Centre Eau Terre et Environnement

MODÉLISATION DES SCÉNARIOS FUTURS DE TEMPÉRATURE DE L'EAU EN MILIEU CÔTIER ET IMPLICATIONS SUR LES INFECTIONS POTENTIELLES PAR *VIBRIO PARAHAEMOLYTICUS* ET *VIBRIO VULNIFICUS* : APPLICATION AUX BANCS COQUILLERS DE L'ESTUAIRE ET DU GOLFE DU SAINT-LAURENT

Par

Habiba Ferchichi

Mémoire présenté pour l'obtention du grade de

Maître ès Sciences (M.Sc.)

en sciences de la Terre

Jury d'évaluation

Examinateur externe	Nicholas Ogden Agence de la santé publique du Canada-ASPC
Examinateur interne	Saeid Homayouni INRS-Eau-Terre-Environnement
Directeur de recherche	André St-Hilaire INRS-Eau-Terre-Environnement Canadian River Institute
Codirecteurs de recherche	Taha B.M.J Ouarda INRS-Eau-Terre-Environnement Benoit Lévesque Institut National de la Santé Publique du Québec-INSPQ

REMERCIEMENTS

Je tiens à exprimer ma gratitude et mes respects les plus sincères à mon directeur de recherche André St-Hilaire de m'avoir accueillie dans son équipe. Merci pour ta disponibilité, ta capacité à l'écoute de ma foule de questions et tes explications rigoureuses. Merci pour ta patience et ta confiance en moi malgré tous les défis que j'ai eu à relever au début.

Je tiens à remercier mon codirecteur Taha B.M.J. Ouarda pour sa rigueur professionnelle, ses critiques constructives et ses encouragements qui m'ont poussée à me dépasser moi-même et à donner une meilleure performance. Merci pour les nombreuses discussions qu'on a eues qui ont contribué à me guider dans mes recherches.

Je tiens à remercier mon codirecteur externe Benoit Lévesque de m'avoir bien m'informer sur un domaine nouveau pour moi (microbiologie). Merci pour tes suggestions pertinentes et l'attention que tu as bien voulu apporter à mes travaux de recherche, ta grande disponibilité et ta patience ainsi que tes réponses et explications détaillées à mes diverses questions.

J'ai été très chanceuse de travailler avec vous tous !!!

Mes remerciements s'adressent également au reste des membres du comité du projet : Estelle Pedneault et Céline Campagna pour leurs réponses à mes questions, leurs lectures et corrections de mon rapport scientifique remis à Ouranos.

Je remercie également M. Yves Gratton, professeur honoraire à l'INRS, d'avoir partagé ses connaissances diverses qui ont alimenté ma réflexion, pour sa disponibilité et ses réponses à mes questions nombreuses.

Mes remerciements vont aussi à Ouranos pour leur soutien financier attribué à ce projet.

J'adresse mes vifs remerciements aux examinateurs, M. Nicholas Ogden et M. Saeid Homayouni, d'avoir gentiment accepté d'évaluer mon mémoire de recherche.

Je tiens à exprimer ma profonde reconnaissance à ma famille pour son encouragement. Merci à Amina, qui devient une véritable sœur, pour m'écouter et me soutenir toujours. Merci également à mes amies, Jocelyne et Jenny, d'être toujours là pour me supporter et me redonner de la motivation.

Finalement, je tiens à remercier mon mari, mon âme sœur, qui n'a pas cessé de m'encourager à faire de mon mieux. Merci d'être toujours là pour moi, malgré la distance qui nous sépare et nous fait souffrir. Merci de me soutenir pendant des périodes de doute et d'avoir cru en moi.

RÉSUMÉ

Les bactéries du type *Vibrio (V)* sont des bactéries présentes dans le milieu maritime et plus particulièrement dans les eaux estuariennes et côtières. Les *Vibrio parahaemolyticus* et *Vibrio vulnificus* sont considérées comme des bactéries pathogènes responsables des infections transmises à l'être humain principalement par la consommation de fruits de mer contaminés. Les *Vibrio parahaemolyticus* sont à l'origine de gastroentérites alors que les *V. vulnificus* provoquent des infections graves (septicémies, infection hépatique…) qui peuvent être mortelles.

La croissance des *Vibrio* pathogènes est liée à la température de l'eau et semble augmenter au-dessus d'un seuil de 15 °C. La propagation des infections par ce genre de bactérie s'accroit progressivement dans le monde entier dû à l'augmentation des températures de l'eau associées au changement climatique. La côte canadienne n'est pas une exception à cette expansion mondiale. Étant donné que la récolte des mollusques est en pleine expansion au Québec et à l'Ile-Prince-Édouard, plus précisément dans l'estuaire et le golfe du Saint-Laurent, la modélisation des futurs scénarios de température de l'eau est primordiale dans l'analyse du risque d'infection, afin de protéger la santé humaine et la qualité des produits maritimes.

Pour atteindre cet objectif, nous avons d'abord expliqué la variation de la température de surface en milieu estuarien et côtier en fonction des variables météorologiques (température de l'air, vitesse du vent), océanographique (marnage), hydrologique (débit du fleuve Saint-Laurent) et de téléconnexion (indices climatiques). Ensuite, un ensemble de modèles d'apprentissage automatique (Réseaux de neurones artificiels et Forêts aléatoires) et un modèle paramétrique (Régression linéaire multiple) ont été testés pour modéliser la température de l'eau. Il est intéressant de noter que l'essai de ces modèles sur un ensemble de séries de données relativement longues, provenant de thermographes et de bouées de l'estuaire et du golfe du Saint-Laurent, a montré que les modèles d'apprentissage automatiques performent mieux que le modèle linéaire dans la prédiction de la température de l'eau. Ensuite, en utilisant les prédicteurs les plus pertinents de la variation de la température de l'eau et le modèle le plus performant, des scénarios futurs de température de l'eau ont été générés pour l'horizon 2040-2100.

Enfin, sur la base de ces scénarios et le seuil théorique des conditions favorables à la croissance de ces bactéries, nous avons modélisé l'indicateur du risque de croissance des

Vibrio pathogènes, i.e. le nombre de jours associés au dépassement de la température minimale de la croissance des *Vibrio* pathogènes (15 °C), pour l'horizon 2040-2100.

Par conséquent, les simulations montrent que la croissance des *Vibrio* pathogènes pourrait augmenter et s'étendre de manière spatiale et même saisonnière pour l'horizon étudié 2040-2100, indépendamment du scénario climatique envisagé (optimiste ou pessimiste).

Mots-Clés : Température de l'eau, bactérie-*Vibrio*, modèles d'apprentissage automatiques, modélisation, prédiction, changement climatique.

ABSTRACT

Vibrio (V) spp. are common bacterial inhabitants of coastal environments, especially the estuarine and coastal ecosystem. *Vibrio parahaemolyticus* and *Vibrio vulnificus* are the leading cause of foodborne infections transmitted to humans mainly through shellfish consumption. The foodborne diseases with these pathogens are associated with gastroenteritis, but can also lead to fatalities in case of *V. vulnificus* infection.

The growth of pathogenic Vibrio is related to ambient water temperature and seems to increase at 15 °C and over. The spread of *Vibrio* infection is increasing worldwide due to the increase of the sea surface temperature associated with climate change. The Canadian Coast is not an exception to *Vibrio* expansion. The shellfish harvesting plays a key role in the economy of the coastal regions in the Estuary and Gulf of St. Lawrence, thus modelling future water temperature will contribute in predicting the future *Vibrio* infection risk and protecting shellfish industry as well as human health.

In order to achieve this goal, we first examined the relationship of the sea surface temperature as a function of meteorological (air temperature, wind velocity), oceanographic (tides), hydrologic (flow) and teleconnection (climatic indices) variables. Then, a set of machine learning models (Artificial Neural Network and Random Forest) and a parametric model (Multiple Linear Regression) were tested to model water temperature. Interestingly, the model implementation of these models on a relatively large dataset gathered from thermographs and buoys in the Estuary and Gulf of St. Lawrence showed the excellent performance of machine learning models in predicting the sea surface temperature. Then, by using the most relevant predictors of the sea surface temperature variation with the best performing machine leaning model, we generated the future water temperatures in the horizon 2040–2100. Finally, based on the predicted future water temperature scenarios and a threshold of 15 °C to determine the conditions favorable to the growth of *Vibrio* bacteria, we modelled the *Vibrio* growth risk indicator, i.e. the number of days exceeding the minimum temperature for Vibrio pathogenic growth (15 °C), in the horizon 2040–2100.

Consequently, simulations show that the *Vibrio* pathogenic growth risk will increase and extend spatially and even seasonally and all the shellfish beds would meet the temperature condition for *Vibrio* growth regardless of the climate scenarios in the horizon 2040–2100.

Keywords: Water temperature, *Vibrio* bacteria, machine learning models, modelling, prediction, climate change.

TABLE DES MATIÈRES

REMERCIEMENTS III

RÉSUMÉ V

ABSTRACT VII

TABLE DES MATIÈRES IX

LISTE DES FIGURES XI

LISTE DES TABLEAUX XIII

LISTE DES ABRÉVIATIONS XV

PREMIÈRE PARTIE SYNTHÈSE 1

1 MISE EN CONTEXTE ET SYNTHÈSE DU MÉMOIRE 1

1.1 MISE EN CONTEXTE 1

1.2 REVUE DE LITTÉRATURE 2

1.2.1 Les facteurs influençant la croissance des souches pathogènes des Vibrio 2

1.2.2 Modélisation de la température de l'eau dans l'estuaire et le golfe du Saint-Laurent 8

1.3 PROJET DE RECHERCHE 9

1.4 QUESTIONS DE RECHERCHE, HYPOTHÈSES ET OBJECTIFS 10

1.5 SYNTHÈSE DES TRAVAUX DE RECHERCHE 12

1.5.1 La modélisation de la température de l'eau par des modèles statistiques (Article1) 12

1.5.2 L'impact des scénarios futurs des températures de l'eau sur la croissance des Vibrio pathogènes (Article2) 18

1.6 CONCLUSION 22

2 RÉFÉRENCES 23

DEUXIÈME PARTIE ARTICLES 27

3 ARTICLE 1: MODELLING COASTAL WATER TEMPERATURE AND IMPLICATIONS FOR POTENTIAL INFECTIONS WITH VIBRIO MARINE BACTERIA 29

3.1 ABSTRACT 31

3.2 ABBREVIATIONS 32

3.3 INTRODUCTION 33

3.4 *MATERIAL AND METHODS* 36

3.4.1 Data description 36

3.4.2 Modelling water temperature 39

3.5 RESULTS 47

3.5.1 Selection of the best predictors of the daily mean sea surface temperature 47

3.5.2 Performance Evaluation for Model Selection 51

3.6 DISCUSSION AND CONCLUSIONS 56
3.7 ACKNOWLEDGEMENTS 57

4 REFERENCES 58

5 APPENDIX I 69

6 APPENDIX II 74

7 ARTICLE 2: IMPACT OF THE FUTURE COASTAL WATER TEMPERATURE SCENARIOS ON THE RISK OF POTENTIAL GROWTH OF PATHOGENIC VIBRIO MARINE BACTERIA 75

7.1 ABSTRACT 77
7.2 ABBREVIATIONS 78
7.3 INTRODUCTION 79
7.4 STUDY AREA 82
7.5 MATERIAL AND METHODS 84
7.5.1 Data collection 84
7.5.2 Modelling water temperature 85
7.5.3 Trend Analysis 89
7.5.4 Mapping future potential risk growth Vibrio 92
7.6 RESULTS 93
7.7 DISCUSSION 104
7.8 ACKNOWLEDGEMENTS 107

8 REFERENCES 108

LISTE DES FIGURES

FIGURE 1.1 LOCALISATION DES STATIONS DE MESURE DE LA TEMPÉRATURE DE L'EAU DANS L'ESTUAIRE MARITIME ET GOLFE DU SAINT-LAURENT 13

FIGURE 3.1 GEOGRAPHIC LOCATIONS OF THE COSTAL THERMOGRAPHS AND BUOYS IN THE ESTUARY AND GULF OF ST. LAWRENCE 37

FIGURE 3.2 THE PAIRWISE CORRELATION MATRIX OF GRANDE-RIVIÈRE VARIABLES 50

FIGURE 3.3 PERFORMANCE OF ARTIFICIAL NEURAL NETWORK (ANN) AND RANDOM FOREST (RF) MODELS FOR GRANDE-RIVIÈRE STATION OBTAINED FROM TESTED DATASET. 53

FIGURE 7.1 GEOGRAPHIC LOCATION OF THE THERMOGRAPHS AND BUOYS IN THE ESTUARY AND GULF OF ST. LAWRENCE 83

FIGURE 7.2 INVERSE DISTANCE WEIGHTING (IDW) INTERPOLATION OF RISK INDICATOR VALUES (NUMBER OF DAYS EXCEEDING THE THRESHOLD [15 °C]) OVER THE ESTUARY AND GULF OF ST. LAWRENCE UNDER PESSIMISTIC AND OPTIMISTIC CLIMATIC SCENARIO FOR THE HORIZONS 2040–2060 AND 2080–2100 IN AUGUST. (A) IDW INTERPOLATION OF THE RISK INDICATOR VALUES IN AUGUST DURING THE HORIZON 2040–2060 UNDER OPTIMISTIC SCENARIO (RCP 4.5). (B) IDW INTERPOLATION OF THE RISK INDICATOR VALUES IN AUGUST DURING THE HORIZON 2040–2060 UNDER PESSIMISTIC SCENARIO (RCP 8.5). (C) IDW INTERPOLATION OF THE RISK INDICATOR VALUES IN AUGUST DURING THE HORIZON 2080–2100 UNDER OPTIMISTIC SCENARIO (RCP 4.5). (D) IDW INTERPOLATION OF THE RISK INDICATOR VALUES IN AUGUST DURING THE HORIZON 2080–2100 UNDER PESSIMISTIC SCENARIO (RCP 8.5) 100

FIGURE 7.3 INVERSE DISTANCE WEIGHTING (IDW) INTERPOLATION OF RISK INDICATOR VALUES (NUMBER OF DAYS EXCEEDING THE THRESHOLD [15 °C]) OVER THE ESTUARY AND GULF OF ST. LAWRENCE UNDER PESSIMISTIC AND OPTIMISTIC CLIMATIC SCENARIO FOR THE HORIZONS 2040–2060 AND 2080–2100 IN SEPTEMBER. (A) IDW INTERPOLATION OF THE RISK INDICATOR VALUES IN SEPTEMBER DURING THE HORIZON 2040–2060 UNDER OPTIMISTIC SCENARIO (RCP 4.5). (B) IDW INTERPOLATION OF THE RISK INDICATOR VALUES IN SEPTEMBER DURING THE HORIZON 2040–2060 UNDER PESSIMISTIC SCENARIO (RCP 8.5). (C) IDW INTERPOLATION OF THE RISK INDICATOR VALUES IN SEPTEMBER DURING THE HORIZON 2080–2100 UNDER OPTIMISTIC SCENARIO (RCP 4.5). (D) IDW INTERPOLATION OF THE RISK INDICATOR VALUES IN SEPTEMBER DURING THE HORIZON 2080–2100 UNDER PESSIMISTIC SCENARIO (RCP 8.5) 101

LISTE DES TABLEAUX

Tableau 1.1 Tableau récapitulatif des seuils thermiques et de salinité en fonction de la croissance des *Vibrio* pathogènes 4

Table 3.1 The results of the linear correlations between the daily SST and the selected predictors for each station 48

Table 3.2 *The results of the linear correlations between the daily SST and the 60-day MA climate indices at each station* 49

Table 3.3 The results of the linear partial correlations between the daily SST and the selected predictors at the Grande-Rivière station 51

Table 3.4 Test performance results for Grande-Rivière data after selecting best features subset using RFE 52

Table 3.5 The RMSE performance results of all the stations computed in the test datasets 55

Table 3.6 The performance results of the RF model in modelling the daily SST using the best subset features at each station 56

Table 7.1 List of the Regional Climate Models (RCMs) used in simulations 85

Table 7.2 Performance criteria results of tested models (RF and ANN) 94

Table 7.3 Theil-Sen's slope for projected daily mean water temperature times series during the horizon 2040–2100 97

Table 7.4 The number of days exceeding the minimum temperature threshold (15 °C) for the growth of pathogenic Vibrio during October in the horizon 2080–2100 under pessimistic scenario (RCP 8.5) 103

Table 7.5 The number of days exceeding the threshold (20 °C) in the horizon 2080–2100 under pessimistic scenario (RCP 8.5) 104

LISTE DES ABRÉVIATIONS

ANN	Artificial Neural Network
CANOPA	CANadian Océan PArallélisé
CART	Classification And Regression Tree
CDC	Centers for Disease Control and Prevention
ENSO	El Niño—Southern Oscillation
IDW	Inverse Distance Weighting
IPE	Île-du-Prince-Édouard
MLR	Multiple Linear Regression
MM-X jours	Moyenne Mobile (glissante) sur X jours
NAO	North Atlantic Oscillation
Nash	Le critère de Nash-Sutcliff
ppt	parts per thousand, ou parties par millier
R^2	Le coefficient de détermination
rBiais	Le biais moyen relatif
RCP	Representative Concentration Pathways
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error (l'erreur quadratique moyenne)
rRMSE	Relative Root Mean Square Error (l'erreur relative quadratique moyenne)

ssp	Species
SST	La température moyenne journalière de la surface d'eau
Tair	La température moyenne journalière de l'air
V.	Vibrio
VNC	Viable Non Cultivable

Première partie

SYNTHÈSE

1 MISE EN CONTEXTE ET SYNTHÈSE DU MÉMOIRE

1.1 Mise en contexte

Les changements climatiques risquent de menacer la santé humaine puisqu'ils vont possiblement influencer l'incidence et la prévalence de certaines maladies. Par exemple, le risque d'infection humaine par certaines bactéries marines nommées « *Vibrio* », suite à la consommation des mollusques crus, pourrait s'accroitre durant le 21e siècle (Baker-Austin *et al.*, 2012; Deeb *et al.*, 2018; Martinez-Urtaza *et al.*, 2010; McLaughlin, 2005; Vezzulli *et al.*, 2013; Vezzulli *et al.*, 2016).

Les bactéries du genre *Vibrio (V.)* sont présentes naturellement dans le milieu maritime et plus particulièrement dans les eaux estuariennes et côtières. On peut distinguer deux espèces de *Vibrio* : Les espèces cholériques causant le choléra et dont l'agent pathogène sont les souches des *Vibrio choléra,* et les espèces non cholériques notamment *les Vibrio parahaemolyticus* et *Vibrio vulnificus*.

Ces *Vibrio* non cholériques sont des pathogènes associés à la consommation des mollusques crus sujets à être contaminés (huîtres, moules, myes, pétoncles, etc.…).

L'infection par ces pathogènes opportunistes provoque des gastroentérites dues aux *V. parahaemolyticus* et des infections sévères qui peuvent être mortelles, notamment des septicémies, dues aux *V. vulnificus*. Les signes cliniques de l'infection par *Vibrio* se manifestent généralement 12 heures après l'ingestion de l'aliment contaminé.

En fait, ces *Vibrio* sont très sensibles à la variation de la température de l'eau, c'est pourquoi les changements anticipés du climat et donc l'augmentation de la température de l'eau pourraient avoir un impact positif sur leur prolifération et la prévalence des infections.

Aux États-Unis, plus de 95 % de décès, liés à la consommation des fruits de mer, sont causés par *V. vulnificus* (Oliver, 2013). Les CDC (Centers for Disease Control and Prevention) estiment une moyenne annuelle d'environ 100 cas d'infections par *V. vulnificus*, en entraînant 50 décès par an.

Au Canada, une étude de surveillance sur la diversité et la dynamique des populations de *Vibrio* a été effectuée récemment pour les côtes canadiennes (Colombie-Britannique [côte pacifique], Nouvelle-Écosse [côte atlantique] et la Gaspésie). Cette étude a mis en évidence l'émergence de *V. choléra* dans les estuaires tempérés du Canada et l'augmentation de la détection des souches pathogènes des *V. parahaemolyticus* dans les mollusques bivalves récoltés au Canada (palourdes, moules et huîtres) durant les mois les plus chauds de 2006-2016 (Banerjee *et al.*, 2018). Elle a également révélé la présence de deux tendances ascendantes du risque de croissance des *V. choléra et V. parahaemolyticus* dans les eaux côtières canadiennes (Banerjee *et al.*, 2018).

Étant donné que la récolte des mollusques (pétoncles, moules, huîtres) est en pleine expansion au Québec et à l'Ile-Prince-Édouard, plus précisément dans l'estuaire et golfe du Saint-Laurent, l'analyse du risque futur de la croissance des *Vibrio* pathogènes via la modélisation de scénarios futurs des températures de l'eau est primordiale pour aider à protéger la santé humaine et la qualité des produits maritimes.

1.2 Revue de littérature

La revue de littérature est divisée en deux parties : la première partie concerne les facteurs influençant la croissance des souches pathogènes des *Vibrio* notamment des variables océanographiques physiques (température de l'eau et la salinité), du changement climatique et des téléconnexions (oscillations climatiques). La deuxième partie présente les méthodes de modélisation de la température de l'eau utilisées dans l'estuaire et le golfe du Saint-Laurent.

1.2.1 Les facteurs influençant la croissance des souches pathogènes des *Vibrio*

1.2.1.1 Effet de la température de l'eau et la salinité de l'eau

La recension bibliographique de la relation entre la croissance des *Vibrio* pathogènes et les seuils de température de l'eau est une étape cruciale pour éventuellement calculer des métriques thermiques. Ces derniers peuvent sévir d'indicateurs du risque d'infection par *Vibrio* que nous allons utiliser dans la cartographie des futures zones à risque d'infection.

Plusieurs études ont affirmé que les *Vibrio* préfèrent les eaux chaudes et moins salées (Baker-Austin *et al.*, 2010; Heng *et al.*, 2017; Kaspar & Tamplin, 1993; Motes *et al.*, 1998; Vezzulli *et al.*, 2013). En effet (Motes *et al.*, 1998) ont trouvé que la densité des *V. vulnificus* dans les huîtres, récoltées à partir des sites de la côte du golfe du Mexique et la côte atlantique des

États-Unis, est fortement corrélée avec la température, qui explique la majorité de la variation de sa densité. En fait, ils ont construit un modèle de régression linéaire pour chercher la liaison de variation de la densité des *V. vulnificus* (organisme/g) dans les huîtres récoltées en fonction de la température et la salinité. Avec ce modèle, ils ont trouvé qu'à peu près 60 % de la variation de la densité des *Vibrio* est expliquée par la variation de la température et 10 % de la variation des *Vibrio* est expliquée par la variation de la salinité.

À partir d'un certain seuil thermique variant entre 13 et 26 °C, les souches des *Vibrio* croissent rapidement (Baker-Austin *et al.*, 2010). En plus de la température, le pouvoir pathogène de ces bactéries n'est déclenché qu'en salinité relativement faible, inférieure à 25 ppt (Kaspar & Tamplin, 1993; Motes *et al.*, 1998).

Motes *et al.* (1998) n'ont pas détecté des cas d'infection des huîtres par les *V. vulnificus* pour un climat froid (T<15 °C) au niveau des sites de culture aux États-Unis, localisés sur la côte Nord du golfe du Mexique et la côte Atlantique. En effet, plus de 95 % des décès chez l'humain, suite à la consommation des huîtres crues aux États-Unis, se produisent lorsque la température de l'eau de culture des mollusques dépasse 20 °C (Nishibuchi & DePaola, 2005; WHO, 2005). En outre, ces bactéries ne peuvent pas supporter des salinités élevées, elles sont donc généralement concentrées dans les eaux saumâtres (Nishibuchi & DePaola, 2005; WHO, 2005). Ainsi, on trouve les *V. vulnificus* principalement dans les zones côtières et les estuaires.

Il en va de même pour les *V. parahaemolyticus.* Ils sont concentrés dans les estuaires et les eaux côtières (Liu *et al.*, 2016). La température minimale d'eau de mer qui déclenche l'infection des huîtres par ce type de *Vibrio* est 15 °C, malgré que cette espèce puisse croître dans une température inférieure à 10 °C (McLaughlin *et al.*, 2005).

Le tableau ci-dessous collige les résultats, confirmés par des recherches antérieures, qui précisent des métriques thermiques et de salinité en fonction de la croissance des *Vibrio* pathogènes soit *in situ* ou *in vitro.*

Tableau 1.1 Tableau récapitulatif des seuils thermiques et de salinité en fonction de la croissance des *Vibrio* pathogènes

Type de *Vibrio*	**Type de culture**	**Source**	**Température de l'eau (°C)**			**Salinité (*ppt*)**		
			Minimale	Maximale	Optimale	Minimale	Maximale	Optimale
Vibrio vulnificus	*In vitro*	Kelly (1982)	13 absence des *Vibrio* pathogènes	42 arrêt de la croissance des *Vibrio*	37		85 arrêt de la croissance des *Vibrio*	≤ 20
		Leonard (2011)	8	43			50	
		Kaspar and Tamplin (1993)	0 à 4	≥ 30	entre 13 et 22	5 ppt	> 25 réduction de la concentration des *Vibrio*	entre 5 et 25
	In situ	Kelly (1982)	12 absence des *Vibrio* pathogènes		≥ 25		>16 réduction de la concentration des *Vibrio*	≤ 16
		Motes and DePaola (1996)			≥ 20			≤30
		Motes *et al.* (1998)	<15 absence des *Vibrio* pathogènes		entre 15 et 26		>32 réduction de la concentration des *Vibrio*	≤25 ppt

		WHO (2005) Nishibuchi and DePaola (2005)			≥20 présence des *Vibrio* pathogènes		>30 réduction de la concentration des *Vibrio*	<30
Vibrio parahaemolyticus	In vitro	Leonard (2011)	5	45.3			100 arrêt de la croissance des *Vibrio*	
		Liu *et al.* (2016)			37	5	90 arrêt de la croissance des *Vibrio*	30
	In situ	McLaughlin (2005)			≥15 apparition des *Vibrio* pathogènes et infections			
		Gilliss *et al.* (2013)	10		>10	10	>34 réduction de la concentration des *Vibrio*	23
		Baker-Austin *et al.* (2010)			≥ 15			≤25
		Motes *et al.* (1998)	<15 absence des *Vibrio* pathogènes		entre 15 et 26			≤25

1.2.1.2 Effet du changement climatique sur le taux d'infection par les *Vibrio*

Il y a plusieurs indications que les changements climatiques ont engendré des évènements climatiques extrêmes incluant la canicule et la hausse de l'intensité des précipitations, en particulier durant le dernier demi-siècle. Ces événements extrêmes deviennent de plus en plus fréquents (IPCC, 2013).

Plusieurs études récentes ont mis en évidence le rôle de ces extrêmes climatiques dans l'augmentation du risque d'infection par les *Vibrio* (Baker-Austin *et al.*, 2012; Vezzulli *et al.*, 2013; Vezzulli *et al.*, 2016). Les auteurs ont mentionné que les canicules en Europe du Nord sur les trois décennies passées ont amené des dépassements du seuil de 18 °C qui correspondent à une augmentation significative des infections de plaies causées par les *V. vulnificus* (Baker-Austin *et al.*, 2012).

Les canicules les plus intenses et anormales jamais enregistrées en Scandinavie se sont produites en 2014 et correspondent aux nombres d'infections humaines par les *Vibrio* les plus élevés en Finlande et Suède (Baker-Austin *et al.*, 2016). De même, pour la Nouvelle-Calédonie, des précipitations intenses ont engendré une diminution de la salinité, donc des conditions favorables pour le développement des bactéries, ce qui a mené à un premier cas d'infection humaine par les *V. vulnificus* en 2008 (Cazorla *et al.*, 2011).

Même un petit réchauffement peut affecter le taux de croissance des *V. parahaemolyticus* et réduira l'effet de restriction associé à la période hivernale sur leur cycle de croissance (Vezzulli *et al.*, 2016). Une augmentation moyenne de 1.5 °C de la température de l'eau va étendre la saison et la zone géographique propices à des proliférations avec une abondance plus répandue et un risque plus élevé (Harvell *et al.*, 2002).

1.2.1.3 Effet des téléconnexions

Les téléconnexions représentent des anomalies qui influencent la variabilité de la circulation atmosphérique et océanographique à grande échelle spatiale et temporelle.

Ces oscillations climatiques engendrent des anomalies (différences importantes par rapport à la valeur moyenne) de la température de l'air, de la température de surface de la mer ainsi que des précipitations. On peut les décrire au moyen de certains indices climatiques. Plusieurs oscillations climatiques ont été responsables de l'apparition de cas d'infections par ces *Vibrio*, en particulier certaines phases de l'ENSO (El Niño—Southern Oscillation), provoquant des anomalies climatiques dans l'océan pacifique.

Pour la première fois, une épidémie associée à *V. parahaemolyticus* au Pérou en 1997 a provoqué des infections humaines inattendues dues à la formation des colonies pandémiques qui ont occupé toute la zone côtière du pays (Martinez-Urtaza *et al.*, 2008).

Les investigations subséquentes ont montré que les conditions océanographiques pendant l'apparition des épidémies péruviennes associées à *V. Choléra* et à *V. parahaemolyticus* ont coïncidé avec les épisodes d'El Niño (Martinez-Urtaza *et al.*, 2010; Martinez-Urtaza *et al.*, 2008). Ce phénomène était caractérisé par le déplacement des eaux chaudes et moins salées du Pacifique Ouest vers les zones côtières d'Amérique du Sud.

Similairement, en Alaska en 2004, plusieurs gastroentérites ont été signalées pour la première fois pendant le mois de juillet suite à la consommation d'huîtres crues (McLaughlin *et al.*, 2005). Ces cas d'infections ont été déclenchés suite à l'arrivée extraordinaire des eaux chaudes (>15 °C) de la côte du Pacifique vers le pôle en passant par les côtes d'Alaska (McLaughlin *et al.*, 2005). Aucune infection n'a été détectée depuis 2005 après le retour des étés typiques et frais d'Alaska (Martinez-Urtaza *et al.*, 2010). De plus, il est important de mentionner que l'absence d'infection peut être due au changement de la culture des huîtres (en eau plus profonde) sous des températures inférieures à 10 °C (Martinez-Urtaza *et al.*, 2010).

Pour les États-Unis, le nombre d'infections humaines par les *Vibrio* augmente depuis 2000 pour plusieurs raisons. D'abord, la fréquence et la sévérité des extrêmes des températures et les précipitations s'élèvent de plus en plus surtout dans le sud-ouest à cause des anomalies climatiques et de l'élévation des températures dans l'ouest du Pacifique (Lau *et al.*, 2008). Ces anomalies climatiques coïncident aussi avec les variations de deux phases d'ENSO : El Niño et la Nina (Higgins *et al.*, 2002). Suite à La Nina en 1998, 11 cas d'infection par les *V. vulnificus* ont été détectés durant le mois de novembre dans le golfe du Mexique (Bell *et al.*, 1999; WHO, 2005).

1.2.2 Modélisation de la température de l’eau dans l’estuaire et le golfe du Saint-Laurent

En se basant sur la littérature, les travaux de recherche dans le golfe du Saint-Laurent ont majoritairement porté sur la modélisation de la température de l’eau dans le golfe par l’intermédiaire des modèles déterministes qui constituent une représentation physique et mathématique des processus climatiques et océaniques. Parmi ces modèles, on trouve un modèle en trois dimensions côte-glace-océan qui incluent les composantes atmosphérique, hydrologique et océanographique pour permettre de mieux comprendre le cycle saisonnier de la formation des masses d’eau dans le golfe (Saucier, 2003). En comparant avec les températures et les salinités entre novembre 1996 et avril 1998, les simulations de la température instantanée présentent une erreur entre 1 et 2 °C. Une autre étude a réussi à expliquer les données satellitaires observées de la température de l’eau et sa variation spatiale et temporelle dans le golfe du Saint-Laurent en augmentant la résolution du modèle tridimensionnel (côte-glace-océan), appelé CANOPA (CANadian Océan PArallélisé) (Long *et al.*, 2015). L’augmentation de la résolution se fait par l’utilisation des sorties du modèle climatique régional canadien sur la période 1960-2069 (Long *et al.*, 2015). En utilisant le scénario climatique A1B, les résultats des simulations montrent une augmentation significative de la température de l’eau, de la température de l’air, de la salinité et de la fonte des glaces (Long *et al.*, 2015).

En contraste avec les études précédentes basées sur des modèles déterministes, une étude, basée sur la modélisation statistique, a montré une forte corrélation entre la température de l’eau et la température de l’air dans le golfe du Saint-Laurent (Galbraith *et al.*, 2012). Cette étude suggère que cette corrélation pourrait permettre les prédictions de la température de l’eau dans le golfe du Saint-Laurent dans un contexte de changement climatique par le truchement d’un modèle statistique (Galbraith *et al.*, 2012). L’utilisation des modèles déterministes peut être un outil couteux qui nécessite beaucoup de temps de calcul et grande base de données en intrants (Caissie *et al.*, 1998). La paucité d’application de modèles statistiques dans le golfe peut être expliquée par la complexité du milieu ainsi que l’interaction entre ses variables environnementales. Dans le cas des milieux hydrauliquement plus simples, comme les rivières, on trouve que les modèles statistiques, notamment la régression linéaire multiple et les réseaux de neurones artificiels sont couramment utilisés dans la modélisation des variables environnementales, incluant la température de l'eau (Bélanger *et al.*, 2005; Caissie *et al.*, 1998; Chenard & Caissie, 2008; Jeong *et al.*, 2013).

1.3 Projet de recherche

La modélisation statistique de la température de l'eau fait l'objet de plusieurs études, particulièrement la modélisation de la température de l'eau en rivière (Bélanger *et al.*, 2005; Caissie *et al.*, 1998; Chenard & Caissie, 2008; Jeong *et al.*, 2013). Rares sont les études qui modélisent la température de l'eau dans les milieux côtiers par des modèles statistiques. Le présent projet veut d'une part procéder à la modélisation de scénarios futurs des températures de surface au niveau des bancs coquillers de l'estuaire et du golfe du Saint-Laurent. D'autre part, il veut également utiliser ces scénarios dans l'évaluation et la modélisation d'un risque plausible de croissance des *Vibrio pathogènes* pouvant infecter des mollusques. Nous avons d'abord voulu identifier les facteurs qui peuvent influencer la croissance des souches pathogènes des *Vibrio* pour mieux comprendre la relation entre celles-ci et l'augmentation de la température de l'eau. L'information acquise a servi à préciser des indicateurs de risque de croissance des *Vibrio* pathogènes. Dans le volet de modélisation statistique, nous avons sélectionné les variables explicatives de la variation de la température de l'eau qui constituent les intrants des modèles testés. Après avoir choisi le modèle statistique le plus performant dans la prédiction de la température de l'eau, nous avons produit les scénarios futurs des températures de l'eau et nous avons déduit par la suite les indicateurs de risque en relation avec la croissance des *Vibrio* pathogènes.

Le mémoire de maitrise est divisé en deux articles (chapitre 2 et 3). Le premier article traite le choix des prédicteurs de la température de l'eau et compare les résultats des performances de la modélisation de la température de l'eau réalisée par un modèle statistique paramétrique (la régression linéaire multiple) avec deux modèles d'apprentissage automatique (Réseaux de neurones artificiels et Forêts aléatoires). Le second concerne la production des scénarios futurs de la température de l'eau, la modélisation du risque plausible de la croissance des *Vibrio* pathogènes dans l'estuaire et le golfe du Saint-Laurent et la cartographie des zones à risque élevé d'infection par ce genre de bactérie.

1.4 Questions de recherche, hypothèses et objectifs

Le sujet de cette étude ne porte pas seulement sur l'application des modèles statistiques dans la modélisation d'une variable environnementale, mais aussi sur l'étude d'impact de la variation de cette variable sur la croissance des bactéries pathogènes pouvant mener à des infections humaines. Les questions qui ont été posées au début de ce projet sont les suivantes :

- Quelle est la nature de la relation entre la croissance des *Vibrio* pathogènes et la variation de la température de l'eau ? Quels sont les autres facteurs, en plus de la température de l'eau, qui sont en relation directe avec la croissance des *Vibrio* pathogènes et l'apparition des infections humaines ? Si la température de l'eau favorise la croissance des bactéries *Vibrio*, à partir de quel seuil thermique se déclenche la croissance des souches pathogènes de ces bactéries ?
- Quelles sont les variables qui peuvent influencer la variation de la température journalière de l'eau dans l'estuaire et le golfe du Saint-Laurent ?
- Après la modélisation des scénarios futurs des températures de l'eau, peut-on calculer et cartographier des indicateurs de risque de croissance des *Vibrio* ?

Suite à une revue bibliographique, les hypothèses ci-dessous sont considérées concernant la liaison entre la température de l'eau et la croissance des *Vibrio* pathogènes :

- La température de l'eau est considérée comme étant le facteur principal qui favorise la croissance des *Vibrio* pathogènes malgré l'influence d'autres variables, notamment la salinité. Cette dernière est exclue vu la difficulté d'obtenir des projections des salinités journalières. En plus, plusieurs études ont montré que l'abondance des *Vibrio* est fortement corrélée avec la température de l'eau, avec des concentrations croissantes et proportionnelles au réchauffement saisonnier des eaux.
- Une température de l'eau de 15 °C est considérée comme une limite inférieure d'alerte d'un risque de croissance des *Vibrio* pathogènes et 20 °C comme une limite supérieure d'un risque plus élevé. Ces seuils de températures ont été observés lors de l'apparition d'infections des mollusques ou même d'infections humaines.

- La prolifération des *Vibrio* pathogènes se fait pendant la période estivale (pour une température supérieure à 15 °C), puis ils entrent dans une phase de VNC (Viable Non Cultivable), c.-à-d. les cellules ne se multiplient pas, mais restent vivantes pendant la période froide (T<5 °C). Donc dans cette étude, nous nous concentrons seulement sur les températures estivales, i.e. entre les mois de juin et octobre.

Concernant la modélisation des températures de l'eau dans l'estuaire et le golfe du Saint-Laurent, nous avons considéré les hypothèses suivantes :

- Ce projet porte sur la modélisation des températures journalières de surface (profondeur moyenne de 1.5 m) au niveau des bancs coquillers de l'estuaire et du golfe du Saint-Laurent. À cette profondeur, on peut trouver des mollusques sauvages (e.g. près des îles), de même que des cultures sur les estrans et en eau peu profonde. Cependant, pour d'autres cultures qui se font parfois en eau plus profonde (comme les huîtres), l'identification des zones de risque, basée sur la valeur d'indicateurs du risque déduit à partir des prédictions des températures de surface, peut être biaisée (surestimation) vu qu'il peut y avoir une stratification thermique importante dans certains sites plus profonds et que la modélisation de la température de surface est moins représentative des conditions en profondeur.
- Vu les phénomènes physiques et océanographiques qui peuvent se produire dans l'estuaire et le golfe du Saint-Laurent, la température de l'eau peut varier en fonction de plusieurs variables notamment des variables météorologiques (température d'air, vitesse du vent), océanographique (niveau des marées), atmo océanographiques (téléconnexions) et hydrologique (débit).

La recherche poursuit donc deux objectifs principaux :

- ✓ La modélisation de la température de l'eau aux sites de suivi (thermographes et des bouées) en fonction des variables explicatives (météorologiques, océanographiques, hydrologiques et indices climatiques).
- ✓ La cartographie des futures zones à risque d'infection par *Vibrio* en fonction des indicateurs de risques, déduits en se basant sur les seuils théoriques en relation avec la croissance des *Vibrio* et les scénarios futurs des températures de l'eau.

1.5 Synthèse des travaux de recherche

Les résultats sont divisés en deux parties : la première partie concerne la modélisation de la température de l'eau par l'intermédiaire des modèles statistiques et la deuxième touche l'impact des futurs scénarios de température de l'eau sur le risque de croissance des *Vibrio pathogènes.*

1.5.1 La modélisation de la température de l'eau par des modèles statistiques (Article1)

1.5.1.1 Site d'étude

Le fleuve du Saint-Laurent est le deuxième plus grand fleuve de l'Amérique du Nord (El-Sabh & Murty, 1990), son débit moyen est d'environ 12 100 m^3/s (Galbraith *et al.*, 2017). Il prend sa source des Grands Lacs et devient un estuaire ou se mélangent les eaux douces du fleuve et les eaux salées provenant de l'océan atlantique. L'estuaire s'écoule sur environ 250 km jusqu'à Pointe-des-Monts où il s'élargit et devient un golfe qui s'ouvre à l'océan atlantique par les détroits de Cabot et Belle-Isle.

Le site d'étude, montré dans la (Figure 1.1), comprend les côtes de l'estuaire maritime (à la hauteur de Rimouski) et du golfe du Saint-Laurent (Côte-Nord, Gaspésie, la Baie-des-Chaleurs, Île-du-Prince-Édouard et Îles-de-la-Madeleine).

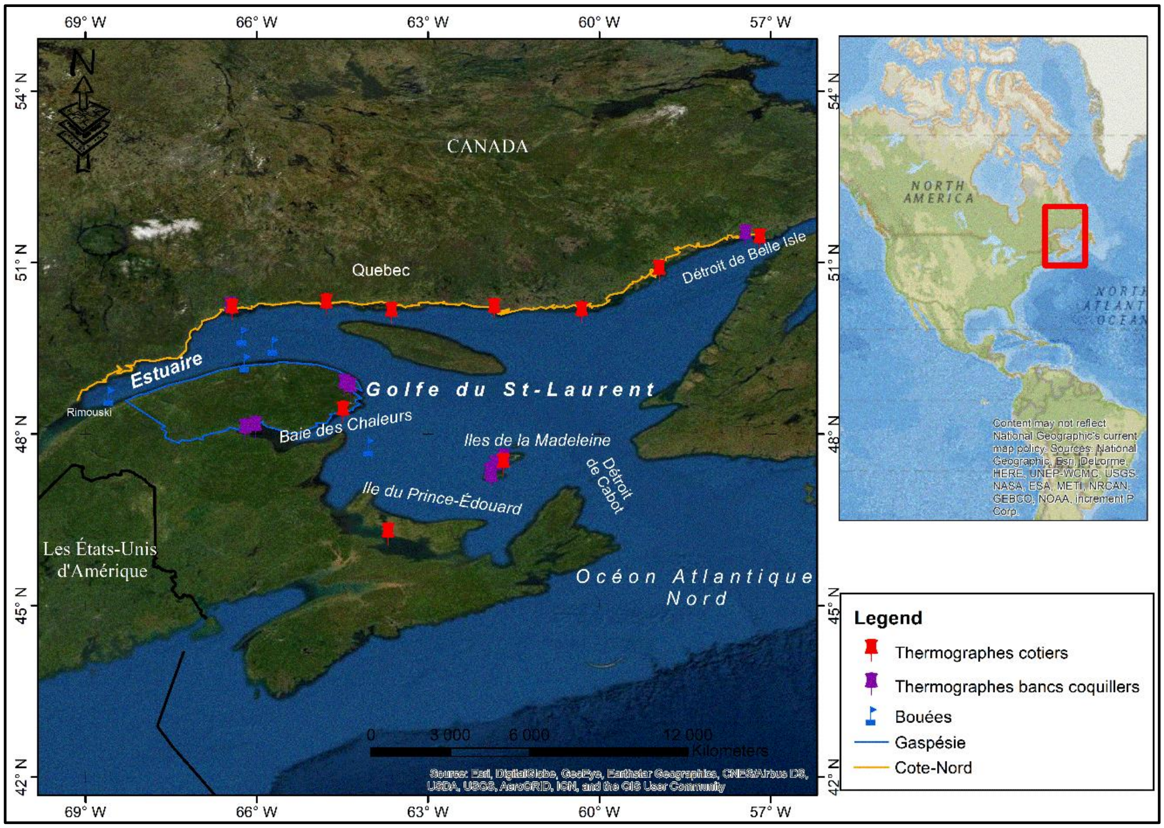


Figure 1.1 Localisation des stations de mesure de la température de l'eau dans l'estuaire maritime et golfe du Saint-Laurent

1.5.1.2 Méthodologie

Afin de sélectionner les prédicteurs de la température de surface de l'eau, nous avons essayé de comprendre les processus physiques et océanographiques qui interviennent dans cette relation. L'océanographie physique et le climat dans l'estuaire et le golfe du Saint-Laurent relient la température de la surface de l'eau (SST), la température de l'air et la vitesse du vent suivant un cycle saisonnier. En été et au printemps, en plus des eaux douces arrivant de l'estuaire du Saint-Laurent (le débit d'eau douce du fleuve Saint-Laurent et la fonte de neige), le rayonnement solaire est une source de chaleur pour la surface des eaux. En conséquence, la température de la surface d'eau atteint son maximum au mois d'août ou juillet. Par contre, en automne et en hiver, la faible température de l'air et la vitesse du vent élevée dissipent la chaleur de la couche superficielle.

En plus de ce cycle météorologique saisonnier, la température de l’eau varie en fonction des marées suivant un cycle journalier, bimensuel et saisonnier. Généralement, la température diminue pendant la marée haute en raison de l’entrée des eaux marines plus froides et plus salées dans les zones côtières. Puis, elle augmente pendant la marée basse. Pour prendre cette relation en considération, nous avons essayé d’introduire le facteur du marnage qui représente la différence entre une pleine mer et une basse mer successive pour détecter la variation de la température de la surface d’eau en fonction du niveau d’eau.

La variabilité et les tendances du climat canadien sont influencées par les oscillations océaniques et atmosphériques à grandes échelles spatiale et temporelle. Les principales oscillations qui ont probablement un effet sur le climat canadien sont : El Niño oscillation australe (ENSO-El Niño—Southern Oscillation) (Bonsal & Shabbar, 2011 ; Shabbar, 2006) par l’intermédiaire des indices Nino 3.3 et Nino 3.4, l’Oscillation Décennale du Pacifique (PDO), la téléconnexion Pacifique-Amérique du Nord (PNA), l’oscillation de l’Atlantique du Nord (NAO), l’oscillation de l’Arctique (AO), l’Oscillation Multidécennale de l’Atlantique (AMO), l’Oscillation d’Est Pacifique/Nord Pacifique (EP/NP) et l’oscillation d’Ouest Pacifique (WP) (Bonsal & Shabbar, 2011 ; Linkin & Nigam, 2008). Certains indices sont calculés au pas du temps journalier et d’autres au pas de temps mensuel. Nous avons sélectionné seulement les indices journaliers (NAO, PNA, ENSO, AO) dans le but de tester la corrélation entre la température de l’eau journalière et ces indices.

Afin de bien identifier les meilleurs prédicteurs, nous avons testé les corrélations croisées linéaires (Pearson) et non linéaires (Spearman), en prenant en considération le décalage dans le temps (lags) entre les Moyennes Mobiles (MM) sur différentes fenêtres (3, 7, 15, 30, 60, 90, 120 jours) des variables explicatives (la température de l’air, la vitesse du vent, le marnage maximal journalier, le marnage moyen journalier, le débit fluvial et les indices climatiques) et la variable d’intérêt (la température moyenne journalière de la surface d’eau).

Les processus physiques expliquant la relation entre la température de l’eau et les variables explicatives (météorologiques, océanographiques, hydrologiques et indices climatiques) sont assez complexes à modéliser à travers un modèle déterministe (Bélanger *et al.*, 2005; Chenard & Caissie, 2008). Les approches statistiques paramétriques (régression linéaire multiple) et non paramétriques (Réseaux de neurones artificiels et Forêts aléatoires) sont des alternatives souvent plus simples pour élaborer des modèles avec un nombre restreint de prédicteurs (Bélanger *et al.*, 2005; Chenard & Caissie, 2008).

Les modèles d'apprentissage automatiques utilisés sont complètement différents. En effet, le modèle des réseaux de neurones artificiels (Artificial Neural Network-ANN, en anglais) est inspiré par l'architecture du système nerveux formé par des couches de neurones. Par contre, le modèle des forêts aléatoires (Random Forest-RF, en anglais) est basé sur le principe de la méthode CART (Classification And Regression Tree), qui est constituée d'arbres de décision binaire.

Pour réduire le problème du surapprentisage des modèles d'apprentissage automatique, l'une des solutions proposées est l'utilisation de la validation croisée (k-fold-cross validation) dans la régularisation des paramètres des modèles. Dans notre étude, nous avons choisi d'utiliser une validation croisée de type « 10 times-10 fold » recommandée par Witten *et al.* (2016) afin d'obtenir un modèle optimal qui minimise l'erreur et réduit le risque de surapprentisage.

Dans cette étude, nous avons utilisé l'algorithme RFE (Recursive Feature Elimination) dans la sélection des variables explicatives (prédicteurs) les plus pertinentes et non redondantes. Durant le processus d'apprentissage, le RFE permet de sélectionner le meilleur ensemble de prédicteurs qui contribue le plus à une meilleure performance prédictive du modèle en optimisant un critère de performance. Dans cette étude, le critère considéré est la racine de l'erreur quadratique moyenne (RMSE, équation 3.11). Pour chaque modèle, la sélection d'attributs pertinents est exécutée pendant la phase d'entrainement sur 80 % de données. Les 20 % de données restantes (les données du test) sont inconnues pour le modèle. Ces données du test sont utilisées pour déterminer la performance de la prédiction de chaque modèle en comparant plusieurs critères tels que la racine de l'erreur quadratique moyenne (RMSE, équation 3.11), la racine relative de l'erreur quadratique moyenne relative (rRMSE, équation 3.12), le coefficient de détermination (R^2, équation 3.13), le coefficient de Nash-Sutcliffe (Nash, équation 3.14) et le biais relatif (rBiais, 3.15).

La comparaison des critères de performance des modèles testés est complétée en utilisant le test d'ANOVA (ANalysis Of Variance) simple (à un facteur) en sélectionnant l'un de ses critères. Le test d'ANOVA est utilisé pour comparer les moyennes d'un ou plusieurs groupes. Lorsque le test d'ANOVA est significatif, on passe à déterminer le ou les groupe(s) différent(s) des autres en les comparant deux à deux. Cette intercomparaison est exécutée au moyen du test Tukey Honestly Significant Difference (Tukey HSD).

1.5.1.3 Résultats

L'analyse de la corrélation pour la plupart des thermographes testés indique que la température de la surface de l'eau (SST) est significativement plus corrélée avec les MM-3 jours de la température de l'air (Tair), MM-30 jours de la vitesse du vent, MM-30 jours du marnage maximal journalier, MM-60 jours de NAO et MM-120 jours du débit fluvial du Saint-Laurent. En effet, la Tair explique plus de 60 % de la variation de SST tandis que la vitesse du vent explique entre 10 et 50 % de sa variation. La dépendance entre SST et le débit fluvial est faible. Le débit fluvial influence surtout la température de l'eau au niveau de l'estuaire maritime en expliquant 10 % de sa variation. Cette dépendance diminue entre 5 et 7 % en s'éloignant vers l'embouchure de l'estuaire et les stations côtières proches. Concernant les indices climatiques, seulement l'indice NAO est significativement corrélé avec SST. Son effet majeur est exercé au niveau des stations de la Côte Nord en expliquant entre 6 et 10 % de la variation de SST.

Nous présentons dans le premier manuscrit les résultats de comparaison des critères de performance des modèles testés au niveau de l'une des stations (Table 3.4). Cette comparaison montre que les modèles d'apprentissage automatique (ANN et RF) performent mieux que le modèle linéaire multiple (Multiple Linear Regression-MLR, en anglais). En particulier, le RF est plus performant que les modèles ANN et MLR grâce aux valeurs les plus faibles des critères RMSE, rRMSE, et les valeurs de Nash et R^2 les plus élevées. En plus, le modèle RF possède un meilleur ajustement et une meilleure précision de prédiction que l'ANN (Figure 3.3).

Par la suite, nous comparons les résultats des performances des modèles testés pour toutes les stations de la zone d'étude à travers le test d'ANOVA. Les modèles testés sont comparés en vérifiant si les moyennes des valeurs de leurs RMSE sont significativement différentes ou pas. Le test d'ANOVA (degré de liberté = 2, statistique-F=5.806) indique une différence significative entre les modèles au seuil de signification de 5 % (valeur de p= 0.006). Ensuite, un test de Tukey Honestly Significant Difference (Tukey HSD) est exécuté entre les différents modèles testés pour l'ensemble des stations. Le test montre qu'il n'y a pas de différence significative entre ANN et MLR (valeur de p = 0.89). En revanche, il existe une différence significative entre le modèle RF et les deux modèles MLR (valeur de p = 0.008) et ANN (valeur de p = 0.026) au seuil de signification de 5 %. Le meilleur sous-ensemble d'entités pour le modèle RF diffère d'une station à l'autre, mais il inclut pratiquement tous les prédicteurs testés. Concernant la performance de RF, la valeur moyenne de RMSE calculée pour l'ensemble de stations est 1.3 °C.

1.5.1.4 Discussion

Dans cette étude, nous avons démontré que des variables explicatives, autres que la température de l'air, expliquent un pourcentage important de la variation de la température de l'eau dans la zone d'étude. Ces prédicteurs incluent une autre variable météorologique (vitesse du vent), une variable océanographique (marnage), hydrologique (débit) et une variable liée aux téléconnexions (NAO). La plupart des travaux, qui portent sur l'indice climatique NAO, montrent l'occurrence de son effet sur la variation de la pression au niveau de la mer durant l'hiver (Hurrell, 1995). Très peu d'études ont montré que l'effet d'indice climatique peut se produire durant l'été (Barnston & Livezey, 1987 ; Portis *et al.*, 2001). Selon nos résultats, il semble pouvoir influencer la variation de la température de l'eau au niveau de la Côte-Nord.

Les modèles d'apprentissage automatique performent mieux que le modèle linéaire dans la modélisation de la température de l'eau dans l'estuaire et le golfe du Saint-Laurent. L'ajout des variables explicatives, autres que la température de l'eau, a amélioré significativement la précision de la prédiction de la température de l'eau. Le modèle RF est le meilleur des modèles testés pour la prédiction de la température de l'eau. Cette approche offre des performances plutôt robustes, grâce à la conception de son algorithme qui permet un compromis faible variance – faible biais. Cependant, il exige un nombre plus important de variables explicatives. Le meilleur sous-ensemble des attributs, qui contribue à une meilleure prédiction, diffère d'une station à l'autre. Cette différence peut être due aux différentes conditions physiques (océanographiques) et climatologiques associées à chaque station.

1.5.2 L’impact des scénarios futurs des températures de l’eau sur la croissance des *Vibrio* pathogènes (Article2)

1.5.2.1 Méthodologie

Dans cette étude, nous présentons le risque futur de croissance des *Vibrio* pathogènes dans l’estuaire et le golfe du Saint-Laurent sous différents scénarios climatiques (optimiste et pessimiste) en nous basant sur la modélisation des scénarios futurs de la température de l’eau et le seuil théorique des conditions favorables à la croissance de ces bactéries. Afin de produire les scénarios futurs des températures de l’eau, nous avons utilisé les prédicteurs les plus pertinents, la température de l’air et la vitesse du vent, qui expliquent plus de 70 % de la variation de la température de l’eau dans la plupart des stations, comme le montre l’étude précédente.

Nous avons choisi deux scénarios climatiques RCP (Representative Concentration Pathways) : un scénario moyen (RCP 4.5) et un scénario pessimiste (RCP 8.5) appelé également « Business as usual ». Ces scénarios conduisent à un réchauffement des températures moyennes de l’air de 2.5 °C à 5 °C vers 2100.

Ouranos, un consortium sur la climatologie régionale et l’adaptation aux changements climatiques au Québec, nous a fourni les projections climatiques de la température de l’air et la vitesse du vent des points de grille du modèle climatique régional canadien situés à proximité des stations météorologiques sélectionnées dans l’estuaire et le golfe du Saint-Laurent. En utilisant la MM-3jours de la Tair et la MM-30 jours de la vitesse du vent, nous avons réentrainé les modèles (RF et ANN) de nouveau et utilisé RFE pour le choix du meilleur sous ensemble des prédicteurs. Dans ce cas, seulement deux combinaisons des prédicteurs sont probables soit la température de l’eau comme intrant soit la température de l’eau et la vitesse du vent. Nous comparons ensuite les températures simulées par les deux modèles d’apprentissage automatique (ANN et RF) avec les observations selon un critère de performance (RMSE) en utilisant le t-test. Nous choisissons le modèle le plus performant dans la prédiction de la température de l’eau pour l’ensemble des stations et nous générons les projections de la température de l’eau journalière pour l’horizon 2040-2100. Par la suite, nous utilisons le test de Mann-Kendall Modifié, qui tient en compte des autocorrélations dans la série temporelle, pour tester la tendance des futures températures de l’eau.

Afin d'évaluer le risque potentiel de la croissance des *Vibrio*, nous avons choisi le nombre de jours au-dessus de la température minimale de croissance des *Vibrio* pathogènes (15 °C), comme métrique thermique. Puis, nous calculons la moyenne mensuelle de cet indicateur durant la période d'étude (juin-octobre). Ensuite, nous calculons la moyenne les valeurs mensuelles obtenues sur une vingtaine d'années durant la période d'étude 2040-2100. Finalement, nous interpolons les valeurs de l'indicateur de risque obtenues pour toutes les stations dans la zone d'étude en utilisant la méthode d'interpolation spatiale basée sur une pondération inverse à la distance (Inverse Distance Weighting-IDW) avec le logiciel ARCGIS.

Des cartes d'interpolations sont subséquemment produites pour les deux horizons futurs, soit 2040-2060 et 2080-2100, et pour les deux scénarios de changements climatiques (optimiste et pessimiste) afin de pouvoir comparer le degré d'expansion possible du risque de croissance des *Vibrio* dans la zone d'étude. L'évaluation de ce risque de la croissance ne se limitait pas à un seul indicateur de risque. Nous avons également calculé le nombre de jours dépassant une limite supérieure de 20 °C pour localiser les zones à risque d'infection plus élevée.

1.5.2.2 Résultats

Les résultats de t-test (DF=13, t-value=1.196) indiquent qu'il n'existe pas une différence significative entre les moyennes des RMSE de deux modèles (Table 7.2) pour l'ensemble des stations (valeur de p=0.24) au seuil de signification de 5 %. La valeur moyenne de RMSE calculée pour l'ensemble de stations est d'environ 1.74 °C pour le RF et 1.8 °C pour ANN. En comparant les stations testées, le meilleur sous ensemble d'attributs pour le modèle RF comprend la température de l'air et la vitesse du vent alors qu'il comprend seulement la température de l'air pour le modèle ANN dans la plupart des stations. Étant donné qu'il n'existe pas une différence significative entre les moyennes de RMSE de deux modèles, le modèle le plus parcimonieux choisi pour la suite des recherches est l'ANN. Des scénarios futurs de températures de l'eau ont été produits à l'aide du modèle ANN pour toutes les stations incluant les stations au niveau des bancs coquillers, fournis par Merinov-Centre d'Innovation de l'aquaculture et des pêches du Québec, en utilisant comme seul prédicteur les projections de la Tair pour l'horizon 2040-2100.

En comparant les résultats du test de la tendance des futures températures de l'eau journalières produites pour chaque station (Table 7.3) dans un scénario pessimiste, on remarque que la pente décennale des futures températures de l'eau varie entre 0.2 °C et 0.7 °C et que les futures températures de l'eau vont augmenter entre 1.25 °C et 4.27 °C vers 2100. En moyennant les pentes des tendances pour toutes les stations, on trouve que la température de l'eau va augmenter de 0.4 °C par décennie, pour un total de 2.5 °C durant la période 2040-2100.

Nous avons présenté comme exemples de la distribution future du risque de la croissance des *Vibrio*, les cartes d'interpolation spatiale d'indicateur du risque durant les mois d'août et septembre (Figure 7.2 et 7.3).

Durant la période 2040-2060 dans un cadre de scénario optimiste (Figure 7.2a), on remarque que les eaux des bancs coquillers aux Îles-de-la-Madeleine, à l'Ile-Prince-Édouard, la Gaspésie et la Baie-des-Chaleurs seront à risque élevé d'infection, car le nombre de jours au-dessus du seuil de 15 °C excède 25 jours. Le nombre de jours dépassant le seuil, calculé aux bancs coquillers de la Côte-Nord, varie quant à lui entre 20 et 25 jours. Pour la même période, mais dans un cadre de scénario pessimiste (Figure 7.2 b), la plupart des bancs coquillers seront à risque élevé d'infection par *Vibrio* avec un nombre de jours au-dessus du seuil > 25 jours occupant environ 67 % de la surface des zones côtières. En comparant les cartes d'interpolations du mois d'août (Figure 7.2c et 7.2 d) durant la période 2080-2100 pour les deux scénarios (optimiste et pessimiste), on remarque que le nombre de jours au-dessus du seuil (15 °C), dépassant 25 jours, occupe entre 64 % (scénario4.5) et 95 % (scénario8.5) de la surface totale des zones côtières. Donc, vers 2100, la plupart des stations localisées dans les bancs coquillers, où se localisent des zones de récolte des mollusques bivalves, seront probablement sous un risque élevé d'infection par *Vibrio* pour les deux scénarios climatiques envisagés. Durant le mois de septembre (Figure 7.3), le nombre de jours au-dessus du seuil, au niveau des bancs coquillers des Îles-de-la-Madeleine et de l'IPE, excède 20 jours pour les deux scénarios climatiques envisagés, et ce, pour les deux horizons. Donc, ils seront sous un risque élevé d'infection par *Vibrio.* Les résultats d'indicateur du risque pour le mois d'octobre (Table 7.4) montrent que le risque d'infection va se prolonger jusqu'au mois d'octobre sur les côtes de l'IPE et des Îles-de-la-Madeleine. Donc, l'expansion du risque d'infection se prolonge spatialement et temporellement.

Durant la période 2080-2100 et dans un cadre de scénario pessimiste (Table 7.5), un seuil de 20 °C dans les zones de culture des Îles-de-la-Madeleine serait dépassé durant environ 31 jours au mois d'août contre une moyenne de 18 jours au mois de septembre. Les bancs coquillers de Gaspésie seront au même niveau de risque que ceux des Iles de la Madeleine avec un nombre de jours au-dessus du seuil de 20 °C dépassant 25 jours tandis que les bancs coquillers de l'IPE et de la Côte-Nord auront un risque moins élevé avec une moyenne de 16 jours.

1.5.2.3 Discussion

En utilisant les prédicteurs les plus pertinents comme intrants aux modèles d'apprentissage automatique (ANN et RF), la température de l'air et la vitesse du vent, nous avons constaté que leurs performances de prédiction de la température de l'eau étaient similaires. Nous avons choisi le modèle ANN dans la prédiction des futures températures de l'eau étant donné sa relative simplicité et le fait qu'il requiert seulement la température de l'air comme intrant pour la plupart des stations.

Les résultats de l'analyse des tendances des températures de l'eau journalière prédites sur l'horizon 2040-2100, à l'aide du test MMK, indiquent que notre zone d'étude présente une tendance à la hausse significative de 2.5 °C en 2100 selon un scénario pessimiste. Cette tendance positive de la température de l'eau implique une augmentation de l'indicateur de risque de croissance de *Vibrio*. Ce résultat est démontré dans les cartes d'interpolation entre les horizons 2040-2060 et 2080-2100. En fait, en comparant les cartes d'interpolation d'août pour les deux horizons du scénario pessimiste (RCP 8.5), nous notons une augmentation du risque de croissance de *Vibrio* de 64 % à 95 % de la totalité des zones côtières de l'estuaire et du golfe du Saint-Laurent. En plus de l'expansion spatiale, les résultats du calcul d'indicateur de risque durant le mois d'octobre montrent que le risque de croissance des *Vibrio* va se prolonger temporellement sur les côtes de l'IPE et des Îles-de-la-Madeleine.

L'évaluation du risque plausible de croissance de *Vibrio* par le calcul d'un indicateur du risque plus élevé, le nombre de jours dépassant un seuil de 20 °C, a permis de localiser les bancs de coquillages présentant un risque plus élevé de croissance de *Vibrio*, précisément ceux des Îles-de-la-Madeleine et de l'IPE.

1.6 Conclusion

Les variables explicatives, autres que la température de l'air, contribuent à expliquer la variation de la température de l'eau. Ces prédicteurs varient selon les stations et les modèles, mais peuvent inclure des prédicteurs météorologique (vitesse du vent), océanographique (marnage), hydrologique (débit) et un indice de téléconnexion (Indice NAO). Les modèles d'apprentissage automatique (ANN et RF) permettent d'expliquer les relations non linéaires complexes qui peuvent exister entre la température de l'eau et ces variables. Nous avons montré que ces modèles performent mieux que le modèle linéaire simple dans la prédiction de la température de l'eau.

Dans le but de modéliser les scénarios futurs de températures de l'eau, on a comparé la performance de la prédiction de deux modèles d'apprentissage automatique (RF et ANN). En utilisant les prédicteurs les plus pertinents en intrants aux modèles testés, Tair et la vitesse du vent, nous avons appliqué le RFE pour la sélection des meilleurs attributs. Suite à la comparaison des modèles ANN et RF, nous n'avons pas trouvé de différence significative entre leur performance de prédiction (RMSE). Étant donné que le modèle ANN est le modèle le plus parcimonieux et simple, en utilisant seulement la température de l'air en intrant, il a été sélectionné pour générer les scénarios de futures températures de l'eau.

En utilisant les projections de la température de l'air en intrant au modèle sélectionné, nous avons pu produire les futures températures de l'eau pour l'ensemble des stations, puis calculer le nombre de jours au-dessus de la température minimale de la croissance des *Vibrio* pathogènes (15 °C), comme indicateur du risque potentiel de croissance.

En produisant les cartes du futur risque potentiel de croissance des *Vibrio pathogènes*, suite à une interpolation spatiale de l'indicateur du risque, nous avons pu conclure que le risque pourrait être élevé au niveau du golfe du Saint-Laurent, notamment aux Îles-de-la-Madeleine, en Gaspésie, à la Baie-des-Chaleurs, à l'IPE et pour une partie de la Côte-Nord selon les deux scénarios climatiques utilisés.

Le risque de croissance des *Vibrio* pathogènes pourrait prendre de l'expansion en allant vers 2100 selon les deux scénarios. En effet, vers 2100, la couverture spatiale du nombre de jours au-dessus du seuil (15 °C) dépassant 25 jours, passe de 64 % (le scénario optimiste) à 95 % (le scénario pessimiste) de la surface des zones côtières de l'estuaire et du golfe du Saint-Laurent. En plus, cette expansion se prolonge temporellement en couvrant même le mois d'octobre pour certaines stations.

La température n'est pas la seule variable qui influence la croissance des *Vibrio*. La salinité est une autre variable importante. Une phase ultérieure de ce projet devrait prendre en compte cette variable. Pour ce faire, un modèle déterministe de circulation des masses d'eau de l'estuaire et du golfe du Saint-Laurent devrait être mis à profit. De plus, le présent projet a porté sur les températures de surface (profondeur moyenne de 1.5 m). Cependant, il peut y avoir une importante stratification thermique dans certains sites côtiers. Cette possible stratification devrait aussi être prise en compte dans une éventuelle seconde phase du projet.

2 RÉFÉRENCES

Baker-Austin C, Stockley L, Rangdale R & Martinez-Urtaza J (2010) Environmental occurrence and clinical impact of Vibrio vulnificus and Vibrio parahaemolyticus: a European perspective. Environmental microbiology reports 2(1):7–18.

Baker-Austin C, Trinanes JA, Salmenlinna S, Löfdahl M, Siitonen A, Taylor NGH & Martinez-Urtaza J (2016) Heat Wave–Associated Vibriois, Sweden and Finland, 2014. Emerging Infectious Diseases 22(7):1216–1220.

Baker-Austin C, Trinanes JA, Taylor NGH, Hartnell R, Siitonen A & Martinez-Urtaza J (2012) Emerging Vibrio risk at high latitudes in response to ocean warming. Nature Climate Change 3(1):73–77.

Banerjee SK, Rutley R & Bussey J (2018) Diversity and dynamics of the Canadian coastal Vibrio community: an emerging trend detected in the temperate regions. Journal of bacteriology 200(15):e00787-00717.

Barnston AG & Livezey RE (1987) Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. Monthly weather review 115(6):1083–1126.

Bélanger M, El-Jabi N, Caissie D, Ashkar F & Ribi JM (2005) Estimation de la température de l'eau de rivière en utilisant les réseaux de neurones et la régression linéaire multiple. Revue des sciences de l'eau 18 (3).

Bell GD, Halpert MS, Kousky VE, Gelman ME, Ropelewski CF, Douglas AV & Schnell RC (1999) Climate Assessment for 1998. Bulletin of the American Meteorological Society 80(5):1040-1040.

Bonsal B & Shabbar A (2011) Large-scale climate oscillations influencing Canada, 1900–2008. Canadian Councils of Resource Ministers, Caissie D, El-Jabi N & St-Hilaire A (1998) Stochastic modelling of water temperatures in a small stream using air to water relations. Canadian Journal of Civil Engineering 25(2):250–260.

Cazorla C, Guigon A, Noel M, Quilici M-L & Lacassin F (2011) Fatal Vibrio vulnificus infection associated with eating raw oysters, New Caledonia. Emerging infectious diseases 17(1):136.

Chenard J-F & Caissie D (2008) Stream temperature modelling using artificial neural networks: application on Catamaran Brook, New Brunswick, Canada. Hydrological Processes 22(17):3361–3372.

Deeb R, Tufford D, Scott GI, Moore JG & Dow K (2018) Impact of Climate Change on Vibrio vulnificus Abundance and Exposure Risk. Estuaries and Coasts 41(8):2289–2303.

El-Sabh MI & Murty TS (1990) Mathematical modelling of tides in the St. Lawrence Estuary. Oceanography of a Large-Scale Estuarine System, Springer. p 10–50.

Galbraith PS, Chassé J, Gilbert D, Larouche P, Brickman D, Pettigrew B, Devine L, Gosselin A, Pettipas R & Lafleur C (2017) Physical oceanographic conditions in the Gulf of St. Lawrence in 2016. Canadian Science Advisory Secretariat, Galbraith PS, Larouche P, Chassé J & Petrie B (2012) Sea-surface temperature in relation to air temperature in the Gulf of St. Lawrence: Interdecadal variability and long term trends. Deep Sea Research Part II: Topical Studies in Oceanography 77–80:10–20.

Gilliss D, Cronquist AB, Cartter M, Tobin-D'Angelo M, Blythe D, Smith K, Lathrop S, Zansky S, Cieslak PR, Dunn J, Holt KG, Lance S, Crim SM, Henao OL, Patrick M, Griffin PM & Tauxe RV (2013) Incidence and Trends of Infection with Pathogens Transmitted Commonly Through Food—Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 1996–2012. MMWR. Morbidity and Mortality Weekly Report 62(15):283–287.

Harvell CD, Mitchell CE, Ward JR, Altizer S, Dobson AP, Ostfeld RS & Samuel MD (2002) Climate Warming and Disease Risks for Terrestrial and Marine Biota. Science 296(5576):2158–2162.

Heng S-P, Letchumanan V, Deng C-Y, Ab Mutalib N-S, Khan TM, Chuah L-H, Chan K-G, Goh B-H, Pusparajah P & Lee L-H (2017) Vibrio vulnificus: An Environmental and Clinical Burden. Frontiers in microbiology 8:997-997.

Higgins RW, Leetmaa A & Kousky VE (2002) Relationships between Climate Variability and Winter Temperature Extremes in the United States. Journal of Climate 15(13):1555–1572.

Hurrell JW (1995) Decadal trends in the north atlantic oscillation: regional temperatures and precipitation. Science 269(5224):676–679.

IPCC (2013) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 1535 p. www.climatechange2013.org

Jeong DI, Daigle A & St-Hilaire A (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(7):805–821.

Kaspar CW & Tamplin ML (1993) Effects of temperature and salinity on the survival of Vibrio vulnificus in seawater and shellfish. Applied and environmental microbiology 59(8):2425–2429.

Kelly MT (1982) Effect of temperature and salinity on Vibrio (Beneckea) vulnificus occurrence in a Gulf Coast environment. Applied and environmental microbiology 44(4):820–824.

Lau N-C, Leetmaa A & Nath MJ (2008) Interactions between the Responses of North American Climate to El Niño–La Niña and to the Secular Warming Trend in the Indian–Western Pacific Oceans. Journal of Climate 21(3):476–494.

Leonard B (2011) Fish and Fishery Products: Hazards and Controls Guidance (4th Ed.). DIANE Publishing Company. https://books.google.tn/books?id=UALJdPmp3GsC

Linkin ME & Nigam S (2008) The North Pacific Oscillation–West Pacific Teleconnection Pattern: Mature-Phase Structure and Winter Impacts. Journal of Climate 21(9):1979–1997.

Liu B, Liu H, Pan Y, Xie J & Zhao Y (2016) Comparison of the Effects of Environmental Parameters on the Growth Variability of Vibrio parahaemolyticus Coupled with Strain Sources and Genotypes Analyses. Frontiers in microbiology 7:994.

Long Z, Perrie W, Chassé J, Brickman D, Guo L, Drozdowski A & Hu H (2015) Impacts of Climate Change in the Gulf of St. Lawrence. Atmosphere-Ocean 54(3):337–351.

Martinez-Urtaza J, Bowers JC, Trinanes J & DePaola A (2010) Climate anomalies and the increasing risk of Vibrio parahaemolyticus and Vibrio vulnificus illnesses. Food Research International 43(7):1780–1790.

Martinez-Urtaza J, Huapaya B, Gavilan RG, Blanco-Abad V, Ansede-Bermejo J, Cadarso-Suarez C, Figueiras A & Trinanes J (2008) Emergence of Asiatic Vibrio Diseases in South America in Phase With El Niño. Epidemiology 19(6):829–837.

McLaughlin (2005) climate anomalies and the increasing risk of Vibrio.

McLaughlin JB, DePaola A, Bopp CA, Martinek KA, Napolilli NP, Allison CG, Murray SL, Thompson EC, Bird MM & Middaugh JP (2005) Outbreak of Vibrio parahaemolyticus gastroenteritis associated with Alaskan oysters. New England Journal of Medicine 353(14):1463–1470.

Motes ML & DePaola A (1996) Offshore suspension relaying to reduce levels of Vibrio vulnificus in oysters (Crassostrea virginica). Applied and environmental microbiology 62(10):3875–3877.

Motes ML, DePaola A, Cook DW, Veazey JE, Hunsucker JC, Garthright WE, Blodgett RJ & Chirtel SJ (1998) Influence of Water Temperature and Salinity on Vibrio vulnificus in Northern Gulf and Atlantic Coast Oysters (Crassostrea virginica). Applied and environmental microbiology 64(4):1459–1465.

Nishibuchi M & DePaola A (2005) Vibrio species. Foodborne pathogens: Microbiology and molecular biology :251-227.

Oliver JD (2013) Vibrio vulnificus: Death on the Half Shell. A Personal Journey with the Pathogen and its Ecology. Microbial ecology 65(4):793–799.

Portis DH, Walsh JE, El Hamly M & Lamb PJ (2001) Seasonality of the North Atlantic oscillation. Journal of Climate 14(9):2069–2078.

Saucier FJ (2003) Modeling the formation and circulation processes of water masses and sea ice in the Gulf of St. Lawrence, Canada. Journal of Geophysical Research 108(C8).

Shabbar A (2006) The impact of El Niño-Southern Oscillation on the Canadian climate. Advances in Geosciences 6:149–153.

Vezzulli L, Colwell RR & Pruzzo C (2013) Ocean warming and spread of pathogenic Vibrio in the aquatic environment. Microbial ecology 65(4):817–825.

Vezzulli L, Grande C, Reid PC, Helaouet P, Edwards M, Hofle MG, Brettar I, Colwell RR & Pruzzo C (2016) Climate influence on Vibrio and associated human diseases during the past half-century in the coastal North Atlantic. Proceedings of the National Academy of Sciences of the United States of America 113(34):E5062-5071.

WHO F (2005) Risk assessment of Vibrio vulnificus in raw oysters: interpretative summary and technical report, microbiological risk assessment series 8: Geneva: World Health Organization. Rome: Food And Agriculture Organization of the United Nations.

Witten IH, Frank E, Hall MA & Pal CJ (2016) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann,

Deuxième partie

ARTICLES

3 ARTICLE 1: MODELLING COASTAL WATER TEMPERATURE AND IMPLICATIONS FOR POTENTIAL INFECTIONS WITH VIBRIO MARINE BACTERIA

Title of article: **Modelling coastal water temperature and implications for potential infections with Vibrio marine bacteria**

Titre de l'article: **Modélisation de la température de l'eau côtière et ses implications sur des infections potentielles par des bactéries marines du genre *Vibrio***

Auteurs et affiliations :

Habiba Ferchichi[1], André St-Hilaire[1, 2], Taha B.M.J Ouarda[1], Benoît Levesque[3]

1. INRS-ETE. 490 De la Couronne St, Quebec City, QC, Canada
2. Canadian Rivers Institute, Fredericton, NB, Canada
3. INSPQ, Quebec City, Canada

Manuscript was submitted to **Environmental Modelling & Software** journal

11th March 2020

Contributions des auteurs :

Les coauteurs A.S., T.B.M.J.O. et B.L ont défini la méthodologie à suivre et révisé le manuscrit. A.S. et T.B.M.J.O. ont proposé les modèles statistiques à tester. L'auteur principal H.F. a réalisé l'analyse, l'interprétation des résultats et la préparation des figures sous la supervision des coauteurs.

3.1 Abstract

Vibrio parahaemolyticus and *Vibrio vulnificus* are zoonotic pathogen bacteria causing foodborne infections associated with the exceedance of Sea Surface Temperature (SST) threshold in estuarine and coastal ecosystems. This study tests the abilities of statistical (Multiple Linear Regression) and machine learning models (Random Forests and Artificial Neural Network) to predict the daily SST, given the selected predictors. Different measures of correlation are used to examine the evolution of the SST as a function of meteorological (air temperature and wind speed), oceanographic (tidal range) and hydrological (flow) variables as well as teleconnection patterns in order to select the best inputs for the tested model. The results show that Random Forests is the best algorithm for properly taking into account the most relevant predictors to predict SST. This model has a root-mean-square error (RMSE) ranging between 0.77 °C and 1.8 °C, with an overall RMSE of 1.3 °C across all the stations.

Key words: Coastal Water temperature, Vibrio bacteria, machine learning models, statistical modelling.

3.2 Abbreviations

AAO, Antarctic Oscillation; ANN, Artificial Neural Networks; ANOVA, ANalysis Of Variance; AO, Arctic Oscillation; Bagging, Bootstrap Aggregating; CANOPA, CANadian Océan PArallélisé ; CART, Classification And Regression Tree; ENSO, El Niño Southern Oscillation; Flow, St. Lawrence freshwater runoff; GSL, Gulf of St. Lawrence; MA, Moving Average; MLP, Multilayer Perceptron; MLR, Multiple Regression Model; MSE, Mean Square Error; NAO, North Atlantic Oscillation; Nash, Nash-Sutcliffe coefficient; Ntry, Number of predictors randomly selected to try at each split of a tree; OOB, Out-Of-Bag; PNA, Pacific/North American; rBias, Relative mean bias; RF, Random Forests; RFE, Recursive Feature Elimination; RMSE, Root Mean Square Error; rRMSE, relative Root Mean Square Error; R2, The coefficient of determination; spp, Species; SST, Sea Surface Temperature; Tair, Air temperature; Tidal range, Maximum daily tidal range; Tukey HSD, Tukey Honestly Significant Difference; The USA, the United States of America; V., Vibrio; VIF, Variation Inflation Factor; WSPD, Wind speed

3.3 Introduction

Vibrio (V.) species (spp), divided into cholera and non-cholera groups, are marine bacteria that are naturally ubiquitous in warm coastal waters (>15 °C) and estuaries with low salinity (<25 ppt) (Baker-Austin et al., 2010; Kaspar and Tamplin, 1993; Motes et al., 1998; Vezzulli et al., 2013). The non-cholera species, especially *Vibrio parahaemolyticus* and *Vibrio vulnificus*, are the leading cause of foodborne infection (Deeb et al., 2018; Semenza et al., 2017) (known generally as vibriosis). Infection by *V. parahaemolyticus* causes gastroenteritis, while *V. vulnificus* leads to serious life-threatening infections, such as septicemia, with subsequent mortality exceeding 50% (Dechet et al., 2008; Kim et al., 2011; Oliver, 2005). The proliferation of most *Vibrio* spp, causing human illness, depends on the exceedance of a threshold of water temperature (about 15 °C) (Baker-Austin et al., 2013; Jacobs et al., 2015; Martinez-Urtaza et al., 2010; McLaughlin et al., 2005). Given that bivalves are filter feeders, *Vibrio* bacteria may accumulate in their tissues. These shellfish could be contaminated with *Vibrio* as the exceeding water temperature threshold promotes *Vibrio* growth. The *Vibrio* bacteria are mostly transmitted to humans through the consumption of raw or undercooked contaminated shellfish (oysters, mussels, clams, etc.) (Baker-Austin et al., 2017; Davis et al., 2017; McLaughlin et al., 2005; Motes and DePaola, 1996; Zimmerman et al., 2007).

Numerous reports showed that the rate of vibriosis is increasing worldwide (Centers for Disease and Prevention, 2013; Martinez-Urtaza et al., 2010; Newton et al., 2012). For example, the average annual incidence of *Vibrio* infections in the USA has increased by 41% between 1996 and 2005 (Centers for Disease, 2006). Most scientists in this field of research believe that the expansion of the prevalence of this pathogen, as well as the rise of infection cases, is linked to SST exceeding certain thresholds (Baker-Austin et al., 2012; Burge et al., 2014; Deeb et al., 2018; Martinez-Urtaza et al., 2010; Vezzulli et al., 2016). The increase in frequency and duration of these exceedances are likely caused by climate change (Baker-Austin et al., 2012; Burge et al., 2014; Deeb et al., 2018; Martinez-Urtaza et al., 2010; Vezzulli et al., 2016). Some scientists have even considered *Vibrio* as microbial barometers of climate change (Baker-Austin et al., 2017). Thus, it is expected that the increasing water temperature, caused by ocean warming, would increase human exposure (Baker-Austin et al., 2017; Vezzulli et al., 2013) to *Vibrio* bacteria, extend the length of their seasonal occurrence and expand their geographical range (Harvell et al., 2002; Vezzulli et al., 2016).

The Gulf of St. Lawrence (Canada), as the rest of worldwide marine ecosystems, is affected by climate change. Indeed, the SST in the Gulf of St. Lawrence (GSL) increased by 1 °C to 1.5 °C for the period 1982–2011 (inter-annual average, May to November) (Galbraith et al., 2012). Shellfish harvesting is an important and dynamic activity (leisure and commercial) in Eastern Canada. Shellfish beds are located in coastal areas of the Estuary and GSL including Rimouski, Bay of Chaleur, the Quebec North Shore, Magdalen Islands and Prince Edward Island.

Given the SST increase in the GSL and the importance of shellfish harvest along Eastern Canadian Coasts, predicting the future *Vibrio* risk of proliferation is of the utmost importance.

The usual developed predictive model of *Vibrio* concentration are based either on only sea surface temperature (SST) (Chu et al., 2011) or both of the SST and salinity (Jacobs et al., 2014; United States Food and Drug Administration (FDA), 2005). Other models are developed for modelling the relation of *Vibrio* infections exposure in response to SST threshold exceedance (Semenza et al., 2017). Given the paucity of *Vibrio* concentration data as well as *Vibrio* infection data, the purpose of this study is limited to evaluate the *Vibrio* growth risk through its relation with the SST threshold exceedance.

This is done through the prediction of future daily mean SST. The first step in predicting the future SST scenarios, detailed in this study, is modelling the historical SST data using available correlated predictors. The model that achieves the best performance can be selected for further studies for predicting the future SST scenarios using the predictors' projections produced by the climate models. Then, a map of the future potential risk areas can be produced using indicators such as the number of days exceeding the selected temperature thresholds.

Modelling SST in GSL has been generally achieved through deterministic models, i.e. physical and mathematical representation of the climatic and ocean processes such as the three-dimensional coastal ice-ocean model called CANOPA (CANadian Océan PArallélisé) (Long et al., 2015; Saucier, 2003).

The predictive model that we use consists of applying a machine learning model or data mining algorithm for predicting the future values of a desired variable using a number of predictors (Shmueli, 2010). The selection of the predictors is based on the statistical association between the dependent and independent variables rather than the causal relationship, adding to the quality and the availability of the data (Shmueli, 2010).

The coastal water temperature depends on a number of variables, including atmospheric, oceanographic, hydrological and climate indices variables. Several studies show that air temperature (Deser et al., 2010; Galbraith et al., 2012), wind speed (Deser et al., 2010; Huang and Qiao, 2009; Ng et al., 2009; Weber et al., 2013), tidal fluctuations (Kim et al., 2010; Lowe et al., 2016; Weber et al., 2013), flow (Materia et al., 2012) and climate indices (Deser et al., 2010; Qu et al., 2012) affect sea water temperature.

The SST in the Estuary and GSL is governed by the interaction of climatic processes following a seasonal cycle. During spring, the solar heating, the ice melt and the freshwater runoff from the St. Lawrence River produce a lower salinity and a rise in SST. The SST reaches its maximum value from mid-July to mid-August. In winter, before the sea ice formation, the cooling and the wind dissipate the heat of the surface layer. The negative relationship between SST and wind speed (Huang and Qiao, 2009; Qu et al., 2012) is due to the fact that wind forcing causes the mixing of the cold thermocline waters and the warm surface water. Strong winds break down the vertical stratification (Wang et al., 1999).

The St. Lawrence River is the second largest river in North America (El-Sabh and Murty, 1990), with an average discharge of 12,100 m^3s^{-1} (Galbraith et al., 2017). Freshwater runoff from the St. Lawrence River decreases the estuarine salinity and increases temperature during spring and early summer. It also contributes more than 80% of freshwater entering the Gulf from the estuary (Koutitonsky, 1991).

In addition to the meteorological seasonal cycle, tides affect SST in daily, fortnightly and seasonal cycles. The temperature decreases during rising tide due to the inflow of cooler and saltier marine waters to warmer coastal areas. Then it increases during the ebb tide. The fluctuation in SST is larger during spring tide than during neap tide as a result of the rise in the tidal range during the former.

A significant portion of Canadian climate variability is due to persistent and occurring patterns of atmospheric and ocean circulation anomalies called teleconnections (Bonsal and Shabbar, 2011; Canadian Science Advisory Secretariat, 2017; Shabbar, 2006; Thiombiano et al., 2017), which also affects the climate variability in the GSL (Bonsal and Shabbar, 2011; Canadian Science Advisory Secretariat, 2017). Numerous studies have discussed the relationship between the climate variability and these large-scale atmospheric and oceanic oscillations widely (Alizadeh-Choobari and Adibi, 2019; Perlwitz et al., 2017). These patterns, measured by climate indices, last from several weeks to several years and even decades over vast areas, and influence air temperatures, SST, wind speed and precipitation (Abhishek et al., 2010; National Weather Service Climate Prediction Center, 2008). El-Nino Southern Oscillation (ENSO) was the cause behind the emergence of *V. parahaemolyticus* epidemic on the entire coast of Peru in 1997 (Martinez-Urtaza et al., 2010; Martinez-Urtaza et al., 2008) and *V. vulnificus* illness in the Gulf of Mexico in 1998 (Bell et al., 1999; WHO, 2005). The anomalous high water temperature, associated with El Niño episodes, caused the large outbreak of *Vibrio* infections in these regions (Martinez-Urtaza et al., 2010).

Although the effect of air temperature along with other meteorological variables on SST variation is known, statistical models have not been used for modelling SST in the GSL, as a potentially simpler, less costly alternative to deterministic models (Caissie et al., 1998). Statistical models (Multiple Regression Model [MLR], Artificial Neural Network [ANN]) have been frequently used in modelling river temperature using meteorological variables, especially air temperature, as predictors (also known as features) (Bélanger et al., 2005; Caissie et al., 1998; Chenard and Caissie, 2008; Jeong et al., 2013). The Random Forests (RF), a machine learning model, has not been tested yet in daily water temperature modelling, although it often displays a better performance than ANN, linear models and other statistical models in different environmental modelling studies (Li et al., 2016; Naing and Htike, 2015; Obringer and Nateghi, 2018; Pang et al., 2017).

3.4 *Material and methods*

3.4.1 *Data description*

Given that Vibrio occurrence is restricted to the warm months, the study period is fixed between June and October. The study database is composed of the daily mean water temperature, at an average depth across all the stations of 1.5 m, and a set of daily predictors. The SST time series at the available stations in GSL (Figure 3.1), where the many shellfish harvesting are located, varied in length between 15 and 25 years.

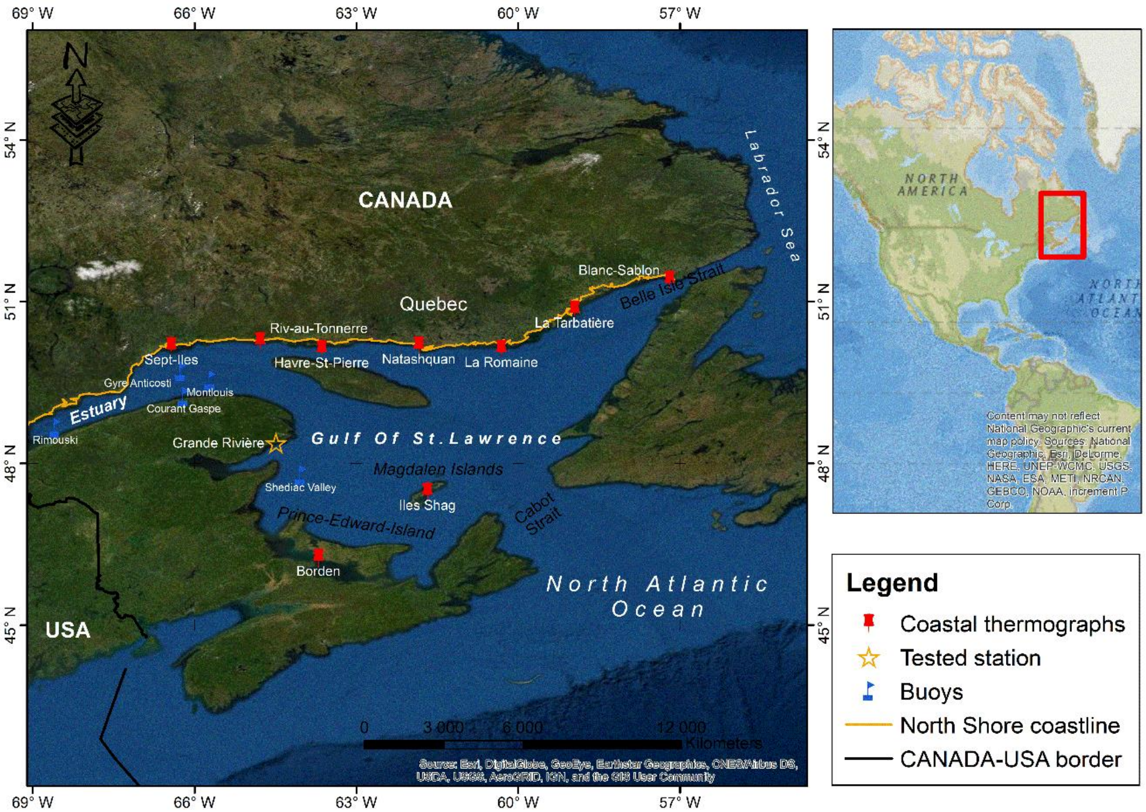


Figure 3.1 Geographic locations of the costal thermographs and buoys in the Estuary and Gulf of St. Lawrence

In addition to air temperature, we use wind speed, flow, tidal range and climate indices as explanatory variables of the sea surface temperature.

The tested climate indices are the North Atlantic Oscillation (NAO), Pacific/North American (PNA), Arctic Oscillation (AO), Antarctic Oscillation (AAO), ENSO (El Niño Oscillation) through the indices Nino 3.4 and Nino3. The aforementioned teleconnections patterns affect climate variability of the Northern Hemisphere (Pozo-Vázquez et al., 2001; Ropelewski and Arkin, 2019; Seager et al., 2010; Thompson and Wallace, 1998) and the Canadian climate in particular (Bonsal and Shabbar, 2008; Bonsal and Shabbar, 2011; Canadian Science Advisory Secretariat, 2017; Shabbar, 2006). Although, the AAO is the dominant mode of the atmospheric variability in the southern hemisphere (Hall and

Visbeck, 2002; Riffenburgh, 2007; Silvestri and Vera, 2003), there are several studies that show its impact on the climate variability of the Northern Hemisphere (Fan and Wang, 2004; Jian-Qi, 2010; Nan and Li, 2003; Wang and Fan, 2007).

The tidal range represents the difference in sea level between low tide and high tide. The tidal cycle in GSL is semi diurnal, composed of two daily high and low tides. We select the maximum tidal range among predictors because it is found to be more correlated with SST in the majority of stations than the mean tidal range.

The Grande-Rivière station is selected as an example due to its relatively long data record (1995–2016).

The used data are gathered from different government agencies and some online resources:

The daily air temperature and wind speed data for Canada are freely available online from the Government of Canada: http://climate.weather.gc.ca/historical_data/search_historic_data_e.html.

The daily sea surface temperatures of the buoys and costal thermographs were supplied by Fisheries and Oceans Canada and the Maurice-Lamontagne Institute (Pettigrew et al., 2016).

The daily water levels are available online from Fisheries and Oceans Canada: http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/twl-mne/inventory-inventaire/index-eng.htm.

The daily climate indices are available online from the Earth System Research Laboratory of the National Oceanic and Atmospheric Administration: ftp://ftp.cpc.ncep.noaa.gov/cwlinks/.

The daily freshwater runoff data of the St. Lawrence River at the LaSalle station are provided online by Environment Canada: https://wateroffice.ec.gc.ca/google_map/google_map_e.html?map_type=historical.

3.4.2 Modelling water temperature

In order to find the best predictors that explain the most variation of the target variable (SST), we first test the linear and non-linear correlations between the daily mean water temperature and the daily values of the predictors. Then, we use a moving average (MA) filter for smoothing the predictors' data in order to get clearer signals. We apply different windows of trailing MA (3, 7, 15, 30, and 90,120–days) on the predictors' signals and test their correlations with the SST variation for different lags times (from 0 to 30 days). For example, we test the correlation of the daily mean water temperature at day j with the trailing 3-day MA air temperature (Tair), i.e. the mean of daily air temperature over the days (j, j-1 and j-2) for different time lags. The selection of the best moving average, as well as lag time for each predictor is made according to the strength of its correlation with the daily mean water temperature.

This step allows to choose between the predictors' daily raw data and the smoothed ones based on their correlation with the daily SST data. The selected predictors are then entered as models inputs.

We split the data into 80% for training and cross validation sets and 20% for the test set. The cross validation is used in order to avoid the overfitting through selecting the best model after evaluating different models with different combinations of hyperparameters and features subsets. The prediction accuracy of the selected final model is tested using the test dataset.

3.4.2.1 **Multiple Linear** Regression **Model**

The Multiple Linear Regression (MLR) is an extension of a simple linear regression which represents the relationship between a continuous dependent variable and two or more independent variables. MLR can be written as follows:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_{p-1} X_{p-1} + \varepsilon \tag{3.1}$$

Where $\hat{Y}$ is the predicted value of the dependent variable, X_1 through X_{p-1} represents the predictor variables, ε is the residual (error) term of the model, β_0 intercept term, and β_1 through β_{p-1} are the estimated regression parameters.

❖ ***Variation Inflation Factor***

The occurrence of multicollinearity bet ween the predictors may pose a problem for linear models. Multicollinearity increases the variance of the estimated coefficients and weakens the statistical power of a linear model to identify the significant predictors (Zuur et al., 2010). Among existing multicollinearity indicators (Dormann et al., 2013), we use the Variation Inflation Factor (VIF) (Belsley, 1993; Hair et al., 1998).

$$VIF_k = \frac{1}{1 - {R_k}^2} \tag{3.2}$$

Where VIF_k is the variation inflation factor for predictor k and R_k^2 represents the coefficient of determination from a linear regression in which the predictor k is the response variable and the rest of predictors as explanatory variables. A VIF value exceeding the critical value (10) indicates the presence of significant multicollinearity.

3.4.2.2 Artificial Neural Network Model

The applied Artificial Neural Network (ANN) model is called a Multilayer Perceptron (MLP), using the backpropagation as training algorithms. This model is formed by three layers: the input layer contains the predictors, which are standardized by subtracting each variable by its mean and dividing by its standard deviation. Then, there is one hidden layer followed by the output layer, which produces the response variable. The ANN principle consists in calculating the weighted sum (Z_i) of the inputs (X_i) plus a bias (b). These weights associated with each input represent the strength of the relationship between the explanatory variables and the variable of interest.

The weighted sum is given by:

$$Z_i = \sum_{i=1}^{n} w_i X_i + b \tag{3.3}$$

Then, we apply a sigmoid activation function ($f(Z_i)$) to generate the desired output given by:

$$f(Z_i) = \frac{1}{1 + e^{-Z_i}} \tag{3.4}$$

We use the multilayer perception network with one hidden layer. A single hidden layer is considered sufficient for approximating any continuous function, but there is no accepted general rule for the number of hidden nodes in the hidden layer (Haykin, 1994; Piotrowski et al., 2015). Using few or many nodes in the hidden layer may result in either underfitting or overfitting the model. Thus, the stability of the ANN depends on the number of hidden nodes in the hidden layer.

There are several methods for ANN architecture selection (Panchal and Panchal, 2014). They include the rules of thumb and trial-error methods. The ANN architecture selection depends on the complexity of the specific problem. Thus, the trial-error method is more appropriate (Zhu et al., 2018). This method is characterized by repeating a serial of attempts until satisfying a specific stopping criterion. In the case of hidden neurons optimization, it consists in testing the model performance by using a grid search of hidden neurons number. We use the forward trial-error method by testing an increasing number of neurons from 1 to 10. This grid is integrated inside of 10 fold cross validation method repeated 10 times (10-times 10-fold cross validation).

This 10-times 10-fold cross validation is recommended by Witten and Frank (2005) for the optimization of hyperparameters and the selection of the best model (Guyon and Elisseeff, 2003; Krstajic et al., 2014; Witten et al., 2016). For each station, we compare the means of the cross validated root mean square error associated with each ANN architecture. The parameter, minimizing the cross validated error, is then used to select the final model.

Testing different ANN architectures shows that adding more than one neuron in the hidden layer induces weak performance improvement (less than 10%) in the majority of stations. Figure II.1, given in appendix II, presents an example of the performance accuracy of different models with different ANN architectures for one of the stations. It shows that using 5 nodes in the hidden layer improves the prediction accuracy by only 2% compared to one hidden node. Thus, we select the ANN architecture with only hidden layer containing one hidden node providing simple model with good performance and fewer parameters.

In addition to the number of hidden neurons, we use a grid search for the weight decay, penalizing large weight values, in order to avoid overfitting.

The ANN is applied by using the nnet function with the Caret package (Kuhn, 2008) in the R programming language (Team, 2013). The nnet method is a feedforward neural network with one hidden layer and uses back propagation as learning algorithm (Ripley et al., 2016). The applied backpropagation is based on an iterative optimization method called BFGS Quasi Newton backpropagation (Ripley et al., 2016). This optimization stops when it converges toward a unique configuration. We identify the adequate number of iterations that assures model convergence by increasing the number of iterations from 50 to 1000. The optimal number of iterations depends on the training stopping criteria. The training is stopped when the network error is below a given threshold (Ripley et al., 2016).

3.4.2.3 Random Forests Model

Random Forests (RF) is a recent machine learning algorithm derived from the Classification and Regression Tree (CART), i.e. decision tree algorithm. It is comprised of decision tree ensembles trained on about 2/3 bootstrap samples of the data (Bootstrap), leaving one third of the dataset for the validation called Out-Of-Bag (OOB). To split each node of the tree, the RF uses a random features selection (Random subspace). The final output is determined by averaging the prediction of all the decision trees (Aggregation) built for the bootstrapped samples from the original dataset (Boostrap). Combining the CART model with the Bagging (Bootstrap Aggregating) and Random subspace aims to reduce high variance of the CART model and helps to avoid overfitting through building independent and stable decision trees, i.e. reducing the correlation between the trees.

The parameters set are: the number of trees to run which is experimentally set after plotting the OOB error as a function of tree numbers, the number of input variables randomly selected to try at each split of a tree (Ntry), which is obtained by dividing the number of input variables by three as recommended by the RF creators (Breiman, 2001) (Ntry=1), and the minimum node size of each tree, which is set at 5. Similar to ANN, we use the trial-error method to confirm the choice of the best tuning hyperparameters. This is done by implementing 10-times 10-fold cross validation.

The prediction error is calculated at the level of the OOB samples by averaging the prediction error of each tree during the training process. In case of a regression model, the OOB error is computed by measuring the following Mean Square Error (MSE):

$$MSE^{OOB} = \frac{1}{n} \times \sum_{i=1}^{n} [\hat{Y}(X_i) - Y_i]^2 \tag{3.5}$$

Where n is the size of the OOB sample, $\hat{Y}(X_i)$ is the output associated with the given input X_i, and Y_i is the actual output. The RF is applied by using the ranger function with the Caret package (Kuhn, 2008) in the R programming language (Team, 2013).

3.4.2.4 Model Validation using K-Fold Cross Validation

The prediction performance assessment can be done using cross validation or split sample validation (Guyon and Elisseeff, 2003). The cross validation, a model validation technique, is one of widely used resampling data methods for calibrating model hyperparameters, assessing model prediction accuracy and avoiding overfitting (Berrar, 2019; Duda et al., 2012; Hastie et al., 2009). The k-fold cross validation is one of the cross validation techniques. It consists of training and validating the model on multiple train-validation splits instead of using one split sample. The training dataset is partitioned randomly into k equal size folds (subsamples). The model is trained on the k-1 subsamples, which represent the training set and validated on the remaining fold (Berrar, 2019). The process is repeated until all the folds have served as validation sets (Berrar, 2019). The cross validated performance (CV_k) is computed by averaging the mean square error (MSE) of the k validation sets (Berrar, 2019).

$$CV_k = \frac{1}{k} \sum_{j=1}^{k} MSE_j \tag{3.6}$$

With the MSE of each fold among the k-folds, is given by:

$$MSE_j = \frac{1}{n}\sum_{i=1}^{n}(y_{i,obs} - y_{i,p})^2 \tag{3.7}$$

Where k represents the number of folds (e.g. 10-fold cross validation), *j* is one of the k folds, *n* is the data size in the fold j, $y_{i,\,obs}$ are the observed data and $y_{i,\,p}$ are the predicted data.

The repeated k-fold cross validation (n-times k-fold cross validation) repeats the same process of k-fold cross validation more than once (*n*-times), so it creates more validation sets (number of folds*number of repetition= k*n) than a single k-fold approach. This method provides high prediction performance accuracy and best choice of model hyperparameters.

3.4.2.5 Feature Selection

The feature selection includes three methods: Filter, Wrapper and Embedded methods (Guyon and Elisseeff, 2003). The Wrapper method (Kohavi and John, 1997) searches the optimal features subset after a performance prediction comparison of the different features combinations (Guyon and Elisseeff, 2003). The Recursive Feature Elimination (RFE), one of the commonly used feature selection methods, is one of the Wrapper methods (Guyon et al., 2002).

The RFE, a backward selection technique, performs feature subset selection during the training process by selecting features contributing the most to model accuracy. This method trains the model by using all the features of training data, it sequentially removes the least important variable based on its contribution to model predictive accuracy, also known as variable importance. Then, it retrains the model associated with the reduced feature set (Guyon et al., 2002).

The RFE selects automatically the best subset of predictors that optimizes the selected performance criteria. We consider the RMSE as the performance criterion to be optimized in the RFE. In order to get the variation due to the feature subset selection, the performance prediction assessments is done with k-fold cross validation (10-fold cross validation). For each feature subset, the goodness of models fit is evaluated through computing the 10-fold cross validated RMSE ($RMSE_{CV}$), i.e. the average of all the ten RMSE achieved over the 10 validation sets:

In general, the k-fold cross-validated RMSE is given by:

$$RMSE_{CV} = \frac{1}{k}\sum_{j=1}^{k} RMSE_j \tag{3.8}$$

Where the RMSE calculated for each fold j is given by:

$$RMSE_j = \sqrt{\sum_{i=1}^{N} \frac{(y_{i,p} - y_{i,obs})^2}{N}} \tag{3.9}$$

Where the k represents the number of folds in the used cross validation (10 in this example), j is one of the k folds, N is the data size in the fold j, $y_{i,\ p}$ are the predicted data and $y_{i,\ obs}$ are the observed data.

After computing the performance prediction criterion ($RMSE_{CV}$) of all possible subsets of predictors, the default best feature subset usually corresponds to the lowest value of $RMSE_{CV}$. Sometimes, there is a very small difference between the different possible subsets. Thus, we decide to add another condition in the selection of the best feature subset. We calculate firstly the performance loss for each subset compared to the performance criterion of the best features subset (Kuhn, 2012) given by:

$$loss = \frac{(X - x)}{x} \times 100 \tag{3.10}$$

Where X is the set of performances for different subsets and x is the lowest value of the selected performance criterion.

This criterion reflects the rate of the RMSE increase in each feature subset, i.e. the decrease in model performance, compared to the default best feature subset. Then, if a certain feature subset contains fewer predictors than the most performant subset and at the same time yields less than 10% increase in the best value of $RMSE_{CV}$, we choose it as the best feature subset.

The RFE is applied by using the RFE function with the Caret package (Kuhn, 2012) in the R programming language (Team, 2013).

3.4.2.6 Performance criteria

After selecting the best model through the cross validation process, we test its prediction performance using the test dataset.

We use five performance criteria to evaluate the performance of the linear and machine learning models: the RMSE, rRMSE, rBias, Nash and R^2.

- Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_{i,obs} - y_{i,p})^2}{n}} \tag{3.11}$$

- relative Root Mean Square Error (rRMSE)

$$rRMSE = 100 \times \sqrt{\sum_{i=1}^{n} \frac{(\frac{y_{i,obs} - y_{i,p}}{y_{i,obs}})^2}{n}} \tag{3.12}$$

- The coefficient of determination (R^2)

$$R^2 = \left(\frac{\sum_{i=1}^{n} (y_{i,obs} - \bar{y}_{obs})(y_{i,p} - \bar{y}_p)}{\sqrt{\sum_{i=1}^{n} (y_{i,obs} - \bar{y}_{obs})^2} \ \sqrt{\sum_{i=1}^{n} (y_{i,p} - \bar{y}_p)^2}} \right)^2 \tag{3.13}$$

- Nash-Sutcliffe coefficient (Nash)

$$Nash = 1 - \frac{\sum_{i=1}^{n} (y_{i,obs} - y_{i,p})^2}{\sum_{i=1}^{n} (y_{i,obs} - \bar{y}_{obs})^2}$$

(3.14)

- Relative mean bias (rBias)

$$rBias = \frac{100}{n} \times \sum_{i=1}^{n} \frac{y_{i,obs} - y_{i,p}}{y_{i,obs}} \quad (3.15)$$

Where $y_{i,obs}$ represents the observed data, $y_{i,p}$ represents the predicted data, $\bar{y}_{obs}$ represents the mean of observed data, $\bar{y}_p$ represents the mean of predicted data and n the number of observations. Generally, higher values of Nash and R^2 indicate higher accuracy prediction. In contrast, lower values of RMSE, rRMSE and rBias indicate better fitted models.

3.4.2.7 Analysis of Variance Test

A one way analysis of variance (ANOVA) is used for comparing the means of more than two groups. This test is used to compare the performances of three models obtained for all the stations. The ANOVA assumptions were checked using the Q-Q plot, the Shapiro test for the residual normality assumption and the Leven test for the homogeneity of variance assumptions. When the ANOVA test is significant, we use the Tukey Honestly Significant Difference (Tukey HSD) for pairwise comparisons between the means of groups.

3.5 Results

3.5.1 Selection of the best predictors of the daily mean sea surface temperature

The linear and non-linear correlation results indicate that the smoothed predictors were highly correlated with the response variable than the predictors' daily values (Figures I.1-I.4, in appendix I). This result suggests that using the trailing MA filter for smoothing the predictors' signals, allows detecting higher correlations between the smoothed predictors and the daily mean SST. Figures I.1-I.4 also show that correlations decrease as the lag time increases. Thus, a lag of zero is selected.

The result of the linear correlation (Pearson) analysis, shown in Table 3.1, indicates that the SST is significantly and most correlated with 3-day MA air temperature (Tair), 30-day MA wind speed (WSPD), 30-day MA tidal range (Tidal range), 60-day MA North Atlantic Oscillation (NAO) and 120-day MA St. Lawrence flow (Flow) for zero day lag time. The 3-day MA air temperature shows the highest correlation with SST and explains more than 60% of its variation for the majority of the stations (nearly 70%). The dependency between the 30-day MA wind speed and SST is moderate. For nearly 90% of the stations, the wind speed explains between 10% and 50% of SST variation. Concerning freshwater runoff, the 120-day MA St. Lawrence flow is weakly correlated with SST. It has a larger influence on estuarine SST, explaining 10% of SST variation. Then, it decreases between 5 and 7% for the buoys located at the mouth of the estuary, and for some coastal stations near it. The 30-day MA tidal range explains less than 10% of SST variation at most stations.

Table 3.1 The results of the linear correlations between the daily SST and the selected predictors for each station

Stations	Tair	WSPD	Tidal range	Runoff	NAO
Rimouski	0.89	-0.33	0.18	0.32	-0.31
Courant-Gaspé	0.91	-0.71	0.28	0.23	-0.32
Gyre-Anticosti	0.57	-0.43	0.24	0.27	-0.29
Montlouis	0.91	-0.43	0.30	0.23	-0.22
Sept-Îles	0.84	-0.13	0.31	0.23	-0.30
Rivière-aux-Tonnerre	0.77	-0.28	0.19	0.27	-0.32
Havre-St-Pierre	0.59	-0.42	Non-sig	0.10	-0.26
Natashquan	0.86	-0.31	0.25	0.21	-0.17
Romaine	0.89	-0.49	-0.15	0.04	-0.29
La Tartbatière	0.61	-0.55	-0.25	-0.04	-0.21
Blanc-Sablon	0.66	-0.58	-0.26	Non-sig	-0.20
Grande-Rivière	0.84	-0.70	0.04	0.10	-0.16
Shediac-Valley	0.94	-0.47	Non-sig	Non-sig	-0.10
Iles-Shag	0.75	-0.44	0.30	-0.10	-0.22

Borden	0.81	-0.50	0.46	Non-sig	-0.16

Non-sig (Non-significant): The linear correlation between the dependent variable (daily SST) and the predictor is not statistically significant at a 5% level.

After testing the correlations between climate indices and SST, it is found that the only index that is significantly correlated with the dependent variable is the 60-day MA North Atlantic Oscillation index (Table 3.2). Its main effect is exerted for stations in the Estuary and the North Shore of GSL, explaining between 6 and 10% of SST variations for the majority of those stations. For the remaining stations, it explains less than 5% of SST variations. The Spearman rank correlation gives similar result to the Pearson correlation regarding the choice of the best moving average filter and lag time for each predictor in relation to the target variable.

Table 3.2 ***The results of the linear correlations between the daily SST and the 60-day MA climate indices at each station***

Stations	PNA	NAO	AO	NINO3.4	NINO3	AAO
Rimouski	-0.23	**-0.31**	-0.16	Non-sig	Non-sig	Non-sig
Courant-Gaspé	-0.21	**-0.32**	-0.20	Non-sig	0.08	0.21
Gyre-Anticosti	-0.29	**-0.29**	-0.26	0.08	0.11	Non-sig
Montlouis	-0.10	**-0.22**	-0.07	Non-sig	0.06	0.08
Sept-Îles	-0.17	**-0.30**	-0.14	0.04	0.09	0.07
Rivière-aux-Tonnerre	-0.13	**-0.32**	-0.22	0.05	0.15	0.16
Havre-St-Pierre	-0.05	**-0.26**	-0.23	0.04	0.08	0.11
Natashquan	-0.16	**-0.17**	-0.08	Non-sig	0.07	0.06
Romaine	-0.04	**-0.29**	-0.19	0.08	0.12	0.05
La Tartbatière	0.05	**-0.21**	-0.18	Non-sig	0.08	0.07
Blanc-Sablon	Non-sig	-0.20	**-0.23**	0.05	0.05	Non-sig
Grande-Rivière	Non-sig	**-0.16**	Non-sig	Non-sig	Non-sig	Non-sig
Shediac-Valley	**-0.12**	-0.10	Non-sig	Non-sig	non	0.06

Iles-Shag	0.08	**-0.22**	Non-sig	Non-sig	Non-sig	Non-sig
Borden	-0.12	**-0.16**	Non-sig	Non-sig	Non-sig	Non-sig

Non-sig (Non-significant): The linear correlation between the daily SST and the tested climate index is not statistically significant at a 5% level.
Bold characters denote the highest linear correlation between the daily SST and 60-day MA of tested climate index among the tested indices.

After selecting the best low-pass filter (MA) for each predictor from the correlation analysis, a pairwise correlation is applied. The pairwise correlation consists in computing the correlation matrix, i.e. the zero-order correlations between each pair of variables. Figure 3.2 presents the linear correlation between the predictors and the target data as well as between the predictors themselves. All the pairwise correlations between the dependent variable (SST) and the selected predictors are statistically significant (p-value <5%).

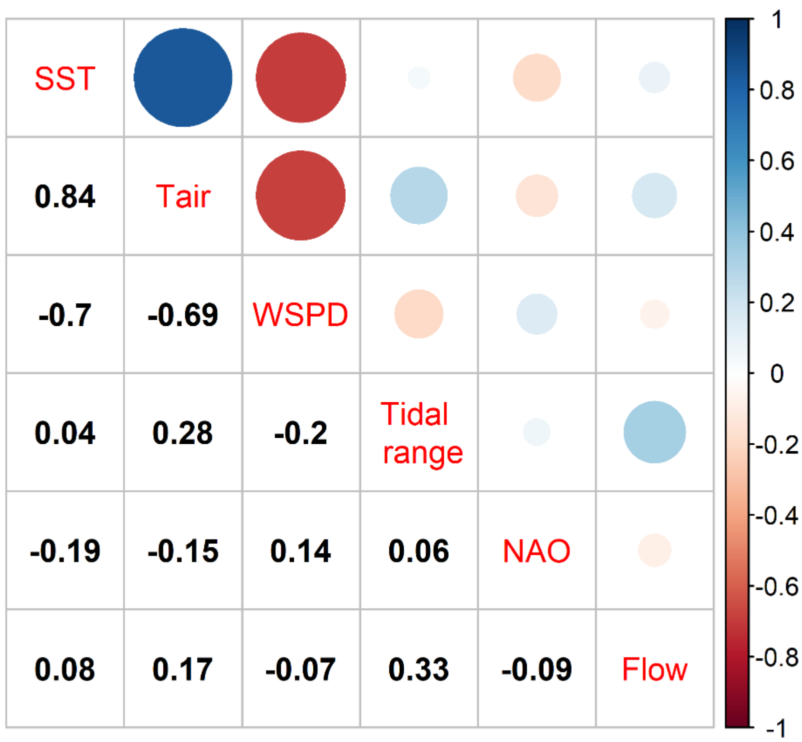


Figure 3.2 The pairwise correlation matrix of Grande-Rivière variables

Figure 3.2 indicates that the SST is highly correlated with Tair and WSPD, while it’s weakly associated with the rest of the predictors. In spite of the relative weakness of the correlation between SST and some predictors, we decide to keep them after testing partial correlations (Table 3.3). This is a measure of the linear relationship between the dependent variable with one of the independent variables while keeping the remaining independent variables in the model constant (Freund et al., 2010).

The partial correlation between SST and these variables, as illustrated in Table 3.3, increases when the selected predictor was included in the model.

Table 3.3 The results of the linear partial correlations between the daily SST and the selected predictors at the Grande-Rivière station

	Tair	WSPD	Tidal_range	Flow	NAO
SST	0.74	-0.32	-0.38	Non-sig	-0.06

These correlation tests provide an initial list for the primary selection of predictors. Then, we use the feature selection in the modelling fit process. This method ensures removing redundant variables taking into account the interactions among them, and extracting the subset of the most relevant variables that yield the best prediction performance and avoid overfitting (Guyon and Elisseeff, 2003).

Regarding the correlation between the predictors, the WSPD is highly correlated with Tair while the Tidal range is moderately correlated with Flow. The occurrence of multicollinearity which may pose a problem for linear models was also verified. The VIF value for each predictor is between 1 and 2, which indicates relatively weak multicollinearity compared to the critical values reported in the literature (Hair et al., 1998; Neter et al., 1989).

3.5.2 Performance Evaluation for Model Selection

We perform a RFE, with 10-fold cross validation, on 80% of data to choose the select the best feature subset. The remaining 20% of the data (test data) is used to determine the performance of each proposed model by comparing several criteria such as Root Mean Square Error (RMSE), coefficient of determination (R^2), relative Root Mean Square Error (rRMSE), the Nash-Sutcliff criterion (Nash) and relative mean Bias (rBiais), as shown in Table 3.4.

Table 3.4 Test performance results for Grande-Rivière data after selecting best features subset using RFE

Models	Selected features	RMSE (°C)	R^2	rRMSE (%)	Nash	rBias (%)
MLR	Tair, Tidal range	1.60	0.74	16.68	0.74	-1,469
ANN	Tair, WSPD, Tidal range Flow, NAO	1.46	0.79	14.67	0.79	-0.987
Random Forest	Tair, WSPD, Tidal range, Flow	1.17	0.86	11.80	0.86	-0.988

The results in Table 3.4 indicate that machine learning methods outperform the parametric statistical model (MLR). The ANN performs slightly better than the MLR with lower values of RMSE, rRMSE, rBias and higher values of Nash and R^2 than those of MLR. The RF performs better than ANN and MLR models, with lower values of RMSE, rRMSE, rBias and higher values of Nash and R^2 than those of ANN and MLR.

Adding more predictors in MLR does not improve the performance due to the non-linear relationship between the predictors and the target variable. In contrast, machine learning models have the ability to capture and model complex non-linear relationships.

Figure 3.3 shows that the RF has higher goodness of fit and better predictive capacity than ANN. This approach yields a rather robust performance, thanks to its low variance–low bias algorithm design.

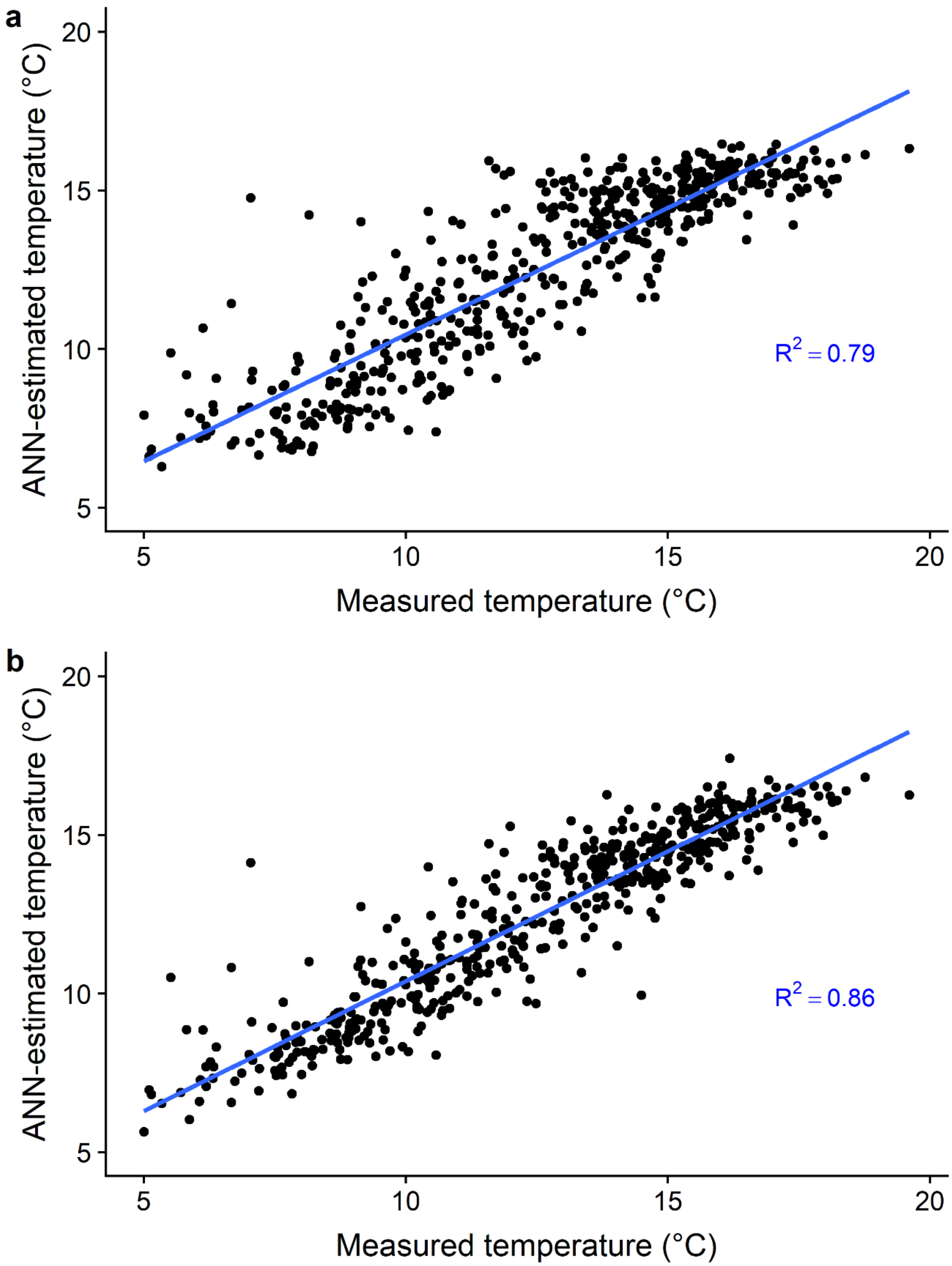


Figure 3.3 Performance of Artificial Neural Network (ANN) and Random Forest (RF) models for Grande-Rivière station obtained from tested dataset.

As demonstrated above, the RF was the best model for predicting the SST in the Grande-Rivière station. However, this result may be specific to this station. Hence, the procedure is repeated for the rest of the stations. For each station, we model the SST using the MLR, ANN and RF after selecting the best set of features for model inputs through the Recursive Feature Elimination. Then, we compute and compare their performance criteria. The models are then compared by verifying if the means of their RMSE values for all stations are significantly different. The one way ANOVA (DF=2, F-value=5,806) test result indicates a significant difference between the models at the 5% significance level (p-value=0.006). Then Tukey Honestly Significant Difference (Tukey HSD) test is computed between the different tested models in each station for multiple pairwise inter-comparisons. The Tukey HSD shows that there is no significant difference between ANN and MLR (p-value=0.89). In contrast, there is a significant difference between the RF model and both MLR (p-value=0.008) and ANN (p-value=0.026) models at the 5% significance level.

Table 3.5 presents the RMSEs of RF and ANN for testing datasets at all stations. The results show that RF is the best performing model for all stations with RMSE ranging between 0.77 °C and 1.8 °C, and an overall RMSE of 1.3 °C across all the stations against 1.7 °C for ANN.

Table 3.5 The RMSE performance results of all the stations computed in the test datasets

Stations	RMSE (°C)	
	Models	
	ANN	RF
Rimouski	1.14	0.99
Courant-Gaspé	1.37	1.37
Gyre Anticosti	1.26	0.77
Montlouis	1.32	1.10
Romaine	1.35	1.06
Sept-Îles	1.96	1.18
Rivière aux Tonnere	2.01	1.54
Natashquan	2.03	1.81
Tarbatière	2.07	1.51
Havre-St-Pierre	2.31	1.80
Ile-Shag	2.02	1.21
Blanc-Sablon	2.26	1.72
Borden	1.73	1.42
Shediac-Valley	1.06	0.77
Grande-Rivière	1.46	1.17
Mean	**1.69**	**1.29**

The best subset of features for RF differs from station to station but includes practically all the tested predictors, as indicated in Table 3.6 below.

Table 3.6 The performance results of the RF model in modelling the daily SST using the best subset features at each station

Stations	Subset features	RMSE (°C)	Rsquared	rRMSE(%)	Nash	rBiais(%)
Rimouski	Tair, Flow, WSPD	0.99	0.87	12.91	0.87	-1.32
Courant-Gaspé	Tair, WSPD, flow	1.37	0.84	16.94	0.84	-1.58
Gyre-Anticosti	Tair, NAO, WSPD, Flow	0.77	0.95	7.82	0.86	-0.54
Montlouis	Tair, WSPD, flow	1.10	0.90	13.85	0.85	-1.56
Sept-Îles	Tair, NAO, Tidal range, Flow	1.18	0.90	13.89	0.70	-2.11
Rivière-aux-Tonnere	Tair, Flow, WSPD, Tidal range	1.54	0.78	22.76	0.63	-5.68
Havre-St-Pierre	Tair, WSPD, , Flow, NAO	1.80	0.62	26.47	0.36	-7.61
Natashquan	Tair, Flow, WSPD	1.81	0.77	22.52	0.71	-4.30
Romaine	Tair, WSPD, Flow, Tidal range	1.06	0.89	17.40	0.82	-2.25
La Tarbatière	Tair, WSPD, Tidal range, Flow, NAO	1.51	0.75	29.49	0.52	-5.08
Blanc-Sablon	Tair, Tidal range, Flow, WSPD	1.72	0.67	26.40	0.32	-7.87
Grande-Rivière	Tair, WSPD, Tidal range, Flow	1.17	0.86	11.8	0.86	-0.99
Shediac-Valley	Tair, WSPD, Flow	0.77	0.95	6.41	0.89	-0.33
Ile-Shag	Tair, Flow, WSPD, NAO	1.21	0.89	9.78	0.68	-1.92
Borden	Tair, WSPD, Flow, Tidal range	1.42	0.80	15.27	0.69	-1.41

3.6 Discussion and conclusions

This study shows the usefulness of predictors other than air temperature in modelling SST, including the climatological (WSPD), oceanographic (Tidal range), hydrologic (Flow) variables and teleconnection pattern (NAO). The NAO is one of the most prominent teleconnection in the Northern Hemisphere during wintertime (Hurrell et al., 2003; Team, 2015; Yu et al., 2019). It explains one third of the sea level pressure variance in the northern hemisphere during winter (Cochran et al., 2019; Feldstein, 2003; Team, 2008). Thus, many studies have focused on the NAO impact during the winter (Benedict et al., 2004; Feldstein, 2003; Franzke et al., 2004; Hurrell, 1995). Very few studies have shown that the NAO effect also occurs during the summer (Barnston and Livezey, 1987; Portis et al., 2001) by affecting the SST variation at some stations.

We focus on determining the most accurate model for predicting the SST in the St. Lawrence Estuary and the GSL. RF was shown to outperform ANN, which is often used in similar studies. Generally, the ANN tends to have high variance in the validation stage due to the overfitting of training data, but a lower bias. Several studies on statistical water temperature modelling, particularly in rivers, show performance (RMSE) between 1 °C and 3 °C (DeWeber and Wagner, 2014; Hadzima-Nyarko et al., 2014; Piotrowski et al., 2015; Zhu et al., 2018). Thus, the performance of this first model is deemed acceptable given the complexity of the coastal environment, but could be improved with additional relevant predictors. The best feature subset selected for the RF model differs among stations. This difference may be due to the different climatological and physical (oceanographic) conditions associated with each station. Overall, this study demonstrates the ability of the RF model to accurately predict the SST, given the current variables. In the context of climate change, implementing RF in producing future water temperature scenarios could be done by using projections of the predictor variables produced by climate models under different greenhouse gas emission scenarios.

In future studies, we plan to generate future water temperature scenarios for optimistic and pessimistic climate change scenarios for each station, using our best model with its best subset of predictors. Then, we plan to map the future potential risk areas of *Vibrio* growth using selected risk indicators in relation to the exceedance of water temperature threshold favorable to *Vibrio* growth, i.e. the number of days exceeding the selected temperature thresholds.

3.7 Acknowledgements

The authors thank OURANOS, a climate-science consortium based in Quebec, for funding this research.

The authors would like to thank Pr. Yves Gratton for helping to find data sources.

4 REFERENCES

Abhishek, A., Lee, J.-Y., Keener, T.C., Yang, Y.J., 2010. Long-Term Wind Speed Variations for Three Midwestern U.S. Cities. Journal of the Air & Waste Management Association 60 (9) 1057-1064.

Alizadeh-Choobari, O., Adibi, P., 2019. Impacts of large-scale teleconnections on climate variability over Southwest Asia. Dynamics of Atmospheres and Oceans 86 41-51.

Baker-Austin, C., Stockley, L., Rangdale, R., Martinez-Urtaza, J., 2010. Environmental occurrence and clinical impact of Vibrio vulnificus and Vibrio parahaemolyticus: a European perspective. Environ Microbiol Rep 2(1) 7-18.

Baker-Austin, C., Trinanes, J., Gonzalez-Escalona, N., Martinez-Urtaza, J., 2017. Non-Cholera Vibrios: The Microbial Barometer of Climate Change. Trends Microbiol 25 (1) 76-84.

Baker-Austin, C., Trinanes, J.A., Taylor, N.G., Hartnell, R., Siitonen, A., Martinez-Urtaza, J., 2013. Emerging Vibrio risk at high latitudes in response to ocean warming. Nature Climate Change 3 (1) 73.

Baker-Austin, C., Trinanes, J.A., Taylor, N.G.H., Hartnell, R., Siitonen, A., Martinez-Urtaza, J., 2012. Emerging Vibrio risk at high latitudes in response to ocean warming. Nature Climate Change 3 (1) 73-77.

Barnston, A.G., Livezey, R.E., 1987. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. Monthly weather review 115 (6) 1083-1126.

Bélanger, M., El-Jabi, N., Caissie, D., Ashkar, F., Ribi, J.M., 2005. Estimation de la température de l'eau de rivière en utilisant les réseaux de neurones et la régression linéaire multiple. Revue des sciences de l'eau 18(3).

Bell, G.D., Halpert, M.S., Kousky, V.E., Gelman, M.E., Ropelewski, C.F., Douglas, A.V., Schnell, R.C., 1999. Climate Assessment for 1998. Bulletin of the American Meteorological Society 80 (5) 1040-1040.

Belsley, D.A., 1993. Conditioning Diagnostics Collinearity and Weak Data in Regression. JOURNAL-OPERATIONAL RESEARCH SOCIETY 44 88-88.

Benedict, J., Lee, S., Feldstein, S., 2004. Synoptic View of the North Atlantic Oscillation. J. Atmos. Sci. 61 121-144.

Berrar, D., 2019. Cross-Validation, In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press: Oxford, pp. 542-545.

Bonsal, B., Shabbar, A., 2008. Impacts of large-scale circulation variability on low streamflows over Canada: a review. Canadian Water Resources Journal 33 (2) 137-154.

Bonsal, B., Shabbar, A., 2011. Large-scale climate oscillations influencing Canada, 1900-2008. Canadian Councils of Resource Ministers.

Breiman, L., 2001. Random forests. Machine learning 45 (1) 5-32.

Burge, C.A., Mark Eakin, C., Friedman, C.S., Froelich, B., Hershberger, P.K., Hofmann, E.E., Petes, L.E., Prager, K.C., Weil, E., Willis, B.L., Ford, S.E., Harvell, C.D., 2014. Climate change influences on marine infectious diseases: implications for management and society. Ann Rev Mar Sci 6 249-277.

Caissie, D., El-Jabi, N., St-Hilaire, A., 1998. Stochastic modelling of water temperatures in a small stream using air to water relations. Canadian Journal of Civil Engineering 25 (2) 250-260.

Canadian Science Advisory Secretariat, 2017. Oceanographic Conditions in the Atlantic Zone in 2016. Fisheries and Oceans Canada, Centre for Science Advice.

Centers for Disease, C., Prevention, 2013. Incidence and trends of infection with pathogens transmitted commonly through food - foodborne diseases active surveillance network, 10 U.S. sites, 1996-2012. MMWR. Morbidity and mortality weekly report 62 (15) 283-287.

Centers for Disease, C.P., 2006. Preliminary FoodNet data on the incidence of infection with pathogens transmitted commonly through food--10 States, United States, 2005. MMWR Morb Mortal Wkly Rep 55 (14) 392-395.

Chenard, J.-F., Caissie, D., 2008. Stream temperature modelling using artificial neural networks: application on Catamaran Brook, New Brunswick, Canada. Hydrological Processes 22 (17) 3361-3372.

Chu, C., Do, Y., Kim, Y., Saito, Y., Lee, S.-D., Park, H., Lee, J.-K., 2011. Mathematical modeling of Vibrio vulnificus infection in Korea and the influence of global warming. Osong public health and research perspectives 2 (1) 51-58.

Cochran, J.K., Bokuniewicz, H.J., Yager, P.L., 2019. Encyclopedia of Ocean Sciences. Elsevier Science.

Davis, B.J.K., Jacobs, J.M., Davis, M.F., Schwab, K.J., DePaola, A., Curriero, F.C., 2017. Environmental determinants of Vibrio parahaemolyticus in the Chesapeake Bay. Appl Environ Microbiol.

Dechet, A.M., Yu, P.A., Koram, N., Painter, J., 2008. Nonfoodborne Vibrio infections: an important cause of morbidity and mortality in the United States, 1997-2006. Clin Infect Dis 46 (7) 970-976.

Deeb, R., Tufford, D., Scott, G.I., Moore, J.G., Dow, K., 2018. Impact of Climate Change on Vibrio vulnificus Abundance and Exposure Risk. Estuaries and Coasts 41 (8) 2289-2303.

Deser, C., Alexander, M.A., Xie, S.-P., Phillips, A.S., 2010. Sea surface temperature variability: Patterns and mechanisms. Ann Rev Mar Sci 2 115-143.

DeWeber, J.T., Wagner, T., 2014. A regional neural network ensemble for predicting mean daily river water temperature. Journal of hydrology 517,187-200.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1) 27-46.

Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern classification. John Wiley & Sons.

El-Sabh, M.I., Murty, T.S., 1990. Mathematical modelling of tides in the St. Lawrence Estuary, Oceanography of a Large-Scale Estuarine System. Springer, pp. 10-50.

Fan, K., Wang, H., 2004. Antarctic oscillation and the dust weather frequency in North China. Geophysical Research Letters 31 (10).

Feldstein, S.B., 2003. The dynamics of NAO teleconnection pattern growth and decay. Quarterly Journal of the Royal Meteorological Society 129 (589) 901-924.

Franzke, C., Lee, S., Feldstein, S., 2004. Is the North Atlantic Oscillation a Breaking Wave. J. Atmos. Sci. 61 145-160.

Freund, R.J., Wilson, W.J., Mohr, D.L., 2010. CHAPTER 8 - Multiple Regression, In: Freund, R.J., Wilson, W.J., Mohr, D.L. (Eds.), Statistical Methods (Third Edition). Academic Press: Boston, pp. 375-471.

Galbraith, P.S., Chassé, J., Gilbert, D., Larouche, P., Brickman, D., Pettigrew, B., Devine, L., Gosselin, A., Pettipas, R., Lafleur, C., 2017. Physical oceanographic conditions in the Gulf of St. Lawrence in 2016. Canadian Science Advisory Secretariat.

Galbraith, P.S., Larouche, P., Chassé, J., Petrie, B., 2012. Sea-surface temperature in relation to air temperature in the Gulf of St. Lawrence: Interdecadal variability and long term trends. Deep Sea Research Part II: Topical Studies in Oceanography 77-80 10-20.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of machine learning research 3 (Mar) 1157-1182.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene Selection for Cancer Classification using Support Vector Machines. Machine learning 46 (1) 389-422.

Hadzima-Nyarko, M., Rabi, A., Šperac, M., 2014. Implementation of Artificial Neural Networks in Modeling the Water-Air Temperature Relationship of the River Drava. Water Resources Management 28 (5) 1379-1394.

Hair, J.F., Anderson, R.E., Tatham, R.L., William, C., 1998. Black(1998), Multivariate data analysis. Upper Saddle River, NJ: Prentice Hall.

Hall, A., Visbeck, M., 2002. Synchronous Variability in the Southern Hemisphere Atmosphere, Sea Ice, and Ocean Resulting from the Annular Mode*. Journal of Climate 15 3043-3057.

Harvell, C.D., Mitchell, C.E., Ward, J.R., Altizer, S., Dobson, A.P., Ostfeld, R.S., Samuel, M.D., 2002. Climate Warming and Disease Risks for Terrestrial and Marine Biota. Science 296 (5576) 2158-2162.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Haykin, S., 1994. Neural networks: a comprehensive foundation. Prentice Hall PTR.

Huang, C., Qiao, F., 2009. The relationship between sea surface temperature anomaly and wind energy input in the Pacific Ocean. Progress in Natural Science 19 (10) 1409-1412.

Hurrell, J.W., 1995. Decadal trends in the north atlantic oscillation: regional temperatures and precipitation. Science 269 (5224) 676-679.

Hurrell, J.W., Kushnir, Y., Ottersen, G., Visbeck, M., 2003. An overview of the North Atlantic oscillation. Geophysical Monograph-American Geophysical Union 134 1-36.

Jacobs, J., Moore, S.K., Kunkel, K.E., Sun, L., 2015. A framework for examining climate-driven changes to the seasonality and geographical range of coastal pathogens and harmful algae. Climate Risk Management 8 16-27.

Jacobs, J.M., Rhodes, M., Brown, C.W., Hood, R.R., Leight, A., Long, W., Wood, R., 2014. Modeling and forecasting the distribution of Vibrio vulnificus in Chesapeake Bay. J Appl Microbiol 117 (5) 1312-1327.

Jeong, D.I., Daigle, A., St-Hilaire, A., 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 (7) 805-821.

Jian-Qi, S., 2010. Possible Impact of the Boreal Spring Antarctic Oscillation on the North American Summer Monsoon. Atmospheric and Oceanic Science Letters 3 (4) 232-236.

Kaspar, C.W., Tamplin, M.L., 1993. Effects of temperature and salinity on the survival of Vibrio vulnificus in seawater and shellfish. Appl Environ Microbiol 59(8) 2425-2429.

Kim, D.-M., Jung, S.-I., Jang, H.-C., Lee, C.S., Lee, S.H., Yun, N.R., Neupane, G.P., Park, K.-H., 2011. Vibrio vulnificus DNA Load and Mortality. Journal of Clinical Microbiology 49 (1) 413-415.

Kim, T.-W., Cho, Y.-K., You, K.-W., Jung, K.T., 2010. Effect of tidal flat on seawater temperature variation in the southwest coast of Korea. Journal of Geophysical Research: Oceans 115 (C2).

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97 (1) 273-324.

Koutitonsky, V., 1991. The physical oceanography of the Gulf of St. Lawrence: a review with emphasis on the synoptic variability of the motion. The Gulf of St. Lawrence: small ocean or big estuary?

Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of cheminformatics 6 (1) 10.

Kuhn, M., 2008. Building predictive models in R using the caret package. Journal of statistical software 28 (5) 1-26.

Kuhn, M., 2012. Variable selection using the caret package. URL http://cran. cermin. lipi. go. id/web/packages/caret/vignettes/caretSelection. pdf.

Li, B., Yang, G., Wan, R., Dai, X., Zhang, Y., 2016. Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China. Hydrology Research 47 (S1) 69-83.

Long, Z., Perrie, W., Chassé, J., Brickman, D., Guo, L., Drozdowski, A., Hu, H., 2015. Impacts of Climate Change in the Gulf of St. Lawrence. Atmosphere-Ocean 54 (3) 337-351.

Lowe, R.J., Pivan, X., Falter, J., Symonds, G., Gruber, R., 2016. Rising sea levels will reduce extreme temperature variations in tide-dominated reef habitats. Science advances 2 (8) e1600825.

Martinez-Urtaza, J., Bowers, J.C., Trinanes, J., DePaola, A., 2010. Climate anomalies and the increasing risk of Vibrio parahaemolyticus and Vibrio vulnificus illnesses. Food Research International 43 (7) 1780-1790.

Martinez-Urtaza, J., Huapaya, B., Gavilan, R.G., Blanco-Abad, V., Ansede-Bermejo, J., Cadarso-Suarez, C., Figueiras, A., Trinanes, J., 2008. Emergence of Asiatic Vibrio Diseases in South America in Phase With El Niño. Epidemiology 19 (6) 829-837.

Materia, S., Gualdi, S., Navarra, A., Terray, L., 2012. The effect of Congo River freshwater discharge on Eastern Equatorial Atlantic climate variability. Climate Dynamics 39.

McLaughlin, J.B., DePaola, A., Bopp, C.A., Martinek, K.A., Napolilli, N.P., Allison, C.G., Murray, S.L., Thompson, E.C., Bird, M.M., Middaugh, J.P., 2005. Outbreak of Vibrio parahaemolyticus gastroenteritis associated with Alaskan oysters. New England Journal of Medicine 353 (14) 1463-1470.

Motes, M.L., DePaola, A., 1996. Offshore suspension relaying to reduce levels of Vibrio vulnificus in oysters (Crassostrea virginica). Appl Environ Microbiol 62(10) 3875-3877.

Motes, M.L., DePaola, A., Cook, D.W., Veazey, J.E., Hunsucker, J.C., Garthright, W.E., Blodgett, R.J., Chirtel, S.J., 1998. Influence of Water Temperature and Salinity on Vibrio vulnificus in Northern Gulf and Atlantic Coast Oysters (Crassostrea virginica). Appl Environ Microbiol 64(4) 1459-1465.

Naing, W.Y.N., Htike, Z.Z., 2015. Forecasting of monthly temperature variations using random forests. APRN J. Eng. Appl. Sci 10 10109-10112.

Nan, S., Li, J., 2003. The relationship between the summer precipitation in the Yangtze River valley and the boreal spring Southern Hemisphere annular mode. Geophysical Research Letters 30 (24).

National Weather Service Climate Prediction Center, 2008. Teleconnection Introduction, availaible at https://www.cpc.ncep.noaa.gov/data/teledoc/teleintro.shtml.

Neter, J., Wasserman, W., Kutner, M.H., 1989. Applied linear regression models.

Newton, A., Kendall, M., Vugia, D.J., Henao, O.L., Mahon, B.E., 2012. Increasing rates of vibriosis in the United States, 1996-2010: review of surveillance data from 2 systems. Clin Infect Dis 54 Suppl 5 S391-395.

Ng, H.G., MatJafri, M.Z., Abdullah, K., Wong, C.J., 2009. The effect of wind speed on SST retrieval, 2009 IEEE Aerospace conference, pp. 1-8.

Obringer, R., Nateghi, R., 2018. Predicting urban reservoir levels using statistical learning techniques. Sci Rep 8 (1) 5164.

Oliver, J.D., 2005. Wound infections caused by Vibrio vulnificus and other marine bacteria. Epidemiology and Infection 133 (3) 383-391.

Panchal, F.S., Panchal, M., 2014. Review on methods of selecting number of hidden nodes in artificial neural network. International Journal of Computer Science and Mobile Computing 3 (11) 455-464.

Pang, B., Yue, J., Zhao, G., Xu, Z., 2017. Statistical Downscaling of Temperature with the Random Forest Model. Advances in Meteorology 2017 1-11.

Perlwitz, J., Knutson, T., Kossin, J.P., LeGrande, A.N., 2017. Large-scale circulation and climate variability, In: Wuebbles, D.J., Fahey, D.W., Hibbard, K.A., Dokken, D.J., Stewart, B.C., Maycock, T.K. (Eds.), Climate Science Special Report: Fourth National Climate Assessment, Volume I. U.S. Global Change Research Program: Washington, DC, USA, pp. 161-184.

Pettigrew, B., Hilbert, D., Desmarais, R., 2016. Thermograph network in the Gulf of St. Lawrence. Fisheries and Oceans Canada= Pêches et océans Canada.

Piotrowski, A.P., Napiorkowski, M.J., Napiorkowski, J.J., Osuch, M., 2015. Comparing various artificial neural network types for water temperature prediction in rivers. Journal of hydrology 529,302-315.

Portis, D.H., Walsh, J.E., El Hamly, M., Lamb, P.J., 2001. Seasonality of the North Atlantic oscillation. Journal of Climate 14 (9) 2069-2078.

Pozo-Vázquez, D., Esteban-Parra, M.J., Rodrigo, F.S., Castro-Díez, Y., 2001. The Association between ENSO and Winter Atmospheric Circulation and Temperature in the North Atlantic Region. Journal of Climate 14 (16) 3408-3420.

Qu, B., Gabric, A.J., Zhu, J.-n., Lin, D.-r., Qian, F., Zhao, M., 2012. Correlation between sea surface temperature and wind speed in Greenland Sea and their relationships with NAO variability. Water Science and Engineering 5 (3) 304-315.

Riffenburgh, B., 2007. Encyclopedia of the Antarctic. Routledge.

Ripley, B., Venables, W., Ripley, M.B., 2016. Package 'nnet'. R package version 7 3-12.

Ropelewski, C.F., Arkin, P.A., 2019. Climate Analysis. Cambridge University Press.

Saucier, F.J., 2003. Modeling the formation and circulation processes of water masses and sea ice in the Gulf of St. Lawrence, Canada. Journal of Geophysical Research 108 (C8).

Seager, R., Kushnir, Y., Nakamura, J., Ting, M., Naik, N., 2010. Northern Hemisphere winter snow anomalies: ENSO, NAO and the winter of 2009/10. Geophysical Research Letters 37 (14).

Semenza, J.C., Trinanes, J., Lohr, W., Sudre, B., Lofdahl, M., Martinez-Urtaza, J., Nichols, G.L., Rocklov, J., 2017. Environmental Suitability of Vibrio Infections in a Warming Climate: An Early Warning System. Environ Health Perspect 125(10) 107004.

Shabbar, A., 2006. The impact of El Niño-Southern Oscillation on the Canadian climate. Advances in Geosciences 6 149-153.

Shmueli, G., 2010. To explain or to predict? Statistical science 25 (3) 289-310.

Silvestri, G., Vera, C., 2003. Antarctic Oscillation signal on precipitation anomalies over southeastern South America. Geophys. Res. Lett 30.

Team, B.A., 2008. Assessment of Climate Change for the Baltic Sea Basin. Springer Berlin Heidelberg.

Team, R.C., 2013. R: A language and environment for statistical computing.

Team, T.B.I.I.A., 2015. Second Assessment of Climate Change for the Baltic Sea Basin. Springer International Publishing.

Thiombiano, A.N., El Adlouni, S., St-Hilaire, A., Ouarda, T.B.M.J., El-Jabi, N., 2017. Nonstationary frequency analysis of extreme daily precipitation amounts in Southeastern Canada using a peaks-over-threshold approach. Theoretical and Applied Climatology 129 (1) 413-426.

Thompson, D.W., Wallace, J.M., 1998. The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. Geophysical Research Letters 25 (9) 1297-1300.

United States Food and Drug Administration (FDA), 2005. Quantitative Risk Assessment on the Public Health Impact of Pathogenic Vibrio parahaemolyticus.

Vezzulli, L., Colwell, R.R., Pruzzo, C., 2013. Ocean warming and spread of pathogenic vibrios in the aquatic environment. Microb Ecol 65 (4) 817-825.

Vezzulli, L., Grande, C., Reid, P.C., Helaouet, P., Edwards, M., Hofle, M.G., Brettar, I., Colwell, R.R., Pruzzo, C., 2016. Climate influence on Vibrio and associated human diseases during the past half-century in the coastal North Atlantic. Proc Natl Acad Sci U S A 113 (34) E5062-5071.

Wang, C., Weisberg, R.H., Yang, H., 1999. Effects of the wind speed–evaporation–SST feedback on the El Niño–Southern Oscillation. Journal of the Atmospheric Sciences 56 (10) 1391-1403.

Wang, H., Fan, K., 2007. Relationship between the Antarctic Oscillation in the western North Pacific typhoon frequency. Chinese Science Bulletin 52 (4) 561-565.

Weber, K., Sturmer, L., Hoover, E., Baker, S., 2013. The role of water temperature in hard clam aquaculture. University of Florida IFAS extension, Gainesville.

WHO, F., 2005. Risk assessment of Vibrio vulnificus in raw oysters: interpretative summary and technical report, microbiological risk assessment series 8: Geneva: World Health Organization. Rome: Food And Agriculture Organization of the United Nations.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

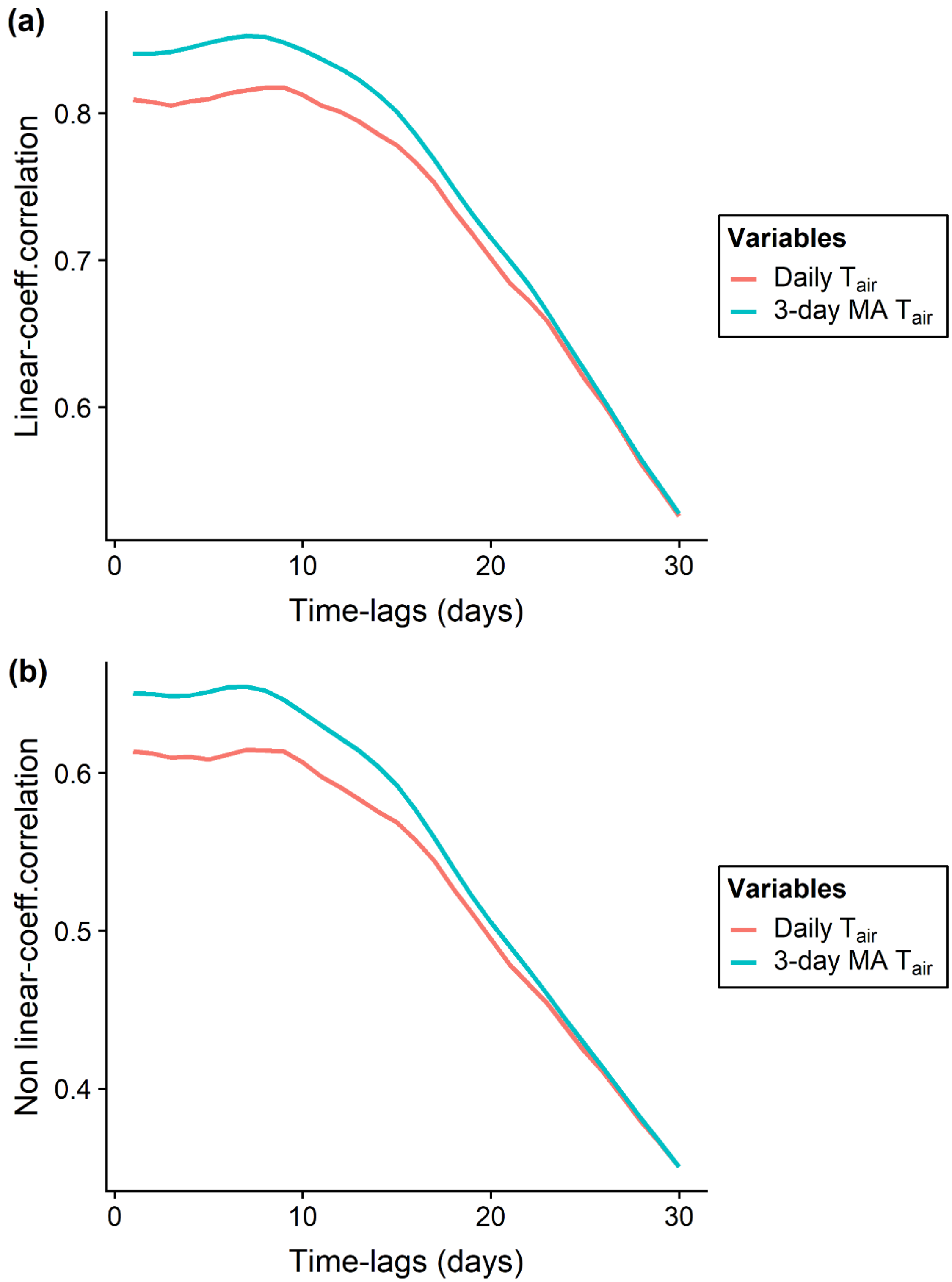
Yu, B., Lin, H., Soulard, N., 2019. A Comparison of North American Surface Temperature and Temperature Extreme Anomalies in Association with Various Atmospheric Teleconnection Patterns. Atmosphere 10,172.

Zhu, S., Nyarko, E.K., Hadzima-Nyarko, M., 2018. Modelling daily water temperature from air temperature for the Missouri River. PeerJ 6 e4894.

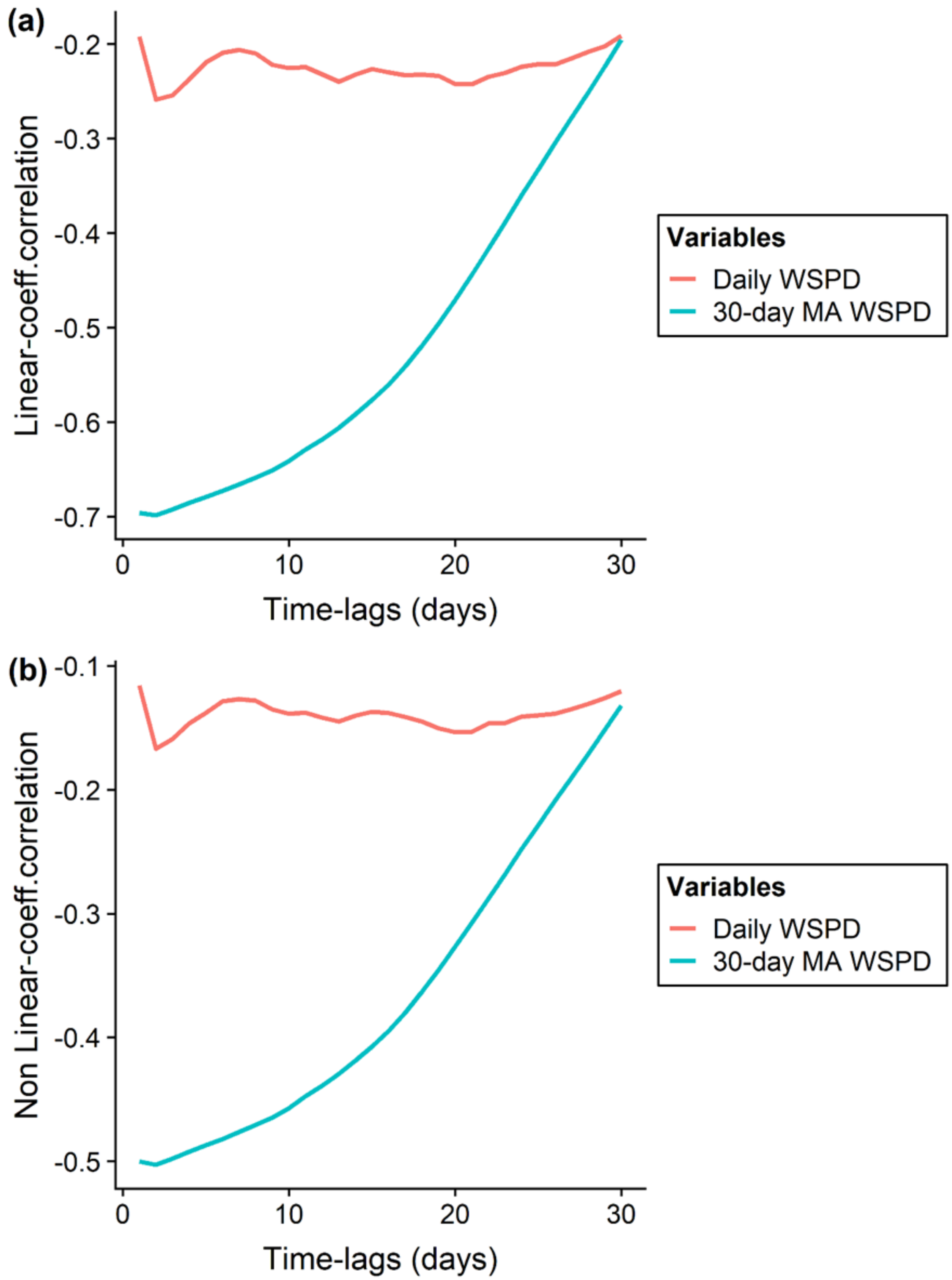
Zimmerman, A.M., DePaola, A., Bowers, J.C., Krantz, J.A., Nordstrom, J.L., Johnson, C.N., Grimes, D.J., 2007. Variability of total and pathogenic Vibrio parahaemolyticus densities in northern Gulf of Mexico water and oysters. Appl Environ Microbiol 73 (23) 7589-7596.

Zuur, A.F., Ieno, E.N., Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. Methods in ecology and evolution 1 (1) 3-14.

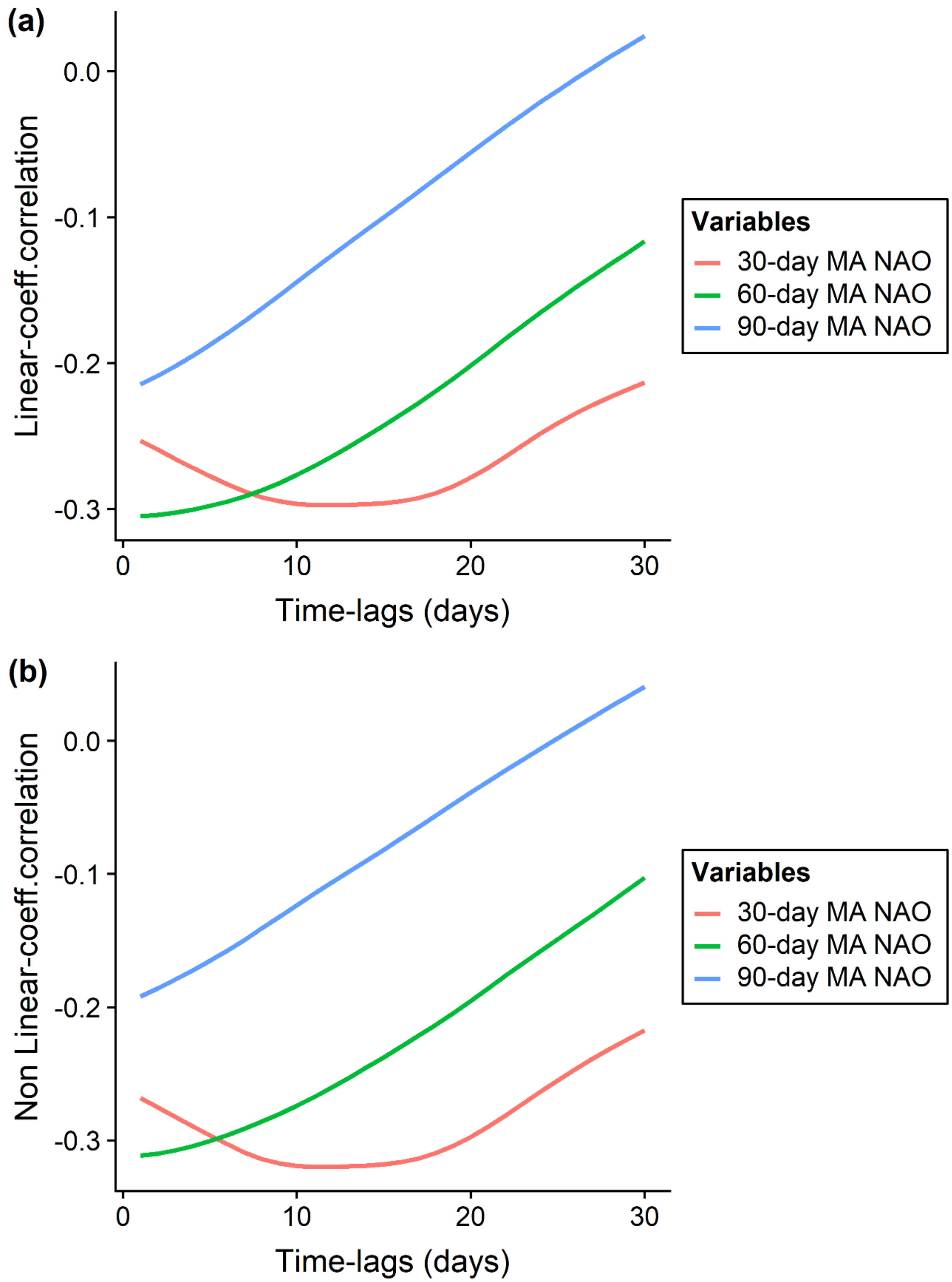
5 APPENDIX I



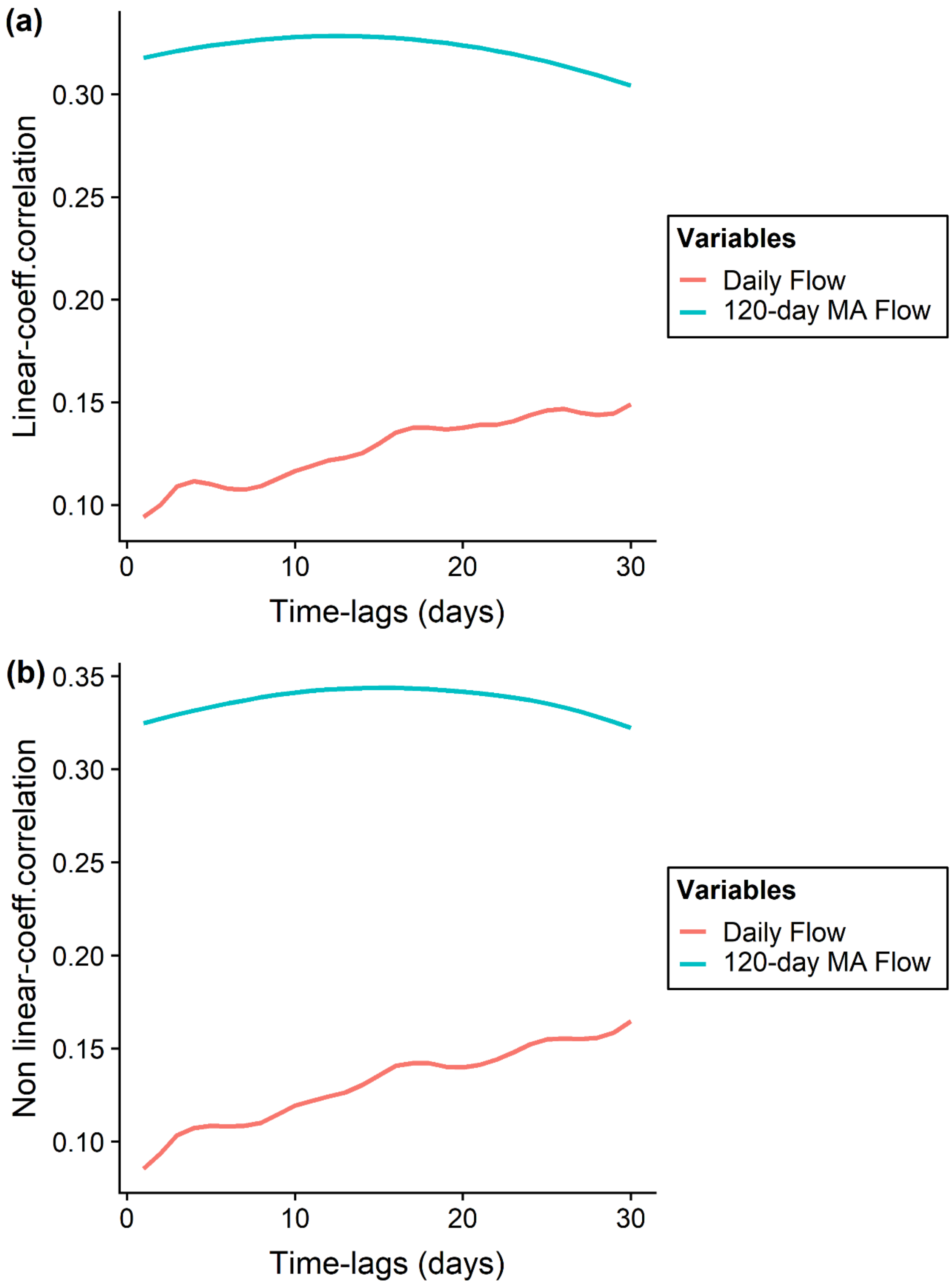
Appendix I.1 Example of the linear (Pearson) and nonlinear (Spearman) correlations between the daily SST and a) the daily mean and b) the 3-day Moving Average (3-day MA) of air temperature for different time lags at the Grande-Rivière station.



Appendix I. 2 Examples of the linear (Pearson) and nonlinear (Spearman) correlations between the daily SST and a) the daily mean and b) the 30-day Moving Average (30-day MA) of wind speed for different time lags at the Grande-Rivière station.

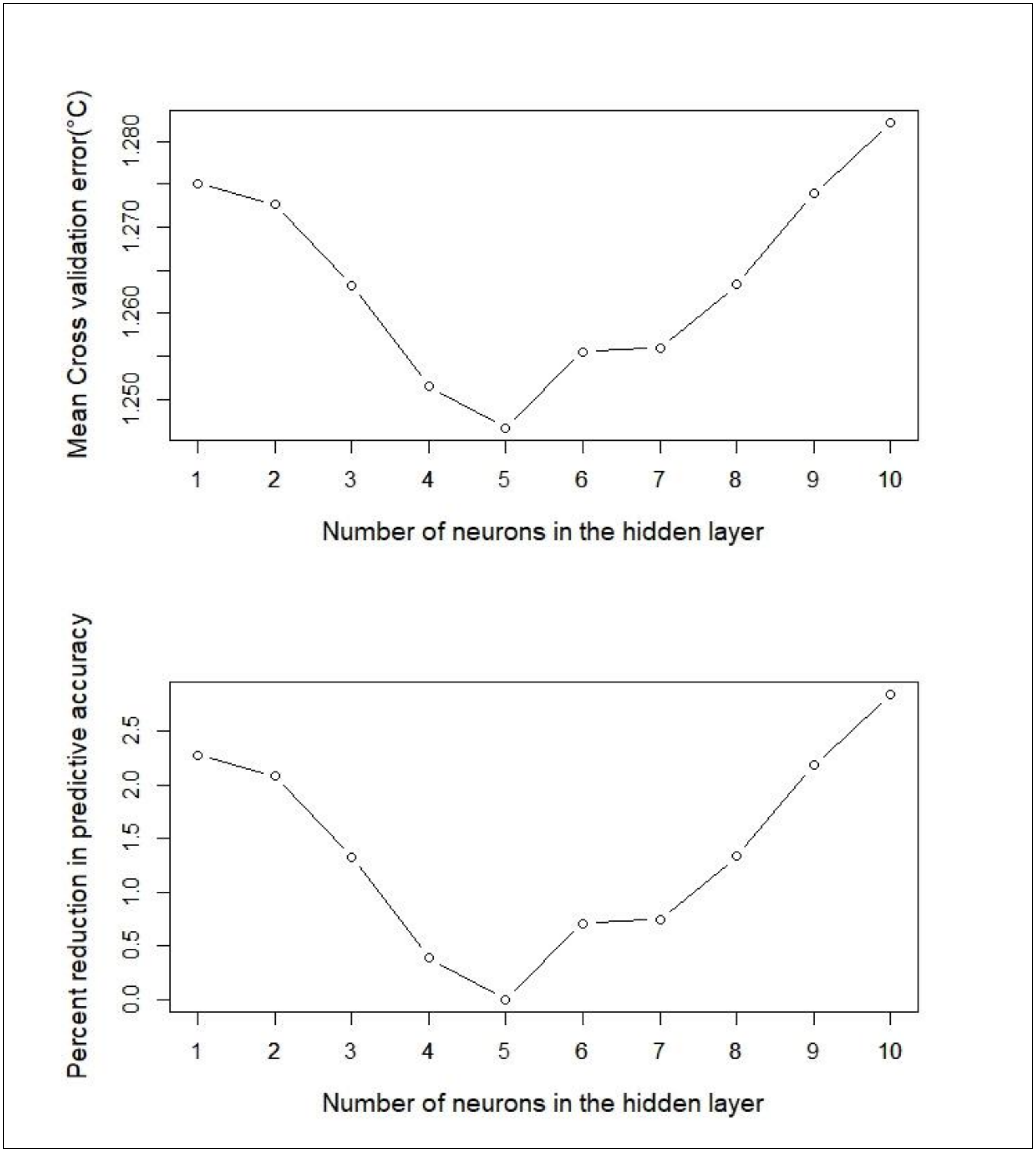


Appendix I. 3 Example of the linear (Pearson) and nonlinear (Spearman) correlations between the daily SST and the (30, 60, 90)-day MA of daily NAO for different time lags at the Rimouski station.



Appendix I. 4 Example of the linear (Pearson) and nonlinear (Spearman) correlations between the daily SST and a) the daily mean and b) the 120-day Moving Average (120-day MA) of St. Lawrence flow for different time lags at the Rimouski station.

6 APPENDIX II



Appendix II. 1 Selection of the optimal number of hidden neurons as a function of its contribution in predictive accuracy at the Courant-Gaspé station.

7 ARTICLE 2: IMPACT OF THE FUTURE COASTAL WATER TEMPERATURE SCENARIOS ON THE RISK OF POTENTIAL GROWTH OF PATHOGENIC VIBRIO MARINE BACTERIA

Title of the article: **Impact of the future coastal water temperature scenarios on the risk of potential growth of pathogenic Vibrio marine Bacteria**

Titre de l'article: **L'impact des futures températures de l'eau côtière sur le risque d'une croissance potentielle des bactéries marines pathogènes du genre *Vibrio***

Auteurs et affiliations :

Habiba Ferchichi[1,*], André St-Hilaire[1, 2], Taha B.M.J Ouarda[1], Benoît Levesque[3]

1. INRS-ETE. 490 De la Couronne St, Quebec City, QC, Canada
2. Canadian Rivers Institute, Fredericton, NB, Canada
3. INSPQ, Quebec City, Canada

*. Corresponding author

This manuscript was submitted to **Estuarine, Coastal and Shelf Science** journal.

14th February 2020

Contributions des auteurs :

Les coauteurs A.S., T.B.M.J.O. et B.L. ont défini la méthodologie à suivre et révisé le manuscrit. L'auteur principal H.F. a réalisé l'analyse, l'interprétation des résultats et la préparation des figures sous la supervision des coauteurs.

Lien entre les articles :

Le deuxième manuscrit implique les variables les plus pertinentes (Tair et vitesse du vent) et les deux modèles les plus performants (ANN et RF), déduits à partir le premier manuscrit, dans la modélisation de la température de l'eau. Après la sélection du meilleur modèle d'apprentissage automatique ainsi que ses intrants, on produit les scénarios futurs des températures de l'eau. En utilisant ces scénarios, on déduit des indicateurs de risque en fonction des seuils de température de l'eau en relation avec la croissance des *Vibrios* pathogènes. Suite à la cartographie d'indicateur de risque, ce manuscrit va présenter la situation future des zones côtières de l'estuaire et golfe du Saint-Laurent vu le risque de la croissance potentielle des *Vibrios* pathogènes.

7.1 Abstract

Vibrio (V), a genus of marine bacteria, are common inhabitants of warm coastal waters and estuaries. Vibrio includes V. parahaemolyticus and V. vulnificus species that can cause human infections through the consumption of contaminated shellfish (as bivalve molluscs). The growth of pathogenic Vibrio is related to ambient water temperature and seems to increase at 15 °C and over. The expansion of Vibrio infection outbreak is increasing worldwide due to the increase of the sea surface temperature as a result of ocean warming. Canada's coast is not an exception to this worldwide Vibrio spread. Faced with this issue, this study focuses on modelling the future potential Vibrio growth risk along the coasts of the St. Lawrence Gulf and Estuary, where the shellfish industry is well developed. This is done using the adequate machine learning model with explanatory variables that include air temperature and wind speed for predicting future water temperatures. Based on the predicted future water temperature scenarios and a threshold of 15 °C to determine the conditions favorable to the growth of Vibrio bacteria, we modelled the Vibrio growth risk indicator, i.e. the number of days exceeding the minimum temperature for Vibrio pathogenic growth (15 °C), in the horizon 2040-2100. Simulations show that the number of days, where the minimum temperature (15 °C) will be reached, will increase spatially and even seasonally and all the shellfish beds would meet the temperature condition for Vibrio growth regardless of the climate scenario (optimistic and pessimistic).

Key words: Coastal Water temperature, Vibrio bacteria, machine learning models, modelling, prediction, climate change.

7.2 Abbreviations

ANN, Artificial Neural Networks; Bagging, Bootstrap Aggregating; CANOPA, CANadian Océan PArallélisé; CART, Classification and Regression Tree; CDC, Centers for Disease Control and Prevention; GHG, Greenhouse Gas GSL; Gulf of St. Lawrence; IDW, Inverse Distance Weighted; MA, Moving Average; MK, Mann-Kendall; MLP, Multilayer Perceptron; MMK, Modified Mann-Kendall; MSE, Mean Square Error; Nash, Nash-Sutcliffe coefficient; Ntry, The number of bootstrap input variables at each split of a tree; OOB, Out-Of-Bag; PEI, Prince Edward Island; ppt, parts per thousand; rBias, Relative mean bias; RCM, Regional Climate Model; RCP, Representative Concentration Pathways; RF, Random Forests; RFE, Recursive feature Elimination; RMSE, Root Mean Square Error; SST, Sea Surface Temperature; V., Vibrio.

7.3 Introduction

Vibrio (*V.*) *parahaem*olyticus and *Vibrio vulnificus* belong to the family of *Vibrionacea*, a group of aquatic microorganisms that includes other human pathogens such as *V. cholera.* They are natural inhabitants of warm coastal waters (>15 °C) and estuaries with low salinity (<25 ppt) (Baker-Austin et al., 2010; Heng et al., 2017; Kaspar and Tamplin, 1993; Motes et al., 1998; Vezzulli et al., 2013). Depending on the environmental conditions, these pathogenic bacteria could be present in high concentrations during summer.

*Vibrio parahaem*olyticus is recognized as a leading cause of gastroenteritis associated with seafood consumption worldwide (Martinez-Urtaza et al., 2010) and it was the cause of significant outbreaks of infections in North America. For example, in the USA between 1996 and 1997, four major outbreaks were reported to the Centers for Disease Control and Prevention (CDC) involving more than 700 illness cases due to *V. parahaem*olyticus infections (Su and Liu, 2007).

Vibrio vulnificus infections are less frequent. However, *V. vulnificus* is a lethal opportunistic human pathogen responsible for the majority of deaths related to seafood consumption worldwide. For instance, in the USA, more than 95% of seafood-related deaths are caused by this bacterium (Oliver, 2013). Consumption of raw or undercooked bivalve shellfish (oysters, mussels, clams, etc.) contaminated with *V. vulnificus* can lead to major infections such as septicemia, with subsequent highest mortality (sometimes exceeding 50%) than any foodborne pathogen (Dechet et al., 2008; Feldhusen, 2000; Oliver, 2005). In the USA, the CDC estimates an average of 100 foodborne infections associated to *V. vulnificus* annually*,* resulting in 50 fatalities per year (Mead et al., 1999).

The growth of pathogenic *Vibrio* species causing human illness is directly related to the exceedance of a threshold of water temperature (about 15 °C) (Baker-Austin et al., 2013; Jacobs et al., 2015; Martinez-Urtaza et al., 2010; McLaughlin et al., 2005). Because most of bivalves are filter feeders, *Vibrio* bacteria may concentrate in their tissues. When the water temperature exceeds a certain threshold, shellfish are more likely to be contaminated with *Vibrio*. These contaminated shellfish transmit, in turn, the *Vibrio* bacteria to humans through consumption of raw or undercooked shellfish (Baker-Austin et al., 2017; Davis et al., 2017; McLaughlin et al., 2005; Motes and DePaola, 1996; Zimmerman et al., 2007).

In order to evaluate and manage the *Vibrio* risk infection, various models have been developed. Among these models, some are related to the prediction of *Vibrio* concentration, which is based either on only sea surface temperature (SST) (Chu et al., 2011) or both SST and salinity (Jacobs et al., 2014; United States Food and Drug Administration (FDA), 2005). Another category includes models developed to explain the relation of *Vibrio* infections exposure in response to SST threshold exceedances (Semenza et al., 2017). Given the paucity of *Vibrio* concentration and *Vibrio* infection data, the aim of this study is evaluating the *Vibrio* growth risk through its relation with SST threshold exceedances.

Several reports and scientific researches show that the incidence of *Vibrio* infections has increased significantly worldwide (Centers for Disease and Prevention, 2013; Martinez-Urtaza et al., 2010; Newton et al., 2012). For instance, the CDC estimated an increase by 41% in the annual average of *Vibrio* infections in the USA between 1996 and 2005 (Centers for Disease, 2006). In addition to the increase of spatial spread worldwide, sudden *Vibrio* outbreaks had emerged in new temperate and even cold regions including Peru (Martinez-Urtaza et al., 2008), Europe (Baker-Austin et al., 2010), Chile (Narjol et al., 2005) and Alaska (McLaughlin et al., 2005). This unusual outbreak emergence of *Vibrio* infections coincides with water temperatures anomalies (Baker-Austin et al., 2017; Martinez-Urtaza et al., 2010). Many microbiologists agree that climate change may explain this increase of *Vibrio* infections spread worldwide as well as the likelihood of its geographical expansion in new areas (Baker-Austin et al., 2012; Deeb et al., 2018; Martinez-Urtaza et al., 2010; McLaughlin, 2005; Vezzulli et al., 2013; Vezzulli et al., 2016). They have even considered *Vibrio* pathogens as microbial barometer of climate change (Baker-Austin et al., 2017).

The global average land-ocean temperature has risen by approximately 0.85 °C since the late nineteenth century (IPCC, 2013). This increase SST, caused by atmospheric warming, is considered as the most severe and pervasive impact of climate change in marine environments, especially in coastal ecosystems (Burge et al., 2014; Halpern et al., 2008). Consequently, above 70% of the world’s coastline are significantly warming (Lima and Wethey, 2012).

The harvesting of molluscs is an important part of the Canadian economy. It is well developed in the provinces of Quebec, and Prince Edward Island (PEI). The shellfish beds are distributed over coastal zones of the Estuary and Gulf of St. Lawrence (GSL) including Rimouski, Gaspe, Baie des Chaleurs, the Quebec North Shore, Magdalen Islands and PEI. The GSL, located in eastern part of Canada, is one of the largest and most diverse marine ecosystems in the world, covering an area of 225,000 km^2. (Halpern et al., 2008; McLaughlin et al., 2005)

As the rest of worldwide marine ecosystems affected by ocean warming, the SST of GSL has increased by 1 to 1.5 °C during 1982-2011 by calculating the annual average of temperatures from May to November (Galbraith et al., 2012). The predicted SST in Eastern Canada, through climate scenarios projections, indicate a possible rise by more than one degree Celsius during the next century (Galbraith et al., 2012). Therefore, this increase of water temperature could lead to the proliferation of *Vibrio* pathogens as well as shellfish contamination and human infections.

In fact, a recent surveillance study on the diversity and dynamics of the *Vibrio* communities in Canada's coasts (British Columbia [Pacific Coast], Nova Scotia [Atlantic Coast] and Gaspe) highlights the emergence of *V. cholerae* in temperate Canadian estuaries and the detection of pathogenic strains of *V. parahaemolyticus* in bivalve molluscs harvested in Canada (clams, mussels and oysters) with increasing trend during the warmest months of 2006-2016 (Banerjee et al., 2018).

In order to protect the shellfish industry as well as human health, modelling the future scenarios of SST in the Estuary and the GSL, with the aim of mapping future potential risk areas, is primordial. Predicting SST in GSL has been generally realized through deterministic model, based on physical and mathematical representation of the climatic and ocean processes, such as the three-dimensional coastal ice-ocean model called CANOPA (CANadian Océan PArallélisé) (Long et al., 2015; Saucier, 2003). Recently, we used machine learning models (Artificial Neuron Networks-ANN, and Random Forest-RF) in predicting daily SST in the GSL by entering a combination of predictors (also known as features) explaining most of SST variation: 3-day trailing Moving Average (MA) of daily mean air temperature (i.e. average of daily mean air temperature of the present and two previous days), the 30 day-MA of daily mean wind speed, the 30 day-MA of maximum daily tidal range, 120 day-MA mean St. Lawrence freshwater runoff and 60 day-MA of

North Atlantic Oscillation (Ferchichi et al., 2019). The MAs are used as filters for smoothing predictors' data and detecting a better association between the dependent (SST) and independent variables. The results showed that Random Forests provided the best SST prediction accuracy of historical SST in the GSL (Ferchichi et al., 2019). In the same study, it has been demonstrated that both of the air temperature and wind speed are the most relevant predictors by explaining more than 70% of SST variation for most of the stations (Ferchichi et al., 2019). A recent study, focusing on coastal water temperature prediction, shows the impact of daily maximum and the average air temperature of previous 1 and 2 days on the daily water temperature variation. Considering this lag time factor and entering these variables as predictors improved significantly the daily coastal temperature prediction (Trinh et al., 2019).

In this paper, we present the future scenarios of *Vibrio* growth risk in the GSL by modelling the future water temperatures under different climate scenarios (optimistic and pessimistic). By entering the most relevant and readily available predictors (air temperature and wind speed) to the machine learning models (ANN and RF), we test their performance predictions and select the best inputs for each model using the backward selection method (Recursive Feature Elimination-RFE). After choosing the best model and entering the climate projections of the selected predictors, we produce the future water temperature in both of optimist and pessimist climatic scenarios. Finally, we map the future potential *Vibrio* growth risk area in the Estuary and GSL by interpolating the calculated risk indicator, in relation to the theoretical proliferation of pathogenic *Vibrio*, over our study area.

7.4 Study Area

The St. Lawrence River is the second largest river in North America (El-Sabh & Murty, 1990), with an average flow of approximately 12,100 m3/s (Galbraith et al., 2017). Originating from the Great Lakes, it reaches a vast estuary, where the fresh water of the river and salt water from the Atlantic Ocean mix. It flows over approximately 250 km to

Pointe-des-Monts where it becomes the Gulf of St-Lawrence, opened to the Atlantic Ocean through the straits of Cabot and Belle-Isle.

The study region, as shown in (Fig 7.1), covers the coastal areas of the Estuary (downstream limit near Rimouski) and the Gulf of St-Lawrence (the Quebec North Shore, Gaspe, Baie-des-Chaleurs, PEI and Magdalen Islands).

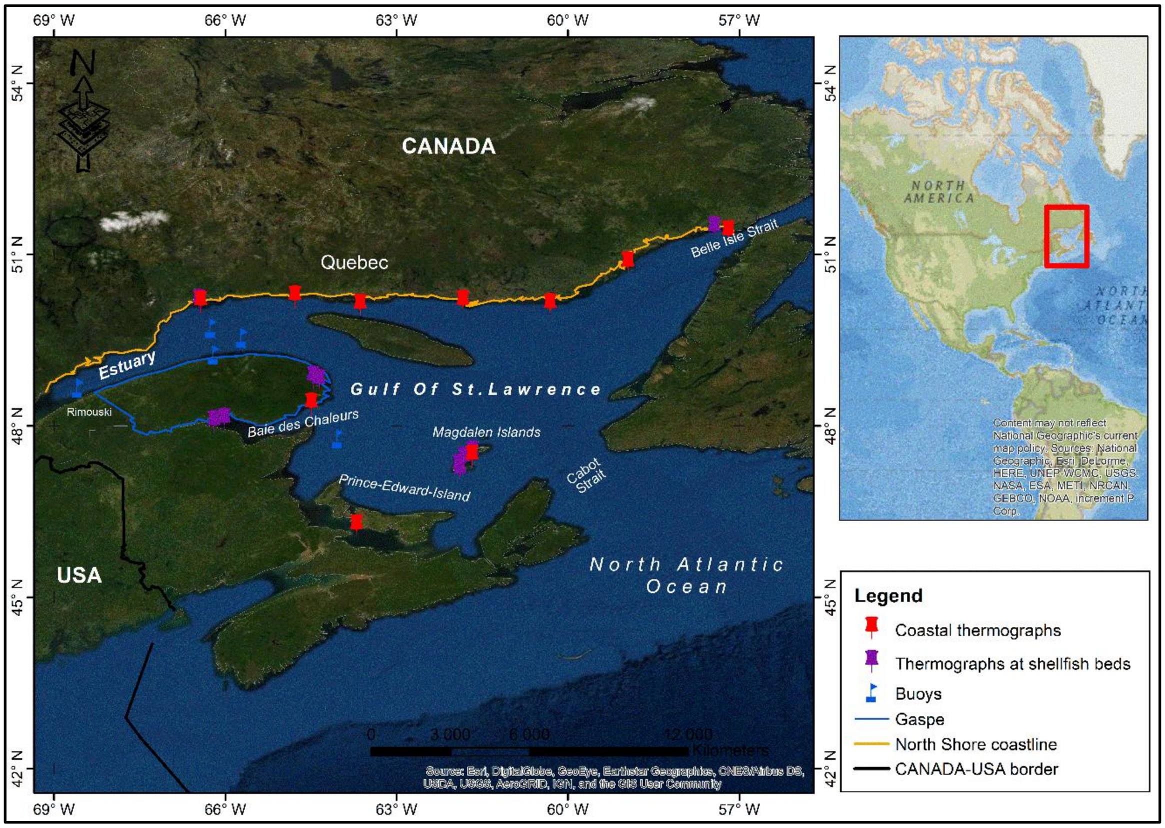


Figure 7.1 Geographic location of the thermographs and buoys in the Estuary and Gulf of St. Lawrence

7.5 Material and methods

7.5.1 Data collection

For modelling daily mean water temperature, we use the 3 day-MA mean air temperature and 30 day-MA wind speed as they present the most relevant explanatory variables of SST variation in the most of stations. The daily air temperature and wind speed data are available online from the Government of Canada through this site: http://climate.weather.gc.ca/historical_data/search_historic_data_e.html.

The daily sea surface temperatures of the buoys and costal thermographs were supplied by Fisheries and Oceans Canada and the Maurice-Lamontagne Institute.

The daily sea surface temperatures of coastal thermographs, located at shellfish beds, were provided by MERINOV-Québec Centre for Innovation in Aquaculture and Fisheries.

The predictor projections (daily air temperature and the 30-day MA wind speed) come from eight climate simulations obtained from Ouranos-a climate-science consortium based in Quebec (Martynov et al., 2013; Šeparović et al., 2013), and the CORDEX program (Giorgi et al., 2009). These simulations are generated through regional climate models driven by global climate model under one of the two Representative Concentration Pathways (RCPs; RCP4.5 (Knutti and Sedláček, 2013) or RCP8.5 (Meinshausen et al., 2011)).

The scenarios used in this study were the average of the regional model outputs mentioned in the Table 7.1.

Table 7.1 List of the Regional Climate Models (RCMs) used in simulations

Source of RCMs	Modelling groups	RCM
Ouranos	_	CRCM5
CORDEX	DMI (Danish Meteorological Institute)	HIRHAM5
	UQUAM (L'Université du Québec à Montréal)	CRCM5
	CCCma (Canadian Centre for Climate Modelling and Analysis)	CanRCM4

7.5.2 Modelling water temperature

We model the target variable (the daily water temperature of each buoy and coastal thermograph) by entering the selected predictors (3 day-MA air temperature and 30-day MA wind speed) into tested models (RF and ANN). The RFE was selected as feature selection method in order to choose the best subset of the predictors. In this case, only two combinations of the predictors were likely to be selected, either the air temperature as the sole input variable, or both air temperature and wind speed. 80% of original data are used for training and the remaining data serve as test data to evaluate the model predictive power. We use the k-fold cross validation (10-fold cross validation) as model validation technique.

7.5.2.1 Artificial Neural Network Model

In this study, we use a Multilayer Perceptron (MLP), a feedforward Artificial Neural Network, trained by using the supervised learning based on the error gradient back-propagation algorithm. This class of model is composed of three layers: the input layer includes the predictors which are standardized by subtracting each variable by its means and dividing by its standard deviations, the output layer produces the response variables and the hidden layer connects both of the input and output layers. At the state of hidden layer, the ANN attributes weights to the set of inputs (X_i), calculates the weighted sum (Z_i) and produces the output after applying an activation function on the weighted sum ($f(Z_i)$) for producing the desired model output. The weighted sum is given by:

$$Z_i = \sum_{i=1}^{n} w_i X_i + b \tag{7.1}$$

We use the sigmoid activation function. This non-linear function is given by:

$$f(Z_i) = \frac{1}{1 + e^{-Z_i}} \tag{7.2}$$

The used MLP is composed by one hidden layer holding only one neuron. A single hidden layer is sufficient to approximate any continuous function, but there is no general rule for selecting the appropriate number of hidden nodes in the hidden layer (Haykin, 1994; Piotrowski et al., 2015). The ANN architecture selection is done via the trial-error method. It tests the model performance by using a grid search of hidden neurons number (1 to 10 hidden nodes). Adding more than one neuron does not improve significantly the prediction.

7.5.2.2 Random Forest Model

Random Forests (RF) is a recent machine learning algorithms (2000) developed by Braiman (Breiman, 2001). The RF is a tree-based ensemble method that randomly selects a subset of predictors to build a binary tree based on bootstrap samples of the training data (Breiman, 2001). The overall prediction is the average of the predictions from all the generated decision trees (Aggregation).

The RF generalization error is estimated by averaging the prediction error of each tree using the Out-Of-Bag (OOB) samples, i.e. samples that are not included in the bootstrap training sets (1/3 bootstrap samples of the training sets). This OOB error is computed with a Mean Square Error (MSE) as shown below:

$$MSE^{OOB} = \frac{1}{n} \times \sum_{i=1}^{n} [\hat{Y}(X_i) - Y_i]^2 \quad (7.3)$$

Where n is the size of the OOB sample, $\hat{Y}(X_i)$ corresponds to the RF output given the input sample X_i, and Y_i represents the actual output.

The parameters set are: the number of trees, the number of bootstrap input variables at each split of a tree (Ntry) and the minimum node size of each tree. The minimum node size of each tree, recommended by RF creators, and used in many studies is 5. The smaller the minimum node size is, the deeper that the tree is. The Ntry, recommended by the RF developers, is the number of input variables divided by three (Breiman, 2001) (Ntry=1). The number of trees would be experimentally set through plotting the OOB error plot in function of tree numbers (number of trees=50).

7.5.2.3 Feature Selection

The Recursive feature Elimination (RFE) presents a backward selection technique used in selecting the best subset of input variables (features) that contribute the most in model accuracy during the training process. The RFE consists in training the model by entering all the features then removing the variables with the lowest contribution in model accuracy, i.e., based on variable importance. Using the new reduced feature subset, it retrains the model (Guyon et al., 2002). The best selected subset is the one that optimizes the most of the chosen performance criteria. In this study, we chose the RMSE as performance criterion for best subset features selection.

We use the k-fold cross validation (k=10) in performance prediction assessment of the possible feature subsets. In fact, the goodness of models fit, according to each feature subset, is assessed by computing the 10-fold cross validated RMSE ($RMSE_{CV}$), i.e. the average of the ten RMSE computed over the 10 validation sets:

The ($RMSE_{CV}$) is given by:

$$RMSE_{CV} = \frac{1}{k}\sum_{j=1}^{k} RMSE_j \tag{7.4}$$

Where the RMSE calculated for each fold j is given by:

$$RMSE_j = \sqrt{\sum_{i=1}^{N} \frac{(y_{i,p} - y_{i,obs})^2}{N}} \tag{7.5}$$

Where the k is the number of folds (10 in this example), j is one of the k folds, N is the sample size of the fold j, $y_{i,p}$ are predicted data and $y_{i,obs}$ are the observed data.

7.5.2.4 Models performance evaluation

In this paper, we compare the prediction accuracy of tested models (ANN and RF) according to three performance criteria: the Root Mean Square Error (RMSE), Nash-Sutcliffe coefficient (Nash) and Relative mean bias (rBias).

- Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_{i,obs} - y_{i,p})^2}{n}} \tag{7.6}$$

- Nash-Sutcliffe coefficient (Nash)

$$Nash = 1 - \frac{\sum_{i=1}^{n}(y_{i,obs} - y_{i,p})^2}{\sum_{i=1}^{n}(y_{i,obs} - \bar{y}_{obs})^2} \tag{7.7}$$

- Relative mean bias (rBias)

$$rBias = \frac{100}{n} \times \sum_{i=1}^{n} \frac{y_{i,obs} - y_{i,p}}{y_{i,obs}} \tag{7.8}$$

Where $y_{i,obs}$ is the observed data, $y_{i,p}$ is the predicted data, $\bar{y}_{obs}$ corresponds to the means of observed data and n is the size of observed data. Generally, a computed value of Nash greater than 0.5 indicates a relative satisfactory model performance, with a value of 1 corresponding to an ideal model (N. Moriasi et al., 2007). Low values of RMSE, rBias indicate better performing models.

7.5.3 Trend Analysis

After selecting the best model for all the stations, we generate the projections of the daily water temperature for the horizon 2040-2100. In order to cover the range of plausible water temperature scenarios, we select two different climatic scenarios: a relatively optimistic scenario (RCP4.5) and a pessimistic scenario (RCP8.5), known as “Business as usual” (i.e. continuous rise in GHG emissions). These scenarios lead to a warming of average air temperatures of 2.5 °C to 5 ° C around 2100. We perform a trend analysis of the predicted water temperature for each station under pessimistic scenario during the horizon 2040-2100 by using the Modified Mann Kendall (MMK) test, which takes into account the serial correlation. Then, we compute the trend slope of each station using the Theil-Sen’s slope estimator.

7.5.3.1 Modified Mann Kendall test (MMK)

The non-parametric Mann-Kendall (MK) test is commonly used for detecting monotonic trends in time series (Kendall, 1975; Mann, 1945). The null hypothesis H0 of this test is that there is no trend in the series.

The MK test statistic S is calculated as follows:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(x_j - x_i) \tag{7.9}$$

Where x_i and x_j denote the values of observations for the respective periods i and j (j> i), *n* is the length of the times series and $sgn(x_j - x_i)$ presents the sign function given by:

$$sgn(x_j - x_i)\begin{cases} 1\ if\ x_j - x_i > 1 \\ 0\ if\ x_j - x_i = 0 \\ -1\ if\ x_j - x_i < 1 \end{cases} \tag{7.10}$$

Mann (1945) and Kendall (1975) have noted that for large values of n (n≥8), the distribution of the S statistic is approximately normal (Kendall, 1975; Mann, 1945), with the mean *E* and variance *V* of the statistic *S*, are defined as follows:

$$E(S) = 0 \tag{7.11}$$

$$V(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^{m} t_i(i-1)(2i+5)}{18} \tag{7.12}$$

Where *m* represents the number of tied groups in the data set and the t_i represents the number of values in the i^th^ tied group.

The standardized statistic Z_S is calculated by:

$$Z_S = \begin{cases} \frac{S-1}{\sqrt{V(s)}} & if\ S > 0 \\ 0 & if\ S = 0 \\ \frac{S+1}{\sqrt{V(s)}} & if\ S < 0 \end{cases}$$

(7.13)

The sign of the statistic Z_S indicates the direction of the trend whether it is upward (positive Z_S) or downward (negative Z_S). The standardized statistic Z_S follows the standard normal distribution with a mean of 0 and variance of 1. The null hypothesis would be rejected, implying the presence of a significant trend, when Z_S is higher than a critical value $Z_{1-\alpha/2}$, where α represents the chosen significance level (5% in this study) and $Z_{1-\alpha/2}$ could be deduced from the standard normal cumulative distribution tables.

In order to account for the autocorrelation that may exist in the time series, Hamed and Rao proposed to modify the variance of the MK test (Hamed and Rao, 1998). The variance is corrected through multiplying by the factor n/n^*, where n^* presents the effective sample size. Yue and Yang have demonstrated that incorporating the effective sample size in variance correction limits effectively the effect of serial correlation on the MK test (Yue and Wang, 2004).

$$\frac{n}{n^*} = 1 + \frac{2}{n(n-1)(n-2)} \sum_{i=1}^{p} (n-i)(n-i-1)(n-i-2)\,\rho_s\,(i)$$

(7.14)

Where n is the actual sample size, n^* is the effective sample size to account for autocorrelations in the data, ρ_s presents the autocorrelation function of the ranks of the observations for lag i and p is the maximum of time lags taking into account.

In this paper, the variance is corrected through considering complete autocorrelations (all lags) in the effective sample size computation, proposed by Yue and Wang (2004), and applied by using the mmky R package.

7.5.3.2 Theil-Sen's slope estimator

Theil-Sen's slope estimator, proposed by (Theil, 1950) and (Sen, 1968), allows to capture the direction and the strength of significant trend slope. It has been considered as robust estimate of the magnitude of trend's slope (Yue and Wang, 2004).

It is given by the following equation: It is given by the following equation:

$$b = median\left(\frac{x_j - x_i}{j - i}\right) \; \forall \; i < j \tag{7.15}$$

7.5.4 Mapping future potential risk growth Vibrio

In order to evaluate the potential risk of Vibrio growth, we chose as thermal metric: the number of days above the minimum known temperature for Vibrio growth (15 °C). This thermal metric, selected as Vibrio growth risk indicator, is computed from the produced daily future water temperatures for both of optimistic and pessimistic scenarios.

We calculate the monthly average of this risk indicator, during the study period (June-October), averaged over twenty years during the study horizon 2040-2100. Subsequently, we interpolate the risk indicator computed for the available stations over the study area using the Inverse Distance Weighting (IDW) method, using ARCGIS. As a result of a comparison between IDW and kriging, we select IDW interpolation as it is the simplest method, given similar interpolation errors to kriging. We produce maps for two selected future horizons 2040-2060 and 2080-2100 under both climate change scenarios (optimistic-RCP4.5 and pessimistic-RCP8.5) in order to compare the level of potential Vibrio expansion risk over the study area.

We compute the root mean squared error (RMSE) of the interpolation using a leave-one-out procedure. The relative RMSE is calculated by dividing the RMSE of the IDW interpolation, produced for every month i of study period, by the areal average of the risk indicator for the same month. Sometimes the relative error is quite strong, which may be due to the large spatial variation of the selected variable. For cases that are too uncertain (relative error>50%), interpolation is useless.

7.5.4.1 Inverse Weighted Distance (IWD) interpolation method

The IDW method, a deterministic spatial interpolation approach, allows to compute an average of a selected variable in ungauged sites using values from nearby weighted sites.

The weights, accorded to gauged locations, are proportional to the distance between the gauged and ungauged sites and determined by the IDW power coefficient. The larger the power coefficient is, the stronger the weights are attributed to the closest locations. The estimated variable at ungauged location (z_j*) is defined by the following equation:

$$z_j^* = \frac{\sum_{i=1}^{n} w_{ij} x_i}{\sum_{i=1}^{n} w_{ij}} \tag{7.16}$$

Where:

$$w_{ij} = \frac{1}{d_{ij}^{p}} \tag{7.17}$$

Where x_i is the variable value of a neighboring gauged site, w_{ij} is the weight assigned to the gauged sites (i), d_{ij} is the distance between the gauged (i) and ungauged sites (j), n is the number of gauged sites and p is the exponent of the distance. In this study, the interpolation was performed using the ArcGIS software and a default value of p = 2 was chosen.

7.6 Results

After applying the Recursive Feature Elimination (RFE) on the selected potential predictors, air temperature and wind speed, we found that the ANN uses only the air temperature as input for most of the stations while the RF uses both predictors.

Table 7.2 presents the results of performance criteria for ANN and RF in SST prediction of the tested dataset.

Table 7.2 Performance criteria results of tested models (RF and ANN)

Models	Stations	RMSE (°C)	Nash	rBiais(%)
	Grande Rivière	1.7	0.708	-2.05
	Borden	1.991	0.585	-1.529
	Blanc Sablon	2.437	0.331	-13.774
	Courant de Gaspé	1.368	0.84	-1.235
	Havre St.Pierre	2.435	0.29	-12.67
	Ile Shag	2.097	0.651	-4.008
ANN	Montlouis	1.323	0.854	-1.647
(Artificial Neural Network)	Natashquan	2.028	0.709	-6.409
	Rimouski	1.139	0.827	-0.779
	Rivière aux Tonnerre	2.015	0.627	-7.809
	Romaine	1.345	0.816	-2.929
	Tabatière	2.342	0.391	-9.577
	Sept-Îles	1.957	0.702	-3.175
	Shediac Valley	1.056	0.893	-0.568
	Mean	**1.802**	**0.659**	**-4.868**
	Grande Rivière	1.654	0.708	-1.877
	Borden	1.837	0.585	-0.66
	Blanc Sablon	2.29	0.331	-10.614
	Courant de Gaspé	1.37	0.84	-1.018

	Havre St. Pierre	2.519	0.29	-10.844
	Ile Shag	1.958	0.685	-3.478
RF	Montlouis	1.418	0.687	-1.407
(Random Forest)	Natashquan	2.082	0.693	-4.756
	Rimouski	1.157	0.827	-0.794
	Rivière aux Tonnerre	2.08	0.627	-6.377
	Romaine	1.404	0.816	-2.496
	Tabatière	2.323	0.391	-7.193
	Sept-Îles	1.982	0.702	-3.676
	Shediac Valley	1.02	0.893	-0.43
	Mean	**1.792**	**0.648**	**-3.973**

By using the paired t-test (DF=13, t-value=0.436) on RMSE values at all of our sites for both models, we note that there is no significant difference between the performances of models in terms of RMSE (p-value=0.67) at a significance level of 5%. The average RMSE performance for all the stations is approximately 1.74 °C for RF and 1.8 °C for ANN. Both models present good performing results in terms of Nash-Sutcliff criterion, i.e. higher than 0.5, and low relative mean bias (<5%). Given that there is no significant difference between RF and ANN, we choose the ANN as it is the most parsimonious model for modelling future water temperatures. Then, we proceed to generate future daily mean water temperature for each station for the horizon 2040-2100 for a pessimistic climate scenario, RCP8.5, and a more optimistic one (RCP4.5), by using the projections of air temperature.

We proceed with a trend analysis of the predicted water temperature for each station under pessimistic scenario during the horizon 2040-2100 by using the Modified Mann Kendall (MMK) test, which takes into account the serial correlation. Significant positive trends (p-value <1%) in future daily mean SST, for the period from June to October in the horizon 2040-2100, were revealed in all the tested stations at a significance level of 5%.

Table 7.3 presents the results of Theil Sen's slope computed for each station after applying the MMK test. By averaging the trend slopes of all the stations over the horizon (2040-2100), we found that the water temperatures are likely to increase by 0.4 °C per decade, for a total of 2.5 °C up to 2100.

Table 7.3 Theil-Sen's slope for projected daily mean water temperature times series during the horizon 2040–2100

Stations	Theil-Sen's slope	[a]Slope-10 years	[b]Slope-60years
Natashquan	4.59E-04	0.711	4.266
Baie Cascapedia	1.69E-04	0.262	1.57
Baie Trascapedia	1.73E-04	0.269	1.613
Sept-Îles	2.42E-04	0.376	2.255
Baie Plaisance	1.34E-04	0.208	1.245
Lagune Havre	3.31E-04	0.513	3.079
Bassin Havre	3.71E-04	0.575	3.447
Lagune Grande	3.16E-04	0.489	2.935
Belles Amours	2.81E-04	0.436	2.618
Blanc Sablon	2.64E-04	0.41	2.458
Borden	1.49E-04	0.232	1.39
Grande Rivière	1.66E-04	0.257	1.54
Iles Shag	1.50E-04	0.233	1.4
Havre St.Pierre	1.75E-04	0.271	1.624
Rivière aux Tonnerre	4.10E-04	0.635	3.813
Romaine	4.49E-04	0.696	4.176

Tarbatière	3.30E-04	0.512	3.072
Rimouski	3.77E-04	0.584	3.506
Courant Gaspé	2.57E-04	0.399	2.391
Gyre Anticosti	2.02E-04	0.312	1.875
Montlouis	2.63E-04	0.407	2.443
Shediac Valley	2.03E-04	0.315	1.889
Mean	**2.67E-04**	**0.414**	**2.482**

[a] **Slope-10 years:** the average of Theil-Sen's slope during 20 years for the time period between June and October

[b] **Slope-60years:** the average of Theil-Sen's slope during 60 years for the time period between June and October

After generating the future daily means of SST, we calculate the number of days exceeding the threshold of 15 °C, the minimum temperature for Vibrio growth, as a risk indicator. Then, we interpolate the values of this risk indicator, averaged for each month over 20 years. We present the results of August and September as examples of spatial interpolation of the risk indicator over the study area in Figures 7.2 and 7.3 respectively.

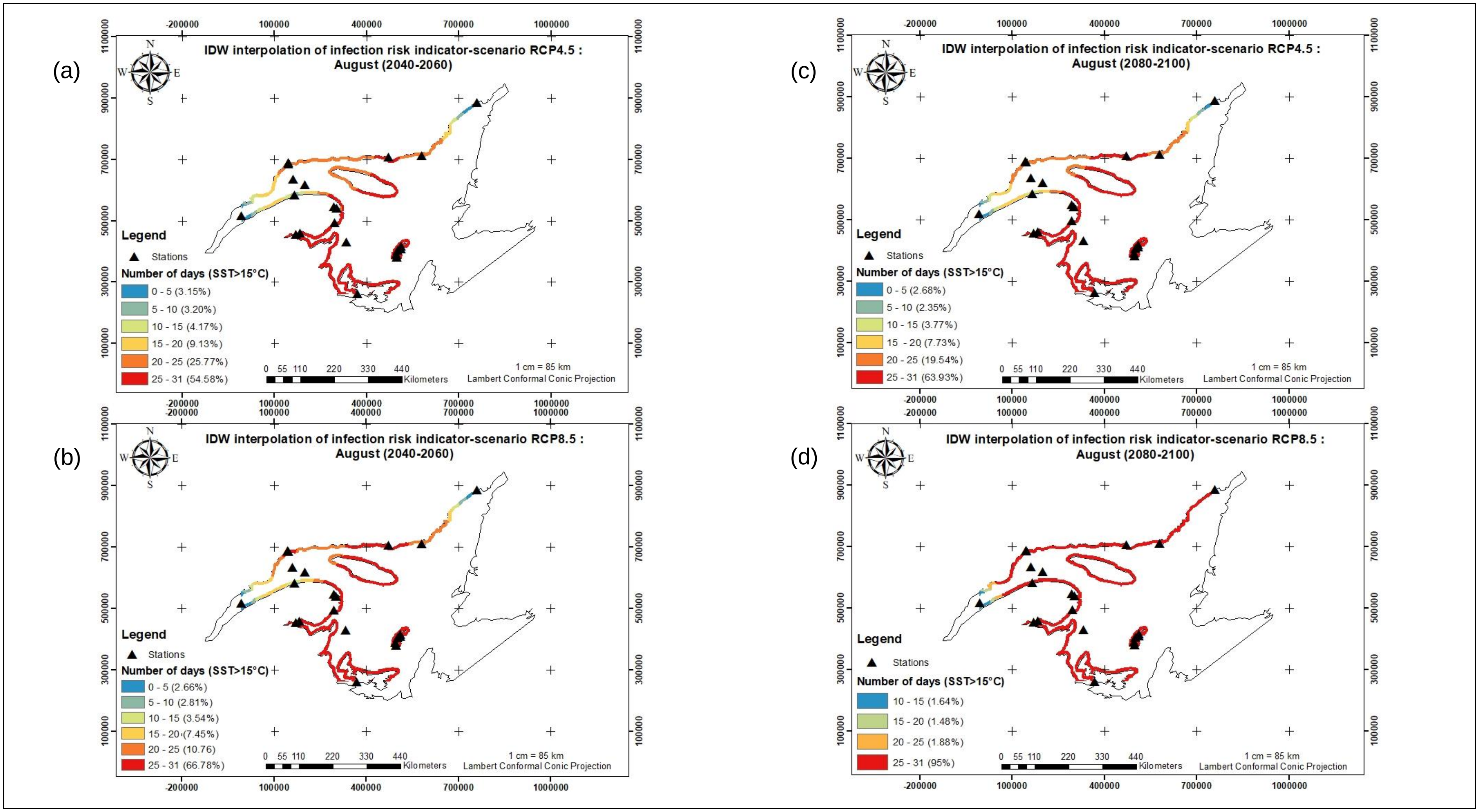


Figure 7.2 Inverse Distance Weighting (IDW) interpolation of risk indicator values (number of days exceeding the threshold [15 °C]) over the Estuary and Gulf of St. Lawrence under pessimistic and optimistic climatic scenario for the horizons 2040–2060 and 2080–2100 in August. (a) IDW interpolation of the risk indicator values in August during the horizon 2040–2060 under optimistic scenario (RCP 4.5). (b) IDW interpolation of the risk indicator values in August during the horizon 2040–2060 under pessimistic scenario (RCP 8.5). (c) IDW interpolation of the risk indicator values in August during the horizon 2080–2100 under optimistic scenario (RCP 4.5). (d) IDW interpolation of the risk indicator values in August during the horizon 2080–2100 under pessimistic scenario (RCP 8.5)

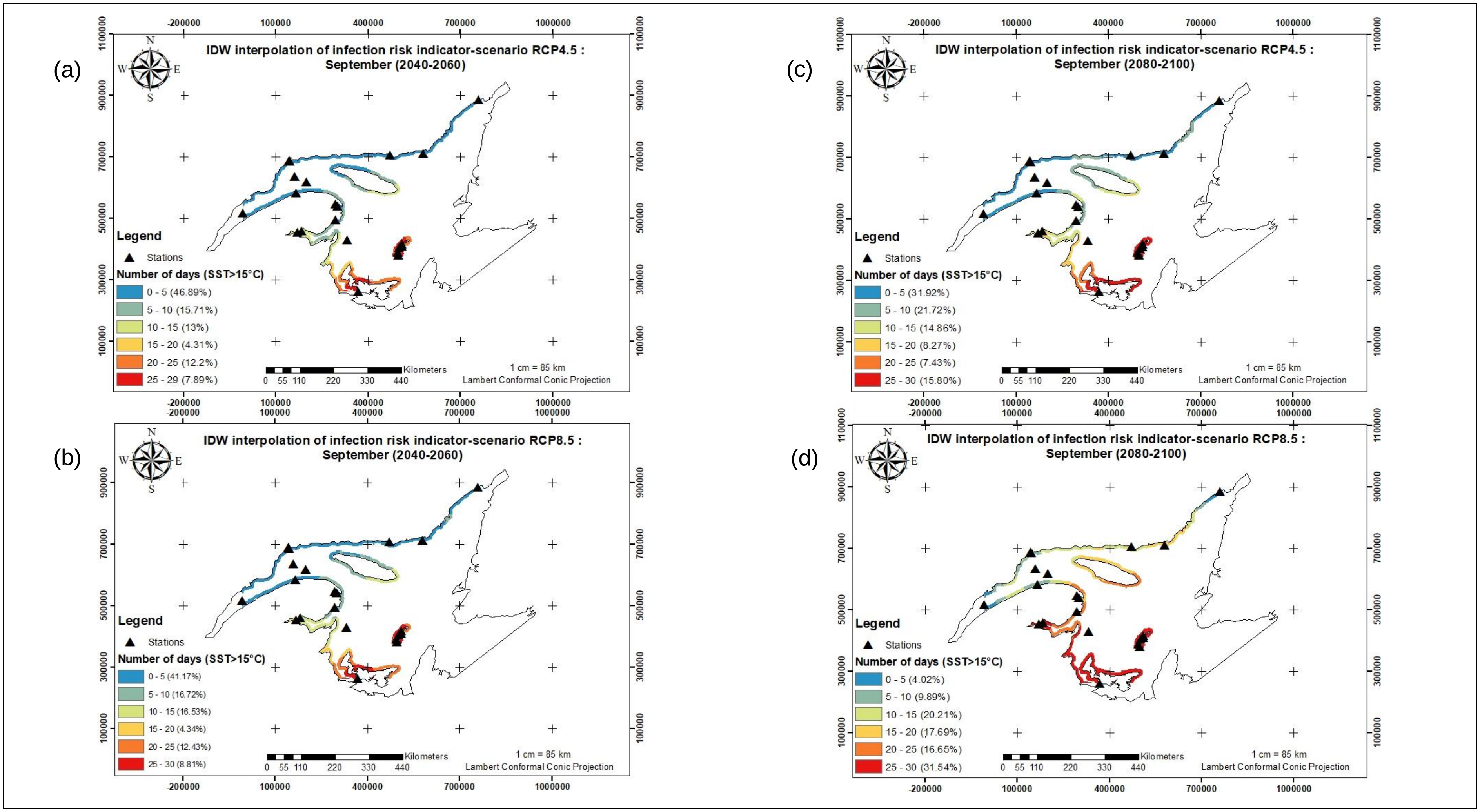


Figure 7.3 Inverse Distance Weighting (IDW) interpolation of risk indicator values (number of days exceeding the threshold [15 °C]) over the Estuary and Gulf of St. Lawrence under pessimistic and optimistic climatic scenario for the horizons 2040–2060 and 2080–2100 in September. (a) IDW interpolation of the risk indicator values in September during the horizon 2040–2060 under optimistic scenario (RCP 4.5). (b) IDW interpolation of the risk indicator values in September during the horizon 2040–2060 under pessimistic scenario (RCP 8.5). (c) IDW interpolation of the risk indicator values in September during the horizon 2080–2100 under optimistic scenario (RCP 4.5). (d) IDW interpolation of the risk indicator values in September during the horizon 2080–2100 under pessimistic scenario (RCP 8.5)

Figure 7.2 (a) shows that, during the horizon 2040-2060, under an optimistic scenario (RCP4.5), the waters in the shellfish beds of Magdalen Islands, PEI, the Gaspe Peninsula and Baie des Chaleurs are likely to be at high risk of infection as the risk indicator (number of days above the 15 °C) exceeds 25 days. The risk indicator of some stations in the Quebec North Shore coast along the GSL varies between 20 and 25 days, so the shellfish beds in this area might also be under high risk of Vibrio growth. For the same horizon, but under a pessimistic scenario as shown in Figure 7.2 (b), most shellfish beds would probably be at high risk of Vibrio growth since the calculated risk indicator exceeds 25 days for approximately 67% of coastal areas.

By comparing the interpolation maps of August (Fig 7.2 (c) and Fig 7.2 (d)) during the horizon 2080-2100 for both scenarios, we note that the risk indicator, exceeding 25 days, covers between 64% (scenario RCP4.5) and 95% (scenario RCP8.5) of the total coastal area. Therefore, by 2100, most of the stations located in shellfish beds, where harvesting occurs, are likely to be at high risk of Vibrio growth whatever the considered scenario.

During September (Fig 7.3), the risk indicator at the shellfish beds of the Magdalen Islands and PEI exceeds 20 days for both climate scenarios and horizons. So, they might be at a higher risk of Vibrio growth. During the horizon (2080-2100) under a pessimistic scenario (Fig 7.3 (d)), we note that in addition to the Magdalen Islands and PEI, the shellfish beds of Gaspe Peninsula and Baie des Chaleurs would probably be at high risk of Vibrio growth as the number of days above the 15 °C threshold exceeds 20 days.

Table 7.4 presents the results of the risk indicator calculated for the month of October for some stations under a pessimistic scenario during the horizon 2080-2100. It was not possible to perform a spatial interpolation because too few stations had non-zero values. Table 7.4 shows that the risk of Vibrio growth may also occur during October on the coasts of PEI and Magdalen Islands. Therefore, the risk of Vibrio growth would probably expand both spatially and temporally (i.e. into the fall for some regions).

Table 7.4 The number of days exceeding the minimum temperature threshold (15 °C) for the growth of pathogenic Vibrio during October in the horizon 2080–2100 under pessimistic scenario (RCP 8.5)

Localization	stations	October
PEI	Borden	19
Magdalen Islands	Baie Plaisance	6
	Bassin Havre	20
	Lagune Havre	16
	Lagune Grande	14
	Ile Shag	4

The risk assessment was not just limited to one risk indicator. We also compute the number of days exceeding a threshold of 20 °C. Table 7.5 presents the stations that would be at high risk of pathogenic Vibrio growth. For the rest of the stations, the number of days exceeding 20 °C is zero so they have not been included in the table. Table 7.5 shows that during 2080-2100 under a pessimistic scenario, the threshold of 20 °C would be exceeded along the coast of the Magdalen Islands for about 31 days in August against an average of 18 days in September. The shellfish beds in the Gaspe Peninsula could be at the same degree of risk with about 30 days in August while the PEI and North Shore would be under a lower risk with an average of 16 days.

Table 7.5 The number of days exceeding the threshold (20 °C) in the horizon 2080–2100 under pessimistic scenario (RCP 8.5)

Localization	stations	August	September
North Shore	Natashquan	17	0
	Bassin Havre	31	22
Magdalen Islands	Lagune Havre	31	16
	Lagune Grande	31	16
Gaspé	Havre Gaspé	30	4
PEI	Borden	16	0

7.7 Discussion

By using one or both of the most relevant predictors, air temperature and wind speed, the results show that the SST prediction performance of ANN and RF were similar. However, RF requires more predictors than ANN to achieve similar prediction performance. Thus, we selected the ANN for SST prediction and used the mean of climate models projections of air temperature, as input for ANN model, without exploring the variability between the projections.

Modelling future water temperature through ANN constitutes a useful tool to predict the plausible future water temperatures in the coastline of the Estuary and GSL where shellfish beds occur. The trend analysis results for daily mean water temperature, using the MMK test, indicates that our study area exhibits a significant increasing trend by 2.5 °C up to 2100 under a pessimistic scenario. This positive trend in water temperature implies a rise in risk indicator of Vibrio growth. This result is demonstrated in the interpolation maps between the horizons 2040-2060 and 2080-2100. In fact, by comparing the interpolation maps in August for both of the horizons under the pessimistic scenario (RCP8.5), we note an expansion of Vibrio growth risk from 64% to 95% of the total coastal area of the Estuary and the Gulf of Saint Lawrence. The risk indicator distribution during July was similar to August so the shellfish beds would be exposed to a similar risk as in August under both scenarios. Whereas, in June the risk of Vibrio growth would be less severe than August and July except over the horizon (2080-2100) under pessimistic scenario, where all the shellfish beds on the coasts of North Shore, Gaspe Peninsula, Baie des Chaleurs, Magdalen Islands and PEI would be under high risk of Vibrio growth. Their risk indicator could exceed an average of 20 days.

The results of risk indicator interpolation in August, suggest that Vibrio growth risk may increase under both of pessimistic or optimistic scenario so all the shellfish beds practically (on the coasts of North Shore, Gaspe Peninsula, Baie des Chaleurs, Magdalen Islands and PEI) would be at risk of Vibrio growth regardless the scenario. In addition to this spatial spread, the Vibrio growth risk would extend seasonally by occurring out of the summer time during September and even October, especially on the coasts of Magdalen Islands and PEI.

The lowest temperature threshold for the Vibrio growth, based on the literature, is 15 °C. Computing the number of days exceeding a higher temperature threshold (20 °C) allowed to locate the shellfish beds that would be at higher risk of Vibrio growth, like the Magdalen Islands and PEI.

This study focused on surface temperatures (average depth of 1.5 m). Wild molluscs can be found or harvested at this depth, e.g. on the foreshore or near the islands. However, in some cases, molluscs harvesting occurs in deeper water (e.g. oysters). The risk maps produced may be biased, i.e. the risk of Vibrio growth may be overestimated, as there may be significant thermal stratification in some coastal sites. Molluscs that are found under the thermocline may not be as much as risks as those in shallow, well mixed areas.

In this study, we focused on water temperature as the main factor affecting Vibrio growth. However, it should be noted that many studies confirm the importance of salinity in Vibrio growth. So, future work on Vibrio risk management should concentrate of the combined effects of water temperature and salinity on the proliferation of pathogenic Vibrio and use the projections of both variables to better locate the potential risk areas followed by sampling of water and shellfish to confirm the presence of Vibrio.

7.8 Acknowledgements

The authors thank Ouranos for funding this research.

We also thank Ouranos for generating and supplying the CRCM5 data.

The authors would like to thank Gabriel Rondeau-Genesse for providing us with the required variable projections.

The Canadian Regional Climate Model (CRCM5) was developed by the ESCER Centre at UQAM (Université du Québec à Montréal) with the collaboration of Environment and Climate Change Canada.

CRCM5 computations were made on the supercomputer guillimin from McGill University, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation For Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and the Fonds de recherché du Québec- Nature et technologies (FRQ-NT).

We also thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and The Canadian Foundation for Climate and Atmospheric Sciences (CFCAS) for the funding of the development of the CRCM5.

We acknowledge the World Climate Research Programme's Working Group on Regional Climate, and the Working Group on Coupled Modelling, former coordinating body of CORDEX and responsible panel for CMIP5. We also thank the climate modelling groups (listed in Table 1 on this paper) for producing and making available their model output. We also acknowledge the U.S. Department of Defense ESTCP for its support of the NA-CORDEX data archive.

8 REFERENCES

Baker-Austin, C., Stockley, L., Rangdale, R., Martinez-Urtaza, J., 2010. Environmental occurrence and clinical impact of Vibrio vulnificus and Vibrio parahaemolyticus: a European perspective. Environmental microbiology reports 2, 7-18.

Baker-Austin, C., Trinanes, J., Gonzalez-Escalona, N., Martinez-Urtaza, J., 2017. Non-Cholera Vibrios: The Microbial Barometer of Climate Change. Trends in microbiology 25, 76-84.

Baker-Austin, C., Trinanes, J.A., Taylor, N.G., Hartnell, R., Siitonen, A., Martinez-Urtaza, J., 2013. Emerging Vibrio risk at high latitudes in response to ocean warming. Nature Climate Change 3, 73.

Baker-Austin, C., Trinanes, J.A., Taylor, N.G.H., Hartnell, R., Siitonen, A., Martinez-Urtaza, J., 2012. Emerging Vibrio risk at high latitudes in response to ocean warming. Nature Climate Change 3, 73-77.

Banerjee, S.K., Rutley, R., Bussey, J., 2018. Diversity and dynamics of the Canadian coastal Vibrio community: an emerging trend detected in the temperate regions. Journal of bacteriology 200, e00787-00717.

Breiman, L., 2001. Random forests. Machine learning 45, 5-32.

Burge, C.A., Mark Eakin, C., Friedman, C.S., Froelich, B., Hershberger, P.K., Hofmann, E.E., Petes, L.E., Prager, K.C., Weil, E., Willis, B.L., Ford, S.E., Harvell, C.D., 2014. Climate change influences on marine infectious diseases: implications for management and society. Annual review of marine science 6, 249-277.

Centers for Disease, C., Prevention, 2013. Incidence and trends of infection with pathogens transmitted commonly through food - foodborne diseases active surveillance network, 10 U.S. sites, 1996-2012. MMWR. Morbidity and mortality weekly report 62, 283-287.

Centers for Disease, C.P., 2006. Preliminary FoodNet data on the incidence of infection with pathogens transmitted commonly through food--10 States, United States, 2005. MMWR Morb Mortal Wkly Rep 55, 392-395.

Chu, C., Do, Y., Kim, Y., Saito, Y., Lee, S.-D., Park, H., Lee, J.-K., 2011. Mathematical modeling of Vibrio vulnificus infection in Korea and the influence of global warming. Osong public health and research perspectives 2, 51-58.

Davis, B.J.K., Jacobs, J.M., Davis, M.F., Schwab, K.J., DePaola, A., Curriero, F.C., 2017. Environmental determinants of Vibrio parahaemolyticus in the Chesapeake Bay. Applied and environmental microbiology.

Dechet, A.M., Yu, P.A., Koram, N., Painter, J., 2008. Nonfoodborne Vibrio infections: an important cause of morbidity and mortality in the United States, 1997-2006. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 46, 970-976.

Deeb, R., Tufford, D., Scott, G.I., Moore, J.G., Dow, K., 2018. Impact of Climate Change on Vibrio vulnificus Abundance and Exposure Risk. Estuaries and Coasts 41, 2289-2303.

Feldhusen, F., 2000. The role of seafood in bacterialfoodborne diseases. Microbes and infection 2, 1651-1660.

Ferchichi, H., St-Hilaire., A., Ouarda., T., Lévesque., B., 2019. Modélisation des scénarios futurs de température de l’eau en milieu côtier et implications sur les infections potentielles par Vibrio parahaemolyticus et Vibrio vulnificus : application aux bancs coquillers de l’estuaire et du golfe du st-laurent (Unpublished master's thesis). Institut National de Recheche Scientifique, Canada, Quebec.

Galbraith, P.S., Larouche, P., Chassé, J., Petrie, B., 2012. Sea-surface temperature in relation to air temperature in the Gulf of St. Lawrence: Interdecadal variability and long term trends. Deep Sea Research Part II: Topical Studies in Oceanography 77-80, 10-20.

Giorgi, F., Jones, C., Asrar, G.R., 2009. Addressing climate information needs at the regional level: the CORDEX framework. World Meteorological Organization (WMO) Bulletin 58, 175.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46, 389-422.

Halpern, B.S., Walbridge, S., Selkoe, K.A., Kappel, C.V., Micheli, F., D'Agrosa, C., Bruno, J.F., Casey, K.S., Ebert, C., Fox, H.E., Fujita, R., Heinemann, D., Lenihan, H.S., Madin, E.M., Perry, M.T., Selig, E.R., Spalding, M., Steneck, R., Watson, R., 2008. A global map of human impact on marine ecosystems. Science 319, 948-952.

Hamed, K.H., Rao, A.R., 1998. A modified Mann-Kendall trend test for autocorrelated data. Journal of hydrology 204, 182-196.

Haykin, S., 1994. Neural networks: a comprehensive foundation. Prentice Hall PTR.

Heng, S.-P., Letchumanan, V., Deng, C.-Y., Ab Mutalib, N.-S., Khan, T.M., Chuah, L.-H., Chan, K.-G., Goh, B.-H., Pusparajah, P., Lee, L.-H., 2017. Vibrio vulnificus: An Environmental and Clinical Burden. Frontiers in microbiology 8, 997-997.

IPCC, 2013. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jacobs, J., Moore, S.K., Kunkel, K.E., Sun, L., 2015. A framework for examining climate-driven changes to the seasonality and geographical range of coastal pathogens and harmful algae. Climate Risk Management 8, 16-27.

Jacobs, J.M., Rhodes, M., Brown, C.W., Hood, R.R., Leight, A., Long, W., Wood, R., 2014. Modeling and forecasting the distribution of Vibrio vulnificus in Chesapeake Bay. Journal of applied microbiology 117, 1312-1327.

Kaspar, C.W., Tamplin, M.L., 1993. Effects of temperature and salinity on the survival of Vibrio vulnificus in seawater and shellfish. Applied and environmental microbiology 59, 2425-2429.

Kendall, M., 1975. Rank Correlation Methods. Griffin, London.

Knutti, R., Sedláček, J., 2013. Robustness and uncertainties in the new CMIP5 climate model projections. Nature Climate Change 3, 369.

Lima, F.P., Wethey, D.S., 2012. Three decades of high-resolution coastal sea surface temperatures reveal more than warming. Nature Communications 3, 704.

Long, Z., Perrie, W., Chassé, J., Brickman, D., Guo, L., Drozdowski, A., Hu, H., 2015. Impacts of Climate Change in the Gulf of St. Lawrence. Atmosphere-Ocean 54, 337-351.

Mann, H.B., 1945. Nonparametric tests against trend. Econometrica: Journal of the Econometric Society, 245-259.

Martinez-Urtaza, J., Bowers, J.C., Trinanes, J., DePaola, A., 2010. Climate anomalies and the increasing risk of Vibrio parahaemolyticus and Vibrio vulnificus illnesses. Food Research International 43, 1780-1790.

Martinez-Urtaza, J., Huapaya, B., Gavilan, R.G., Blanco-Abad, V., Ansede-Bermejo, J., Cadarso-Suarez, C., Figueiras, A., Trinanes, J., 2008. Emergence of asiatic vibrio diseases in south america in phase with El Niño. Epidemiology 19, 829-837.

Martynov, A., Laprise, R., Sushama, L., Winger, K., Šeparović, L., Dugas, B., 2013. Reanalysis-driven climate simulation over CORDEX North America domain using the Canadian Regional Climate Model, version 5: model performance evaluation. Climate Dynamics 41, 2973-3005.

McLaughlin, 2005. climate anomalies and the increasing risk of Vibrio.

McLaughlin, J.B., DePaola, A., Bopp, C.A., Martinek, K.A., Napolilli, N.P., Allison, C.G., Murray, S.L., Thompson, E.C., Bird, M.M., Middaugh, J.P., 2005. Outbreak of Vibrio parahaemolyticus gastroenteritis associated with Alaskan oysters. New England Journal of Medicine 353, 1463-1470.

Mead, P.S., Slutsker, L., Dietz, V., McCaig, L.F., Bresee, J.S., Shapiro, C., Griffin, P.M., Tauxe, R.V., 1999. Food-related illness and death in the United States. Emerging infectious diseases 5, 607.

Meinshausen, M., Smith, S.J., Calvin, K., Daniel, J.S., Kainuma, M.L.T., Lamarque, J.-F., Matsumoto, K., Montzka, S.A., Raper, S.C.B., Riahi, K., Thomson, A., Velders, G.J.M., van Vuuren, D.P.P., 2011. The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. Climatic Change 109, 213.

Motes, M.L., DePaola, A., 1996. Offshore suspension relaying to reduce levels of Vibrio vulnificus in oysters (Crassostrea virginica). Applied and environmental microbiology 62, 3875-3877.

Motes, M.L., DePaola, A., Cook, D.W., Veazey, J.E., Hunsucker, J.C., Garthright, W.E., Blodgett, R.J., Chirtel, S.J., 1998. Influence of Water Temperature and Salinity on Vibrio vulnificus in Northern Gulf and Atlantic Coast Oysters (Crassostrea virginica). Applied and environmental microbiology 64, 1459-1465.

N. Moriasi, D., G. Arnold, J., W. Van Liew, M., L. Bingner, R., D. Harmel, R., L. Veith, T., 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Transactions of the ASABE 50, 885-900.

Narjol, G.-E., Viviana, C., Claudia, A., María, L.R., Juan, A.V., Felipe, C., Jaime, R., Romilio, T.E., 2005. Vibrio parahaemolyticus Diarrhea, Chile, 1998 and 2004. Emerging Infectious Disease journal 11, 129.

Newton, A., Kendall, M., Vugia, D.J., Henao, O.L., Mahon, B.E., 2012. Increasing rates of vibriosis in the United States, 1996-2010: review of surveillance data from 2 systems. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 54 Suppl 5, S391-395.

Oliver, J.D., 2005. Wound infections caused by Vibrio vulnificus and other marine bacteria. Epidemiology and Infection 133, 383-391.

Oliver, J.D., 2013. Vibrio vulnificus: Death on the Half Shell. A Personal Journey with the Pathogen and its Ecology. Microbial ecology 65, 793-799.

Piotrowski, A.P., Napiorkowski, M.J., Napiorkowski, J.J., Osuch, M., 2015. Comparing various artificial neural network types for water temperature prediction in rivers. Journal of Hydrology 529, 302-315.

Saucier, F.J., 2003. Modeling the formation and circulation processes of water masses and sea ice in the Gulf of St. Lawrence, Canada. Journal of Geophysical Research 108.

Semenza, J.C., Trinanes, J., Lohr, W., Sudre, B., Lofdahl, M., Martinez-Urtaza, J., Nichols, G.L., Rocklov, J., 2017. Environmental Suitability of Vibrio Infections in a Warming Climate: An Early Warning System. Environmental health perspectives 125, 107004.

Sen, P.K., 1968. Estimates of the regression coefficient based on Kendall's tau. Journal of the American statistical association 63, 1379-1389.

Šeparović, L., Alexandru, A., Laprise, R., Martynov, A., Sushama, L., Winger, K., Tete, K., Valin, M., 2013. Present climate and climate change over North America as simulated by the fifth-generation Canadian regional climate model. Climate Dynamics 41, 3167-3201.

Su, Y.-C., Liu, C., 2007. Vibrio parahaemolyticus: a concern of seafood safety. Food microbiology 24, 549-558.

Theil, H., 1950. A rank-invariant method of linear and polynominal regression analysis (parts 1-3), Ned. Akad. Wetensch. Proc. Ser. A, pp. 1397-1412.

Trinh, N.X., Trinh, T.Q., Phan, T.P., Thanh, T.N., Thanh, B.N., 2019. Water Temperature Prediction Models in Northern Coastal Area, Vietnam.

United States Food and Drug Administration (FDA), 2005. Quantitative Risk Assessment on the Public Health Impact of Pathogenic Vibrio parahaemolyticus.

Vezzulli, L., Colwell, R.R., Pruzzo, C., 2013. Ocean warming and spread of pathogenic vibrios in the aquatic environment. Microbial ecology 65, 817-825.

Vezzulli, L., Grande, C., Reid, P.C., Helaouet, P., Edwards, M., Hofle, M.G., Brettar, I., Colwell, R.R., Pruzzo, C., 2016. Climate influence on Vibrio and associated human diseases during the past half-century in the coastal North Atlantic. Proceedings of the National Academy of Sciences of the United States of America 113, E5062-5071.

Yue, S., Wang, C., 2004. The Mann-Kendall Test Modified by Effective Sample Size to Detect Trend in Serially Correlated Hydrological Series. Water Resources Management 18, 201-218.

Zimmerman, A.M., DePaola, A., Bowers, J.C., Krantz, J.A., Nordstrom, J.L., Johnson, C.N., Grimes, D.J., 2007. Variability of total and pathogenic Vibrio parahaemolyticus densities in northern Gulf of Mexico water and oysters. Applied and environmental microbiology 73, 7589-7596.