Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space

C. Shu[1] and T. B. M. J. Ouarda[1]

Received 2 May 2006; revised 9 February 2007; accepted 21 February 2007; published 24 July 2007.

[1] Models based on canonical correlation analysis (CCA) and artificial neural networks (ANNs) are developed to obtain improved flood quantile estimates at ungauged sites. CCA is used to form a canonical physiographic space using the site characteristics from gauged sites. Then ANN models are applied to identify the functional relationships between flood quantiles and the physiographic variables in the CCA space. Two ANN models, the single ANN model and the ensemble ANN model, are developed. The proposed approaches are applied to 151 catchments in the province of Quebec, Canada. Two evaluation procedures, the jackknife validation procedure and the split sample validation procedure, are used to evaluate the performance of the proposed models. Results of the proposed models are compared with the original CCA model, the canonical kriging model, and the original ANN models. The results indicate that the CCA-based ANN models provide superior estimation than the original ANN models. The ANN ensemble approaches provide better generalization ability than the single ANN models. The CCA-based ensemble ANN model has the best performance among all models in terms of prediction accuracy.

Citation: Shu, C., and T. B. M. J. Ouarda (2007), Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space, *Water Resour. Res.*, *43*, W07438, doi:10.1029/2006WR005142.

1. Introduction

[2] Regional flood frequency analysis has been widely used to improve flood quantile estimation at catchments where streamflow records are either short or not available. A number of regionalization techniques have been developed for this purpose [see, e.g., *Wiltshire*, 1986; *Burn*, 1990a, 1990b; *Cavadias*, 1990; *Zrinji and Burn*, 1994; *Castellarin et al.*, 2001; *Ouarda et al.*, 2001; *Chokmani and Ouarda*, 2004; *Shu and Burn*, 2004]. *Groupe de Recherche en Hydrologie Statistique* (*GREHYS*) [1996a, 1996b] provided an extensive review and comparative evaluation of different regionalization techniques.

[3] Identification of homogeneous regions is one of the major steps in regional flood frequency analysis. The purpose of this step is to select a group of sites that are hydrologically similar to the target site. Traditionally, homogeneous regions are formed on the basis of geographic or administrative boundaries [*Matalas et al.*, 1975; *Beable and McKerchar*, 1982]. However, this approach is not hydrologically sound, since regions formed using this approach are seldom homogeneous in terms of their hydrologic response [*Cunnane*, 1988].

[4] ''Site focused'' regionalization techniques, where each site has a potential unique set of catchments forming the homogeneous region for the site, has received much attention because of its flexibility and effectiveness. The region of influence (ROI) method [*Burn*, 1990a, 1990b] laid the foundation for this technique. The site focused approach is also known as the hydrological neighborhood approach [*GREHYS*, 1996a; *Ouarda et al.*, 2000, 2001]. The comparison studies by *GREHYS* [1996b] concluded that the neighborhood approach has superior performance than the fixed region approach.

[5] Canonical correlation analysis (CCA) [*Ouarda et al.*, 2000, 2001] is also a frequently used approach to define hydrological neighborhoods. CCA was introduced by *Cavadias* [1990] to flood quantile estimation where the regions are formed on the basis of visual judgment of clustering patterns. *Ouarda et al.* [2000] applied the CCA approach to estimate extreme flood quantiles in Quebec, Canada. *Ouarda et al.* [2001] presented additional improvements to the method and proposed the detailed algorithms to delineate homogeneous regions for gauged and ungauged sites using CCA.

[6] *Chokmani and Ouarda* [2004] presented a CCA-based kriging approach, named canonical kriging, for flood quantile estimation at ungauged sites. CCA is introduced by the authors to construct a projected physiographical space. Ordinary kriging is then used for the interpolation of flood quantiles over the physiographical space defined by CCA. The application of the method to data from the province of Quebec, Canada showed that canonical kriging can provide comparable results to the traditional CCA-based flood estimation method. The physiographic space defined using the CCA method is more feasible to provide hydrological variable estimation than using other methods, such as principle component analysis (PCA).

[7] Different quantile estimation methods can be used with the CCA approach [*GREHYS*, 1996a, 1996b; *Ouarda et al.*, 2001]. Regional regression is frequently integrated

[1]Eau, Terre, et Environnement, Institut National de la Recherche Scientifique, University of Quebec, Quebec, Quebec, Canada.

0043-1397/07/2006WR005142$09.00

with the CCA approach to provide quantile estimation, especially at ungauged sites, from site physiographic characteristics. The frequently used regional regression model has the following generalized form [*Thomas and Benson*, 1970]

$$Q_T = a x_1^{\theta_1} x_2^{\theta_3} \cdots x_i^{\theta_i} \cdots x_n^{\theta_n} \quad (1)$$

where Q_T is the flood quantile at the site of interest; x_i is the ith site characteristic used for flood quantile estimation; θ_i is the ith model parameter which needs to be estimated using statistical analysis; n is the total number of site characteristics used in the model; and a is the multiplicative error term. A log transformation is frequently used to estimate the parameters of equation (1). The solution obtained by linear regression methods is theoretically unbiased in the logarithmic domain, but is biased in the real flood flow domain [*McCuen et al.*, 1990]. *Pandey and Nguyen* [1999] and *Grover et al.* [2002] compared a wide range of regression techniques applied to regional flood frequency analysis, and the results indicated that nonlinear regression methods directly solving equation (1) can provide more precise estimates than linear regression techniques.

[8] As an alternative to standard nonlinear regression methods, artificial neural networks (ANNs) and ANN ensemble models are introduced by *Shu and Burn* [2004] for index flood and flood quantile estimation. *Seidou et al.* [2006] applied ANNs to the regional estimation of lake ice thickness in ungauged sites. ANNs are nonparametric approaches which require no assumptions about the form of the true underlying function being estimated. The application to selected catchments in the United Kingdom (UK) indicates that the nonlinearity introduced by ANN models allows them to outperform multiple linear regression methods. The generalization ability of a single ANN can be improved by using a properly designed ANN ensemble. *Dawson et al.* [2006] applied ANNs to flood quantile and index flood estimation for 870 catchments across the UK. The results obtained from the ANNs are comparable in accuracy with those obtained by the Flood Estimation Handbook (FEH) [*Reed and Robson*, 1999] models.

[9] In the present paper, regional flood quantile estimation methods based on CCA and ANN are proposed. CCA is used to define a transformed physiographical space. An ANN is then used to establish the nonlinear relationships between the site physiographical variables in the CCA space and hydrological variables to be estimated. To improve the generalization ability of a single ANN, the ANN ensemble technique is used. Since only physiographical and climatic data are required as input to the ANN models, the proposed approaches are feasible for flood estimation at ungauged sites. A comparison study is carried out between the proposed approaches and several other approaches using data from the province of Quebec, Canada.

[10] The remainder of this paper is organized as follows. In section 2, a general introduction to CCA and ANNs is provided and the methodology for integrating the two techniques for flood quantile estimation is presented. In section 3, the details for designing the ANNs, the estimation models to be compared, and the evaluation methodology are presented. In section 4, a description of the study area is provided. In section 5, the results obtained by applying the proposed approaches are presented and discussed. Finally, in section 6, the conclusions of this work and recommendations for further research are presented.

2. ANN Models in the CCA Physiographical Space

[11] In the proposed approach, site characteristics including physiographical and climatic data are projected in the canonical space. The canonical variables in the physiographical space are then fed to ANN models to generate flood quantile estimates. CCA preserves the character of the original data by omitting nonessential data [*Razavi et al.*, 2005]. Models built upon the data processed using the CCA analysis could lead to better generalization ability. *Chokmani and Ouarda* [2004] compared two dimensional reduction techniques, PCA and CCA, and the results indicate that CCA leads to a much better performance than PCA. A brief description of the CCA and ANN techniques is provided in sections 2.1 and 2.2, respectively. The methodology of integration of the two techniques for regional flood frequency analysis at ungauged sites is provided in section 2.3.

2.1. Canonical Correlation Analysis

[12] Canonical correlation analysis (CCA) is a way of explaining the linear relationship between two sets of variables. Consider X and Y are two random variables, CCA computes two sets of basis vectors (canonical variables), one for X and the other for Y, such that the correlations between the projections of the variables onto these basis vectors are mutually maximized [*Muirhead*, 1982]. The maximum number of canonical variable pairs is equal to or less than the smallest dimensionality of the two variables.

[13] Let W and V be linear combinations of X and Y, respectively,

$$W = \alpha' X \quad (2)$$

$$V = \beta' Y \quad (3)$$

Let Σ be the covariance matrix of variables X and Y, defined as

$$\sum = \mathrm{cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{bmatrix} \sum_X & \sum_{XY} \\ \sum_{YX} & \sum_Y \end{bmatrix} \quad (4)$$

The correlation between W and V can then be calculated as

$$\rho = \frac{\alpha' \sum_{XY} \beta}{\sqrt{\alpha' \sum_X \alpha \beta' \sum_Y \beta}} \quad (5)$$

The goal of CCA is to find the vectors of α and β maximizing ρ subject to the constraint that W and V must have unit variances. Once the first pair of canonical variables is obtained, other pairs of canonical variables can be obtained in the uncorrelated directions to the previous ones by maximizing equation (5) subject to the constraint of unit variance.

[14] CCA was recently used by *Chokmani and Ouarda* [2004] to construct a transformed space defined by the

physiographical and meteorological characteristics. The hydrological variables (flood quantiles in our case) are generally not continuous in the geographical space. However, they are continuous in the canonical physiographical space [*Chokmani and Ouarda*, 2004]. This characteristic is crucial for flood estimation at ungauged sites. Because the physiographic variables and the meteorological variables are generally available at the ungauged sites, one can easily locate an ungauged site in the physiographical space constructed by these variables. For more detailed information regarding CCA, the readers are referred to *Ouarda et al.* [2001].

2.2. ANN and ANN Ensemble

[15] An ANN is an information processing system which is designed to mimic certain structures and functions of biological neural networks of the human brain. Given sufficient parameters, an ANN can be used for creating nonlinear mathematical models for universal approximation. This extraordinary capability has enabled ANNs to solve large complex problems such as pattern recognition, nonlinear modeling, classification, and control.

[16] Multilayer perceptrons (MLPs) represent the most commonly used and well researched class of ANNs, originally because of work by *Rumelhart and McClelland* [1986]. This type of ANN implements a feed forward supervised paradigm. A MLP consists of an input layer, one or more hidden layers, and an output layer. The input layer receives values of the input variables for a given problem. The output layer provides the ANN prediction and represents model output. Layers lying between the input and output layer are called hidden layers. Nodes in each layer are interconnected through weighted acyclic arcs from each preceding layer to the following, without lateral or feedback connections.

[17] To improve the generalization ability and stability of a single ANN, an ANN ensemble can be used. To construct an ANN ensemble, a number of ANNs are trained to tackle a given problem, and the results produced by these individual networks are combined to generate a unique output. Each network in an ensemble is first trained using the training instances. Then, for each example, the predicted output of each of these networks is combined to produce the output of the ensemble.

[18] Using ensemble ANN to improve model generalization performance is an active research topic [*Dietterich*, 1997]. For many real world problems, ensemble models can outperform the best base models. There have been some successful applications of ANN ensemble models in hydrology. *Cannon and Whitfield* [2002] used a bootstrap aggregated ANN ensemble to predict changes in streamflow conditions. Results showed that the prediction obtained by the ANN ensemble model was better than a stepwise linear regression model. Furthermore, by adopting the ensemble approach, some commonly encountered problems when applying ANNs in hydrology can be easily solved. *Shu and Burn* [2004] introduced the ANN ensemble methods to estimate the index flood and flood quantile at ungauged sites. *Shu and Burn* [2004] evaluated three methods (randomization, bagging and boosting) for generating the member networks and two methods (averaging and stacking) for integrating the member networks. The results showed that properly designed ANN ensemble models can significantly reduce prediction error when compared with parametric regression methods. *Anctil and Lauzon* [2004] compared five ANN generalization approaches for streamflow prediction: stop training, Bayesian regularization, stacking, bagging and boosting. The application to six selected catchments indicated that the performance of standard ANNs can be improved by using any of the generalization approaches. The ANN ensemble methods of stacking, bagging and boosting provided better improvement than the other two generalization approaches.

[19] The task of using ANN ensembles to model a given problem can be broken down into the following two questions: (1) how to generate the component ANN constructing the ensemble? and (2) how to combine the multiple outputs from the component networks to generate a unique output? [*Merz*, 1998]. To benefit from the ensemble approach, member networks in an ensemble should have diverse generalization ability. A number of methods have been proposed for this purpose. The frequently used methods for generating ensemble ANNs include manipulating the set of initial random weights, using different network topology, training component networks using different training algorithms, and manipulating the training set [*Sharkey*, 1999]. The methods of manipulating the training set using bagging [*Breiman*, 1996] and boosting [*Schapire*, 1990; *Freund and Schapire*, 1996] have been most frequently used. Many approaches have been proposed for integrating the multiple outputs from the component networks [*Sharkey*, 1999; *Ahmad and Zhang*, 2002]. The two frequently used methods are averaging and stacked generalization [*Wolpert*, 1992].

[20] The bagging procedure is selected in this paper to generate the individual member networks. Simple averaging is selected in this paper to combine the outputs from each individual ANN. This method is a simple and effective way to generate ensemble output [*Shu and Burn*, 2004].

2.3. Integrating CCA and ANNs for Regional Flood Frequency Analysis at Ungauged Sites

[21] For ungauged sites, no historical flood records are available to directly estimate the hydrological variables such as flood quantiles. However, by establishing a functional relationship between the physiographical variables and the hydrological variables, the hydrological variables can be indirectly estimated. The model used for the estimation is usually calibrated using data from the gauged sites. In the approaches proposed in the present research work, the physiographical variables are projected into the canonical space, and the projected variables are then fed to the ANN models to generate estimates of the hydrological variables.

[22] Suppose a set of physiographic and climatic variables, X, and hydrological variables, Y, are associated with each gauged site. Using CCA, canonical variables W and V can be obtained as a linear combination of X and Y, respectively. The coefficients used for the combination are computed so that the correlation between the variables W and V is maximized. Knowing the combination coefficients, the physiographical variable X_u for an ungauged site can be easily projected into the CCA space to obtain the physiographical variable W_u in the CCA space.

[23] The goal of the ANN model is to approximate the functional relationship between the canonical variables W

and the hydrologic variables Y which act as the input and output of an ANN, respectively. The canonical variables V are not used in the ANN training and estimation phase. To achieve this goal, the ANN must be trained using the samples from the gauged sites in the study a.rea. During the training process, network parameters must be updated so as to minimize the estimation error made by the network. The error of a particular configuration of the network can be determined by running all the training cases through the network and comparing the actual generated output with the desired or target outputs. The differences are combined together by an error function to give the network error. Several learning algorithms exist for determining the network parameters. The most well known is the back propagation algorithm [see *Haykin*, 1994; *Fausett*, 1994]. It uses gradient descent techniques to minimize the network error function. There are also other training algorithms which use techniques for nonlinear function optimization. These methods include the conjugate gradient algorithm, the quasi-Newton algorithm and the Levenberg-Marquardt algorithm [see *Bishop*, 1995]. They are collectively known as second-order training algorithms.

[24] After an ANN model is trained using data from gauged sites, obtaining the estimation of hydrological variables for an ungauged site is straightforward. Applying the projected physiographic data to the ANN input layer, the estimation can be obtained directly from the output layer.

[25] The approach described above uses a single ANN (SANN) to estimate flood quantiles at ungauged sites in the CCA physiographic space. The abbreviation SANN-CCA will be used in the remainder of the paper to represent this model.

[26] To improve the generalization ability and the stability of the single ANN, the ANN ensemble model is used. Component networks in the ANN ensemble are generated using the bagging approach and the resulting networks are combined using simple averaging. Bagging stands for bootstrap aggregation. The bagging approach was developed by *Breiman* [1996] to improve the accuracy of predictions in classification and regression problems. The algorithm is based on the bootstrap resampling technique [*Efron and Tibshirani*, 1993]. Bagging can be implemented in parallel, and the method is easy to use and has been shown to effectively improve the generalization ability of the single network [*Cannon and Whitfield*, 2002; *Shu and Burn*, 2004; *Anctil and Lauzon*, 2004]. Each member ANN of the ensemble is trained by only a subset of the training set. The subset is drawn from the original training set T with replacement using bootstrap sampling. Training instances in the training set have equal chance of being drawn. The number of training instances in the subset is the same as the training set. Thus some data in the training set appear more than once in the subset, and the probability an individual training sample from T will not be part of a bootstrap resampled training set is $(1 - 1/\mathrm{N})^N \approx 0.37$, where N is the number of training samples in T. Suppose the process is repeated K times, and each time an ANN is trained on the basis of the training subset. Then, K member networks can be generated with each network trained with a different random sampling of the original training set. After all the member networks are generated, a unique output for the ensemble can be derived by averaging the outputs from member networks. Suppose, for the site i, that the predicted flood quantile using the kth member ANN is $\hat{q}_i^k (k = 1, \ldots K)$. The ensemble output can be calculated using

$$\hat{q}_i = \frac{1}{K} \sum_{k=} \hat{q}_i^k \tag{6}$$

[27] The approach described above uses an ANN ensemble to estimate flood quantiles at ungauged sites in the CCA physiographic space. The abbreviation EANN-CCA will be used in the remainder of this work to represent this model.

3. Methodology

3.1. ANN Model Structure

[28] For the SANN-CCA approach, a MLP having one output layer, one hidden layer and one input layer is used. The system inputs are the canonical variables in the physiographical space derived using CCA. The outputs of the system are the specific quantiles. The tan-sigmoid transfer function is used for neurons in the hidden layers. The use of the nonlinear transfer function extends the nonlinear approximation ability of the ANN. A linear transfer function is used for the output layer. A linear transfer function for the output neuron has the advantage of a potentially unbounded output [*Shu and Burn*, 2004].

[29] Determining the number of hidden neurons in the hidden layer is an important task when designing an ANN. Too many hidden neurons may lead to the problem of overfitting which is caused by not having enough training cases to adequately train all the neurons in an ANN. Too few neurons in the hidden layer may cause the problem of underfitting which occurs when an ANN does not have sufficient complexity to fully represent the functional relationship between the system input and output. Thus some compromise must be made between too many and too few neurons in the hidden layer. As a rule of thumb, the number of hidden neurons should be less than twice the input layer size. *Shu and Burn* [2004] showed that a MLP with five hidden neurons in the hidden layer provided sufficient generalization ability when it is applied to provide flood estimation from catchment characteristics. In this paper, a sensitivity analysis is carried out to identify the optimal number of hidden neurons. By varying the number of hidden neurons from three to eight, ANNs with five hidden neurons are identified to provide most accurate estimation when they are applied to estimate the selected specific quantiles. Five hidden neurons are finally used in the hidden layer. The training algorithm selected in this work is the Levenberg-Marquardt (LM) algorithm [*Hagan and Menhaj*, 1994]. This algorithm is much faster than the gradient descent method to find optimal solutions for various problems. The scalar parameter μ is required to implement the algorithm [*Demuth and Beale*, 2003]. When the value of μ is large, the LM algorithm behaves as a gradient descent method with a small step size. However, when the value of μ is small, the optimization follows Gauss-Newton method which is faster and more accurate near an error minimum. An initial value must be set for μ, and it is given as 0.005. The value of μ changes during the ANN training process on the basis of the performance function of the ANN. If a training epoch decreases the performance function, the

value of μ is multiplied by $\mu_d = 0.1$. If a training epoch increases the performance function, the value of μ is multiplied by $\mu_i = 10$. A maximum value of $\mu_{\max} = 1 \times 10^6$ is set for μ to stop the training algorithm. During the ANN training, the transfer functions of the hidden neurons operate increasingly in nonlinear parts of the sigmoid functions which enables the network to produce more and more nonlinear mapping. In the same time, the number of the effective parameters and number of degrees of freedom in the network also increase which could lead to the problem of overtraining. To avoid the overtraining problems in ANN, one of the two effective approaches, regularization and early stopping, can be generally applied [*Bishop*, 1995]. However, two problems are triggered if early stopping is used [*Shu and Burn*, 2004]. First, a validation set needs to be extracted from the training set, which may lead to insufficient data being available to successfully train an ANN. Secondly, how to optimally separate the validation set still remains a major challenge. The regularization technique, which is free of these problems, is selected in this paper. In the regularization algorithm, the error function which is minimized during the training phase is augmented with additional terms that penalize the complexity of the model. *Shu and Burn* [2004] provided the background information related to the implementation of the regularization technique.

[30] For the EANN-CAA approach, each component ANN uses the same configuration as the SANN-CAA model; however it is trained on bootstrap sampled data. The identification of the size of an ensemble is important. If the size is too small, the improvement in generalization is not apparent; if the size is too large, it will increase the training time and the effort to establish the ensemble. Previous studies by *Hansen and Salamon* [1990] and *Agrafiotis et al.* [2002] suggested that using ten networks can achieve significant reduction in classification error. Experiments conducted by *Opitz and Maclin* [1999] showed that, when the ensemble size increases to ten or fifteen, the generalization ability of the ensemble can be noticeably improved. Recent studies by *Shu and Burn* [2004] suggest that a network size of ten is necessary to attain sufficient generalization ability. The authors also found out that a network size of fourteen achieved best results when applied to the United Kingdom data. Different ensemble sizes ranging from two to twenty are applied to the study area, and results indicate that estimation error gradually decreases when the ensemble size increases to eleven, while with further increase of the ensemble size, very little change in the estimation error can be observed. Beyond a size of 14, virtually no improvement in the estimation is observed. An ensemble size of 14 is used in this paper.

3.2. Selection of Methods for the Comparison

[31] The SANN-CCA model and EANN-CAA model described in section 2 are used to estimate the 10, 50 and 100 year flood quantiles at the study area catchments. To evaluate the relative performance of these two models, they are compared to the following four models:

[32] 1. The traditional CCA model (Tradition-CCA) [*Ouarda et al.*, 2001]. On the basis of CCA analysis, the optimal hydrological neighborhood for each individual site is determined. Multiple regression is used for regional flood estimation.

[33] 2. The canonical Kriging model (Kriging-CCA) [*Chokmani and Ouarda*, 2004]. The CCA method is used to define the physiographical space, and the geostatistical method of ordinary kriging is used to obtain regional flood estimates by interpolating the flood quantile over the canonical physiographical space. The method was shown to produce flood estimates as precise as the traditional CCA model; however the computation is less complicated [*Chokmani and Ouarda*, 2004].

[34] 3. The original single ANN model (SANN-Origin) [*Shu and Burn*, 2004]. An ANN model is used to directly establish the relationship between site characteristics and the flood quantile of interest. As opposed to the implementation of SANN-CCA approach, the physiographical and meteorological variables are not projected into the CCA space, but are directly fed to the inputs of an ANN.

[35] 4. The original ensemble neural network model (EANN-Origin) [*Shu and Burn*, 2004]. In this approach, an ensemble ANN model is used to improve the generalization ability of the SANN-Origin model. The component ANNs composing the ANN ensemble are created using the bagging approach. The ensemble output is generated by combining the outputs from the individual networks using simple averaging.

3.3. Evaluation Criteria

[36] Each regional flood frequency analysis model is assessed using the following five indices: the Nash criterion (*NASH*), the root mean squared error (*RMSE*), the relative root mean squared error (*RMSEr*), the mean bias (*BIAS*), and the relative mean bias (*BIASr*). The indices are computed according to the following equations:

$$NASH = 1 - \frac{\sum_{i=1}^{n} (q_i - \hat{q}_i)^2}{\sum_{i=1}^{n} (q_i - \bar{q})^2} \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (q_i - \hat{q}_i)^2} \tag{8}$$

$$RMSEr = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{q_i - \hat{q}_i}{q_i} \right)^2} \tag{9}$$

$$BIAS = \frac{1}{n} \sum_{i=1}^{n} (q_i - \hat{q}_i) \tag{10}$$

$$BIASr = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{q_i - \hat{q}_i}{q_i} \right) \tag{11}$$

where n is the total number of sites being modeled, q_i is the at-site estimation for site i, $\hat{q}_i$ is the estimation obtained

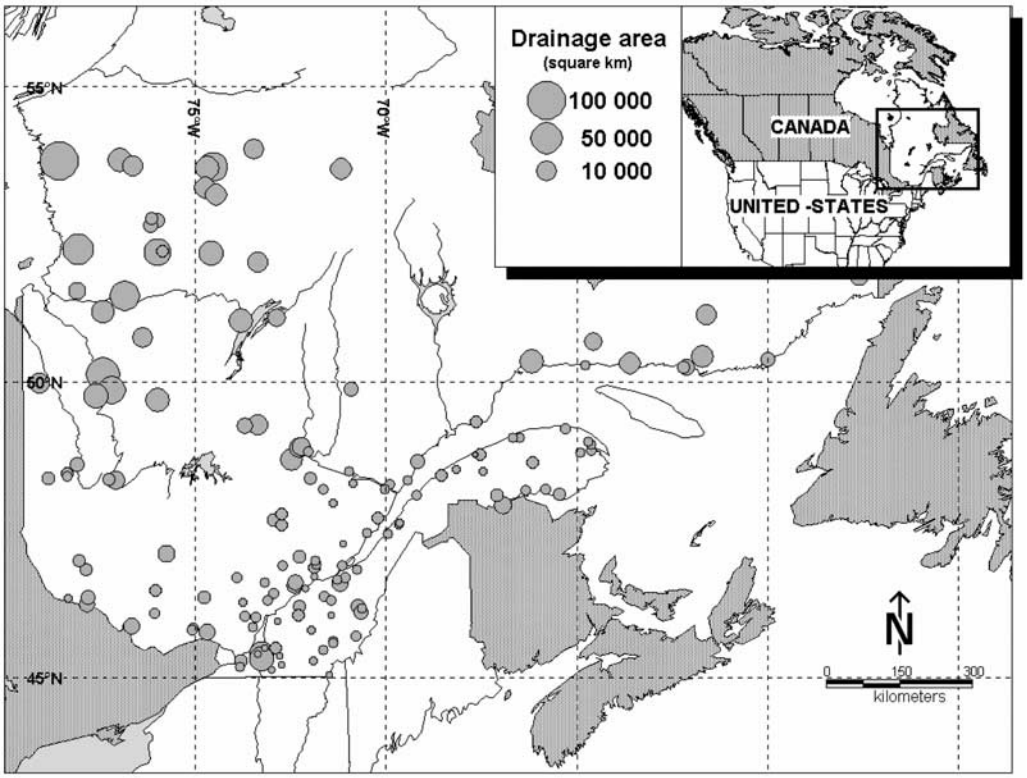

Figure 1. Hydrometric stations across the province of Quebec, Canada.

from the regional flood frequency model for site i, and $\bar{q}$ is the mean of at-site estimation of the n sites.

3.4. Evaluation Procedure

[37] The jackknife resampling procedure is used to compare the relative performances of the regional flood estimation methods. In the jackknife resampling procedure, the flood records of each catchment in the study area are temporarily removed from the database, thus the site is assumed to be ''ungauged''. Then each regional flood frequency analysis model is calibrated using the data of the remaining sites. Regional estimates can be obtained for the ''ungauged site'' using the calibrated models, and they are evaluated against its at-site estimates.

4. Case Study

[38] The hydrometric station network of southern Quebec, Canada is chosen as case study for this work. According to the following three criteria, 151 hydrometric stations managed by the ministry of the environment of Quebec (MENVIQ) services are selected:

[39] 1. To get reliable at-site estimation, a historical flood record of 15 years or longer is required.

[40] 2. The gauged river should present a natural flow regime.

[41] 3. The historical data at the gauging stations must pass the tests of homogeneity, stationarity and independence.

[42] The selected stations are located between 45°N and 55°N. The area of these catchments ranges from 200 km^2 to 100000 km^2. The locations of these hydrometric stations are shown in Figure 1. Figure 2 illustrates the distribution of the number of years of observations for the stations of the case study.

[43] Three types of data, physiographical, meteorological, and hydrological are used in this study. The physiographical and hydrological data were extracted from the MENVIQ hydrological database and from the topographic digital maps of Quebec. Meteorological variables were obtained using interpolated historical data from the MENVIQ meteorological network across the province of Quebec.

[44] Five variables including three physiographical variables and two meteorological variables are selected in this work on the basis of the previous study by *Chokmani and Ouarda* [2004]. The three physiographical variables are basin area (AREA), mean basin slope (MBS) and the fraction of the basin area covered with lakes (FAL). The two meteorological variables are annual mean total precipitations (AMP) and annual mean degree days over 0°C (AMD). The summary statistics of these variables are presented in Table 1.

[45] At-site flood quantile estimates for all the gauging stations in the study area were extracted from the database compiled by *Kouider et al.* [2002]. The flood quantile estimates for each site were computed by fitting the most appropriate statistical distribution to the historical flood record. Scale effects may have a negative impact on modeling the underlying physical mechanism of drainage systems and should be eliminated from experiment data [*Eaton et al.*, 2002]. Thus specific quantiles (flood quantiles standardized by basin area) are used to minimize the scale effect. Three typical specific flood quantiles, the 10-year (q10), the 50-year (q50), and the 100-year (q100) specific quantiles are selected for this study.

[46] The scatterplots between the specific quantiles and the selected physiographical and meteorological variables are shown in Figure 3. From Figure 3, we can observe that

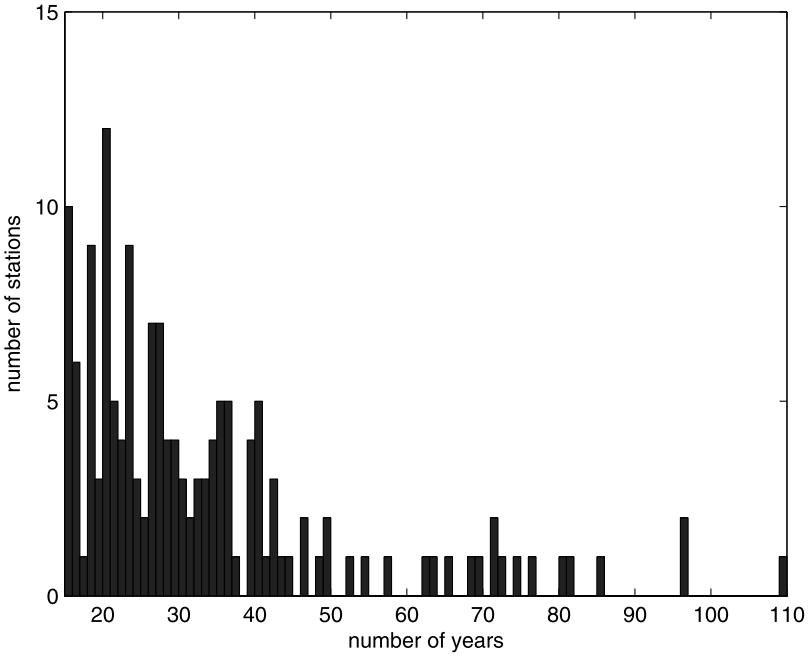

Figure 2. Length of data series.

the catchment descriptors MBS and AMP are positively correlated with the specific quantiles; while the catchment descriptors AREA, FAL and AMD are negatively correlated with the specific quantiles. CCA requires all variables be transformed for normality and standardized. Significant asymmetry exists in the physiographical, meteorological and hydrological variables in the study region [*Chokmani and Ouarda*, 2004]. Thus a logarithmic transformation is used for the variables, q10, q50, q100, AREA, MBS, AMP and AMD, and a root transformation is used for FAL. All variables were also standardized prior to CCA.

5. Results and Discussion

[47] The two approaches proposed in this paper and the four models used for comparison purposes are applied to the study area database. The results obtained using the jackknife validation procedure are presented in Table 2. For each cell of Table 2, bold font denotes the best performing approach.

[48] A model can be claimed to produce perfect estimation if the *NASH* criterion is equal to 1. Normally a model can be considered as accurate if the *NASH* criterion is greater than 0.8. The six models, ranked according to their performance in the *NASH* criterion from the highest to lowest, are listed as follows: EANN-CCA, SANN-CCA, Kriging-CCA, EANN-Origin, Tradition-CCA, and SANN-Origin. The *NASH* values obtained using the SANN-CCA and EANN-CCA approaches for the estimation of the three specific quantiles are all very close to or above 0.8. This indicates that the ANN models in the CCA space can provide satisfactory estimates.

[49] *RMSE* and *RMSEr* indices provide assessment of prediction accuracy in absolute and relative scale, respectively. The EANN-CCA model has the best performance among all the models according to these two indices. The CCA-based ANN approaches show significantly better generalization ability than the ANN approaches applied in the original physiographical space. The proposed approach which combines the advantages of linear and nonlinear methods seems to lead to a performance improvement. Furthermore, ANNs are nonparametric approaches which have strong limitations for the extrapolation beyond the range of observed data. The combination with a parametric approach seems to help the performance of the ANNs. The relative performance of all models ranked using both *RMSE* and *NASH* indicators are the same. A similar pattern can generally be observed using the *RMSEr* indicator. The Kriging-CCA model underperforms the Tradition-CCA model with the *RMSEr* indicator, however it outperforms the Traditional-CCA model with the *RMSE* and *NASH* indices. This indicates that the Traditional-CCA model and the CCA-based ANN models provide optimal estimates to minimize the absolute prediction error as indicated by the *RMSE* indicator, without sacrificing the performance in the relative measure as indicated by the *RMSEr* indicator.

[50] The *BIAS* and *BIASr* indices provide indication on whether a model tends to overestimate or underestimate. The analysis based on the *BIAS* index suggests that ANN models generally overestimate flood quantiles and the magnitude is larger than the Tradition-CCA model. However,

Table 1. Descriptive Statistics of Hydrological, Physiographical, and Meteorological Variables

Variables	Min	Mean	Max	STD
MBS, %	0.96	2.43	6.81	0.99
FAL, %	0.00	7.72	47.00	7.99
AMP, mm	646	988	1534	154
AMD, degree day	8589	16346	29631	5382
AREA, km^2	208	6255	96600	11716
q10, $m^3/s.km^2$	0.03	0.22	0.53	0.13
q50, $m^3/s.km^2$	0.03	0.28	0.77	0.18
q100, $m^3/s.km^2$	0.03	0.31	0.94	0.20

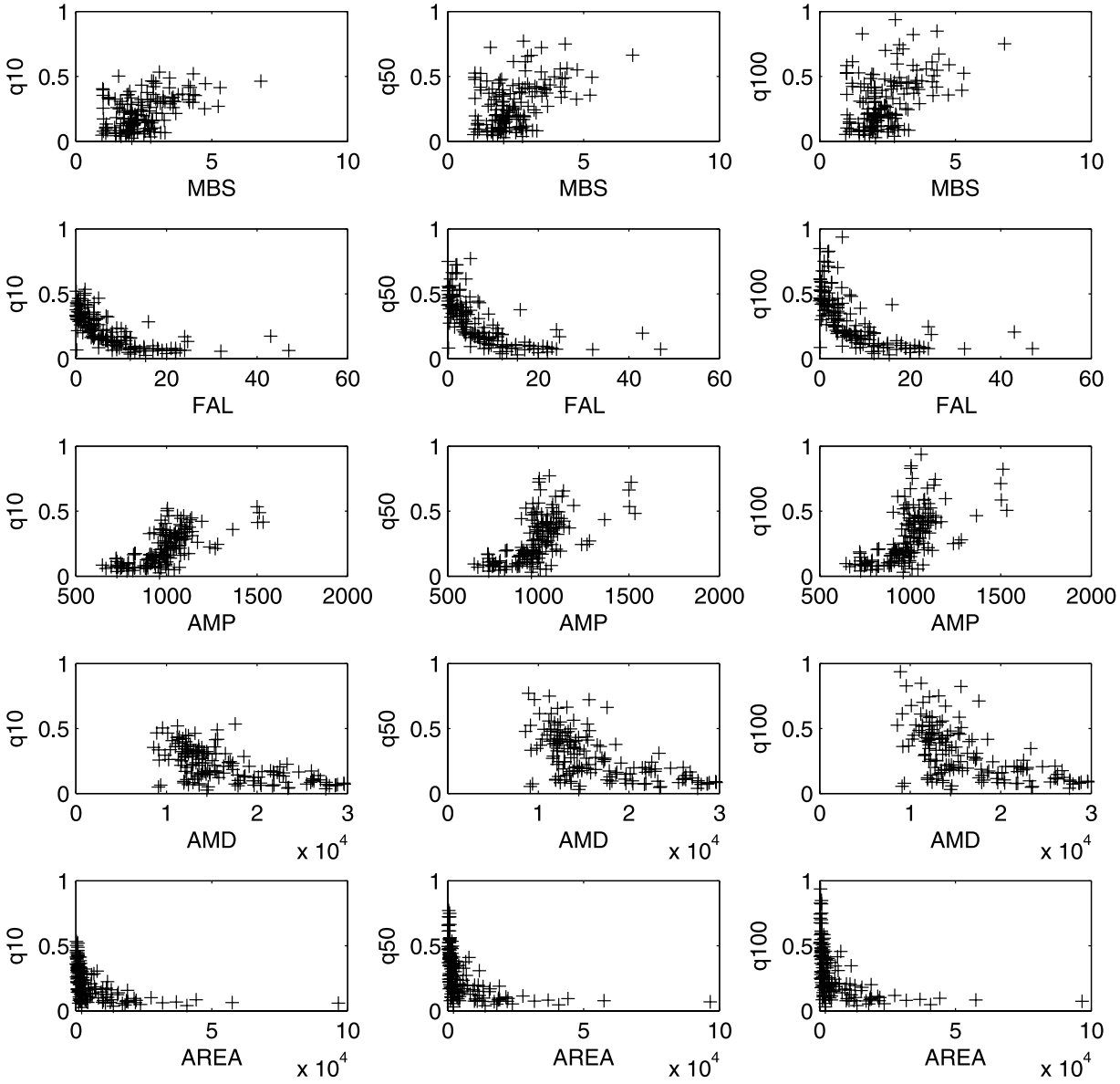

Figure 3. Scatterplot of site characteristics and specific flood quantiles.

Table 2. Jackknife Validation Results

	Hydrological Variables	SANN-CCA	EANN-CCA	Kriging-CCA	Tradition-CCA	SANN-Origin	EANN-Origin
NASH	q10	0.82	**0.84**	0.78	0.78	0.75	0.78
	q50	0.78	**0.80**	0.72	0.72	0.69	0.72
	q100	0.77	**0.78**	0.70	0.68	0.66	0.69
RMSE, $m^3/s.km^2$	q10	0.053	**0.050**	**0.050**	0.059	0.060	0.058
	q50	0.082	**0.079**	0.093	0.094	0.098	0.093
	Q100	0.095	**0.093**	0.110	0.112	0.115	0.109
RMSEr, %	Q10	38	**37**	51	43	47	44
	Q50	44	**43**	64	49	55	53
	Q100	46	**45**	70	**51**	64	60
BIAS, $m^3/s.km^2$	Q10	0.006	0.005	−0.004	**0.001**	0.006	0.004
	Q50	0.009	0.009	−0.007	**0.005**	0.010	0.009
	q100	0.013	0.012	−0.008	**0.007**	0.015	0.013
BIASr, %	q10	**−5**	**−5**	−16	−9	−7	−7
	q50	−7	**−5**	−21	−11	−8	−8
	q100	−7	**−6**	−23	−11	−11	−10

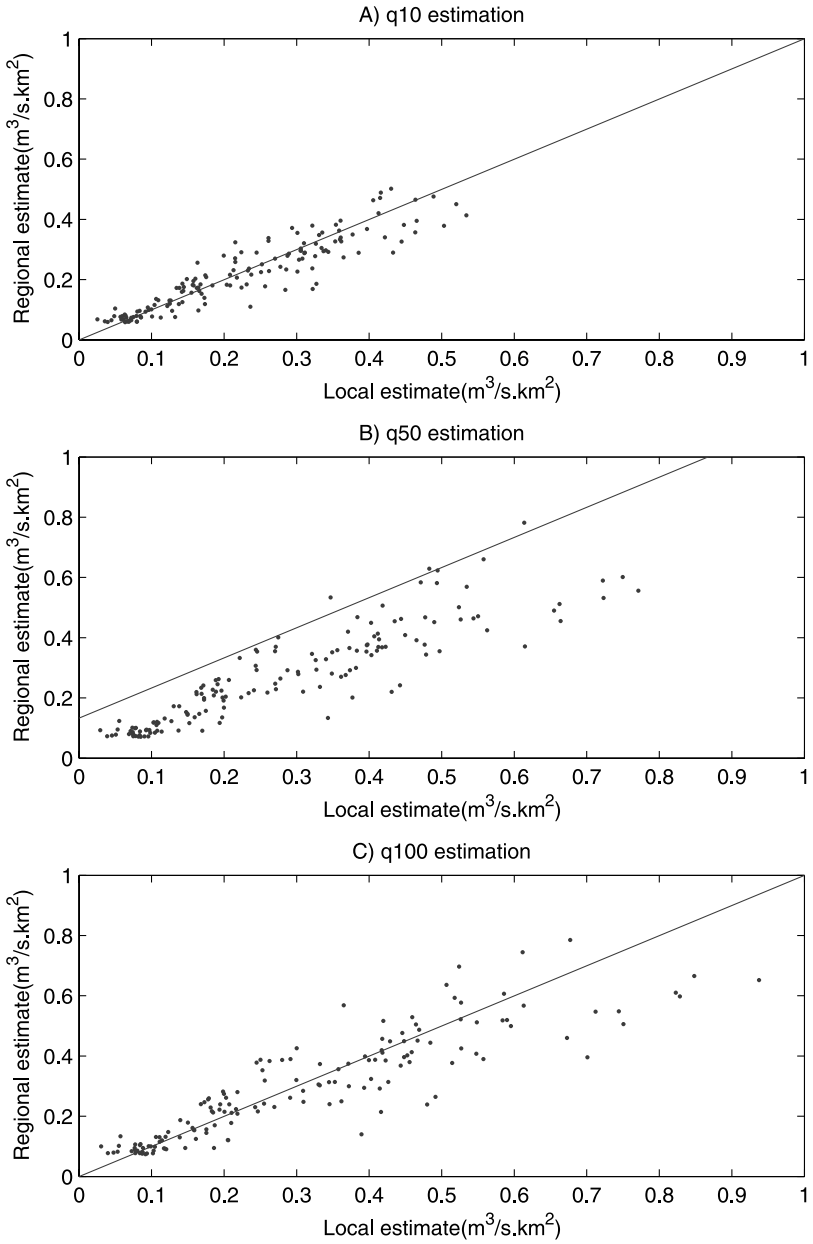

Figure 4. Jackknife estimation using the SANN-CCA approach.

when the *BIASr* index is used, both the ANN models and the Tradition-CCA model underestimate flood quantiles. Estimates obtained by ANN models in the CCA physiographical space have the lowest bias.

[51] Overall, the SANN-CCA model leads to a much better performance with *NASH, RMSE, RMSEr* indices than the SANN-Origin model. The EANN-CCA model shows better performance than the EANN-Origin model. These results indicate that applying ANN models in the CCA physiographical space can greatly improve the performance of ANN models than in the original physiographical space. *Chokmani and Ouarda* [2004] concluded that the CCA technique is more capable to characterize the physiographical space for conducting flood quantile estimation. The research results of this paper are consistent with their conclusions.

[52] The ANN ensemble approaches outperform the single ANN approach in both the original physiographical space and the CCA physiographical space according to most performance indices. These results are not surprising, as the ensemble approach can be used to improve the generalization ability of a single ANN [*Shu and Burn*, 2004].

[53] The regional estimates using the jackknife validation procedure for specific quantiles q10, q50, q100 using the SANN-CCA and EANN-CCA are shown in Figures 4 and 5, respectively. *Chokmani and Ouarda* [2004] provided the results using other CCA-based approaches. From Figures 4 and 5, we can observe that the estimation error and bias tend to increase with the return period. The CCA-based ANN models and the Tradition-CCA model tend to provide a better estimation than the Kriging-CCA approach for sites with specific quantiles lower than 0.15 $m^3/s.km^2$. All models underestimate at sites with higher values of specific quantiles (over 0.45 for the q10 estimate, over 0.6 for q50

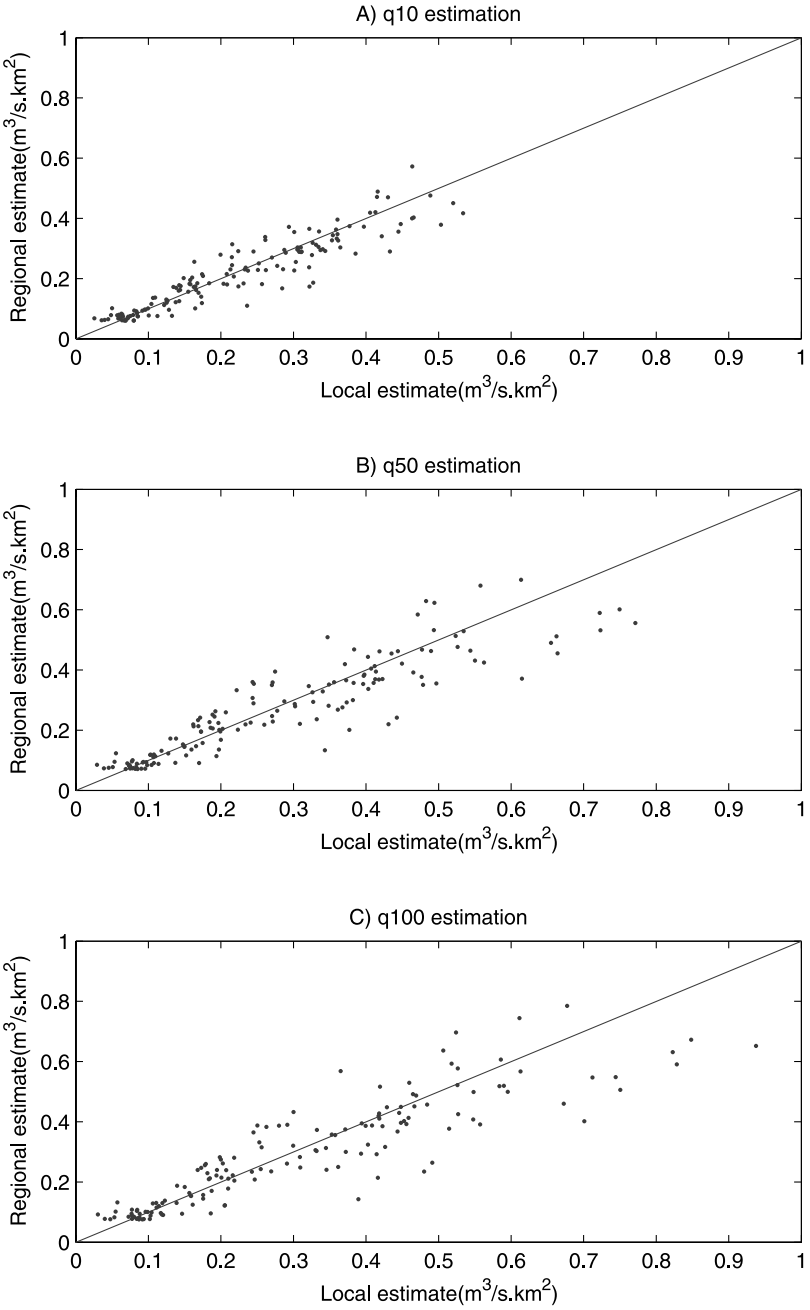

Figure 5. Jackknife estimation using the EANN-CCA approach.

and over 0.65 for q100). These sites generally represent smaller basins for which the hydrological response is very sharp (large specific quantiles). These basins are underrepresented in this case study. Indeed, there are only 9 sites in the database that represent basins with an area smaller than 500 km^2. Thus less training data is available in the variable space occupied by these smaller sites, which also increases the difficulty to provide precise estimation for small basins.

[54] In order to evaluate the contribution of each individual variable to flood quantile estimation, an additional experiment is conducted. A number of methods [*Olden et al.*, 2004] have been developed over the past few years to evaluate the relative importance of each input variable on the contribution to the estimation of the outputs in ANNs. *Olden et al.* [2004] compared nine methods for quantifying variable importance in ANN, and the results indicated that the connection weight approach [*Olden and Jackson*, 2002] is the best methodology. This approach is adopted in this paper. In this approach, the products of the input-hidden and hidden-output connection weights between each input neuron and output neuron are first calculated, then the products are summed across all hidden neurons to generate the importance of each input. Since the ANN model in the CCA physiographical space involves an input space projection, calculation of the contribution of each input variable to the estimation can be very complicated, and the connection weight approach cannot be used directly to provide the measurements. Thus the SANN-Origin model is selected to do the analysis. The relative importance of each input variable for the estimation of each specific flood quantile

Table 3. Relative Importance of the Five Input Variables for the Estimation of the Specific Flood Quantiles

	Relative Importance, %			
Input Variables	q10	q50	q100	Rank
MBS	15.7	15.9	16.8	3
FAL	34.6	35.3	35.9	1
AMP	22.5	21.9	20.7	2
AMD	13.4	12.7	12.1	5
AREA	13.8	14.2	14.5	4

is presented in Table 3. FAL and AMP are identified as the most important variables, and are followed by the variable MBS. AREA ranks fourth among the five input variables, and its relative importance ranges between 13.8% and 14.5%. The relative importance of AREA increases with the increase of the return period. AMD is the least important variable among all inputs.

6. Conclusions

[55] The methodology of integrating the CCA technique and ANNs for flood quantile estimation at ungauged sites is presented in this paper. CCA is used to project the site characteristics into the canonical physiographic space. ANN models are then used to approximate the functional relationship between flood quantiles and the projected physiographic variables. Two CCA-based ANN models, using a single network and an ensemble network, respectively, are developed and applied to the data of the case study.

[56] The jackknife validation is used to assess the performance of each model. The comparison with the other four regional flood frequency models shows that the proposed approaches can provide an estimation with relatively higher accuracy. The proposed CCA-based ANN models lead to a much better performance than the original ANN models, which tends to indicate that the CCA space is more appropriate for flood quantile estimation. The ensemble ANN approach outperforms the single ANN approach, which demonstrates that the generalization ability of a single ANN can be improved using the ensemble approach.

[57] Compared with the traditional CCA approach, the CCA-based ANN approaches are much easier to apply. In the traditional CCA approach, a procedure is required to optimally determine the value of the parameter α for each site which is directly related to the size of a hydrological neighborhood [*Ouarda et al.*, 2000]. In the CCA-based ANN approaches, once the ANN structure is specified, no interference is required in the training and estimation phase of the models.

[58] Although the CCA-based ANN approaches proposed in this paper lead to a better performance than the other methods, all methods tend to underestimate flood quantiles for catchments with very high specific quantiles (catchment with a small drainage area). Close inspection of these sites indicates that they locate in the variable space where less training data is available. Future attention should focus on the estimation of extreme basins (very small and very large) for which regional methodologies do not generally lead to very reliable estimates. To correctly estimate flood quantiles at these sites, further research is still required to increase the extrapolation ability of the current models.

[59] Six models for regional flood frequency analysis are compared in this work. These models are developed using three estimation techniques (ANN, kriging, multiple regression) in two physiographical spaces. The research results indicate that diverse generalization abilities are demonstrated in these models. For example, the EANN-CCA model shows better prediction accuracy than the Tradition-CCA model, while the Tradition-CCA model leads to a less biased estimation than the EANN-CCA model.

[60] The research in this paper is based on one type of ANN model, the MLP model. The method developed in this paper can be extended to use other types of ANNs such as the Radial Basis Function (RBF) network and the generalized regression network.

[61] **Acknowledgments.** The authors are grateful to the Editor, Tom Torgersen, the Associate Editor, and the three anonymous reviewers whose comments and suggestions greatly improved the quality of the manuscript. The financial support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chair Program, and the "Centre d'Études Nordiques" of Laval University is gratefully acknowledged. The authors wish also to thank K. Chokmani for his assistance.

References

Agrafiotis, D. K., W. Cedeño, and V. S. Lobanov (2002), On the use of neural network ensembles in QSAR and QSPR, *J. Chem. Inf. Compu. Sci.*, *42*, 903–911.

Ahmad, Z., and J. Zhang (2002), A comparison of different methods for combining multiple neural network models, paper presented at 2002 International Joint Conference on Neural Networks, World Congr. on Comput. Intell., Honolulu, Hawaii.

Anctil, F., and N. Lauzon (2004), Generalisation for neural networks through data sampling and training procedures, with applications to streamflow predictions, *Hydrol. Earth Syst. Sci.*, *8*, 940–958.

Beable, M. E., and A. I. McKerchar (1982), Regional flood estimation in New Zealand, *Water Soil Tech. Publ. 20*, 139 pp., Minist. of Works and Develop., Wellington, New Zealand.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford Univ. Press, New York.

Breiman, L. (1996), Bagging predictors, *Mach. Learn.*, *26*, 123–140.

Burn, D. H. (1990a), An appraisal of the "region of influence" approach to flood frequency analysis, *Hydrol. Sci. J.*, *35*, 149–165.

Burn, D. H. (1990b), Evaluation of regional flood frequency analysis with a region of influence approach, *Water Resour. Res.*, *26*, 2257–2265.

Cannon, A. J., and P. H. Whitfield (2002), Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models, *J. Hydrol.*, *259*, 136–151.

Castellarin, A., D. H. Burn, and A. Brath (2001), Assessing the effectiveness of hydrological similarity measures for flood frequency analysis, *J. Hydrol.*, *241*, 270–285.

Cavadias, G. S. (1990), The canonical correlation approach to regional flood estimation, *IAHS Publ.*, *191*, 171–178.

Chokmani, K., and T. B. M. J. Ouarda (2004), Physiographical space-based kriging for regional flood frequency estimation at ungauged sites, *Water Resour. Res.*, *40*, W12514, doi:10.1029/2003WR002983.

Cunnane, C. (1988), Methods and merits of regional flood frequency analysis, *J. Hydrol.*, *100*, 269–290.

Dawson, C. W., R. J. Abrahart, A. Y. Shamseldin, and R. L. Wilby (2006), Flood estimation at ungauged sites using artificial neural networks, *J. Hydrol.*, *319*, 391–409.

Demuth, H., and M. Beale (2003), *Matlab Neural Network Toolbox*, *Ver. 4*, Math Works, Natick, Mass.

Dietterich, T. G. (1997), Machine learning research: Four current directions, *AI Mag.*, *18*(4), 97–136.

Eaton, B., M. Church, and D. Ham (2002), Scaling and regionalization of flood flows in British Columbia, Canada, *Hydrol. Processes*, *16*, 3245–3263.

Efron, B., and T. J. Tibshirani (1993), *An Introduction to the Bootstrap*, CRC Press, Boca Raton, Fla.

Fausett, L. (1994), *Fundamentals of Neural Networks*, Prentice-Hall, Upper Saddle River, N. J.

Freund, Y., and R. E. Schapire (1996), Experiments with a new boosting algorithm, in *Proceedings of the Thirteenth International Conference on Machine Learning*, edited by L. Saitta, pp. 148–156, Morgan Kaufmann, San Francisco, Calif.

Groupe de Recherche en Hydrologie Statistique (GREHYS) (1996a), Presentation and review of some methods for regional flood frequency analysis, *J. Hydrol.*, *186*, 63–84.

Groupe de Recherche en Hydrologie Statistique (GREHYS) (1996b), Intercomparison of regional flood frequency procedures for Canadian rivers, *J. Hydrol.*, *186*, 85–103.

Grover, P. L., D. H. Burn, and J. M. Cunderlik (2002), A comparison of index flood estimation procedures for ungauged catchments, *Can. J. Civ. Eng.*, *29*, 734–741.

Hagan, M. T., and M. Menhaj (1994), Training feedforward networks with the Marquardt algorithm, *IEEE Trans. Neural Networks*, *5*(6), 989–993.

Hansen, L., and P. Salamon (1990), Neural network ensembles, *IEEE Trans. Pattern Anal. Mach. Intell.*, *12*, 993–1001.

Haykin, S. (1994), *Neural Networks—A Comprehensive Foundation*, 852 pp., Macmillan Coll., New York.

Kouider, A., H. Gingras, T. B. M. J. Ouarda, Z. Ristic-Rudolf, and B. Bobée (2002), Analyse fréquentielle locale et régionale et cartographie des crues au Québec, *Rep. R-627-el*, Eau, Terre, et Environ., Inst. Natl. de la Rech. Sci., Ste-Foy, Que., Canada.

Matalas, N. C., J. R. Slack, and J. R. Wallis (1975), Regional skew in search of a parent, *Water Resour. Res.*, *11*, 815–826.

McCuen, R. H., R. B. Leahy, and P. A. Johnson (1990), Problems with logarithmic transformations in regression, *J. Hydraul. Eng.*, *116*, 414–428.

Merz, C. J. (1998), Classification and regression by combining models, Ph.D. thesis, Dep. of Inf. and Comput. Sci., Univ. of Calif., Irvine.

Muirhead, R. J. (1982), *Aspect of Multivariate Statistical Theory*, John Wiley, Hoboken, N. J.

Olden, J. D., and D. A. Jackson (2002), Illuminating the "black box:" Understanding variable contributions in artificial neural networks, *Ecol. Modell.*, *154*, 135–150.

Olden, J. D., M. K. Joy, and R. G. Death (2004), An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data, *Ecol. Modell.*, *178*, 389–397.

Opitz, D., and R. Maclin (1999), Popular ensemble methods: An empirical study, *J. Artif. Intell. Res.*, *11*, 169–198.

Ouarda, T. B. M. J., M. Haché, P. Bruneau, and B. Bobée (2000), Regional flood peak and volume estimation in a northern Canadian basin, *J. Cold Reg. Eng.*, *14*, 176–191.

Ouarda, T. B. M. J., C. Girard, G. S. Cavadias, and B. Bobée (2001), Regional flood frequency estimation with canonical correlation analysis, *J. Hydrol.*, *254*, 157–173.

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

Razavi, A. R., H. Gill, H. Åhlfeldt, and N. Shahsavar (2005), A data preprocessing method to increase efficiency and accuracy in data mining, in *Proceedings of the* 10th *Conference on Artificial Intelligence in Medicine*, edited by S. Miksch, J. Hunter, and E. Keravnou, pp. 434–443, Springer, Berlin.

Reed, D. W., and A. J. Robson (1999), *Flood Estimation Handbook*, vol. 3, Inst. of Hydrol., Wallingford, U. K.

Rumelhart, D. E., and J. L. McClelland (Eds.) (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, *Foundations*, MIT Press, Cambridge, Mass.

Schapire, R. E. (1990), The strength of weak learnability, *Mach. Learn.*, *5*, 197–227.

Seidou, O., B. M. J. T.Ouarda, L. Bilodeau, M. Hessami, A. St-Hilaire, and P. Bruneau (2006), Modeling ice growth on Canadian lakes using artificial neural networks, *Water Resour. Res.*, *42*, W11407, doi:10.1029/2005WR004622.

Sharkey, A. J. C. (Ed.) (1999), *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, Springer, New York.

Shu, C., and D. H. Burn (2004), Artificial neural network ensembles and their application in pooled flood frequency analysis, *Water Resour. Res.*, *40*, W09301, doi:10.1029/2003WR002816.

Thomas, D. M., and M. A. Benson (1970), Generalization of streamflow characteristics from drainage-basin characteristics, *U.S. Geol. Surv. Water Supply Pap., 1975*.

Wiltshire, S. E. (1986), Regional flood frequency analysis I: Homogeneity statistics, *Hydrol. Sci. J.*, *31*, 321–333.

Wolpert, D. H. (1992), Stacked generalization, *Neural Networks*, *5*, 241–259.

Zrinji, Z., and D. H. Burn (1994), Flood frequency analysis for unguaged sites using a region of influence approach, *J. Hydrol.*, *153*, 1–21.

T. B. M. J. Ouarda and C. Shu, ETE, INRS, University of Quebec, 490 de la Couronne, Quebec, QC, Canada G1K 9A9. (taha_ouarda@ete.inrs.ca; chang.shu@ete.inrs.ca)