WATER TEMPERATURE MONITORING IN EASTERN CANADA: A CASE STUDY FOR NETWORK OPTIMIZATION

ANDRÉ ST-HILAIRE[1,2], CLAUDINE BOYER[1], NORMAND BERGERON[1] & ANIK DAIGLE[1,3]
[1]INRS-ETE, Canada
[2]Canadian Rivers Institute, Canada
[3]Garneau College, Canada

ABSTRACT

Water temperature is a key variable affecting important water quality parameters such as dissolved oxygen. In Eastern Canada, iconic fish species such as Atlantic salmon (*Salmo salar*) can be affected by increase in temperature associated with climate change. A major endeavour is underway to establish and optimize a water temperature monitoring network in this region. This network, called RivTemp, includes temperature data from over 600 stations in 277 streams or rivers. These data are being used to develop/adapt methods network optimization, temperature interpolation and modelling/forecasting. Different approaches to interpolate water temperature at ungauged sites using data from monitoring stations are being compared. More recently, two regression approaches that are often used when collinearity is present among predictors, the ridge regression and the LASSO regression were compared. Results show that the LASSO regression is more parsimonious than the ridge regression and provides adequate estimates of daily average water temperature.

Keywords: water temperature, network, monitoring model regression.

1 INTRODUCTION

During the late 20th century, water temperature monitoring in Eastern Canada consisted mostly of spot (e.g. monthly) measurements on a few rivers. In the late 1990s, there was growing interest in the study of thermal processes and water temperature modelling [1]. In recent years, this interest has expanded considerably, and monitoring of this variable has become a priority for water resources managers. In 2013, a workshop on the development and implementation of a water temperature monitoring network for rivers in Eastern Canada was held in Québec City [2]. Workshop participants concluded that there was a need to compile a comprehensive inventory of historical temperature data. In addition, the implementation of data collection with proper protocols and methods to optimize the network were required. This paper present the progress made in the establishment of the RivTemp network and a comparison of two regression approaches to allow for the estimation of temperature at ungauged sides, as a step towards network optimization.

2 RIVTEMP NETWORK

The RivTemp network (www.rivtemp.ca) is bringing together key partners concerned with water temperature issues. The initial focus is on Atlantic salmon rivers in Eastern Canada. This fish species has known temperature *preferenda*, including optimal growth between 16 and 20°C. Climate change scenarios predict that the frequency of exceedance of the 20°C threshold will be increasing during the forthcoming decades (e.g. [3], [4]). This will also have repercussions on other water quality variables.

The stated objectives of RivTemp are to share data and increase the knowledge about river temperature; to develop management tools and communicate relevant statistics; and to optimize the number of temperature monitoring stations within specific areas. The network is a partnership between universities, federal and provincial governments, watershed groups

WIT Transactions on Ecology and the Environment, Vol 228, © 2018 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/WP180251

and organisations, many of which are working for the conservation of Atlantic salmon. The centralized database collates water temperature data from the network. These data have been used to produce thermal "report cards" with relevant descriptive statistics (seasonal or monthly maximums, means, threshold exceedances, degree-days, etc.). Currently, the database includes data from over 600 stations in 277 Eastern Canadian Rivers (Fig. 1). Most stations are seasonal (June to October). Nearly half (296) of the stations have less than two years of data, while 244 have between two and ten years of continuous (hourly) temperature measurements. 61 stations have relatively long time series (≥ 10years), with the longest time series of the region having 26 years of data (Trinité River) and is still in operation.

A number of researchers are currently using these data to develop methods of network optimization. As can be seen in Fig. 1, some drainage basins have a relatively dense network of stations. In such cases, there is a strong risk of redundancy in the information content of temperature series from stations that are highly correlated with one another. The next section presents a case study from one region with many monitoring stations, the Gaspé Peninsula in Québec, Canada.

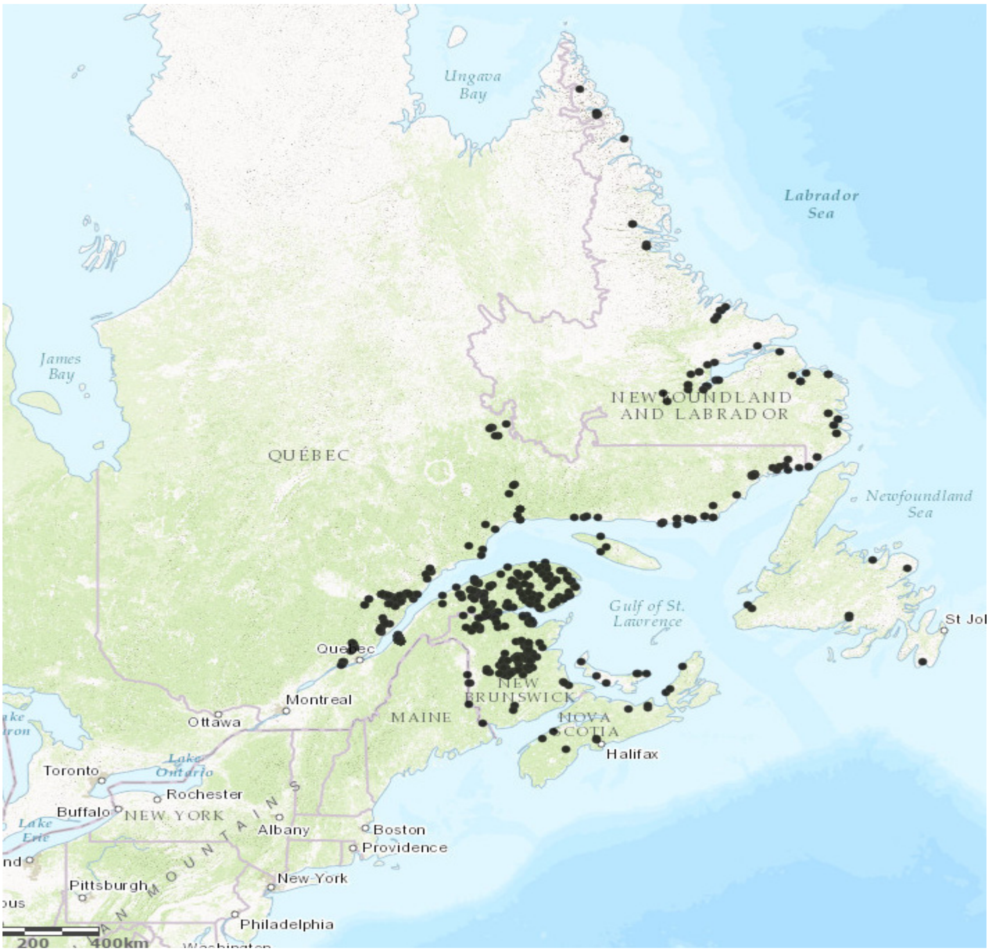


Figure 1: Spatial distribution of RivTemp temperature monitoring stations.

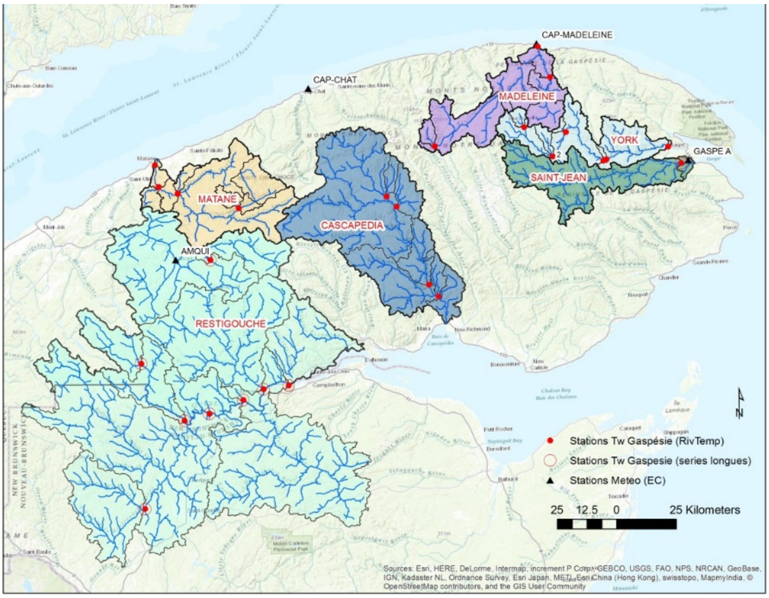
3 ESTIMATION OF WATER TEMPERAURE AT UNGAUGED SITES:A CASE STUDY

One approach in network optimization/rationalization is to eliminate from the network stations for which water temperature can be estimated using neighbouring stations. This approach is implemented in two phases in this case study. As a first step, relatively homogenous groups of stations are defined using the air-water temperature relationship. Subsequently, a simple statistical model (regression method that minimizes the impact of correlation between predictors) is used to verify that water temperature can be adequately estimated at stations where monitoring ceases from stations with ongoing monitoring. Hence, the regression approach, used within each homogenous group, allows to identify stations that could be eliminated without important loss of information.

3.1 Grouping of stations

The study region is located in the eastern part of the province of Québec (Canada) and is called the Gaspé Peninsula. Water temperature loggers were initially deployed in 2014 and concomitant data from 2014 and 2015 in 32 stations were used in the present study. Stations are located in six different drainage basins (Fig. 2).

Physical variables characterizing each sub-basin (including geological information and surficial deposits) were used to group stations that may have similar thermal regimes. The water-air temperature relationship was also used to determine if the thermal signature differs from one station to another. Linear regression coefficients were compared using an Analysis of Covariance (ANCOVA). Principal Component analysis was performed on the physical variables and the slopes of the air-water temperature regressions. This multivariate approach allows to determine if certain physical proxies can assist in grouping stations with similar thermal behaviour.



Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Figure 2: Position of meteorological stations (black triangles) and water temperature stations (Tw, red dots).

3.2 Regressions methods

The following step consisted of developing models that can estimate water temperature at sites that would be deemed redundant and potentially eliminated from the network. Two regression approaches were compared in the present study: Multiple Ridge Regression (MRR; [5]) and the Least Absolute Shrinkage and Selection Operator, or LASSO regression [6]. Both parametric methods are typically used to minimize the impact of multicolinearity (correlation between predictors) used to estimate water temperature (i.e. water temperature from other stations and air temperature). Classical linear multiple regression uses the following general model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_n x_{ki} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), \tag{1}$$

where Y_{i} is the i$^{\mathrm{th}}$ water temperature observation (dependent variable) and x_{ji} is the i$^{\mathrm{th}}$ observation of the j$^{\mathrm{ith}}$ (j=1.k) predictor. The coefficients (β_{j}) quantifies the portion of predictor variability explained by the j$^{\mathrm{th}}$ predictor. These coefficients are estimated by a least square approach:

$$\beta = (\mathrm{XX}')^{-1}\mathrm{X}'\mathrm{Y}. \tag{2}$$

β estimates become unstable in the presence of multicolinearity between predictors, i.e., small changes in the data can lead to considerable changes in the values of the regression coefficients. One way to partially alleviate this problem is to use a biased estimate of the parameters to impart more stability in their estimation. This is the approach used in both the ridge and the LASSO regressions. For the ridge regression, eqn (2) is modified as follows:

$$\beta = (\mathrm{XX}' + k\mathrm{I})^{-1}\mathrm{X}'\mathrm{Y}, \tag{3}$$

where k is the ridge parameter, obtained from graphing the so-called ridge trace, which shows the variation of the regression coefficient values as a function of k. The selected ridge parameter value is the smallest one that minimizes the error while at the same time providing stable regression coefficient estimates. Hence, the ridge regression solution minimizes the following equation:

$$\textstyle\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + k\sum_j \beta_j^2. \tag{4}$$

However, this approach does not allow to eliminate predictors, i.e. it does not yield values of β_j= 0 [6].

The second type of regression, the LASSO, estimates coefficients conditional to a different bias parameter t:

$$\beta = arg\ min\left\{\textstyle\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2\right\} \text{ conditional to } \textstyle\sum_j |\beta_j| \le t,\ t > 0. \tag{5}$$

The t parameter has the same function as the ridge parameter k. It biases the parameter estimation in order to gain more stability in the coefficient estimates. t values are such that some coefficients β_{j} can be set to zero (this is referred to as “shrinkage”) and hence the LASSO regression may yield more parsimonious models than the ridge regression.

Both Ridge and LASSO regressions were applied at each site within each homogenous group of stations. In each case, the model was calibrated using a leave-one-out procedure and RMSE are estimated for each individual model as a goodness of fit measure.

4 RESULTS

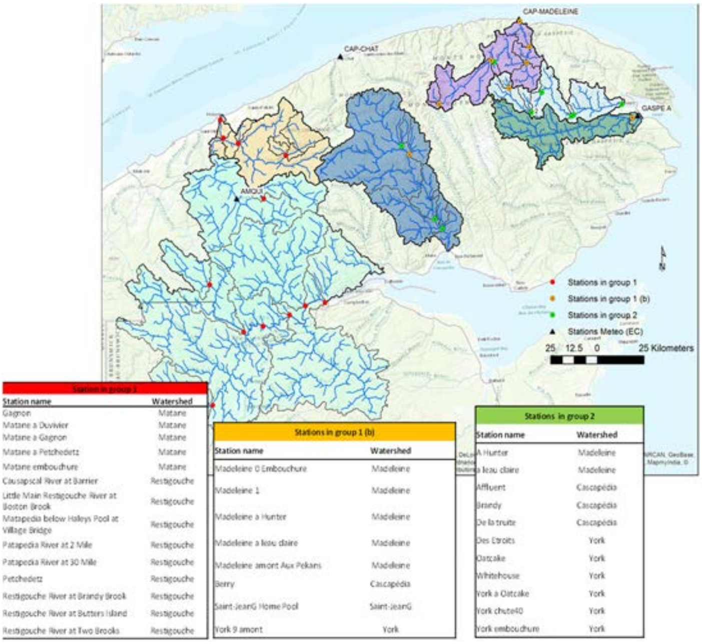
4.1 ANCOVA results

The analysis of covariance of the relation between air and water temperature resulted in the creation of two groups of stations (Fig. 3). All of the stations from the Restigouche and Matane drainage basins are in Group 1, but those from the Madeleine, Cascapédia and York basins, are found in both groups, indicating that the slope of the air-water temperature relationship is not solely dependent on geographic location.

4.2 Regression results

For each station, potential predictors were initially identified by correlation analysis between water temperature of each stations (r >0.7), with the condition that only one station per river be selected. In addition, air temperature was included as a potential predictor for each model. This pre-selection limits the number of potential predictors to one by river in each group. Both Ridge and LASSO regressions were applied by station group. In each case, potential predictors in the multiple regression included air temperature and water temperature from all highly correlated sites in the same group.

As an example, Fig. 4(a) shows the Ridge trace (*k*) obtained for the three predictors identified by correlation analysis used to estimate water temperature at the Matane



Station in group 1

Station name	Watershed
Gagnon	Matane
Matane a Duvivier	Matane
Matane a Gagnon	Matane
Matane a Petchedetz	Matane
Matane embouchure	Matane
Causapscal River at Barrier	Restigouche
Little Main Restigouche River at Boston Brook	Restigouche
Matapedia below Halleys Pool at Village Bridge	Restigouche
Patapedia River at 2 Mile	Restigouche
Patapedia River at 30 Mile	Restigouche
Petchedetz	Restigouche
Restigouche River at Brandy Brook	Restigouche
Restigouche River at Butters Island	Restigouche
Restigouche River at Two Brooks	Restigouche

Stations in group 1 (b)

Station name	Watershed
Madeleine 0 Embouchure	Madeleine
Madeleine 1	Madeleine
Madeleine a Hunter	Madeleine
Madeleine a leau claire	Madeleine
Madeleine amont Aux Pekans	Madeleine
Berry	Cascapédia
Saint-JeanG Home Pool	Saint-JeanG
York 9 amont	York

Stations in group 2

Station name	Watershed
A Hunter	Madeleine
a leau claire	Madeleine
Affluent	Cascapédia
Brandy	Cascapédia
De la truite	Cascapédia
Des Etroits	York
Oatcake	York
Whitehouse	York
York a Oatcake	York
York chute40	York
York embouchure	York

Figure 3: Stations grouping based in ANCOVA results. Stations in orange are those that could be in either group.

Embouchure Station. The trace indicates how parameter estimations change as a function of the selected *k* value (eqn (3)). It can be seen that when *k* reaches a value of 300, all regression coefficients are relatively stable, although their value continue to decrease slowly with increasing *k*. Therefore, at *k* = 300, the ridge regression model is relatively robust. However, although one coefficient has a smaller value than the other two when *k*=300, this value is not 0.

Variations in the LASSO regression coefficients as a function of the *t* parameter are shown in Fig. 4(b). It can be seen that for *t* = 0.85 (dotted blue line), one regression coefficient has a value of 0, which shows that one of the three potential predictors are deemed redundant and excluded from the final LASSO model. Hence the LASSO regression can be more parsimonious that the Ridge regression. The Root Mean Square Error of the Ridge regression at that station was 0.85°C, while it was smaller (0.64°C) for the LASSO regression. Values of regression coefficients for the two most important predictors (Tw Madeleine and Tw Restigouche) were of the order of 0.3 for the Ridge model. The third predictor, Tair, has a coefficient value of about 0.2. For the LASSO, the coefficient values at *t* = 0.85 is 0.35 for Tw Madeleine and 0.4 for Tw Restigouche, while Tair as a coefficient of 0.

The same comparison was performed on all stations. For Group 1, RMSE for the Ridge regression varied between 0.47 and 0.8°C, with an average of 0.69°C. For the same group, LASSO RMSE varied between 0.39 and 0.56°C, with a mean of 0.5°C. Of the 22 stations for which the model was applied in Group 1, 15 LASSO regressions were more parsimonious than the Ridge regression, eliminating one (10 stations) or two (5 stations) predictors that were included in the Ridge model.

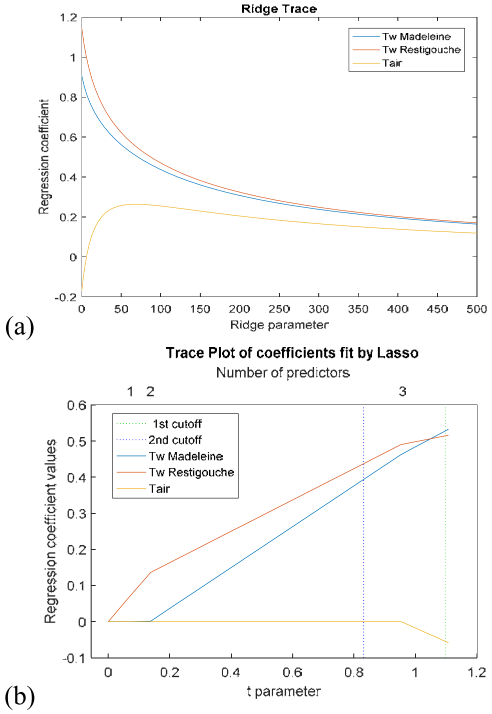


Figure 4: (a) Ridge Trace for the Matane Embouchure Station; (b) LASSO regression coefficient values as a function of *t* (eqn (5)).

For Group 2, fewer predictors were available (only three river basins, see Fig. 3), which may explain in part the higher RMSE values found for the regression models. Ridge regression RMSE values ranged between 0.54 and 3.5°C (mean = 1.69°C). LASSO regressions have RMSE that varied between 0.26 and 0.99°C (mean = 0.49°C). The LASSO models were therefore more performant. In addition, they were more parsimonious than the Ridge regression for six of the 11 stations.

5 CONCLUSION

The construction of the Rivtemp database in Eastern Canada offers water resources managers a unique opportunity to enhance their knowledge of the thermal regime of rivers and its spatio-temporal variability in this region. Network optimization/rationalization must be accomplished with minimum loss of information. This can be achieved by ensuring that stations that are selected to remain in the network are good predictors for those that are considered for elimination. The LASSO regression is a relatively simple, efficient and robust tool for estimating water temperature at those sites.

ACKNOWLEDGEMENT

This work was funded in part by a grant from the ENGAGE program of the Natural Sciences and Engineering Council (NSERC) of Canada, in partnership with ENGLOBE.

REFERENCES

[1] Caissie, D., El-Jabi, N. & St-Hilaire, A., Modelling of stream water temperatures in a small stream using different air to water relations. *Canadian Journal of Civil Engineering*, **25**(2), pp. 250–260, 1998.

[2] Benyahya, L. et al., Workshop on the development & implementation of a water temperature monitoring network for Atlantic salmon (*Salmo salar*) rivers in Eastern Canada held in Quebec City, Quebec, 22–23 January 2014: Abstracts and proceedings. *Can. Manuscr. Rep. Fish. Aquat. Sci.,* **3045**, p. 14, 2014.

[3] Jeong, D.I., Daigle, A. & St-Hilaire, A., Development of a stochastic water temperature model and projection of future water temperature and extreme events in the Ouelle river basin in Québec, Canada. *River Research and Applications,* **29**, pp. 805–821, 2013. DOI: 10.1002/rra.2574.

[4] Daigle, A., Jeong, D.I. & Lapointe, M.F., Climate change and resilience of tributary thermal refugia for salmonids in Eastern Canadian rivers. *Hydrological Sciences Journal*, 2014. DOI: 10.1080/02626667.2014.898121.

[5] Hoerl, A. & Kennard, R., Ridge regression: biased estimation for nonorthogonal problems. *Technometrics,* **12**(1), pp. 55–67, 1970.

[6] Tibhsirani, R., Regression shrinkage and selection via de Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), pp. 267–288, 1996.