Evaluation of additional physiographical variables characterising drainage network systems in regional frequency analysis, a Quebec watersheds case-study

A. Msilini[1,*], T.B.M.J. Ouarda[1] and P. Masselot[2]

[1]*Canada Research Chair in Statistical Hydro-Climatology, INRS-ETE, 490 de la Couronne, Québec, QC, G1K 9A9, Canada.*

[2] *London School of Hygiene & Tropical Medicine (LSHTM), 15-17 Tavistock Pl, London, WC1H 9SH, United Kingdom.*

[*] Corresponding author: Amina Msilini (amina.msilini.m@gmail.com)

Acknowledgments and Data

Financial support for the present study was graciously provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research chairs program (CRC) and the University Mission of Tunisia in Montreal (MUTAN). The authors would like to thank Christian Charron for his valuable help and input. The authors are grateful to Natural Resources Canada (https://www.nrcan.gc.ca/earth-sciences/geography/topographic-information/download-directory-documentation/17215) and the USGS (https://earthexplorer.usgs.gov/) services for the employed DEM and NHN data. The authors would like also to thank the Ministry of Sustainable Development, Environment, and Fight Against Climate Change (MDDELCC) services for the used dataset (STA). The authors are grateful to the Editor-in-Chief, Dr. George Christakos, and to three anonymous reviewers for their comments which helped improve the quality of the manuscript.

Conflicts of Interest: The author declares no conflict of interest.

August, 2021

Abstract

Regional Frequency Analysis (RFA) relies on a wide range of physiographical and meteorological variables to estimate hydrological quantiles at ungauged sites. However, additional catchment characteristics related to its drainage network are not yet fully understood and integrated in RFA procedures. The aim of the present paper is to propose the integration of several physiographical variables characterizing the drainage network systems in RFA, and to evaluate their added value in predicting quantiles at ungauged sites. The proposed extended dataset (EXTD) includes several variables characterising drainage network characteristics. To evaluate the new variables, a number of commonly used RFA approaches are applied to the extended data representing 151 stations in Quebec (Canada) and compared to a standard dataset (STA) that excludes the new variables. The considered RFA approaches include the combination of two neighborhood methods namely the canonical correlation analysis (CCA) and the region of influence (ROI) with two regional estimation (RE) models which are the log-linear regression model (LLRM) and the generalized additive model (GAM). The RE models are also applied without the hydrological neighborhood. Results show that regional models using the extended dataset lead to significantly better flood quantile predictions, especially for large basins. Indeed, the variable selection performed with EXTD consistently includes some of the new variables, in particular the drainage density, the stream length ratio, and the ruggedness number. Two other new variables are also identified and included in the DHR step: the circularity ratio and the texture ratio. This leads to better predictions with relative errors about 29% for EXTD, versus around 42% for STA in the case of the best

combination of RFA approaches. Thus, the proposed new variables allow for a better representation of the physical dynamics within the watersheds.

Keywords : Drainage network characteristics; Ungauged basin; Canonical correlation analysis; Region of influence; Generalized Additive Model, Regional frequency analysis.

Abbreviations

BH	Basin relief
BIAS	Mean bias
CCA	Canonical correlation analysis
DD	Drainage density
DDBZ	Mean annual degree days below 0 °C
DEM	Digital elevation model
DHR	Delineation of homogenous regions
Edf	Estimated smooth degree of freedom
EXTD	Extended dataset
FS	Stream frequency
GAM	Generalized additive model
IF	Infiltration number
LATC	Latitude of the centroid of the basin
LLRM	Log-linear regression model
LONGC	Longitude of the centroid of the basin
LU	Stream length
MALP	Mean annual liquid precipitation
MASP	Mean annual solid precipitation
MATP	Mean annual total precipitation
MBS	Mean basin slope
MCL	Main channel length
MRB	Mean bifurcation ratio
MRL	Mean stream length ratio
NASH	Nash efficiency criterion
NHN	National Hydro Network
PFOR	Percentage of the area occupied by forest
PLAKE	Percentage of the area occupied by lakes
PL1	Percentage of first-order stream lengths
PN1	Percentage of first-order streams
QS_T	Specific quantile associated to the return period T
Q_T	At-site flood quantile corresponding to return period T
R^2	Coefficient of determination
RB	Bifurcation ratio
RBIAS	Relative mean bias
RC	Circularity ratio
RE	Regional estimation
RFA	Regional frequency analysis
RL	Stream length ratio
RMSE	Root-mean-square error
RN	Ruggedness number
ROI	Region of influence
RRMSE	Relative root-mean-square error
RT	Texture ratio
STA	Standard dataset
U	Stream order
Var	Explanatory variable
WMRB	Weighted mean bifurcation ratio
ρ	RHO coefficient
ρ_{WMRB}	RHO WMRB coefficient

1. Introduction

Regional frequency analysis (RFA) procedures are commonly used in hydrology to estimate flood and low-flow quantiles at sites where little or no hydrological data is available. Generally, RFA includes two main steps: delineation of homogenous regions (DHR) and regional estimation (RE) (e.g. Chebana et al. 2014; Chebana and Ouarda 2007; Ouarda 2016). In this context, climatic, morphometric and physiographic characteristics of the watershed are widely used to describe geomorphic processes (e.g. Baumgardner 1987; Hadley and Schumm 1961; Marchi and Dalla Fontana 2005; Tramblay et al. 2010) in order to predict hydrological variables using RFA approaches (e.g. Dawson et al. 2006; Dodangeh et al. 2014; Goswami et al. 2007; Seidou et al. 2006; Tsakiris et al. 2011).

A number of physio-meteorological variables, such as basin area, basin slope, precipitation characteristics and land occupation are commonly used in the field of hydrology and more precisely in the RFA procedures. They are considered as the most relevant variables for these studies based on their high correlation with the hydrological variables (Chokmani and Ouarda 2004). In addition to the commonly considered variables (a more exhaustive list is in Table 1), drainage network characteristics (Jung et al. 2017) and tectonic setting (e.g. Ahmadi et al. 2006; Hamed et al. 2014) may have a strong impacts on hydrological dynamics, and are consequently related to flood quantiles. However, they are not yet well investigated and integrated in RFA studies. Indeed, the assessment of morphometric and physiographic variables requires the analysis of a number of stream characteristics (e.g. ordering of the streams, bifurcation ratio, texture ratio, stream length ratio, etc.). These variables characterize the basin shape as well as the drainage system, and can be useful to model the hydrological dynamics. Youssef et al. (2011) also indicated that the circularity ratio, number of orders and drainage density have a direct impact on the hydrological risk. Hence, the integration of these variables in the procedures for the regionalization of extreme hydrological events may contribute to the enhancement of RFA results. Variables related to drainage network systems are already used in several morphometric and hydrologic studies (e.g. Ameri et al. 2018; Biswas et al. 1999; Kaliraj et al. 2015; Pareta and Pareta 2011; Rai et al. 2017; Ratnam et al. 2005; Reddy et al. 2004; Sivasena Reddy and Janga Reddy 2013; Vijith and Satheesh 2006; Youssef et al. 2011) and they can eventually be useful in regionalization studies. These variables can be extracted based on classical approaches such as topographic maps and field examination or with advanced techniques using remote sensing and Digital Elevation Models (DEM). Remote sensing techniques coupled with the potential of GIS tools are increasingly popular. Indeed, they make it possible to calculate the various characteristics of the basin very quickly and more efficiently based on a DEM which is not possible in the past.

During the last decades, the focus in RFA has been mainly on the development of new delineation and estimation methods (e.g. Durocher et al. 2015; Ouali et al. 2016; Wazneh et al. 2016). Meanwhile, the list of physiographical and meteorological variables used as predictors has seen little evolution. In the present study, a number of commonly used RFA approaches are applied to test and evaluate the potential improvements that may result from the adoption of new physiographic variables.

The objective of this work is to propose the use of new physiographical variables related to the basin shape and drainage network and argue about their usefulness. To evaluate their added value for quantile prediction in RFA, they are computed and used for a set of 151 basins in Quebec (Canada). More specifically, the objective is to use both the standard and extended databases to predict quantiles associated to several return periods, and compare their prediction performances. In this work, standard RFA methods are considered for quantile prediction, namely Canonical correlation analysis (CCA) (Ouarda et al. 2000) and the region of influence (ROI) (Burn 1990) for DHR, including a case with no DHR, as well as the log-linear regression model (LLRM) and the generalized additive model (GAM) (Hastie 1986) for RE.

The present paper is structured as follows: Section 2 offers a review of the new physiographic and morphometric variables proposed in this work by detailing their characteristics. Section 3 briefly presents the theoretical background of the CCA and the ROI approaches for the delineation of neighborhoods and the LLRM and the GAM for the regional estimation. The adopted methodology and the developed regional models are detailed in section 4. Section 5 describes the study area and the used datasets. The results are presented and discussed in section 6, and the conclusions of the work are summarized in the last section.

2. Variables characterizing drainage networks

Drainage network characteristics and evolution depend closely on the prevailing climatic, physiographic, and topographic conditions of the basin (Jung et al. 2015). These conditions determine the drainage network configuration which, in turn, can affect the hydrological response of the watershed (Howard 1990), and consequently hydrological quantile estimation. The new physiographical variables considered in this work are presented herein. Table 2 summarizes the definitions and standard mathematical equations used to determine these variables.

2.1 Stream order (U)

The stream order of a basin is the highest stream order within the basin, where an order one is a stream starting at the source. A number of stream ordering systems are available in the hydrological literature. The simplest and most used one is the Strahler system originally introduced by Horton (1945) and then modified by Strahler (1952). This method is based on a hierarchical ranking of streams. When two first order streams join, an order two is formed and so on. Several researchers have directly correlated the stream order with stream flow (e.g. Blyth and Rodda 1973; Stall and Fok 1967). Blyth and Rodda (1973) also observed that during dry periods, first-order streams present less than 20% of the total length of the drainage network. At the maximum development of the drainage network, the total length of first-order streams constitutes over 50% of the total basin stream length. Thus, stream order frequency, especially the frequency of the first-order streams, may be well correlated with the hydrological response of the watershed.

2.2Texture ratio (*RT*)

The texture ratio (*RT*) allows characterizing the basin drainage texture and is one of the most important factors in the drainage morphometric analysis due to its high relationship with the underlying lithology, the infiltration ability and the topographic characteristics of the terrain (Schumm 1956). High *RT* levels indicate the presence of soft rocks with high sensitivity to erosion (Ameri et al. 2018), and consequently a high and speedy surface runoff.

2.3 Circularity ratio (*RC*)

The circularity ratio (*RC*) is defined as the ratio between the areas of a catchment to the area of the circle having the same perimeter of the catchment. It is an important variable that helps characterize the basin shape. It is affected by the length and frequency of streams, geological structures, land use and cover, and the slope of the catchment (Dar et al. 2014; Vijith and Satheesh 2006). *RC* values range between 0 and 1. Basins with *RC* values close to 1 are characterized by circular form and a low concentration time and then a high peak flow. Low *RC* values are associated with strongly elongated basins and with lower runoff.

2.4 Stream length ratio (*RL*)

The stream length ratio (*RL*) was defined by Horton (1945) as the ratio between the mean length of the streams of a given order and the next lower order. It is based on Horton's law (1945) of stream length that indicates the existence of a direct geometric relationship between the mean length of the streams of a given order and the next lower order. The *RL* between successive stream orders changes under the effect of the topographic and slope variability, and has a significant relationship with surface runoff and the erosional stage of the watershed (Sreedevi et al. 2005).

2.5 Mean bifurcation ratio (*MRB*) and weighted mean bifurcation ratio (*WMRB*)

The bifurcation ratio (*RB*) is defined as the ratio between the stream's number of a given order and those of the next-higher order in a drainage network. It permits the characterization of the impacts of the geological structures on the drainage network. Strahler (1957) indicated that the *RB* shows a slight range of variation for different regions except where the impact of the geological control is important. Chow (1964), Strahler (1964) and Verstappen (1983) indicated that, in general, the geological structures have a negligible impact on drainage networks, if the mean bifurcation ratio (*MRB*) of the watershed is comprised between 3 and 5. A higher value of this variable indicates a sort of geological control (Agarwal 1998). This variable can also characterize the watershed's shape. A high *RB* value is, generally, associated with an elongated basin, while a low *RB* value is likely to be associated with a circular basin (Gajbhiye 2015; Taofik et al. 2017). Strahler (1953) proposed a more representative bifurcation number measure, called weighted mean bifurcation ratio (*WMRB*). It consists in multiplying the ordinary RB identified for each successive order by the total number of streams involved in the ratio and subsequently taking the mean of these values. Schumm (1956) used this approach to determine the *WMRB* of the drainage system of the Perth Amboy (N.J). Pareta and Pareta (2011) and Bajabaa et al. (2014) also used this variable in hydrologic and morphometric analysis studies.

2.6 RHO coefficient (ρ)

The RHO coefficient (ρ) is defined as the ratio between the *RL* and the *RB* of the watershed. It characterizes the relationship between the physiographic development of the watershed and the drainage density, and permits the assessment of the storage capacity of the drainage network (Horton 1945). This variable is affected by several climatic, geologic, biologic, geomorphologic and anthropogenic factors (Mesa 2006).

2.7 Drainage density (*DD*)

The drainage density (*DD*) was introduced by Horton (1932) in the hydrological literature as the total length of stream networks per unit area. *DD* express the closeness of the spacing of streams, and provides a quantitative measurement of landscape dissection and runoff potential (Magesh et al. 2011). It is a result of interacting factors controlling the surface runoff such as, the infiltration capacity, the climatic conditions and the vegetation cover of the watershed (Máčka 2001; Patton 1988; Reddy et al. 2004; Verstappen 1983).

2.8 Stream frequency (*FS*)

The stream frequency (*FS*) is the number of stream segments of all orders per unit area (Horton 1932; Horton 1945). It depends on the rock characteristics, infiltration capacity, vegetation cover, relief, amount of rainfall and subsurface permeability (Hajam et al. 2013), and reflects the texture of the drainage network (Magesh et al. 2011). In general, a high *FS* is associated with impermeable subsurface, sparse vegetation, high relief conditions and low infiltration capacity (Reddy et al. 2004; Shaban et al. 2005).

2.9 Infiltration number (*IF*)

The infiltration number (*IF*) is defined by Faniran (1968) as the product of the *DD* and the *FS*. It allows the characterization of the watershed infiltration capacity (Hajam et al. 2013). This variable is inversely proportional to the infiltration capacity of the basin.

The higher the *IF* values, the lower will be the infiltration and the higher will be the runoff (Pareta and Pareta 2011).

2.10 Ruggedness number (*RN*)

The ruggedness number (*RN*) is often used to evaluate the flood potential of streams (Patton and Baker 1976) and it usually combines the impact of slope steepness with its length (Strahler 1964). This variable allows describing the structural complexity of the terrain. Watersheds characterized by high *RN* values are highly subject to erosion and therefore susceptible to an increased peak flow (Sreedevi et al. 2013).

3. Theoretical background

In this section, we briefly present the statistical approaches adopted in the present work. We define a RFA model as a two-step procedure beginning with a neighborhood identification method and then performing regional estimation. We hereby consider two different methods for each step, which are described below.

3.1 Delineation of homogeneous regions

3.1.1 Canonical correlation analysis (CCA)

CCA method is detailed in Ouarda et al. (2001) in the context of RFA, and commonly used in this context to identify group of basins having the same hydrological response. This method consists of space reduction by establishing pairs of canonical variables based on a linear transformation of two groups of random variables. Let two sets of random variables $\mathrm{X}=(\mathrm{X}_1, \mathrm{X}_2,\ldots,\mathrm{X}_\mathrm{m})$ and $\mathrm{Y}=(\mathrm{Y}_1, \mathrm{Y}_2,\ldots,\mathrm{Y}_n)$ containing, respectively, the m physio-meteorological variables and the n hydrological variables of N gauged sites. Based on these variables, the linear combinations V_i and Z_i of the variables X and Y and the canonical correlation coefficients $\lambda_1, \ldots, \lambda_p$ (with $\lambda_i = \text{corr}(V_i, Z_i)$) can be computed.

Using the CCA method, the considered basins can be represented as points in a spaces of the uncorrelated canonical variables (V_i , Z_j); where $i \neq j$. Then, it will be possible to examine the similarity of the point patterns in these spaces, i.e., the ability of the physio-meteorological canonical variables to predict the hydrological variables. The point patterns that are sufficiently similar are associated with sub-group of basins that belongs to the same statistical population and vice versa. The similarity between the basins are measured based on a Mahalanobis distance.

3.1.2 Region of influence (ROI)

As the CCA, the ROI method (Burn 1990) allows the identification of a hydrological neighborhood for a given target-site based on a Euclidean distance, generally a weighted Euclidean distance. This distance determines the similarity of watersheds in a multidimensional space of physio-meteorological variables. A more detailed description of the approach can be found for example in Burn (1990) and GREHYS (1996).

3.2 Regional estimation approaches

3.2.1 Linear Regression Model

The linear regression model or the log-linear regression model (LLRM) is commonly used to find a linear relationship between the hydrological variable (such as the flood quantile Q_T corresponding to a return period T) and the physio-meteorological characteristics of a watershed (X_1, X_2, …, X_m), and it is defined as (e.g. Girard et al. 2004; Pandey and Nguyen 1999) :

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

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-square method (Pandey and Nguyen 1999) and ε is the model error.

3.2.2 Generalized Additive Model

GAM was developed by Hastie and Tibshirani (1986). It is an extension of the generalized linear model (GLM). This model allows for a response distribution other than Gaussian and for a non-linear relationship between response and explanatory variables through smooth functions (Hastie 1986; Wood 2006), which may lead to a more close description of the hydrological processes involved. The GAM formula is given by Wood (2006):

$$g\left(E(Y/X)\right) = \beta_0 + \sum_{j=1}^{m} S_j(X_j) + \varepsilon \qquad (2)$$

where g is a monotonic link function and S_j are smooth functions of explanatory variables X_j.

The estimation of the smooth functions S_j is carried out using splines, which are piecewise polynomial functions linked at points named knots. Generally, the smooth functions S_j are defined as follows:

$$S_j(x)=\sum_{i=1}^{q}\beta_{ji}\, b_{ji}(x) \tag{3}$$

where β_{ji} are unknown parameters and b_{ji} are the spline basis functions.

4. Methodology

4.1 Regional models

In this study, we apply all combinations of the two DHR methods (CCA, ROI) in conjunction with the RE models (LLRM and GAM) presented in section 3. The RE models are also considered with all stations (i.e. without defining any neighborhood). This result in six possible combinations for each dataset (STA and EXTD). Thus, the following regionalization approaches are evaluated (Fig.1):

- ALL/LLRM (STA and EXTD): LLRM used without neighborhoods (all stations) and with variables selected from the STA and the EXTD datasets using the backward stepwise procedure.
- ALL/GAM (STA and EXTD): GAM used without neighborhoods (all stations) and with variables selected from the STA and the EXTD datasets using the backward stepwise procedure.
- CCA/LLRM (STA and EXTD): LLRM used with neighborhoods identified by the CCA method and with variables selected from the STA and the EXTD datasets using the backward stepwise procedure.

- CCA/GAM (STA and EXTD): GAM used with neighborhoods identified by the CCA method and with variables selected from the STA and the EXTD datasets using the backward stepwise procedure.
- ROI/LLRM (STA and EXTD): LLRM used with neighborhoods identified by the ROI method and with variables selected from the STA and the EXTD datasets using the backward stepwise procedure.
- ROI/GAM (STA and EXTD): GAM used with neighborhoods identified by the ROI method and with variables selected from the STA and the EXTD datasets using the backward stepwise procedure

The CCA and ROI methods are used in the DHR considering two different sets of physio-meteorological variables. The first group includes variables from the STA dataset, namely the area (AREA), mean basin slope (MBS), percentage of the area occupied by lakes (PLAKE), mean annual total precipitation (MATP), mean annual degree days below 0 °C (DDBZ) and the longitude of the centroid of the catchment (LONGC). The second one comprises variables from the EXTD dataset, which are PLAKE, MATP, DDBZ, LONGC, *RT* and *RC*. The selection of these variables is carried out based on their correlation level with the hydrological variables (Table 3) as the principle of the CCA is based on correlations. For the aim of simplicity and to be consistent with the CCA, variables selected for the ROI are also based on correlation levels.

The classical procedures of ROI and CCA lead to neighbourhoods with highly variable sample sizes from a target site to another. Indeed, considering a given threshold value, sites located near the centre of the cloud of points determined by the Euclidean space for ROI and the canonical space for CCA are expected to include more sites within their neighbourhoods than sites located on the edge of the cloud of points (Leclerc and Ouarda 2007). Since the accuracy of the estimates obtained by regression models is sensible to the sample size, it was decided to fix the neighbourhood size for all target stations. This size is chosen with a standard jackknife procedure and optimized using the optimization procedure of Ouarda et al. (2001) developed in the Matlab environment.

LLRM and GAM are used in this study as RE models. GAM was developed based on the R package mgcv (Wood 2006). In this work, the thin plate regression spline is considered as basis b_{ji} (.) in the smoothing function $S_j(.)$ in equation (3). This basis function is considered due to its advantages. The thin plate regression spline is characterized by its reduced calculation time, its flexibility and it comprises a lower number of parameters compared to other smoothing functions (Wood 2006). The considered link function g in (2) is the identity function since the log-transformed quantiles are approximately normal (as in Ouali et al. (2017)).

4.2 Selection of explanatory variables

Variable selection procedure is different for the two RFA steps; a correlation-based selection is considered for DHR and a stepwise method is used for RE as a standard approach in the RFA studies. Based on correlation level between physio-meteorological variables and hydrological variables (Table3), six variables are identified for DHR (see above).

For the RE step, four variable selection methods are firstly tested namely forward, backward, stepwise and shrinkage approaches (Heinze et al. 2018) in this study. Table 4 presents the results obtained from each variable selection approach applied for QS_{10} that can be considered as the most reliable quantile. It can be seen that, regardless of the considered selection method, several new variables are selected in the final model. This suggests that new variables in the EXTD are potentially useful for RFA.

To evaluate whether the new variables are predictive of target quantiles, the backward stepwise selection procedure is adopted for both LLRM and GAM. It has already been successfully applied previously with the same dataset (STA) and in the same context by Chebana et al. (2014), Ouarda et al. (2018) and more recently by Msilini et al. (2020). Backward stepwise selection procedure consists in a progressive elimination of variables having the highest p value (based on the hypothesis that the coefficients in equation (1) for LLRM or the smooth terms in equation (3) for GAM are null) from an initial model comprising all available variables. The procedure stops when the number of variables remaining in the model drops below a specific number (Fig.2). This number is chosen as the one minimizing the RRMSE estimated by jackknife.

4.3 Models validation

For each RFA model, a jackknife procedure (also called leave one-out cross validation procedure) is used to evaluate its performance. It consists in considering, in turn, each gauged site as an ungauged one and comparing thereafter the regional estimate to the observed value. This comparison is performed through several criteria: first, the Nash criterion (NASH) gives an evaluation of the degree of adequacy and a global assessment of the prediction quality. Second, the root mean squared error (RMSE) provides information about the accuracy of the prediction in an absolute scale, and the relative RMSE (RRMSE) removes the impact of each site's order of magnitude from the

RMSE values and gives information about the accuracy of the prediction in a relative scale. Finally, the bias (BIAS) and the relative bias (RBIAS) give a measure of the magnitude of the systematic overestimation or underestimation of a model. The formulations of these criteria are given as follows:

Nash:

$$\mathrm{NASH} = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{4}$$

Root-mean-square error :

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{5}$$

Relative root-mean-square error :

$$\mathrm{RRMSE} = 100 \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left[\frac{(y_i - \hat{y}_i)}{y_i}\right]^2} \tag{6}$$

Mean bias :

$$\mathrm{BIAS} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i) \tag{7}$$

Relative mean bias :

$$\mathrm{RBIAS} = 100\, \frac{1}{N}\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)}{y_i} \tag{8}$$

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.

5. Case study and datasets

The data used in this study includes two datasets, the STA and the EXTD, covering 151 stations located in the southern part of Quebec, Canada (Fig. 3). The STA was considered in previous studies with geographical coordinates of the stations and commonly used physio-meteorological variables (e.g. Durocher et al. 2015; Shu and Ouarda 2007; Wazneh et al. 2016). The EXTD dataset combining STA dataset with less common variables representing drainage network properties. The stations are operated by the Ministry of Sustainable Development, Environment, and Fight Against Climate Change.

The considered hydrological variables (Y in the theoretical background) are at-site quantiles standardized by the basin area (specific quantiles), denoted by QS_{10}, QS_{50} and QS_{100} with 10, 50 and 100 are the return periods. Descriptive statistics of hydrological and physio-meteorological variables of the STA (not presented here to avoid repetition) can be found for example in Durocher et al. (2015). The hydrological variables were identified in Kouider et al. (2002a) using a local Frequency Analysis in each gauged site. Data series with at least 15 years of measurement were considered for the analysis. The basic assumptions of stationarity, homogeneity and independence were verified and the appropriate statistical distributions were fitted to data. The appropriate probability distributions identified, are mainly the inverse gamma and Log-Normal with two parameters. For more details about this study, reader may refer to the report of Kouider et al. (2002b). The new physiographical variables, considered in the EXTD, are summarized in Table 5. These variables are identified from drainage networks extracted using the D8 method based on the DEMs (Jenson and Domingue 1988; O'Callaghan and Mark 1984). This technique is implemented in Arc Gis (Arc Hydro).

The D8 method is based on a digital elevation model (DEM) which is basically a grid of elevation values. For each cell, it is considered that water flows in direction of the steepest slope among the eight neighbors of a given DEM cell. The direction grid can then be used to estimate flow accumulation which is obtained by summing the weight of all grid cells following into each downslope cell in the output grid, i.e. simulating the flow path. Based on the obtained flow accumulation grid, the drainage networks can be extracted with the stream head locations corresponding to accumulation values below a constant threshold value (see for instance (Tarboton et al. 1991)).

In this work, the DEMs were hydrologically corrected based on information from the National Hydro Network (NHN). This correction was carried out using the DEM Reconditioning process, which is an implementation of the "AGREE" method. It consists in adjusting the DEM by imposing linear features as a reference. The reference in this case is the (NHN).

The used DEMs have a spatial resolution of ~ 20 m grid cells and are obtained from the Natural Resources Canada database (https://www.nrcan.gc.ca/earth-sciences/geography/topographic-information/download-directory-documentation/17215). Note that, drainage networks of six cross-border watersheds are extracted using the United States Geological Survey (USGS) data distributed with ~ 30 m grid cells (https://earthexplorer.usgs.gov/).

CCA requires the normality of all variables. Hence, some variables need to be transformed. The normality of each variable is visually assessed with a normal probability plot. This technique plots empirical quantiles versus theoretical Gaussian quantiles and should be approximately linear in the case of actual normality. The logarithmic transformation is considered for the hydrological variables, AREA, MBS, MATP, DDBZ and *RT*, and a square root transformation for PLAKE and *RC*. The LONGC is used without transformation since it is approximately normal.

6. Results and Discussion

A correlation analysis is carried out in order to investigate the relationships between variables. Table 3 shows the list of the variables selected for the DHR step based on their high correlation level with the hydrological variables. One can see the existence of relatively high negative correlations between the hydrological variables and the AREA, PLAKE, DDBZ and RT. We also note the presence of important positive correlations between the response variables and the MATP and RC variables. The linear correlation coefficients between the variable RT, which is one of the most important new variables, and the specific quantiles QS_{10} and QS_{100} are -0.53 and -0.51 respectively. However, those between the RT variable and the at-site flood quantiles Q_{10} and Q_{100} are 0.87 and 0.86 respectively. Positive and high correlation values indicate that the increase in RT is associated with a relatively fast and high hydrologic response and consequently an increased risk of erosion. This is consistent with what is stated in Ameri et al. (2018). The second important new variable in terms of correlation level is the RC characterizing the basin shape. Higher RC values (close to 1) are associated with circular basins with low concentration time and high hydrological response hence the positive correlation.

The identification of the neighborhood requires the determination of the optimal number of stations to be used in the RE step. To this end, the optimization procedure of Ouarda et al. (2001) is used. Based on a selected criterion such as RMSE, RRMSE, BIAS or RBIAS the optimal size of neighborhoods can be identified. The optimal size of the neighborhoods should be large enough to ensure that RE can be carried out effectively, but not too large in order to maintain an acceptable degree of homogeneity within the neighborhoods. In this study, we obtain n^{opt} (STA) = 85 sites and n^{opt} (EXTD) = 78 sites with respect to the RRMSE, which is the most important criterion (Hosking and Wallis 2005), for the CCA approach. For the ROI method, the obtained optimum sizes are n^{opt} (STA) = 54 sites and n^{opt} (EXTD) = 44 sites with respect to the same criterion.

The backward stepwise selection method is considered for each quantile (QS_{10}, QS_{50} and QS_{100}) and for each model (LLRM and GAM). In the present study, the optimal number of variables in GAM, which is the most complex model, is found to be seven. Table 6 shows the seven selected variables for each quantile and model combination. We note the selection of three new variables (*RN*, *MRL* and *DD*).

The jackknife procedure results for all considered combinations are presented in Table 7. The best overall performances are obtained with the EXTD, especially with ROI/GAM/EXTD followed by the CCA/GAM/EXTD approaches. Based on the high NASH values (0.79) and the lowest RRMSE values (29.24 % for QS_{100}), the ROI/GAM/EXTD combination gives the most precise estimates compared to all other approaches. According to RBIAS, all models underestimate flood quantiles but the least biased model is ROI/LLRM/EXTD (-1.38 % for QS_{100}). However, compared to the ROI/GAM/EXTD approach, the difference is low (around -1.8 % for QS_{100}).

Note that, GAM applied to EXTD (with and without the neighborhoods) outperforms LLRM applied to EXTD and STA. This may be explained by the ability of GAM to take into account the possible nonlinear connections between predictor and response variables, and also by the important impact of the new variables.

We also notice that the use of the EXTD leads to even more important improvements when adopting the ROI method compared to the CCA approach. Wazneh et al. (2016) have also obtained better results with the ROI than with the CCA approach.

To further explain the previous results, the relative errors as a function of the stations ordered according to their area corresponding to the best combinations (ROI/GAM and CCA/GAM) are given in Fig. 4 and Fig. 5 respectively. It can be seen that the EXTD performs well especially for large basins. Indeed, for the large watersheds the relative errors decrease considerably with the EXTD. This result may also be confirmed by Fig. 6, where one can note that the lowest specific quantiles, which are usually associated to sites with large basin areas, are well estimated with the EXTD. A significant improvement can also be seen for some specific sites that have exceptionally large relative errors with STA. Four such sites (030401, 030402, 041903 and 042607) were identified previously by Chokmani and Ouarda (2004), Durocher et al. (2015) and Ouali et al. (2017) as particular stations with underestimated areas. The integration of more accurate variables dealing with the drainage network, improves considerably the quantile estimates corresponding to these sites.

Jackknife estimates using the ROI/GAM and CCA/GAM approaches (for QS_{100}) are illustrated, respectively, in Fig. 7 and Fig. 8. One can see that these models combined with the EXTD show better performances compared to the STA. The points associated to the scatter diagram of the at-site and regional estimates are less dispersed when using the EXTD than the STA. In addition, the coefficient of determination R^2 values show that the linearity between the local and the regional specific quantile estimates is better explained when using the EXTD than the STA.

Results also indicate that sites with high specific quantile values (more than 0.7 $m^3/s.km^2$), which are generally associated to small basins with an area less than 800 km², are underestimated using the two datasets. This may suggest the usefulness of developing specific regional models for small basins. This result can be explained by the fact that traditional neighborhood approaches (CCA and ROI) lead to an underestimation for sites with small basin areas as shown in Wazneh et al.(2016). This may be the cause of the obtained negative RBIAS values in this work.

Fig. 9 and Fig. 10 present the smooth functions of the response variable $\log(QS_{100})$ as a function of the STA and the EXTD explanatory variables respectively. We notice that the variables PLAKE, DDBZ, AREA and *DD* show a complex nonlinear relationship (nonlinear smooth function curves and high edf values), while the variables LONGC; MALP, MCL, MBS and *MRL* present linear relations.

A particular case of interest from the EXTD that can be observed concerns the relationship between the hydrological variable and the *DD* values. One can see that the higher the *DD* values are the lower the hydrological response will be. This result is in contradiction with what is commonly observed in practice (Melton 1957). In fact, the correlation between the *DD* variable and specific quantile is negative (-0.11) while the correlation between flood quantile and the variable *DD* is positive (0.13). Thus, this variable depends on the size of the watershed, for this reason its effect is reversed in this study case because the specific quantile is used.

We also notice that the *MRL* and MCL variables are found to be inversely proportional to the hydrological response. An increase of these variables is associated with a decrease of the MBS and hence a decrease of the hydrological response.

It can also be seen that the relationship between log(QS_{100}) and PLAKE is decreasing for the majority of PLAKE values, but increases for the highest values of PLAKE. However, the number of points is very limited in the high PLAKE range and more effort will be required to understand the effect of this variable on the flow regime for this range. In general, lakes act as a sponge absorbing the excess water during extreme events, which explains the decreasing relationship between log(QS_{100}) and PLAKE.

The LONGC in this study is an indicator of the station proximity to the Atlantic Ocean and thereafter reflects the influence of the ocean on the local climate. Finally, the variability in the relationship between the DDBZ values and the hydrological response may indirectly reflect the seasonality impact of the temperature on the flow regime. The same patterns were observed previously by Chebana et al. (2014) for the DDBZ and PLAKE variables.

7. Conclusions

Through a case study in the province of Quebec, the present study shows the relevancy of considering drainage network characteristics for quantile prediction in RFA. This result is outlined by the variable importance in RFA models which shows that five new variables, namely RT, RC, DD, MRL and RN are found particularly useful for the specific case of Quebec. Prediction accuracy is also improved using the new variables, especially when considering small neighbourhoods and nonlinear models as shown by the superior accuracy of the ROI/GAM/EXTD combination. This result seems also more important for large basins.

By focusing on the drainage network and basin shape, the new physiographical variables allow integrating more information about the underlying hydrogeological flows and thus, indirectly, to make the link between the groundwater and the surface water flows. This added information allows for a better description of the hydrological dynamics involved and consequently to better flood quantile estimates.

The present study paves the way for several perspectives. In particular, drainage network characteristics should be evaluated further in a wider variety of settings including different climate and catchment geology. The increasing complexity of databases used in RFA to which this research participate, also outlines the need for methodological development that allow a more efficient use of this extensive information, as classical approaches may be limited in this regard. Future research should thus focus on studying how to take advantage of the interaction between the newly proposed variables on quantile estimation, as well as the potential nonlinear impact of the considered variables.

References

Agarwal C (1998) Study of drainage pattern through aerial data in Naugarh area of Varanasi district, UP Journal of the Indian Society of Remote Sensing 26:169-175 doi:https://doi.org/10.1007/BF02990795

Ahmadi R et al. (2006) The geomorphologic responses to hinge migration in the fault-related folds in the Southern Tunisian Atlas Journal of Structural Geology 28:721-728

Alobaidi MH, Marpu PR, Ouarda TB, Chebana F (2015) Regional frequency analysis at ungauged sites using a two-stage resampling generalized ensemble framework Advances in Water Resources 84:103-111 doi:https://doi.org/10.1016/j.advwatres.2015.07.019

Ameri AA, Pourghasemi HR, Cerda A (2018) Erodibility prioritization of sub-watersheds using morphometric parameters analysis and its mapping: A comparison among TOPSIS, VIKOR, SAW, and CF multi-criteria decision making models Science of The Total Environment 613:1385-1400 doi:https://doi.org/10.1016/j.scitotenv.2017.09.210

Aziz K, Rahman A, Fang G, Shrestha S (2014) Application of artificial neural networks in regional flood frequency analysis: a case study for Australia Stochastic environmental research and risk assessment 28:541-554

Bajabaa S, Masoud M, Al-Amri N (2014) Flash flood hazard mapping based on quantitative hydrology, geomorphology and GIS techniques (case study of Wadi Al Lith, Saudi Arabia) Arabian Journal of Geosciences 7:2469-2481 doi:https://doi.org/10.1007/s12517-013-0941-2

Baumgardner RW (1987) Morphometric Studies of Subhumid and Semiarid Drainage Basins, Texas Panhandle and Northeastern New Mexico. vol 163. Bureau of Economic Geology, University of Texas at Austin,

Beck HE, Dijk AI, Miralles DG, Jeu RA, McVicar TR, Schellekens J (2013) Global patterns in base flow index and recession based on streamflow observations from 3394 catchments Water Resources Research 49:7843-7863 doi:https://doi.org/10.1002/2013WR013918

Biswas S, Sudhakar S, Desai V (1999) Prioritisation of subwatersheds based on morphometric analysis of drainage basin: A remote sensing and GIS approach Journal of the Indian society of remote sensing 27:155-166 doi:https://doi.org/10.1007/BF02991569

Blyth K, Rodda J (1973) A stream length study Water Resources Research 9:1454-1461 doi: https://doi.org/10.1029/WR009i005p01454

Burn DH (1990) Evaluation of regional flood frequency analysis with a region of influence approach Water Resources Research 26:2257-2265 doi: https://doi.org/10.1029/WR026i010p02257

Castellarin A (2014) Regional prediction of flow-duration curves using a three-dimensional kriging Journal of hydrology 513:179-191 doi:https://doi.org/10.1016/j.jhydrol.2014.03.050

Castiglioni S, Castellarin A, Montanari A (2009) Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation Journal of hydrology 378:272-280 doi:https://doi.org/10.1016/j.jhydrol.2009.09.032

Chebana F, Charron C, Ouarda TBMJ, Martel B (2014) Regional frequency analysis at ungauged sites with the generalized additive model Journal of Hydrometeorology 15:2418-2428 doi:https://doi.org/10.1175/JHM-D-14-0060.1

Chebana F, Ouarda TB (2007) Multivariate L-moment homogeneity test Water resources research 43 doi:doi:10.1029/2006WR005639
Chokmani K, Ouarda TB (2004) Physiographical space-based kriging for regional flood frequency estimation at ungauged sites Water Resources Research 40
Chow V (1964) Handbook of applied hydrology: a compendium of water-resources technology vol 1. McGraw-Hill,
Dar RA, Romshoo SA, Chandra R, Ahmad I (2014) Tectono-geomorphic study of the Karewa Basin of Kashmir Valley Journal of Asian Earth Sciences 92:143-156 doi:https://doi.org/10.1016/j.jseaes.2014.06.018
Dawson CW, Abrahart RJ, Shamseldin AY, Wilby RL (2006) Flood estimation at ungauged sites using artificial neural networks Journal of hydrology 319:391-409 doi:https://doi.org/10.1016/j.jhydrol.2005.07.032
Dodangeh E, Soltani S, Sarhadi A, Shiau JT (2014) Application of L-moments and Bayesian inference for low-flow regionalization in Sefidroud basin, Iran Hydrological Processes 28:1663-1676 doi:https://doi.org/10.1002/hyp.9711
Durocher M, Chebana F, Ouarda TB (2015) A nonlinear approach to regional flood frequency analysis using projection pursuit regression Journal of Hydrometeorology 16:1561-1574 doi:https://doi.org/10.1175/JHM-D-14-0227.1
Faniran A (1968) The index of drainage intensity-A provisional new drainage factor Australian Journal of Science 31:328-330
Flavell D (2012) Design flood estimation in Western Australia Australian Journal of Water Resources 16:1-20
Gajbhiye S (2015) Morphometric analysis of a Shakkar river catchment using RS and GIS International Journal of u-and e-Service, Science and Technology 8:11-24 doi:http://dx.doi.org/10.14257/ijunesst.2015.8.2.02
Girard C, Ouarda TB, Bobée B (2004) Étude du biais dans le modèle log-linéaire d'estimation régionale. Canadian Journal of Civil Engineering 31:361-368 doi:https://doi.org/10.1139/l03-099
Goswami M, O'connor K, Bhattarai K (2007) Development of regionalisation procedures using a multi-model approach for flow simulation in an ungauged catchment Journal of Hydrology 333:517-531 doi:https://doi.org/10.1016/j.jhydrol.2006.09.018
GREHYS (1996) Presentation and review of some methods for regional flood frequency analysis Journal of hydrology(Amsterdam) 186:63-84
Griffis V, Stedinger J (2007) The use of GLS regression in regional hydrologic analyses Journal of Hydrology 344:82-95 doi:https://doi.org/10.1016/j.jhydrol.2007.06.023
Haddad K, Rahman A (2012) Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework–Quantile Regression vs. Parameter Regression Technique Journal of Hydrology 430:142-161 doi:https://doi.org/10.1016/j.jhydrol.2012.02.012
Hadley R, Schumm S (1961) Sediment sources and drainage basin characteristics in upper Cheyenne River basin US Geological Survey Water-Supply Paper 1531:198
Hailegeorgis TT, Alfredsen K (2017) Regional flood frequency analysis and prediction in ungauged basins including estimation of major uncertainties for mid-Norway Journal of Hydrology: Regional Studies 9:104-126 doi:https://doi.org/10.1016/j.ejrh.2016.11.004

Hajam RA, Hamid A, Bhat S (2013) Application of morphometric analysis for geo-hydrological studies using geo-spatial technology-A case study of Vishav Drainage Basin Hydrol Current Res 4:2 doi:http://dx.doi.org/10.4172/2157-7587.1000157

Hamed Y et al. (2014) Use of geochemical, isotopic, and age tracer data to develop models of groundwater flow: A case study of Gafsa mining basin-Southern Tunisia Journal of African Earth Sciences 100:418-436

Hastie TaRT (1986) Generalized additive models. Stat Sci 1:297–310 doi:doi:10.1214/ss/1177013604.

Heinze G, Wallisch C, Dunkler D (2018) Variable selection–a review and recommendations for the practicing statistician Biometrical Journal 60:431-449

Horton RE (1932) Drainage-basin characteristics Eos, transactions american geophysical union 13:350-361 doi:https://doi.org/10.1029/TR013i001p00350

Horton RE (1945) Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology Geological society of America bulletin 56:275-370 doi:https://doi.org/10.1130/0016-7606(1945)56[275:EDOSAT]2.0.CO;2

Hosking JRM, Wallis JR (2005) Regional frequency analysis: an approach based on L-moments. Cambridge University Press,

Howard AD (1990) Theoretical model of optimal drainage networks Water Resources Research 26:2107-2117 doi:https://doi.org/10.1029/WR026i009p02107

Jenson SK, Domingue JO (1988) Extracting topographic structure from digital elevation data for geographic information system analysis Photogrammetric engineering and remote sensing 54:1593-1600

Jung K, Marpu PR, Ouarda TB (2015) Improved classification of drainage networks using junction angles and secondary tributary lengths Geomorphology 239:41-47 doi:https://doi.org/10.1016/j.geomorph.2015.03.004

Jung K, Marpu PR, Ouarda TB (2017) Impact of river network type on the time of concentration Arabian Journal of Geosciences 10:546 doi:https://doi.org/10.1007/s12517-017-3323-3

Kaliraj S, Chandrasekar N, Magesh N (2015) Morphometric analysis of the River Thamirabarani sub-basin in Kanyakumari District, South west coast of Tamil Nadu, India, using remote sensing and GIS Environmental Earth Sciences 73:7375-7401 doi:https://doi.org/10.1007/s12665-014-3914-1

Kouider A, Gingras H, Ouarda T, Ristic-Rudolf Z, Bobée B (2002a) Analyse fréquentielle locale et régionale et cartographie des crues au Québec Rep R-627-el

Kouider A, Ouarda T, Bobée B (2002b) Analyse fréquentielle locale des crues au Québec doi:http://espace.inrs.ca/365/1/T000342.pdf

Latt ZZ, Wittenberg H, Urban B (2015) Clustering hydrological homogeneous regions and neural network based index flood estimation for ungauged catchments: an example of the Chindwin River in Myanmar Water resources management 29:913-928 doi:https://doi.org/10.1007/s11269-014-0851-4

Leclerc M, Ouarda TB (2007) Non-stationary regional flood frequency analysis at ungauged sites Journal of hydrology 343:254-265 doi:https://doi.org/10.1016/j.jhydrol.2007.06.021

Máčka Z (2001) Determination of texture of topography from large scale contour maps.

Magesh N, Chandrasekar N, Soundranayagam JP (2011) Morphometric evaluation of Papanasam and Manimuthar watersheds, parts of Western Ghats, Tirunelveli district, Tamil Nadu, India: a GIS approach Environmental Earth Sciences 64:373-381 doi:https://doi.org/10.1007/s12665-010-0860-4

Marchi L, Dalla Fontana G (2005) GIS morphometric indicators for the analysis of sediment dynamics in mountain basins Environmental Geology 48:218-228 doi:https://doi.org/10.1007/s00254-005-1292-4

Melton MA (1957) An analysis of the relations among elements of climate, surface properties, and geomorphology. COLUMBIA UNIV NEW YORK,

Mesa L (2006) Morphometric analysis of a subtropical Andean basin (Tucuman, Argentina) Environmental Geology 50:1235-1242 doi:https://doi.org/10.1007/s00254-006-0297-y

Miller VC (1953) quantitative geomorphic study of drainage basin characteristics in the Clinch Mountain area, Virginia and Tennessee Technical report (Columbia University Department of Geology); no 3

Msilini A, Masselot P, Ouarda TBMJ (2020) Regional Frequency Analysis at Ungauged Sites with Multivariate Adaptive Regression Splines Journal of Hydrometeorology:1-1 doi:10.1175/jhm-d-19-0213.1

Muttiah RS, Srinivasan R, Allen PM (1997) PREDICTION OF TWO-YEAR PEAK STREAM-DISCHARGES USING NEURAL NETWORKS JAWRA Journal of the American Water Resources Association 33:625-630 doi:https://doi.org/10.1111/j.1752-1688.1997.tb03537.x

O'Callaghan JF, Mark DM (1984) The extraction of drainage networks from digital elevation data Computer vision, graphics, and image processing 28:323-344 doi:https://doi.org/10.1016/S0734-189X(84)80011-0

Odry J, Arnaud P (2017) Comparison of Flood Frequency Analysis Methods for Ungauged Catchments in France Geosciences 7:88 doi:https://doi.org/10.3390/geosciences7030088

Ouali D, Chebana F, Ouarda TBMJ (2016) Non-linear canonical correlation analysis in regional frequency analysis Stochastic environmental research and risk assessment 30:449-462 doi:https://doi.org/10.1007/s00477-015-1092-7

Ouali D, Chebana F, Ouarda TBMJ (2017) Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites Journal of Advances in Modeling Earth Systems 9:1292-1306 doi:https://doi.org/10.1002/2016MS000830

Ouarda TBMJ (2016) Regional flood frequency modeling chapter 77, in: VP Singh, (Ed) Chow's Handbook of Applied Hydrology,3rd Edition, Mc-Graw Hill, New York:pp. 77.71-77.78, ISBN 978-970-907-183509-183501

Ouarda TBMJ, Charron C, Hundecha Y, St-Hilaire A, Chebana F (2018) Introduction of the GAM model for regional low-flow frequency analysis at ungauged basins and comparison with commonly used approaches Environmental Modelling & Software 109:256-271 doi:https://doi.org/10.1016/j.envsoft.2018.08.031

Ouarda TBMJ, Girard C, Cavadias GS, Bobée B (2001) Regional flood frequency estimation with canonical correlation analysis Journal of Hydrology 254:157-173 doi:https://doi.org/10.1016/S0022-1694(01)00488-7

Ouarda TBMJ, Haché M, Bruneau P, Bobée B (2000) Regional flood peak and volume estimation in northern Canadian basin Journal of cold regions engineering 14:176-191 doi:https://doi.org/10.1061/(ASCE)0887-381X(2000)14:4(176)

Pandey G, Nguyen V-T-V (1999) A comparative study of regression based methods in regional flood frequency analysis Journal of Hydrology 225:92-101 doi:https://doi.org/10.1016/S0022-1694(99)00135-3

Pareta K, Pareta U (2011) Quantitative morphometric analysis of a watershed of Yamuna basin, India using ASTER (DEM) data and GIS International journal of Geomatics and Geosciences 2:248

Patton PC (1988) Drainage basin morphometry and floods Flood Geomorphology John Wiley & Sons New York 1988 p 51-64 11 fig, 1 tab, 67 ref

Patton PC, Baker VR (1976) Morphometry and floods in small drainage basins subject to diverse hydrogeomorphic controls Water Resources Research 12:941-952 doi:https://doi.org/10.1029/WR012i005p00941

Rahman A (2005) A quantile regression technique to estimate design floods for ungauged catchments in south-east Australia Australian Journal of Water Resources 9:81-89 doi:https://doi.org/10.1080/13241583.2005.11465266

Rahman A, Charron C, Ouarda TBMJ, Chebana F (2018) Development of regional flood frequency analysis techniques using generalized additive models for Australia Stochastic Environmental Research and Risk Assessment 32:123-139 doi:https://doi.org/10.1007/s00477-017-1384-1

Rai PK, Mishra VN, Mohan K (2017) A study of morphometric evaluation of the Son basin, India using geospatial approach Remote Sensing Applications: Society and Environment 7:9-20 doi:https://doi.org/10.1016/j.rsase.2017.05.001

Ratnam KN, Srivastava Y, Rao VV, Amminedu E, Murthy K (2005) Check dam positioning by prioritization of micro-watersheds using SYI model and morphometric analysis—remote sensing and GIS perspective Journal of the Indian society of remote sensing 33:25 doi:https://doi.org/10.1007/BF02989988

Reddy GPO, Maji AK, Gajbhiye KS (2004) Drainage morphometry and its influence on landform characteristics in a basaltic terrain, Central India–a remote sensing and GIS approach International Journal of Applied Earth Observation and Geoinformation 6:1-16 doi:https://doi.org/10.1016/j.jag.2004.06.003

Requena AI, Ouarda TBMJ, Chebana F (2018) Low-flow frequency analysis at ungauged sites based on regionally estimated streamflows Journal of Hydrology doi:https://doi.org/10.1016/j.jhydrol.2018.06.016

Ridolfi E, Rianna M, Trani G, Alfonso L, Di Baldassarre G, Napolitano F, Russo F (2016) A new methodology to define homogeneous regions through an entropy based clustering method Advances in Water Resources 96:237-250 doi:https://doi.org/10.1016/j.advwatres.2016.07.007

Schumm SA (1956) Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey Geological society of America bulletin 67:597-646 doi:https://doi.org/10.1130/0016-7606(1956)67[597:EODSAS]2.0.CO;2

Seckin N (2011) Modeling flood discharge at ungauged sites across Turkey using neuro-fuzzy and neural networks Journal of hydroinformatics 13:842-849 doi:https://doi.org/10.2166/hydro.2010.046

Seidou O, Ouarda T, Barbet M, Bruneau P, Bobee B (2006) A parametric Bayesian combination of local and regional information in flood frequency analysis Water resources research 42 doi: https://doi.org/10.1029/2005WR004397

Shaban A, Khawlie M, Abdallah C, Awad M (2005) Hydrological and watershed characteristics of the El-Kabir River, North Lebanon Lakes & Reservoirs: Research & Management 10:93-101 doi:https://doi.org/10.1111/j.1440-1770.2005.00262.x

Shu C, Ouarda T (2008) Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system Journal of Hydrology 349:31-43 doi:https://doi.org/10.1016/j.jhydrol.2007.10.050

Shu C, Ouarda TBMJ (2007) Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space Water Resources Research 43 doi:doi:10.1029/2006WR005142

Sivasena Reddy A, Janga Reddy M (2013) Identification of homogenous regions in rain-fed watershed using Kohonen neural networks ISH Journal of Hydraulic Engineering 19:55-66 doi:https://doi.org/10.1080/09715010.2013.763408

Smith A, Sampson C, Bates P (2015) Regional flood frequency analysis at the global scale Water Resources Research 51:539-553 doi: https://doi.org/10.1002/2014WR015814

Sreedevi P, Sreekanth P, Khan H, Ahmed S (2013) Drainage morphometry and its influence on hydrology in an semi arid region: using SRTM data and GIS Environmental earth sciences 70:839-848 doi:https://doi.org/10.1007/s12665-012-2172-3

Sreedevi P, Subrahmanyam K, Ahmed S (2005) The significance of morphometric analysis for obtaining groundwater potential zones in a structurally controlled terrain Environmental Geology 47:412-420 doi:https://doi.org/10.1007/s00254-004-1166-1

Stall JB, Fok Y-S Discharge as related to stream system morphology. In: International Association of Scientific Hydrology Symposium on River Morphology, Publication, 1967. pp 224-235

Strahler (1953) Revisions of Horton's quantitative factors in erosional terrain Trans Am Geophys Union 34:345

Strahler (1957) Quantitative analysis of watershed geomorphology Eos, Transactions American Geophysical Union 38:913-920 doi:https://doi.org/10.1029/TR038i006p00913

Strahler (1964) Part II. Quantitative geomorphology of drainage basins and channel networks Handbook of Applied Hydrology: McGraw-Hill, New York:4-39

Strahler A (1952) Hypsometric (area-altitude) analysis of erosional topography Geological Society of America Bulletin 63:1117-1142 doi:https://doi.org/10.1130/0016-7606(1952)63[1117:HAAOET]2.0.CO;2

Taofik OK, Innocent B, Christopher N, Jidauna GG, James AS (2017) A Comparative Analysis of Drainage Morphometry on Hydrologic Characteristics of Kereke and Ukoghor Basins on Flood Vulnerability in Makurdi Town, Nigeria Hydrology 5:32 doi:doi: 10.11648/j.hyd.20170503.11

Tarboton DG, Bras RL, Rodriguez-Iturbe I (1991) On the extraction of channel networks from digital elevation data Hydrological processes 5:81-100

Tramblay Y, Ouarda TBMJ, St-Hilaire A, Poulin J (2010) Regional estimation of extreme suspended sediment concentrations using watershed characteristics Journal of Hydrology 380:305-317 doi:https://doi.org/10.1016/j.jhydrol.2009.11.006

Tsakiris G, Nalbantis I, Cavadias G (2011) Regionalization of low flows based on canonical correlation analysis Advances in water resources 34:865-872 doi:https://doi.org/10.1016/j.advwatres.2011.04.007
Verstappen H (1983) Applied Geomorphology. Elsevier, Amsterdam-Oxford-New York,
Vijith H, Satheesh R (2006) GIS based morphometric analysis of two major upland sub-watersheds of Meenachil river in Kerala Journal of the Indian Society of Remote Sensing 34:181-185 doi:https://doi.org/10.1007/BF02991823
Wan Jaafar W, Liu J, Han D (2011) Input variable selection for median flood regionalization Water Resources Research 47 doi:10.1029/2011WR010436,
Wazneh H, Chebana F, Ouarda TB (2016) Identification of hydrological neighborhoods for regional flood frequency analysis using statistical depth function Advances in water resources 94:251-263 doi:https://doi.org/10.1016/j.advwatres.2016.05.013
Wood (2006) Generalized additive models: an introduction with R. CRC press,
Youssef AM, Pradhan B, Hassan AM (2011) Flash flood risk estimation along the St. Katherine road, southern Sinai, Egypt using GIS based morphometry and satellite imagery Environmental Earth Sciences 62:611-623 doi:https://doi.org/10.1007/s12665-010-0551-1

List of tables

List of Figures

Table 1 Predictor variables used in a number of previous regionalization studies.

References	Country	Predictor variables adopted
(Muttiah et al. 1997)	USA	Catchment areas, mean annual rainfall, and mean basin elevation.
(Rahman 2005)	Australia	Catchment area, design rainfall intensity, mean annual rainfall, mean annual rain days, mean annual Class A pan evaporation, mainstream slope, lemniscate shape, river bed elevation at the gauging station, maximum elevation difference in the basin, stream density, forest cover, and fraction quaternary sediment area.
(Dawson et al. 2006)	United Kingdom	Catchment area, base flow index, standard percentage runoff, index of flood attenuation attributable to reservoirs and lakes, standard period (1961-1990) average annual rainfall, median annual maximum 1-day rainfall, median annual maximum 2-day rainfall, median annual maximum 1-h rainfall, mean Soil Moisture Deficit for 1941–1970, proportion of time when Soil Moisture Deficit <6 mm during 1961–1990, longest drainage path, mean distance between each node (on a regular 50 m grid) and catchment outlet, mean altitude of catchment above sea level, mean of all inter-nodal slopes in the catchment, invariability of slope directions, extent of urban and suburban land cover in 1990.
(Leclerc and Ouarda 2007)	Canada	Catchment area, gauging station latitude, gauging station longitude, mean total winter/spring precipitation, mean winter/spring maximum air temperature.
(Leclerc and Ouarda 2007)	USA	Catchment area, mean annual rainfall, runoff measured, mainstream slope, main-channel length, forest cover, and storage measured as the percent of the catchment area.
(Griffis and Stedinger 2007)	Canada	Catchment area, mean annual rainfall, mean basin slope, the fraction of the basin area covered with lakes and annual mean degree days below 0 °C.
(Shu and Ouarda 2008) (Alobaidi et al. 2015) (Durocher et al. 2015) (Ouali et al. 2016) (Wazneh et al. 2016)	Mexico	Drainage area, mean annual precipitation, final altitude of the mainstream and slope of the main stream.
(Castiglioni et al. 2009)	Italy	Drainage area, main channel length, the percentage of permeable area, maximum, minimum and mean elevations, average elevation relative to the minimum elevations, concentration time, mean annual temperature and mean annual temperature precipitation.
(Wan Jaafar et al. 2011)	England	Catchment area, longest flow path, basin length, basin perimeter, form factor, average slope, maximum relief, relief ratio, drainage density, stream frequency, bifurcation ratio, length of overland flow, land use (agriculture), land use (forest), land use (residential), land use (water and wetland), soil type (coarse), soil type (medium), soil type (medium fine), soil type (fine), soil type (peat soil) and rainfall.
(Seckin 2011)	Turkey	Drainage area, elevation, latitude, longitude and return period.
(Flavell 2012)	Australia	Catchment area, mean annual rainfall, mainstream slope, main-channel length, and 12 and 24 h statistical rainfall totals.
(Haddad and Rahman 2012)	Australia	Catchment area, design rainfall intensity, mean annual rainfall, mean annual evapotranspiration, stream density, mainstream slope, stream length, and forest cover.
(Beck et al. 2013)	3394 basins around the world.	Humidity index, mean annual precipitation, precipitation seasonality, mean annual potential evaporation, potential evaporation seasonality, seasonal correlation between water supply and demand, mean annual air temperature, mean snow water equivalent depth, mean elevation, mean surface slope, fraction of open water, fraction of forest, mean Normalized Difference Vegetation Index (NDVI), mean permeability of consolidated and unconsolidated geologic units below the soil, mean gravel content, mean sand content, mean silt content, mean clay content.
(Aziz et al. 2014)	Australia	Catchment area, design rainfall intensity values I(tc) with where ARI = 2, 5, 10, 20, 50 and 100 years return period (tc = time of concentration), mean annual rainfall, mean annual areal evapotranspiration, and mainstream slope.
(Castellarin 2014)	Italy	Drainage area, mainstream length, maximum, mean, and minimum elevations, mean annual temperature, net annual precipitation, annual potential evapotranspiration, coefficients of L variation of the net annual precipitation, annual potential evapotranspiration, percentage of previous area, the long-term mean daily stream flow standardized by the catchment area, and the daily stream flow associated with a duration of 355 days standardized by catchment.
(Latt et al. 2015)	Myanmar	Catchment area, mean basin elevation, basin slope, basin length, shape factor, soil conservation curve number, time of concentration, mean annual rainfall.
(Smith et al. 2015)	Several basins across the world.	Catchment area, average annual rainfall and the upstream catchment slope.
(Ridolfi et al. 2016)	Italy	Catchment area, the previous area, the maximum and mean altitudes, the gauge elevation, the mean slope, the length and the slope of the longest drainage path (LDP), annual mean precipitation, and the coordinates of each site.
(Odry and Arnaud 2017)	France	Aridity index, annual mean evapotranspiration, annual mean solid precipitation, annual mean liquid precipitation, annual mean temperature, annual mean soil moisture, mean soil moisture prior to a rainy event (>20 mm), mean duration of rainfall events, mean number of rainfall events per season, mean intensity of rainfall events, river network density, mean elevation, mean slope, capacity of the production reservoir of a lumped rainfall-runoff model, presence of sand bedding, presence of rock bedding, low infiltration capacity class, medium infiltration capacity class, high infiltration capacity class, forest cover, arable cover, grassland cover, catchment area, catchment eastening (X) and catchment northening (Y).
(Hailegeorgis and Alfredsen 2017)	Mid-Norway	Catchment area,
(Requena et al. 2018)	Canada	Catchment area, fraction of the catchment controlled by lakes, fraction of the catchment occupied by forest, annual mean degree-days below 0 °C, summer mean liquid precipitation, curve number and average number of days with mean temperature greater than 27 °C.
(Rahman et al. 2018)	Australia	Catchment area, catchment shape factor, main stream slope, stream density, percentage of catchment covered by forest, rainfall intensity (6 h duration and 2 year return period), mean annual rainfall and mean annual potential evapotranspiration.

Table 2 Morphometric variables definitions.

Morphometric variables	Formula / Relationship	Reference
Stream order (u) (*)	Hierarchical order	(Horton 1945; Strahler 1957)
Stream Length (Lu) (*)	Length of stream	(Horton 1945)
Texture ratio (RT)	$RT = \frac{N_1}{P}$, where N_1= the number of first order streams and P = perimeter (km).	(Schumm 1956)
Circularity Ratio (RC) (*)	$RC = 4\pi\left(\frac{A}{P^2}\right)$, where A = area of the basin (km^2), P = perimeter of the basin (km) and $\pi = 3.1415$.	(Miller 1953)
Stream length ratio (RL)	$RL = \frac{MLu}{Lu\ -1}$, where MLu = the average stream length of a given order u (km) and MLu-1 = the average stream length of the next lower order (km).	(Horton 1945)
Mean stream length ratio (MRL)	MRL = Average of the stream length ratio of all orders	(Horton 1945)
Bifurcation ratio (RB)	$RB = \frac{Nu}{Nu+1}$, Nu = the number of stream segments of order u, Nu + 1 = the number of stream segments of the next higher order.	(Horton 1945)
Mean bifurcation ratio (MRB) (*)	MRB = Average of bifurcation ratios of all orders	(Strahler 1957)
Weighted mean bifurcation ratio (WMRB)	$WMRB = \frac{\sum RB_{\left(\frac{u}{u+1}\right)}(N_u+N_{u+1})}{\sum N}$, where $RB_{\left(\frac{u}{u+1}\right)}$ = the bifurcation ratio between each successive pair of orders, Nu = the total number of stream segments of order u and $\sum N$ = the total number of streams involved in the ratio.	(Schumm 1956; Strahler 1953)
RHO coefficient (ρ)	$\rho = \frac{RL}{RB}$	(Horton 1945)
RHO $_{WMRB}$ coefficient (ρ_{WMRB})	$\rho_{WMRB} = \frac{RL}{WMRB}$	(Horton 1945; Schumm 1956; Strahler 1953)
Drainage density (DD) (*)	$DD = \frac{L}{A}$, where L = total stream length of all orders (km), A = area of the basin (km^2).	(Horton 1932; Horton 1945)
Stream frequency (FS) (*)	$FS = \frac{N}{A}$, where N = total number of streams of all orders and A = area of the basin (km^2).	(Horton 1932; Horton 1945)
Infiltration number (IF)	$IF = DD \times FS$	(Faniran 1968)
Basin Relief (BH)	The highest elevation of the basin - Lowest elevation of the basin (km)	(Schumm 1956; Strahler 1957)
Ruggedness number (RN)	$RN = BH \times DD$, where BH = Basin relief and DD = Drainage density.	(Melton 1957)
PN1	Percentage of first-order streams	(Patton and Baker 1976)
PL1	Percentage of first-order stream lengths	(Blyth and Rodda 1973)

(*) Variables previously used in regional hydrological frequency analysis studies, but not used with the Quebec data base.

Table 3 Correlation between hydrological and physiographical variables.

	QS_{10}	QS_{50}	QS_{100}
AREA	-0.46	-0.45	-0.44
MBS	0.47	0.46	0.46
PLAKE	-0.67	-0.65	-0.63
MATP	0.68	0.64	0.62
DDBZ	-0.60	-0.60	-0.59
LONGC	0.47	0.45	0.44
RT	-0.53	-0.52	-0.51
RC	0.68	0.66	0.65

Table 4 Variables selection results for QS_{10} case (with different methods).

	STA								**EXTD**							
Models / Variables	**LLRM**				**GAM**				**LLRM**				**GAM**			
	Fd	**Bd**	**Sw**	**Sh**	**Fd**	**Bd**	**Sw**	**Sh**	**Fd**	**Bd**	**Sw**	**Sh**	**Fd**	**Bd**	**Sw**	**Sh**
AREA	*	*	*	*	*	*	*	*	*	*	*	*				
MCL				*				*				*		*		*
MCS				*	*							*				
MBS	*	*	*	*	*		*	*					*			
PFOR	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*
PLAKE	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
MATP	*				*			*	*				*	*		
MALP	*	*	*	*	*	*		*	*	*	*	*	*		*	
MASP					*				*	*	*					*
MALPS				*			*					*				*
DDBZ	*	*	*	*	*	*		*	*	*	*	*		*		
LATC														*		
LONGC	*	*	*	*	*	*	*	*	*	*	*	*			*	*
RT									*		*		*		*	*
RC												*	*	*	*	*
MRL									*	*	*	*	*	*	*	*
MRB									*	*	*		*			
WMRB									*	*	*	*	*		*	*
ρ_{WMRB}																
DD									*	*	*	*	*	*	*	*
FS										*				*		
IF										*			*	*		
RN									*	*	*	*	*	*		*
PN1												*		*		*
PL1										*			*			

Fd is Forward ; Bd is backward; Sw is stepwise selection and Sh is shrinkage approach selection.

Table 5 Descriptive statistics of new physiographical variables.

Variable	Min	Mean	Max	STD.dev
DD (Km^{-1})	2.41	2.96	4.73	0.34
FS (Km^{-2})	7.34	9.74	11.86	0.97
IF (Km^{-3})	17.69	29.26	67.09	6.56
RT (Km^{-1})	8.09	32.11	131.84	21.41
MRB	1.67	2.40	17.27	2.08
WMRB	1.95	2.08	4.14	0.24
MRL	0.85	0.97	1.11	0.05
ρ_{WMRB}	0.23	0.47	0.55	0.04
RN	0.20	1.89	7.48	1.03
RC	0.06	0.18	0.46	0.08
PN1 (%)	50.12	50.41	52.50	0.30
PL1 (%)	44.09	52.89	66.36	4.10

Table 6 Explanatory variables selected for the various regression models.

Regional models	Quantile	Selected predictor variables
ALL/LLRM/STA,CCA/LLRM/STA,ROI/LLRM/STA	QS_{10}	AREA, MBS, PFOR, PLAKE, MALP, DDBZ, LONGC
	QS_{50}	AREA, MBS, PFOR, PLAKE, MALP, DDBZ, LONGC
	QS_{100}	AREA, MBS, PFOR, PLAKE, MATP, MALP, LONGC
ALL/LLRM/EXTD,CCA/LLRM/EXTD,ROI/LLRM/EXTD	QS_{10}	AREA, PFOR, PLAKE, MALP, ***DD, MRL***, LONGC
	QS_{50}	AREA, PFOR, PLAKE, MALP, ***DD, MRL***, LONGC
	QS_{100}	AREA, PFOR, PLAKE, MALP, ***DD, MRL***, LONGC
ALL/GAM/STA,CCA/GAM/STA,ROI/GAM/STA	QS_{10}	AREA, MBS, PLAKE, MALP, MASP, DDBZ, LONGC
	QS_{50}	AREA, MCL, MBS, PLAKE, MALP, DDBZ, LONGC
	QS_{100}	AREA, MCL, MBS, PLAKE, MALP, DDBZ, LONGC
ALL/GAM/EXTD,CCA/GAM/EXTD,ROI/GAM/EXTD	QS_{10}	MCL, PLAKE, MATP, DDBZ, ***DD, RN***, LATC
	QS_{50}	MCL, PLAKE, MALP, DDBZ, ***DD, MRL***, LONGC
	QS_{100}	MCL, PLAKE, MALP, DDBZ, ***DD, MRL***, LONGC

Variables dealing with drainage network characteristics are in bold character.

Table 7 Jackknife Validation Results.

	Quantile	ALL/LLRM		ALL/GAM		CCA/LLRM		CCA/GAM		ROI/LLRM		ROI/GAM	
		STA	**EXTD**	**STA**	**EXTD**	**STA**	**EXTD**	**STA**	**EXTD**	**STA**	**EXTD**	**STA**	**EXTD**
NASH	QS_{10}	0,669	0.641	0.774	0.802	0.799	0.808	0.797	0.837	0.807	0.804	0.829	**0.865**
	QS_{50}	0,620	0.587	0.745	0.754	0.731	0.743	0.762	0.775	0.754	0.750	0.796	**0.816**
	QS_{100}	0.609	0.556	0.715	0.725	0.680	0.706	0.723	0.742	0.703	0.720	0.762	**0.791**
RMSE $[(m^3/s)km^{-2}]$	QS_{10}	0,073	0.076	0.060	0.056	0.057	0.056	0.057	0.051	0.056	0.056	0.053	**0.047**
	QS_{50}	0,109	0.113	0.089	0.087	0.092	0.089	0.086	0.080	0.087	0.088	0.080	**0.076**
	QS_{100}	0.125	0.133	0.107	0.105	0.113	0.108	0.105	0.101	0.109	0.106	0.097	**0.091**
RRMSE (%)	QS_{10}	43.528	41.202	40.937	34.970	37.412	32.760	37.163	30.619	34.418	31.584	34.690	**27.974**
	QS_{50}	48.518	44.891	49.420	36.659	42.232	36.520	43.333	35.086	39.251	34.034	39.365	**27.818**
	QS_{100}	50.682	46.918	51.832	38.630	46.259	38.426	45.678	37.416	41.497	35.214	41.661	**29.235**
BIAIS $[(m^3/s)km^{-2}]$	QS_{10}	0.004	**0.003**	0.005	0.005	0.007	0.008	0.006	0.007	0.008	0.009	**0.003**	0.004
	QS_{50}	0.008	**0.006**	0.008	0.008	0.015	0.017	0.015	0.015	0.013	0.015	**0.006**	0.009
	QS_{100}	0.011	**0.007**	0.011	0.011	0.020	0.022	0.020	0.020	0.015	0.019	0.009	0.012
RBIAIS (%)	QS_{10}	-6,161	-5.936	-5.461	-4.179	-6.023	-4.587	-5.555	-3.871	-3.040	**-0.932**	-4.177	-2.836
	QS_{50}	-7,338	-6.892	-7.047	-4.954	-6.293	-4.238	-5.632	-3.513	-4.358	**-1.175**	-5.487	-2.892
	QS_{100}	-7.782	-7.431	-7.663	-5.472	-6.623	-4.305	-5.780	-3.714	-4.881	**-1.381**	-5.816	-3.172

Best results are in bold character.

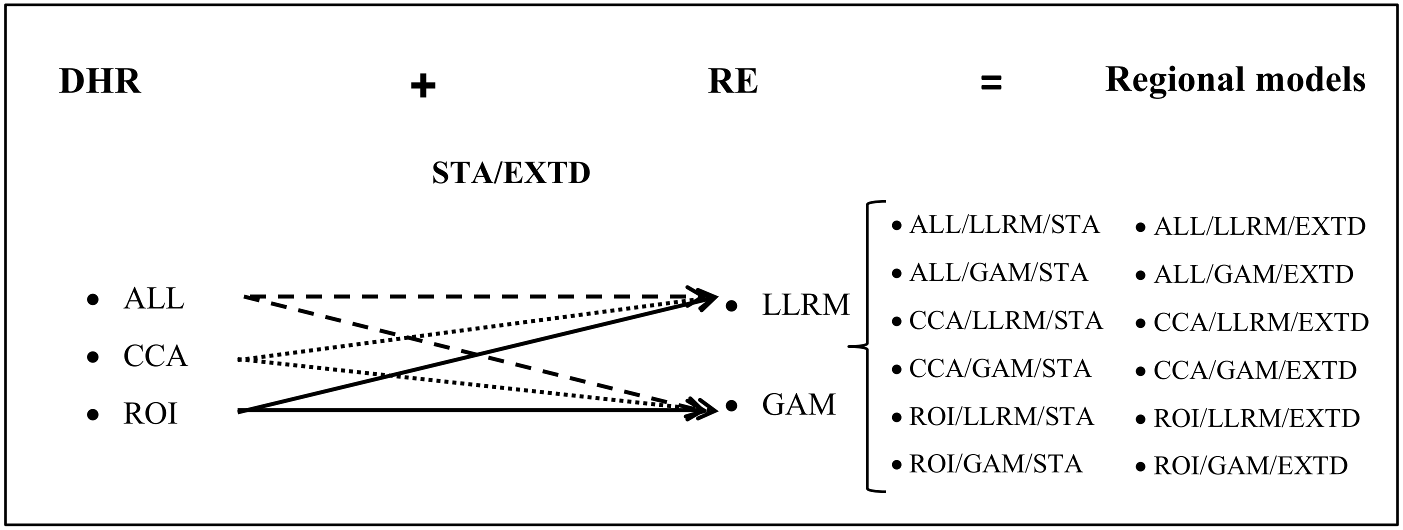


Fig.1 Different combinations and considered models.

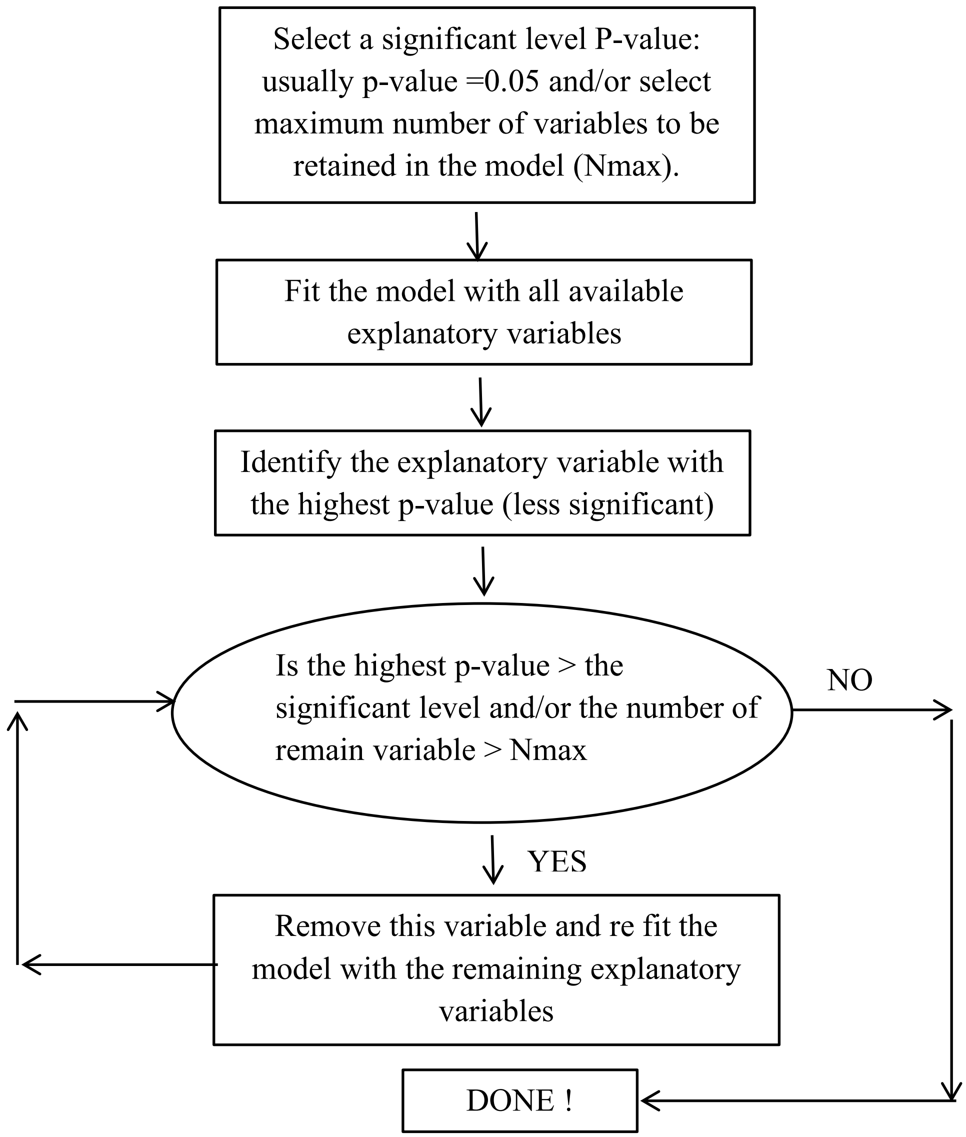
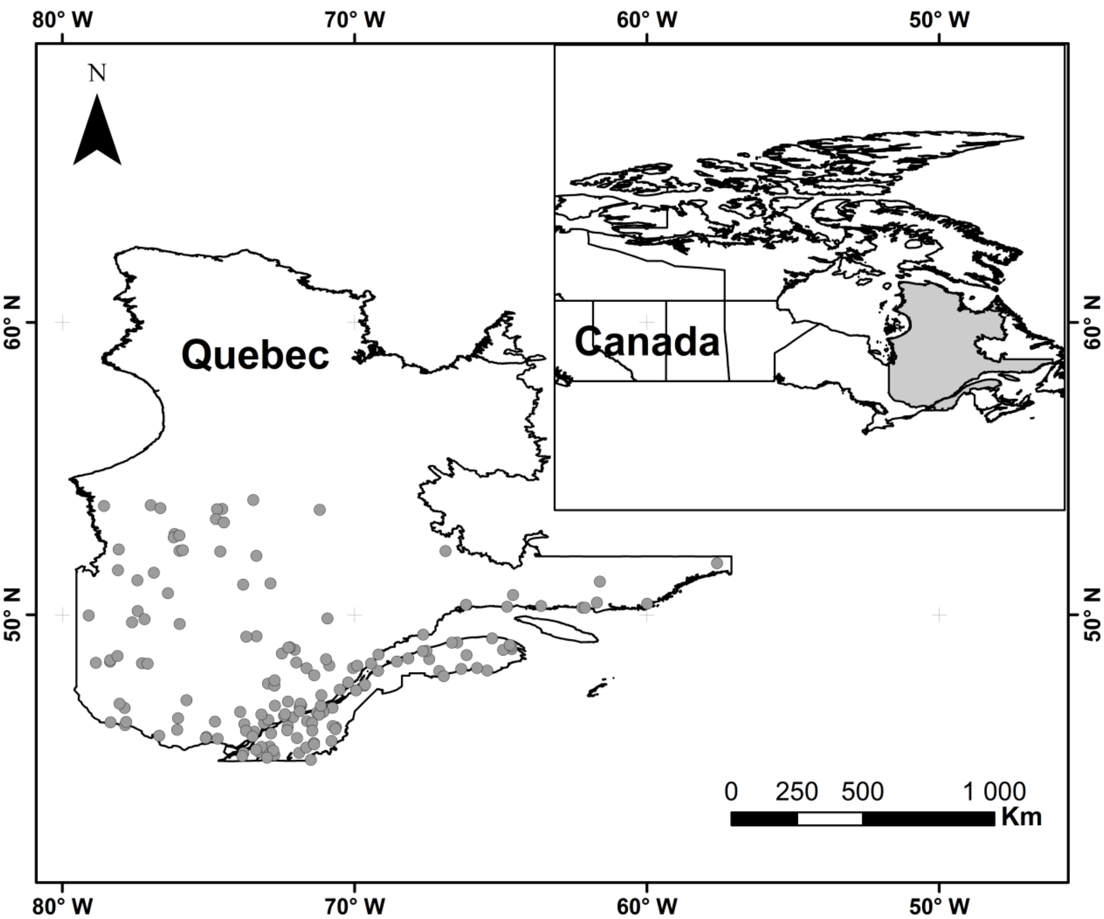
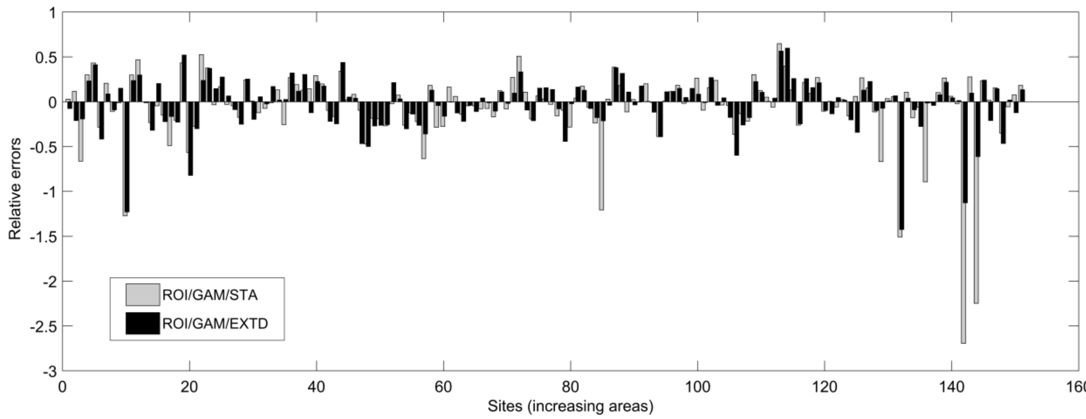
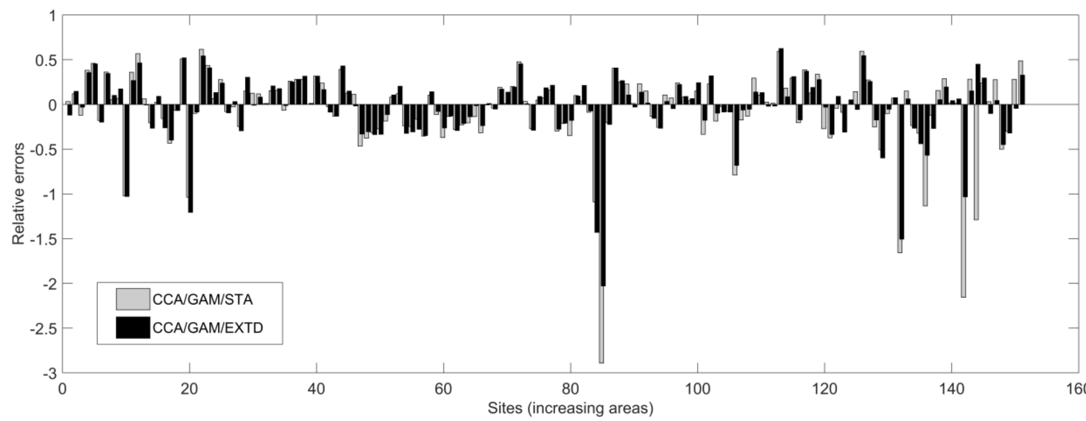


Fig.2 Backward elimination process.

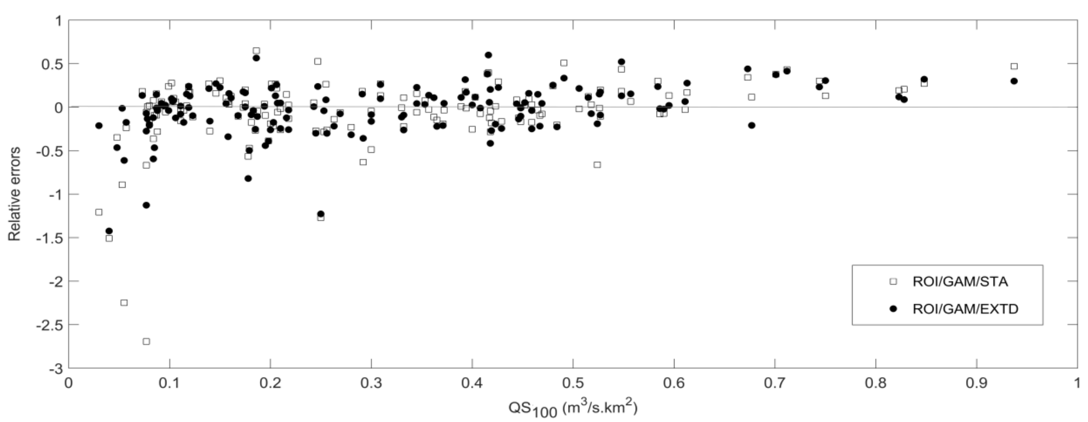

80° W
70° W
60° W
50° W
N
Quebec
Canada
60° N
50° N
0 250 500 1 000
Km

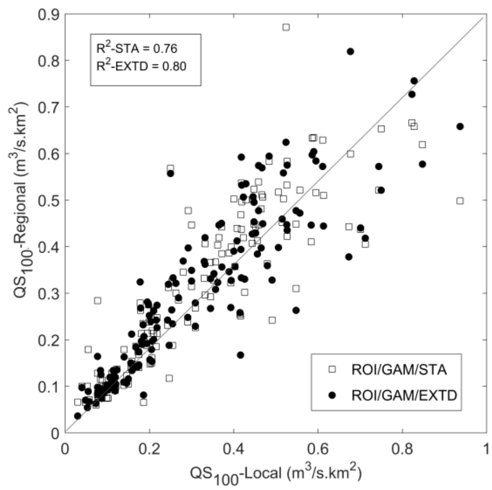


ROI/GAM/EXTD.

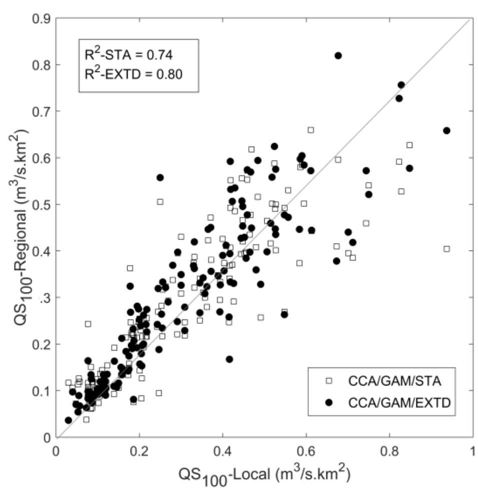


CCA/GAM/EXTD.

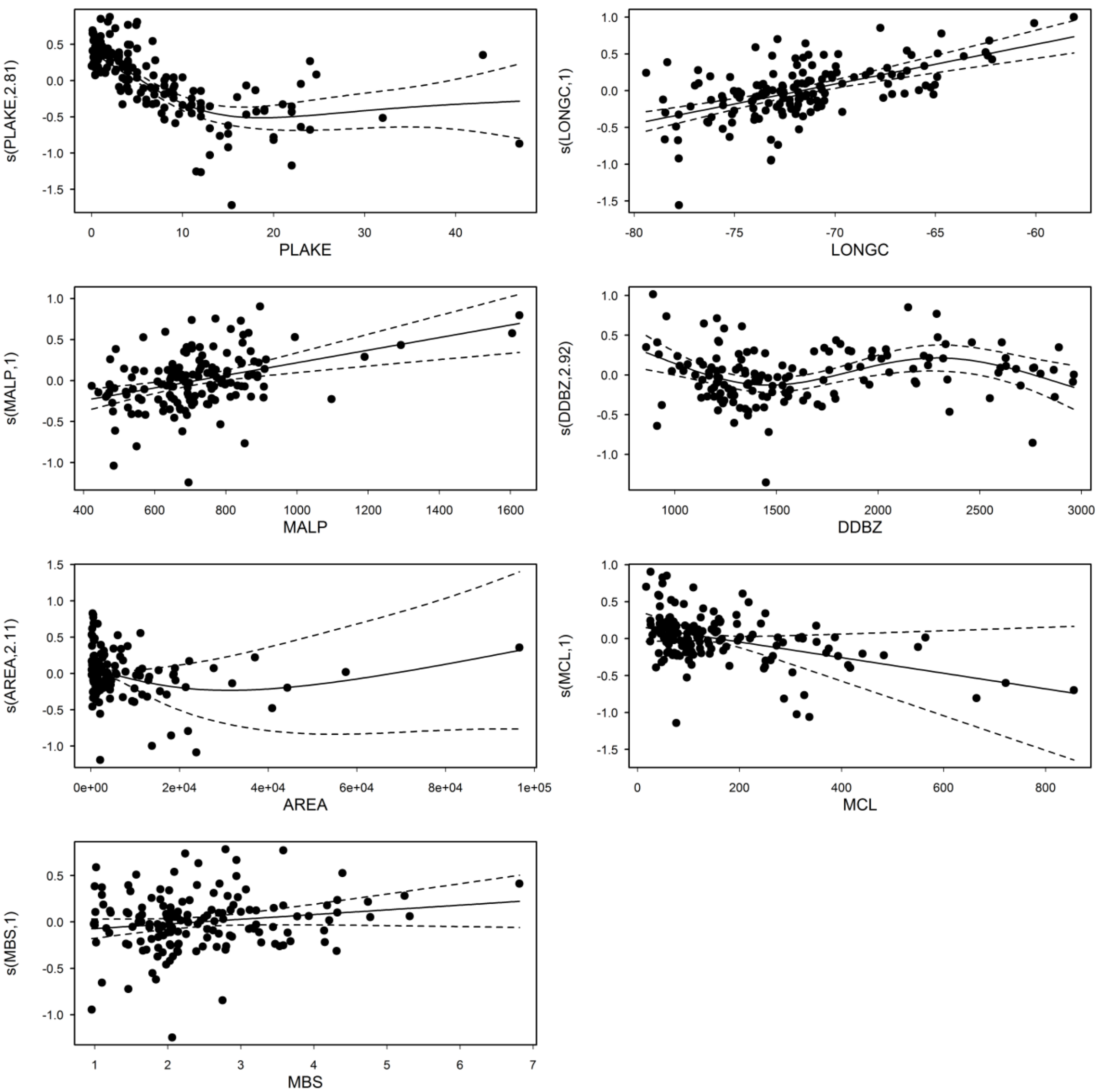




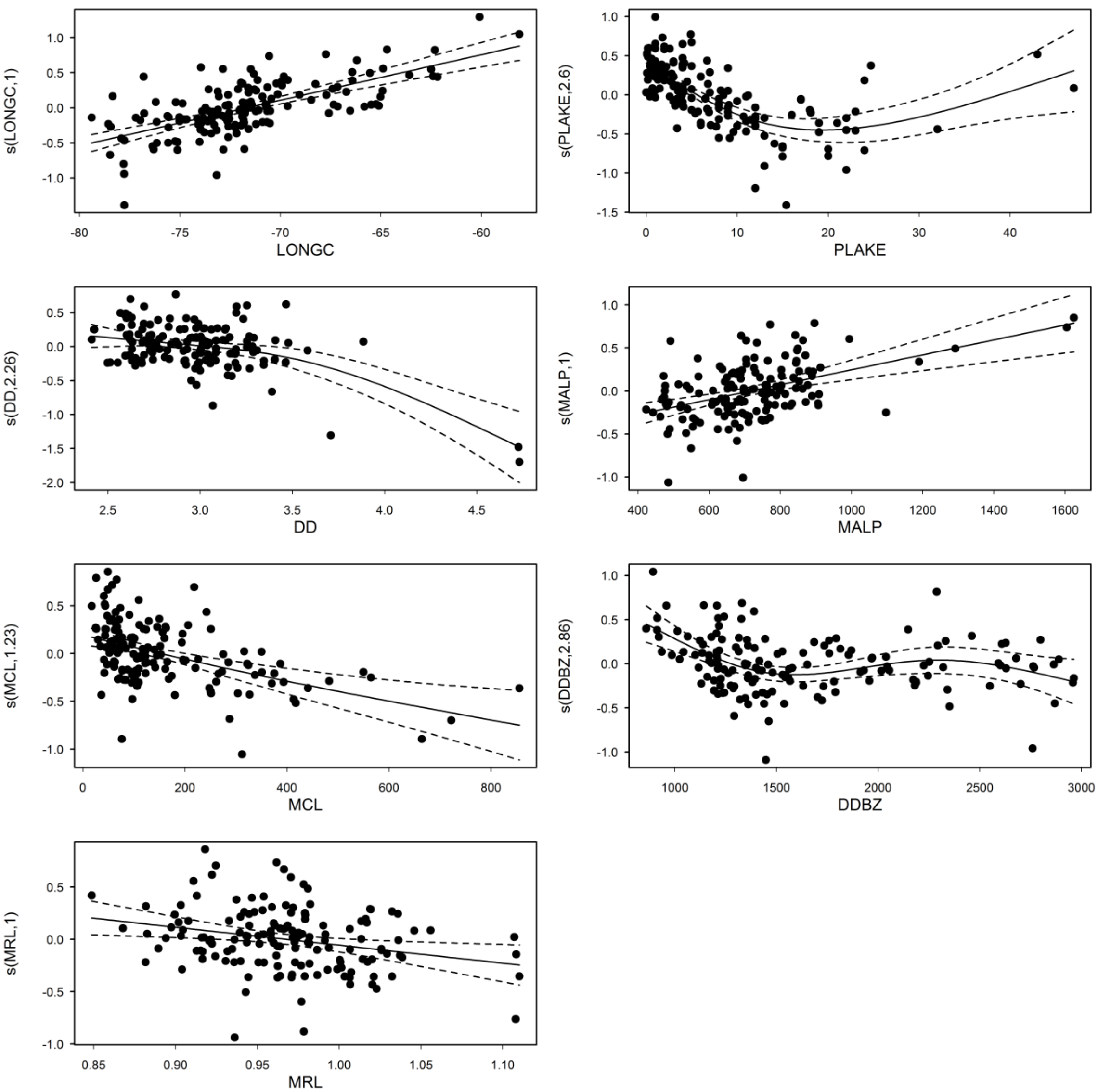
approaches for QS_{100}.



approaches for QS_{100}.



ALL/GAM/STA, CCA/GAM/STA and ROI/GAM/STA. The dotted lines represent the 95% confidence intervals. The vertical axes denote the spline of each explanatory variable.



ALL/GAM/EXTD, CCA/GAM/EXTD and ROI/GAM/EXTD. The dotted lines represent the 95% confidence intervals. The vertical axes denote the spline of each explanatory variable.