Stoch Environ Res Risk Assess (2016) 30:449–462
DOI 10.1007/s00477-015-1092-7

ORIGINAL PAPER

Non-linear canonical correlation analysis in regional frequency analysis

D. Ouali[1] · F. Chebana[1] · T. B. M. J. Ouarda[1,2]

Published online: 12 June 2015

Abstract Hydrological processes are complex non-linear phenomena. Canonical correlation analysis (CCA) is frequently used in regional frequency analysis (RFA) to delineate hydrological neighborhoods. Although non-linear CCA (NL-CCA) is widely used in several fields, it has not been used in hydrology, particularly in RFA. This paper presents an overview of techniques used to reproduce non-linear relationships between two sets of variables. The approaches considered in this work are based on NL-CCA using neural networks (CCA-NN), coupled to a log-linear regression model for flood quantile estimation. In order to demonstrate the usefulness of these approaches in RFA, a comparative study between the latter and linear CCA is performed using three different databases from North America. Results show that CCA-NN is more robust and can better reproduce the non-linear relationship structures between physiographical and hydrological variables. This reflects the high flexibility of this approach. Results indicate that for all three databases, it is more advantageous to proceed with the non-linear CCA approach.

Keywords Non-linear canonical correlation analysis · Neural network · Regional frequency analysis · Homogeneous region · Hydrological neighborhood · Ungauged basins

Abbreviations

DHR	Delineation of homogeneous regions
RE	Regional estimation
CCA & LR	CCA associated to a log-linear regression
CCA-NN & LR	Non-linear CCA based on Neural Network in DHR step associated to a log-linear regression in the RE step
CCA-NN & CLR	Non-linear CCA based on Neural Network in DHR step associated to a log-linear regression in the canonical space in the RE step

1 Introduction and literature review

One of the main objectives of regional frequency analysis (RFA) is the estimation of extreme event quantiles (e.g. floods and droughts) at sites where little or no hydrological data is available. In general, RFA procedures have two main steps, namely the delineation of homogeneous regions (DHR) and regional estimation (RE) (e.g. Chebana and Ouarda 2007, 2008; Ouarda et al. 2008a). For each of these two steps, a large number of methodologies have been proposed (Ouarda et al. 2008b). Canonical correlation analysis (CCA) is one of the most commonly used methods for DHR where it consists in identifying linear combinations of variables within the same group, for which the canonical correlation is maximal. Ouarda et al. (2008a) demonstrated the advantages of CCA by comparing its performance to other techniques such as the hierarchical cluster analysis approach. However, note that in Shu and Ouarda (2007), CCA was used not for the DHR step, but to form a canonical physiographic space over which an

✉ D. Ouali
dhouha.ouali@ete.inrs.ca

[1] Institut National de la Recherche Scientifique, Centre Eau Terre et Environnement, 490, rue de la Couronne, Quebec, QC G1K 9A9, Canada

[2] Institute Centre for Water Advanced Technology and Environmental Research, P.O. Box 54224, Abu Dhabi, UAE

artificial neuronal network (ANN) is then employed to estimate flood quantile.

CCA is an important statistical tool for multivariate data analysis. However, it presents a drawback in the interpretation of results, which seems to be often difficult. In addition, this approach is based on a linear foundation and, hence, is not able to adequately describe non-linear relationships between variables. Therefore, CCA may not be suitable for representing hydrological processes in the DHR step. Two groups of variables are usually considered in RFA: (i) hydrological variables and (ii) meteorological and/or physiographical characteristics of the watersheds (Ouarda 2013). Hydrological processes are relatively complex because of the variability in the response of watersheds which does not generally result from a linear relationship between the hydrological and the physiographical characteristics (e.g. Chen et al. 2008; Xu et al. 2010; Chebana et al. 2014). Hydrological processes and their inherent non-linearities could not be adequately represented by linear relationships. One aspect of the non-linearity is represented by the rainfall-runoff relationship. Indeed, the variations of meteorological variables and flows are linked by a non-linear relationship (Riad and Mania 2004). This non-linear behavior depends strongly on the physiographic characteristics of the watersheds. For instance, surface runoff is strongly influenced by the soil storage capacity and soil infiltration.

A number of statistical tools have been proposed in the literature to deal with the additional complexity associated to non-linearity in a variety of fields (e.g. Bolton et al. 2003; Yin 2007). Among the proposed techniques, we can mention non-linear principal component analysis (NL-PCA) (Rumelhart et al. 1985; Kramer 1991) and non-linear CCA (NL-CCA) (Dauxois and Nkiet 1998; Hsieh 2000). NL-PCA has been applied in various fields such as chemistry (Kramer 1991), image processing (Botelho et al. 2005) and atmospheric sciences (e.g. Sengupta and Boyle 1995; Monahan 2000). Sengupta and Boyle (1995) applied NL-PCA to average monthly rainfall data in the United States. Compared to conventional PCA, results showed that the non-linear approach is a more effective data reduction tool. It was also demonstrated that NL-PCA represented better the variation of variables than ordinary PCA. However, this method presents some technical drawbacks (Malthouse 1998).

Although the above constraints of the NL-PCA also persist for NL-CCA (Hsieh 2000), the latter seems to provide better results than the CCA. NL-CCA was used in several fields, such as analysis of voice conversion (e.g. Zhihua and Zhen 2010), biomedicine (e.g. Campi et al. 2013), medicine (e.g. Wang et al. 2005) and sociology (e.g. Frie and Janssen 2009). A number of techniques related to NL-CCA have been proposed in the literature. For instance, Dauxois and Nkiet (1998) introduced measures of association between two random variables based on NL-CCA. Among the most studied non-linear methods associated to CCA, we can mention the neural network approach (NN) (Hsieh 2000), genetic algorithms (GA) (Kruger et al. 2004) and Kernel based methods (Akaho 2001; Hardoon and Shawe-Taylor 2009). Recently, Nagai (2013) proposed an optimization approach based on cross validation to optimize the NL-CCA parameters. In terms of applications, the non-linear method based on NN was adopted in a number of studies in meteorology and climatology. For example, Wu and Hsieh (2002) studied the El Nino Southern oscillation using NL-CCA based on the NN approach (CCA-NN). They showed the ability of CCA-NN to detect non-linearity between surface wind stress and sea surface temperature. Hsieh (2001) also applied CCA-NN to study the relationship between sea level pressure in the tropical Pacific and sea surface temperature. Results revealed the ability of this model to characterize non-linearity between variables, which was not the case with the conventional CCA.

Other studies in the past were interested by treating non-linear aspects of categorical variables (qualitative). Gifi (1990) presented two different techniques and algorithms, mainly OVERALS and CANALS to deal with such qualitative variables. However, the treated variables in RFA are quantitative and continuous. Therefore, the latter methods are not applicable in the context of the present study. In Table 1, all non-linear approaches discussed previously are summarised including their advantages and drawbacks. Note that methods designed for quantitative variables are more flexible than those for categorical ones.

Despite strong evidence concerning the non-linearity of hydrological processes, NL-CCA approaches have not yet been considered in hydrology. In RFA, non-linear approaches can account for possible non-linearities in order to determine the most representative homogeneous regions and lead to a better regional estimation. The purpose of the present paper is to deal with the issue of non-linearity in RFA by introducing NL-CCA in the DHR strep in order to improve its performance and representativeness.

The present paper is organized as follows: In the following section, the potential of NL-CCA in the DHR step is developed. In order to verify and validate the usefulness of the NL-CCA approach for the modelling of hydrological processes, a comparative study is carried out in Sect. 3 using three different datasets from North America (Quebec, Arkansas and Texas). These approaches are used in the delineation of hydrological neighborhoods where the obtained results are presented and discussed in Sect. 4. The conclusions of this work are reported in Sect. 5.

Table 1 Summary of common methods of NL-CCA

Variables	Method	Advantages	Drawbacks
Categorical	CANALS		Requires only two sets of variables Allows only a small number of possible values
	OVERALS	Ability to treat **k** groups of variables	Allows only a small number of possible values
Quantitative	CCA-NN		Significant computation time Black box A fairly complex mathematical structure
	CCA-K	Flexibility Low computation time No local optima	Difficult to interpret
	CCA-GA	A parsimonious technique Easier to interpret	

2 Background and methodology

In this section we present a brief description of the use of CCA in RFA, as well as a description of the NL-CCA method and its application to RFA.

2.1 Canonical correlation analysis in RFA

CCA is a multivariate analysis method used to identify the correlations that may exist between two groups of variables. It has been applied in a number of fields, such as seasonal climate forecasting (e.g. Barnett and Preisendorfer 1987), management science (e.g. Tishlert and Lipovetsky 1996), forecasting of accident risk modeling (e.g. Michael and Raymond 2003), river thermal regime modeling (e.g. Guillemette et al. 2009), water quality estimation (e.g. Khalil et al. 2011) and especially flood frequency estimation (e.g. Ouarda et al. 2001).

As mentioned above, in RFA, variables of interest are mainly hydrological and physiographical variables. We denote Y the vector describing hydrological variables, and X the vector containing meteorological and/or physiographical variables. Considering linear combinations of variables $X_1, X_2, \ldots, X_q$ and $Y_1, Y_2, \ldots, Y_r$, we obtain a new canonical space composed by canonical vectors U_i and V_i such as:

$$U_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{iq}X_q \tag{1}$$

$$V_i = b_{i1}Y_1 + b_{i2}Y_2 + \cdots b_{ir}Y_r \tag{2}$$

where $i = 1, \ldots, p$ with $p = \min(r, q)$. The canonical space is built under constraints of unit variance and maximum correlation between pairs of canonical variables. Let Λ be a p-by-p diagonal matrix composed of canonical correlation coefficients given by:

$$\lambda_i = corr(U_i, V_i); \quad i = 1, \ldots, p \tag{3}$$

Once the first pair of canonical variables $(U_1, V_1)_p$ is obtained, other canonical pairs are obtained subject to the constraint $corr\left(U_i, V_j\right) = 0$ for $i \neq j$. Note that all distinct hydrological canonical variables (as well as distinct physiographical variables) are also uncorrelated (Ouarda et al. 2001).

In order to improve quantile estimations in RFA, CCA is commonly used for the determination of neighborhoods of target sites. For an ungauged site, the canonical meteorological-physiological information U_0 is usually known but the hydrological information V_0 is not available. The hydrological mean position of the target site S is given by ΛU_0. Hence, a 100 (1−α) % confidence level neighborhood is identified by the Mahalanobis distance. It is considered between the mean position of target site ΛU_0 and positions of other sites V, such that:

$$(V - \Lambda U_0)'\left(I_p - \Lambda^2\right)^{-1}(V - \Lambda U_0) \leq \chi^2_{\alpha,p} \tag{4}$$

where $P(\chi^2_p \leq \chi^2_{\alpha,p}) = 1 - \alpha$ and χ^2_p has a Chi squared distribution with p degrees of freedom. Expression (4) is used to define an ellipsoid representing the neighborhood region for the ungauged site associated to ΛU_0 (Ouarda et al. 2001).The equation of the ellipsoid has the following form:

$$\frac{(V_1 - \Lambda_1 U_{01})^2}{a^2} + \frac{(V_2 - \Lambda_2 U_{02})^2}{b^2} = 1 \tag{5}$$

where V_1 and V_2 denote the hydrological canonical variables, Λ_1 and Λ_2 are the canonical correlation coefficients, $(\Lambda_1 U_{01}, \Lambda_2 U_{02})$ are the coordinates of the center of the ellipsoid and a and b denote respectively the semi-major axis (or focal) and the semi-minor axis (Ballard 1981). Expression (5) is the equation of an ellipsoid in an orthonormal base (two orthogonal unit vectors), where axes are parallel to the coordinate system axes.

2.2 Nonlinear CCA using a neural network approach (CCA-NN)

An artificial neuron network ANN is a fairly simple mathematical model compared to the natural biological evolution, with a running-inspired design of biological neurons (Bishop 1995). It consists essentially in several neurons generally organized in layers. The output of each neuron results from the weighted sum of inputs, and transformed by a transfer function. Different transfer functions can be used (Duch and Jankowski 1999). ANNs have been widely used in a number of fields, such as in geology where Li et al. (2014) utilized the back-propagation (BP) neural network approach to forecast the geological hazard linked to bank destruction and landslides, and in hydrology where Zaier et al. (2010) used ANNs to model lake ice thickness, and Chen et al. (2014) used ANNs to model the rainfall-runoff relationship. As previously indicated, ANNs were integrated in RFA for instance by Ouarda and Shu (2009) and by Aziz et al. (2014) for the estimation of flood quantiles at ungauged sites.

In the meteorological field, Hsieh (2000) developed a NL-CCA version based on ANN (CCA-NN). The CCA-NN approach consists on establishing non-linear combinations between groups of original variables (X and Y) and the new canonical variables (U and V) via a transfer function. Consider the following hidden layer:

$$h_k^{(x)} = f\left(\left(W^{(x)}x + b^{(x)}\right)_k\right); \quad k \,\text{and}\, n = 1, \ldots, l \tag{6}$$

$$h_n^{(y)} = f\left(\left(W^{(y)}y + b^{(y)}\right)_n\right) \tag{7}$$

where $W^{(x)}$ and $W^{(y)}$ are weight matrices, $b^{(x)}$ and $b^{(y)}$ are vectors of biased parameters, k and n denote respectively the indexes of the vector's elements $h^{(x)}$ and$h^{(y)}$ and l denotes the number of hidden neurons. The transfer function f, the same for x and y, is generally set to the hyperbolic tangent function (Hsieh 2000):

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{8}$$

Multivariate canonical neurons U and V are determined from a linear combination of respective neurons $h^{(x)}$ and $h^{(y)}$(but from a non-linear combination with respect to x and y):

$$U = w^{(x)}h^{(x)} + \overline{b}^{(x)} \tag{9}$$

$$V = w^{(y)}h^{(y)} + \overline{b}^{(y)} \tag{10}$$

Without loss of generality, U and V are assumed to have zero mean. Thus, we have

$$\overline{b}^{(x)} = -\left\langle w^{(x)}h^{(x)}\right\rangle \quad \text{and} \quad \overline{b}^{(y)} = -\left\langle w^{(y)}h^{(y)}\right\rangle \tag{11}$$

where $\langle z\rangle$ is the empirical mean of variable z.

A limitation of the CCA-NN is that, once applied to the original data, it provides only one pair of canonical variables, i.e. one for the physiographical variables and one for the hydrological variables. This may lead to ignoring a part of the information since it is not guaranteed that the first pair of canonical variables covers a significant part of the explained variance. To overcome this problem, the notion of modes was considered (Hsieh 2000). It consists in applying CCA-NN on the original datasets. The obtained result, denoted x', is related to the first mode. For the second mode, the CCA-NN is applied to the initial data, i.e. the set x, excluding the first mode. In other words, we determine the unexplained information in the previous mode by reapplying the procedure on the new variables:

$$I_2 = x - x' \tag{12}$$

Based on Eq. (12) we get:

$$J_2 = y - y' \tag{13}$$

where y' is the result of the first iteration, y is the matrix of original data.

The same procedure applies for higher order modes by considering each time the residual of the previous mode as input. The number of iterations, m, should be at least equal to the lowest number of variables, p in our case. The final result consists in summing up the results of all considered iterations:

$$x_{estimated} = x' + x'' + \cdots + x^m \tag{14}$$

where x^m is the result of the m^i th iteration, $m \geq p$. Therefore, the use of several modes may increase the percentage of the information contained in the resulting canonical variables.

2.3 Adaptation of CCA-NN to regional frequency analysis

For more clarity and to avoid confusion, it is important to note that in the approach proposed by Shu and Ouarda (2007), a CCA-based ANN model is used for flood quantile estimation without considering the DHR step and in which the employed CCA is the linear one. The aim of the linear CCA in Shu and Ouarda (2007) is to filter the signal from the original data and apply the ANN model on the canonical variables. However, in the present work the non linear version of CCA using ANN (CCA-NN) is introduced in order to identify homogeneous regions, while a log linear regression model is used in the RE step.

Several versions of CCA-NN may be considered depending on the selected cost functions (canonical correlation, mean square error MSE, mean absolute error MAE). Indeed, Cannon (2008) introduced a robust version of CCA-NN based on the biweight midcorrelation coefficient as a new measure of correlation instead of the Pearson correlation. After choosing the cost functions, canonical variables can be obtained and hence one can determine the hydrological neighborhood for an ungauged site. In the non-linear case, the variables V_1 and V_2 denote the hydrological canonical variables of the first and second mode, respectively, and Λ_1 and Λ_2 are the canonical correlation coefficients of the two modes. Identifying the physiographical coordinates of an ungauged site, U_{01} and U_{02}, is performed using relation (9).

Similarly to the neighborhood of the linear case, the non-linear one can be obtained using the same constraint. However, the equation of the ellipsoid is different from the linear case (5), since the axes are not parallel to those of the coordinate system.

Let Y denote an array of hydrological data and V the corresponding canonical variable, thus we can write:

$$Y = h(V) \tag{15}$$

Therefore by substituting (15) in (13) we obtain:

$$h_2(V_2) = h_1(V_1) - y' \tag{16}$$

Note that h, h_1 and h_2 are known non-linear functions.Hence, the angle $\theta = (V_1, V_2)$ is different from $\pi/2$. Since the axes of the ellipsoid are always perpendicular, the ellipsoid is then rotated through an angle ϕ relative to the coordinate system(V_1, Z). As illustrated in Fig. 1, (V_1, Z) is an orthonormal basis with $Z = \sin(\theta)\ V_2$. The equation of the ellipsoid in the non-linear canonical space is given by:

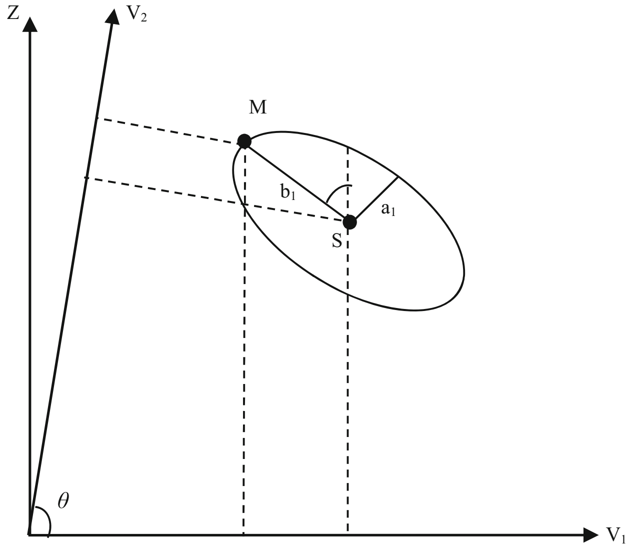


Fig. 1 Geometrical definition of the homogeneous region in the non-linear canonical space

$$\frac{(P_1 - \Lambda_1 U_{01})^2}{a_1^2} + \frac{(P_2 - \Lambda_2 U_{02})^2}{b_1^2} = 1 \tag{17}$$

where:

$$P_1 = V_1 \cos\phi - Z \sin\phi \quad \text{and} \quad P_2 = V_1 \sin\phi + Z \cos\phi \tag{18}$$

Note that the angle is the same for all sites and with different values of α. It depends only on θ: $\phi = f(\theta)$. Equation (5) related to the linear CCA is a special case of (17) with a zero angle of rotation ϕ and $\theta = \pi/2$.

Similarly to CCA, the objective of NL-CCA consists in reducing the dimensions of hydrological and physiographical/meteorological spaces by taking into account the relationships between the considered variables. However, the construction of CCA reflects only linear relationships. The use of NL-CCA is necessary especially in the presence of non-linear structures. Note that the non-linearity in the hydrological processes is related to the non-linearity treated in NL-CCA.

To get a clear view of the correlation structure, it is essential to locate the source of interactions between variables. Note that the non-linearity in NL-CCA exists between the canonical and original variables of the same set, e.g. between U and physiographic variables. However, the non-linearity that occurs through the hydrological process is between hydrological variables Y and physiographical ones X. We show that these two types of non-linearities are connected. Indeed, in the NL-CCA context, the canonical variables can be written as:

$$U_i = f_1(X_i) \quad \text{and} \quad V_i = f_2(Y_i) \tag{19}$$

where f_1 and f_2 are non-linear functions (or linear in the case of CCA) and $i = 1, \ldots, p$. The simplest situation is the linear case, where more complex relations load to the same correlation:

$$U_i \approx \lambda_i V_i \tag{20}$$

The symbol ≈ indicates that both sides are approximately equal. Using relation (19), we obtain:

$$U_i \approx \lambda_i f_2(Y_i) \approx h(Y_i) \tag{21}$$

Substituting Eq. (19) into (21), we get:

$$h(Y_i) \approx f_1(X_i) \tag{22}$$

which leads to

$$Y_i \approx k(X_i) \tag{23}$$

where k (.) is a general function (if h is invertible k would be equal to $h^{-1} of_1$).

Thus non-linear relations described by (19) are equivalent to non-linear relationships between the two groups of original variables (23). On the other hand, the presence of non-linearity in hydrological processes, between X and Y, leads to a non-linearity between canonical variables. Therefore, it is necessary to use the nonlinear approach in the context of RFA.

2.4 Regional estimation

Among the various RE methods, the most popular ones are the index-flood and regression models (Ouarda 2013). In this paper we focus on the multivariate log-linear regression model, since it is more appropriate to use with CCA and with the available datasets. The relationship between flood quantiles (Y) and the physiographical/meteorological characteristics (X) is generally described by a power product model. With a log-transformation, the following log-linear model is obtained:

$$\log(Y) = \beta \log(X) + \varepsilon \tag{24}$$

where β is a vector of parameters and $\boldsymbol{\varepsilon}$ represents the error (see Pandey and Nguyen (1999) for instance).

2.5 Evaluation criteria

To assess the performance of the proposed techniques, different criteria are used. Each model is evaluated using the following five indices: the Nash criterion (NASH) which provides a general evaluation of the quality estimation, the root mean squared error (RMSE) providing information about the accuracy of the estimator in an absolute scale, the relative RMSE (RMSEr) which is related to the relative scale, the mean bias (BIAS) and the relative mean bias (BIASr) provide a measure of the magnitude of overestimation or underestimation of a model. These indices are estimated based on a jackknife resampling procedure (e.g. Ouarda et al. 2001). It consists in removing temporarily each site and considering it as an ungauged one. The regional estimate is thus compared to the local estimate and the ability of each method is then evaluated.

The correlation coefficient and the proportion of explained variance are also used as evaluation criteria in the present work. The explained variance is deduced from the correlations between canonical components and initial variables, (Van Den Wollenberg 1977):

$$\sigma_E^2(U_i) = \frac{1}{q}\sum_{j=1}^{q}\left[corr\left(U_i, X_j\right)\right]^2 \tag{25}$$

In a similar way, expression (25) is also valid for hydrological variables $Y_{j,j} = 1,\ldots,r$ and canonical variables $V_{i,i} = 1,\ldots,2$.

3 Case study

3.1 Data

The data used in this study covers three regions in North America, namely the province of Quebec (Canada), and the states of Arkansas and Texas (USA). The data from Arkansas and Texas are available in Tasker et al. (1996).

The first region includes 151 hydrometric stations and is located in the southern part of the province of Quebec, between 45° and 55° N. The considered physiographical and meteorological variables are those used previously by Chokmani and Ouarda (2004): the mean basin slope (PMBV), the basin area (BV), the proportion of the basin area covered with lakes (PLAC), the annual mean total precipitation (PTMA) and the annual mean degree-days (DJBZ). Hydrological variables are at-site flood quantiles standardized by basin area to eliminate the scale effect (specific quantiles), denoted Q_{ST} for a return period T. For each site, the most appropriate statistical distribution has been identified in order to estimate the quantiles corresponding to different return periods. Two specific quantiles are selected for this study, namely the 10-year and the 100-year quantiles.

The second case-study concerns data from the state of Arkansas in the southern United States. Data stems from a hydrometric network composed of 204 gauging stations with drainage areas ranging from 0.13 to 6890 km^2. The same data was used by Tasker et al. (1996), namely the area (A), the slope of the main channel (S), the mean annual precipitation (P), the mean elevation of the watershed (EL), the length of the main stream (L), and estimated flood quantiles, Q_{ST}, corresponding to return periods of T = 2, 5, 10, 25 and 50 years.

The last region covers a hydrometric network of 69 stations in the state of Texas. Basin areas range between 86 and 101,000 km^2. The variables used are those indicated in Tasker et al. (1996) i.e. five physiographic variables (A, S, P, EL and L) and five flood quantiles which are the same as those considered in the Arkansas case study.

3.2 Model design

In order to determine the homogeneous region, both CCA and CCA-NN analysis were carried out in the DHR step using $r = 2$ hydrological variables and $q = 5$ physiographical variables for all case studies (Quebec, Arkansas and Texas).

To build a model able to provide flood quantile estimation using the neighborhood approach, the CCA and CCA-NN approaches are coupled to a log-linear regression

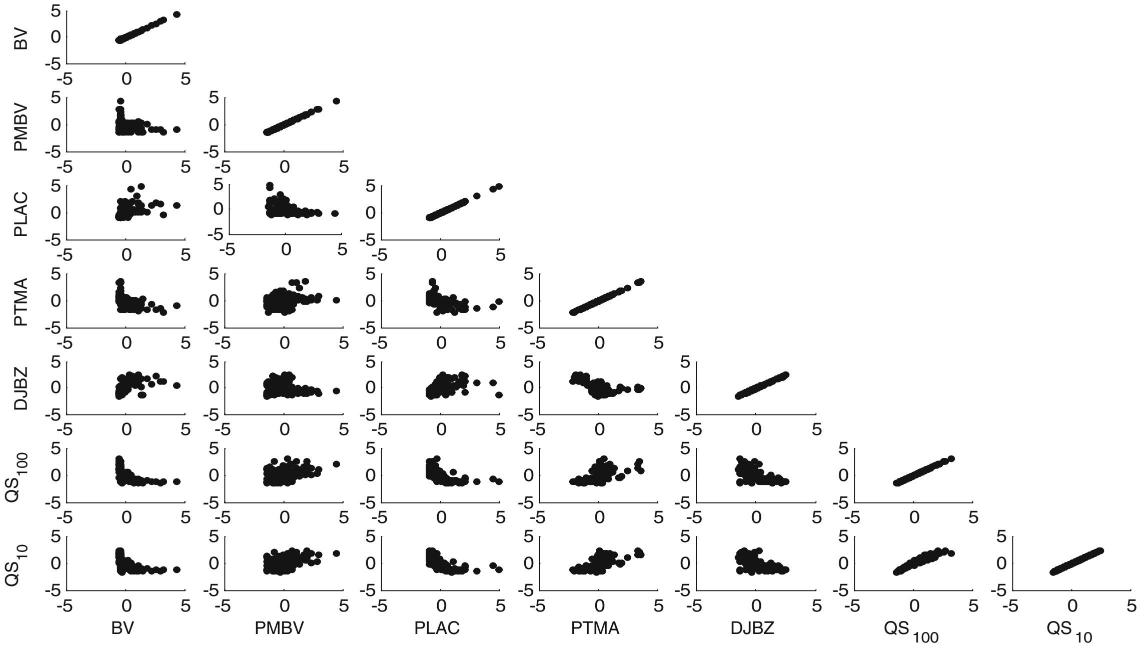


Fig. 2 Scatter plot of physiographical and hydrological variables—Quebec

(24) in the RE step (denoted CCA & LR and CCA-NN & LR respectively). For comparison purposes, two regression models are considered in the non-linear case, according to the explanatory input variables, either directly using the initial data (*X*) or using the physiographical canonical variables (U_1, U_2). The latter is denoted CCA-NN & CLR and has the advantage of considering only the useful information with a smaller number of variables.

To compare the obtained results with different approaches presented in Chebana and Ouarda (2008), we discuss essentially results related to Quebec. Results associated to the other two regions will be presented briefly. Actually, several versions of CCA-NN with different cost functions were treated (Correlation coefficient/Mean absolute error COR/MAE, biweight midcorrelation coefficient/Mean absolute error BICOR/MAE and biweight midcorrelation coefficient/Mean square error BICOR/MSE). In the section below, only the results associated to BICOR/MSE are presented and discussed since this version provides the lowest evaluation criteria values. This finding is in concordance with the conclusion presented by Cannon (2008). In addition, it should be noted that the choice of the transfer function is an important step in ANN modeling, as it can significantly affect the results. In the hydrological literature, the sigmoid and the hyperbolic tangent functions are most commonly used as nonlinear transfer functions (Dawson and Wilby 2001; Yonaba et al. 2010). In this regard, several transfer functions belonging to the sigmoid function class were tested (the arctangent, the hyperbolic tangent and the sigmoid), and the hyperbolic tangent function yielded the best results. Hence, this transfer function (8) is employed for all case studies in the neurons of the hidden layers. The outputs of this model are canonical variables when the model is designed to forward mapping, and original variables in the case of inverse mapping. In the current application, three NNs were considered where the first ensures the forward mapping, while the second and the third are relative to the inverse mapping.

After extracting the first CCA-NN mode, the extraction of second mode is carried out by taking the residual as input, i.e., the original data minus the first CCA-NN mode, as in (12). Hence, we obtain the canonical variables in the non-linear space. Based on the Mahalanobis distance (4), the hydrological neighborhood of each ungauged site is determined.

4 Results

In this section, we present the results of the regional flood estimation procedure where the CCA-NN approach is considered for the DHR step. First, preliminary results are presented in order to study the relationships between variables. Figure 2 presents scatter plots of flood quantiles and physiographical/meteorological variables for Quebec. The examination of the scatter plots shows different forms of relationships between variables. We note, for instance, the existence of non-linear relations. The most notable ones are those between the variable basin area (BV) and the rest of the variables. Table 2 presents the correlation coefficients

Table 2 Correlation between hydrological and physiographical variables-Quebec

	QS_{100}	QS_{10}
BV	−0.43	−0.46
PMBV	0.45	0.47
PLAC	−0.63	−0.67
PTMA	0.61	0.68
DJBZ	−0.59	−0.60

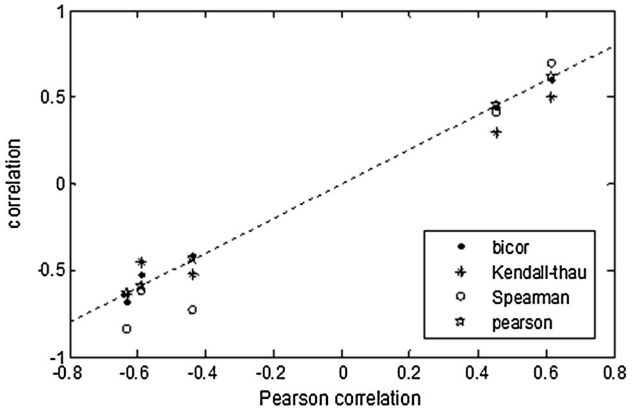


Fig. 3 Empirical comparison between the Pearson correlation and other measures of correlation (the Kendall tau, the Spearman Rho and the biweight midcorrelation)—Quebec

between the hydrological and the physiographical variables. Despite the existence of a relatively strong positive correlation between flood quantiles and PLAC on one hand, and negative linear correlation between quantiles and PTMA on the other hand, we can observe from Fig. 2 that these structures are rather non-linear. Further correlation measures are also evaluated between these variables. Figure 3 shows the correlation coefficients obtained by other correlation measures with respect to the Pearson correlation. This empirical comparison shows differences between measures, expressed by values higher or lower than those based on Pearson correlation. These behaviors indicate the existence of other dependence structures that are more complex than linearity.

By carrying out a linear CCA, the canonical correlation coefficients (3) are $\lambda_1 = 0.81$ and $\lambda_2 = 0.27$. In Chebana and Ouarda (2008), representations of data in the canonical spaces (not presented here to avoid repetition) show that the relationship between the first two canonical variables (U_1, V_1) can be considered to be linear, unlike variables (U_2, V_2) where linearity is relatively low.

In the following, results related to the CCA-NN are presented and discussed. Figure 4 presents the scatterplot of the study sites in the non-linear canonical spaces: physiographical (U_1, U_2) and hydrological (V_1, V_2). It is also convenient to present data in the spaces (U_1, V_1) and (U_2, V_2) to get prior information about the estimation error (Chebana and Ouarda 2008). This is illustrated in Fig. 5 for the non-linear case. A nearly linear relationship is observed between the two canonical variables (U_1, V_1). This is not the case for the couple (U_2, V_2). However, the CCA-NN scatterplot seems to be more linear than the scatterplot of the data set in the linear space (U_2, V_2) presented in Chebana and Ouarda (2008). This may be explained by the fact that the canonical correlation coefficients obtained from CCA-NN ($\lambda_1 = 0.90$ and $\lambda_2 = 0.36$ using (3) and (20)) are higher than their counter parts deduced from CCA.

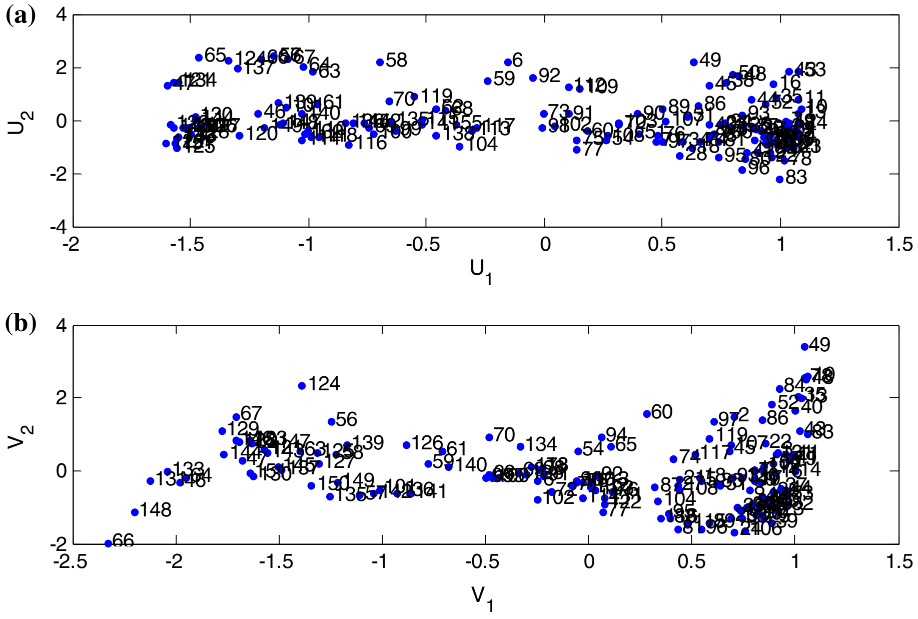


Fig. 4 Data set in the non-linear canonical spaces: **a** physiographical and **b** hydrological—Quebec

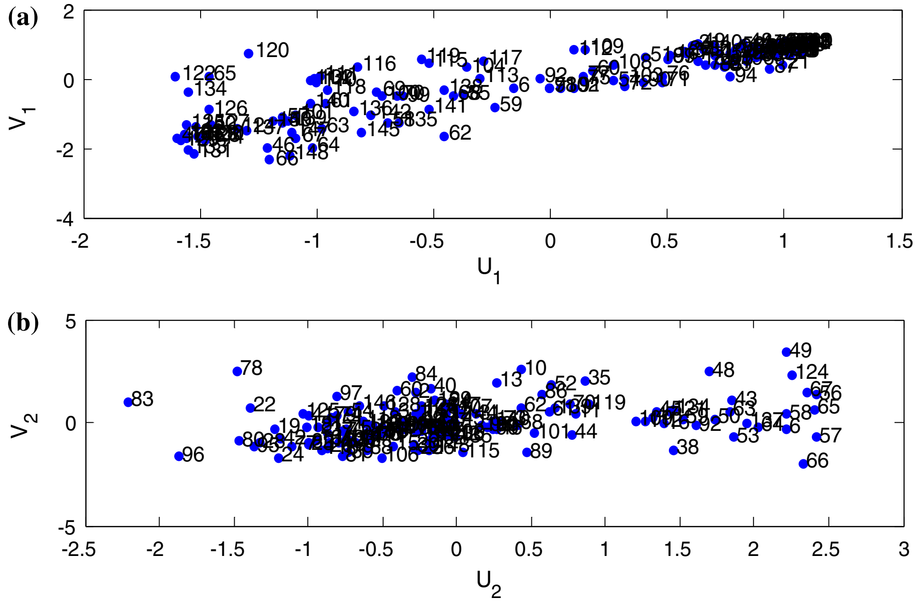


Fig. 5 Data set in the non-linear canonical spaces: **a** (U_1, V_1) and **b** (U_2,V_2)—Quebec

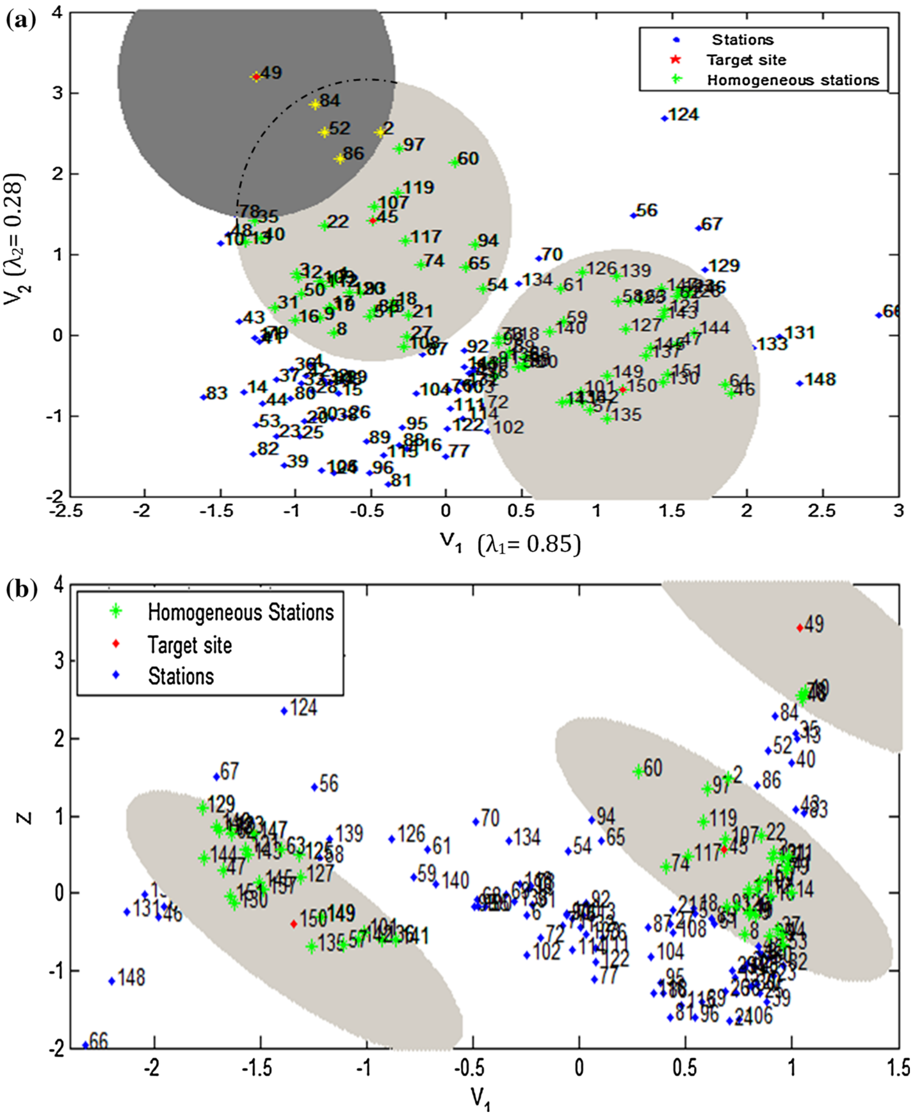
The explained variance (25), for the two first components, is respectively 51.16 and 97.36 % (vs 56.92 and 99 % in the linear CCA). Therefore, the canonical variables deduced from the linear CCA explain slightly better the variance of variables than those corresponding to CCA-NN. This may be due to the linearity induced by the correlation coefficient in the expression of the explained variance. However this does not affect the results significantly since the selection of canonical variables is based essentially on the canonical correlation coefficients.

In the following we study the difference between the linear and non-linear approaches in identifying the hydrological neighborhood. The neighborhoods of selected stations are presented for both CCA and CCA-NN approaches in Fig. 6. We observe a remarkable difference between the two approaches. Indeed, using the CCA, the neighborhood of each site is an ellipsoid with a zero angle of rotation. The non-linear method identified a rotated ellipsoid with a rotation angle $\phi \sim 21°$. Unlike CCA, the orientation of the CCA-NN ellipsoid tends to follow the shape of the data dispersion. For instance, the non-linear neighborhood of station 030340 (n = 45) identified 31 neighboring stations while the linear one identified a classical neighborhood with 39 stations, for the same value of α, $\alpha_{CCA-NN} = 0.2$. This means that the CCA-NN requires a smaller number of stations to reach the same RMSE as CCA. The optimal value of α corresponds to the minimum RMSEr. Figure 7 presents the variation of RMSEr for different values of α using CCA-NN. It can be seen that the optimal value $\alpha_{\mathrm{CCA\text{-}NN}}$ is 0.2. Note that for high values of α, the performance criteria tends to infinity.

To assess the magnitude of obtained results and their impact on RFA, we proceed to the RE step. Table 3 illustrates the jackknife results for all considered approaches through the criteria cited above. It can be seen that the NASH of the linear and non-linear models are substantially equal and sufficiently high to present acceptable results. For instance, for a return period of 100 years, the NASH of CCA is equal to 0.70 while it is equal to 0.71 for the non-linear case. Results indicate also that the RMSE of CCA-NN & LR and CCA & LR are almost equal whereas the RMSEr of the estimates computed by the CCA-NN & LR model are considerably lower than the linear model. By comparing the results with those obtained with the iterative procedure in Chebana and Ouarda (2008) and Wazneh et al. (2013) for the same data set, it can be seen that the proposed model, CCA-NN & LR, leads to best results among all models in terms of RMSEr. Indeed, while the linear approaches resulted in an RMSEr value of about 38 % for the quantile QS_{10} and 44 % for the quantile QS_{100}, the CCA-NN & LR RMSEr values are around 34 % for the quantile QS_{10} and 41 % for the quantile QS_{100}. It is also observed that the CCA-NN & LR results in both spaces, canonical and original, are very similar and are significantly better than the other models, i.e. the linear approach and the iterative procedure.

For all considered models, the BIAS is very close to zero with a slight improvement with the CCA-NN & LR approach. According to the BIASr criterion, the CCA-NN & CLR leads to the best results. However, in comparison with results reported in Wazneh et al. (2013), the BIASr of the proposed models is higher (for values of QS_{100} and QS_{10} BIASr values are about −6 and −7 % respectively using the CCA-NN & LR, versus around −2 and −3 %

Fig. 6 DHR results shown for stations 030340, 030420 and 02717 using: **a** CCA and **b** CCA-NN approaches, n = 45, 49 and 150 respectively—Quebec



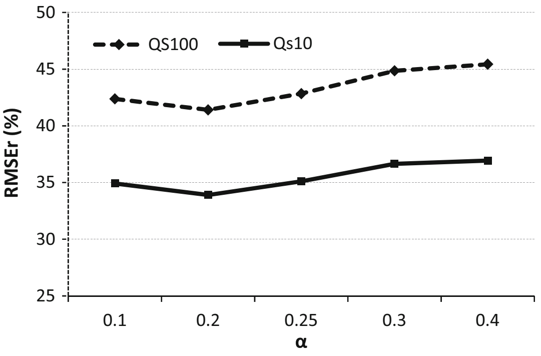


Fig. 7 RMSEr variation as a function of the α parameter for hydrological variables QS_{10} and QS_{100}—Quebec

with the iterative procedure). This may be explained by the choice of the ANN parameters in the CCA-NN method. In fact, different parameters must be fixed from the beginning to guarantee optimum solution, such as penalty parameters which are chosen in such a way to avoid over-fitting. Optimization of these parameters is performed based on the RMSEr criterion. Consequently the model loses in terms of BIASr but this latter remains in the same order of magnitude as the linear approaches.

Figure 8 presents the estimation error for flood quantiles QS_{100}, and QS_{10} using both the CCA & LR and the CCA-NN & LR models. One can observe that, overall, the CCA-NN & LR leads to smaller estimation errors than the linear model, CCA & LR. Particularly, the improvement for some sites is significant. For instance, for site 66 which has a particular location in both linear and non-linear canonical spaces, the estimation error goes from −4.13 using CCA & LR to −2.3 using CCA-NN & LR.

In the following, selected results related to Arkansas and Texas are presented. Without loss of generality, we will

Table 3 Jackknife validation results-Quebec

	Variables	CCA-NN & CLR	CCA-NN & LR	CCA and LR
NASH	QS_{100}	0.672	**0.710**	0.700
	QS_{10}	0.728	**0.793**	0.790
RMSE (m^3/s.km^2)	QS_{100}	0.114	**0.107**	0.109
	QS_{10}	0.066	**0.058**	**0.057**
RMSEr (%)	QS_{100}	42.250	**41.400**	51.030
	QS_{10}	35.696	**33.903**	44.870
BIAS (m^3/s.km^2)	QS_{100}	0.014	**0.010**	0.017
	QS_{10}	0.006	**0.002**	0.005
BIASr (%)	QS_{100}	−7.953	**−7.747**	−8.390
	QS_{10}	−6.114	**−6.026**	−7.880

Best results are shown in bold character

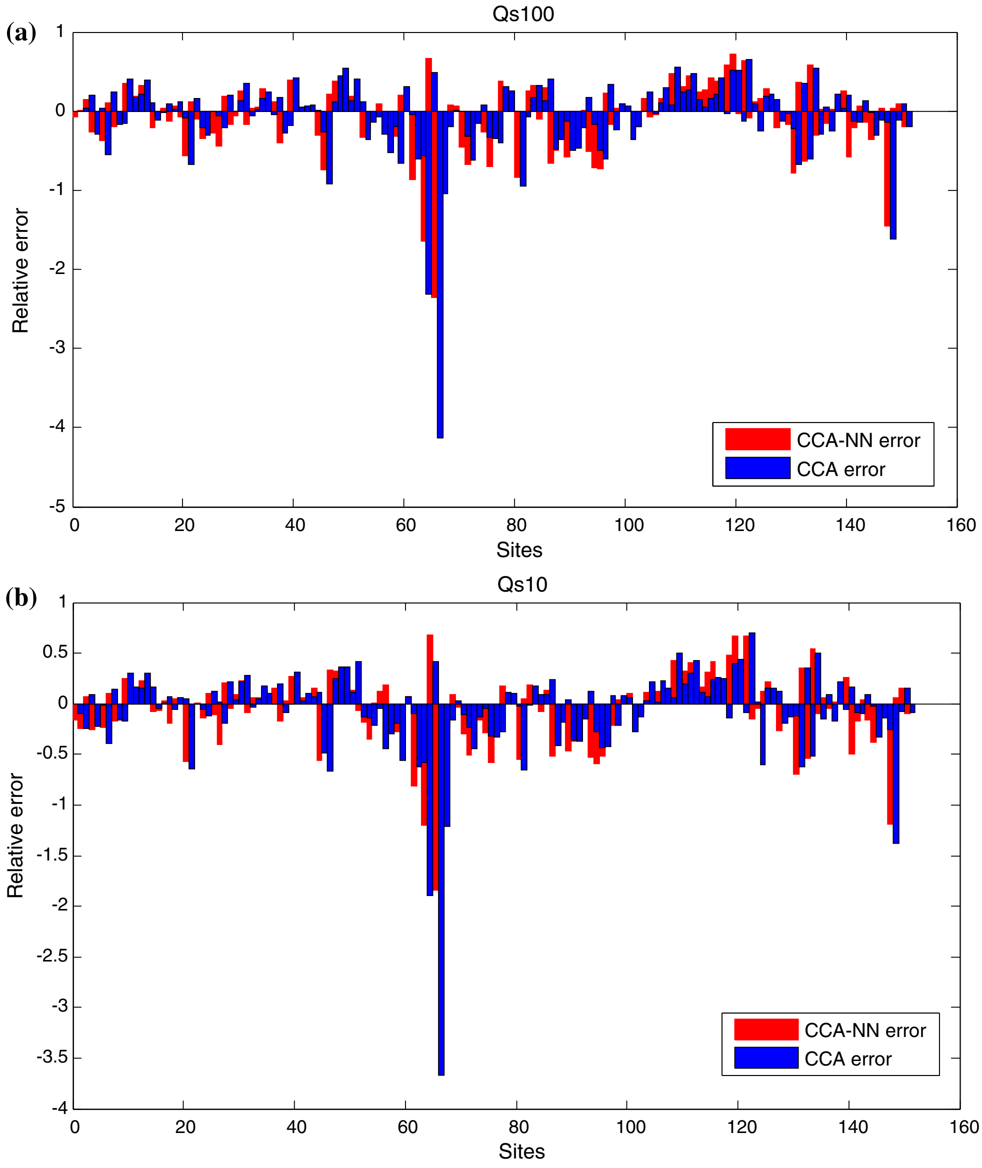


Fig. 8 Estimation error resulting from the CCA & LR and CCA-NN & LR models—Quebec

Table 4 Correlation coefficients and percentage of explained variance for CCA & LR and CCA-NN & LR relative to Arkansas and Texas

	Arkansas		Texas	
	CCA-NN & LR	CCA & LR	CCA-NN & LR	CCA & LR
Correlations				
(U_1,V_1)	0.96	0.93	0.90	0.90
(U_2,V_2)	0.45	0.37	0.61	0.50
Explained variance (%)				
U_1	39.96	46.09	40.93	42.19
V_1	79.97	65.11	61.29	62.99

Table 5 Jackknife validation results

Variables	Region					
	Arkansas (USA)			Texas (USA)		
	CCA-NN & CLR	CCA-NN & LR	CCA & LR	CCA-NN & CLR	CCA-NN & LR	CCA & LR
NASH						
QS_{50}	0.733	**0.748**	0.735	**0.552**	0.389	0.136
QS_{10}	0.732	**0.761**	0.755	**0.577**	0.499	0.351
RMSE ($m^3/s.km^2$)						
QS_{50}	2.923	**2.839**	2.913	**0.255**	0.298	0.355
QS_{10}	1.685	**1.592**	1.610	**0.119**	0.129	0.147
RMSEr (%)						
QS_{50}	**55.104**	59.308	61.360	**39.309**	50.757	54.887
QS_{10}	**46.786**	47.083	47.705	**35.599**	42.114	44.759
BIAS ($m^3/s.km^2$)						
QS_{50}	0.790	**0.610**	0.627	0.017	**0.005**	0.008
QS_{10}	0.464	**0.336**	0.337	**0.000**	0.002	0.008
BIASr (%)						
QS_{50}	**1.557**	−3.759	−5.762	−5.168	−11.225	−4.119
QS_{10}	3.390	**−1.371**	−3.046	**−5.841**	−6.261	−7.567

Best results are shown in bold character

focus on specific quantiles corresponding to return periods of 10 and 50 years.

Table 4 presents canonical correlation coefficients as well as percentages of explained variance for these two regions resulting from linear and non-linear CCA. Results indicate that, similarly to the region of Quebec, the canonical correlation coefficients are more important using a CCA-NN than using a CCA. This means that the non-linear components capture more information than the linear ones. However, as it was the case for Quebec case study, the explained variance of CCA is slightly higher than that of CCA-NN.

Table 5 summarises the results of the jackknife procedure using linear and non-linear analysis for these two regions. These results confirm the superiority of the non-linear approach. Indeed, when proceeding with CCA-NN & CLR applied to data of Arkansas, this model improves the RMSEr of QS_{10} by about 2 % over the linear model CCA-LR and about 10 % for QS_{50}. Similarly, results for the Texas region indicate that non-linear models perform better than CCA. The improvement of the RMSEr is even more important for Texas than for the Arkansas case study, with a significant improvement of BIASr.

5 Conclusions

This study has focused on the use of CCA-NN & LR methods in the context of RFA. The CCA approach has been successfully used for the delineation of homogeneous regions in RFA. However, this approach is not capable of representing the possible non-linear relationships between the variables of interest. To overcome the CCA limitations, several non-linear methods have been developed and used in other fields. CCA-

NN and CCA-K are among the most prominent and most commonly used non-linear CCA methods.

In the current work, CCA-NN is presented and adapted to the RFA context. The method is also applied to three different regions to study its robustness in dealing with the nonlinearity of hydrological processes. In order to assess the performance of this method, its results are compared to those of linear CCA. Results show that CCA-NN can be adopted to represent the non-linear behavior of hydrological process and provide a more accurate and flexible delineation of homogeneous neighborhoods leading to a better regional estimation. However, this method has a number of drawbacks similarly to other ANN-based approaches, such as the identification of optimum parameters and the selection of the transfer function. This latter requires the non-linear relationship to be empirical, i.e., dependent on the data, whereas in the current work and previous works, the hyperbolic tangent function was considered.

Acknowledgments The authors thank M.A. Ben Alaya, for his valuable help and input into the present work. The authors wish also to express their appreciation to the reviewers, the Associate Editor and the Editor in Chief for their invaluable comments and suggestions. Financial support for the present study was provided by the National Sciences and Engineering Research Council of Canada (NSERC). To get access to the data used in this study, reader may refer to the report of A. Kouider (http://espace.inrs.ca/365/1/T000342.pdf).

References

Akaho S (2001) A kernel method for canonical correlation analysis. In: Proceedings of the International Meeting of Psychometric Society (IMPS). University Convention Center-Osaka, Japan

Aziz K, Rahman A, Fang G, Shrestha S (2014) Application of artificial neural networks in regional flood frequency analysis: a case study for Australia. Stoch Environ Res Risk Assess 28(3):541–554

Ballard DH (1981) Generalizing the Hough transform to detect arbitrary shapes. Pattern Recognit 13(2):111–122

Barnett TP, Preinsendorfer R (1987) Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. Mon Weather Rev 115:1825–1850

Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, Clarendon

Bolton RJ, Hand DJ, Webb AR (2003) Projection techniques for nonlinear principal component analysis. Stat Comput 13(3): 267–276

Botelho SSDC, Bem RAD, ÍLD Almeida, Mata MM (2005) C-nlpca: extracting non-linear principal components of image datasets. Anais XII Simposio Brasileiro de Sensoriamento Remoto, Goiania, Brasil, pp 3495–3502

Campi C, Parkkonen L, Hari R, Hyvärinen A (2013) Non-linear canonical correlation for joint analysis of MEG signals from two subjects. Front Neurosci 7:107

Cannon AJ (2008) Multivariate statistical models for seasonal climate prediction and climate downscaling. Atmospheric Science, University of British Columbia. Doctor of philosophy, p 141

Chebana F, Ouarda T (2007) Multivariate L-moment homogeneity test. Water Resour Res 43(8)

Chebana F, Ouarda TBMJ (2008) Depth and homogeneity in regional flood frequency analysis. Water Resour Res 44(11)

Chebana F, C Charron, TBMJ Ouarda, B Martel (2014) Regional frequency analysis at ungauged sites with the generalized additive model. In: press J Hydrometeorol

Chen C-S, Liu C-H, Su H-C (2008) A nonlinear time series analysis using two-stage genetic algorithms for streamflow forecasting. Hydrol Process 22:3697–3711

Chen L, Singh VP, Guo S, Zhou J, Ye L (2014) Copula entropy coupled with artificial neural network for rainfall–runoff simulation. Stoch Environ Res Risk Assess 28(7):1755–1767

Chokmani K, Ouarda TBMJ (2004) Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resour Res 40(12)

Dauxois J, Nkiet GM (1998) Nonlinear canonical analysis and independence tests. Ann Stat 26(4):1254–1278

Dawson C, Wilby R (2001) Hydrological modelling using artificial neural networks. Prog Phys Geogr 25(1):80–108

Duch W, Jankowski N (1999) Survey of neural transfer functions. Neural Comput Surv 2(1):163–212

Frie KG, Janssen C (2009) Social inequality, lifestyles and health–a non-linear canonical correlation analysis based on the approach of Pierre Bourdieu. Int J Public Health 54(4):213–221

Gifi A (1990) Nonlinear multivariate analysis. Wiley, Chichester, p 579

Guillemette N, St-Hilaire A, Ouarda TB, Bergeron N, Robichaud É, Bilodeau L (2009) Feasibility study of a geostatistical modelling of monthly maximum stream temperatures in a multivariate space. J Hydrol 364(1):1–12

Hardoon DR, Shawe-Taylor J (2009) Convergence analysis of kernel canonical correlation analysis: theory and practice. Mach Learn 74:23–38

Hsieh WW (2000) Nonlinear canonical correlation analysis by neural networks. Neural Netw 13:1095–1105

Hsieh WW (2001) Nonlinear canonical correlation analysis of the tropical Pacific climate variability using a neural network approach. J Clim 14:2528–2539

Khalil B, Ouarda T, St-Hilaire A (2011) Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. J Hydrol 405(3):277–287

Kramer MA (1991) Nonlinear principal component analysis using autoassociative neural networks. Am Inst Chem Eng J 37(2):233–243

Kruger U, Sharma SK, Irwin GW (2004) Improved nonlinear canonical correlation analysis using genetic strategies. UKACC control. University of Bath, Bath

Li C, Tang H, Ge Y, Hu X, Wang L (2014) Application of back-propagation neural network on bank destruction forecasting for accumulative landslides in the three Gorges Reservoir Region, China. Stoch Environ Res Risk Assess 28(6):1465–1477

Malthouse EC (1998) Limitations of nonlinear PCA as performed with generic neural networks. IEEE Trans Neural Netw 9(1):165–173

Michael AG, Raymond CP (2003) Using traffic conviction correlates to identify high accident-risk drivers. Accid Anal Prev 35(6):903–912

Monahan AH (2000) Nonlinear principal component analysis by neural networks: theory and application to the Lorenz system. J Clim 13:821–835

Nagai I (2013) Optimization using cross-validation for penalized nonlinear canonical correlation analysis. Graduate School of Science and Technology, Kwansei Gakuin University 2-1 Gakuen, Sanda, pp 669–1337

Ouarda TBMJ (2013) Hydrological frequency analysis, regional. Encycl Environ. doi:10.1002/9780470057339.vnn9780470057043

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

Ouarda TBMJ, Girard C, Cavadias GS, Bobée B (2001) Regional flood frequency estimation with canonical correlation analysis. J Hydrol 254(1–4):157–173

Ouarda T, Bâ K, Diaz-Delgado C, Cârsteanu A, Chokmani K, Gingras H, Quentin E, Trujillo E, Bobée B (2008a) Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. J Hydrol 348(1):40–58

Ouarda TBMJ, St-Hilaire A, Bobée B (2008b) Synthèse des développements récents en analyse régionale des extrêmes hydrologiques. Revue des sciences de l'eau 21(2):219–232

Pandey G, Nguyen V-T-V (1999) A comparative study of regression based methods in regional flood frequency analysis. J Hydrol 225(1):92–101

Riad S, Mania J (2004) Rainfall-runoff model using an artificial neural network approach. Math Comput Model 40:839–846

Rumelhart DE, GE Hinton, RJ Williams (1985) Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, P. R. Group, 1: 318–362

Sengupta S, Boyle J (1995) Non-linear principal component analysis of climate data. PCMDI

Shu C, Ouarda TBMJ (2007) Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resour Res 43(07)

Tasker GD, Hodge SA, Barks CS (1996) Region of influence regression for estimating the 50-year flood at ungauged sites. Water Resour Res 1(32):163–170

Tishlert A, Lipovetsky S (1996) Canonical correlation analyses for three data sets: a unified framework with application to management. Comput Oper Res 23(7):667–679

Van Den Wollenberg AL (1977) Redundancy analysis an alternative for canonical correlation analysis. Psychometrika 42(2):207–219

Wang D, Shi L, Yeung DS, Tsang E (2005) Nonlinear canonical correlation analysis of fMRI signals using HDR models. 27th Annual international conference of the IEEE engineering in medicine and biology society, 2005. IEEE-EMBS 2005

Wazneh H, Chebana F, Ouarda T (2013) Optimal depth-based regional frequency analysis. Hydrol Earth Syst Sci 17(6): 2281–2296

Wu A, Hsieh WW (2002) Nonlinear canonical correlation analysis of the tropical Pacific wind stress and sea surface temperature. Clim Dyn 19:713–722

Xu J, Li W, Ji M, Lu F, Dong S (2010) A comprehensive approach to characterization of the nonlinearity of runoff in the headwaters of the Tarim River, Western China. Hydrol Process 24:136–146

Yin H (2007) Nonlinear dimensionality reduction and data visualization: a review. Int J Autom Comput 4(3):294–303

Yonaba H, Anctil F, Fortin V (2010) Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting. J Hydrol Eng 15(4):275–283

Zaier I, Shu C, Ouarda T, Seidou O, Chebana F (2010) Estimation of ice thickness on lakes using artificial neural network ensembles. J Hydrol 383(3):330–340

Zhihua J, Zhen Y (2010) On using non-linear canonical correlation analysis for voice conversion based on Gaussian mixture model. J Electron 27(1):1–7