Optimal depth-based regional frequency analysis

H. Wazneh*[1], F. Chebana[1] and T.B.M.J. Ouarda[2,1]

[1]INRS-ETE, 490 rue de la Couronne, Québec (QC),
Canada G1K 9A9
[2]Masdar Institute of science and technology
P.O.Box 54224, Abu Dhabi, UAE

***Corresponding author:** Tel: +1 (418) 654 2530#4461
Email: hussein.wazneh@ete.inrs.ca

May 24th 2013

Abstract:

Classical methods of regional frequency analysis (RFA) of hydrological variables face two drawbacks: 1) the restriction to a particular region which can lead to a loss of some information and 2) the definition of a region that generates a border effect. To reduce the impact of these drawbacks on regional modeling performance, an iterative method was proposed recently, based on the statistical notion of the depth function and a weight function φ. This depth-based RFA (DBRFA) approach was shown to be superior to traditional approaches in terms of flexibility, generality and performance. The main difficulty of the DBRFA approach is the optimal choice of the weight function φ (e.g., φ minimizing estimation errors). In order to avoid subjective choice and naïve selection procedures of φ, the aim of the present paper is to propose an algorithm-based procedure to optimize the DBRFA and automate the choice of φ according to objective performance criteria. This procedure is applied to estimate flood quantiles in three different regions in North America. One of the findings from the application is that the optimal weight function depends on the considered region and can also quantify the region homogeneity. By comparing the DBRFA to the canonical correlation analysis (CCA) method, results show that the DBRFA approach leads to better performances both in terms of relative bias and mean square error.

Keywords: regional frequency analysis; statistical depth function; floods estimation; optimization; canonical correlation analysis; hydrology.

1. Introduction

Due to the large territorial extents and the high costs associated to installation and maintenance of monitoring stations, it is not possible to monitor hydrologic variables at all sites of interest. Consequently, hydrologists have often to provide estimates of design events quantiles *QT*, corresponding to a large return period *T* at ungauged sites. In this situation, regionalization approaches are commonly used to transfer information from gauged sites to the target site (ungauged or partially gauged) [e.g., Burn, 1990b; Dalrymple, 1960; Ouarda et al., 2000]. A number of estimation techniques in regional frequency analysis (RFA) have been proposed and applied in several countries [De Michele and Rosso, 2002; Haddad and Rahman, 2012; Madsen and Rosbjerg, 1997; Nguyen and Pandey, 1996; Ouarda et al., 2001].

In general, RFA consists of two main steps: (1) grouping stations with similar hydrological behavior (delineation of hydrological homogeneous regions) [e.g., Burn, 1990a] and (2) regional estimation within each homogenous region at the site of interest [e.g., GREHYS, 1996a; Ouarda et al., 2001; Ouarda et al., 2000]. The two main disadvantages of this type of regionalization methods are: i) a loss of information due to the exclusion of a number of sites in the step of delineation of hydrological homogeneous region, and ii) a border effect problem generated by the definition of a region.

To reduce or eliminate the negative impact of these disadvantages on the estimation quality, a number of regional methods have been proposed that combine the two stages (delineation and estimation) and use all stations [e.g., Ouarda et al., 2008; Shu and Ouarda, 2007; Shu and Ouarda, 2008]. One of these regional methods was developed recently by Chebana and Ouarda [2008]. This RFA method is based on statistical depth functions (denoted by DBRFA for depth-based RFA). The DBRFA approach focuses directly on quantile estimation using the weighted least squares (WLS) method to estimate parameters and avoids the delineation step. It employs the multiple regression (MR) model that describes the relation between hydrological and physio-meteorological variables of sites [Girard et al., 2004].

After Chebana and Ouarda [2008], statistical depth functions are used in a number of hydrological and environmental studies. For instance, Chebana and Ouarda [2011a] used these functions in an exploratory study of a multivariate sample including location, scale, skewness and kurtosis as well as outlier detection. In another study, Chebana and Ouarda [2011b] combined depth functions with the orientation of observations to identify the extremes in a multivariate sample. Bardossy and Singh [2008] used the statistical notion of depth to detect unusual events in order to calibrate hydrological models. Recently, some studies present further developments of the approach that calibrate hydrological models by a depth function [e.g., Krauße and Cullmann, 2012; Krauße et al., 2012].

The DBRFA method consists generally of ordering sites by using the statistical notion of depth functions [Zuo and Serfling, 2000]. This order is based on the similarity between each gauged site and the target one. Accordingly, a weight is attributed to each gauged site using a weight function denoted φ. This function, with a suitable shape, eliminates the border effect and includes all the available sites proportionally to their hydrological similarity to the target site. Note that classical RFA approaches correspond to a special weight function with value 1 inside the region and 0 outside. The definition of a region in the classical RFA approaches becomes rather a question of choice of weight function φ according to a given criterion (e.g., relative root mean square error RRMSE).

By construction, the estimation performance in the MR model using the DBRFA approach depends on the choice of the weight functions φ. Chebana and Ouarda [2008] applied several families of functions φ, where the corresponding coefficients were chosen arbitrary and after several trials. In addition, even though the obtained results are improvement of the traditional approaches, they are not necessarily the best ones.

The aim of the present paper is to propose a procedure to optimize the DBRFA approach over φ. This aim has theoretical as well as practical considerations. This procedure allows an optimal choice of the weight function φ and makes the DBRFA approach automatic and objective. It should be noted that Ouarda et al. [2001] determined the optimal homogenous neighborhood of a target site in the Canonical Correlation Analysis (CCA) based approach. In Ouarda et al [2001] the optimization corresponds to the selection of the neighborhood coefficient, denoted by α, according to the bias or the squared error. The optimal choice of weight functions has been the topic of numerous studies in the field of statistics [e.g., Chebana, 2004].

To optimize the choice of φ, suitable families of functions as well as algorithms are required. In the present context, four families of φ are considered: Gompertz (φ_G) [Gompertz, 1825], logistic ($\varphi_{\text{logistic}}$) [Verhulst, 1838], linear (φ_{Linear}) and indicator (φ_I). The three families φ_G, $\varphi_{\text{logistic}}$ and φ_{Linear} are regular, flexible, S-shaped and have other suitable properties.

Several appropriate algorithms can be considered [Wright, 1996]. They are appropriate when the objective function ξ (criterion to be optimized) is not differentiable or the gradient is unavailable and must be calculated by a numerical method (e.g., finite differences). Among these algorithms, the most commonly used are: the simplex method

[Nelder and Mead, 1965], the pattern search method of Hooke and Jeeves [Hooke and Jeeves, 1961; Torczon, 2000] and the Rosenbrock methods [Rao, 1996; Rosenbrock, 1960]. These methods are used successfully in several domains, and are particularly popular in chemistry, engineering and medicine. Specifically, in this paper the simplex and the pattern search algorithms are used because of their advantages. Indeed, they are very robust [e.g., Dolan et al., 2003; Hereford, 2001; Torczon, 2000], simple in terms of programming, valid for nonlinear optimization problems with real coefficients [McKinnon, 1999] and helpful in solving optimization problems with and without constraints [e.g., Lewis and Torczon, 1999; Lewis and Torczon, 2002].

In this study, the proposed optimization procedure is applied to the flood data from three different regions of the United States and Canada (Texas, Arkansas and southern Quebec). For each region, the obtained results are compared with those of the CCA approach.

The present paper is organized as follows. Section 2 describes the used technical tools including depth functions, the WLS method and the definitions of the considered weight functions. Section 3 describes the proposed procedure. Then section 4 presents the application to the three case studies as well as the obtained results. The last section is devoted to the conclusions of this work.

2. Background

In this section, the background elements required to introduce and apply the optimization procedure of the DBRFA approach are briefly presented. This section contains a number of basic notions.

2.1. Mahalanobis depth function

The absence of a natural order to classify multivariate data led to the introduction of the depth functions [Tukey, 1975]. They are used in many research fields, and were introduced in water science by Chebana and Ouarda [2008]. Several depth functions were introduced in the literature [Zuo and Serfling, 2000]. Depth functions have a number of features that fit well with the constraint of RFA [Chebana and Ouarda, 2008].

In this study, the Mahalanobis depth function is used to sort sites where the deeper the site is the more it is hydrologically similar to the target site. This function is used for its simplicity, value interpretability, and for the relationship with the CCA approach used in RFA. The Mahalanobis depth function is defined on the basis of the Mahalanobis distance given by $d_A^2(x,y)=(x-y)' A^{-1}(x-y)$ between two points $x,y \in R^d$ $(d \geq 1)$ where A is a positive definite matrix [Mahalanobis, 1936]. This distance is used by Ouarda et al. [2001] in the development of the CCA approach. The Mahalanobis depth of x with respect to μ is given by:

$$MHD(x;F)=\frac{1}{1+d_A^2(x,\mu)} \qquad x \text{ in } R^d \tag{1}$$

for a cumulative distribution function F characterized by a location parameter μ and a covariance matrix A. Note that the Mahalanobis depth function has values in the interval $[0,1]$.

An empirical version of the Mahalanobis depth of x with respect μ is defined by replacing F by a suitable empirical function $\hat{F}_N$ for a sample of size N [Liu and Singh, 1993]. In the context of the present paper, the notation in (1) is replaced by:

$$MHD_{\hat{A}}(x;\hat{\mu}) = \frac{1}{1 + d_{\hat{A}}^2(x, \hat{\mu})} \qquad (2)$$

where $\hat{\mu}$ and $\hat{A}$ are respectively the location and covariance matrix estimated from the observed sample.

2.2. Weight functions

Below are the definitions of the four families of weight functions φ_G, $\varphi_{\text{logistic}}$, φ_{Linear} and φ_I considered in this paper along with special cases of functions φ for comparison purposes.

2.2.1. Gompertz function

The Gompertz function is usually employed as a distribution in survival analysis. This function was originaly formulated by Gompertz [1825] for modeling human mortality. A number of authors contributed to the studies of the characterization of this distribution [e.g., Chen, 1997; Wu and Lee, 1999]. In the field of water resources, the Gompertz function was adopted by Ouarda et al. [1995] to estimate the flood damage in the residential sector. The function φ_G is increasing, flexible and continuous [Zimmerman and Núñez-Antón, 2001]. The Gompertz distribution has different formulations one of which is given by:

$$\varphi_G(x) = c \exp\left\{-ae^{-bx}\right\} \quad a,b,c > 0 \; ; \; x \in R \qquad (3)$$

where c is its upper limit, a and b are two coefficients which respectively allow to translate and change the spread of the curve. Figure 1 shows the effects of these coefficients on the form of φ_G. Note that this function starts at zero (starting phase), then increases exponentially (growth phase) and finally stabilizes by approaching the upper limit c (stationary phase) with $0 \le \varphi_G(x) \le c$. The inflection point of this function is $\left(\frac{\ln a}{b}, \frac{c}{e}\right)$.

2.2.2. Logistic function

Verhulst [1838] proposed this function to study population growth. It is given by:

$$\varphi_{\text{logistic}}(x) = \frac{c}{1 + ae^{-bx}} \qquad a, b, c > 0; x \in R \tag{4}$$

where the coefficients c, a and b play the same role as in φ_G.

This function has similar properties to those of φ_G (increasing, flexible, continuous and with three phases). However, $\varphi_{\text{logistic}}$ is symmetric around its inflection point $\left(\frac{\ln a}{b}, \frac{c}{2}\right)$ which is not the case for φ_G.

2.2.3. Linear function

It is a simple function, linear over three pieces corresponding to the three previous phases. Explicitly it is given by:

$$\varphi_{Linear}(x) = \begin{cases} 0 & \text{if } x \le d_1 \\ \dfrac{x - d_1}{d_2 - d_1} & \text{if } d_1 \le x \le d_2, \\ 1 & \text{if } x \ge d_2 \end{cases} \qquad d_2 > d_1 > 0 \tag{5}$$

This function is considered as a weight function in the study of Chebana and Ouarda [2008].

2.2.4. Indicator function

This function is given by:

$$\varphi_I(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{6}$$

where A is a subset in R (set of real numbers), such as an interval. The subset A represents the neighborhood or the region in the classical RFA approaches. The weight is equal to 1 if the site is included in the region, otherwise, it is 0.

In the case where the set A is the interval $\left[C_{\alpha,p},1\right]$ with $C_{\alpha,p} = \dfrac{1}{1+\chi^2_{\alpha,p}}$ and $\chi^2_{\alpha,p}$ is the $(1-\alpha)$ quantile associated to the chi-squared distribution with p degrees of freedom, the DBRFA reduces to the traditional CCA approach [e.g., Bates et al., 1998]. The corresponding weight function is denoted by φ_{CCA}.

If $A=[0,1]$ i.e. $\alpha=0$, then the DBRFA represents the uniform approach which includes all available sites with similar importance. The corresponding weight function is denoted by φ_U.

2.3. Weighted Least Squares Estimation

In the RFA framework, the MR model is generally used to describe the relationship between the hydrological variables and the physiographical and climatic variables of the sites of a given region. This model has the advantage to be simple, fast, and not requiring the same distribution for hydrological data at each site within the region [Ouarda et al., 2001].

Let QT be the quantile corresponding to the return period T. It is often assumed that the relationship between QT, as the hydrological variable, and the physio-meteorological variables and basin characteristics $A_1, A_2, \ldots. A_r$ takes the form of a power function [Girard et al., 2004]:

$$QT = \beta_0 A_1^{\beta_1} A_2^{\beta_2} \ldots A_r^{\beta_r} e \qquad (7)$$

where e is the model error.

Let s be the number of quantiles QT corresponding to s return periods and N be the total number of sites in the region. A matrix of hydrological variables $Y = \left(QT_1, QT_2, ..., QT_s\right)$ of dimension $N \times s$ is then constructed. With a log-transformation in (7) we obtain the multivariate log-linear model in the following form:

$$\log Y = \left(\log X\right)\beta + \varepsilon \qquad (8)$$

where $\log X = (1, \log A_1, \log A_2, ..., \log A_r)$ is the $N \times \left(r+1\right)$ matrix formed by (r) physio-meteorological variables series, β is the $(r+1)\times s$ matrix of parameters and $\varepsilon = \left(\varepsilon^1, ..., \varepsilon^s\right)$ is the $N \times s$ matrix that represents the model error (residual) with null mean vectors and variance-covariance matrix Γ:

$$E(\varepsilon) = \left(0,..,0\right) \quad \text{and} \quad Var(\varepsilon) = \Gamma = \begin{pmatrix} Var\left(\varepsilon^1\right) & \ldots & Cov\left(\varepsilon^1, \varepsilon^s\right) \\ \vdots & \ddots & \vdots \\ Cov\left(\varepsilon^s, \varepsilon^1\right) & \cdots & Var\left(\varepsilon^s\right) \end{pmatrix} \qquad (9)$$

The parameter matrix β can be estimated, using the WLS estimation, by:

$$\begin{aligned} \hat{\beta}_w &= \arg\min_{\beta} \left(logY - logX\,\beta\right)' \Omega \left(logY - logX\,\beta\right) \\ &= \left((\log X)'\Omega \log X\right)^{-1} (\log X)'\Omega \log Y \end{aligned} \qquad (10)$$

where $\Omega = \text{diag}(w_1, ..., w_N)$ is the diagonal matrix with diagonal elements w_i where w_i is the weight for the site i. The matrix Γ is estimated by:

$$\hat{\Gamma}_w = \frac{\left(\log Y - \log X\,\hat{\beta}_w\right)' \left(\log Y - \log X\,\hat{\beta}_w\right)}{N - r - 1} \qquad (11)$$

Note that the log-transformation induces generally a bias in the estimation of QT [Girard et al., 2004].

3. Methodology

This section describes a general procedure for optimizing the DBRFA approach and treats special cases where this procedure is applied using the weight functions defined in section 2.2.

3.1. General procedure

In order to find the optimal weight function $\varphi_{Optimal}$ in the DBRFA approach, the procedure is composed of three main steps. They are summarized as follows:

i. For a given class of weight functions φ and a set of gauged sites (region), use a jackknife procedure to assess the regional flood quantile estimators (Eq. 8) for the sites of the region using the DBRFA approach. These estimators depend on the weight function φ through its coefficients;

ii. For a pre-selected criterion, calculate its value to quantify the performance of the estimates obtained from step i;

iii. Using an optimization algorithm, optimize the criterion (objective function) calculated in step ii. The parameters of the optimization problem are the coefficients of the weight function. The outputs of this step are $\varphi_{Optimal}$ and the value of the selected criterion.

3.2 Description of the procedure

In the first step of the procedure, we use a jackknife resampling procedure to assess the regional flood quantile estimators for the sites of the region. This jackknife procedure consists in considering each site l ($l = 1,...,N$) in the region as an ungauged one by removing it temporarily from the region (i.e. we assume that the hydrological variable Y_l of site l is unknown and the physio-meteorological variable X_l is known since it can be easily estimated from existing physiographic maps and climatic data). Then we calculate the regional estimator $\left(\hat{Y}_l\right)_\varphi$ of site l by the iterative WLS regression, using the $N-1$ remaining sites, which is related to the given weight function φ. The parameters of the starting estimator (initial point) of DBRFA, denoted by $\hat{\beta}_{1,l}$ and $\hat{\Gamma}_{1,l}$, are calculated by assuming that $X = X^{<-l>}$, $Y = Y^{<-l>}$ and $\Omega = I_{N-1}$ in (10) and (11), where $X^{<-l>}$ represents the matrix of physio-meteorological variables excluding site l, $Y^{<-l>}$ is the matrix of hydrological variables excluding site l and I_{N-1} is the identity matrix of dimension $(N-1)\times(N-1)$. The starting estimator $\left(\hat{Y}_{1,l}\right)_\varphi$ is obtained by replacing β with $\hat{\beta}_{1,l}$ in (8). Then for each depth iteration k, $k = 2,3,..,k_{iter}$, we calculate the Mahalanobis depth (2) of the gauged site i, $i = 1,...,N-1$, with respect to the ungauged site l denoted by $\left(D_{k,(i,l)}\right)_\varphi = MHD_{\left(\hat{\Gamma}_{k-1,l}\right)_\varphi}\left(\log Y_i;\left(\log \hat{Y}_{k-1,l}\right)_\varphi\right)$. The number of iterations k_{iter} is fixed to ensure the convergence of the depth function (generally k_{iter} = 25 is appropriate). The weight matrix at iteration k is defined by applying the function φ to the depth calculated at this iteration. The parameters of the MR model at the k^{th} iteration are estimated by:

$$\left(\hat{\beta}_{k,l}\right)_\varphi = \left(\left(\log X^{<-l>}\right)'\left(\Omega_{k,l}\right)_\varphi\left(\log X^{<-l>}\right)\right)^{-1}\left(\log X^{<-l>}\right)'\left(\Omega_{k,l}\right)_\varphi \log Y^{<-l>} \quad (12)$$

$$\left(\hat{\Gamma}_{k,l}\right)_\varphi = \frac{\left(\log Y^{<-l>} - \left(\log X^{<-l>}\right)\left(\hat{\beta}_{k,l}\right)_\varphi\right)'\left(\log Y^{<-l>} - \left(\log X^{<-l>}\right)\left(\hat{\beta}_{k,l}\right)_\varphi\right)}{(N-1)-r-1} \quad (13)$$

where $\left(\Omega_{k,l}\right)_{\varphi}$ is a N-1 diagonal matrix with elements:

$$\varphi\left[\left(D_{k,(1,l)}\right)_{\varphi}\right],.......,\varphi\left[\left(D_{k,(N-1,l)}\right)_{\varphi}\right] \tag{14}$$

Note that all these parameters depend on φ. Then, the regional quantile estimator for the site l in this iteration is:

$$\left(\hat{Y}_{k,l}\right)_{\varphi} = \exp\left[\left(\log X_{l}\right)\left(\hat{\beta}_{k,l}\right)_{\varphi}\right] \tag{15}$$

In the second step of the procedure, we use the regional estimators at the last iteration since their associated estimation errors are the minimum possible by construction. Consequently, in order to simplify the notations in the rest of this paper, we denote $\left(\hat{Y}_1\right)_{\varphi} = \left(\hat{Y}_{k_{iter},1}\right)_{\varphi},...,\left(\hat{Y}_l\right)_{\varphi} = \left(\hat{Y}_{k_{iter},l}\right)_{\varphi},..,\left(\hat{Y}_N\right)_{\varphi} = \left(\hat{Y}_{k_{iter},N}\right)_{\varphi}$.

After calculating $\left(\hat{Y}_l\right)_{\varphi}$, $l=1,..,N$ in step i, we consider and evaluate one or several performance criteria in step ii. The considered criteria are employed as objective functions in the optimization step iii.

The relative bias (RB) and the relative root mean square error (RRMSE) are widely used in hydrology, particularly in RFA, as criteria to evaluate model performances. These two criteria are defined using an element-by-element division by:

$$RB_{\varphi} = 100\times\frac{1}{N}\sum_{l=1}^{N}\left(\frac{Y_l-\left(\hat{Y}_l\right)_{\varphi}}{Y_l}\right) \tag{16}$$

$$RRMSE_{\varphi} = 100\times\sqrt{\frac{1}{N-1}\sum_{l=1}^{N}\left(\frac{Y_l-\left(\hat{Y}_l\right)_{\varphi}}{Y_l}\right)^2} \tag{17}$$

where Y_l is the local quantile estimation for the l^{th} site, $\left(\hat{Y}_l\right)_\varphi$ is the regional estimation by DBRFA approach according to φ and excluding site l, and N is the number of sites in the region. The RB_φ measures the tendency of quantile estimates to be uniformly too high or too low across the whole region and the RRMSE_φ measures the overall deviation of estimated quantiles from true quantiles [Hosking and Wallis, 1997]. Note that other criteria can also be considered such as the Nash criterion (NASH) and the coefficient of determination (R^2). In the hydrological framework, the previously defined criteria are used as key performance indicators (KPI) to compare different RFA approaches [e.g., Gaál et al., 2008].

Finally in step iii, we apply an optimization algorithm on the selected and evaluated criterion in step ii. The algorithms to be considered are indicated in the introduction section. The formulation of the criteria to be optimized, generally complex and non-explicit, suggests the use of zero-order algorithms. The application of these algorithms allows to find the optimal function $\varphi_{Optimal}$ with respect to selected criteria. An overview diagram summarizing the optimization procedure of the DBRFA approach is illustrated in Figure 2.

The procedure described above aims to calculate $\varphi_{Optimal}$ according to the desired criterion. In order to estimate the quantile Y_u of an ungauged site u using the optimal DBRFA approach, the user simply repeats step i of the procedure without excluding any site and while fixing the weight function, i.e. step i with $\varphi = \varphi_{Optimal}$.

Based on the optimization procedure of the DBRFA approach described previously, the parameters of the optimization problem are the coefficients of the weight function.

Consequently, reducing the number of coefficients in φ can make the algorithm more efficient and less expensive in terms of memory and computing time. If the weight function is one of the two functions Gompertz (3) or logistic (4), the coefficient c represents the upper limit of these functions. As in the DBRFA approach, the upper limit of φ is 1, namely the gauged site is completely similar to the target site, hence the value $c=1$ is fixed. In this case, the problem is reduced to find the couple $\left(\hat{a}_N, \hat{b}_N\right)$ that optimizes one of the pre-selected criteria, such as (16) and (17).

Moreover, in the classes $\varphi = \varphi_G$ or $\varphi = \varphi_{\text{logistic}}$, the optimization problem is applied in semi-bounded domain (i.e. $a > 0$ and $b > 0$) and without other constraints (linear or nonlinear). In this case, the Nelder-Mead algorithm can also be applied as well as the Pattern search one [Luersen and Le Riche, 2004].

On the other hand, in the case where $\varphi = \varphi_{Linear}$ (5), the inequality constraint $d_2 > d_1 > 0$ is imposed. Therefore, the Nelder-Mead algorithm can not be considered.

Theoretically and generally, the two optimization algorithms used in this paper (i.e. the Nelder-Mead and the pattern search algorithms) converge to a local minimum (or maximum) according to the initial point. To overcome this problem and make the algorithm more efficient, two solutions are proposed in the literature: a) for each objective function, use several starting points and calculate the optimum for each of these points; the optimum of the function will be the best value of these local optima [Bortolot and Wynne, 2005]; or b) use a single starting point and each time the algorithm converges, the optimization algorithm restarts again using the local optimum as a new starting point. This procedure is repeated until no improvement in the optimal value of the objective function is obtained [Press et al., 2002].

4. Data sets for case studies

In this section we present the data sets on which the DBRFA approach will be applied the following section. These data come from three geographical regions located in the states of Arkansas and Texas (USA) and in the southern part of the province of Quebec (Canada). The first region is located between 45 ° N and 55 ° N in the southern part of Quebec, Canada. The data-set of this region is composed of 151 stations, each with station has a flood record of more than 15 years. The conditions of application of frequency analysis (i.e. homogeneity, stationary and independence) are tested on the historical data of these stations in several studies [Chokmani and Ouarda, 2004; Ouarda and Shu, 2009; Shu and Ouarda, 2008]. Three types of variables are considered: physiographical, meteorological and hydrological. The selected variables for the regional modeling are also used in Chokmani and Ouarda [2004]. The selected physiographical variable are: the basin area (AREA) in km^2, the mean basin slope (MBS) in % and the fraction of the basin area covered with lakes (FAL) in %. The meteorological variables are the annual mean total precipitation (AMP) in mm and the annual mean degree days over 0°C (AMD) in degree-day. The selected hydrological variables are represented by at-site specific flood quantiles (QST) in m^3/km^2s, corresponding to return periods $T = 10$ and 100 years.

The two other considered regions correspond to a database of the United States Geological Survey (USGS). This database, called Hydro-Climatic Data Network (HCDN), consists of observations of daily discharges from 1659 sites across the United States and its Territories [Slack et al., 1993]. The sites included in this database contain at least 20 years of observations. As part of the HCDN project, the United States are divided into 21 large hydrological regions.

In this study, the data of the states of Arkansas and Texas (USA) are used for comparison purposes. The applicability conditions of frequency analysis as well as the variables to consider are justified in the study of Jennings et al., [1994]. The physiographical and climatological characteristics are the area of drainage basin (AREA) in km^2, the slope of main channel (SC) in m/km, the annual mean precipitation (AMP) in cm, the mean elevation of drainage basin (MED) in m and the length of main channel (LC) in km. The selected hydrological variables in these two regions are the at-site flood quantiles (*QT*), in m^3/s, corresponding to the return periods $T = 10$ and 50 years.

The data-set of the states of Arkansas is composed of 204 sites. These data and the at-site frequency analysis are published in the study of Hodge and Tasker [1995]. Tasker et al. [1996] used these data to estimate the flood quantiles corresponding to the 50 year return period by the region of influence method [Burn, 1990b].

The Texas data base is composed of 90 sites but due to the lack of some explanatory variables at several sites, modeling was performed with only 69 stations. The data-set used in this region is the same used by Tasker and Slade [1994].

5. Results

The results obtained from the CCA-based approach are first presented and then compared to those obtained by the optimized DBRFA approach.

The variations of the two performance criteria RB and RRMSE, obtained by the CCA approach, as a function of the coefficient α (neighborhood coefficient) for the three regions are presented in Figure 3. The complete variation range of α is the interval [0, 1].

However, in this application, the range is [0, 0.30] for Quebec and Arkansas regions and [0, 0.17] for the Texas region. These upper bounds of α are fixed to ensure that all neighborhoods of the sites contain sufficient stations to allow the estimation by the MR model. Note that it is appropriate to have at least three times more stations than the number of parameters in the MR model [Haché et al., 2002]. Figure 3 indicates that, for a given region, the same value of α optimizes the two criteria for the various return periods, even though this is not a general result [Ouarda et al., 2001]. The optimal α values are 0.25, 0.01 and 0.05 respectively for Quebec, Arkansas and Texas.

The coefficients λ_1 and λ_2 correspond respectively to the correlations of the first and the second couples of the canonical variables. Their values for Arkansas ($\lambda_1 = 0.973$, $\lambda_2 = 0.470$) and Texas ($\lambda_1 = 0.923$, $\lambda_2 = 0.402$) are larger than those of Quebec ($\lambda_1 = 0.853$, $\lambda_2 = 0.281$). This corresponds to a large optimal value of α for the latter region. Indeed, the higher the canonical correlation, the smaller the size of the ellipse defining the homogeneous neighborhood [Ouarda et al., 2001]. The value of α should be small enough so that the neighborhood contains an appropriate number of stations to perform the estimation in the MR model, and large enough to ensure an adequate degree of homogeneity within the neighborhood.

Figure 4 shows the projection sites of the three regions in the two canonical spaces (V1, W1) and (V2, W2) corresponding respectively to λ_1 and λ_2. This figure shows that for these three regions, the relationship between V1 and W1 is approximately linear, in contrast to V2 and W2. The presentation of a site in the space (V1, W1) is useful for an a priori information on the estimation error of this site. For example, in the Quebec region, the two sites 66 and 122 are poorly estimated. By fitting a linear model between V1 and

W1 for each region, it is seen that the linearity assumption is more respected in Arkansas and Texas than in Quebec (R^2_{Arkansas} = 0.94, R^2_{Texas} = 0.85 et R^2_{Quebec} = 0.73).

The previous results show that the values of λ_1, λ_2, α and R^2 can be used as indicators of the quality of the homogeneity in a given region. In this application, the lower values of λ_1, λ_2 and R^2 as well as the higher value of α for Quebec compared to the values of the other two regions indicate that the Quebec region is less homogeneous than the two others. This conclusion needs to be verified by other criteria or statistical tests.

The DBRFA approach is applied by using the Mahalanobis depth function (2). The optimal weight functions, from each one of the three considered families, are obtained on the basis of the indicated optimization algorithms (i.e. φ_G and $\varphi_{\text{logistic}}$ using Nelder-Mead and φ_{Linear} using pattern search). They are presented in Figure 5. The corresponding results are summarized in Table 1. The optimization is made with respect to the RB and RRMSE criteria. Note that, for a given region, the regional flood quantile estimation is more accurate for small return periods. This result is valid for local as well as regional frequency analysis approaches [Hosking and Wallis, 1997]. In addition, Table 1 shows that the worst estimates are obtained using the uniform approach (weight function φ_U). This justifies the usefulness of considering the regional approaches. Note that for all regions, DBRFA with $\varphi_{Optimal}$ leads to more accurate estimates in terms of RB and RRMSE than those obtained using the CCA approach with optimal α. These results show also that the optimal coefficients of a given weight function depend on the chosen criterion (objective function). Finally, for the southern Quebec region, the results of Chebana and Ouarda (2008) are very close to those in the present paper (Table 1). The reason for this closeness is that the above authors forced the DBRFA approach to provide good results by trying several different combinations of values of φ coefficients (i.e. iteration loop of coefficients). Consequently, their trials took a long time and did not ensure the optimality of the approach which is not the case for the present study.

According to Figure 5, the form of optimal weight function depends on the considered region. For instance, the steep S-curve (with long upper extremity) of the two regions Arkansas and Texas depicts a large number of gauged sites similar to the target one; however, the high S-curve (with short upper extremity) of Quebec shows a small number of gauged sites similar to the target one. This result supports the previously mentioned conclusion about the homogeneity level for these regions.

In order to visualize the influence of gauged sites on the regional estimation of a target site in the DBRFA and CCA approaches, assume that Texas site number 25 is a target site and has to be estimated using the remaining 68 gauged sites. Figure 6 illustrates the weights allocated to each gauged site in the canonical hydrological space (W1, W2) instead of the geographical space. The estimate is made with the optimal α for the CCA approach and the optimal φ_G for the DBRFA approach. We observe that the influence of a gauged site on the estimation of the target site in the DBRFA approach is proportional to the hydrological similarity between these two sites. Hence, the weight function takes a bell shape in a 3D presentation (Figure 6b). However, with the CCA approach, the weight function (6) takes only two values, 1 within the neighborhood of the target-site or 0 otherwise (Figure 6a).

To study the impact of depth iterations on the performance of the DBFRA method, this approach is applied to the three regions but without iterations on the Mahalanobis depth (i.e. $k_{\text{iter}} = 2$ in step i in the DBRFA optimization procedure). The outputs of this application, with $\varphi = \varphi_G$ and $\zeta(.) = RRMSE$, are shown in Table 2. These results indicate that the optimal weight function changes depending on the case (with or without iterations) but keeps the S shape (for space limitation, the associated figure is not presented). In addition, using the iterations, we observe an improvement in the performance of the DBRFA method. This improvement varies from one region to another where it is more significant in Quebec than in Texas and Arkansas (Table 2). This is another result indicating a difference between Quebec and the two other regions. Note that similar results are found for other families of weight functions and for different optimization criteria. In conclusion, the depth iterative step in the DBRFA before weight optimization is important.

In order to examine the convergence speed in terms of the performance criteria, we present the variations of these criteria as a function of depth iteration for different weight functions (Figure 7). The employed coefficient values of the weight functions are those minimizing the RRMSE (Table 1). We observe a rapid convergence (5 iterations) to the RRMSE values in Table 1 for Arkansas and Texas (Figure 7b and 7c), whereas, for Quebec (Figure 7a) it requires more than 20 iterations to converge to the results in Table 1. These results could be again due to the level of homogeneity in the region.

To compare the relative errors of flood quantile estimates obtained by different approaches for the three regions, Figure 8 illustrates these errors with respect to the logarithm of basin area. The weight functions used are those optimizing the RRMSE. It is generally observed that the DBRFA relative errors are lower than those obtained with the

CCA approach. We also observe large negative errors for some sites, such as number 64 and 66 in the southern Quebec, 180 and 175 in Arkansas and 62 and 69 in Texas.

In this paper, the optimal DBRFA approach is mainly compared with the basic formulation of one of the most popular RFA approaches, that is the CCA approach. However, different variants of the latter are developed and are available in the literature, such as the Ensemble Artificial Neural Networks-CCA approach (EANN-CCA) [Shu and Ouarda, 2007] and the Kriging-CCA approach [Chokmani and Ouarda, 2004]. In order to insure the optimality of the optimal DBRFA, it is of interest to expend the above comparison to those approaches. A comprehensive comparison requires presentation of these approaches as well a number of data sets for the considered regions. Some of the data sets are not available for the regions of Texas and Arkansas, e.g. at-site peak flows to estimate at-site quantiles as hydrological variables. However, all these approaches are already applied to the region of Quebec in different studies. Table 3 summarizes the obtained results for all those methods along with those of the DBRFA approach. The results indicate that the optimal DBRFA performs better than the available approaches both in terms of RB and RRMSE, except a very slight difference of 1% in the RRMSE of QS10 with EANN-CCA. This could be related to the numerical approximations in the computational algorithms.

6. Conclusions

In the present paper, a procedure is proposed to optimize the selection of a weight function in the DBRFA approach. This procedure automates the optimal choice of the weight function φ with respect to a given criterion. Therefore, aside from leading to optimal estimation results, it allows the DBRFA approach to be more practical and usable without the user's subjective intervention. The user has only to select one or several objective performance criteria to obtain the model, the estimated performance and the weight functions for a specific region. One of the findings is that the optimal weight function can be seen as characterization of the associated region.

General and flexible families of weight function are considered, as well as two optimization algorithms to find $\varphi_{Optimal}$. The used algorithms can handle cases with or without constraints on the definition domain of the function φ.

The obtained results, from three regions in North America, show the utility to consider the DBRFA method in terms of performance as well as the efficiency and flexibility of the proposed optimization procedure.

The study of the three regions shows an association between the level of the homogeneity of the region, the form of the optimal weight function and the computation convergence speed. This result deserves to be developed in future work.

Acknowledgments

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canada Research Chair Program. The authors are grateful to the Editor and the anonymous reviewers for their valuable comments and suggestions.

References

Bárdossy, A., Singh, S.K., 2008. Robust estimation of hydrological model parameters. Hydrology and Earth System Sciences, 12(6): 1273-1283.

Bates, B.C., Rahman, A., Mein, R.G., Weinmann, P.E., 1998. Climatic and physical factors that influence the homogeneity of regional floods in southeastern Australia. Water Resources Research, 34(12): 3369-3381.

Bortolot, Z.J., Wynne, R.H., 2005. Estimating forest biomass using small footprint LiDAR data: An individual tree-based approach that incorporates training data. ISPRS Journal of Photogrammetry and Remote Sensing, 59(6): 342-360.

Burn, 1990a. An appraisal of the "region of influence' approach to flood frequency analysis. Hydrological Sciences Journal/Journal des Sciences Hydrologiques, 35(2): 149-165.

Burn, 1990b. Evaluation of regional flood frequency analysis with a region of influence approach. Water Resources Research, 26(10): 2257-2265.

Chebana, F., 2004. On the optimization of the weighted Bickel-Rosenblatt test. Statistics and Probability Letters, 68(4): 333-345.

Chebana, F., Ouarda, T.B.M.J., 2008. Depth and homogeneity in regional flood frequency analysis. Water Resources Research, 44(11).

Chebana, F., Ouarda, T.B.M.J., 2011a. Depth-based multivariate descriptive statistics with hydrological applications. Journal of Geophysical Research D: Atmospheres, 116(10).

Chebana, F., Ouarda, T.B.M.J., 2011b. Multivariate extreme value identification using depth functions. Environmetrics, 22(3): 441-455.

Chen, Z., 1997. Parameter estimation of the Gompertz population. Biometrical Journal, 39(1): 117-124.

Chokmani, K., Ouarda, T.B.M.J., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resources Research, 40(12): 1-13.

Dalrymple, T., 1960. Flood frequency methods. Water Supply Paper No. 1543 A.

De Michele, C., Rosso, R., 2002. A multi-level approach to flood frequency regionalisation. Hydrology and Earth System Sciences, 6(2): 185-194.

Dolan, E.D., Michael Lewis, R., Torczon, V., 2003. On The Local Convergence Of Pattern Search. SIAM J. OPTIM, 14(2): 567-583.

Gaál, L., Kyselý, J., Szolgay, J., 2008. Region-of-influence approach to a frequency analysis of heavy precipitation in Slovakia. Hydrology and Earth System Sciences, 12(3): 825-839.

Girard, C., Ouarda, T.B.M.J., Bobée, B., 2004. Study of bias in the log-linear model for regional estimation. Étude du biais dans le modèle log-linéaire d'estimation régionale, 31(2): 361-368.

Gompertz, B., 1825. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philos. Trans. R. Soc. Lond., 115: 513-585.

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

Haché, M., Ouarda, T.B.M.J., Bruneau, P., Bobée, B., 2002. Regional estimation by canonical correlation analysis: Hydrological variable analysis. Estimation régionale par la méthode de l'analyse canonique des corrélations : Comparaison des types de variables hydrologiques, 29(6): 899-910.

Haddad, K., Rahman, A., 2012. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework - Quantile Regression vs. Parameter Regression Technique. Journal of Hydrology, 430-431: 142-161.

Hereford, J., 2001. Comparison of four parameter selection techniques. Proceedings of SoutheastCon 2001, Clemson, SC, 30 March-1 April 2001: 11-16.

Hodge, S.A., Tasker, G.D., 1995. Magnitude and Frequency of Floods in Arkansas. U.S. Geological Survey Water-Resources Investigations Report.

Hooke, R., Jeeves, T.A., 1961. Direct search solution of numerical and statistical problems. Journal of the Association for Computing Machinery, 8(2): 212-229.

Hosking, J.R.M., Wallis, J.R., 1997. Regional frequency analysis: an approach based on L-moments. Cambridge University Press, Cambridge.

Jennings, M.E., Thomas W.O, Jr., Riggs, H.C., 1994. Nationwide summary of U.S. geological survey regional regression equations for estimating magnitude and frequency of floods for ungaged sites, 1993. USGS Water-Resources Investigations Rep. 94-4002.

Krauße, T., Cullmann, J., 2012. Towards a more representative parametrisation of hydrologic models via synthesizing the strengths of Particle Swarm Optimisation and Robust Parameter Estimation. Hydrology and Earth System Sciences, 16(2): 603-629.

Krauße, T., Cullmann, J., Saile, P., Schmitz, G.H., 2012. Robust multi-objective calibration strategies – possibilities for improving flood forecasting. Hydrology and Earth System Sciences, 16(10): 3579-3606.

Lewis, R.M., Torczon, V., 1999. Pattern search algorithms for bound constrained minimization. SIAM Journal on Optimization, 9(4): 1082-1099.

Lewis, R.M., Torczon, V., 2002. A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds. SIAM Journal on Optimization, 12(4): 1075-1089.

Liu, R.Y., Singh, K., 1993. A quality index based on data depth and multivariate rank tests. J. Amer. Statist. Assoc., 88(421): 252-260.

Luersen, M.A., Le Riche, R., 2004. Globalized nelder-mead method for engineering optimization. Computers and Structures, 82(23-26): 2251-2260.

Madsen, H., Rosbjerg, D., 1997. Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling. Water Resources Research, 33(4): 771-781.

Mahalanobis, P.C., 1936. On the generalized distance in statistics. Calcutta Statist. Assoc. Bull., 14: 9.

McKinnon, K.I.M., 1999. Convergence of the Nelder-Mead simplex method to a nonstationary point. SIAM Journal on Optimization, 9(1): 148-158.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J., 7: 308-313.

Nguyen, V.T.V., Pandey, G., 1996. A new approach to regional estimation of floods in Quebec. Proceedings of the 49th Annual Conference of the CWRA: 587-596.

Ouarda et al., 2008. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. Journal of Hydrology, 348(1-2): 40-58.

Ouarda, El-Jabi, N., Ashkar, F., 1995. Flood damage estimation in the residential sector. Water Resources and Environmental Hazards: Emphasis on Hydrologic and Cultural insight in the Pacific Rim, AWRA Technical Publication series (1995): 73-82.

Ouarda, Shu, C., 2009. Regional low-flow frequency analysis using single and ensemble artificial neural networks. Water Resources Research, 45(11).

Ouarda, T.B.M.J., Girard, C., Cavadias, G.S., Bobée, B., 2001. Regional flood frequency estimation with canonical correlation analysis. Journal of Hydrology, 254(1-4): 157-173.

Ouarda, T.B.M.J., Hache, M., Bruneau, P., Bobee, B., 2000. Regional flood peak and volume estimation in northern Canadian basin. Journal of Cold Regions Engineering, 14(4): 176-191.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 2002. Numerical recipes in C: the art of scientific computing. 2nd ed.

Rao, S.S., 1996. Engineering Optimization-Theory and Practice, 3rd Ed., 9: 621-622.

Rosenbrock, H.H., 1960. An automatic method for finding the greatest or least value of a function. Comput. J., 3(3): 175-184.

Shu, C., Ouarda, T.B.M.J., 2007. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resources Research, 43(7).

Shu, C., Ouarda, T.B.M.J., 2008. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. Journal of Hydrology, 349(1-2): 31-43.

Slack, J.R., Lumb, A.M., Landwehr, J.M., 1993. Hydro-Climatic Data Network (HCDN): Streamflow data set, 1874-1988. Hydro-climatic Data Network (HCDN): A U.S. Geological Survey Streamflow Data Set for the United States for the Study of Climate Variations, 1874-1988.

Tasker, G.D., Hodge, S.A., Barks, C.S., 1996. Region of influence regression for estimating the 50-year flood at ungaged sites. Journal of the American Water Resources Association, 32(1): 163-170.

Tasker, G.D., Slade, R.M., 1994. An interactive regional regression approach to estimating flood quantiles. Water Policy and Management: Solving the Problems, ASCE Proceedings of the 21st Annual Conference of the Water Resources Planning and Management Division: 782-785.

Torczon, V., 2000. On the Convergence of Pattern Search Algorithms. SIAM Journal on Optimization, 7(1): 1-25.

Tukey, J.W., 1975. Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians, 2: 523-531.

Verhulst, P.F., 1838. Notice sur la loi que la population pursuit dans son accroissement.

Wright, M.H., 1996. Direct search methods: Once scorned, now respectable. Dundee Biennial Conf. Numer.

Wu, J.W., Lee, W.C., 1999. Characterization of the mixtures of Gompertz distributions by conditional expectation of order statistics. Biometrical Journal, 41(3): 371-381.

Zimmerman, D.L., Núñez-Antón, V., 2001. Parametric modelling of growth curve data: An overview. Test, 10(1): 1-73.

Zuo, Y., Serfling, R., 2000. General notions of statistical depth function. Annals of Statistics, 28(2): 461-482.

Table 1. Quantile estimation result with the various approaches

		Region														
		Southern Quebec (Canada)					Arkansas (United States)					Texas (United States)				
			QS10		QS100			Q10		Q50			Q10		Q50	
Objective function ζ	Weight function φ	Optimal coefficients	RB (%)	RR MSE (%)	RB (%)	RR MSE (%)	Optimal coefficients	RB (%)	RR MSE (%)	RB (%)	RR MSE (%)	Optimal coefficients	RB (%)	RR MSE (%)	RB (%)	RR MSE (%)
-	φ_U	-	-8.60	55.00	-11.0	64.00	-	-13.2	65.48	-15.1	73.34	-	-9.70	46.50	-13.8	61.00
RRMSE or RB	φ_{CCA}	$\alpha = 0.25$	-7.54	44.62	-8.14	51.84	$\alpha = 0.01$	-7.80	48.16	-9.31	59.50	$\alpha = 0.05$	-1.20	42.30	-7.40	57.40
RRMSE	φ_G	$a = 30.5$ $b = 7$	-3.55	**38.70**	-2.20	**44.50**	$a = 97$ $b = 25$	-6.00	41.50	-6.33	47.70	$a = 129.7$ $b = 35.4$	-1.01	36.86	-6.00	50.79
	$\varphi_{\text{logistic}}$	$a = 2537.5$ $b = 14.8$	-3.85	39.20	-2.80	44.90	$a = 11863$ $b = 54.149$	-6.18	41.53	-6.52	47.65	$a = 3618$ $b = 50.1$	-0.90	**36.84**	-5.00	**49.50**
	φ_{Linear}	$C1 = 0.30$ $C2 = 0.80$	-3.60	38.94	-2.25	44.65	$C1 = 0.157$ $C2 = 0.162$	-5.90	**40.90**	-6.37	**47.11**	$C1 = 0.116$ $C2 = 0.152$	-2.81	38.20	-6.37	49.51
RB	φ_G	$a = 55$ $b = 9$	-3.50	39.10	-2.30	44.90	$a = 23.950$ $b = 13.661$	**-5.80**	41.52	**-6.29**	47.70	$a = 2134$ $b = 43$	-0.80	37.90	-6.20	52.17
	$\varphi_{\text{logistic}}$	$a = 2791$ $b = 15$	-3.70	39.30	-2.70	45.00	$a = 19593.7$ $b = 58.417$	-6.10	41.67	-6.49	47.70	$a = 3618.2$ $b = 50.3$	**-0.80**	37.70	**-4.90**	50.90
	φ_{Linear}	$C1 = 0.296$ $C2 = 0.768$	**-3.20**	38.90	**-1.90**	44.70	$C1 = 0.093$ $C2 = 0.267$	-5.87	41.67	-6.35	47.74	$C1 = 0.100$ $C2 = 0.112$	-0.90	39.20	-5.50	50.95

Best results for each region are in bold character.

Table 2. Results of the DBRFA Approach With and Without Depth Iterations using $\zeta(.) = RRMSE$ and $\varphi = \varphi_G$

	Region														
	Southern Quebec (Canada)					Arkansas (United States)					Texas (United States)				
		QS10		QS100			Q10		Q50			Q10		Q50	
	Optimal coefficients	RB (%)	RRMSE (%)	RB (%)	RRMSE (%)	Optimal coefficients	RB (%)	RRMSE (%)	RB (%)	RRMSE (%)	Optimal coefficients	RB (%)	RRMSE (%)	RB (%)	RRMSE (%)
With iteration	*a* = 30.5 *b* = 7	-3.55	38.70	-2.20	44.50	*a* = 97 *b* = 25	-6.00	41.50	-6.33	47.70	*a* = 129.7 *b* = 35.4	-1.01	36.86	-6.00	50.79
Without iteration	*a* = 66.50 *b* = 14.25	-6.60	47.05	-7.52	55.07	*a* = 721 *b* = 81	-7.24	42.87	-8.64	50.34	*a* = 186.7 *b* = 42.65	-1.60	38.29	-6.29	51.00

Table 3. Quantile estimation result for Quebec with available approaches and their references

Approach	Reference	QS10		QS100	
		RB (%)	RRMSE (%)	RB (%)	RRMSE (%)
Linear regression (LR)	Table 1 above	-9	55	-11	64
Nonlinear regression (NLR)	Shu and Ouarda [2008]	-9	61	-12	70
NLR with regionalisation approach	Shu and Ouarda [2008]	-19	67	-24	79
CCA	Table 1 above	-7	44	-8	52
Kriging-CCA space	Chokmani and Ouarda [2004]	-20	66	-27	86
Kriging-Principal Component Analysis space	Chokmani and Ouarda [2004]	-16	51	-23	70
Adaptive Neuro-Fuzzy Inference Systems (ANFIS)	Shu and Ouarda [2008]	-8	57	-14	64
Artificial Neural Networks (ANN)	Shu and Ouarda [2008]	-8	53	-10	60
Single ANN-CCA (SANN-CCA)	Shu and Ouarda [2007]	-5	38	-4	46
Ensemble ANN (EANN)	Shu and Ouarda [2007]	-7	44	-10	60
Ensemble ANN-CCA (EANN-CCA)	Shu and Ouarda [2007]	-5	**37**	-6	45
Optimal DBRFA	Table 1 above	**-3**	38	**-2**	**44**

Best results are in bold character

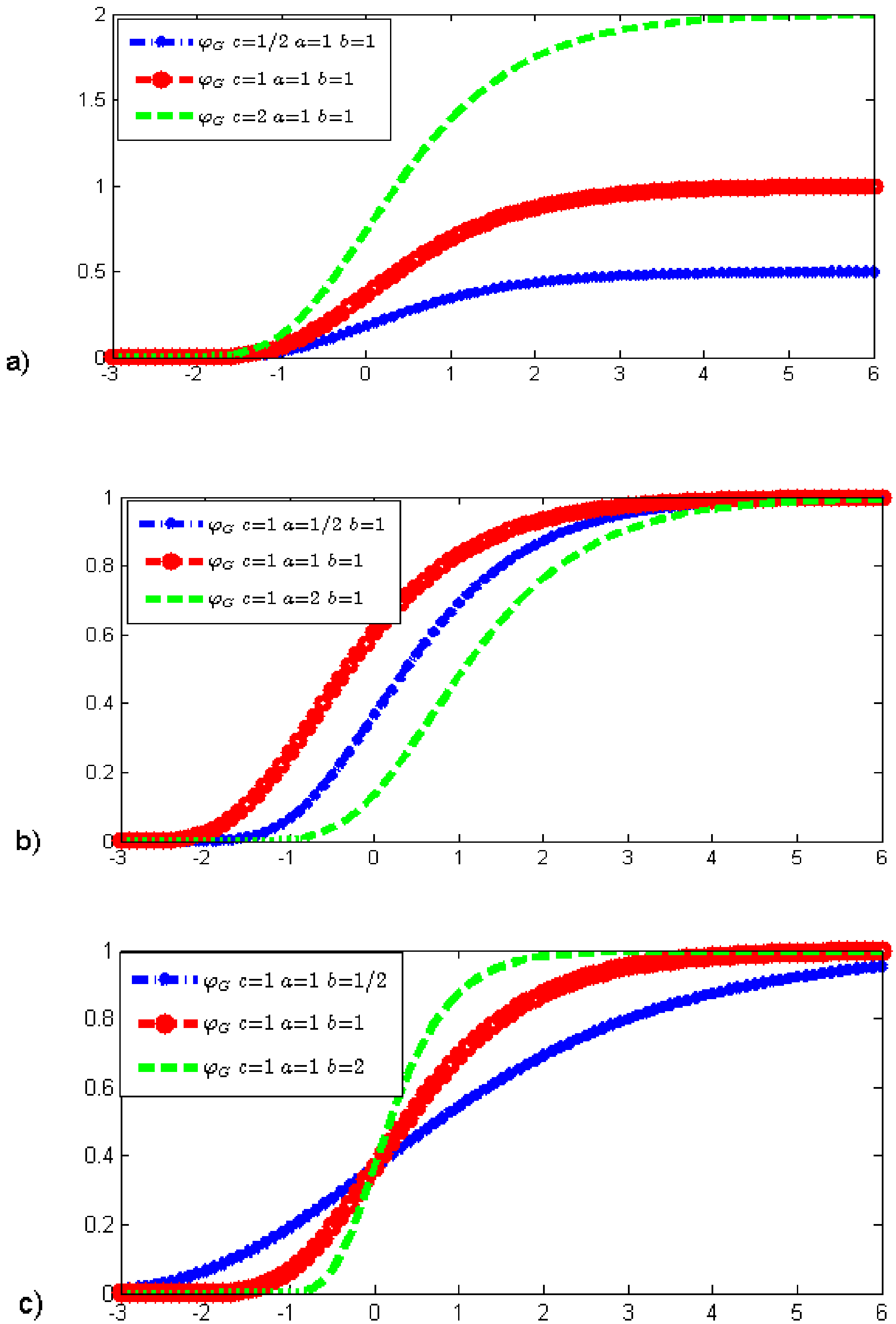

Figure 1. Illustration of Gompertz function: (a) *c* varies with fixed *a* and *b*, (b) *a* varies with fixed *b* and *c* and (c) *b* varies with fixed *a* and c.

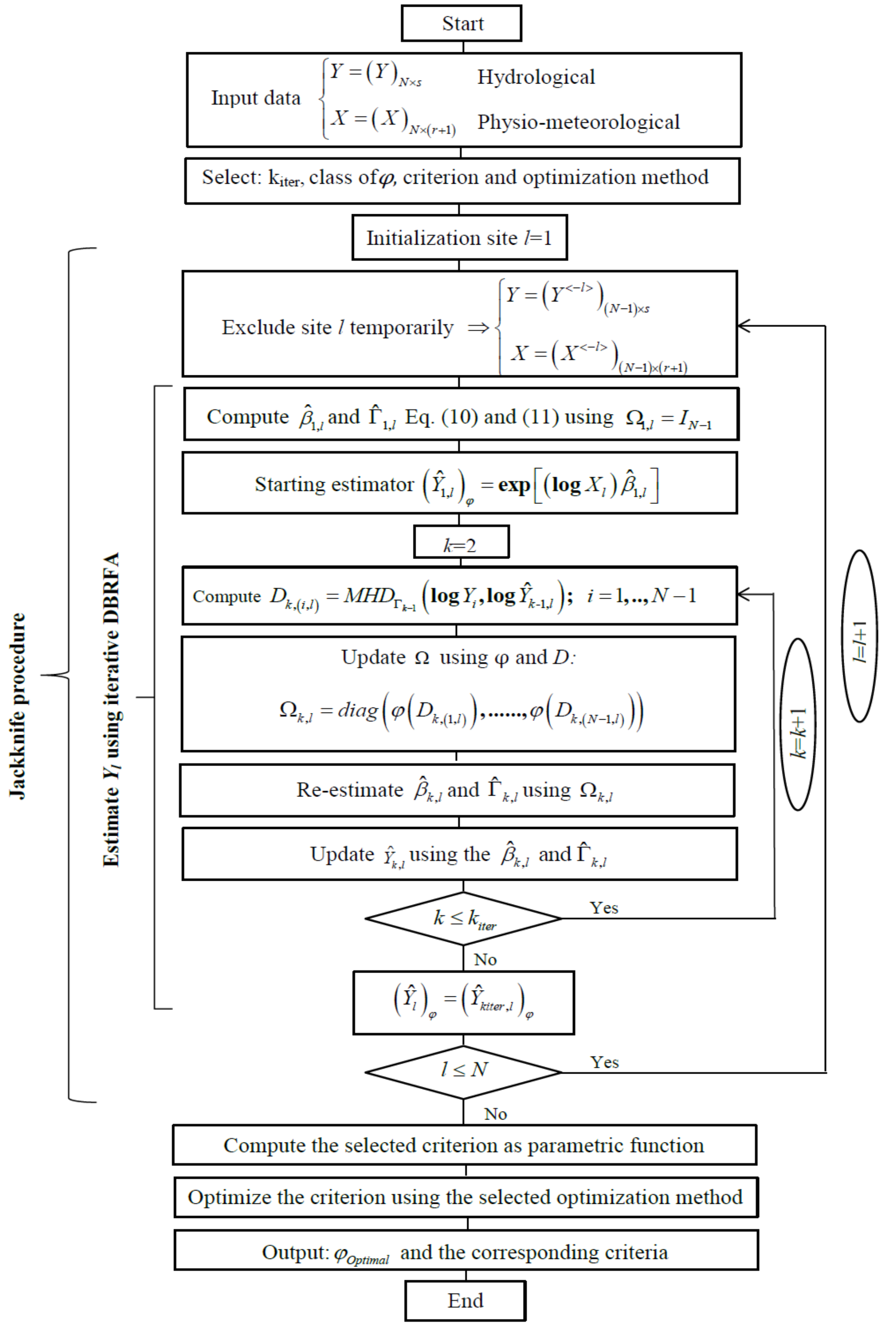

Figure 2. An overview diagram summarizing the optimization procedure of the DBRFA approach.

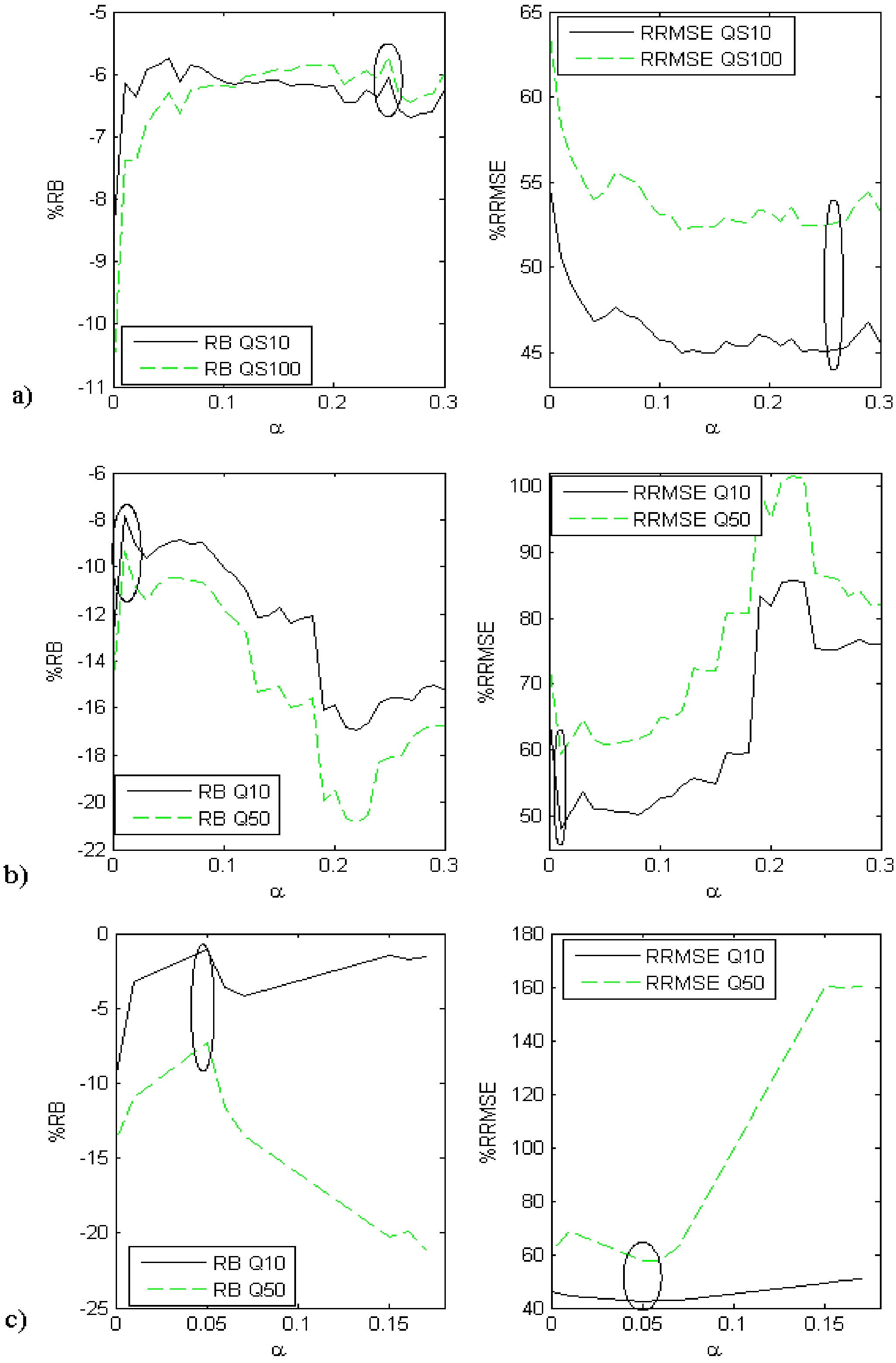

Figure 3. Optimal value of the neighborhood coefficient α for the CCA approach for: (a) Southern Quebec, (b) Arkansas and (c) Texas. The first column illustrates the RB and the second column illustrates the RRMSE.

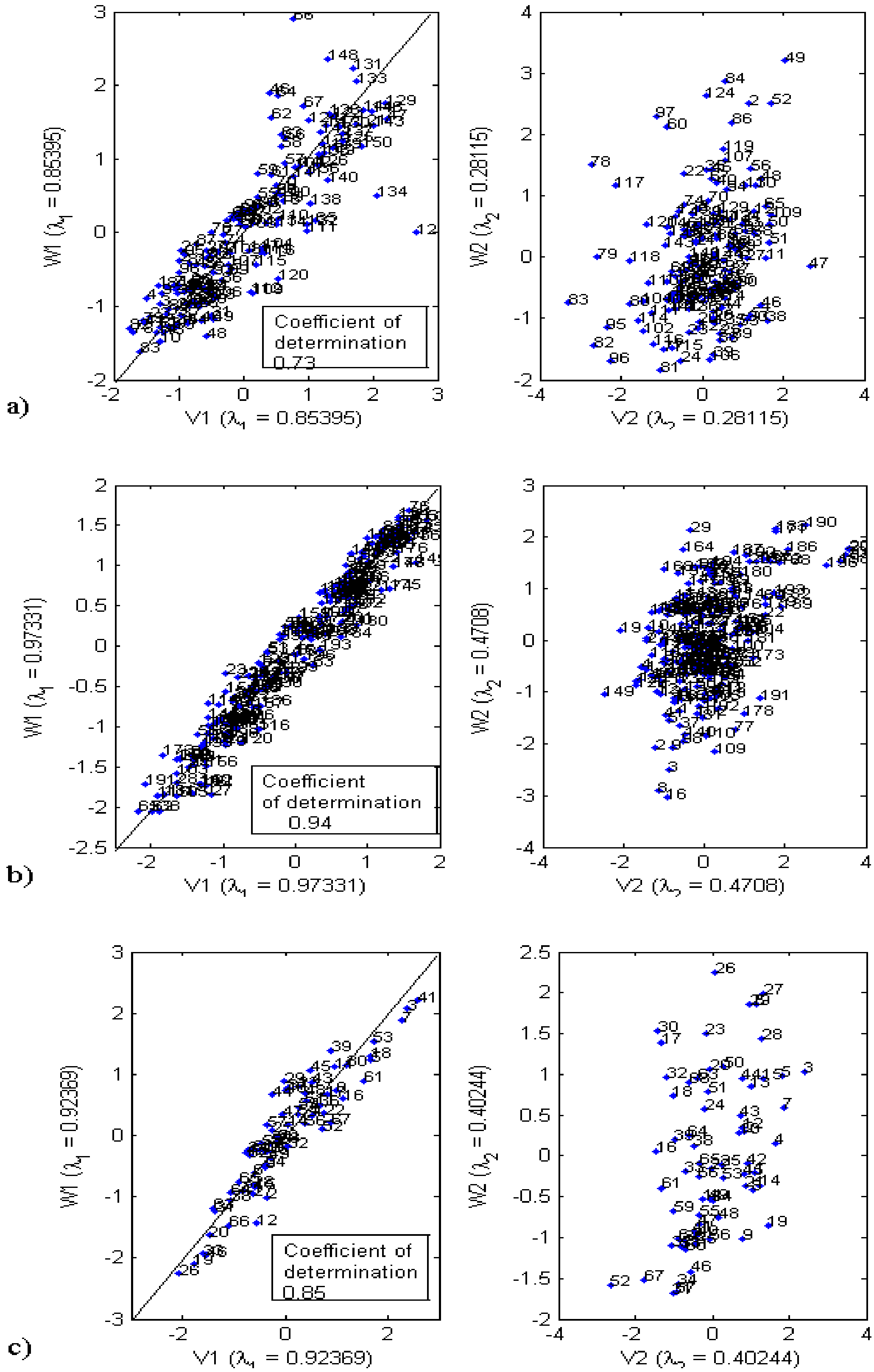

Figure 4. Scatterplot of sites in the canonical spaces (V1, W1) and (V2, W2) for: (a) Southern Quebec, (b) Arkansas and (c) Texas. The first column illustrates the canonical (V1, W1) space and the second column illustrates the (V2, W2) space.

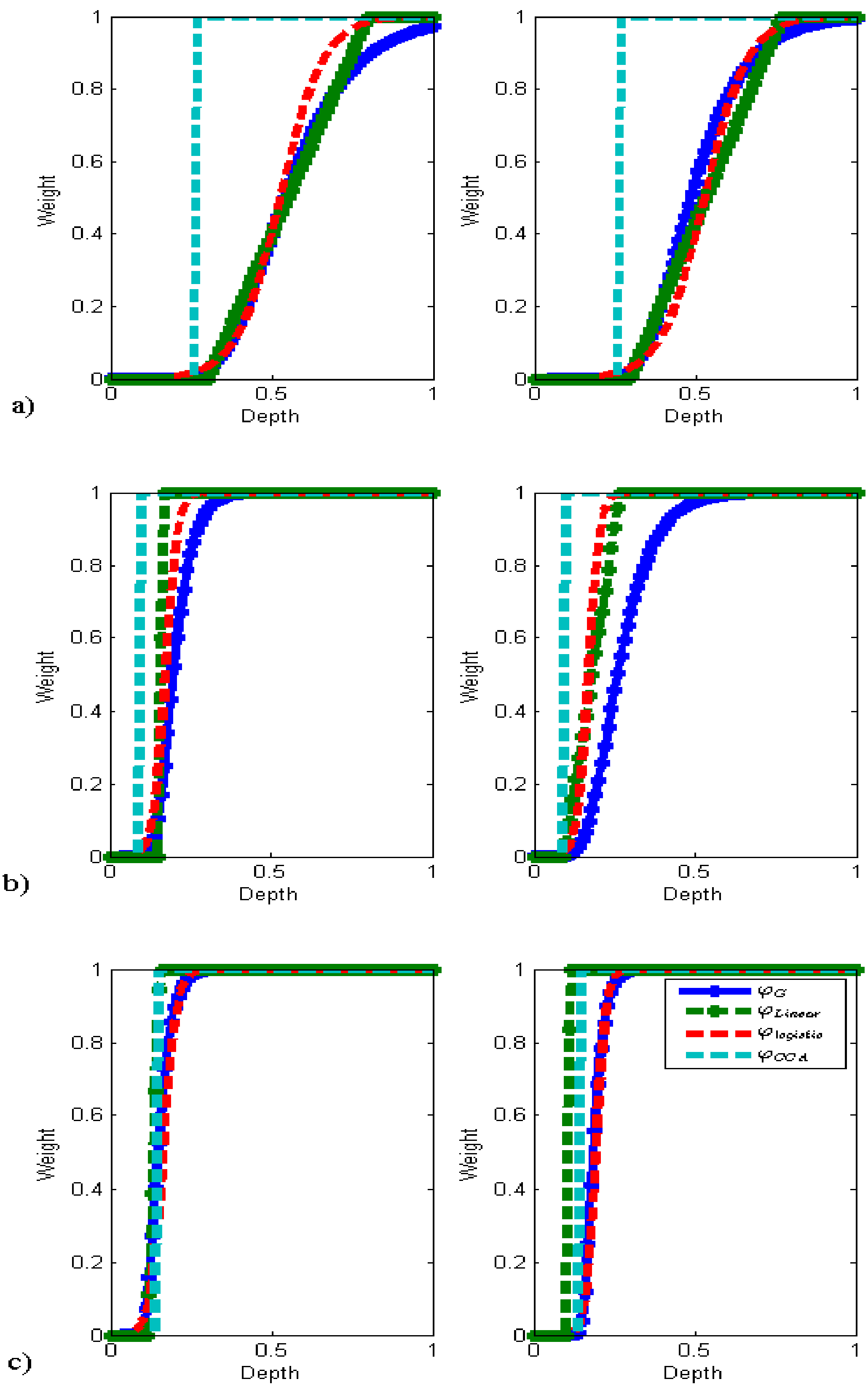

Figure 5. Optimal weight functions for: (a) Southern Quebec, (b) Arkansas and (c) Texas. The first column illustrates the weight functions optimal with respect to RRMSE and the second column illustrates the weight functions optimal with respect to RB.

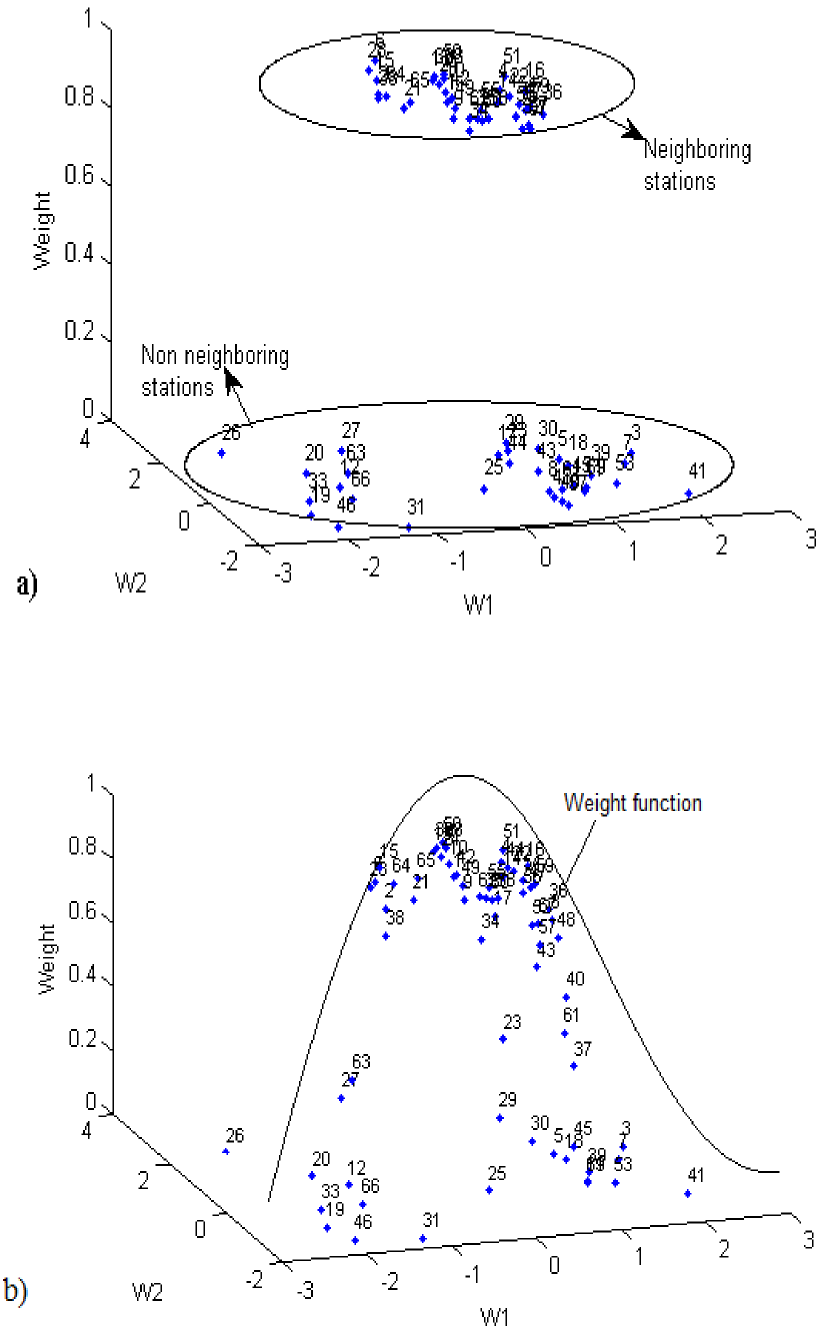

Figure 6. Weight allocated to each gauged-site to estimate the target-site number 25 in the Texas region in the Canonical hydrological space (W1, W2) using: (a) CCA with optimal α and (b) the DBRFA approach with optimal φ_G .

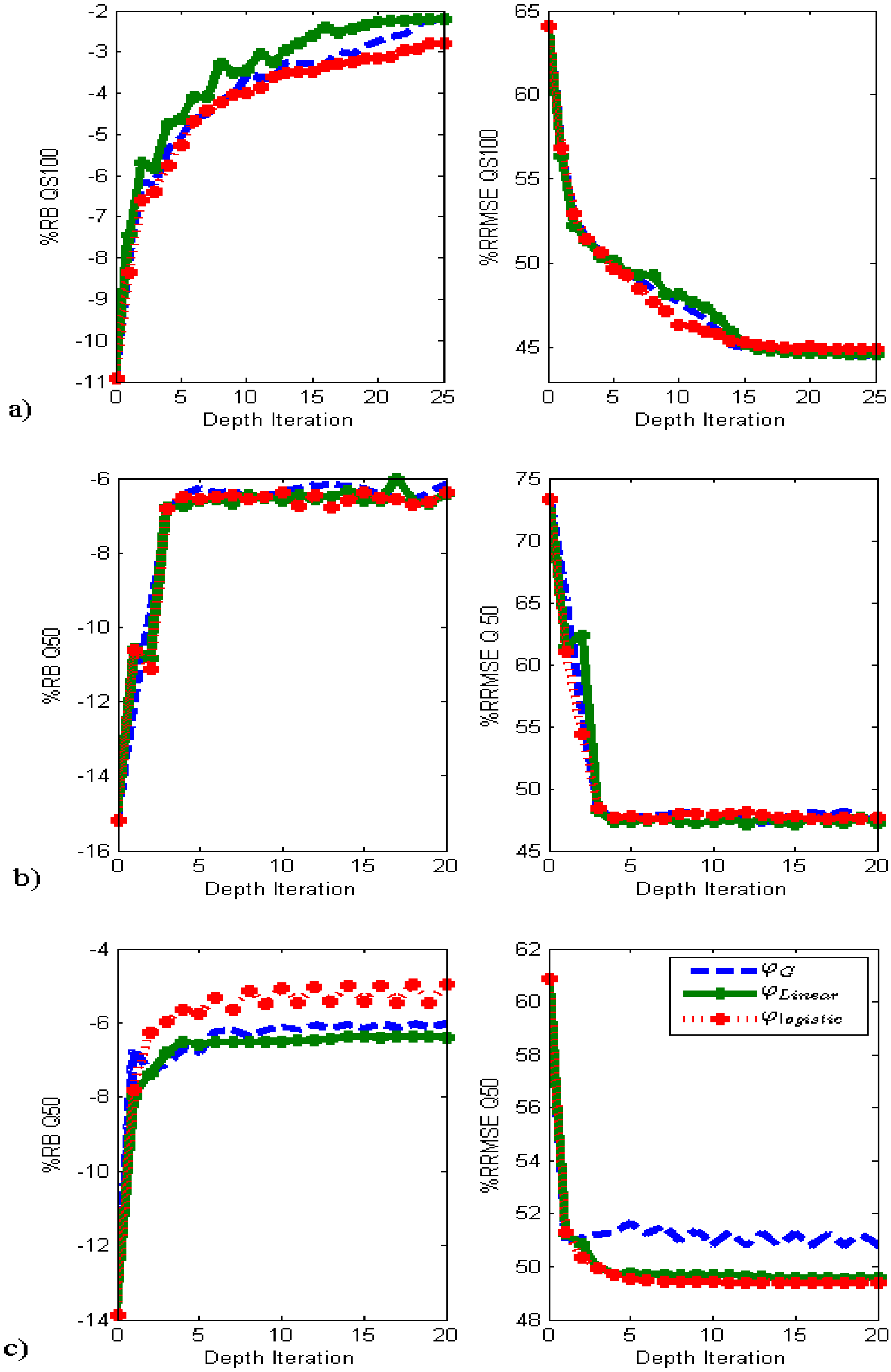

Figure 7. Variation of criteria (RB and RRMSE) as a function of the depth iteration number for the estimation of (a) QS100-Southern Quebec, (b) Q50-Arkansas and (c) Q50-Texas.

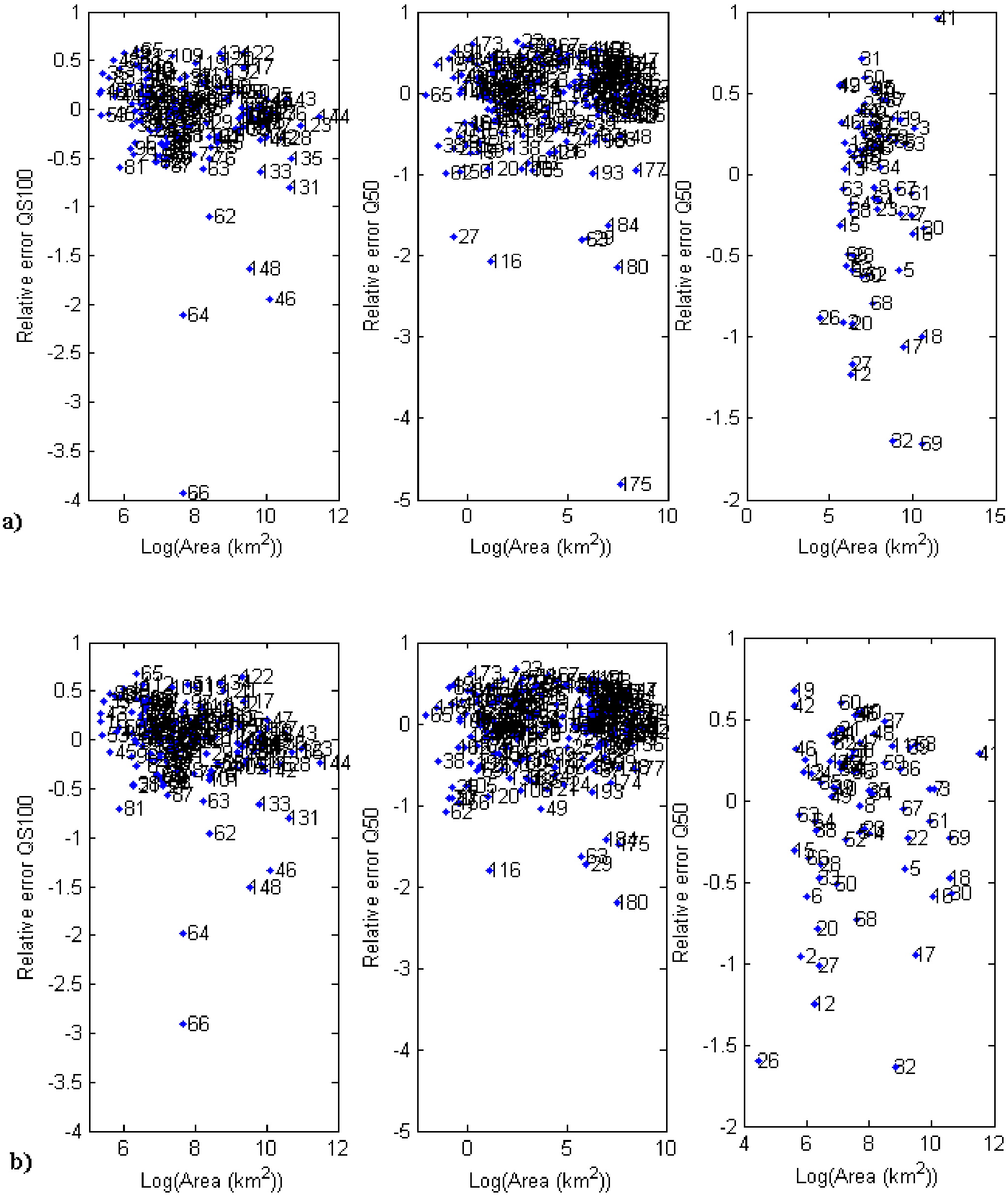

Figure 8. Relative quantile errors using: (a) φ_{CCA} and (b) φ_G . The first column illustrates the error of QS100 in southern Quebec, the second column illustrates the errors of Q50 in Arkansas and the third column illustrates the errors of Q50 in Texas.