available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jhydrol

Feasibility study of a geostatistical modelling of monthly maximum stream temperatures in a multivariate space

Nicolas Guillemette [a,*], André St-Hilaire [a,d], Taha B.M.J. Ouarda [a], Normand Bergeron [b,d], Élaine Robichaud [c], Laurent Bilodeau [c]

[a] *Chair in Statistical Hydrology, INRS-ETE, Université du Québec, 490 de la Couronne, Québec, Canada G1K 9A9*
[b] *INRS-ETE, Université du Québec, 490 de la Couronne, Québec, Canada G1K 9A9*
[c] *Hydro-Québec, 855 Ste-Catherine Street East, Montreal, Québec, Canada H2L 1A4*
[d] *Centre Interuniversitaire de recherche sur le saumon Atlantique Sacré-Cœur, Québec, Canada*

Received 28 February 2008; received in revised form 29 September 2008; accepted 6 October 2008

KEYWORDS
Modelling;
Temperature;
River;
Multivariate;
Geostatistical;
Kriging

Summary Healthy river conditions through optimal thermal regime controls water quality as well as the availability and distribution of fish habitat. A multivariate and geostatistical approach was developed to estimate maximum stream temperatures at a large basin scale. The methodology relies on the construction of a physiographical space using characteristics of gauging stations by testing two multivariate methods: principal components analysis (PCA) and canonical correlation analysis (CCA). Within the physiographical space, a geostatistical technique called ordinary kriging was then used to interpolate stream temperatures. Data from 12 temperature monitoring stations during July 1996 and July 1997 were used to estimate monthly maximum temperature. Results from the proposed approach were evaluated by comparing kriging performance obtained using both multivariate methods. Cross-validation technique has been performed on both approaches and satisfactory results were obtained. Kriging in the CCA physiographical space leads to better results because this approach seems more adapted to link physiographical information with specific water temperature. In addition, CCA requires less physiographical information than PCA (i.e. 10 metrics for PCA vs 8 metrics for CCA) to provide more satisfactory results (up to 15% decrease in RMSEr). In physiographical space, the gauging stations were found to cluster, potentially providing information to improve the accuracy of interpolation in that space. An example is provided to illustrate how to estimate one of the stream temperature properties at ungauged stations

[*] Corresponding author. Tel.: +418 654 2530 4457; fax: +418 654 2600.
E-mail addresses: nico_guillemette@hotmail.com, nicolas.guillemette@ete.inrs.ca (N. Guillemette).

doi:10.1016/j.jhydrol.2008.10.002

using the PCA algorithm. The relevance of the results regarding the quality of fish habitats of the Moisie river is discussed.

Introduction

In recent years, global climate change is being considered as a potential threat for several fish species (Sinokrot et al., 1995; Eaton and Scheller, 1996; Crozier and Zabel, 2006). The integrity of many physical and bio-chemical characteristics of an ecosystem is mainly controlled by temperature. The thermal regime of rivers controls water quality as well as the availability and distribution of fish habitat (Caissie, 2006). During warm months, unusually high stream temperature can occur naturally or as a result of human activities. Such extreme events are often linked to flow reduction, increased incident radiation and high air temperature (Sinokrot and Gulliver, 2000). Deforestation (Johnson and Jones, 2000; Chen et al., 1998) has often been identified as having a negative impact on stream temperature (e.g. Holtby, 1988; St-Hilaire et al., 2000).

A modified thermal regime can affect a number of poikilotherm fish such as Atlantic salmon (*Salmo salar*). For example, parr can be submitted to a thermal sublethal stress when water temperature exceeds 23 °C through interrupted mRNA induction, which is an important process in protein synthesis (Lund et al., 2002). Hodgson and Quinn (2002) found that Sockeye salmon (*Oncorhynchus nerka*) spawning can be interrupted or delayed when water temperature rises above a threshold of 19 °C in northwestern USA. When this threshold is reached, adults start to seek thermal refugia. Therefore, improving our understanding of the thermal regime of rivers and our capacity to predict and simulate high temperatures by developing more accurate and flexible models applicable across watersheds are essential steps to identify the best management framework for aquatic resources and fisheries managers.

Many models have been developed and used to estimate stream temperature. Two broad categories are usually identified: (i) deterministic or physical models (e.g. Caissie et al., 2007; St-Hilaire et al., 2003a; Gu and Li, 2002) and (ii) statistical/stochastic models (e.g. Caissie et al., 1998; Mohseni et al., 1998; Ahmadi-Nedushan et al., 2007). Deterministic models use a conceptual approach which is based on the thermal exchange between atmosphere, the body of water and sometimes the river bed. Meteorological parameters such as air temperature, wind velocity and solar radiation are used as inputs to calculate energy budget equations and are also important to predict water temperature variations. For this reason, deterministic models are quite flexible in terms of input parameters modification, but also quite demanding in terms of model development and data requirement. As an alternative, statistical/stochastic approaches are based on a mathematical relationship between water temperature and independent variables such as air temperature and flow. This second model category requires fewer input data than deterministic models and model development can be relatively simple (e.g. linear or nonlinear regressions; Stefan and Preud'homme, 1993; Mohseni et al., 1998). Statistical/stochastic models are most often developed for a specific point or station on a river. As such, they cannot easily be transferred to another point in the river or another stream. Recent development of low cost temperature recorders makes it possible to sample water temperature at many locations with relatively good accuracy and high frequency.

Previous work on abiotic variables, such as water temperature, have attempted to make use of a number of independent physiographical/climatic variables via statistical analyses that explained the spatial and temporal variability (Collings, 1973; Mosley, 1982; Miyazawa et al., 1982; Hawkins et al., 1997; Arscott et al., 2001; Scott et al., 2002). Water temperature has also been characterized by spatial correlation (Peterson and Sickbert, 2006). One approach for modelling this type of spatial correlation is kriging (Isaaks and Srivastava, 1989). Ordinary kriging, a very popular geostatistical approach, consists of quantifying the spatial correlation structure between stations as a function of separation distances. The spatial interpolation at any point uses a weighted combination of neighbours. For example, Gardner et al. (2003) have considered a set of temperature recorders located throughout of Beaverkill Watershed in southwestern New York to estimate river temperature at ungauged points on the same system. They used kriging to interpolate directly in a physical space by considering three metrics to calculate separation distances using a total of 72 temperature loggers. This resulted in a one-dimensional model of water temperature as a function of one of these three metrics: the shortest path between loggers, distances calculated along the stream network (river kilometer) and distances weighted by stream order.

The aim of the present study is to expand from the work of Gardner et al. (2003) and include a larger number of metrics to interpolate water temperature. In contrast with previous studies, a relatively sparse network of loggers is placed on a large scale hydrographical system and a physiographical space rather than a geographical interpolation space has been created by using multivariate approaches (Manly, 2004). This concept, first developed by Chokmani and Ouarda (2004) for flood quantiles, consists in combining physiographical and climatic information from drainage basin using principal component analysis (PCA) or canonical correlation analysis (CCA) to define a multivariate orthogonal interpolation space and then use ordinary kriging to interpolate water temperature in the newly created principal components (PCA) or canonical correlation (CCA) space. Therefore, the objectives of this study are: (1) to elaborate a reliable spatial stream temperature interpolation model on a large scale basin using a multivariate geostatistical approach. (2) To compare results between two different multivariate approaches (PCA and CCA) to determine which is the most suitable for water resources managers in predicting stream temperature based on model performance.

Data and study site

Thermographs were deployed on the Moisie and Ste-Marguerite rivers drainage basins on Québec North Shore, Canada (Fig. 1), which are respectively large drainage basins of 19197 km^2 and 6711 km^2. The majority of the loggers were deployed on the drainage basin of the Moisie River which discharges a mean annual flow of 466 m^3 s^{-1} into the St-Lawrence Estuary. It drains Lake Menistouc, which is located in the upper part of the drainage basin and then runs for 363 km. Main triburaries are Carheil, Nipissis and Aux Pécans rivers. In the province of Québec (Canada) the Moisie River is home to the most important spawning grounds for Atlantic salmon. Moreover, it is considered by anglers as one of the most important salmon rivers in the province because of the high average weight of individual adult salmon.

A total of 20 temperature monitoring stations were established by Hydro-Québec, the main provincial hydroelectric authority, during the period of 1989—1998. Daily temperatures were calculated from hourly observations provided by Hydro-Québec. As a test case for the methodology, we decided to focus on observations periods corresponding to warmer months of the year. There were a number of missing values during the first part of the observation period (1989—1992) and for this reason we decided to focus on the second part of the observation period (1992—1998). Two years with the greatest number of concomitant time series for a minimum of 10 stations were selected for this study (1996 and 1997). Descriptive statistics used in the models are: monthly maximums of daily temperatures and maximum daily range, the latter being only used in CCA (see Section 'Kriging in CCA physiographical space'). The focus on maxima stems from the potential interest of having a method that can be useful in defining areas of thermal stress or thermal refugia. Twelve stations with monthly maximum temperatures of July 1996 and July 1997 have been selected as test cases for modelling in multivariate space. Table 1 shows the station name with their associated monthly maximum temperatures for both time periods.

By extracting a large number of metrics, more information is known on the watershed. Multiple combinations of physiographical variables can then be used to influence the spatial variation of the thermal regime and to interpolate water temperatures. The choice of those variables is simply based on data availability and their potential to influence stream temperature across the watershed. In the present study, and in accordance with watershed management terminology, we identified geographical metrics such as the latitude, the longitude and the azimuth, physical variables such as the drainage area and the mean slope, and hydrographical variables such as the river kilometer and the stream order. Thus, a total of 18 physiographical variables characterizing each station were estimated using a Geographic Information System. They are reported in Table 2.

Statistical methods

In this study, the interpolation of stream temperature is based on the use of the basins coordinates in a physiographical space rather than a geographical space. Stream temperatures in the geographical space may change dramatically

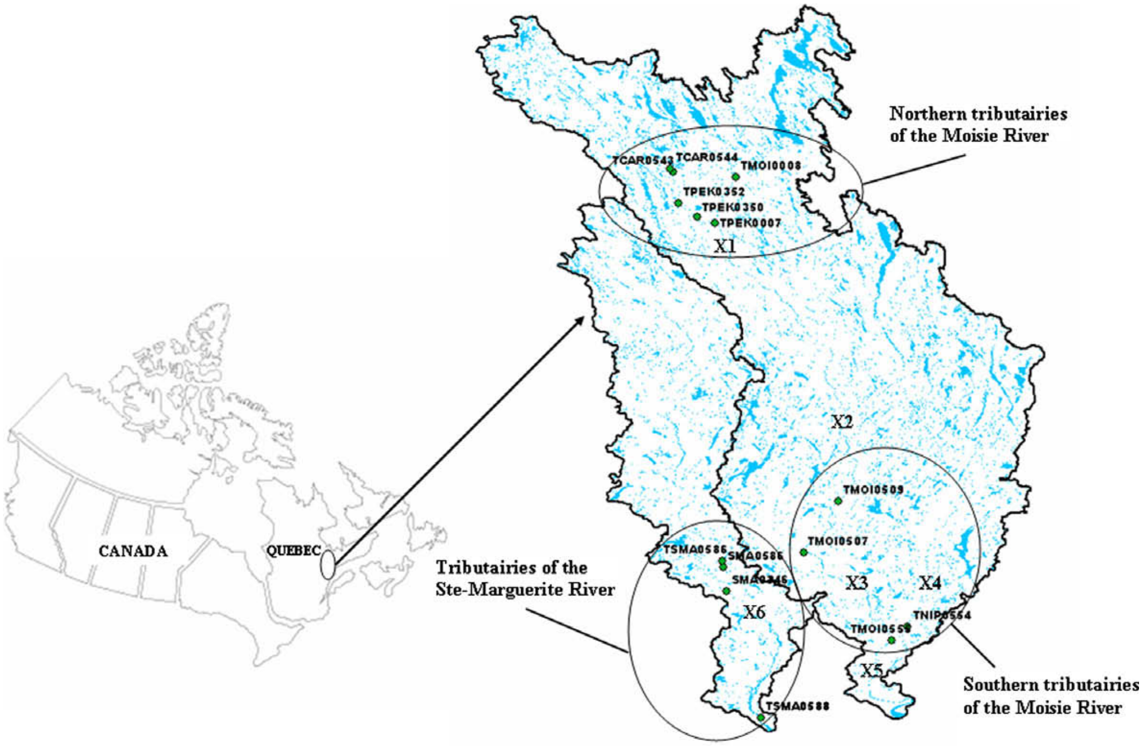

Figure 1 Location and map of the Moisie and Ste-Marguerite rivers drainage basins with temperature monitoring stations grouped into three broad geographical regions. X marks indicate locations of ungauged stations.

Table 1 Station names and their associated broad geographical regions and monthly maximum temperatures. Data used to construct kriged maps of July 1996 and 1997.

July 1996			July 1997		
Station name	Geographical region	*T* (°C)	Station name	Geographical region	*T* (°C)
TCAR_0543	North Moisie	18.36	TCAR_0544	North Moisie	18.35
TPEK_0352	North Moisie	17.50	TCAR_0543	North Moisie	18.67
TPEK_0350	North Moisie	18.55	TPEK_0352	North Moisie	17.14
TPEK_0007	North Moisie	18.29	TPEK_0350	North Moisie	17.99
TMOI_0008	North Moisie	18.06	TPEK_0007	North Moisie	18.42
TMOI_0509	South Moisie	17.19	TMOI_0008	North Moisie	18.50
TMOI_0507	South Moisie	16.58	TMOI_0509	South Moisie	17.72
TMOI_0555	South Moisie	17.63	TMOI_0507	South Moisie	16.41
TNIP_0554	South Moisie	17.49	TNIP_0554	South Moisie	16.88
TSMA_0586	Ste-Marguerite	17.03	TSMA_0586	Ste-Marguerite	17.43
SMA_0586	Ste-Marguerite	16.99	SMA_0586	Ste-Marguerite	17.42
SMA_0346	Ste-Marguerite	17.22	TSMA_0588	Ste-Marguerite	17.06

Table 2 Physiographical variables characterizing each station estimated using GIS.

Metrics	Units	Notation
Latitude	UTM	LAT
Longitude	UTM	LONG
Drainage area	km^2	DA
Distance to the closest tributary from station	m	CT
Distance to the closest lake from station	m	CL
Area of the closest lake	m^2	ACL
Mean stream azimut	°	MSA
Mean stream azimut of all upstream tributaries	°	MSAT
Azimut between station and closest lake	°	ASL
Altitude at the station	m	A
Maximum altitude of drainage basin	m	MAXA
Mean altitude of the drainage basin	m	MEANA
Mean slope of the drainage basin	%	MS
Local slope at the station	%	S
Stream order		SO
River kilometer	km	RK
Forest cover on drainage basin	%	FC
Percentage of area covered by lakes and marches	%	PLM

over adjacent drainage basins. In fact, while stream temperatures are discontinuous in the geographical space, they can be regarded as continuous variables in the physiographical space, thereby permitting the use of interpolation techniques. Different approaches are possible to construct the physiographical space. Chokmani and Ouarda (2004) proposed two multivariate approaches called principal component analysis (PCA) and canonical correlation analysis (CCA) to realize a study on regional flood frequency estimation. They used both methods respectively to simplify complex data sets as well as describing relationship of dependence existing between hydrological and physiographical variables. These approaches had never been tested on abiotic habitat variables such as water temperature. Therefore this study focuses on the feasibility of using PCA and CCA to establish a multivariate coordinate system.

Principal component analysis (PCA) is a statistical technique that linearly transforms an original set of variables into a substantially smaller set of uncorrelated (orthogonal) variates, called principal components, that represent most of the information of the original data set (Dunteman, 1989). Each principal component is a linear combination of the original variables. Geometrically, the first principal component is the line of closest fit to the *n* observations in the *p*-dimensional variable space. Algebraically, the first principal component, PC1, is a linear combination of the original standardized variables x_i.

$$\mathrm{PC}_1 = \sum_{i=1}^{p} a_{1i} x_i \tag{1}$$

In Eq. (1), the variance explained by PC1 is maximized under the constraint that the sum of the squared weights (e.g. a_{1i}) is equal to one.

$$\sum_{i=1}^{p} a_{1i}^2 = 1 \tag{2}$$

The basics statistics of principal components analysis are the k variances (eigenvalues) and the associated variable weight vectors (eigenvectors) $a_1, \ldots, a_k$. The correlations of the variables with a particular principal component are called the loadings. The sum of the squared correlations for each principal component equals the amount of variance explained by each PC. Therefore, the first principal component accounts for the greatest percentage of the variation in the original variables and the explained variance gets smaller as successive principal components are calculated.

In some multivariate analyses, the variables divide naturally into two groups. A canonical correlation analysis (CCA) can then be used to examine the relationship between these two sets of random variables. Given a set of hydrological variables, X, (stream temperature in our case) and a set of physiographical variables, Y, characterizing each station location of X, CCA aims to link two sets using vectors of canonical variates. It involves searching for linear combinations of X variables $(U_1, U_2, \ldots, U_\mathrm{i})$ that have the maximum possible correlation with linear combinations of Y variables $(V_1, V_2, \ldots, V_\mathrm{i})$ (Manly, 2004). This is somewhat similar to the concept behind a principal components analysis, except that in this case, correlation is maximized instead of explained variance. Algebraically, a pair of sample canonical variates is a pair of linear combinations U and V (Eq. (3)) having unit sample variances that maximize the ratio called the sample canonical correlation (Eq. (4))

$$\begin{aligned} U &= a'X \\ V &= b'Y \end{aligned} \tag{3}$$

$$r_{U,V} = \frac{a'S_{12}b}{\sqrt{a'S_{11}a}\sqrt{b'S_{22}b}} \tag{4}$$

In Eq. (3), a and b represent the canonical coefficients vectors for the first and second set of random variables (X and Y), respectively. In addition, S_{12}, S_{11} and S_{22} are the sample covariance matrices consistent with the case of the initial variables. In general, the kth pair of sample canonical variates is the pair of linear combinations U_k, V_k among those linear combinations uncorrelated with the previous $k-1$ sample canonical variates (Johnson and Wichern, 2007).

Once PCA and CCA spaces are obtained, monthly maximum temperatures can be projected into these physiographical spaces to be interpolated using kriging. This geostatistical approach quantifies the spatial correlation structure between stations as a function of separation distance. An experimental correlogram, or its inverse, the semivariogram, is first established using water temperature measurements and Euclidian distances measured using the coordinates establishes by the first two PCA or CCA variates. The semi-variance (or covariance) structure is estimated by the experimental semivariogram:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (z_{x_i+h} - z_{x_i})^2 \tag{5}$$

where $N(h)$ is the number of data pairs at a separation distance h which have an observed value z_{x_i}. A model is then fitted to the experimental semivariogram such as spherical, exponential, or Gaussian functions with three parameters: the nugget effect (C_0), the sill (c) and the range (a). The nugget effect describes the occurrence of discontinuity at the origin of the semivariogram that may be caused by dissimilar sample values at short inter-station distances (Isaaks and Srivastava, 1989). The sill is the plateau reached by $\hat{\gamma}h$, which indicates a value of semi-variance that is a threshold beyond which there is essentially no spatial structure in the data. Finally, the range represents the distance over which the observed values are correlated.

The kriging estimator is a weighted average of the observed values $z(x_\mathrm{i})$ which is used to estimate the value of $z(x_0)$, identified at a specific location x_0 where there are no measured values. The model is denoted

$$\hat{z}(x_0) = \sum_{i=1}^{N(h)} \lambda_i z(x_i) \tag{6}$$

Where λ_i are the weights of the estimator that minimize the variance of the estimation error (ordinary kriging weights). By using the spatial structure defined by the theoretical semivariogram, a kriging system of linear equations combining neighbouring information can be defined as

$$\sum_{i=1}^{N(h)} \lambda_i \gamma(x_i - x_j) + \upsilon = \gamma(x_i - x_0) \tag{7}$$

under the constraint on weights:

$$\sum_{i=1}^{N(h)} \lambda_i = 1 \tag{8}$$

where υ is the Lagrange multiplier for the constraint on the weights. The values of λ_i are obtained by solving this linear kriging system. In the present study, the geostatistical software GS+ (Gamma Design Software, 2007) has been used to solve the kriging system.

Results

During July 1996 and July 1997, 12 stations were available to perform the multivariate analysis (Table 1). In order, to construct the CCA and PCA spaces, the first step was to choose a maximum number of physiographical variables significantly correlated with monthly maximum temperatures for these two specific time periods. As an example, significant (P-value < 0.05) correlation coefficients for July 1997 are reported in Table 3 and a maximum of 11 physiographical variables were selected for this month. Similar results were found for July 1996. Smaller errors and stable models for both years were obtained by using 10 of the most significant correlated variables for PCA and the eight most significantly correlated variables for CCA.

Kriging in PCA physiographical space

Figs. 2–4 illustrate the PCA results. Fig. 2 shows the loadings, i.e. the projection of the physiographical information on the PCA space of the two first principal components, which explain respectively 60.4 percent and 26.2 percent of the total variance, for July 1996. Similarly, the two first principal components for July 1997 explain respectively 63.6 percent and 20.7 percent of the total variance. PC1 (the x axis) is dominated in both cases by the geographical

Table 3 Correlation coefficients between physiographical variables and monthly maximum water temperatures of July 1997 and associated *P*-values.

Metrics	Notation	Correlation coefficient	*P*-value
Latitude	LAT	0.74	0.0055
Longitude	LONG	−0.59	0.0417
Drainage area	DA	−0.61	0.0337
Area of the closest lake	ACL	0.70	0.0109
Altitude	A	0.58	0.0489
Maximum altitude of drainage basin	MAXA	−0.66	0.0198
Mean slope of the drainage basin	MS	−0.67	0.0176
Stream order	SO	−0.60	0.0392
River kilometer	RK	−0.65	0.0220
Forest cover on drainage basin	FC	−0.66	0.0189
Percentage of area covered by lakes and marches	PLM	0.69	0.0139

coordinates (LAT and LONG) and by drainage basin characteristic such as river kilometer (RK) and drainage area (DA). On the other hand, PC2 in both cases is most strongly associated with forest cover (FC) and drainage basin topography such as altitude at the station (A) and the maximum altitude of the drainage basins (MAXA).

In order to determine and measure the multivariate spatial structure of maximum water temperatures, the isotropic experimental semivariograms were calculated in accordance with work set scores, i.e. the projection of temperature monitoring station on the PCA physiographical space of the two first principal components. Fig. 3 presents an example of the semivariogram calculated for July 1996 using separation distances in PCA coordinates. The adjusted fitted semivariogram is an exponential function. Monthly maximum water temperatures for each station were projected in PCA space for each period and ordinary kriging was performed to obtain interpolated temperatures for the entire PCA space. Fig. 4 shows the results obtained for July 1996 and July 1997 where stations have been grouped into three broad geographical regions

- Triangle: tributaries of the Ste-Marguerite River
- Circle: northern tributaries of the Moisie River
- Square: southern tributaries of the Moisie River

Note: the three broad geographical regions are identified on Fig. 1.

The position of each station in the PCA space helps to define which stations or broad geographical regions are more associated with warmer or colder maximum water temperature in July. This exercise can help characterizing the area of thermal stress or identifying the area of thermal refuge under various hydrological conditions. The kriged maps (Fig. 4) provide the basis for estimating water temperature maxima for any location in the drainage basins under study and probably on adjacent basins within region. For example, in July 1996, lower maximum temperature conditions, which may eventually become thermal refugia for poïkilotherm fish, are found for PC1 values ranging between 2.08 and 3.36 and for PC2 values located between −0.73 and −2.18. PC1 values ranging between −2.40 and −1.12 and PC2 values between −0.52 and 0.52 characterize areas with the highest temperatures that could become sites of potential thermal stress. Conversely, Fig. 1 shows a few ungauged points (X mark) of the river system where the stream temperatures were not monitored. By extracting physiographical information associated with each ungauged stations of Fig. 1, it becomes possible to find the PCA coordinates (score values) needed to estimate monthly maximum water temperatures using kriged maps. Table 4 shows the physiographical variables extracted as well as the estimated monthly maximum water temperatures at each ungauged points using PCA algorithm and kriged maps. The range of estimated temperatures at ungauged stations is in accordance with measured values in the geographical sub-region of July 1996 (Table 1).

Kriging in CCA physiographical space

In comparison to the construction of the PCA space, the CCA space is relatively different. First, because CCA involves searching for linear combinations between two sets of variables by maximising the correlation instead of variance, it becomes impossible to use all the significantly correlated physiographical variables. Canonical correlation analysis must restrict the number of degrees of freedom. In fact, model parsimony dictates that the number of metrics be as small as possible without deteriorating the correlation structure between the two groups of variables (Stevens, 1986). Best results were obtained using 10 original variables for July 1996 and 11 original variables for July 1997 to construct the CCA space. In comparison with PCA analysis where score values are obtained only with physiographical variables, CCA analysis requires water temperatures and associated variates (first set of variables) as well as physiographical variables (second set of variables) to extract score values. The July 1996 water temperature variables used in CCA analysis were monthly maximum temperature and monthly maximum daily range. Because a more comprehensive data set was available in 1997, the July 1997 CCA uses the monthly maximum water temperature of July, August and September as the first set of variables. In both cases, the 8 physiographical variables used as second group of variables to compute the canonical correlation analysis were the 8 most significantly correlated variables (smallest *P*-value) of Table 3.

Sample canonical correlations (i.e. correlations between the linear combinations of the two sets of variables) are always largest for the first pair of canonical variates (U_1, V_1).

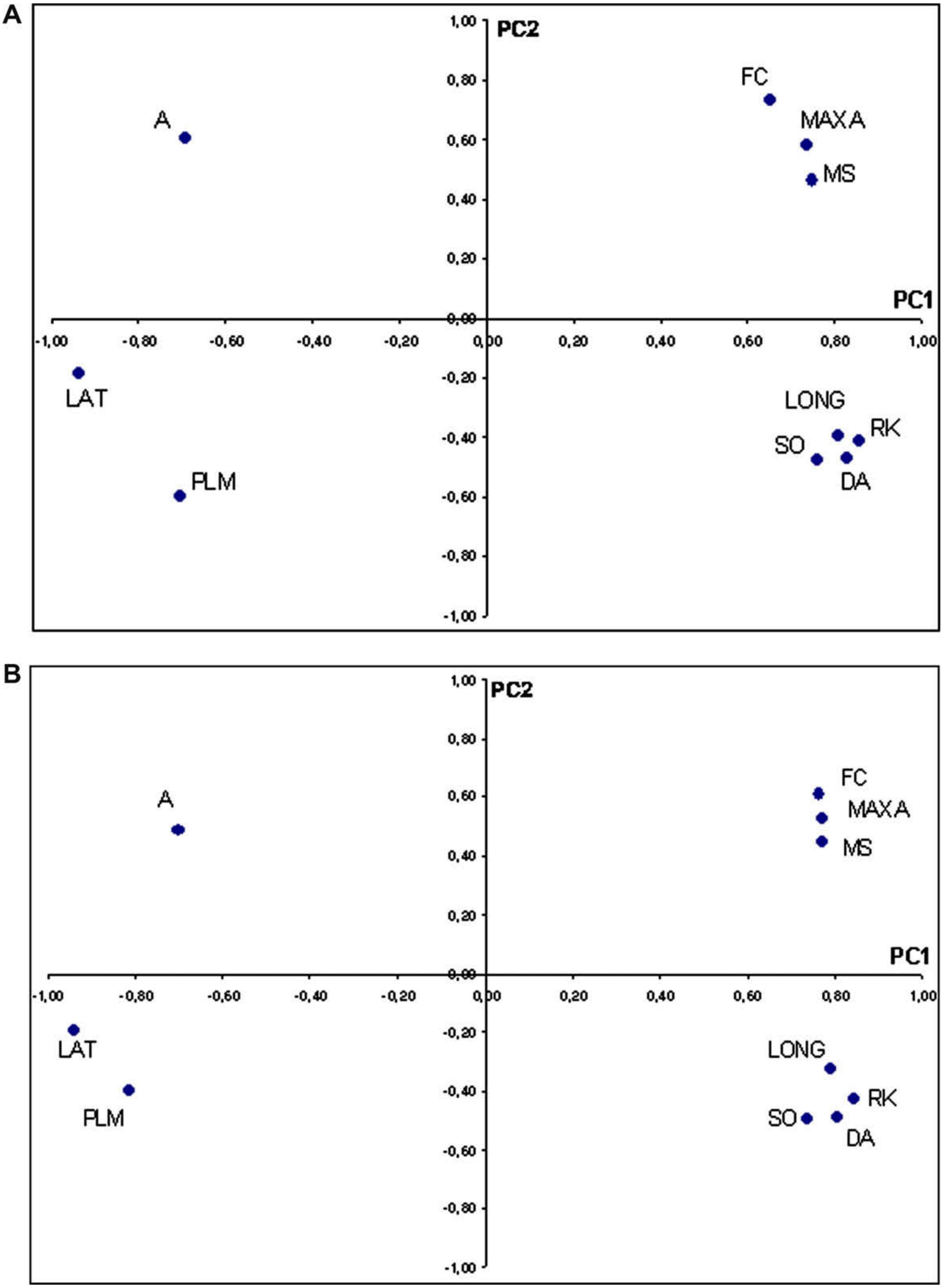

Figure 2 Projection of significantly correlated physiographical variables in PCA space. Work set loading. (A) July 1996 and (B) July 1997.

Canonical correlations were 0.99 for July 1996 and July 1997. For the second pair of canonical variates (U_2, V_2), the canonical correlations were 0.86 and 0.98 for July 1996 and July 1997 respectively. Hence, the V_1 and V_2 cartesian space is highly correlated with water temperature and it was selected as the CCA kriging space.

As for PCA, semivariograms were calculated using the inter-station distances in CCA space. Fig. 5 shows the semi-

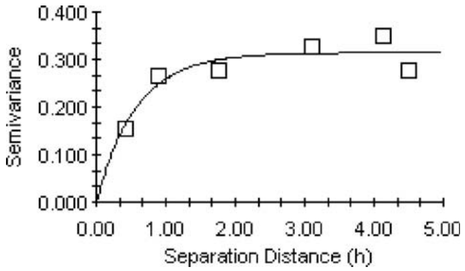

Figure 3 Experimental semivariogram of July 1996 used to krige in PCA space. Exponential model with nugget = 0.0001, range = 1.725 and sill = 0.314.

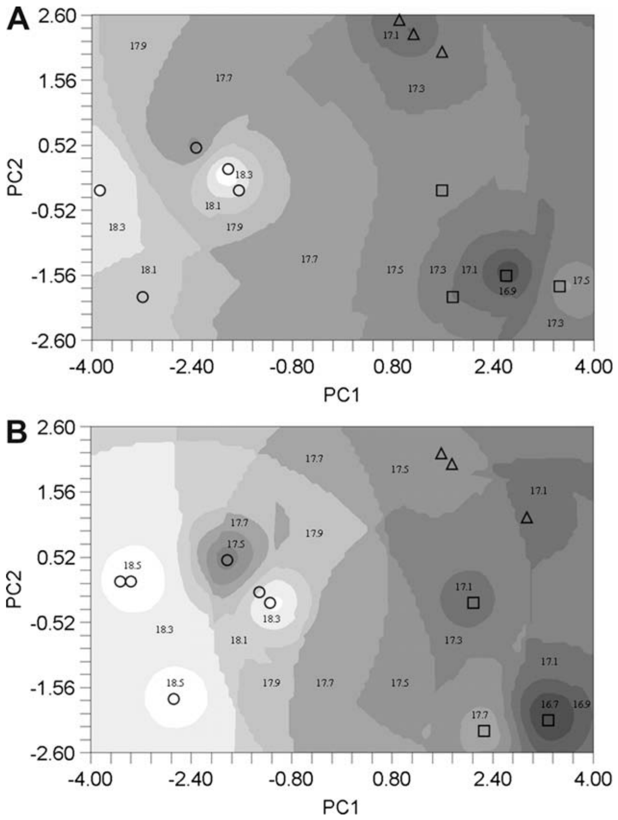

Figure 4 Interpolated monthly maximum water temperature in PCA space performed with ordinary kriging. Triangle: tributaries of the Ste-Marguerite River. Circle: northern tributairies of the Moisie River. Square: southern tributaries of the Moisie River (see Fig. 1 to identify geographical regions). (A) July 1996 and (B) July 1997. *Note*: station locations were not identical for both years (see Table 1).

variogram obtained for July 1996, using a gaussian function as fitted model. Ordinary kriging was performed for the entire CCA space by projecting monthly maximum water temperatures for July 1996 and July 1997. Fig. 6 presents the results with groups of stations (tributaries of the Ste-Marguerite River, northern tributairies of the Moisie River and southern tributaries of the Moisie River) identified and projected in the interpolation space. The spatial pattern of stations is less structured than in PCA space. The three broad geographical regions are still recognizable but the three clusters are less defined with CCA. The cluster which represents the northern tributaries of the Moisie River (circle) is well circumscribed in contrast with the other stations characterizing the southern tributaries of the Moisie River (square) and the tributaries of Ste-Marguerite River (triangle) which are less ordered. However, they are still located in a same area of the CCA space.

Kriging performances

A cross validation using leave-one-out resampling was used to estimate the error associate with the interpolation of monthly maximum water temperature in PCA and CCA space. This validation technique eliminates temporarily a station from the sample and the value for this observation is then estimated using remaining stations. This procedure was repeated for the whole station set. The relative mean bias (BIASr) as well as the relative root mean square error (RMSEr) were used as performance evaluation criteria. These indicators are defined as follows:

$$\mathrm{BIASr} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i - \hat{z}_i}{\Delta z_{\max}}\right) \qquad (9)$$

$$\mathrm{RMSEr} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i - \hat{z}_i}{\Delta z_{\max}}\right)^2} \qquad (10)$$

where z_i and $\hat{z}_i$ are respectively the observed value and the estimated value at station i. Because the range of observed maximum temperature was relatively small in July 1996 and July 1997, the rBIAS and the rRMSE were redefined using this maximum range ($\Delta z_{\max} = z_{i\ \max} - z_{i\ \min}$) of both years as the denominator. Eqs. (9) and (10) are a more conservative interpretation of error than the usual definition of BIASr and RMSEr, because $\Delta z_{\max}$ is of typically less than 1.5 °C, while the typical temperature in July was more than 10 times $\Delta z_{\max}$. According to Table 5, BIASr and RMSEr varied, respectively, between -2.5% and 1.1% and 5.2% and 20.9% of the range of measured values (i.e. $\Delta z_{\max}$, see Eqs. (9) and (10)). Kriging in CCA space produced lower RMSEr than PCA, especially for July 1996.

Discussion

This study focused on the characterization of the river thermal regime at the drainage basin scale using multivariate and geostatistical approach. The use of PCA and CCA was shown by Chokmani and Ouarda (2004) to provide an adequate means of summarizing climatic and/or physiographical variability of drainage basins, while producing at the same time a cartesian space in which interpolation of hydrological extremes can be performed. The results shown in the present study confirm that a similar space can be used to interpolate water temperature extremes, which can be of interest for many water resources management issues such as fish habitat. The present study focused only on monthly maximum temperatures of July 1996 and July 1997 as a test case to construct kriged maps. However, other temperature statistics could be used such as daily mean, daily minimum or daily range to characterize any temperature events of short or long duration.

Table 4 Physiographical variables extracted to estimate monthly maximum stream temperatures using the PCA algorithm at ungauged stations (Fig. 1) for July 1996.

Stations	LAT	LONG	DA	A	MAXA	MS	SO	RK	FC	PLM	PC1 (score)	PC2 (score)	*T* (°C)
X1	5771721	652543	6272.4	325.0	904	5.30	7	5967.2	82.1	12.6	−3.36	−1.32	18.1
X2	5683390	689416	10507.7	152.0	915	6.79	7	10451.4	85.2	10.6	−0.61	−1.50	17.6
X3	5623021	693833	13786.4	46.0	986	9.09	7	14154.6	86.6	10.2	1.24	−1.22	17.4
X4	5606334	709213	4124.7	46.0	901	12.69	6	4573.5	87.8	10.9	−0.16	0.11	17.6
X5	5590114	700899	18871.3	18.0	1009	10.50	7	19594.0	87.1	10.2	2.55	−1.49	16.7
X6	5607757	666115	5196.9	65.0	1055	11.36	6	5893.8	7.2	91.6	0.67	2.37	17.1

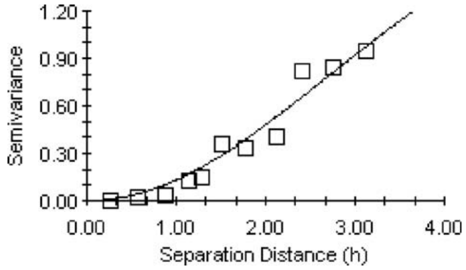

Figure 5 Experimental semivariogram of July 1996 used to krige in CCA space. Gaussian model with nugget = 0.001, range = 6.639 and sill = 2.011. Search radius was limited to $h < 3.2$.

PCA kriging appears to be somewhat less accurate than interpolation in CCA space. This result is similar to those obtained by Fernandez and Saenz (2003) as well as Chokmani and Ouarda (2004) with hydrological variables. However, as expected PCA provides more definite clusters of stations because the underlying criterion is the maximization of the explained physiographical variance. An examination of Fig. 4 reveals the three clusters which define specific areas of the drainage basin of the Moisie and Ste-Marguerite rivers. Therefore, the PCA space is a better representation of physiographical contrasts between regions. When comparing July 1996 and July 1997, it can be seen that each cluster was characterized by a specific range of temperatures. Stations from the same cluster have a similar thermal regime associated with a specific area of the drainage basin. The first two broad geographical regions constituted by the stations from tributaries of the Ste-Marguerite River (triangle) and southern Moisie River (square) are located on both maps of Fig. 4 in a part of the PCA space where the monthly maximum temperature varies between 16.6 and 17.7 °C. On the other hand, the region defined by the stations from the northern tributaries of the Moisie River (circle) were found to be projected in an area where the monthly maximum water temperatures were warmer than the two other clusters with temperatures oscillating between 17.3 and 18.5 °C. This reverse north to south gradient can be partly explained by the proximity of the northern stations to the upper lakes which become relatively warm during hot summer months. Moreover, the southern stations from Ste-Marguerite and Moisie rivers are more influenced by a good number of upstream cold tributaries during the same period.

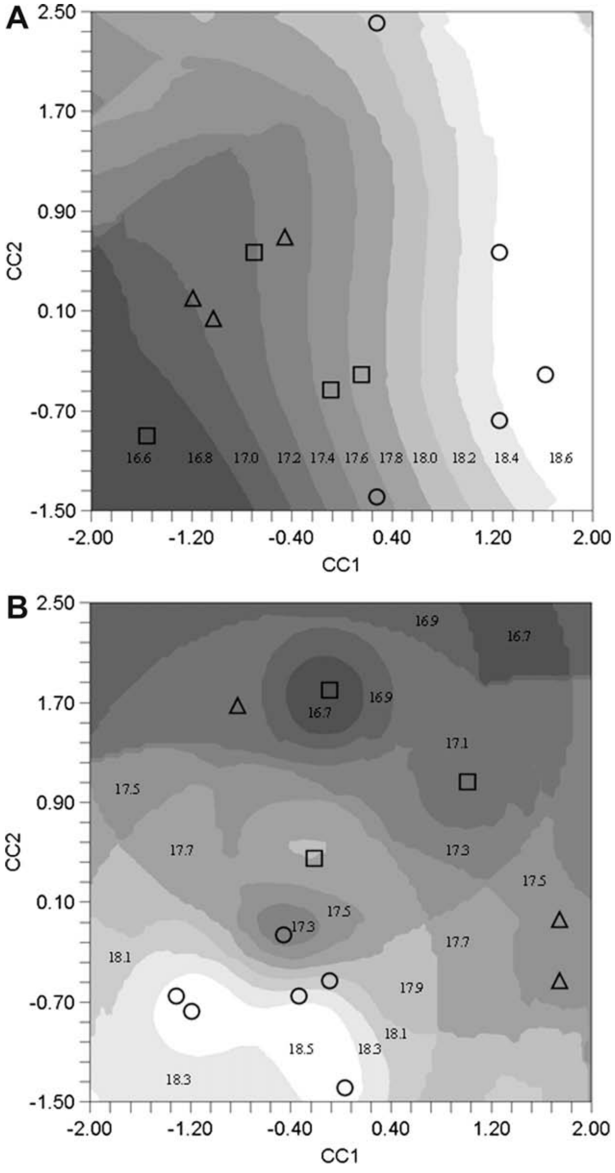

Figure 6 Interpolated monthly maximum water temperature in CCA space performed with ordinary kriging. Triangle: tributaries of the Ste-Marguerite River. Circle: northern tributairies of the Moisie River. Square: southern tributaries of the Moisie River (see Fig. 1 to identify geographical regions). (A) July 1996b and (B) July 1997. *Note*: station locations were not identical for both years (see Table 1).

Table 5 Cross–validation results.

	Kriging in the PCA physiographical space		Kriging in the CCA physiographical space	
	July 1996	July 1997	July 1996	July 1997
BIASr (%)	−0.38	−2.03	1.14	−2.54
RMSEr (%)	20.89	19.71	5.17	18.49

For July 1996 and 1997, PCA reveals that the latitude (LAT) and the river kilometer (RK) are most strongly associated with PC1, while the percentage of forest cover (FC) as well as metrics related to altitude (A and MAXA) are associated to PC2. Hence the geographical position as well as the degree of ramification of stream within each drainage basin appears to be suitable metrics for the interpolation of temperature maxima. This conclusion is in accordance with the results of Gardner et al. (2003) who used similar metrics in their 1D interpolation. In addition, the percentage of forest cover, strongly associated with PC2 is in fact, an important metric to characterize the quantity of solar radiation. On the other hand, the maximum altitude of the drainage basin (MAXA) and the altitude at the station (A) are two metrics that influence directly air temperature and therefore indirectly affect water temperature. Such additional information could not readily be included in a 1D interpolation scheme.

It is interesting to note that, despite of the fact that CCA does not maximize explained physiographical variance, the CCA kriged maps of July 1996 and 1997 (Fig. 6) generally show the same station clusters than those obtained with PCA. Therefore, both approaches were able to reproduce three regions with distinct thermal signatures. For both time periods analyzed in our study, CCA presented RMSEr values lower than those obtained with PCA (Table 5). One potential reason for this difference is that PCA is restrained to maximizing the variance along the physiographical space irrespectively of water temperature variability. For this reason, the CCA technique is more adapted to link physiographical information with specific stream temperature. Canonical correlations can be a very useful tool for a priori estimation of the links between physiographic features and thermal regime. It is interesting to note that CCA kriging produced better results with 8 metrics while PCA required 10 explanatory variables. The extraction of metrics using GIS can be time consuming and thus, it becomes relatively advantageous to optimize the number of metrics used and to reduce the degree of model complexity. It should be noted that attempts to use less than 8 or 10 metrics, respectively, for CCA and PCA lead to a decrease in model performance, in terms of RMSEr values. In fact, model parsimony dictates that the number of metrics must be as small as possible without deteriorating the correlation structure which is directly associated with semivariogram calibration and kriging performances.

To demonstrate the improvement of this modelling technique over more traditional approaches, we performed multiple linear regressions to estimate maximum water temperature for July of both years using the same independent variables (Table 3) than those selected for PCA and CCA. We found respectively RMSEr and BIASr values of 99% and −42% for July 1996 and values of 47% and 10.9% for July 1997. Moreover, in the study realized by Joseph et al. (2007), it was shown that kriging in CCA space with 21 stations led to better results than a simple linear regression (RMSEr values of 54.4% versus 59.6% and BIASr of −13% versus 20%) for the estimation of the mean annual streamflow. In both cases, kriging in multivariate space with a relatively low number of stations gives better result than regression analysis in terms of error estimations and shows less bias in the model.

In this present study, kriging was performed with a sparse network of stations (12) on a relatively large territory. With so few measurements, it becomes difficult fitting a semivariogram model. Attempts to further decrease the number of stations led to a very poor fit of the theoretical semivariogram and therefore no modelling capability. In fact, the variability in the performances of the model observed for CCA between 1996 and 1997 is explained by the variation of the quality of the semivariogram fitting between these two years. Even if all metrics were significantly correlated with monthly maximum temperatures, with so few measurements, even one station with poorer metric measures is sufficient to introduce outliers in the experimental semivariograms. Therefore, the kriging performances can decrease largely between years because the correlation structure is highly dependant on all observations with only 12 stations.

In addition to the cross validation provided, Table 4 illustrates how to predict stream temperature at ungauged stations. By extracting physiographical information for each ungauged stations and by using the PCA algorithm of July 1996 constructed with only 12 stations in our case, PCA coordinates (score values) can be calculated for any location and stream temperatures can be estimated. In fact, once the physiographical information is known it becomes possible to use the PCA algorithm or the CCA algorithm for water temperature estimation anywhere in the study area, at any time period for which we have a minimum (e.g. 12 in our case) number of measurements. There were no temperature measurements to validate the estimated temperatures in Table 4. However, estimated values appeared to be in the same range as those observed in the nearby station belonging to the same cluster. For this reason we can assume that each cluster represents a relatively homogenous geographical region in terms of physiographical characteristics. For ungauged stations outside of these 3 regions, water temperature estimation seems to be appropriate, i.e. no unreasonable values were predicted.

Finally, it is important to recall that one of the main sources of uncertainty in the interpolation stage is network density and spatial distribution of stations, which can be related to kriging variance (St-Hilaire et al., 2003b). As seen in Fig. 1, the stations used in the present study were not uniformly dispersed on the drainage basins but rather formed clusters with relatively large regions with no measurements in between these clusters. This has direct implications on neighbourhood definition and semivariogram calibration. The user must find the appropriate trade-off between good local interpolation and high variance in low density areas.

Conclusions and future work

This study focused on developing a new stream temperature model by using a multivariate geostatistical approach. A physiographical space-based estimation technique was used for the interpolation of stream temperature rather than the usually employed geographical space. It was demonstrated that both multivariate methods, i.e. PCA and CCA, can be employed to construct the physiographical space and then used to build semivariograms for characterizing the correlation structure and ultimately, to perform spatial interpolation. A better performance was observed using the CCA algorithm. In addition, CCA required less information to provide more satisfactory results. Future work should include the possibility of testing the method on a much denser network of temperature loggers and compare the performance of each algorithm as network density decreases. The optimal choice of metrics may also change as a function of network density and a sensitivity analysis may lead to an optimal design.

The present study focused on two relatively large contiguous drainage basins (Moisie and Ste-Marguerite) located on the Quebec North shore. Hence the physiographical features of both basins are somewhat similar. A future study should test the approach when there is potentially greater physiographical variability (e.g. various drainage basin areas, different climate, etc.). At the other end of the physiographical spectrum, the method could also be tested on a sub basin or river reach for which the selection of variables may be different.

Other multivariate approaches should be considered in subsequent analyses to construct a more informative space in which kriging could be performed. As example, nonlinear methods such as Principal Curves Analysis, Curvilinear Component Analysis, Nonlinear Canonical Correlation Analysis, Nonlinear Redundancy Analysis or Nonlinear Principal Predictor Analysis could be used (Yin, 2007). Most of these methods are explicitly designed for dimensionality reduction.

It is believed that this model can become a powerful tool to understand thermal regime of rivers which is often essential for best management of aquatic resources including fisheries.

Acknowledgements

The authors are grateful to J. Poulin for the extraction of physiographical data and to the two anonymous reviewers and associate editor for their insightful comments. This project was financed by NSERC and Hydro-Québec.

References

Ahmadi-Nedushan, B., St-Hilaire, A., Ouarda, T.B.M.J., Bilodeau, L., Robichaud, E., Thiemonge, N., Bobee, B., 2007. Predicting river water temperatures using stochastic models: case study of the Moisie River (Québec, Canada). Hydrological Processes 21, 21–34.

Arscott, D.B., Tockner, K., Ward, J.V., 2001. Thermal heterogeneity along a braided floodplain river (Tagliamento River, northeastern Italy). Canadian Journal of Fisheries and Aquatic Sciences 58, 2359–2373.

Caissie, D., 2006. The thermal regime of rivers: a review. Journal of Freshwater Biology 51, 1389–1406.

Caissie, D., Satish, M.G., El-Jabi, N., 2007. Predicting water temperatures using a deterministic model: application on Miramichi River catchments (New Brunswick, Canada). Journal of Hydrology 336, 303–315.

Caissie, D., El-Jabi, N., St-Hilaire, A., 1998. Stochastic modelling of water temperatures in a small stream using air to water relations. Canadian Journal of Civil Engineering 25, 250–260.

Chen, Y.D., McCutcheon, S.C., Norton, D.J., Nutter, W.L., 1998. Stream temperature simulation of forested Riparian areas: II. Model application. Journal of Environmental Engineering 124, 316–328.

Chokmani, K., Ouarda, T.B.M.J., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resources Research 40 (12), 1–13 (art. no. W12514).

Collings, M.R., 1973. Generalization of stream temperature data in Washington. US Geological Survey Water-Supply Paper 2029-B, B1–B45.

Crozier, L., Zabel, R.W., 2006. Climate impacts at multiple scales: evidence for differential population responses in juvenile Chinook salmon. Journal of Animal Ecology 75, 1100–1109.

Dunteman, G.H., 1989. Principal Component Analysis. Sage Publications, Newbury Park, California, 96 pp.

Eaton, J.G., Scheller, R.M., 1996. Effects of climate warming on fish thermal habitat in streams of the United States. Limnology and Oceanography 41 (5), 1109–1115.

Fernandez, J., Saenz, J., 2003. Improved field reconstruction with the analog method: searching the CCA space. Climate Research 24 (3), 199–213.

Gamma Design Software, 2007. GS+ Geostatistics for the Environmental Sciences, Professional Edition, vers. 7.0, Gamma Design Software, Plainwell, Michigan.

Gardner, B., Sullivan, P.J., Lembo, A.J., 2003. Predicting stream temperatures: geostatistical model comparison using alternative distance metrics. Canadian Journal of Fisheries and Aquatic Sciences 60, 344–351.

Gu, R.R., Li, Y., 2002. River temperature sensitivity to hydraulic and meteorological parameters. Journal of Environmental Management 66, 43–56.

Hawkins, C.P., Hogue, J.N., Decker, L.M., Feminella, J.W., 1997. Channel morphology, water temperature, and assemblage structure of stream insects. Journal of the North American Benthological Society 16, 728–749.

Hodgson, S., Quinn, T.P., 2002. The timing of adult sockeye salmon migration into fresh water: adaptations by populations to prevailing thermal regimes. Canadian Journal of Zoology 80, 542–555.

Holtby, L.B., 1988. Effects of logging on stream temperatures in Carnation Creek, British Columbia, and associated impacts on the coho salmon (Oncorhynchus kisutch). Canadian Journal of Fisheries and Aquatic Sciences 45, 502–515.

Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York, 561 pp.

Johnson, R.A., Wichern, D.W., 2007. Applied Multivariate Statistical Analysis. Pearson Education, Inc., Upper Saddle River, New Jersey.

Johnson, S.L., Jones, J.A., 2000. Stream temperature responses to forest harvest and debris flows in western Cascades Oregon. Canadian Journal of Fisheries and Aquatic Sciences 57 (Suppl. 2), 30—39.

Joseph, G., Chokmani, K., Ouarda, T.B.M.J., St-Hilaire, A., 2007. Une evaluation de la robustesse de la méthode du krigeage canonique pour l'analyse régionnale des débits. Revue des Sciences de l'Eau 20 (4), 367—380.

Lund, S.G., Caissie, D., Cunjak, R.A., Vijayan, M.M., Tufts, B.L., 2002. The effects of environmental heat stress on heat-shock mRNA and protein expression in Miramichi Atlantic salmon (*Salmo salar*) parr. Canadian Journal of Fisheries and Aquatic Sciences 59, 1553—1562.

Manly, B.F.J., 2004. Multivariate Statistical Methods: A Primer, third ed. Chapman and Hall, CRC Press Company, New York, 228 pp.

Miyazawa, T., Yamashita, S., Kitagawa, M., Sekine, K., 1982. An evaluation and mapping of topographical factors affecting river water temperature in the upstream area of the Nagara River/ Japan. Beitrage zur Hydrologie 3, 83—95.

Mohseni, O., Stefan, H.G., Erickson, T.R., 1998. A nonlinear regression model for weekly stream temperatures. Water Resources Research 34 (10), 2685—2692.

Mosley, M.P., 1982. New Zealand river temperature regimes. In: Water and Soil Miscellaneous Publication No. 36. Water and Soil Division Ministry of Works and Development for the National Water and Soil Conservation Organisation, Christchurch.

Peterson, E.W., Sickbert, T.B., 2006. Stream water bypass through a meander neck, laterally extending the hyporheic zone. Hydrogeology Journal 14, 1443—1451.

Scott, M.C., Helfman, G.S., McTammany, M.E., Benfield, E.F., Bolstad, P.V., 2002. Multiscale influences on physical and chemical stream conditions across Blue Ridge landscapes. Journal of the American Water Resources Association 38, 1379—1392.

Sinokrot, B.A., Gulliver, J.S., 2000. In-stream flow impact on river water temperatures. Journal of Hydraulic Research 38, 339—349.

Sinokrot, B.A., Stefan, H.G., McCormick, J.H., Eaton, J.G., 1995. Modeling of climate change effects on stream temperatures and fish habitats below dams and near groundwater inputs. Climate Change 30, 181—200.

Stefan, H., Preud'homme, E., 1993. Stream temperature estimation from air temperature. Water Resources Bulletin 29, 27—45.

Stevens, J., 1986. Applied Multivariate Statistics for the Social Sciences. Erlbaum, Hillsdale, NJ, 660 pp.

St-Hilaire, A., El-Jabi, N., Caissie, D., Morin, G., 2003a. Sensitivity analysis of a deterministic water temperature model to forest canopy and soil temperature in Catamaran Brook (New Brunswick, Canada). Hydrological Processes 17, 2033—2047.

St-Hilaire, A., Morin, G., El-Jabi, N., Caissie, D., 2000. Water temperature modelling in a small forested stream: implication of forest canopy and soil temperature. Canadian Journal of Civil Engineering 27, 1095—1108.

St-Hilaire, A., Ouarda, T.B.M.J., Lachance, M., Bobée, B., Gaudet, J., Gignac, C., 2003b. Assessment of the impact of meteorological network density on the estimation of basin precipitation and runoff: a case study. Hydrological Processes 17, 3561—3580.

Yin, H., 2007. Nonlinear dimensionality reduction and data visualization: a review. International Journal of Automation and Computing 4 (3), 294—303.