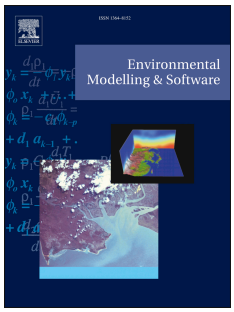
Journal Pre-proof



Regional frequency analysis of stream temperature at ungauged sites using non-linear canonical correlation analysis and generalized additive models

Zina Souaissi, Taha B.M.J. Ouarda, André St-Hilaire, Dhouha Ouali

PII: S1364-8152(23)00068-3

DOI: https://doi.org/10.1016/j.envsoft.2023.105682

Reference: ENSO 105682

To appear in: *Environmental Modelling and Software*

Received Date: 12 September 2022

Revised Date: 13 March 2023

Accepted Date: 14 March 2023

Please cite this article as: Souaissi, Z., Ouarda, T.B.M.J., St-Hilaire, André., Ouali, D., Regional frequency analysis of stream temperature at ungauged sites using non-linear canonical correlation analysis and generalized additive models, *Environmental Modelling and Software* (2023), doi: https://doi.org/10.1016/j.envsoft.2023.105682.

Regional frequency analysis of stream temperature at ungauged sites using non-linear canonical correlation analysis and generalized additive models.

Zina Souaissi[1, 2*], Taha B.M.J. Ouarda[1], André St-Hilaire[1,3], Dhouha Ouali[4]

[1] Canada Research Chair in Statistical Hydro-Climatology, Institut national de la recherche scientifique, Centre Eau Terre Environnement. INRS-ETE, 490 De la Couronne, Québec City, QC, Canada.

[2] Département des Sciences de la Terre et de l'atmosphère, Université du Québec à Montréal, Pavillon Président-Kennedy, Montréal, QC H2X 3Y7, Canada.

[3] Canadian Rivers Institute, University of New Brunswick, Fredericton, Canada

[4] Pacific Climate Impacts Consortium, University of Victoria, Victoria, British Columbia, Canada

[*]Corresponding author: Zina.Souaissi@inrs.ca

Abbreviations

ALL	All stations
ANN	Artificial neural network
BIAS	Mean bias
CCA	Canonical correlation analysis
DHR	Delineation of homogeneous regions
GAM	Generalized additive model
LFA	Local frequency analysis
MCN	Mean curve number
Mean-El	Mean elevation
MLR	Multiple linear regression
NASH	Nash efficiency criterion
NLCCA	Non-linear canonical correlation analysis
PAGR	Percentage of the area covered by agriculture
PCC	Pearson correlation coefficient
PFOR	Percentage of area occupied by forest
PSTERIL	Percentage of sterile area
RE	Regional estimation
RFA	Regional frequency analysis
RMSE	Root-mean-square error
RRMSE	Relative root-mean-square error
Tw_T	Water temperature quantile corresponding to return period T

1. Introduction and review

River temperature is an important indicator of an aquatic ecosystem's health. It influences the metabolic activity of aquatic organisms (Demars 2011), their reproduction, their survival (Connor et al. 2003), and their growth rate (Edwards et al. 1979; Elliott et al. 1995; Elliott and Hurley 1997). Changes in the thermal regime of rivers also affect the distribution of species in rivers (Dugdale et al. 2016; Edwards and Cunjak 2007; Elliott and Hurley 1997; Howell et al. 2010). There is a specific range of temperatures that aquatic organisms can tolerate, and high temperatures can adversely affect fisheries' resources by limiting their habitat or even causing fish mortality (Caissie et al. 2007; Elliott and Hurley 2001; Lund et al. 2002; Sundt-Hansen et al. 2018). Therefore, the thermal regime of rivers has become a critical variable for assessing and modeling their health.

Modeling can enhance our understanding of river thermal regimes and provides tools for assessing environmental effects. Several models have been developed to predict thermal regimes in river systems based on various climate variables and watershed characteristics. Modeling to estimate water temperature at different spatial and temporal scales has traditionally been approached using two types of models: deterministic and statistical (Benyahya et al. 2007; Caissie 2006). Deterministic models focus on mathematical relationships that characterize physical heat fluxes and mass transfer processes and relate stream temperature to other hydrologic factors (Caissie et al. 2007; Hebert et al. 2011; Ouellet et al. 2014; Sinokrot and Stefan 1993; St-Hilaire et al. 2003). However, these models require a large number of input variables (Benyahya et al. 2007). Statistical or empirical models are widely used to predict water temperature since they require fewer parameters than deterministic models. Over the past few years, linear regression models

(Arismendi et al. 2014; Krider et al. 2013) and non-linear regression models (Qiu et al. 2020; Saadi et al. 2022; St-Hilaire et al. 2012) have been adopted for different time scales.

The lack of data at sites of interest motivated the use of regional frequency analysis (RFA). This approach aims to determine the return period of extreme events at ungauged sites where few or no observations are available. In general, RFA consists of two main steps: (1) the delineation of homogeneous regions (DHR) to identify gauged sites similar to the target one, and (2) regional estimation (RE) to transfer the information from the gauged sites to the target one.

Different approaches to RE have been widely considered in RFA studies. For instance, multiple linear regression (MLR) assumes a linear relationship between the response variables and explanatory variables (GREHYS 1996; Ouarda et al. 2008a). However, this assumption is not always verified. Alternative methods were employed to account for the presence of potential non-linearities. As an illustration, due to their considerable flexibility, generalized additive models (GAM) are commonly used in the RFA of floods, low flows, etc. (Ouarda et al. 2018; Rahman et al. 2018). An artificial neural network (ANN) is a non-parametric mathematical model whose design is based on the biological functioning of brain neurons (Bishop 1995). It was considered in several RFA studies, for instance, to estimate floods (Ouali et al. 2017; Shu and Ouarda 2007) and low flows Ouarda and Shu (2009). The random forest (RF) technique is also gaining popularity in several areas due to its powerful non-linear and non-parametric nature. A recent study by Desai and Ouarda (2021) applied the RF to estimate flood flows at ungauged sites. The multivariate adaptive regression splines (MARS) model that considers non-linearity and interactions between variables was introduced in RFA by Msilini et al. (2020). A procedure based on the

Temperature-Duration-Curve approach was developed by Ouarda et al. (2022) to estimate the daily river water temperature at ungauged locations.

These RE methods are applied within a homogeneous region. Homogenous regions were, for example, defined as geographically contiguous regions or non-contiguous hydrological neighborhoods. Among the methods used for delineating geographically non-contiguous regions was the hierarchical cluster analysis (HCA) (Statsoft 1995). HCA identifies similar sites based on the distance between stations within the physiographic and meteorological space (Ouarda et al. 2018). On the other hand, the neighborhood approach assumes that each target site has its own homogeneous region. This means that each site will have its own unique set of stations within its neighborhood. The neighborhood approach can be based on the region of influence principle (ROI) (Burn 1990) or the use of canonical correlation analysis (CCA) (Ouarda et al. 2001). CCA is an important statistical tool for the analysis of multivariate data. CCA allows the establishment of linear combinations of variables within the group, for which the canonical correlation is maximum (Ouarda et al. 2000). According to Ouarda et al. (2008b), the CCA method is more robust than other methods, such as HCA for DHR.

Nonetheless, this method has some limitations when modeling the thermal regime of rivers. It is not robust to the non-linear relationships between response variables and watershed physiographic and meteorological characteristics. Natural factors such as topography, soil structure, geological formations, and climate affect the variability of water temperatures. This leads to a natural complexity, which has been widely recognized and documented in the literature (Beechie et al. 2010; Caissie 2006; Hewlett and Fortson 1982; Lisi et al. 2013;

Wahli et al. 2008). Thus, ignorance of the non-linear structure of the relevant explanatory variables in the DHR step can result in large uncertainty in the estimation step.

In the literature, few studies have focused on integrating non-linear approaches into the delineation step. For example, in Lin and Chen (2006), the self-organizing map was trained using an unsupervised competitive learning algorithm for DHR. Durocher et al. (2016) proposed to carry out the delineation of homogenous regions using hydrological variables predicted by projection pursuit regression instead of using physiographic caracteristics. Results showed clear improvements in neighbourhood definitions and quantile estimates in comparison to linear approaches. Wazneh et al. (2016) used a similarity measure derived from depth functions to calculate similarities between target sites and those gauged in the DHR. Ouali et al. (2016) introduced the non-linear canonical correlation analysis (NLCCA) approach to identify hydrological neighborhoods. In this approach, the authors coupled CCA with ANN. The obtained results demonstrated the importance of considering non-linearity in the delimitation step, which also improved estimation performance.

Although there is strong evidence for the non-linearity of river thermal regime processes, the NLCCA approach has not yet been considered for modeling river temperature. It is important to note that no study in the literature has considered non-linear methods simultaneously in both RFA steps to estimate river water temperature quantiles. The present paper aims to address the issue of non-linearity in the DHR step using NLCCA to improve performance. Then, different combinations of non-linear approaches in the two RFA steps are considered. Another objective is to identify which step is most affected by the non-linearity.

This paper is organized as follows: Section 2 presents a brief theoretical overview of the regionalization approaches used. The case study and data used are presented in Section 3. The methodology is described in Section 4. Section 5 provides the results obtained and their discussions. Finally, the conclusions of the study are summarized in the last section.

2. Theoretical background

In this section, the adopted statistical tools are briefly described and discussed.

2.1. Methods for the delineation of homogeneous regions

The following section presents a brief description of the CCA and NLCCA methods used in RFA.

2.1.1 Canonical correlation analysis

CCA is a multivariate analysis approach for identifying possible correlations between two groups of variables (Hotelling 1935). It consists of a linear transformation of two groups of random variables into pairs of canonical variables, which are arranged to maximize the correlations between the pairs. Let *X = (X1, X2, ..., Xq)* and *Y = (Y1, Y2, ..., Ys)* be sets of random variables including respectively, the q physio-meteorological variables and the s thermal variables. By considering linear combinations of variables X and Y, we can obtain canonical variables U_i and V_i as follows:

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

$$V_i = b_{i1}Y_1 + b_{i2}Y_2 + \cdots + b_{is}Y_s \quad (2)$$

where $i = 1, \ldots, p$, with $p = min(q, s)$. The first weight vectors a_1 and b_1 maximize the correlation coefficients between the resulting canonical variables, i.e., $\lambda_1 = corr\ (U_1, V_1)$,

under unit variance constraints. Once the first pair of canonical variables is identified, other pairs (U_i , $V_i, i > 1$) can be obtained under the constraint $corr\ (U_i\ , V_j)\ =\ 0$ for $i \neq j$.

CCA is commonly used to determine target site neighborhoods. In a target site, the canonical physio-meteorological information is generally known, and the thermal data are not available. We denote u_0 the canonical score corresponding to the physiographic values of a target basin, that is, the corresponding values of the canonical variables for the target basin with respect to its physiographic scores. The thermal mean position of the target site is given by $\Lambda\ u_0$, where $\Lambda\ =\ diag\ (\lambda_1, \dots, \lambda_p)$. Hence, a neighborhood with a confidence level of 100 (1−α) % can be identified by the Mahalanobis distance between the mean position of the target site $\Lambda\ u_0$ and the positions of the other sites *V*, such that:

$$(\mathrm{V} - \Lambda u_0\ )'(I_P - \Lambda^2)^{-1}(\mathrm{V} - \Lambda u_0\ ) \leq\ \chi^2_{\alpha\ ,\ \mathrm{p}} \qquad (3)$$

where I_P is a $p \times p$ identity matrix, and $\chi^2_{\alpha\ ,\ p}$ is such that, for an observed Mahalanobis distance χ^2, we have: $P\left(\chi^2 \leq\ \chi^2_{\alpha\ ,\ p}\right) = 1 - \alpha$. Eq. (3) describes the interior of an ellipsoidal region in the canonical space *V*. The use of CCA for the delineation of hydrological neighborhoods is described in detail in (Ouarda et al. 2001).

2.1.2 Non-linear canonical correlation analysis

ANN is considered a non-parametric model that is a universal approximator (Geman et al. 1992; Hornik et al. 1989). This model was applied to solve large complex problems such as pattern recognition, non-linear modeling, classification, and control (see for instance Alobaidi et al. 2014). The basic architecture of an ANN (Fig. 1) consists of a layer of input neurons linked to one or more layers of "hidden" neurons, which are themselves linked to a layer of output neurons. An ANN is widely used in many fields to solve regression and

classification problems, and it can be coupled with other multivariate methods such as NLCCA. NLCCA is an approach developed by Hsieh (2000) in the meteorological domain based on ANN (CCA-NN). The approach has since been used in a number of fields, such as environmental modeling and renewable energy assessment (see for instance Woldesellasse et al. 2020). In NLCCA, we follow the same procedures as in CCA, except for the linear mappings in Eqs. (1 and 2) which are substituted by non-linear mapping functions based on ANN. Mappings from the original variables (*X* and *Y*) to the new canonical variables (*U* and *V*) are represented by the double-barreled ANN on the left half of Fig. 1. The transfer function f maps the inputs to the neurons in the hidden layer $h^{(x)}$ and $h^{(y)}$:

$$h_k^{(x)} = f\left[(W^{(x)}x + b^{(x)})_k\right]; \qquad k \in \{1, \ldots, l\} \tag{4}$$

$$h_n^{(y)} = f\left[(W^{(y)}y + b^{(y)})_n\right]; \qquad n \in \{1, \ldots, l\} \tag{5}$$

where $W^{(x)}$ and $W^{(y)}$ are weight matrices; $b^{(x)}$ and $b^{(y)}$ are bias parameter vectors; *k*, *n*, are respectively the indices of the vector elements $h^{(x)}$ and $h^{(y)}$, and l is the number of hidden neurons. The transfer function *f*, identical for *x* and *y*, is generally fixed to the hyperbolic tangent function (Hsieh 2001; Wu et al. 2003) (See Bishop (1995b), section 4.3, discusses the choice of transfer functions).

The canonical variables neurons *U* and *V* are determined from a linear combination of hidden neurons $h^{(x)}$ and $h^{(y)}$ (but from a non-linear combination with respect to *x* and *y*).

$$U = w^{(x)}h^{(x)} + \bar{b}^{(x)} \tag{6}$$

$$V = w^{(y)}h^{(y)} + \bar{b}^{(y)} \tag{7}$$

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

$$\bar{b}^{(x)} = -\langle w^{(x)} h^{(x)} \rangle \text{ and } \bar{b}^{(y)} = -\langle w^{(y)} h^{(y)} \rangle \quad (8)$$

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

Once the canonical variables are determined, the inverse mapping of the canonical variables to the original variables is developed. The mapping of canonical variables *(U, V)* to hidden layers ($h^{(U)}$, $h^{(V)}$) is represented in Fig. 1. Then the final mapping is from the hidden layers ($h^{(U)}$, $h^{(V)}$) to the model output (X', Y') (Fig. 1).

The limitation of NLCCA is that it provides only one set of canonical variables when applied to the original data, namely, one for the physiographic variables and another for the thermal variables. This may exclude some information since there is no guarantee that this first pair *(U, V)* of canonical variables represents the most significant part of the explained variance. To solve this problem, we use the notion of modes or iterations introduced by Hsieh (2000). Having determined the first mode X' from the initial data *X*, the NLCCA method can be applied again to the residue to determine the second mode. In other words, we determine the unexplained information from the previous mode by reapplying the procedure to the new variables:

$$I_2 = X - X' \quad (9)$$

$$J_2 = Y - Y' \quad (10)$$

where X' and Y' are results of the first mode and X and Y are the original data.

The same procedure is followed for the higher-order modes, taking each time the residue of the previous mode as input. There must be a minimum number of iterations equal to the smallest number of variables. The final result is the sum of all considered modes:

$$X_{estimated} = X' + X'' + \cdots X^n \quad (11)$$

where X^n is the result of the n^{ith} mode. As a result, multiple modes can increase the amount of information embodied in the canonical variables.

In this work, we present the NLCCA approach for identifying the neighborhood of a non-gauged site. Following the extraction of the first NLCCA mode, the second mode is extracted by taking the residue as input as in Eqs (9) and (10). Consequently, we obtain the canonical variables in the non-linear space. In the non-linear case, V_1 and V_2 are the canonical thermal variables of the first and second modes, respectively, and Λ_1 and Λ_2 are the canonical correlation coefficients of the two modes. In order to identify physiographic scores U_{01} and U_{02} at a target site, Eq. (6) is used. Then, using the Mahalanobis distance calculated with Eq. (3), the hydrological neighborhood of each ungauged site is determined. Similarly to the linear case, the non-linear case can also be obtained using the same constraint. Nevertheless, the ellipsoid equation differs from the linear case since the axes are not parallel to the coordinate systems. A detailed discussion of the theoretical background and the application of NLCCA for DHR is presented in Ouali et al. (2016).

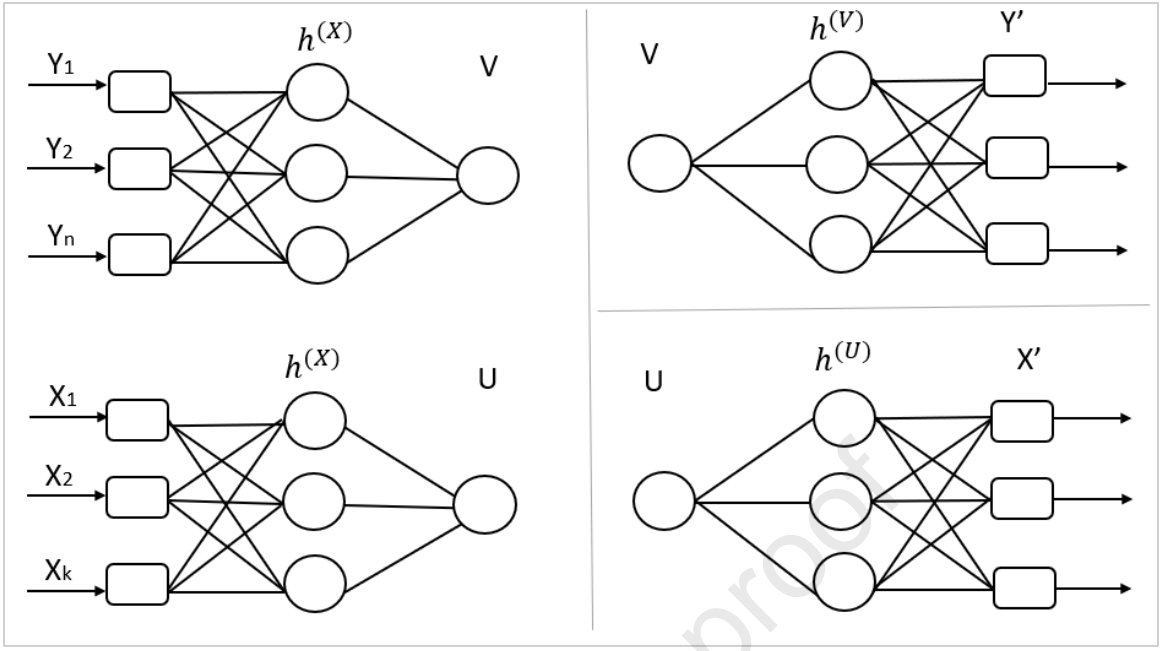


Fig. 1 The neural networks (NNs) used to perform the NLCCA. Double-barrel NNs on the left map from (*Y*, *X*) to canonical variables (*V*, *U*) on the right.

2.2. Regional thermal quantile estimation

The methods described below are used to transfer information from the neighborhood stations to the target site.

2.2.1. Multiple linear regression

The MLR method is commonly used in a large number of fields to determine a linear relationship between the response variable Y (such as the quantile of water temperature Tw_T corresponding to a period T) and one or more random variables X, called explanatory variables (X1, X2, ..., Xn) (Hosking and Wallis 1993; Pandey and Nguyen 1999) and is defined as follows:

$$Y = \beta_0 + \sum_{j=1}^{n} \beta_j X_j + \varepsilon \tag{12}$$

where X is a matrix whose columns correspond to a set of n explanatory variables, β_0 and β_j are unknown parameters and ε is the model error.

A logarithmic transformation is usually applied to linearize the relationship in Eq (12):

$$log\, Y = log\, \beta_0 + \sum_{j=1}^{n} \beta_j log\, X_j + \varepsilon \quad (13)$$

This transformation introduces an additional bias (Girard et al. 2004). The coefficients β_j of the model are usually determined using the ordinary least squares method (Thomas and Benson 1970).

2.2.2. Generalized Additive Model

The GAM was first introduced by Hastie and Tibshirani (1987) and has since gained wide popularity in a large number of fields (see for instance Ouarda et al. 2016). GAM is an extension of the general linear model (GLM) in which the linear predictor is replaced with a set of smooth functions of explanatory variables. This model allows for a non-Gaussian response distribution and a non-linear relationship between the response variable Y and the explanatory variables X using smooth functions (Wood 2006). For a response variable Y, the GAM can be expressed as:

$$g(Y) = \alpha + \sum_{i=1}^{m} f_i\,(X_i) + \varepsilon, \quad (14)$$

where g is a monotonic link function and f_i are smooth functions giving the relationship between the explanatory variables X_i and the response Y. The parameter α is the intercept,

and ε is the error term. The structure of Eq (14) permits the interpretation of each explanatory variable.

In GAM, a spline is used to estimate smooth functions f_i. A spline is a curve composed of piecewise polynomial functions, joined at points called nodes. Several types of splines have been proposed in the literature, such as cubic splines, P-splines, B-splines, etc. (Wahba 1990). In a regression spline, the number of nodes is significantly reduced. Generally, a smooth function f_i can be described by a linear combination of q basis functions $b_{ij}(X)$:

$$f_i(X) = \sum_{j=1}^{q} \beta_{ij} + b_{ij}(X) \tag{15}$$

where β_{ij} are smoothing coefficients.

3. Case study and datasets

The network of water temperature stations in Switzerland is chosen as a case study for this work. This study is based on the daily water temperature data provided by the Swiss Federal Office for the Environment (FOEN). The data used in the present study consist of 24 river temperature stations studied in Souaissi et al. (2021). The 24 stations selected met the following three criteria: First, the river must have a natural flow regime. Secondly, the station must have at least 15 years of historical record. The station's historical data must also meet the basic assumptions of independence, stationarity, and homogeneity, i.e., the data series of seasonal or annual maxima must be independent and identically distributed (iid). To test independence, the Wald and Wolfowitz (1943) test is employed, the Mann (1945) test is used to test stationarity, and the Wilcoxon (1946) test is employed to test homogeneity. Fig. 2 illustrates the location of the gauging stations that are selected for this

study. The diameters of the circles are proportional to the areas of the basins, which vary between 3 and 6299 km^2. The stations cover a large area of Switzerland.

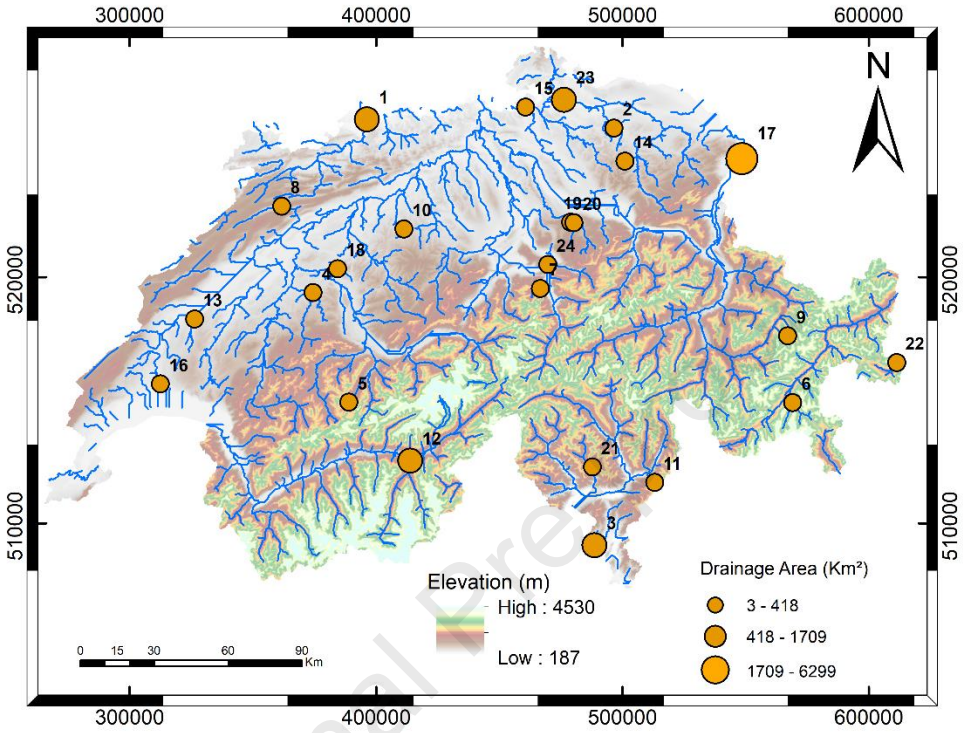


Fig. 2 Geographic distribution of the water temperature stations used in the study.

This study uses the results of the local frequency analysis of maximum summer water temperatures of Souaissi et al. (2021) to estimate regional thermal quantiles corresponding to different return periods: Tw_2, Tw_5, Tw_{10}, Tw_{20}, Tw_{50}, and Tw_{100}. Tw_T represents the water temperature quantile corresponding to return period T. Local quantiles are distributed between a minimum of 12.75 °C and a maximum of 30.84 °C (Fig. 3). The appropriate probability distributions identified are mainly the two-parameter Weibull distribution (W2), the normal distribution (N) and the inverse Gamma (IG) distribution.

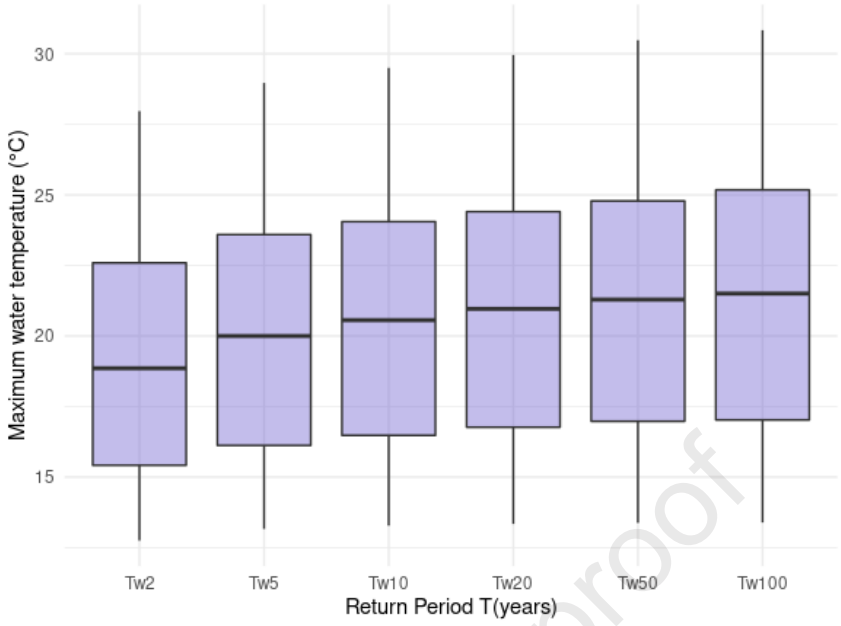


Fig. 3 Box plots of maximum water temperature corresponding to different return periods.

The physiographic and meteorological variables for each watershed in the study area are extracted using ArcGIS. The variables related to topography are calculated through the use of a digital elevation model (DEM), such as the mean elevation (Mean-El) and the mean slope of a watershed (MSLP). The other explanatory variables are determined using the land use and geological maps that are available on the FOEN website in Switzerland, such as the percentage of forest (PFOR), the percentage of agriculture (PAGR), the percentage of barren areas (PSTERIL), and the mean curve number (MCN). The descriptive statistics of the physio-meteorological variables used in the data set are presented in Table 1.

Table 1. Descriptive statistics of the physiographic-meteorological variables.

Notation	Variable	Unit	Min	Max	Median	Mean
Physiographic- meteorological variables						
AREA	Catchment area	km^2	2.76	6299.19	94.51	522.03
DD	Drainage density	Km^{-1}	1.7	4.64	2.416	2.52
LATC	Latitude of the centroid of the basin	m	5089623	5219930	5089623.83	5089623.83
LONGC	Longitude of the centroid of the basin	m	487475	566530	527002.62	527002.62

MCN	Mean curve number	-	71	86	73.5	74.95
Mean-El	Mean elevation	m	501.97	2698.32	1052.81	1310.41
MSLP	Mean slope of the watershed	°	1.36	36.67	15.24	17.614
PAGR	Percentage of the area covered by agriculture	%	8.17	79.41	43.45	42.18
PFOR	Percentage of area occupied by forest	%	3.05	62.6	29.28	31.21
PGLACIAL	Percentage of area covered by glacial deposits	%	0	26.02	0	2.40
PLAKE	Percentage of area occupied by lakes	%	0	8.14	0	0.543
PSTERIL	Percentage of sterile area	%	0	54.70	1.29	14.83
PURBAN	Percentage of urban area	%	0	28.30	2.86	5.35
PWetland	Percentage of area occupied by peatlands and marshes	%	0	28.4	1.75	4.99
Max-T_{air}	Maximum annual air temperature	°C	16.75	32.1	29.14	28.04

4. Methods

4.1 Regional models

In this study, we combine the two DHR methods, CCA and NLCCA, with the RE models MLR and GAM presented in Section 2. We also consider using all stations without definition of homogeneous regions (denoted ALL).

This leads to the following linear combinations:

- ALL/MLR: MLR is used without neighborhoods.
- CCA/MLR: MLR is used with neighborhoods identified using the CCA linear method.

The following semi-linear combinations:

- ALL/GAM: GAM is used without neighborhoods.
- CCA/GAM: GAM is used with neighborhoods identified with the CCA linear method.

- NLCCA/MLR: MLR is used with neighborhoods identified using the non-linear NLCCA method.

And one non-linear combination:

- NLCCA/GAM: GAM is used in conjunction with non-linear NLCCA methods of identifying neighborhoods.

NLCCA has the same objective as CCA: reduce the variables *X* and *Y* into canonical variates *U* and *V* such that *cor(U, V)* is maximized. Unlike the linear compression of the CCA, which is achieved by weighted sums of the original variables (Eqs. 1 and 2), the non-linear compression of the NLCCA is accomplished by neural networks (Eqs. 4 and 5). These two approaches are applied to DHR using sets of physio-meteorological variables. When performing a CCA or a NLCCA, the relevant variables are selected using a stepwise process. A description of this approach is provided in subection 4.2.

The CCA and NLCCA procedures generate neighborhoods with varying sample sizes from one site to another. Note that using CCA, each site's neighborhood is an ellipsoid with a zero rotation angle (Leclerc and Ouarda 2007; Ouarda et al. 2001). In the NLCCA method, an ellipsoid was identified with a rotation angle of $\phi \sim 21°$. As opposed to CCA, the orientation of the ellipsoid in the NLCCA method tends to follow the shape of the data dispersion (Ouali et al. 2016). Since the sample size is an essential factor for the reliability of the estimates obtained by MLR or GAM, it was decided that the region size for each target station would be increased until an optimal size was reached by applying a jackknife procedure.

MLR and GAM are used in this study as RE methods. GAM was implemented in R using the mgcv package (Wood 2006). Due to their theoretical origins, thin plate regression splines, which are a generalization of cubic splines, can be considered as the $b_{ij}(.)$ basis of the smoothing functions f_i as described in Eq. (15). This smoothing function has the advantage of having a high computational speed and a limited number of parameters compared to other smoothing functions (Wood 2003). The considered link function g in Eq. (14) is the identity function since the log-transformed quantiles are approximately normally distributed (as in Msilini et al. 2022; Ouali et al. 2017).

4.2 Stepwise regression

The stepwise procedure is adopted in this work to select the optimal explanatory variables in the two RFA steps. For this approach, a regression method (MLR or GAM) is first applied with a model incorporating all the explanatory variables. During each step, the variable with the highest p-value for the null hypothesis is removed, namely the parameter (for MLR) or smooth term (for GAM). The procedure ends once the p-values for all remaining variables are below a given threshold (5%).

4.3 Validation

For each regional model mentioned in subsection 4.1, a jackknife method (cross-validation) is employed. In this procedure, a gauged site is temporarily removed, i.e., considered non-gauged, to make a regional estimate. It is a process of comparing the regional estimate with the observed value. The following four evaluation criteria are used to assess the performance of each regional model: Nash coefficient (NASH), root mean square error (RMSE), relative root mean square error (RRMSE), and mean bias (BIAS). These criteria are expressed as:

$$NASH = 1 - \left(\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{N} (y_i - \bar{y})^2 \right) \quad (16)$$

$$RMSE = \sqrt{{}^{1}/_{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \quad (17)$$

$$RRMSE = 100 \sqrt{{}^{1}/_{N} \sum_{i=1}^{N} \left[{}^{(y_i - \hat{y}_i)}/_{y_i} \right]^2} \quad (18)$$

$$BIAS = {}^{1}/_{N} \sum_{i=1}^{N} (y_i - \hat{y}_i) \quad (19)$$

where N is the number of stations; y_i and $\hat{y}_i$ are the local and regional quantile estimates at station i; and $\bar{y}$ is the local mean of the thermal variable.

5. Results

5.1 Selection of optimal variables for DHR and RE

The selection of variables is based on a stepwise procedure in the two RFA steps. This approach led to the identification of four explanatory variables for DHR. Fig. 4 depicts the dispersion of the thermal quantiles Tw_5 and Tw_{10}, the physiographic variables selected for DHR, and the Pearson correlation coefficients (PCC) associated with these variables. Therefore, for DHR with CCA and NLCCA, we used q=4 physiographic variables and s=2 thermal variables. Examining the scatterplots in Fig. 4 reveals various types of relationships between the variables. In this figure, we can observe non-linear relations. The most notable non-linear relationships are detected between MCN, PSTERIL, and the other variables.

For the RE step, a stepwise selection is considered for each thermal quantile corresponding to different return periods (Tw_2, Tw_5, Tw_{10}, Tw_{20}, Tw_{50}, and Tw_{100}). This is carried out for each MLR and GAM estimation model. The variables selected for each quantile and model are shown in Table 2. Following Eq. 13, the MLR model is used to estimate regional thermal quantiles as follows:

$$\log(Tw_2) = \beta_0 + \beta_1 \, log\,(Mean - EL) + \beta_2 \, log\,(PSTERIL) + \beta_3 \, log\,(MCN), \quad (20)$$

$$\log(Tw_5) = \beta_0 + \beta_1 \, log\,(Mean - EL) + \beta_2 \, log\,(PSTERIL) + \beta_3 \, log\,(PFOR), \quad (21)$$

Mean-EL is the most crucial variable in Eqs. (20) and (21), respectively. Mean-EL, PSTERIL, and PFOR are important variables for quantiles corresponding to all return periods except Tw_2. MCN is used instead of PFOR for Tw_2.

GAM is likely to have a different selection of variables. Table 2 summarizes the final variables for each GAM combination. The model used for extreme water temperatures within the models ALL + GAM, CCA + GAM, and NLCCA + GAM is then written as:

$$\log(Tw_T) = \alpha + f_1(Mean - EL) + f_2\,(PAGR) + f_3\,(PSTERIL) + \varepsilon, \quad (22)$$

Mean-EL, PAGR, and PSTERIL, are three significant predictors for all thermal quantiles.

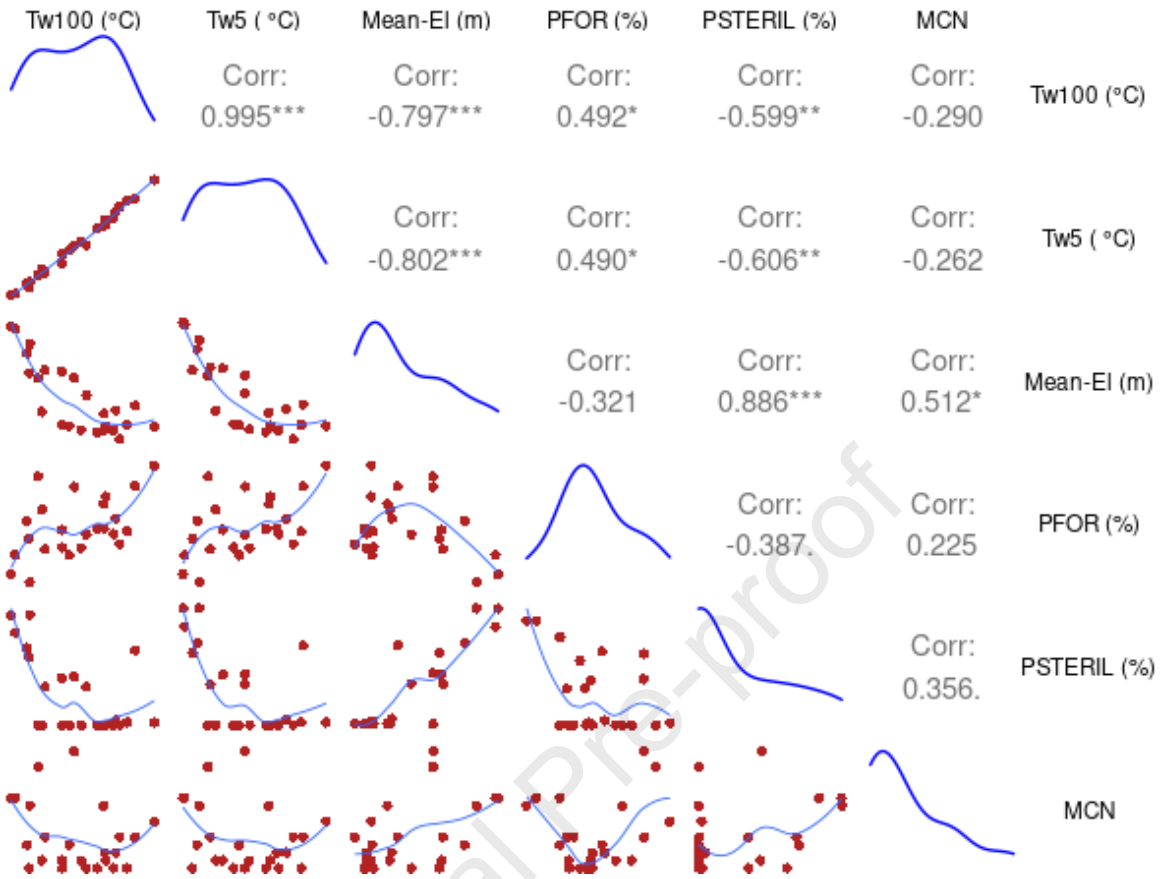


Fig. 4 Scatter plots of physiographical and thermal variables with their correlation coefficients

Table 2. Variables selected for each regional model.

Regional models	Quantiles	Selected predictor variables
ALL+MLR, CCA+MLR, NLCCA+MLR	Tw_2	Mean-El, PSTERIL, MCN
	Tw_5	Mean-El, PSTERIL, PFOR,
	Tw_{10}	Mean-El, PSTERIL, PFOR
	Tw_{20}	Mean-El, PSTERIL, PFOR
	Tw_{50}	Mean-El, PSTERIL, PFOR
	Tw_{100}	Mean-El, PSTERIL, PFOR
ALL+GAM, CCA+GAM, NLCCA+GAM	Tw_2	Mean-El, PAGR, PSTERIL
	Tw_5	Mean-El, PAGR, PSTERIL
	Tw_{10}	Mean-El, PAGR, PSTERIL
	Tw_{20}	Mean-El, PAGR, PSTERIL
	Tw_{50}	Mean-El, PAGR, PSTERIL
	Tw_{100}	Mean-El, PAGR, PSTERIL

Fig. 5 illustrates the smooth functions obtained for the thermal quantile Tw_5. Smooth functions allow us to assess the influence of each variable independently of the others. The Mean-EL is perfectly linear with Tw_5 with narrow confidence intervals and small residuals. Smoothing functions of Mean-EL have a negative slope. This explains that the higher the temperature, the lower the altitude, i.e., the low-altitude areas are more heated than the high-altitude areas (Wahli et al. 2008). In terms of PAGR, the results are not as far from linearity with Tw_5. The slope of the smoothing functions of the PAGR is negative, which explains why water temperatures increase after logging of riparian forests or agriculture (Steedman et al. 1998; Zeni et al. 2019). The PSTERIL variable exhibits complex non-linear behavior. In particular, the relationship between Tw_5 and PSTERIL increases for low PSTERIL values, decreases for medium values and increases again for high values. Therefore, GAM demonstrates that even if the first two variables are linear, the third variable exhibits a non-linear relationship. A negative PCC is observed between the thermal quantiles corresponding to different return periods and PSTERIL in Table 3, whereas the slope of the smooth function is positive. Also, the PCC is positive for PAGR, while the slope of the smooth functions is negative. The positive correlation between thermal quantiles and PAGR can be explained by the fact that much agricultural activity occurs at high altitudes in Switzerland. Altitude may hence be a confounding factor (Table 3). Hence, the conclusions that can be drawn from the PCC are different from those that can be drawn from the GAM model. GAM allows for the interpretation of the impact of a given explanatory variable on the response variable independently of the other explanatory variables. Conclusions based solely on correlations can hence be misleading.

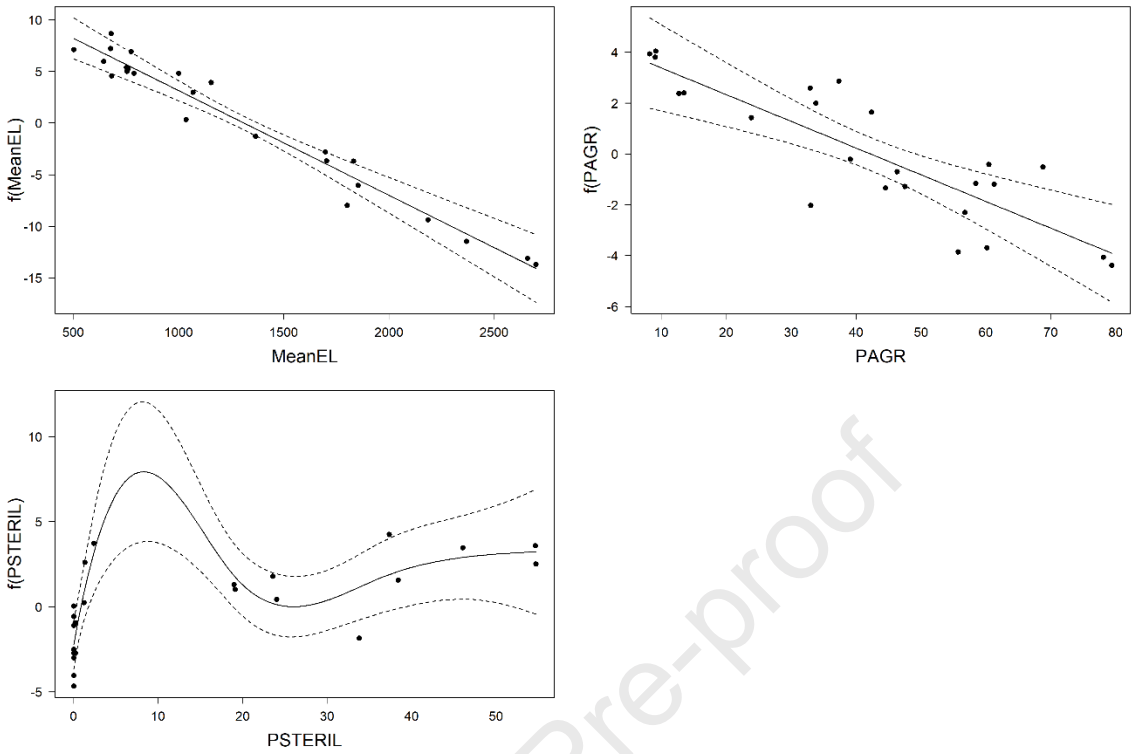


Fig. 5 Smooth functions of thermal quantile Tw_5 for the explanatory variables. The dashed lines represent the 95% confidence intervals, and the dots are the residuals.

Table 3. Pearson correlation coefficients between quantiles and selected physiographic variables.

Physiographic variables	Tw_2	Tw_5	Tw_{10}	Tw_{20}	Tw_{50}	Tw_{100}
MCN	-0.25	-0.26	-0.27	-0.28	-0.29	-0.29
Mean-El	-0.80	-0.80	-0.80	-0.80	-0.80	-0.80
PAGR	0.36	0.37	0.37	0.37	0.37	0.37
PFOR	0.48	0.49	0.49	0.49	0.49	0.49
PSTERIL	-0.61	-0.61	-0.60	-0.60	-0.60	-0.60

5.2 Delineation of regions with CCA and NLCCA

A neighborhood is defined for the target site with the CCA and NLCCA methods. In both DHR approaches, similar variables are selected, such as Mean_El, PSTERIL, PFOR, and

MCN as physiographic variables and Tw_5 and Tw_{100} as thermal variables. A neighborhood size of 18 stations is optimal for the CCA and NLCCA methods according to RMSE results. Both delineation approaches require that the thermal variables and the explanatory variables be normal. Therefore, some variables are transformed to achieve normality. In CCA and NLCCA, the physiographic and thermal variables are logarithmically transformed, whereas in NLCCA, a square root transformation was more appropriate for PSTERIL.

Fig. 6 illustrates the point cloud of the study sites in the non-linear canonical spaces: physiographic (U_1, U_2) and thermal (V_1, V_2) spaces. In contrast, Fig. 7 depicts the point cloud in physiographic and thermal linear spaces with CCA. These two figures demonstrate that each space is nearly symmetrical to the other, i.e., the non-linear physiographic space is symmetrical to the linear physiographic space, and likewise for the thermal spaces. On the basis of the spatial location of the stations, as shown in Fig. 2, the high-altitude stations are located within the same area in the canonical space. At the same time, the low-altitude stations are also located within the same area.

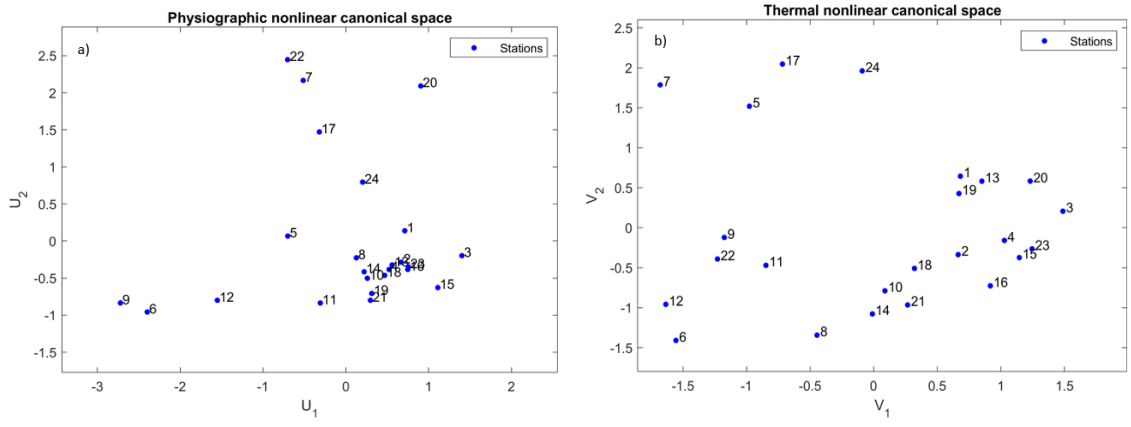


Fig. 6 Data set in non-linear canonical spaces: a) the physiographic non-linear space and b) the thermal non-linear space.

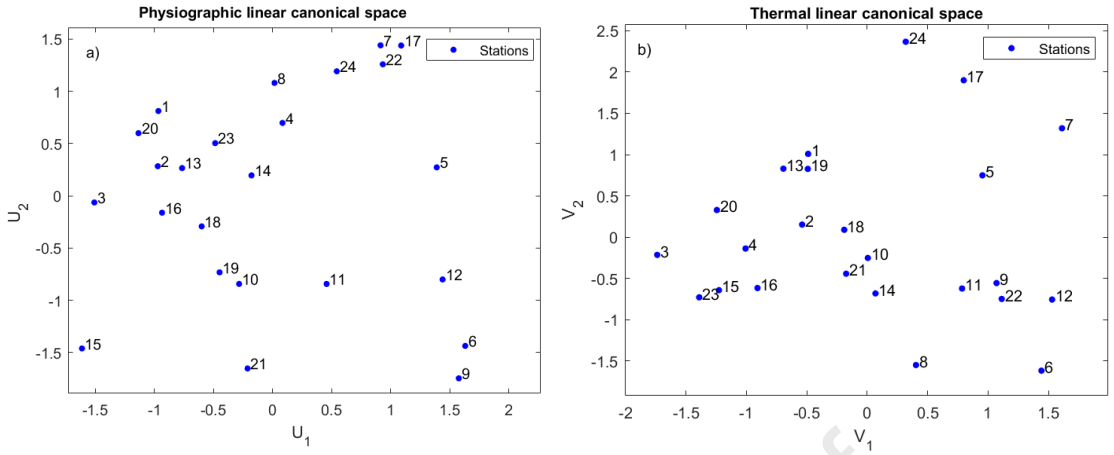
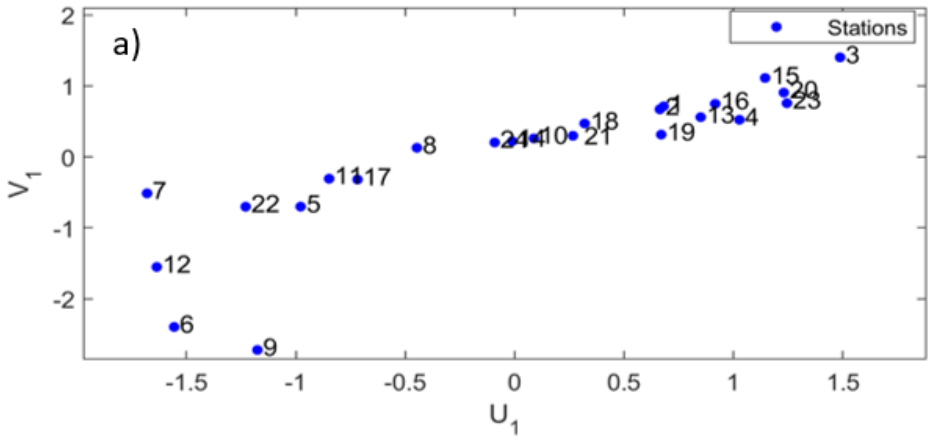


Fig. 7 Data set in linear canonical spaces: a) the physiographic linear space and b) the thermal linear space.

Additionally, it is convenient to present the data in spaces (U_1, V_1) and (U_2, V_2) to obtain information regarding the estimation error. This is displayed in Figs. 8 and 9 for the non-linear and linear cases, respectively. There is a linear relationship between the two canonical variables (U_1, V_1). That does not appear to be the case for the couple (U_2, V_2) (Chebana and Ouarda 2008; Ouali et al. 2016). As a result, the point cloud in the non-linear case seems to be more linear in the two pairs (U_1, V_1) and (U_2, V_2) (Fig.8) than in the linear case (Fig.9). Moreover, the canonical correlation coefficients obtained from the NLCCA (U_1, V_1 = 0.968; U_2, V_2 = 0.761) method are higher than those obtained from CCA (U_1, V_1 = 0.906; U_2, V_2 = 0.445).



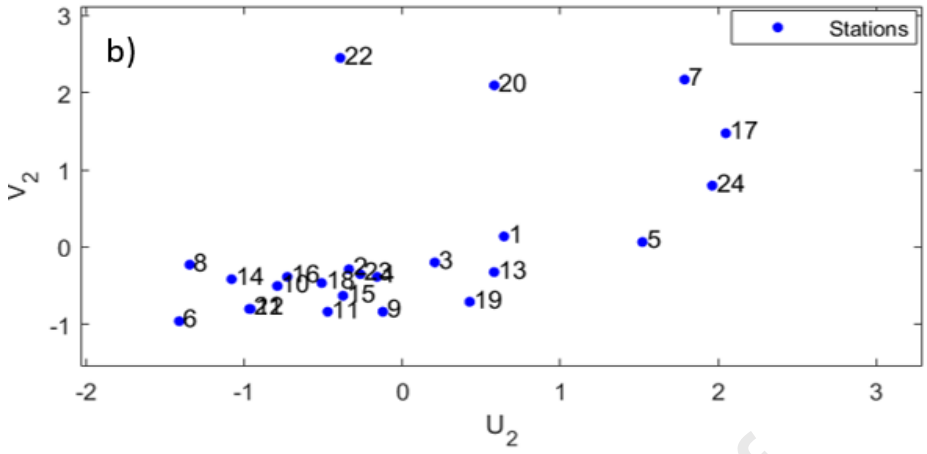


Fig. 8 Data set in the non-linear canonical spaces: a (U_1, V_1) and b (U_2, V_2).

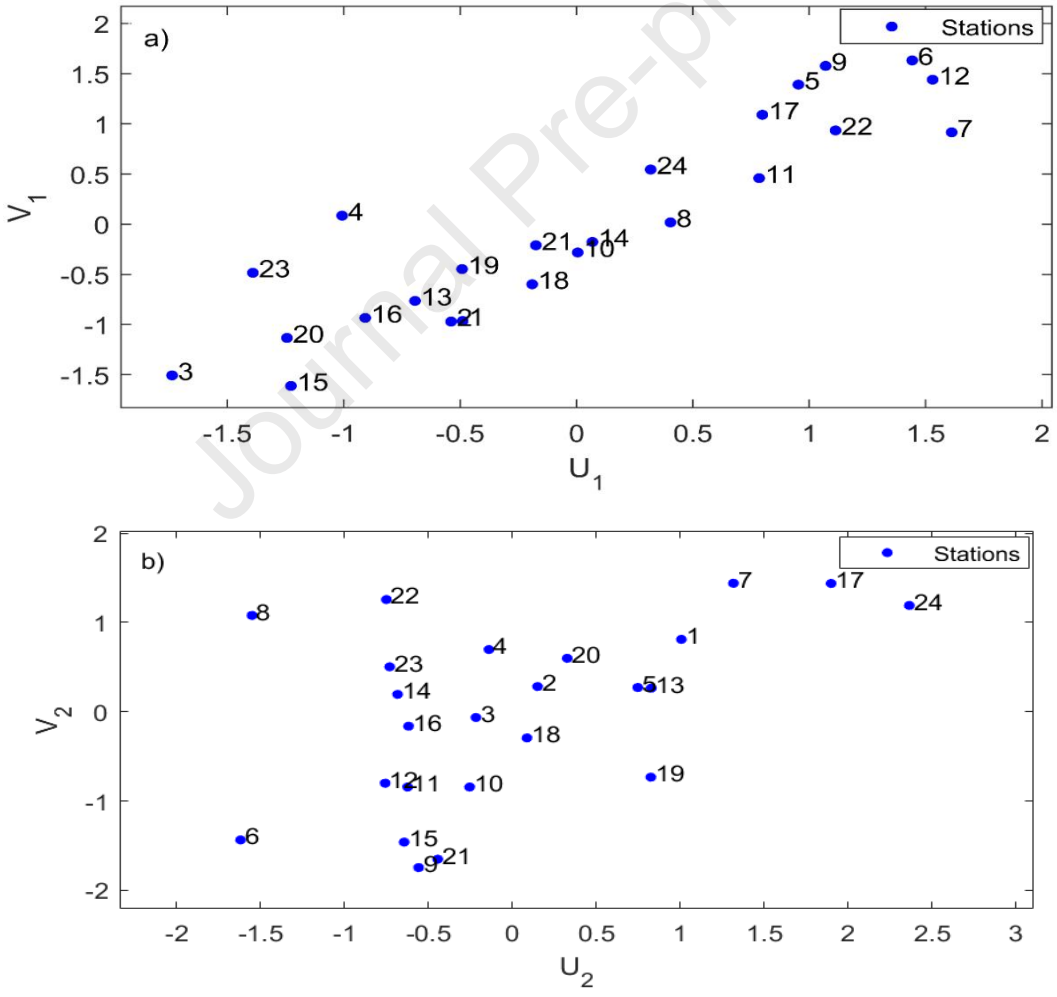
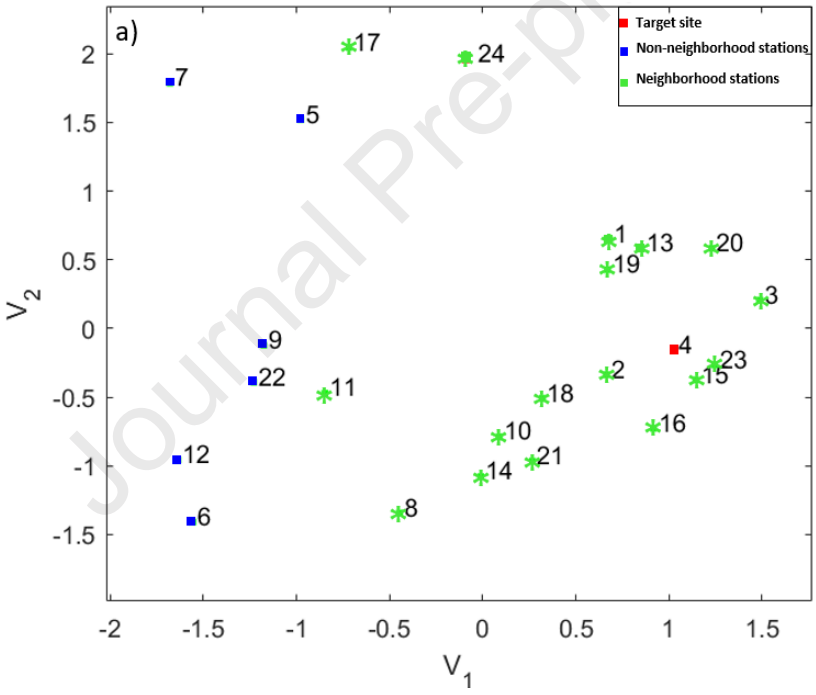


Fig. 9 Data set in the linear canonical spaces: a (U_1, V_1) and b (U_2, V_2).

The CCA and NLCCA procedures are also applied with coefficients α ranging in the interval [0, 1]. The optimal coefficient α value is determined using the minimum RMSE values of the jackknife resampling procedure, as explained in Ouarda et al. (2001). The optimal value is determined to be $\alpha = 10^{-5}$ for CCA, and $\alpha = 10^{-14}$ for NLCCA. Fig. 10 shows that in the non-linear case (n = 4), 17 neighboring stations are identified. In contrast, in the linear case, 22 neighboring stations are identified (Fig.10), although the α parameter is less optimized with NLCCA. As a result, the NLCCA requires fewer stations to obtain the same RMSE as the CCA.



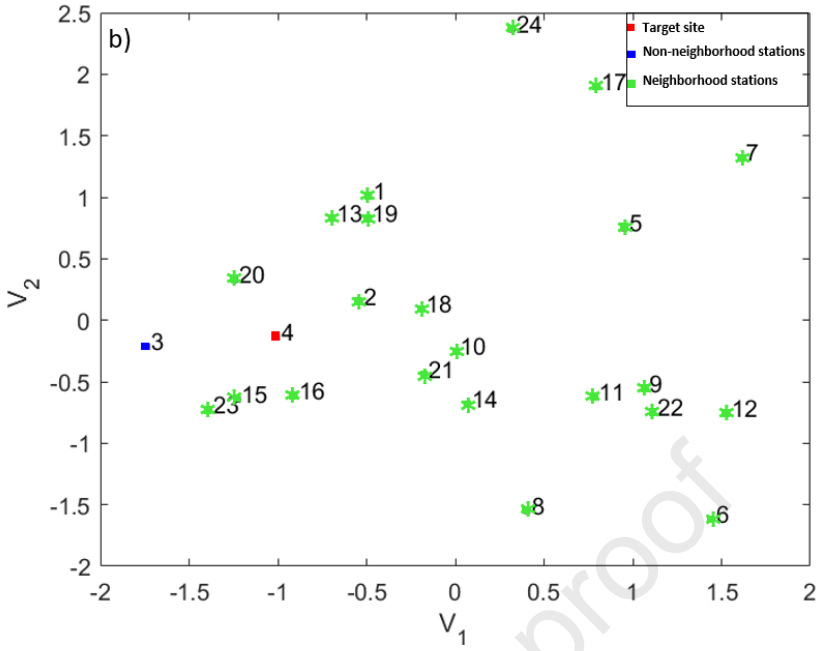


Fig. 10 Delineation of homogeneous regions results shown for station 4 using: a) NLCCA and b) CCA approaches.

5.3 Comparison of regional models

To evaluate the results obtained for DHR and their impact on the final RFA estimates, we proceed to the RE step. As discussed in Section 4, a jackknife resampling procedure is employed to compare the different approaches considered in this study. The performance indices for thermal quantile estimations obtained from this procedure are presented in Table 4. The bold font in Table 4 indicates the optimal approach for each cell. The NLCCA + GAM methods give us the best results in terms of total performance. A NASH criterion of 1, would indicate that the model produces a perfect estimate. The obtained NASH results are higher than 0.7. According to this criterion, the various models are ranked in order of performance from highest to lowest: non-linear combinations NLCCA + GAM, semi-linear combinations NLCCA + MLR, ALL + GAM, and CCA + GAM, and linear combinations CCA + MLR, and ALL + MLR (Table 4). In the non-linear combinations, all the NASH

values are near or above 0.9. This suggests that the GAM model can provide more accurate estimates in the NLCCA space.

The RMSE and BIAS indices represent a measure of prediction accuracy on an absolute scale. Both these indices indicate that the method based on the neighborhood approach in conjunction with MLR (NLCCA + MLR) performs better than the methods based on the CCA + MLR and ALL + MLR approaches (Table 4). Furthermore, the combination of the GAM model and NLCCA leads to the best RMSE and BIAS performance among all models used. For instance, the highest RMSE for the non-linear combination NLCCA + GAM is 1.8°C, and the highest BIAS is -0.029°C. It is also worth noting that Both GAM and MLR approaches applied in the non-linear canonical space lead to better results than in the linear canonical space. As a result, these results demonstrate the importance of incorporating non-linearity into the DHR step.

Regarding the RE method, the results obtained with GAM using the same delineation methods are more accurate than those obtained with MLR. This is most likely due to GAM's flexibility and the way it accounts for non-linearities between predictor and response variables.

Based on the relative RRMSE index, NLCCA + GAM also results in better performances (Table 4). For instance, RRMSE for the Tw_5 quantile is 7% with NLCCA + GAM, while it is 10% with CCA + GAM and ALL + GAM. The combination of NLCCA with MLR provides an RRMSE of 8%, which is more accurate than the results obtained with CCA + MLR and ALL + MLR. According to these results, the NLCCA approach improves significantly regional water temperature estimates in comparison to the CCA or the non-neighborhood approaches.

Table 4. Results of cross-validation of all regionalization methods for stream temperatures.

		ALL		CCA (Neighborhood = 18)		NLCCA (Neighborhood = 18)	
Quantiles		GAM	MLR	GAM	MLR	GAM	MLR
NASH							
	Tw_2	0.837	0.750	0.835	0.747	**0.916**	0.856
	Tw_5	0.820	0.727	0.813	0.826	**0.907**	0.877
	Tw_{10}	0.815	0.725	0.796	0.821	**0.902**	0.876
	Tw_{20}	0.812	0.724	0.78	0.814	**0.897**	0.874
	Tw_{50}	0.807	0.721	0.752	0.806	**0.889**	0.871
	Tw_{100}	0.805	0.719	0.732	0.799	0.868	**0.869**
RMSE (°C)							
	Tw_2	1.768	2.188	1.775	2.201	**1.265**	1.659
	Tw_5	1.939	2.393	1.977	1.906	**1.390**	1.604
	Tw_{10}	2.018	2.46	2.12	1.987	**1.465**	1.651
	Tw_{20}	2.079	2.519	2.247	2.064	**1.535**	1.697
	Tw_{50}	2.156	2.294	2.444	2.164	**1.633**	1.759
	Tw_{100}	2.209	2.650	2.586	2.238	1.817	**1.807**
BIAS (°C)							
	Tw_2	-0.093	0.034	-0.268	-0.174	**-0.029**	-0.067
	Tw_5	-0.106	-0.144	-0.316	-0.040	**-0.017**	0.090
	Tw_{10}	-0.112	-0.142	-0.374	-0.043	**-0.005**	0.101
	Tw_{20}	-0.116	-0.140	-0.382	-0.045	**0.007**	0.112
	Tw_{50}	-0.121	-0.136	-0.390	-0.046	**-0.010**	0.126
	Tw_{100}	-0.123	-0.133	-0.424	-0.047	**-0.020**	0.136
RRMSE (%)							
	Tw_2	9.792	10.533	9.407	10.249	**7.441**	8.045
	Tw_5	10.140	13.773	9.996	10.535	**7.722**	8.855
	Tw_{10}	10.280	13.85	10.488	10.769	**7.917**	8.939
	Tw_{20}	10.394	13.943	10.905	11.021	**8.121**	9.042
	Tw_{50}	10.573	14.097	11.613	11.367	**8.444**	9.213
	Tw_{100}	10.698	14.228	12.133	11.632	**9.057**	9.352

Fig. 11 illustrates the local and regional estimates of Tw_5 for the six combinations considered in this research. According to Fig. 11, the full non-linear model NLCCA + GAM shows better overall performances, followed by the semi-linear model NLCCA + MLR. Therefore, the regional and local estimates are close because the points are less dispersed around the diagonal in the case of the combined models with the NLCCA. At the same time, the variance is larger when using the CCA neighborhood method or even without the neighborhood. These results clearly illustrate the importance of using the non-linear tools in both steps of the RFA process.

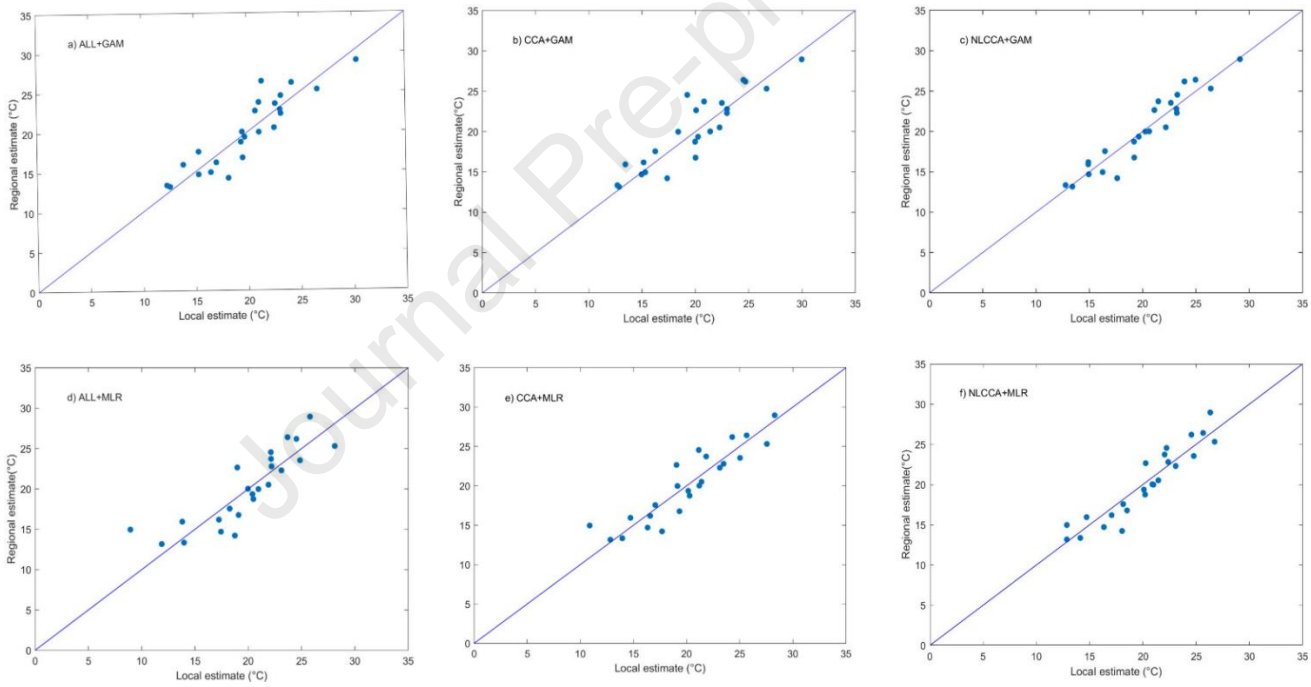


Fig. 11 Regional and local estimates for Tw_5 quantile at sites for river temperature. a) ALL + GAM, b) CCA + GAM, c) NLCCA + GAM, d) ALL + MLR, e) CCA + MLR, f) NLCCA + MLR.

Fig. 12 shows the Tw_5 and Tw_{100} regional thermal quantile maps from the best combination of NLCCA and GAM. There is generally an increase in temperature toward the north on both maps. This indicates that the highest thermal quantiles are found in low-altitude regions, whereas the lowest values are located in high-altitude areas. The diameters of the

circles are proportional to the absolute residual error rates of the Tw_5 and Tw_{100} quantiles obtained from NLCCA+GAM. The findings indicate that the highest residual errors are generally associated with lower altitudes. Consequently, the lower the altitude, the greater the temperature, and the greater the error.

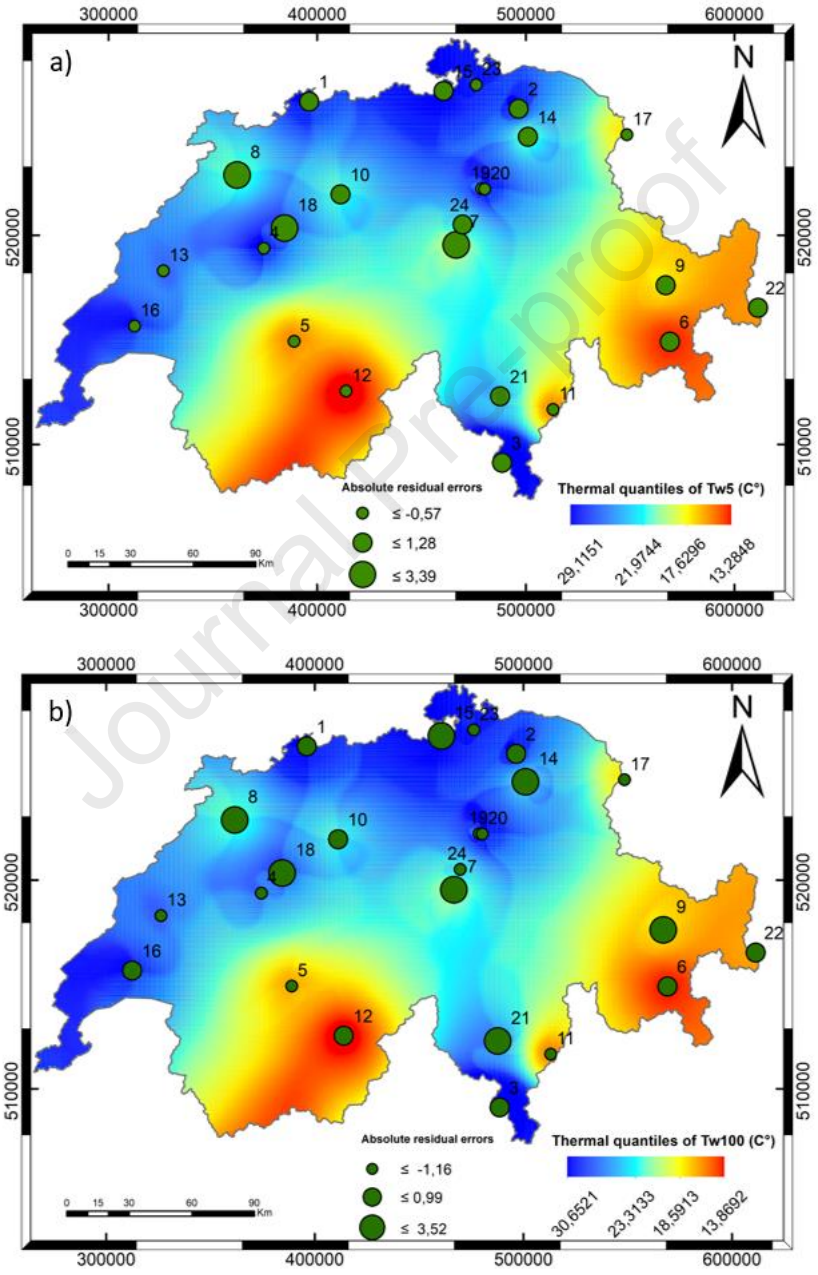


Fig. 12 Maps of thermal quantiles of Tw_5 (a) and Tw_{100} (b), and the absolute residual error in Switzerland using the NLCCA+GAM combination. Station numbers are also provided.

In addition, Table 5 provides cross-validation results of the α-optimization process. Compared to the results of an 18-station neighborhood, the optimization case showed no improvement. This may be related to the fact that the data set considered is not large enough (24 watersheds). Table 5 indicates that the best overall results are achieved with the combination of NLCCA and GAM, when we consider either one of the following performance criteria NASH, RMSE, or RRMSE. On the other hand, the CCA + MLR combination leads to the lowest values of the BIAS. The comparison of the two results reveals that the fixed neighborhood approach is more accurate than the approach based on the optimization of the neighborhood through the parameter α. We can hence conclude that optimizing the parameter α for CCA and NLCCA does not significantly improve the results when the database is composed of a limited number of stations.

Table 5. Cross-validation results for all regionalization methods where the neighborhood determined by the optimization procedure α.

		CCA ($\alpha = 10^{-5}$)		NLCCA ($\alpha = 10^{-14}$)	
	Quantiles	GAM	MLR	GAM	MLR
NASH					
	Tw_2	0.774	0.781	**0.904**	0.853
	Tw_5	0.753	0.828	0.898	**0.899**
	Tw_{10}	0.741	0.824	**0.896**	0.895
	Tw_{20}	0.74	0.819	**0.895**	0.891
	Tw_{50}	0.729	0.813	**0.894**	0.886
	Tw_{100}	0.724	0.808	**0.893**	0.882
RMSE (°C)					
	Tw_2	2.083	2.048	**1.351**	1.678
	Tw_5	2.276	1.895	1.457	**1.455**

	Tw_{10}	2.386	1.968	**1.515**	1.518
	Tw_{20}	2.444	2.036	**1.56**	1.578
	Tw_{50}	2.559	2.125	**1.592**	1.656
	Tw_{100}	2.625	2.190	**1.632**	1.715
BIAS (°C)					
	Tw_{2}	**-0.122**	-0.216	0.333	-0.153
	Tw_{5}	-0.140	**0.036**	0.344	0.108
	Tw_{10}	-0.175	**0.031**	0.320	0.098
	Tw_{20}	-0.217	**0.027**	0.302	0.089
	Tw_{50}	-0.208	**0.025**	0.240	0.079
	Tw_{100}	-0.216	**0.023**	0.201	0.071
RRMSE (%)					
	Tw_{2}	10.126	9.742	**7.102**	7.822
	Tw_{5}	10.745	10.471	**7.405**	7.948
	Tw_{10}	11.141	10.619	**7.604**	8.131
	Tw_{20}	11.292	10.788	**7.766**	8.326
	Tw_{50}	11.754	11.04	**7.986**	8.6
	Tw_{100}	12.018	11.234	**8.140**	8.802

The best statistics are in bold characters.

6. Discussions

The motivation for this study was the lack of work that investigates methods for the regional estimates of water temperature quantiles. It is found that the proposed procedure produces positive results. According to the cross-validation results, the procedure is more robust when GAM+NLCCA is used.

The RMSE and BIAS values found are low, with the highest being 1.8°C for the non-linear combination (GAM+NLCCA). NASH values are relatively high when compared to other regional estimation methods. There is a consistent improvement in these values over the semi-linear and linear combinations. It is important to note that using the GAM approach in RFA produces positive results. This result is consistent with previous studies conducted

on hydrological variables (Ouarda et al. 2018). Recently, Abidi et al. (2022) demonstrated that the GAM model produces better results for estimating regional water temperatures. The main reason for this is that the GAM model allows for modelling the nonlinear relationships between the response variables and the predictors. The GAM was also used successfully by Laanaya et al. (2017) to estimate daily water temperatures.

In other words, for the DHR step, the case where the NLCCA approach is used shows the best results, including NLCCA+GAM, followed by NLCCA + MLR. This study's results indicate that applying the ANN approach in the physiographic and thermal CCA space can significantly improve the estimation performance over the linear CCA approach or combine all stations without neighbourhood. Ouali et al. (2016) concluded that NLCCA performs better than CCA to estimate flood quantiles. Shu and Ouarda (2007) demonstrated that this approach can characterize the physiographic space better to estimate flood quantiles. The results of our research are consistent with those of this study. Therefore, better results could be obtained by using a more advanced NLCCA parameterization, and the difference between linear and nonlinear approaches to neighbourhood identification in this study is evident. It is in line with the suggestion made by Ouali et al. (2017), who also suggested that using fully nonlinear models (in both RFA steps) is the most appropriate as they provide the best performance and a more realistic description of the physical processes.

On the other hand, semi-linear models that account for non-linearity in the delineation or estimation steps showed little improvement over linear models such as NLCCA+MLR, ALL+GAM and CCA+GAM. For most predictor variables, complex relationships that deviate from linearity can be observed. A similar issue has been raised in previous studies

dealing with regional water temperature estimation. There is an analogy between this research and the research of Ouarda et al. (2022) who focused on the estimation of daily temperatures at ungauged sites.

Nevertheless, the limitations of the NLCCA approach are that it has relatively greater complexity than linear models and requires an extensive database, unlike our study. However, despite these shortcomings, it has produced acceptable results. This explains the robustness of the method.

7. Conclusions and future work

The present study focused on integrating the CCA technique coupled with ANN (NLCCA) for DHR. To evaluate the performance of this method, we compared it with the linear CCA and the ALL neighborhood-free method. Both CCA and ALL approaches are unable to represent the possibility of non-linear relationships between the variables of interest. Within each delimited region, either the GAM or the MLR were used to transfer thermal information. In total, six regionalization models were compared. This study examined the potential of non-linear approaches in both steps of the RFA process simultaneously. A stepwise regression method was employed to select the optimal variables to include in the regional models based on the study dataset. Results indicate that the regional model comprising the NLCCA approach for the DHR step and the GAM approach for the RE (NLCCA+GAM) is the most appropriate, followed by the NLCCA+MLR model. These results show that using a non-linear approach in DHR can significantly improve the performance of regional stream temperature frequency analysis approaches. However, the best results for estimating stream temperature quantiles at ungauged sites are obtained when non-linear approaches are adopted in both steps of the RFA.

Another DHR result derived from a neighborhood optimized through the α parameter for CCA and NLCCA is compared with results obtained using the fixed 18-station neighborhood. Results indicate that the α parameter optimization leads to a marginal performance improvement over the fixed neighborhood for some quantiles in the NLCCA+GAM combination. It is hence concluded that the use of the optimization procedure is not recommended when the database is composed of a limited number of stations.

In future efforts, it may be interesting to investigate the impact of adopting other statistical estimation techniques such as the Random Forest and the multivariate adaptive regression splines models in conjunction with linear and non-linear delineation approaches for estimating water temperature quantiles.

Acknowledgments

Financial support for this work was graciously provided by the National Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chair Program (CRC), and the University Mission of Tunisia in Montreal (MUTAN). The authors are grateful to the Swiss Federal Office for the Environment (FOEN) for the employed data. The authors thank M. Christian Charron for his assistance with some of the codes. The authors also thank the editor, Dr. Dan Ames, and the anonymous reviewers for their comments that improved the quality of the manuscript.

References

Abidi O, St-Hilaire A, Ouarda TBMJ, Charron C, Boyer C, Daigle A (2022) Regional thermal analysis approach: A management tool for predicting water temperature metrics relevant for thermal fish habitat Ecological Informatics 70:101692 doi:https://doi.org/10.1016/j.ecoinf.2022.101692

Alobaidi MH, Marpu PR, Ouarda TBMJ, Ghedira H (2014) Mapping of the Solar Irradiance in the UAE Using Advanced Artificial Neural Network Ensemble IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7:3668-3680 doi:10.1109/JSTARS.2014.2331255

Arismendi I, Safeeq M, Dunham JB, Johnson SL (2014) Can air temperature be used to project influences of climate change on stream temperature? Environmental Research Letters 9:084015 doi:https://doi.org/10.1088/1748-9326/9/8/084015

Beechie TJ et al. (2010) Process-based Principles for Restoring River Ecosystems BioScience 60:209-222 doi:10.1525/bio.2010.60.3.7

Benyahya L, Caissie D, St-Hilaire A, Ouarda TBMJ, Bobée B (2007) A Review of Statistical Water Temperature Models Canadian Water Resources Journal / Revue canadienne des ressources hydriques 32:179-192 doi:10.4296/cwrj3203179

Bishop CM (1995) Neural Networks for Pattern Recognition. Oxford University Press,

Burn DH (1990) Evaluation of regional flood frequency analysis with a region of influence approach Water Resources Research 26:2257-2265 doi:https://doi.org/10.1029/WR026i010p02257

Caissie D (2006) The thermal regime of rivers: a review Freshwater Biology 51:1389-1406 doi:https://doi.org/10.1111/j.1365-2427.2006.01597.x

Caissie D, Satish MG, 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 doi:https://doi.org/10.1016/j.jhydrol.2007.01.008

Chebana F, Ouarda TBMJ (2008) Depth and homogeneity in regional flood frequency analysis Water Resources Research 44 doi:https://doi.org/10.1029/2007WR006771

Connor WP, Piston CE, Garcia AP (2003) Temperature during Incubation as One Factor Affecting the Distribution of Snake River Fall Chinook Salmon Spawning Areas Transactions of the American Fisheries Society 132:1236-1243 doi:https://doi.org/10.1577/T02-159

Demars BO, Russell manson, J., Ólafsson, J.S., Gíslason, G.M., Gudmundsdóttir, R., Woodward, G., Reiss, J., Pichler, D.E., Rasmussen, J.J. and friberg, N. (2011) Temperature and the metabolic balance of streams Freshwater Biology 56:1106-1121 doi:https://doi.org/10.1111/j.1365-2427.2010.02554.x

Desai S, Ouarda TBMJ (2021) Regional hydrological frequency analysis at ungauged sites with random forest regression Journal of Hydrology 594:125861 doi:https://doi.org/10.1016/j.jhydrol.2020.125861

Dugdale SJ, Franssen J, Corey E, Bergeron NE, Lapointe M, Cunjak RA (2016) Main stem movement of Atlantic salmon parr in response to high river temperature Ecology of Freshwater Fish 25:429-445 doi:https://doi.org/10.1111/eff.12224

Durocher M, Chebana F, Ouarda TBMJ (2016) Delineation of homogenous regions using hydrological variables predicted by projection pursuit regression Hydrol Earth Syst Sci 20:4717-4729 doi:10.5194/hess-20-4717-2016

Edwards PA, Cunjak RA (2007) Influence of water temperature and streambed stability on the abundance and distribution of slimy sculpin (Cottus cognatus) Environmental Biology of Fishes 80:9-22 doi:10.1007/s10641-006-9102-8

Edwards R, Densem J, Russell P (1979) An assessment of the importance of temperature as a factor controlling the growth rate of brown trout in streams The Journal of Animal Ecology:501-507 doi:https://doi.org/10.2307/4176

Elliott J, Hurley M, Fryer R (1995) A new, improved growth model for brown trout, Salmo trutta Functional ecology:290-298 doi:https://doi.org/10.2307/2390576

Elliott J, Hurley MA (2001) Modelling growth of brown trout, Salmo trutta, in terms of weight and energy units Freshwater Biology 46:679-692 doi:https://doi.org/10.1046/j.1365-2427.2001.00705.x

Elliott JM, Hurley MA (1997) A functional model for maximum growth of Atlantic Salmon parr, Salmo salar, from two populations in northwest England Functional Ecology 11:592-603 doi:https://doi.org/10.1046/j.1365-2435.1997.00130.x

Geman S, Bienenstock E, Doursat R (1992) Neural Networks and the Bias/Variance Dilemma Neural Computation 4:1-58 doi:https://doi.org/10.1162/neco.1992.4.1.1

Girard C, Ouarda TB, Bobée B (2004) Étude du biais dans le modèle log-linéaire d'estimation régionale Canadian Journal of Civil Engineering 31:361-368 doi:https://doi.org/10.1139/l03-099

GREHYS GDRES (1996) Presentation and review of some methods for regional flood frequency analysis Journal of hydrology(Amsterdam) 186:63-84

Hastie T, Tibshirani R (1987) Generalized Additive Models, Cubic Splines and Penalized Likelihood. STANFORD UNIV CA DEPT OF STATISTICS,

Hebert C, Caissie D, Satish MG, El-Jabi N (2011) Study of stream temperature dynamics and corresponding heat fluxes within Miramichi River catchments (New Brunswick, Canada) Hydrological Processes 25:2439-2455 doi:https://doi.org/10.1002/hyp.8021

Hewlett JD, Fortson JC (1982) Stream temperature under an inadequate buffer strip in the southeast piedmont1 Jawra Journal of the American Water Resources Association 18:983-988 doi:https://doi.org/10.1111/j.1752-1688.1982.tb00105.x

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators Neural Networks 2:359-366 doi:https://doi.org/10.1016/0893-6080(89)90020-8

Hosking JRM, Wallis JR (1993) Some statistics useful in regional frequency analysis Water Resources Research 29:271-281 doi:https://doi.org/10.1029/92WR01980

Hotelling H (1935) The most predictable criterion vol 26. doi:10.1037/h0058165

Howell PJ, Dunham JB, Sankovich PM (2010) Relationships between water temperatures and upstream migration, cold water refuge use, and spawning of adult bull trout from the Lostine River, Oregon, USA Ecology of Freshwater Fish 19:96-106 doi:https://doi.org/10.1111/j.1600-0633.2009.00393.x

Hsieh WW (2000) Nonlinear canonical correlation analysis by neural networks Neural Networks 13:1095-1105 doi:https://doi.org/10.1016/S0893-6080(00)00067-8

Hsieh WW (2001) Nonlinear Canonical Correlation Analysis of the Tropical Pacific Climate Variability Using a Neural Network Approach Journal of Climate 14:2528-2539 doi:https://doi.org/10.1175/1520-0442(2001)014

Krider LA, Magner JA, Perry J, Vondracek B, Ferrington Jr. LC (2013) Air-Water Temperature Relationships in the Trout Streams of Southeastern Minnesota's Carbonate-Sandstone Landscape JAWRA Journal of the American Water Resources Association 49:896-907 doi:https://doi.org/10.1111/jawr.12046

Laanaya F, St-Hilaire A, Gloaguen E (2017) Water temperature modelling: comparison between the generalized additive model, logistic, residuals regression and linear regression models Hydrological Sciences Journal 62:1078-1093 doi:10.1080/02626667.2016.1246799Leclerc M, Ouarda TBMJ (2007) Non-stationary regional flood frequency analysis at ungauged sites Journal of Hydrology 343:254-265 doi:https://doi.org/10.1016/j.jhydrol.2007.06.021

Lin G-F, Chen L-H (2006) Identification of homogeneous regions for regional frequency analysis using the self-organizing map Journal of Hydrology 324:1-9 doi:https://doi.org/10.1016/j.jhydrol.2005.09.009

Lisi PJ, Schindler DE, Bentley KT, Pess GR (2013) Association between geomorphic attributes of watersheds, water temperature, and salmon spawn timing in Alaskan streams Geomorphology 185:78-86 doi:https://doi.org/10.1016/j.geomorph.2012.12.013

Lund SG, Caissie D, Cunjak RA, Vijayan MM, Tufts BL (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 doi:https://doi.org/10.1139/f02-117

Mann B (1945) H.(1945) Non-Parametric Test Against Trend Econometrica 13 doi:https://doi.org/10.2307/1907187

Msilini A, Masselot P, Ouarda TBMJ (2020) Regional Frequency Analysis at Ungauged Sites with Multivariate Adaptive Regression Splines Journal of Hydrometeorology 21:2777-2792 doi:https://doi.org/10.1175/JHM-D-19-0213.1

Msilini A, Ouarda TBMJ, Masselot P (2022) Evaluation of additional physiographical variables characterising drainage network systems in regional frequency analysis, a Quebec watersheds case-study Stochastic Environmental Research and Risk Assessment 36:331-351 doi:https://doi.org/10.1007/s00477-021-02109-7

Ouali D, Chebana F, Ouarda TBMJ (2016) Non-linear canonical correlation analysis in regional frequency analysis Stochastic Environmental Research and Risk Assessment 30:449-462 doi:10.1007/s00477-015-1092-7

Ouali D, Chebana F, Ouarda TBMJ (2017) Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites Journal of Advances in Modeling Earth Systems 9:1292-1306 doi:https://doi.org/10.1002/2016MS000830

Ouarda T, Charron C, St-Hilaire A (2022) Regional estimation of river water temperature at ungauged locations Journal of Hydrology X:100133 doi:https://doi.org/10.1016/j.hydroa.2022.100133

Ouarda T, St-Hilaire A, Bobée B (2008a) Synthèse des développements récents en analyse régionale des extrêmes hydrologiques Revue des sciences de l'eau / Journal of Water Science 21:219-232 doi:https://doi.org/10.7202/018467ar

Ouarda TB, Haché M, Bruneau P, Bobée B (2000) Regional flood peak and volume estimation in northern Canadian basin Journal of cold regions engineering 14:176-191 doi:https://doi.org/10.1061/(ASCE)0887-381X(2000)14:4(176)

Ouarda TBMJ et al. (2008b) Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study Journal of Hydrology 348:40-58 doi:https://doi.org/10.1016/j.jhydrol.2007.09.031

Ouarda TBMJ, Charron C, Hundecha Y, St-Hilaire A, Chebana F (2018) Introduction of the GAM model for regional low-flow frequency analysis at ungauged basins and comparison with commonly used approaches Environmental Modelling & Software 109:256-271 doi:https://doi.org/10.1016/j.envsoft.2018.08.031

Ouarda TBMJ, Charron C, Marpu PR, Chebana F (2016) The Generalized Additive Model for the Assessment of the Direct, Diffuse, and Global Solar Irradiances Using SEVIRI Images, With Application to the UAE IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9:1553-1566 doi:10.1109/JSTARS.2016.2522764

Ouarda TBMJ, Girard C, Cavadias GS, Bobée B (2001) Regional flood frequency estimation with canonical correlation analysis Journal of Hydrology 254:157-173 doi:https://doi.org/10.1016/S0022-1694(01)00488-7

Ouarda TBMJ, Shu C (2009) Regional low-flow frequency analysis using single and ensemble artificial neural networks Water Resources Research 45 doi:https://doi.org/10.1029/2008WR007196

Ouellet V, Secretan Y, St-Hilaire A, Morin J (2014) DAILY AVERAGED 2D WATER TEMPERATURE MODEL FOR THE ST. LAWRENCE RIVER River Research and Applications 30:733-744 doi:https://doi.org/10.1002/rra.2664

Pandey GR, Nguyen VTV (1999) A comparative study of regression based methods in regional flood frequency analysis Journal of Hydrology 225:92-101 doi:https://doi.org/10.1016/S0022-1694(99)00135-3

Qiu R, Wang Y, Wang D, Qiu W, Wu J, Tao Y (2020) Water temperature forecasting based on modified artificial neural network methods: Two cases of the Yangtze River Science of The Total Environment 737:139729 doi:https://doi.org/10.1016/j.scitotenv.2020.139729

Rahman A, Charron C, Ouarda TBMJ, Chebana F (2018) Development of regional flood frequency analysis techniques using generalized additive models for Australia Stochastic Environmental Research and Risk Assessment 32:123-139 doi:https://doi.org/10.1007/s00477-017-1384-1

Saadi AM, Msilini A, Charron C, St-Hilaire A, Ouarda TBMJ (2022) Estimation of the area of potential thermal refuges using generalized additive models and

multivariate adaptive regression splines: A case study from the Ste-Marguerite River River Research and Applications 38:23-35 doi:https://doi.org/10.1002/rra.3886

Shu C, Ouarda TBMJ (2007) Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space Water Resources Research 43 doi:https://doi.org/10.1029/2006WR005142

Sinokrot BA, Stefan HG (1993) Stream temperature dynamics: Measurements and modeling Water Resources Research 29:2299-2312 doi:https://doi.org/10.1029/93WR00540

Souaissi Z, Ouarda TBMJ, St-Hilaire A (2021) River water temperature quantiles as thermal stress indicators: Case study in Switzerland Ecological Indicators 131:108234 doi:https://doi.org/10.1016/j.ecolind.2021.108234

St-Hilaire A, El-Jabi N, Caissie D, Morin G (2003) 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 doi:https://doi.org/10.1002/hyp.1242

St-Hilaire A, Ouarda TBMJ, Bargaoui Z, Daigle A, Bilodeau L (2012) Daily river water temperature forecast model with a k-nearest neighbour approach Hydrological Processes 26:1302-1310 doi:https://doi.org/10.1002/hyp.8216

Statsoft I (1995) Statistica for Windows (Computer program manual) Tulsa, USA

Steedman RJ, France RL, Kushneriuk RS, Peters RH (1998) Effects of riparian deforestation on littoral water temperatures in small boreal forest lakes Boreal Environment Research 3:161-170

Sundt-Hansen LE et al. (2018) Modelling climate change effects on Atlantic salmon: Implications for mitigation in regulated rivers Science of The Total Environment 631-632:1005-1017 doi:https://doi.org/10.1016/j.scitotenv.2018.03.058

Thomas DM, Benson MA (1970) Generalization of streamflow characteristics from drainage-basin characteristics

Wahba G (1990) CBMS-NSF Regional Conference Series in Applied Mathematics Based on a series of 10:23-27

Wahli T, Bernet D, Segner H, Schmidt-Posthaus H (2008) Role of altitude and water temperature as regulating factors for the geographical distribution of Tetracapsuloides bryosalmonae infected fishes in Switzerland Journal of Fish Biology 73:2184-2197 doi:https://doi.org/10.1111/j.1095-8649.2008.02054.x

Wald A, Wolfowitz J (1943) An Exact Test for Randomness in the Non-Parametric Case Based on Serial Correlation The Annals of Mathematical Statistics 14:378-388

Wazneh H, Chebana F, Ouarda TBMJ (2016) Identification of hydrological neighborhoods for regional flood frequency analysis using statistical depth function Advances in Water Resources 94:251-263 doi:https://doi.org/10.1016/j.advwatres.2016.05.013

Wilcoxon F (1946) Individual Comparisons of Grouped Data by Ranking Methods Journal of Economic Entomology 39:269-270 doi:10.1093/jee/39.2.269

Woldesellasse H, Marpu PR, Ouarda TBMJ (2020) Long-term forecasting of wind speed in the UAE using nonlinear canonical correlation analysis (NLCCA) Arabian Journal of Geosciences 13:962 doi:10.1007/s12517-020-05981-9

Wood SN (2003) Thin plate regression splines Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65:95-114 doi:https://doi.org/10.1111/1467-9868.00374

Wood SN (2006) Generalized Additive Models: An Introduction with R. CRC Press,

Wu A, Hsieh WW, Zwiers FW (2003) Nonlinear Modes of North American Winter Climate Variability Derived from a General Circulation Model Simulation Journal of Climate 16:2325-2339 doi:https://doi.org/10.1175/2776.1

Zeni JO, Pérez-Mayorga MA, Roa-Fuentes CA, Brejão GL, Casatti L (2019) How deforestation drives stream habitat changes and the functional structure of fish assemblages in different tropical regions Aquatic Conservation: Marine and Freshwater Ecosystems 29:1238-1252 doi:https://doi.org/10.1002/aqc.3128

Highlights

- Improve the estimation of water temperature extremes at ungauged sites.
- Incorporate non-linearities in the homogenous region delineation step using NLCCA.
- Consider non-linear models in the whole estimation procedure (NLCCA+GAM).
- Compare fully and partially non-linear approaches for water temperature regionalization.
- The results underline the importance of considering the non-linearity of thermal processes.

Software and/or data availability

Software availability

Program language: MATLAB and R

Developers: Zina Souaissi, Dhouha Ouali, Christian Charron

Hardware requirements: PC

Availability: Codes are available from the author

Data availability

Data will be made available on request.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: