

## Checklist

- Composition jury
- Dédicace(?)
- Remerciements
- Avant-propos
- Résumé de la thèse (fr)
- Résumé de la thèse (en)
- Liste des abréviations(?)
- Introduction générale
- Chapitre 1 : Ecology Letters paper
- Chapitre 2 : Frontiers in Marine Sciences - Drivers
- Chapitre 3 : iEat
- Chapitre 4 : Cumulative impacts on food webs
- Conclusion générale
- Annexe 1 : Supplementary information chapitre 1
- Annexe 2 : Supplementary information chapitre 2
- Annexe 3 : Supplementary information chapitre 3
- Annexe 4 : Supplementary information chapitre 4
- References

## Introduction générale

- Réviser introduction générale
- Mettre à jour avec la littérature récente
- Mettre à jour les objectifs de la thèse

## Chapitre 1

- Contexte scientifique (Thèse)
- Publication associée (Thèse)
- Traduction du résumé de l'article publié (Thèse)
- Proposal letter Ecology Letters - Ideas and Perspectives
- Cover letter and novelty statement
- Conflict of interest statement
- Statement of authorship
- Data accessibility statement
- Reviewers
- Keywords
- Abstract
- Introduction
- Of food webs and multiple disturbances (concept)
- Simulations
- Sensitivity
- Amplification
- Food web sensitivity & amplification
- Conclusions
- Acknowledgements
- References
- Figure 1 - Concept
- Figure 2 - Sensitivity
- Figure 3 - Amplification
- Figure 4 - Food web scores table
- Figure 5 - Topological ~ Realised scores
- Figure 6 - Scores ~ Trophic level & degree
- Table S1 - Systems of equations

- Article formatting

## Chapitre 2

- Retranscrire article
- Contexte scientifique (Thèse)
- Publication associée (Thèse)
- Traduction du résumé de l'article publié (Thèse)
- Références
- Table
- Figures
- Box
- keywords

## Chapitre 3

- Retranscrire article
- Contexte scientifique (Thèse)
- Publication associée (Thèse)
- Traduction du résumé de l'article publié (Thèse)
- Références
- keywords

## Chapitre 4

- Données
- Catalogue interactions
- HMSC
- Probabilité d'exposition

- Rapport nouveaux développements en océanographie
- Plan article
- Article de soumission

### **Conclusion générale**

- Plan conclusion générale
- Brainstorm



Université du Québec  
à Rimouski

## **ÉVALUATION DES IMPACTS CUMULÉS SUR LES RÉSEAUX TROPHIQUES**

**Le cas de l'estuaire et du golfe du Saint-Laurent**

Thèse présentée  
dans le cadre du Programme de doctorat en océanographie  
en vue de l'obtention du grade de Philosophiae Doctor

PAR

**©DAVID BEAUCHESNE**

**Décembre 2019**



**Composition du jury :**

**QQ1, président du jury, Université du Québec à Rimouski**

**Phillipe Archambault, directeur de recherche, Université Laval**

**Dominique Gravel, codirecteur de recherche, Université de Sherbrooke**

**Jean-Claude Brêthes, examinateur interne, Université du Québec à Rimouski**

**QQ1, examinateur externe, QQpart**

UNIVERSITÉ DU QUÉBEC À RIMOUSKI

Service de la bibliothèque

Avertissement

La diffusion de ce mémoire ou de cette thèse se fait dans le respect des droits de son auteur, qui a signé le formulaire « *Autorisation de reproduire et de diffuser un rapport, un mémoire ou une thèse* ». En signant ce formulaire, l'auteur concède à l'Université du Québec à Rimouski une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de son travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, l'auteur autorise l'Université du Québec à Rimouski à reproduire, diffuser, prêter, distribuer ou vendre des copies de son travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de la part de l'auteur à ses droits moraux ni à ses droits de propriété intellectuelle. Sauf entente contraire, l'auteur conserve la liberté de diffuser et de commercialiser ou non ce travail dont il possède un exemplaire.

à qui tu veux,

## ***REMERCIEMENTS***

La j'ai fait des mercis personnels

## ***AVANT-PROPOS***

La j'ai fait des merci formels!

## ***RÉSUMÉ***

La biogéographie est l'étude des mécanismes et des processus qui influencent la répartition géographique des êtres vivants.

Mots clés: Biogéographie, interactions biotiques, réseaux écologiques, contraintes abiotiques, co-occurrence, théorie de la biogéographie des îles, théorie métabolique de l'écologie.

## ***ABSTRACT***

Biogeography is the study of the mechanisms and processes that control the geographical distribution of plants and animals.

Keywords: Biogeography, biotic interactions, ecological networks, abiotic constraints, co-occurrence, theory of island biogeography, metabolic theory of ecology.

## ***TABLE DES MATIÈRES***

Checklist . . . . .	i
Introduction générale . . . . .	i
Chapitre 1 . . . . .	ii
Chapitre 2 . . . . .	iii
Chapitre 3 . . . . .	iii
Chapitre 4 . . . . .	iii
Conclusion générale . . . . .	iv
REMERCIEMENTS . . . . .	x
AVANT-PROPOS . . . . .	xi
RÉSUMÉ . . . . .	xii
ABSTRACT . . . . .	xiii
TABLE DES MATIÈRES . . . . .	xiv
LISTE DES TABLEAUX . . . . .	xv
LISTE DES FIGURES . . . . .	xvi
LISTE DES ABRÉVIATIONS . . . . .	xvii
INTRODUCTION GÉNÉRALE . . . . .	1
Introduction . . . . .	1
ARTICLE 1	
TITRE CHAPITRE 1 . . . . .	2
Résumé en français du deuxième article . . . . .	2
Contexte scientifique . . . . .	2
Publication associée . . . . .	2
Traduction du résumé de l'article publié . . . . .	2
Introduction . . . . .	2
Mat & Meth . . . . .	2

ARTICLE 2	
GESTION DE NOUVELLE GÉNÉRATION - STRUCTURER ET PARTAGER LES DONNÉES DE PRESSIONS ENVIRONNEMENTALES POUR LE SY- TÈME DU SAINT-LAURENT . . . . .	3
Résumé en français du deuxième article . . . . .	3
Contexte scientifique . . . . .	3
Publication associée . . . . .	3
Traduction du résumé de l'article publié . . . . .	3
Title . . . . .	3
Authors . . . . .	3
Abstract . . . . .	4
Introduction . . . . .	5
Materials and Methods . . . . .	7
Estuary and Gulf of St. Lawrence . . . . .	7
Drivers . . . . .	8
Cumulative exposure . . . . .	9
Driver interactions . . . . .	10
Threat complexes . . . . .	10
Results and Discussion . . . . .	12
Cumulative exposure . . . . .	12
Driver interactions . . . . .	13
Threat complexes . . . . .	14
Open Knowledge Platform: <u>eDrivers</u> . . . . .	16
Unity and inclusiveness . . . . .	16
Findability, accessibility, interoperability and reusability . . . . .	18
Adaptiveness . . . . .	19
Recognition . . . . .	20
Perspectives . . . . .	20
Acknowledgements . . . . .	22

Author contributions statement . . . . .	22
Conflict of interest statement . . . . .	23
Listings . . . . .	24
Figures . . . . .	26
Tables . . . . .	32
<b>ARTICLE 3</b>	
PRÉDIRE LES INTERACTIONS BIOTIQUES AU SEIN DE MILIEUX PAUVRES EN DONNÉES . . . . .	36
Résumé en français du deuxième article . . . . .	36
Contexte scientifique . . . . .	36
Publication associée . . . . .	36
Traduction du résumé de l'article publié . . . . .	36
Title . . . . .	36
Authors . . . . .	36
Abstract . . . . .	36
Introduction . . . . .	37
Methods . . . . .	39
Biotic interaction catalogue . . . . .	39
Unsupervised machine learning . . . . .	40
Predicting interactions . . . . .	41
Algorithm prediction accuracy . . . . .	42
Results . . . . .	44
Biotic interaction catalogue . . . . .	44
Algorithm predictive accuracy . . . . .	44
Southern Gulf of St. Lawrence . . . . .	45
Discussion . . . . .	46
Algorithm accuracy . . . . .	46
Usefulness of taxonomic relatedness . . . . .	46
Interactions classification . . . . .	47

Southern Gulf of St. Lawrence . . . . .	48
Perspectives . . . . .	49
Acknowledgements . . . . .	50
Box 1 . . . . .	51
Figures . . . . .	53
ARTICLE 4	
TITRE CHAPITRE 4 . . . . .	58
Résumé en français du deuxième article . . . . .	58
Contexte scientifique . . . . .	58
Publication associée . . . . .	58
Traduction du résumé de l'article publié . . . . .	58
Introduction . . . . .	58
Mat & Meth . . . . .	58
CONCLUSION GÉNÉRALE . . . . .	59
Brainstorm . . . . .	59
Une belle conclusion générale . . . . .	59
To conclude . . . . .	59
ANNEXE I	
ANNEXE I . . . . .	60
ANNEXE II	
GESTION DE NOUVELLE GÉNÉRATION - STRUCTURER ET PARTAGER LES DONNÉES DE PRESSIONS ENVIRONNEMENTALES POUR LE SY- TÈME DU SAINT-LAURENT . . . . .	61
Drivers description . . . . .	61
Climate . . . . .	62
Coastal . . . . .	70
Fisheries . . . . .	74
Marine traffic . . . . .	76
Driver intensity and distribution . . . . .	78
Cumulative exposure . . . . .	81

Threat complexes . . . . .	82
Clustering . . . . .	82
Inter-cluster dissimilarity . . . . .	83
Intra-cluster similarity . . . . .	84
ANNEXE III	
ANNEXE III . . . . .	85
ANNEXE IV	
ANNEXE IV . . . . .	92
RÉFÉRENCES . . . . .	92

## ***LISTE DES TABLEAUX***

1	List of drivers currently available on eDrivers and used for the analyses presented in this paper. . . . .	32
2	List of drivers currently available on eDrivers along with their respective acronym used in the figures in the supplementary material. . . . .	61
3	List of functional groups included in the dataset presented in ? with their taxa composition. Only taxa that were at least at the scale of the family were used to predict interactions. List adapted from ?. . . . .	87
4	List of functional groups for which interactions were predicted by the algorithm, but not observed in ? (b). . . . .	90
5	List of functional groups for which observed interactions in ? were not predicted by the algorithm (c) . . . . .	91

## ***LISTE DES FIGURES***

1	Description of the Estuary and Gulf of St. Lawrence in Eastern Canada . . . . .	26
2	Distribution of cumulative footprint in the St. Lawrence System. . . . .	27
3	Distribution of cumulative hotspots in the St. Lawrence System. . . . .	28
4	Interaction between the intensity of hypoxia and demersal destructive fisheries in the St. Lawrence. An index of hypoxia ( <i>A</i> ) was created using bottom-water dissolved oxygen between 2013 and 2017 (Blais et al., 2018). Demersal destructive fisheries ( <i>i.e.</i> trawl and dredges) ( <i>B</i> ) intensity was evaluated from fisheries catch data collected between 2010 and 2015 used to measure annual area weighted total biomass ( <i>kg</i> ) in $1 \text{ km}^2$ grid cells (DFO, 2016b). See supplementary materials for more information on specific methodologies. Relative hypoxic stress and demersal destructive fisheries intensity was summed ( <i>C</i> ) to visualize their combined spatial distribution and intensity. Finally, individual density and the co-intensity of hypoxia and demersal destructive fisheries was investigated with a two-dimensional kernel analysis ( <i>D</i> ). . . . . .	29
5	Distribution of threat complexes in the Estuary and Gulf of St. Lawrence (upper panel). Threat complexes (a term coined by Bowler et al., 2019) are areas with similar cumulative driver exposure regimes. Mean intensity of all coastal (red), climate (green), fisheries (blue) and marine traffic (purple) drivers within each threat complex (lower panel). Refer tp Table 1 for acronym meaning and to the Supplementary Materials for more details. . . . . . . . . . . . . . . . .	30

- 6 Diagram of the platform structure. Community input in the form of raw data is accessed through the St. Lawrence Global Observatory (SLGO; <https://ogsl.ca/en>) repository - the platform host - or through open access repositories (*e.g.* NASA data). The raw data are then processed through a workflow hosted on the *eDrivers* GitHub organization (<https://github.com/orgs/eDrivers/>). Data processing may be as simple as data rescaling (*e.g.* night lights) or make use of more complex methodologies (*e.g.* acidification). All data is then hosted on SLGO's geoserver and accessible through their API. We developed a R package called *eDrivers* to access the driver layers through R and we are actively developing a second R package called *eDriversEx* that will include analytical tools to explore drivers data. Finally, we have developed a Shiny application that allows users to explore drivers data interactively (<https://david-beauchesne.shinyapps.io/eDriversApp/>). All R components of the project are hosted and available on the *eDrivers* GitHub organization. . . . .
- 7 Description of 17 logical steps (S1-S17) used by the algorithm to suggest a list of candidate resources ( $C_R$ ) for each consumer taxa ( $T_C$ ) in a set of  $N_1$  for which interactions are predicted, using a set of taxa  $N_0$  with empirically described interactions. Interactions between consumer and resource taxa are denoted as  $I(T_C, T_R)$ .  $K$  is the number of most similar neighbours selected for the KNN algorithm;  $t$  stands for tanimoto in equation 1;  $w_t$  is the weight given to sets of resources and consumers in equation 2; the minimum threshold is a value setting the minimal similarity value accepted for taxa to be considered as close neighbours in the KNN algorithm; the weight is the value added to a candidate resource each time it is added to  $C_R$ ; the minimum weight is the minimal weight value accepted for candidate resources to be selected as predicted sources in the algorithm. . . . .

31

54

- 8 Representation of the three statistics (*i.e.*  $Score_y$ ,  $Score_{\neg y}$  and TSS) used to evaluate the accuracy of the algorithm as a function of  $K$  values tested (*i.e.* 2, 4, 6 and 8 most similar neighbours, top  $x$ -axis) and weight for taxonomy (bottom  $x$ -axis), which varies between 0 and 1. A weight of 0 means that similarity is measured only using set of resources/consumers for each taxa, while a weight of 1 means that similarity is based solely on taxonomy. For each statistic, the topmost panel presents prediction accuracy for the complete algorithm, the middle panel corresponds to predictions made through the predictive portion of the algorithm (Steps S4-S16; Figure 7) and the bottom panel presents the catalogue contribution for the algorithm (Steps S1-S3; Figure 7). Note that the sum of the predictive and catalogue contributions can be over 100% as there is overlap between predictions made through both. The 7 datasets used for this analysis contained over 50 taxa (???????) . . . . . 55
- 9 Representation of  $Score_y$  as a function of catalogue comprehensiveness, *i.e.* the amount of information on sets of consumer and resources available in the catalogue. The sensitivity of the algorithm to data accuracy was evaluated with the arctic food web from ?. This food web was highly detailed taxonomically. Once removed from the catalogue, almost 100% of its taxa still had information available on sets of consumers and resources, which is necessary for testing the impact of catalogue comprehensiveness on prediction accuracy. A random percentage of data available in the catalogue for taxa in the food web (*i.e.* 0 to 100%) was iteratively removed ( $n = 50$  randomizations) before generating new predictions with the algorithm.  $w_t$  values of 0.5 and 1 were evaluated to verify the usefulness of taxonomy in supporting predictive accuracy. The topmost panel presents prediction accuracy for the complete algorithm, the middle panel corresponds to predictions made through the predictive portion of the algorithm (Steps S4-S16; Figure 7) and the bottom panel presents the catalogue contribution for the algorithm (Steps S1-S3; Figure 7). Note that the sum of the predictive and catalogue contributions can be over 100% as there is overlap between predictions made through both. . . . . 56
- 10 Example of predicted interactions with the network of the southern Gulf of St. Lawrence (?), centered around the interactions of the capelin (*Mallotus villosus*) and piscivorous small pelagic feeders (*e.g.* *Scomber scombrus* and *Illex illecebrosus*). Edge with colors green were both predicted and observed (26), black were observed only (3) and blue were predicted only (19). Arrows are pointed towards consumers. . . . . 57
- 11 Frequency distribution of the untransformed data for all driver layers. . . . . 79

12	Distribution and intensity of transformed and normalized drivers in the Estuary and Gulf of St. Lawrence available on *eDrivers*. . . . .	80
13	Frequency distribution of cumulative exposure ( <i>i.e.</i> sum of normalized driver intensity in each grid cell) and percent contribution of each driver to the frequency distribution of cumulative exposure in the Estuary and Gulf of St. Lawrence. . . . .	81
14	Validation procedure for the <i>k-medoids</i> and <i>k-means</i> clustering algorithms based on the number of cluster that maximizes average silhouette width (upper panels; Kaufman and Rousseeuw, 1990) and minimizes the total within-cluster sum of squares (WSS; lower panels). . . . .	82
15	Evaluation of inter-cluster dissimilarity using a similarity percentage analysis (SIMPER) with Manhattan distance (Clarke, 1993). The figure diagonal presents the distribution of the 6 clusters identified using the <i>k-medoids</i> clustering algorithm. The lower triangle shows all combinations of inter-cluster dissimilarity with circular barplots showing the percent contribution to total dissimilarity of each driver and with the total inter-cluster dissimilarity in the center of the barplots. The upper triangle shows the average relative intensity of each driver for all driver combinations, with barplots to the left and the right representing the row and columns clusters, respectively. . . . .	83
16	Evaluation of intra-cluster similarity using the Manhattan distance transformed to a similarity index. The distribution of the 6 clusters is presented along with circular barplots showing the percent contribution to total similarity of each driver and with the total intra-cluster similarity in the center of the barplots. . . . .	84

## ***LISTE DES ABRÉVIATIONS***

**DOI** : *Digital Object Identifier*; identifiant numérique d'objet.

**GIEC** : Groupe d'experts Intergouvernemental sur l'Évolution du Climat.

**IPBES** : *Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*; Plateforme intergouvernementale sur la biodiversité et les services écosystémiques.

## **INTRODUCTION GÉNÉRALE**

### **Introduction**

A citation (?)

## **ARTICLE 1**

### **TITRE CHAPITRE 1**

**Résumé en français du deuxième article**

**Contexte scientifique**

**Publication associée**

**Traduction du résumé de l'article publié**

**Introduction**

**Mat & Meth**

## **ARTICLE 2**

# **GESTION DE NOUVELLE GÉNÉRATION - STRUCTURER ET PARTAGER LES DONNÉES DE PRESSIONS ENVIRONNEMENTALES POUR LE SYSTÈME DU SAINT-LAURENT**

**Résumé en français du deuxième article**

**Contexte scientifique**

**Publication associée**

**Traduction du résumé de l'article publié**

**Title**

Next Generation Planning - Structuring and Sharing Environmental Drivers Data for the St. Lawrence System

**Authors**

David Beauchesne, Rémi M. Daigle, Steve Vissault, Dominique Gravel, Andréane Bastien, Simon Bélanger, Pascal Bernatchez, Marjolaine Blais, Hugo Bourdages, Clément Chion, Peter S. Galbraith, Benjamin Halpern, Camille Lavoie, Christopher W. McKinsey, Alfonso Mucci, Simon Pineault, Michel Starr, Anne-Sophie Ste-Marie, Philippe Archambault

## Abstract

The St. Lawrence is a vast and complex socio-ecological system providing a wealth of services sustaining numerous economic sectors. These ecosystems are subject to significant human pressures that overlap and potentially interact with climate driven environmental changes. Our objective in this paper is to systematically characterize the distribution and intensity of drivers in the St. Lawrence System. To do so, we launch eDrivers, an open knowledge platform gathering experts committed to structuring, standardizing and sharing knowledge on drivers in support of science and management. We gathered data on 22 coastal, climate, fisheries and marine traffic drivers through collaborations, existing environmental initiatives and open data portals. We show that few areas of the St. Lawrence are free of cumulative exposure. The Estuary, the Anticosti Gyre and coastal areas are particularly exposed, especially in the vicinity of urban centers. We identified 6 areas of distinct cumulative exposure regime that show that certain drivers typically co-occur in different regions of the St. Lawrence and that coastal areas are exposed to all driver types. Of particular concern are two threat complexes capturing most exposure hotspots that show the convergence of contrasting exposure regimes at the head of the Laurentian Channel. eDrivers was built on a series of guiding principles upholding existing data management and open science standards. We therefore expect it to evolve through time to address knowledge gaps and refine current driver layers. Ultimately, we believe that eDrivers represents a much needed solution that could radically influence broad scale research and management practices by increasing knowledge accessibility and interoperability.

**Keywords:** ocean observing systems, St. Lawrence, environmental drivers, cumulative exposure, threat complex, multiple stressors, global change

## Introduction

The St. Lawrence System, formed by one of the largest estuaries in the world, the St. Lawrence Estuary, and a vast interior sea, the Gulf of St. Lawrence, is a complex social-ecological system characterized by highly variable environmental conditions and oceanographic processes (???). It constitutes a unique and heterogeneous array of habitats suited for the establishment of diverse and productive ecological communities (?). As a result, the St. Lawrence System provides a wealth of ecosystem services that have historically and contemporarily benefited the Canadian economy. It sustains a rich fisheries industry targeting more than 50 species, serves as the gateway to eastern North-America by granting access to more than 40 ports and the most densely populated Canadian region, hosts a booming tourism industry and an expanding aquaculture production, fosters emerging activities and boasts a yet untapped hydrocarbon potential (???). With major investments recently made and more forthcoming in economic and infrastructure development and research (e.g. ??), an intensification of the human footprint is likely in the St. Lawrence System.

As elsewhere in the world (see ?), this intensifying human footprint will likely result in increasingly intricate environmental exposure regimes, i.e. suites of overlapping and potentially interacting environmental drivers threatening ecosystems, habitats or ecological communities. Drivers, often referred to as stressors or pressures, are any externalities that affect environmental processes and disturb natural systems. Drivers may originate from natural or human-induced biophysical processes (e.g. sea surface temperature anomalies and hypoxia) or directly from anthropogenic activities (e.g. fisheries and marine pollution). The potential for complex interactions between drivers is the largest uncertainty when studying or predicting environmental impacts (??). Multiple drivers can combine non-linearly and result in effects that are greater (synergistic effect) or lower (antagonistic effect) than the sum of individual effects (???).

The uncertainty associated with complex driver interactions must therefore be taken

into account when investigating environmental impacts (?), yet most research on driver effects in marine environments remains overwhelmingly focused on single driver assessments (?). Increasing exposure and the experiences of past ecological tragedies such as the collapse of cod fisheries (??) and the decline of the beluga and right whale populations (?) together urge the need to characterize the distribution, intensity and overlap between drivers in the St. Lawrence System. This will provide critical information on areas most exposed to cumulative drivers and on the interaction potential of drivers in the St. Lawrence. It is also a necessary step towards the holistic and integrated management of the St. Lawrence System.

Gathering environmental data for large scale, systematic initiatives can, however, be a very challenging and time consuming – not to say painful – process. On one hand, there is an overwhelming and expanding wealth of data available. Such information overload may inhibit our ability to make decisions based on scientific information, promote massive effort duplication, disproportionately appropriate research funds to certain sectors, and obscure knowledge gaps amid a sea of information (?). On the other hand, crucial data are lacking and remain largely unavailable or inaccessible for a variety of reasons, including proprietary rights, lack of organizational time, capacity and training, and in some rare cases unwillingness to share, curtailing our ability for appropriate decision-making.

There are now initiatives that address this issue by assembling, organizing and sharing environmental knowledge, such as the Ocean Biogeographic Information System (OBIS; ?) for biotic data and Bio-ORACLE (?) for abiotic data. However, equivalent platforms for environmental drivers have largely focused on single drivers (e.g. Global Fishing Watch) and platforms collating data and knowledge on multiple drivers in a comparable and interoperable way remain conspicuously missing (but see ?). This is in spite of integrated management and assessment approaches needing efficient data reporting, standardized data management practices and tools tailored to the study of the effects of multiple drivers (??). An additional objective thus emerged in the process of addressing

our initial goals: sharing the knowledge gathered through the description of drivers in the St. Lawrence.

Our main objective in this study is to characterize the distribution and intensity of drivers in the St. Lawrence. More specifically, we aim to: 1) identify areas of high cumulative exposure, 2) identify drivers that are likely to interact in the St. Lawrence and 3) characterize areas with similar cumulative exposure regimes. Here we focus solely on drivers, one of four elements required to evaluate environmental impacts, *i.e.* drivers, ecosystem components of interest (*e.g.* habitats or species) and the exposure and the vulnerability of the ecosystem components to drivers. We achieve these objectives with the development of an open knowledge platform named eDrivers. The platform was designed to facilitate collaboration, real-time assessments of cumulative exposure and to evolve with the addition of information and threats to the St-Lawrence ecosystems.

## Materials and Methods

### Estuary and Gulf of St. Lawrence

The St. Lawrence System is formed by the Estuary and the Gulf of St. Lawrence (Figure 1). The Estuary is defined by the seawater limit close to Îles d’Orléans to the west and by its connexion to the Gulf near Pointe-des-Monts. Estuarine circulation characterizes the Estuary. The surface layer is composed of a freshwater seaward outflow primarily from the Great Lakes Basin and through the St. Lawrence River, upstream of the Estuary. Atlantic waters are flowing mainly from the Cabot Strait landward through the Laurentian Channel (see below).

The Gulf is an interior sea connected to the Atlantic by the Cabot and Belle-Isle Straits to the south and north of Newfoundland, respectively. The topology of the Northern Gulf is characterized by three deep channels (250-500 m). The Laurentian Channel is the main channel connecting the Estuary to the Atlantic through the Cabot Strait.

The Esquiman and Anticosti Channels are two secondary channels that branch off from the Laurentian Channel to the north towards the Arctic and the north of Anticosti Island, respectively. The Southern Gulf is characterized by the Magdalen Shallows, a vast area with an average depth of ~50 m. The Gulf is composed of a seasonal cold intermediate layer that separates the surface and deep layers. Seasonal sea ice affects the circulation of the St. Lawrence. Finally, three islands impact the physical dynamics of the St. Lawrence: the Anticosti Island to the north, the Magdalen Islands in the middle of the Magdalen Shallows and Prince Edward Island to the south. See ? and ? for more information on the physical oceanography of the St. Lawrence.

The St. Lawrence drains over 25% of global freshwater reserves through its connexion to the Great Lakes Basin. The Basin is home to over 45 million North-Americans, i.e. 15 and 30 million in Canada and the United States, respectively (?). The St. Lawrence itself boasts a much lower population of approximately 1 million Canadians living within 10 km of the coast, with populations mainly located in a few coastal cities in the Estuary and the Southern Gulf (?).

## Drivers

The list of drivers for which we sought data was informed by a global cumulative impact assessment initiative (??), regional holistic evaluations of the state of the St. Lawrence (??), and communications with regional experts. We integrated datasets from regional experts and also use global data from the global cumulative impact assessment initiative (??) available from the National Center for Ecological Analysis and Synthesis (NCEAS) online data repository (Table 1; ?). We selected global data that were unavailable at the regional scale and that were available at a resolution adequate for use at the scale of the St. Lawrence (e.g. marine pollution).

We characterized the intensity and distribution of 22 drivers (Table 1;). Drivers incorporated in the analyses are varied in origin, i.e. from terrestrial (e.g. nutrient input)

to marine (e.g. shipping), and from large scale biophysical processes (e.g. temperature anomalies) to localized anthropogenic activities (e.g. fisheries). Drivers were divided into 4 groups: coastal, climate, fisheries and marine traffic (Table 1). All data layers and methodologies are described in the supplementary materials.

Drivers with non-normal frequency distributions were log-transformed (Figure S1) and all drivers were scaled between 0 and 1 to allow driver comparisons. The 99th quantile of individual driver distribution was used as the upper bound for scaling to control for extreme values and produce maps of individual drivers (Figure S2). All drivers were embedded in a regular grid composed of 245604  $1\text{km}^2$  hexagonal cells to construct the integrated dataset used for the analyses.

### Cumulative exposure

Areas with high cumulative exposure (objective 1) were identified by comparing areas on the basis of the number and relative intensity of drivers in each grid cell. Throughout the text, we use exposure when describing areas exposed to drivers and we use footprint when describing the distribution and intensity of drivers.

Cumulative footprint ( $F$ ) was defined as the sum of the scaled intensity of all drivers in each grid cell:

$$F_x = \sum_{i=1}^n D_{i,x}$$

where  $x$  is a grid cell,  $i$  is a driver and  $D$  is the scaled intensity of driver  $i$ . The cumulative footprint provides an estimate of the total relative footprint in each grid cell. A grid cell with a high  $F$  value is either characterized by multiple drivers at low relative intensity, limited drivers at high relative intensity, or both.

We also identified cumulative hotspots ( $H$ ) to explore the the distribution of cumulative exposure in the St. Lawrence (objective 1). Cumulative hotspots ( $H$ ) were defined as the

number of drivers in each grid cell with scaled intensity contained over their respective 80th percentile:

$$H_x = \sum_{i=1}^n \mathbb{1}(D_{i,x} \in P_{80,D_i})$$

where,  $x$  is a grid cell,  $i$  is a driver and  $D$  is the scaled intensity of driver  $i$  and  $P_{80,D_i}$  is the 80th percentile of driver  $i$ . Hotspots thus identify areas where drivers are co-occurring at high relative intensities.

### **Driver interactions**

Interactions among drivers were investigated using the cumulative footprint ( $F$ ) between pairs of drivers (objective 2). The intensity at which pairs of drivers co-occur was evaluated using a two-dimensional kernel density estimate. As there are 231 pairwise combinations between 22 drivers, we focus on a single example using hypoxia and demersal destructive fisheries, two drivers known to occur mainly in deeper areas of the St. Lawrence and, hence, an interaction between the effects of the two drivers could be anticipated.

### **Threat complexes**

In order to better capture potential interactions between drivers (objective 2) and to identify areas with similar exposure regimes (objective 3), we identify threat complexes using a clustering approach (e.g. see ?). We use the term clusters in presenting the methods, but use threat complex when discussing the results on cumulative exposure regimes.

## Clustering

Threat complexes were identified using a partitional k-medoids clustering algorithm, CLARA (CLustering for Large Applications; ?), which was designed for large datasets. The CLARA algorithm uses the PAM (Partition Around Medoids) algorithm on a sample from the original dataset to identify a set of  $k$  objects that are representative of all other objects, i.e. medoids and that are central to the cluster they represent. The goal of the algorithm is to iteratively minimize intra-cluster dissimilarity. Iterations are compared on the basis of the average dissimilarity between cluster objects and representative medoid to select the optimal set of  $k$  medoids that minimizes average dissimilarity. We used the clustering algorithm with the Manhattan distance since this measure is less affected by extreme values (?), as is the k-medoids clustering algorithm (?). We used 100 iterations using samples of 10000 observations (i.e. ~5% of observations) to identify clusters. Analyses were performed using the cluster R package (?). Partitional clustering algorithms require a user-defined number of clusters. Values of  $k$  ranging from 2 to 10 were tested and validated by selecting the number of clusters that maximized the average silhouette width (?) and minimized the total within-cluster sum of squares (Figure S4).

## Inter-cluster dissimilarity

The difference between clusters was explored by measuring the total inter-cluster dissimilarity and the contribution of each driver to the total inter-cluster dissimilarity using a similarity percentage analysis (SIMPER) with Manhattan distance (Figure S5; ?). The Manhattan distance was again preferred for continuity with the clustering analysis and to ensure that outliers did not have a strong influence the analysis. As the drivers dataset is large (~ 250000 observations), we used a bootstrap procedure for the SIMPER analysis, randomly selecting 5% of each cluster to run the analysis and repeating the process over 300 iterations. We also compared the mean intensity of each

driver within each cluster to better capture the inter-cluster dissimilarity.

### Intra-cluster similarity

Intra-cluster similarity was evaluated calculating the intra-cluster Manhattan distance and by transforming the mean contribution to distance ( $M_c$ ) of each driver by  $.1/(.1 + M_c)$  to obtain a similarity measure for each driver ( $S_d$ ). The total similarity is the sum of all  $S_d$ . (Figure S6). As with the inter-cluster dissimilarity, we used a bootstrap procedure for the intra-cluster similarity, randomly selecting 25% of each cluster observation to run the analysis and repeating the process over 50 iterations. We only used the bootstrapping procedure for clusters with less than 40000 observations since computation time was manageable.

## Results and Discussion

### Cumulative exposure

Apart from the northeastern Gulf, the cumulative footprint of drivers is ubiquitous in the St. Lawrence (Figure 2). Cumulative exposure is generally highest along the coast (Figure 2), with hotspots located in the vicinity of coastal cities (Figure 3). In general, offshore areas are less exposed to cumulative drivers, with the Estuary and the Anticosti Gyre being notable exceptions (Figures 2 and 3). This is not to say that offshore areas are free of exposure, as most of the St. Lawrence is exposed to multiple overlapping drivers (Figures 2 and 3). For example, it is worthy to note high cumulative footprint observed at the heads of the Anticosti and Esquiman Channels (Figure 2).

These results are consistent with observations elsewhere in the world, where cumulative driver exposure conspicuously arises from and markedly intensifies close to coastal cities and at the mouth of rivers draining highly populated areas (e.g. ????). These are areas where human activities (e.g. coastal development and shipping) and footprint

(e.g. pollution runoff) are the most intense (?), and on which is overlaid a background of natural disturbances (?). They are also the areas in which the most dramatic increases in exposure are expected, with populations increasing more rapidly along the coast than inland (?). In the St. Lawrence, large coastal cities are mostly located along the Estuary and the southwestern Gulf, while the northeastern Gulf is largely uninhabited or home to small coastal communities. Certain smaller coastal communities with high cumulative footprint are characterized by large industries (e.g. Sept-Îles and Charlottetown).

As for offshore exposure, the Estuary, along with the St. Lawrence River, provide access to and serve as the primary drainage outflow of the Great Lakes Basin, which is home to over 45 million North-Americans and is the most densely populated region in Canada (??). Most marine traffic thus converges to the Estuary.

While we cannot ascertain that high exposure areas are the most impacted, we can safely predict that these are the areas where studying ecosystem state will be the most complex due to the uncertainty associated with driver interactions, an uncertainty bound to increase rapidly with the number of interacting drivers (?).

## Driver interactions

Hypoxia is mainly distributed in the Laurentian, Anticosti and Esquiman Channels, with the head of the Channels most exposed (Figure 4A). Demersal destructive fisheries are located along the Laurentian Channel, the heads of the Anticosti and Esquiman Channels and around the Magdalen Islands (Figure 4B). By combining both drivers, we can observe that hypoxia and demersal destructive fisheries overlap mostly at high relative intensity (Figure 4D) in the vicinity of the Anticosti Gyre and the heads of the Esquiman and Anticosti Channels (Figure 4C, Box 1).

Fisheries in the St. Lawrence have historically affected biodiversity distribution and habitat quality (?). Concurrently, hypoxia decreases overall habitat quality, but triggers

species-dependent responses ranging from adaptation (e.g. northern shrimp Pandalus borealis and Greenland halibut Reinhardtius hippoglossoides; ?) to reduced growth rates (?) and avoidance of oxygen-depleted habitats (e.g. Atlantic cod Gadus morhua; ?) to increased mortality (e.g. sessile benthic invertebrates; ???). Certain species may thus be adversely affected by fisheries and withstand hypoxia but still experience a decrease in prey availability, while others may be deleteriously affected by the compounded effect of both drivers (?).

### Threat complexes

While informative, the hypoxia-fisheries example focuses on a single pair of drivers and falls short of the number of drivers overlapping at high intensities throughout the St. Lawrence (Figure 3). The number of drivers overlapping in the St. Lawrence increases with cumulative exposure (Figure S3). Areas with high exposure such as the Estuary, the Anticosti Gyre and the southwestern Gulf coastline (Figure 2 and 3) are thus areas where driver interactions are most likely, and where they can arise between a host of different drivers. The identification of threat complexes provides a crucial tool to simplify the multi-dimensional complexity of overlapping drivers to areas exposed to similar suites of drivers (?). This may prove critical for a better understanding the state of species, habitats and ecosystems located within or moving through threat complexes and exposed to the combined effects of all drivers typical to those areas.

Six distinct threat complexes were identified in the St. Lawrence using the k-medoids algorithm (Figures S4, S5). Based on their distribution and representative drivers, threat complexes can be divided into 3 offshore and 3 coastal complexes (Figures 5, S6 and S7). Coastal threat complexes (1 to 3; Figure 5) include all types of drivers besides hypoxia and are the most exposed threat complexes, both in terms of driver overlap and intensity. Threat complex 1 encompasses the coastline and is characterized by higher direct human impact (i.e. population density). Threat complex 2 is differentiated from

other complexes by the presence of aquaculture sites. Threat complex 3 is the most exposed complex and has a distribution similar to the most exposed coastal hotspots (Figure 3). This complex is characterized by high intensities of land-based drivers (e.g. nutrient input), demersal non-destructive high-bycatch fisheries (e.g. trap fishing), climate drivers and marine traffic drivers in the vicinity of ports.

Offshore threat complexes (4 to 6; Figure 5) are generally characterized by high intensity climate and marine traffic drivers. Threat complex 4 is differentiated by demersal non-destructive high-bycatch fisheries, higher marine traffic drivers compared to complex 5 and generally corresponds to the whole Southern Gulf. Threat complex 5 is characterized by more fisheries types (i.e. demersal destructive and pelagic high-bycatch), generally lower intensity marine traffic drivers and is located almost exclusively in the Northern Gulf. Finally, threat complex 6 is the most exposed offshore threat complex and includes all offshore hotspots (Figure 3). It is characterized by high intensity hypoxia, marine traffic and pollution, as well as demersal destructive and pelagic high-bycatch fisheries. This threat complex corresponds primarily to the Laurentian Channel and incorporates parts of the Esquiman and Anticosti Channels.

Of particular concern are threat complexes 3 and 6, which are the two most exposed complexes in the St. Lawrence and are characterized by distinct cumulative exposure regimes. Between them, they capture most of the coastal and offshore hotspots identified in the St. Lawrence. They also offer some insight into the potential importance of considering spatial dynamics in areas intersecting multiple threat complexes. For example, threat complexes 3 and 6 meet at the mouth of the River Saguenay. This area is particularly dynamic, with deep Atlantic waters advected through estuarine circulation mixing with surface waters from the St. Lawrence River and the Saguenay River (?). This results in the convergence of climate drivers from the bottom of the Laurentian Channel and marine traffic drivers (threat complex 6) with terrestrial run-off from river outflows and direct human impacts (i.e. population density; threat complex 3).

## Open Knowledge Platform: eDrivers

Sharing the knowledge acquired through the description of drivers in the St. Lawrence emerged as a priority to curtail the need to reach dozens of experts across multiple organizations and over extensive periods of time to assemble the data needed to apply integrated research and management. It is also a requirement to ensure that this manuscript will not quickly become an outdated snapshot of drivers distribution and intensity in the St. Lawrence System, but rather serve as a stepping stone towards an adaptive and ever-improving collection of knowledge.

As such, we are launching eDrivers, an open knowledge platform focused on sharing knowledge on the distribution and intensity of drivers and on gathering a community of experts committed to structuring, standardizing and sharing knowledge on drivers in support of science and management. In launching this initiative, our objective is to uphold the highest existing standards of data management and open science. We identified four guiding principles to meet this objective and that guide the structure of the platform (Figure 6).

### Unity and inclusiveness

**Why:** Operating over such large scales in time, space and subject matter requires a vast and diverse expertise that cannot possibly be possessed by any one individual or organization. Consequently, we envision an initiative that seeks to mobilize all individuals and entities with relevant expertise.

**How:** By promoting, consolidating and working with experts involved in existing and highly valuable environmental initiatives already in place in the St. Lawrence. Notable examples of environmental initiatives are the annual review of physical (?), chemical, and biological (?) oceanographic conditions in the St. Lawrence, the fisheries monitoring program (?), the annual groundfish and shrimp multidisciplinary survey (?), the

characterization of benthic (?), epipelagic and coastal (?) habitats of the St. Lawrence, and Canada's shoreline classification (?). There are also nascent efforts to share information on several human activities in the St. Lawrence such as the Marine Spatial Data Infrastructure portal, which provides data on zoning, shipping, port activities, and other human activities in Canadian waters, including the St. Lawrence system (?).

By working with existing data portals whose objective is to share environmental data. We are thus collaborating actively with the St. Lawrence Global Observatory (SLGO) to develop the initiative and to host the platform on their web portal. The mission of SLGO is to promote and facilitate the accessibility, dissemination and exchange of official and quality data and information on the St. Lawrence ecosystem through the networking of organizations and data holders to meet their needs and those of users, to improve knowledge and to assist decision-making in areas such as public safety, climate change, transportation, resources and biodiversity conservation. SLGO is also one of three regional associations spearheading the Canadian Integrated Ocean Observing System (CIOOS<sup>1</sup>), which will focus on integrating oceanographic data from multiple sources to make them accessible to end-users and to enable the national coordination of ocean observing efforts by integrating isolated or inaccessible data, and by identifying gaps or duplications in observations and research efforts. We are also developing collaborations with the Portal on water knowledge<sup>2</sup>, an initiative from the Québec provincial government. This portal aims at collecting and sharing accurate, complete and updated resources on water and aquatic ecosystems to support the mandate of relevant actors and stakeholders working in water and aquatic ecosystems management.

By actively inviting, seeking, and developing collaborations as well as encouraging constructive criticism from the inception and throughout the lifetime of the platform.

---

1. <https://cioos.ca>

2. <http://www.environnement.gouv.qc.ca/eau/portail/>

By inviting external community contributions (Figure 6). External researchers or entities wishing to submit marine data will be able to do so through SLGO web portal. Submissions through other data portals will also be accepted either through the development of data sharing agreements or with the caveat that shared data are under an open source license and that they adhere to the platform data standards.

### **Findability, accessibility, interoperability and reusability**

**Why:** Open knowledge has been propelled to the forefront of scientific research in an era of open, collaborative and reproducible science. By moving towards large scale, cross-disciplinary research and management projects, there is a growing need to increase the efficiency of data discovery, access, interoperability and analysis (??). Our goal is to foster efficient and functional open science by creating a fully open, transparent and replicable open knowledge platform.

**How:** By building an infrastructure adhering to the FAIR Data Principles, which states that data and metadata must be Findable, Accessible, Interoperable and Reusable. These principles focus on the ability of humans and machines to automatically find and (re)use data and knowledge (?).

By making data and associated tools accessible through a variety of ways: the SLGO web portal, two R packages called [eDrivers](#)<sup>3</sup> and [eDriversEx](#)<sup>4</sup> to access the data through SLGO's API and to provide analytical tools to explore data, respectively, and a Shiny application<sup>5</sup> to explore drivers data interactively (Figure 6). Note that the data are currently contained within and accessible through the [eDrivers](#) R package only, as we

3. <https://github.com/orgs/eDrivers/eDrivers>

4. <https://github.com/orgs/eDrivers/eDriversEx>

5. <https://david-beauchesne.shinyapps.io/eDriversApp/>

are actively working to allow users to download selected layers from SLGO's web portal and geoserver. The functions available in [eDrivers](#) to access the data have however been developed to ensure forward compatibility once the data migrate to SLGO's geoserver.

By defining clear data and metadata standards and specifications to support the regional standardization of current and future protocols and practices and to favour interoperability with national and international initiatives like the Essential Ocean Variables (EOV) identified by the Global Ocean Observing System (GOOS). As such, we will adopt the metadata standard currently targeted for CIOOS, [i.e.](#) the North American Profile of ISO 19115:2014 - Geographic information - Metadata, a schema favoured for geospatial data in Canada and the United States.

By providing version control and code access to the workflows set up to generate driver layers from raw data, the R packages and the Shiny application through a GitHub organization called [eDrivers](#)<sup>6</sup>.

## Adaptiveness

**Why:** In the face of uncertainty and in an effort to address impending environmental changes, adaptive management has been identified as the chief strategy to guide efficient decision-making ([e.g.](#) [????](#)) and has already been discussed in the context of multi-drivers and cumulative impact assessments ([????](#)). Adaptive management can only be truly achieved through a commitment to adaptive monitoring and data reporting ([???](#)). We further contend that adaptive management requires the development of adaptive monitoring tools and infrastructures, which we seek to address through a continuously-evolving platform.

**How:** By setting up mechanisms structuring cyclic reviews of platform content, for the integration of new material ([e.g.](#) data and methods) as it becomes available or

---

6. <https://github.com/orgs/eDrivers/>

accessible, and by striving to provide time-series data that are crucial to assess temporal trends and potentially early-warning signals of ecosystem change (Figure 6).

## Recognition

**Why:** Like peer-reviewed publications, data must also be given its due importance in scientific endeavours and thus be considered as legitimate citable products contributing to the overall scientific output of data providers (??). Appropriate citations should therefore be provided for all data layers used and shared by the platform.

**How:** By adhering to the Data Citation Principles (?), which focus on citation practices that provide appropriate credit to data products.

## Perspectives

Understanding how ecosystem state will be affected by global change requires a comprehensive understanding of how threats are distributed and interact in space and time, which in turn hinges on appropriate data tailored to multi-driver studies (???). In the St. Lawrence, we found that few areas are free of cumulative exposure and that the whole Estuary, the Anticosti Gyre and coastal southwestern Gulf are particularly exposed to cumulative drivers, especially close to urban areas. We also identified six geographically distinct threat complexes that display similar cumulative exposure regimes. These complexes reveal that coastal areas are particularly exposed to all types of drivers and that multiple drivers typically co-occur in space. These results allow us to efficiently identify areas in need of heightened scrutiny from a science and management perspective.

Through eDrivers, these observations will be iteratively improved towards an increasingly robust assessment of cumulative exposure and threat complexes as gaps in knowledge are addressed or approaches to describe drivers are refined. Arguably, the most

meaningful benefit anticipated from eDrivers will be the gain in efficient access to comparable data-based knowledge on the exposure of ecosystems to multiple threats. This could pay quick scientific and management dividends by drawing on the knowledge and efforts of a wide range of contributors, by expanding avenues of scientific inquiry, by decreasing overall effort duplication and research costs, and by increasing research efficiency (?).

Critically, eDrivers will allow the scientific and governmental communities to identify key knowledge gaps that will assist in prioritizing and optimizing research efforts. Ultimately, we believe that eDrivers will operationalize evidence-based decision-making by streamlining data management and research, allowing science output to be available and interpretable on a time scale relevant to management (see ??). The platform will thus greatly facilitate the application of broad scale, holistic research and management approaches such as marine spatial planning, ecosystem-based management, marine spatial planning and strategic environmental assessments (e.g. ???).

While the focus of our paper has been on the description of drivers in the St. Lawrence, drivers are but one of the 4 elements required to properly evaluate environmental impacts. This single knowledge node should be weaved with other, comparable, nodes required for integrated management approaches such as species distribution and marine protected areas. Ultimately, all of these knowledge nodes could be weaved together through social-ecological metanetworks (?).

Significant effort is still needed to bring our vision to fruition. Foremost is to maintain our efforts to foster collaborations, develop platform content and identify key knowledge gaps. A fair and efficient organizational structure will be developed in order to manage eDrivers as a community and appropriate funding must be secured to continue building this community and ensure the long-term viability of the initiative, although the partnership with SLGO partly addresses this issue.

Finally, terrestrial and coastal environments must be incorporated, as sources of stress

within those habitats extend to the marine environments. Moreover, despite coastal areas being recognized as the most exposed to environmental threats, we continue to delineate terrestrial and marine realms, considering coastlines as an impermeable barrier. While there is a sensible rationale for this division, we must strive to eliminate it if we are to appropriately study and predict the impacts of global change (e.g. see ?).

Despite the challenges and work ahead, we are hopeful that this initiative will be very successful. Ultimately, eDrivers represents a much needed solution to address important issues in data management that could radically shift broad scale research and management practices towards efficient, adaptive and holistic ecosystem-based management in the St. Lawrence and elsewhere in the world. All it requires to be successful is for the scientific and political communities to fully commit to open knowledge, adaptive monitoring and, most of all, an integrated vision of ecosystem management.

## Acknowledgements

We thank the Fond de Recherche Québécois Nature et Technologie (FRQNT) and the Natural Science and Engineering Council of Canada (CRSNG) for financial support. This project is also supported by Québec Océan, the Quebec Centre for Biodiversity Science (QCBS), Takuvik, and the Notre Golfe networks. This research is also sponsored by the NSERC Canadian Healthy Oceans Network and its Partners: Department of Fisheries and Oceans Canada and INREST (representing the Port of Sept-Îles and City of Sept-Îles).

## Author contributions statement

DB, RD, DG and PA conceived the manuscript and the underlying objectives. DB prepared/formatted the data, performed the analyses, was in charge of technical developments and lead the drafting of the manuscript. All co-authors contributed to data,

analyses and writing based on their respective expertise and contributed to the revision of the manuscript.

### **Conflict of interest statement**

The authors declare that the submitted work was carried out in the absence of any personal, professional or financial relationships that could potentially be construed as a conflict of interest.

## Listings

Box 1. Code snippet demonstrating how to use the `eDrivers` R package to reproduce figure 4.

```
# Install and load eDrivers package
devtools::install_github('eDrivers/eDrivers')
library(eDrivers)

# Load data
drivers <- fetchDrivers(drivers = c('hypoxia','fishDD'))

# Get data from `eDrivers` class object
driverData <- getData(drivers)

# Normalize data
driverData <- driverData / cellStats(driverData, 'max')

# Visualize data and combination
plot(driverData$fishDD)    # Demersal destructive fisheries
plot(driverData$hypoxia)   # Hypoxia
plot(sum(driverData))      # Combination

# Identify values > 0 and not NAs
driverData$fishDD[driverData$fishDD < 0] <- NA
driverData$fishDNH[driverData$hypoxia < 0] <- NA
id0 <- !is.na(values(driverData$fishDD)) &
  !is.na(values(driverData$hypoxia))

# 2D kernel for driver co-intensity
library(MASS)
```

```
coInt <- kde2d(x = values(driverData$fishDD)[id0] ,  
                 y = values(driverData$hypoxia)[id0] ,  
                 n = 500 , lims = c(0 , 1 , 0 , 1))  
  
image(coInt , zlim = c(0 , max(coInt$z)))  
  
# Driver density distribution  
plot(density(driverData$fishDD[id0]))    # Demersal destructive  
plot(density(driverData$hypoxia[id0]))    # Hypoxia
```

## Figures

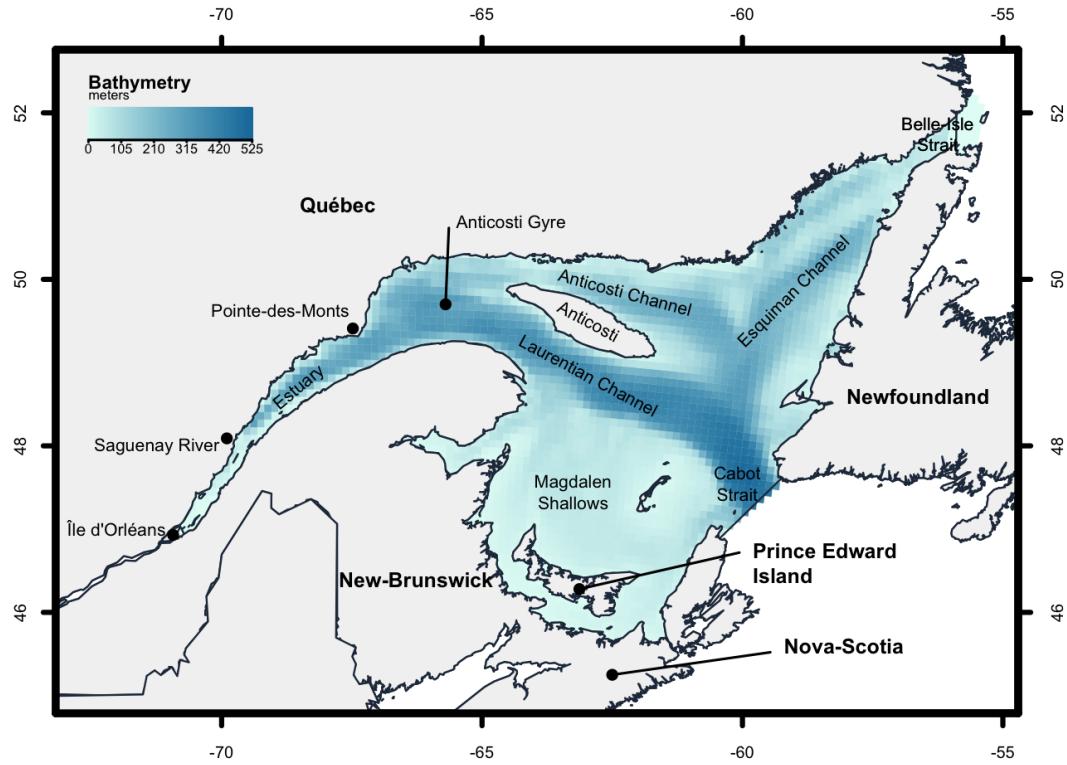


Figure 1: Description of the Estuary and Gulf of St. Lawrence in Eastern Canada

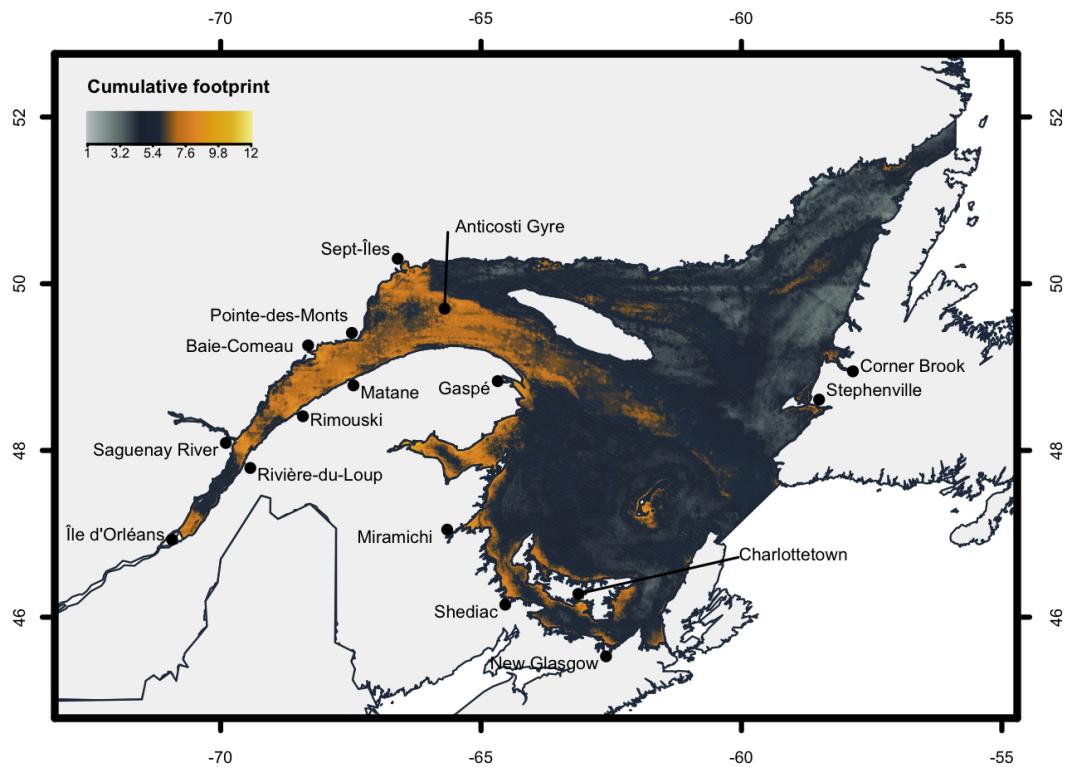


Figure 2: Distribution of cumulative footprint in the St. Lawrence System.

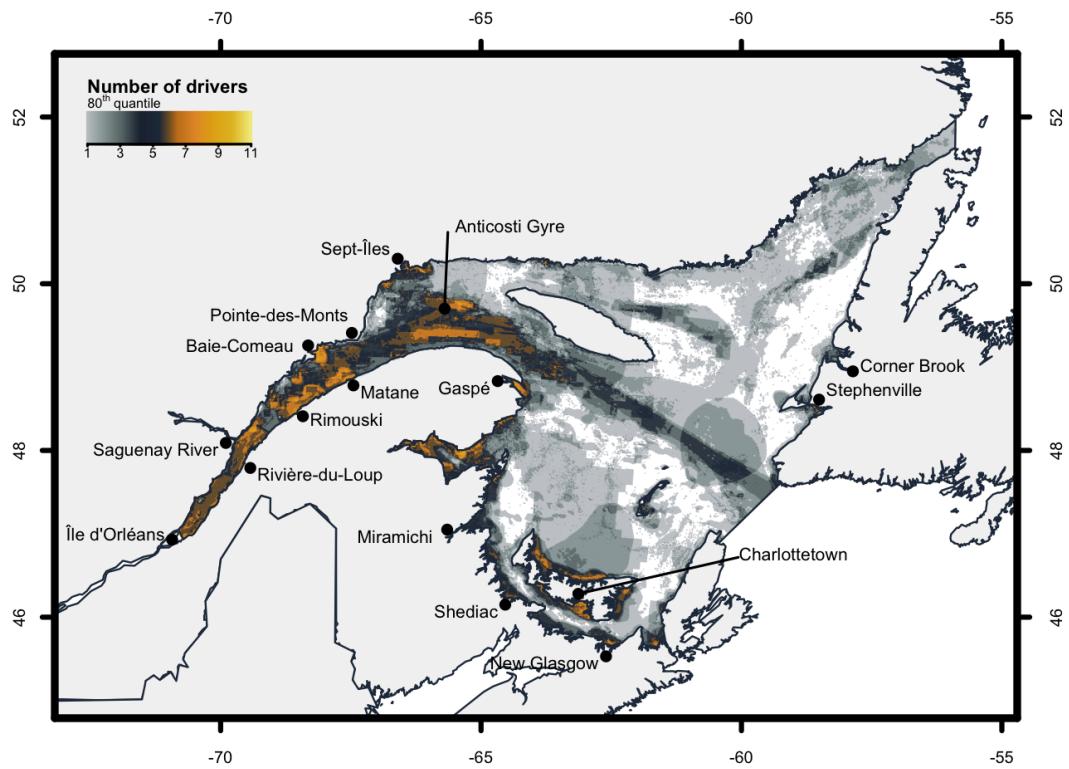


Figure 3: Distribution of cumulative hotspots in the St. Lawrence System.

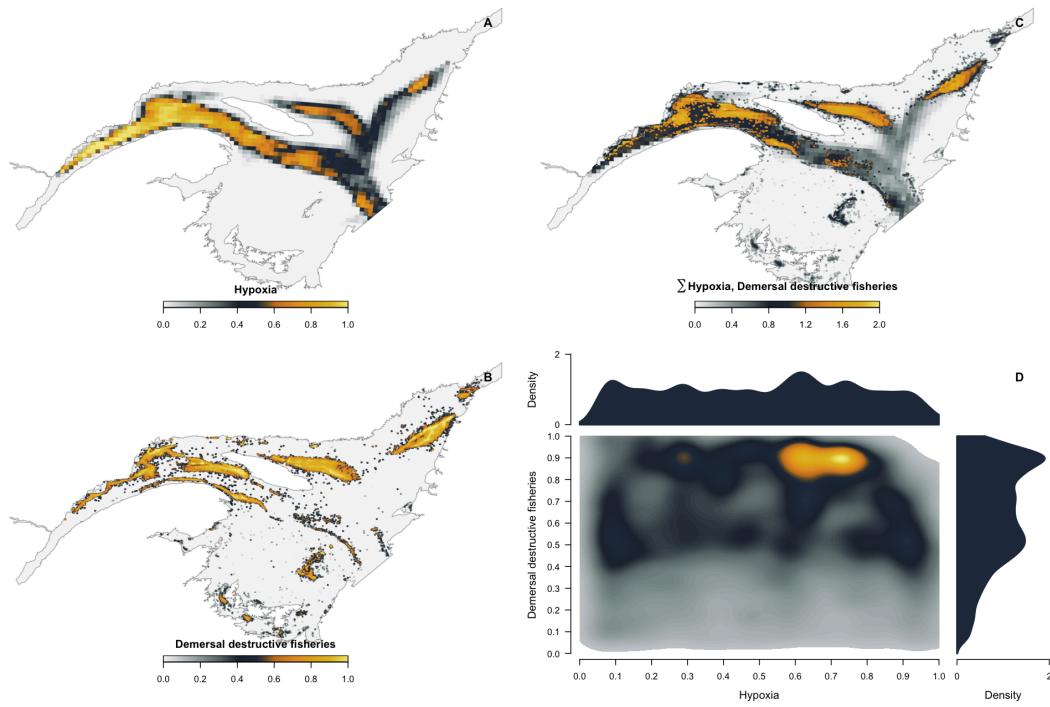


Figure 4: Interaction between the intensity of hypoxia and demersal destructive fisheries in the St. Lawrence. An index of hypoxia (A) was created using bottom-water dissolved oxygen between 2013 and 2017 (Blais et al., 2018). Demersal destructive fisheries (*i.e.* trawl and dredges) (B) intensity was evaluated from fisheries catch data collected between 2010 and 2015 used to measure annual area weighted total biomass ( $kg$ ) in  $i\ km^2$  grid cells (DFO, 2016b). See supplementary materials for more information on specific methodologies. Relative hypoxic stress and demersal destructive fisheries intensity was summed (C) to visualize their combined spatial distribution and intensity. Finally, individual density and the co-intensity of hypoxia and demersal destructive fisheries was investigated with a two-dimensional kernel analysis (D).

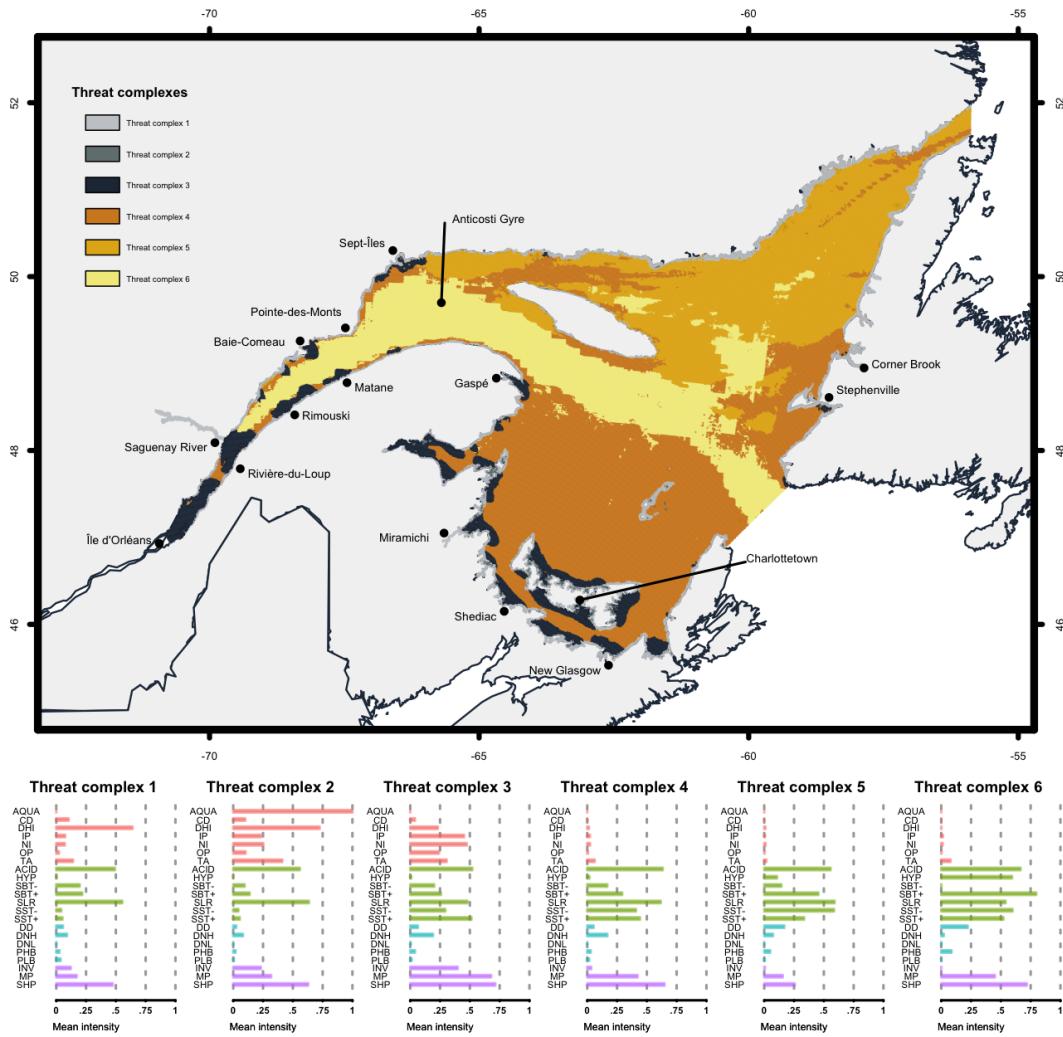


Figure 5: Distribution of threat complexes in the Estuary and Gulf of St. Lawrence (upper panel). Threat complexes (a term coined by Bowler et al., 2019) are areas with similar cumulative driver exposure regimes. Mean intensity of all coastal (red), climate (green), fisheries (blue) and marine traffic (purple) drivers within each threat complex (lower panel). Refer to Table 1 for acronym meaning and to the Supplementary Materials for more details.

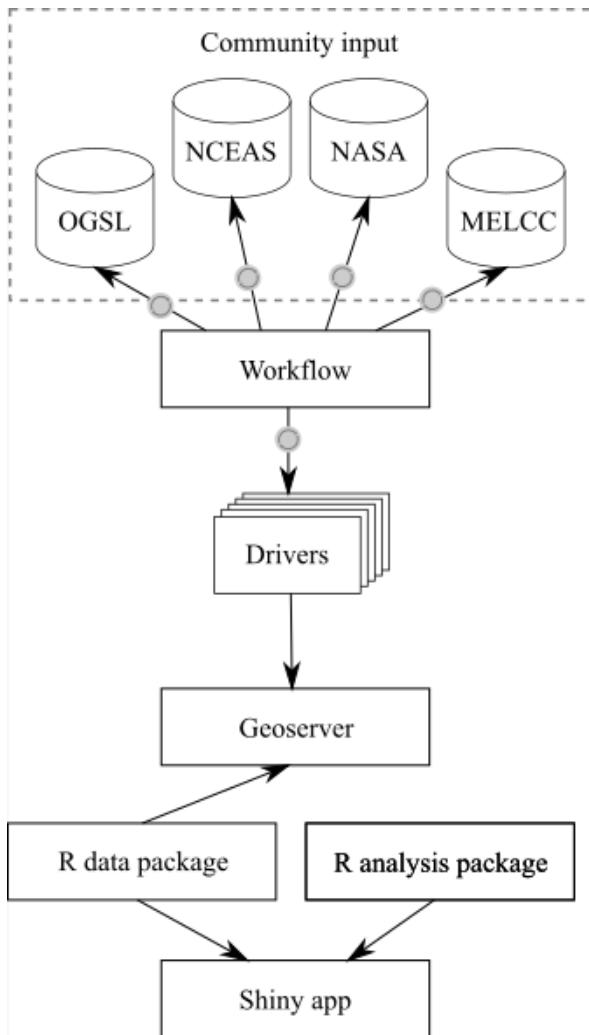


Figure 6: Diagram of the platform structure. Community input in the form of raw data is accessed through the St. Lawrence Global Observatory (SLGO; <https://ogsl.ca/en>) repository - the platform host - or through open access repositories (*e.g.* NASA data). The raw data are then processed through a workflow hosted on the *eDrivers* GitHub organization (<https://github.com/orgs/eDrivers/>). Data processing may be as simple as data rescaling (*e.g.* night lights) or make use of more complex methodologies (*e.g.* acidification). All data is then hosted on SLGO's geoserver and accessible through their API. We developed a R package called *eDrivers* to access the driver layers through R and we are actively developing a second R package called *eDriversEx* that will include analytical tools to explore drivers data. Finally, we have developed a Shiny application that allows users to explore drivers data interactively (<https://david-beauchesne.shinyapps.io/eDriversApp/>). All R components of the project are hosted and available on the *eDrivers* GitHub organization.

## Tables

Table 1: List of drivers currently available on [eDrivers](#) and used for the analyses presented in this paper.

Groups	Drivers	Acronym	Spatial	Temporal	Years	Units	Source
			resolution	resolution			
Climate	Aragonite	ACID	Lat/long	August-	2018	$\Omega$	(?)
				September			
Climate	Hypoxia	HYP	Lat/long	August-	2018	$ml\ L^{-1}$	(?)
				September			
Climate	Sea bottom temperature	SBT-	$\sim 2\ km^2$	Monthly	1981-2010	negative	(?)
					vs. 2013-	anomalies	
					2017		
Climate	Sea bottom temperature	SBT+	$\sim 2\ km^2$	Monthly	1981-2010	positie	(?)
					vs. 2013-	anomalies	
					2017		
Climate	Sea level rise	SLR	Modeled 0.25 degree	10 days	1992-2012	$mm$	(?)

Groups	Drivers	Acronym	Spatial	Temporal	Years	Units	Source
			resolution	resolution			
Climate	Sea surface temperature	SST-	$\sim 2 \text{ km}^2$	Monthly	1981-2010 vs. 2013-2017	negative anomalies	(?)
Climate	Sea surface temperature	SST+	$\sim 2 \text{ km}^2$	Monthly	1981-2010 vs. 2013-2017	positive anomalies	(?)
Coastal	Aquaculture	AQUA	Lat/long	-	Variable, between	<i>presence – absence</i>	(?????)
					1990-2016		
Coastal	Coastal development	CD	15 arc-second	Annual	2015-2016	<i>nanoWatts cm<sup>-2</sup> sr<sup>-1</sup></i>	(?)
Coastal	Direct human impact	DHI	Dissemination areas	Annual	2016	population	(?)
Coastal	Inorganic pollution	IP	Modeled $\text{km}^2$	Annual	2000-2001	-	(?)
Coastal	Nutrient import	NI	Modeled $\text{km}^2$	Annual	2007-2010	<i>t fertilizer</i>	(?)

Groups	Drivers	Acronym	Spatial	Temporal	Years	Units	Source
			resolution	resolution			
Coastal	Organic pollution	OP	Modeled $km^2$	Annual	2007-2010	$t$ pesticide	(?)
Coastal	Toxic algae	TA	-	-	-	Expert based	(?)
Fisheries	Demersal, destructive	DD	Lat/long	Event based	2010-2015	$kg$	(?)
Fisheries	Demersal, non-destructive, high-bycatch	DNH	Lat/long	Event based	2010-2015	$kg$	(?)
Fisheries	Demersal, non-destructive, low-bycatch	DNL	Lat/long	Event based	2010-2015	$kg$	(?)
Fisheries	Pelagic, high-bycatch	PHB	Lat/long	Event based	2010-2015	$kg$	(?)
Fisheries	Pelagic, low-bycatch	PLB	Lat/long	Event based	2010-2015	$kg$	(?)

Groups	Drivers	Acronym	Spatial	Temporal			Source
			resolution	resolution	Years	Units	
Marine traffic	Invasive species	INV	Modeled $km^2$	Annual	2011	$t$ port volume	(?)
Marine traffic	Marine pollution	MP	Modeled $km^2$	Event based & annual	2003-2011 & 2011	$n$ lanes + $t$ port volume	(?)
Marine traffic	Shipping	SHP	0.1 degree based	Event	2003-2011	$n$ lanes	(?)

## **ARTICLE 3**

### **PRÉDIRE LES INTERACTIONS BIOTIQUES AU SEIN DE MILIEUX PAUVRES EN DONNÉES**

**Résumé en français du deuxième article**

**Contexte scientifique**

**Publication associée**

**Traduction du résumé de l'article publié**

**Title**

Thinking outside the box – predicting biotic interactions in data-poor environments

**Authors**

David Beauchesne, Philippe Desjardins-Proulx, Philippe Archambault, Dominique Gravel

**Abstract**

Large networks of ecological interactions, such as food webs, are complex to characterize, be it empirically or theoretically. The former requires exhaustive observations, while the latter generally requires ample data to be validated. We therefore wondered whether readily available data, namely empirically described interactions in a variety of

ecosystems, could be combined to predict species interactions in data deficient ecosystems. To test this, we built a biotic interactions catalogue from a collection of 94 empirical food webs, detailed predator-prey interaction databases and interactions from the Global Biotic Interactions (GloBI) database. We used an unsupervised machine learning method to predict interactions between any given set of taxa, given pairwise taxonomic proximity and known consumer and resource sets found in the interaction catalogue. Results suggest that pairwise interactions can be predicted with high accuracy. Although conclusions are seemingly dependent on the comprehensiveness of the catalogue knowledge of taxonomy was found to complement well the catalogue and improve predictions, especially when empirical information available is scarce. Given its high accuracy, this methodology could promote the use of food webs and network level descriptors in certain fields of ecological science in which data is typically hard to gather and in remote and frontier location where empirical data is hard to gather. Network characteristics could then be efficiently evaluated and correlated to levels of environmental stressors in order to improve vulnerability assessments of ecosystems to global changes, opening promising avenues for further research and for management initiatives.

**Interactions, machine learning, food webs, K-nearest neighbour, taxonomy, St. Lawrence**

## **Introduction**

Large networks of ecological interactions, such as food webs, are complex to characterize (???). Empirical descriptions require exhaustive observations, while theoretical inference generally requires ample data to be validated. For this reason, studies focusing on communities of interacting species remain understudied, even though we acknowledge the importance of considering the reticulated nature of complex networks (??). When time is of the essence, the long term studies required quickly become impractical and

the use of network level approaches relegated to the sideline.

Alternatively, an approach currently gaining in popularity is to predict interactions using proxies such as functional traits, phylogenies and spatial distributions (e.g. ??). For example, multiple traits can play a significant role in community dynamics and influence the presence and intensity of biotic interactions, like the influence of body size on predator-prey interactions, a literal take on big fish eats small fish (????). However, the time required to gather the necessary data to apply those methods may still be restrictive, or the data be unavailable altogether, so much so that other methods such as imputation techniques have been developed to fill the gaps in knowledge (e.g. ??).

We therefore wondered whether more readily available data could be used to infer interactions in data deficient ecosystems. There is an increasing amount of data describing worldwide species interactions, some freely available through the Global Biotic Interactions (GloBI) database (?). Similarly, while phylogenies can be challenging to construct and require ample data, a taxonomical description of species is easily accessible through initiatives like the World Register of Marine Species (WoRMS; ?). More than simple nomenclature, evolutionary processes are thought to influence and shape consumer-resource relationships (??) so that taxonomically related species would be more likely to share similar types of both consumers and resources (???). Based on that assumption, taxonomy might be a useful surrogate in predicting interactions for species lacking detailed information on their biology, but which have a taxonomically related species for which such information is available.

The objective of this work is thus to combine empirical biotic interactions originating from a variety of ecosystems with taxonomic relatedness to predict interactions in data deficient ecosystems. The concept underlying our methodology is that instead of constraining ourselves to a specific environment, we would look to other environments – outside the box – to glean insights as to the inner workings of an area of interest. As an example, we compare the observed interactions in the southern Gulf of St. Lawrence

in Canada (SGSL; ?) with predictions made using our approach.

## Methods

The objective of our methodology is to predict the interactions between all pairs of taxa within an arbitrary set  $N_1$ , using a set of taxa  $N_0$  with empirically described interactions from which we can extract pairs of consumers and resources and their taxonomy. We couple the use of empirical data with an unsupervised machine learning method to achieve this.

### Biotic interaction catalogue

We built a biotic interaction catalogue to serve as a set of taxa  $N_0$  for with empirically described interactions. The empirical data used to construct the interaction catalogue was gathered in two successive steps. The first consisted of gathering data from a collection of 94 empirical food webs from which we extracted pairwise taxa interactions (see ??? for more information). We also used a detailed predator-prey interaction database describing trophic relationships between marine fishes and their prey (?). From these datasets, only interactions between taxa at the taxonomic scale of the family or higher were selected for inclusion in the catalogue. Data used came exclusively from marine and coastal ecosystems and encompassed a wide variety of organisms: fungi, algae, parasites, phytoplankton, zooplankton, benthic and pelagic invertebrates, demersal and pelagic fishes, marine birds and marine mammals.

As empirical food webs are vastly dominated (96%) by unobserved or absent interactions ("0", hereafter referred to non-interactions), these datasets yielded a highly skewed distribution of interactions vs non-interactions. To counterbalance this, the second step of data compilation consisted of extracting observed interactions from the Global Biotic Interaction (GloBI) database (?), which describes binary interactions for a wide range of taxa worldwide. We extracted all trophic interactions available on GloBI for

species belonging to the families of taxa identified through step 1. Interactions were extracted using the rGloBI package in R (?). As per step 1, only interactions between taxa at the taxonomic scale of the family or higher were retained.

The nomenclature used between datasets and food webs varied substantially. Taxa names thus had to be verified, modified according to the scientific nomenclature and validated. This process was performed using the Taxize package in R (??) and manually verified for errors. The same package was used to extract the taxonomy of all taxa for which interactions were obtained in previous steps. The complete R code and data used to build the catalogue is available at [https://github.com/david-beauchesne/Interaction\\_catalog](https://github.com/david-beauchesne/Interaction_catalog).

## Unsupervised machine learning

We use the  $K$ -nearest neighbor (KNN) algorithm (?) to predict pairwise interactions for a set of taxa  $S$ . The KNN algorithm predicts missing entries or proposes additional entries by a majority vote based on the  $K$  nearest (i.e. most similar) entries (see Box 1 for an example). In this case, taxa are described by a set of resources when considered as a consumer, a set of consumers when considered as a resource and their taxonomy (i.e. kingdom, phylum, class, order, family, genus, species). Similarity between taxa was evaluated using the Tanimoto similarity measure, which compares two vectors  $x$  and  $y$  with  $n = |\mathbf{x}| = |\mathbf{y}|$  elements, and is defined as the size of the intersection of two sets divided by their union:

$$\text{tanimoto}(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|}, \quad (3.1)$$

where  $\cap$  is the intersect and  $\cup$  the union of the vectors. Adding a weighting scheme, we can measure the similarity using two different sets of vectors  $\{\mathbf{x}, \mathbf{y}\}$  and  $\{\mathbf{u}, \mathbf{v}\}$ :

$$\text{tanimoto}_t(x, y, u, v, w_t) = w_t \text{tanimoto}(\mathbf{x}, \mathbf{y}) + (1 - w_t) \text{tanimoto}(\mathbf{u}, \mathbf{v}), \quad (3.2)$$

where  $w_t$  the weight (in  $[0; 1]$ ). For our analyses, the first element on the right-hand side of (3.2) is the Tanimoto similarity measured using the taxonomy of two taxa. The second is the Tanimoto similarity between the sets of resources (or consumers) of the same taxa. When  $w_t = 0$  only resource or consumer sets are used to compute similarity, while  $w_t = 1$  solely uses taxonomy. This approach to consider the relative contribution of two sets of vectors to the Tanimoto similarity was developed by ?.

### Predicting interactions

The algorithm was built on a series of logical steps that ultimately predicts a candidate resources list  $C_R$  for each taxon in  $N_1$  based on empirical data available and the similarity among consumers and among resources (Figure 7). For all consumer taxa  $T_C$  in  $N_1$ , the algorithm first verifies, for all resources in resource set  $T_R$ , if they are found the  $N_0$  (Step S1, Figure 7). When it does, all  $T_R$  taxa that are also in  $N_1$  are added as predicted resources for  $T_C$  (Steps S2 and S3). This corresponds to what we refer to as the catalogue contribution to resource predictions. In essence, two taxa in  $N_1$  that are known to interact through empirical data in the catalogue are automatically assumed to interact in  $N_1$ .

Otherwise, the algorithm passes to what we refer to as the predictive contribution to resource predictions (Steps S4 to S16), with candidate resources for  $T_{Ci}$  (focal taxa for explanation) identified with the KNN algorithm. For each resource in  $T_R$  that were not in  $N_1$  (Step S2), K most similar resources  $T_{R'}$  are identified from  $N_1$  (Step S4). If similar resources  $T_{R'}$  have a similarity value above a minimal similarity threshold set to 0.3 in our analysis, they are added to  $C_R$  as candidate resources. If not, they are automatically discarded (Steps S5 to S7). This minimal threshold is an arbitrary parameter used to avoid predicting resources that have very small and insignificant

similarity and hence is very unlikely to share consumers and resources with the taxa it is being compared to.

Then for all consumer taxa  $T_C$  in  $N_1$ , K most similar consumers  $T_{C'}$  are identified from  $N_0$ . This step aims at extracting sets of potential resources  $T_R$  from similar types of consumers found in the catalogue (Step S8). Resources  $T_R$  are added to candidate resources  $C_R$  for  $T_{Ci}$  if they are also found in  $N_1$  (Steps S10 to S12). Otherwise, Steps S4 to S7 are duplicated to identify potential similar resources for  $T_{Ci}$  in  $N_1$  from the set of resources  $T_R$  of similar consumers  $T_{C'}$  (Steps S13 to S16). A simple working example is presented at Box 1. A comprehensive mathematical description of the algorithm and the parameters used is however available through Figure 7 and the complete R code and data used for the algorithm is available at [https://github.com/david-beauchesne/Predict\\_interactions](https://github.com/david-beauchesne/Predict_interactions).

### Algorithm prediction accuracy

We used datasets including more than 50 taxa (??????) to assess the prediction accuracy of the algorithm. Testing accuracy of a particular dataset was done by first removing from the catalogue all pairwise interacting taxa originating from that dataset. Accuracy was evaluated using three different statistics:

1.  $Score_y$  is the fraction of interactions correctly predicted:

$$Score_y = \frac{a}{a + c} \quad (3.3)$$

2.  $Score_{\neg y}$  is the fraction of non-interactions correctly predicted:

$$Score_{\neg y} = \frac{d}{b + d} \quad (3.4)$$

3. TSS, The True Skilled Statistics (TSS) evaluated prediction success by considering both true and false predictions, returning a value ranging from 1 (perfect

predictions) to -1 (inverted predictions; [?](#)):

$$TSS = \frac{(ad - bc)}{(a + c)(b + d)} \quad (3.5)$$

where  $a$  is the number of interactions correctly predicted (*i.e.* true positives),  $b$  is the number of non-interactions predicted as interactions (*i.e.* false positives),  $c$  is the number of observed interactions predicted as non-interactions (*i.e.* false negatives) and  $d$  is the number of non-interactions correctly predicted (*i.e.* true negatives). These three statistics give a different perspective on prediction accuracy, focusing in turn on true interactions and non-interactions, and on both true and false predictions. It is however important to note that false positives and true negatives are solely representative of the datasets used rather than the environment itself. However extensive the datasets may be, unobserved interactions may not necessarily mean a true absence of interaction.

For each statistic, we evaluated prediction accuracy 1) for the complete algorithm, 2) for predictions made through the predictive portion of the algorithm (Steps S4-S16; Figure 7) and 3) for the catalogue contribution of the algorithm (Steps S1-S3; Figure 7). We evaluated these steps separately in order to partition the relative contribution of the catalogue and of the predictions made using the KNN algorithm to the overall predictive accuracy of the algorithm. Multiple  $w_t$  values were also tested to evaluate whether taxa similarity measured as a function of resource/consumer sets or taxonomy contributed more significantly towards increased predictive accuracy. The same was done with multiple  $K$  values.

Finally, we evaluated the influence of the comprehensiveness of the catalogue on prediction accuracy. We selected the arctic marine food web from [?](#) as a test. This food web was selected as it is highly detailed taxonomically. Furthermore, once removed from the catalogue, almost 100% of its taxa still had information available on sets of consumers and resources, which necessary for testing the impact of catalogue comprehensiveness on prediction accuracy. We iteratively and randomly ( $n = 50$  randomizations) removed

a percentage of empirical data describing the food web taxa from the catalogue before generating new predictions with the algorithm. We also tested  $w_t$  values of 0.5 and 1 to evaluate whether taxonomic similarity could support predictive accuracy in cases when empirical data for species in  $N_1$  in the catalogue is unavailable.

## Results

### Biotic interaction catalogue

The data compilation process allowed us to build an interaction catalogue composed of 276708 pairwise interactions (interactions = 72110; non-interactions = 204598). A total of 9712 taxa (Superfamily = 15; Family = 591; Subfamily = 29; Tribe = 8; Genus = 1972; Species = 7097) are included in the catalogue, 4159 of which have data as consumers and 4375 as resources.

### Algorithm predictive accuracy

The overall predictive accuracy of the algorithm ranges between 80% to almost 100% in certain cases (Figure 8). Both interactions and non-interactions are well predicted by the algorithm. TSS scores are lower than  $Score_y$  and  $Score_{\neg y}$  due to misclassified interactions and non-interactions. This can also be observed through the effect of varying  $K$  values, which increases the number of potential candidate resources for each taxa in the predictive portion of the algorithm. Prediction accuracy increases for interactions, while it decreases for non-interactions, as  $K$  values increase.

Similarity being predominantly measured with resource/consumer sets ( $w_t$  closer to 0) yielded better predictions than when measured with taxonomy ( $w_t$  closer to 1; Figure 8). Resource/consumer sets therefore appears to serve as a better measure of similarity between taxa for interactions predictions. It is nonetheless interesting to note that although the predictive contribution of the algorithm decreases as  $w_t$  increases, an

increased mean and decreased variability values for the TSS and  $Score_y$  statistics is also observed (Figure 8). This suggests that while resource/consumer similarity yields higher predictive accuracy, taxonomy better complement the catalogue contribution by predicting interactions not captured through empirical data, effectively increasing the predictive accuracy of the complete algorithm.

The partitioning of the catalogue and predictive portions of the algorithm reveals the importance of the comprehensiveness of the catalogue in prediction accuracy (Figures 8, 9). As the amount of empirical data available in the catalogue increases so does the overall accuracy of the algorithm (Figures 9). While prediction accuracy of the predictive portion of the algorithm is somewhat lower, it nonetheless supports high prediction efficiency when the catalogue comprehensiveness is lower (Figures 9). Prediction accuracy still remains around 75% with only 40% of  $N_1$  taxa found in the catalogue (Figures 9). Furthermore, the use of taxonomy for similarity measurements is more efficient when empirical data is scarcer and no different than resource/consumer sets for the complete algorithm when ample data is available (Figures 9).

### Southern Gulf of St. Lawrence

As an example, we predict interactions in the southern Gulf of St. Lawrence (SGSL) in eastern Canada. The empirical data and taxa list come from ?. They present a list of 29 functional groups for a total of 80 taxa presented at least at taxonomical scale of the family. Other coarser functional groups were not used for this example (see Table S1 in Supplementary information (SI) and ? for a complete description of documented groups). We used the algorithm to predict interactions between all 80 taxa selected. As their interaction data are reported for functional groups rather than taxa, we then aggregated them back to their original functional groups to compare with interactions presented in ?. In total, there were empirical data available in the catalogue for 78% of SGSL taxa (62/80). The algorithm correctly predicted close to 80% of interactions (*a*

$= 135/170$ ) and non-interactions ( $d = 354/455$ ) extracted from ?. It also predicted an additional 101 interactions that were not noted in ? and failed to predict 36 observed interactions that were, resulting in a TSS score of 0.57. A visual comparison of results obtained from the algorithm with interactions noted in ? is available at Figure 10. The network presented is centered on the observed and predicted interactions of the capelin (*Mallotus villosus*) and piscivorous small pelagic feeders (e.g. *Scomber scombrus* and *Illex illecebrosus*).

## Discussion

### Algorithm accuracy

We show that out of the box interaction inference for a set of taxa with incomplete or unavailable preexisting information can be achieved with high accuracy using a combination of empirical data describing biotic interactions and taxonomic relatedness. Although the efficiency of the algorithm is dependent on the comprehensiveness of the interactions catalogue, taxonomic proximity acts as a complement to increase the number of observed interactions correctly predicted. Taxonomic proximity also supports the efficiency of the algorithm when information gleaned through the catalogue is scarce.

### Usefulness of taxonomic relatedness

We found that taxonomy can be highly useful in complementing predictions made using empirical data. Much like the findings from ?, evolutionary history provides a significant background from which inferences on network structure can be made. Nonetheless, while evolutionary history plays a significant role in influencing consumer-resource trait matching and food web structure (??), phylogenetic constraints do not necessarily account efficiently for certain traits such as body size (?). Complementing our methodology with additional, higher-order information such as functional traits (e.g.

metabolism and body size) could thus yield even more efficient results, especially in cases where the catalogue lacks data on taxa for which interactions have to be predicted. Similarly, using phylogenies rather than taxonomy could enhance the resolution at which evolutionary history is considered. This could be achieved through recent efforts to extensively describe all-encompassing phylogenies (e.g. ?). Complementing our approach by making it more data dependent could undermine the premise under which this method was built and which constitutes its main strength, *i.e.* predicting interactions in data deficient environments using readily available data. The flexibility of our methodology would however easily allow for the inclusion of alternate sources of data. Therefore, high-order data such as phylogenies could and should be used in instances where ample data is available, making the use of this methodology broader than simply in instances when data is unavailable.

### **Interactions classification**

That  $Score_y$  and  $Score_{\neg y}$  are inversely proportional means that non-interactions are misclassified as interactions in the process of increasing  $Score_y$ , consequently decreasing  $Score_{\neg y}$ . This could either stem from the algorithm poorly predicting non-interactions or from the empirical data itself. Accuracy evaluation assumes that non-interactions from empirical food web are observed data, yet it is usually not the case. Most empirical webs have a strong focus attributed to higher order consumer species and often uneven effort made to thoroughly detail species interactions (?). Furthermore, the methodologies used to obtain consumer-resource data, often relying on gut content analyses, which is efficient at observing interactions, may be inefficient to detect absence of interactions in natural systems (?). This is especially true with our methodology, where we predict interactions between species whose co-occurrence may have been observed in the other ecosystems we are using to predict interactions. Misclassified interactions could thus be real, albeit unobserved through empirical data available.

## Southern Gulf of St. Lawrence

The St Lawrence example (Figure 10 and SI) provides adequate material to discuss predictions in greater detail. The algorithm fails to predict 20% of interactions presented in ?. Interactions that failed to be predicted were mainly centered on invertebrate species (e.g. polychaetes and mollusks) and taxonomically diverse functional groups described by coarse taxonomic categories (e.g. diatoms) alongside few species in ? (e.g. piscivorous small pelagic feeders; Table S3). As we focused on the taxa at least at the scale of family, it is likely that their functional groups had a broader range of possible interactions included than what the algorithm could predict using only a few taxa. Furthermore, the efficiency of the algorithm greatly depends on the underlying empirical data that defines the catalogue. If the empirical data used to build the catalogue focuses on higher order consumers, it should come as no surprise that the algorithm would be afflicted by the same limitations.

On the other hand, the algorithm also predicts substantially more interactions than those presented in ? (Figure 10; Table S2). For instance, an important number of additional interactions were predicted for small piscivorous pelagic feeders as consumers (Figure 10). When considering that these species are typically considered as resources, it should be unsurprising that the broad range of interactions composing the catalogue and from which predictions are made results in new consumer interactions being predicted for those species. An ecological interpretation can therefore be easily provided to explain these additional interactions, such as small piscivorous pelagic feeders consuming cod, likely representing a consumption of cod eggs and/or juveniles. This greatly exemplifies the point we made in the previous section with regards to misclassified interactions being real rather than false positives. The resulting TSS score is therefore greatly diminished by classifying additional interactions as false positives. We therefore believe that the TSS score for the St. Lawrence analysis represents an underestimation of the efficiency of our methodology to predict interactions.

## Perspectives

We show that out of the box interaction inference can be achieved with high accuracy using readily available data, suggesting that ecological networks are characterized by a degree of predictability and that this predictive value can be recovered through learning (see ?? for other examples). This adds weight to claims that regularities can be observed and predicted in network structure (?).

We believe that our methodology offers promising avenues for further applied research and management initiatives. The flexibility of our methodology allows it to take advantage of multiple types of data. Complementing and testing our methodology with additional ecological information such as functional traits and phylogenies would therefore be highly valuable. Interaction strength and species co-occurrence are additional major attributes affecting the probability of observing interactions and the resulting network structure. Interaction strength is instrumental to understanding community dynamics, stability and robustness (??), while the co-occurrence of species encloses valuable information on interactions and is obviously a pre-requisite for interactions to exist (?). Considering them in our methodology would be highly valuable to correctly assess interactions in a given ecosystem and predict the spatial distribution of interaction networks.

The significance of this approach also extends to other areas of ecological research where gathering data can be highly difficult, such as the reconstruction of interaction networks forming palaeocommunities (e.g. ??). Predicted networks of taxa known to co-occur could be used in hindsight to evaluate the influence of major events such as biodiversity collapse or significant climatic regime shifts on the structure of past ecological communities.

Ultimately, given its high efficiency and simplicity, our methodology could help in promoting the use and the accessibility of food webs and network level descriptors for integrative management initiatives such as cumulative impacts assessments and systematic

planning (??), especially for remote locations and frontier areas where empirical data is hard to gather. Network characteristics could be efficiently evaluated and correlated to levels of multiple environmental stressors to assess the vulnerability of ecosystems to global changes (?). We believe that the development of such predictive approaches could represent the first much needed steps towards the use of ecological networks in systematic impacts assessments.

### **Acknowledgements**

We thank the Fond de Recherche Québécois Nature et Technologie (FRQNT) and the Natural Science and Engineering Council of Canada (CRSNG) for financial support. This project is also supported by Québec Océan, the Quebec Centre for Biodiversity Science (QCBS), and the Notre Golfe and CHONeII networks. We also wish to thank K. Cazelles for the help, constructive comments and suggestions. We also thank David Bohan and an anonymous reviewer for their constructive comments and suggestions.

**Box 1**

The algorithm follows a series of logical steps to predict resources for all taxa in an arbitrary set of taxa  $N_1$  using a set of taxa  $N_0$  with empirically described interactions from which we can extract sets of consumers and resources and their taxonomy. In this example, we are predicting interactions for a fictitious  $N_1 = \{T_1, T_9, T_{10}, T_{11}, T_{12}\}$  using  $N_0$  with information on 12 taxa. This catalogue holds information on consumer or resource for 10 taxa and the taxonomy for all 12 taxa in the list.

$N_0$ taxa ID	taxonomy	resource	consumer
$T_1$	$\{a, b, c\}$	$\{T_2, T_3, T_{12}\}$	$\{T_4\}$
$T_2$	$\{e, f, g\}$		$\{T_1, T_5\}$
$T_3$	$\{i, j, k\}$		$\{T_5\}$
$T_4$	$\{m, n, o\}$	$\{T_1, T_5\}$	
$T_5$	$\{a, b, d\}$	$\{T_8, T_9\}$	$\{T_4\}$
$T_6$	$\{i, q, r\}$	$\{T_2, T_8\}$	$\{T_4\}$
$T_7$	$\{e, f, h\}$		$\{T_1, T_6\}$
$T_8$	$\{s, t, u\}$		$\{T_5, T_6\}$
$T_9$	$\{s, t, v\}$		$\{T_5\}$
$T_{10}$	$\{i, j, l\}$		
$T_{11}$	$\{m, n, p\}$		
$T_{12}$	$\{q, r, s\}$		$\{T_1\}$

Similarity between all pairs of taxa in  $N_0$  is measured for consumer, resource and taxonomic proximity using equation 1. The upper triangular matrix represents similarity measured with taxa sets of resources/consumers, while the lower triangular represents taxonomic similarities. For consumer/resource set similarities, values of 0 mean that similarity equals 0 for both similarity measurements.

$$\text{tanimoto}(T_C x, T_C y) / \text{tanimoto}(T_R x, T_R y)$$

	$T_1$	$T_2$	$T_3$	$T_4$	$T_5$	$T_6$	$T_7$	$T_8$	$T_9$	$T_{10}$	$T_{11}$	$T_{12}$
$T_1$	-	0	0	0	0/1	0.3/1	0	0	0	0	0	0
$T_2$	0	-	0/0.5	0	0	0	0/0.3	0/0.3	0/0.5	0	0	0/0.5
$T_3$	0	0	-	0	0	0	0	0/0.5	0/1	0	0	0
$T_4$	0	0	0	-	0	0	0	0	0	0	0	0
$T_5$	0.5	0	0	0	-	0.3/1	0	0	0	0	0	0
$T_6$	0	0	0.2	0	0	-	0	0	0	0	0	0
$T_7$	0	0.5	0	0	0	0	-	0/0.3	0	0	0	0/0.5
$T_8$	0	0	0	0	0	0	0	-	0	0	0	0
$T_9$	0	0	0	0	0	0	0	0.5	-	0	0	0
$T_{10}$	0	0	0.5	0	0	0.2	0	0	0	-	0	0
$T_{11}$	0	0	0	0.5	0	0	0	0	0	0	-	0
$T_{12}$	0	0	0	0	0	0.5	0	0.2	0.2	0	0	-

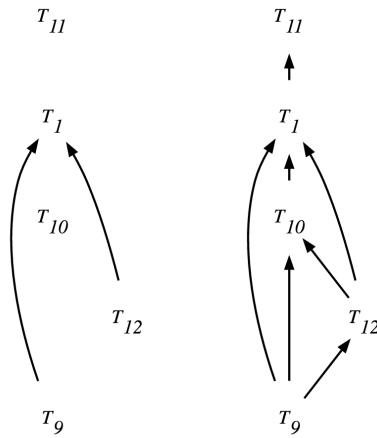
$$\text{tanimoto}(T_Tx, T_Ty)$$

From these, the algorithm goes through logical steps (Figure 7) to identify a candidate resource list  $C_R$  for each taxon in  $N_1$  using either empirical data directly or  $K$  most similar taxa with equation 2. Going through the process for  $T_1$ , using  $K = 1$  and  $w_t = 1$ :

The logical steps allow us to predict a set of resources for  $T_1 = \{T_9, T_{10}, T_{12}\}$ . Doing it for all taxa in  $N_1$  with  $w_t = 0$  and 1 predicts the following networks:

$$w_t = 0 \quad w_t = 1$$

Steps		Catalogue	Prediction
1	$I(T_1, T_R)$ in $N_0$ ?		
2	$T_R$ in $N_1$ ?		
4-7	$T_2 = \text{no} \rightarrow t(T_2, T_{R'}, w_t) = \text{NA}$	{}	{}
4-7	$T_3 = \text{no} \rightarrow t(T_2, T_{R'}, w_t) = T_{10} = 0.5$	{}	{ $T_{10}$ }
3	$T_{12} = \text{yes}$	{ $T_{12}$ }	{ $T_{10}$ }
8	$t(T_1, T_{C'}, w_t) = T_5 = 0.5$		
9	$I(T_5, T_R)$ in $N_1$ ?		
13-16	$T_8 = \text{no} \rightarrow t(T_8, T_{R'}, w_t) = T_9 = 0.5$	{ $T_{12}$ }	{ $T_9, T_{10}$ }
10-12	$T_9 = \text{yes}$	{ $T_9, T_{12}$ }	{ $T_9, T_{10}$ }



## Figures

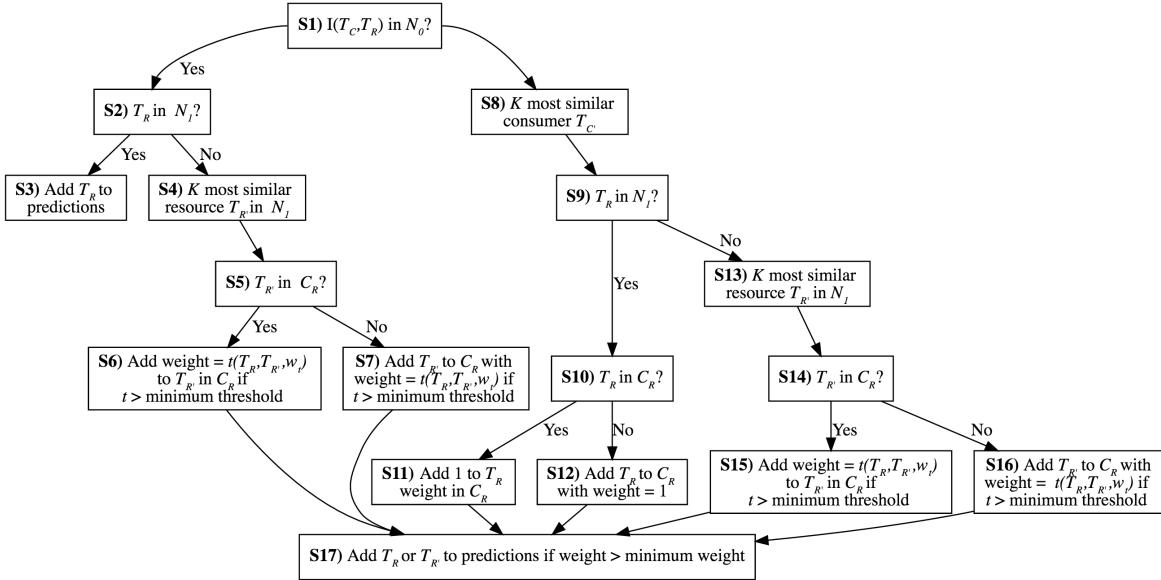


Figure 7: Description of 17 logical steps (S1-S17) used by the algorithm to suggest a list of candidate resources ( $C_R$ ) for each consumer taxa ( $T_C$ ) in a set of  $N_1$  for which interactions are predicted, using a set of taxa  $N_0$  with empirically described interactions. Interactions between consumer and resource taxa are denoted as  $I(T_C, T_R)$ .  $K$  is the number of most similar neighbours selected for the KNN algorithm;  $t$  stands for tanimoto in equation 1;  $w_t$  is the weight given to sets of resources and consumers in equation 2; the minimum threshold is a value setting the minimal similarity value accepted for taxa to be considered as close neighbours in the KNN algorithm; the weight is the value added to a candidate resource each time it is added to  $C_R$ ; the minimum weight is the minimal weight value accepted for candidate resources to be selected as predicted sources in the algorithm.

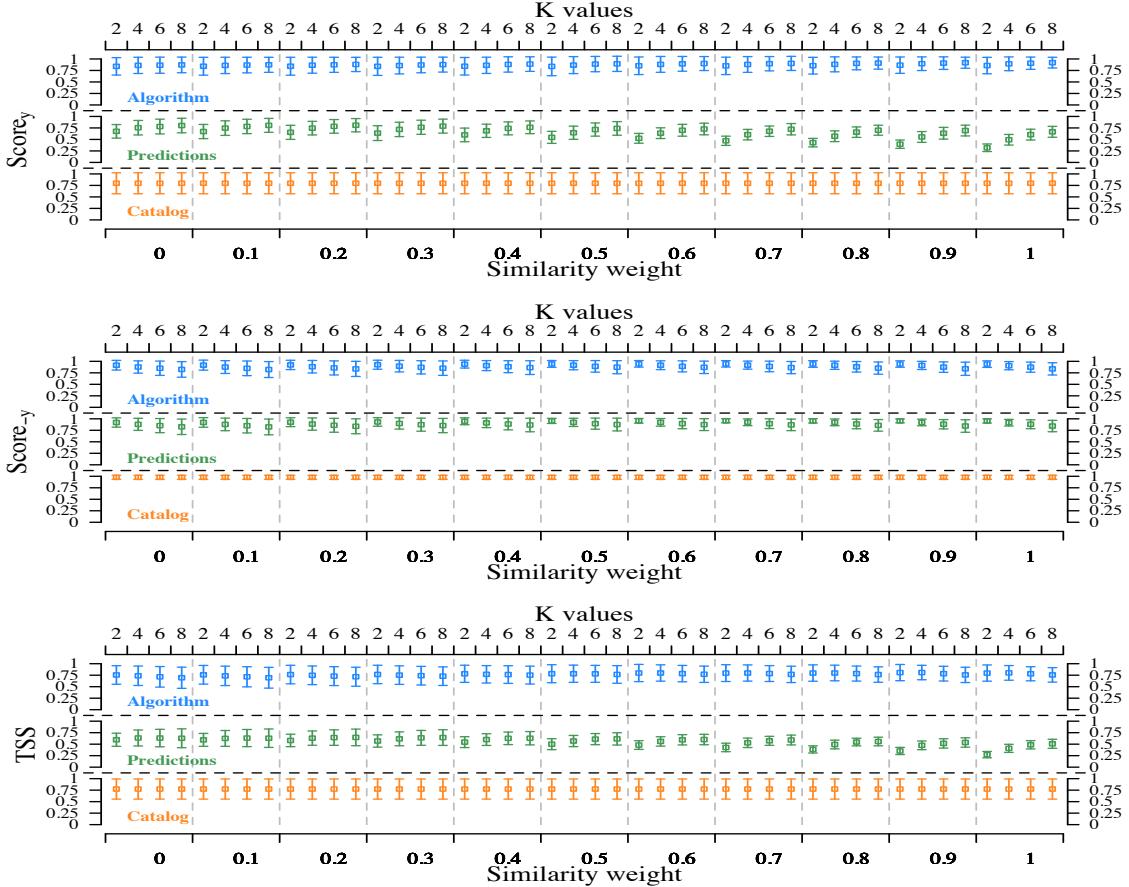


Figure 8: Representation of the three statistics (*i.e.*  $Score_y$ ,  $Score_{-y}$  and TSS) used to evaluate the accuracy of the algorithm as a function of  $K$  values tested (*i.e.* 2, 4, 6 and 8 most similar neighbours, top  $x$ -axis) and weight for taxonomy (bottom  $x$ -axis), which varies between 0 and 1. A weight of 0 means that similarity is measured only using set of resources/consumers for each taxa, while a weight of 1 means that similarity is based solely on taxonomy. For each statistic, the topmost panel presents prediction accuracy for the complete algorithm, the middle panel corresponds to predictions made through the predictive portion of the algorithm (Steps S4-S16; Figure 7) and the bottom panel presents the catalogue contribution for the algorithm (Steps S1-S3; Figure 7). Note that the sum of the predictive and catalogue contributions can be over 100% as there is overlap between predictions made through both. The 7 datasets used for this analysis contained over 50 taxa (???????)

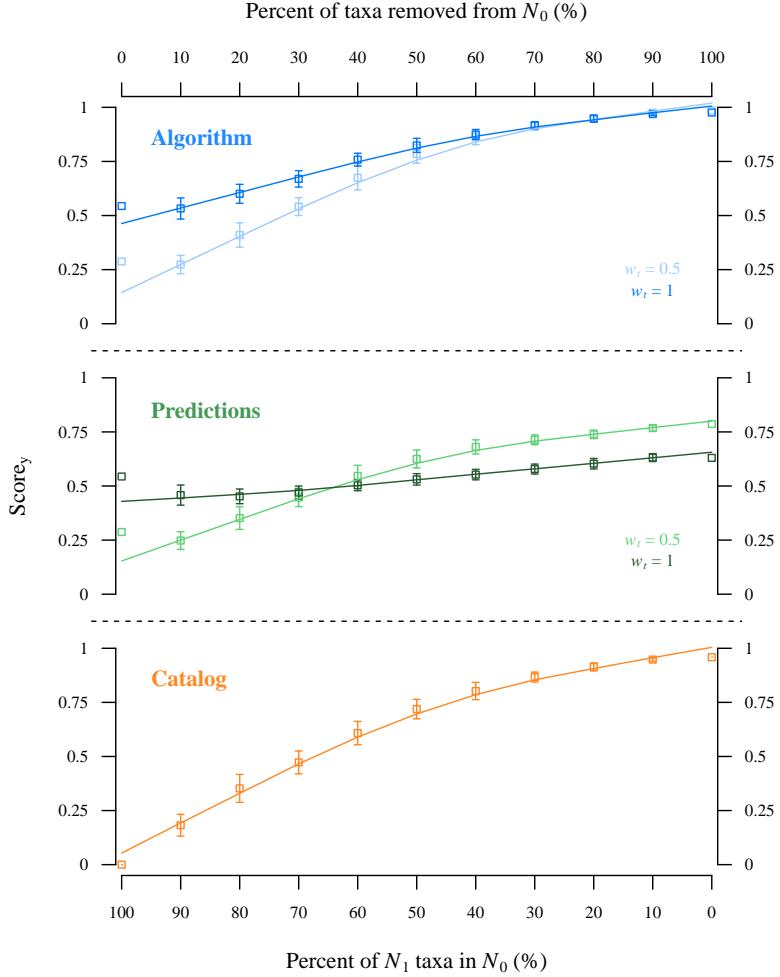


Figure 9: Representation of  $Score_y$  as a function of catalogue comprehensiveness, *i.e.* the amount of information on sets of consumer and resources available in the catalogue. The sensitivity of the algorithm to data accuracy was evaluated with the arctic food web from ?. This food web was highly detailed taxonomically. Once removed from the catalogue, almost 100% of its taxa still had information available on sets of consumers and resources, which is necessary for testing the impact of catalogue comprehensiveness on prediction accuracy. A random percentage of data available in the catalogue for taxa in the food web (*i.e.* 0 to 100%) was iteratively removed ( $n = 50$  randomizations) before generating new predictions with the algorithm.  $w_t$  values of 0.5 and 1 were evaluated to verify the usefulness of taxonomy in supporting predictive accuracy. The topmost panel presents prediction accuracy for the complete algorithm, the middle panel corresponds to predictions made through the predictive portion of the algorithm (Steps S4-S16; Figure 7) and the bottom panel presents the catalogue contribution for the algorithm (Steps S1-S3; Figure 7). Note that the sum of the predictive and catalogue contributions can be over 100% as there is overlap between predictions made through both.

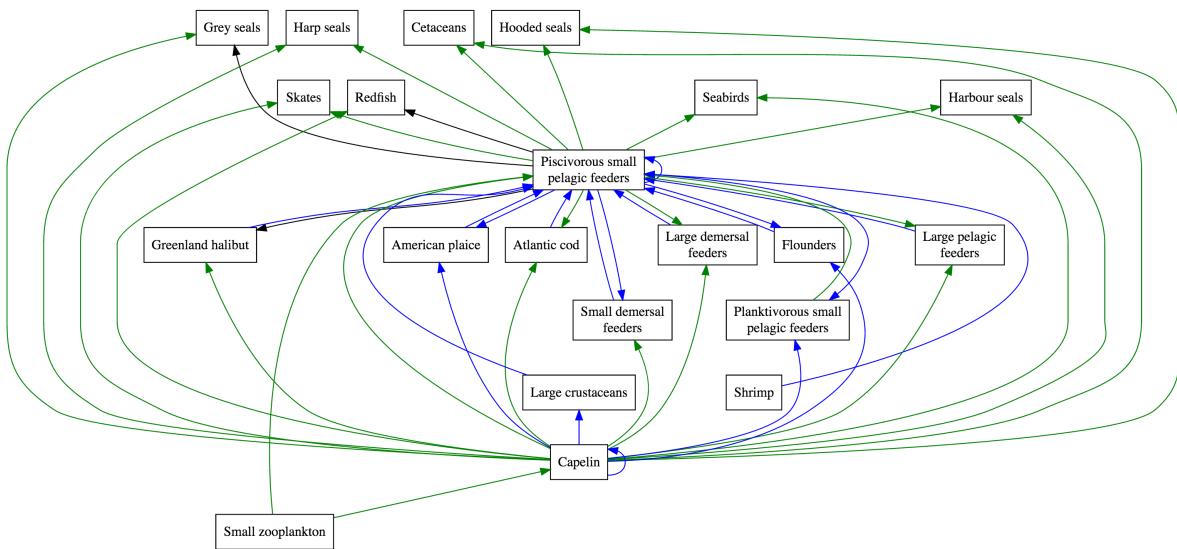


Figure 10: Example of predicted interactions with the network of the southern Gulf of St. Lawrence (?), centered around the interactions of the capelin (*Mallotus villosus*) and piscivorous small pelagic feeders (e.g. *Scomber scombrus* and *Illex illecebrosus*). Edge with colors green were both predicted and observed (26), black were observed only (3) and blue were predicted only (19). Arrows are pointed towards consumers.

## **ARTICLE 4**

### **TITRE CHAPITRE 4**

**Résumé en français du deuxième article**

**Contexte scientifique**

**Publication associée**

**Traduction du résumé de l'article publié**

**Introduction**

**Mat & Meth**

## **CONCLUSION GÉNÉRALE**

### **Brainstorm**

#### **Éléments discussion :**

- Scénarios de décisions de gestion
- Données ouvertes
- Caractérisation food webs
- Services écosystémiques
- Conservation (AMP)
- Planification spatiale marine

#### **Une belle conclusion générale**

Une petite conclusion générale du même format que l'introduction

#### **To conclude**

## **ANNEXE I**

## **ANNEXE I**

## ANNEXE II

### GESTION DE NOUVELLE GÉNÉRATION - STRUCTURER ET PARTAGER LES DONNÉES DE PRESSIONS ENVIRONNEMENTALES POUR LE SYSTÈME DU SAINT-LAURENT

#### Drivers description

Table 2: List of drivers currently available on eDrivers along with their respective acronym used in the figures in the supplementary material.

Groups	Drivers	Acronym	Source
Climate	Aragonite	ACID	(?)
Climate	Hypoxia	HYP	(?)
Climate	Sea bottom temperature	SBT-	(?)
Climate	Sea bottom temperature	SBT+	(?)
Climate	Sea level rise	SLR	(?)
Climate	Sea surface temperature	SST-	(?)
Climate	Sea surface temperature	SST+	(?)
Coastal	Aquaculture	AQUA	(?????)
Coastal	Coastal development	CD	(?)

Groups	Drivers	Acronym	Source
Coastal	Direct human impact	DHI	(?)
Coastal	Inorganic pollution	IP	(?)
Coastal	Nutrient import	NI	(?)
Coastal	Organic pollution	OP	(?)
Coastal	Toxic algae	TA	(?)
Fisheries	Demersal, destructive	DD	(?)
Fisheries	Demersal, non-destructive, high-bycatch	DNH	(?)
Fisheries	Demersal, non-destructive, low-bycatch	DNL	(?)
Fisheries	Pelagic, high-bycatch	PHB	(?)
Fisheries	Pelagic, low-bycatch	PLB	(?)
Marine	Invasive species traffic	INV	(?)
Marine	Marine pollution traffic	MP	(?)
Marine	Shipping traffic	SHP	(?)

## Climate

### Acidification

Oceans are the largest reservoir and sink of atmospheric carbon dioxide ( $CO_2$ ). Its uptake increases seawater acidity and the lowers saturation state with respect to calcite

( $\Omega_C$ ) and aragonite ( $\Omega_A$ ), the two most common  $CaCO_3$  polymorphs constituting the shells and skeleton of many marine organisms (?). This can have deleterious effects on carbonate-secreting organisms (e.g. mollusks and crustaceans) and certain physiological processes in non-calcifying organisms (??).

When  $\Omega_C$  or  $\Omega_A$  decrease below 1, water becomes undersaturated and corrosive to the skeletal minerals of carbonate-secreting organisms. Note that if  $\Omega_A < 1$ , the waters may still be supersaturated with respect to calcite since it is 50% more soluble than aragonite. However, organisms have wide ranging responses to  $\Omega_C$  and  $\Omega_A$  saturation state. For example, most corals stop calcifying at  $\Omega_A < 2$  (??), while other organisms may have adapted to precipitate  $CaCO_3$  even when seawater is undersaturated with respect to calcite ( $\Omega_C < 1$  or  $\Omega_A < 0.65$ ; e.g. ?).

We used  $\Omega_A$  saturation state to characterize ocean acidification in the bottom waters of the St. Lawrence. The data come from the Department of Fisheries and Oceans' (DFO) Atlantic Zone Monitoring Program (AZMP; ?) surveys and DFO's multispecies surveys (?) collected in August-September of 2017 (?). The carbonate chemistry was determined through pH and TA measurements. Samples for pH and total alkalinity (TA) were collected under a bubble free and no head space conditions into 500 mL borosilicate glass flasks, 250  $\mu L$  of the saturated  $HgCl_2$  solution was added to sample, and processed following the "Guide to best practices for Ocean CO<sub>2</sub> Measurements" (?).

$pH_T$  was determined spectrophotometrically using the indicator dye m-cresol purple (Sigma-Aldrich). Absorbance was measured at 730, 578 and 434 nm before and after dye addition in 10 cm quartz cells thermostated at  $25 \pm 0.05^\circ C$  (?). A similar procedure was carried out before each set of sample measurements using a TRIS (Tris (hydroxymethyl)-aminomethane) buffer prepared at a practical salinity (S) of approximately 30 (?). Certified Reference Material (CRM) (supplied by Professor Andrew Dickson, Scripps Institution of Oceanography, San Diego, USA) was used for quality control of our pH TRIS buffer.

TA was determined by potentiometric titration in an open cell using an automated Radiometer potentiometric titrator (Titrlab 865) and a pH combination electrode (pHC2001) in a continuous titrant addition mode, an algorithm specifically designed for shallow end-point detection (?). The dilute HCl titrant 0.1M in a solution of NaCl of 0.6M was calibrated using CRM provided by Professor Andrew G. Dickson.

The carbonate system parameters (including  $\Omega_A$ ) were corrected for in situ pressure and temperature using the algorithm CO2SYS (?) with measured pH, total alkalinity, soluble reactive and silicate concentrations as input parameters.

$\Omega_A$ , pH, dissolved oxygen ( $O_2$ ) and pressure were sampled at 117 stations in the Estuary and Gulf of St. Lawrence in the summer of 2018. The majority of sampling stations were located in the Southern Gulf ( $n = 84$ ), compared to the Estuary and Northern Gulf ( $n = 33$ ).

To account for low sample size in the Northern Gulf, we used the correlation between  $O_2$  and  $CO_2$ , which are linked through the stoichiometry of the respiration reaction in waters that are isolated from the atmosphere (?). Metabolic  $CO_2$  increases in water through biotic processes such microbial respiration of organic matter (?). In deep waters, variations of  $\Omega_A$  are largely governed by the addition of metabolic  $CO_2$ , whereas near the air-sea interface  $O_2$  is replenished faster than  $CO_2$  can escape (?). The impact of respiration on carbonate chemistry is highest in hypoxic regions, where metabolic  $CO_2$  accumulates to high concentrations (?).

We explored the correlation between  $\Omega_A$  and  $O_2$  in the data and found no significant correlation at the scale of the whole St. Lawrence (p-value: 0.08;  $R^2$ : 0.03). Nevertheless, a correlation between  $\Omega_A$  and  $O_2$  exists in the Northern St. Lawrence (p-value: < 0.01;  $R^2$ : 0.49), but not in the Southern St. Lawrence (p-value: 0.85;  $R^2$ : < 0.01). Furthermore, the correlation observed in the north between  $\Omega_A$  and  $O_2$  is especially strong when deep Channels (Deep model: p-value: < 0.01;  $O_2$ : 0.20;  $R^2$ : 0.97) are separated from shallower regions (Shallow model: p-value: < 0.01;  $O_2$ : 0.07;  $R^2$ : 0.96),

with  $\Omega_A$  increasing more steeply with increasing  $O_2$  in deep Channels. This stronger correlation is likely caused by the lack of gas exchange with the atmosphere or the surface mixed layer and the importance of metabolic  $CO_2$  in deep waters, some of which may have mean ages of up to 20 years (?). In the Northern Gulf, the acidification is following depth and distance from the Cabot Strait along the Laurentian, Anticosti and Esquiman Channels due to the progressive oxygen depletion and metabolic carbon dioxide accumulation (?).

There are likely different dynamics at play in the Southern Gulf, such as waters with lowered alkalinity and salinity caused by the St. Lawrence plume. Such water masses are have lower buffering capacities to decreasing pH and  $\Omega_A$ . Also, the accumulation of metabolic carbon dioxide produced by rapid oxic degradation of organic matter in the sandy and permeable sediments of the Magdalen Shelf or through the St. Lawrence plume could decrease  $\Omega_A$  in the Southern Gulf (Starr and Chassé, personal communication).

We thus divided the analyses between the Southern and Northern St. Lawrence. The St. Lawrence was divided using the 350m isobath on the southern slope of the Laurentian Channel, from the Cabot Strait to the tip of the Gaspé Peninsula.  $\Omega_A$  distribution in the Southern St. Lawrence was modelled using the exponential kriging model.

We predicted  $\Omega_A$  using  $O_2$  for the Northern St. Lawrence. Whereas there are few published datasets on  $\Omega_A$  in the St. Lawrence, there is a robust time-series on oxygen concentration in the water column from the Department of Fisheries and Oceans' (DFO) Atlantic Zone Monitoring Program (AZMP; ?). We used oxygen concentration data collected between 2013 and 2017 (?) and interpolated oxygen % saturation using cokriging with depth as a covariate, as done in (?). Bathymetric data come from ? and have a  $100 \text{ km}^2$  resolution. For more details, refer to the hypoxia description. The Deep model was used to transform  $O_2$  values to  $\Omega_A$  in grid cells with depth  $> 350\text{m}$ , while the Shallow model was applied to cells with depth  $< 350\text{m}$ .

Based on the wide range of  $\Omega_A$  levels at which organism are negatively affected (see above), we built an index of acidification stress ( $A_s$ ) ranging between 0 and 1 that slowly begins increasing as  $\Omega_A \approx 2$ , increases more rapidly close to  $\Omega_A = 1$  and reaches its peak at  $\Omega_A \approx 0.5$ :

$$A_s = \frac{-3}{.99 + e^{-2*\Omega_A}} + 3$$

Hence, the higher the acidification stress, the higher the index of acidification becomes.

## Hypoxia

The data used to characterize hypoxia uses oxygen saturation in % and come from the Department of Fisheries and Oceans' (DFO) Atlantic Zone Monitoring Program (AZMP; ?) surveys in late spring and fall, as well as from annual DFO's multispecies surveys for the northern Gulf in August and for the Magdalen Shallows in September. We provide a brief summary of data and methods to describe hypoxia in this document. For more details, refer to ?.

Oxygen concentration is measured at every station using an oxygen probe (Sea-Bird SBE43) mounted on the CTD. The probe is calibrated against seawater samples that are analyzed by Winkler titration on every cast (see ?, for calibration procedure). The data used is the last depth sampled on the CTD profile, which typically corresponds to ~10m above the bottom. Oxygen concentration ( $\mu M L^{-1}$ ) is converted to oxygen saturation taking into account the salinity and temperature at the selected depth.

We interpolated oxygen saturation using cokriging with depth as a covariate, as done in (?). Bathymetric data come from ? and have a  $100 km^2$  resolution. According to ?, severe hypoxia can be observed when dissolved oxygen falls below  $2 ml L^{-1}$ , which corresponds to  $62.5 \mu mol L^{-1}$  and ~20% saturation. This is considered as the level necessary to maintain most animal life. Similarly, ? studied the effects of hypoxia on

the energy budget of cod in the St. Lawrence and found that behavioural effects began manifesting below 70% oxygen saturation and that survival becomes jeopardized below ~20% saturation. We used these observations to create an index of hypoxia using an inverted logistic curve that slowly increases below 70% oxygen saturation, increases acutely as it approaches the 20% oxygen saturation threshold and reaches its peak intensity between 20 and 30% oxygen saturation:

$$H_s = \frac{-1}{.99 + 200 * e^{-0.15*O_2}} + 1$$

Hence, the higher the hypoxic stress, the higher the index of hypoxia becomes.

### **Bottom temperature anomalies**

The data used to characterize bottom temperature anomalies come from the Department of Fisheries and Oceans' (DFO) Atlantic Zone Monitoring Program (AZMP; ?). We provide a brief summary of data and methods to characterize the bottom temperature climatology and anomalies in this document. For more details, refer to ?.

Bottom temperatures are interpolated in the Gulf using conductivity-temperature-depth (CTD) sampling performed annually through DFO's multispecies surveys for the northern Gulf in August and for the Magdalen Shallows in September. Using this sampling survey, temperatures are horizontally interpolated at each 1 m depth layer on a 2 km resolution grid. Bottom temperatures are then extracted by using a bathymetry layer from the Canadian Hydrographic Survey (?) and selecting the interpolated temperature from the layer corresponding to the bottom depth at each grid point.

We used temperature anomalies, i.e. deviations from long-term normal conditions, to measure an annual index of stress associated with extreme temperatures between 2013 and 2017. Temperature anomalies were calculated using the difference between grid cell values with 1981-2010 climatological averages. Anomaly time series were normalized by

their standard deviation (SD) to allow comparison between areas of the St. Lawrence with different temperature ranges. For example, temperatures observed in deep channels are less variable than in shallower regions of the St. Lawrence. Hence, if anomalies were expressed in degrees Celsius, it would underestimate the relative importance of anomalies in deep channels when compared to shallower regions. Grid cells whose monthly value exceeded  $\pm 0.5$  standard deviation (SD) from the long-term average were considered as anomalous (?). Outliers in the data were defined as those that fell beyond the interquartile range \* 3, identified as extreme outliers by ?. Outlier values were capped to correspond to the 5th and 95th percentiles. Anomalies were divided into positive and negative anomalies and the absolute value of anomalies were used as an indicator of the intensity of bottom temperature anomalies. The mean anomaly intensity between 2013 and 2017 for each grid cell was used to generate the final index of bottom temperature anomalies.

### **Surface temperature anomalies**

The data used to characterize surface temperature anomalies come from the Department of Fisheries and Oceans' (DFO) Atlantic Zone Monitoring Program (AZMP; ?). We provide a brief summary of data and methods to characterize the surface temperature climatology and anomalies in this document. For more details, refer to ?.

The surface layer is characterized using sea surface temperature (SST) monthly composites from Advanced Very High Resolution Radiometer (AVHRR) satellite images obtained from the National Oceanic and Atmospheric Administration (NOAA) and European Organization for the Exploitation of Meteorological Satellites (EUMETSAT). Images used are from DFO's Maurice Lamontagne Institute at a 1 km resolution from 1985-2013 and from DFO's Bedford Institute of Oceanography (BIO) Operational Remote Sensing group at a 1.5 km resolution since 2014. Monthly anomalies were constructed as the difference between monthly averages and the 1985-2010 climatological

mean for each month.

Surface temperature anomalies were characterized following the same method used for bottom temperature anomalies. Only the months of May to November were included to avoid biases associated with the presence of ice cover. Monthly anomalies from May to November values were summed to obtain an indicator of annual surface temperature anomaly intensity in each grid cell. The mean anomaly intensity between 2013 and 2017 for each grid cell was used to generate the final index of surface temperature anomalies.

### **Sea level rise**

The data used to characterize sea level rise risk comes from the global cumulative impacts assessment on habitats (??) and available on the NCEAS online data repository (?). We provide a brief summary of data and methods in this document. For more details, refer to ?.

Sea level rise was characterized by ? using NASA's satellite altimetry data (Topex/-Poseidon, Jason-1&2, GFO, ERS-1&2, and Envisat missions) and available at <http://www.aviso.altimetry.fr/en/data/products/ocean-indicatorsproducts/mean-sea-level/products-images.html>

The rate of sea level rise (*mm/year*) was measured between 1992 and 2012 and transformed as a net change value (*mm*) by multiplying by the number of years considered. Only positive values were selected under the assumption that only positive sea level rise is likely to cause environmental stress.

For the St. Lawrence, we overlaid the raw data layers (?) with our  $1 \text{ km}^2$  grid cell using weighted area average.

## **Coastal**

### **Aquaculture**

Aquaculture data comes from a variety of sources in the St. Lawrence because aquaculture sites are mostly managed at the provincial level. We therefore had to gather the data on aquaculture sites from the 5 provinces dividing the St. Lawrence (?????).

Invertebrates aquaculture is especially important in the southern and western Gulf. Fish and algae aquaculture, on the other hand, remains marginal. Considering this, we only considered invertebrates aquaculture for the aquaculture driver layer. However, if fish or algae farming were to become more important, these driver should be incorporated in future analyses as individual layers, as impacts vary between types of aquaculture.

Aquaculture activities are highly localized and potential effects do not or rarely extend beyond the location of the farms. We therefore only considered the actual location of sites to characterize the distribution of this driver. We were unable to characterize site production in terms of biomass farmed, which could provide an indication of the intensity of aquaculture activities. As such, we considered aquaculture as binary presence-absence data in our analyses.

### **Coastal development**

We used lights at night as a proxy of coastal infrastructure development, as terrestrial stable lights at night represent light from human settlements and industrial sites with electricity.

The data comes from the Nighttime Lights Time Series. Nighttime light products are compiled by the Earth Observation Group at the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Information (NCEI). They use globally available nighttime data obtained from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) of the Defense Meteorological

Satellite Program (DMSP) to characterize global average radiance ( $nanoWatts\ cm^{-2}\ sr^{-1}$ ) composite images at a 15 arc-second (~200 m) resolution.

We used the annual Version 1 Nighttime VIIRS DNB composites between 2015 and 2016 (?) to characterize coastal development in coastal areas of the St. Lawrence. As the effects of coastal development are likely acute in its direct vicinity, we extracted average radiance values using a 2 km buffer around grid cells within 2 km of the coast. We used a weighted area average to extract the radiance values.

### **Direct human impact**

As in ? and ?, we used the sum of coastal populations as a proxy of direct human impact. We used Statistics Canada dissemination area population count from the 2016 census to obtain coastal population size around the St. Lawrence (?). Dissemination areas are the smallest standard geographic area in which census data are disseminated and they combine to cover all of Canada. The census provides population count within the boundary of each dissemination area, which we used to evaluate total coastal population.

As the effects of direct human impacts are likely acute mostly in coastal areas we calculated total population in grid cells within 2 km of the coast. Total population was measured in a 10 km buffer around each coastal cell. The total population in each buffer was the sum of intersecting dissemination areas divided by the intersection area between buffers and dissemination areas:

$$DHI_j = \sum_{k=1}^{n_j} P_k * \frac{A_{j,k}}{Atot,k}$$

where  $j$  is a buffered grid cell,  $k$  is a dissemination area intersecting  $j$ ,  $P$  is the population in  $k$ ,  $A$  is the area of the  $k$  overlapping with  $j$  and  $A_{tot}$  is the total area of  $k$ . This approach was favoured to reduce the effects of very large dissemination areas overlapping with buffers on a very small percentage of their total area.

## Inorganic pollution

The data used to characterize inorganic pollution comes from the global cumulative impacts assessment on habitats (??) and available on the NCEAS online data repository (?). We provide a brief summary of data and methods in this document. For more details, refer to ?.

Inorganic pollution was modelled using impervious surface area (i.e. artificial surfaces such as paved roads) under the assumption that most of this pollution source comes from urban runoff. Inorganic pollution originating from point-sources or in areas lacking paved roads is therefore not captured by this layer. The data obtained was aggregated at the watershed scale and spread into coastal and marine environments was modelled using a diffusive plume model from each watershed pourpoints (e.g. river mouths).

For the St. Lawrence, we overlaid the raw data layers (?) with our  $1 \text{ km}^2$  grid cell using weighted area average.

## Nutrient pollution

The data used to characterize nutrient pollution comes from the global cumulative impacts assessment on habitats (??) and available on the NCEAS online data repository (?). We provide a brief summary of data and methods in this document. For more details, refer to ?.

Annual fertilizer use in tonnes ( $t$ ) was used as a proxy of nutrient pollution. The data used came from the Food and Agriculture Organization of the United Nations (FAO). Gaps in data were modelled using a linear regression between fertilizer and pesticides or agricultural gross domestic product (GDP). Dasymetric maps were then used to distribute fertilizer data over the landscape using 2009 data from the Moderate Resolution Imaging Spectroradiometer (MODIS) at ~500 m resolution and aggregated to watersheds. Diffusive plume models from each watershed pourpoint (e.g. river

mouths) were then used to model the distribution and intensity of nutrient pollution in coastal and marine environments.

For the St. Lawrence, we overlaid the raw data layers (?) with our  $1 \text{ km}^2$  grid cell using weighted area average.

### Organic pollution

The data used to characterize organic pollution comes from the global cumulative impacts assessment on habitats (??) and available on the NCEAS online data repository (?). We provide a brief summary of data and methods in this document. For more details, refer to ?.

Annual pesticide use in tonnes ( $t$ ) was used as a proxy of organic pollution. The data used came from the FAO and gaps in data were modelled using a linear regression between pesticides and fertilizers or agricultural GDP. The same methodology as that used to characterized nutrient pollution was then applied to organic pollution.

For the St. Lawrence, we overlaid the raw data layers (?) with our  $1 \text{ km}^2$  grid cell using weighted area average.

### Toxic algae

The data we use to describe the risk of toxic algae comes from an expert based map delineating the areas where coastal areas are at risk from five different toxins (?). The map presents coastal areas at risk from 5 different toxins: 1) paralytic shellfish poisoning (PSP) toxins from the regular presence of the dinoflagellate Alexandrium catenella (previously Alexandrium tamarense) at high concentrations, 2) amnesic shellfish poisoning (ASP) toxins from domoic acid 3) diarrhetic shellfish poisoning (DSP) toxins, 4) spirolides and 5) pectenotoxins, two toxins produced by dinoflagellates occurring in the St. Lawrence.

The information provided on this expert map on the 5 toxins (?). was georeferenced and transformed as vectorized objects. We calculated a toxic algae risk ( $T$ ) index for each cell ( $x$ ) in the  $1 \text{ km}^2$  study grid. For each toxin ( $t$ ), a value of 1 was attributed to all grid cells overlapping with areas identified at risk on the expert map and a value of 0.5 for grid cells overlapping with areas where ASP and DSP toxins were observed without exceeding legal thresholds. The value for all 5 toxins was then summed for all grid cells:

$$TA_{i,x} = \sum_{i=1}^5 i_x$$

## Fisheries

The impacts of fisheries activities in the St. Lawrence are evaluated using DFO's fisheries logbooks program (?). While logbooks are not mandatory for all fisheries in the St. Lawrence, they still provide a very thorough overview of the spatial distribution and intensity of fishing activities in the St. Lawrence. The data we used spans 6 years from 2010 to 2015 and details 218323 fishing events ( $36387 \pm 3147$  fishing events per year). There were 31 targeted species and a total of 53 caught species in the dataset.

Fishing activities are performed using a variety of gear types: trap, trawl, dredge, drift-net, hand line, longline, scuba diving, purse seine, seine, beach seine and jig fishing. Intensity of fishing activities was divided among gear types and based on their respective types of environmental impacts (Table S2). For example, traps and trawls have very different effects on a system. Gear classification was done using the classification presented in ? and ? and is broken down into 5 distinct classes: demersal destructive (DD), demersal, non-destructive, low-bycatch (DNL), demersal, non-destructive, high-bycatch (DNH), pelagic, low-bycatch (PLB) and pelagic, high-bycatch (PHB). This categorization therefore divides the fisheries data into 5 distinct driver layers characterizing fishing activities.

Gear types can also be further classified into fixed or mobile engines based on their mobility. We used these two mobility classes to generate a buffer of impact around each fishing activity coordinates to consider potential spatial uncertainty associated with locations and the fact that mobile engines can be tracted over several kilometers during fishing activities and that we do not have the beginning and end points of mobile fishing events. Buffer sizes for fixed and mobile engine was of 200 and 2000 meters, respectively.

Table S3.2. Classification of gear types in the fisheries dataset based on their environmental impact and mobility

Gear type (EN)	Classification	Mobility
Trap	DNH	Fixed
Trawl	DD	Mobile
Dredge	DD	Mobile
Driftnet	PHB	Fixed
Hand lines	PLB	Fixed
Longline	PHB	Fixed
Scuba diving	DNL	Fixed
Purse seine	PLB	Fixed
Seine	DNH	Fixed
Beach seine	DNH	Fixed
Trap	DNH	Fixed
Jig fishing	PLB	Fixed

In order to characterize the intensity of fishing activities ( $FI$ ), we used a biomass yield density index. We multiplied the total annual biomass captured in each grid cell  $j$ , regardless of species, by the proportion of fishing area in each grid cell:

$$FI_j = \sum_{k=1}^{n_j} B_{tot,k} * \frac{A_{j,k}}{A_{tot,k}}$$

where  $j$  is a grid cell,  $k$  is a fishing event,  $B_{tot}$  is the total biomass of a fishing event  $k$ ,  $A$  is the area of a fishing event  $k$  overlapping a cell  $j$  and  $A_{tot}$  is the total area of the fishing event  $k$ . This formula gives an intensity measurement in biomass units, which is kg in our case. Since we measure the intensity within a  $1\ km^2$  grid cell, the intensity evaluation is in  $kg * km^{-2}$ . This metric distributes the biomass captured within each grid cell as a function of overlapping fishing area and provides an overview of how impacted each grid cell is in terms of extracted biomass.

## Marine traffic

### Shipping

The data used to characterize shipping comes from the global cumulative impacts assessment on habitats (??) and available on the NCEAS online data repository (?). We provide a brief summary of data and methods in this document. For more details, refer to ?.

Two data sources were used to characterize shipping. The first set of data is gathered as part of the World Meteorological Organization Voluntary Observing Ships' (VOS) scheme. Ships participating in the program gather meteorological data along with observation location as part of an open-ocean climate dataset. The data spans 20 years and annually covers 10-20% of ships worldwide. Data used spanned 2003 to 2011.

The second set of data comes from the Automatic Identification System (AIS), an initiative launched in 2002 that sought to improve marine safety by providing mariners with real-time vessel traffic. Through the International Maritime Organization SOLAS agreement, all vessels of over 300 gross tonnage on international voyages and those carrying passengers are now required to be equipped with AIS transceivers. These

transceivers use Global Positioning System technology to locate vessels every 10 minutes. The data used was from November 2010 to December 2011.

Data used come mostly from vessels that move globally (*i.e.* cargo, tanker and passenger), as they are required to carry AIS transceivers, but also include data from fishing, high-speed, pleasure and support classes. Shipping intensity was evaluated as the number of fishing tracks at a 0.1 decimal degrees resolution. For more details on data and methods used, consult (?).

For the St. Lawrence, we overlaid the raw data layers (?) with our  $1 \text{ km}^2$  grid cell using weighted area average.

### Invasive species

The data used to characterize invasive species risk comes from the global cumulative impacts assessment on habitats (??) and available on the NCEAS online data repository (?). We provide a brief summary of data and methods in this document. For more details, refer to supplementary materials provided in ? and ?.

Cargo volume was used as a proxy of invasion risk under the assumption that risk of invasion is proportional to tonnes of goods transferred through ports. Cargo throughput in metric tonnes for the year 2011 was accessed through a variety of sources (see supplementary material in ?, for more details) and cross-matched with entries in the World Port Index database (WPI; available from the National Geospatial-Intelligence Agency). A gap-filling procedure using linear regression and sets of predictors related to port volume and available in the WPI dataset was then applied to the WPI dataset to predict missing cargo volume entries. Finally, volume data was distributed in marine environments adjacent to ports using a diffusive plume model with an exponential decay function that set the maximum spread distance to approximately 1000 km. The plume model was then clipped to areas less than 60 m deep, as invasive species are more likely to invade shallow areas.

For the St. Lawrence, we overlaid the raw data layers (?) with our  $1 \text{ km}^2$  grid cell using weighted area average.

### Marine pollution

The data used to characterize marine pollution risk comes from the global cumulative impacts assessment on habitats (??) and available on the NCEAS online data repository (?). Marine pollution was considered to be mainly driver by the shipping industry. As such, the driver layer was constructed by combining the shipping (*i.e.* shipping lanes) and invasive species (*i.e.* cargo volume) layers. invasive. For more details, refer to supplementary materials provided in ? and ?.

For the St. Lawrence, we overlaid the raw data layers (?) with our  $1 \text{ km}^2$  grid cell using weighted area average.

### Driver intensity and distribution

We evaluated the frequency distribution of each drivers to verify whether data should be transformed (Figure S11). In light of this, we log-transformed the following driver layers

- Coastal development
- Direct human impact
- All fisheries data
- Hypoxia
- Inorganic pollution
- Invasive species
- Nutrient pollution
- Organic pollution
- Sea bottom temperature anomalies

— Shipping

To allow for relative intensity comparison, all driver layers were subsequently normalized between 0 and 1 using the 99th quantile to further control for extreme values (Figure S12).

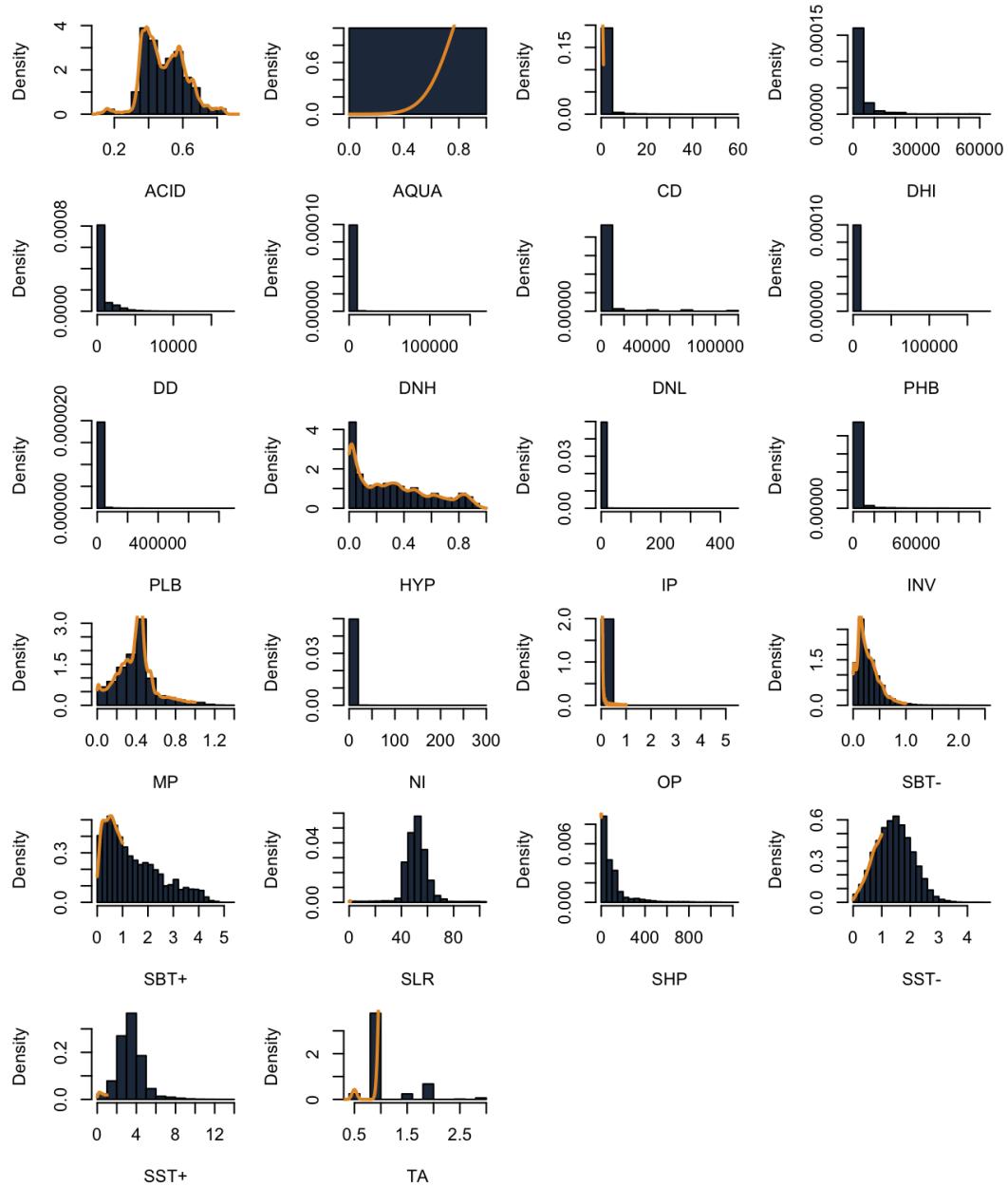


Figure 11: Frequency distribution of the untransformed data for all driver layers.

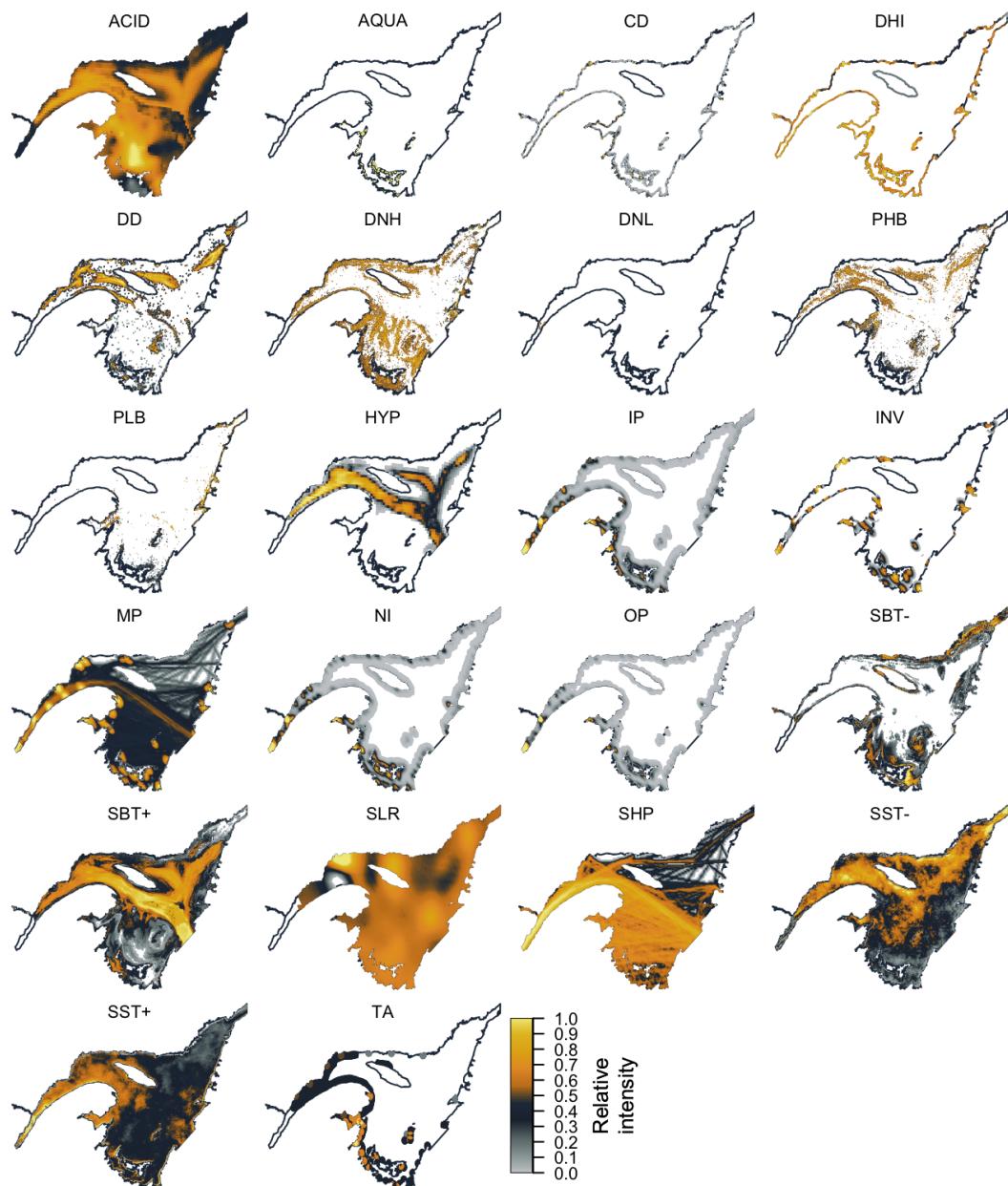


Figure 12: Distribution and intensity of transformed and normalized drivers in the Estuary and Gulf of St. Lawrence available on \*eDrivers\*.

## Cumulative exposure

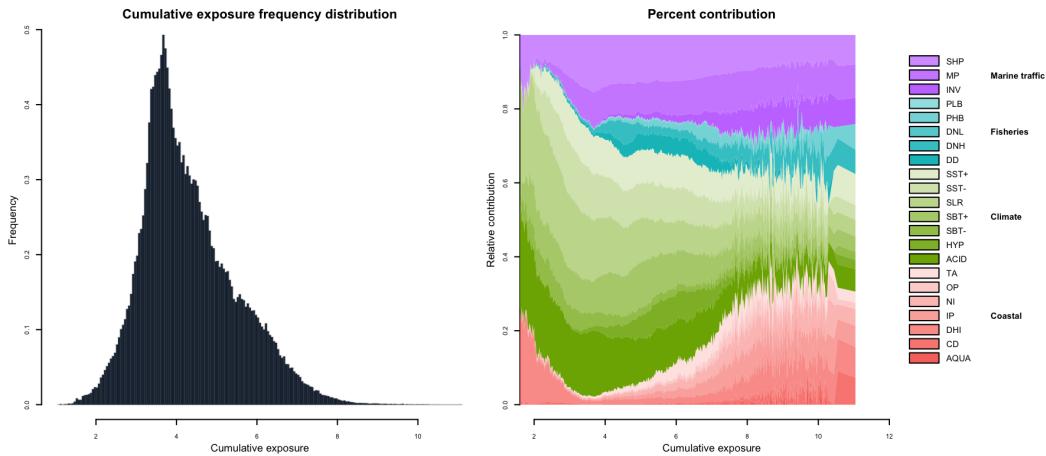


Figure 13: Frequency distribution of cumulative exposure (*i.e.* sum of normalized driver intensity in each grid cell) and percent contribution of each driver to the frequency distribution of cumulative exposure in the Estuary and Gulf of St. Lawrence.

## Threat complexes

### Clustering

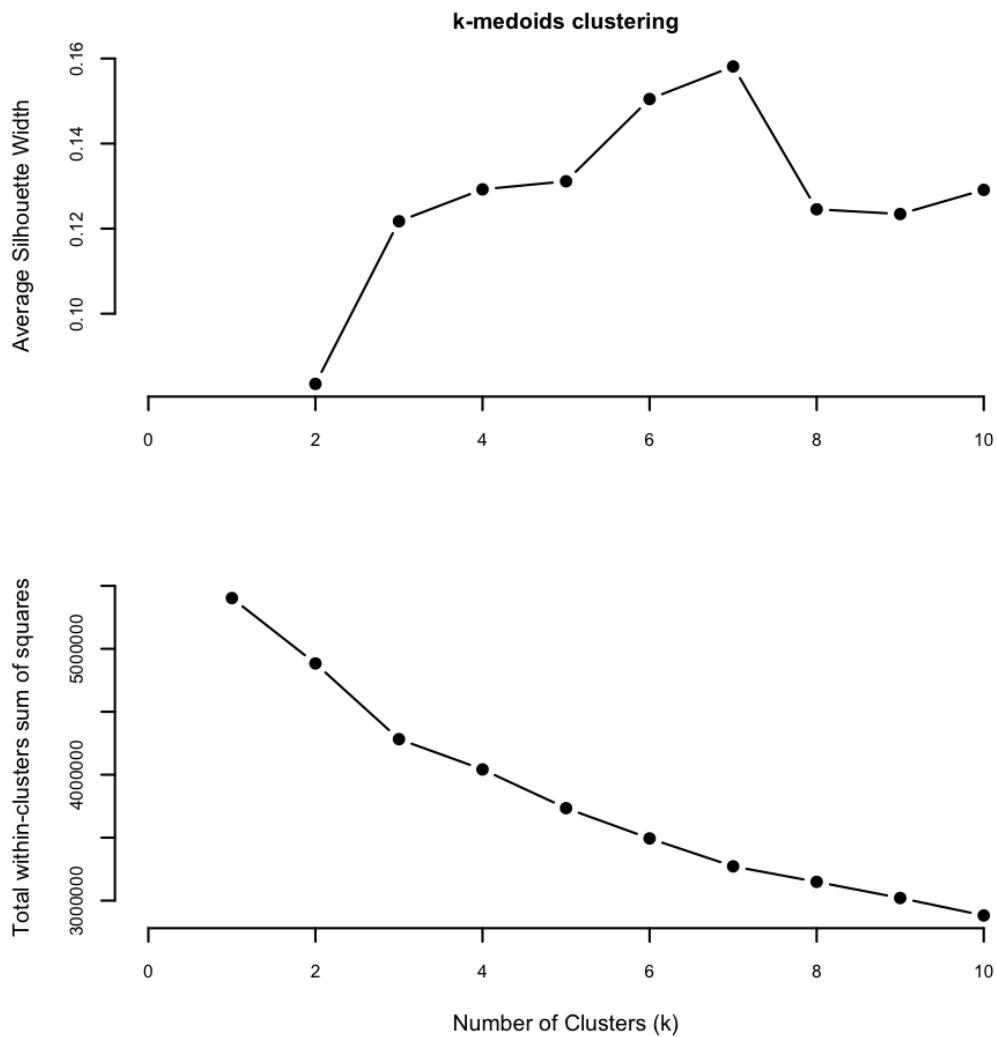


Figure 14: Validation procedure for the  $k$ -medoids and  $k$ -means clustering algorithms based on the number of cluster that maximizes average silhouette width (upper panels; Kaufman and Rousseeuw, 1990) and minimizes the total within-cluster sum of squares (WSS; lower panels).

## Inter-cluster dissimilarity

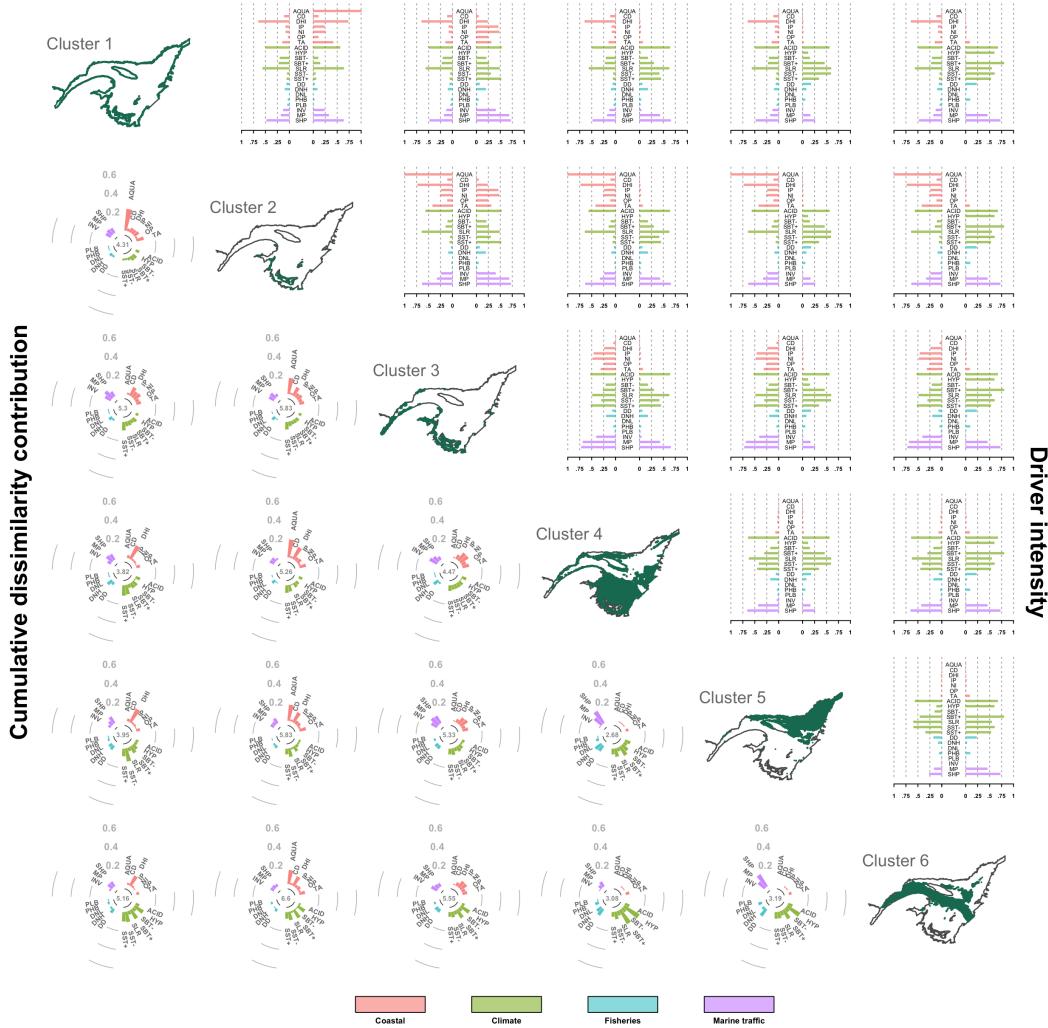


Figure 15: Evaluation of inter-cluster dissimilarity using a similarity percentage analysis (SIMPER) with Manhattan distance (Clarke, 1993). The figure diagonal presents the distribution of the 6 clusters identified using the *k-medoids* clustering algorithm. The lower triangle shows all combinations of inter-cluster dissimilarity with circular barplots showing the percent contribution to total dissimilarity of each driver and with the total inter-cluster dissimilarity in the center of the barplots. The upper triangle shows the average relative intensity of each driver for all driver combinations, with barplots to the left and the right representing the row and columns clusters, respectively.

## Intra-cluster similarity

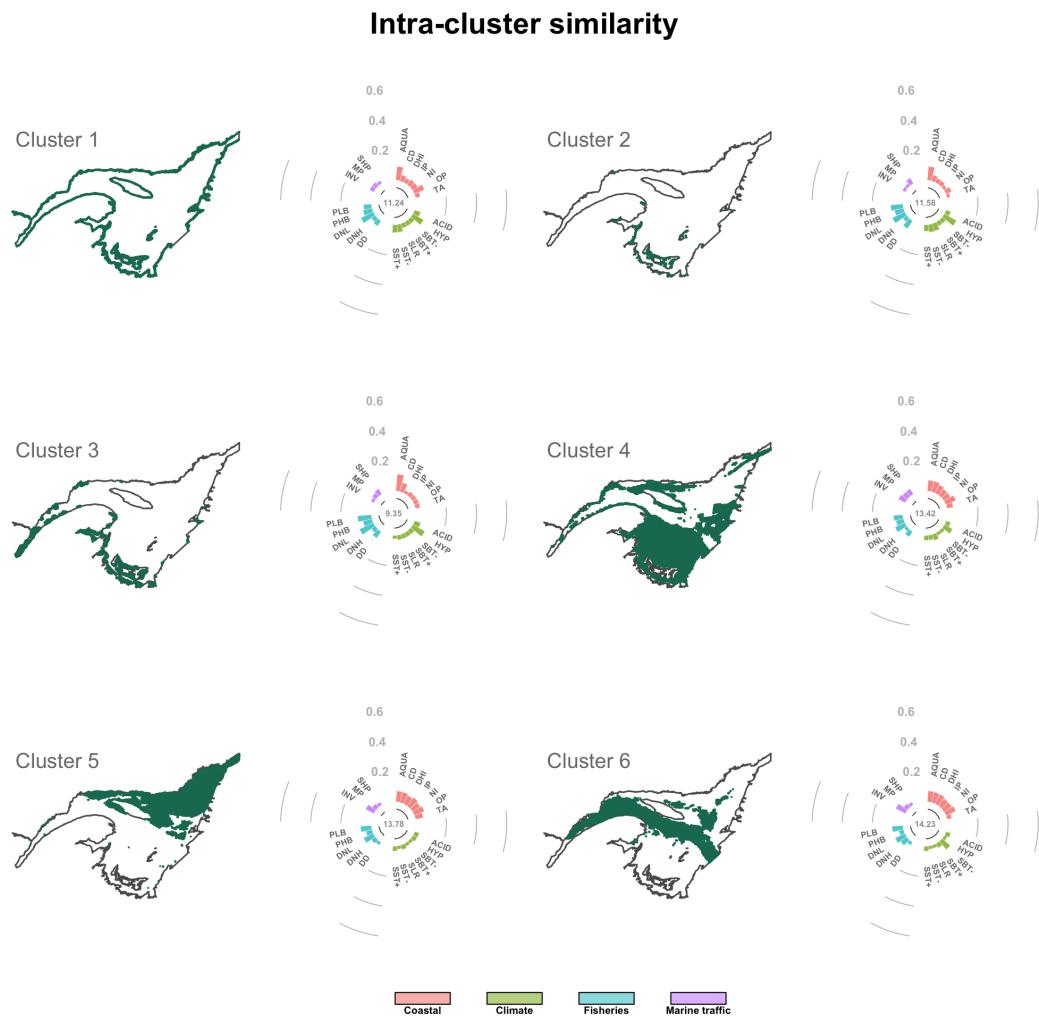


Figure 16: Evaluation of intra-cluster similarity using the Manhattan distance transformed to a similarity index. The distribution of the 6 clusters is presented along with circular barplots showing the percent contribution to total similarity of each driver and with the total intra-cluster similarity in the center of the barplots.

## **ANNEXE III**

### **THINKING OUTSIDE THE BOX - PREDICTING BIOTIC INTERACTIONS IN DATA-POOR ENVIRONMENTS - SUPPLEMENTARY INFORMATION**

Functional group name	Functional group main taxa composition
Cetaceans	<i>Balaenoptera physalus, B. acutorostrata, Megaptera novaeangliae, Phocoena phocoena, Lagenorhynchus acutus, L. albirostris</i>
Harp seals	<i>Pagophilus groenlandicus</i>
Hooded seals	<i>Cystophora cristata</i>
Grey seals	<i>Halichoerus grypus</i>
Harbour seals	<i>Phoca vitulina</i>
Seabirds	<i>Phalacrocorax carbo, P. auritus, Larus delawarensis, L. argentatus, L. marinus, Sterna hirundo, S. paradisaea, Cephus grylle, Oceanodroma leucorhoa, Morus bassanus, Rissa tridactyla, Uria aalge, Alca torda, Fratercula arctica</i>
Atlantic cod	<i>Gadus morhua</i>
Greenland halibut	<i>Reinhardtius hippoglossoides</i>
American plaice	<i>Hippoglossoides platessoides</i>
Flounders	<i>Limanda ferruginea, Glyptocephalus cynoglossus, Pseudopleuronectes americanus</i>
Skates	<i>Amblyraja radiata, Malacoraja senta, Leucoraja ocellata</i>
Redfish	<i>Sebastodes mentella, S. fasciatus</i>
Large demersal feeders	<i>Urophycis tenuis, Melanogrammus aeglefinus, Centroscyllium fabricii, Anarhichassp., Cyclopterus lumpus, Lycolessp., Macrouridae, Zoarcidae, Lophius americanus, Hippoglossus hippoglossus</i>
Small demersal feeders	<i>Myoxocephalusssp., Tautogolabrus adspersus, Zoarcesamericanus, large demersal juveniles</i>
Capelin	<i>Mallotus villosus</i>
Large pelagic feeders	<i>Squalus acanthias, Pollachius virens, Merluccius bilinearis, Cetorhinus maximus</i>
Piscivorous small pelagic feeders	<i>Scomber scombrus, Illex illecebrosus, piscivorous myctophids and other mesopelagics, piscivorous large pelagic juveniles</i>
Planktivorous small pelagic feeders	<i>Clupea harengus harengus, Scomberesox saurus, Gonatussp., planktivorous myctophids and other mesopelagics, planktivorous large pelagic juveniles</i>
Shrimp	<i>Argis dentata, Eualus macilentus, E. gaimardi, Pandalus montagui</i>

Functional group name	Functional group main taxa composition
Molluscs	<i>Mesodesma deauratum, Cyrtodaria siliqua</i>
Polychates	<i>Parexogone hebes</i>
Small zooplankton	<i>Oithona similis, Temora longicornis, Pseudocalanusspp., Calanus finmarchicus</i> , tunicates, meroplankton, heterotrophic protozoa
Phytoplankton	<i>Chaetoceros affinis, Chaetocerosspp., Leptocylindrus minimus, Thalassiosira nordenskioldii, Thalassiosirasp., Fragilariopsissspp.</i> , other diatoms, mixture of autotrophic and mixotrophic organisms including Cryptophytes, dinoflagellates, Prasinophytes and Prymnesiophytes

Table 3: List of functional groups included in the dataset presented in ? with their taxa composition. Only taxa that were at least at the scale of the family were used to predict interactions. List adapted from ?.

Consumer	Resource
Skates	Skates
Atlantic cod	Skates
Hooded seals	Shrimp
Piscivorous small pelagic feeders	Shrimp
Planktivorous small pelagic feeders	Phytoplankton
Planktivorous small pelagic feeders	Large crustaceans
Hooded seals	Large crustaceans
Echinoderms	Large crustaceans
Flounders	Large crustaceans
Seabirds	Large crustaceans
Greenland halibut	Large crustaceans
Piscivorous small pelagic feeders	Large crustaceans
Redfish	Large crustaceans
Planktivorous small pelagic feeders	Planktivorous small pelagic feeders
American plaice	Planktivorous small pelagic feeders
Echinoderms	Echinoderms
Large demersal feeders	Echinoderms
Planktivorous small pelagic feeders	Atlantic cod
American plaice	Atlantic cod
Flounders	Atlantic cod
Greenland halibut	Atlantic cod
Piscivorous small pelagic feeders	Atlantic cod
Cetaceans	American plaice
Planktivorous small pelagic feeders	American plaice
Hooded seals	American plaice
American plaice	American plaice
Flounders	American plaice
Harbour seals	American plaice
Piscivorous small pelagic feeders	American plaice
Redfish	American plaice
Large pelagic feeders	American plaice

Consumer	Resource
Cetaceans	Small zooplankton
Large crustaceans	Small zooplankton
Large pelagic feeders	Small zooplankton
Large demersal feeders	Small zooplankton
Atlantic cod	Seabirds
Seabirds	Seabirds
Large demersal feeders	Seabirds
Harbour seals	Harbour seals
Skates	Greenland halibut
Cetaceans	Greenland halibut
Planktivorous small pelagic feeders	Greenland halibut
Atlantic cod	Greenland halibut
American plaice	Greenland halibut
Flounders	Greenland halibut
Small demersal feeders	Greenland halibut
Harbour seals	Greenland halibut
Piscivorous small pelagic feeders	Greenland halibut
Redfish	Greenland halibut
Large pelagic feeders	Greenland halibut
Planktivorous small pelagic feeders	Piscivorous small pelagic feeders
American plaice	Piscivorous small pelagic feeders
Flounders	Piscivorous small pelagic feeders
Small demersal feeders	Piscivorous small pelagic feeders
Piscivorous small pelagic feeders	Piscivorous small pelagic feeders
Atlantic cod	Redfish
Harp seals	Redfish
Seabirds	Redfish
Redfish	Redfish
Large pelagic feeders	Redfish
Skates	Large pelagic feeders
Planktivorous small pelagic feeders	Large pelagic feeders

Consumer	Resource
Redfish	Large pelagic feeders
Large pelagic feeders	Large pelagic feeders
Large demersal feeders	Large pelagic feeders
Skates	Large demersal feeders
Cetaceans	Large demersal feeders
Planktivorous small pelagic feeders	Large demersal feeders
Atlantic cod	Large demersal feeders
American plaice	Large demersal feeders
Flounders	Large demersal feeders
Small demersal feeders	Large demersal feeders
Seabirds	Large demersal feeders
Greenland halibut	Large demersal feeders
Piscivorous small pelagic feeders	Large demersal feeders
Redfish	Large demersal feeders
Large pelagic feeders	Large demersal feeders
Large demersal feeders	Large demersal feeders

Table 4: List of functional groups for which interactions were predicted by the algorithm, but not observed in ? (b).

Consumer	Resource
Grey seals	Skates
Seabirds	Skates
Harbour seals	Skates
Cetaceans	Shrimp
Shrimp	Phytoplankton
Mollusks	Phytoplankton
Polychaetes	Phytoplankton
Grey seals	Large crustaceans
Flounders	Planktivorous small pelagic feeders
Flounders	Echinoderms
Small demersal feeders	Echinoderms
Grey seals	American plaice
Hooded seals	Flounders
Harp seals	Flounders
Skates	Mollusks
Large crustaceans	Mollusks
Atlantic cod	Mollusks
American plaice	Mollusks
Flounders	Mollusks
Small demersal feeders	Mollusks
Harbour seals	Mollusks
Seabirds	Small zooplankton
Skates	Polychaetes
Shrimp	Polychaetes
Large crustaceans	Polychaetes
Atlantic cod	Polychaetes
American plaice	Polychaetes
Flounders	Polychaetes
Small demersal feeders	Polychaetes
Polychaetes	Polychaetes
Large pelagic feeders	Polychaetes

## **ANNEXE IV**

## **ANNEXE IV**

## RÉFÉRENCES

- AAF (2016). Marine Aquaculture Site Mapping Program (MASMP). Department of Agriculture, Aquaculture and Fisheries (AAF), Government of New-Brunswick, New-Brunswick, Canada. Available at: [https://www2.gnb.ca/content/gnb/en/departments/10/aquaculture/\\_content/-masmp.html](https://www2.gnb.ca/content/gnb/en/departments/10/aquaculture/_content/-masmp.html). Technical report, Department of Agriculture, Aquaculture and Fisheries.
- Archambault, P., Schloss, I. R., Grant, C., and Plante, S., editors (2017). Les Hydrocarbures Dans Le Golfe Du Saint-Laurent - Enjeux Sociaux, Économiques et Environnementaux. Notre Golfe, Rimouski, Qc, Canada.
- Bates, S., Beach, D., Comeau, L., Haigh, N., Lewis, N., Locke, A., Martin, J., McCaron, P., McKenzie, C., Michel, C., Miles, C., Quilliam, M., Rourke, W., Scarratt, M., Starr, M., and Wells, T. (2019). Marine harmful algal blooms and phycotoxins of concern to Canada. *Can. Tech. Rep. Fish. Aquat. Sci.* In revision. Technical report, Department of Fisheries and Oceans.
- Beauchesne, D., Grant, C., Gravel, D., and Archambault, P. (2016). L'évaluation des impacts cumulés dans l'estuaire et le golfe du Saint-Laurent : vers une planification systémique de l'exploitation des ressources. *Le Naturaliste Canadien*, 140(2):45–55.
- Belley, R., Archambault, P., Sundby, B., Gilbert, F., and Gagnon, J.-M. (2010). Effects of hypoxia on benthic macrofauna and bioturbation in the Estuary and Gulf of St. Lawrence, Canada. *Continental Shelf Research*, 30(12):1302–1313.
- Benoît, H. P., Gagné, J. A., Savenkoff, C., Ouellet, P., and Bourassa, M.-N. (2012). State of the Ocean Report for the Gulf of St. Lawrence Integrated Management (GOSLIM). Technical Report 2986, Department of Fisheries and Oceans.
- Blais, M., Galbraith, P. S., Plourde, S., Scarratt, M., Devine, L., and Lehoux, C. (2019). Chemical and Biological Oceanographic Conditions in the Estuary and Gulf of St. Lawrence during 2017. DFO Can. Sci. Advis. Sec. Res. Doc. 2019/009. iv + 56 p. Technical Report 2018/037, Department of Fisheries and Oceans.
- Bove, C. B., Ries, J. B., Davies, S. W., Westfield, I. T., Umpanhowar, J., and Castillo, K. D. (2019). Common Caribbean corals exhibit highly variable responses to future acidification and warming. *Proceedings of the Royal Society B: Biological Sciences*, 286(1900):20182840.
- Bowler, D., Bjorkmann, A., Dornelas, M., Myers-Smith, I., Navarro, L., Niamir, A., Supp, S., Waldock, C., Vellend, M., Blowes, S., Boehning-Gaese, K., Bruelheide, H., Elahi, R., Antao, L., Hines, J., Isbell, F., Jones, H., Magurran, A., Cabral, J., Winter,

- M., and Bates, A. (2019). The geography of the Anthropocene differs between the land and the sea. *bioRxiv*.
- Chabot, D. and Claireaux, G. (2008). Environmental hypoxia as a metabolic constraint on fish: The case of Atlantic cod, *Gadus morhua*. *Marine Pollution Bulletin*, 57(6):287–294.
- Chion, C., Turgeon, S., Cantin, G., Michaud, R., Ménard, N., Lesage, V., Parrott, L., Beaufils, P., Clermont, Y., and Gravel, C. (2018). A voluntary conservation agreement reduces the risks of lethal collisions between ships and whales in the St. Lawrence Estuary (Québec, Canada): From co-construction to monitoring compliance and assessing effectiveness. *PLOS ONE*, 13(9):e0202560.
- Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18(1):117–143.
- Côté, I. M., Darling, E. S., and Brown, C. J. (2016). Interactions among ecosystem stressors and their importance in conservation. *Proceedings of the Royal Society B: Biological Sciences*, 283(1824):20152592.
- Crain, C. M., Kroeker, K., and Halpern, B. S. (2008). Interactive and cumulative effects of multiple human stressors in marine systems. *Ecology Letters*, 11(12):1304–1315.
- Dafforn, K. A., Johnston, E. L., Ferguson, A., Humphrey, C. L., Monk, W., Nichols, S. J., Simpson, S. L., Tulbure, M. G., and Baird, D. J. (2016). Big data opportunities and challenges for assessing multiple stressors across scales in aquatic ecosystems. *Marine and Freshwater Research*, 67(4):393–413.
- Darling, E. S. and Côté, I. M. (2008). Quantifying the evidence for ecological synergies. *Ecology Letters*, 11(12):1278–1286.
- De Leo, F. C., Gauthier, M., Nephin, J., Mihály, S., and Juniper, S. K. (2017). Bottom trawling and oxygen minimum zone influences on continental slope benthic community structure off Vancouver Island (NE Pacific). *Deep Sea Research Part II: Topical Studies in Oceanography*, 137:404–419.
- Dee, L. E., Allesina, S., Bonn, A., Eklöf, A., Gaines, S. D., Hines, J., Jacob, U., McDonald-Madden, E., Possingham, H., Schröter, M., and Thompson, R. M. (2017). Operationalizing Network Theory for Ecosystem Service Assessments. *Trends in Ecology & Evolution*, 32(2):118–130.
- Dempsey, D. P., Gentleman, W. C., Pepin, P., and Koen-Alonso, M. (2018). Explanatory Power of Human and Environmental Pressures on the Fish Community of the Grand Bank before and after the Biomass Collapse. *Frontiers in Marine Science*, 5.

- DFO (2016a). Prince Edward Island Aquaculture Leases. Aquaculture Division. Department of Fisheries and Oceans Canada (DFO). Prince-Edward-Island, Canada. Available at: <http://www.arcgis.com/home/item.html?id=16aa8830c7084a8a92ce066b525978b4>. Technical report, Department of Fisheries and Oceans.
- DFO (2016b). Zonal Interchange File Format (ZIFF) data. A compilation of landing data from logbook data between 2010 and 2015. Gestion des données, Institut Maurice Lamontagne, Department of Fisheries and Oceans (DFO) Mont-Joli, Canada. Technical report, Department of Fisheries and Oceans.
- Diaz, R. J. and Rosenberg, R. (1995). Marine benthic hypoxia: A review of its ecological effects and the behavioural responses of benthic macrofauna. Oceanography and Marine Biology. An annual review, 33:245–303.
- Dickson, A. G., Sabine, C. L., and Christian, J. R. (2007). Guide to Best Practices for Ocean CO<sub>2</sub> Measurements. Report, North Pacific Marine Science Organization.
- Dufour, R. and Ouellet, P. (2007). Estuary and Gulf of St. Lawrence marine ecosystem overview and assessment report. Can. Tech. Rep. Fish. Aquat. Sci. 2744E: Vii + 112 p. Technical Report 2744E.
- Dupont-Prinet, A., Vagner, M., Chabot, D., and Audet, C. (2013). Impact of hypoxia on the metabolism of Greenland halibut (*Reinhardtius hippoglossoides*). Canadian Journal of Fisheries and Aquatic Sciences, 70(3):461–469.
- Dutil, J.-D., Proulx, S., Chouinard, P.-M., and Borcard, D. (2011). A hierarchical classification of the seabed based on physiographic and oceanographic features in the St. Lawrence. Can. Tech. Rep. Fish. Aquat. Sci. 2916: Vii + 72 p. Technical Report 2916.
- Dutil, J.-D., Proulx, S., Galbraith, P. S., Chassé, J., Lambert, N., and Laurian, C. (2012). Coastal and epipelagic habitats of the estuary and Gulf of St. Lawrence. Can. Tech. Rep. Fish. Aquat. Sci. 3009: Ix + 87 pp. Technical Report 3009.
- Earth observation group (2019). Version 1 VIIRS Day/Night Band Nighttime Lights. Technical report, NOAA National Centers for Environmental Information (NCEI).
- Eby, L. A., Crowder, L. B., McClellan, C. M., Peterson, C. H., and Powers, M. J. (2005). Habitat degradation from intermittent hypoxia: Impacts on demersal fishes. Marine Ecology Progress Series, 291:249–262.
- ECCC (2018). Environment and Climate Change Canada's (ECCC) Atlantic Shoreline Classification Available from <https://open.canada.ca/data/en/dataset/30449352-2556-42df-9ffe-47ea8e696f91> Accessed 2019-09-19.

- El-Sabh, M. I. and Silverberg, N. (1990). Oceanography of a Large-Scale Estuarine System. Springer New York.
- Eppler, M. J. and Mengis, J. (2004). The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. The Information Society, 20(5):325–344.
- FA (2016). Aquaculture Site Mapping Tool. Department of Fisheries and Aquaculture (FA), Department of Agriculture, Government of Nova-Scotia. Nova-Scotia, Canada. Available at: <https://novascotia.ca/fish/aquaculture/site-mapping-tool/>. Technical report, Department of Fisheries and Aquaculture.
- Fabry, V. J., Seibel, B. A., Feely, R. A., and Orr, J. C. (2008). Impacts of ocean acidification on marine fauna and ecosystem processes. ICES Journal of Marine Science, 65(3):414–432.
- Feist, B. E. and Levin, P. S. (2016). Novel Indicators of Anthropogenic Influence on Marine and Coastal Ecosystems. Frontiers in Marine Science, 3.
- FFA (2016). Description of aquaculture sites in Newfoundland. Department of Fisheries, Forestry and Agrifoods (FFA), Government of Newfoundland and Labrador. Newfoundland, Canada. Technical report, Department of Fisheries, Forestry and Agrifoods.
- FORCE11 (2014). Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA. Technical report.
- Frank, K. T., Petrie, B., Choi, J. S., and Leggett, W. C. (2005). Trophic Cascades in a Formerly Cod-Dominated Ecosystem. Science, 308(5728):1621–1623.
- Franzoni, C. and Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. Research Policy, 43(1):1–20.
- Galbraith, P. S., Chassé, J., Caverhill, C., Nicot, P., Gilbert, D., Lefavre, D., and Lafleur, C. (2018). Physical Oceanographic Conditions in the Gulf of St. Lawrence during 2017. DFO Can. Sci. Advis. Sec. Res. Doc. 2018/050. v + 79 p. Technical Report 2018/050, Department of Fisheries and Oceans.
- Gilbert, D., Chabot, D., Archambault, P., Rondeau, B., and Hébert, S. (2007). Apauvrissement en oxygène dans les eaux profondes du Saint-Laurent marin: Causes possibles et impacts écologiques. Naturaliste Canadien, 131:67–75.
- Government of Canada (2018). Marine Spatial Data Infrastructure portal Available from <http://msdi-idsm.maps.arcgis.com/home/index.html> Accessed 2019-09-19.
- Government of Québec (2015). Stratégie maritime, The maritime strategy by the year 2030. 2015-2020 action Plan. Technical report.

- Halpern, B. S., Frazier, M., Potapenko, J., Casey, K. S., Koenig, K., Longo, C., Lowndes, J. S., Rockwood, R. C., Selig, E. R., Selkoe, K. A., and Walbridge, S. (2015a). Cumulative human impacts: Raw stressor data (2008 and 2013). Knowledge Network for Biocomplexity. <https://knb.ecoinformatics.org/#view/doi:10.5063/F1S180FS>.
- Halpern, B. S., Frazier, M., Potapenko, J., Casey, K. S., Koenig, K., Longo, C., Lowndes, J. S., Rockwood, R. C., Selig, E. R., Selkoe, K. A., and Walbridge, S. (2015b). Spatial and temporal changes in cumulative human impacts on the world's ocean. *Nature Communications*, 6(1).
- Halpern, B. S., Longo, C., Hardy, D., McLeod, K. L., Samhouri, J. F., Katona, S. K., Kleisner, K., Lester, S. E., O'Leary, J., Ranelletti, M., Rosenberg, A. A., Scarborough, C., Selig, E. R., Best, B. D., Brumbaugh, D. R., Chapin, F. S., Crowder, L. B., Daly, K. L., Doney, S. C., Elfes, C., Fogarty, M. J., Gaines, S. D., Jacobsen, K. I., Karrer, L. B., Leslie, H. M., Neeley, E., Pauly, D., Polasky, S., Ris, B., Martin, K. S., Stone, G. S., Sumaila, U. R., and Zeller, D. (2012). An index to assess the health and benefits of the global ocean. *Nature*, 488(7413):615–620.
- 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. P., Perry, M. T., Selig, E. R., Spalding, M., Steneck, R., and Watson, R. (2008). A Global Map of Human Impact on Marine Ecosystems. *Science*, 319(5865):948–952.
- Jones, F. C. (2016). Cumulative effects assessment: Theoretical underpinnings and big problems. *Environmental Reviews*, 24(2):187–204.
- Kaufman, L. and Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. page 342. Wiley, New York.
- Keith, D. A., Martin, T. G., McDonald-Madden, E., and Walters, C. (2011). Uncertainty and adaptive management for biodiversity conservation. *Biological Conservation*, 144(4):1175–1178.
- Kleypas, J. A., Feely, R. A., Fabry, V. J., Langdon, C., Sabine, C. L., and Robbins, L. L. (2006). Impacts of Ocean Acidification on Coral Reefs and Other Marine Calcifiers: A Guide for Future Research, report of a workshop held 18–20 April 2005, St. Petersburg, FL, sponsored by NSF, NOAA, and the US Geological Survey, 88 pp. Technical report.
- Kroeker, K. J., Kordas, R. L., Crim, R., Hendriks, I. E., Ramajo, L., Singh, G. S., Duarte, C. M., and Gattuso, J.-P. (2013). Impacts of ocean acidification on marine organisms: Quantifying sensitivities and interaction with warming. *Global Change Biology*, 19(6):1884–1896.

- Legendre, P. and Legendre, L. F. J. (2012). *Numerical Ecology*. Elsevier.
- Lewis, E., Wallace, D., and Allison, L. J. (1998). Program developed for CO<sub>2</sub> system calculations. Technical Report ORNL/CDIAC-105, Brookhaven National Lab., Dept. of Applied Science, Upton, NY (United States); Oak Ridge National Lab., Carbon Dioxide Information Analysis Center, TN (United States).
- Lubchenco, J. and Grorud-Colvert, K. (2015). Making waves: The science and politics of ocean protection. *Science*, 350(6259):382–383.
- Mach, M. E., Wedding, L. M., Reiter, S. M., Micheli, F., Fujita, R. M., and Martone, R. G. (2017). Assessment and management of cumulative impacts in California's network of marine protected areas. *Ocean & Coastal Management*, 137:1–11.
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2018). *Cluster: Cluster Analysis Basics and Extensions*.
- MAPAQ (2016). Description des sites maricoles de la province de Québec. Sous-ministéariat aux pêches et à l'aquaculture commerciales, Direction régionale des Îles-de-la-Madeleine. Ministère de l'Agriculture, des Pêcheries et de l'Alimentation du Québec (MAPAQ). Québec, Canada. Technical report, Ministère de l'Agriculture, des Pêcheries et de l'Alimentation du Québec.
- Margules, C. R. and Pressey, R. L. (2000). Systematic conservation planning. <https://www.nature.com/articles/35012251>.
- Micheli, F., Heiman, K. W., Kappel, C. V., Martone, R. G., Sethi, S. A., Osio, G. C., Fraschetti, S., Shelton, A. O., and Tanner, J. M. (2016). Combined impacts of natural and human disturbances on rocky shore communities. *Ocean & Coastal Management*, 126:42–50.
- Millero, F. J. (1986). The pH of estuarine waters. *Limnology and Oceanography*, 31(4):839–847.
- Moritz, C., Gravel, D., Savard, L., McKindsey, C. W., Brêthes, J.-C., and Archambault, P. (2015). No more detectable fishing effect on Northern Gulf of St Lawrence benthic invertebrates. *ICES Journal of Marine Science*, 72(8):2457–2466.
- Mucci, A., Levasseur, M., Gratton, Y., Martias, C., Scarratt, M., Gilbert, D., Tremblay, J.-É., Ferreyra, G., and Lansard, B. (2017). Tidally induced variations of pH at the head of the Laurentian Channel. *Canadian Journal of Fisheries and Aquatic Sciences*, 75(7):1128–1141.
- Mucci, A., Starr, M., Gilbert, D., and Sundby, B. (2011). Acidification of Lower St. Lawrence Estuary Bottom Waters. *Atmosphere-Ocean*, 49(3):206–218.

- Nicholls, R. J. and Cazenave, A. (2010). Sea-Level Rise and Its Impact on Coastal Zones. *Science*, 328(5985):1517–1520.
- OBIS (2019). Ocean Biogeographic Information System. Intergovernmental Oceanographic Commission of UNESCO. [www.iobis.org](http://www.iobis.org). Accessed 2019-04-19.
- O'Brien, A. L., Dafforn, K. A., Chariton, A. A., Johnston, E. L., and Mayer-Pinto, M. (2019). After decades of stressor research in urban estuarine ecosystems the focus is still on single stressors: A systematic literature review and meta-analysis. *Science of The Total Environment*.
- Pillet, M., Dupont-Prinet, A., Chabot, D., Tremblay, R., and Audet, C. (2016). Effects of exposure to hypoxia on metabolic pathways in northern shrimp (*Pandalus borealis*) and Greenland halibut (*Reinhardtius hippoglossoides*). *Journal of Experimental Marine Biology and Ecology*, 483:88–96.
- Plourde, S., Galbraith, P. S., Lesage, V., Grégoire, F., Bourdages, H., Gosselin, J.-F., McQuinn, I., and Scarratt, M. (2014). Ecosystem perspective on changes and anomalies in the Gulf of St. Lawrence: A context in support of the management of the St. Lawrence beluga whale population. DFO Can. Sci. Advis. Sec. Res. Doc. 2013/129. v + 29 p. Technical Report 2013/129.
- Reichman, O. J., Jones, M. B., and Schildhauer, M. P. (2011). Challenges and Opportunities of Open Data in Ecology. *Science*, 331(6018):703–705.
- Rice, J. (2011). Managing fisheries well: Delivering the promises of an ecosystem approach. *Fish and Fisheries*, 12(2):209–231.
- RQM (2018). Réseau Québec Maritime (RQM). Available from <http://rqi.quebec/en/home/>. Accessed 2019-09-19.
- Saucier, F. J., Roy, F., Gilbert, D., Pellerin, P., and Ritchie, H. (2003). Modeling the formation and circulation processes of water masses and sea ice in the Gulf of St. Lawrence, Canada. *Journal of Geophysical Research: Oceans*, 108(C8).
- Savenkoff, C., Vézina, A. F., Roy, S., Klein, B., Lovejoy, C., Therriault, J. C., Legendre, L., Rivkin, R., Bérubé, C., Tremblay, J. E., and Silverberg, N. (2000). Export of biogenic carbon and structure and dynamics of the pelagic food web in the Gulf of St. Lawrence Part 1. Seasonal variations. *Deep Sea Research Part II: Topical Studies in Oceanography*, 47(3–4):585–607.
- Schloss, I. R., Archambault, P., Beauchesne, D., Cusson, M., Ferreyra, G., Levasseur, M., Pelletier, É., St-Louis, R., and Tremblay, R. (2017). Cumulative potential impacts of the stress factors associated with human activities on the St. Lawrence marine ecosystem. In Archambault, P., Schloss, I. R., Grant, C., and Plante, S., editors,

Hydrocarbon in the Gulf of St. Lawrence - Social, Economic and Environmental Issues, pages 133–165. Notre Golfe, Rimouski, Qc, Canada.

Starr, M. and Chassé, J. (2019). Distribution of omega aragonite in the Estuary and Gulf of St. Lawrence in eastern Canada. Department of Fisheries and Oceans. Technical report, Department of Fisheries and Oceans.

Statistics-Canada (2017). Population and Dwelling Count Highlight Tables. 2016 Census. Statistics Canada Catalogue no. 98-402-X2016001. Ottawa. Released February 8, 2017. Technical report, Statistics Canada.

Stock, A., Haupt, A. J., Mach, M. E., and Micheli, F. (2018). Mapping ecological indicators of human impact with statistical and machine learning methods: Tests on the California coast. Ecological Informatics, 48:37–47.

Sutherland, W. J., Pullin, A. S., Dolman, P. M., and Knight, T. M. (2004). The need for evidence-based conservation. Trends in Ecology & Evolution, 19(6):305–308.

Task Group on Data Citation Standards and PractOut of Cite, Out of Mind: The Current Sices, C.-I. and PractOut of Mind: The Current Sices, C.-I. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal, 12(0):CIDCR1–CIDCR7.

Tukey, J. W. (1977). Exploratory Data Analysis. Reading, Massachusetts: Addison-Wesley.

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., and Clerck, O. D. (2012). Bio-ORACLE: A global environmental dataset for marine species distribution modelling. Global Ecology and Biogeography, 21(2):272–281.

Uthicke, S., Ebert, T., Liddy, M., Johansson, C., Fabricius, K. E., and Lamare, M. (2016). Echinometra sea urchins acclimatized to elevated pCO<sub>2</sub> at volcanic vents outperform those under present-day pCO<sub>2</sub> conditions. Global Change Biology, 22(7):2451–2461.

Walbridge, S. (2013). Assessing Ship Movements Using Volunteered Geographic Information. University of California, Santa Barbara.

White, L. and Johns, F. (1997). Marine Environmental Assessment of the Estuary and Gulf of St. Lawrence. Fisheries and Oceans Canada Toxic Chemicals Program 1997. Department of Fisheries and Oceans.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S.,

Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018.

Zeebe, R. E. and Wolf-Gladrow, D. (2001). CO<sub>2</sub> in Seawater: Equilibrium, Kinetics, Isotopes. Number 65. Gulf Professional Publishing.