REGIONAL ANALYSIS AND MODELLING OF WATER TEMPERATURE METRICS FOR ATLANTIC SALMON (*SALMO SALAR*) IN EASTERN CANADA

C. Charron[1], A. St-Hilaire[1,2], C. Boyer[1], T. B.M.J. Ouarda[1]
A. Daigle[1,2], N.E. Bergeron[1,2],

1. INRS-ETE, Québec City, 490 de la Couronne, Québec City, Qc, G1K 9A9
2. Centre Interuniversitaire de recherche sur le saumon Atlantique

May 2019

ISBN : 978-2-89146-919-7 (version électronique)
Dépôt légal - Bibliothèque et Archives nationales du Québec, 2019
Dépôt légal - Bibliothèque et Archives Canada, 2019

TABLE OF CONTENTS

1 INTRODUCTION 4

2 METHODS 4

2.1 Selection of water temperature monitoring stations 6

2.2 water temperature metrics 8

2.3 climatic data 9

2.4 Physiographic data 10

3 RESULTS 13

4 DISCUSSION 21

5 CONCLUSION 23

6 REFERENCES 23

7 APPENDIX A: SCATTER PLOTS OF OBSERVED VS ESTIMATED WATER TEMPERATURE METRICS 25

8 APPENDIX B: **HIERARCHICAL CLUSTER ANALYSIS** 26

9 APPENDIX C: REGION OF INFLUENCE METHOD 27

10 APPENDIX D: GENERALIZED ADDITIVE MODEL 28

Reference to be cited:

Charron, C., A. St-Hilaire, C. Boyer, T. B.M.J. Ouarda, A. Daigle, N.E. Bergeron. 2019. Regional analysis and modelling of water temperature metrics for Atlantic salmon (*salmo salar*) in eastern Canada. INRS Scientific Report #1855. 29 pages.

LIST OF FIGURES

Figure 1. Spatial distribution of the water temperature stations used in this study. 8

Figure 2. Position of Canadian weather stations used to interpolate climate data at 10 km resolution (Source: Lepage and Bourgeois, 2011). 10

Figure 3. Example of a hierarchical classification tree for the parameter Gaussian_*a* where only the first 30 leaf nodes are presented. 13

Figure 4. Spatial distribution of the regions with HCA. 14

Figure 5. Smooth functions for MaxWaterTmax. The dashed lines represent the 95% confidence intervals and dots are the residuals. 16

Figure 6. Smooth functions for MaxNumDay. The dashed lines represent the 95% confidence intervals and dots are the residuals. 17

Figure 7. Smooth functions for the Gaussian_*a* parameter. The dashed lines represent the 95% confidence intervals and dots are the residuals. 18

Figure 8. Smooth functions for the Gaussian_*b* parameter. The dashed lines represent the 95% confidence intervals and dots are the residuals. 19

Figure 9. Smooth functions for the Gaussian_*c* parameter. The dashed lines represent the 95% confidence intervals and dots are the residuals. 20

LIST OF TABLES

Table 1. Numbers of station and rivers monitors by province (including Environment Canada stations). 7

Table 2. List of the water temperature metrics and explanatory variables with descriptive statistics. 9

Table 3. Physiographic data compiled in the RivTemp database tables and source documents used. 11

Table 4. List of the water temperature metrics and explanatory variables with descriptive statistics. 12

Table 5. Explanatory variables selected using the stepwise forward regression procedure. **Erreur ! Signet non déf**

Table 6. Performance statistics for the different approaches using the leave-one-out validation. 21

1 INTRODUCTION

This report presents the initial findings of a project led by INRS-ETE and members of the *Centre Interuniversitaire de Recherche sur le saumon Atlantique* (CIRSA) on the regional analysis of some water temperature metrics deemed relevant for Atlantic salmon habitat. It is a well-known fact that of all abiotic habitat variables, water temperature is fundamental for stenotherm fish such as salmon. The river thermal regime has important impacts on the growth of juvenile Atlantic salmon (e.g. Nicieza *et al.*, 1997; Elliott and Elliott, 2010; Sundt-Hansen *et al.*, 2018). Thermal stresses are known to have detrimental impacts on juvenile salmon (Corey et al., 2017) and high temperatures may lead to important fish movement as they seek thermal refugia (Dugdale et al., 2017). These refugia are likely to become key habitat components in the context of climate change (Jeong et al., 2013; Daigle et al., 2015).

Although the temperature monitoring effort is on the rise in Eastern Canada (Boyer et al., 2016), there is still a relative paucity of thermal data on many Atlantic salmon rivers. In this context, the development of modelling and analysis tools that allow to estimate relevant water temperature metrics (i.e. descriptive statistics deemed important for Atlantic salmon) in rivers where there is little or no data is of the utmost importance.

To achieve this, the approach used in hydrology for Regional Frequency Analysis (RFA) for flood or low flow quantile estimation was adapted to water temperature. With the objective of estimating high or low flow extremes at ungauged sites, RFA includes two main steps: 1) Defining groups (regions, neighborhoods) of rivers with relatively homogenous hydrological behaviour; and 2) transferring the information from gauged sites to the target ungauged location in order to estimate flow quantiles of interest at this site. The same two main steps are developed in the present project, with the exception that flow quantiles are replaced by temperature metrics, as described in the following section.

2 METHODS

The first methodological step consists in defining groups of rivers that are characterized by a relatively similar thermal behaviour. In RFA, a number of methods have been used to establish these homogenous groups of rivers. One simple approach could be to define geographic groups of contiguous drainage basins with similar temperature regimes. An alternative to this approach is to define groups based on similarities other than location, using different climatic and physiographic characteristics of the drainage basins. This approach was used in the present study and potentially non-contiguous regions were constructed using Hierarchical Clustering

Analysis (HCA), which groups rivers based on calculating a statistical (Euclidian) distance in the multidimensional space defined by selected climatic and physiographic descriptors. HCA minimizes within-group differences and maximizes between-group differences. It can be done in ascending (i.e. starting with all stations in separate groups and coalescing them) or descending order, i.e. starting with one group including all stations and separating them (Johnson, 1967). The latter method was used. The choice of the number of classes is generally made visually from the dendrogram, which is a tree diagram of possible groupings as a function of Euclidean distance. A brief description of the HCA technique is presented in Appendix B.

As an alternative to the definition of contiguous or non-contiguous regions, the Region of Influence (ROI, Burn, 1990) approach was also tested. ROI allows to define, for each target station, the potentially unique set of stations to be used in the estimation of the thermal metrics at the target sites. A brief description of the ROI approach is presented in Appendix C of the present report.

Once homogenous groups of stations have been established, temperature metrics can be estimated using independent variables known to influence water temperature. For each temperature metrics of interest, a statistical model is built to establish the link between the temperature metric and independent variables at gauged sites within the group or ROI. Two statistical models were tested in the present study: Multiple Linear Regression (MLR) and the Generalized Additive Model (GAM).

MLR is a parametric linear model and probably represents the simplest method that can be used for information transfer to the target site. It can be expressed as:

$$Tw(t) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i(t) + \varepsilon \qquad (1)$$

where *Tw* (t) is the temperature metric of interest, β are coefficients to be adjusted, x_i: *xn* are the independent variables (or predictors) and ε is an error term. In the present work, the selection of predictors was performed using a forward stepwise procedure.

The GAM, is an extension of the linear model (McCullagh & Nelder, 1989). It assumes no specific form of dependency between predictand and predictor and hence, the relationship is not necessarily linear. GAM is defined by:

$$g(E(Tw) = f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \varepsilon \qquad (2)$$

where *g* is the link function, E(*Tw)* is the expected value of the predictand (in our case, a temperature metric), x_j is the j^{th} predictor and f_j is the associated smooth nonlinear function (often combination of cubic splines). A more detailed introduction to the method is presented in Appendix D of the present report.

As a first attempt to implement RFA for water temperature metrics in Eastern Canada, six different models were compared: 1) MLR and GAM when all stations are used with no subdivision into groups of rivers or ROI; 2) MLR and GAM applied to distinct groups of rivers identified using hierarchical clustering analysis; 3) MLR and GAM implemented using the ROI approach.

Model performances were compared using a leave-one-out cross validation approach. This is done by estimating the parameters of the model using all stations but one in the region. Using the fitted model, the water temperature metric is estimated for the station that was left out and compared to the metric calculated from observations. This is repeated for each station in the region. Three performance metrics were used. The coefficient of determination (R^2), the root-mean-square error (RMSE) and the bias. Equations for the latter two are:

$$Bias = \frac{1}{n}\sum_{t=1}^{n}\left(\hat{y}_t - y_t\right) \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\left(\sum_{t=1}^{n}\left(\hat{y}_t - y_t\right)^2\right)} \tag{4}$$

where *n* is the sample size, y_t is the simulated temperature metric for the period *t* and $\hat{y}_t$ is the metric calculated from observations during the period *t*.

2.1 Selection of water temperature monitoring stations

RivTemp is a partnership between universities, provincial and federal governments, watershed groups and organizations dedicated to Atlantic salmon conservation (http://rivtemp.ca; Boyer et al., 2016). These partners operate a network of river temperature monitoring stations. The

network also includes many stations monitored by Environment Canada in Atlantic provinces rivers.

Data coming from the monitoring stations are processed in this unique centralized database. RivTemp contains daily water temperature metrics for 787 stations installed into 389 rivers across Québec and the Atlantic Provinces (Table 1).

Table 1. Number of stations and rivers monitored by province (including Environment Canada stations).

	# Stations	# **Rivers**
QC	456	161
NB	153	81
NL	153	124
NS	19	17
PE	6	6
Total	787	389

For this study, we have selected stations installed on rivers and for which at least 5 years of data are available. In the database, 122 stations currently meet this criterion (Figure 1). It should be noted that there is a much higher density of stations in Newfoundland than in other provinces. In addition, some rivers are over represented, i.e. the number of thermographs deployed on these watercourses is higher than in other rivers (Ouelle, Ste-Marguerite, Restigouche, and Miramichi).

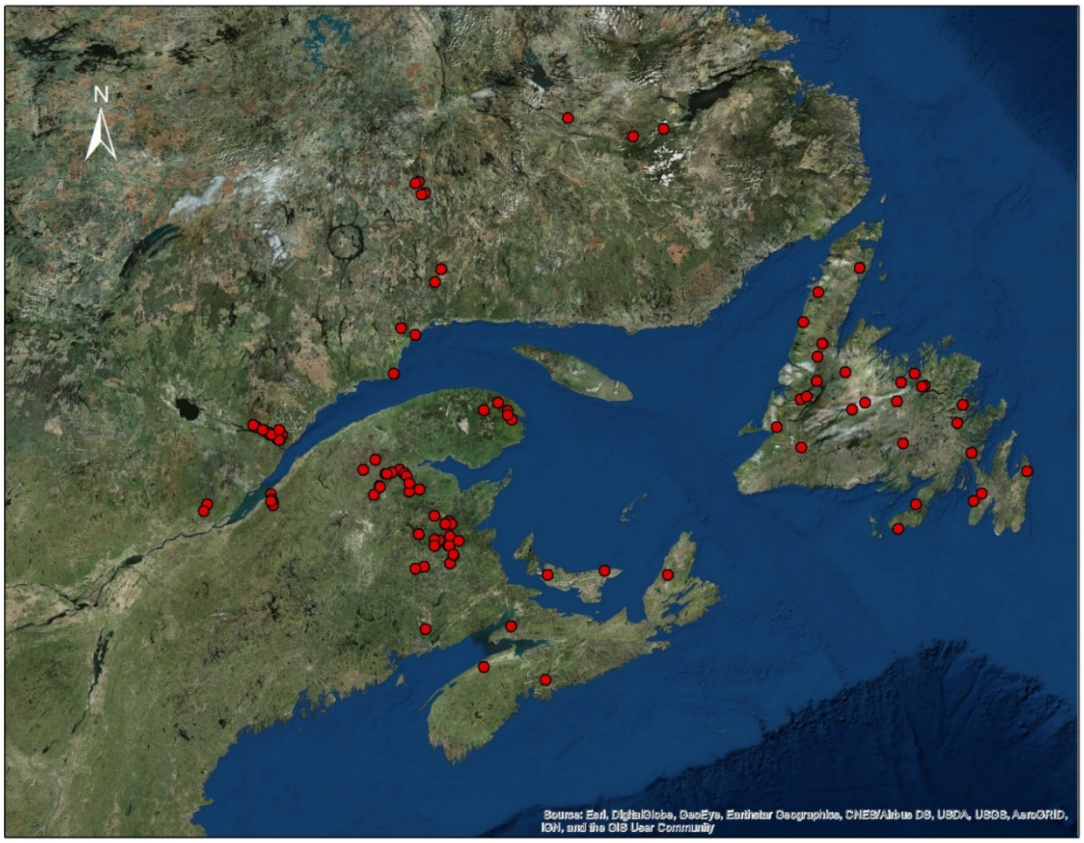


Figure 1. Spatial distribution of the water temperature stations used in this study.

2.2 WATER TEMPERATURE METRICS

To test the approach, a limited number of water temperature metrics known to be relevant for Atlantic salmon were selected. They include:

The *annual water temperature maximum (MaxWaterTmax)* and the *annual maximum number of consecutives days (MaxNumDay)* over a stressful threshold. This stressful threshold was defined by DFO. The water temperature conditions used to close the sport fishery in certain rivers were used, i.e. thresholds of daily Tmin>20°C and Tmax>25°C.

In addition to these two metrics, three other variables were used to test the RFA approach, based on a study completed by Daigle et al. (2019, submitted) in which the thermal regimes of rivers in Québec were characterized using a Gaussian function fitted to the interannual mean daily temperature *T*, as a function of the day of the year *d* (= [1,365]):

$$\hat{T}(d) = a \exp\left(-\frac{1}{2}\left(\frac{d-c}{b}\right)^2\right) \quad (5)$$

Where parameter *a* is a scale factor representing the annual maximum value, *b* is the standard deviation which is a measure of the duration of the warm period, and *c* is the day of occurrence of the maximum. All three parameters were also used as predictands in the RFA models. Table 2 provides the complete list of temperature metrics and their descriptive statistics for the selected stations.

Table 2. List of the water temperature metrics and explanatory variables with descriptive statistics.

Notation	Description	Unit	Mean	Median	Min	Max
MaxWaterTmax	Maximum annual value of Tmax	°C	23.95	23.83	16.96	29.22
MaxNumDay	Maximum number of days where Tmax > 25 °C and Tmin > 20 °C	day	1.09	0.15	0.00	6.92
Gaussian_*a*	Parameter *a* of the Gaussian model for the interannual mean daily water temperature	°C	18.93	18.95	14.08	23.64
Gaussian_*b*	Parameter *b* of the Gaussian model for the interannual mean daily water temperature	day	56.45	56.12	40.20	80.35
Gaussian_*c*	Parameter *c* of the Gaussian model for the interannual mean daily water temperature	day	213.27	212.60	204.94	238.39

2.3 CLIMATIC DATA

As stated previously, climatic data are used to determine groups of homogenous rivers and as predictors to estimate the water temperature metrics. The climatic data used for this study were extracted from the ANUSPLIN database (Hutchinson et al., 2009). These data are interpolated on a 10 km x 10 km grid derived from observations at Canadian meteorological stations (Figure 2). Interpolation from this network is achieved despite a non-uniform spatial distribution of stations (e.g. lack of representation of mountainous areas) and the fact that the northern part of the country is poorly covered. The uncertainty of interpolated data is therefore greater in areas where the number of stations is reduced (Hutchinson, 2009). The interpolated data available are daily air temperature maximum (AirTmax), minimum (AirTmin) and total daily precipitation (TotPrecip).

Daily climate data were extracted using the closest ANUSPLIN grid point to each water temperature station. For the analysis presented in this study, the annual Tairmax (mean, maximum and minimum) values were calculated form daily values.

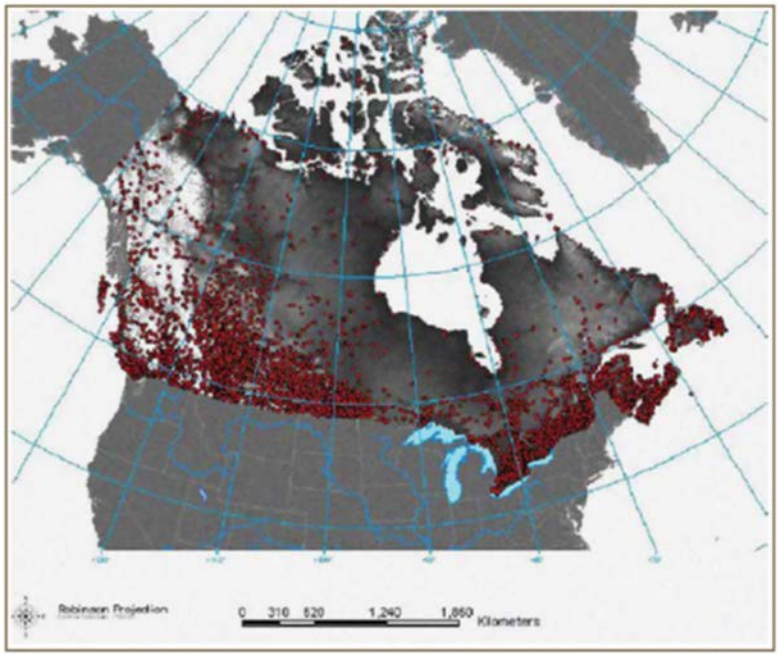


Figure 2. Position of Canadian weather stations used to interpolate climate data at 10 km resolution (Source: Lepage and Bourgeois, 2011).

2.4 PHYSIOGRAPHIC DATA

The physical variables characterizing the watersheds associated with each station of the network have been compiled and added to the RivTemp database. The selected physiographic data were chosen because they may influence water temperature at various levels and explain the spatial variability of this key variable.

These data were extracted from documents accessible through provincial databases and Natural Resources Canada's data dissemination sites, the Geological Survey of Canada and the Consortium for Spatial Information (CGIAR-CSI) (Table 3). The scale of the available products differs from one province to the other. We have made an effort to reduce the loss of information that may be induced by these differences. The largest differences were observed for the surficial deposits.

Table 3. Physiographic data compiled in the RivTemp database tables and source documents used.

VARIABLES	SOURCE DOCUMENT	
Delimitation of drainage area for each station	DEM for Arc Hydro processing	QC : (Ministère de l'environnement et de la lutte aux changements climatiques) - Reconstituted model (25 m) for rivers in the Bas-St-Laurent/Gaspésie and Capitale Nationale regions - Reconstituted model (50 m) for rivers: for rivers in the Côte Nord region NB : Digital Elevation Model of Canada (MNEC) NS : Nova Scotia Department of Natural Resources -Reconstituted model (20 m) NL : Digital Elevation Model of Canada (MNEC) PE : Digital Elevation Model of Canada (MNEC)
Drainage area for each station		
Elevation	DEM	Digital Elevation Model of Canada (MNEC) 50K from Natural Resources Canada
River slope		
Drainage density (Σlength of rivers/area)	WaterLinearFlow1 and Waterbody_2	Canvec 50K from Natural Resources Canada
Lakes total area		
Vegetation	2010 Land Cover of North America at 30 meters (Commission for Environmental Cooperation (CEC))	
Surficial deposits	Main source of data: Geological Survey of Canada (scale varies by basin) QC: Geological Survey of Canada, Bas St-Laurent/Gaspésie, Scale: 1/250 000; other regions, scale: 1/5 000 000 NB: Digital data were not available, the map was digitalised (map scale: 1/500 000). Area are subject to under estimation due to the digitalised and polygon conversion processes. NS: Geosciences Atlas, Nova Scotia Department of Natural Resources, Scale: 1/500 000. NL: Geosciences Atlas, Newfoundland and Labrador Geological Survey, Scale: 1/500 000 for the island and 1/1 000 000 for Labrador. PE: Dept. of agriculture and fisheries, Scale: 1/500 000.	

Table 4. List of the explanatory variables with descriptive statistics.

Notation	Description	Unit	Mean	Median	Min	Max
Physiographic variables						
BasinArea	Catchment area	km^2	1652	597	1.75	41167
Xcentroid	X-axis location of the catchment centroid	m	8119148	8024825	7728342	8977876
Ycentroid	Y-axis location of the catchment centroid	m	1779127	1639584	1428465	2361764
LakeArea	Total lake area	%	4.48	3.38	0.00	19.88
DrainageDensity	Drainage density of the hydrological network	m^{-2}	1.56	1.37	0.53	3.02
MinElevation	Minimum elevation of the catchment	m	102.58	73.50	-4.04	543.84
MaxElevation	Maximum elevation of the catchment	m	616.14	652.00	68.00	1125.00
MeanElevation	Mean elevation of the catchment	m	356.18	346.09	37.00	775.50
ElevationStation	Elevation at the station	m	105.02	76.50	1.00	544.83
Slope	River slope	%	0.0103	0.0047	0.0001	0.1635
Climatic variables						
TotPrecip	Total annual precipitation over the catchment area	mm	980.41	968.25	666.88	1294.26
MeanAirTmax	Annual mean of maximum air temperatures at the nearest grid point	°C	8.37	8.78	1.93	12.49
MaxAirTmax	Annual maximum of maximum air temperatures at the nearest grid point	°C	29.66	29.45	23.68	34.08
MinAirTmin	Annual minimum of minimum air temperatures at the nearest grid point	°C	-28.11	-29.13	-41.98	-14.70
Land cover						
Shrubland	Percentage of shrubland area	%	6.64	2.53	0.00	47.76
Grassland	Percentage of grassland area	%	1.96	1.33	0.00	14.42
Wetland	Percentage of wetland area	%	1.84	0.80	0.00	13.42
Forest	Percentage of forest area	%	80.52	88.23	5.85	99.69
Surface deposits						
Glacial Deposits	Percentage of area covered by glacial deposits	%	70.16	82.84	0.00	100.00
Rock	Percentage of area covered by rock	%	5.08	0.02	0.00	57.72
Fluvio-Glacial Deposits	Percentage of area covered by fluvio-glacial deposits	%	3.56	0.62	0.00	34.39

The initial delineation of regions performed using HCA was completed independently for each water temperature metric. Figure 3 provides an example of a dendrogram for one of the temperature metrics: the *a* parameter of the Gaussian function. When the truncation is done at a Euclidian distance of 17, two homogenous regions are identified. Figure 4 shows the result of the same approach for each water temperature metric. For Tmax, one region includes most of the stations located in Labrador, Quebec North shore and Gaspé Peninsula, while the remaining stations are all included in the second group. MaxnumDay (Maximum number of days where Tmax > 25 °C and Tmin > 20 °C) also shows two groups: one group is essentially limited to Newfoundland and Labrador, with a few stations on the Quebec side of the Labrador border. The second group includes all other stations except for a few in the Gaspé Peninsula. For the Gaussian function parameters, the Newfoundland stations constitute one homogenous group (with the addition of some stations in Nova Scotia and PEI for parameters *b* and *c*). The remainder of the stations are all in the same second group.

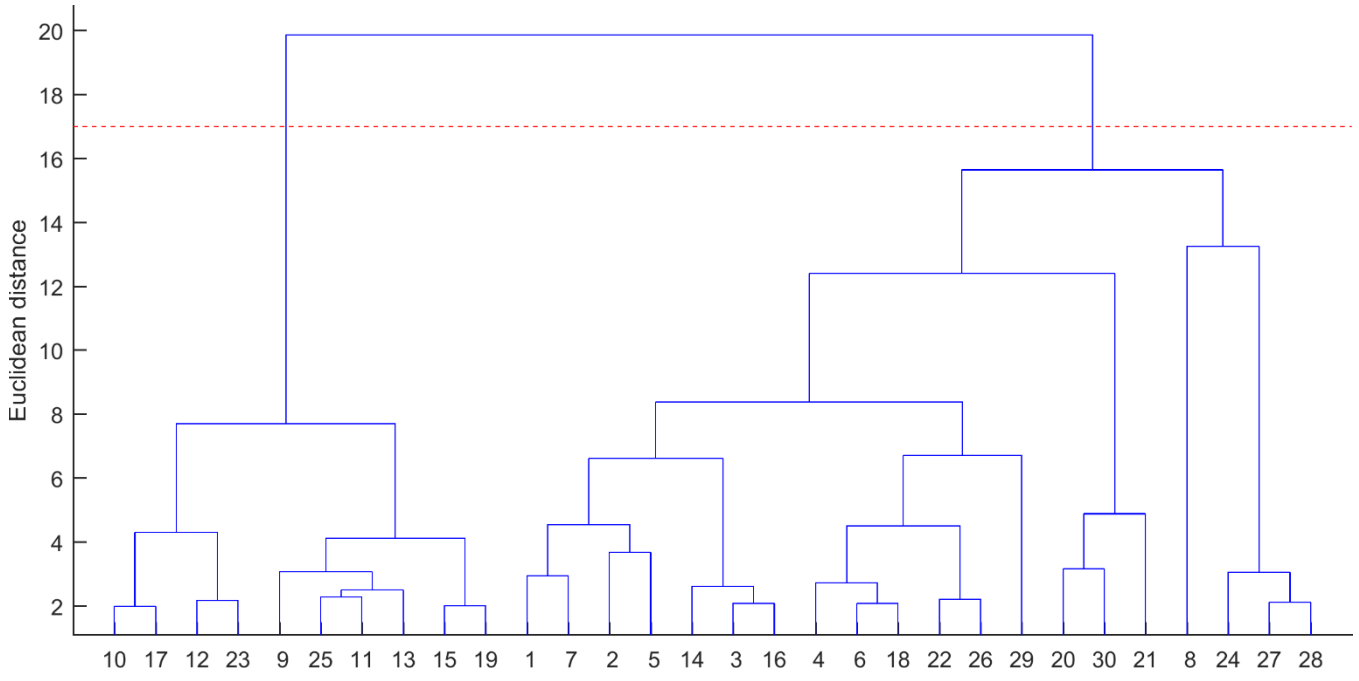


Figure 3. Example of a hierarchical classification tree for the parameter Gaussian_*a* where only the first 30 leaf nodes are presented.

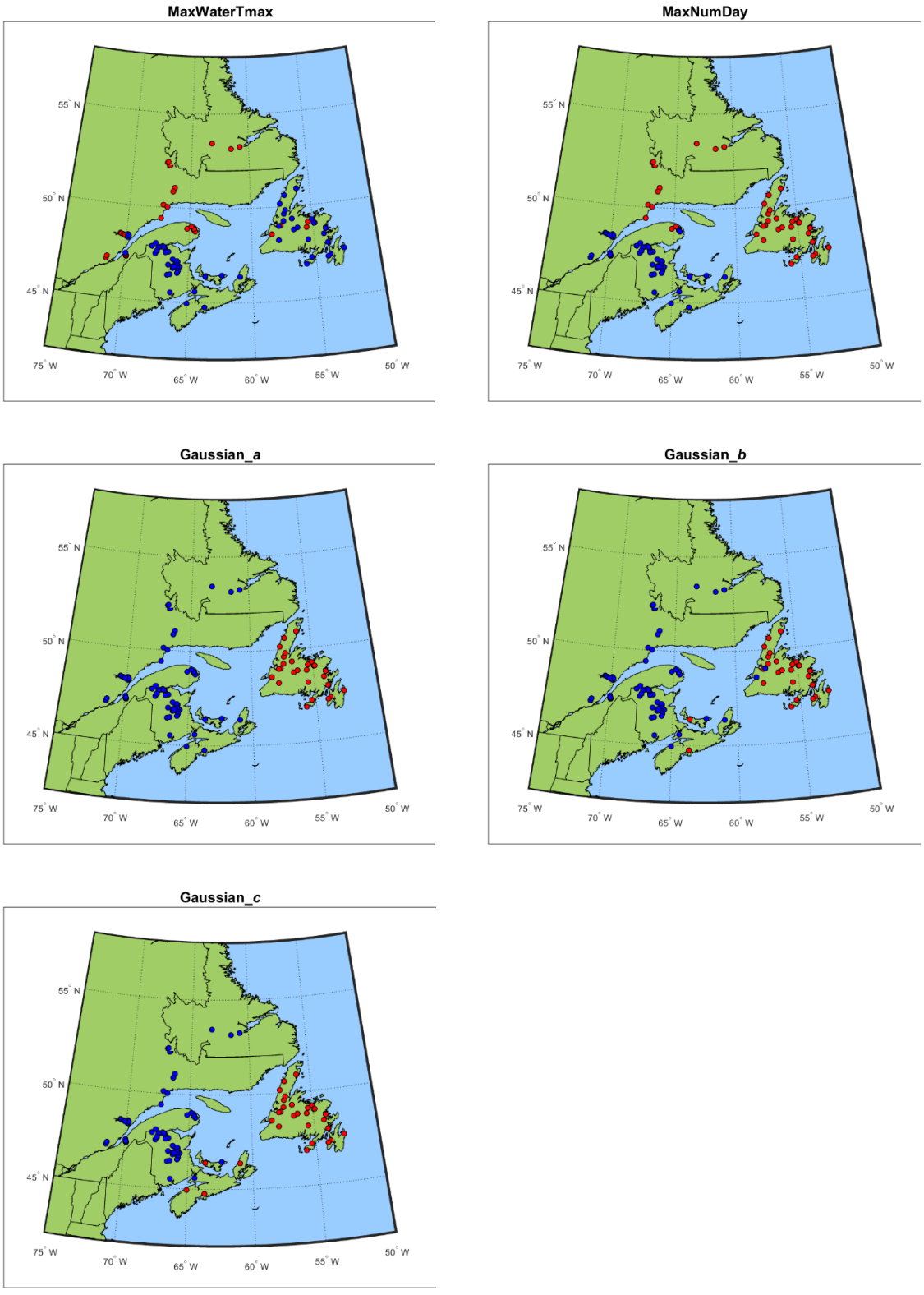


Figure 4. Spatial distribution of the regions with HCA.

The multiple linear regression (MLR) model and GAM were used for information transfer to the target site in this study. Using the forward stepwise regression procedure, the first six most important predictors were selected from the list in Table 4. The list of the independent variables selected for each metric and each model is given in Table 5. As expected, air temperature (max or min) is selected as one of the predictors for each metric. Elevation is also a common predictor for most metrics. Ycentroid is a predictor selected for four of the five metrics with GAM. The relationships between each selected predictor and temperature metrics are shown in figures 5 to 9. The fact that most of these relationships are non-linear already indicates that the GAM is a model that is better adapted to estimate the selected water temperature metrics than the MLR. This is confirmed in Table 6. The two models are compared 1) when all stations are pooled in a single region, 2) when the regions are defined using HCA and 3) when the ROI approach is used. The highest performance metrics are indicated in bold in the table. As expected from the observation of figures 5 to 9, the GAM systematically outperforms the MLR. In addition, this initial study indicates that there is merit in grouping stations in sub-regions. Indeed, creating two sub-regions using HCA provides a better performance than having a single regions or using the ROI approach. This result is rather novel as all studies carried out previously on hydrological variables (mainly flood and low flow quantiles) have shown neighborhood-based approaches (such as ROI) to be more flexible and to lead to better performances than fixed non-continuous regions (such as HCA) according to all performance criteria (see for instance Ouarda et al., 2008). Explanatory variables selected using the stepwise forward regression procedure.

Table 5. Explanatory variables selected using the stepwise forward regression procedure.

Metric	Explanatory variables					
MLR						
MaxWaterTmax	MeanAirTmax	Forest	Slope	Rock	Shrubland	FluvioGlacial Deposits
MaxNumDay	MaxAirTmax	Rock	MeanAirTmax	BasinArea	TotPrecip	MinElevation
Gaussian_*a*	MaxAirTmax	BasinArea	Rock	MeanAirTmax	Forest	LakeArea
Gaussian_*b*	MeanElevation	TotPrecip	Forest	BasinArea	MeanAirTmax	MaxAirTmax
Gaussian_*c*	Ycentroid	MinElevation	MaxElevation	Forest	MinAirTmin	LakeArea
GAM						
MaxWaterTmax	MeanAirTmax	BasinArea	Rock	TotPrecip	Forest	Grassland
MaxNumDay	MaxAirTmax	Ycentroid	BasinArea	ElevationStation	Rock	MaxElevation
Gaussian_*a*	Ycentroid	BasinArea	Rock	MinAirTmin	Shrubland	MinElevation
Gaussian_*b*	MeanElevation	MinAirTmin	Ycentroid	Forest	BasinArea	Wetland
Gaussian_*c*	Ycentroid	ElevationStation	Slope	Xcentroid	Forest	MaxElevation

This peculiar result may be linked to the over-representation of certain rivers in the database. Scatter plots of the estimated vs observed metrics are shown in Appendix A.

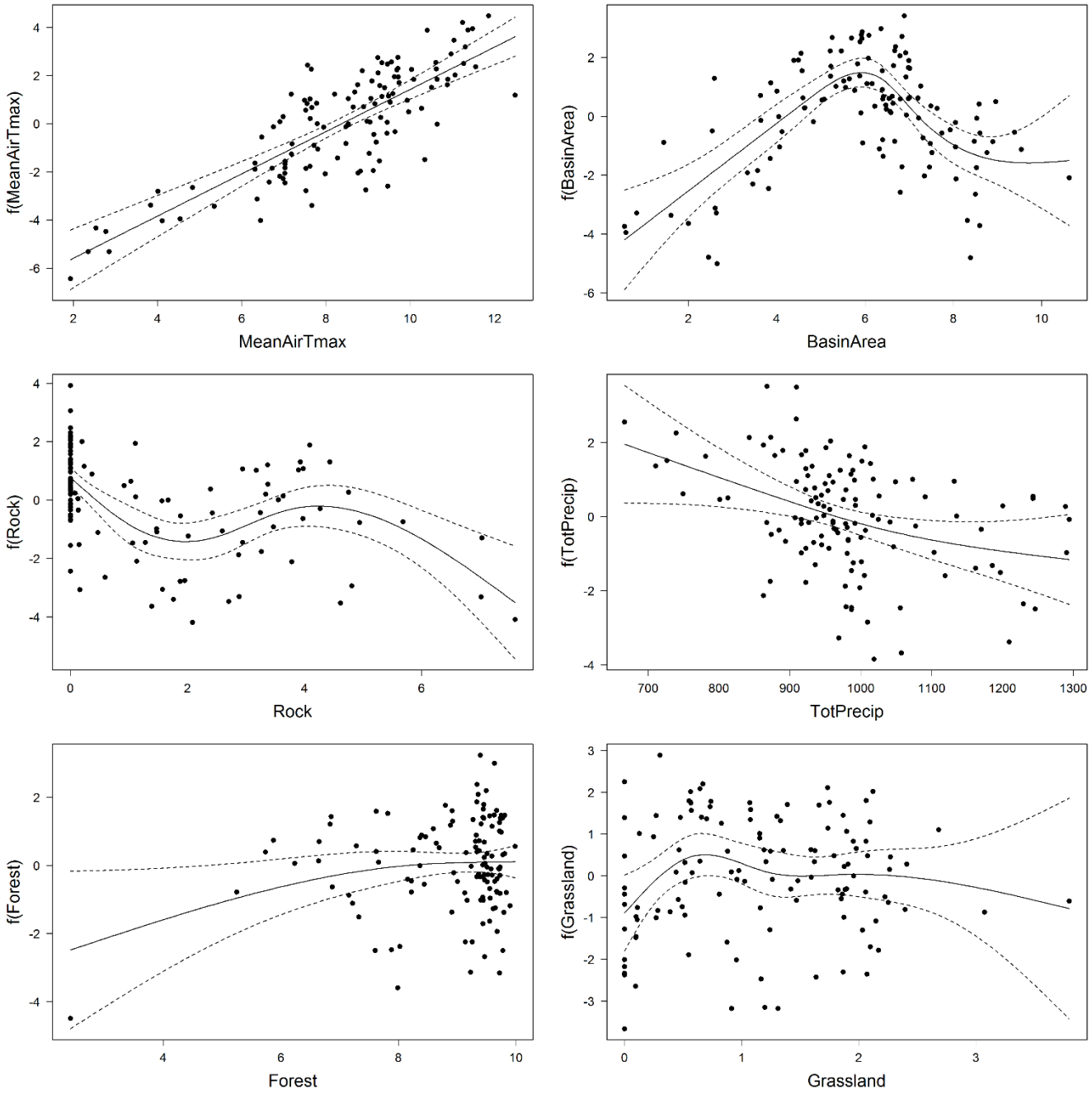


Figure 5. Smooth functions for MaxWaterTmax. The dashed lines represent the 95% confidence intervals and dots are the residuals.

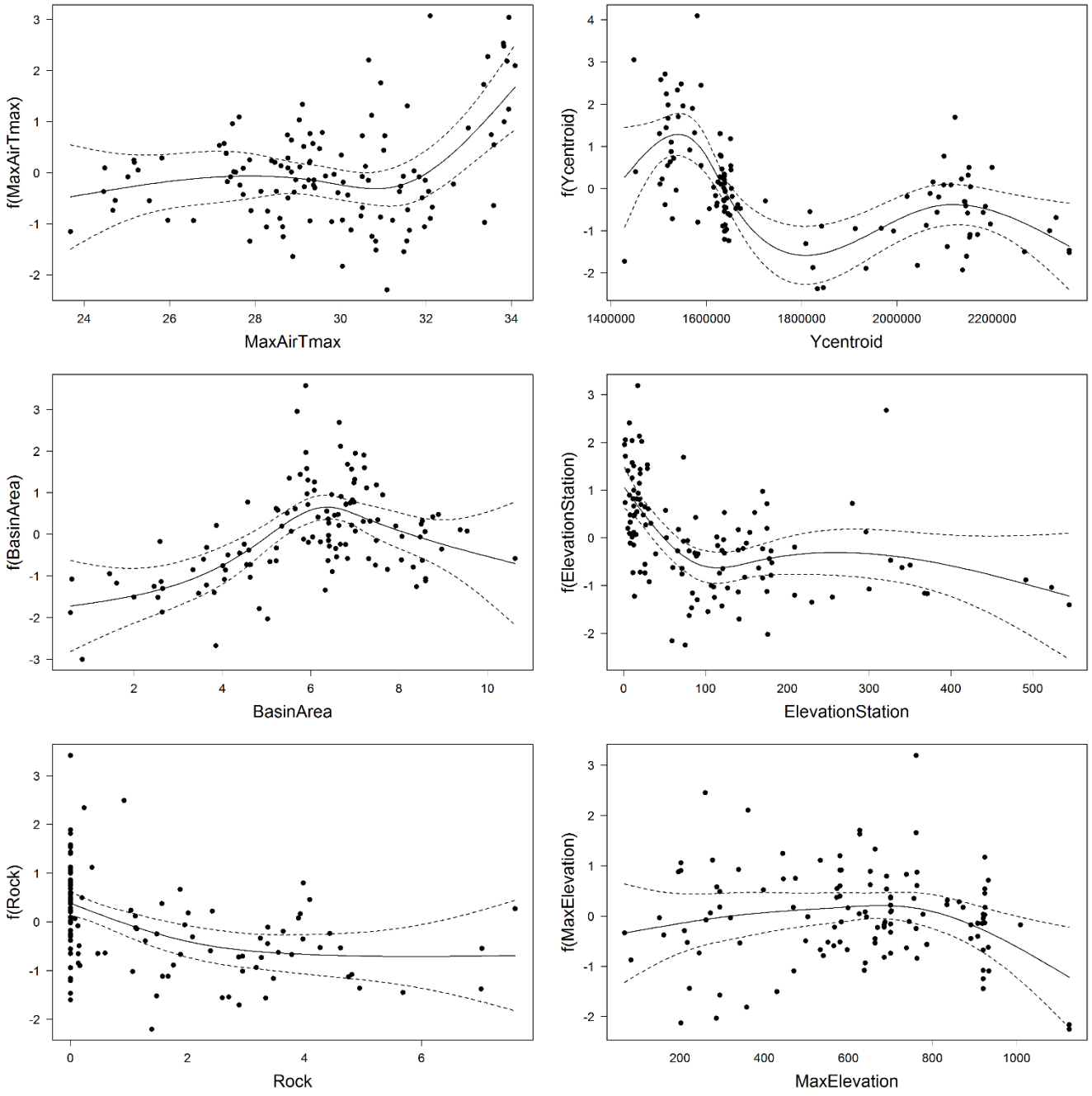


Figure 6. Smooth functions for MaxNumDay. The dashed lines represent the 95% confidence intervals and dots are the residuals.

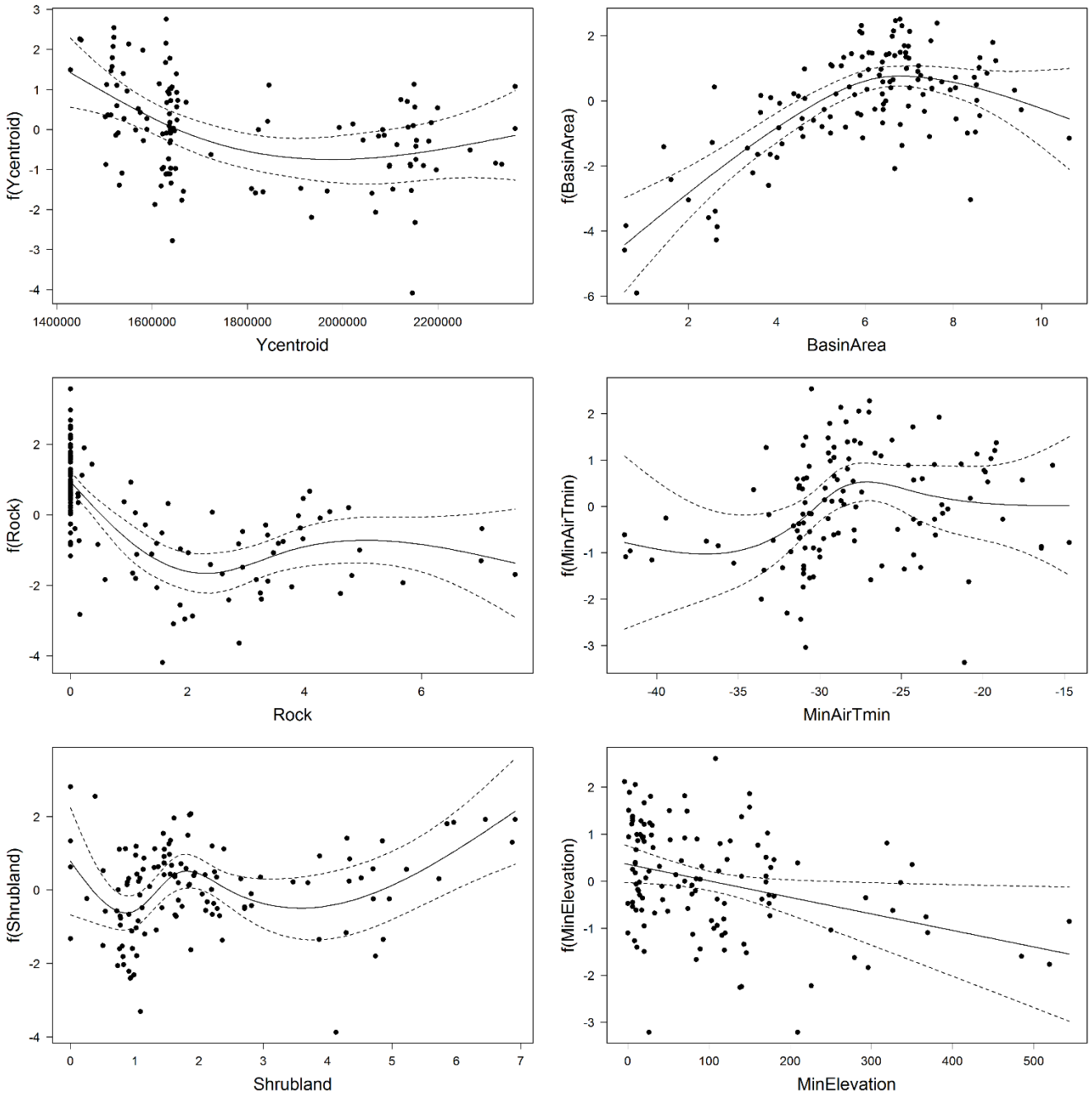


Figure 7. Smooth functions for the Gaussian_*a* parameter. The dashed lines represent the 95% confidence intervals and dots are the residuals.

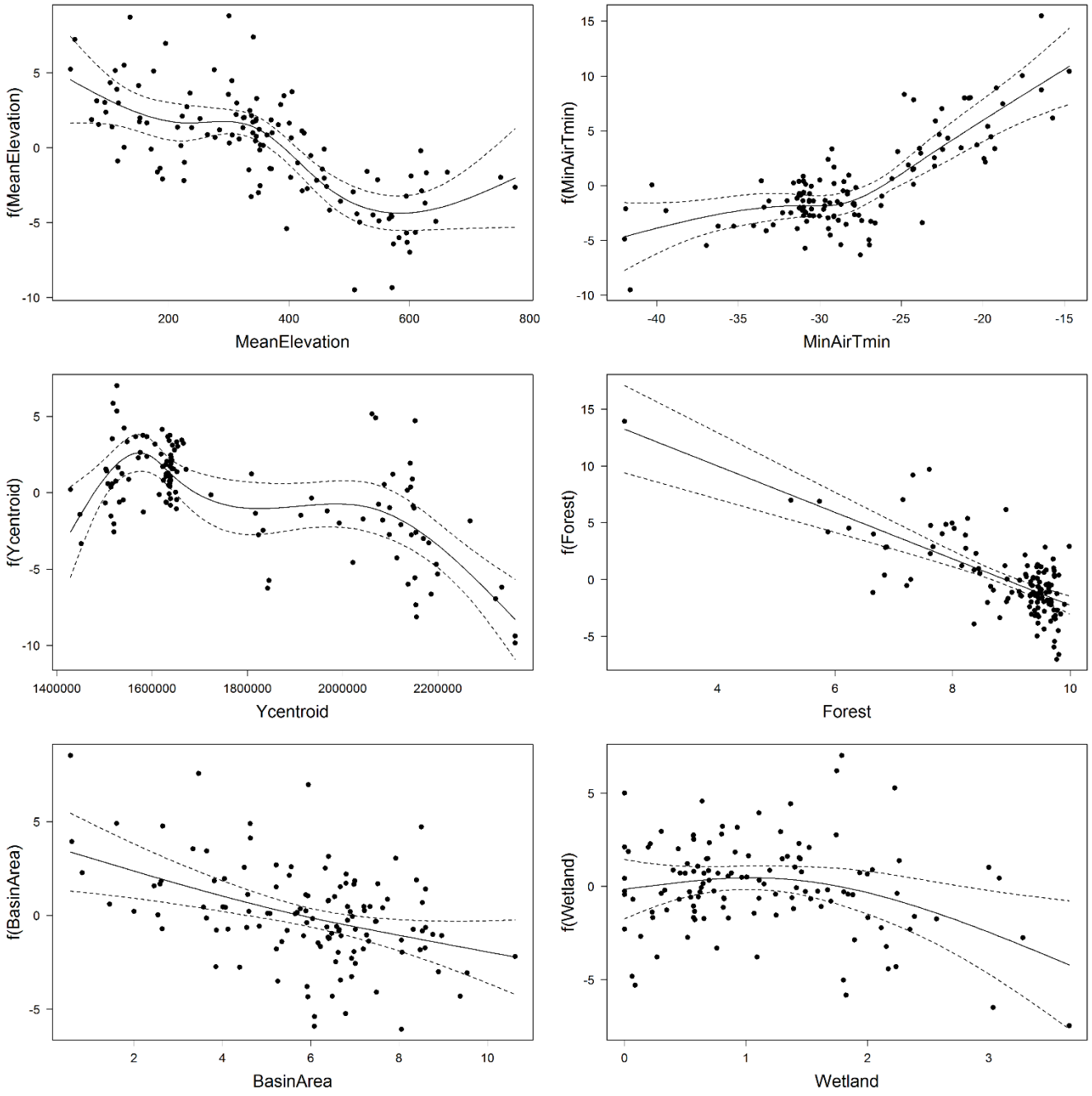


Figure 8. Smooth functions for the Gaussian_*b* parameter. The dashed lines represent the 95% confidence intervals and dots are the residuals.

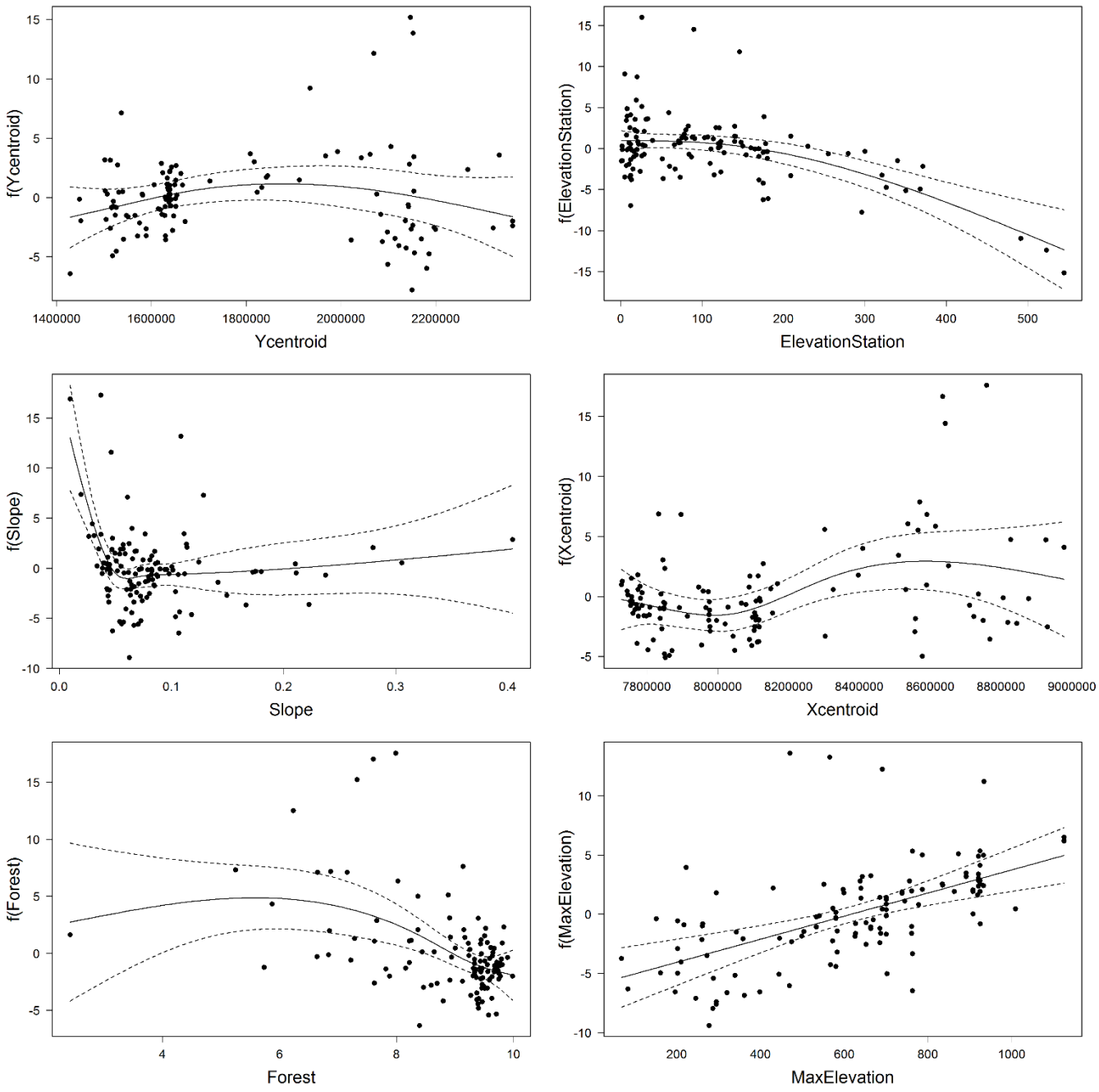


Figure 9. Smooth functions for the Gaussian_*c* parameter. The dashed lines represent the 95% confidence intervals and dots are the residuals.

Table 5. Performance statistics for the different approaches using the leave-one-out validation.

Metric	ALL			HCA			ROI		
	R^2	Bias	RMSE	R^2	Bias	RMSE	R^2	Bias	RMSE
MLR									
MaxWaterTmax	0.426	**-0.015**	2.012	0.448	0.146	1.973	**0.479**	0.185	**1.917**
MaxNumDay	0.499	**0.108**	1.229	**0.567**	0.130	**1.143**	0.520	0.361	1.203
Gaussian_*a*	0.313	**0.005**	1.637	0.399	0.052	1.530	**0.457**	0.100	**1.455**
Gaussian_*b*	0.806	**-0.014**	3.125	**0.853**	-0.043	**2.722**	0.800	-0.414	3.176
Gaussian_*c*	0.368	**0.043**	4.330	**0.397**	0.091	**4.229**	0.131	0.832	5.076
GAM									
MaxWaterTmax	0.634	0.037	1.606	**0.734**	**-0.022**	**1.371**	0.579	0.121	1.723
MaxNumDay	0.645	0.110	1.036	**0.800**	**0.088**	**0.777**	0.629	0.102	1.057
Gaussian_*a*	0.591	0.030	1.263	**0.656**	0.048	**1.158**	0.534	**0.025**	1.348
Gaussian_*b*	0.832	-0.049	2.907	**0.842**	**-0.018**	**2.819**	0.841	-0.048	2.828
Gaussian_*c*	0.422	**-0.026**	4.141	**0.442**	-0.058	**4.070**	0.348	-0.003	4.397

4 DISCUSSION

This first attempt at defining thermally homogenous regions in Eastern Canada was affected by the limits imposed by the minimum sample size prescribed. Indeed, selecting stations with at least five years of temperature data led to a relative over representation of Newfoundland rivers as well as three river systems outside of this province: the Ouelle, Ste-Marguerite and Miramichi rivers. Given the high density of Newfoundland stations, it is not surprising that the HCA segregated two thermal river groups: Newfoundland and elsewhere. To further divide the remaining stations in sub-regions, a lower truncation level (i.e. selecting a lower Euclidean distance to define groups) could be used in the HCA. For instance, a truncation level of 15 would yield three groups of stations. However, the drawback of increasing the number of homogenous groups is a decrease of the number of stations within each group, yielding a smaller sample size to establish the transfer models for water temperature metrics. An alternative that should be tested would be to relax the selection criterion for stations. The RivTemp database includes a

large number of stations that have between 1 and four years of data. Including water temperature stations with fewer than five years of data would increase station density outside of Newfoundland, thereby allowing to define a larger number of thermally homogenous regions, while potentially having a sufficiently large number of stations within each region to produce models with acceptable uncertainties. Another alternative that should be tested would be to select fewer stations on over-represented river systems.

However, it is important to ensure that this would not result in the calculation of the water temperature metrics with too few data, as uncertainty would likely increase. In fact, Daigle et al. (2019) have shown that when the parameters of the Gaussian function are calculated on fewer than five years of data, the error on the estimation of the parameters increases significantly.

Only two models were tested in this first study. The non-linear GAM was systematically superior to the simpler multiple linear regression for all five water temperature metrics. This result is consistent with previous hydrological studies (see for instance Ouarda et al., 2018). This result confirms also the non-linear nature of the water temperature dynamics and the need to increasingly adopt non-linear estimation methods when modeling the thermal regimes of rivers. Other statistical models could be used, including random forest regression models that were used by Maheu et al. (2015) in the U.S. to classify thermal regimes. The results of the present study cannot be generalised without further efforts. Additional studies need to be carried out for other climate characteristics, and for other conditions (including data quality and quantity conditions) in order to test the robustness of the proposed modeling approaches and the generality of the conclusions of the present work. Future work can also focus on the combination of local and regional information at gauged sites. This would allow improving the estimation of water temperature metrics in sites with measurements. Future work should also focus on applying multivariate regional estimation models to this problem in order to carry out the estimation of a number of water temperature metrics at the same time while taking into consideration the linkages between these metrics.

Other water temperature metrics can also be modelled using the same approach. One key metric under development is the Potential Growth Thermal Index, or PGTI (Ouellet-Proulx et al., in prep). This index is based on known temperature ranges for which growth is initiated, becomes optimal and beyond which it ceases. The analyses of Ouellet-Proulx et al. have shown that the PGTI has the potential to explain juvenile salmon size at age at regional to continental scales. Multivariate reginal frequency analysis approaches can also be used to model water temperature metrics along with other habitat metrics and model the relationships between them.

5 CONCLUSION

A first regional analysis of water temperature metrics was completed in Eastern Canada. Using water temperature monitoring stations with 5 or more years of data, two relatively homogenous thermal regions were defined for five different thermal metrics. The generalized additive model outperformed multiple linear regression as a tool to estimate the thermal metrics at ungauged sites within each region, as demonstrated by a leave-one-out validation procedure. Initial results are promising but further work is required to better define the thermally homogenous regions.

6 REFERENCES

Acreman, M.C., 1987. Regional flood frequency analysis in the U.K.: Recent research-new ideas, Report of the Institute of Hydrology, Wallingford, UK

Boyer. C., ·A. St-Hilaire, · N. Bergeron R.A. Curry, D. Caissie, ·C.-A. Gillis. 2016. Technical Report:RivTemp: A Water Temperature Network for Atlantic salmon rivers in Eastern Canada. Water News, Canadian Water Association Newsletter, spring edition.

Burn, D.H. 1990. Evalutation of regional flood frequency analysis with a Region of Influence Approach. Water Resources Research 26(10): 2257-2265

Corey E, Linnansaari T, Cunjak RA, Currie S. 2017. Physiological effects of environmentally relevant , multi-day thermal stress on wild juvenile Atlantic salmon (Salmo salar). *Conservation Physiology* **5** (January): 10.1093/conphys/cox014 DOI: 10.1093/conphys/cox014.

Daigle, A. D.I. Jeong, M.F. Lapointe. 2015. Climate change and resilience of tributary thermal refugia for salmonids in eastern Canadian rivers *Hydrological Sciences Journal* 60(6): 1044-1063.

Daigle, A., C. Boyer, A. St-Hilaire. 2019. A standardized characterization of river thermal regimes in Québec. Submitted to the *Journal of Hydrology*.

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** (3): 429–445 DOI: 10.1111/eff.12224

Elliott JM, Elliott JA. 2010. Temperature requirements of Atlantic salmon Salmo salar, brown trout Salmo trutta and Arctic charr Salvelinus alpinus: Predicting the effects of climate change. *Journal of Fish Biology* **77** (8): 1793–1817 DOI: 10.1111/j.1095-8649.2010.02762.

Hastie, T., Tibshirani, R., 1986. Generalized Additive Models. Statistical Science, 1(3): 297-310.

Hutchinson, M. F., D. W. Mckenney, K. Lawrence, J. H. Pedlar, R. F. Hopkinson, E. Milewska and P. Papadopol (2009). Development and Testing of Canada-Wide Interpolated Spatial Models of Daily Minimum-Maximum Temperature and Precipitation for 1961-2003. Journal of Applied Meteorology and Climatology 48(4): 725-741.

Jeong, D.I., A. Daigle, A. St-Hilaire. 2013. 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* 29:805-821. DOI: 10.1002/rra.2574.

Johnson, S. C., 1967. *Hierarchical clustering schemes.* Springer ed. s.l.:Psychometrika.

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** (7): 1078–1093 DOI: 10.1080/02626667.2016.1246799.

Lepage, M.-P. et G. Bourgeois. 2011. Le réseau québécois de stations météorologiques et l'information générée pour le secteur agricole. Rapport du CRAAC. Publication no PAGR0101 ISBN 978-2-7649-0234-9.

Maheu, A., l. poff, A. St-Hilaire.2015. A classification of stream water temperature regimes in the conterminous United States. Publié en ligne dans *River Research and Applications*. 32(16) : 896-906 DOI: 10.1002/rra2906

McCullagh P & Nelder JA (1989) *Generalized linear models.* CRC press, Nicieza AG, Metcalfe NB, Dec N. 1997. Growth Compensation in Juvenile Atlantic Salmon : Responses to Depressed Temperature and Food Availability. **78** (8): 2385–2400

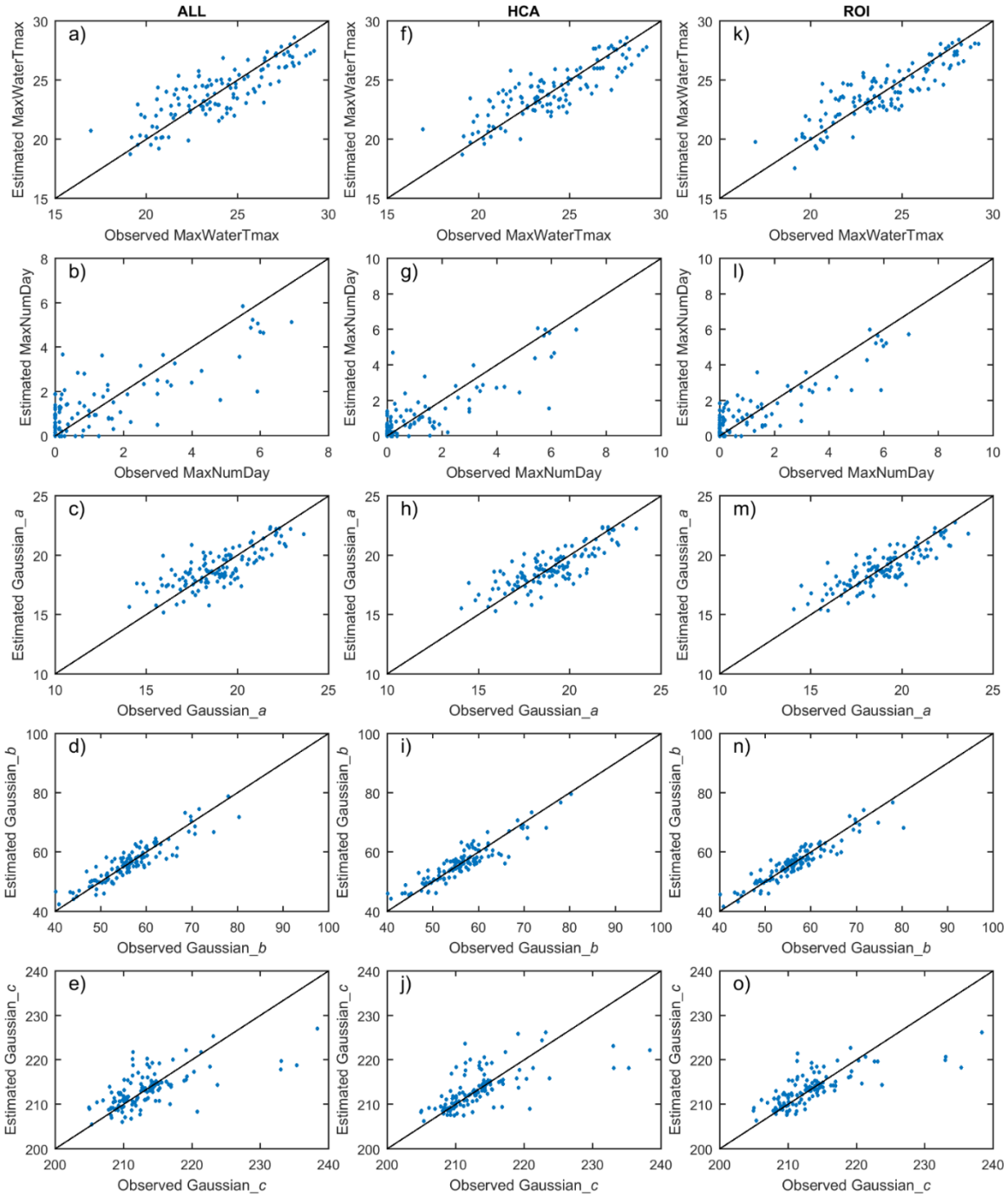
Ouarda, T.B.M.J., Ba, K.M., Diaz-Delgado, C., Carstenu, A., Chokmani, K., Gingras, H., Quentin, E., Trujillo. E., and B. Bobée, 2008. Regional flood frequency estimation at ungauged sites in the Balsas River Basin, Mexico. Journal of Hydrology. doi: 10.1016/j.hydrol.2007.09.031, 348: 40-58.

Ouarda, T.B.M.J., Charron, C., Hundecha, Y., St-Hilaire, A. and 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:10.1016/j.envsoft.2018.08.031

Ouellet-Proulx, S., A. Daigle, A.-St-Hilaire, M. Clément, C.-A. Gillis, T. Linaansaari, G. Dauphin, N. Bergeron. in prep. The effect of thermal regime on juvenile Atlantic salmon growth.

Sundt-Hansen LE, Hedger RD, Ugedal O, Diserud OH, Finstad AG, Sauterleute JF, Tøfte L, Alfredsen K, Forseth T. 2018. Modelling climate change effects on Atlantic salmon: Implications for mitigation in regulated rivers. *Science of the Total Environment* **631**–**632**: 1005–1117 DOI: 10.1016/j.scitotenv.2018.03.058

7 APPENDIX A: SCATTER PLOTS OF OBSERVED VS ESTIMATED WATER TEMPERATURE METRICS



The hierarchical cluster analysis method can be divided in the following three distinct steps:

1. Quantification of the similarity between each pair of basins: This step consists in computing a given distance statistic (e.g. the Euclidean distance) between every pairs of basins in the space defined by a set of selected physiographic and/or meteorological variables. These variables are selected based on the knowledge of their impacts on the variable of interest.

The distance used in this study is the standardized Euclidean distance defined by:

$$d^2(r,s) = (x_r - x_s)D^{-1}(x_r - x_s)^T \quad (1)$$

where x_r and x_s are the vectors of coordinates in the physiographical/meteorological space for basin *r* and *s* respectively and *D* is the diagonal matrix for which the diagonal elements v_j^2 are the variances of the respective variables. This statistic is similar to the Euclidean distance but where each variable is scaled by its variance.

2. Grouping of stations into a hierarchical cluster tree: This step consists in grouping pairs of basins that are close based on the measure given by a linkage function. This function uses the distance information generated in step 1 to determine the proximity of basins. As the basins are paired into binary clusters, the newly formed clusters are grouped into larger clusters. This is repeated until only one cluster is reached. The result can be displayed with a cluster tree diagram. In this study, the distance between clusters are obtained using the Ward's method. This method computes the following sum of squared distances:

$$WSS_p = \sum_{i=1}^{n_p} d^2(x_{pi}, \bar{x}_p) \quad (2)$$

where n_p is the size of the cluster *p* and $\bar{x}_p$ is the centroid of the cluster *p*. The distance between cluster *p* and *q* is given by:

$$d_W(p,q) = WSS_{p+q} - (WSS_p + WSS_q) = \frac{n_p n_q d^2(\bar{x}_p, \bar{x}_q)}{n_p + n_q} \quad (3)$$

3. Identification of clusters: At this step, clusters are identified using the hierarchical tree obtained in the previous steps. This can be achieved either by detecting natural groupings in the hierarchical tree or by cutting off the tree at an arbitrary level which may be determined by the targeted number of clusters.

9 APPENDIX C: REGION OF INFLUENCE METHOD

Acreman (1987) proposed a method to identify homogeneous neighbourhoods at target sites. This method select stations within a certain critical distance from each target site. It was later adopted by Burn (1990) for the regionalization of flood flows and was termed the "region of influence" method (ROI). In this method, the identification of a neighborhood is based on a Euclidean distance in a multidimensional space defined by a set of hydrological attributes of a site and/or a set of physiographical and meteorological attributes of the contributing basin. In the case of ungauged sites, only physiographical and meteorological catchment attributes are used.

The Euclidean distance D_{ij} between stations *i* and *j* is given by the following Euclidean distance:

$$D_{ij} = \left[\sum_{k=1}^{K}\left(C_k^i - C_k^j\right)^2\right]^{1/2} \quad (4)$$

where C_k^i and C_k^j are the standardized values of attribute *k* for stations *i* and *j* respectively, and *K* is the number of attributes. The selection of the attributes used to define the Euclidean space are selected based on the knowledge of their impacts on the variable of interest. The stations to be included into the ROI for a given target site are all those within a given threshold distance δ_i:

$$ROI_i = \{k: D_{ik} \leq \delta_i\}. \quad (5)$$

δ_i is fixed in such a way that there is a good compromise between the number of stations in the neighbourhood and the hydrological homogeneity of the selected stations. Burn (1990) presents different options for the selection of a value of δ_i for the aim of regionalisation.

Generalized linear models (GLM) are a generalization of linear models in which the response variable can follow any distribution of the exponential family and where a link function relates the response variable to the linear predictor function. Generalized additive models (GAM) were introduced in Hastie and Tibshirani (1986) as extensions of generalized linear models (GLM) in which the linear predictor function is replaced by a set of smoothed functions of the explanatory variables. GAMs are thus more flexible than linear models by allowing a non-linear relationship between the response variable and each of the explanatory variables. For a response variable *Y*, GAMs can be expressed by:

$$g\big(E(Y|\mathbf{X})\big) = \alpha + \sum_{j=1}^{p} f_j\,(X_j) \quad , \tag{10}$$

where f_j is a smooth function for the *j*th explanatory variable, **X** is a matrix whose columns are the set of *p* explanatory variables, α is an intercept and $g(.)$ is a monotonic link function.

The smooth function f_j can be defined by a linear combination of *q* basis functions:

$$f_j(x) = \sum_{i=1}^{q} \beta_{ji} b_{ji}(x) \tag{11}$$

where β_{ji} are smoothing coefficients and $b_{ji}(x)$ is a basis function. Spline is a convenient basis for smooth functions. A spline is a curve composed of piecewise polynomial functions joined together at points called knots.

A problem that arises with spline basis function is overfitting. Penalized regression spline avoids this problem by introducing a penalty parameter. Also, with spline basis, the location of the knots needs then to be chosen. However, with penalized regression splines, the exact location as well as the number of the knots are not as important. GAMs with penalized regression splines are usually optimized by maximizing the penalized log-likelihood:

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \tfrac{1}{2}\sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}' \mathbf{S}_j \boldsymbol{\beta}, \tag{12}$$

where $\boldsymbol{\beta}$ is a matrix of smoothing coefficients, $\boldsymbol{\beta}'$ is the transpose of $\boldsymbol{\beta}$, $l(\boldsymbol{\beta})$ is the log-likelihood function, λ_j is the smoothing parameter of the *j*-th smooth function, and $\mathbf{S}_j$ is a matrix of known coefficients (Wood, 2008). The parameter λ_j, ranging from 0 to 1, controls the degree of smoothness of the smooth function where 0 is the un-penalized case and 1 is the completely smoothed case. The optimum value of λ_j is a good compromise between optimization and smoothness. $\boldsymbol{\lambda}$ is found iteratively according to a criterion such as the generalized cross validation (GCV; Wahba, 1985), unbiased risk estimator (UBRE; Craven and Wahba, 1978) or maximum likelihood (ML). At each step, the function $l_p(.)$ is solve for a given vector of

smoothing parameters $\boldsymbol{\lambda}$, by the penalized iteratively reweighted least squares method (P-IRLS; Wood, 2004).