 1996. Frequency analysis of low-flows using the Akaike information criterion. Can. J. Civ. Eng. 23, 1180–1189. https://doi.org/10.1139/l96-927.

Leclerc, M., Ouarda, T.B.J.M., 2007. Non-stationary regional flood frequency analysis at ungauged sites. J. Hydrol. 343(3-4), 254-265. https://doi.org/10.1016/j.jhydrol.2007.06.021

Lee, T., Modarres, R., Ouarda, T.B.M.J., 2013. Data-based analysis of bivariate copula tail dependence for drought duration and severity. Hydrol. Process. 27(10), 1454-1463. https://doi.org/10.1002/hyp.9233.

Lee, T., Ouarda, T.B.M.J., Yoon, S., 2017. KNN-based local linear regression for the analysis and simulation of low flow extremes under climatic influence. Clim. Dyn. 49(9), 3493-3511. https://doi.org/10.1007/s00382-017-3525-0.

Leitte, A.M., Petrescu, C., Franck, U., Richter, M., Suciu, O., Ionovici, R., Herbarth, O., Schlink, U., 2009. Respiratory health, effects of ambient air pollution and its modification by air humidity in Drobeta-Turnu Severin, Romania. Sci. Total Environ. 407(13), 4004-4011. https://doi.org/10.1016/j.scitotenv.2009.02.042.

Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environ. Model. Softw. 25(8), 891-909. https://doi.org/10.1016/j.envsoft.2010.02.003.

Marra, G., Wood, S.N., 2011. Practical variable selection for generalized additive models. Comput. Stat. Data Anal. 55(7), 2372-2387. https://doi.org/10.1016/j.csda.2011.02.004.

Nathan, R.J., McMahon, T.A., 1990. Practical aspects of low-flow frequency analysis. Water Resour. Res. 26, 2135–2141. https://doi.org/10.1029/WR026i009p02135.

Ouali, D., Chebana, F., Ouarda, T.B.M.J., 2016a. Non-linear canonical correlation analysis in regional frequency analysis. Stoch. Environ. Res. Risk Assess. 30(2), 449-462. https://doi.org/10.1007/s00477-015-1092-7.

Ouali, D., Chebana, F., Ouarda, T.B.M.J., 2016b. Quantile regression in regional frequency analysis: A better exploitation of the available information. J. Hydrometeorol. 17(6), 1869-1883. https://doi.org/10.1175/JHM-D-15-0187.1.

Ouali, D., Chebana, F., Ouarda, T.B.M.J., 2017. Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites. J. Adv. Model. Earth Syst. 9(2), 1292-1306. https://doi.org/10.1002/2016MS000830.

Ouarda, T.B.M.J., 2016. Regional flood frequency modeling, Chap. 77. In: Chow's Handbook of Applied Hydrology, 2nd Edn. Edited by Singh, V. P., McGraw-Hill Education, New York, pp. 77.1-77.8.

Ouarda, T.B.M.J., Ba, K.M., Diaz-Delgado, C., Carsteanu, A., Chokmani, K., Gingras, H., Quentin, E., Trujillo, E., Bobee, B., 2008a. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. J. Hydrol. 348(1-2), 40-58. https://doi.org/10.1016/j.jhydrol.2007.09.031.

Ouarda, T.B.M.J., Charron, C., Marpu, P.R., Chebana, F., 2016. The generalized additive model for the assessment of the direct, diffuse, and global solar irradiances using SEVIRI images, with application to the UAE. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 9(4), 1553-1566. https://doi.org/10.1109/jstars.2016.2522764.

Ouarda, T.B.M.J., Charron, C., St-Hilaire, A., 2008b. Statistical models and the estimation of low flows. Can. Water Resour. J. 33(2), 195-205.

Ouarda, T.B.M.J., C. Girard, G.S. Cavadias, Bobée, B., 2001. Regional flood frequency estimation with canonical correlation analysis. J. Hydrol. 254, 157-173. https://doi.org/10.1016/S0022-1694(01)00488-7.

Ouarda, T.B.M.J., Haché, M., Bruneau, P., Bobée, B., 2000. Regional flood peak and volume estimation in northern Canadian basins. J. Cold Reg. Eng. 14(4), 176-191. https://doi.org/10.1061/(ASCE)0887-381X.

Ouarda, T. B. M. J. and C. Shu 2009. Regional low-flow frequency analysis using single and ensemble artificial neural networks. Water Resour. Res. 45(11), W11428. https://doi.org/10.1029/2008wr007196.

Rahman, A., Charron, C., Ouarda, T.B.M.J., Chebana, F., 2018. Development of regional flood frequency analysis techniques using generalized additive models for Australia. Stoch. Environ. Res. Risk Assess. 32(1), 123-139. https://doi.org/10.1007/s00477-017-1384-1.

Rao, A.R., Hamed, K.H., 2000. Flood Frequency Analysis. CRC Press, New York.

Rees, H.G., Holmes, M.G.R., Fry, M.J., Young, A.R., Pitson, D.G., Kansakar, S.R., 2006. An integrated water resource management tool for the Himalayan region. Environ. Model. Softw. 21(7), 1001-1012. https://doi.org/10.1016/j.envsoft.2005.05.002.

Rencher, A.C., Christensen, W.F., 2012. Methods of multivariate analysis, third ed. John Wiley & Sons, New York.

Requena, A.I., Ouarda, T.B.M.J., Chebana, F., 2018. Low-flow frequency analysis at ungauged sites based on regionally estimated streamflows. J. Hydrol. 563, 523-532. https://doi.org/10.1016/j.jhydrol.2018.06.016.

Requena, A.I., Ouarda, T.B.M.J., Chebana, F., 2017. Flood frequency analysis at ungauged sites based on regionally estimated streamflows. J. Hydrometeorol. 18(9), 2521-2539. https://doi.org/10.1175/jhm-d-16-0143.1.

Russell, D.S.O., 1992. Estimating flows from limited data. Can. J. Civ. Eng. 19(1), 51–58. https://doi.org/10.1139/l92-005.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6(2), 461-464. https://doi.org/10.2307/2958889.

Seidou, O., Ouarda, T.B.M.J., Barbet, M., Bruneau, P., Bobée, B., 2006. A parametric Bayesian combination of local and regional information in flood frequency analysis. Water Resour. Res. 42(11), W11408. https://doi.org/10.1029/2005wr004397.

Shu, C., Ouarda, T.B.M.J., 2007. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resour. Res. 43(7), W07438. https://doi.org/10.1029/2006WR005142.

Shu, C., Ouarda, T.B.M.J., 2012. Improved methods for daily streamflow estimates at ungauged sites. Water Resour. Res. 48(2), W02523. https://doi.org/10.1029/2011wr011501.

Smakhtin, V.U., 2001. Low flow hydrology: a review. J. Hydrol. 240, 147–186. https://doi.org/10.1016/S0022-1694(00)00340-1.

Smakhtin, V.U., Eriyagama, N., 2008. Developing a software package for global desktop assessment of environmental flows. Environ. Model. Softw. 23(12), 1396-1406. https://doi.org/10.1016/j.envsoft.2008.04.002.

Smith, W.H.F., Wessel, P., 1990. Gridding with continuous curvature splines in tension. Geophysics. 55(3), 293-305. https://doi.org/10.1190/1.1442837.

Stedinger, J.R., Tasker, G.D., 1985. Regional hydrologic analysis 1: ordinary, weighted and generalized least squares compared. Water Resour. Res. 21(9), 1421-1432. https://doi.org/10.1029/WR021i009p01421.

Tasker, G.D., 1980. Hydrologic regression with weighted least squares. Water Resour. Res. 16(6), 1107-1113. https://doi.org/10.1029/WR016i006p01107.

Thomas, D.M., Benson, M.A., 1970. Generalization of streamflow characteristics from drainage basin characteristics. USGS, Water-Supply Pap. No 1975.

Tsakiris, G., Nalbantis, I., Cavadias, G., 2011. Regionalization of low flows based on Canonical Correlation Analysis. Adv. Water Resour. 34(7), 865-872. http://dx.doi.org/10.1016/j.advwatres.2011.04.007.

USDA, 1986. Urban hydrology for small watersheds. Natural Resources Conservation Service. Tech Release 55.

Vieira, V., Webster, T., Weinberg, J., Aschengrau, A., 2009. Spatial analysis of bladder, kidney, and pancreatic cancer on upper Cape Cod: an application of generalized additive models to case-control data. Environ. Health. 8(1), 3. https://dx.doi.org/10.1186/1476-069X-8-3.

Vogel, R.M., Kroll, C.N., 1990. Generalized low flow frequency relationships for ungaged sites in Massachusetts. Water Resour. Bull. 26(2), 241-253. http://dx.doi.org/10.1111/j.1752-1688.1990.tb01367.x.

Vogel, R.M., Kroll, C.N., 1992. Regional geohydrologic-geomorphic relationships for the estimation of low flow statistics. Water Resour. Res. 28(9), 2451-2458. https://doi.org/10.1029/92wr01007.

Wahba, G., 1985. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Ann. Stat. 13(4), 1378-1402. https://doi.org/10.1214/aos/1176349743.

Wald, A., Wolfowitz, J., 1943. An exact test for randomness in the non-parametric case based on serial correlation. Ann. Math. Stat. 14(4), 378-388.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845.

Wazneh, H., Chebana, F., Ouarda, T.B.M.J., 2013. Depth-based regional index-flood model. Water Resour. Res. 49(12), 7957-7972. https://doi.org/10.1002/2013wr013523.

Wazneh, H., Chebana, F., Ouarda, T.B.M.J., 2016. Identification of hydrological neighborhoods for regional flood frequency analysis using statistical depth function. Adv. Water Resour. 94, 251-263. https://doi.org/10.1016/j.advwatres.2016.05.013.

Wen, L., Rogers, K., Saintilan, N., Ling, J., 2011. The influences of climate and hydrology on population dynamics of waterbirds in the lower Murrumbidgee River floodplains in Southeast Australia: Implications for environmental water management. Ecol. Model. 222(1), 154-163. https://doi.org/10.1016/j.ecolmodel.2010.09.016.

Wessel, P., Smith, W.H.F., Scharroo, R., Luis, J., Wobbe, F., 2013. Generic Mapping Tools: Improved Version Released. Eos Trans. Am. Geophys. Union. 94(45), 409-410. https://doi.org/10.1002/2013EO450001.

Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Am. Stat. Assoc. 99(467), 673-686. https://doi.org/10.1198/016214504000000980.

Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC Press, London.

Wood, S.N., 2008. Fast stable direct fitting and smoothness selection for generalized additive models. J. R. Stat. Soc. Ser. B-Stat. Methodol. 70, 495-518. https://doi.org/10.1111/j.1467-9868.2007.00646.x.

Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol. Model. 157(2–3), 157-177. https://doi.org/10.1016/S0304-3800(02)00193-X.

Wu, W., Dandy, G.C., Maier, H.R., 2014. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. Environ. Model. Softw. 54, 108-127. https://doi.org/10.1016/j.envsoft.2013.12.016.

Zalants, M.G., 1991. Low-flow frequency and flow duration of selected south Carolina streams through 1987. USGS Water-Resour. Investig. Rep. 91-4170.

Table 1. Descriptive statistics of the physiographic-meteorological variables and hydrological variables.

Variable	Unit	Notation	Mean	Median	Max	Min	CV	Skewness	Kurtosis
Catchment area	km^2	AREA	5646	1387	96600	0.69	2.07	4.53	29.72
Catchment mean slope	degree	MSLP	2.40	2.21	6.95	0.13	0.46	0.92	4.74
% occupied by lakes	%	PLAKE	6.33	4.00	32.00	0.00	1.04	1.32	4.32
% occupied by forest	%	PFOR	85.78	90.30	100.00	6.50	0.19	-2.24	8.68
Mean annual total precipitation	mm	PTMA	1018	1010	1520	646	0.17	0.64	3.94
Mean annual liquid precipitation (summer-fall)	mm	PLMS	465	460	664	306	0.17	0.36	2.79
Mean curve number	-	MCN	45.08	44.00	78.20	21.00	0.28	0.32	2.24
Mean number of days where the temperature is > 27 °C	day	NDH27	12.28	12.20	36.60	0.80	0.62	0.60	3.20
Mean annual degree-days < 0 °C	degree-day	DDBZ	1635	1428	2963	921	0.32	0.99	2.89
Mean annual degree-days > 13 °C	degree-day	DDH13	323	329	734	70	0.46	0.32	2.75
Latitude of the catchment centroid	°N	LAT	48.40	47.87	54.35	45.01	0.05	0.73	2.51
Longitude of the catchment centroid	°W	LONG	71.41	71.83	78.56	58.11	0.05	-0.93	3.97
Summer low-flow quantile of 30 days and 5-yr return period	m^3/s	$Q_{30,5}$	70.44	6.83	1280	0.0055	2.37	4.26	25.53
Summer low-flow quantile of 7 days and 2-yr return period	m^3/s	$Q_{7,2}$	85.62	7.38	1560	0.0044	2.38	4.27	25.80
Summer low-flow quantile of 7 days and 10-yr return period	m^3/s	$Q_{7,10}$	58.91	4.3	1080	0.0032	2.44	4.26	25.16
Winter low-flow quantile of 30 days and 5-yr return period	m^3/s	$Q_{30,5}$	26.46	6.2855	369	0.0044	2.10	4.00	21.49
Winter low-flow quantile of 7 days and 2-yr return period	m^3/s	$Q_{7,2}$	28.91	6.8585	406	0.0048	2.16	3.96	20.83
Winter low-flow quantile of 7 days and 10-yr return period	m^3/s	$Q_{7,10}$	22.85	4.705	341	0.0034	2.23	4.11	22.38

CV denotes the coefficient of variation.

Table 2. Pearson correlation coefficients between quantiles and physiographic-meteorological variables.

	Summer			Winter		
	$Q_{30,5}$	$Q_{7,2}$	$Q_{7,10}$	$Q_{30,5}$	$Q_{7,2}$	$Q_{7,10}$
AREA	**0.986**	**0.985**	**0.974**	**0.941**	**0.941**	**0.942**
MSLP	-0.103	-0.104	-0.103	-0.182	-0.168	-0.164
PLAKE	**0.531**	**0.541**	**0.530**	**0.587**	**0.584**	**0.583**
PFOR	-0.029	-0.031	-0.031	-0.071	-0.063	-0.064
PTMA	**-0.496**	**-0.495**	**-0.489**	**-0.487**	**-0.488**	**-0.484**
PLMS	**-0.432**	**-0.429**	**-0.426**	**-0.428**	**-0.427**	**-0.424**
MCN	**-0.203**	**-0.214**	**-0.212**	**-0.178**	**-0.188**	**-0.187**
NDH27	**-0.344**	**-0.341**	**-0.343**	**-0.309**	**-0.302**	**-0.299**
DDBZ	**0.575**	**0.572**	**0.566**	**0.557**	**0.558**	**0.556**
DDH13	**-0.403**	**-0.395**	**-0.394**	**-0.372**	**-0.370**	**-0.367**
LAT	**0.541**	**0.535**	**0.529**	**0.521**	**0.524**	**0.521**
LONG	-0.140	-0.156	-0.150	-0.187	-0.212	-0.214

Bold characters denote significant correlations at a level of 5%.

Table 3. Cross-validation results of all the regionalization methods for the summer low flows.

	Quantiles	HCA+MLR	ROI+MLR	CCA+MLR	HCA+GAM	ROI+GAM	CCA+GAM	ALL+MLR	ALL+GAM	SI
NASH	$Q_{30,5}$	0.936	0.921	0.925	0.967	0.958	0.954	0.907	0.937	**0.982**
	$Q_{7,2}$	0.892	0.935	0.931	0.970	0.968	0.938	0.914	0.923	**0.979**
	$Q_{7,10}$	0.875	0.895	0.903	0.955	0.960	0.964	0.883	0.917	**0.968**
BIAS	$Q_{30,5}$	1.48	2.03	-3.80	1.22	2.94	**0.87**	-4.48	-1.17	-3.33
(m^3/s)	$Q_{7,2}$	3.23	0.47	-4.68	0.98	1.45	1.89	-5.88	**0.33**	-4.46
	$Q_{7,10}$	1.76	1.79	-3.94	0.65	**-0.45**	-0.55	-4.22	-1.19	-3.70
RMSE	$Q_{30,5}$	42.26	46.87	45.67	30.50	34.40	36.04	50.87	41.86	**22.05**
(m^3/s)	$Q_{7,2}$	66.65	51.83	53.31	35.26	36.28	50.73	59.61	56.56	**29.76**
	$Q_{7,10}$	50.49	46.35	44.45	30.39	28.84	27.34	48.80	41.23	**25.68**
rBIAS	$Q_{30,5}$	8.46	**4.90**	8.87	5.04	5.58	9.72	8.55	8.21	14.26
(%)	$Q_{7,2}$	8.71	5.73	8.64	**3.08**	5.26	9.18	8.92	7.81	13.59
	$Q_{7,10}$	11.84	7.74	11.05	**5.61**	7.80	13.12	12.45	11.85	19.03
rRMSE	$Q_{30,5}$	47.12	**36.33**	43.89	36.82	37.05	45.74	45.76	45.88	59.84
(%)	$Q_{7,2}$	49.31	38.45	45.08	**33.04**	36.78	44.77	46.88	44.63	58.27
	$Q_{7,10}$	58.36	45.31	52.72	45.12	**45.11**	56.16	56.60	56.76	84.56

Best statistics are in bold characters.

Table 4. Cross-validation results of all the regionalization methods for the winter low flows.

	Quantiles	HCA+MLR	ROI+MLR	CCA+MLR	HCA+GAM	ROI+GAM	CCA+GAM	ALL+MLR	ALL+GAM	SI
NASH	$Q_{30,5}$	0.872	0.886	0.881	**0.925**	0.909	0.895	0.872	0.883	0.915
	$Q_{7,2}$	0.874	0.891	0.883	**0.947**	0.929	0.899	0.876	0.894	0.919
	$Q_{7,10}$	0.856	0.883	0.886	**0.912**	0.907	0.894	0.875	0.890	0.912
BIAS	$Q_{30,5}$	-0.83	-1.43	-1.04	-0.95	-1.75	-0.91	-3.11	**-0.32**	-0.73
(m^3/s)	$Q_{7,2}$	-1.09	-1.44	-1.41	-0.89	-1.58	-0.56	-3.39	**-0.55**	-0.87
	$Q_{7,10}$	-0.23	-0.87	-0.98	-0.86	-1.50	-0.48	-2.82	**0.13**	-0.79
RMSE	$Q_{30,5}$	19.81	18.77	19.12	**15.27**	16.81	18.05	19.82	19.03	16.16
(m^3/s)	$Q_{7,2}$	22.09	20.48	21.26	**14.42**	16.63	19.80	21.90	20.27	17.72
	$Q_{7,10}$	19.22	17.38	17.13	15.13	15.53	16.59	17.93	16.86	**15.08**
rBIAS	$Q_{30,5}$	5.56	0.93	6.18	1.01	**-0.20**	4.87	5.03	4.85	6.12
(%)	$Q_{7,2}$	4.70	0.74	5.77	0.92	**-0.34**	3.58	4.58	3.77	6.56
	$Q_{7,10}$	6.79	1.19	8.59	1.94	**1.09**	7.23	6.90	5.83	8.52
rRMSE	$Q_{30,5}$	37.01	27.94	32.81	**23.70**	24.19	34.07	34.20	32.54	29.75
(%)	$Q_{7,2}$	32.28	25.67	30.74	**21.37**	21.79	30.94	32.24	27.79	29.94
	$Q_{7,10}$	39.51	30.61	38.18	**27.36**	28.63	43.86	40.58	35.55	37.66

Best statistics are in bold characters.

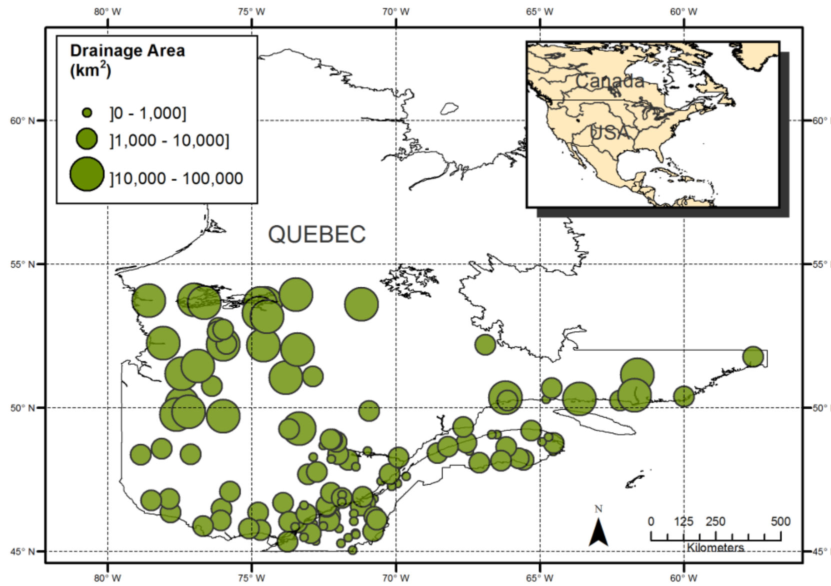

Fig. 1. Location of hydrometric stations across the province of Quebec (Canada).

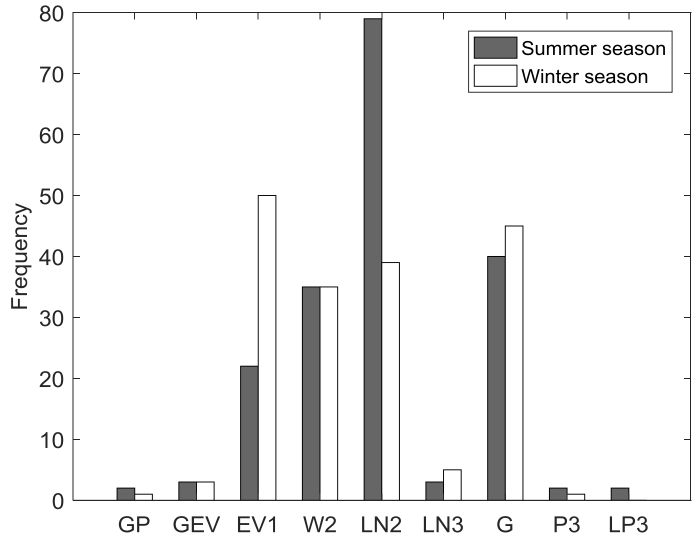

Fig. 2. Frequency with which different at-site distributions were selected for 7-day low flows.

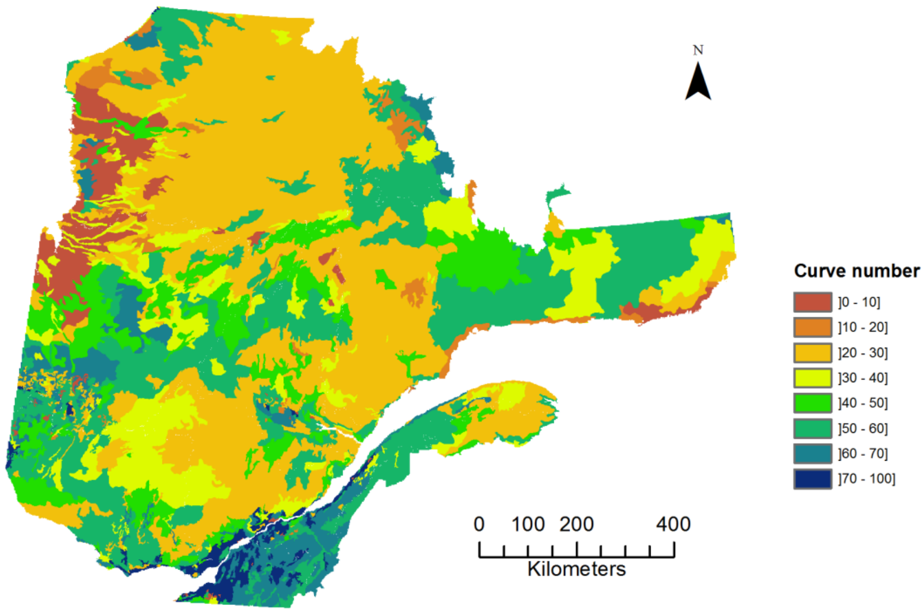

Fig. 3. Distribution of CN values within the study area.

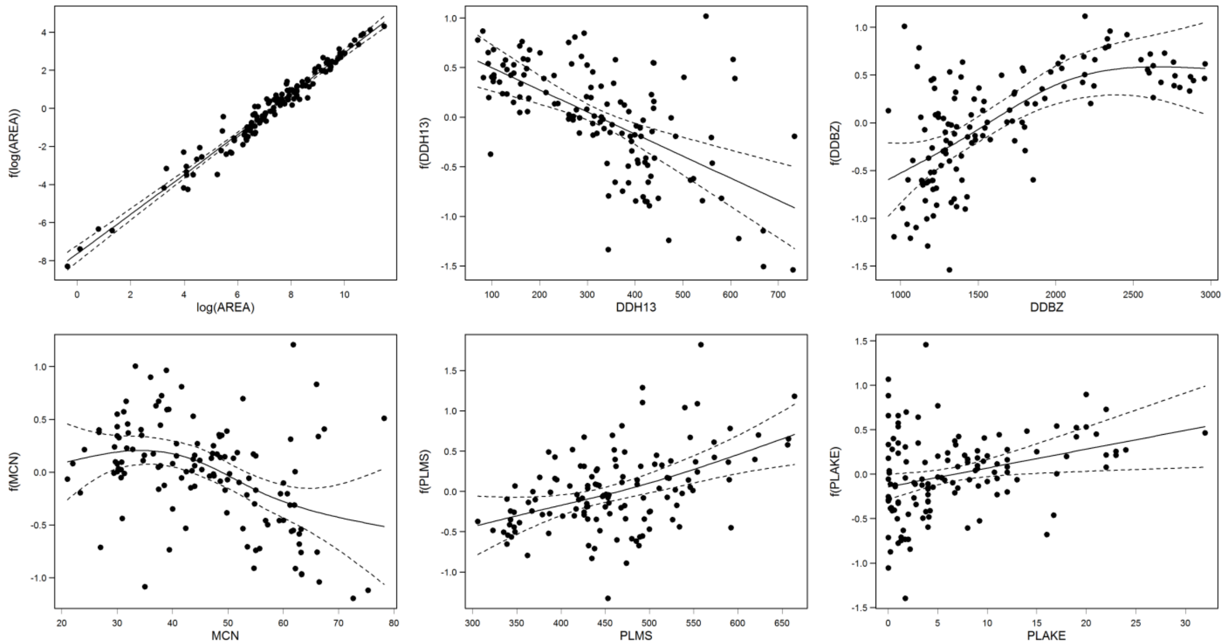

Fig. 4. Smooth functions of summer $Q_{7,10}$ for the explanatory variables. The dashed lines represent the 95% confidence intervals and dots are the residuals.

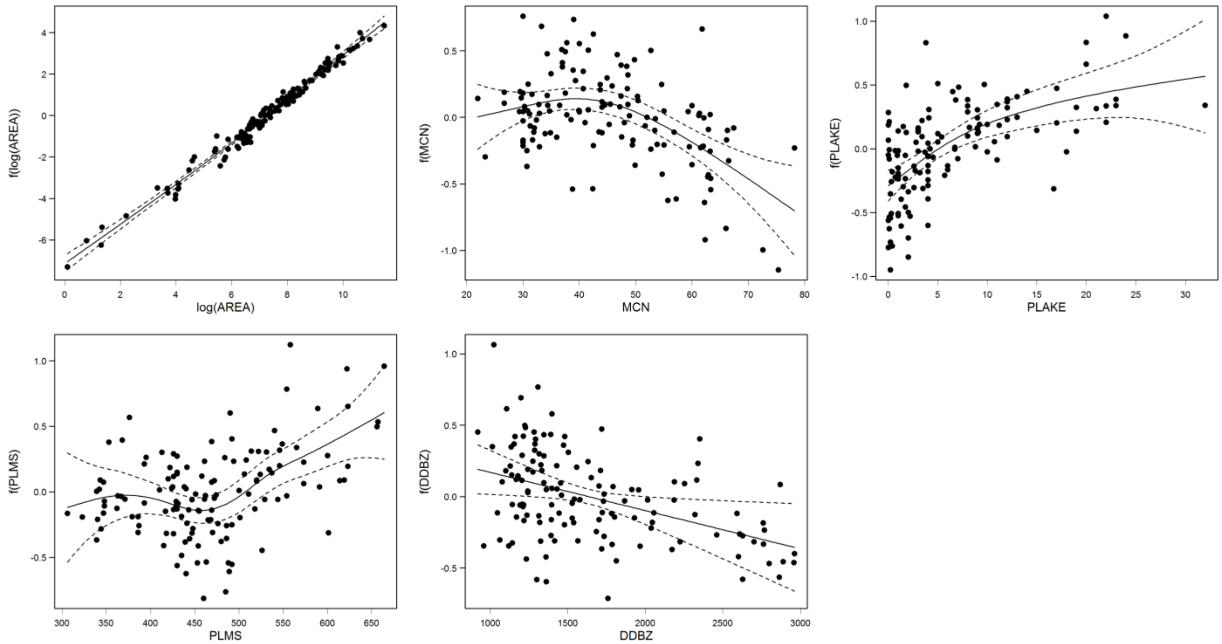

Fig. 5. Smooth functions of winter $Q_{7,10}$ for the explanatory variables. The dashed lines represent the 95% confidence intervals and dots are the residuals.

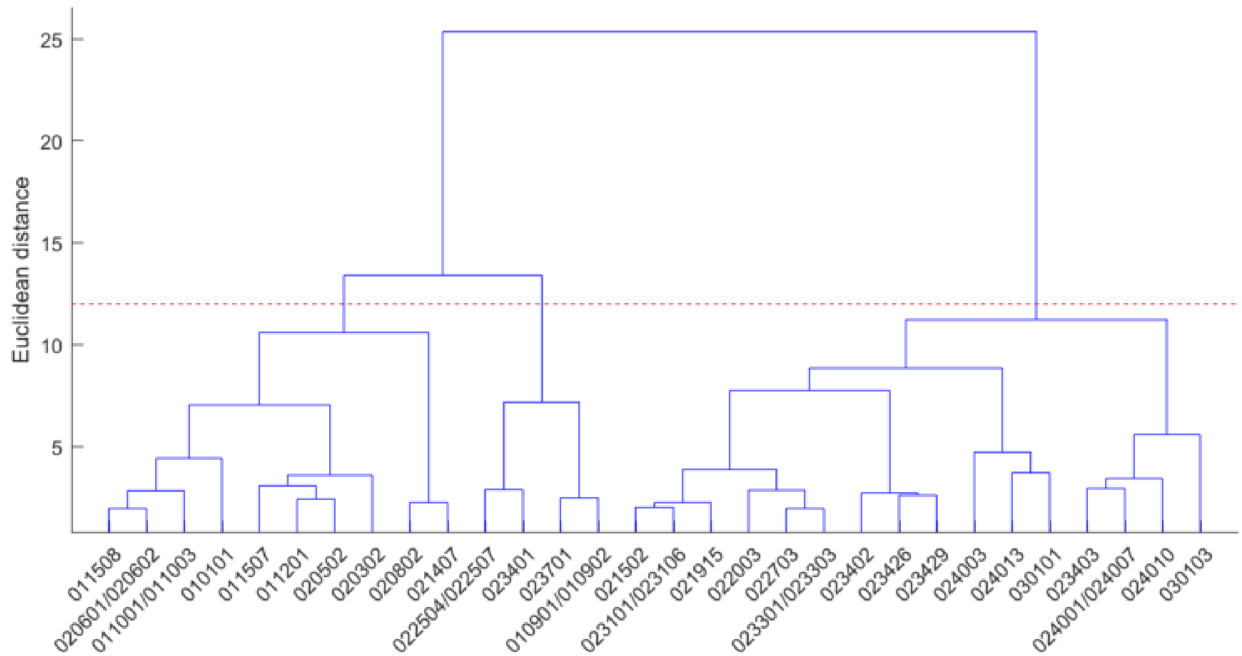

Fig. 6. Dendrogram corresponding to hierarchical clustering for summer low flows for which only 30 leaf nodes are presented. The red line indicates the cut-off distance.

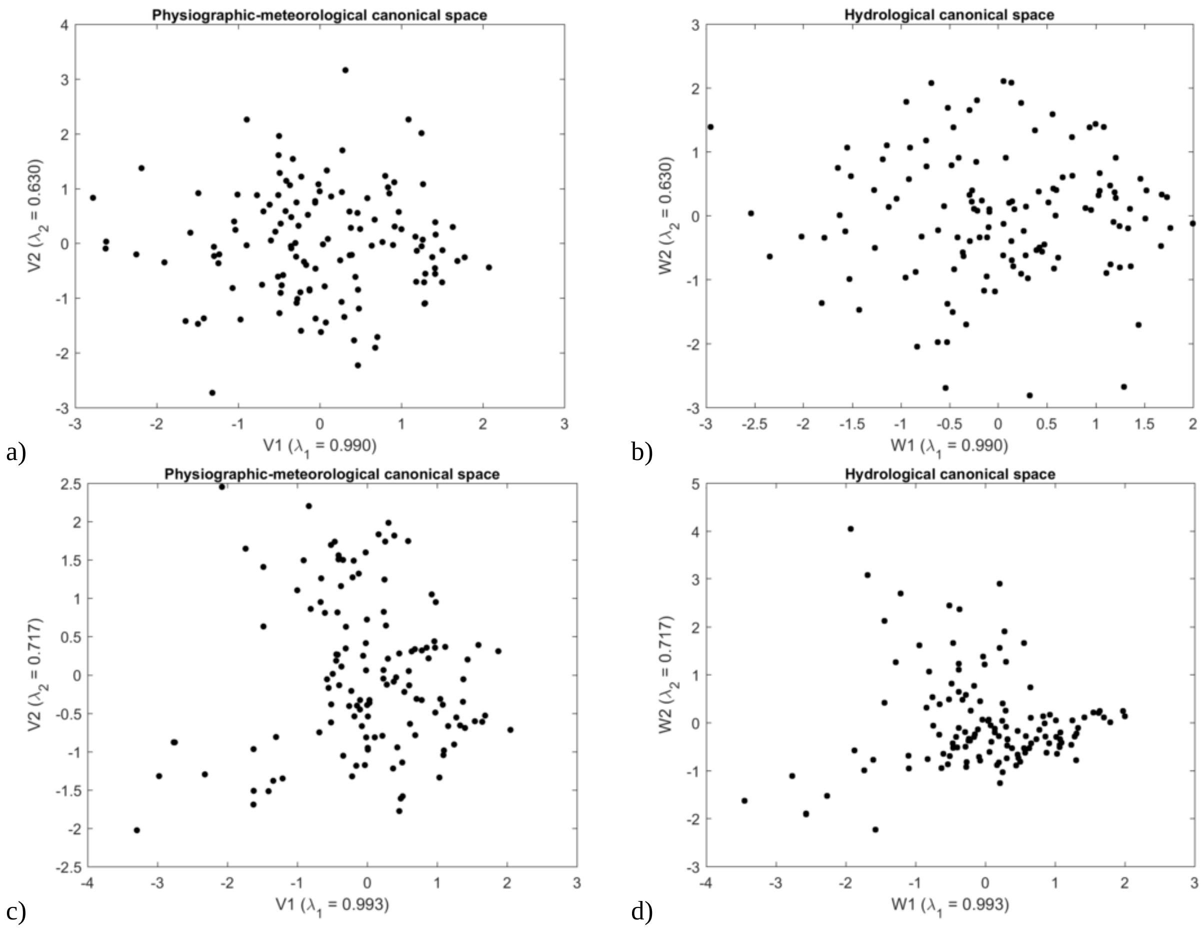

Fig. 7. The physiographic-meteorological canonical space and the hydrological canonical space for the summer season (a and b) and for the winter season (c and d).

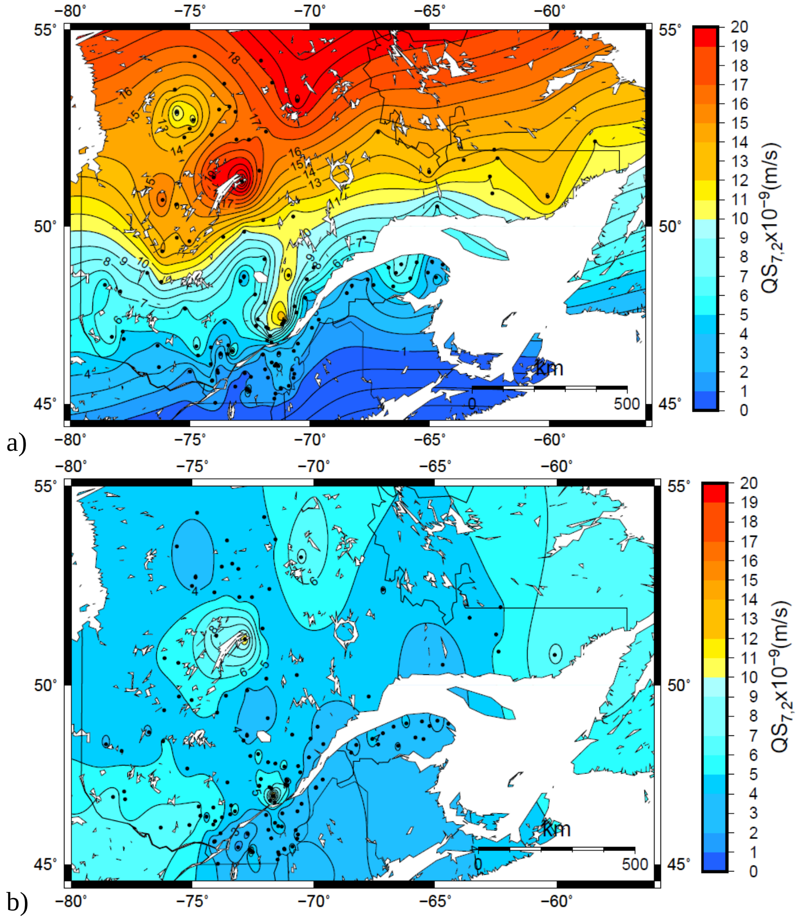

Fig. 8. Contour maps of specific quantiles of $Q_{7,2}$ ($QS_{7,2}$) in the province of Quebec using the method SI for (a) summer low flows and (b) winter low flows. Basin centroids coordinates are represented with dots.

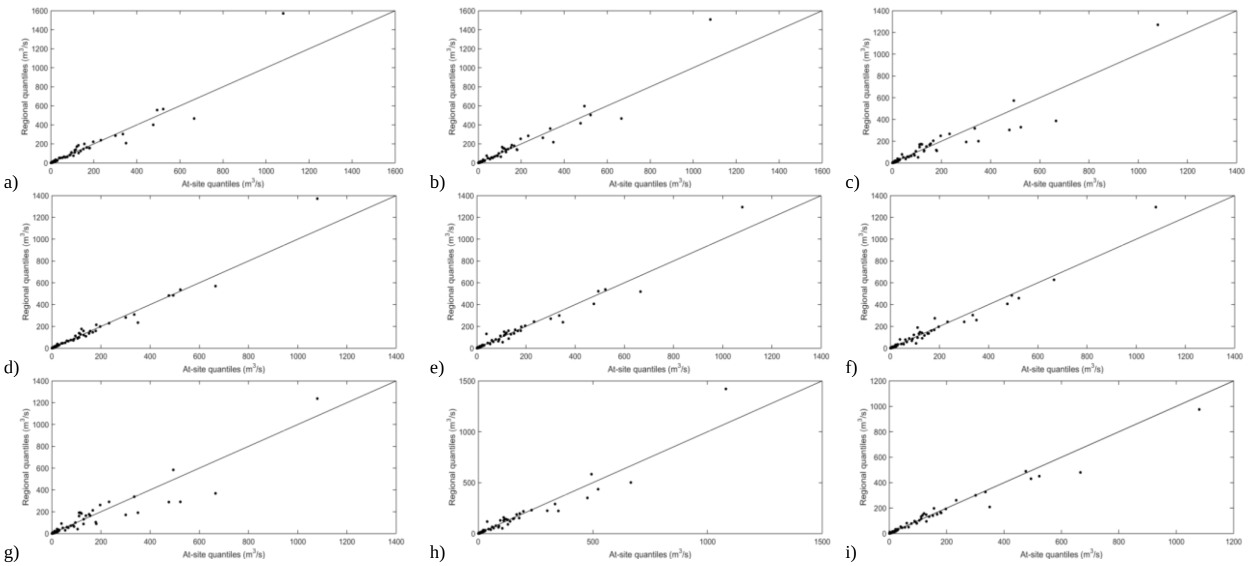

Fig. 9. Regional versus at-site quantiles $Q_{7,10}$ for summer low flows. a) HCA+MLR, b) ROI+MLR, c) CCA+MLR, d) HCA+GAM, e) ROI+GAM, f) CCA+GAM, g) ALL+MLR, h) ALL+GAM and i) SI.

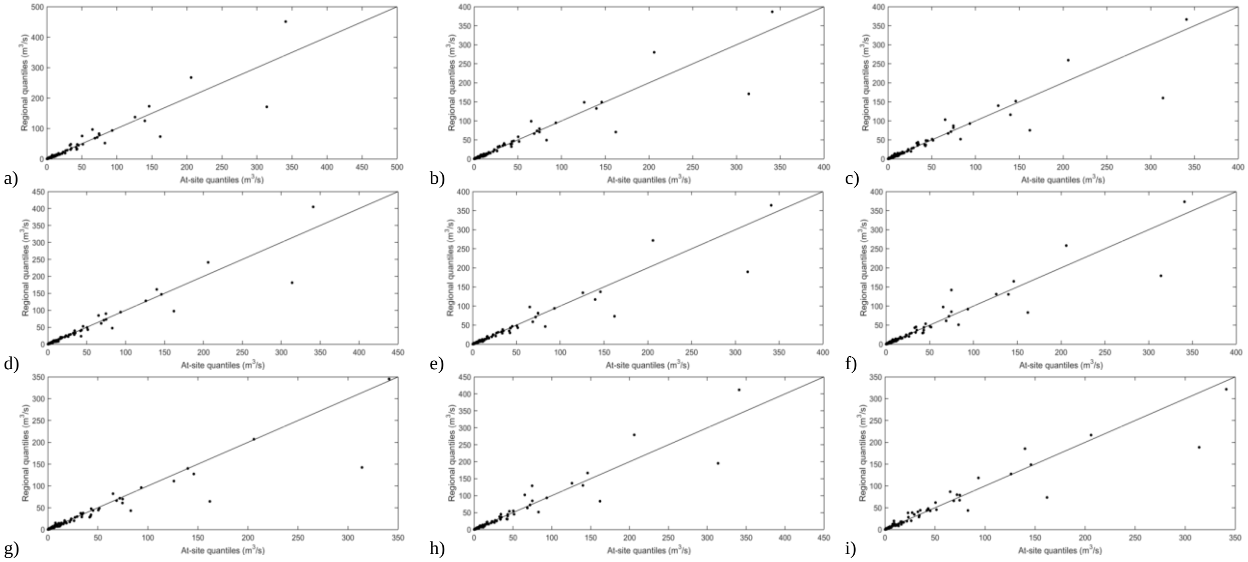

Fig. 10. Regional versus at-site quantiles $Q_{7,10}$ for winter low flows. a) HCA+MLR, b) ROI+MLR, c) CCA+MLR, d) HCA+GAM, e) ROI+GAM, f) CCA+GAM, g) ALL+MLR, h) ALL+GAM and i) SI.

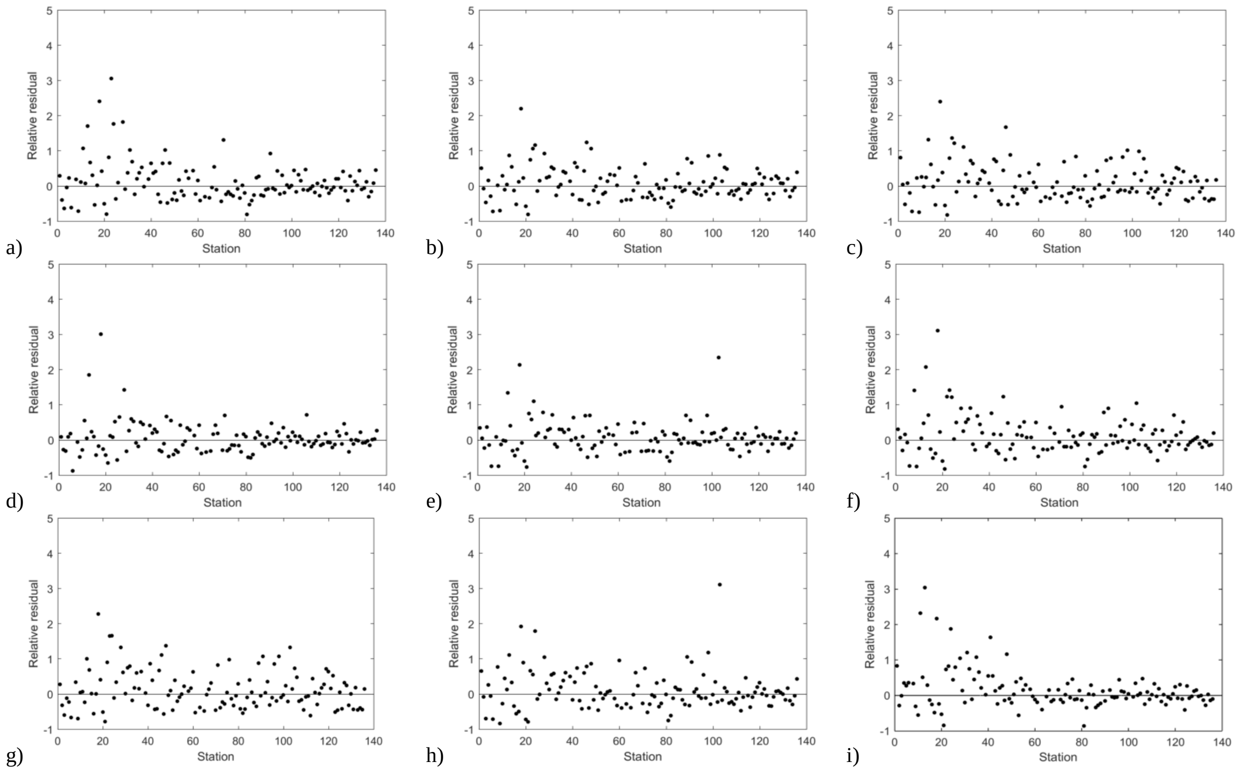

Fig. 11. Relative residuals for summer low-flow quantiles $Q_{7,10}$. a) HCA+MLR, b) ROI+MLR, c) CCA+MLR, d) HCA+GAM, e) ROI+GAM, f) CCA+GAM, g) ALL+MLR, h) ALL+GAM and i) SI. Stations are sorted from the one with the smallest catchment area to the one with the largest catchment area.

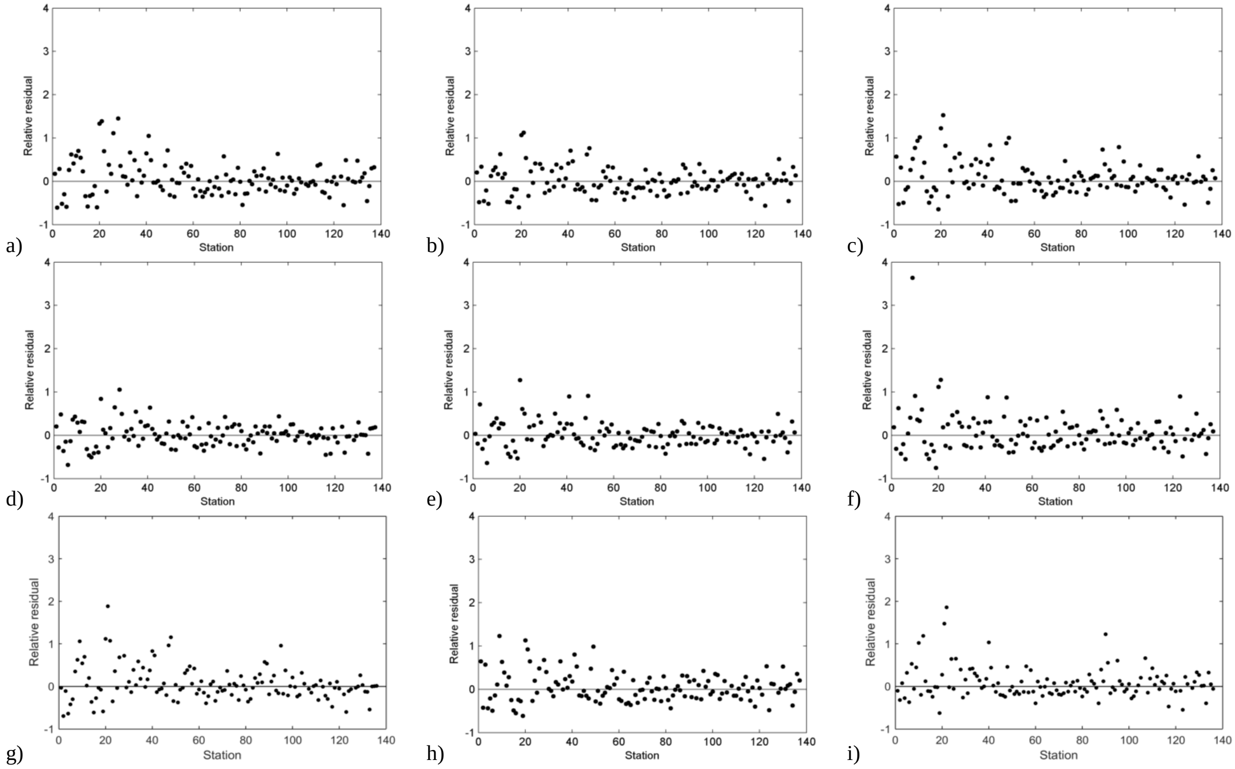

Fig. 12. Relative residuals for winter low-flow quantiles $Q_{7,10}$. a) HCA+MLR, b) ROI+MLR, c) CCA+MLR, d) HCA+GAM, e) ROI+GAM, f) CCA+GAM, g) ALL+MLR, h) ALL+GAM and i) SI. Stations are sorted from the one with the smallest catchment area to the one with the largest catchment area.