REGIONAL FREQUENCY ANALYSIS AT UNGAUGED SITES WITH THE GENERALIZED ADDITIVE MODEL

F. Chebana*[1], C. Charron[2], T.B.M.J. Ouarda[2,1] and B. Martel[1]

[1]INRS-ETE, University of Quebec, 490 de la Couronne, Québec (Qc), Canada, G1K 9A9

[2] Institute Center for Water and Environment (iWATER),
Masdar Institute of Science and Technology,
P.O.Box 54224, Abu Dhabi, UAE

*Corresponding author

Email: fateh.chebana@ete.inrs.ca

Tel: +1 418 654 2542

Revised version

July 2014

Abstract

The log-linear regression model is one of the most commonly used models to estimate flood quantiles at ungauged sites within the regional frequency analysis (RFA) framework. However, hydrological processes are naturally complex in several aspects including nonlinearity. The aim of the present paper is to take into account this nonlinearity by introducing the generalized additive model (GAM) in the estimation step of RFA. A neighbourhood approach using canonical correlation analysis (CCA) is used to delineate homogenous regions. GAMs possess a number of advantages such as flexibility in shapes of the relationships as well as the distribution of the output variable. The regional model is applied on a dataset of 151 hydrometrical stations located in the province of Québec, Canada. A stepwise procedure is employed to select the appropriate physio-meteorological variables. A comparison is performed based on different elements (regional model, variable selection and delineation). Results indicate that models using GAM outperform models using the log-linear regression as well as other methods applied to this dataset. In addition, GAM is flexible and allows including and showing non linear effects of explanatory variables, in particular basin area effect (scale). Another finding is the reduced effect of CCA delineation when combined with GAM.

Keywords

Regional frequency analysis; ungauged basin; Flood; Generalized additive model; GAM; Non-linear model, Canonical correlation analysis.

1. Introduction

Knowledge of flood characteristics is very important for resource management and design of hydraulic structures. Estimation of design flows is often needed at locations where little or no information is available. In this case, regional frequency analysis (RFA) is often used for the estimation of flow characteristics. Ouarda et al. (2008) presented a detailed review of the various available RFA methods (Blöschl et al. 2013). Generally, RFA is composed of two main steps: the identification of groups of hydrologically homogeneous basins and the application of a regional estimation method within each delineated region (GREHYS 1996a; Ouarda 2013). Since flow characteristics are highly dependent upon physiographical and meteorological basin characteristics, these can be used to estimate flood quantiles at un-gauged sites. The hydrological literature abounds with studies dealing with the development and evaluation of methods for the delineation of hydrological regions and for the study of their homogeneity. However, much less attention has been dedicated to the development of new regional estimation methods.

In the present study, canonical correlation analysis (CCA) is used to delineate homogenous regions. In GREHYS (1996b), it was shown that this method produced the best performances in comparison to other ones. Among RFA estimation methods, regression models and index-flood models are commonly used. GREHYS (1996b) showed that their performances are equivalent and are superior to other models. Generally, regression models such as linear regression models (LRM) or log-linear regression models (LLRM) are preferred for their simplicity and rapidity, as well as their performances. LLRM has been used in conjunction with CCA in many studies (Chokmani and Ouarda 2004; Ouarda et al. 2001). Linear models imply that the relations between the dependent variable (hydrologic) and the predictors (physio-meteorological) are linear. This is generally not realistic and can be problematic in some situations such as the effect of the basin size on flood quantiles, where it is documented that small basins behave differently than large ones. The basin hydrologic response is also not linearly related to the slope of the basin, as larger basin slopes (which are often associated to smaller size basins) lead to much more intense flood responses and very extreme specific peak values.

The generalized additive models, GAMs (Hastie and Tibshirani 1986) allow to take into account possible nonlinearities which is not possible through linear models or by using simple variable transformations such as log, power or square root. The use of a nonlinear model is justified by the fact that hydrological processes are naturally nonlinear (Kundzewicz and Napiórkowski 1986; Wittenberg 1999). Pandey and Nguyen (1999) compared a number of regional flood quantile estimation methods for the power regression model (equivalently log-linear) and found that nonlinear estimation methods (within the same power model) outperformed the log-linear one. Shu and Ouarda (2007) used an artificial neural network approach, which represents a nonlinear model, and obtained better results than with linear regression methods.

GAMs are an extension of the generalized linear models, GLMs (Nelder and Wedderburn 1972). The latter brought flexibility to regression methods by allowing non-normal residuals as well as a general link between predictors and the response variable. In addition, GAMs use non-parametric smooth functions to link the dependant variable to the predictors. Therefore, they are more flexible and can capture more realistically the relation between variables. GAMs have been attracting high attention in statistical developments as well as in practical applications (Hastie and Tibshirani 1986; Kauermann and Opsomer 2003; Marx and Eilers 1998; Morlini 2006; Schindeler et al. 2009; Wood 2003). Recently, additional methodological developments and the availability of implemented computer programs made GAMs increasingly popular in practical research, mainly in the public health and epidemiology fields (Bayentin et al. 2010; Cans and

Lavergne 1995; Leitte et al. 2009; Rocklöv and Forsberg 2008; Vieira et al. 2009) and in environmental studies (Borchers et al. 1997; Wen et al. 2011; Wood and Augustin 2002). In the field of meteorology, GAMs were used to model the effect of traffic and meteorology on air quality (Bertaccini et al. 2012), to predict air temperature from satellite surface temperature (Kloog et al. 2012), as well as to model mean temperature in mountainous regions (Guan et al. 2009). In hydrological modeling, very few studies employed GAMs. For instance, Tisseuil et al. (2010) used GLM and GAM for the statistical downscaling of general circulation model outputs to local-scale river flows. GAMs were used to estimate nonlinear trends in water quality by Morton and Henderson (2008) and in hydrological extreme series modeling by Ramesh and Davison (2002). Recently, Asquith et al. (2013) employed GAMs to develop readily implemented procedures for the estimation of discharge and velocity from selected predictors at ungauged stream locations. However, to the author's best knowledge, GAMs have never been used in the context of RFA of hydrological variables.

The objective of the present study is to introduce GAMs in a complete regional model to estimate flood quantiles. A set of 151 basins in the province of Québec, Canada, is considered as case study. It is used in combination with the neighborhood approach using CCA. A cross validation is used to evaluate performances. In previous studies dealing with the estimation of flood quantiles with the same dataset (Chokmani and Ouarda 2004; Kamali Nezhad et al. 2010; Shu and Ouarda 2007), explanatory variables have been selected based on correlation with specific quantiles. In the present study an attempt is made to select optimal variables with a stepwise method. The regional model adopting GAM is compared with a model using LLRM, which is commonly used in RFA. Comparisons are also carried out for models with and without the delineation of homogenous regions with CCA, and also with and without the use of the stepwise method for the selection of variables. The latter is important to separate the impacts of using the GAM model and the stepwise variable selection procedure.

This paper is organized as follows. Section 2 presents the theoretical background on linear regression models, GAMs and the CCA approach for the delineation of neighborhoods in RFA. The considered dataset as well as the study design are presented in section 3. Section 4 includes the obtained results, while the last section contains the conclusions of the study.

2. Theoretical Background

In this section, the required statistical tools are briefly presented and their use in RFA is discussed.

2.1. Linear regression models

Regression analysis is used to find a relationship between a random variable *Y*, called the response variable or dependant variable, and one or several random variables *X*, called the explanatory or predictor variables (or independent variables). Let us define **X**, a matrix whose columns are $X_1, X_2, \ldots, X_m$, a set of *m* explanatory variables. The linear regression model is defined by:

$$Y = \beta_0 + \sum_{j=1}^{m} \beta_j X_j + \varepsilon \qquad (1)$$

where β_0 and β_j are unknown parameters and ε is the error term which is assumed to be normally distributed $N\left(0, \sigma^2\right)$. The model parameters are often estimated by the least squares estimator $\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'Y$.

A power product model is generally used to express the relationship between flood quantiles and explanatory variables (Ouarda et al. 2008; Pandey and Nguyen 1999). A log transformation allows expressing this model as follows (log-linear model):

$$Y = \log(\beta_0) + \sum_{j=1}^{m} \beta_j \log\left(X_j\right) + \varepsilon \quad (2)$$

Note that the log transformation introduces a bias in the prediction since the aim is the estimation of the variable expectation rather than its logarithm (Girard et al. 2004).

2.2. Generalized additive models

The generalized linear models (GLMs) are a generalization of the well-known ordinary linear model presented previously. They allows for a response distribution other than normal and for a degree of nonlinearity in the model structure (Wood 2006). The GLM can be expressed as follows:

$$g(Y) = \beta_0 + \sum_{j=1}^{m} \beta_j X_j + \varepsilon \quad (3)$$

where g is a monotonic link function, and Y could have whatever distribution from the exponential family which includes, for instance, Poisson, Binomial and Normal distributions.

For more flexibility, GLMs are themselves extended to GAMs by allowing non-parametric fits of the X_j where the linear forms are replaced by smooth functions f_j (Hastie and Tibshirani 1986; Wood 2006):

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

GAM has several advantages over linear models. It is more flexible due to the smooth functions f_j where there is no need for a transformation to achieve linearity. Hence, it is possible to identify more realistically the effect of each explanatory variable X_j on Y.

In order to estimate the smooth function f_j, a spline is used. A spline is a curve composed of piecewise polynomial functions, joined together at points called knots. A number of spline types have been proposed in the literature, such as cubic splines, P-splines and B-splines. The thin plate regression splines have some advantages such as fast computation, lack of requirement for a choice of knot locations, and optimality in approximation of the smoothing, for more details see (Wood 2003, 2006). In the present study, the latter splines are considered.

In general, a smooth function f_j can be defined by a set of q spline basis functions $b_{ji}(x)$ such that:

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

where β_{ji} represents the smoothing coefficients related to the jth function. To avoid overfitting, the estimator $\hat{\beta}$ of β is obtained by maximizing the penalized log-likelihood:

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_{j=1}^{m} \lambda_j \beta^T \mathbf{S}_j \beta \qquad (6)$$

where $l_p(.)$ is the log-likelihood function, λ_j is the smoothing parameter of the j^{th} smooth function f_j and $\mathbf{S}_j$ is a matrix with known coefficients (Wood 2008). The parameter λ_j controls the smoothness degree of the curve f_j. Its value ranges from 0 to 1, with 0 corresponding to the un-penalised case and 1 to the completely smoothed curve. The optimum value of λ_j is a right balance between best fitting and smoothing. The function $l_p(.)$ is maximized by the penalized iteratively reweighted least squares, P-IRLS (Wood 2004). The smoothing parameter λ can be selected according to a criterion such as the generalized cross validation, GCV (Wahba 1985), unbiased risk estimator, UBRE (Craven and Wahba 1978) or maximum likelihood (ML).

2.4. CCA Approach in RFA

This section briefly presents the CCA approach and its connection to the delineation step of RFA. This method is explained in more details in Ouarda et al. (2001) in the RFA context. Let us define two sets of random variables $\mathbf{X}=\{X_1,X_2,...,X_r\}$ and $\mathbf{Y}=\{Y_1,Y_2,...,Y_s\}, s\geq r$. In the present study, the set **X** contains basin physiographical and meteorological variables, e.g. drainage area and mean annual precipitation, and **Y** contains basin hydrological variables such as flood quantiles. In general, all variables should be standardized and transformed for normality. Mainly, CCA aims to identify the dominant linear modes of covariability between the vectors **X** and **Y**, and then make inference about **Y** given the vector **X**.

Consider the linear combinations **V** and **W** of the variables of **X** and **Y**:

$$\mathbf{V}=a_1X_1+a_2X_2+\cdots+a_rX_r=\mathbf{a}'\mathbf{X} \text{ and } \mathbf{W}=b_1Y_1+b_2Y_2+\cdots+b_sY_s=\mathbf{b}'\mathbf{Y} \quad (7)$$

CCA allows to identify vectors **a** and **b** for which $\delta_{i,CCA}=corr(V_i,W_i)\quad i=1,...,p$ are maximized as well as $corr(W_i,V_j)=0,\quad i\neq j$ with unit variance.

For each basin B_k, $k=1,\ldots,K$ within a given set of basins *B*, the corresponding values for $\mathbf{V}_i$ and $\mathbf{W}_i$ are denoted as $\mathbf{v}_{i,k}$ and $\mathbf{w}_{i,k}$. Let $\mathbf{v}_0$ denote the physio-meteorological canonical score for a target site, associated to the obtained canonical variables. The vector $\mathbf{v}_0$ is known whereas the interest is the estimation of the unknown hydrological canonical score $\mathbf{w}_0$. The approximation can be obtained through $\Lambda \mathbf{v}_0$ such that $\Lambda = diag(\delta_{1,CCA}, ..., \delta_{p,CCA})$. This leads to the definition of the 100(1-α)% confidence level neighbourhood for $\Lambda \mathbf{v}_0$ containing sites with realizations $\boldsymbol{w}$ of $\boldsymbol{W}$ such that:

$$(\boldsymbol{w} - \Lambda \boldsymbol{v}_0)^T (I_p - \Lambda^2)^{-1} (\boldsymbol{w} - \Lambda \boldsymbol{v}_0) \leq \chi^2_{\alpha,p} \quad (8)$$

where I_p is the $p \times p$ identity matrix and $\chi^2_{\alpha,p}$ is such that $P(\chi^2 \leq \chi^2_{\alpha,p}) = 1 - \alpha$. All the aspects related to the CCA in the RFA context are developed in Ouarda et al. (2001).

3. Dataset and study design

The considered dataset has already been studied in the context of RFA in a number of previous studies (Chebana and Ouarda 2008; Chokmani and Ouarda 2004; Kamali Nezhad et al. 2010; Shu and Ouarda 2007), which provides an opportunity for comparative evaluation of the results. The dataset consists of 151 hydrometric stations located in the southern half of the province of Québec (between 45°N and 55°N), Canada. The hydrological variables are represented by specific flood quantiles (quantiles divided by the basin area), denoted by QS_{10}, QS_{50} and QS_{100}. The physiographical and meteorological variables, available for each basin, are summarized in Table 1. To avoid redundancy with the previously mentioned studies, details concerning the dataset are not reported here. The reader is referred to the references listed above for information concerning the geographic location of the stations and the scatter plots of the basins in the canonical spaces.

The CCA in conjunction with LLRM has been proven to perform well (GREHYS 1996b). However, it is suspected that the more general GAM approach can improve the estimations. In this study, LLRM and GAM are compared as regional estimation models. The fitting of data for

GAM is performed with the *R* package *mgcv* (Wood 2004). Smooth parameters, λ_j in (6), are estimated with the P-IRLS procedure where the ML score is employed as criterion

Homogenous regions are delineated with the CCA method on the basis of the variables *BV, PMBV, PLAC, PTMA* and *DJBZ*. These variables are selected on the basis of maximizing correlations with the hydrological variables. Since CCA requires normality, these variables are transformed for the regional analysis as in the previous studies for this region, i.e. a logarithmic transformation for the hydrological variables, PMBV, PTMA and DJBZ, and a square root transformation for PLAC. Figure 3 (not reported here to avoid repetition) in Shu and Ouarda (2007) shows clear nonlinearities in different levels for some variables. This represents a motivation for the use of the GAM model with the present dataset.

The design of the present study aims to check the performance of three elements: i) adoption of the CCA delineation step or considering all stations, ii) consideration of the nonlinearity in the regression model through either LLRM or GAM during the regional estimation step and iii) the variable selection method (stepwise or correlation). This leads to 8 combinations denoted as follows:

- LLRM|ALL|CORR: LLRM with all stations (no delineation) and with the 5 selected variables (from correlation);
- LLRM|ALL|STPW: LLRM with all stations (no delineation) and variables selected using the stepwise method;
- LLRM|CCA|CORR: LLRM with homogeneous regions defined by CCA and with the 5 selected variables (from correlation);
- LLRM|CCA|STPW: LLRM with homogeneous regions defined by CCA and variables selected using the stepwise method;

- GAM|ALL|CORR: GAM with all stations (no delineation) and with the 5 selected variables (from correlation);
- GAM|ALL|STPW: GAM with all stations (no delineation) and variables selected using the stepwise method;
- GAM|CCA|CORR: GAM with homogeneous regions defined by CCA and with the 5 selected variables (from correlation);
- GAM|CCA|STPW: GAM with homogeneous regions defined by CCA and variables selected using the stepwise method.

The selection method used in this study is the backward stepwise selection method. It starts with an initial model including all available variables. The regression method is then applied with the current model and the variable with the highest *p*-value is excluded, corresponding to the hypothesis that $\beta_j = 0$ in (5) where *j* is the *j*th variable. At each step, one variable is excluded. The procedure ends when the *p*-values of all the remaining and significant variables are under a given threshold (5%).

Once a model is established, its performance can be evaluated. A jackknife procedure is applied to assess the performance of the models. In this procedure, gauged sites are in turn considered ungauged in order to carry out regional estimation. This procedure allows assessing the following performance criteria:

the coefficient of determination $$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \quad (9)$$

the root mean square error $$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2} \quad (10)$$

the relative root mean square error $$\text{rRMSE} = 100\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(z_i - \hat{z}_i)/z_i\right]^2} \qquad (11)$$

the mean bias $$\text{BIAS} = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i) \qquad (12)$$

the relative mean bias $$\text{rBIAS} = 100\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)/z_i \qquad (13)$$

where z_i and $\hat{z}_i$ are respectively the local (at site) and regional quantile estimates at station *i*, $\overline{z}$ is the local mean of the hydrological variable and *n* is the number of stations.

4. Results and discussion

The CCA is applied on the dataset with the normalized variables *BV, PMBV, PLAC, PTMA* and *DJBZ*. An optimal value of $\alpha = 0.05$ is obtained with the optimisation procedure of Ouarda et al. (2001). This optimal value is used to delineate the neighborhood at each station. Each regional model, when considering CCA delineation, uses the same neighbourhood for a given station. When CCA is applied to the whole dataset, the two physiographical-meteorological canonical variables are defined as:

$$V_1 = 0.24\log(BV) - 0.07\log(PMBV) + 0.58\sqrt{PLAC} - 0.33\log(PTMA) - 0.03\log(DJBZ) \qquad (14)$$

$$V_2 = 0.48\log(BV) - 0.25\log(PMBV) - 0.45\sqrt{PLAC} + 1.05\log(PTMA) + 1.10\log(DJBZ) \qquad (15)$$

and the two hydrological canonical variables are defined as:

$$W_1 = 2.14\log(QS_{10}) - 13.14\log(QS_{50}) + 10.03\log(QS_{100}) \qquad (16)$$

$$W_2 = 6.27\log(QS_{10}) + 2.45\log(QS_{50}) - 8.84\log(QS_{100}) \qquad (17)$$

The non-negligible values of the *BV* coefficient in V_1 and V_2 confirm the need to include *BV* in the CCA despite the fact that specific hydrological quantiles are used.

The stepwise selection of variables is applied for each specific quantile separately and for each regression model LLRM and GAM. Table 2 indicates that the selected variables are the same for a given model and a given selection method, independently of whether CCA is used for homogeneous region delineation. Therefore, the delineation step seems not to have an effect on the selected variables.

The results of the application of the jackknife procedure for the performance evaluation of each regional model are presented in Table 3. The best overall performances are obtained with GAM|ALL|STPW and GAM|CCA|STPW with CCA leading to slightly better performances. More precisely and in particular based on the rRMSE, GAM always performs better than LLRM for combinations using the same variable selection approach and the same delineation approach (CCA or ALL).

The use of CCA to delineate hydrologically homogeneous regions generally leads to improvements in regional estimation in comparison to the ALL approach for the same selection of variables and the same regression model (GAM or LLRM). However, when GAM is used, the difference between CCA and ALL is not significant especially when using the stepwise procedure for the selection of variables. These results show that the use of GAM makes the procedure more robust and compensates for the advantages of using CCA. This is not the case for LLRM where the use of CCA was shown to lead to significant improvements, see e.g. Chokmani and Ouarda (2004). In other words, this indicates that the use of GAM reduces the importance of delineating the appropriate hydrological neighborhood. A possible interpretation for this result is that the consideration of non-linear formulations in the relation between the explanatory physiographical and meteorological variables on one side and the hydrological variables on the other side leads to a reduction of the weight of basins that are not hydrologically similar to the target site.

The stepwise method for variable selection improves quantile estimations in comparison to those obtained with the fixed 5 variables. This can be explained by the fact that the correlation-based selection of physiographical and meteorological variables to be used in the model is mainly based on a linear relationship between variables. It must also be noted that the variables are originally selected for CCA purposes (delineation) rather than for regression modeling (estimation).

Figures 1 and 2 present the smooth functions f_j of the response variable log(QS100) with the explanatory variables of the fitted models GAM|ALL|CORR and GAM|ALL|STPW respectively. It can be seen that the variables BV, PLAC, LAT and DJBZ show nonlinear relations. Furthermore, the nonlinear relation is more complex for some variables. For instance, the relationship between log(QS100) and DJBZ decreases for small values of DJBZ, increases for midrange values and decreases again for high values of DJBZ. This result reflects the seasonality effect of temperature, through DJBZ, on the flood regime. Another particular example of interest concerns the BV variable. Indeed, it can be seen that small basins have a different effect than moderate basins. This result is important since nonlinearity allows appropriately including the variable BV in the model which eliminates the need to develop specific models for small, moderate or large basins. Variables PMBV, LONG, PLMA and PTMA have approximately linear relations.

In the present study, the proposed approach based on GAM is mainly compared with the basic formulation of one of the most popular RFA approaches, which is the log-linear estimation model combined with the CCA delineation approach. The comparison can be extended to other regional flood frequency models, such as the ensemble artificial neural networks-CCA approach (EANN-

CCA) (Shu and Ouarda 2007; Shu and Ouarda 2008), the kriging-CCA approach (Chokmani and Ouarda 2004), and the depth-based approach (Chebana and Ouarda 2008; Wazneh et al. 2013a, 2013b). In order to widen the comparison, results corresponding to the above approaches are considered since they are already available for the data set considered in the present study. Table 4 summarizes the obtained results for all these methods. The results indicate that the GAM-based approach outperforms significantly all the above listed approaches in terms of rRMSE. In terms of rBIAS, the optimal depth-based approach seems to lead to slightly better results, although the difference is not significant.

5. Conclusions

GAM is commonly used in health, epidemiological and environmental studies. However, it remains unutilized in the field of hydrology, especially in RFA. The multiple linear regression model is the most employed estimation model in RFA mainly because of its simplicity. However, it assumes a log linear relationship between the response variable and the explanatory variables. This assumption is not always true and does not reflect the complexity of the hydrological processes involved. The purpose of the present study is first to introduce GAM in RFA and then to compare its results with those obtained by LLRM. GAM is a flexible model that relaxes the assumptions of the LLRM model (normality and linearity).

Results of this study indicate that significantly better estimations are obtained from regional models with GAM. For some explanatory variables, the logarithmic relationship of the response variable with the explanatory variables is not linear. Smooth curves allow for a more realistic understanding of the true relationship between response and explanatory variables. The performance gain is not significant using CCA in conjunction with GAM compared to LLMR. This indicates that GAM is robust and is efficient in RFA even without use of a neighborhood approach. Further efforts are required to generalize this conclusion and to test the benefits of GAM modeling in other hydrological applications.

In summary, the use of GAM in RFA is valuable not only in terms of performance but also in terms of other practical aspects (e.g. explicit formulation of the smooth functions, flexibility, reduced number of assumptions, and less subjective choices).

Acknowledgments

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada, under funding to the Canada Research Chair on the Estimation of Hydro-Meteorological variables. The authors thank the ministry of the environment of Quebec (MENVIQ) services for the employed data sets. For confidentiality reasons, the data cannot be released. The authors are grateful to the Editor and the anonymous reviewers for their valuable comments and suggestions.

References

Asquith, W. H., G. R. Herrmann, and T. G. Cleveland, 2013: Generalized Additive Regression Models of Discharge and Mean Velocity Associated with Direct-Runoff Conditions in Texas: Utility of the U.S. Geological Survey Discharge Measurement Database. *Journal of Hydrologic Engineering*, **18,** 1331-1348.

Bayentin, L., S. El Adlouni, T. B. M. J. Ouarda, P. Gosselin, B. Doyon, and F. 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**.

Bertaccini, P., V. Dukic, and R. Ignaccolo, 2012: Modeling the short-term effect of traffic and meteorology on air pollution in turin with generalized additive models. *Advances in Meteorology*, **2012**.

Blöschl, G., M. Sivapalan, T. Wagener, A. Viglione, and H. Savenije, 2013: *Runoff prediction in ungauged basins. Synthesis across processes, places and scales.* Cambridge University Press.

Borchers, D. L., S. T. Buckland, I. G. Priede, and S. Ahmadi, 1997: Improving the precision of the daily egg production method using generalized additive models. *Canadian Journal of Fisheries and Aquatic Sciences*, **54,** 2727-2742.

Cans, C., and C. Lavergne, 1995: De la régression logistique vers un modèle additif généralisé : un exemple d'application. *Revue de Statistique Appliquée*, **43,** 77-90. .

Chebana, F., and T. B. M. J. Ouarda, 2008: Depth and homogeneity in regional flood frequency analysis. *Water Resour. Res.*, **44,** W11422.

Chokmani, K., and T. B. J. M. Ouarda, 2004: Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. *Water Resour. Res.*, **40,** W12514.

Craven, P., and G. Wahba, 1978: Smoothing noisy data with spline functions. *Numer. Math.*, **31,** 377-403.

Girard, C., T. B. M. J. Ouarda, and B. Bobée, 2004: Study of the bais in the log-linear regional estimation model. *Can. J. Civ. Eng.*, **31,** 361-368.

GREHYS, 1996a: Presentation and review of some methods for regional flood frequency analysis. *Journal of Hydrology*, **186,** 63-84.

——, 1996b: Inter-comparison of regional flood frequency procedures for Canadian rivers. *Journal of Hydrology*, **186,** 85-103.

Guan, B. T., H. W. Hsu, T. H. Wey, and L. S. Tsao, 2009: Modeling monthly mean temperatures for the mountain regions of Taiwan by generalized additive models. *Agricultural and Forest Meteorology*, **149,** 281-290.

Hastie, T., and R. Tibshirani, 1986: Generalized Additive Models. *Statistical Science*, **1,** 297-310.

Kamali Nezhad, M., K. Chokmani, T. Ouarda, M. Barbet, and P. Bruneau, 2010: Regional flood frequency analysis using residual kriging in physiographical space. *Hydrological Processes*, **24,** 2045-2055.

Kauermann, G., and J. D. Opsomer, 2003: Local Likelihood Estimation in Generalized Additive Models, **30,** 317-337.

Kloog, I., A. Chudnovsky, P. Koutrakis, and J. Schwartz, 2012: Temporal and spatial assessments of minimum air temperature using satellite surface temperature measurements in Massachusetts, USA. *Science of the Total Environment*, **432,** 85-92.

Kundzewicz, Z. W., and J. J. Napiórkowski, 1986: Non linear models of dynamic hydrology. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, **31,** 163-185.

Leitte, A. M., and Coauthors, 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,** 4004-4011.

Marx, B. D., and P. H. C. Eilers, 1998: Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, **28,** 193-209.

Morlini, I., 2006: On Multicollinearity and Concurvity in Some Nonlinear Multivariate Models. *Statistical Methods & Applications*, **15,** 3-26.

Morton, R., and B. L. Henderson, 2008: Estimation of nonlinear trends in water quality: An improved approach using generalized additive models. *Water Resour. Res.*, **44**.

Nelder, J. A., and R. W. M. Wedderburn, 1972: Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, **135,** 370-384.

Ouarda, T. B. M. J., 2013: Regional Hydrological Frequency Analysis. *Encyclopedia of Environmetrics*, John Wiley & Sons, Ltd.

Ouarda, T. B. M. J., A. St-Hilaire, and B. Bobée, 2008: Synthèse des développements récents en analyse régionale des extrêmes hydrologiques / A review of recent developments in regional frequency analysis of hydrological extremes. *Revue des sciences de l'eau / Journal of Water science*, **21,** 219-232.

Ouarda, T. B. M. J., C. Girard, G. S. Cavadias, and B. Bobee, 2001: Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology*, **254,** 157-173.

Pandey, G. R., and V. T. V. Nguyen, 1999: A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology*, **225,** 92-101.

Ramesh, N. I., and A. C. Davison, 2002: Local models for exploratory analysis of hydrological extremes. *Journal of Hydrology*, **256,** 106-119.

Rocklöv, J., and B. Forsberg, 2008: The effect of temperature on mortality in Stockholm 1998-2003: A study of lag structures and heatwave effects. *Scandinavian Journal of Public Health*, **36,** 516-523.

Schindeler, S., D. Muscatello, M. Ferson, K. Rogers, P. Grant, and T. Churches, 2009: Evaluation of alternative respiratory syndromes for specific syndromic surveillance of influenza and respiratory syncytial virus: a time series analysis. *BMC Infectious Diseases*, **9,** 190.

Shu, C., and T. B. J. M. Ouarda, 2007: Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. *Water Resour. Res.*, **43,** W07438.

Shu, C., and T. B. M. J. Ouarda, 2008: Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. *Journal of Hydrology*, **349,** 31-43.

Tisseuil, C., M. Vrac, S. Lek, and A. J. Wade, 2010: Statistical downscaling of river flows. *Journal of Hydrology*, **385,** 279-291.

Vieira, V., T. Webster, J. Weinberg, and A. Aschengrau, 2009: Spatial analysis of bladder, kidney, and pancreatic cancer on upper Cape Cod: an application of generalized additive models to case-control data. *Environmental Health*, **8,** 3.

Wahba, G., 1985: A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Stat.*, **13,** 1378-1402.

Wazneh, H., F. Chebana, and T. B. M. J. Ouarda, 2013a: Optimal depth-based regional frequency analysis. *Hydrology and Earth System Sciences*, **17,** 2281-2296.

——, 2013b: Depth-based regional index-flood model. *Water Resour. Res.*, **In presss**.

Wen, L., K. Rogers, N. Saintilan, and J. 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,** 154-163.

Wittenberg, H., 1999: Baseflow recession and recharge as nonlinear storage processes. *Hydrological Processes*, **13,** 715-726.

Wood, S. N., 2003: Thin plate regression splines. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, **65,** 95-114.

——, 2004: Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99,** 673-686.

——, 2006: *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC Press, 392 pp.

——, 2008: Fast stable direct fitting and smoothness selection for generalized additive models. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, **70,** 495-518.

Wood, S. N., and N. H. Augustin, 2002: GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, **157,** 157-177.

Table 1. Descriptive statistics of hydrological variables and physio-meteorological variables.

Variable	Unit	Notation	Min	Moy	Max	SD
Specific flood of 10 year return period	$m^3/s.km^2$	QS_{10}	0.03	0.22	0.53	0.13
Specific flood of 50 year return period	$m^3/s.km^2$	QS_{50}	0.03	0.28	0.77	0.18
Specific flood of 100 year return period	$m^3/s.km^2$	QS_{100}	0.03	0.31	0.94	0.20
Area of Watershed	km^2	*BV*	208	6 265	96 600	11 713
Length of main channel	km	*LCP*	17	157	855	142
Slope of main channel	m/km	*PCP*	0.20	3.23	23.60	3.22
Mean slope of watershed	°	*PMBV*	0.96	2.43	6.81	0.99
Percentage of the basin occupied by forest	%	*PFOR*	18.00	83.05	99.80	16.61
Percentage of the basin occupied by lakes	%	*PLAC*	0.03	7.72	47.00	7.99
Mean annual total precipitations	mm	*PTMA*	646	988	1 534	154
Mean annual liquid precipitations	mm	*PLMA*	423	717	1625	176
Mean annual solid precipitations	cm	*PSMA*	166	302	720	86
Mean annual liquid precipitations during summer and fall		*PLME*	306	455	664	72
Mean annual degree-days over 0°C	dgr-day	*DJBZ*	8 589	16 346	29 631	5 385
Latitude of the station	°	*LAT*	45	48	54	2
Longitude of the station	°	*LONG*	58	72	79	4
Altitude of the station	m	*ALT*	5	157	555	125

Table 2. Variables selected for each regional model.

Regional Models	Quantile	Selected explanatory variables
[LLRM|ALL|STPW], [LLRM|CCA| STPW]	QS_{10}	BV, PMBV, PFOR, PLAC, PLMA, DJBZ, LONG
	QS_{50}	BV, PMBV, PFOR, PLAC, PLMA, LONG
	QS_{100}	BV, PLAC, PLMA, LONG
[GAM|ALL|STPW], [GAM|CCA|STPW]	QS_{10}	BV, PFOR, PLAC, PTMA, LAT, LONG
	QS_{50}	BV, PLAC, PLMA, LAT, LONG
	QS_{100}	BV, PLAC, PLMA, LAT, LONG
[LLRM|ALL|CORR], [LLRM|CCA|CORR], [GAM|ALL|CORR], [GAM|ALL|CORR]	QS_{10}	BV, PMBV, PLAC, PTMA, DJBZ
	QS_{50}	BV, PMBV, PLAC, PTMA, DJBZ
	QS_{100}	BV, PMBV, PLAC, PTMA, DJBZ

Table 3. Performances obtained with the eight combinations (model, delineation and variable selection).

		LLRM				GAM			
		ALL		CCA		ALL		CCA	
	Quantiles	CORR	STPW	CORR	STPW	CORR	STPW	CORR	STPW
R^2	QS_{10}	0.62	0.63	0.76	0.78	0.77	**0.82**	0.79	**0.82**
	QS_{50}	0.56	0.63	0.68	0.72	0.68	0.75	0.73	**0.76**
	QS_{100}	0.53	0.53	0.64	0.65	0.65	**0.72**	0.69	0.67
RMSE (m3/s.km2)	QS_{10}	0.078	0.077	0.062	0.060	0.061	**0.054**	0.059	**0.054**
	QS_{50}	0.117	0.108	0.100	0.094	0.099	0.088	0.092	**0.087**
	QS_{100}	0.137	0.137	0.120	0.118	0.118	**0.106**	0.112	0.115
rRMSE (%)	QS_{10}	51.4	48.7	44.2	41.5	41.4	37.6	39.1	**33.7**
	QS_{50}	56.4	55.5	48.5	48.9	47.0	**41.0**	43.4	43.5
	QS_{100}	58.9	60.0	50.7	50.9	49.3	42.1	45.6	**37.0**
BIAS (m3/s.km2)	QS_{10}	-0.006	-0.005	-0.012	-0.009	0.007	**0.004**	0.009	0.009
	QS_{50}	-0.010	-0.011	-0.021	-0.015	0.013	0.009	0.018	**-0.003**
	QS_{100}	-0.013	-0.015	-0.026	-0.022	0.016	**0.011**	0.023	0.043
rBIAS (%)	QS_{10}	7.6	7.4	5.6	5.3	-5.4	-5.1	-4.8	**-3.5**
	QS_{50}	8.9	8.8	6.0	7.5	-6.8	-6.1	**-4.7**	-11.4
	QS_{100}	9.6	10.0	6.3	7.7	-7.6	-6.5	-4.9	**3.4**

Best performances are in bold character for each criterion and quantile

Table 4. Results of several RFA approaches applied to the same data set considered in this study

Method	References	QS_{10}		QS_{100}	
		rBIAS (%)	rRMSE (%)	rBIAS (%)	rRMSE (%)
Linear regression	Table 3 above	-9	55	-11	64
Nonlinear regression	Shu and Ouarda 2008	-9	61	-12	70
Nonlinear regression with regionalization approach	Shu and Ouarda 2008	-19	67	-24	79
Linear regression-CCA	Table 3 above	-7	44	-8	52
Kriging in the CCA Physiographical Space	Chokmani and Ouarda 2004	-20	66	-27	86
Kriging in the PCA Physiographical Space	Chokmani and Ouarda 2004	-16	51	-23	70
Adaptive Neuro-Fuzzy Inference Systems	Shu and Ouarda 2008	-8	57	-14	64
Artificial Neural Networks	Shu and Ouarda 2008	-8	53	-10	60
Single Artificial Neural Networks-CCA space	Shu and Ouarda 2007	-5	38	-4	46
Ensemble Artificial Neural Networks	Shu and Ouarda 2007	-7	44	-10	60
Ensemble Artificial Neural Networks -CCA space	Shu and Ouarda 2007	-5	37	-6	45
Optimal depth-based approach	Wazneh et al. 2013a	**-3**	38	**-2**	44
GAM|CCA|STPW	Table 3 above	**-3.5**	**33.7**	3.4	**37**

Best results are in bold character

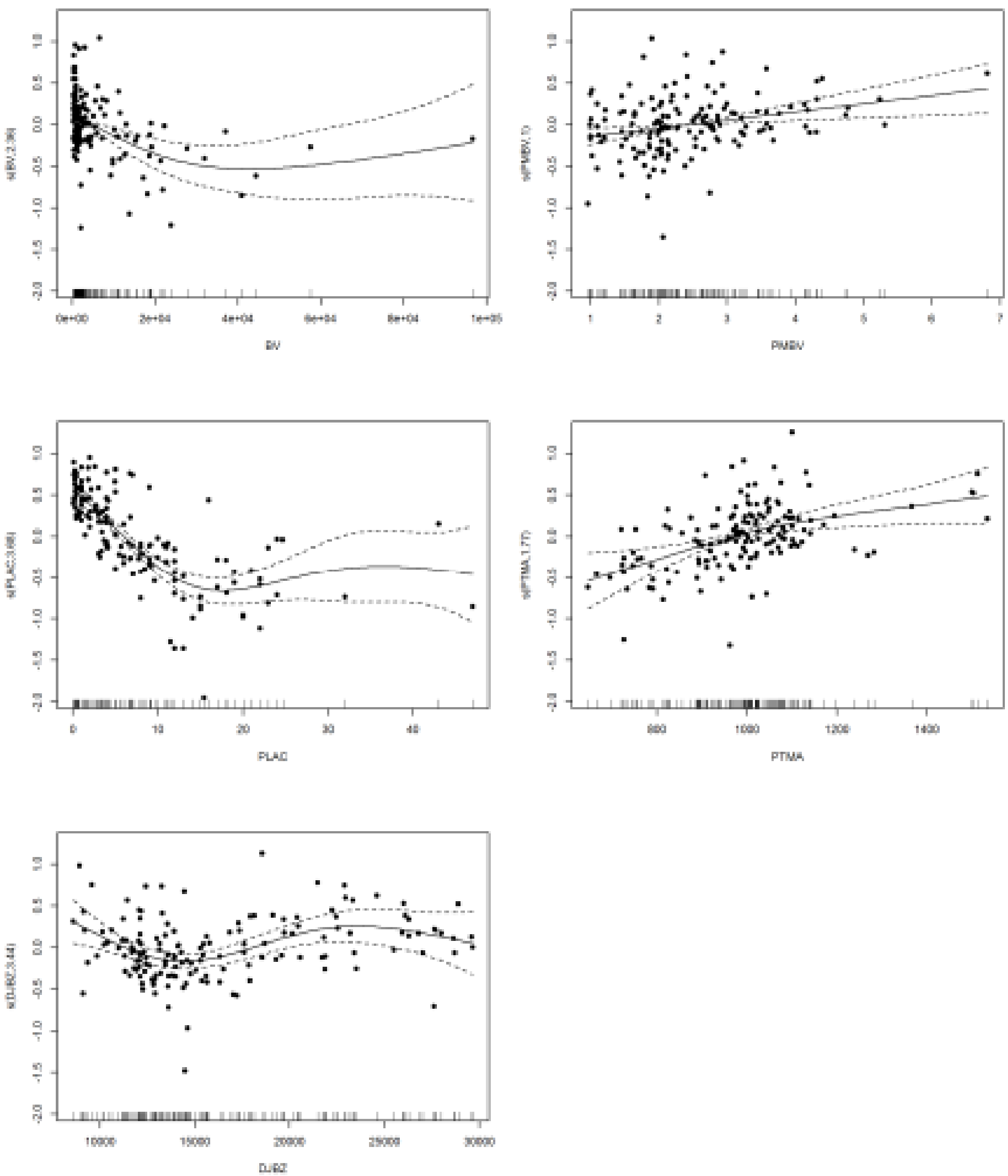

Figure 1. Smooth functions of QS_{100} for the explanatory variables included in the regional model GAM|ALL|CORR. The dotted lines represent the 95% confidence intervals. The y-axes are named s(*var*,*edf*) where *var* is the name of the explanatory variable and *edf* is the estimated degree of freedom of the smooth.

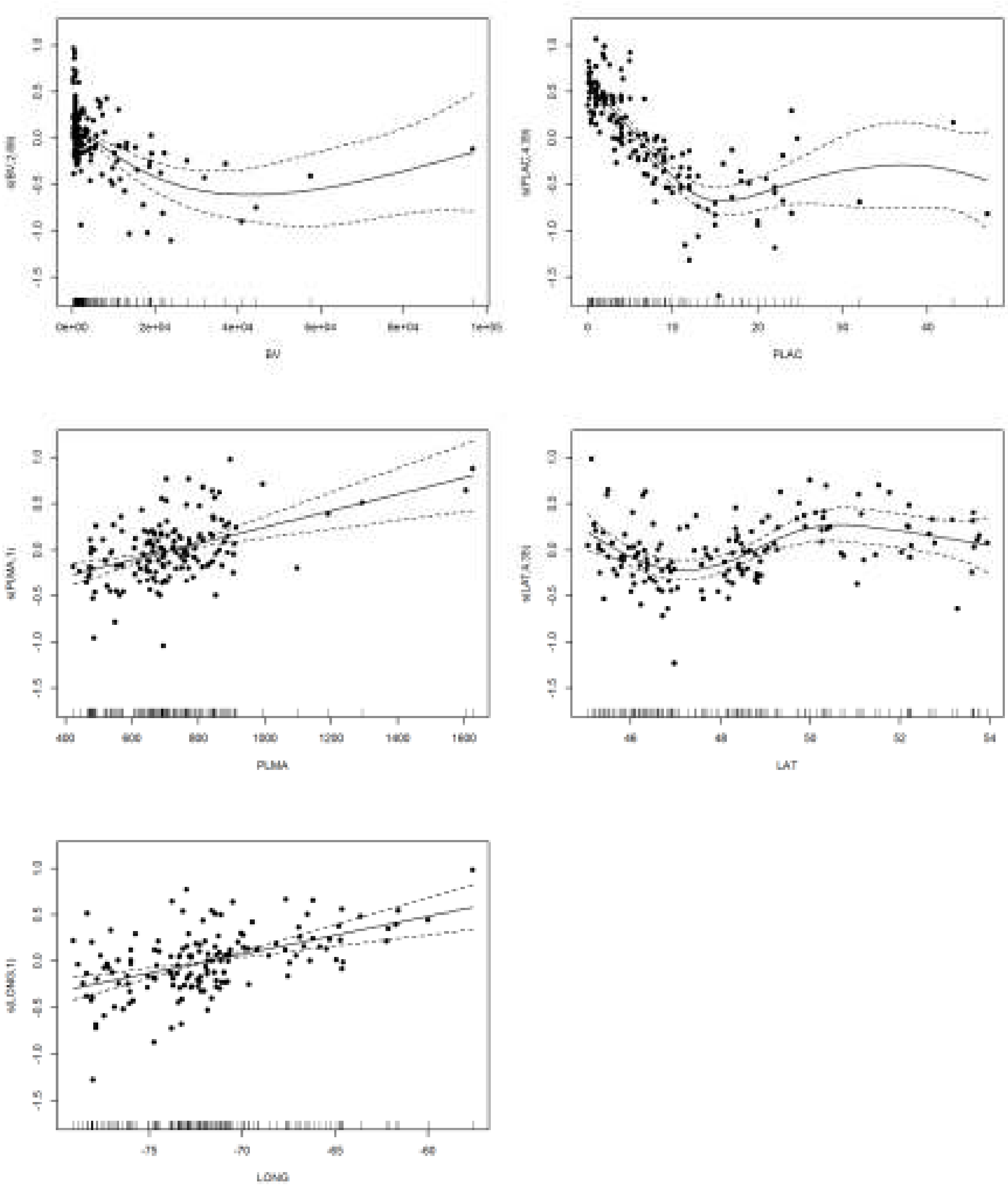

Figure 2. Smooth functions of QS_{100} for the explanatory variables included in the regional model GAM|ALL|STPW. The dotted lines represent the 95% confidence intervals. The y-axes are named s(*var*,*edf*) where *var* is the name of the explanatory variable and *edf* is the estimated degree of freedom of the smooth.