Taylor & Francis
Taylor & Francis Group

Canadian Water Resources Journal / Revue canadienne des ressources hydriques

ISSN: 0701-1784 (Print) 1918-1817 (Online) Journal homepage: http://www.tandfonline.com/loi/tcwr20

An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds

Tara Razavi & Paulin Coulibaly

To cite this article: Tara Razavi & Paulin Coulibaly (2017) An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 42:1, 2-20, DOI: 10.1080/07011784.2016.1184590

To link to this article: https://doi.org/10.1080/07011784.2016.1184590

Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 2017
Vol. 42, No. 1, 2–20, http://dx.doi.org/10.1080/07011784.2016.1184590

An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds

Tara Razavi* and Paulin Coulibaly

McMaster University, Department of Civil Engineering, School of Geography and Earth Sciences, Hamilton, Canada

(*Received 17 April 2015; accepted 27 April 2016*)

Regionalization – the process of transferring hydrological information from gauged to ungauged watersheds – has the potential to perform significantly better if these watersheds are classified in advance. In this study, we demonstrate the benefits of classification by a systematic combination of watershed classification techniques, regionalization methods, and rainfall-runoff models. Basins are first classified, then regionalization methods are applied, for continuous daily streamflow estimation at ungauged watersheds in Ontario, Canada. Nonlinear data-driven methods are used as regionalization and watershed classification schemes to transfer the parameters of two conceptual hydrologic models – namely McMaster University Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) and Sacramento Soil Moisture Accounting (SAC-SMA) – from gauged to ungauged watersheds. Our results suggest that a certain combination of watershed classification technique, regionalization method and hydrologic model can significantly improve the estimation of continuous streamflow at ungauged basins by improving the accuracy of estimated daily mean, low and peak flows. However, some combinations do not provide a clear improvement when compared to the scenario of unclassified basins. For example, the MAC-HBV model coupled with a counter propagation neural network as a regionalization technique provides a clear improvement in estimated daily mean, low and peak flow when the watersheds are first classified using a nonlinear principal component analysis method. Interestingly, a higher improvement is achieved for low flow as well, which is usually difficult to estimate in ungauged basins.

La régionalisation – processus de transfert de l'information hydrologique de bassins jaugés à bassins non jaugés – a le potentiel de donner des performances significativement meilleures si ces bassins sont classés à l'avance. Dans cette étude, nous démontrons les avantages de la classification par une combinaison systématique des techniques des bassins de classification, des méthodes de régionalisation et des modèles pluie-écoulements. Les bassins sont d'abord classés, puis des méthodes de régionalisation sont appliquées pour l'estimation des écoulements par jour en continu à des bassins versants non jaugés en Ontario au Canada. Les méthodes non linéaires orientées données sont utilisées comme systèmes de régionalisation et de classification des bassins versants pour transférer les paramètres de deux modèles hydrologiques conceptuels – à savoir MAC-HBV (McMaster University Hydrologiska Byråns Vattenbalansavdelning) et SAC-SMA (Sacramento Soil Moisture Accounting) – de bassins versants jaugés à bassins versants non jaugés. Nos résultats suggèrent qu'une certaine combinaison de la technique de classification des bassins versants, la méthode de régionalisation et le modèle hydrologique peuvent considérablement améliorer l'estimation des écoulements en continu à bassins non jaugés en améliorant la précision des flux journaliers moyens, faibles et débits de pointe estimés. Par exemple, le modèle MAC-HBV couplé avec un compteur de propagation à réseau de neurones comme technique de régionalisation fournit une nette amélioration du débit moyen, faible et de pointe journalier estimé lorsque les bassins versants sont d'abord classés en utilisant une méthode d'analyse à composantes principales non linéaires. Fait intéressant, une grande amélioration est atteinte pour un faible débit aussi bien, ce qui est habituellement difficile à estimer dans les bassins non jaugés.

Keywords: continuous daily streamflow; neural networks; regionalization; watershed classification

Introduction

Continuous streamflow series are fundamental for water resources management and the design of various hydraulic infrastructures. With sustainable management decisions, governments and agencies can help to maintain existing water resources for future generations and protect human life from catastrophic flood events. At a detailed level, continuous daily streamflow series are useful for the estimation of daily flow peaks, low flow and flow duration curves. Unfortunately, streamflow data are not available for many of the world's watersheds (Mishra and Coulibaly 2009). In Ontario, Canada, most of the natural flow regime basins (more than 60%) within the province's 1 million km^2 area are still ungauged or poorly gauged (Samuel et al. 2012a). In the United States, less than 25,000 (10%) river basins out of 250,000 are gauged by the United States Geological Survey (Besaw et al. 2010). This picture gets worse for many developing countries.

*Corresponding author: E-mail: razaviz@mcmaster.ca

For gauged and/or ungauged watersheds, streamflow series are usually predicted using rainfall-runoff models or data-driven methods such as regression-based approaches, in which streamflow series are not estimated through hydrologic models but are based on watershed physiographic and/or meteorological characteristics and data-driven models. A detailed discussion on these methods can be found in Razavi and Coulibaly 2012. Since observed streamflow series in ungauged or poorly gauged watersheds are not available to calibrate these prediction models, the model parameters of the gauged watersheds are usually transferred to the ungauged ones. Some studies have explored the value of available data in poorly gauged watersheds to reconstruct continuous streamflow series for ungauged watershed (e.g. Drogue and Plasse 2014; Viviroli and Seibert 2015). This process is known as regionalization in hydrology, and is expected to be more reliable if the watersheds are similar in some respects (Blöschl and Sivapalan 1995). There are a number of different approaches that can be used for regionalization, such as spatial proximity, physical similarity, and regression-based methods including linear and non-linear regression.

To the best of our knowledge, watershed classification prior to regionalization has not yet been systematically evaluated in large and diverse watersheds such as Canada's river basins. In this study, we aim to investigate the benefits of continuous daily streamflow regionalization after systematic watershed classification using nonlinear data-driven approaches such as artificial neural networks.

Some conventional regionalization techniques are inherently involved with watershed classification. For example, in physical similarity or spatial proximity approaches, hydrologic responses are transferred from gauged to ungauged watersheds in clusters of similar physical attributes or similar location. Other types of regionalization approaches, such as linear regression or artificial neural networks, can be applied on either homogeneous groups of watersheds or unclassified ones. Several studies have investigated the potential of improving hydrological predictions in ungauged watersheds after classification. These mostly present a procedure to identify homogeneous regions based on watersheds' physical attributes to estimate hydrological responses which can also be used in ungauged watersheds (e.g. Nathan and McMahon 1990a; Burn and Boorman 1993; Cavadias et al. 2001; Ilorme and Griffis 2013). The focus of past studies has mostly been on the classification framework and the possibility of its application for ungauged basins or regionalization itself, rather than the impacts of classification techniques on the performance of regionalization. Very few studies have investigated the latter. For example, Prinzio et al. (2011) investigated the performance of estimating streamflow indices (such as mean annual runoff, flood quartiles and mean annual flood) in ungauged watersheds after classification by applying a self-organizing map (SOM) on physical attributes. They found that watershed classification using a SOM could reduce the uncertainty of hydrological predictions in ungauged sites. Kileshye et al. (2012) performed principal component analysis (PCA)| and tree cluster analysis on landform attributes to find physiographic similarities between sub-catchments, and finally predicted the long-term monthly mean discharges using linear and nonlinear regression models. They discovered the existence of two major groups of sub-catchments using PCA, concluding that the non-linear model provides better prediction of the flows compared to the linear one.

The present study proposes a more comprehensive analysis, including an investigation of systematic watershed classification using nonlinear classification techniques prior to regionalization, by applying two rainfall-runoff models to 90 basins in Ontario. Watershed classification can be based on either the physiographic or climate characteristics of watersheds, or streamflow metrics. Since streamflow series are not available in ungauged watersheds and climate data might not be available as well, a classification based on watershed physiographic attributes is most plausible. In this study, we will consider three watershed classification scenarios. In each scenario, clusters are identified as the homogeneous regions using nonlinear clustering techniques including SOMs, standard non-linear principal component analysis (NLPCA), and compact non-linear principal component analysis (CNLPCA) on watershed physiographic attributes to classify 90 watersheds into four clusters. The performance of these nonlinear techniques in watershed classification is compared with PCA and k-means clustering based on runoff signatures (applicable to gauged watersheds) as linear benchmark techniques in a previous study (Razavi and Coulibaly 2013). Our results suggest that the nonlinear classification techniques applied on watershed attributes could be reliable alternative methods for the classification of gauged and/or ungauged watersheds.

In this study, we apply inverse distance weighted (IDW), multi-layer perceptron (MLP), counter propagation neural network (CPNN), and support vector machine (SVM) as regionalization techniques to pre-identified homogeneous clusters of watersheds (as well as unclassified watersheds) to investigate potential improvements in continuous daily streamflow regionalization. The main objective is to investigate these potential improvements by applying nonlinear data-driven approaches to systematically pre-classified watersheds along with different rainfall-runoff models and the combination of them.

Study area and data

The study area covers 90 natural watersheds across the Province of Ontario, Canada (Figure 1), with annual mean precipitation of 400–600 mm in the northern region and 800–1200 mm in the southern part. In the northern region, the average air temperature ranges approximately between −20°C (in January) and 17°C (in July); in the southern regions, between −10°C (in January) and 19°C (in July). Most of the natural watersheds in the northern region are covered with coniferous forest, with gaps of swamp, muskeg and small lakes, whereas the southern region is dominated by mixed forests (*Atlas of Canada*, available at http://atlas.nrcan.gc.ca).

Meteorological data, including daily precipitation and air temperature, were obtained from Canadian daily climate data using the climate stations which were nearest to the watersheds' centroid and had less than 20% missing data (provided by Environment Canada at http://climate.weather.gc.ca). Daily flow data (1976–1994) were obtained from the HYDAT database (hydrometric data, Environment Canada 2004). Climate and streamflow data from 1976 to 1985 (10 years) were used for model calibration, while data from 1986 to 1994 (9 years) were used for model validation. The areas of the watersheds range from approximately 100 to 100,000 km^2, representing different types and sizes of watersheds. The watershed attributes used in this study are similar to the ones used in Samuel et al. (2011) and Razavi and Coulibaly (2013). To select the catchment attributes, we considered data availability and the catchment attributes mostly used in regionalization studies. They can be classified as follows:

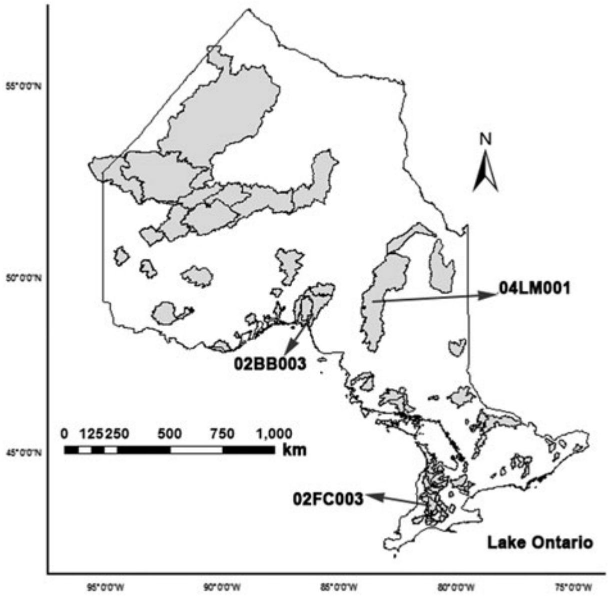

Figure 1. Location map of selected Ontario watersheds and sample watersheds.

- Location of the centroid of the watersheds (latitude and longitude);
- Morphology (mean elevation, mean slope and area);
- Percentage of area covered by water (proportion of lakes);
- Land use (proportion of forests);
- Water drainage (the sum of the percentage of the area covered by rapid and moderate drainage classes);
- Rooting depth (associated with soil depth and available water, i.e. the portion [fractional] of area covered by vegetation with root depth deeper than 150 cm); and
- The surficial geology (the percentage of the region covered by glaciofluvial sediments, glacio-deposits and rock).

The details of the catchment attributes and range of their values for the 90 Ontario watersheds (study area) are presented in Table 1. All catchment attributes were derived using the digital maps and digital elevation database. Data were obtained from the Shuttle Radar Topography Mission (SRTM; available at http://www2.jpl.nasa.gov/srtm/cbanddataproducts.html). SRTM data are organized into individual rasterized cells, each covering 1° by 1° in latitude and longitude, with approximately 90 m resolution. The digital elevation model data were used to delineate catchments, compute the morphology of the catchments and determine the location of the centroid of the catchments. Information on the proportions of forests, lakes, water drainages, rooting depths, glaciofluvial sediments, glacio-deposits and rocks is available from a large database known as ArcCanada (ESRI 1997) in the form of digital maps. The data sets were created for the entire province at a scale of 1:1000,000 and based on the North American datum of 1927. These data were then combined with the catchment boundaries to derive areal portions of forests, lakes, water drainages, rooting depths, glaciofluvial sediments, glacio-deposits and rocks (Samuel et al. 2011).

Methodology

An overview of the methodology is provided in a flowchart (see Figure 2). The method includes four regionalization approaches – specifically, three types of neural networks (MLP, SVM, CPNN), and a spatial proximity method (IDW) – that are applied to both scenarios, the unclassified (all 90 watersheds) and classified watersheds, using SOM, NLPCA and CNLPCA techniques to transfer the hydrologic model parameters of gauged

Table 1. Catchment characteristics used in this study and their range of values for the 90 Ontario basins (modified after Samuel et al. 2011).

Catchment attributes			Unit	Range of values	Notes
Location of watershed (centroid)	Latitude		Degrees	44.2–55.3	
	Longitude		Degrees	−94.4–−74.6	
Morphology of basins	Area		km^2	85.5–91,802	
	Mean watershed slope		%	1.5–208.3	
	Mean elevation		m	13.5 – 785.6	
Percentage of area covered by	Lakes		%	0 – 0.2	Percentage of area covered by lakes
	Forest		%	0–100	Sum of percentage of area covered by coniferous, deciduous and mixed forest
	Rapid drainage class		%	0–100	Sum of percentage of area covered by rapid, well and moderately well drained drainage classes
	Rooting depth deeper than 150 cm		%	0–100	Percentage of area covered by rooting depth deeper than 150 cm
	Surficial geology	Glaciofluvial sediments	%	0–55	Sum of percentage of area covered by glaciofluvial sediments
		Glacio-deposits	%	5–100	Sum of percentage of area covered by till blanket and till veneer
		Rocks	%	0–66	Percentage of area covered by undivided bedrocks

watersheds to the ungauged ones. The 90 watersheds were considered a combination of gauged and ungauged watersheds. The classification techniques used watershed physical characteristics, and therefore can be applied on any gauged or ungauged watershed. These classification techniques are described in detail in a previous study (Razavi and Coulibaly 2013). When regionalization techniques were applied, gauged watersheds were used in the model training while hypothetical ungauged watersheds were used in the validation step. To evaluate the performance of models for all watersheds, for the IDW technique a leave-one-out cross validation and for neural networks a 3-fold cross validation technique was used, such that two thirds of watersheds were used in model training and the remaining one third in model validation. The two hydrologic models used to estimate continuous daily streamflow are the conceptual rainfall-runoff models described in the next section, followed by a brief description of the regionalization techniques and model performance evaluation criteria.

Rainfall-runoff models

MAC-HBV

The McMaster University Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV; Samuel et al. 2011) is a lumped conceptual rainfall-runoff model, following the structure of the HBV model (Bergström 1976) which has been widely used in hydrological studies, and in particular many regionalization studies. The MAC-HBV uses a concept of the HBV model similar to what was presented earlier by Merz and Blöschl (2004) and a modified routing routine following Seibert (1999) with a simplified Thornwaite formula to account for daily potential evapotranspiration (SMHI 2005). The model consists of a snow routine, a soil moisture routine, a response function and a routing routine. The snow routine represents changes in the snowpack using a simple degree-day concept. The soil moisture routine represents the soil moisture accounting or changes in soil moisture storage in the top soil layer. The response function estimates the amount of runoff from the upper zone and lower zone based on the current water storage and the maximum storage. For channel routing, an equilateral triangular weighting function is used to obtain the final streamflow. The parameters of this model are presented in Table 2. A detailed description of the MAC-HBV model can be found in Samuel et al. (2011).

SAC-SMA

The Sacramento Soil Moisture Accounting (SAC-SMA) is a conceptual watershed model (Burnash et al. 1973) used by the National Weather Service for operational streamflow forecasting and flood warning throughout the United States. This hydrologic model is a conceptual system for modeling the headwater portion of the hydrologic cycle. The first component of the model, rainfall, occurring over the basin is considered as falling on two basic areas: the pervious area and impervious area. Pervious area refers to the permeable portion of the soil mantle, while impervious area is the portion of the soil mantle covered by streams, lake surfaces, marshes, or other impervious material. This model consists of six

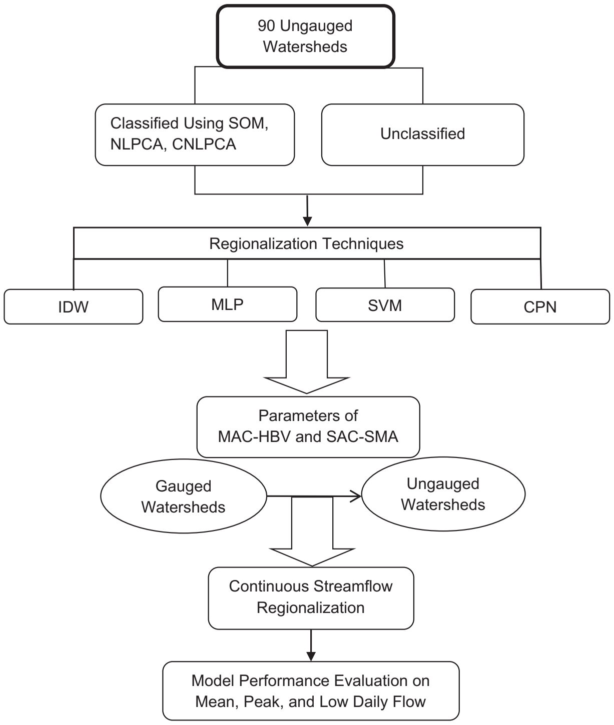
90 Ungauged Watersheds
Classified Using SOM, NLPCA, CNLPCA
Unclassified
Regionalization Techniques
IDW
MLP
SVM
CPN
Parameters of MAC-HBV and SAC-SMA
Gauged Watersheds
Ungauged Watersheds
Continuous Streamflow Regionalization
Model Performance Evaluation on Mean, Peak, and Low Daily Flow

Figure 2. Flowchart of methodology showing the four regionalization techniques – inverse distance weighted (IDW), multilayer perceptron (MLP), support vector machine (SVM), and counter propagation neural network (CPNN) – used to transfer the parameters of two hydrologic models (McMaster University Hydrologiska Byråns Vattenbalansavdelning [MAC-HBV] and Sacramento Soil Moisture Accounting [SAC-SMA]) from gauged to ungauged watersheds considering two scenarios of classified and unclassified watersheds using non-linear principal component analysis (NLPCA), CNLPCA (compact NLPCA), and self-organizing map (SOM) techniques.

state variable reservoirs representing the accumulation of water in two soil zones (upper and lower) in the form of both "tension" and "free" water. Tension water is considered water which is closely bound to soil particles and is available for evapotranspiration, while free water is the portion of water which is not bound to soil particles and so is free to descend to deeper portions of the soil and move laterally through the soil due to gravitational and pressure forces.

The state variables include: additional impervious area content (ADIMC), upper-zone tension water storage content (UZTWC), upper-zone free water storage content (UZFWC), lower-zone tension water storage content (LZTWC), lower-zone free primary water storage content (LZFPC) and lower-zone free secondary water storage content (LZFSC). The structure of the SAC-SMA model used herein is illustrated in Figure 3, while the optimized maximum and minimum ranges of each model parameter are presented in Table 2. The routing approach used in this model is the Nash cascade method, and the same snow component and evapotranspiration calculation methods used in the MAC-HBV model are added to this model.

Model parameter optimization

The two rainfall-runoff models were calibrated against observed daily streamflow time series, with 1976–1985 (10 years) used for model calibration to obtain the optimized parameters, while 1986–1994 (9 years) were used for model validation and comparison. Data of the first year were used for model warm-up. The optimization algorithms – including particle swarm optimization (PSO; Clerc 2006; Eberhart and Kennedy 1995), shuffle complex efficiency (SCE; Duan et al. 1994), and non-sorted genetic algorithm II (NSGA II; Deb et al. 2002) – and a Monte Carlo simulation were used to optimize the parameters of two models. In the Monte Carlo simulation, 100,000 uniformly distributed random values of the model parameters were selected in their initial ranges and the parameter set which produced the highest model performance was selected as the optimized set of parameters. The criterion of performance evaluation used for all the optimization algorithms was the objective function NVE (combined Nash Sutcliffe efficiency and volume error) used by Samuel et al. (2011) which addresses mean, low and high flows at the same time:

$$NVE = 0.5NSE - 0.1VE + 0.25NSE_{log} + 0.25NSE_{sqr} \tag{1}$$

where Nash Sutcliffe efficiency (NSE) is:

$$NSE = 1 - \left[\frac{\sum_{i=1}^{N}(Q_{obs} - Q_{sim})^2}{\sum_{i=1}^{N}(Q_{obs} - Q_{obs})^2}\right] \tag{2}$$

and volume error (VE) is:

$$VE = \frac{\sum_{i=1}^{N} Q_{sim} - \sum_{i=1}^{N} Q_{obs}}{\sum_{i=1}^{N} Q_{obs}} \tag{3}$$

Values of NSE based on square and logarithm of discharge can also be calculated as follows:

$$NSE_{sqr} = 1 - \left[\frac{\sum_{i=1}^{N}(Q^2_{sim} - Q^2_{obs})^2}{\sum_{i=1}^{N}(Q^2_{obs} - \overline{Q^2_{obs}})^2}\right] \tag{4}$$

$$NSE_{log} = 1 - \left[\frac{\sum_{i=1}^{N}(\log Q_{sim} - \log Q_{obs})^2}{\sum_{i=1}^{N}(\log Q_{obs} - \overline{\log Q_{obs}})^2}\right] \tag{5}$$

where Q_{sim} and Q_{obs} are the simulated and observed streamflow, respectively, while $\bar{Q}_{\mathrm{obs}}$ is the average of observed streamflow values and N is the number of data points. The $\mathrm{NSE}_{\mathrm{log}}$ is better at reflecting the accuracy of low flow, while $\mathrm{NSE}_{\mathrm{sqr}}$ is better at reflecting the

Table 2. Parameters of McMaster University Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) and Sacramento Soil Moisture Accounting (SAC-SMA) models and optimized ranges (using particle swarm optimization [PSO] algorithm).

MAC-HBV				
Parameter	**Description**	**Unit**	**Initial range**	**Optimized range**
tr	Upper threshold temperature to distinguish between rainfall and snowfall	°C	−1.5–2.5	0–2.5
scf	snowfall correction factor	_	0.4–1.6	0.44–1.55
ddf	Degree day factor	mm/(day °C)	0–5	0.5–5
athorn	A constant for Thornthwaite's equation	_	0.1–0.3	0.1–0.3
fc	Maximum soil box water content	mm	50–800	111–800
lp	Limit for potential evaporation	mm/mm	0.1*fc–0.9*fc	5–717
beta	A non-linear parameter controlling runoff generation	_	0–10	0.25–10
k0	Flow recession coefficient in an upper soil reservoir (for soil moisture exceeds a threshold lsuz value)	days	1–30	1–30
lsuz	A threshold value used to control response routing on an upper soil reservoir	mm	1–100	1–100
k1	Flow recession coefficient in an upper soil reservoir	days	2.5–100	30–100
cperc	A constant percolation rate parameter	mm/day	0.01–6	0.01–6
k2	Flow recession coefficient in a lower soil reservoir	days	20–1000	100–500
maxbas	A triangle weighting function for modelling a channel routing routine	days	1–20	1–17
rcr	Rainfall correction factor	_	0.5–1.5	0.65–1.5
α1	An exponent in relation between outflow and storage representing non-linearity of storage – discharge relationship of lower reservoir	_	0.5–20	0.6–1.5
SAC-SMA				
UZTWM	Upper-zone tension water maximum storage	mm	1–150	1–150
UZFWM	Upper-zone free water maximum storage	mm	1–150	17–145
LZTWM	Lower-zone tension water maximum storage	mm	1–500	1–446
LZFPM	Lower-zone free water primary maximum storage	mm	1–1000	1–996
LZFSM	Lower-zone free water supplemental maximum storage	mm	1–1000	1–1000
ADIMP	Additional impervious area	-	0–0.4	0–0.4
UZK	Upper-zone free water lateral depletion rate	day^{-1}	0.1–0.5	0.1–0.5
LZPK	Lower-zone primary free water lateral depletion rate	day^{-1}	0.0001–0.025	0.0001–0.025
LZSK	Lower-zone supplemental free water lateral depletion rate	day^{-1}	0.01–0.25	0.01–0.25
ZPERC	Maximum percolation rate	-	2.5–100	30–100
REXP	Exponent of the percolation equation	-	0.01–6	0.01–6
PCTIM	Impervious fraction of the watershed area	-	20–1000	100–500
PFREE	Fraction percolating from upper to lower zone free water storage	days	1–20	1–17
Rq	Routing coefficient	_	0.5–1.5	0.65–1.5
ddf	Degree day factor	_	0.5–20	0.6–1.5
scf	Snowfall correction factor	-	0.4–1.6	0.5–1.5
tr	Upper threshold temperature to distinguish between rainfall and snowfall	°C	−1.5–2.5	0–2.5
athorn	A constant for Thornthwaite's equation	-	0.1–0.3	0.1–0.3
rcr	Rainfall correction factor	-	0.5–1.5	0.7–1.5
Constant numbers				
RIVA	Riparian vegetation area	-	0	
SIDE	Ratio of the deep recharge to channel baseflow	-	0	
RSERV	Fraction of lower zone free water not transferable to tension water	-	0.3	

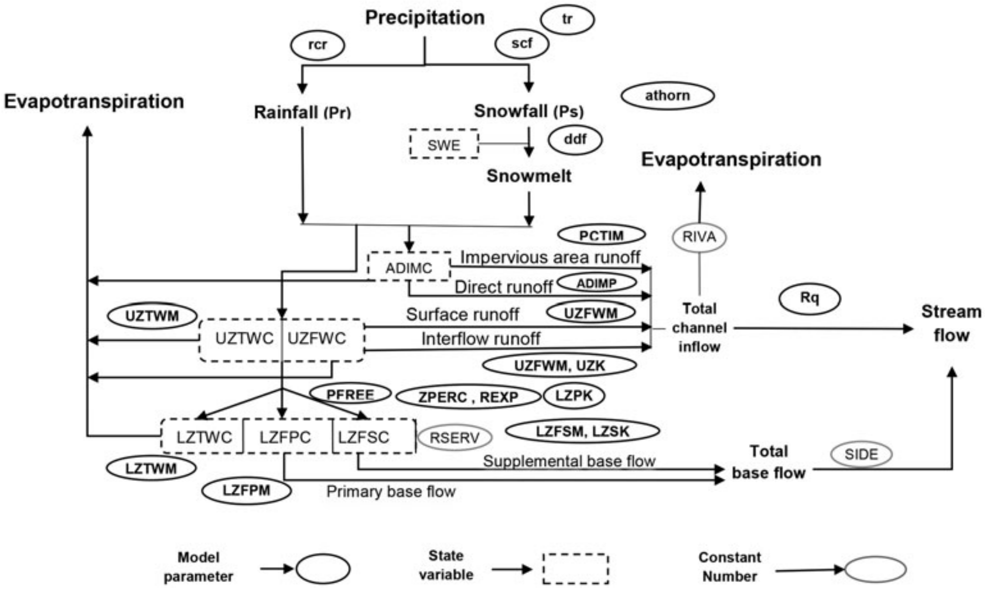

Figure 3. Schematic representation of the Sacramento Soil Moisture Accounting (SAC-SMA) model as used in this study showing the model state parameters used. Abbreviations are presented in Table 2. Arrows indicate fluxes between components and the streamflow at the watershed outlet is shown (adapted from Vrugt et al. 2006).

accuracy of high flows. Using the objective function in Equation (1), the single-objective algorithms can be useful as a multi-objective one. The model with the highest performance should produce a value close to 1 for NVE and NSE (Nash and Sutcliffe 1970) and a value close to 0 for VE.

Regionalization techniques

Inverse distance weighted (IDW)

IDW is an interpolation technique based on the inverse spatial distance between watershed centroids. This method is coupled with the physical similarity approach as in Samuel et al. (2011), and is recognized as the best regionalization method among the other investigated approaches (regression-based and physical similarity) for the study area. The spatial distance between the watersheds is calculated using the latitude and longitude of the watersheds' centroids. The IDW equation (Shepard 1968) used in this study to estimate model parameters in ungauged watersheds is:

$$P_j = \sum_{i=1}^{n} W_i p_i \tag{6}$$

where n is the number of gauged watersheds; $p_i$is the model parameter of gauged watersheds, P_j is the model parameter of an ungauged watershed and W_i is the weight function of each gauged watershed and is calculated as follows:

$$W_i = \frac{(d_i^{-2})}{\sum_{i=1}^{n}(d_i^{-2})} \tag{7}$$

where d_{i} is distance from the centroid of the gauged watersheds to the centroid of the ungauged watershed. In the selection of gauged watersheds additional criteria such as the NSE value can be considered; however, the number of available gauged watersheds can be a constraint.

First, before watershed classification, each of the 90 watersheds was assumed to be ungauged in turn. After calculating the weights of other watersheds based on their distance, the model parameters of the ungauged watershed were obtained using the parameters of the gauged ones. After classifying the watersheds, the weights were calculated within each cluster and model parameters were obtained for each watershed assumed as ungauged based on the distance from other watersheds in the cluster.

Multi-layer perceptron (MLP)

MLP is a feed-forward neural network which maps input data sets to output or the network target. As the most widely used data-driven model in hydrologic applications (Maier et al. 2010), it was selected as a benchmark method. The MLP used in this study has one input layer, one hidden layer and one output layer. Physiographic watershed attributes (12 attributes) were used as the network input's vector, and each hydrologic model parameter of the MAC-HBV and the SAC-SMA models, presented in Table 2, is used as the network output in separate networks for 90 samples (watersheds). Data from two thirds of the watersheds were used to train the network, while the model parameters of the remaining watersheds were obtained from the validation period. To cover all the watersheds, this procedure is repeated three times. Data from 60 watersheds were used as gauged watersheds to train the network, and, using the same network architecture, the model parameters of the remaining 30 watersheds (as ungauged watersheds) were estimated using their attributes as the network's input. This procedure was repeated three times so that all the watersheds were considered ungauged once. For regionalization with classification, the same procedure was performed within each cluster separately. Two thirds of the watersheds in each cluster were considered gauged while the remaining one third were considered ungauged watersheds, and the analysis was performed three times to encompass all the watersheds in each cluster.

The architecture of the neural network with the best performance was achieved by taking the average of the mean square error of the network's output. The network was trained using the Bayesian regularization backpropagation training algorithm (MacKay 1992). Bayesian regularization is a network training process that updates the weight and bias values using Levenberg-Marquardt optimization. It minimizes a combination of squared errors and weights, and then determines the correct combination so as to produce a network that generalizes well. A comparison between the networks trained by regular Levenberg-Marquardt and Bayesian regularization algorithms indicated better performance for the latter. The smallest network mean square error was achieved for three hidden units for regionalization without classification and two hidden units for regionalization with classification. A tangent sigmoid ("tansig") function was used as a transfer function in both the hidden and output layers. The networks were trained 100 times and the outputs with highest performance on the training data set were selected.

Support vector machine (SVM)

SVM is a neural-network-based algorithm for solving multidimensional function estimation problems, initially developed by Vapnik (1995) for pattern recognition problems and later extended to solve non-linear regression estimation problems by the introduction of Vapnik's ϵ-insensitive loss function (Vapnik et al. 1996). Therefore, SVM can be used for classification (support vector classifier, or SVC) and regression (support vector regression, or SVR). This tool is especially useful for high-dimensional input space (in our case, 12 catchment attributes) where decision functions are based on nonlinear elements. SVM applies the structural risk minimization principle, which minimizes an upper bound of the generalization error rather than minimizing the training error. Generalization error is bounded by the sum of the training error and a confidence interval term. Therefore, SVM is expected to result in better generalization performance than other neural network models. Furthermore, the solution of SVM is unique because the training of SVM is equivalent to solving linearly constrained quadratic programming. As well, it is optimal and unaffected by local minima, unlike other network training which requires non-linear optimization and involves the risk of getting stuck in local minima. The regression function which maps a set of data points $\{(xi, di)\}_{i=1}^{n}$ in which xi is the input vector, d_{i} is the desired target value, and n is the total number of data patterns estimated by SVM can be approximated by (Tay and Cao 2001):

$$y = f(x) = w^{T\sim}\emptyset(x) + b \tag{8}$$

where $\emptyset(x)$ maps the input x to a vector in multi-feature space.

$$R_{SVMs\,(C)} = C\frac{1}{n}\sum_{i=1}^{n} L_{\varepsilon}(di, yi) + \frac{1}{2}||w||^2 \tag{9}$$

where $C\frac{1}{n}\sum_{i=1}^{n} L_{\varepsilon}(di, yi)$ is the empirical error (risk) measured by the ε-insensitive loss function$(L_{\varepsilon}(d, y))$ given in Equation (10), $\frac{1}{2}||w||^2$ is the regularization term and C is referred to as the regularized constant.

$$L_{\varepsilon}(d, y) = \begin{cases} |d-y| - \varepsilon\, |d-y| \geq \varepsilon \\ 0 \quad otherwise \end{cases} \tag{10}$$

The MATLAB code of LS-SVM (Least Square- Support Vector Machine) (Brabanter et al. 2011) was used in this study for support vector regression model development. A cross-validation procedure was used as a performance measure to determine tuning parameters (regularization and kernel parameters) in two steps. First, a global optimization technique, coupled simulated annealing (CSA), determines suitable parameters according to some criterion and parameters. These are then given to a second optimization procedure (simplex or grid search) to perform a fine-tuning step. SVR has been investigated for hydrological prediction in previous studies but rarely investigated for streamflow prediction in ungauged watersheds. In this study, the SVM is further investigated as a regionalization technique for both classified and unclassified basins, and compared with more advanced techniques such as the CPNN.

Counter propagation neural network (CPNN)

Introduced by Hecht-Nielsen (1987), CPNN consists of an input layer, a Kohonen layer, and an output layer called Grossberg outstar. The input layer performs the mapping of the multidimensional input data into a lower-dimensional array (most often two-dimensional). The mapping is performed by the use of competitive learning – often called a "winner-takes-it-all" strategy. The counter-propagation algorithm is executed in two phases: a training phase and an operational phase (classification/prediction). The training process of the CPNN connects the input vector with N variables ($x_s = x_{s,1},\ldots, x_{s,i},\ldots, x_{s,N}$) with the weight vector ($w_j = wj,_1,\ldots, w_{j,i}, \ldots, w_{j,N}$) of the neurons in the Kohonen layer. The winning (or central) neuron c is first found among the neurons in the Kohonen layer, then the weights of both Kohonen and output layers are adjusted according to the pairs of input and target vectors (x, y) using suitably selected learning rate η(t) and neighborhood function $f_{(dj-dc)}$(Kuzmanovski and Novic 2008):

$$W_{j,i}^{new} = W_{j,i}^{old} + \eta_t.f_{(dj-dc)}.(x_i - W_{j,i}^{old}) \tag{11}$$

$$u_{j,i}^{new} = u_{j,i}^{old} + \eta_t.f_{(dj-dc)}.(y_i - u_{j,i}^{old}) \tag{12}$$

where the difference $(dj - dc)$ is the topological distance between the winning neuron c and the neuron j, the weights of which are adjusted. $W_{j,i}^{old}$ and $W_{j,i}^{new}$are weights of the Kohonen layer before and after its adjustments were performed, while $u_{j,i}^{old}$ and $u_{j,i}^{new}$ are the weights of the output layer before and after the performed adjustments. The CPNN MATLAB code developed by Kuzmanovski and Novic (2008) was adapted and used in this study.

Similar to the MLP, the data set was divided into two parts: training and validation. Attributes and model parameters of two thirds of the watersheds were used for the network's training while the remaining one third was used for validation. Performing the same process three times, all the watersheds were considered as ungauged once. The attributes and parameters were normalized using the maximum and minimum values of attributes and parameters, respectively. The best values of parameters of CPNN for regression, width and length of network, parameters of rough and fine-tuning training, and shape of network were determined by a standard trial-and-error approach.

Model performance evaluation

To evaluate the performance of the regionalization models (combination of regionalization and classification techniques), in addition to mean daily streamflow (i.e. the derived daily baseflow time series), peak flow values were evaluated. The description of the estimation and evaluation methods are described in the following sections.

Evaluation of daily streamflow

To evaluate the performance of the regionalization models in daily streamflow prediction, three criteria values for daily streamflow were calculated: NSE: Equation (2); VE: Equation (3); and root mean square error (RMSE): Equation (13). NSE values close to 1, VE values close to 0, and RMSE values close to 0 indicate better model performance.

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}(Qsim - Qobs)^2}{n}} \tag{13}$$

where Q_{sim} and Q_{obs} are the simulated and observed streamflow, respectively, and N is the number of data points.

Evaluation of daily baseflow

Baseflow is derived from the streamflow series to evaluate model performance in low-flow estimation. Baseflow was separated from total streamflow using a recursive digital filter (Lyne and Hollick 1979; Nathan and McMahon 1990b) as follows:

$$f_n = a \times f_{n-1} + 0.5(1 + a)(Q_n - Q_{n-1}) \tag{15}$$

$$Qb_n = Q_n - f_n \tag{16}$$

where Qb_n, f_n and Q_n are the baseflow, the filtered quick response and the original streamflow at the n_{th} event, respectively, and a is the filter parameter which varies in the 0.92–0.99 range (Nathan and McMahon 1990b) and is set to 0.925 in this study (following Samuel et al. 2012b). NSE values are calculated for daily baseflow to evaluate the models' performance in daily baseflow prediction.

Evaluation of daily peak flow

To evaluate the performance of models in peak flow prediction, usually some threshold values are considered and model error in days with flow over that threshold is calculated. Examples of high-flow thresholds include flow-duration percentile (describing the daily mean discharge that is exceeded a given percentage of the time) or long-term median flow. We consider 33% duration flow a threshold value and the streamflow values above this threshold are considered to be high flows. The reason for selecting this threshold is to have almost the same number of days with high flow for all watersheds (highest one third of data length) and also to obtain a reasonable number of days for error calculation. VE values for the flow values above the threshold are calculated for all models.

Results and discussion

Hydrologic model parameter optimization

To calibrate hydrologic models, optimization algorithms including SCE, PSO, NSGA-II and Monte Carlo simulation were applied, with MAC-HBV and SAC-SMA against observed daily streamflow time series of gauged watersheds for 1976–1985 (10 years) as the calibration period and evaluated for 1986–1994 (9 years) as the validation period. All variable parameters of MAC-HBV (the 15 parameters) and all variable parameters (not constant values) of SAC-SMA (the 19 parameters) presented in Table 2 were optimized by changing their initial ranges while other parameters were kept constant at their average possible value. The initial and optimized ranges of the parameters of the two hydrologic models are presented in Table 2.

The box plots of NSE and VE values of the simulated daily streamflow for the validation period using optimized parameters for 90 watersheds for the two hydrologic models are presented in Figure 4. Results from the PSO and the SCE algorithms indicated equally superior mean and median of NSE and VE values with fewer outlier values for the PSO algorithm compared to other optimization methods. Further, analysis on the performance of the PSO and SCE algorithms was performed by calculating NSE values for daily baseflow and VE values for peak flows. The results indicated similar performances, but slightly better results for the PSO algorithm. Therefore, the PSO algorithm was selected as the hydrologic model optimization method for the regionalization study. Optimized parameters from the PSO algorithm for the calibration period (1976–1985) were selected for the regionalization and were transferred to hypothetical ungauged watersheds. To select the best-performing optimization algorithm for the two hydrologic models and all 90 basins, we considered their general performance and superiority based on mean and median values of NSE and VE. Given the computational cost of optimization algorithms, this mass analysis was preferred, as it can show which algorithm most likely performs better when MAC-HBV or SAC-SMA models are used.

Continuous streamflow regionalization

Daily streamflow

NSE, RMSE and VE values of simulated daily streamflow using the two hydrologic models (SAC-SMA and MAC-HBV) coupled with the four regionalization techniques (IDW, MLP, CPNN and SVM) were calculated and applied to the unclassified and classified homogeneous watersheds using the classification techniques (SOM, NLPCA and CNLPCA) for the validation period (1986–1994). Tables 3 and 4 present the statistics of NSE values of daily streamflow and daily baseflow for MAC-HBV and SAC-SMA models coupled with IDW, MLP, CPNN and SVR techniques on unclassified and classified watersheds. The mean and median of NSE values of daily streamflow for the unclassified watersheds using the IDW technique were slightly higher than the corresponding

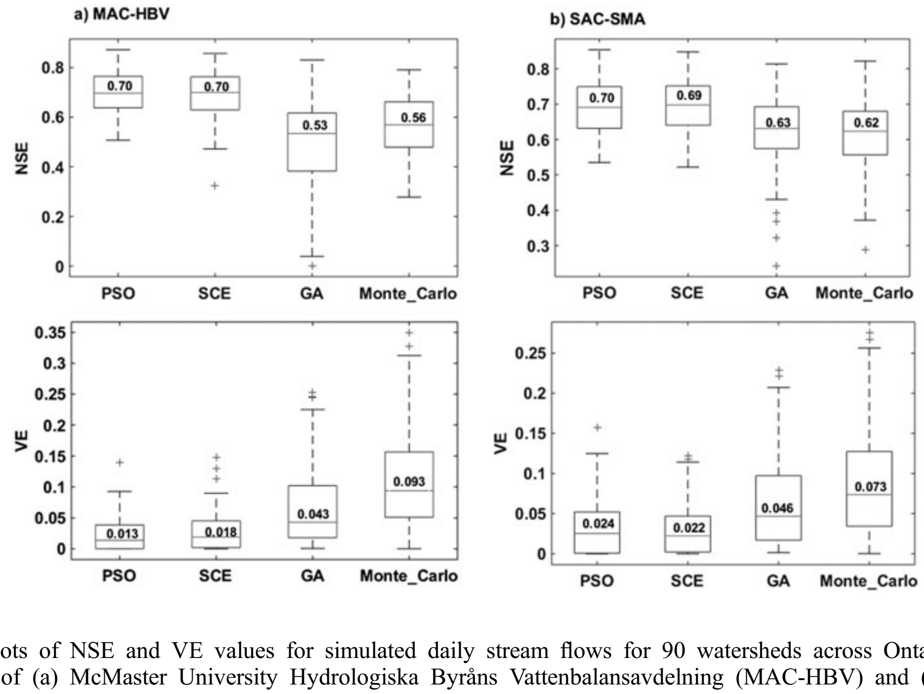

Figure 4. Box plots of NSE and VE values for simulated daily stream flows for 90 watersheds across Ontario using calibrated model parameters of (a) McMaster University Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) and (b) Sacramento Soil Moisture Accounting (SAC-SMA). These resulted from the following optimization algorithms: particle swarm optimization (PSO), shuffle complex efficiency (SCE), non-sorted genetic algorithm II (NSGA II), and Monte Carlo simulation for validation period 1986–1994. Nash Sutcliffe efficiency (NSE) values above 0.5 and volume error (VE) values between −0.1 and 0.1 are considered to be models with good performance.

Table 3. Statistics of Nash Sutcliffe efficiency (NSE) values of estimated daily streamflow using McMaster University Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) and Sacramento Soil Moisture Accounting (SAC-SMA) models coupled with the following regionalization techniques: inverse distance weighted (IDW), multi-layer perceptron (MLP), counter propagation neural network (CPNN) and support vector machine (SVM). These were applied to unclassified watersheds (Unc) and classified watersheds with self-organizing map (SOM), non-linear principal component analysis (NLPCA) and CNPLCA (compact NLPCA) for the validation period from 1986 to 1994.

Regionalization technique		**IDW**				**MLP**				**CPNN**				**SVR**			
Classification technique		Unc	SOM	NLPCA	CNLPCA	Unc	SOM	NLPCA	CNLPCA	Unc	SOM	NLPCA	CNLPCA	Unc	SOM	NLPCA	CNLPCA
MAC-HBV	Minimum	−0.75	−0.48	−0.63	−0.14	−2.20	−0.02	−0.25	−0.28	−1.21	−0.47	−0.13	−1.11	−0.82	−0.95	−2.80	−0.89
	Mean	0.44	0.45	0.45	0.47	0.35	0.47	0.43	0.41	0.26	0.41	0.43	0.42	0.29	0.31	0.34	0.33
	Median	0.49	0.48	0.51	0.50	0.43	0.51	0.47	0.45	0.32	0.44	0.46	0.47	0.33	0.36	0.44	0.43
	Maximum	0.68	0.73	0.69	0.69	0.64	0.68	0.69	0.68	0.72	0.69	0.70	0.70	0.66	0.64	0.68	0.70
SAC-SMA	Min	−1.50	−1.28	−0.41	−0.38	−0.68	−1.07	−0.53	−1.97	−2.52	−1.03	−1.66	−0.70	−3.05	−1.07	−2.85	−0.91
	Mean	0.40	0.45	0.45	0.48	0.40	0.44	0.43	0.45	0.26	0.38	0.39	0.42	0.29	0.34	0.31	0.36
	Median	0.52	0.53	0.53	0.53	0.47	0.52	0.50	0.50	0.40	0.48	0.48	0.49	0.41	0.46	0.46	0.47
	Maximum	0.70	0.71	0.72	0.71	0.70	0.68	0.70	0.68	0.71	0.69	0.70	0.69	0.68	0.66	0.69	0.67

values for the same model coupled with the MLP technique, and clearly higher than the SVM and the CPNN techniques. This shows that the MLP technique is very competitive with the IDW approach. The CPNN and SVM techniques indicated competitive performance when they were applied to the classified watersheds. For both hydrologic models, coupled with the four regionalization techniques, the average results after watershed classification were improved, although this might not be the case for some watersheds. For example, both MAC-HBV and SAC-SMA models coupled with the IDW technique reached their highest performance for classified watersheds using the CNLPCA technique with NSE mean/medians of 0.47/0.50 and 0.48/0.53, respectively. For the watersheds with negative values of NSE there is still considerable room for progress in flow prediction. The low performance of the studied techniques for some watersheds is in part due to the limitations of the technique and in part due to the low quality of available observed streamflow and catchments' physical attributes. Some modifications in the techniques such as considering additional criteria in the selection of gauged donor catchments and different architectures of neural networks might help to improve the performance of regionalization; however, this issue needs more study. We also found the percentage of improvement and deterioration of models' performance after watershed classification. Deterioration implies that watershed classification had a negative impact on regionalization performance. Table 5(a) presents the percentage of basins with more than −20% deterioration and more than 20% improvement in RMSE of daily streamflow, normalized by long-term mean daily streamflow after watershed classification. For example, the performance of the MAC-HBV model coupled with the MLP technique indicated more than −20% of deterioration in normalized RMSE after CNLPCA classification for about 17% of watersheds, while improvement was greater than 20% for 20% of the watersheds. When using CPNN as a regionalization method, MAC-HBV and SAC-SMA models reached more than 20% improvement in RMSE of daily streamflow for about 39% and 13% of basins after watershed classification using the NLPCA technique. In general, the results in Table 5(a) indicate that some combinations of hydrologic model, regionalization technique and basin classification method can yield higher improvement (> 20%) in daily streamflow estimation in most ungauged basins, while some other combinations result in a deteriorating performance. Two of the combinations (CPNN-NLPCA and CPNN-SOM), which indicated consistent improvement in daily mean, low and peak flow regionalization using MAC-HBV and SAC-SMA models in the majority of the watersheds, were further analyzed.

Figures 5 and 6 show the spatial distribution of improvements in daily streamflow regionalization using the NLPCA and SOM watershed classification techniques, combined with the regionalization technique CPNN and the MAC-HBV and SAC-SMA models, respectively. The basins which indicated a consistent improvement of > 20% in daily mean, low and peak flow regionalization are specified with a circle. Furthermore, the hydrographs of observed and simulated daily streamflows using the two hydrologic models coupled with the CPNN technique on unclassified and classified watersheds for three sample watersheds (specified in Figure 1) are presented in Figure 7. This figure shows a generally better performance of models after watershed classification.

Daily baseflow and peakflow

According to Table 4, the MLP and CPNN techniques become very competitive with the IDW technique (on average) when applied to the classified basins. In general, the performance of models was better for daily baseflow compared to daily streamflow. Similar to daily streamflow, the improvements in NSE values of daily baseflow regionalization on average were more significant for the CPNN and SVR techniques. Table 5(b) presents the percentage of watersheds with more than −20% deterioration and more than 20% improvement in VEs of daily baseflow for regionalization models after watershed classification. The baseflow regionalization was improved more than 20% in a higher percentage of watersheds (more than 40%) when using the MLP, CPNN and SVR techniques after watershed classification. Table 5(c) indicates the percentage of watersheds with more than −20% deterioration and more than 20% improvement in VE of daily peak flow. More than 20% improvement in VE of daily peak flow can be achieved in about 61% and 47% of the watersheds when the CPNN and SVM are applied, respectively, to basins classified with the NLPCA method, while there will be a deterioration of more than −20% in 26 and 19% of watersheds, respectively. Figures 5 and 6 show the spatial variability of improvements in VE of daily baseflow and peak flow regionalization using watershed classification techniques NLPCA and SOM, combined with regionalization techniques and the CPNN, MAC-HBV and SAC-SMA models.

Discussion and hydrologic implications

The hydrologic behavior of all basins for the period 1976–1994 (19 years) was evaluated by determining the timing of monthly low and peak flow and the FDC (Flow Duration Curve) slope, Q95/Q5 (95% flow duration, 5% flow duration). The shape of FDC has been shown to depend on some watershed attributes such as hydrogeology (Patel 2006) and drainage for agriculture and urban land use (Smakhtin and Toulouse 1998; Booker and Snelder 2012). The timing of low and peak

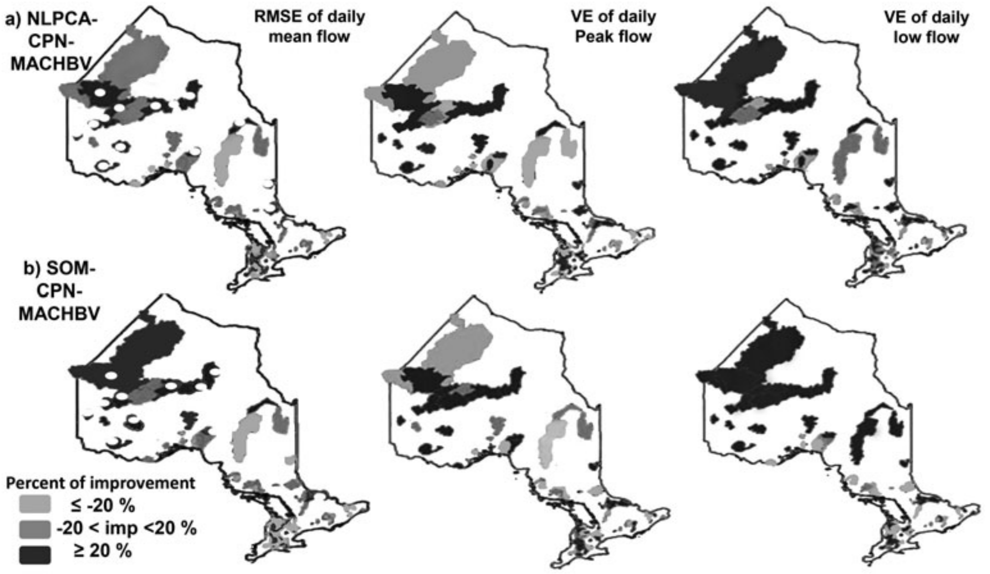

Figure 5. Schematic maps of the spatial distribution of improvements in Root Mean Square Error (RMSE) and Volume Error (VE) of daily streamflow, low flow and peak flow regionalization using the counter propagation neural network (CPNN) technique on the McMaster University Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) model after (a) non-linear principal component analysis (NLPCA) and (b) self-organizing map (SOM) classification techniques. Small circles identify the basins with a consistent improvement of > 20% in daily mean, low and peak flow regionalization after watershed classification.

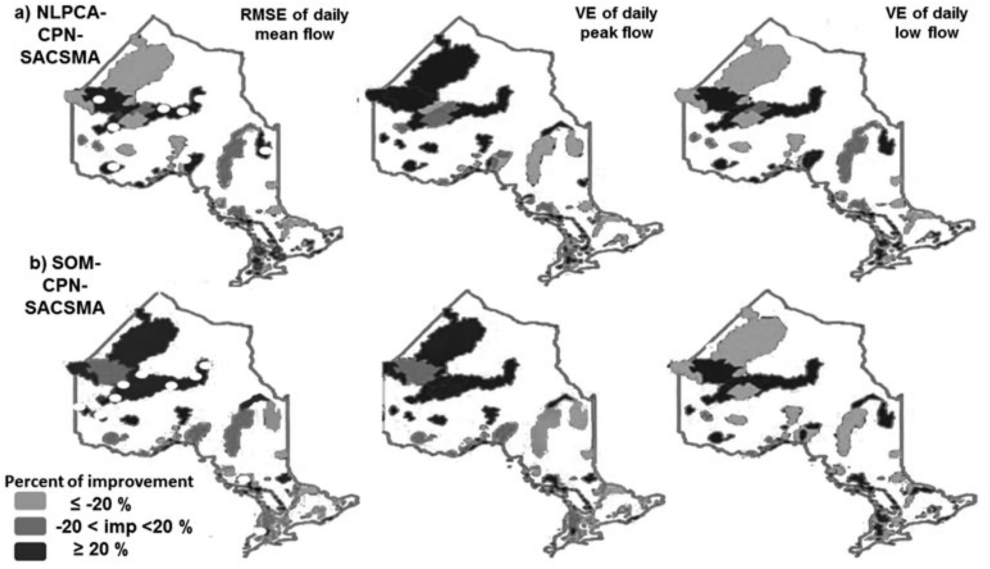

Figure 6. Schematic maps of the spatial distribution of improvements of Root Mean Square Error (RMSE) and Volume Error (VE) in daily streamflow, low flow and peak flow regionalization, using the counter propagation neural network (CPNN) regionalization technique on the Sacramento Soil Moisture Accounting (SAC-SMA) model after (a) non-linear principal component analysis (NLPCA) and (b) self-organizing map (SOM) classification techniques. Small circles identify the basins with a consistent improvement of > 20% in daily mean, low and peak flow regionalization after watershed classification.

flow is governed by the climate, which acts as a first-order control in Ontario (Stainton and Metcalfe 2007), while hydrograph shape is reflected in the low and peak points of the mean monthly flow hydrograph. The hydrographs of monthly mean flow (Figure 8) indicate that in small southern watersheds, the lowest mean monthly flow generally occurs in July/August with an early spring snowmelt peak in March/April. In northern watersheds, the lowest mean monthly flow generally happens in March and spring snowmelt reaches a peak in May/June.

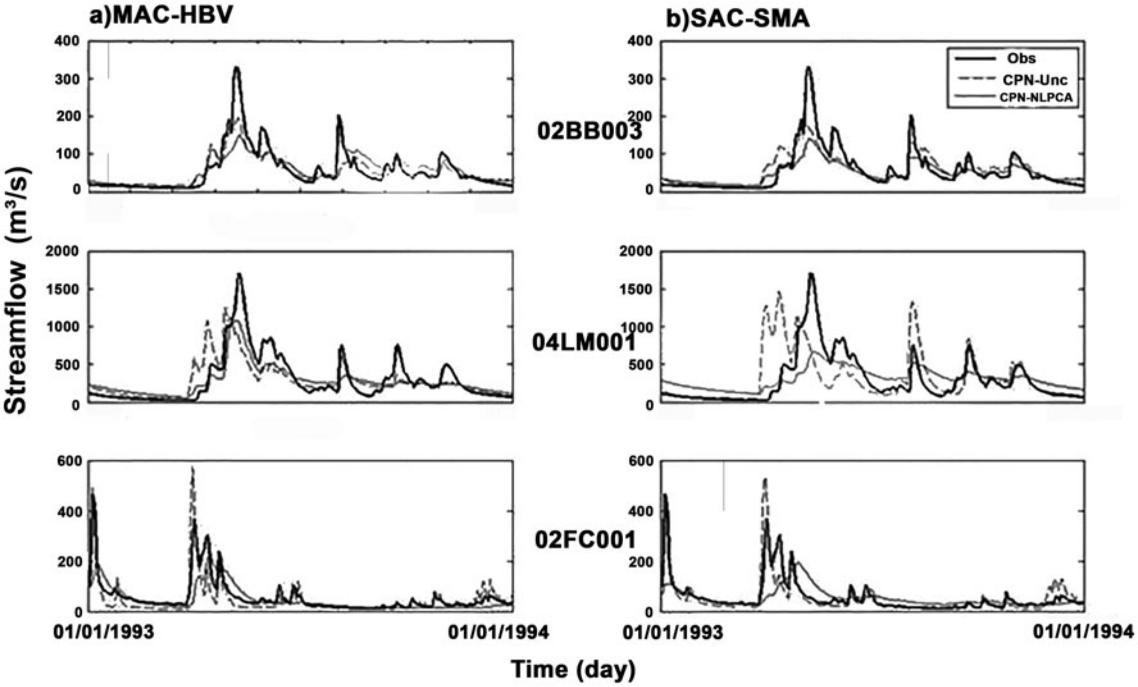

Figure 7. Observed and simulated streamflow using the following hydrologic models: (a) McMaster University Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) and (b) Sacramento Soil Moisture Accounting (SAC-SMA) models coupled with counter propagation neural network (CPNN) technique on unclassified (CPN-Unc) and classified watersheds using non-linear principal component analysis (NLPCA) (CPN-NLPCA) on three sample watersheds specified in Figure 1.

Finally, in central watersheds, the lowest mean monthly flow generally happens in February/March and the spring snowmelt peak occurs in April/May.

Figure 9 shows the spatial distribution of watershed land cover and the FDC slope. Considering the maps in Figures 5 and 6, the spatial variability of the FDC slope (Figure 9) indicates that in northern watersheds and some southern watersheds with a higher FDC slope, a higher improvement in regionalization can be achieved after watershed classification. The spatial variability of watershed land cover (Figure 9) indicates that in southern watersheds with a lower percentage of forest, a higher improvement in streamflow regionalization can be achieved compared to the other southern basins. In central watersheds where the most deterioration was apparent, forest cover is relatively high; in watersheds with a higher percentage of area covered by rapid drainage areas and glacio-deposits, more improvement in regionalization was achieved after watershed classification. Therefore, it can be concluded that among the investigated approaches, nonlinear watershed classification techniques, and the SOM and NLPCA coupled with the CPNN as a regionalization technique, are more likely to improve daily streamflow regionalization in watersheds displaying these characteristics: monthly low flow in March, spring snowmelt peak flow in May/June, high FDC slope, less area covered by forest, and more area covered by rapid drainage class and glacio-deposits.

According to Figures 5 and 6, regardless of the hydrologic model used, most of the northern watersheds (except the largest one) reach > 20% improvement after watershed classification consistently for mean, low and high flows. In small southern watersheds, this improvement is less frequent, while in most of the central watersheds and some southern watersheds, deterioration is more apparent. The early results of our study (not presented) indicate that although the IDW technique outperforms the other machine-learning techniques on average, and for unclassified watersheds, it failed to perform satisfactorily for the large sparse northern watersheds. This could be foreseen, since the IDW technique is based on the distance between watersheds and it is expected to perform better in regions with dense watersheds such as southern Ontario. We also found that the applied watershed classification techniques could improve the performance of machine-learning regionalization techniques for sparse large northern watersheds more than for the dense southern watersheds. This implies that, in general, for areas with dense and more similar watersheds, IDW can outperform machine-learning techniques, while machine-learning techniques such

Table 4. Nash Sutcliffe efficiency (NSE) values of estimated daily baseflow using McMaster University Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) and Sacramento Soil Moisture Accounting (SAC-SMA) models coupled with the following regionalization techniques: inverse distance weighted (IDW), multi-layer perceptron (MLP), counter propagation neural network, (CPNN) and support vector machine (SVM). These are applied to unclassified watersheds (Unc) and classified watersheds with self-organizing map (SOM), non-linear principal component analysis (NLPCA) and CNPLCA (compact NLPCA) for the validation period from 1986 to 1994.

Regionalization technique		**IDW**				**MLP**				**CPNN**				**SVR**			
Classification technique		Unc	SOM	NLPCA	CNLPCA	Unc	SOM	NLPCA	CNLPCA	Unc	SOM	NLPCA	CNLPCA	Unc	SOM	NLPCA	CNLPCA
MAC-HBV	Minimum	−0.31	−0.43	−0.65	−0.44	−0.73	−0.32	−0.80	−0.29	−1.15	−0.66	−0.40	−0.68	−0.69	−0.55	−1.09	−0.69
	Mean	0.48	0.48	0.48	0.50	0.37	0.50	0.45	0.44	0.27	0.44	0.46	0.47	0.32	0.36	0.38	0.40
	Median	0.51	0.52	0.54	0.54	0.47	0.52	0.51	0.49	0.33	0.47	0.51	0.53	0.36	0.43	0.49	0.49
	Maximum	0.60	0.73	0.72	0.70	0.66	0.71	0.74	0.70	0.73	0.72	0.72	0.73	0.67	0.67	0.70	0.71
SAC-SMA	Minimum	−0.91	−0.37	−0.92	−0.36	−1.26	−1.33	−1.41	−1.20	−1.89	−1.32	−0.83	−0.98	−1.37	−0.96	−1.69	−0.80
	Mean	0.46	0.50	0.39	0.52	0.42	0.47	0.45	0.48	0.33	0.42	0.44	0.45	0.34	0.40	0.39	0.41
	Median	0.57	0.58	0.53	0.58	0.52	0.57	0.54	0.52	0.51	0.51	0.53	0.56	0.43	0.50	0.50	0.51
	Maxiimum	0.75	0.76	0.78	0.76	0.71	0.72	0.72	0.71	0.73	0.72	0.73	0.72	0.71	0.68	0.70	0.70

Table 5. Percentage of basins with deterioration more than −20% and improvement greater than 20% in (a) root mean square error (RMSE) of daily streamflow, (b) volume error (VE) of daily baseflow, (c) VE of daily peak flow, using the following classification techniques: non-linear principal component analysis (NLPCA), CNLPCA (compact NLPCA), and self-organizing map (SOM), applied prior to the following regionalization techniques: inverse distance weighted (IDW), multi-layer perceptron (MLP), counter propagation neural network (CPNN) and support vector machine (SVM), and on hydrologic models McMaster University Hydrologiska Byråns Vattenbalansavdelning (MAC-HBV) and Sacramento Soil Moisture Accounting (SAC-SMA).

(a) RMSE of daily streamflow

Regionalization	**IDW**			**MLP**			**CPNN**			**SVM**		
Classification	**NLPCA**	**CNLPCA**	**SOM**	**NLPCA**	**CNLPCA**	**SOM**	**NLPCA**	**CNLPCA**	**SOM**	**NLPCA**	**CNLPCA**	**SOM**
MAC-HBV												
< −20%	8	3	9	3	17	17	7	43	12	13	14	11
> 20%	4	3	3	16	20	20	39	10	37	16	23	18
SAC-SMA												
< −20%	12	1	3	7	7	9	8	10	14	11	9	18
>20%	2	11	8	4	10	6	13	17	19	10	9	12
(b) VE of daily baseflow												
MAC- HBV												
< −20%	34	16	17	32	38	36	33	44	36	33	34	38
> 20%	41	23	16	53	47	56	56	41	54	51	49	41
SAC-SMA												
< −20%	51	26	23	34	38	35	33	44	44	31	37	42
> 20%	11	24	23	48	50	41	49	36	48	46	49	46
(c) VE of daily peak flow < −20% or > 20%												
MAC-HBV												
< −20%	26	11	12	41	30	22	26	39	20	19	21	29
> 20%	31	19	16	31	43	47	61	36	52	47	47	36
SAC-SMA												
< −20%	32	14	17	26	29	18	27	29	44	31	37	42
> 20%	14	12	11	39	29	42	37	37	48	46	49	46

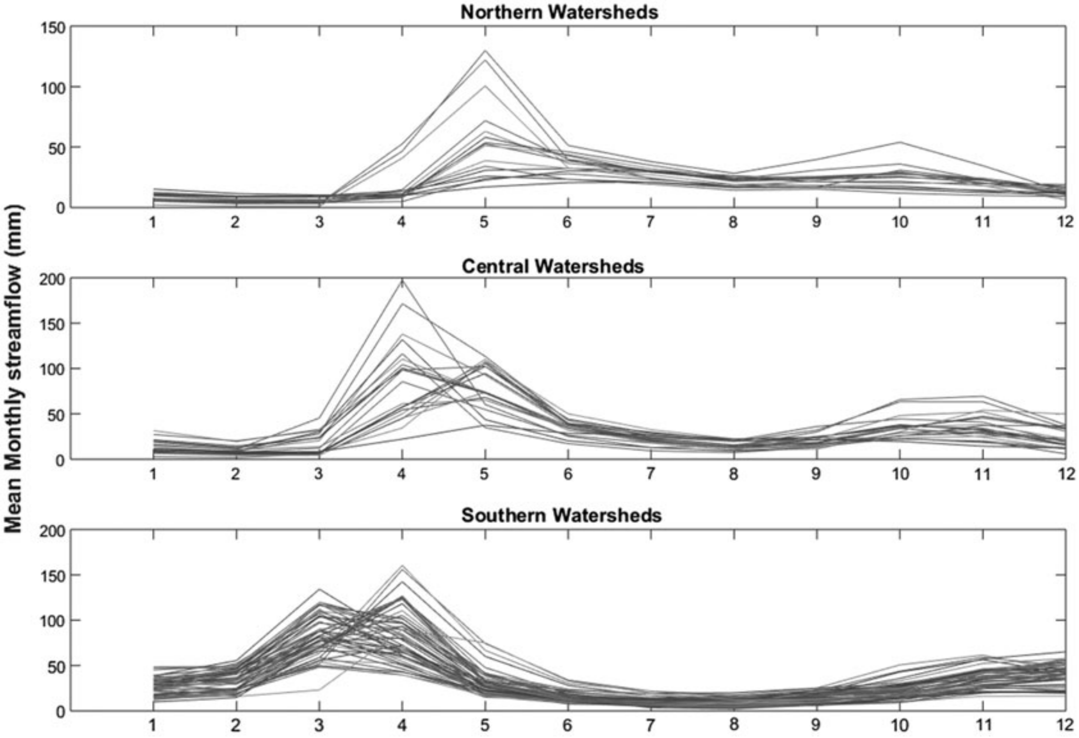

Figure 8. The hydrograph of mean monthly flow in northern, central and southern watersheds of Ontario.

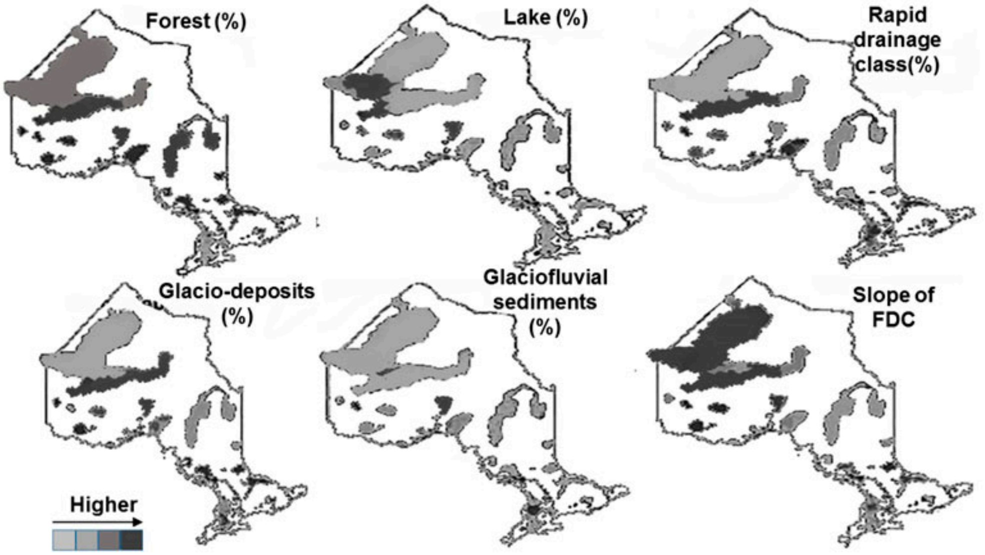

Figure 9. Spatial variability of percentage of land covered by: forest, lake, rapid drainage class, glacio-deposit, and glaciofluvial sediments and slope of FDC (Flow Duration Curve) in selected Ontario watersheds.

as MLP can perform better for sparse and large watersheds. Furthermore, improvement in regionalization performance after watershed classification is more evident for the latter.

Conclusion

In the current study, 90 watersheds across Ontario (Canada) were used to assess the benefit of classified homogeneous basins in the regionalization of continuous daily streamflow. Four regionalization techniques (IDW, MLP, CPNN, SVM) with two hydrologic models (MAC-HBV, SAC-SMA) were applied to the watersheds classified with nonlinear data-driven classification techniques (NLPCA, CNLPCA, SOM) and to the unclassified watersheds.

The study results indicate that the MLP technique was very competitive with the IDW technique, which was identified in a previous study as the best regionalization method in the study area. The more complicated types of neural networks, CPNN and SVR, became competitive when they were applied to the classified watersheds. It is shown that the combination of watershed classification and regionalization techniques for a certain hydrologic model can improve the performance of daily streamflow, baseflow and peak flow regionalization in most of the watersheds. However, this combination may not be the best one for some watersheds. For example, each of the hydrologic models coupled with the CPNN technique in combination with the NLPCA or SOM as a classification technique revealed a clear improvement in daily streamflow, baseflow and peak flow regionalization. The results of this study reveal that in general the nonlinear data-driven techniques are more likely to improve the performance of daily streamflow regionalization in watersheds with high FDC slope, less area covered by forest, more area covered by rapid drainage and glacio-deposits, monthly low flow in March and spring snowmelt peak flow in May/June.

Moreover, the improvement of regionalization results for daily baseflow was higher compared to daily streamflow and peak flows. This can have positive implications for environmental flow determination, which is based on baseflow. The accurate baseflow estimation for ungauged basins is still a challenging task. This study suggests that an appropriate combination of regionalization technique, hydrologic model and basin classification method can provide substantially improved streamflow and baseflow estimates at ungauged basins across Ontario. Furthermore, it appears that neural networks as dynamic nonlinear methods are capable of accounting for nonstationarity due to urbanization and climate variability in the hydrological modelling for ungauged watersheds. The potential of neural networks for nonstationary hydrological time series modeling has been documented by Coulibaly and Baldwin (2005). However, the data quality and availability issue has always been a main concern in applying data-driven methods. The selected 90 basins used in this study have been passed through a quality control of available physiographic attributes and hydro-meteorological data. Overall, it is shown that a good combination of methods can provide improved streamflow estimation at ungauged basins across Ontario, which can be particularly useful for water resources managers.

Acknowledgements

This work was partially funded by the Natural Science and Engineering Research Council (NSERC) of Canada, and the Ontario Ministry of Natural Resources. The authors would like to thank Dr. Jos Samuel and Dr. Robert Metcalfe for their help in collecting the required data. The authors thank the reviewers and the editor for their comments and suggestions that helped to improve the manuscript substantially.

Funding

This work was supported by the Natural Science and Engineering Research Council (NSERC) of Canada.

References

Bergström, S. 1976. Development and application of a conceptual runoff model for Scandinavian catchments. Lund,Sweden: Lund Institute of Technology/Univ. of Lund, A (52).

Besaw, L. E., M. R. Donna, P. R. Bierman, and R. H. William. 2010. Advances in ungauged streamflow prediction using artificial neural networks. *Journal of Hydrology* 386 (1–4): 27–37.

Blöschl, G., and M. Sivapalan. 1995. Scale issues in hydrological modelling: A review. *Hydrological Processes* 9 (3–4): 251–290.

Booker, D. J., and T. H. Snelder. 2012. Comparing methods for estimating flow duration curves at ungauged sites. *Journal of Hydrology* 434–435: 78–94.

Brabanter, K. D., P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. Moor, J. Vandewalle, and J. A. K. Suykens. 2011. *LS-SVMlab Toolbox User's Guide version 1.8*, Katholieke Universiteit Leuven ESAT-SISTA. Technical Report: 10–146.

Burn, D. H., and D. B. Boorman. 1993. Estimation of recharge and runoff volumes from ungauged catchments in eastern Australia. *Journal of Hydrology* 143: 429–454.

Burnash, R. J. C., R. L. Ferral, and R. McGuire. 1973. *A generalized streamflow simulation system: Conceptual modeling for digital computers*. Technical report, Joint Fed. State River Forecast Center, Sacramento.

Cavadias, G. S., T. Ouarda, B. Bobee, and C. Girard. 2001. A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins. *Hydrological Sciences* 46 (4): 499–512.

Clerc, M. 2006. *Particle swarm optimization*. ISTE (International Scientific and Technical Encyclopedia) doi:10.1002/9780470612163.

Coulibaly, P., and C. K. Baldwin. 2005. Nonstationary hydrological time series forecasting using nonlinear dynamic methods. *Journal of Hydrology* 307: 164–174.

Deb, K., A. Pratap, S. Agarwal, and T. A. M. T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation, IEEE Transactions on* 6 (2): 182–197.

Di Prinzio, M., A. C. Castellarin, and E. Toth. 2011. Data-driven catchment classification: Application to the pub problem. *Hydrology and Earth System Sciences* 15 (6): 1921–1935.

Drogue, G., and J. Plasse. 2014. How can a few streamflow measurements help to predict daily hydrographs at almost ungauged sites? *Hydrological Sciences Journal* 59: 12. doi:10.1080/02626667.2013.86503.

Duan, Q., S. Sorooshian, and V. K. Gupta. 1994. Optimal use of the SCEUA global optimization models. *Journal of Hydrology* 158: 265–284.

Eberhart, R.C., and J. Kennedy. 1995. A new optimizer using particle swarm theory. In *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, 39–43. Piscataway, N. J.: IEEE Press.

Environment Canada. 2004. *HYDAT CD-ROM Version 2.04: Surface water and sediment data*. Saskatoon, Saskatchewan: Univ. of Saskatchewan.

Environmental Systems Research Institute Canada, Inc. (ESRI). 1997. *ESRI ArcCanada: Schools and libraries Version 1.0* [Computer files]. Redlands, CA: ESRI.

Hecht-Nielsen, R. 1987. Counterpropagation networks. *Applied Optics* 26 (23): 4979–4984.

Ilorme, F., and V. W. Griffis. 2013. A novel procedure for delineation of hydrologically homogeneous regions and the classification of ungauged sites for design flood estimation. *Journal of Hydrology* 492: 151–162.

Kileshye Onema, J. M., A. E. Taigbenu, and J. Ndiritu. 2012. Classification and flow prediction in a data-scarce watershed of the equatorial Nile region. *Hydrology and Earth System Sciences* 16 (5): 1435–1443.

Kuzmanovski, I., and M. Novič. 2008. Counter-propagation neural networks in Matlab. *Chemometrics and Intelligent Laboratory Systems* 90: 84–91.

Lyne, V. D., and M. Hollick. 1979. *Stochastic time-variable rainfall runoff modeling, hydrology and water resources symposium*, 89–92. Perth: Institution of Engineers Australia.

MacKay, D. J. C. 1992. A practical bayesian framework for backpropagation networks. *Neural Computation* 4 (3): 448–472.

Maier, H. R., A. Jain, G. C. Dandy, and K. P. Sudheer. 2010. Methods used for the development of neural networks for the prediction of water resources variables: Current status and future directions. *Environmental Modelling and Software* 25: 891–909.

Merz, R., and G. Blöschl. 2004. Regionalisation of catchment model parameters. *Journal of Hydrology* 287: 95–123.

Mishra, A. K., and P. Coulibaly. 2009. Development in hydrometric networks design: A Review. *Reviews of Geophysics* 47 (2): 1–24.

Nash, J., and J. Sutcliffe. 1970. River flow forecasting through conceptual models, part 1: A discussion of principles. *Journal of Hydrology* 10 (3): 282–290.

Nathan, R. J., and T. A. McMahon. 1990a. Identification of homogeneous regions for the purposes of regionalization. *Journal of Hydrology* 121: 217–238.

Nathan, R. J., and T. A. McMahon. 1990b. Evaluation of automated techniques for base flow and recession analyses. *Water Resources Research* 26 (7): 1465–1473.

Patel, J. A. 2006. Evaluation of low flow estimation techniques for ungauged catchments. *Water and Environment Journal* 1–6.

Razavi, T., and P. Coulibaly. 2012. Streamflow estimation in ungauged basins: Review of regionalization methods. *Journal of Hydrologic Engineering* 18 (8): 958–975.

Razavi, T., and P. Coulibaly. 2013. Classification of Ontario watersheds based on physical attributes and streamflow series. *Journal of Hydrology* 493: 81–94.

Samuel, J., P. Coulibaly, and R. A. Metcalfe. 2011. Estimation of continuous streamflow in Ontario ungauged basins: Comparison of regionalization methods. *Journal of Hydrologic Engineering* 16 (5): 447–459.

Samuel, J., P. Coulibaly, and R. A. Metcalfe. 2012a. Evaluation of future flow variability in ungauged basins: Validation of combined methods. *Advances in Water Resources* 35: 121–140.

Samuel, J., P. Coulibaly, and R. A. Metcalfe. 2012b. Identification of rainfall–runoff model for improved baseflow estimation in ungauged basins. *Hydrological Processes* 26 (3): 356–366.

Seibert, J. 1999. Regionalisation of parameters for a conceptual rainfall runoff model. *Agricultural and Forest Meteorology* 98–99: 279–293.

Shepard, D. 1968. A two-dimensional interpolation function for irregularly-spaced data. Proc., ACM National Conf., New York: 517–524.

Smakhtin, V. Y., and M. Toulouse. 1998. Relationships between low-flow characteristics of South African streams. *Water SA* 24 (2): 107–112.

Stainton, R. T., and R. A. Metcalfe. 2007. Characterisation and classification of flow regimes of natural rivers in Ontario to support the identification of potential reference basins. Waterpower Project Science Transfer Report 7.0, Ontario Ministry of Natural Resources, Ontario, Canada.

Sveriges Meteorologiska och Hydrologiska Institut (SMHI). 2005. *Integrated hydrological modelling system (IHMS) Manual Version 5.8*. Norrköping, Sweden: SMHI, 109.

Tay, F. E., and L. Cao. 2001. Application of support vector machines in financial time series forecasting. *Omega* 29 (4): 309–317.

Vapnik, V. 1995. *The nature of statistical learning theory.* N.Y.: Springer Verlag, 189.

Vapnik, V. N., S. E. Golowich, and A. J. Smola. 1996. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems* 9: 281–287.

Viviroli, D., and J. Seibert. 2015. Can a regionalised model parameterisation be improved with a limited number of runoff measurements? *Journal of Hydrology* 529 (1): 49–61. doi:10.1016/j.jhydrol.2015.07.009.

Vrugt, J. A., H. V. Gupta, S. C. Dekker, S. Sorooshian, T. Wagener, and W. Bouten. 2006. Application of stochastic parameter optimization to the Sacramento soil moisture accounting model. *Journal of Hydrology* 325 (1–4): 288–307.