Flood frequency analysis at ungauged catchments with the GAM and MARS approaches in the Montreal region, Canada

Amina Msilini*[1], Christian Charron[1], Taha B.M.J. Ouarda[1] and Pierre Masselot[2]

[1] Canada Research Chair in Statistical Hydro-Climatology, Institut national de la recherche scientifique, Centre Eau Terre Environnement, 490 de la Couronne, Québec, QC, G1K 9A9, Canada.

[2] London School of Hygiene & Tropical Medicine (LSHTM), Keppel Street London, WC1E 7HT, United Kingdom.

*Corresponding author: Amina Msilini (Amina.Msilini@ete.inrs.ca; amina.msilini.m@gmail.com).

January, 2022

Flood frequency analysis at ungauged catchments with the GAM and MARS approaches in the Montreal region, Canada

Abstract

Regional frequency analysis (RFA) aims to estimate quantiles of extreme hydrological variables (e.g. floods or low-flows) at sites where little or no hydrological data is available. This information is of interest for the optimal planning and management of water resources. A number of regional estimation models are evaluated and compared in this study and then used for regional estimation of flood quantiles at ungauged catchments located in the Montreal region in southern Quebec, Canada. In this study, two neighborhood approaches using canonical correlation analysis (CCA) and the region of influence (ROI) method are applied to delineate homogenous regions. Three regression methods namely log-linear regression model (LLRM), generalized additive models (GAM), and multivariate adaptive regression splines (MARS), recently introduced in the RFA context, are considered for regional estimation. These models are also applied considering all stations (ALL). The considered models, especially MARS, have never been used previously in a concrete application. Results indicate that MARS and GAM have comparable predictive performances, especially when applied with the whole dataset. Results also show that MARS used in combination with the CCA approach provide improved performances compared to all considered regional approaches. This may reflect the flexibility of the combination of these two approaches, their robustness, and their ability to better reproduce the hydrological phenomena, especially in real-world conditions when limited data are available.

Résumé

L'analyse fréquentielle régionale (AFR) vise à estimer les quantiles de variables hydrologiques extrêmes (par exemple, les crues ou les étiages) sur des sites avec peu ou aucune information hydrologique de disponible. Ces informations sont intéressantes pour la planification et la gestion optimales des ressources en eau. Un certain nombre de modèles d'estimation régionale ont été évalués et comparés dans cette étude, puis utilisés pour l'estimation régionale des quantiles de crue dans des bassins versants non jaugés situés dans la région de Montréal dans le sud du Québec, Canada. Dans cette étude, deux approches d'identification de voisinage utilisant l'analyse canonique de corrélation (CCA) et la méthode de la région d'influence (ROI) sont appliquées pour délimiter des régions homogènes. Trois méthodes de régression, à savoir le modèle de régression log-linéaire (LLRM), les modèles additifs généralisés (GAM) et la régression multivariée par spline adaptative (MARS), récemment introduite dans le contexte de l'AFR, sont prises en compte pour l'estimation régionale. Ces modèles sont également appliqués en considérant toutes les stations (ALL). Les modèles considérés, en particulier MARS, n'ont jamais été utilisés auparavant dans une application concrète. Les résultats indiquent que MARS et GAM ont des performances prédictives comparables, en particulier lorsqu'ils sont appliqués à l'ensemble de la base de données. Les résultats montrent également que MARS utilisé en combinaison avec l'approche de CCA offre de meilleures performances par rapport à toutes les approches régionales considérées. Cela peut refléter la flexibilité de la combinaison de ces deux approches, leur robustesse et leur capacité à mieux reproduire les phénomènes hydrologiques, en particulier dans des conditions réelles lorsque des données limitées sont disponibles.

Keywords: Multivariate adaptive regression splines; Generalized additive models; Montreal region (Canada); Ungauged basin; Regional frequency analysis; Drainage network characteristics.

1. Introduction

Knowledge of the frequency and the magnitude of extreme hydrological events (e.g. floods and low-flows) are of interest for water resources management and hydrological design. Estimation of extreme flows is often required at sites where little or no hydrological data is available. To this end, regional frequency analysis (RFA) approaches are commonly used to estimate and assess extreme hydrological event characteristics. Generally, RFA includes two main steps: i) delineation of homogenous regions (DHR) to group gauged sites with hydrological behavior similar to the target one and ii) regional estimation (RE) to transfer the information from gauged sites to the target one within the same homogeneous region (e.g. Chebana and Ouarda 2008). Various methods have been suggested and documented for each of these two steps (e.g. Ouarda 2016). In practice, two DHR methods are often considered, namely : the region of influence (ROI) (Burn 1990) and the canonical correlation analysis (CCA) (Ouarda et al. 2001). The geographical proximity of the catchments has also long been recognized and considered in RFA to group sites with similar characteristics (Han et al. 2020). It is especially convenient for practical purposes.

For the RE step, two main approaches have commonly been used to regionalize flood characteristics. The first one includes regression-based approaches, where log-linear regression models (LLRM) are the most used because of their simplicity and good predictive performances. The second approach includes the index-flood models (Dalrymple 1960), where it is assumed that in a given homogeneous region, all local data normalized by a central position indicator (e.g. mean or median) have the same distribution.

Hydrological processes represent complex and nonlinear natural phenomena (Xu et al. 2010). They depend on a large number of interactive physio-meteorological catchment attributes such as the climate of the region, the topographic variability of the catchments, their soil characteristics, and their geological formations. The log-linear method commonly used in the RE step assumes that the relation between the response variable and the explanatory variables is linear. This assumption is generally not satisfied in such complex non-linear processes. To deal with the natural complexity of the hydrological events and account for the presence of non-linearity between the explanatory and the response variables, a number of non-linear approaches have been suggested in the literature such as the artificial neural networks (ANNs) and the Generalized Additive Models (GAMs) (e.g. Khalil, Ouarda, and St-Hilaire 2011; Ouarda et al. 2018). The use of ANNs to regionalize extreme hydrological characteristics has become increasingly popular (Ouarda and Shu 2009). However, it presents a major drawback which is the tendency to overfit the data (e.g. Gal and Ghahramani 2016; Lawrence and Giles 2000). Furthermore, their calibration is somewhat a complex task that requires some subjective choices.

The use of GAM has also become increasingly popular in a number of fields such as hydro-climatology and environmental modelling (e.g. Wen et al. 2011; Rahman et al. 2018), public health (e.g. Leitte et al. 2009; Bayentin et al. 2010), renewable energy assessment (e.g Ouarda et al. 2016) and hydrology (e.g. Rahman et al. 2018; López-Moreno and Nogués-Bravo 2005). It has been recently introduced in the RFA context by Chebana et al. (2014), where the authors found that GAM performs better than the classical linear regression model. However, the method can be computationally intensive and difficult to fit to high-dimensional databases (large number of explanatory variables).

The reliability of the regional flood characteristic estimates depends strongly on the amount of available gauged sites data used in the regional estimation. In practice, it is often the case that rivers are poorly monitored and/or they have a short time series. Msilini et al. (2020) suggested that it may be possible to perform a reliable regional estimation with MARS even using a few data in the RFA context. The application of MARS in a real case study has never been performed.

The aim of the present paper is to develop and test a number of approaches listed above in a practical real-world case study, with limited number of stations, consisting in the estimation of flood quantiles at 11 ungauged sites of interest in the Montreal region (Canada). Such quantiles are essential for the municipality to established flood maps within the region. The catchments of the considered study region are often of small areas and they are characterized by their high urbanized and agricultural areas which allow for a very high runoff. Moreover, the hydrological response in the Montreal catchments is known to have a higher degree of variability and non-linearity. Hence, the adoption of non-linear RE models, especially, MARS in predicting flood discharge at ungauged catchments in such conditions may be relevant.

In this study, the LLRM, GAM and MARS models are used in conjunction with/and without the delineation of homogeneous region methods (ROI and CCA). Calibrations of the regional models are performed with catchments located within a radius of 250 kilometres around the target area to ensure certain similarity between their catchment characteristics. The performances of the different approaches are compared and the best identified models are used to predict flood quantiles at the 11 target ungauged sites.

This paper is organized as follows. Section 2 presents a brief theoretical background of the different RFA approaches adopted in this work. The considered methodology is discussed in section 3. The case study and the dataset are described in section 4. The obtained results are illustrated and discussed in section 5. Finally, the conclusions of the study are summarized in section 6.

2. Theoretical background

2.1 Delineation of homogeneous region approaches

2.1.1 Canonical correlation analysis (CCA)

CCA is a technique commonly used to identify the possible correlations between two groups of random variables. Let $X=(X_1,X_2,...,X_r)$ and $Y=(Y_1,Y_2,...,Y_s)$ be sets of random variables of respectively r physio-meteorological variables and s hydrological variables of n gauged sites. CCA allows identifying the dominant linear modes of covariability between the vectors X and Y so that it is possible to do inference about Y knowing X. Let V_i and W_i be linear combinations (called canonical variables) of the sets X and Y, i.e.:

$$V_i = A_{i1}X_1 + A_{i2}X_2 + \ldots + A_{ir}X_r \quad (1)$$

$$W_i = B_{i1}Y_1 + B_2Y_2 + \ldots + B_{is}Y_s \quad (2)$$

where $i = 1,\ldots,d$ and $d = \min(r, s)$. CCA allows for the identification of vectors A and B in such a way that the correlation coefficients between the canonical variables, i.e. $\lambda_i = \mathrm{corr}(V_i, W_j)$ where $i = j$, is maximized and $\mathrm{corr}(V_i, W_j) = 0$ where $i \neq j$ under constraints of unit variance.

In the RFA, the hydrological neighborhood for a given target ungauged site at $100(1-\alpha)\%$ confidence level is defined by the set of K sites such that the canonical hydrological score w_k, $k = 1, \ldots, K$, is close to the canonical physio-meteorological score of the target site v_0. The closeness is measured using a Mahalanobis distance calculated between the hydrological mean position of the site of interest Λv_0 and the positions of other sites w_k such that :

$$(W - \Lambda V_0)^T (I_d - \Lambda^2)^{-1} (W - \Lambda V_0) \leq \chi^2_{\alpha,d} \qquad (3)$$

where $\chi^2_{\alpha,d}$ is defined such that Prob $(\chi^2 \leq \chi^2_{\alpha,d}) = 1-\alpha$, I_d is the d×d identity matrix and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. For more details the reader is referred to Ouarda et al. (2001).

2.1.2 Region of influence (ROI)

The ROI approach was introduced by Burn (1990). As the CCA technique, the ROI can be used in the RFA to identify the neighborhood of a given target site. In this method, the identification of the neighborhood is carried out based on the similitude between catchment characteristics. The similitude is measured based on the Euclidean distance calculated in the physio-meteorological space (e.g. Burn 1990; Tasker, Hodge, and Barks 1996) i.e.:

$$\mathrm{ROI_i}= \left\{ \text{sites } j \in (1,\ldots,n);\; D_{ij}= [\sum_{k=1}^{r} W_k\, (X_{k,i}\text{-}X_{k,j})^2]^{\frac{1}{2}} \leq \Theta \right\} \tag{4}$$

where D_{ij} is the weighted Euclidean distance between the target site i and the gauged one, j = 1,…, n, $X_{k,j}$ (k = 1,…, r) is the standardized value of the k^{th} physio-meteorological variable at site j, W_k is the weight associated with the k^{th} physio-meteorological variable, and Ɵ represents the threshold value. For more details, the reader is referred to (e.g. Burn 1990; GREHYS 1996).

2.2 Regional estimation approaches

2.2.1 Log Linear Regression Model (LLRM)

The log-linear regression model (LLRM) is one of the most common regional estimation models. It consists in establishing a linear relationship between the hydrological variable Y and the physio-meteorological characteristics of a given catchment (X_1, X_2, …, X_m) (e.g. Pandey and Nguyen, 1999) :

$$\log\left(E(Y/X)\right)= \beta_0+\sum_{j=1}^{m} \beta_j \log\left(X_j\right)+ \varepsilon \tag{5}$$

Where X is a matrix whose columns correspond to a set of m explanatory variables, β_0 and β_j are unknown parameters to be estimated using the least-squares method, and ε is the model error.

2.2.2 Generalized Additive Model (GAM)

GAM (Hastie and Tibshirani 1987) is a non-linear model that is able to model a large variety of nonlinear relationships and it allows to consider non-Gaussian response variables (Wood 2006). This model uses flexible non-linear smooth functions to model the response variable (i.e. the hydrological variable). A GAM can then be defined as (Wood 2006):

$$g\left(E(Y/X)\right)= \alpha+ \sum_{j=1}^{m} f_j\left(X_j\right)+ \varepsilon \tag{6}$$

where g is a monotonic link function, X is a matrix whose columns correspond to a set of m explanatory variables, and f_j are smooth functions giving the relationship between the explanatory variables X_j and the response variable Y, α is the intercept and ε is the error term. Because of the additive property of GAM, one can separately analyze the impact of each explanatory variable on the response variable.

The smooth non-linear functions f_j are expressed as:

$$f_j(X)= \sum_{i=1}^{q} \beta_{ji}\, b_{ji}(X) \tag{7}$$

where β_{ji} are parameters to be estimated and b_{ji} are the spline basis functions. Further information on GAM can be found in Wood (2006) and Wood (2017).

2.2.3 Multivariate adaptive regression splines (MARS)

Friedman (1991) introduced MARS as a flexible non-parametric regression approach able to model complex and non-linear relationship often hidden in high-dimensional data. The MARS model f(X) can be defined as a linear combination of basis functions and their interactions as:

$$f(X) = \beta_0 + \sum_{n=1}^{r} \beta_n B_n(X) \qquad (8)$$

where β_0 is the intercept, and β_n are regression coefficients of the basis functions ($B_n(X)$).

Three forms can be taken by the $B_n(X)$ terms in the MARS model: i) a constant term which represents the intercept, ii) a linear spline functions on a given variable X_j namely hinge function ($h_m(X_j) = (t_m - X_j)_+$ or $h_m(X_j) = (X_j - t_m)_+$ where t is a knot) or iii) a product of two or more $h_m(X_j)$ which represents the interaction between the variables. The $B_n(X)$ are defined in pairs of $h_m(X_j)$ and are separated by a knot between the range of a given variable.

MARS algorithm builds a model in two main steps: the first step is the forward pass where the model starts with the intercept and iteratively adds the B_ns. At each time, the most significant variable and knot yielding the largest decrease in the error of the model are chosen. This step results in a large model that usually overfits the data. The second step is the backward pass which allows improving the predictive performance of the built model by deleting the less significant B_ns. This later step continues until obtaining the best sub-models having the lowest Generalized Cross Validation (GCV) score. For more details, the reader is referred to Msilini, Masselot, and Ouarda (2020).

3. Methodology

3.1 Regional models

In this work, two methods for neighborhood identification (CCA and ROI) are applied in combination with the LLRM, GAM and MARS for regional estimation. Three other approaches are also assessed by applying the LLRM, GAM and MARS using all stations (ALL). Table 1 summarizes the used combinations.

The CCA and the ROI techniques are applied in the DHR step to improve the degree of homogeneity, and hence the accuracy of the predictions of the RE models. For these methods, the relevant variables in terms of explaining the flooding process need to be identified. In this work, the appropriate variables selected for the LLRM with a stepwise procedure approach are adopted in each of the neighborhood methods such as in Ouarda et al. (2018). Then, the optimal number of sites in the neighborhood (optimum threshold distance) is identified based on a jackknife procedure. This distance is the one that minimizes a given performance criterion of the log-linear model applied in each neighborhood.

GAM is fitted using the R package mgcv (Wood 2006). The thin plate regression spline is considered in this study as a basis in the smoothing function. The adopted link function is the identity function because of the approximately Gaussian log-transformed quantiles (see Chebana et al. (2014), for instance).

MARS is built using the R package earth (Milborrow 2018). To this end, three main parameters need to be tuned: the maximum number of terms to be reached in the model in the forward phase (N_k), the degree of interaction between the variables (degree) which allows including interaction terms between multiple hinge functions when its value is greater than 1, and the maximum number of terms to be retained after the backward phase (N_{prune}). These parameters are optimized based on the GCV, the residual sum of squares (RSS) and the coefficient of determination (R^2) criteria of the fitted models. Imposing termination conditions for the forward pass is necessary to save calculation time and to avoid the generation of terms with arbitrary knots. This allows optimizing the model more efficiently. In this study, the parameter N_k is optimized to avoid that the final model includes a large number of variables. This may allow obtaining more reliable estimates within the neighborhood.

For each regional model, different sets of physio-meteorological variables are considered. A backward stepwise technique is used in this work to select the most significant explanatory variables for each RE models (LLRM, GAM and MARS). The presentation of this approach is given in the next section.

3.2 Variable selection

The backward stepwise selection procedure is used in this study to identify the optimal combination of explanatory variables as in Ouarda et al. (2018). This technique consists in removing iteratively the least significant variable from an initial full model containing all available variables. At each step, the deleted variable is the one associated with the highest *p*-value for the null hypothesis that the coefficients β_j in Eq. (5) (for the LLRM) and the smooth terms (for GAM) are null. In the case of MARS, the removed variables are those yielding to the most significant decrease in the GCV score. For the aim of simplicity, the predictor variables selected with the backward stepwise regression approach applied to the quantile associated to the smallest return period are considered as predictor variables to estimate the other quantiles. Ouarda et al. (2018) suggested that the quantile with the smallest return period can be considered as the most reliable quantile.

3.3 Validation

The performances of each considered RFA combination are assessed using a jackknife procedure. This method consists in considering, in turn, each gauged site as the target site and performs RE. This process is repeated for each gauged site. Then, the regional estimate is compared to its corresponding observed value. Based on the jackknife procedure, a number of standard performance criteria can be used to evaluate the prediction power of each regional model:

Nash- Sutcliffe Efficiency index:

$$\mathrm{NASH} = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{9}$$

Root-mean-square error :

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{10}$$

Relative root-mean-square error :

$$\mathrm{RRMSE} = 100\sqrt{\frac{1}{N}\sum_{i=1}^{N} \left[\frac{(y_i - \hat{y}_i)}{y_i}\right]^2} \tag{11}$$

Mean bias :

$$\mathrm{BIAS} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i) \tag{12}$$

Relative mean bias :

$$\text{RBIAS} = 100\ \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i - \hat{y}_i)}{y_i} \tag{13}$$

where y_i and $\hat{y}_i$ are, respectively, the local and regional quantile estimates at site i, $\bar{y}$ is the mean of the local quantile estimates, and N is the number of stations.

Based on the computed performance criteria, the best models can be identified and then used to make predictions in the ungauged sites of the study case.

4. Case study and datasets

Considering a number of physio-meteorological variables (Table 2), the considered regional approaches are applied to a group of hydrometric stations located in the southern part of Quebec (Canada) within a radius of 250 kilometres around the city of Montreal (Figure 1). The objective is to estimate the specific flood quantiles QS_T (with T = 10, 50 and 100 years) for the spring season (January-June) at ungauged sites. The considered region is characterized by its low number of hydrometric stations.

In this study, we focus on the spring season because maximum annual floods in the study area often occur on this season. Figure 2 illustrates the variation of the annual mean of the day's indices associated to the maximum annual flow as a function of the sites. It can be seen that annual floods occur generally during the spring season and especially between the April and May months, hence the choice to focus on this season.

The hydrological variables are calculated from daily flows acquired by the Quebec Water Expertise Center (CEHQ) available at (https://www.cehq.gouv.qc.ca/hydrometrie/historique_donnees/default.asp). Considering a number of selection criteria such as the minimum size of the sample series at the station (15 years), their monitoring levels (proximity to a natural regime with a maximum of an influence on a daily basis) and their geographical proximity to the target stations, 63 hydrometric stations are retained for the estimation of the local quantiles.

A local frequency analysis (FA) is carried out in each gauged site. This involves the verification of the basic assumptions (independence and stationarity) and the identification of the adequate distributions. The distributions that are found to best fit the observed data are essentially the two-parameter distribution functions such as gamma, Weibull and the log normal. Finally, 57 stations are retained for the analysis of the QS_T for the spring season.

The physio-meteorological variables used in this study come from widely validated and used dataset covering the South of the province of Quebec (e.g. Shu and Ouarda 2007; Durocher, Chebana, and Ouarda 2015; Wazneh, Chebana, and Ouarda 2016; Ouali, Chebana, and Ouarda 2016) and are given in Table 2. The characteristics of catchments corresponding to each gauged station are computed using the ArcHydro and HecGeoHms tools implemented in the ArcGIS environment. These tools comprise functionalities for catchment delineation and drainage network extraction from Digital Elevation Models (DEMs). The DEMs used here are obtained from the Natural Resources Canada database (https://www.nrcan.gc.ca/earth-sciences/geography/topographic-information/download-directory-documentation/17215) distributed with a spatial resolution of ~ 20 m grid cells. The DEMs of the United States Geological Survey (USGS) (https://earthexplorer.usgs.gov/) are used for the cross-border catchments. These data have a spatial resolution of ~ 30 m grid cells.

The catchment limit features are used to calculate the spatial average of the physio-meteorological variables. The variables characterizing the drainage network systems are extracted using the D8 method (O'Callaghan and Mark 1984; Jenson and Domingue 1988). The variables related to the land cover, are calculated based on the digital maps of Quebec also available in the Natural Resources Canada database. The meteorological variables are computed using spatial interpolation of the meteorological data of the Ministry of the Environment and the Fight against Climate Change (MELCC). The meteorological stations which retained in this study had at least 15 years of data. The universal kriging method (Isaaks and Srivastava 1989) is used in this work for the spatial interpolation of the meteorological data. This technique gave the most accurate predictions based on a cross validation method. Descriptive characteristics of the considered hydrological and physio-meteorological variables are summarized in Table 3. It should be noted that, in this work a specific RT (RT standardized by basin area) is used to eliminate the scale effect as RT is a variable that is highly correlated with the basin area.

5. Results and discussion

5.1 Delineation with CCA and ROI

The CCA and the ROI approaches are applied in this study in the DHR using a set of explanatory variables selected by a stepwise procedure. Given the complexity of GAM and the small number of stations, 6 variables are used to model the spring flood quantiles and 3 knots are considered in this model in the smooth functions. Based on these parameters, the optimal threshold distance for the CCA and the ROI neighborhoods is fixed at 5×10^{-6} and 6, respectively.

CCA requires the normality of the hydrological and physio-meteorological variables. To achieve normality, some variables need to be transformed. The normality of each variable is visually evaluated with a normal probability plot. This technique plots empirical quantiles versus theoretical Gaussian quantiles and the plot should be approximately linear in the case of actual normality. Visual inspection of transformed variables indicates that the logarithmic transformation is applied to the flood quantiles, RT and MBS and a square root transformation is used for PLAKE.

5.2 Selection of optimal explanatory variables

To avoid overfitting and optimise the predictive power of the methods, we perform variable selection through backward stepwise techniques. The optimal variables selected for the LLRM are RT, PLAKE, LONGC, ρ_{WMRB}, MBS, LATC and FS. For GAM, the most relevant explanatory variables were found to be somewhat different than those obtained for the LLRM because in this case selected predictors present non-linear links with the response variables. These variables are namely, MCL, MBS, PFOR, PLAKE, MASP and ρ_{WMRB}. Finally, the significant explanatory variables selected for MARS are AREA, PLAKE, MALPS, RT, PFOR, MASP, ρ_{WMRB} and WMRB. The definition of these variables is given in Table 2.

The selected variables mainly include: i) variables dealing with drainage network characteristics such as RT, ρ_{WMRB} and WMRB. These variables have a high relationship with the underlying lithology, the infiltration ability and the topographic characteristics of the terrain which allow integrating more information about the underlying hydrogeological flows (Msilini, Ouarda, and Masselot 2021); ii) Precipitations (MALPS and MASP) and variables related to the local climate conditions such as LONGC and LATC; and iii) variables characterising the land cover such as PLAKE acting like a sponge absorbing the excess of water during the extreme events and PFOR variable controlling the soil erosion phenomenon and the infiltration ability of the basins.

5.3 Comparison of regional models

Table 4 shows the jackknife validation results for each regional model. Accordingly, the lowest RRMSE values are associated with the CCA/MARS approach, followed by MARS and GAM applied with all datasets. With ALL, MARS has a comparable or even superior performance than GAM. One can also see that, applying the LLRM model within the neighborhoods gives considerably improved results. However, it did not improve significantly the predictive ability of non-linear RE models, especially GAM. This may be attributed to the fact that the amount of data used in this study is not sufficiently large. On the other hand, the use of the neighborhood approaches often leads to significant improvement in the RE in comparison with ALL. In this study, when non-linear RE models are used, especially GAM, the difference between ALL and neighborhood approaches is negligible. This result indicates that the use of non-linear RE models may make the analyses more satisfactory and robust by compensating the benefits of using the neighborhoods approaches which is not the case for the LLRM. Therefore, non-linear RE models, especially GAM, seem especially useful for smaller datasets. The use of these models may reduce the importance of using the neighborhood approaches.

In this work, the considered limited amount of data may also be the cause of the high variance observed for the different models. It can also be seen that the NASH obtained with the different approaches is not sufficiently high, especially for QS_{50} and QS_{100}. This result may also be explained by the small size of the used data as the NASH is a criterion that is highly sensitive to the sample size (McCuen, Knight, and Cutter 2006).

Figure 3 shows the variability of the relative error as a function of the sites associated to the best models ALL/GAM, ALL/MARS and CCA/MARS for QS_{50} (QS_{10} and QS_{100} are not presented here because of the similarity of the results). Overall, CCA/MARS performs slightly better than the other approaches, especially for two specific sites that have exceptionally large relative errors. The first site (050701) was also previously identified by Ouali, Chebana, and Ouarda (2017) as a problematic station with atypically large relative errors; the second site (030919) is a cross-border catchment. In this study, the physiographical variables of the cross-border catchments are extracted based on data come from the USGS database, which have a different resolution than the DEMs obtained from the Natural Resources Canada database. This difference in measurement might therefore explain this observation different behaviour.

The best models identified in this study are used to do predictions in the 11 ungauged sites of the study case (see Figure 1). The estimations of the quantiles obtained by CCA/MARS are found to be higher compared to those obtained by ALL/GAM and ALL/MARS. This may be explained by the fact that the CCA/MARS approach presented a positive RBIAS, and then it overestimates the target quantiles.

6. Conclusions

In this study, the performances of a number of commonly used regional approaches are compared for the estimation of spring flood quantiles at 11 ungauged sites of interest located in the Montreal region (Canada). The objective is to test the robustness of the various methods by testing them on a real world case study with less than ideal conditions: limited number of stations and moderate data quality. Different RE models (LLRM, GAM and MARS) are considered with and without delineation methods (CCA and ROI). These models are calibrated and validated on a group of catchments from the study area. The best models are selected and used to estimate the flood quantiles at the target ungauged sites.

Results indicate that it is possible and important to use the proposed non-linear regional models in practice (GAM and MARS) because performances are improved when these models are used instead of LLRM. The CCA/MARS combination was found to be the best combination of DHR and RE with respect of the RMSE and RRMSE for this case study. The neighborhood approaches considered in conjunction with GAM do not lead to improved performances. This may be explained by the fact that the calibration of GAM requires a large dataset which is not the case for the present study area. The different models are also found to have a high variance compared to the bias, which may also be attributed to the size and quality of the used dataset.

In future efforts, it may be of interest to enlarge the database by considering other stations with short time series. Procedures for the combination of local and regional information can then be used and their performance assessed (see for instance, Seidou et al. (2006)).

These procedures have been proposed in the literature but are almost never used in practice. Their application to a real-world case study may help demonstrate their potential and increase their use in practical hydrological estimation studies. One important aspect that can also be considered in future work is the integration of climate change influence in the modeling of the hydrological response. Indeed, it would be of interest to test the proposed statistical model using flood quantiles estimated under a changing climate. In future efforts it may also be useful to assess and compare the predictions that were obtained with the considered models with those obtained with deterministic models such as HYDROTEL or CEQUEAU. In this work, we assessed and applied the different RE models (LLRM, GAM and MARS) in combination with linear neighborhood models (CCA and ROI). In further work, it should be of interest to evaluate and apply these models in conjunction with non-linear neighborhood approaches such as the non-linear canonical correlation analysis model (Ouali, Chebana, and Ouarda 2016) and the nonlinear neighborhood approach based on statistical depth functions (Wazneh, Chebana, and Ouarda 2016).

Acknowledgments

Financial support for this work was graciously provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chairs program (CRC), the Regional County Municipality of Vaudreuil-Soulanges (MRC-VS), the Communauté Métropolitaine de Montréal (CMM) and the Tunisian University Mission in North America (MUTAN). The authors are grateful to Natural Resources Canada and the USGS services for the DEM and digital data used, and to the info-climat service of the Ministry of the Environment and the Fight against Climate Change of Quebec (MELCC) for the hydrological and meteorological data used. The authors would also like to thank Michel Leclerc, Simon Bellemare (MRC-VS) and Pierre Dupuis (CMM) for their comments and contributions throughout the study.

References

Bayentin, Lampouguin, Salaheddine El Adlouni, Taha BMJ Ouarda, Pierre Gosselin, Bernard Doyon, and Fateh Chebana. 2010. "Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989-2006 in Quebec, Canada." *International journal of health geographics* 9 (1):5. doi: https://doi.org/10.1186/1476-072X-9-5.

Burn, D. H. 1990. "Evaluation of regional flood frequency analysis with a region of influence approach." *Water Resources Research* 26 (10):2257-65. doi: https://doi.org/10.1029/WR026i010p02257.

Chebana, Fateh, Christian Charron, T. B.M.J Ouarda, and Barbara Martel. 2014. "Regional frequency analysis at ungauged sites with the generalized additive model." *Journal of Hydrometeorology* 15 (6):2418-28. doi: https://doi.org/10.1175/JHM-D-14-0060.1.

Chebana, Fateh, and Taha BMJ Ouarda. 2008. "Depth and homogeneity in regional flood frequency analysis." *Water Resources Research* 44 (11). doi: https://doi.org/10.1029/2007WR006771.

Dalrymple, T. 1960. *Flood-frequency analyses*. US Geological Survey Water Supply Paper 1543A: USGPO.

Durocher, Martin, Fateh Chebana, and Taha BMJ Ouarda. 2015. "A nonlinear approach to regional flood frequency analysis using projection pursuit regression." *Journal of Hydrometeorology* 16 (4):1561-74. doi: https://doi.org/10.1175/JHM-D-14-0227.1.

Friedman, Jerome H. 1991. "Multivariate adaptive regression splines." *The annals of statistics*:1-67.

Gal, Yarin, and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. Paper presented at the Advances in neural information processing systems.

GREHYS. 1996. "Presentation and review of some methods for regional flood frequency analysis." *Journal of hydrology(Amsterdam)* 186 (1-4):63-84.

Han, Xudong, Taha BMJ Ouarda, Ataur Rahman, Khaled Haddad, Rajeshwar Mehrotra, and Ashish Sharma. 2020. "A Network Approach for Delineating Homogeneous Regions in Regional Flood Frequency Analysis." *Water Resources Research* 56 (3):e2019WR025910.

Hastie, and R. Tibshirani. 1987. "Generalized Additive Models: Some Applications." *Journal of the American statistical Association* 82 (398):371-86. doi: 10.1080/01621459.1987.10478440.

Isaaks, EH, and RM Srivastava. 1989. "An Introduction to Applied Geostatistics, New York: Oxford Univ." In.: Press.

Jenson, Susan K, and Julia O Domingue. 1988. "Extracting topographic structure from digital elevation data for geographic information system analysis." *Photogrammetric engineering and remote sensing* 54 (11):1593-600.

Khalil, B, TBMJ Ouarda, and A St-Hilaire. 2011. "Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis." *Journal of Hydrology* 405 (3-4):277-87.

Lawrence, Steve, and C Lee Giles. 2000. Overfitting and neural networks: conjugate gradient and backpropagation. Paper presented at the Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium.

Leitte, Arne Marian, Cristina Petrescu, Ulrich Franck, Matthias Richter, Oana Suciu, Romanita Ionovici, Olf Herbarth, and Uwe Schlink. 2009. "Respiratory health, effects of ambient air pollution and its modification by air humidity in Drobeta-Turnu Severin, Romania."

Science of The Total Environment 407 (13):4004-11. doi: https://doi.org/10.1016/j.scitotenv.2009.02.042.

López-Moreno, Juan I, and David Nogués-Bravo. 2005. "A generalized additive model for the spatial distribution of snowpack in the Spanish Pyrenees." *Hydrological Processes: An International Journal* 19 (16):3167-76.

McCuen, Richard H, Zachary Knight, and A Gillian Cutter. 2006. "Evaluation of the Nash–Sutcliffe efficiency index." *Journal of hydrologic engineering* 11 (6):597-602.

Milborrow, Stephen. 2018. "Derived from MDA: mars by Trevor Hastie and Rob Tibshirani.Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. Earth: Multivariate Adaptive Regression Splines . R package version 4.6.3."

Msilini, A, TBMJ Ouarda, and P Masselot. 2021. "Evaluation of additional physiographical variables characterising drainage network systems in regional frequency analysis, a Quebec watersheds case-study." *Stochastic environmental research and risk assessment*:1-21.

Msilini, A., P. Masselot, and T.B.M.J. Ouarda. 2020. "Regional Frequency Analysis at Ungauged Sites with Multivariate Adaptive Regression Splines." *Journal of Hydrometeorology*:1-. doi: 10.1175/jhm-d-19-0213.1.

O'Callaghan, John F, and David M Mark. 1984. "The extraction of drainage networks from digital elevation data." *Computer vision, graphics, and image processing* 28 (3):323-44. doi: https://doi.org/10.1016/S0734-189X(84)80011-0.

Ouali, Dhouha, Fateh Chebana, and T.B.M.J Ouarda. 2016. "Non-linear canonical correlation analysis in regional frequency analysis." *Stochastic environmental research and risk assessment* 30 (2):449-62. doi: https://doi.org/10.1007/s00477-015-1092-7.

———. 2017. "Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites." *Journal of Advances in Modeling Earth Systems* 9 (2):1292-306. doi: https://doi.org/10.1002/2016MS000830.

Ouarda, T.B.M.J, Claude Girard, George S Cavadias, and Bernard Bobée. 2001. "Regional flood frequency estimation with canonical correlation analysis." *Journal of Hydrology* 254 (1):157-73. doi: https://doi.org/10.1016/S0022-1694(01)00488-7.

Ouarda, T.B.M.J. 2016. "Regional flood frequency modeling." *chapter 77, in: V.P. Singh, (Ed). Chow's Handbook of Applied Hydrology,3rd Edition, Mc-Graw Hill, New York*:pp. 77.1-.8, ISBN 978-0-07-183509-1.

Ouarda, T.B.M.J., Christian Charron, Yeshewatesfa Hundecha, André St-Hilaire, and Fateh Chebana. 2018. "Introduction of the GAM model for regional low-flow frequency analysis at ungauged basins and comparison with commonly used approaches." *Environmental Modelling & Software* 109:256-71. doi: https://doi.org/10.1016/j.envsoft.2018.08.031.

Ouarda, T.B.M.J., Christian Charron, Prashanth R Marpu, and Fateh Chebana. 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9 (4):1553-66. doi: https://doi.org/10.1109/jstars.2016.2522764.

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

Rahman, Ataur, Christian Charron, Taha BMJ Ouarda, and Fateh Chebana. 2018. "Development of regional flood frequency analysis techniques using generalized additive models for Australia." *Stochastic environmental research and risk assessment* 32 (1):123-39. doi: https://doi.org/10.1007/s00477-017-1384-1.

Seidou, O, TBMJ Ouarda, M Barbet, P Bruneau, and B Bobee. 2006. "A parametric Bayesian combination of local and regional information in flood frequency analysis." *Water Resources Research* 42 (11). doi: https://doi.org/10.1029/2005WR004397.
Shu, Chang, and T.B.M.J. Ouarda. 2007. "Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space." *Water Resources Research* 43 (7). doi: doi:10.1029/2006WR005142.
Tasker, Gary D, Scott A Hodge, and C Shane Barks. 1996. "REGION OF INFLUENCE REGRESSION FOR ESTIMATING THE 50-YEAR FLOOD AT UNGAGED SITES." *JAWRA Journal of the American Water Resources Association* 32 (1):163-70. doi: https://doi.org/10.1111/j.1752-1688.1996.tb03444.x.
Wazneh, Hussein, Fateh Chebana, and Taha BMJ Ouarda. 2016. "Identification of hydrological neighborhoods for regional flood frequency analysis using statistical depth function." *Advances in water resources* 94:251-63. doi: https://doi.org/10.1016/j.advwatres.2016.05.013.
Wen, Li, Kerrylee Rogers, Neil Saintilan, and Joanne Ling. 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." *Ecological Modelling* 222 (1):154-63. doi: https://doi.org/10.1016/j.ecolmodel.2010.09.016.
Wood. 2006. *Generalized additive models: an introduction with R*: CRC press.
———. 2017. *Generalized additive models: an introduction with R*: CRC press.
Xu, Jianhua, Weihong Li, Minhe Ji, Feng Lu, and Shan Dong. 2010. "A comprehensive approach to characterization of the nonlinearity of runoff in the headwaters of the Tarim River, western China." *Hydrological Processes: An International Journal* 24 (2):136-46. doi: https://doi.org/10.1002/hyp.7484.

Table 1 Considered regional models.

Regional model \ Step	DHR	RE
ALL/LLRM	ALL (all stations)	LLRM
ALL/GAM	ALL (all stations)	GAM
ALL/MARS	ALL (all stations)	MARS
CCA/LLRM	CCA	LLRM
CCA/GAM	CCA	GAM
CCA/MARS	CCA	MARS
ROI/LLRM	ROI	LLRM
ROI/GAM	ROI	GAM
ROI/MARS	ROI	MARS

Table 2 List of variables used in the present study.

Notation	Variable
QS_T	Spring specific flood quantiles associated to the return period T
AREA	Basin area
MCL	Main channel length
MCS	Main channel slope
MBS	Mean basin slope
PFOR	Percentage of the area occupied by forest
PLAKE	Percentage of the area occupied by lakes
MATP	Mean annual total precipitation
MALP	Mean annual liquid precipitation
MASP	Mean annual solid precipitation
MALPS	Mean annual liquid precipitation (summer–fall)
DDBZ	Mean annual degree days below 0 °C
LATC	Latitude of the centroid of the basin
LONGC	Longitude of the centroid of the basin
RT	Texture ratio
RC	Circularity ratio
MRL	Mean stream length ratio
MRB	Mean bifurcation ratio
WMRB	Weighted mean bifurcation ratio
ρ_{WMRB}	RHO WMRB coefficient
DD	Drainage density
FS	Stream frequency
IF	Infiltration number
RN	Ruggedness number
PN1	Percentage of first-order streams
PL1	Percentage of first-order stream lengths

Table 3 Descriptive statistics of the hydrological and physio-meteorological variables.

Variable	Min	Mean	Max	Std. dev
AREA (km^2)	26.30	1045.85	5440	1196.13
MCL (km)	12.51	68.98	225.80	46.13
MCS (m/km)	0.77	4.23	21.06	3.94
MBS (degree)	0.26	3.00	9.72	2.05
PFOR (%)	5.15	67.45	96	24.99
PLAKE (%)	0.00	3.93	21.28	3.86
MATP (mm)	923	1066.47	1239	78.33
MALP (mm)	669	828.68	1097	73.23
MASP (cm)	166	252.61	343	41.46
MALPS (mm)	426	504.64	664	45.30
DDBZ (degree-day)	859	1167.19	1578	184.25
LATC (°N)	44.88	45.97	47.43	0.66
LONGC (°W)	70.65	72.88	75.12	1.14
RT (km^{-1})	2.42	16.40	45.25	9.84
RC	0.08	0.21	0.39	0.07
MRL	0.48	0.84	1.14	0.20
MRB	1.69	2.12	5.78	0.74
WMRB	1.85	2.06	2.86	0.18
ρ_{WMRB}	0.18	0.41	0.55	0.10
DD (km^{-1})	2.10	2.84	3.48	0.29
FS (km^{-2})	7.04	9.10	11.42	1.21
IF (km^{-3})	16.04	26.20	39.73	6.06
RN	0.05	1.56	3.70	0.85
PN1 (%)	50.16	50.39	51.20	0.22
PL1 (%)	38.78	51.59	60.43	4.47
QS_{10} ($m^3/s\ km^{-2}$)	0.080	0.272	0.482	0.093
QS_{50} ($m^3/s\ km^{-2}$)	0.108	0.346	0.679	0.135
QS_{100} ($m^3/s\ km^{-2}$)	0.119	0.377	0.772	0.156

Table 4 Jackknife validation results (Best results are in bold character).

Quantile		LLRM			GAM			MARS		
		ALL	**CCA**	**ROI**	**ALL**	**CCA**	**ROI**	**ALL**	**CCA**	**ROI**
NASH	QS10	0.426	0.604	0.522	0.705	0.681	0.706	0.731	**0.743**	0.652
	QS50	0.409	**0.593**	0.464	0.589	0.534	0.579	0.551	0.586	0.559
	QS100	0.383	**0.575**	0.419	0.521	0.468	0.510	0.426	0.558	0.486
RMSE [$(m^3/s)km^{-2}$]	QS10	0.070	0.058	0.064	0.050	0.052	0.050	0.048	**0.046**	0.054
	QS50	0.103	**0.085**	0.099	**0.085**	0.091	0.086	0.089	**0.085**	0.088
	QS100	0.122	**0.101**	0.119	0.107	0.113	0.108	0.117	**0.102**	0.111
RRMSE (%)	QS10	29.020	25.610	26.712	21.796	21.290	21.353	19.023	**17.189**	22.607
	QS50	29.933	27.785	29.078	28.476	28.637	28.196	26.857	**21.385**	26.532
	QS100	31.456	29.508	30.933	31.863	32.017	31.602	32.514	**23.723**	29.529
BIAS [$(m^3/s)km^{-2}$]	QS10	0.003	0.009	0.005	0.004	0.007	**0.001**	0.007	0.013	-0.004
	QS50	0.004	0.013	0.005	0.008	0.012	0.005	0.013	0.015	**0.001**
	QS100	**0.005**	0.015	**0.005**	0.011	0.016	0.007	**0.005**	0.021	0.006
RBIAS (%)	QS10	-3.819	-1.792	-2.535	-1.488	**-0.815**	-2.843	1.199	2.577	-4.377
	QS50	-4.049	-2.218	-3.677	-2.331	-1.967	-4.005	**-0.496**	1.546	-4.628
	QS100	-4.390	-2.584	-4.350	-2.946	-2.552	-4.834	-4.498	**1.541**	-3.865

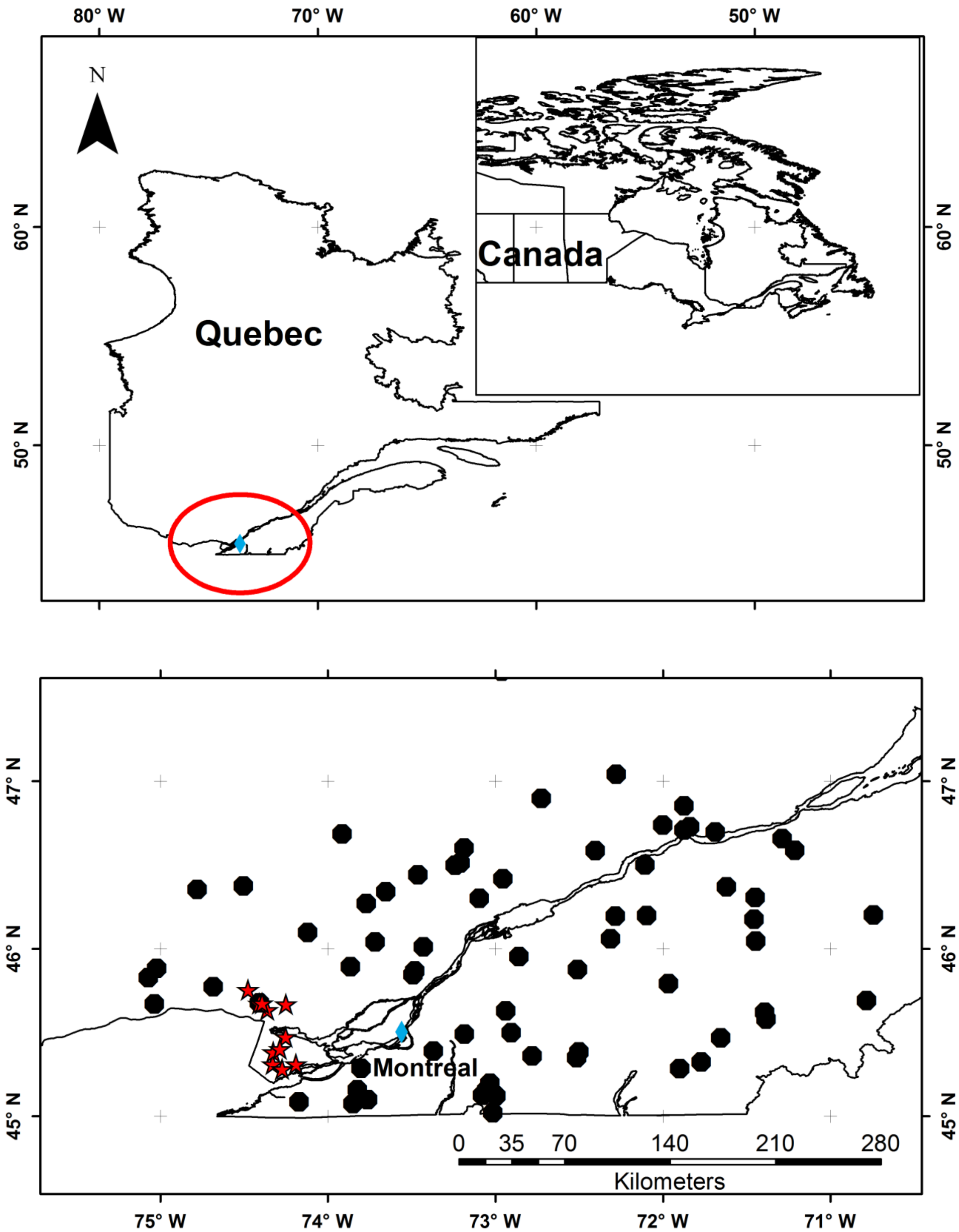

Figure 1 Location of the hydrometric stations across the study area (black circles), the red stars present the ungauged sites. The blue diamond refers to the location of the study area (Montreal region).

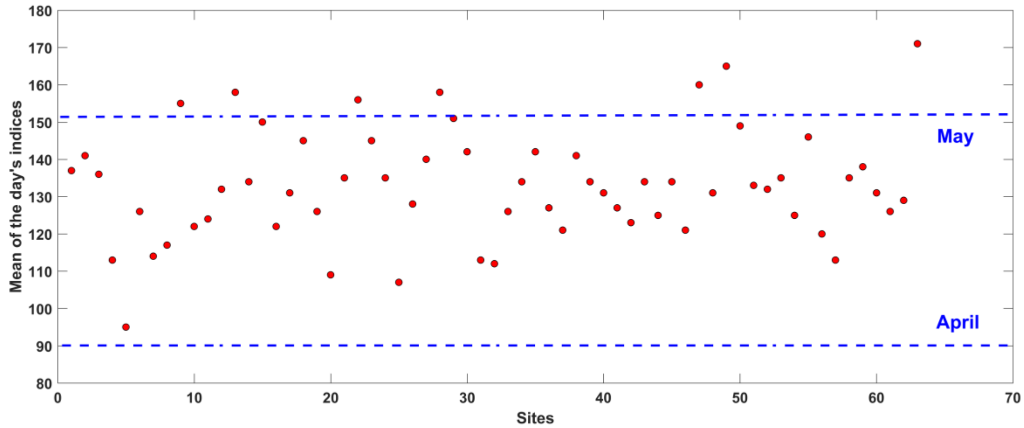

Figure 2 Annual mean of the day's indices (MDI) associated to the maximum annual flow as a function of sites. The dotted blue lines represent the limit of the April-May months. The red circles are the MDI values (annual floods) which are mostly observed in the April-May months.

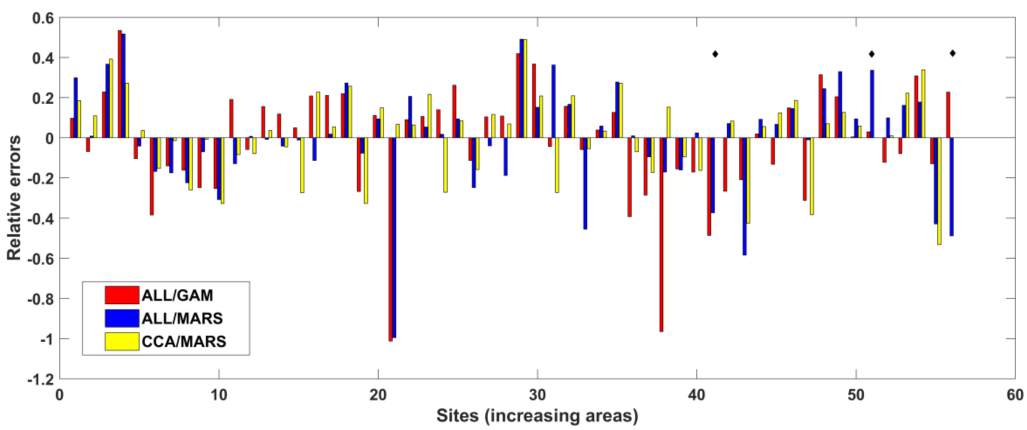

Figure 3 Relative errors associated to the at site quantile QS50 calculated using ALL/GAM; ALL/MARS and CCA/MARS. Diamond refers to sites with (or not) a small neighborhood. Sites are ordered according to their areas.