Regional Frequency Analysis at Ungauged Sites with Multivariate Adaptive Regression Splines.

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

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

* Corresponding author: Amina Msilini (Amina.Msilini@ete.inrs.ca).

March, 2020

Abstract

Hydrological systems are naturally complex and nonlinear. A large number of variables, many of which not yet well considered in regional frequency analysis (RFA), have a significant impact on hydrological dynamics and consequently on flood quantile estimates. Despite the increasing number of statistical tools used to estimate flood quantiles at ungauged sites, little attention has been dedicated to the development of new regional estimation (RE) models accounting for both nonlinear links and interactions between hydrological and physio-meteorological variables. The aim of this paper is to simultaneously take into account non-linearity and interactions between variables by introducing the multivariate adaptive regression splines (MARS) approach in RFA. The predictive performances of MARS are compared with those obtained by one of the most robust RE models: the generalized additive model (GAM). Both approaches are applied to two datasets covering 151 hydrometric stations in the province of Quebec (Canada): a standard dataset (STA) containing commonly used variables and an extended dataset (EXTD) combining STA with additional variables dealing with drainage network characteristics. Results indicate that RE models using MARS with the EXTD outperform slightly RE models using GAM. Thus MARS seems to allow for a better representation of the hydrological process and an increased predictive power in RFA.

1. Introduction and literature review

The main objective of regional frequency analysis (RFA) is the estimation of the return period of extreme hydrological events at target sites where little or no hydrological data is available. Examples of these events include floods and low-flow quantiles which are crucial for infrastructure design and management. In general, RFA comprises two main steps: i) the delineation of homogenous region (DHR) to determine gauged sites similar to the target one and ii) regional estimation (RE) to transfer the information from sites determined in the DHR step to the target one (e.g. Chebana and Ouarda, 2008). Various methods have been suggested for each of these two steps (e.g. Ouarda, 2016).

Among the most common DHR methods, we can mention the region of influence (ROI) (Burn, 1990a) and the canonical correlation analysis (CCA) (Ouarda et al., 2001). Recently, several advanced non-linear neighborhood approaches were suggested (e.g. Ouali et al., 2016; Wazneh et al., 2016). Among the commonly used RE approaches, we can distinguish the regression-based models and the index-flood models. Among the former, the log-linear regression models are the most commonly used ones in practice, because of their simplicity and good predictive performances. We focus here on regression-based models in the RE step.

Hydrological processes depend from a large number of variables, such as the topographic variability of the basins, their soil structure and texture, their geological formations and the climatology. This leads to a natural complexity, which has been widely recognized and documented in the hydrological literature (e.g. Ibbitt and Woods, 2004; Sivakumar, 2007; W. Wang et al., 2008; Xu et al., 2010). In statistical terms, this complexity manifests itself through three aspects: i) the high number of explanatory variables necessary to paint a realistic picture of the processes, ii) the nonlinear impact of these explanatory variables and iii) the important interaction between the different explanatory variables. It is thus important that the RE step in RFA accounts for these three aspects in order to yield accurate estimations of the target site's quantiles of interest.

In RFA studies, the RE step usually requires a large number of explanatory variables to result in satisfactory predictive performances. This number usually exceeds five, as in Ouarda et al. (2018), but should increase in the future with the discovery of new potential variables. For instance, evidence is growing that drainage network characteristics have a strong impact on hydrological dynamics, and are consequently linked to flood quantiles (Jung et al., 2017).Thus, integrating additional characteristics related to the drainage network may lead to more accurate estimates of the regional quantiles. Hence, there is a need to propose efficient approaches that are able to manage such high-dimensional databases.

Another consequence of the natural complexity of hydrological processes is the nonlinearity between explanatory variables and the at-site quantiles. To handle this problem and better reproduce the dynamics of hydrological processes, various non-linear approaches have been proposed (e.g. Shu and Burn, 2004). The classical log-linear method used in the RE step assumes that the relation between the logarithm of the response variable (hydrological) and explanatory variables (physio-meteorological) is linear, which is too simplistic for such complex non-linear processes. Therefore, several RE approaches, such as random forest (RF), artificial neural network (ANN), and generalized additive models (GAM) have been proposed in the literature to account for the possible nonlinear links between variables (e.g. Aziz et al., 2014; Khalil et al., 2011; Ouali et al., 2017; Ouarda et al., 2018; Saadi et al., 2019).

Random forest (Breiman, 2001), is a powerful nonlinear and non-parametric method commonly used to handle regression and classification problems based on decision trees. Due to its good performance, it has been applied in several fields, such as hydrology (e.g. Diez-Sierra and del Jesus, 2019; Muñoz et al., 2018; Z. Wang et al., 2015), ecology (e.g. Cutler et al., 2007; Prasad et al., 2006) environmental modeling (e.g. Masselink et al., 2017; Pourghasemi and Kerle, 2016) and RFA (e.g. Booker and Woods, 2014; Brunner et al., 2018). Despite its predictive power, RF suffers from major limitations such as the difficulty of interpretation and the large memory requirements for storing the model when used with a large dataset (Geurts et al., 2009).

The ANN is a nonparametric mathematical model, whose design is inspired by the biological functioning of brain neurons (Bishop, 1995). It was considered in several RFA studies for the estimation of flood and low-flow quantiles at ungauged sites (e.g. Aziz et al., 2014; Ouarda and Shu, 2009). However, ANNs present a major common problem which is the tendency to overfit (e.g. Gal and Ghahramani, 2016; Lawrence and Giles, 2000). In addition, their calibration is relatively complex, especially for debutant users, which requires some subjective choices since no explicit regression equations can be given (Ouali et al., 2017).

GAMs do not suffer the same drawbacks as ANNs. GAMs are flexible nonlinear regression models (Hastie and Tibshirani, 1987), that have been introduced in the RFA context by Chebana et al. (2014). The authors found that the GAM-based methods present the best performances when compared to the classical log-linear model and other common methods. GAMs are increasingly being adopted in several fields such as hydro-climatology and environmental modeling (e.g. Rahman et al., 2018; Wen et al., 2011), public health (e.g. Bayentin et al., 2010; Leitte et al., 2009), and renewable energy (e.g Ouarda et al., 2016). However, it still presents a number of disadvantages. Indeed, the method can be computationally intensive, especially when a large number of variables is involved. It can, then, be difficult to fit GAM to high-dimensional databases because of memory limitations imposed by the numerical complexities of this model (Leathwick et al., 2006). More importantly, GAMs do not cope well with the interaction between variables (e.g. Ramsay et al., 2003), which is difficult to integrate in the model.

The interaction between physiographical variables within the watershed has long been recognized (e.g. Niehoff et al., 2002). Thus, the inclusion of the terms of interactions between the explanatory variables used to model the hydrological dynamics seems to be essential for better estimates of flood quantiles. However, this aspect is difficult to take into account in the RE models due to the high complexity that it may add to the models (see above for the specific example of GAMs). This affects the quality of the estimates and makes it less accurate. Hence, the motivation behind the present paper is to propose and explore alternative techniques able to realistically reproduce the hydrological process while avoiding the problems mentioned above.

The method considered here is multivariate adaptive regression splines (MARS), a procedure designed to build complex nonlinear regression models in a high dimensional setting. It is attractive in the RFA context since it actually addresses the three issues developed above which are: high number of variables, nonlinearity, and interactions. Indeed, MARS is efficient in a high dimensional setting and naturally selects the relevant predictors in this context. In addition, it does not require assumptions about the form of the relationships between the response and the explanatory variables (Friedman, 1991). MARS also allows the modelling of complex structures between variables, which are often hidden in high-dimensional data, without imposing strong model assumptions. Hence, it can easily include interactions between variables, allowing any degree of interaction to be considered (Lee et al., 2006).

All of these desirable properties lead to a very flexible approach able to adapt well to the hydrological phenomenon. Due to its simplicity and capacity to capture complex nonlinear relationships, it has been successfully applied in several fields such as ecology and environment (e.g. Balshi et al., 2009; Bond and Kennard, 2017; Leathwick et al., 2006; Leathwick et al., 2005), finance (e.g. Lee and Chen, 2005; Lee et al., 2006), geology (e.g. Zhang and Goh, 2016; Zhang et al., 2015), energy (e.g. Li et al., 2016; Roy et al., 2018) and hydrology (e.g. Bond and Kennard, 2017; Deo et al., 2017; Emamgolizadeh et al., 2015; Kisi, 2015; Kisi and Parmar, 2016). Despite the extensive use of the MARS model in various frameworks and contexts, its potential has never been exploited and investigated in the context of RFA of extreme hydrological events.

The main objective of the present study is to introduce the MARS approach in the RFA context to estimate flood quantiles and evaluate its predictive potential when it is applied to an extensive database. It is hereby applied in combination with the DHR with the CCA and the ROI approaches. MARS is also applied without DHR to test its performance when applied to all stations without consideration of hydrological neighborhoods. A jackknife procedure is used to evaluate the model performances, with GAMs used as a benchmark.

This paper is structured as follows. Section 2 presents the theoretical background of MARS and the other RFA approaches adopted. The considered methodology is outlined in section 3. Section 4 describes the case study and the considered datasets. The obtained results are presented and discussed in section 5. The conclusions of the study are summarized in the last section.

2. Theoretical background

In this section, the adopted statistical tools are briefly presented and discussed.

2.1 Neighborhood identification approaches

Here we present the two most commonly considered neighborhood identification approaches as a necessary step before the RE one.

2.1.1 Canonical correlation analysis (CCA) approach

CCA (Hotelling, 1935) is a multivariate analysis technique used to identify the possible correlations between two groups of variables. It consists of a linear transformation of two groups of random variables into pairs of canonical variables, which are established in such a way that the correlations between each pair are maximized.

Let $X = (X_1, X_2, \dots, X_r)$ and $Y = (Y_1, Y_2, \dots, Y_s)$ be sets of random variables including, respectively, the r physio-meteorological variables and the s hydrological variables of

n gauged sites. The objective of CCA is to construct linear combinations V_i and W_i (called canonical variables) of the variables X and Y, i.e.:

$$V_i = A_{i1}X_1 + A_{i2}X_2 + \cdots + A_{ir}X_r \tag{1}$$

$$W_i = B_{i1}Y_1 + B_2Y_2 + \cdots + B_{is}Y_s \tag{2}$$

where i = 1,…,p, with p = min (r, s). The first weights vectors $A_1$and B_1 maximize the correlation coefficients between resulting canonical variables, i.e. λ_1= corr (V_1 ,W_1), under constraints of unit variance. Once the first pair of canonical variables is identified, other pairs ($V_i, W_i, i > 1$) can be obtained under the constraint corr (V_i ,W_j) = 0 (where $i \neq j$).

For neighbourhood delineation in RFA, the considered X_r are physio-meteorological variables while the Y_S are the flood quantiles of interest. CCA is then used to construct canonical variables W_i that correlate well with physio-meteorological variables. The neighbourhood is the set of sites such that the canonical hydrological score w_k , k =1, . . . , K, is close to the canonical physio-meteorological score of the target ungauged site v_0. The distance is measured by a Mahalanobis distance between the hydrological mean position of the target site Λv_0 and the positions of other sites w_k, where $\Lambda = diag\ (\lambda_1, \dots \lambda_p)$ and v_0 is the physio-meteorological canonical score of the target site Provided the X variables are approximately normal, the Mahalanobis distance converges to a χ^2 distribution with p degrees of freedom. The size of the neighborhood is controlled by the parameter α that represent the $(1-\alpha)$ χ_p^2 quantile above which sites are excluded from the neighborhood. As extreme cases, all stations are considered if α =

0, and no station is included in the neighborhood when $\alpha = 1$. For more details the reader is referred to Ouarda et al. (2001).

2.1.2 Region of influence (ROI) approach

The ROI approach was introduced by Burn (1990b), to identify the neighborhood of a given target-site based on the similitude between watersheds characteristics. The similitude is measured using an Euclidean distance in the multidimensional physio-meteorological space (e.g. Burn, 1990b; Tasker et al., 1996) i.e.:

$$ROI_i = \left\{ sites\ j\ \epsilon\ (1, \dots, n);\ \ D_{ij} = [\sum_{k=1}^{r} W_k\ (X_{k,i} - X_{k,j})^2]^{\frac{1}{2}} \leq \theta \right\} \tag{3}$$

where D_{ij} is the weighted Euclidean distance between the target site i and the gauged one, $j = 1,\dots, n$, $X_{k,j}$ ($k = 1,\dots, r$) is the standardized value of the k^{th} variable at site j, W_k is the weight associated with the k^{th} variable, and θ represents the threshold value. The threshold value is defined for each site in such a way that it permits a compromise between the amount of information to be used and the degree of hydrological homogeneity of the neighborhood (Ouarda et al., 1999). For more details, the reader is referred to (e.g. Burn, 1990b; GREHYS, 1996).

2.2 Regional estimation approaches

Once a neighborhood is identified, the methods described below are used to transfer information from the neighborhood stations to the target site.

2.2.1 Generalized Additive Model (GAM)

GAM (Hastie and Tibshirani, 1987) is a flexible class of nonlinear models that is able to efficiently model a wide variety of nonlinear relationships. In addition, it allows for non-gaussian response variables (Wood, 2006) making it relevant for streamflow data. Thus,

GAM allows a more realistic description of the hydrological phenomenon because of the flexible non-parametric fitting of the smooth functions.

Formally, a GAM is defined as (Wood, 2006):

$$g\,(Y) = \alpha + \sum_{j=1}^{m} f_j\,(X_j) + \varepsilon \qquad (4)$$

where g is a monotonic link function and f_j are smooth functions giving the relationship between the explanatory variables X_j and the response Y. α is the intercept and ε is the error term. The structure of eq. 4 allows for a distinct interpretation of each explanatory variable.

To estimate the model, the smooth functions f_j are expressed as a set of q spline basis functions, a common choice for smoothing (Wahba, 1990). They are expressed as:

$$f_j(X) = \sum_{i=1}^{q} \beta_{ji}\, b_{ji}(X) \qquad (5)$$

where β_{ji} are unknown parameters to be estimated and b_{ji} are the spline basis functions. The expansion in (5) allows linearizing the model that can then be estimated through backfitting (Hastie and Tibshirani, 1987) or simple penalized least-squares (Wood, 2004).

For more details, the reader is referred to (e.g. Wood, 2006; Wood, 2017).

2.2.2 Multivariate adaptive regression splines (MARS)

MARS was introduced by Friedman (1991) as a flexible non-parametric regression approach able to deal with high-dimensional data. The MARS model $f(X)$ can be seen as a flexible extension of GAM, in that it is expressed as a linear combination of basis functions and their interactions as:

$$f(X) = \beta_0 + \sum_{n=1}^{r} \beta_n B_n(X) \tag{6}$$

where β_0 is the intercept, β_n are regression coefficients of the basis functions ($B_n(X)$). In the MARS model, the $B_n(X)$ terms can take one of the following forms: i) a constant (just one term) which represent the intercept, ii) a linear spline functions on a single variable X_j called hinge function, i.e. of the form $h_m(X_j) = (t_m - X_j)_+$ or $h_m(X_j) = (X_j - t_m)_+$ where t is a knot and iii) a products of two or more hinge functions, e.g. $B_n(X) = h_m(X_j)h_{m'}(X_k)$ where $j \neq k$. The latter represent interaction between two or more variables. The $B_n(X)$ are defined in pairs and separated by a knot which represents an inflection point along the range of a given explanatory variable (see Figure 1). Allowing the product of several linear spline terms $h_m(X_j) = (t_m - X_j)_+$ as basis functions further allows the integration of interaction in the model, an aspect GAMs are not well designed for.

In mathematical terms, the hinge functions $h_m(X_j)$ are defined as (Rounaghi et al., 2015):

$$(t - X_j)_+ = \begin{cases} t - X_j\,, & if\ t > X_j \\ 0, & otherwise \end{cases} \qquad (7)$$

$$(X_j - t)_+ = \begin{cases} X_j - t\,, & if\ X_j > t \\ 0, & otherwise \end{cases} \qquad (8)$$

where t is the knot position.

The main difference of MARS with GAM is in the estimation algorithm. Where the spline bases are defined *a priori* in GAM, they are iteratively constructed in MARS, adapting hence to the data. Indeed, building the model in (6) is carried out through two phases: i) a forward addition of linear spline terms (i.e. of the form (7) and (8)) to build a large model and ii) a backward deletion to delete irrelevant terms. The forward phase begins with an empty model containing only the intercept β_0. B_ns are then iteratively added to the model, each time choosing the variable and knot yielding the largest decrease in the residual error of the model. This process of adding B_ns continues until the model reaches some predetermined maximum number, leading to a large model which may over-fit the data. A backward deletion phase is then performed to improve the model performance by removing the less significant B_ns until obtaining the best sub-models. Comparison of sub-models is made based on the Generalized Cross Validation (GCV). Figure 2 illustrates the details of the MARS model algorithm.

Another interesting feature of MARS is the assessment of the variable importance for the prediction of the response. Variable importance can be measured in two different ways: i) the number of sub-models that include the variable, or ii) the increase in GCV caused by deleting the considered variables from the final MARS model (e.g. Roy et al., 2018).

3. Methodology

3.1 Regional models

In this study, the methods presented in section 2 for neighborhood delineation (CCA and ROI) are used in combination with the regional estimation models GAM and MARS for transfer of hydrological information. As mentioned in section 1, other evaluated models are obtained by applying the GAM and MARS using all stations, i.e. without defining any neighborhoods. Table 1 summarizes all six resulting combinations.

The two most commonly used neighborhood approaches, the CCA and the ROI (Ouarda, 2016) are applied to the DHR using two sets of variables. For these methods, the relevant variables are selected based on their correlation degree with the hydrological variables.

Considering the classical procedures used to define the threshold in ROI and CCA, the density of stations in the neighborhoods can vary considerably from one region to another. Indeed, for a given fixed threshold, stations located near the center of the cloud points defined by the canonical space for CCA or the Euclidean space for ROI will have more stations within their neighbourhoods and vice versa (Leclerc and Ouarda, 2007). Since, the sample may affect the accuracy of the estimates obtained by regression models, it was decided that for each target station, the size of the region is increased until a selected optimal size is reached. The optimal number of stations to be considered in the DHR step is chosen based on the optimization procedure of Ouarda et al. (2001). The optimal number of sites in the neighborhood is the one that minimizes a given performance criterion of the log-linear model applied in each neighborhood.

MARS is fitted using the R package earth (Milborrow, 2018).The application of MARS needs the tuning of three main parameters (see Figure 2): the maximum number of terms in the model in the forward phase (N_k), the degree of interaction (degree), and the maximum number of terms in the Backward phase (N_prune). A range of values of these parameters was tested and evaluated in order to optimize them based on the GCV,the residual sum of squares (RSS) and the coefficient of determination (R^2) criteria of the fitted models.

GAM is also implemented on R, through the package mgcv (Wood, 2006). The thin plate regression spline is used in this study as basis b_{ji} in the smoothing function f_j in (5). The latter is selected due to its advantages, i.e. low calculation time, flexibility and fewer number of parameters compared to other smoothing functions (Wood, 2003). The used link function g in (4) is the identity function because of the approximately normal log-transformed quantiles such as considered in Ouali et al. (2017).

Different physio-meteorological variables are considered in each regional model. A backward stepwise approach is applied in this study to select the relevant explanatory variables to be used in each RE models (GAM and MARS). This method is presented in the next section.

3.2 Variable selection

The backward stepwise selection procedure is applied in this work to select the optimal explanatory variables as in Ouarda et al. (2018) and Chebana et al., (2014). It consists in a progressive deleting of the least effective variables from an initial full model containing all available variables. At each step, the removed variable is the one having either the highest p value for the null hypothesis that the smooth term for GAM is zero or those whose consideration yields the most significant increase in the GCV score of the model for MARS.

Note that the MARS algorithm naturally includes a variable selection feature since it builds a sparse model and a variable for which no term is added is by default discarded. This is not the case for GAM within which an automatic backward stepwise procedure was specially developed for this study.

3.3 Validation

For each RFA combination in Table 1, performances are evaluated using a leave-one-out cross validation, commonly called jackknife procedure in the field of hydrology. It consists in deleting temporarily each site to consider it the target one and perform RE. This process is repeated for each gauged sites. Then, the regional estimate is compared to its observed values. Note that, in statistics, the validation with the jackknife technique is carried out on the retained data not on the data removed as in the leave-one-out cross validation (Quenouille, 1949). However, we will retain the jackknife term for ease of presentation.

Based on the jackknife procedure, several standard performance criteria are used to evaluate the prediction power of each regional model (e.g. Ouali et al., 2016). First, the Nash criterion (NASH) gives a global evaluation 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 computation. Finally, the bias (BIAS) and the relative bias (RBIAS) provide a measure of the magnitude of the systematic overestimation or underestimation of a model.

4. Case study and datasets

The dataset considered in the present paper consists in 151 hydrometric stations located in the southern part of the province of Quebec, Canada (Figure 3). Two versions of the datasets with different variables are considered. The first is a standard one (STA) with only well-known variables used in previous RFA studies (e.g. Shu et al*.*, 2007, Chebana et al*.*, 2014, Durocher et al., 2016, Ouali et al., 2016, Wazneh et al., 2013; 2015 and 2016). Note that geographical coordinates of the stations are considered instead of the geographical coordinates of the centroids. The second is an extended dataset (EXTD) combining STA with less common variables characterizing the drainage network systems. Table 2 lists all variables considered as well as whether they are in the EXTD dataset and thorough definitions of the new variables can be found in (e.g. Adhikary and Dash, 2018). These new variables are calculated based on drainage networks extracted using the D8 approach implemented in Arc Gis (Arc Hydro) using the digital elevation models; DEMs (Jenson and Domingue, 1988; O'Callaghan and Mark, 1984; Tarboton et al., 1991). This method consists in calculating the flow direction and the flow accumulation layers based on the direction of the steepest slope among the eight neighbors of a given DEM. Using this information, the drainage networks can be defined considering a constant threshold value which represents the stream head locations (O'Callaghan and Mark, 1984). Descriptive statistics of the new variables used in the EXTD dataset (Msilini et al., 2020) are given in Table 3. In both datasets the considered hydrological response variables are at-site specific flood quantiles, chosen to match the specific return periods of 10, 50 and 100 years. These quantiles are thus denoted by QS_{10}, QS_{50} and QS_{100}.

To ensure the convergence of the Mahalanobis distance to a χ^2 distribution in CCA, note that the logarithmic transformation is used for the following variables to achieve approximate normality: AREA, MBS, MATP, DDBZ and RT and a square root transformation for PLAKE and RC. After transformation normal q-q plot indicate that all variables are approximately normal.

5. Results and Discussion

5.1 Region delineation with CCA and ROI

The CCA and the ROI are applied to the DHR using two sets of variables. The first set contains variables from STA, which are 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 basin (LONGC). The second one includes variables from the EXTD, namely PLAKE, MATP, DDBZ, LONGC, texture ratio (RT) and circularity ratio (RC).

The obtained optimum sizes of the neighborhood are n^{opt} (STA) = 85 sites and n^{opt} (EXTD) = 78 sites according to the RRMSE for the CCA method. For the ROI approach, we obtain n^{opt} (STA) = 54 sites and n^{opt} (EXTD) = 44 sites according to the same criterion. Thus, these neighborhood sizes are used for each target station.

5.2 Selection of optimal variables

The selection of significant explanatory variables is applied for each specific quantile (QS_{10}, QS_{50} and QS_{100}) and for each estimation model (GAM and MARS). Table 4 summarizes the final variables for each datasets (STA and EXTD). Following the application of the backward technique with GAM and MARS, we note the selection of the same new variables for the two models (RN, MRL and DD). The definition of these variables can be found for example in Adhikary and Dash (2018). For each quantile and for each model, different combinations of variables are selected. The variables that seem to be the most important are AREA, PLAKE, MCL and LONGC.

5.3 MARS model results

Figure 4 shows the variable importance graph for QS_{100} obtained using the EXTD (we present only the results of QS_{100} to avoid repetitions). The variable with the most influence for the QS_{100} is the percentage of the area occupied by lakes, PLAKE. Indeed, lakes act as a sponge absorbing the excess water during extreme events. Thus they may have a significant effect on flood peaks.

Figure 5 shows the GCV R^2 (GRSq) value for the QS_{100} predictions versus the number of terms in the final MARS model. The GCV R^2 statistic is equivalent to the ordinary R^2 statistic calculated with the variance for error replaced with the GCV statistic.

It allows quantifying the goodness-of-fit for models that use unobserved data. The vertical dashed lined at 12 indicates the optimal number of terms retained where marginal increases in GCV R^2 are less than 0.001. The twelve final terms include seven variables in this case. Five terms are related to interaction effects.

5.4 Comparison between MARS and GAM models

Table 5 shows the jackknife results for each model combination. The comparison of GAM and MARS models confirms that the simple linear spline fitting generated by MARS captures more information from the EXTD than the more sophisticated smoothing functions used in GAM. Indeed, MARS adds the terms in an iterative way leading to a simple and performant model including the effects of interactions. This model performs well with the ROI which contains a smaller number of stations than CCA. Thus, based on the results of our case study MARS seems applicable in small neighborhoods even with complex terms (interaction effects) and able to give good predictions with fewer stations than GAM.

The response functions fitted by GAM and MARS models for selected explanatory variables are given in Figure 6. It can be seen that the smoothing functions fitted by MARS approximate closely the more continuous smooth curves fitted by GAM, in a simpler way. This result has been observed by Leathwick et al. (2006) in a comparative study made between GAM and MARS applied in the field of ecology. The smooth curves generated by GAM add degrees of freedom to the model which makes it relatively more complex. This may be the reason for the better prediction results obtained by MARS than GAM.

Figure 7 illustrates the interaction effects between some explanatory variables fitted by GAM and MARS models. Note that we considered the same interactions automatically identified by MARS to be able to make the comparison. The interaction surface generated by both models is also close. GAM gives more continuous and complex interaction effects, which lead to a large model with a large number of coefficients. This makes it difficult or impossible to integrate the interaction effects with GAM if we have a large number of explanatory variables in the model. For example, for the QS_{100}, the integration of the same interactions identified by MARS to GAM considering the same variables gives a model with 79 coefficients, versus only 12 using MARS. In addition, MARS searches for and integrates interaction effects automatically into the model, which allows obtaining flood quantile estimates overall better than those obtained by GAM. We take as a simple example of interaction the first effect illustrated in Figure 7 which represents the predicted response (specific quantile) as DD and LONGC vary. It can be seen that the LONGC affects little the hydrological variable level unless the DD is high where a nonlinear effect is seen.

5.5 Comparison of regional models

According to Table 5 (see above), the highest NASH values (0.80) and the lowest RRMSE values (28.30 % for QS_{100}) are given by the ROI/MARS/EXTD, which leads to the most accurate estimates compared to all other combinations. It can also be seen that, with ALL, MARS has a comparable performance to GAM considering both databases. However, using the neighborhoods, especially the ROI, MARS overall outperforms GAM in terms of RRMSE and RBIAS criteria. This may be attributable to the flexibility of MARS and its generalization ability in small size neighborhoods.

Figure 8 illustrates the relative error, which is the most important criterion (Hosking and Wallis, 2005), as a function of the sites ordered according to their area associated to the best models (ROI/MARS/EXTD and ROI/GAM/EXTD). One can notice that, overall, MARS with the EXTD performs better than GAM. The figure also shows that the performances at the level of extreme size basins are much worse than those obtained at the level of medium size basins.

Figure 9 presents the differences between relative errors of MARS and GAM calculated using ROI/EXTD. One can notice that, in terms of RRMSE, MARS outperforms GAM in 84 sites out of 151, which represents 56% of the total number of sites. Accordingly, MARS is shown to be a simple performant model that can be considered as an alternative RE model.

6. Conclusions

The aim of this study is to introduce MARS in the RFA of extreme hydrological variables and to compare its performance to GAM. The MARS model is able to model complex relationship between physio-meteorological variables, including variables dealing with drainage network characteristics, and flood quantiles at ungauged sites.

MARS is hereby compared to the GAM which is gaining popularity in RFA and is one of the best performing models. Results show that slightly better flood quantile estimates are obtained from regional models that combine MARS with the EXTD including a STA with additional variables dealing with drainage network proprieties. Results indicate also that better performances are obtained with the ROI which includes low density of stations than CCA. This suggests that MARS is able to transfer hydrological information adequately even with fewer data than GAM. Further efforts are required to generalize this conclusion and to evaluate the benefits of MARS in other study areas and with other hydrological variables.

Although MARS is an effective and simple tool for estimation that can be used in RFA, there are some constraints such as the maximum number of terms and the maximum allowable degree of interaction in the forward pass that have to be specified by the user. These depend on the problem at hand and should be considered carefully. In addition, MARS does not cope well with missing data and, like many machine learning algorithms, is prone to overfitting. Note however that the backward deletion phase is meant to address this drawback

Aside from the above-mentioned shortcomings, MARS is easy-to-use as shown in this work. It is able to addresses the issues of high number of variables, nonlinearity, and interactions involved in the hydrological phenomena. This yields flood quantile estimates that compete with those obtained from GAM, while being simpler and more applicable to smaller datasets. Flood quantiles represent important information that is used in the design of hydraulic structures (e.g. dams). The construction of these structures is very expensive. The availability of simple and sophisticated tools for the reliable estimation of flood quantiles is crucial for hydraulics engineers.

In this work we considered linear neighborhood approaches (CCA and ROI), which are the most used methods in RFA. Future efforts can focus on the assessment of the performance of the MARS model in combination with non-linear neighborhood approaches such as the non-linear canonical correlation analysis (Ouali et al., 2016) and the nonlinear neighborhood based on the statistical depth function (Wazneh et al., 2016).

Acknowledgments

Financial support for the present study was graciously provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research chairs program (CRC). The authors are grateful to Natural Resources Canada and the USGS services for the employed DEM data. The authors would like also to thank the Ministry of Sustainable Development, Environment, and Fight Against Climate Change (MDDELCC) services for the employed dataset (STA). The authors would like to thank the Editor, Prof. Andrew Wood and three anonymous reviewers for their comments which helped improve the quality of the manuscript.

Appendix

Abbreviations

ANN	Artificial neural network
AREA	Basin area
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
GCV	Generalized cross validation
IF	Infiltration number
LATC	Latitude of the centroid of the basin
LONGC	Longitude of the centroid of the basin
MALP	Mean annual liquid precipitation
MALPS	Mean annual liquid precipitation (summer–fall)
MARS	Multivariate adaptive regression splines
MASP	Mean annual solid precipitation
MATP	Mean annual total precipitation
MBS	Mean basin slope
MCL	Main channel length
MCS	Main channel slope
MRB	Mean bifurcation ratio
MRL	Mean stream length ratio
NASH	Nash efficiency criterion
NL-CCA	Nonlinear canonical correlation analysis
PFOR	Percentage of the area occupied by forest
PL1	Percentage of first-order stream lengths
PLAKE	Percentage of the area occupied by lakes
PN1	Percentage of first-order streams
QS_T	Specific quantile associated to the 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
RSS	Residual sum of squares
RT	Texture ratio
STA	Standard dataset
WMRB	Weighted mean bifurcation ratio

References

Adhikary, & Dash. (2018). Morphometric analysis o f Katra Watershed of Eastern Ghats: A GIS approach. *Int. J. Curr. Microbiol. App. Sci, 7*(3), 1651-1665.

Aziz, Rahman, Fang, & Shrestha. (2014). Application of artificial neural networks in regional flood frequency analysis: a case study for Australia. *.Stochastic environmental research and risk assessment., 28*(3), 541-554. doi:https://doi.org/10.1007/s00477-013-0771-5

Balshi, McGUIRE, Duffy, Flannigan, Walsh, & Melillo. (2009). Assessing the response of area burned to changing climate in western boreal North America using a Multivariate Adaptive Regression Splines (MARS) approach. *Global Change Biology, 15*(3), 578-600. doi:https://doi.org/10.1111/j.1365-2486.2008.01679.x

Bayentin, El Adlouni, Ouarda, Gosselin, Doyon, & 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

Bishop. (1995). *Neural networks for pattern recognition*: Oxford university press.

Bond, & Kennard. (2017). Prediction of hydrologic characteristics for ungauged catchments to support hydroecological modeling. *Water Resources Research, 53*(11), 8781-8794. doi:https://doi.org/10.1002/2017WR021119

Booker, & Woods. (2014). Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. *Journal of Hydrology, 508*, 227-239.

Breiman. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brunner, Furrer, Sikorska, Viviroli, Seibert, & Favre. (2018). Synthetic design hydrographs for ungauged catchments: a comparison of regionalization methods. *Stochastic environmental research and risk assessment, 32*(7), 1993-2023.

Burn. (1990a). An appraisal of the "region of influence" approach to flood frequency analysis. *.Hydrological sciences journal., 35*(2), 149-165. doi:https://doi.org/10.1080/02626669009492415

Burn. (1990b). Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research, 26*(10), 2257-2265. doi: https://doi.org/10.1029/WR026i010p02257

Chebana, Charron, Ouarda, & Martel. (2014). Regional frequency analysis at ungauged sites with the generalized additive model. *Journal of Hydrometeorology, 15*(6), 2418-2428. doi:https://doi.org/10.1175/JHM-D-14-0060.1

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

Cutler, Edwards Jr, Beard, Cutler, Hess, Gibson, & Lawler. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783-2792.

Deo, Kisi, & Singh. (2017). Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. *Atmospheric Research, 184*, 149-175. doi:https://doi.org/10.1016/j.atmosres.2016.10.004

Diez-Sierra, & del Jesus. (2019). Subdaily rainfall estimation through daily rainfall downscaling using random forests in Spain. *Water, 11*(1), 125.

Emamgolizadeh, Bateni, Shahsavani, Ashrafi, & Ghorbani. (2015). Estimation of soil cation exchange capacity using genetic expression programming (GEP) and multivariate adaptive regression splines (MARS). *Journal of Hydrology, 529*, 1590-1600. doi:https://doi.org/10.1016/j.jhydrol.2015.08.025

Friedman. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1-67.

Gal, & Ghahramani. (2016). *A theoretically grounded application of dropout in recurrent neural networks.* Paper presented at the Advances in neural information processing systems.

Geurts, Irrthum, & Wehenkel. (2009). Supervised learning with decision tree-based methods in computational and systems biology. *Molecular Biosystems, 5*(12), 1593-1605.

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

Hastie, & Tibshirani. (1987). Generalized Additive Models: Some Applications. *Journal of the American statistical Association, 82*(398), 371-386. doi:10.1080/01621459.1987.10478440

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

Hotelling. (1935). Canonical correlation analysis (cca). *Journal of Educational Psychology*, 10.

Ibbitt, & Woods. (2004). Re-scaling the topographic index to improve the representation of physical processes in catchment models. *Journal of Hydrology, 293*(1-4), 205-218. doi:https://doi.org/10.1016/j.jhydrol.2004.01.016

Jenson, & Domingue. (1988). Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric engineering and remote sensing, 54*(11), 1593-1600.

Jung, Marpu, & Ouarda. (2017). Impact of river network type on the time of concentration. *Arabian Journal of Geosciences, 10*(24), 546. doi:https://doi.org/10.1007/s12517-017-3323-3

Khalil, Ouarda, & 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), 277-287.

Kisi. (2015). Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology, 528*, 312-320. doi:https://doi.org/10.1016/j.jhydrol.2015.06.052

Kisi, & Parmar. (2016). Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. *Journal of Hydrology, 534*, 104-112. doi:https://doi.org/10.1016/j.jhydrol.2015.12.014

Lawrence, & 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.

Leathwick, Elith, & Hastie. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling, 199*(2), 188-196. doi:https://doi.org/10.1016/j.ecolmodel.2006.05.022

Leathwick, Rowe, Richardson, Elith, & Hastie. (2005). Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish.

Freshwater Biology, 50(12), 2034-2052. doi:https://doi.org/10.1111/j.1365-2427.2005.01448.x
Leclerc, & Ouarda. (2007). Non-stationary regional flood frequency analysis at ungauged sites. *Journal of Hydrology, 343*(3-4), 254-265. doi:https://doi.org/10.1016/j.jhydrol.2007.06.021
Lee, & Chen. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28*(4), 743-752. doi:https://doi.org/10.1016/j.eswa.2004.12.031
Lee, Chiu, Chou, & Lu. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis, 50*(4), 1113-1130. doi:https://doi.org/10.1016/j.csda.2004.11.006
Leitte, Petrescu, Franck, Richter, Suciu, Ionovici, et al. (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), 4004-4011. doi:https://doi.org/10.1016/j.scitotenv.2009.02.042
Li, He, Su, & Shu. (2016). Forecasting the daily power output of a grid-connected photovoltaic system based on multivariate adaptive regression splines. *Applied Energy, 180*, 392-401. doi:https://doi.org/10.1016/j.apenergy.2016.07.052
Masselink, Temme, Giménez Díaz, Casalí Sarasíbar, & Keesstra. (2017). Assessing hillslope-channel connectivity in an agricultural catchment using rare-earth oxide tracers and random forests models. *Cuadernos de Investigación Geográfica 2017, nº 43 (1), pp. 19-39*.
Milborrow. (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, Ouarda, & Masselot. (2020). Evaluation of additional physiographical variables characterising drainage network systems in regional frequency analysis, a Quebec watersheds case-study. *Manuscript submitted for publication*.
Muñoz, Orellana-Alvear, Willems, & Célleri. (2018). Flash-flood forecasting in an Andean mountain catchment—Development of a step-wise methodology based on the random forest algorithm. *Water, 10*(11), 1519.
Niehoff, Fritsch, & Bronstert. (2002). Land-use impacts on storm-runoff generation: scenarios of land-use change and simulation of hydrological response in a meso-scale catchment in SW-Germany. *Journal of Hydrology, 267*(1-2), 80-93. doi:https://doi.org/10.1016/S0022-1694(02)00142-7
O'Callaghan, & Mark. (1984). The extraction of drainage networks from digital elevation data. *Computer vision, graphics, and image processing, 28*(3), 323-344. doi:https://doi.org/10.1016/S0734-189X(84)80011-0
Ouali, Chebana, & Ouarda. (2016). Non-linear canonical correlation analysis in regional frequency analysis. *Stochastic environmental research and risk assessment, 30*(2), 449-462. doi:https://doi.org/10.1007/s00477-015-1092-7
Ouali, Chebana, & Ouarda. (2017). Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites. *Journal of Advances in Modeling Earth Systems, 9*(2), 1292-1306. doi:https://doi.org/10.1002/2016MS000830
Ouarda. (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. 77.71-77.78, ISBN 978-970-907-183509-183501.
Ouarda, Charron, Hundecha, St-Hilaire, & 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*, 256-271. doi:https://doi.org/10.1016/j.envsoft.2018.08.031

Ouarda, Charron, Marpu, & 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), 1553-1566. doi:https://doi.org/10.1109/jstars.2016.2522764

Ouarda, Girard, Cavadias, & Bobée. (2001). Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology, 254*(1), 157-173. doi:https://doi.org/10.1016/S0022-1694(01)00488-7

Ouarda, Lang, Bobée, Bernier, & Bois. (1999). Synthèse de modèles régionaux d'estimation de crue utilisée en France et au Québec. *Revue des sciences de l'eau/Journal of Water Science, 12*(1), 155-182. doi:https://doi.org/10.7202/705347ar

Ouarda, & 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

Pourghasemi, & Kerle. (2016). Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran. *Environmental Earth Sciences, 75*(3), 185.

Prasad, Iverson, & Liaw. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems, 9*(2), 181-199.

Quenouille. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics, 20*(3), 355-375. doi:10.1214/aoms/1177729989

Rahman, Charron, Ouarda, & Chebana. (2018). Development of regional flood frequency analysis techniques using generalized additive models for Australia. *Stochastic environmental research and risk assessment, 32*(1), 123-139. doi:https://doi.org/10.1007/s00477-017-1384-1

Ramsay, Burnett, & Krewski. (2003). The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology, 14*(1), 18-23.

Rounaghi, Abbaszadeh, & Arashi. (2015). Stock price forecasting for companies listed on Tehran stock exchange using multivariate adaptive regression splines model and semi-parametric splines technique. *Physica A: Statistical Mechanics and its Applications, 438*, 625-633. doi:https://doi.org/10.1016/j.physa.2015.07.021

Roy, Roy, & Balas. (2018). Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of MARS and ELM. *Renewable and Sustainable Energy Reviews, 82*, 4256-4268. doi:https://doi.org/10.1016/j.rser.2017.05.249

Saadi, Oudin, & Ribstein. (2019). Random Forest Ability in Regionalizing Hourly Hydrological Model Parameters. *Water, 11*(8), 1540. doi:https://doi.org/10.3390/w11081540

Shu, & Burn. (2004). Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resources Research, 40*(9). doi:https://doi.org/10.1029/2003WR002816

Sivakumar. (2007). Nonlinear determinism in river flow: prediction as a possible indicator. *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group, 32*(7), 969-979. doi: https://doi.org/10.1002/esp.1462

Tarboton, Bras, & Rodriguez-Iturbe. (1991). On the extraction of channel networks from digital elevation data. *Hydrological Processes, 5*(1), 81-100. doi: https://doi.org/10.1002/hyp.3360050107

Tasker, Hodge, & 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), 163-170. doi:https://doi.org/10.1111/j.1752-1688.1996.tb03444.x

Wahba. (1990). *Spline models for observational data* (Vol. 59): Siam.

Wang, Chen, Shi, & Van Gelder. (2008). Detecting changes in extreme precipitation and extreme streamflow in the Dongjiang River Basin in southern China. *Hydrology and Earth System Sciences Discussions, 12*(1), 207-221.
Wang, Lai, Chen, Yang, Zhao, & Bai. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology, 527*, 1130-1141.
Wazneh, Chebana, & Ouarda. (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
Wen, Rogers, Saintilan, & 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), 154-163. doi:https://doi.org/10.1016/j.ecolmodel.2010.09.016
Wood. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(1), 95-114. doi:https://doi.org/10.1111/1467-9868.00374
Wood. (2004). Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American statistical Association, 99*(467), 673-686. doi:10.1198/016214504000000980
Wood. (2006). *Generalized additive models: an introduction with R*: CRC press.
Wood. (2017). *Generalized additive models: an introduction with R*: CRC press.
Xu, Li, Ji, Lu, & 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), 136-146. doi:https://doi.org/10.1002/hyp.7484
Zhang, & Goh. (2016). Evaluating seismic liquefaction potential using multivariate adaptive regression splines and logistic regression. *Geomech Eng, 10*(3), 269-280. doi:http://dx.doi.org/10.12989/gae.2016.10.3.269
Zhang, Goh, Zhang, Chen, & Xiao. (2015). Assessment of soil liquefaction based on capacity energy concept and multivariate adaptive regression splines. *Engineering Geology, 188*, 29-37. doi:https://doi.org/10.1016/j.enggeo.2015.01.009

Table 1 Adopted regional models.

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

Table 2 Variables used in the STA and the EXTD.

QS_T	Specific quantile associated to the return period T ; (T = 10, 50 and 100 years.)	*	+	
AREA	Basin area	*	+	**Log**
MCL	Main channel length	*	+	
MCS	Main channel slope	*	+	
MBS	Mean basin slope	*	+	**Log**
PFOR	Percentage of the area occupied by forest	*	+	
PLAKE	Percentage of the area occupied by lakes	*	+	√.
MATP	Mean annual total precipitation	*	+	**Log**
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	*	+	**Log**
LATC	Latitude of the centroid of the basin	*	+	
LONGC	Longitude of the centroid of the basin	*	+	**---**
RT	Texture ratio		+	**Log**
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	+

(*) Variables considered in the standard dataset (STA).

(+) Variables considered in the extended dataset (EXTD).

The variables considered in the neighborhoods and their transformations are presented in bold character.

Table 3 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 4 Explanatory variables selected for the various regression models.

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

Table 5 Jackknife Validation Results (STD and EXTD).

		STA						EXTD					
		ALL		CCA		ROI		ALL		CCA		ROI	
	Quantile	GAM	MARS	GAM	MARS	GAM	MARS	GAM	MARS	GAM	MARS	GAM	MARS
NASH	QS_{10}	0.774	0.788	0.797	0.771	0.829	**0.866**	0.802	0.820	0.837	0.797	0.865	0.859
	QS_{50}	0.745	0.648	0.762	0.749	0.796	0.785	0.754	0.742	0.775	0.748	**0.816**	0.802
	QS_{100}	0.715	0.643	0.723	0.679	0.762	0.752	0.725	0.625	0.742	0.682	0.791	**0.803**
RMSE $[(m^3/s)km^{-2}]$	QS_{10}	0.060	0.058	0.057	0.060	0.053	**0.047**	0.056	0.054	0.051	0.057	**0.047**	**0.047**
	QS_{50}	0.089	0.104	0.086	0.088	0.080	0.081	0.087	0.089	0.080	0.088	**0.076**	**0.076**
	QS_{100}	0.107	0.119	0.105	0.113	0.097	0.099	0.105	0.122	0.101	0.112	0.091	**0.089**
RRMSE (%)	QS_{10}	40.937	40.781	37.163	35.316	34.690	25.950	34.970	32.065	30.619	30.435	27.974	**24.423**
	QS_{50}	49.420	51.552	43.333	43.086	39.365	30.439	36.659	35.214	35.086	35.282	**27.818**	29.210
	QS_{100}	51.832	47.953	45.678	42.298	41.661	37.775	38.630	41.215	37.416	38.818	29.235	**28.298**
BIAS $[(m^3/s)km^{-2}]$	QS_{10}	0.005	0.004	0.006	0.004	**0.003**	0.007	0.005	0.005	0.007	0.008	0.004	0.008
	QS_{50}	0.008	0.008	0.015	0.014	**0.006**	0.009	0.008	**0.006**	0.015	0.015	0.009	0.009
	QS_{100}	0.011	0.008	0.020	0.014	0.009	0.011	0.011	0.007	0.020	0.016	0.012	**0.001**
RBIAIS (%)	QS_{10}	-5.461	-4.650	-5.555	-5.095	-4.177	-1.682	-4.179	-4.003	-3.871	-2.818	-2.836	**-0.250**
	QS_{50}	-7.047	-8.563	-5.632	-5.778	-5.487	-3.154	-4.954	-4.862	-3.513	-3.514	-2.892	**-2.176**
	QS_{100}	-7.663	-8.451	-5.780	-6.291	-5.816	-5.275	-5.472	-5.767	-3.714	-4.465	**-3.172**	-3.583

Best results are in bold character.

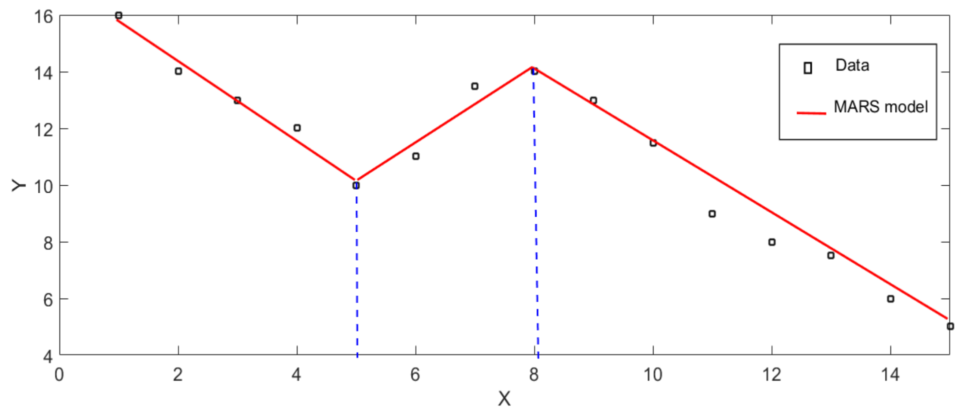

Figure 1 Knots and linear splines for a simple example of MARS.

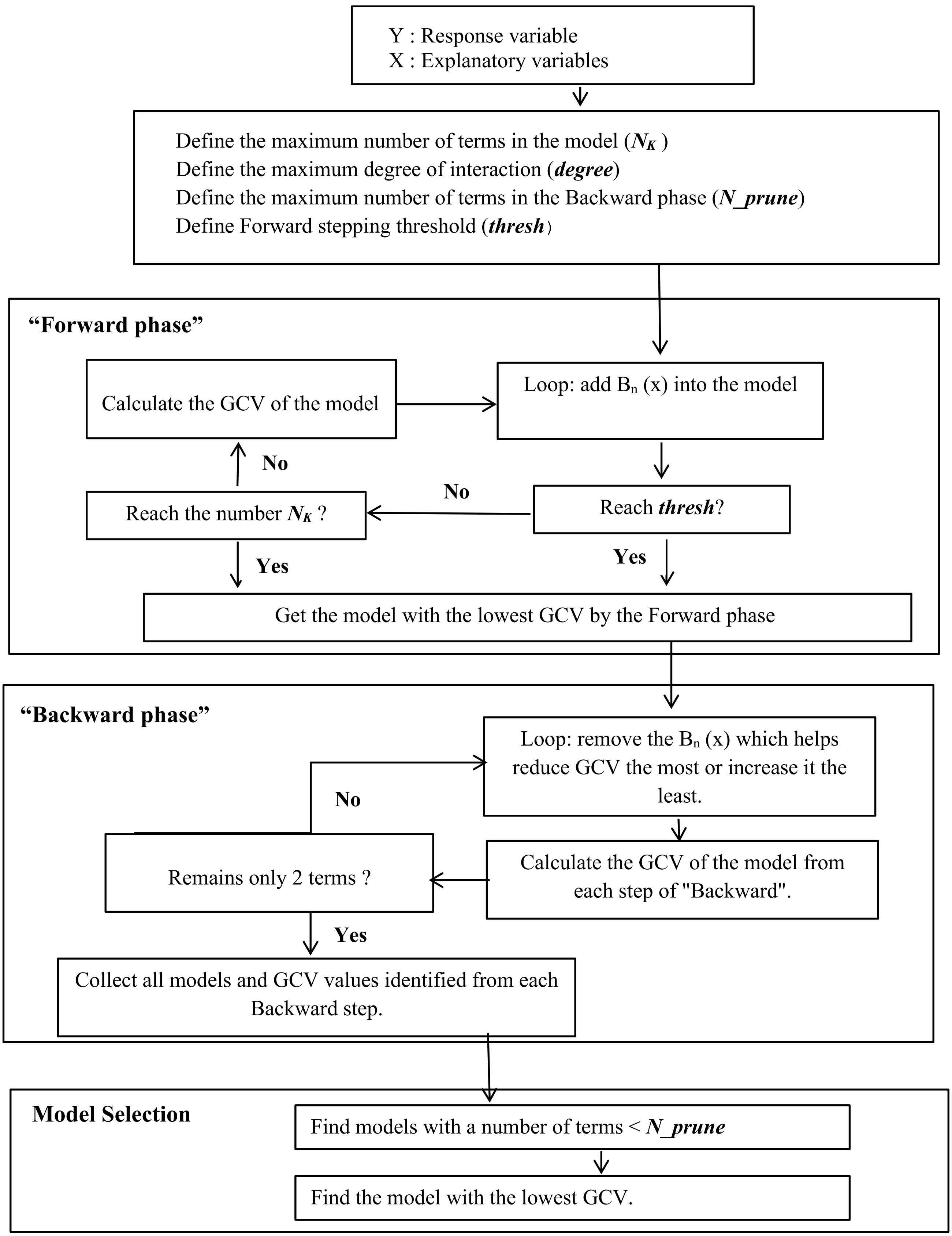

Figure 2 Graph of MARS modelling process.

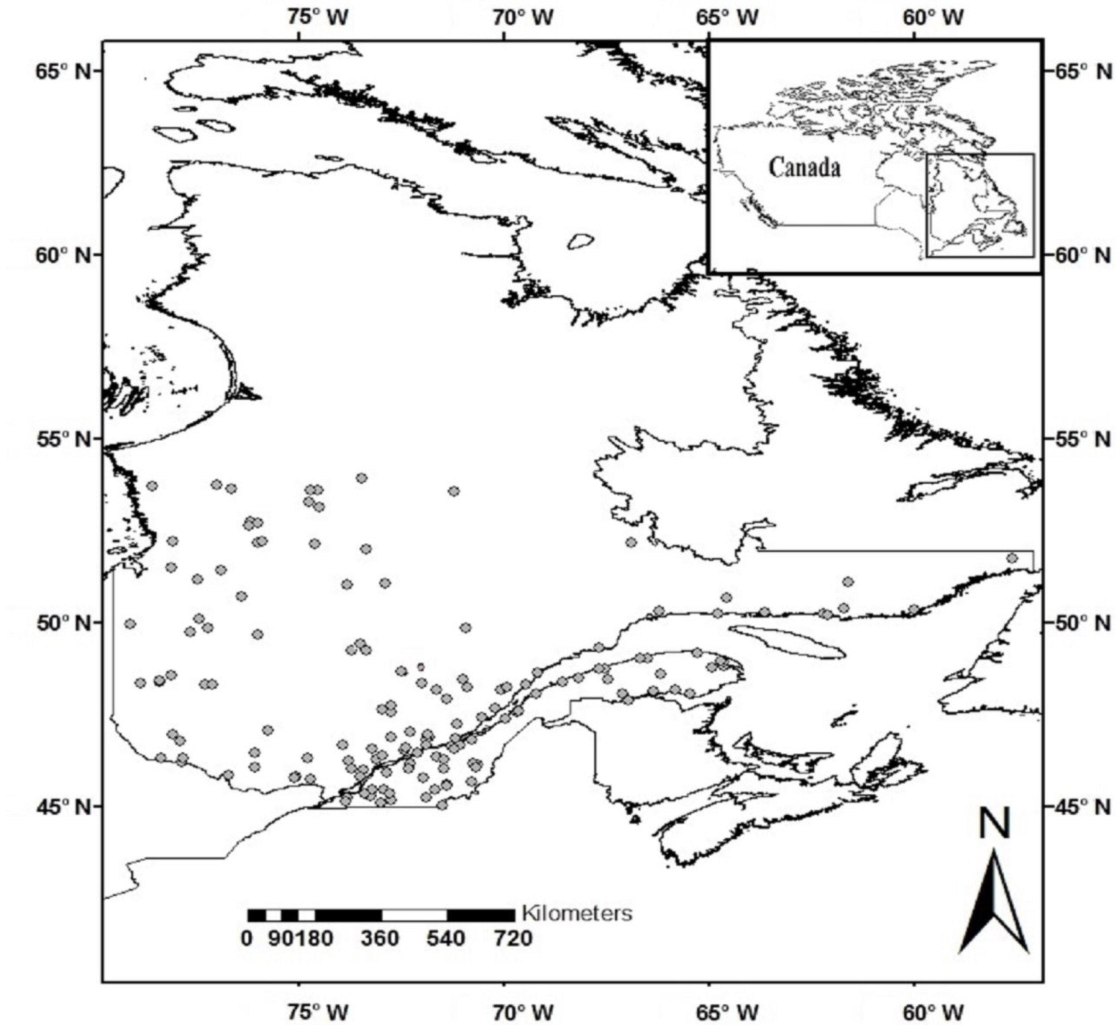

Figure 3 Geographical location of the studied sites in the southern part of the province of Quebec, Canada.

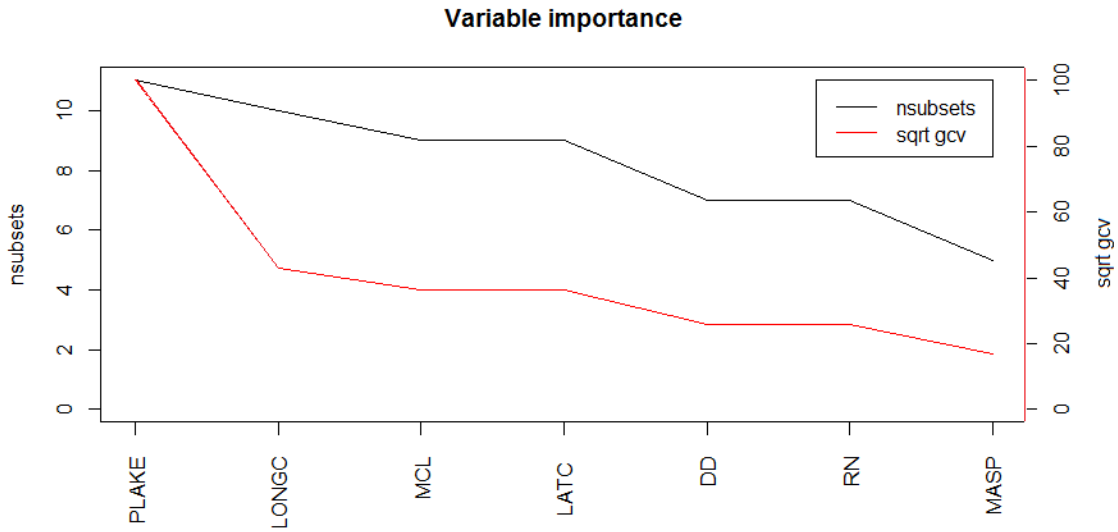

Figure 4 Variable Importance while predicting QS_{100}. The Redline represents the variation of the sqrt GCV values caused by the removal of a given variable from the MARS model during the backward phase. The black line represents the variation of the number of sub-models including a given variable.

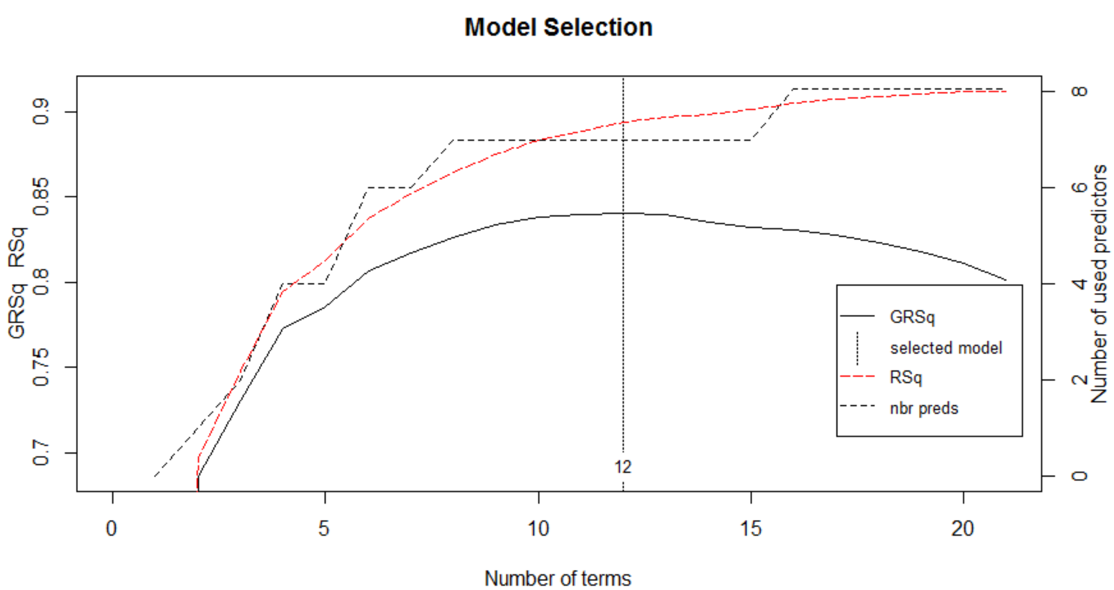

Figure 5 MARS model selection for QS_{100}. The gray line and the red dashed line represent, respectively, the variation of the GCV R^2 (GRSq) and the R^2 (RSq) values in the backward phase. For this model, 12 terms were retained which are based on 7 predictors (nbr preds).

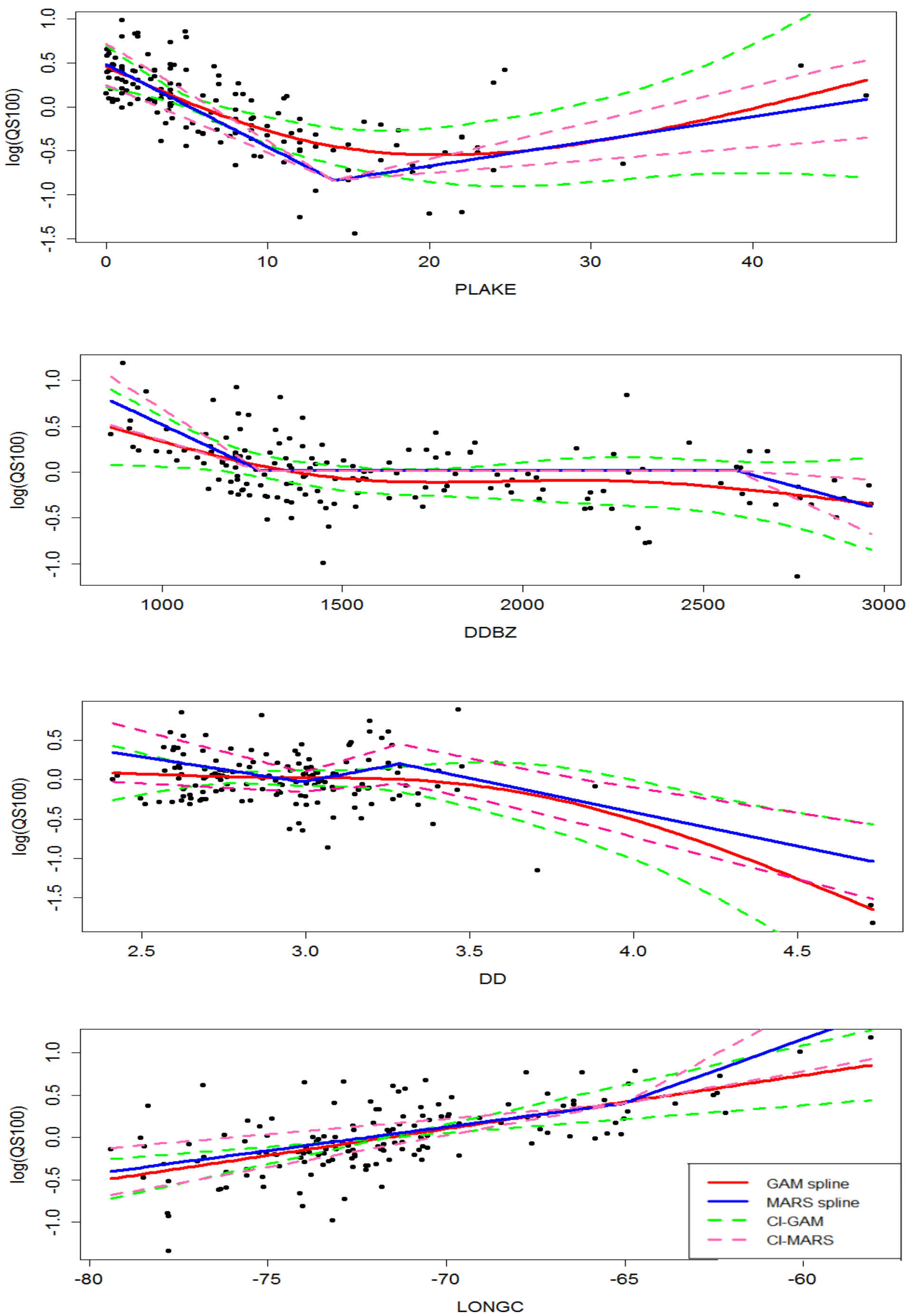

Figure 6 Examples of smoothing functions produced by the GAM and MARS models for some explanatory variables. Dashed lines represent the 95% confidence intervals (CI). A Bayesian approach to variance estimation is used to calculate the CI for GAM. For MARS, the approach considered to identify the CI for MARS is the one that we can use for a linear regression model as it is simply a linear regression of linear basis functions. All the terms are estimated with a sum to zero constraint, leading to lower uncertainty associated with the mean in the plots.

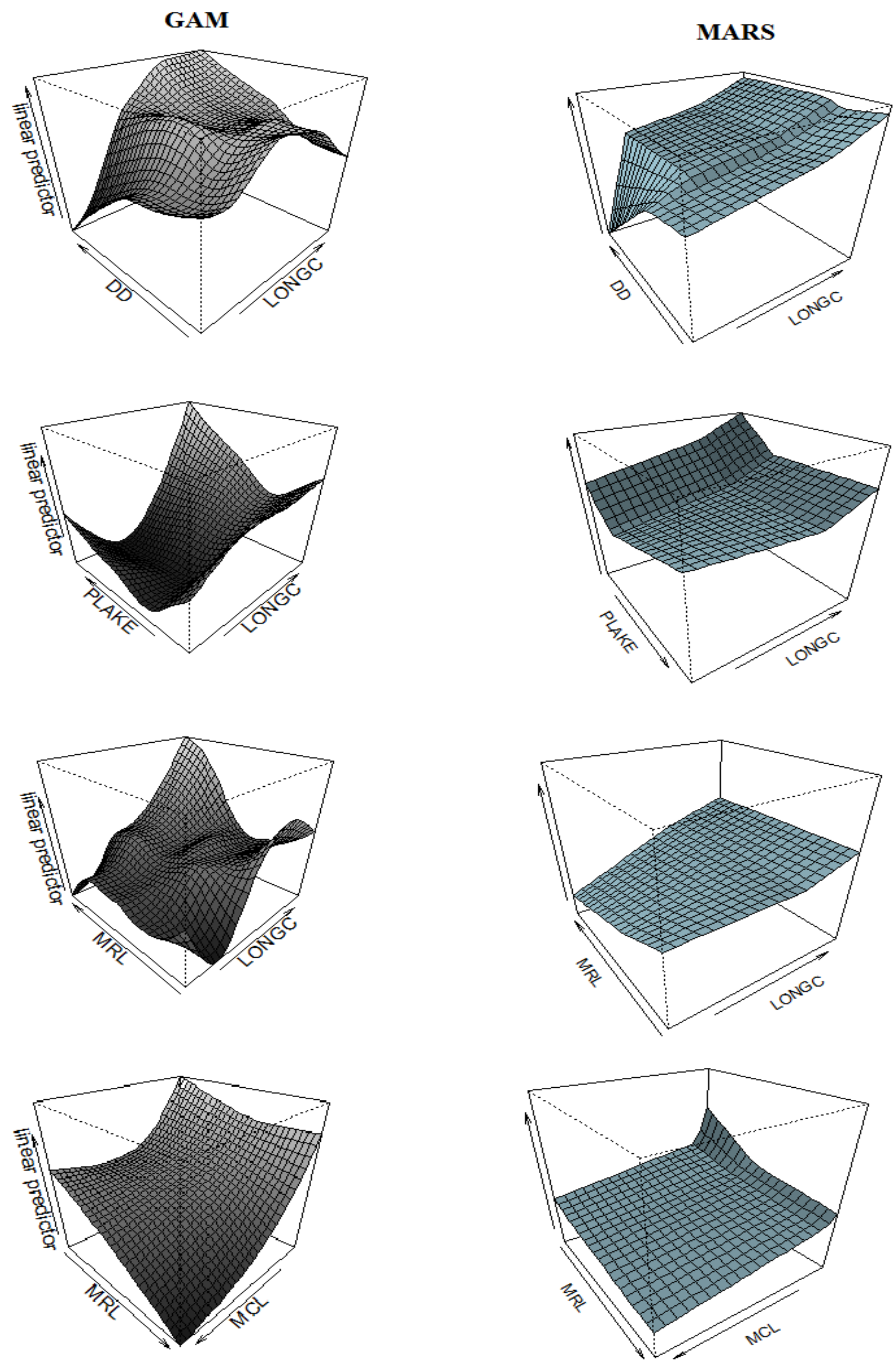

Figure 7 Examples of the multivariate effects of some explanatory variables produced by the GAM and MARS models on the response variable (interactions).

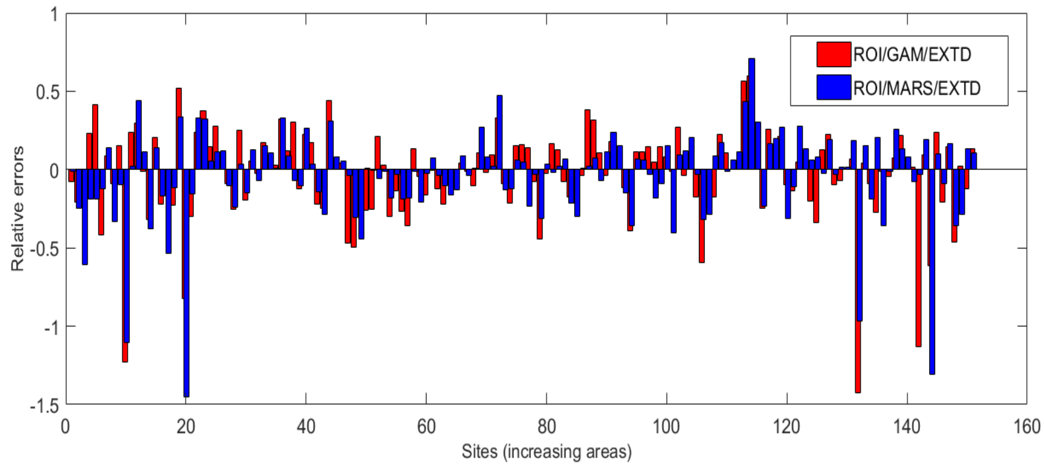

Figure 8 Relative errors associated to the at site quantile QS_{100} calculated using ROI/GAM/EXTD and ROI/MARS/EXTD.

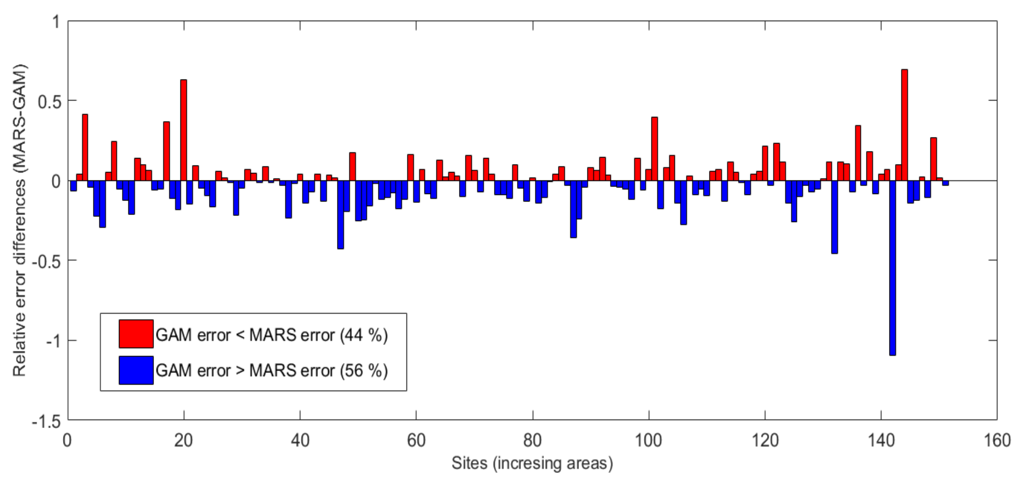

Figure 9 Relative errors differences associated to the at site quantile QS100 calculated between MARS and GAM. The considered combinations are ROI/GAM/EXTD and ROI/MARS/EXTD.