Distribution Modelling of Sea Pens, Sponges, Stalked Tunicates and Soft Corals from Research Vessel Survey Data in the Gulf of St. Lawrence for Use in the Identification of Significant Benthic Areas

F.J. Murillo, E. Kenchington, L. Beazley, C. Lirette, A. Knudby, J. Guijarro, H. Benoît, H. Bourdages, B. Sainte-Marie

Ocean and Ecosystem Sciences Division
Maritimes Region
Fisheries and Oceans Canada

Bedford Institute of Oceanography
PO Box 1006
Dartmouth, Nova Scotia
Canada B2Y 4A2

2016

Canadian Technical Report of Fisheries and Aquatic Sciences 3170

Fisheries and Oceans Canada Pêches et Océans Canada

Canada

Canadian Technical Report of Fisheries and Aquatic Sciences

Technical reports contain scientific and technical information that contributes to existing knowledge but which is not normally appropriate for primary literature. Technical reports are directed primarily toward a worldwide audience and have an international distribution. No restriction is placed on subject matter and the series reflects the broad interests and policies of Fisheries and Oceans Canada, namely, fisheries and aquatic sciences.

Technical reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report is abstracted in the data base *Aquatic Sciences and Fisheries Abstracts*.

Technical reports are produced regionally but are numbered nationally. Requests for individual reports will be filled by the issuing establishment listed on the front cover and title page.

Numbers 1-456 in this series were issued as Technical Reports of the Fisheries Research Board of Canada. Numbers 457-714 were issued as Department of the Environment, Fisheries and Marine Service, Research and Development Directorate Technical Reports. Numbers 715-924 were issued as Department of Fisheries and Environment, Fisheries and Marine Service Technical Reports. The current series name was changed with report number 925.

Rapport technique canadien des sciences halieutiques et aquatiques

Les rapports techniques contiennent des renseignements scientifiques et techniques qui constituent une contribution aux connaissances actuelles, mais qui ne sont pas normalement appropriés pour la publication dans un journal scientifique. Les rapports techniques sont destinés essentiellement à un public international et ils sont distribués à cet échelon. II n'y a aucune restriction quant au sujet; de fait, la série reflète la vaste gamme des intérêts et des politiques de Pêches et Océans Canada, c'est-à-dire les sciences halieutiques et aquatiques.

Les rapports techniques peuvent être cités comme des publications à part entière. Le titre exact figure au-dessus du résumé de chaque rapport. Les rapports techniques sont résumés dans la base de données *Résumés des sciences aquatiques et halieutiques.*

Les rapports techniques sont produits à l'échelon régional, mais numérotés à l'échelon national. Les demandes de rapports seront satisfaites par l'établissement auteur dont le nom figure sur la couverture et la page du titre.

Les numéros 1 à 456 de cette série ont été publiés à titre de Rapports techniques de l'Office des recherches sur les pêcheries du Canada. Les numéros 457 à 714 sont parus à titre de Rapports techniques de la Direction générale de la recherche et du développement, Service des pêches et de la mer, ministère de l'Environnement. Les numéros 715 à 924 ont été publiés à titre de Rapports techniques du Service des pêches et de la mer, ministère des Pêches et de l'Environnement. Le nom actuel de la série a été établi lors de la parution du numéro 925.

Canadian Technical Report of
Fisheries and Aquatic Sciences 3170

2016

Distribution Modelling of Sea Pens, Sponges, Stalked Tunicates and Soft Corals from Research Vessel Survey Data in the Gulf of St. Lawrence for Use in the Identification of Significant Benthic Areas

by

F.J. Murillo[1], E. Kenchington[1], L. Beazley[1], C. Lirette[1], A. Knudby[2], J. Guijarro[1], H. Benoît[3], H. Bourdages[4], B. Sainte-Marie[4]

[1]Ocean and Ecosystem Sciences Division
Maritimes Region
Fisheries and Oceans Canada
Bedford Institute of Oceanography
P.O. Box 1006
Dartmouth, N.S.
Canada B2Y 4A2

[2]Department of Geography, Environment and Geomatics
University of Ottawa
Ottawa, Ontario
Canada K1N 6N5

[3]Fisheries and Oceans Canada
Gulf Fisheries Centre
PO Box 5030
Moncton, New Brunswick
Canada E1C 9B6

[4]Benthic Sciences,
Demersal and Benthic Sciences Branch
Maurice Lamontagne Institute
850, route de la Mer, C. P. 1000
Mont-Joli, Québec
Canada G5H 3Z4

© Her Majesty the Queen in Right of Canada, 2016.
Cat. No. Fs97-6/3170E-PDF ISBN 978-0-660-05664-7 ISSN 1488-5379

Correct citation for this publication:

Murillo, F.J., Kenchington, E., Beazley, L., Lirette, C., Knudby, A., Guijarro, J., Benoît, H., Bourdages, H., Sainte-Marie, B. 2016. Distribution Modelling of Sea Pens, Sponges, Stalked Tunicates and Soft Corals from Research Vessel Survey Data in the Gulf of St. Lawrence for Use in the Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3170: vi + 132 p.

TABLE OF CONTENTS

ABSTRACT ..... *v*
RÉSUMÉ ..... *vi*
INTRODUCTION *1*
MATERIALS AND METHODS *2*
Study Area 2
Response Data 2
Environmental Data 4
Temperature relationship between model data (GLORYS) and *in situ* data (CTD) 4
Distribution Modelling 10
Random Forest 10
Generalized Additive Model 11
Model evaluation 11
Ecological interpretation 13
RESULTS *14*
Relationship between Catch Weight and Trawl Duration 14
Temperature relationship between model data (GLORYS) and *in situ* data (CTD) 14
Distribution Modelling 18
Sea Pens (Pennatulacea) 18
Data Sources and Distribution 18
Model 1 – Balanced species prevalence 19
Model 2 – Unbalanced species prevalence 25
Model Selection 30
Prediction of Sea Pen Biomass Using Random Forest (RF) 30
Prediction of Sea Pen Biomass Using Generalized Additive Models (GAM) 39
Comparison between GAM and RF 47
Sponges (Porifera) 47
Data Sources and Distribution 47
Model 1 – Unbalanced species prevalence 48
Prediction of Sponge Biomass Using Random Forest 53
Stalked tunicates (*Boltenia ovifera*) 62
Data Sources and Distribution 62
Model 1 – Balanced species prevalence 64
Model 2 – Unbalanced species prevalence 70
Correlation between Biomass and Presence Probability 75
Model Selection 75
Prediction of Stalked Tunicates Biomass Using Random Forest 76
Soft corals (*Gersemia rubiformis*) 85
Data Sources and Distribution 85
Model 1 – Balanced species prevalence 86
Model 2 – Unbalanced species prevalence 92
Model Selection 97
Prediction of *G. rubiformis* Biomass Using Random Forest 97
DISCUSSION *106*

APPENDIX 1 ..... *114*
Prediction of Sponges, Stalked Tunicates and Soft Corals Biomass Using GAM Models 114
Sponges (Porifera) 114
Stalked tunicates (*Boltenia ovifera*) 120
Soft corals (*Gersemia rubiformis*) 126
APPENDIX 2 ..... *132*
Distribution of Solitary Cup Corals (*Flabellum alabastrum*) 132

ABSTRACT

Murillo, F.J., Kenchington, E., Beazley, L., Lirette, C., Knudby, A., Guijarro, J., Benoît, H., Bourdages, H., Sainte-Marie, B. 2016. Distribution Modelling of Sea Pens, Sponges, Stalked Tunicates and Soft Corals from Research Vessel Survey Data in the Gulf of St. Lawrence for Use in the Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3170: vi + 132 p.

Models of probability of occurrence and predicted biomass distribution have been created using random forest (RF) machine learning techniques for different invertebrate taxa in the Gulf of St. Lawrence. Response data were derived from by-catch data collected from DFO research vessel trawl surveys following a stratified random design based on depth and geographic region. Predictors were drawn from 78 environmental data layers. Occurrence models performed very well for sea pens and stalked tunicates and better than those for soft corals and sponges, with cross-validated AUC (area under the receiver operating characteristic curve) values ranging from 0.71 to 0.91. For the models based on biomass, soft corals and sea pens had the highest R^2 values (0.42 and 0.37, respectively) in the southern Gulf of St. Lawrence and stalked tunicates and sea pens in the north (0.41 and 0.27, respectively). Sponges had R^2 values less than 0.1 in both areas indicating poor model performance. Biomass models from RF were compared with Generalized Additive Models (GAM). In most of the cases RF and GAM models provided similar results and were both good options, although the fewer assumptions required for RF makes this method more convenient. These results could be used to identify the potential distribution of some vulnerable marine ecosystems indicator taxa and help to refine the borders of the significant benthic area polygons defining significant concentrations of these taxa as identified through the kernel density analyses. In particular these models can be used to extrapolate to areas not covered by the research vessel surveys.

RÉSUMÉ

Murillo, F.J., Kenchington, E., Beazley, L., Lirette, C., Knudby, A., Guijarro, J., Benoît, H., Bourdages, H., Sainte-Marie, B. 2016. Modélisation de la répartition des pennatules, des éponges, des tuniciers lobés et des coraux mous selon les données des relevés par navire scientifique dans le golfe du Saint-Laurent aux fins d'utilisation dans la désignation des zones benthiques importantes. Can. Tech. Rep. Fish. Aquat. Sci. 3170: vi + 132 p.

Des modèles de la probabilité de présence et de la répartition de la biomasse prévue ont été créés au moyen de techniques d'apprentissage machine avec forêts aléatoires pour différents taxons d'invertébrés dans le golfe du Saint-Laurent. Les données de réponse ont été dérivées des données sur les prises accessoires recueillies à partir des relevés au chalut par navire scientifique de Pêches et Océans Canada, conformément à un plan aléatoire stratifié fondé sur la profondeur et la région géographique. Des facteurs de prévision ont été élaborés en fonction de 78 couches de données environnementales. Les modèles de présence donnaient un très bon rendement pour les pennatules et les tuniciers lobés et étaient meilleurs que ceux pour les coraux mous et les éponges, selon des valeurs contre-validées de l'aire sous la courbe de la fonction d'efficacité du récepteur variant de 0,71 à 0,91. Pour les modèles fondés sur la biomasse, les coraux mous et les pennatules avaient les plus grandes valeurs de R^2 (0,42 et 0,37, respectivement) dans le sud du golfe du Saint-Laurent, tandis que les tuniciers lobés et les pennatules avaient les plus grandes valeurs dans le nord (0,41 et 0,27, respectivement). Les éponges avaient des valeurs de R^2 inférieures à 0,1 dans les deux zones, ce qui indique un faible rendement du modèle. Les modèles de la biomasse avec forêts aléatoires ont été comparés à des modèles additifs généralisés. Dans la plupart des cas, les modèles avec forêts aléatoires et les modèles additifs généralisés donnaient des résultats semblables et étaient tous les deux de bonnes options; toutefois, le modèle avec forêts aléatoires est plus pratique comme il nécessite moins d'hypothèses. Ces résultats pourraient servir à déterminer l'aire de répartition possible de certains taxons indicateurs des écosystèmes marins vulnérables et aider à peaufiner les limites des polygones des zones benthiques importantes pour définir les concentrations importantes de ces taxons, conformément aux analyses des noyaux de densité. En particulier, ces modèles peuvent servir à extrapoler à des zones qui ne sont pas couvertes par les relevés par navire scientifique.

INTRODUCTION

Ecologists and conservation managers are increasingly using predictive models as a means of estimating species distributions, and they have become a significant component of conservation planning (e.g., Davis and Guinotte, 2011; Porfirio et al., 2014; Rooper et al., 2014). Species distribution models (SDM) estimate the relationship between species records at different locations and the environmental and/or spatial characteristics of those sites (Franklin, 2009). SDMs are often used to extend typically sparse point observations to create continuous predictions of species' distribution or habitat type. SDMs have been extensively used in both terrestrial and marine environments to make contemporary distribution maps (e.g., Dolan et al., 2008; Davies et al., 2008; Guinan et al., 2009; Tittensor et al., 2009; Davies and Guinotte, 2011; Yesson et al., 2012; Knudby et al. 2013 a,b,c; Rooper et al., 2014), to predict species/habitat responses to climate change (e.g., Lawler et al., 2009) and to predict the future range of invasive species (Peterson and Robins, 2003; Peterson 2003). They can also provide knowledge of the differences between actual and potential species distribution, making them a very useful tool for habitat restoration in fisheries and ocean management.

Many different SDM models and modelling approaches are currently available (Guisan and Zimmerman, 2000). In recent years, machine learning methods have become increasingly common. These comprise a series of non-parametric techniques capable of synthesizing regression or classification functions based on the available data. Random Forest (RF) (Breiman, 2001) is one such method that can be used for regression or classifications and is considered superior to most commonly-used methods (Cutler et al., 2007). Others have used regression based techniques such as generalized additive modelling (e.g., Rooper et al., 2014) which offer a different theoretical basis for the production of the interpolated surfaces.

SDM using RF was recently applied in the Northwest Atlantic Fisheries Organization (NAFO) Regulatory Area (Knudby et al., 2013 a,b,c; NAFO, 2015) to refine the Vulnerable Marine Ecosystems polygons based on corals and sponges (NAFO, 2015) previously identified through Kernel Density Analysis (KDE) (Kenchington et al., 2014). In 2010, Kenchington et al. (2010) identified significant concentrations of corals and sponge on the east coast of Canada including the Gulf of St. Lawrence region, using kernel density estimation (KDE). This analysis has been updated with the most recent survey data (Kenchington et al., 2016). The Estuary and Gulf of St. Lawrence are amongst the largest and most productive estuarine/marine ecosystems in the world (Dufour and Ouellet, 2007). Here we present SDM-based maps of sea pens, sponges, stalked tunicates and cauliflower coral in the Gulf of St. Lawrence using RF. Additionally, generalized additive models (GAM) were used to model sea pen biomass, the results of which were compared to those from RF. These models will complement the kernel density surface polygons provided in Kenchington et al. (2016) towards Department of Fisheries and Oceans' (DFO) implementation of the Policy for Managing the Impact of Fishing on Sensitive Benthic Areas for the east coast of Canada.

MATERIALS AND METHODS

Study Area

The 'Gulf Region' used herein (Figure 1), is a DFO Marine Protected Area (MPA) Network Planning Area for the Estuary and Gulf of St. Lawrence and corresponds to the area used as the spatial boundary for the construction of environmental variables in Beazley et al. (2016). This area combines three of DFO's six administrative regions across Canada, the Gulf Region in the southern portion, the Quebec Region in the northern portion, and the Newfoundland Region to the east. A 5-km buffer was placed around all land to avoid its inclusion in the models. The total area covered in the study extent is approximately 196,494 km^2 based on a NAD 1983 UTM Zone 20N projection.

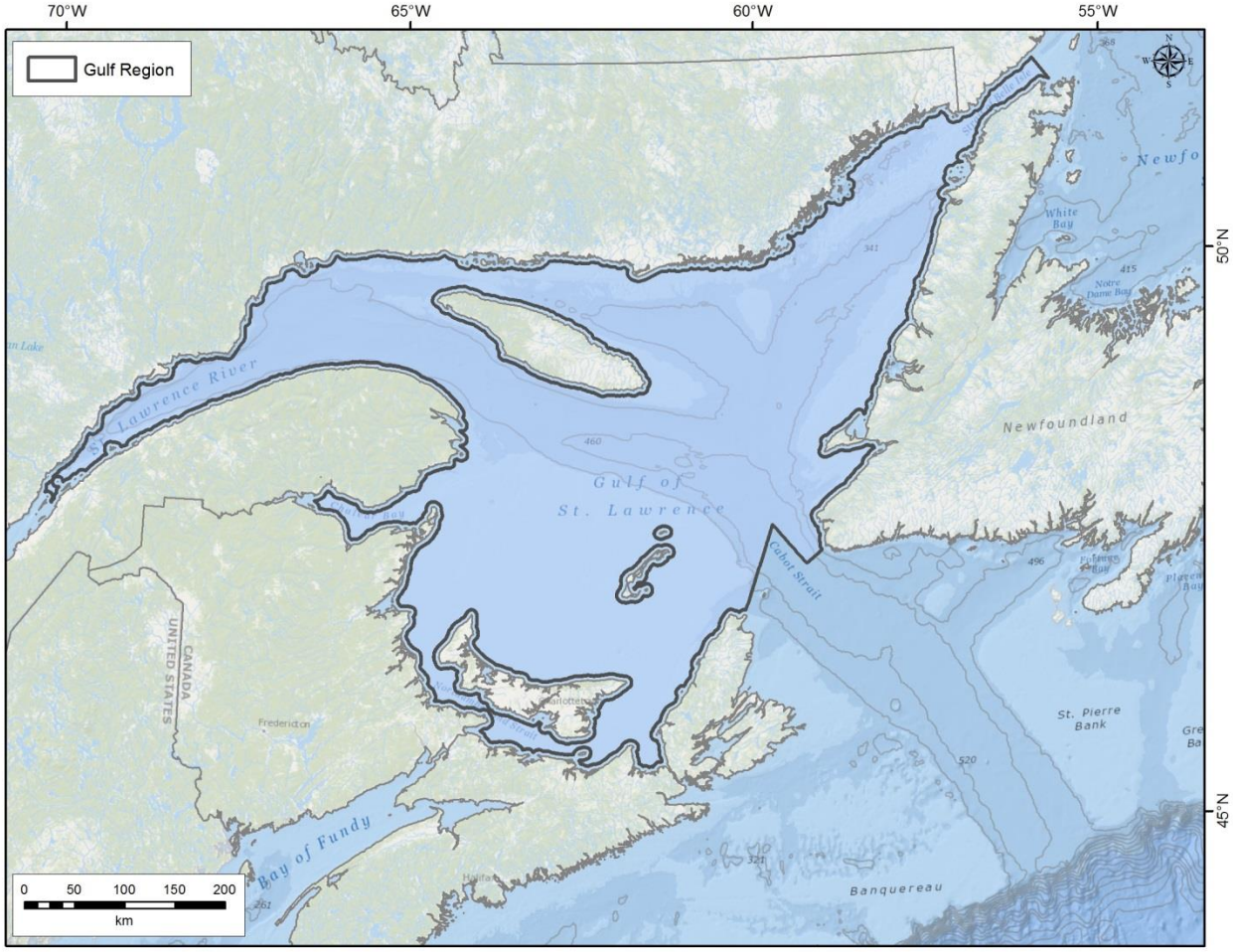

Figure 1. Extent of the boundary used for species distribution modelling in the Gulf of St. Lawrence.

Response Data

Presence-absence and biomass records for four taxonomic groups (see Table 1 for species/taxa included in each group) were derived from by-catch data collected from DFO research vessel trawl surveys following a stratified random design based on depth and geographic region (Kulka et al., 2006). We included analyses on the soft corals (Nephtheidae), most likely the cauliflower coral *Gersemia rubiformis* [syn. *Eunephthya rubiformis*], and on the stalked tunicates, mostly

Boltenia ovifera, as these were identified by oceans managers as being of interest. The latter taxon, the stalked tunicates, is recognized by NAFO as a vulnerable marine ecosystem indicator (NAFO, 2013).

Trawl surveys in the Gulf Region were conducted on the *CCGS Alfred Needler* or *Wildfred Templeman* using a Western IIA trawl gear with tow duration of 30 minutes in the southern Gulf of St. Lawrence. In Quebec Region surveys were conducted on the *CCGS Teleost* using a Campelen 1800 trawl gear with tow duration of 15 minutes in the northern Gulf of St. Lawrence.

Table 1. Taxon name and species code included in each of the taxonomic group studied for the northern and southern Gulf of St. Lawrence.

Taxonomic group	Region	Taxon name	Species code
Porifera (sponges)	Northern Gulf	Porifera	1101
		Stylocordila borealis	1112
	Southern Gulf	*Asconema foliata*	8365
		Biemna variantia	8617
		Geodia spp.	8364
		Halichondria panicea	8623
		Halichondria sitiens	8620
		Haliclona oculata	8621
		Haliclona sp.	8618
		Iophon sp.	8614
		Mycale lingua	8616
		Phakellia spp.	8366
		Phakellia ventilabrum	8624
		Polymastia mammillaris	8611
		Polymastia sp.	8610
		Porifera	8600
		Suberites ficus	8613
		Tentorium semisuberites	8612
Pennatulacea (sea pens)	Northern Gulf	*Anthoptilum grandiflorum*	2218
		Halipteris finmarchica	2217
		Pennatula aculeata	2203
		Pennatula grandis	2210
		Pennatulacea	2201
	Southern Gulf	*Anthoptilum grandiflorum*	8631
		Pennatulacea	8318
Gersemia (cauliflower coral)	Northern Gulf	*Gersemia rubiformis*	2184
	Southern Gulf	*Eunephthya rubiformis* (unaccepted) (syn.*Gersemia rubiformis)*	8324
Boltenia ovifera (stalked tunicate)	Northern Gulf	*Boltenia ovifera*	8792
	Southern Gulf	Sea potato	1823

Response data for each taxonomic group were filtered so that only one presence or absence record occurred within a single environmental data raster cell, with presence records taking precedence over an absence record when both co-occurred in the same cell. The temporal range and number of presence and absence records are detailed individually for each taxonomic group. The associated biomass values were averaged across all tows within a single cell. There is overlap between the surveys along the southern margin of the Laurentian Channel which in theory would allow for the calculation of a conversion factor to one or other gear type and permit a single analysis of the data. However, the area of overlap is small and specific to one habitat type (slope) and consequently any attempt to apply a conversion factor derived from those data would introduce other biases. Therefore, two different models were performed with the biomass data given the different tow lengths and gears; although the entire region was modeled with the presence-absence data where catchability was deemed to be less of an issue.

In fisheries ecology and management, fish catch is normally standardized to CPUE (catch per unit effort) assuming a linear relationship between fish abundance and time of trawling (DFO, 2007). Previously, we have found no relationship between trawl length and the biomass of these highly aggregated taxa (Kenchington et al., 2014). However, to verify that conclusion for the present data, by-catch weight of each taxon from the surveys carried out in the northern and southern Gulf of St. Lawrence was plotted in relation to trawl length in order to evaluate if biomass should be standardized to CPUE or analysed as kg/tow instead.

Environmental Data

Dutil et al. (2011) compiled different environmental variables for use in habitat classification for the Gulf of St. Lawrence. They produced continuous interpolated surfaces for each variable through natural neighbours, although no quantitative assessment of the quality of the underlying data or interpolation was provided. Therefore new continuous interpolated surfaces of 113 environmental variables were produced for potential use in species distribution models. Variables were chosen based on their availability and assumed relevance to the distribution of benthic fauna. Specific details on the methods used for the spatial interpolation of these variables as well as quantitative assessment of the quality of the interpolations are documented in Beazley et al. (2016). From the 113 continuous interpolated surfaces obtained, 78 were used as predictor variables in the models (Table 2). All variables except depth and slope were derived from long-term modelled oceanographic or remote-sensing data and were spatially interpolated across the study area using ordinary kriging in ArcMap 10.2.2 software (ESRI, 2011).

Temperature relationship between model data (GLORYS) and *in situ* data (CTD)

Independent data from CTDs obtained during DFO research vessel trawl surveys were used to evaluate the interpolated mean bottom temperature obtained from GLORYS data. Surveys in the southern region were mainly carried out in September and data were available for the period 2004 to 2011. Whereas, surveys in the northern region were mainly carried out in August and data were available for the period 2006 to 2011. Temperature values from the mean bottom temperature surface obtained from the GLORYS data were extracted to the CTD locations in ArcGIS 10.2 (ESRI, 2011) and the correlation between CTD and GLORYS temperature was evaluated. In order to study the spatial correlation between the two variables, different bottom temperature interpolated surfaces from CTD data were created for each year and region

following the same methodology than Beazley et al. (2016). Data from 2006 and 2011 are presented.

Table 2. Summary of the 78 environmental variables used as predictor variables in the modelling. N/A = Not applicable.

Variable	Data source	Temporal range	Unit	Native resolution
Depth	CHS-ABC	N/A	metres	15 arc-sec (~500 m)
Slope	CHS-ABC	N/A	degrees	15 arc-sec (~500 m)
Bottom Salinity Mean	GLORYS2V1	1993 - 2011	N/A	¼ º
Bottom Salinity Average Minimum	GLORYS2V1	1993 - 2011	N/A	¼ º
Bottom Salinity Average Maximum	GLORYS2V1	1993 - 2011	N/A	¼ º
Bottom Salinity Average Range	GLORYS2V1	1993 - 2011	N/A	¼ º
Bottom Temperature Mean	GLORYS2V1	1993 - 2011	ºC	¼ º
Bottom Temperature Average Minimum	GLORYS2V1	1993 - 2011	ºC	¼ º
Bottom Temperature Average Maximum	GLORYS2V1	1993 - 2011	ºC	¼ º
Bottom Temperature Average Range	GLORYS2V1	1993 - 2011	ºC	¼ º
Bottom Current Speed Mean	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ º
Bottom Current Speed Average Minimum	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ º
Bottom Current Speed Average Maximum	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ º
Bottom Current Speed Average Range	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ º
Bottom Shear Mean	GLORYS2V1	1993 - 2011	Pa	¼ º
Bottom Shear Average Minimum	GLORYS2V1	1993 - 2011	Pa	¼ º
Bottom Shear Average Maximum	GLORYS2V1	1993 - 2011	Pa	¼ º
Bottom Shear Average Range	GLORYS2V1	1993 - 2011	Pa	¼ º
Surface Salinity Mean	GLORYS2V1	1993 - 2011	N/A	¼ º
Surface Salinity Average Minimum	GLORYS2V1	1993 - 2011	N/A	¼ º
Surface Salinity Average Maximum	GLORYS2V1	1993 - 2011	N/A	¼ º
Surface Salinity Average Range	GLORYS2V1	1993 - 2011	N/A	¼ º

Surface Temperature Mean	GLORYS2V1	1993 - 2011	ºC	¼ º
Surface Temperature Average Minimum	GLORYS2V1	1993 - 2011	ºC	¼ º
Surface Temperature Average Maximum	GLORYS2V1	1993 - 2011	ºC	¼ º
Surface Temperature Average Range	GLORYS2V1	1993 - 2011	ºC	¼ º
Surface Current Speed Mean	GLORYS2V1	1993 - 2011	m s^{-1}	¼ º
Surface Current Speed Average Minimum	GLORYS2V1	1993 - 2011	m s^{-1}	¼ º
Surface Current Speed Average Maximum	GLORYS2V1	1993 - 2011	m s^{-1}	¼ º
Surface Current Speed Average Range	GLORYS2V1	1993 - 2011	m s^{-1}	¼ º
Maximum Average Mixed Layer Depth Fall	GLORYS2V1	1993 - 2011	metres	¼ º
Maximum Average Mixed Layer Depth Winter	GLORYS2V1	1993 - 2011	metres	¼ º
Maximum Average Mixed Layer Depth Spring	GLORYS2V1	1993 - 2011	metres	¼ º
Maximum Average Mixed Layer Depth Summer	GLORYS2V1	1993 - 2011	metres	¼ º
Fall Chlorophyll *a* Mean	MODIS	2002 - 2012	mg m^{-3}	2 km
Fall Chlorophyll *a* Minimum	MODIS	2002 – 2012	mg m^{-3}	2 km
Fall Chlorophyll *a* Maximum	MODIS	2002 – 2012	mg m^{-3}	2 km
Fall Chlorophyll *a* Range	MODIS	2002 – 2012	mg m^{-3}	2 km
Spring Chlorophyll *a* Mean	MODIS	2002 – 2012	mg m^{-3}	2 km
Spring Chlorophyll *a* Minimum	MODIS	2002 – 2012	mg m^{-3}	2 km
Spring Chlorophyll *a* Maximum	MODIS	2002 – 2012	mg m^{-3}	2 km
Spring Chlorophyll *a* Range	MODIS	2002 – 2012	mg m^{-3}	2 km
Summer Chlorophyll *a* Mean	MODIS	2002 – 2012	mg m^{-3}	2 km
Summer Chlorophyll *a* Minimum	MODIS	2002 – 2012	mg m^{-3}	2 km
Summer Chlorophyll *a* Maximum	MODIS	2002 – 2012	mg m^{-3}	2 km
Summer Chlorophyll *a* Range	MODIS	2002 – 2012	mg m^{-3}	2 km
Annual Chlorophyll *a* Mean	MODIS	2002 – 2012	mg m^{-3}	2 km
Annual Chlorophyll *a* Minimum	MODIS	2002 – 2012	mg m^{-3}	2 km

Annual Chlorophyll *a* Maximum	MODIS	2002 – 2012	mg m^{-3}	2 km
Annual Chlorophyll *a* Range	MODIS	2002 – 2012	mg m^{-3}	2 km
Fall Chlorophyll *a* Mean	MERIS	2003 - 2011	mg m^{-3}	1 km
Fall Chlorophyll *a* Minimum	MERIS	2003 - 2011	mg m^{-3}	1 km
Fall Chlorophyll *a* Maximum	MERIS	2003 - 2011	mg m^{-3}	1 km
Fall Chlorophyll *a* Range	MERIS	2003 - 2011	mg m^{-3}	1 km
Spring Chlorophyll *a* Mean	MERIS	2003 - 2011	mg m^{-3}	1 km
Spring Chlorophyll *a* Minimum	MERIS	2003 - 2011	mg m^{-3}	1 km
Spring Chlorophyll *a* Maximum	MERIS	2003 - 2011	mg m^{-3}	1 km
Spring Chlorophyll *a* Range	MERIS	2003 - 2011	mg m^{-3}	1 km
Summer Chlorophyll *a* Mean	MERIS	2003 - 2011	mg m^{-3}	1 km
Summer Chlorophyll *a* Minimum	MERIS	2003 - 2011	mg m^{-3}	1 km
Summer Chlorophyll *a* Maximum	MERIS	2003 - 2011	mg m^{-3}	1 km
Summer Chlorophyll *a* Range	MERIS	2003 - 2011	mg m^{-3}	1 km
Annual Chlorophyll *a* Mean	MERIS	2003 - 2011	mg m^{-3}	1 km
Annual Chlorophyll *a* Minimum	MERIS	2003 - 2011	mg m^{-3}	1 km
Annual Chlorophyll *a* Maximum	MERIS	2003 - 2011	mg m^{-3}	1 km
Annual Chlorophyll *a* Range	MERIS	2003 - 2011	mg m^{-3}	1 km
Spring Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Summer Primary Production Mean	SeaWiFS Level-3 with	2006 – 2010	mg C m^{-2} day^{-1}	9 km

	other input parameters			
Summer Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Summer Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Summer Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km

Distribution Modelling

Random Forest

Random Forest (Breiman, 2001), is a non-parametric machine learning technique, where multiple regression or classification trees (usually ≥ 500) are built using random subsets of the data (Figure 2). Each tree is fit to a bootstrap sample of the biological observations (i.e. the 'in-bag' observations), and the best split at each node is selected based on a randomly-chosen subset of predictor variables. Regression trees are used for response variables consisting of continuous data and classification trees for categorical variables. RF is a robust statistical method requiring no distributional assumptions on the covariate relationship to the response in comparison to other classical statistical models such as generalized linear models (GLMs) or generalized additive models (GAMs).

For classification with presence-absence response data, RF can be used to predict the probability of a species' presence in non-sampled areas by identifying areas with similar environmental conditions. Whereas, for regression with biomass response data, RF can be used to predict the species' biomass in non-sampled areas. The models were built in the statistical computing software package R (R Development Core Team, 2015) using the 'randomForest' package (Liaw and Wiener, 2002). Default values were used for RF parameters, and 500 trees were constructed.

RF classification models on response presence/absence and RF regression models on response biomass, were built individually for sea pens, sponges, stalked tunicates and cauliflower coral.

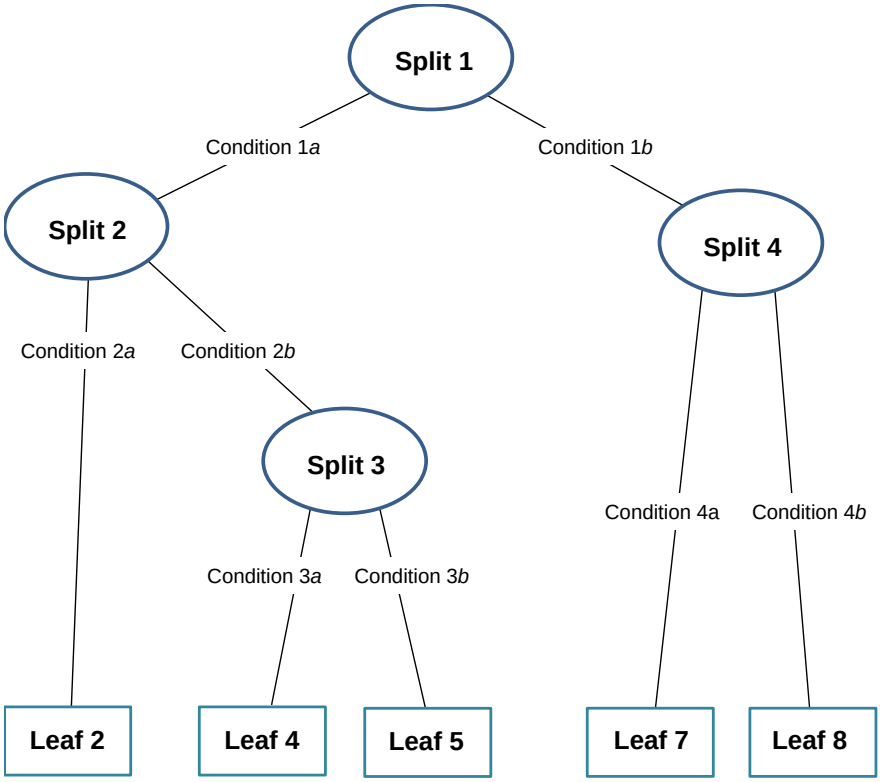

Figure 2. An example of a regression model tree (modified from Kuhn & Johnson, 2013)

Generalized Additive Model

A generalized additive model (GAM) (Hastie and Tibshirani, 1986) is a generalized linear model in which the linear predictor involves the sum of unknown smooth functions of some predictor variables. In general the model has a structure such as:

$$g(\mathrm{E}(\mathrm{Y})) = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + ...+ f_m(x_m)$$

where an exponential family distribution is specified for Y along with a link function g. The functions $f_{j(}x_{j)}$ are smooth functions that can be specified by non-parametric means. The model allows for somewhat flexible specification of the dependence of the response on the covariates. This flexibility provides potential for better fits to data than purely parametric models, but brings the cost to represent the smooth functions in some way and to choose how smooth they should be.

The mgcv package in R (Wood, 2006) was used to construct GAM models to predict the biomass in order to compare with the RF models. The top ten and top fifteen most important environmental variables obtained from the RF model based on biomass of sea pens were used as covariates in these models as well as the environmental variables correlated less than 0.7. Highly correlated variables were removed following the approach described in Knudby et al. (2013a). The autocorrelation of residuals was studied for the best of these models and in the case where it was significant latitude and longitude were included in the best model as a tensor product (i.e. te(lat, long)). Additional GAM models for the rest of the taxa with the environmental variables correlated less than 0.7 (Model 1) and including location (Model 2) are presented in Appendix I. The full model follows the formula:

y =s(var.1)+s(var.2)+…+s(var.n) +te(lat,long)

where y was specified as a Tweedie distribution and s indicate a thin plate regression spline smoothing function. A Tweedie model is an expansion of a compound Poisson model derived from the stochastic process where the weight of the counted objects has a gamma distribution. This model has the advantage of handling the zero-catch data in a unified way and the statistical performance seems to be rather better than that of a Delta lognormal model (Shono, 2008). Tweedie factor was estimated inside the model.

Model evaluation

Presence-Absence response data – Classification model (RF)

The by-catch records of some taxonomic groups are characterized by a higher number of absences relative to presences (i.e. unbalanced species prevalence). The distribution of these two classes may be biased spatially and/or environmentally across the study area. Classification accuracy in random forest is prone to bias when the categorical response variable is highly imbalanced (Chen et al., 2004). This is due to over-representation of the majority class in the bootstrap sample leading to a higher frequency in which the majority class is drawn, therefore skewing predictions in that favour (Evans et al., 2011). Several different approaches have been used to address imbalanced data: 1) assign a high cost to misclassification of the minority class, 2) down-sample the majority class, and 3) up-sample the minority class (Evans et al., 2011). Although several studies suggest a balanced modelling prevalence of 0.5 (McPherson et al.,

2004; Liu et al., 2005), this approach may result in a loss of information particularly for rare species, and may not be necessary when the model training data is reliable and not biased spatially and/or environmentally (Jiménez-Valverde and Lobo, 2006). Another widely-used approach is to adjust the threshold used to divide the probabilistic predictions of occurrence into discrete predictions of presence or absence, to match modelling prevalence (Liu et al., 2005). The latter approach has shown to produce constant error rates and optimal model accuracy measures compared to balancing modelling prevalence (Liu et al., 2005; Hanberry and He, 2013).

For each taxonomic group we assessed the number of presences and absences and their spatial distribution across the study area. We employed two different random forest modelling methods. The first method was to model the response data with a balanced species prevalence and threshold of 0.5. Here the absence records were randomly down-sampled to match the number of presences prior to modelling. In the second method we used all presence and absence records and used species prevalence as the threshold. The appropriateness of each modelling approach on the response data was assessed based on the model accuracy measures (see explanation below of model accuracy measures) and the spatial pattern of the predictions of presence probability in relation to the response data.

Accuracy measures were obtained using 10-fold cross validation (10 resamples over which performance estimates were obtained). In 10-fold cross validation the response data are randomly split into 10 equal-sized groups and the model is trained on a combination of 9, while validated on the remaining group.

Three measures of accuracy were used to assess model performance: 1) sensitivity, 2) specificity, and 3) AUC. In a classification model with two classes (e.g. presence and absence), there are four possible predicted outcomes: 1) true positive, where observed presences are predicted as presences, 2) false negative, where observed presences are predicted as absences, 3) true negative, where observed absences are predicted as absences, and 4) false positive, where observed absences are predicted as presences (Fawcett, 2006). Sensitivity measures the proportion of observed presences correctly predicted as presence (i.e. the true positive rate) (McPherson et al., 2004; Fawcett, 2006). Low sensitivity indicates high omission error (i.e. false negative rate). Specificity measures the proportion of observed absences correctly predicted as absence (i.e. the true negative rate). Low specificity indicates high commission error (i.e. the false positive rate). Both sensitivity and specificity are derived from a two-by-two confusion matrix of the tabulated predicted outcomes.

The AUC is a threshold-independent measure of model accuracy that is calculated from the combination of true positive rate (sensitivity) and false positive rate (1 – specificity), and equals the probability that the model will rank a randomly-chosen presence instance higher than a randomly-chosen absence instance (Fawcett, 2006). Its value ranges from 0 to 1, with values larger than 0.5 indicating performance better than random (Fawcett, 2006). It was calculated using 10-fold cross validation.

For models generated using a balanced species prevalence and threshold of 0.5, 10 data subsets were created with an equal number of presence and absences (balanced data) and 10 models

were run. AUC was determined by averaging AUC values between folds within each run. The model with the highest average AUC was considered the most accurate in predicting the validated data and was used as the final model in which predicted presence probabilities of the response data were generated. The predicted outcomes from the two-by-two confusion matrices were summed across all 10 folds to give a complete confusion matrix for each model from which sensitivity and specificity were calculated. For models generated using all presence and absence data and a threshold equal to species prevalence, only one model was considered and the AUC was determined by averaging AUC values between folds. The predicted outcomes from the two-by-two confusion matrices were summed across all 10 model folds to give one confusion matrix from which sensitivity and specificity were calculated.

Biomass response data – Regression model

RF

Models were validated using 10-fold cross validation. Data were split using the createFolds function in R. This function performs stratified partitioning into k groups in order to evenly distributed the biomass within splits. Models were built using each calibrated and validated dataset and accuracy measures were calculated for each corresponding dataset. The accuracy measures used to validate the models included the goodness-of-fit statistic R^2, the root mean squared error (RMSE) value and the percentage of variance explained. RMSE was normalised to a percentage of the range of observed biomass values ($y_{max} - y_{min}$) for each specific response (NRMSE) to facilitate the comparison between responses in the different models. The correlation between biomass and presence probability for stalked tunicates was also evaluated. Cross-validation gives an average of the accuracy measures used, but can also be used to estimate the variability around the mean to evaluate the stability of the model fit, and to check for the arbitrary effects from subsampling data for calibrate and validate the model.

GAM

Residual plots to evaluate the fitness of the model can be generated with the function gam.check of the mgcv package. However, an artifact of the link function shows exact zeros as a band along the residuals vs. linear predictor plot, making it difficult to see whether residuals show heteroskedasticity. In order to avoid this issue randomised quantile residuals (Dunn and Smyth, 1996) were generated using the rqgam.check function of the dsm package in R (Miller et al. 2015). Randomised quantile residuals transform the residuals to be exactly normally distributed making the residuals vs. linear predictor plot much easier to interpret as it therefore doesn't include the artifacts generated by the link function. The goodness-of-fit statistic R^2, the percentage of variance explained, was used to evaluate the performance of the models as well as the prediction map derivate of the model in comparison to the real data.

Ecological interpretation

Ecological interpretation of the models was aided by predictor variable importance measures and partial dependence plots. In classification random forest, variable importance is measured as the mean decrease in Gini value, otherwise known as Gini impurity. When the response data are split into two child nodes based on a randomly-chosen variable, the data in the two descendent nodes are more homogeneous than that of the parent node. This difference in homogeneity between parent and child nodes is measured by the Gini index, where the increase in homogeneity equals in a decrease in Gini value. The sum of all decreases in Gini index for each variable in each tree

is averaged across all trees in the model 'forest' and then across all 10 repetitions of each model fold. The variable with the highest mean decrease in Gini value is considered the most important variable in the model. Variable importance in regression random forest is measured by the mean decrease in the residual sum of squares when the variable is included in a tree split.

Partial dependence plots using the partialPlot function in R were generated for the 6 highest variable importance scores. Partial dependence plots show the relationship between a particular predictor variable and log-transformed predicted probabilities of presence (for classification models) or the biomass regression function (for regression models),while the other predictor variables were held constant at their mean observed value and are useful in showing general trends in model accuracy's dependence on the predictors (Herrick et al., 2013). For classification models, the *y* axis ranges from $-\infty$ to ∞ and quantifies the log-odds of a positive classification for the total range of values in x. Log-odds are logarithmic transformations of the probabilities for values in x (Hastie et al., 2005). These values were transformed to the original presence probability scale using $p = \exp(y) / (1 + \exp(y))$, where p = the probability of presence, and y is the log-odds of presence, the standard output from the partialPlot function.

RESULTS

Relationship between Catch Weight and Trawl Duration

Analyses of the catch weight distributions between trawl durations (length) indicated that there is no relationship between the catch weight and length of the tow (Figure 3). This is consistent with the aggregated distribution of the VME indicator taxa and the need for alignment between the trawl path and the maximum dimension of the aggregation to create a linear relationship. For this reason biomass was analysed as kg/tow instead of standardized to CPU.

Temperature relationship between model data (GLORYS) and *in situ* data (CTD)

Bottom temperature recorded from CTD during bottom trawl surveys carried out in August in the northern region and in September in the southern region from 2006 to 2011 was used to study the Spearman correlation with the mean bottom temperature obtained from GLORYS data for the same month and year in each region. This was done to evaluate the GLORYS data so that we could judge whether ecological inferences could be made from those modeled variables. In the northern region the average ($\pm$ SD) correlation between years was 0.761 $\pm$ 0.020 and when the full period was considered, was 0.81. In the southern region the average ($\pm$ SD) correlation between years was 0.836 $\pm$ 0.003 and when the full period was considered, was 0.81. For the entire Gulf region the average ($\pm$ SD) correlation between years was 0.838 $\pm$ 0.002 and when the full period was considered, was 0.839 (Figure 4).

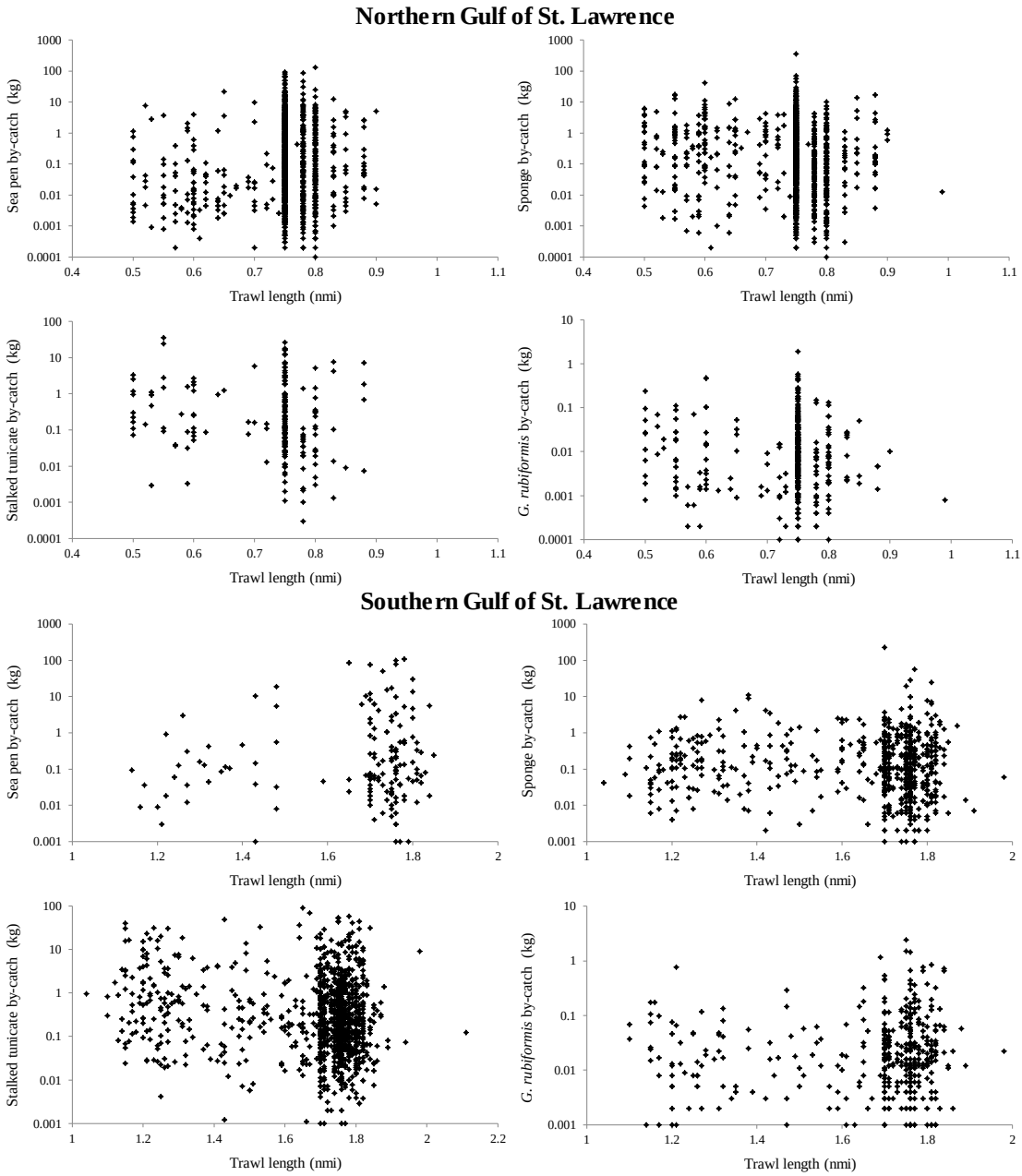

Figure 3. By-catch weight (kg) of each taxon in relation of trawl length (nmi) for the surveys carried out in the northern (upper panel) and southern (lower panel) Gulf of St. Lawrence. Note that the *y* axis is in logarithmic scale.

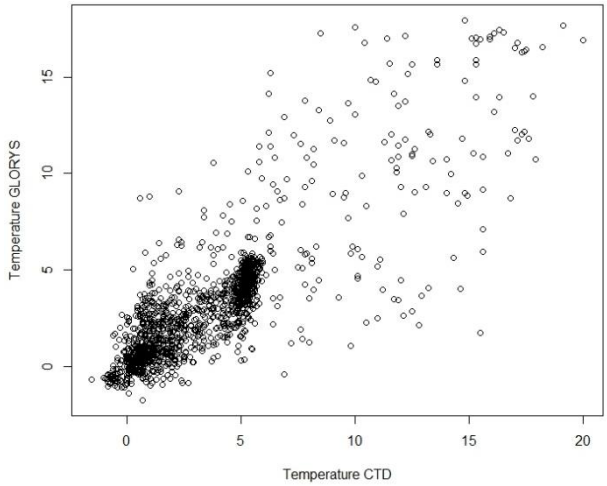

Figure 4. Scatterplot between mean bottom GLORYS temperature and CTD bottom temperature for the Gulf Region (2006-2011 period).

Two years (2006 and 2011) were used to study the spatial correlation between the mean bottom temperature based on GLORYS data and the bottom temperature based on CTD data obtained for the same month and year in each of the regions. Figures 5 to 8 show a high spatial correlation between both data types for both regions. The only discrepancy between the figures is found around Magdalen Islands in the southern Gulf Region (Figures 7, 8) where the GLORYS data predict higher mean bottom temperature than the CTD data. However, this area is not surveyed by the cruises, therefore the pattern observed in the CTD figures are also product of the CTD interpolation data.

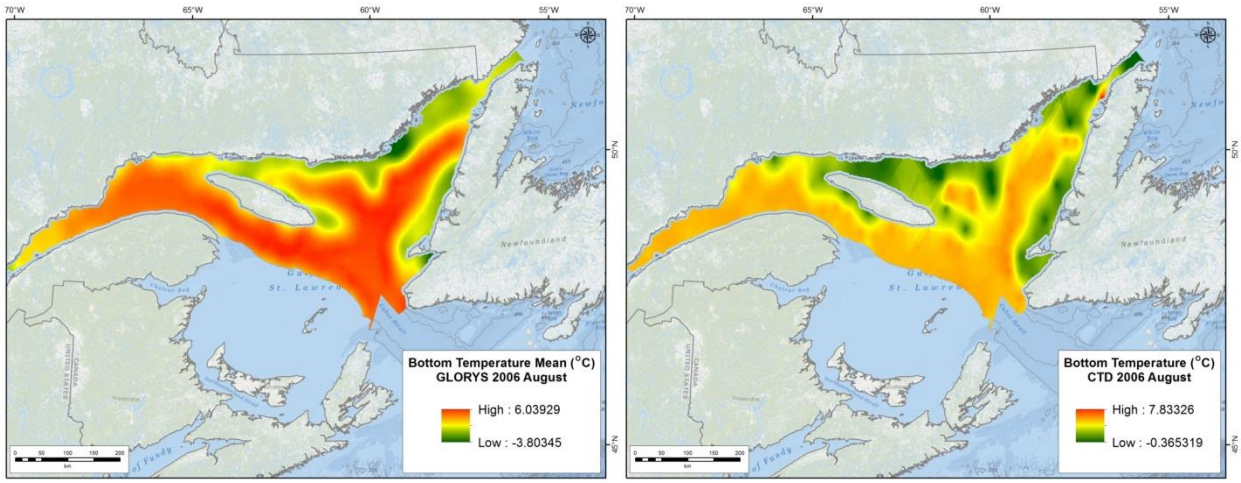

Figure 5. Interpolated prediction surface of Bottom Temperature (˚C) based on GLORYS data (left panel) and CTD data (right panel) in the northern Gulf of St. Lawrence for 2006 August.

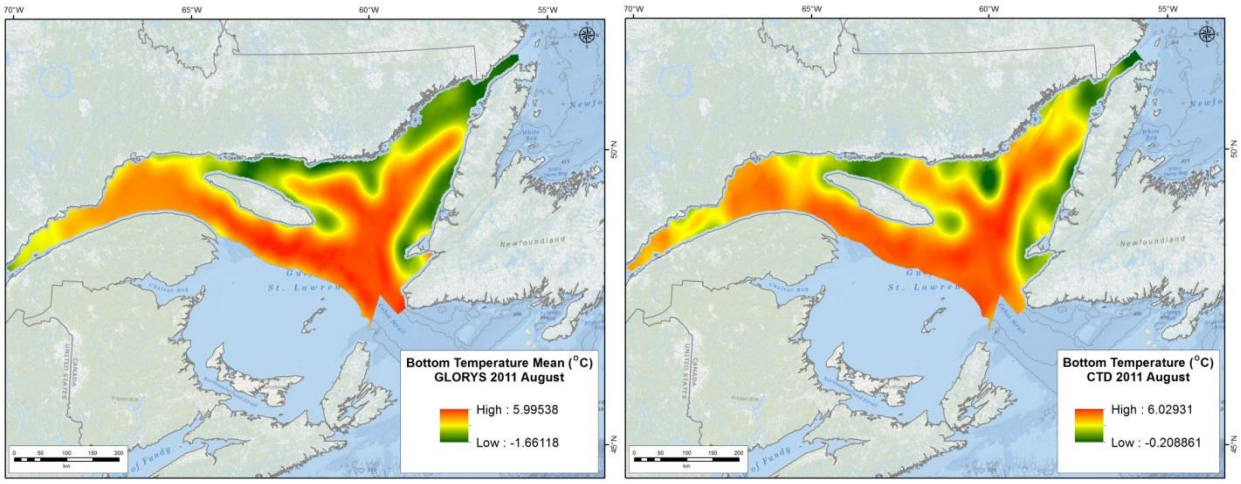

Figure 6. Interpolated prediction surface of Bottom Temperature (˚C) based on GLORYS data (left panel) and CTD data (right panel) in the northern Gulf of St. Lawrence for 2011 August.

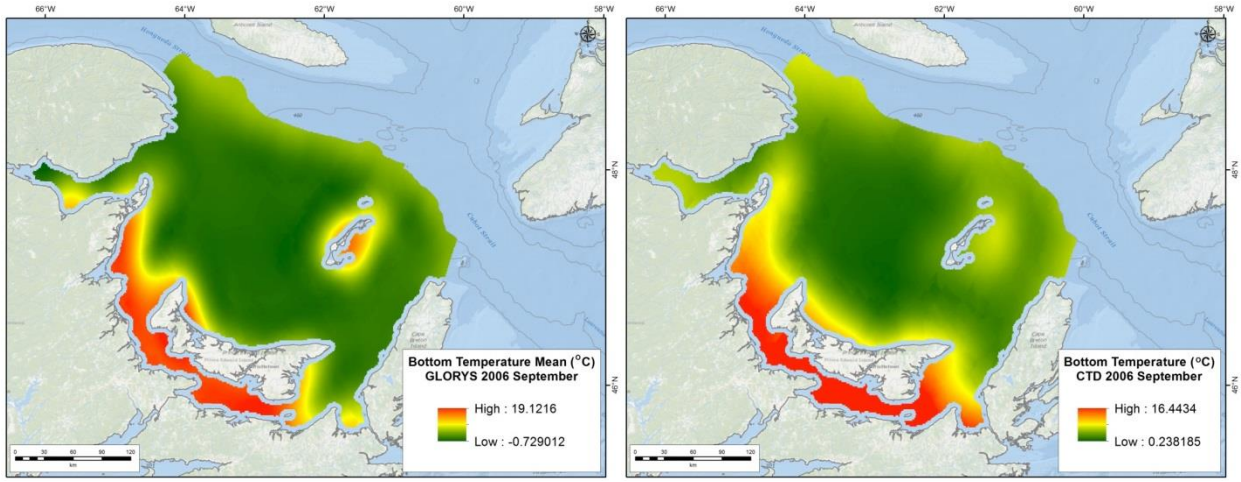

Figure 7. Interpolated prediction surface of Bottom Temperature (˚C) based on GLORYS data (left panel) and CTD data (right panel) in the southern Gulf of St. Lawrence for 2006 September.

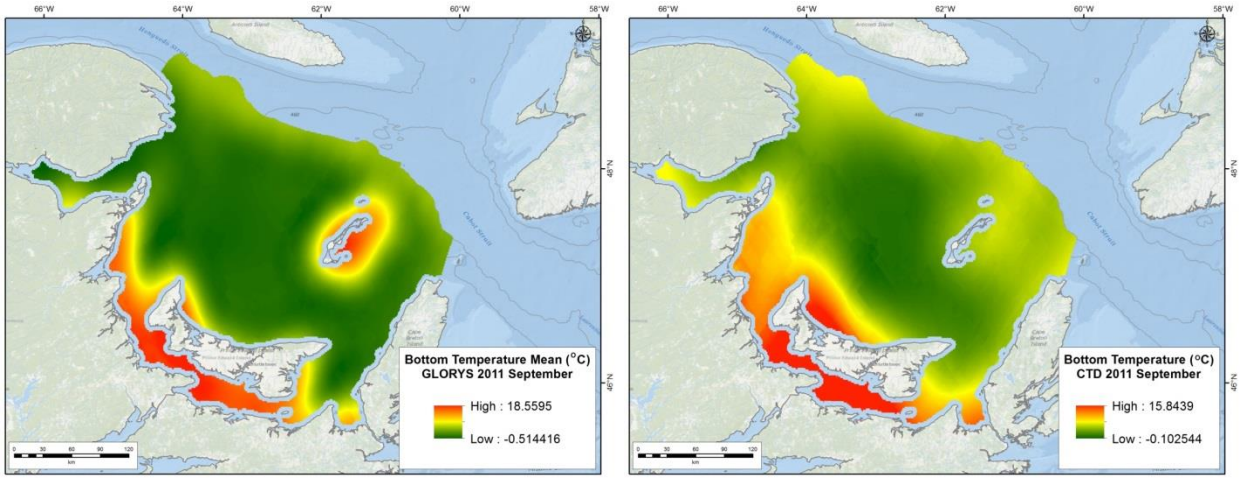

Figure 8. Interpolated prediction surface of Bottom Temperature (˚C) based on GLORYS data (left panel) and CTD data (right panel) in the southern Gulf of St. Lawrence for 2011 September.

Distribution Modelling

Sea Pens (Pennatulacea)

Data Sources and Distribution

The data available for analysis of sea pens in the Gulf Region was collected over a period of 13 years from 2003 to 2015 (Table 3). The data were filtered so that only one presence or absence record occurs within the same environmental data raster cell (approximately 1 km^2; with presence records taking precedence over an absence record when both occurred in the same cell), giving a total of 238 presence and 1595 absence records for the southern Gulf and 1034 presence and 769 absence records for the northern Gulf. Sea pen biomass (kg) was averaged across each catch occurring within the same cell. Sea pens presence and absence records were distributed unevenly across the Gulf Region (Figure 9). Presences were predominant on the northern Gulf and absences in the southern. Highest biomass occurred mainly along the Laurentian Channel.

Table 3. Total number of presences and absences of sea pen catch data recorded from DFO trawl surveys from 2003 to 2015 in the Gulf Region. P, presences; A, absences.

	Southern Gulf		**Northern Gulf**	
Year	P	A	P	A
2003	5	77	-	-
2004	17	202	2	126
2005	28	190	-	-
2006	18	147	80	101
2007	31	136	105	74
2008	24	151	121	69
2009	28	119	117	49
2010	28	98	100	40
2011	19	103	105	51
2012	10	126	98	74
2013	14	109	97	46
2014	16	137	103	69
2015	-	-	106	70
TOTAL	**238**	**1595**	**1034**	**769**

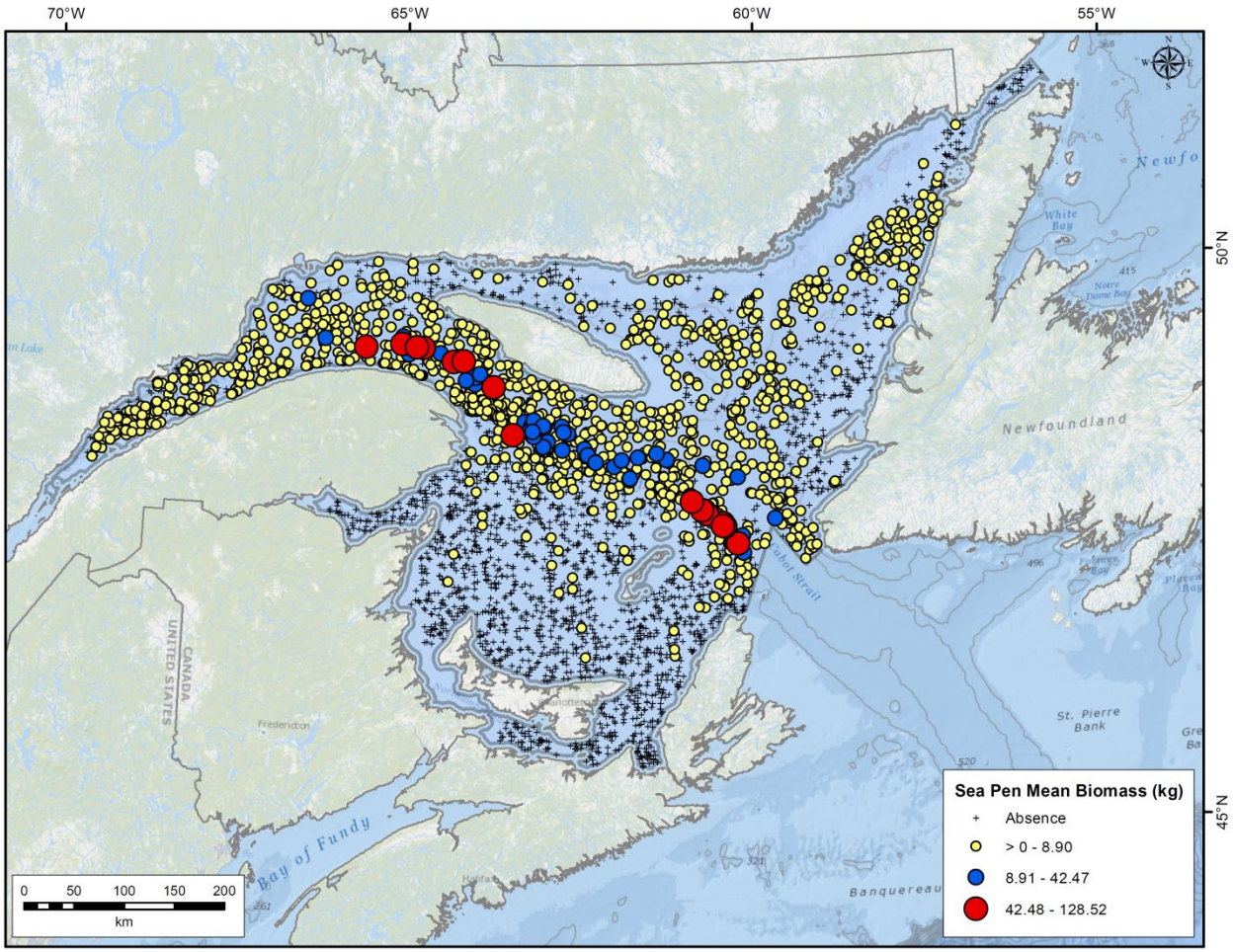

Figure 9. Mean biomass (kg) per grid cell of sea pen by-catches recorded from DFO groundfish trawl surveys from 2003 to 2015 within the Gulf Region.

Model 1 – Balanced species prevalence

Accuracy measures (mean AUC, sensitivity and specificity) for the random forest model on balanced species prevalence data are presented in Table 4. The highest AUC was 0.912 associated with Model Run 3 which is considered the optimal model for the prediction of sea pen response data. The sensitivity and specificity measures of this model were 0.840 and 0.822, respectively. The confusion matrix of the optimal model is also presented in Table 4. Class error for both classes was below 0.180.

Table 4. Accuracy measures for all 10 model repetitions of 10-fold cross validation from random forest modelling of presence and absence of sea pens within the Gulf Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 3) which is considered the optimal model for predicting sea pens probability in the Region.

Model Run	AUC	Sensitivity	Specificity
1	0.902	0.843	0.811
2	0.901	0.845	0.806
3	**0.912**	**0.840**	**0.822**
4	0.907	0.853	0.814
5	0.908	0.851	0.818
6	0.909	0.857	0.812
7	0.906	0.836	0.818
8	0.907	0.838	0.818
9	0.905	0.842	0.811
10	0.909	0.840	0.825
Mean	**0.907**	**0.845**	**0.815**
SD	**0.003**	**0.007**	**0.006**

Confusion matrix of model with maximum AUC:

Observations	Predictions		Total n	Class error
	Absence	**Presence**		
Absence	1046	226	1272	0.178
Presence	204	1068	1272	0.160

The presence probability prediction surface of sea pens is presented in Figure 10. The highest predictions of presence probability occurred in the deeper areas along the Laurentian Channel. Areas of higher presence probability corresponded well with the spatial distribution of presence records (see Figure 11). The area around Prince Edward Island is predicted to have zero or low presence probability of sea pens.

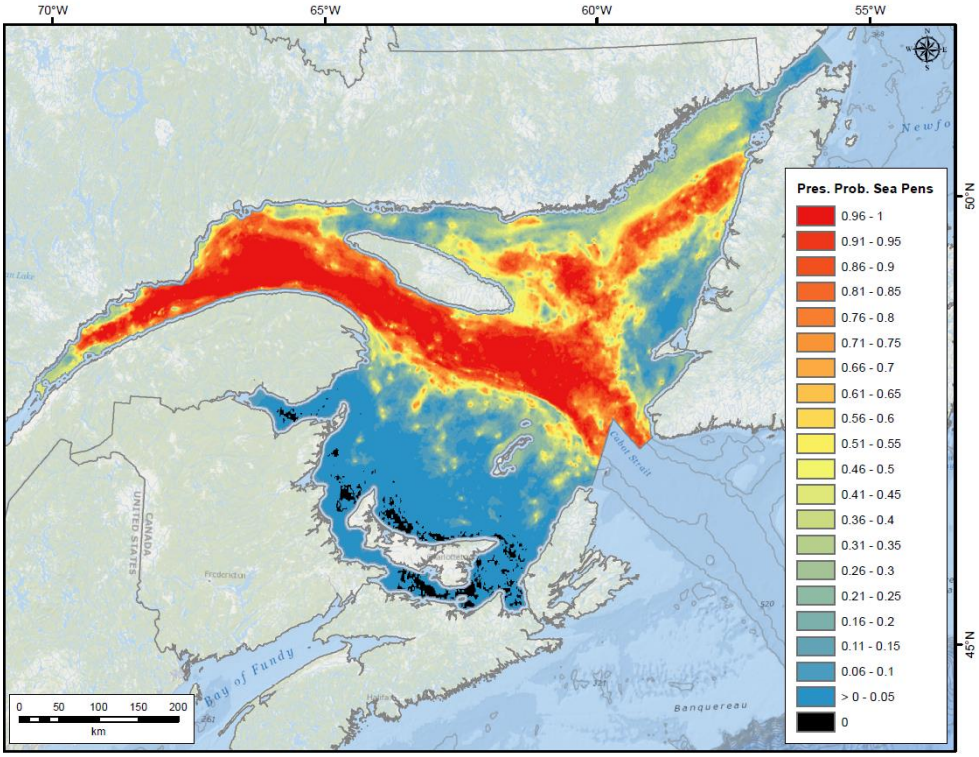

Figure 10. Predictions of presence probability (pres. prob.) of sea pens from the optimal random forest model (see Table 4) based on balanced data.

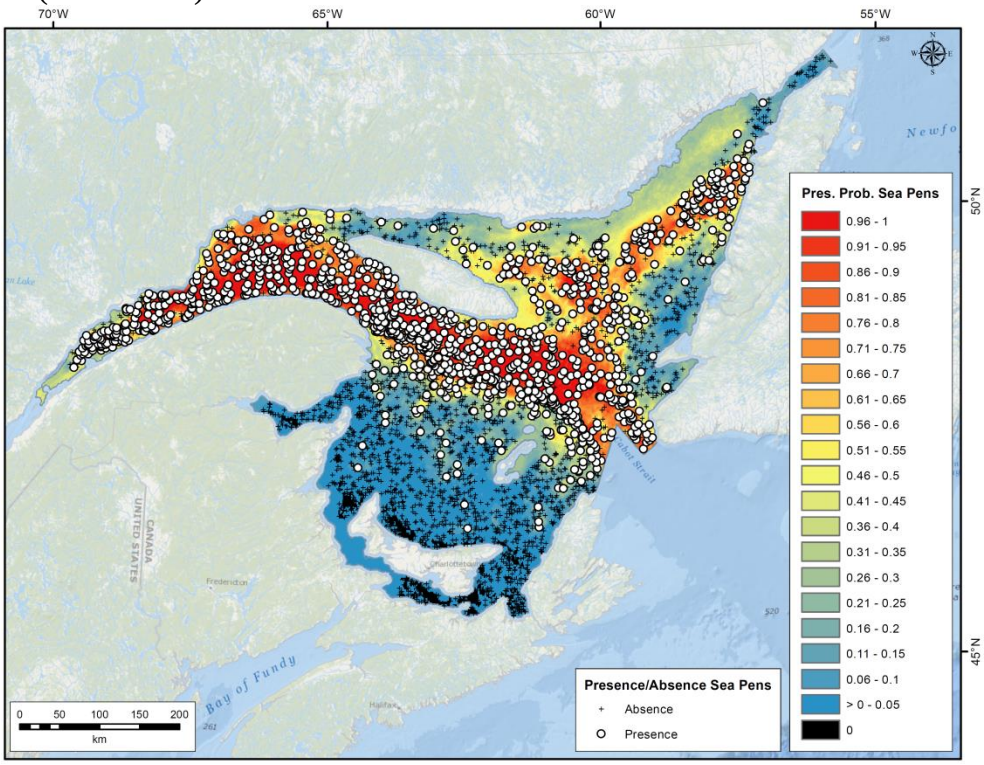

Figure 11. Total presence and absence observations and predictions of presence probability (pres. prob.) of sea pens from the optimal random forest model (see Table 4) based on balanced data.

The actual presence and absence data observations (1272 presences and 1272 absences) used in the optimal model did not show important differences (Figure 12) with respect of the distribution of all the absence records across the study area (Figure 11). In this figure are also showed the areas of model extrapolation (i.e. areas where at least one environmental variable has values beyond its sampled range). The south of Prince Edward Island and the shallow portion in the northeast, before the Strait of Belle Isle are considered extrapolated area, as well as smaller areas scattered across the study area.

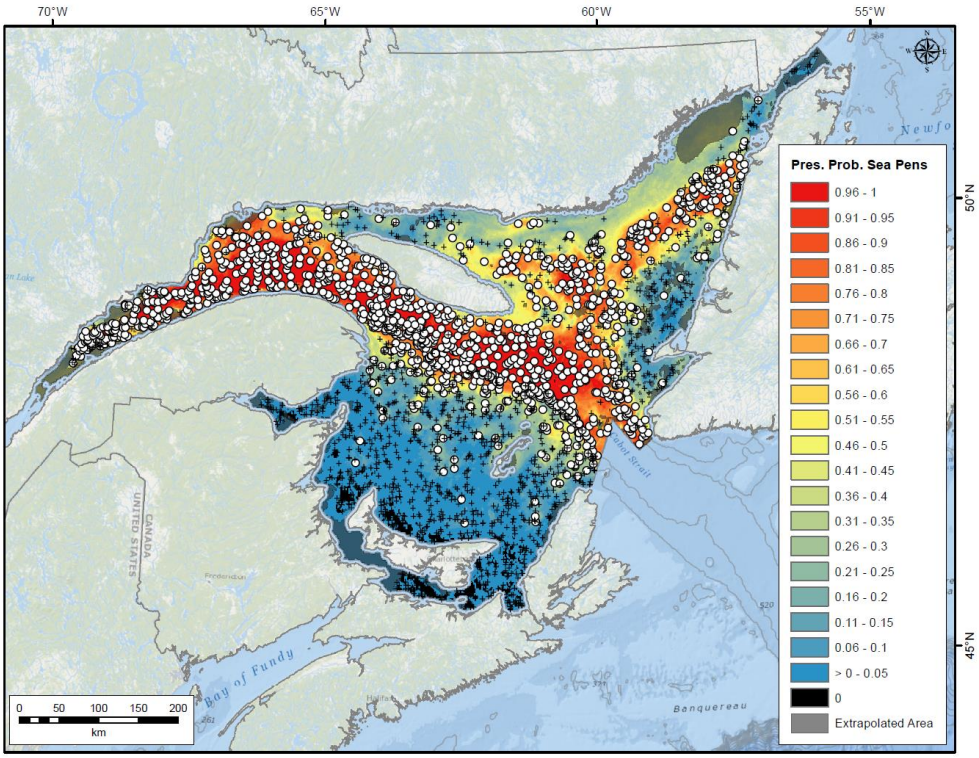

Figure 12. Presence and balanced absence observations and predictions of presence probability (pres. prob.) of sea pens from the optimal random forest model (see Table 4) based on balanced data. The grey areas of extrapolation are also shown.

Of all 78 environmental predictor variables used in the model, Bottom Salinity Average Minimum was the most important for the classification of the sea pen presence and absence data (Figure 13). This variable was followed in terms of its Mean Decrease in Gini Value by Bottom Salinity Mean and Bottom Salinity Average Maximum. Bottom Temperature Average Minimum and Depth variables had high importance in this model too. The partial dependence plots for the top 6 most important predictors are shown in Figure 14. Presence probability of sea pens rapidly increased at a salinity average minimum at bottom of almost 31, and depths values of ~60-70 m.

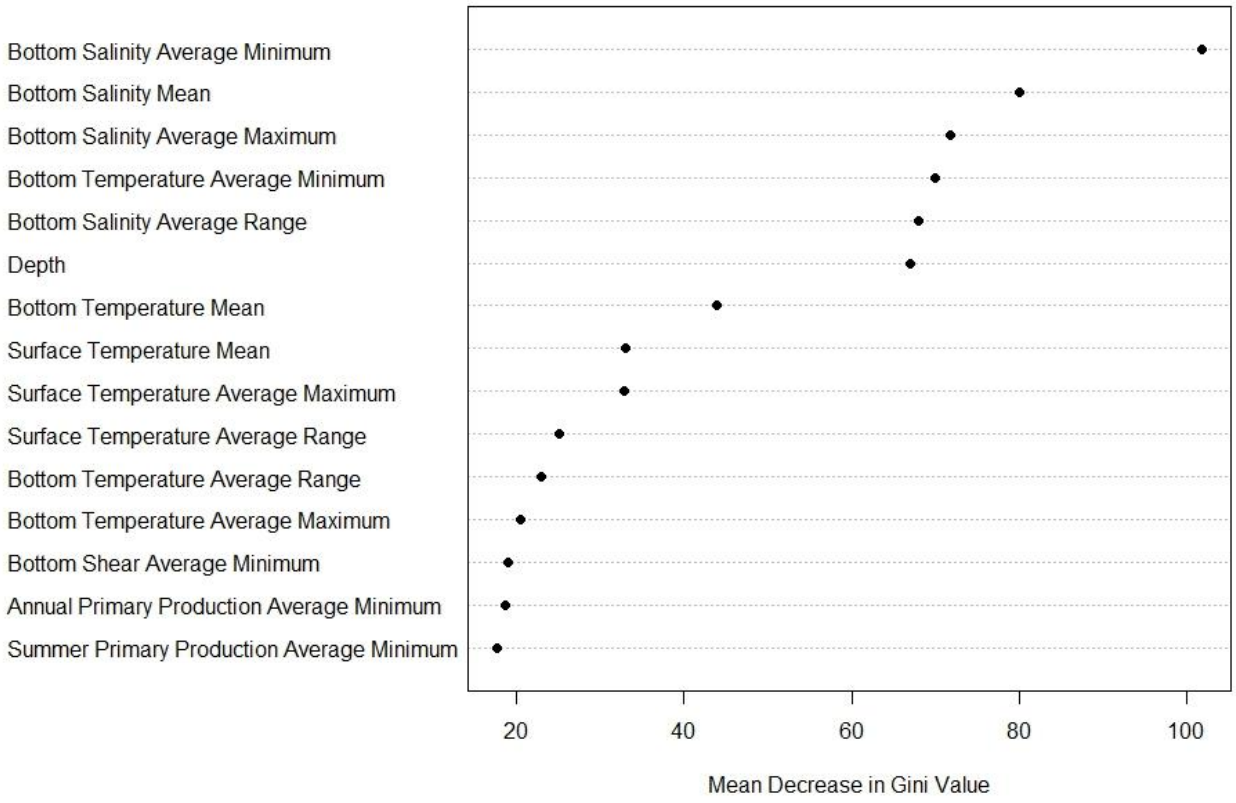

Figure 13. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model of sea pens presence and absence data within the Gulf Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

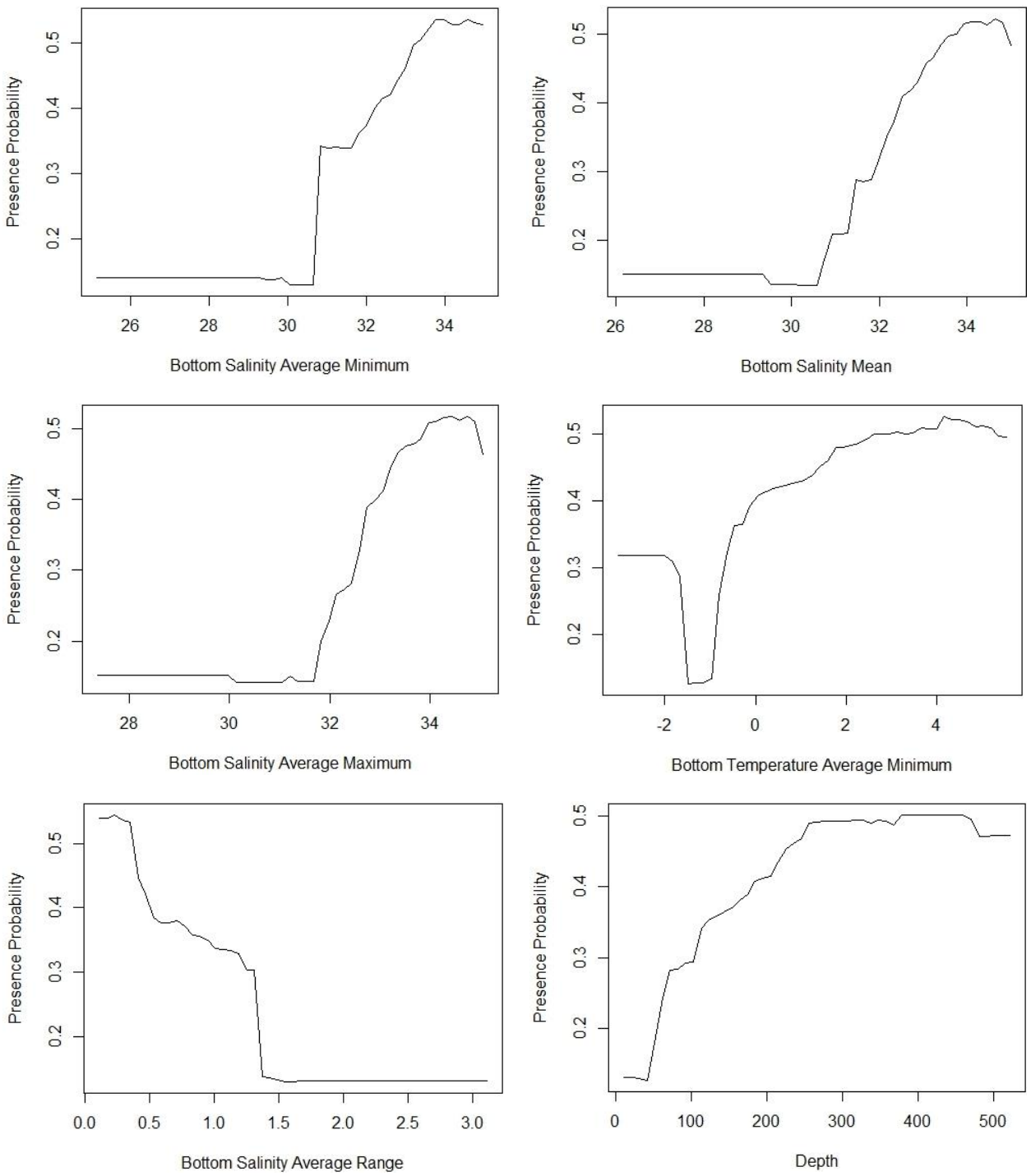

Figure 14. Partial dependence plots of the top six predictors from the optimal random forest model of sea pens presence and absence data collected within the Gulf Region, ordered left to right from the top.

Model 2 – Unbalanced species prevalence

Table 5 shows the accuracy measures for the random forest model on sea pen presence and absence data (1272 presences and 2364 absences) and a threshold equal to species prevalence (0.35). The AUC calculated from Model 2 was 0.905, lower than that of Model 1 (0.912). Class error of the presence class was the same to that of Model 1 (0.160), however class error for the absence class was slightly higher (0.185 compared to 0.178 from Model 1). Sensitivity was the same for Model 2 (0.840), and Specificity of Model 2 was slightly lower than that of Model 1 (0.815 versus 0.822).

Table 5. Accuracy measures and confusion matrix from 10-fold cross validation from random forest modelling of presence and absence of sea pens within the Gulf Region with a cut-off threshold of 0.35.

Observations	Predictions		Total n	Class error	AUC	Sensitivity	Specificity
	Absence	**Presence**					
Absence	1926	438	2364	0.185	0.905	0.840	0.815
Presence	203	1069	1272	0.160			

The predicted presence probability surface of sea pens generated from Model 2 is shown in Figure 15. The highest presence probability of sea pens is found along the Laurentian Channel where are found most of the sea pen presence records (Figure 16). The area around Prince Edward Island is predicted to have zero or low presence probability of sea pens.

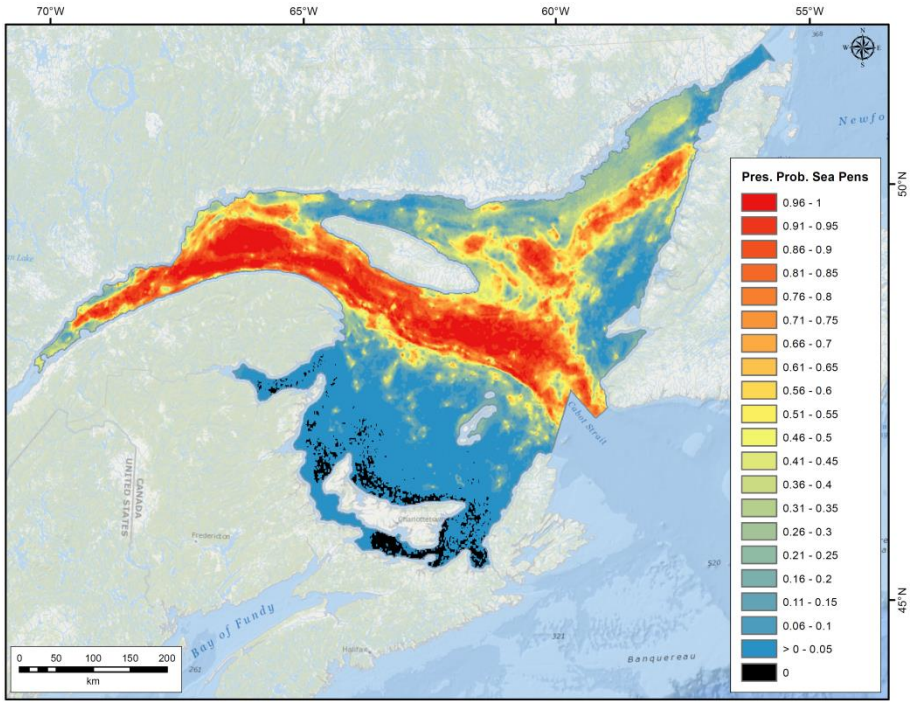

Figure 15. Predictions of presence probability (pres. prob.) of sea pens from the unbalanced random forest model (see Table 5).

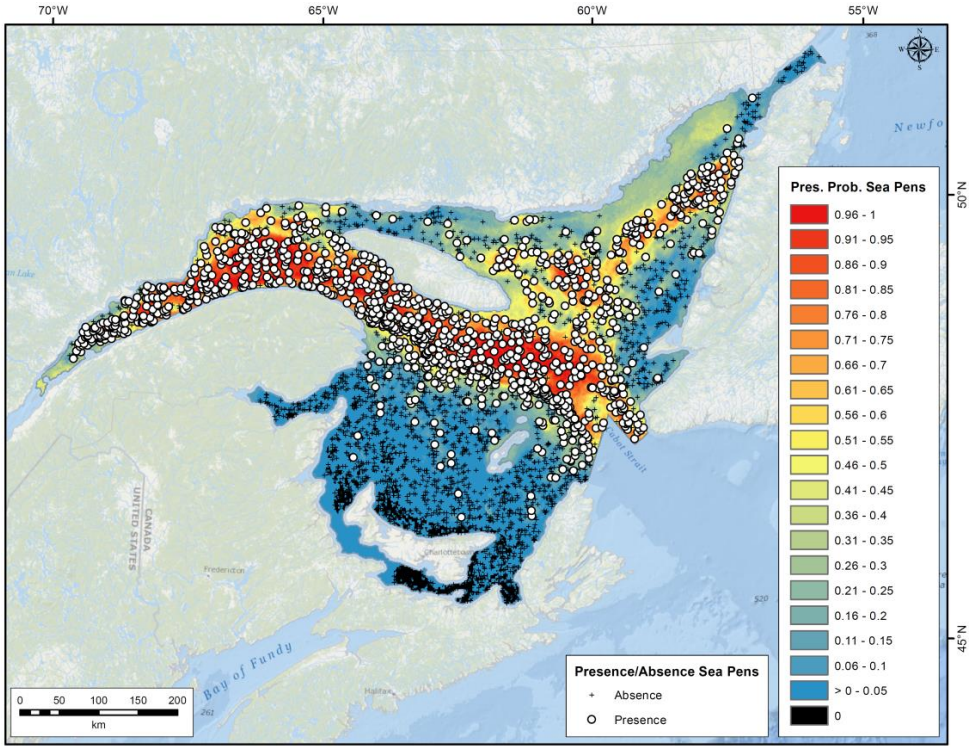

Figure 16. Presence and absence observations and predictions of presence probability (pres. prob.) of sea pens from the unbalanced random forest model (see Table 5).

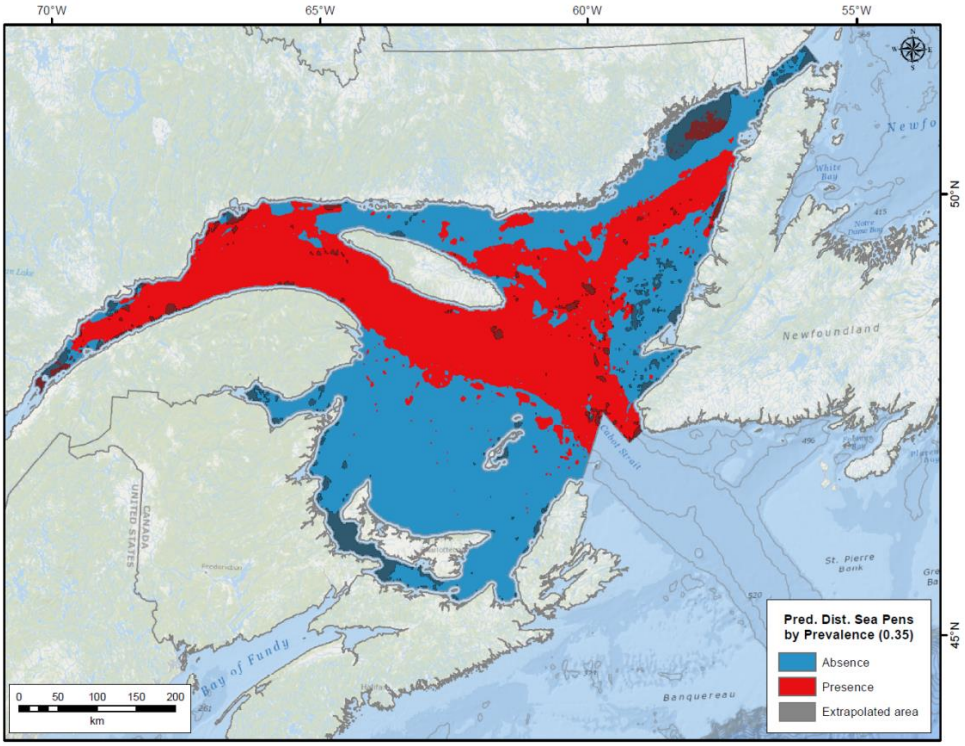

Figure 17. Predicted distribution of sea pens from the unbalanced random forest model (see Table 5) using the prevalence cut-off threshold of 0.35. The grey areas of extrapolation are also shown, they appear dark red when overlain on the red presence surface.

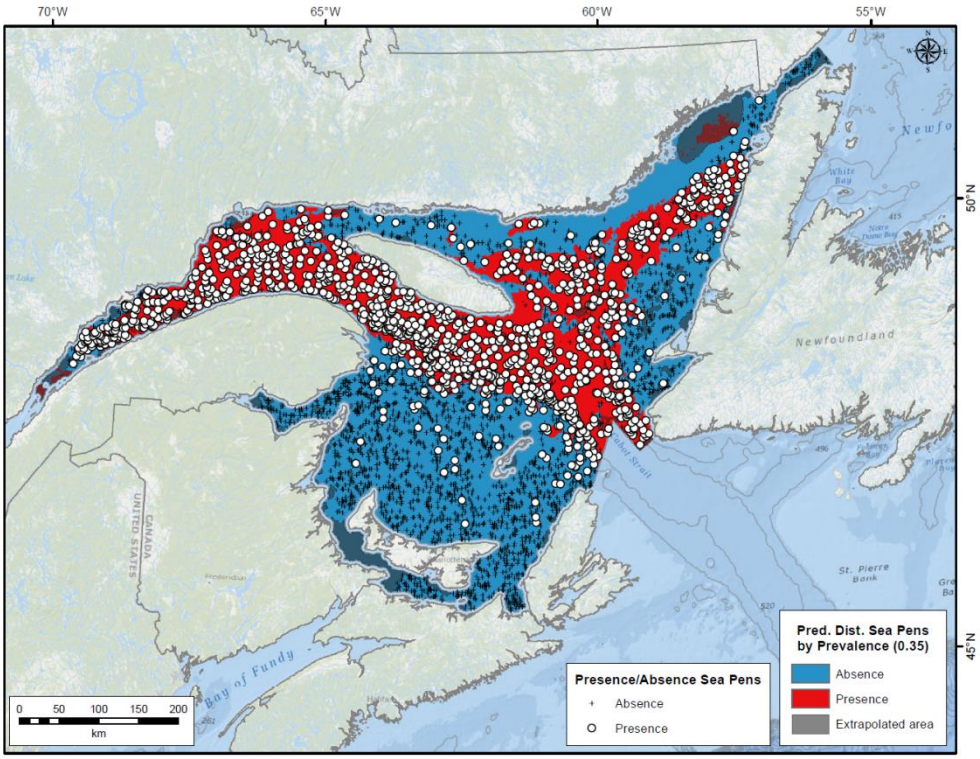

Figure 18. Presence and absence observations and predicted distribution of sea pens from the unbalanced random forest model (see Table 5) using the prevalence cut-off threshold of 0.35. The grey areas of extrapolation are also shown, they appear dark red when overlain on the red presence surface.

The six most important environmental predictor variables in Model 2 were the same than that of Model 1 (Figure 19), although in this case Depth was the most important variable in Model 2 compared to Bottom Salinity Average Minimum in Model 1. Partial dependence of sea pens presence and absence data on the top 6 predictor variables is shown in Figure 20. Presence probability of sea pens increased at ~50 m depth and reaches is maximum around ~300 m.

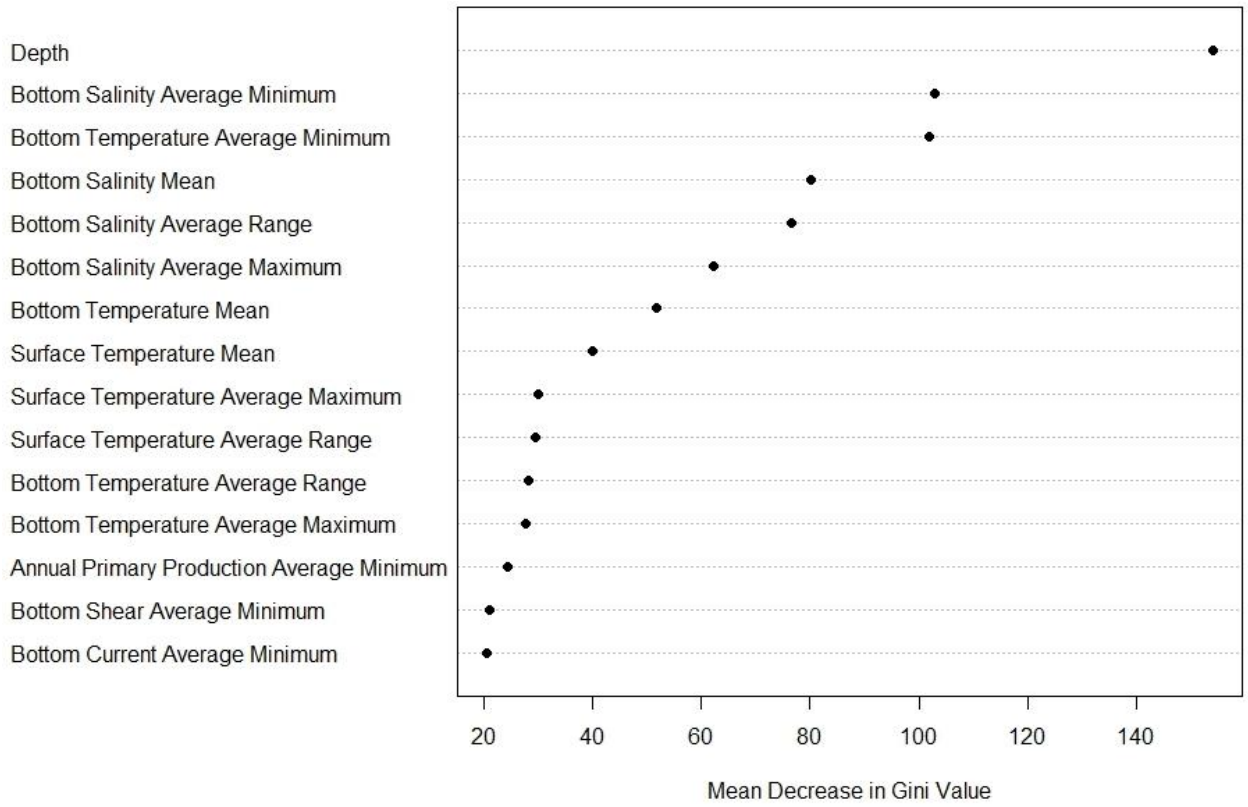

Figure 19. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the unbalanced random forest model of sea pen presence and absence data within the Gulf Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

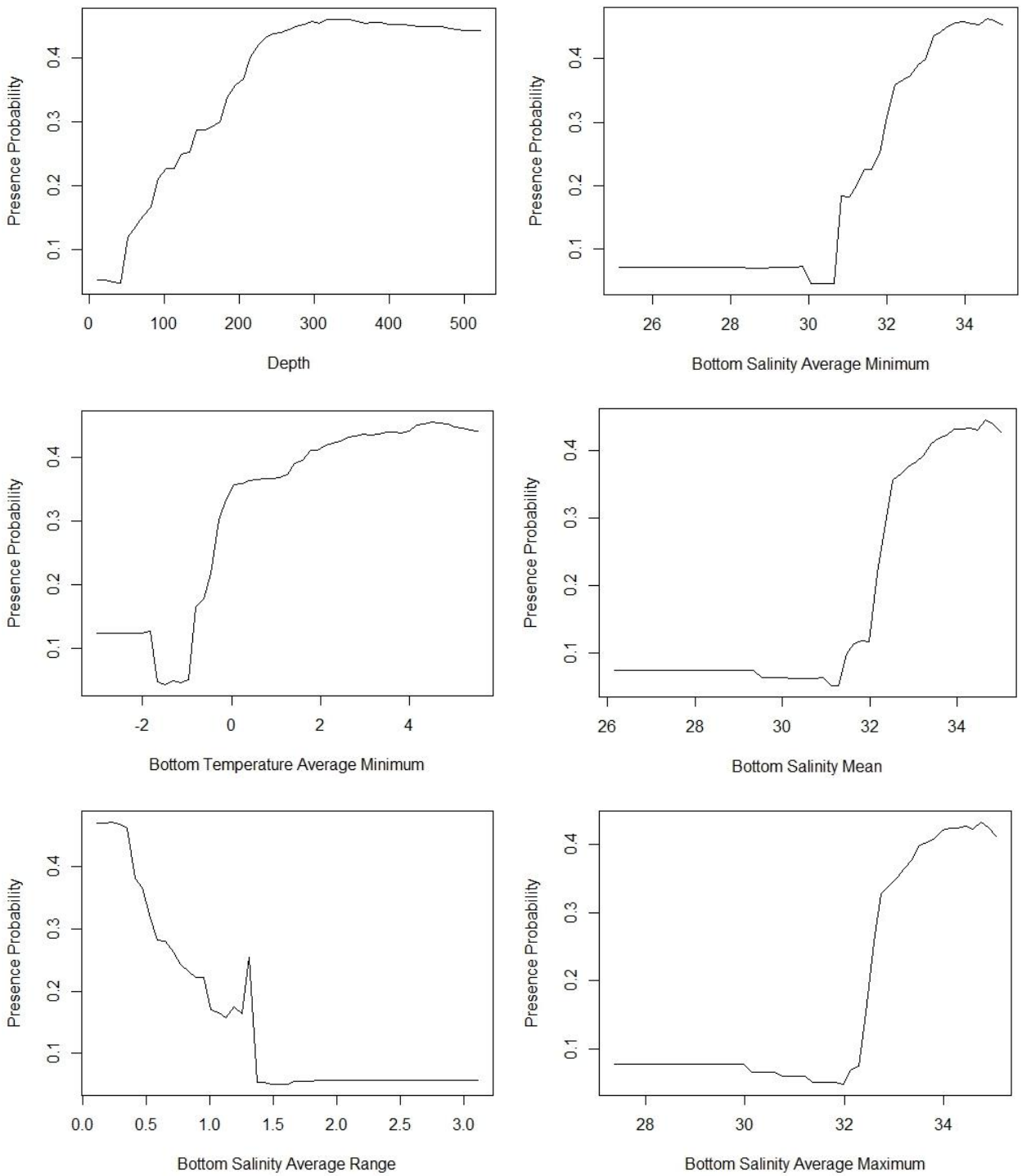

Figure 20. Partial dependence plots of the top six predictors from the unbalanced random forest model of sea pens presence and absence data collected within the Gulf Region, ordered left to right from the top.

Model Selection

Model accuracy measures were very similar for both models, with AUC values higher than 0.9 in both cases (0.912 for Model 1 and 0.905 for Model 2). The selection of Model 1, based on balanced data, allows for the use of the presence probabilities considering values higher than 0.5 as presences and lower as absences to maintain fairly constant the error rates.

Prediction of Sea Pen Biomass Using Random Forest (RF)

a) Southern Gulf of St. Lawrence

The accuracy measures of the regression random forest model on mean sea pen biomass per grid cell for the southern Gulf of St. Lawrence are presented in Table 6. The average R-square (R^2) value was 0.370 ± 0.217, the average Root-Mean-Square Error (RMSE) was 3.540 ± 1.815, and the average percent variance explained was 32.439 ± 7.728. The high variability around the mean for both of these metrics indicates high variability of the model fit.

Table 6. Accuracy measures from 10-fold cross validation from random forest modelling of biomass of sea pens within the southern Gulf of St. Lawrence. NRMSE, is the normalized root-mean square deviation (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	%variance explained
1	0.519	1.374	0.015	32.910
2	0.777	4.227	0.046	23.550
3	0.287	3.895	0.042	36.870
4	0.263	7.290	0.079	17.810
5	0.375	2.830	0.031	34.580
6	0.389	2.596	0.028	33.470
7	0.016	3.647	0.039	42.570
8	0.549	1.561	0.017	38.530
9	0.127	2.479	0.027	38.330
10	0.399	5.502	0.060	25.770
Mean	**0.370**	**3.540**	**0.038**	**32.439**
SD	**0.217**	**1.815**	**0.020**	**7.728**

Figures 21 and 22 show the surface of sea pen biomass (kg) predictions per grid cell generated from the random forest model in the southern Gulf of St. Lawrence. The maximum biomass was predicted in the Laurentian Channel portion sampled by these surveys. Around Magdalen Islands some areas were predicted with high biomass, however, this part correspond to an extrapolated area (Figure 21), so we caution that this prediction could be consequence of the lack of data.

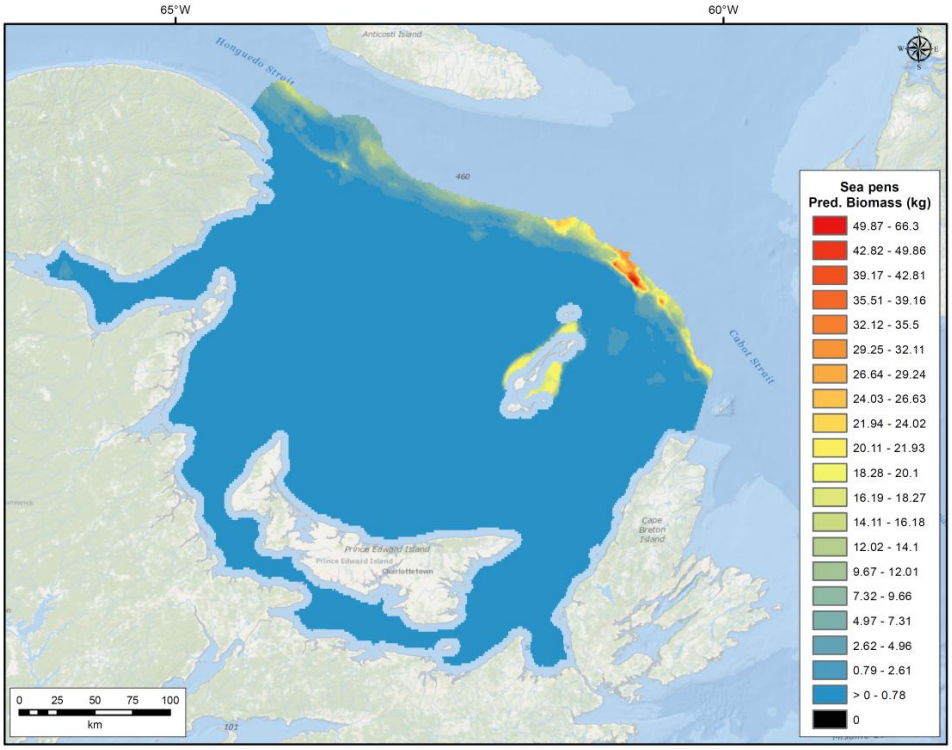

Figure 21. Prediction of sea pen biomass (kg) from the random forest (see Table 6) within the southern Gulf of St. Lawrence.

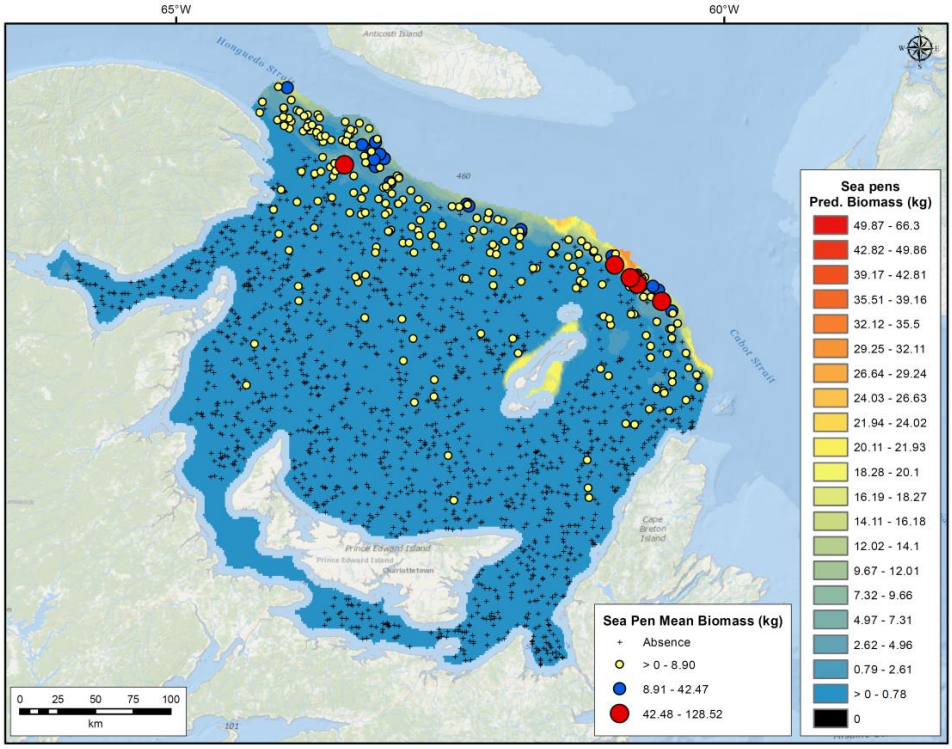

Figure 22. Sea pen biomass catches overlaid to the sea pen biomass (kg) predicted from the random forest model (see Table 6) within the southern Gulf of St. Lawrence.

The top 15 most important environmental variables for predicting sea pen biomass in the southern Gulf of St. Lawrence are shown in Figure 23. Spring Primary Production Average Minimum was the most important variable in the model, followed more distantly by Bottom Salinity Average Range and the other variables in the model. The partial dependence of sea pen biomass on the top 6 most important variables is shown in Figure 24. Spring Primary Production Average Minimum predicted highest biomass at primary production values lower than 550 mg C m^{-2} day^{-1}.

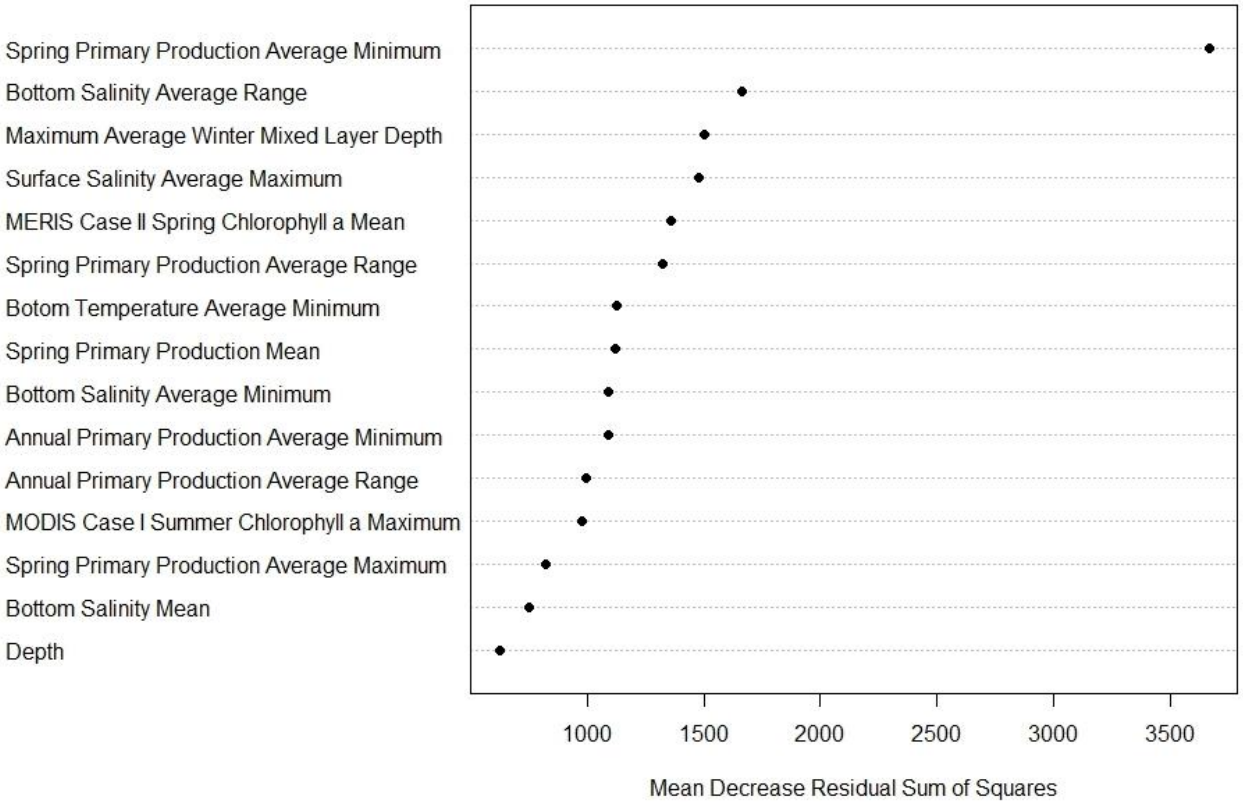

Figure 23. Importance of the top 15 predictor variables measured as the Mean Decrease Residual Sum of Squares of the random forest model (see Table 6) of sea pen biomass data within the southern Gulf of St. Lawrence.

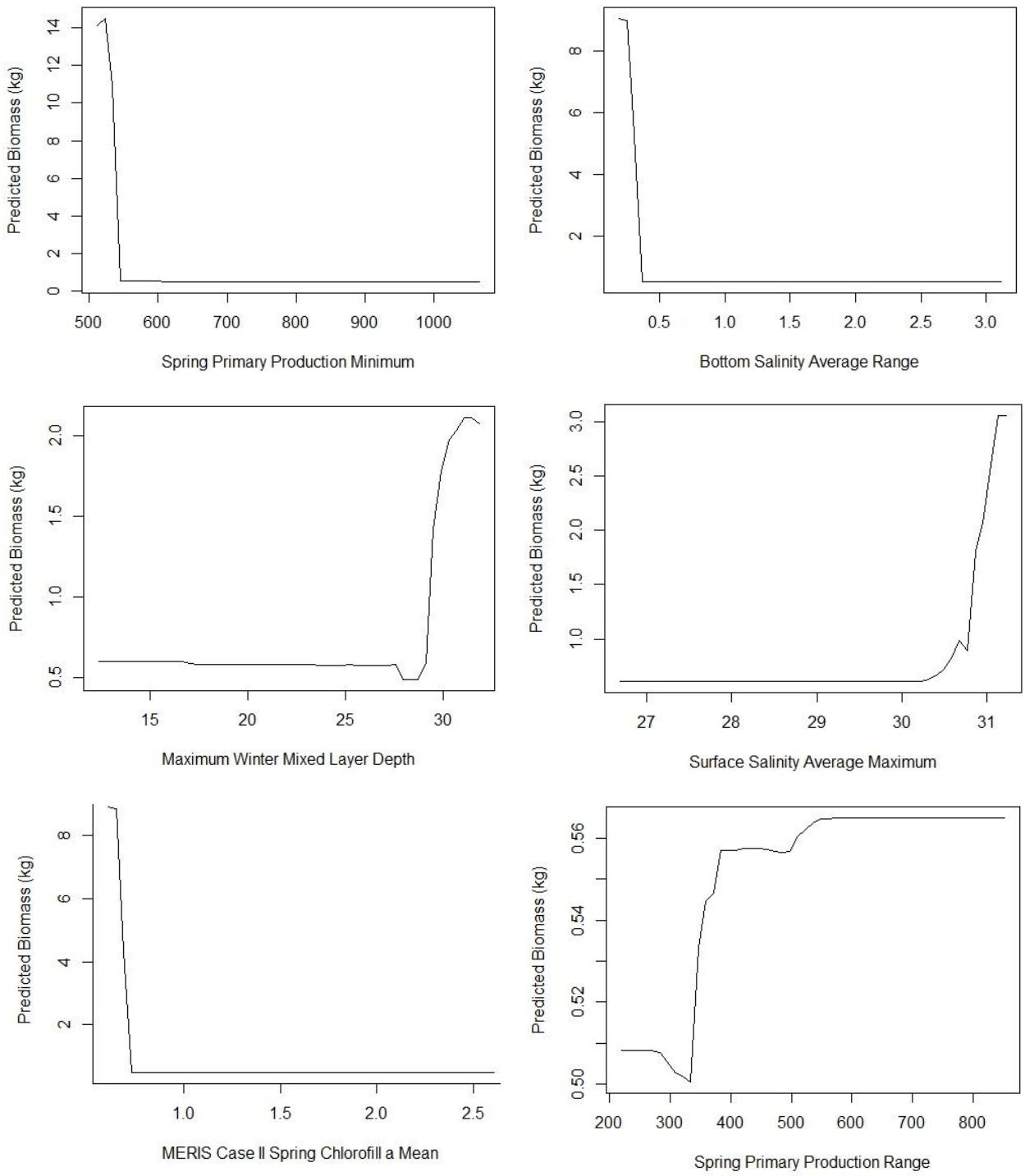

Figure 24. Partial dependence plots of the top six predictors from the random forest model based on sea pen biomass data collected within the southern Gulf of St. Lawrence, ordered left to right from the top.

b) Northern Gulf of St. Lawrence

The accuracy measures of the regression random forest model on mean sea pen biomass per grid cell for the northern Gulf of St. Lawrence are presented in Table 7. The average R-square (R^2) value was 0.273 ± 0.137, the average Root-Mean-Square Error (RMSE) was 5.030 ± 2.737, and the average percent variance explained was 22.362 ± 2.998. The variability around the mean for both of these metrics was lower than in the southern Gulf but still high, indicating variability of the model fit.

Table 7. Accuracy measures from 10-fold cross validation from random forest modelling of biomass of sea pens within the northern Gulf of St. Lawrence. NRMSE, is the normalized root-mean square deviation (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	%variance explained
1	0.096	2.861	0.022	25.790
2	0.283	2.754	0.021	22.650
3	0.502	3.199	0.025	25.370
4	0.279	11.599	0.090	17.880
5	0.314	3.788	0.029	22.380
6	0.419	7.585	0.059	20.620
7	0.027	4.276	0.033	26.360
8	0.275	3.643	0.028	21.400
9	0.303	4.891	0.038	17.970
10	0.234	5.702	0.044	23.200
Mean	**0.273**	**5.030**	**0.039**	**22.362**
SD	**0.137**	**2.737**	**0.021**	**2.998**

Figures 25 and 26 show the surface of sea pen biomass (kg) predictions per grid cell generated from the random forest model in the northern Gulf of St. Lawrence. The maximum biomass was predicted in two areas of the Laurentian Channel, with an area of lower biomass between them.

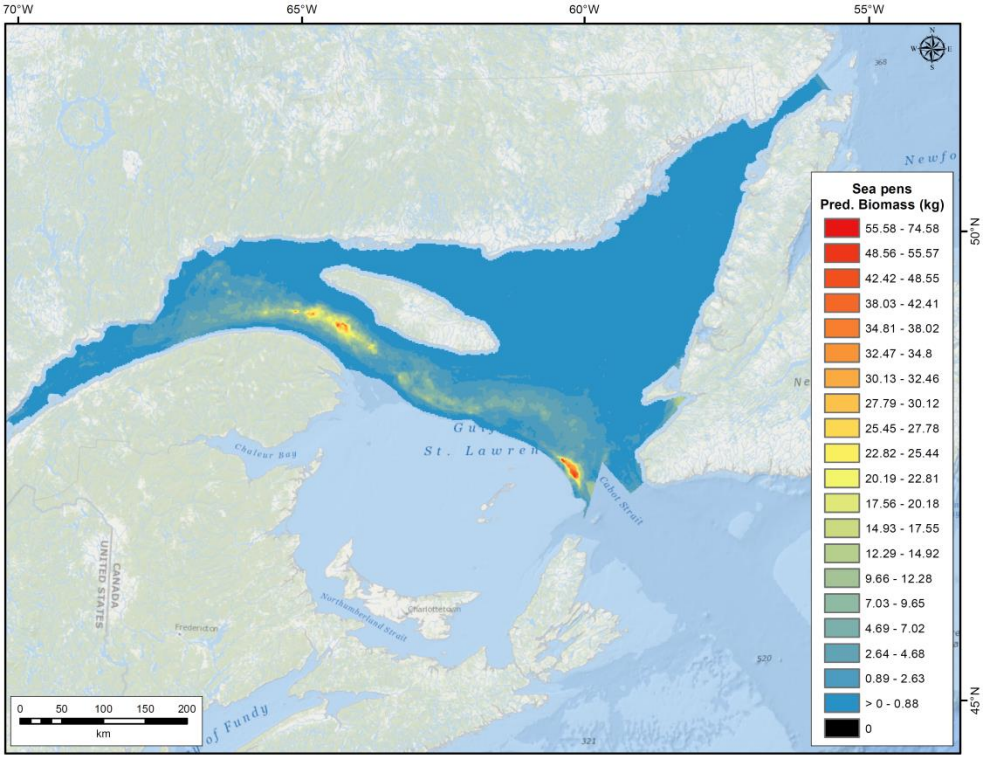

Figure 25. Prediction of sea pen biomass (kg) from the random forest model (see Table 7) within the northern Gulf of St. Lawrence.

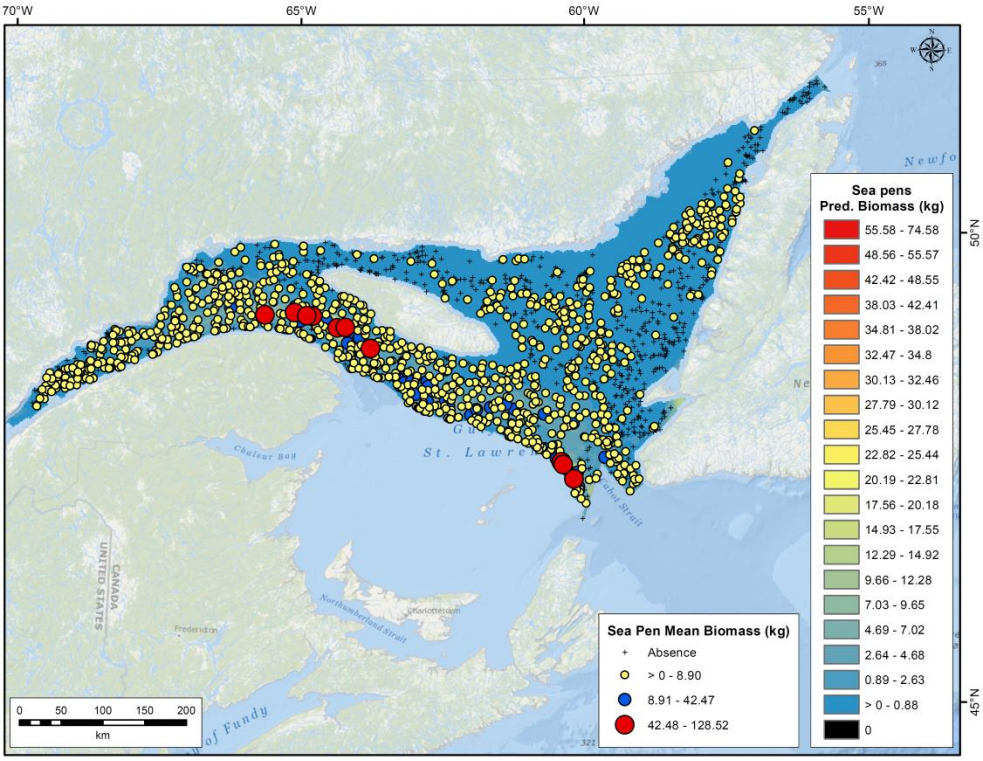

Figure 26. Sea pen biomass catches overlaid to the sea pen biomass (kg) predicted from the random forest model (see Table 7) within the southern Gulf of St. Lawrence.

The top 15 most important environmental variables for predicting sea pen biomass in the northern Gulf of St. Lawrence are shown in Figure 27. Surface Temperature Average Maximum was the most important variable in the model, followed distantly by Depth and the other variables in the model. The partial dependence of sea pen biomass on the top 6 most important variables is shown in Figure 28. Surface Temperature Average Maximum predicted highest biomass at temperature values higher than 17 °C.

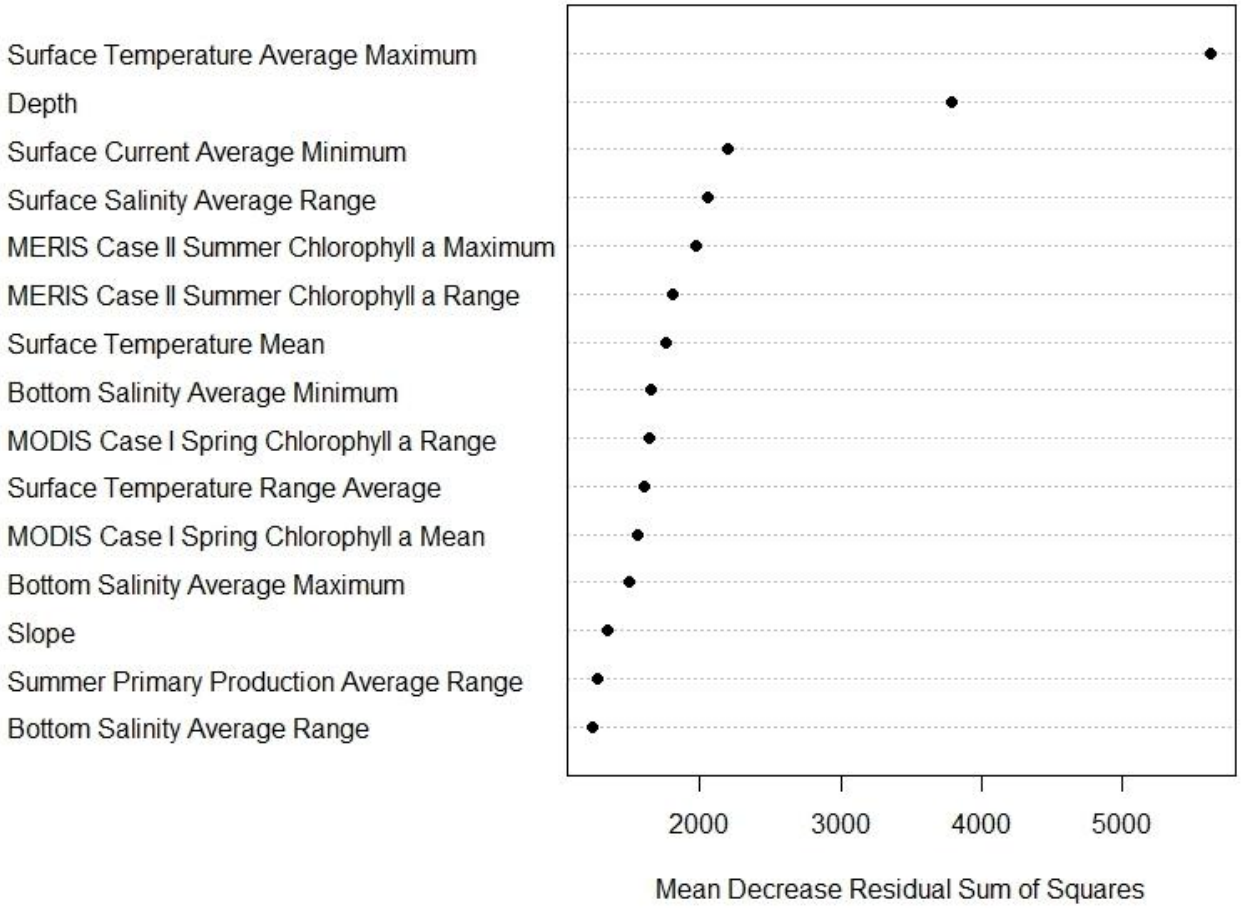

Figure 27. Importance of the top 15 predictor variables measured as the Mean Decrease Residual Sum of Squares of the random forest model (see Table 7) of sea pen biomass data within the northern Gulf of St. Lawrence.

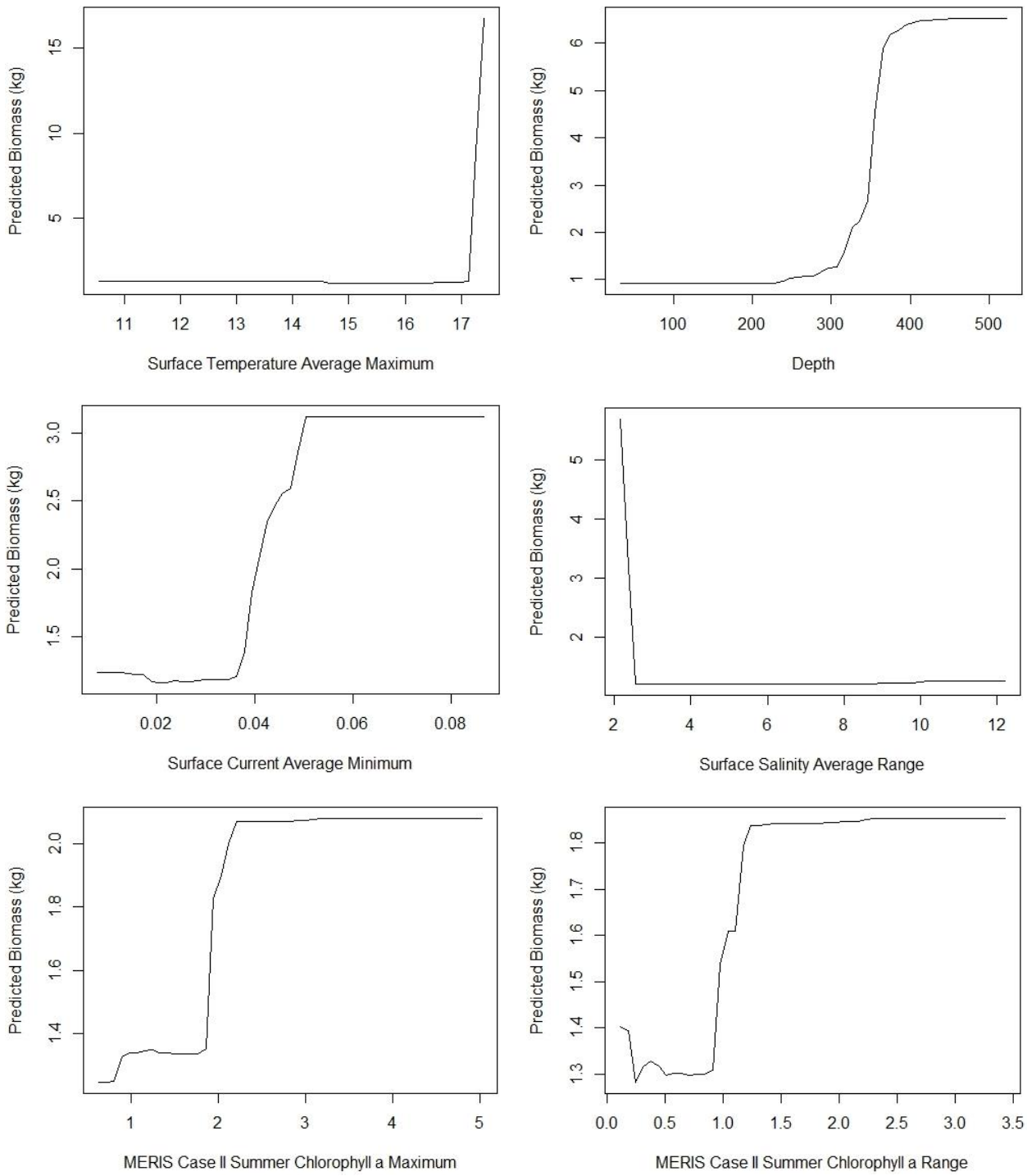

Figure 28. Partial dependence plots of the top six predictors of the random forest model (see Table 7) of sea pen biomass data within the southern Gulf of St. Lawrence on, ordered left to right from the top.

Figures 29 and 30 show the surface of sea pen biomass (kg) predictions per grid cell generated from the random forest model in the entire Gulf of St. Lawrence.

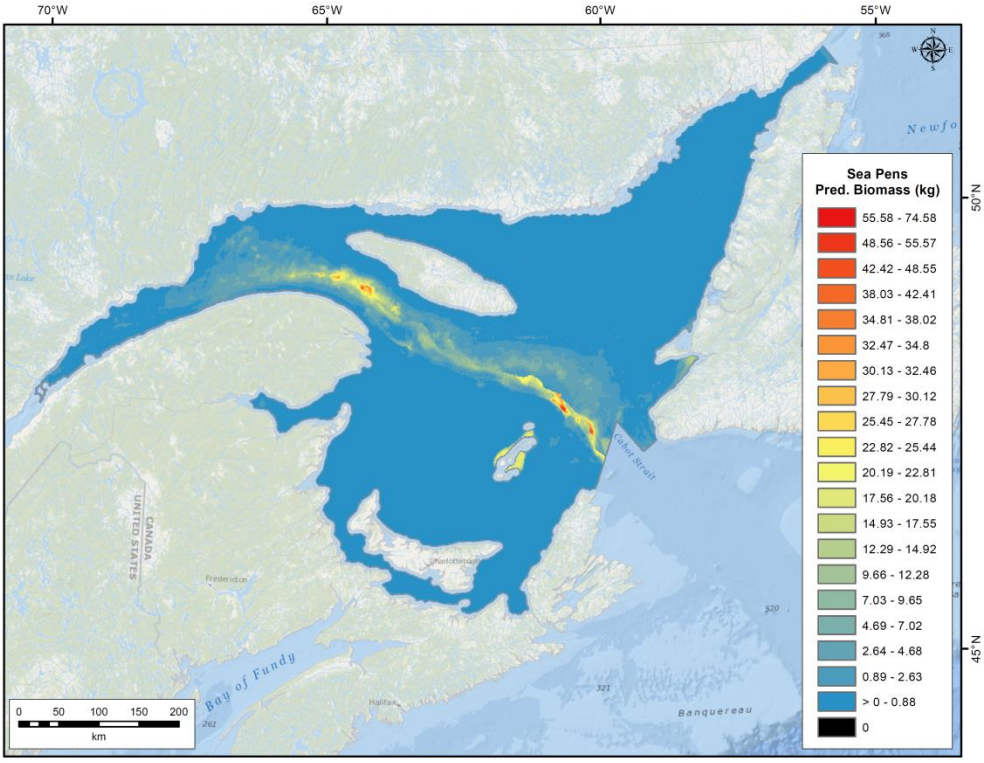

Figure 29. Predictions of sea pen biomass from the random forest model within the Gulf Region.

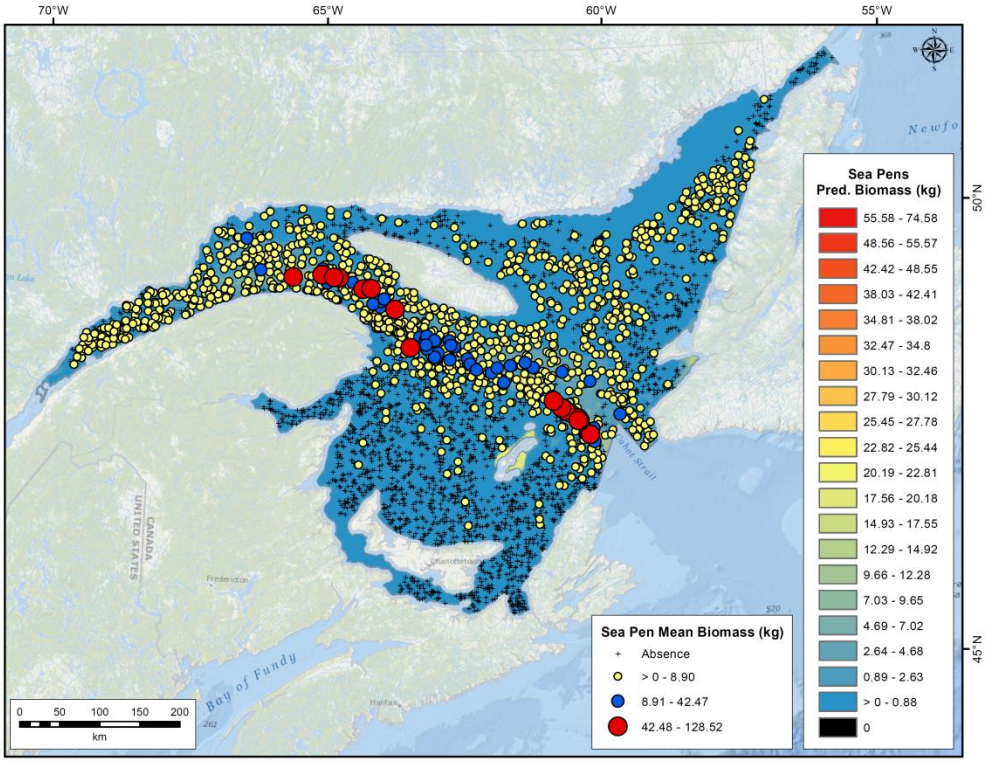

Figure 30. Sea pen biomass catches overlaid to the sea pen biomass (kg) predicted from the random forest model within the Gulf Region.

Prediction of Sea Pen Biomass Using Generalized Additive Models (GAM)

a) Southern Gulf of St. Lawrence

The GAM results on mean sea pen biomass per grid cell for the southern Gulf of St. Lawrence are presented in Table 8. The maximum R-square (R-sq) value was 0.58 for the model with environmental variables correlated less than 0.7 and including the latitude and longitude. Figure 31 shows that the autocorrelation of the residuals for Model 1 when latitude and longitude were not considered in the model was significant and therefore latitude and longitude were considered as a tensor product for Model 2.

Table 8. GAM results. Significant variables ($p < 0.05$) are listed in order of their significance (highest to lowest) in the model, the deviance explained and the estimated degrees of freedom (edf) and the R^2 are also provided. Variables are Spring Primary Production Average Minimum (pp_spr_min_avg), Depth, Bottom Temperature Average Maximum (b_tmp_max_avg), MODIS Case I Fall Chrolophyll *a* Range (chl_fall_ran), Summer Primary Production Average Range (pp_sum_ran_avg), Maximum Summer Mixed Layer Depth (mld_sum_max_avg), Spring Primary Production Average Range (pp_spr_ran_avg), Summer Primary Production Average Maximum (pp_sum_max_avg), Bottom Current Average Range (b_cur_ran_avg), Slope, Surface Current Average Range (s_cur_ran_avg), MODIS Case I Spring Chrolophyll *a* Range (chl_spr_ran), location (lat, long), MERIS Case II Fall Chlorophyll *a* Minimum (M_chl_fall_min).

Model	Significant variables	Deviance explained	edf	R^2
1 (0.7 correlation)	s(pp_spr_min_avg) + s(depth) + s(b_tmp_max_avg) + s(chl_fall_ran) + s(pp_sum_ran_avg) + s(mld_sum_max_avg) + s(pp_spr_ran_avg) + s(pp_sum_max_avg) + s(b_cur_ran_avg) + s(slope) + s(s_cur_ran_avg) + s(chl_spr_ran)	82.40%	1; 1; 1.9; 1; 4.7; 1; 3.7; 1; 1.8; 1; 1; 1; 3	0.36
2 (0.7 correlation with lat and long)	te(lat, long) + s(chl_fall_ran) + s(depth) + s(M_chl_fall_min) + s(slope) + s(chl_spr_ran)	87.90%	63.6, 1; 1.7; 1; 1; 2.6	0.58

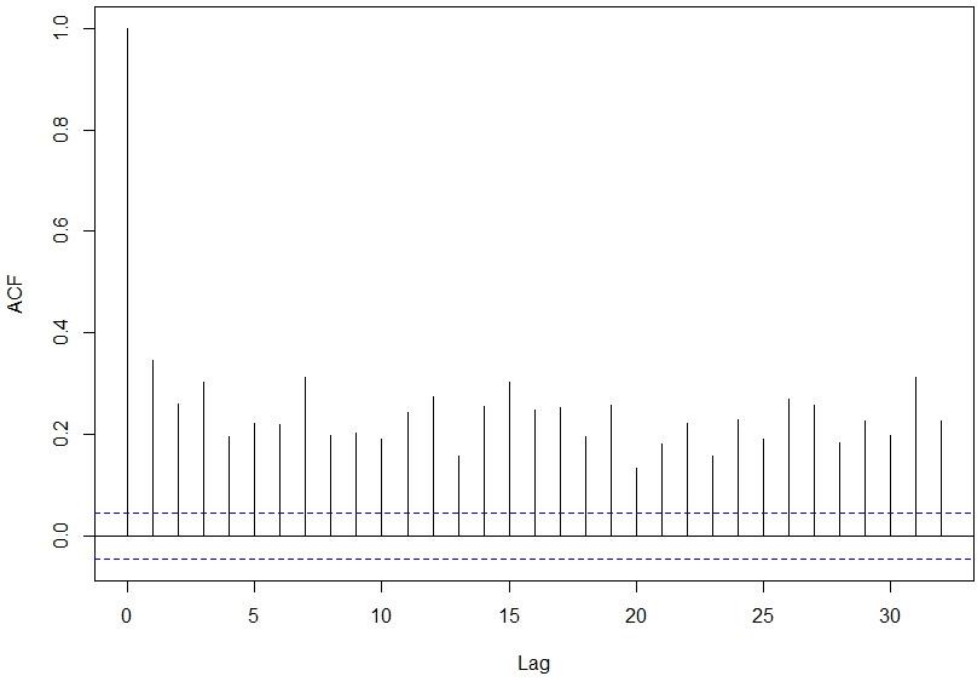

Figure 31. Autocorrelation for residuals of GAM Model 1

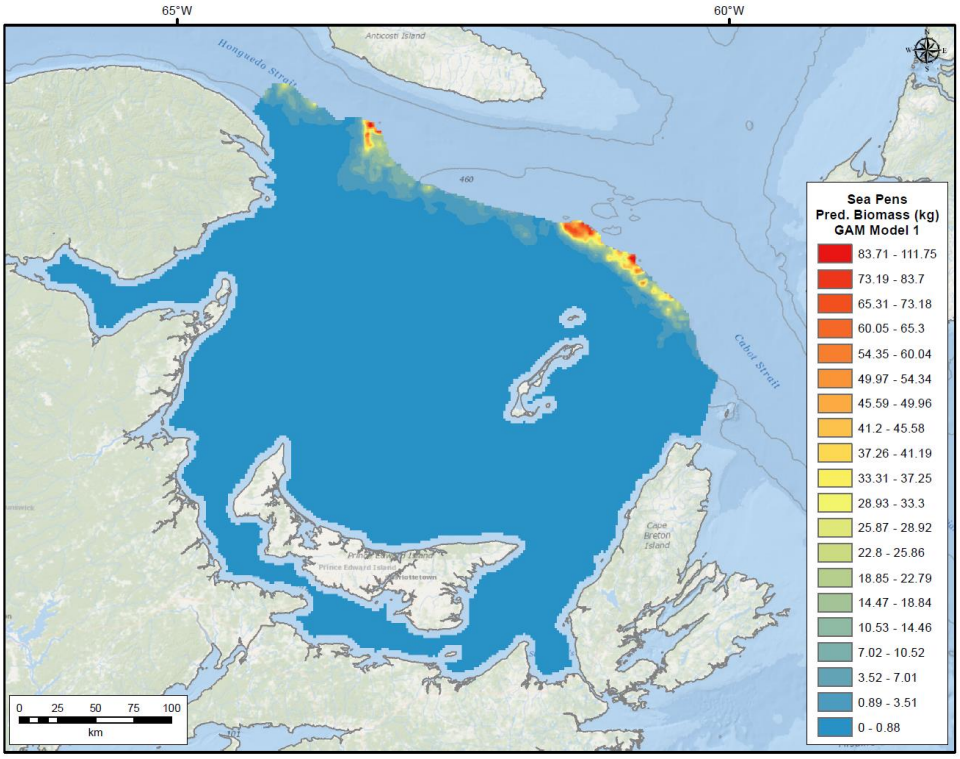

Figure 32. Prediction of sea pen biomass (kg) from the GAM Model 1 (see Table 8) within the southern Gulf of St. Lawrence.

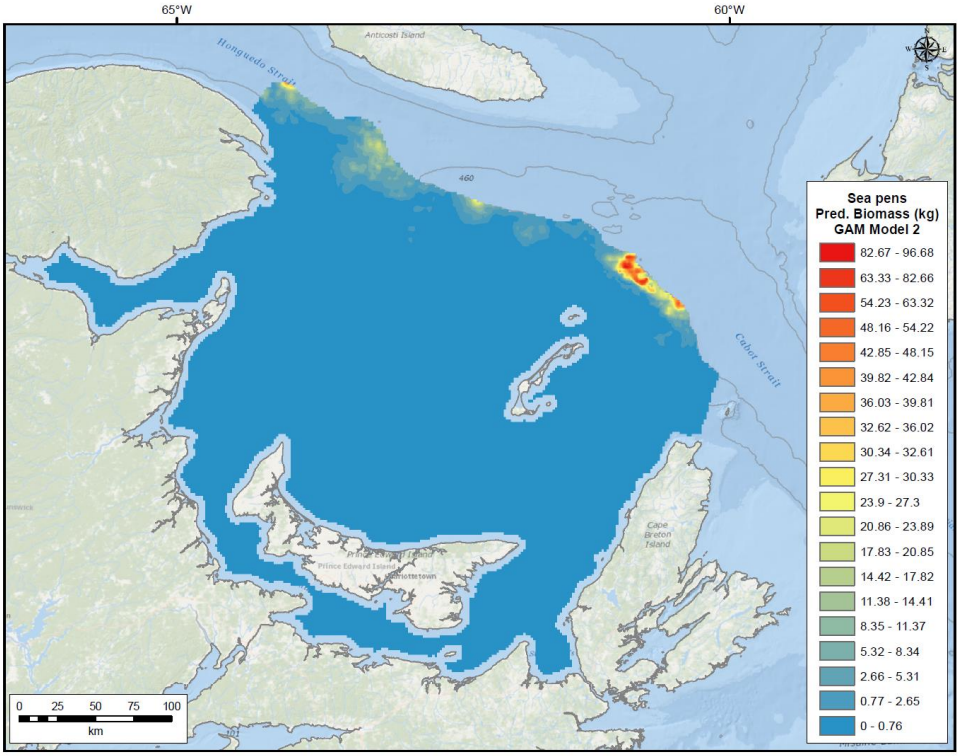

Figure 33. Prediction of sea pen biomass (kg) from the GAM Model 2 (see Table 8) within the southern Gulf of St. Lawrence.

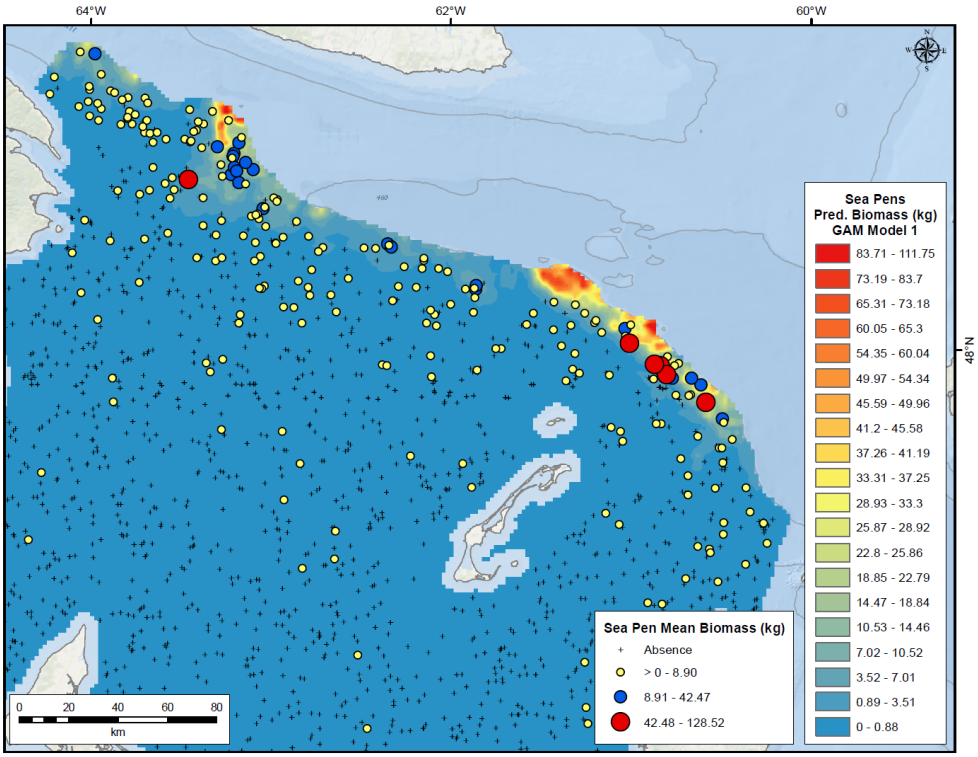

Figure 34. Sea pen biomass catches overlaid to the sea pen biomass (kg) predicted from the GAM Model 1 (see Table 8) within the southern Gulf of St. Lawrence.

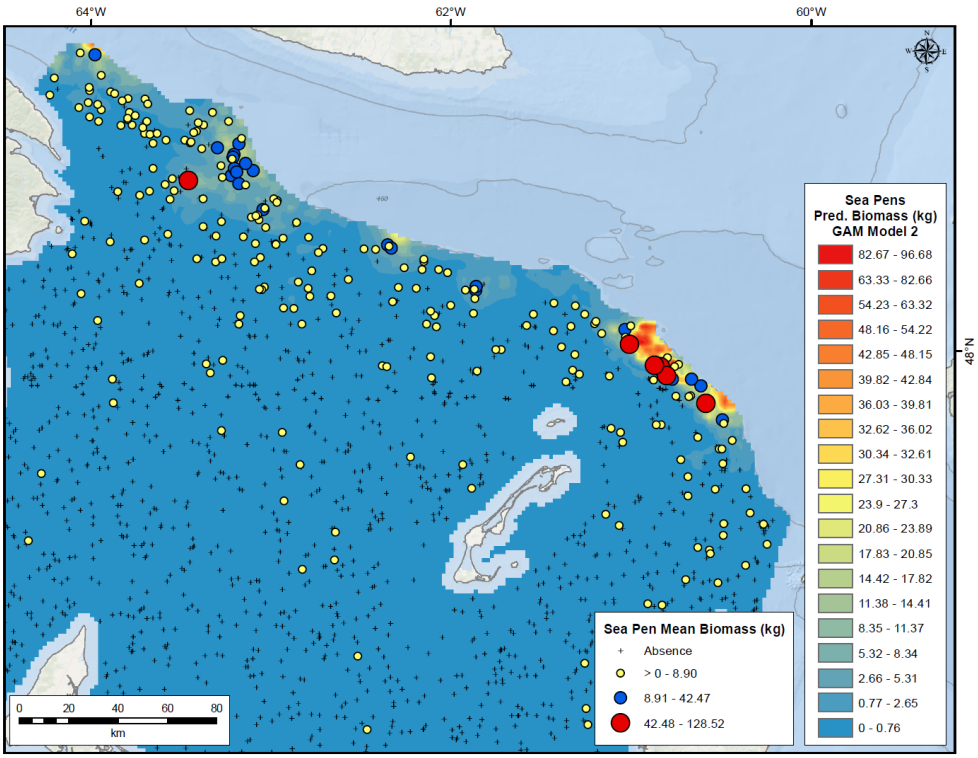

Figure 35. Sea pen biomass catches overlaid to the sea pen biomass (kg) predicted from the GAM Model 2 (see Table 8) within the southern Gulf of St. Lawrence.

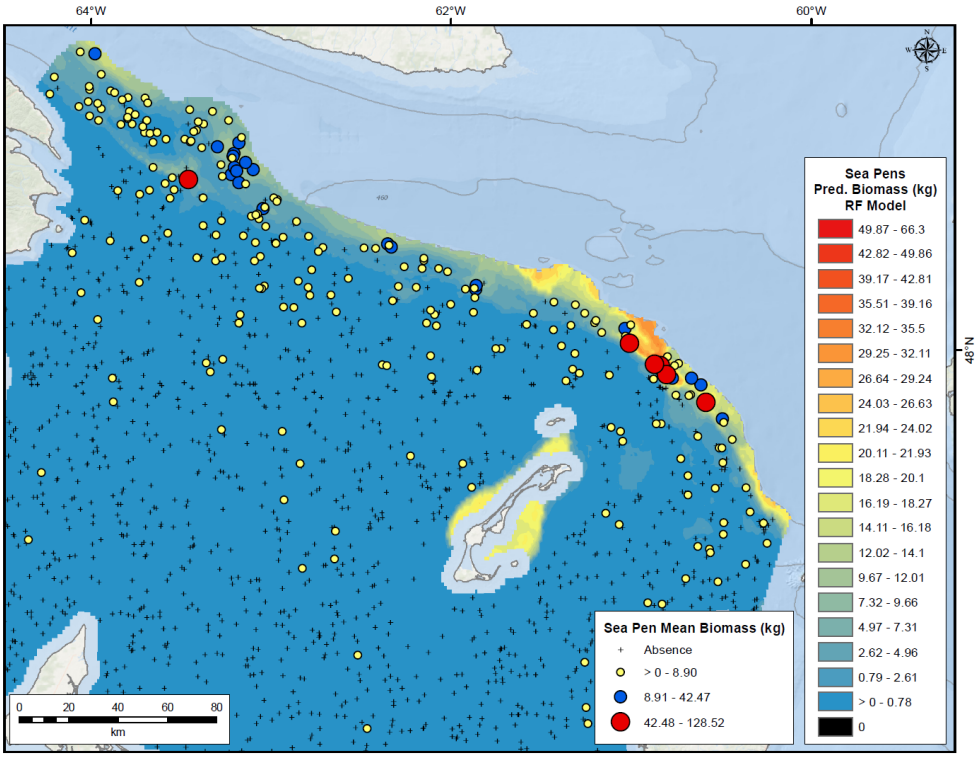

Figure 36. Sea pen biomass catches overlaid to the sea pen biomass (kg) predicted from the RF model (see Table 6) within the southern Gulf of St. Lawrence.

a) Northern Gulf of St. Lawrence

The GAM results on mean sea pen biomass per grid cell for the northern Gulf of St. Lawrence are presented in Table 9. The maximum R-square (R-sq) value was 0.33 for the model with environmental variables correlated less than 0.7 and including the latitude and longitude. Figure 37 shows that the autocorrelation of the residuals for Model 3 when latitude and longitude were not considered in the model was significant and therefore latitude and longitude were considered as a tensor product for Model 4.

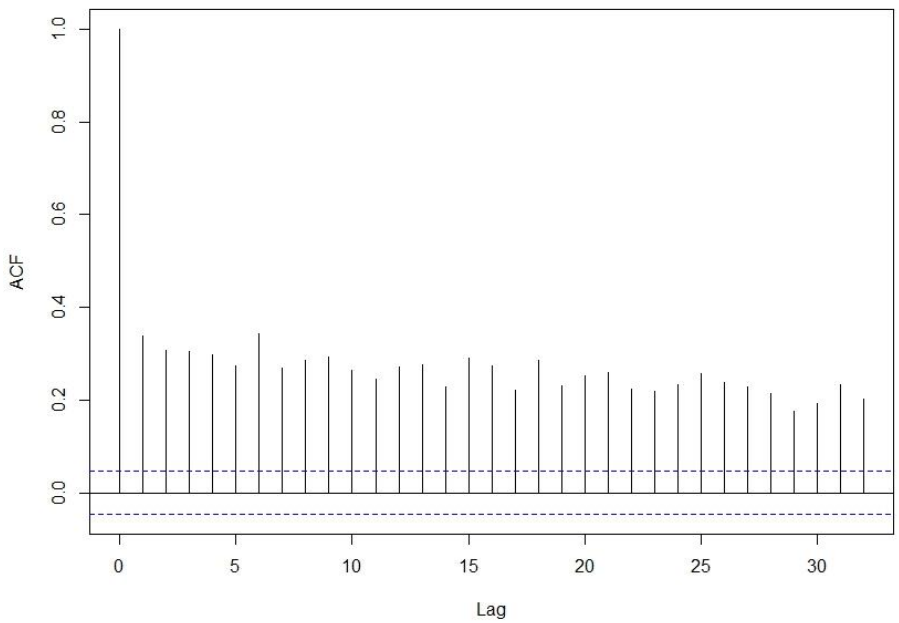

Figure 37. Autocorrelation for residuals of GAM Model 3.

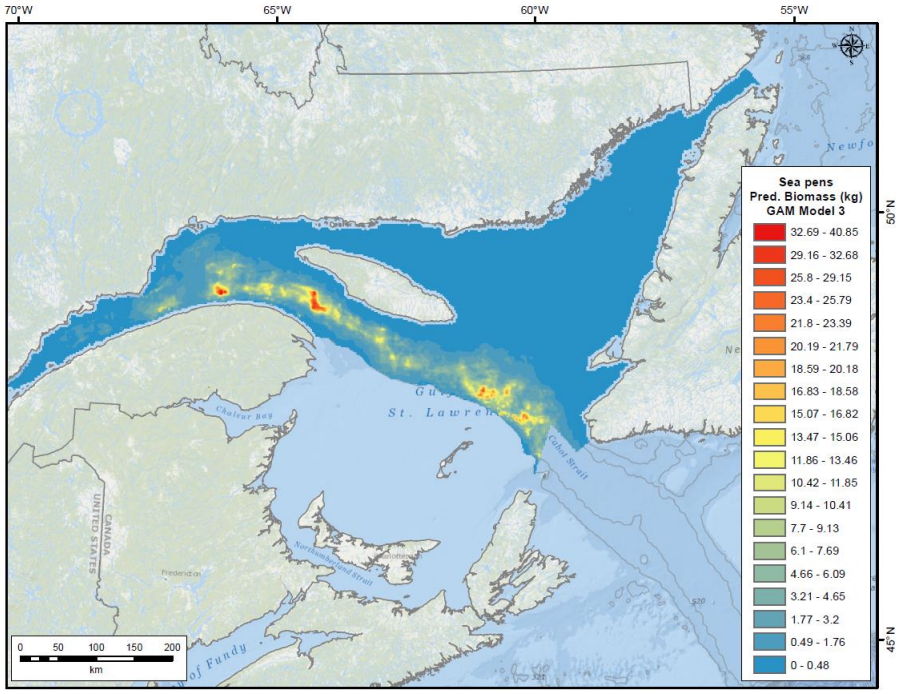

Figure 38. Prediction of sea pen biomass (kg) from the GAM Model 3 (see Table 9) within the northern Gulf of St. Lawrence.

Table 9. GAM results. Significant variables ($p < 0.05$) are listed in order of their significance (highest to lowest) in the model, the deviance explained and the estimated degrees of freedom (edf) and the R^2 are also provided. Variables are Depth, Surface Temperature Average Maximum (s_tmp_max_avg), Surface Temperature Average Range (s_tmp_ran_avg), Bottom Salinity Average Range (b_sal_ran_avg), Surface Salinity Average Range (s_sal_ran_avg), Summer Primary Production Average Range (pp_sum_ran_avg), Surface Current Average Minimum (s_cur_min_avg), MODIS Case I Spring Chrolophyll *a* Mean (chl_spr_mean), Surface Temperature Mean (s_tmp_mean), MERIS Case II Summer Chlorophyll *a* Maximum (M_chl_sum_max), Bottom Salinity Average Maximum (b_sal_max_avg), MERIS Case II Summer Chlorophyll *a* Range (M_chl_sum_ran), Slope, Surface Temperature Average Minimum (s_tmp_min_avg), Spring Primary Production Average Minimum (pp_spr_min_avg), Anuual Primary Production Average Minimum (pp_ann_min_avg), Bottom Temperature Average Maximum (b_tmp_max_avg), Summer Primary Production Average Maximum (pp_sum_max_avg), Bottom Current Average Range (b_cur_ran_avg), Maximum Summer Mixed Layer Depth (mld_sum_max_avg), Spring Primary Production Average Range (pp_spr_ran_avg), Summer Primary Production Average Minimum (pp_sum_min_avg), Bottom Temperature Average Range (b_tmp_ran_avg), MODIS Case I Fall Chrolophyll *a* Range (chl_fall_ran), MODIS Case I Spring Chrolophyll *a* Range (chl_spr_ran), location (lat, long).

Model	Significant variables	Deviance explained	edf	R^2
1 15 RF Biomass	s(depth) + s(s_tmp_max_avg) + s(s_tmp_ran_avg) + s(b_sal_ran_avg) + s(s_sal_ran_avg) + s(pp_sum_ran_avg) + s(s_cur_min_avg) +s(chl_spr_mean) +s(s_tmp_mean) + s(M_chl_sum_max) + s(b_sal_max_avg) + s(M_chl_sum_ran) +s(slope)	64.40%	1; 4.3; 1; 2; 4.8; 4.6; 4.6; 4.5; 3.8; 3.8; 1.9; 3.5; 1	0.01
2 10 RF Biomass	s(depth) + s(s_tmp_max_avg) + s(b_sal_min_avg) + s(s_tmp_ran_avg) + s(s_sal_ran_avg) + s(M_chl_sum_max) + s(chl_spr_ran) + s(s_tmp_mean) + s(M_chl_sum_ran) + s(s_cur_min_avg)	60.90%	1; 4.6; 1.9; 1; 4.6; 3.8; 3.3; 4.2; 3.1; 4.3	-0.06
3 0.7 correlation	s(depth) + s(s_tmp_min_avg) + s(slope) + s(pp_spr_min_avg) + s(pp_ann_min_avg) + s(pp_sum_ran_avg) + s(b_tmp_max_avg) + s(pp_sum_max_avg) + s(b_cur_ran_avg) + s(mld_sum_max_avg)+ s(pp_spr_ran_avg) + s(s_cur_min_avg) + s(M_chl_sum_ran) + s(pp_sum_min_avg) + s(b_tmp_ran_avg) + s(chl_fall_ran) + s(chl_spr_ran)	61.30%	1.5; 4.9; 1.9; 1; 1.8; 4.6; 1.9; 1.9; 1.9; 4.6; 4.6; 3.3; 2.5; 2.5; 1.9; 3.7; 3.1	0.1
4 0.7 correlation with lat and long	te(x,y) + s(depth) + s(b_tmp_ran_avg) + s(s_cur_min_avg) + s(pp_sum_ran_avg) + s(M_chl_sum_ran)	76.20%	115; 1; 1; 4; 4; 3	0.33

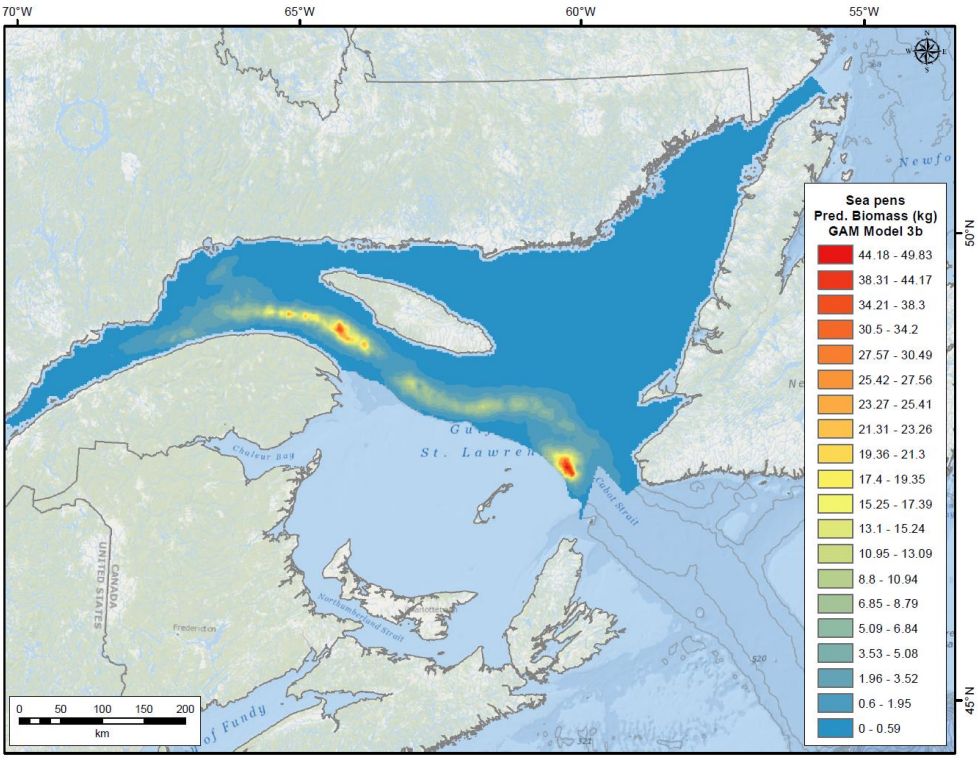

Figure 39. Prediction of sea pen biomass (kg) from the GAM Model 4 (see Table 9) within the northern Gulf of St. Lawrence.

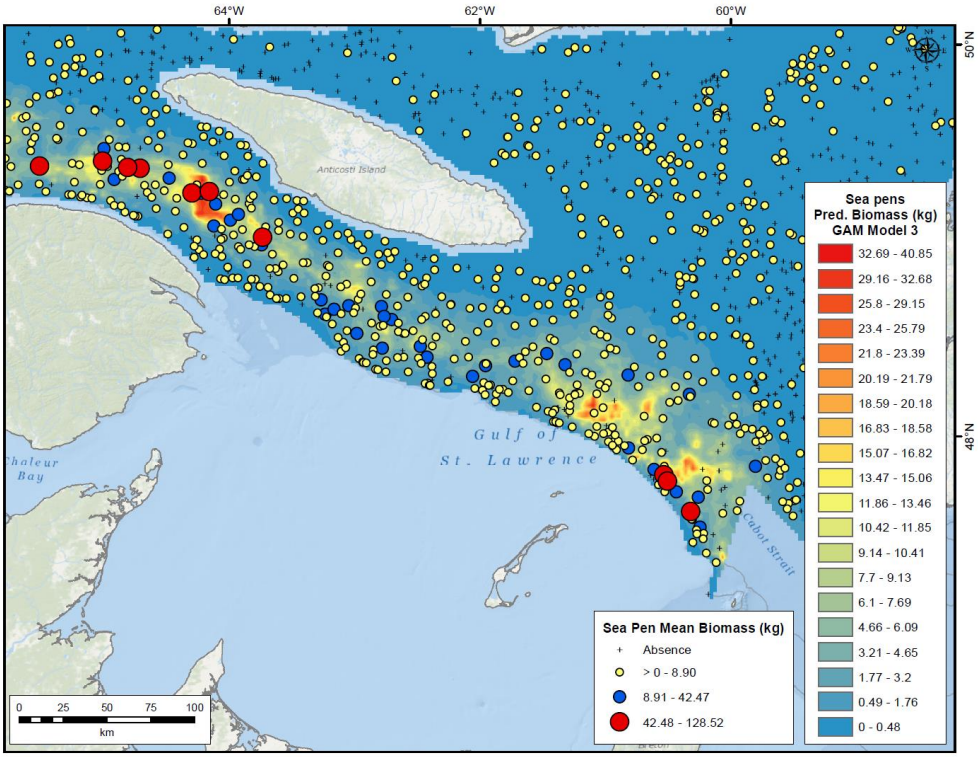

Figure 40. Sea pen biomass catches overlaid to the sea pen biomass (kg) predicted from the GAM Model 3 (see Table 9) within the northern Gulf of St. Lawrence.

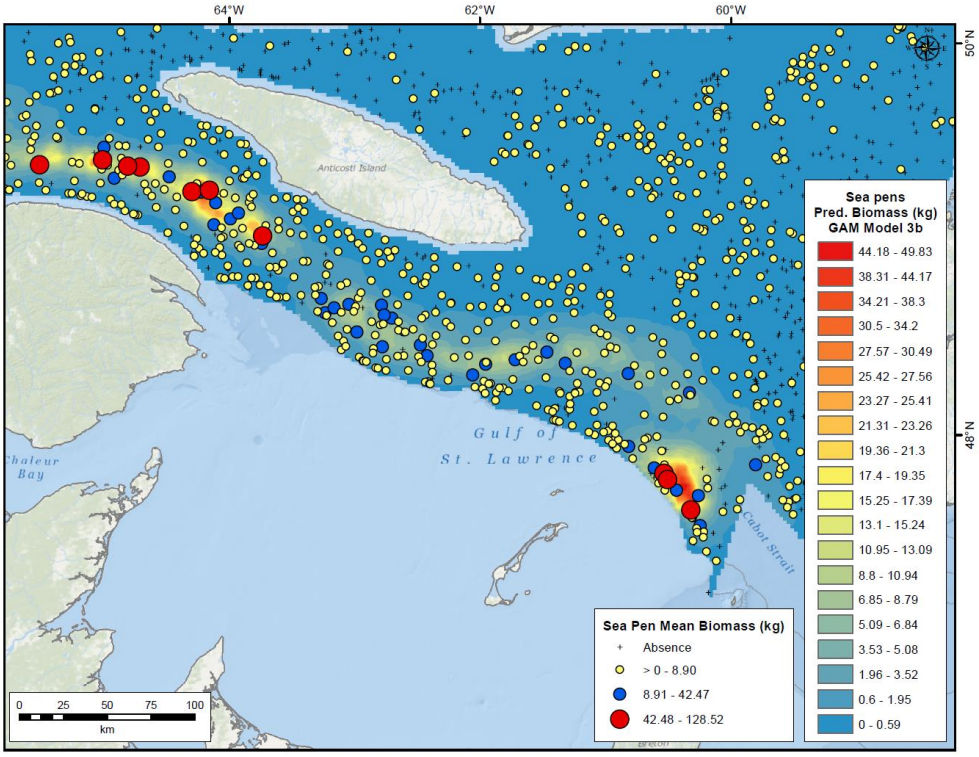

Figure 41. Sea pen biomass catches overlaid to the sea pen biomass (kg) predicted from the GAM Model 4 (see Table 9) within the northern Gulf of St. Lawrence.

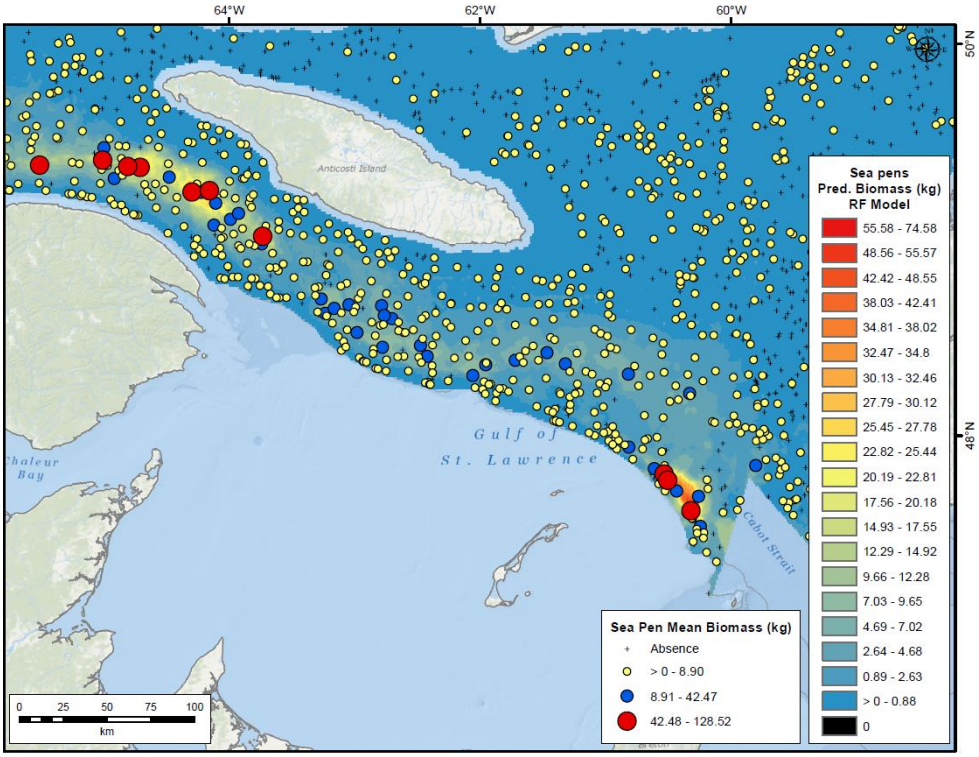

Figure 42. Sea pen biomass catches overlaid to the sea pen biomass (kg) predicted from the RF model (see Table 7) within the northern Gulf of St. Lawrence.

Comparison between GAM and RF

RF models performed better (in terms of R^2) than GAM models in the northern and southern Gulf of St. Lawrence, except when the latitude and longitude was considered as a tensor product in the GAM model. A RF model for the northern Gulf with the environmental variables correlated less than 0.7 was created but it did not improve the performance than when all the predictors were consider (R^2=0.22 vs 0.27). These models were tested in the southern Gulf and the highest R^2 (0.13) was obtained with the GAM Model 3 (predictors correlated less than 0.7), followed by GAM Model 4 (R^2=0.10). The RF with the predictors correlated less than 0.7 gave better performance than when all the predictors were considered in the model (R^2=0.01 vs 0.006).

Sponges (Porifera)

Data Sources and Distribution

The data available for analysis sponges in the Gulf Region was collected over a period of 13 years from 2003 to 2015 (Table 10). The data were filtered so that only one presence or absence record occurs within the same environmental data raster cell (approximately 1 km^2; with presence records taking precedence over an absence record when both occurred in the same cell), giving a total of 1125 presence and 709 absence records for the southern Gulf and 1089 presence and 569 absence records for the northern Gulf. Sponge biomass (kg) was averaged across each catch occurring within the same cell. Sponges presences are distributed practically across all the Region, with some absence locations along the Laurentian Channel (Figure 44). Two areas with higher biomass are found in shallow depths on the southern Gulf and on the mouth of the Saint Lawrence River.

Table 10. Total number of presences and absences of sponges by-catch data recorded from DFO trawl surveys from 2003 to 2015 in the Gulf Region. P, presences; A, absences.

	Southern Gulf		**Northern Gulf**	
Year	P	A	P	A
2003	53	27	-	-
2004	130	83	-	-
2005	120	99	-	-
2006	117	47	89	92
2007	104	61	94	83
2008	104	71	139	50
2009	107	41	115	55
2010	82	52	106	31
2011	78	47	113	39
2012	71	69	114	59
2013	80	44	88	53
2014	79	68	119	48
2015	-	-	112	59
TOTAL	**1125**	**709**	**1089**	**569**

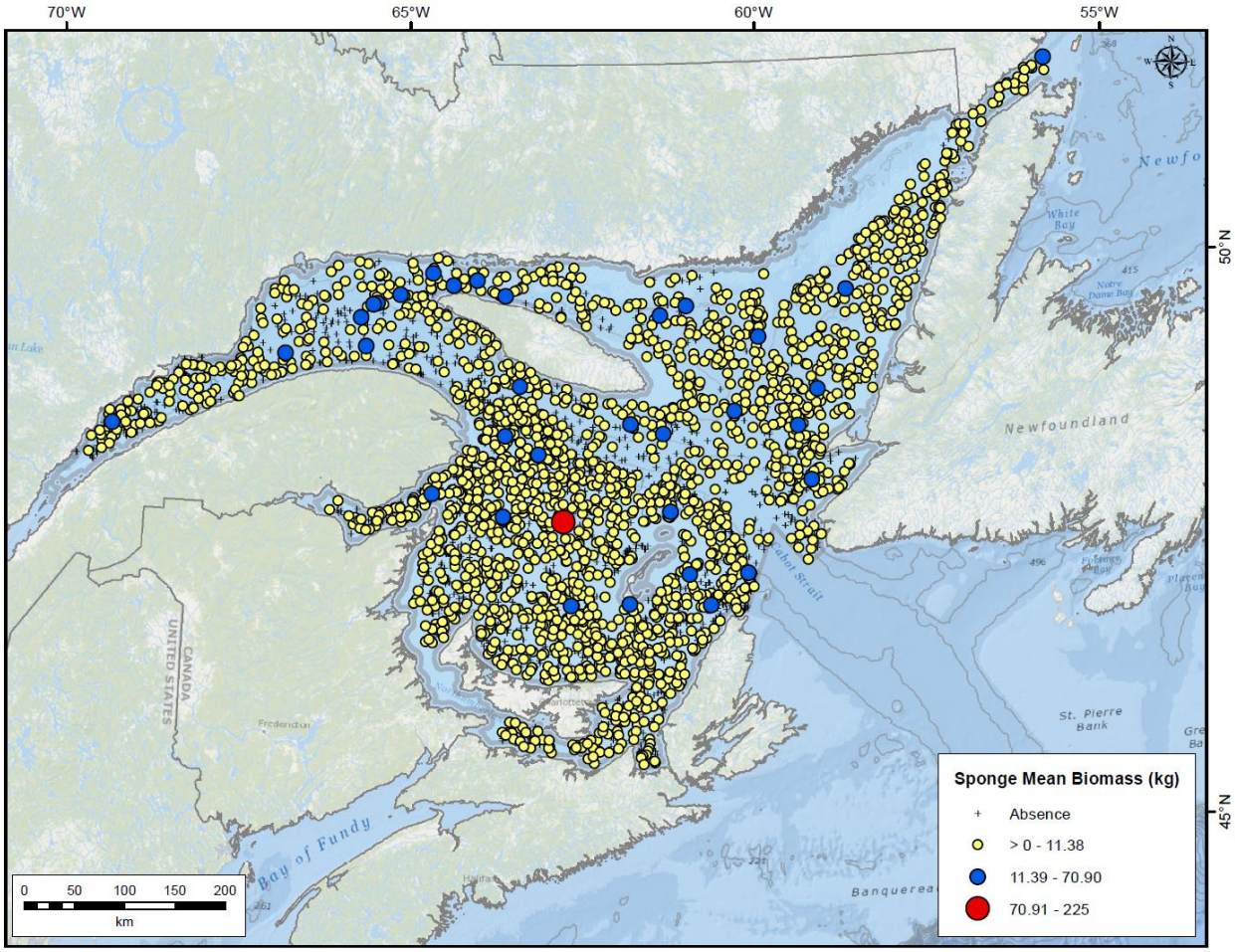

Figure 44. Mean biomass (kg) per grid cell of sponge by-catches recorded from DFO groundfish trawl surveys from 2003 to 2015 within the Gulf Region.

Model 1 – Unbalanced species prevalence

Table 11 shows the accuracy measures for the random forest model on sponge presence and absence data (2214 presences and 1278 absences) and a threshold equal to species prevalence (0.63). The AUC calculated was 0.708. Class error of the presence and absence classes were relatively higher, 0.354 and 0.338, respectively. Sensitivity and specificity were low, 0.646 and 0.662, respectively.

Table 11. Accuracy measures and confusion matrix from 10-fold cross validation from random forest modelling of presence and absence of sponges within the Gulf Region with a cut-off threshold of 0.63.

Observations	Predictions		Total n	Class error	AUC	Sensitivity	Specificity
	Absence	Presence					
Absence	846	432	1278	0.338	0.708	0.646	0.662
Presence	783	1431	2214	0.354			

The predicted presence probability surface of sponges generated is shown in Figure 45. Most of the Region presents high probability, mainly the northern Gulf of St. Lawrence. Lower probability was found along the Laurentian Channel.

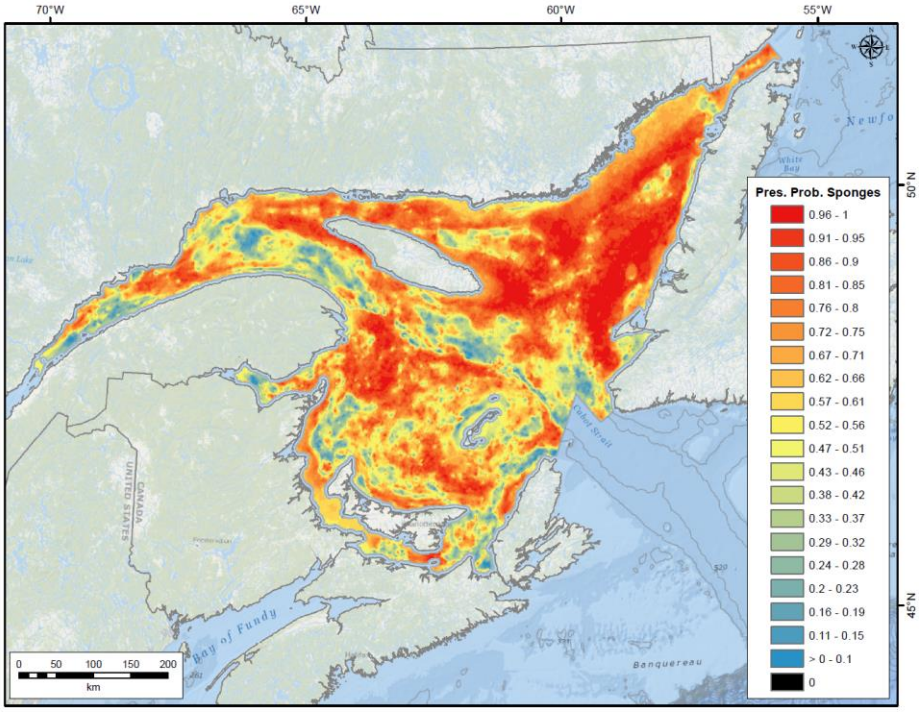

Figure 45. Predictions of presence probability (pres. prob.) of sponges from the unbalanced random forest model (see Table 11).

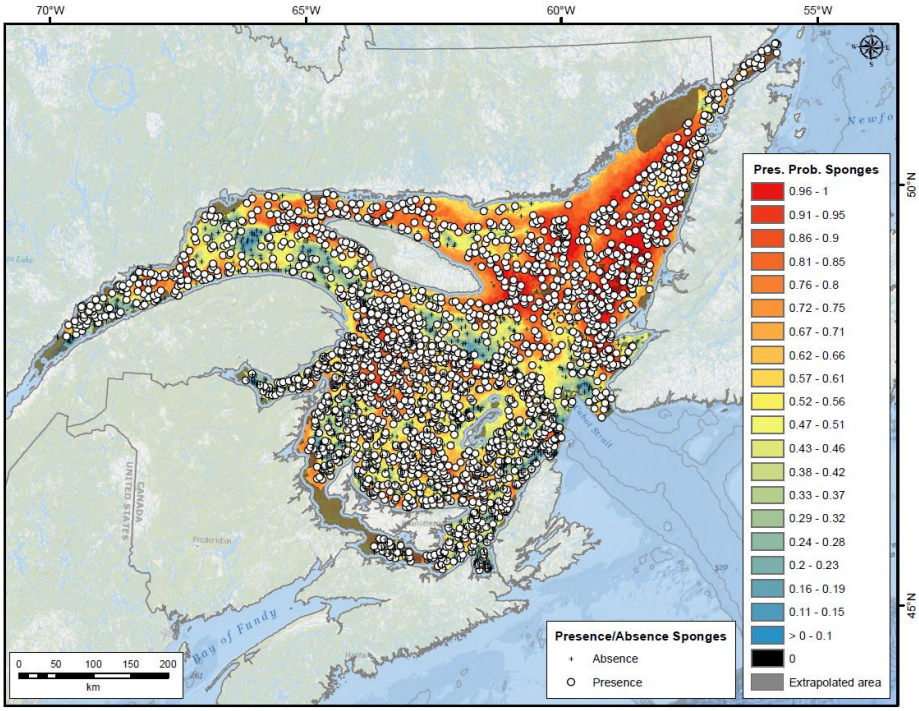

Figure 46. Presence and absence observations and predictions of presence probability (pres. prob.) of sponges from the unbalanced random forest model (see Table 11). The grey areas of extrapolation are also shown.

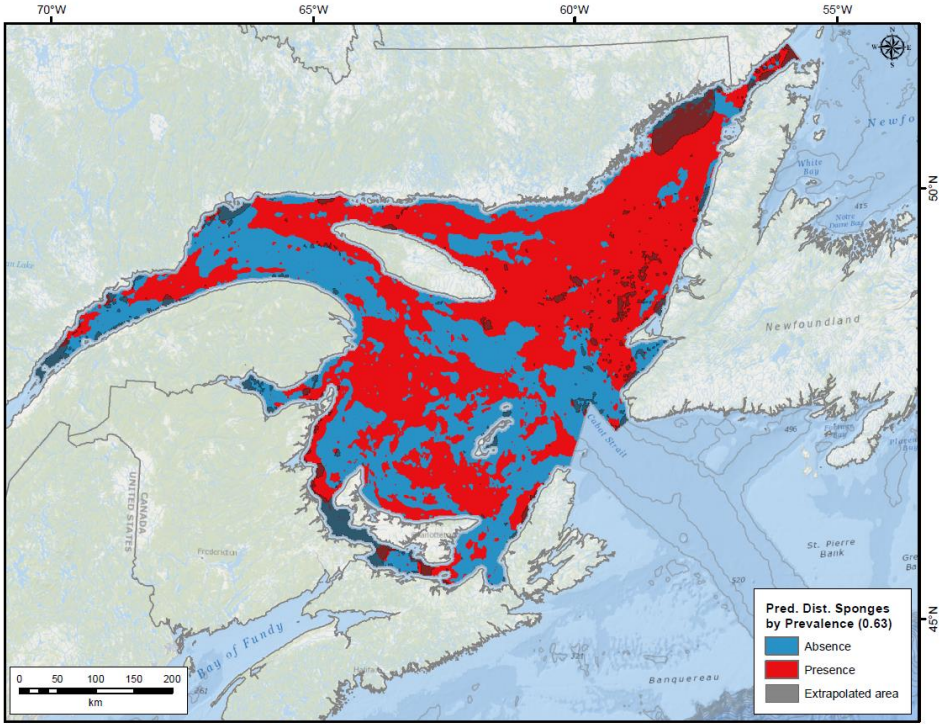

Figure 47. Predicted distribution of sponges from the unbalanced random forest model (see Table 11) using the prevalence cut-off threshold of 0.63. The grey areas of extrapolation are also shown, they appear dark red when overlain on the red presence surface.

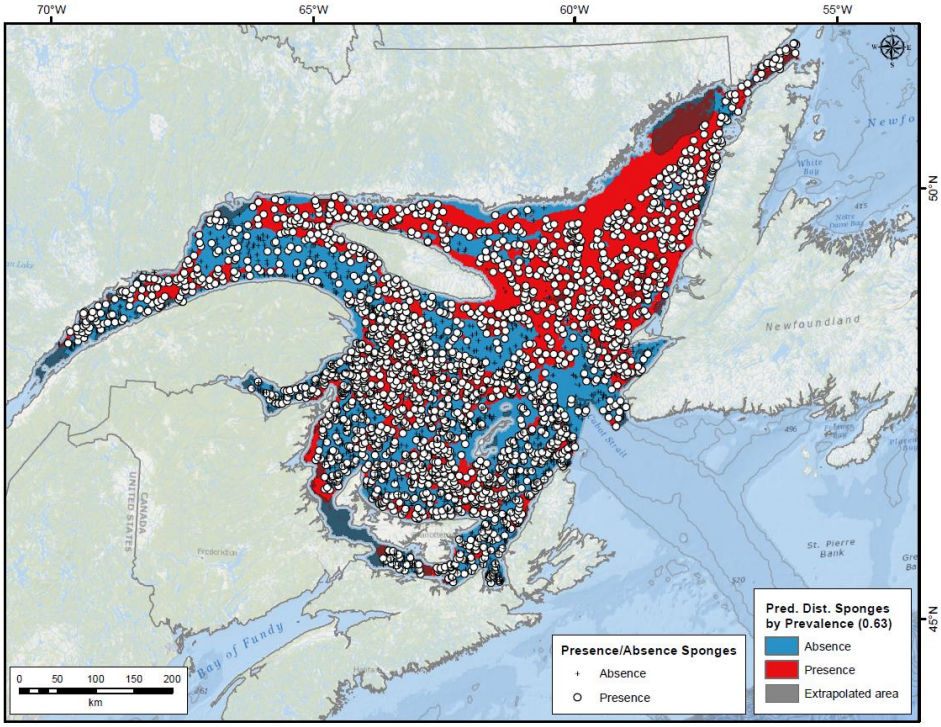

Figure 48. Presence and absence observations and predicted distribution of sponges from the unbalanced random forest model (see Table 11) using the prevalence cut-off threshold of 0.63. The grey areas of extrapolation are also shown, they appear dark red when overlain on the red presence surface.

Of all 78 environmental predictor variables used in the model, Depth was the most important for the classification of the sponge presence and absence data (Figure 49). This variable was followed more distantly in terms of its Mean Decrease in Gini Value by Slope and Surface Salinity Average Maximum. The partial dependence plots for the top 6 most important predictors are shown in Figure 50. Presence probability of sponges started to decrease below 200 m depth.

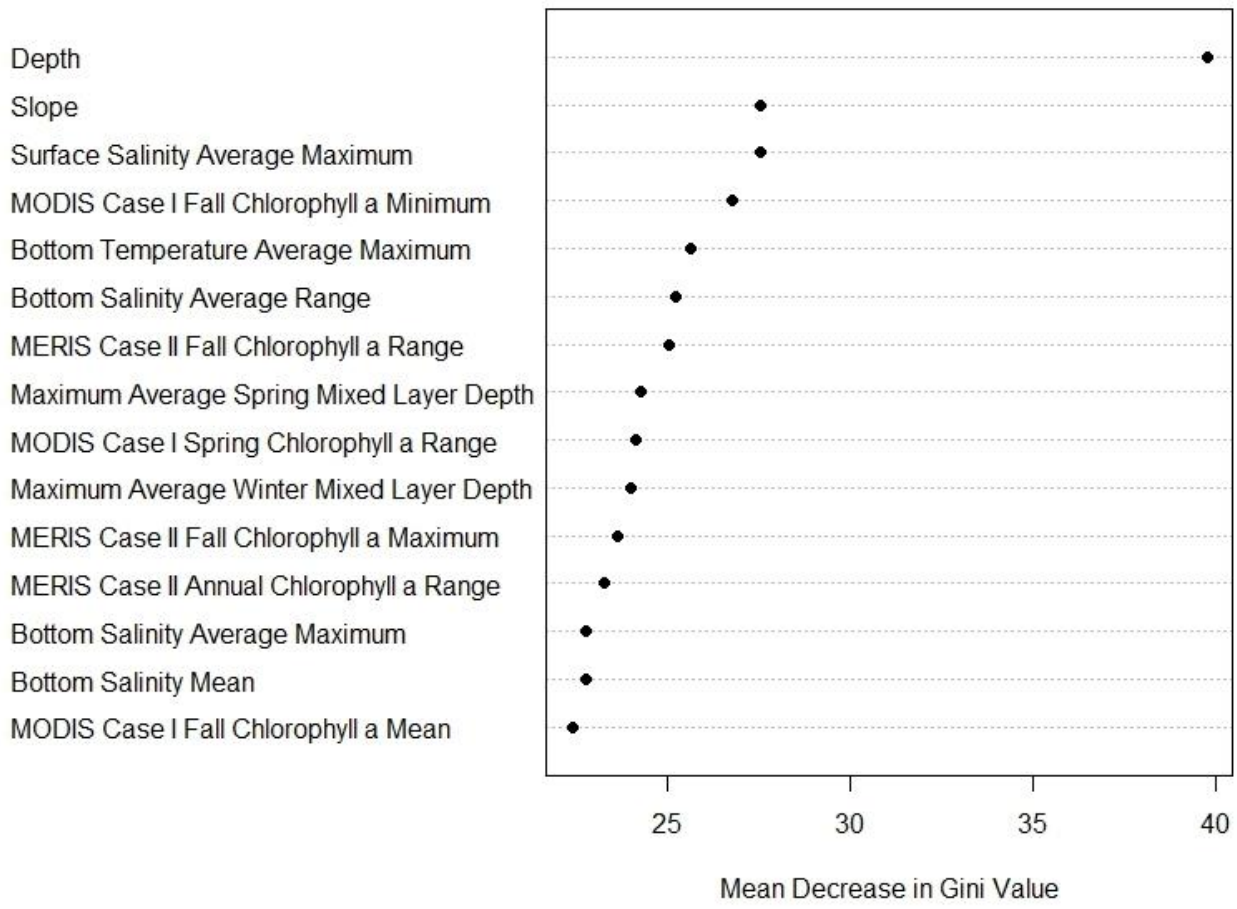

Figure 49. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the unbalanced random forest model of sponge presence and absence data within the Gulf Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

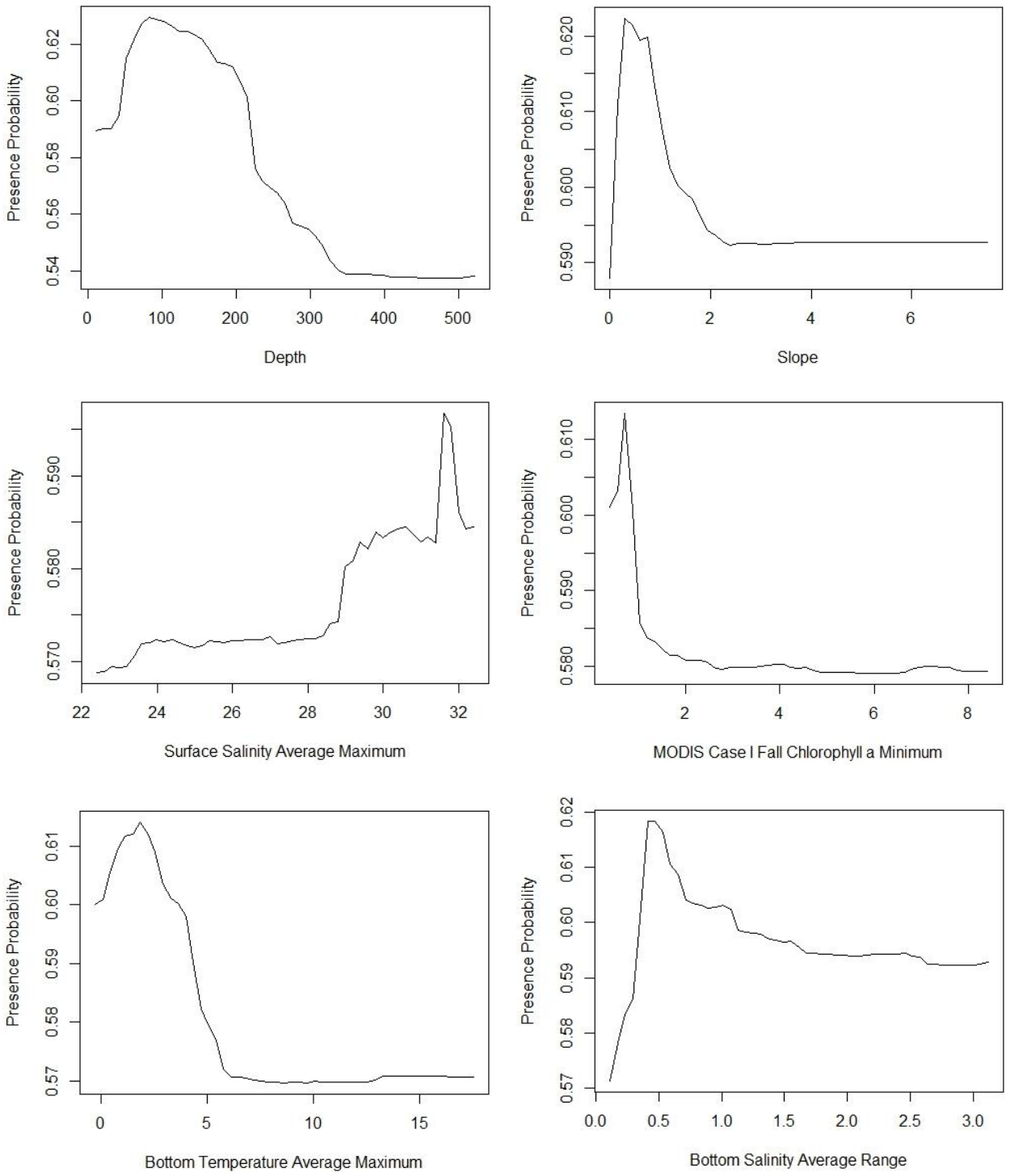

Figure 50. Partial dependence plots of the top six predictors of the unbalanced random forest model of sponges presence and absence data collected within the Gulf Region, ordered left to right from the top.

Prediction of Sponge Biomass Using Random Forest

a) Southern Gulf Region

The accuracy measures of the regression random forest model on mean sponge biomass per grid cell for the southern Gulf of St. Lawrence are presented in Table 12. The average R-square (R^2) value was 0.066 ± 0.130, the average Root-Mean-Square Error (RMSE) was 3.794 ± 4.677, and the average percent variance explained was negative. These low accuracy metrics indicate poor predictive performance of the model.

Table 12. Accuracy measures from 10-fold cross validation from random forest modelling of biomass of sponges within the southern Gulf of St. Lawrence. NRMSE, is the normalized root-mean square deviation (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	%variance explained
1	0.091	1.070	0.005	-4.2
2	0.018	1.184	0.005	-2.58
3	0.428	2.756	0.012	-9.35
4	0.004	1.619	0.007	-6.97
5	0.026	4.942	0.022	-5.7
6	0.001	4.356	0.019	-6.95
7	0.020	1.088	0.005	-4.37
8	0.005	3.180	0.014	-3.07
9	0.007	1.261	0.006	-4.29
10	0.063	16.487	0.073	-2
Mean	**0.066**	**3.794**	**0.017**	**-4.948**
SD	**0.130**	**4.677**	**0.021**	**2.290**

Figures 51 and 52 show the surface of sponge biomass (kg) predictions per grid cell generated from the random forest model in the southern Gulf of St. Lawrence. The maximum biomass was predicted in a small area in the middle of the area where a high catch of 225 kg was found.

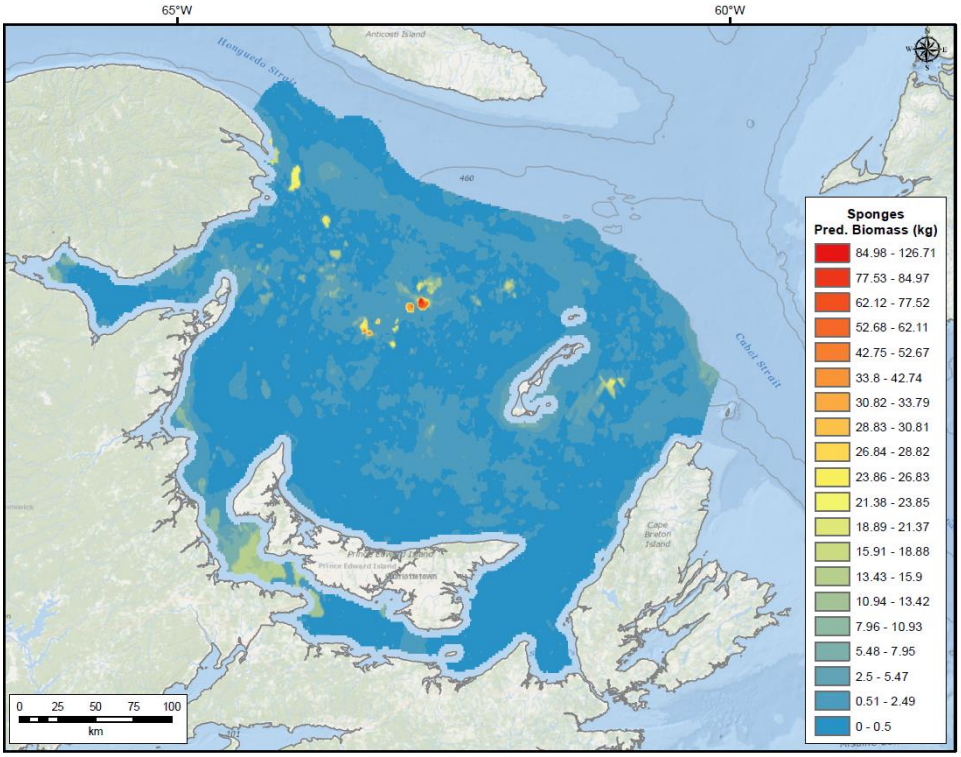

Figure 51. Prediction of sponge biomass (kg) from the random forest model (see Table 12) within the southern Gulf of St. Lawrence.

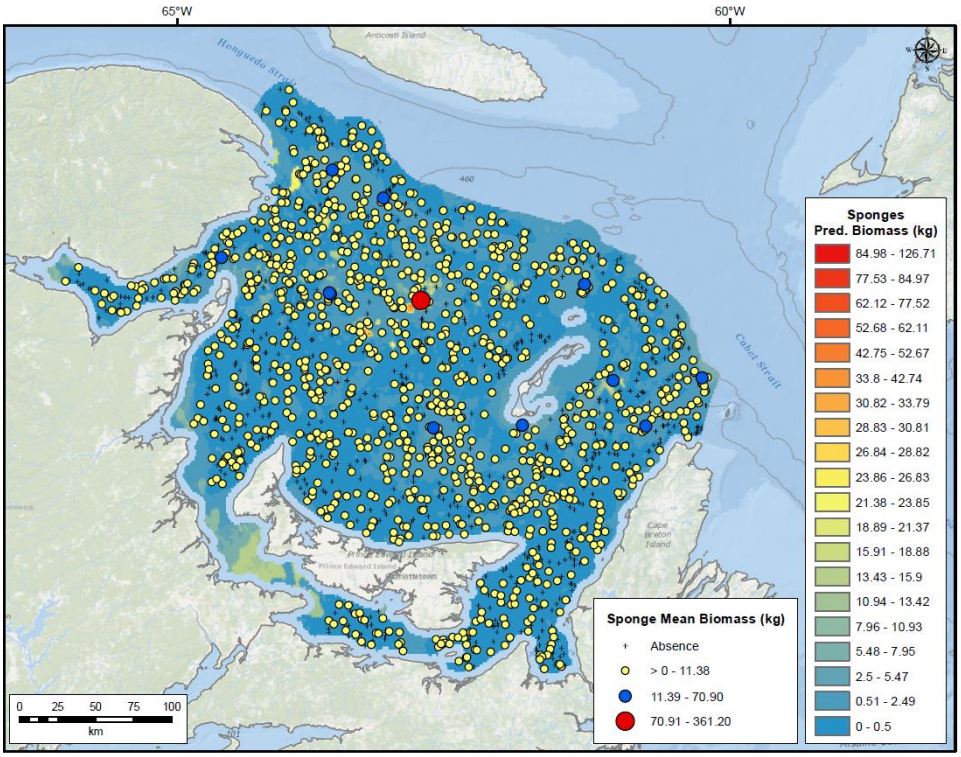

Figure 52. Sponge biomass catches overlaid to the sponge biomass (kg) predicted from the random forest model (see Table 12) within the southern Gulf of St. Lawrence.

The top 15 most important environmental variables for predicting sponge biomass in the southern Gulf of St. Lawrence are shown in Figure 53. MERIS Case II Spring Chlorophyll *a* Range was the most important variable in the model, followed by MODIS Case I Spring Chlorophyll *a* and Bottom Temperature Average Range. The partial dependence of sponge biomass on the top 6 most important variables is shown in Figure 54. MERIS Case II Spring Chlorophyll *a* Range predicted highest biomass at Chlorophyll *a* range of 2.0 mg m^{-3}.

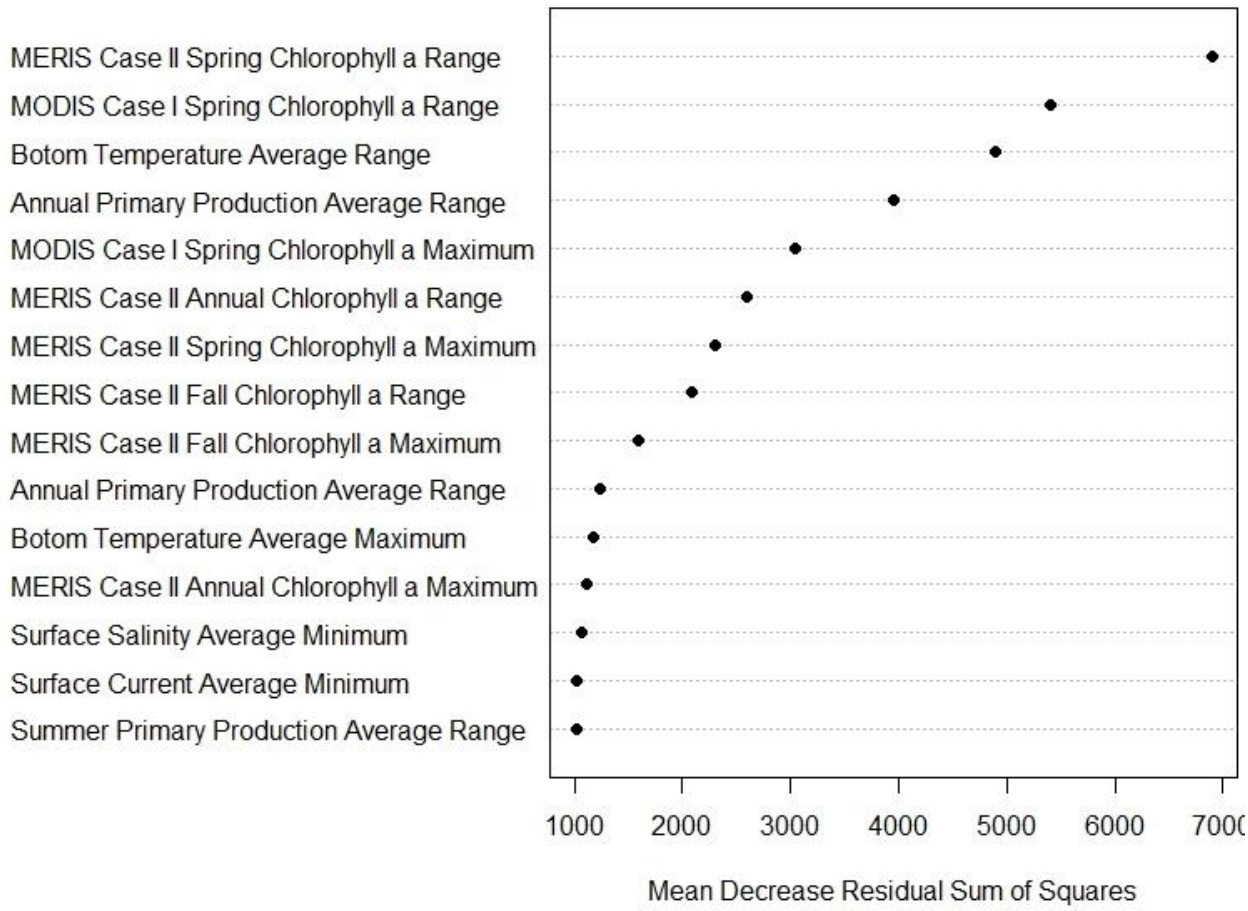

Figure 53. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the unbalanced random forest model of sponge presence and absence data within the southern Gulf of St. Lawrence. The higher the Mean Gini value the more important the variable is for predicting the response data.

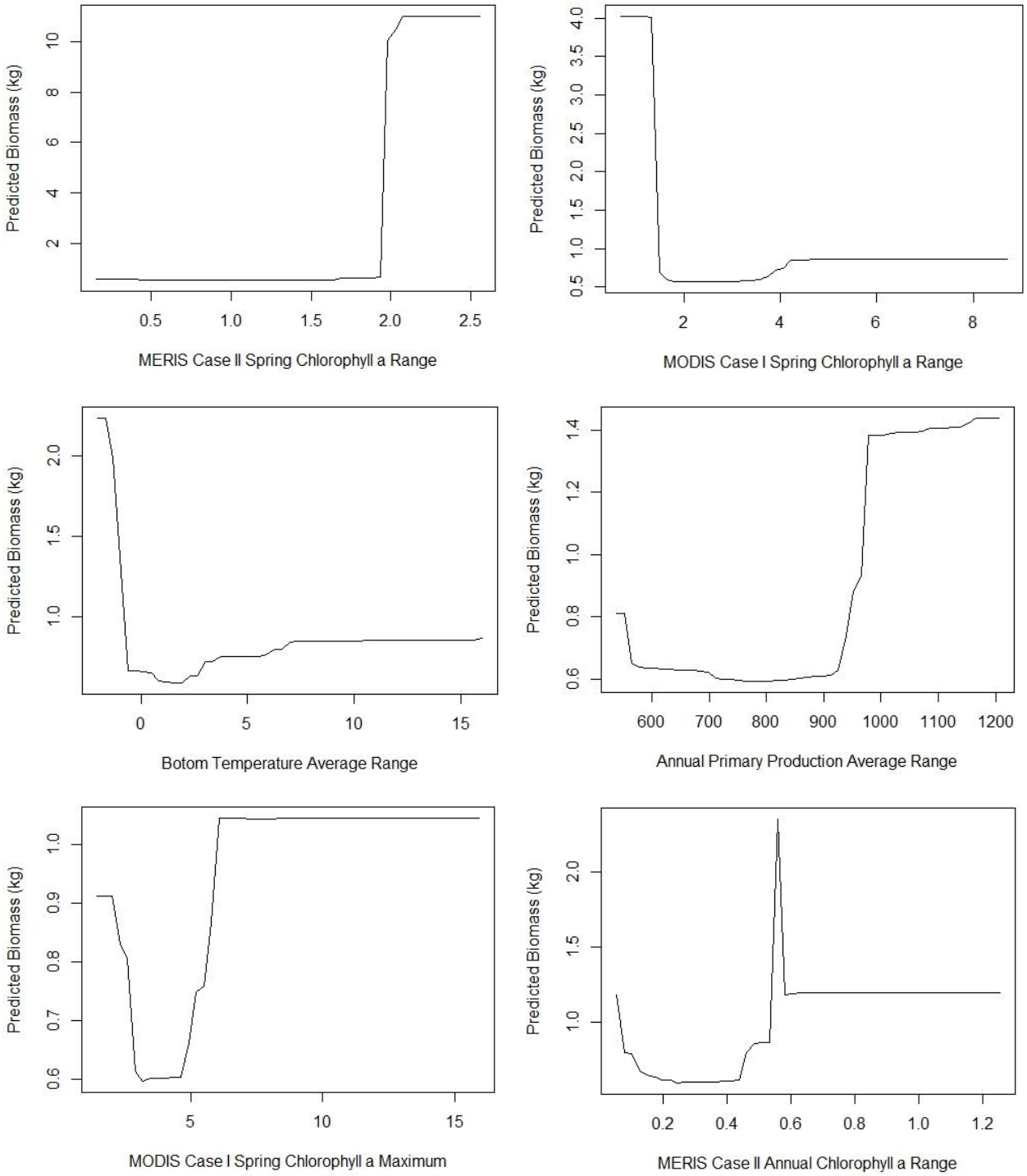

Figure 54. Partial dependence plots of the top six predictors of from the unbalanced random forest model of sponge presence and absence data collected within the southern Gulf of St. Lawrence, ordered left to right from the top.

b) Northern Gulf Region

The accuracy measures of the regression random forest model on mean sponge biomass per grid cell for the northern Gulf of St. Lawrence are presented in Table 13. The average R-square (R^2) value was 0.033 ± 0.041, the average Root-Mean-Square Error (RMSE) was 3.109 ± 1.503, and the average percent variance explained was negative. These low accuracy metrics indicate poor predictive performance of the model.

Table 13. Accuracy measures from 10-fold cross validation from random forest modelling of biomass of sponges within the northern Gulf of St. Lawrence. NRMSE, is the normalized root-mean square deviation (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	%variance explained
1	0.013	1.908	0.027	-2.28
2	0.044	1.888	0.027	-1.69
3	0.005	6.202	0.087	-0.15
4	0.023	2.574	0.036	-1.88
5	0.024	2.330	0.033	-2.67
6	0.019	4.874	0.069	-3.76
7	0.145	3.614	0.051	-4.83
8	0.024	1.749	0.025	-4.55
9	0.001	3.888	0.055	-3.17
10	0.032	2.059	0.029	-3.56
Mean	**0.033**	**3.109**	**0.044**	**-2.854**
SD	**0.041**	**1.503**	**0.021**	**1.421**

Figures 55 and 56 show the surface of sponge biomass (kg) predictions per grid cell generated from the random forest model in the northern Gulf of St. Lawrence. No large areas with high biomass were predicted, although the maximum biomass predicted can be found at the mouth of the Saint Lawrence River.

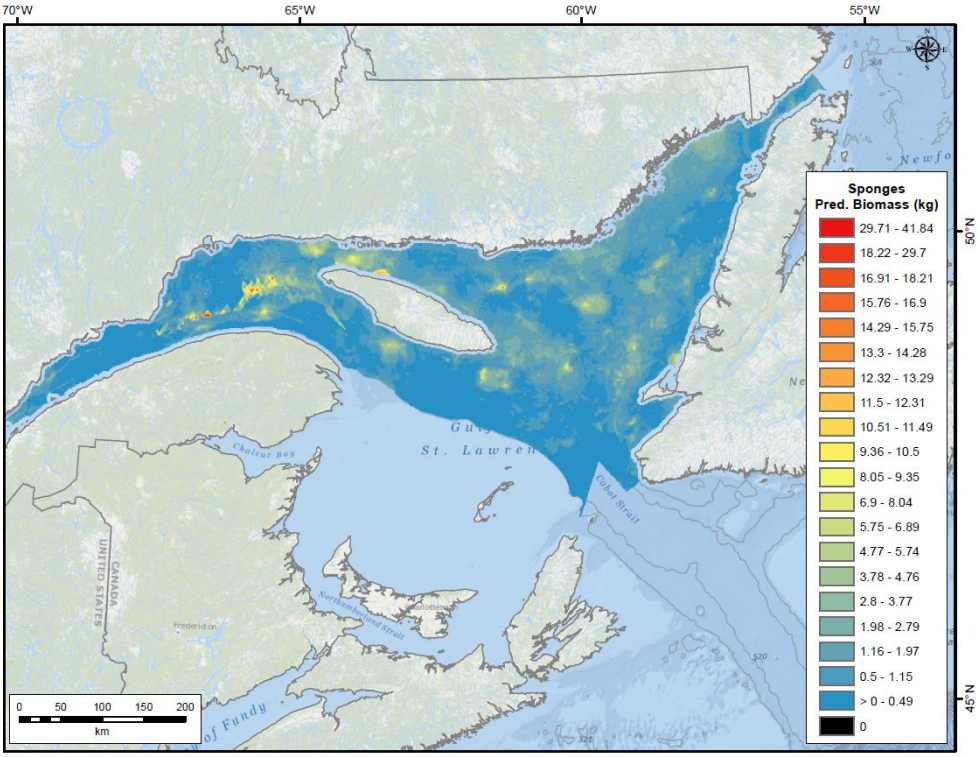

Figure 55. Predictions of sponge biomass (kg) from the random forest model (see Table 13) within the northern Gulf of St. Lawrence.

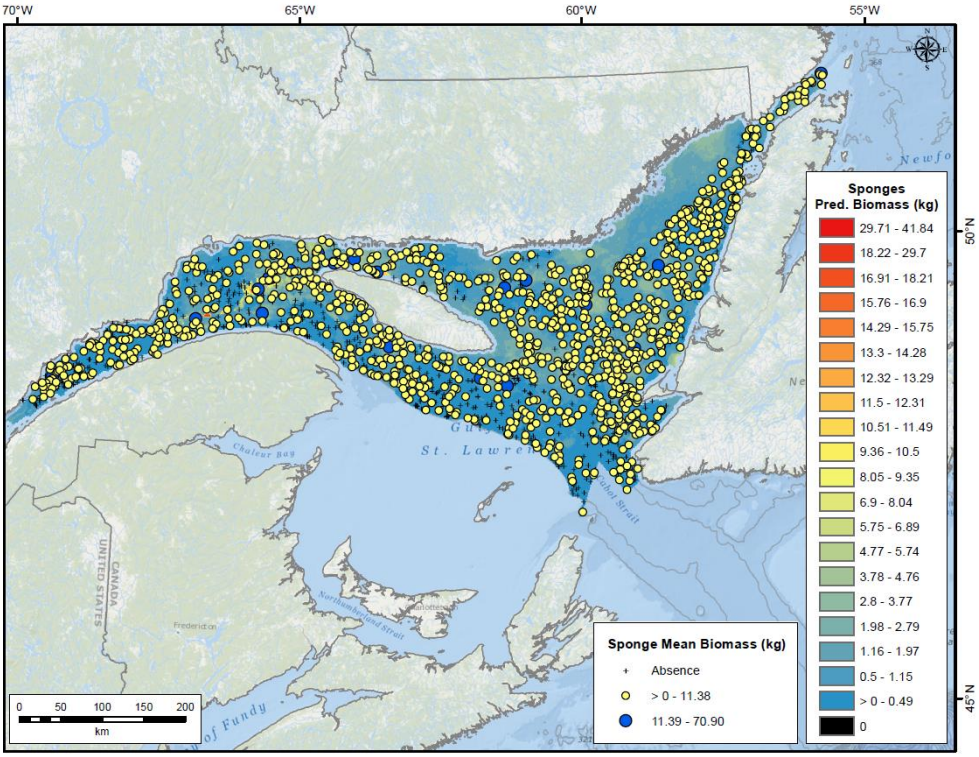

Figure 56. Sponge biomass catches overlaid to the sponge biomass (kg) predicted from the random forest model (see Table 13) within the northern Gulf of St. Lawrence.

The top 15 most important environmental variables for predicting sponge biomass in the northern Gulf of St. Lawrence are shown in Figure 57. Maximum Average Summer Mixed Layer Depth was the most important variable in the model, followed distantly by Annual Primary Production Average Maximum. The partial dependence of sponge biomass on the top 6 most important variables is shown in Figure 58. Biomass (kg) of sponges predicted decreased at Maximum Average Summer Mixed Layer Depth of 11 m.

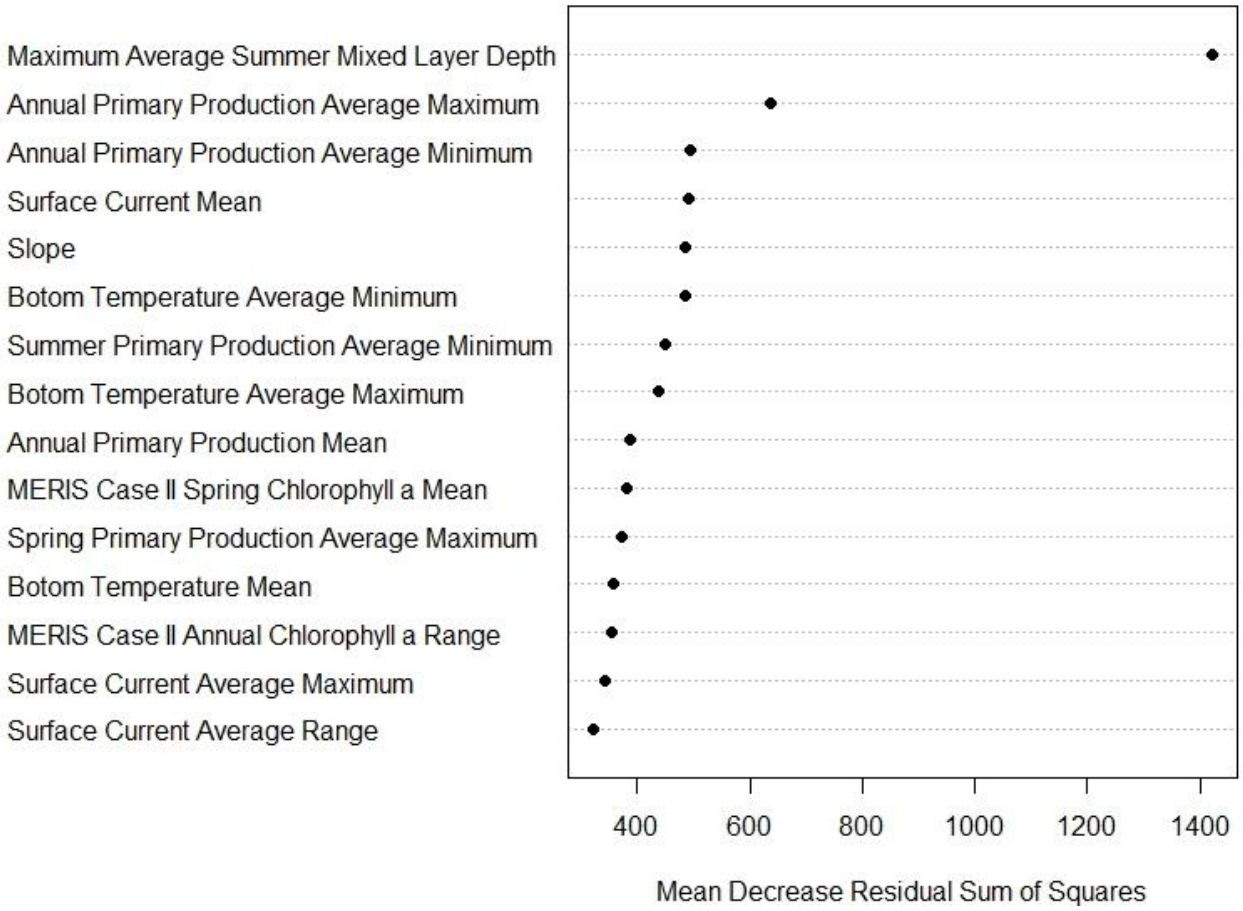

Figure 57 Importance of the top 15 predictor variables measured as the Mean Decrease Residual Sum of Squares of the random forest model (see Table 13) of sponge biomass data within the northern Gulf of St. Lawrence.

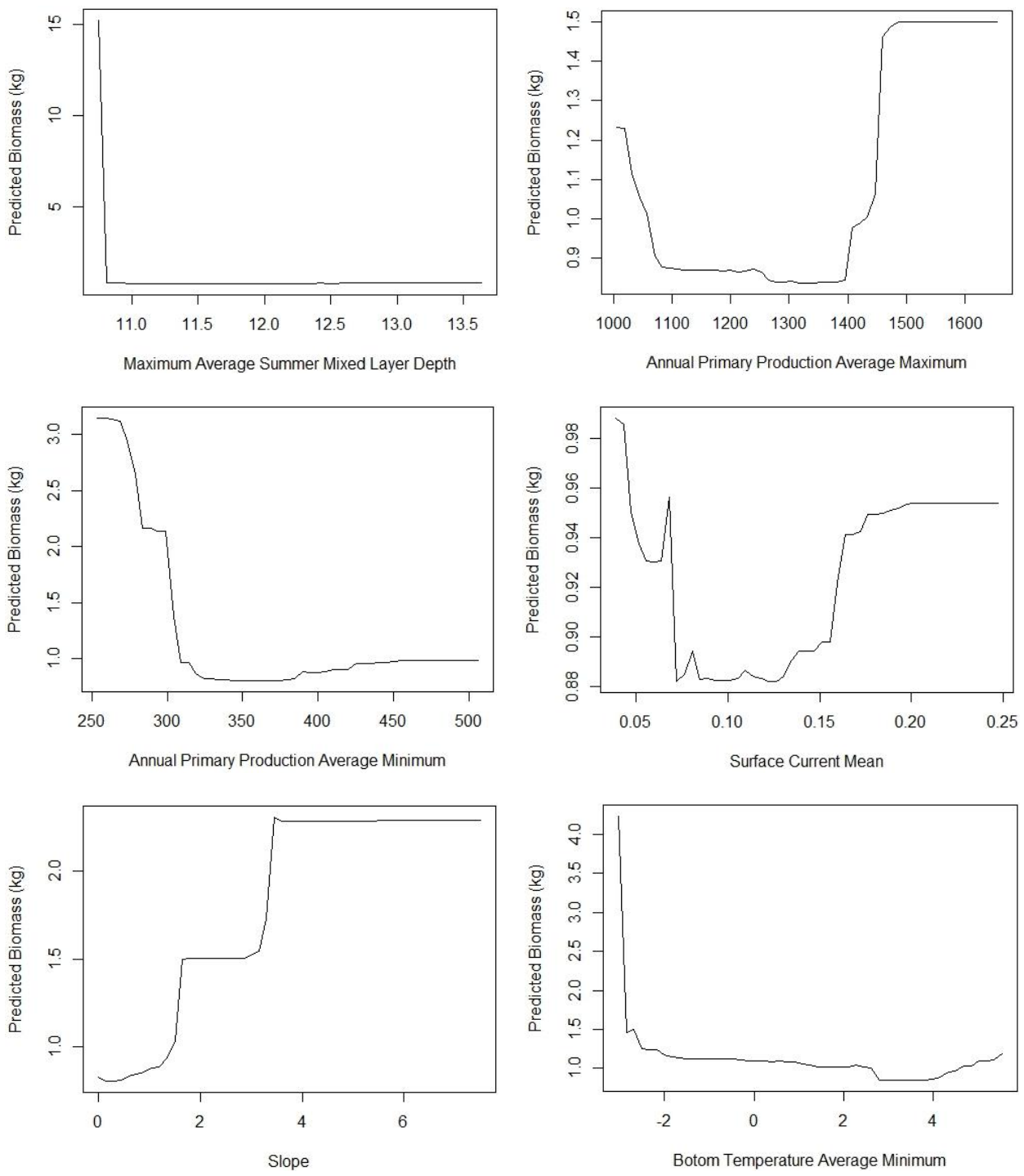

Figure 58. Partial dependence plots of the top six predictors of the random forest model (see Table 13) of sponge biomass data within the northern Gulf of St. Lawrence on, ordered left to right from the top.

Figures 59 and 60 show the surface of sponge biomass (kg) predictions per grid cell generated from the random forest model in the entire Gulf of St. Lawrence.

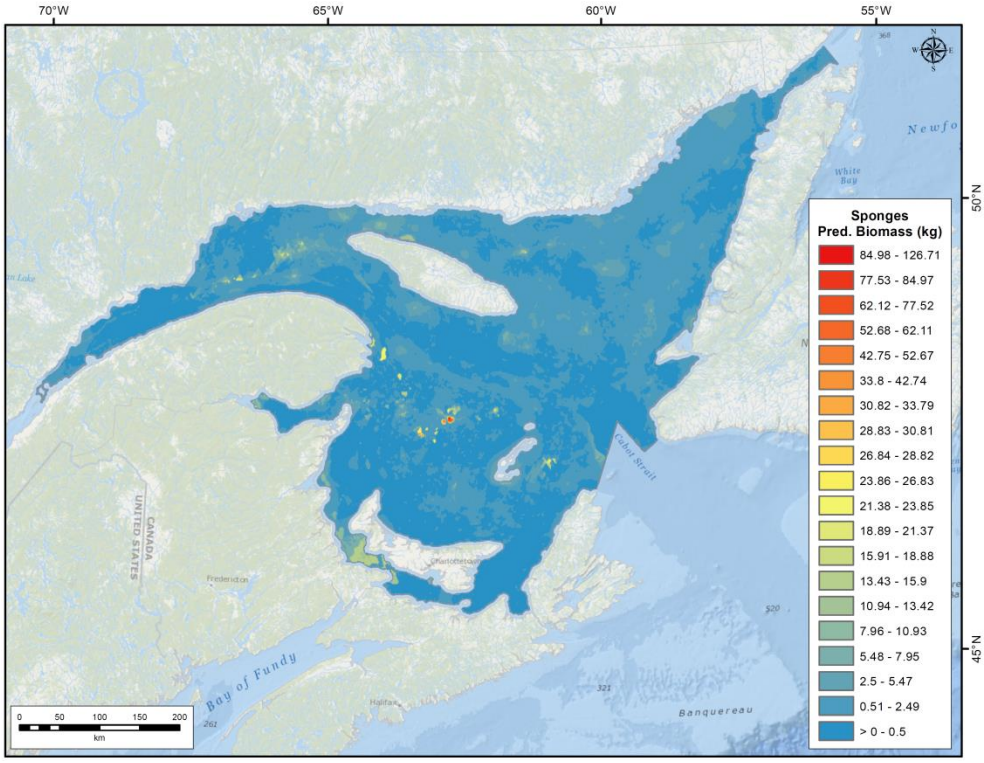

Figure 59. Predictions of sponge biomass from the random forest model within the Gulf Region.

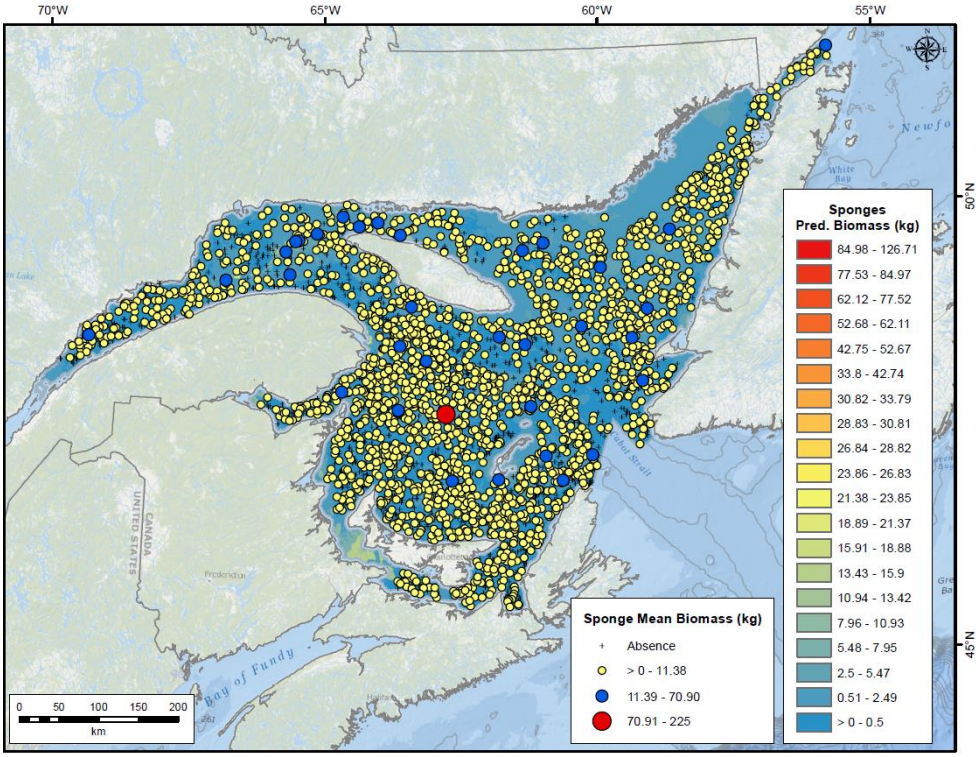

Figure 60. Sponge biomass catches overlaid to the sponge biomass predicted from the random forest model within the Gulf Region.

Stalked tunicates (*Boltenia ovifera*)

Data Sources and Distribution

The data available for analysis of stalked tunicates in the Gulf Region was collected over a period of 15 years from 2001 to 2015 (Table 14). The data were filtered so that only one presence or absence record occurs within the same environmental data raster cell (approximately 1 km^2; with presence records taking precedence over an absence record when both occurred in the same cell), giving a total of 1036 presence and 1126 absence records for the southern Gulf of St. Lawrence and 167 presence and 1495 absence records for the northern. Stalked tunicate biomass (kg) was averaged across each catch occurring within the same cell.

In order to study the different catchability of both gears, an analysis of the presence and absence of stalked tunicates from the tow sets carried out by both surveys in a common area in the Laurentian Channel was made. 168 tow sets were analyzed for the southern Gulf of St. Lawrence, whereas 155 were analyzed for the northern Gulf, finding 58 presences of stalked tunicates in the southern Gulf and 0 in the northern (Figure 61). Most of all the presences were small catches preceded by large catches carried out in shallower waters, which suggests some kind of contamination from the shallower tow sets to the deeper ones in the Channel. This contamination is an artefact that could affect the SDM and therefore 52 of these presence records were removed from the analysis; the remaining 6 presences being along the edge of the shelf and in an area with no evidence of contamination from precedent tow sets was found. Therefore, the final data used are presented in Table 14 and include a total of 984 presence and 1178 absence records for the southern Gulf of St. Lawrence and 167 presence and 1495 absence records for the northern.

Table 14. Data date range, gear type, area, and the total number of presences and absences of stalked tunicates catch data recorded from DFO trawl surveys.

	Southern Gulf		**Northern Gulf**	
Year	P	A	P	A
2001	49	95	-	-
2002	67	111	-	-
2003	27	55	-	-
2004	106	114	-	-
2005	100	122	-	-
2006	77	87	23	160
2007	84	82	21	158
2008	80	95	20	170
2009	66	82	15	157
2010	59	67	16	121
2011	63	62	12	140
2012	59	76	16	156
2013	68	56	7	133
2014	79	74	24	144
2015	-	-	13	156
TOTAL	**984**	**1178**	**167**	**1495**

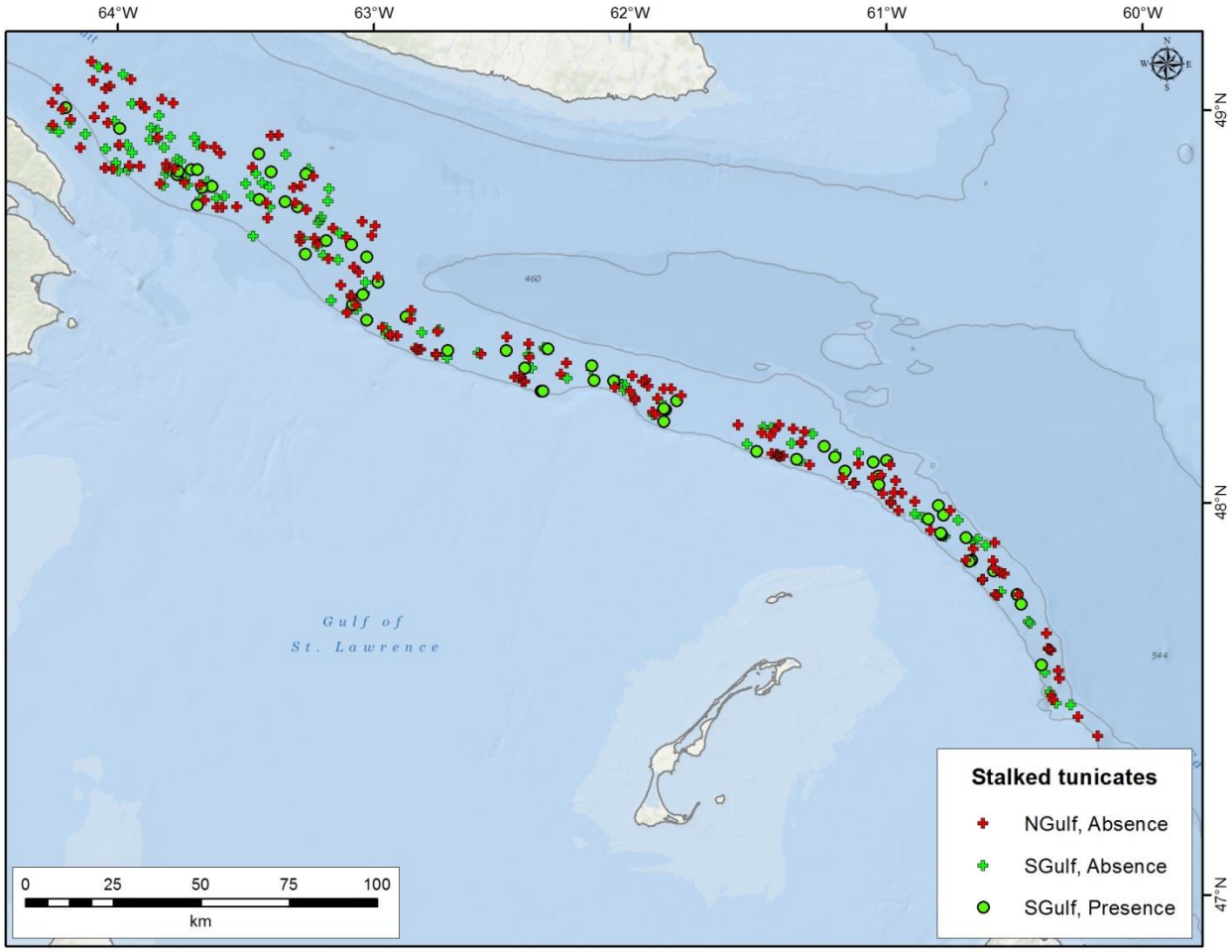

Figure 61. Stalked tunicates presence and absence from the common area surveyed by the southern and northern research vessel trawl surveys from 2001 to 2015 within the Gulf Region.

Stalked tunicate presence and absence records were distributed unevenly across the Gulf Region and are practically absent along the Laurentian Channel (Figure 62). Presences were concentrated mainly on the southern Gulf of St. Lawrence in shallow waters and scattered along the northern border, while absences were concentrated in the deeper areas. The highest biomass records occurred northeast of Prince Edward Island.

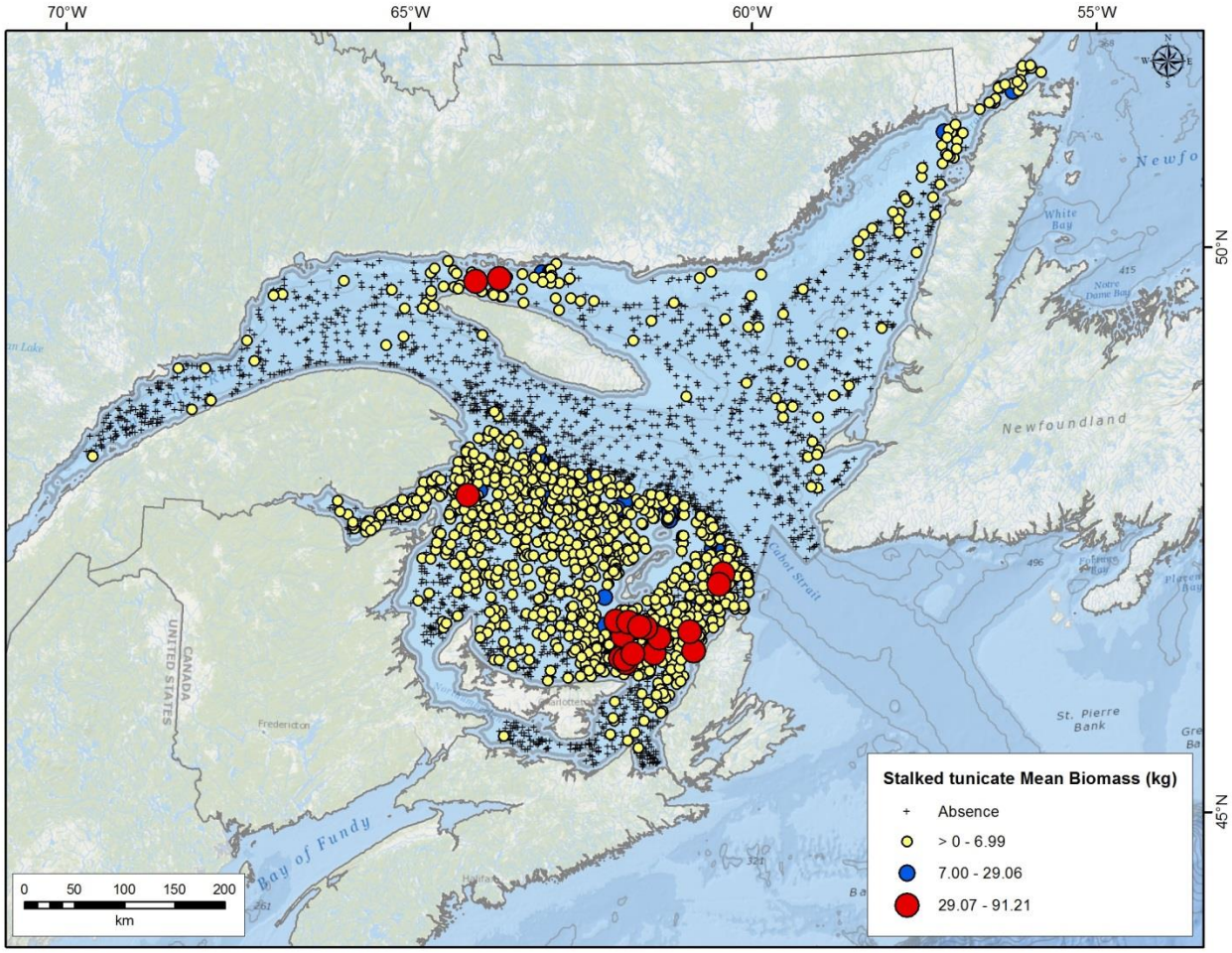

Figure 62. Mean biomass (kg) per grid cell of stalked tunicate by-catches recorded from DFO research vessel trawl surveys from 2001 to 2015 within the Gulf Region.

Model 1 – Balanced species prevalence

Accuracy measures (AUC, sensitivity, and specificity) for the random forest model on balanced species prevalence (1151 presences and 1151 absences) are presented in Table 15. The highest AUC was 0.891 associated with Model Run 7 which was considered the optimal model for the prediction of the stalked tunicate response data. The sensitivity and specificity measures of this model were 0.837 and 0.805, respectively. The confusion matrix of the optimal model is also presented in Table 15. Class error for the presence class was low (0.163).

Table 15. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of stalked tunicates within the Gulf Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 7) which is considered the optimal model for predicting the presence probability of stalked tunicates in the region.

Model Run	AUC	Sensitivity	Specificity
1	0.872	0.826	0.772
2	0.869	0.811	0.769
3	0.875	0.829	0.790
4	0.875	0.822	0.776
5	0.881	0.835	0.785
6	0.872	0.818	0.781
7	**0.891**	**0.837**	**0.805**
8	0.877	0.811	0.831
9	0.877	0.827	0.782
10	0.883	0.832	0.785
Mean	**0.877**	**0.825**	**0.788**
SD	**0.007**	**0.009**	**0.018**

Confusion matrix of model with maximum AUC:

Observations	Predictions		Total n	Class error
	Absence	**Presence**		
Absence	927	224	1151	0.195
Presence	188	963	1151	0.163

The presence probability prediction surface of stalked tunicates is presented in Figure 63. The highest predictions of presence probability occurred in the southern Gulf and in the shallower areas in the north. Areas of higher presence probability corresponded well with the spatial distribution of presence records (see Figure 64). Some areas in the Laurentian Channel are predicted to have zero or low presence probability of stalked tunicates.

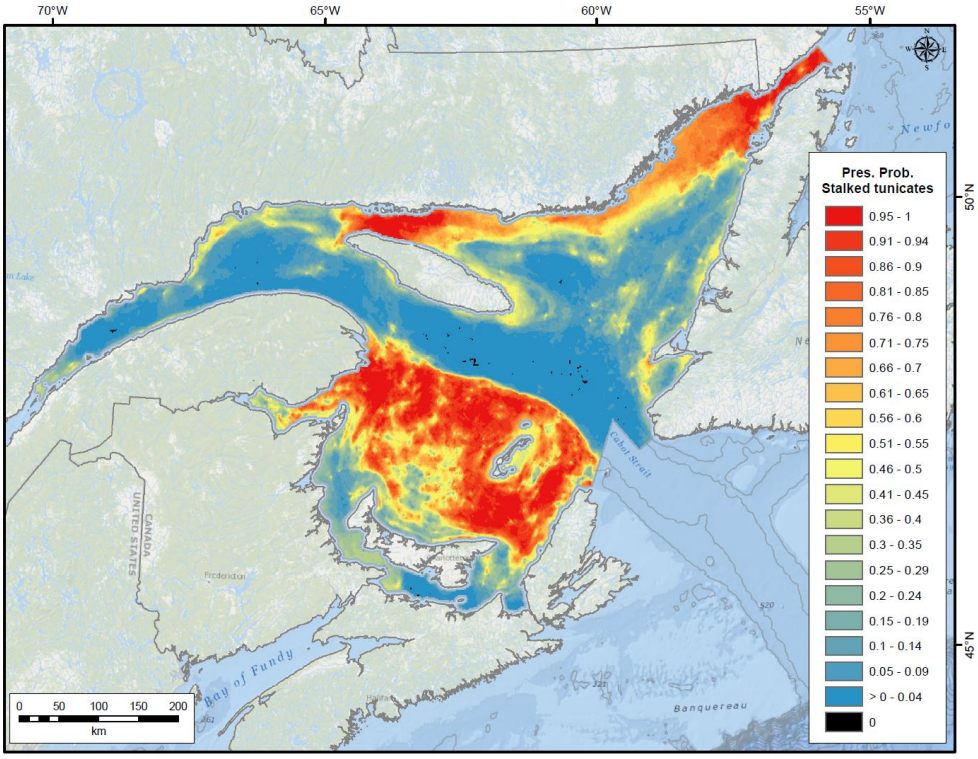

Figure 63. Predictions of presence probability (pres. prob.) of stalked tunicates from the optimal random forest model (see Table 15) based on balanced data.

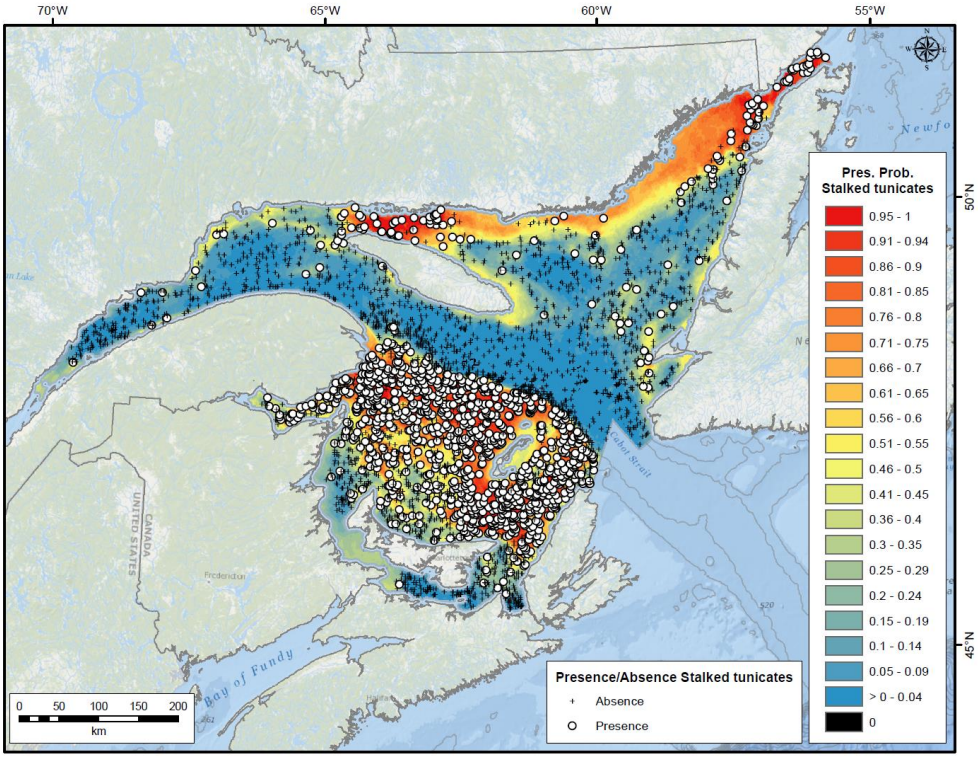

Figure 64. Total presence and absence observations and predictions of presence probability (pres. prob.) of stalked tunicates from the optimal random forest model (see Table 15) based on balanced data.

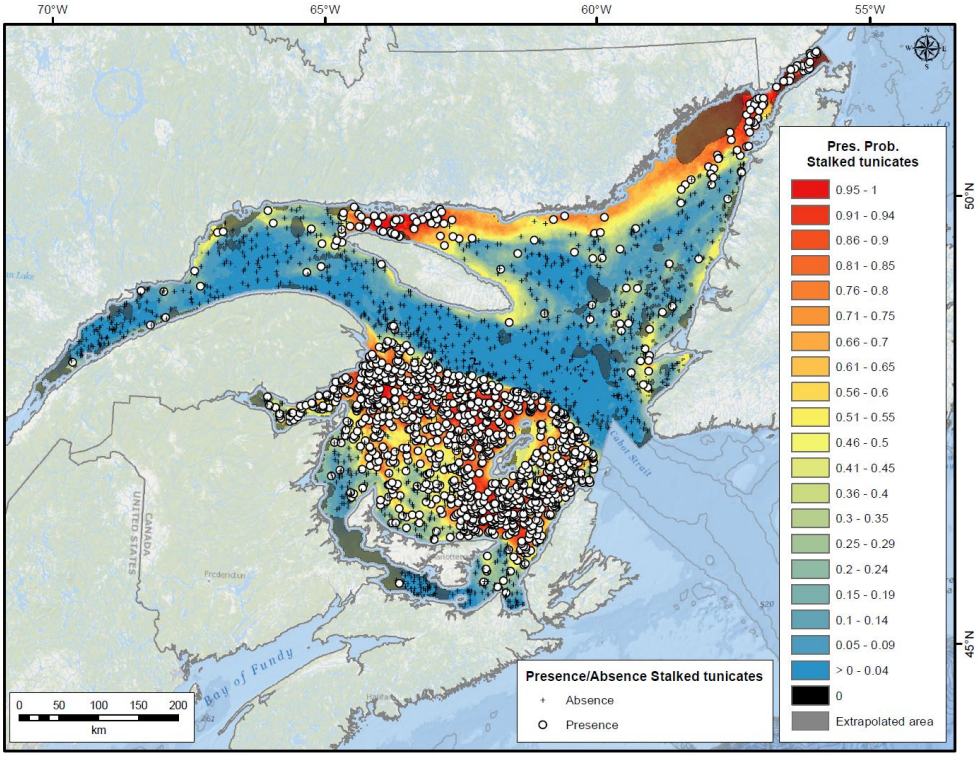

Figure 65. Presence and balanced absence observations and predictions of presence probability (pres. prob.) of stalked tunicates from the optimal random forest model (see Table 15) based on balanced data. The grey areas of extrapolation are also shown.

The actual presence and absence data observations (1151 presences and 1151 absences) used in the optimal model did not show important differences (Figure 65) with respect of the distribution of all the absence records across the study area (Figure 64). In this figure are also shown the areas of model extrapolation (i.e. areas where at least one environmental variable has values beyond its sampled range). The south of Prince Edward Island and the shallow portion in the northeast, before the Strait of Belle Isle are considered extrapolated area, as well as smaller areas scattered across the study area.

Of all 78 environmental predictor variables used in the model, Bottom Temperature Mean was the most important for the classification of the stalked tunicate presence and absence data (Figure 66). This variable was followed closely in terms of its Mean Decrease in Gini Value by Depth. Bottom Salinity Average Minimum and Bottom Salinity Average Maximum variables had high importance in this model too. The partial dependence plots for the top 6 most important predictors are shown in Figure 67. Presence probability of stalked tunicates was maximum at low values of Bottom Temperature Mean.

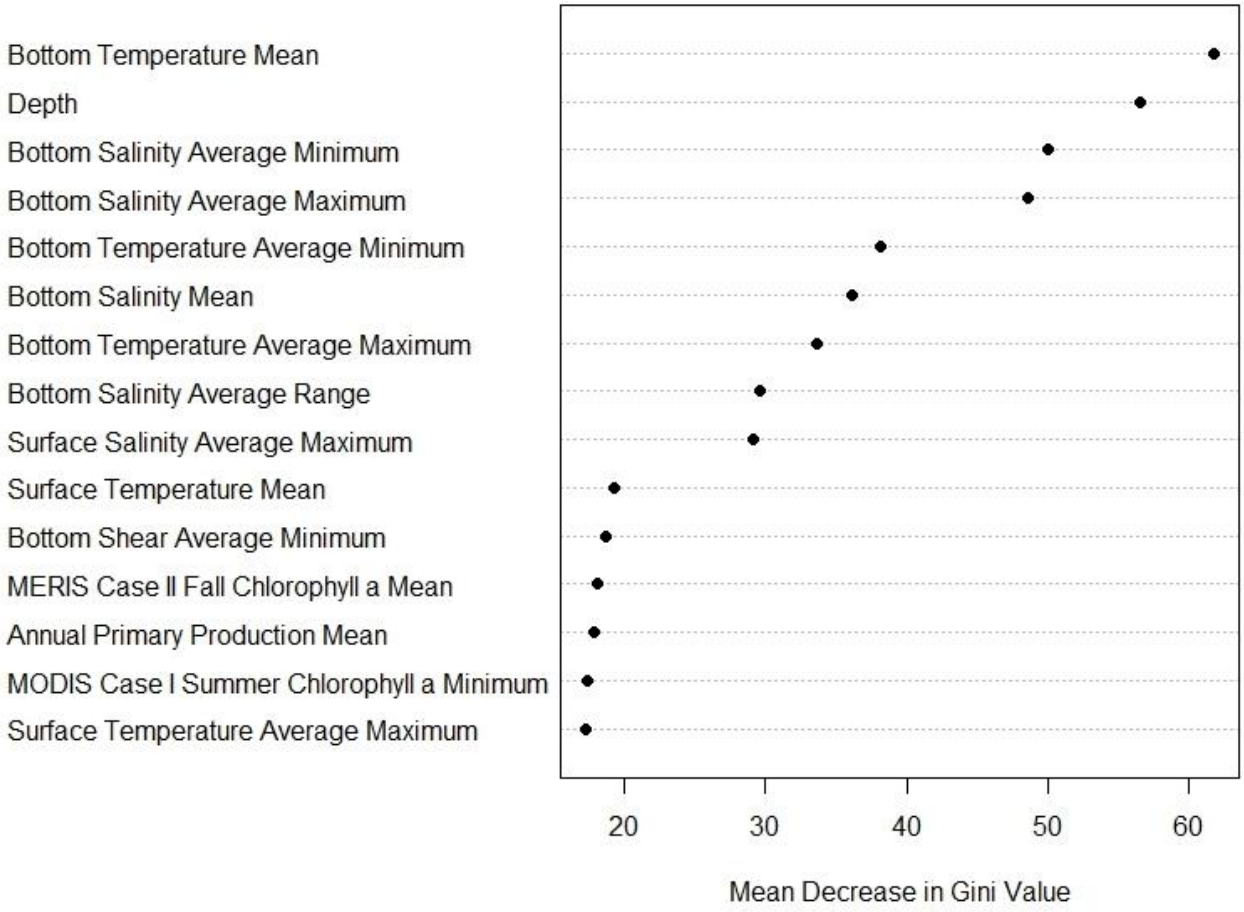

Figure 66. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model of stalked tunicate presence and absence data within the Gulf Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

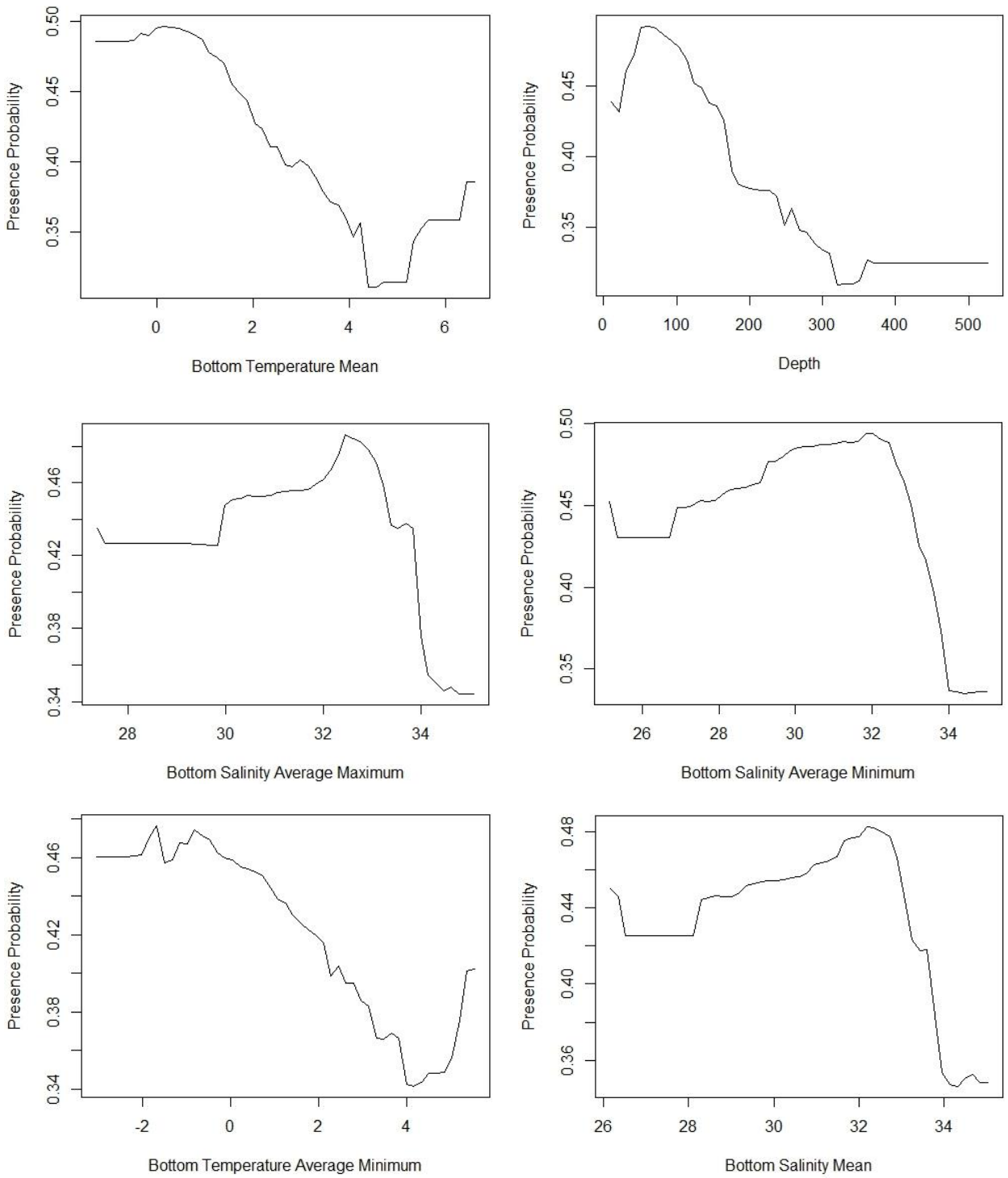

Figure 67. Partial dependence plots of the top six predictors from the optimal random forest model of stalked tunicate presence and absence data collected within the Gulf Region, ordered left to right from the top.

Model 2 – Unbalanced species prevalence

Table 16 shows the accuracy measures for the random forest model on stalked tunicates presence and absence data (1151 presences and 2673 absences) and a threshold equal to species prevalence (0.31). The AUC calculated from Model 2 was 0.874, slightly lower than that of Model 1 (0.891). Class error of the presence class was higher than that of Model 1 (0.163), as well as the class error for the absence class (0.218 compared to 0.195 from Model 1). Sensitivity and Specificity were also slightly lower.

Table 16. Accuracy measures and confusion matrix from 10-fold cross validation from random forest modelling of presence and absence of stalked tunicates within the Gulf Region with a cut-off threshold of 0.31.

Observations	Predictions		Total n	Class error	AUC	Sensitivity	Specificity
	Absence	Presence					
Absence	2089	584	2673	0.218	0.874	0.821	0.782
Presence	206	945	1151	0.179			

The predicted presence probability surface of stalked tunicates generated from Model 2 is shown in Figure 68 and it is very similar than Model 1, with the highest predictions of presence probability occurring in the southern Gulf and in the shallower areas in the north. However, more areas, mainly in the Laurentian Channel are predicted to have zero probability of stalked tunicates.

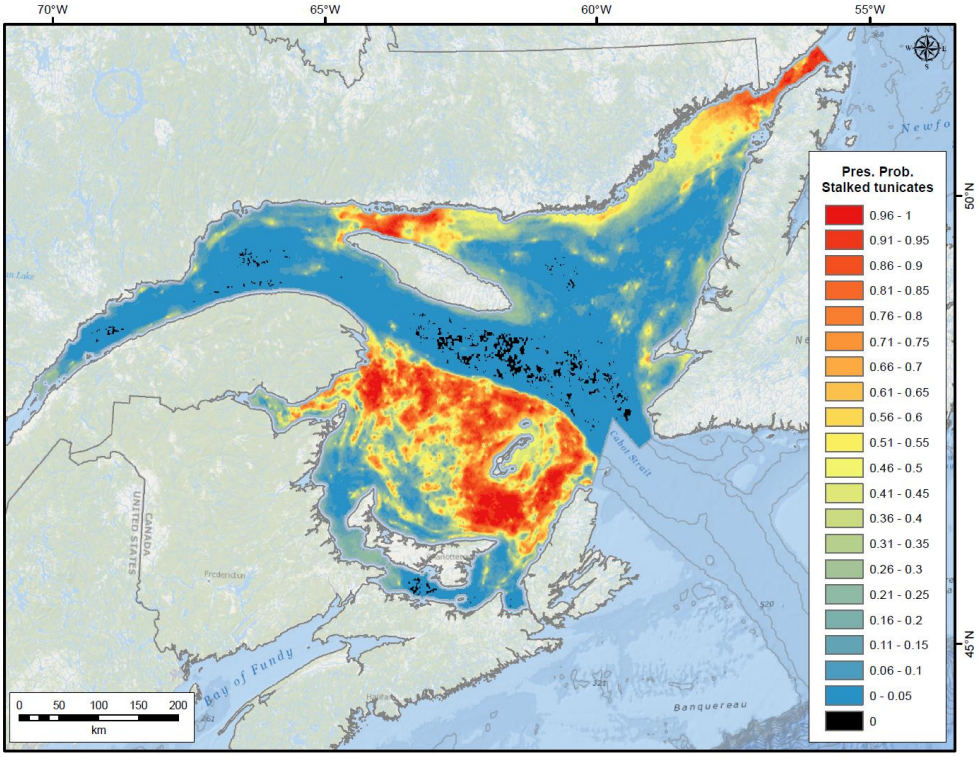

Figure 68. Predictions of presence probability (pres. prob.) of stalked tunicates from the unbalanced random forest model (see Table 16).

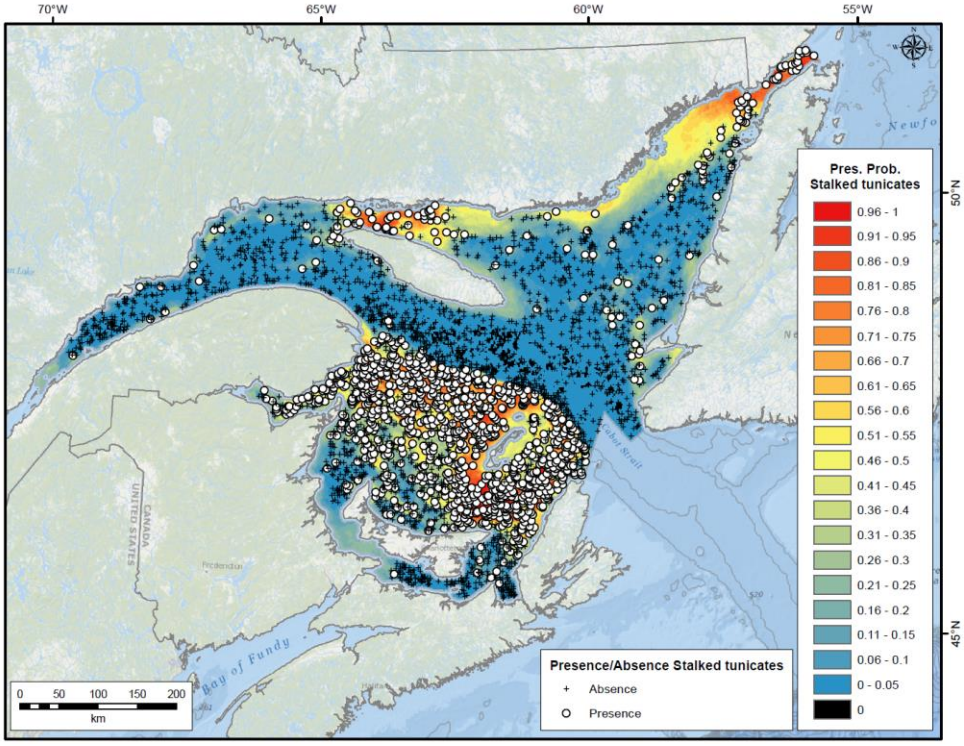

Figure 69. Presence and absence observations and predictions of presence probability (pres. prob.) of stalked tunicates from the unbalanced random forest model (see Table 16). The grey areas of extrapolation are also shown.

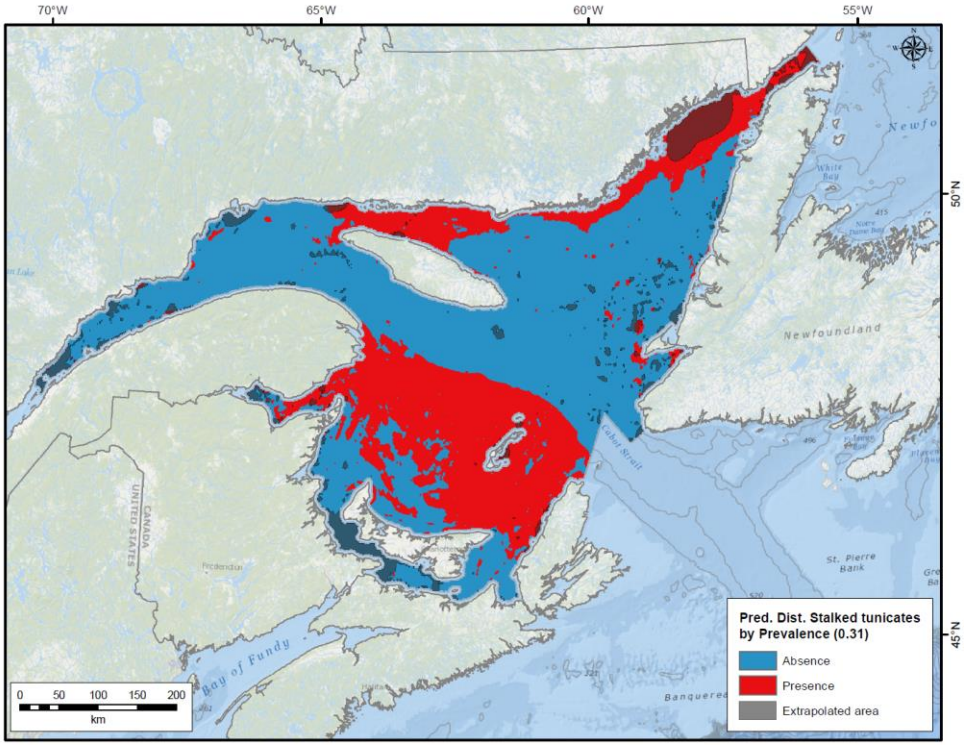

Figure 70. Predicted distribution of stalked tunicates from the unbalanced random forest model (see Table 16) using the prevalence cut-off threshold of 0.31. The grey areas of extrapolation are also shown, they appear dark red when overlain on the red presence surface.

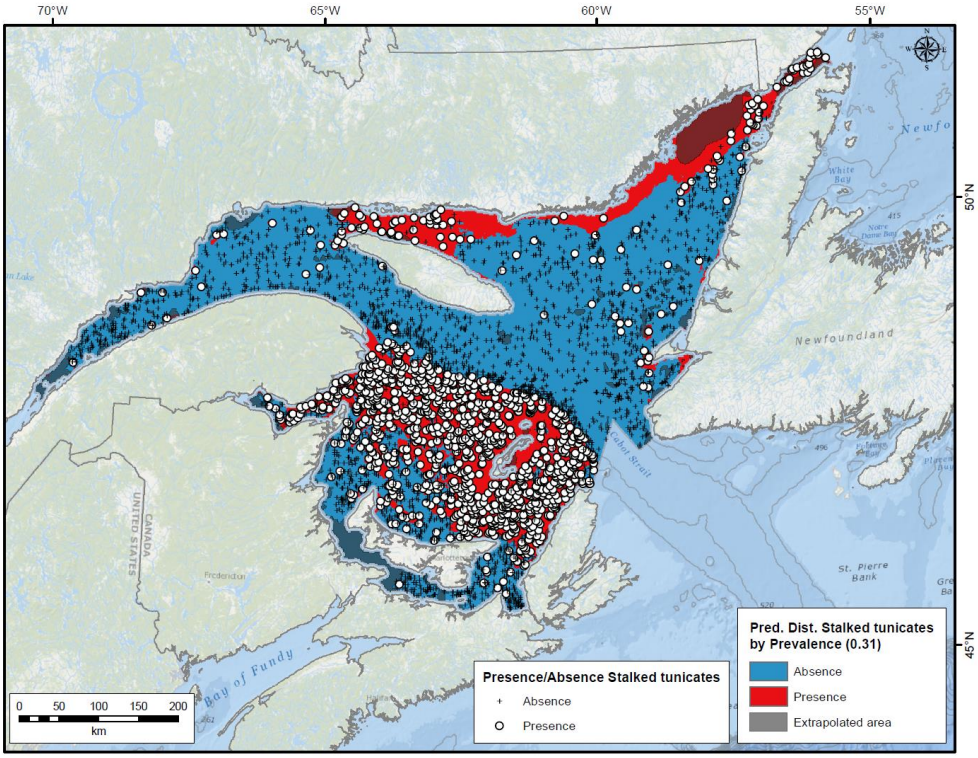

Figure 71. Presence and absence observations and predicted distribution of stalked tunicates from the unbalanced random forest model (see Table 16) using the prevalence cut-off threshold of 0.31. The grey areas of extrapolation are also shown, they appear dark red when overlain on the red presence surface.

Five of the six most important environmental predictor variables in Model 2 were the same than that of Model 1 (Figure 72), and Bottom Temperature Mean and Depth were also the most important variables for the classification of stalked tunicates in Model 2. The partial dependence plots for the top 6 most important predictors are shown in Figure 73 and show similar results to those for Model 1.

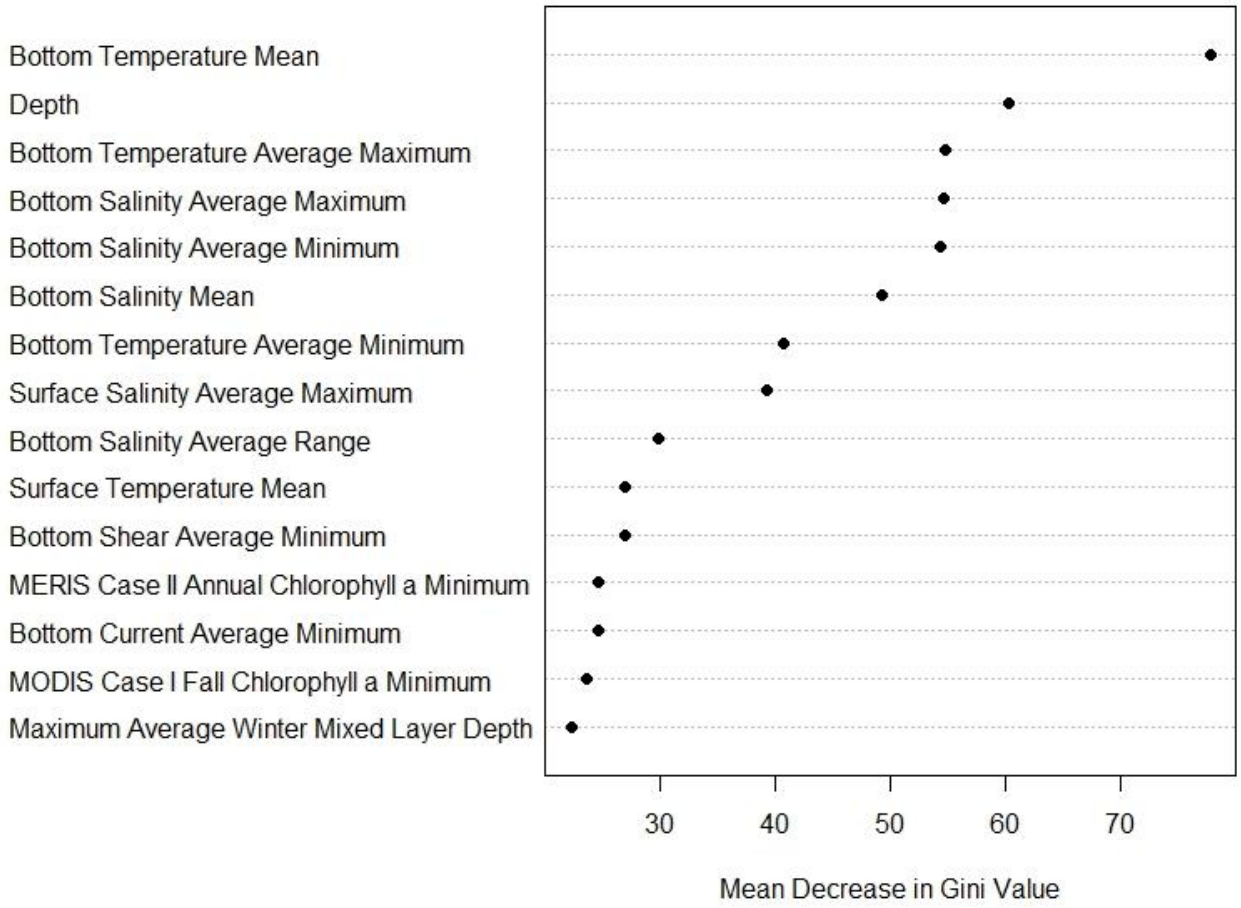

Figure 72. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the unbalanced random forest model of stalked tunicate presence and absence data within the Gulf Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

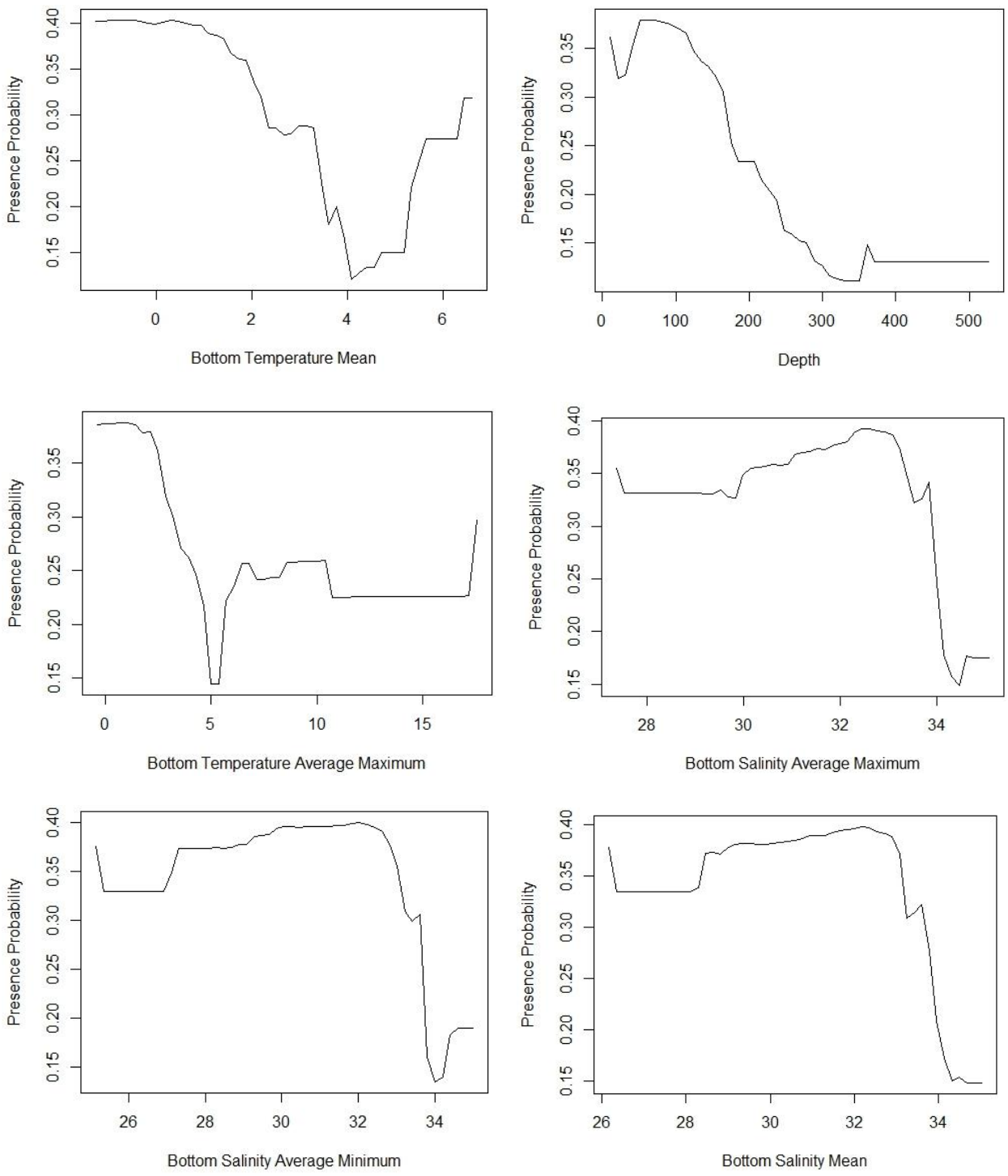

Figure 73. Partial dependence plots of the top six predictors of from the unbalanced random forest model of stalked tunicate presence and absence data collected within the Gulf Region, ordered left to right from the top.

Correlation between Biomass and Presence Probability

Biomass (kg) of stalked tunicates was positively correlated (ρ=0.73) to the presence probability of stalked tunicates based on balanced data, showing that areas with higher probabilities tend to present higher biomass (Figure 74).

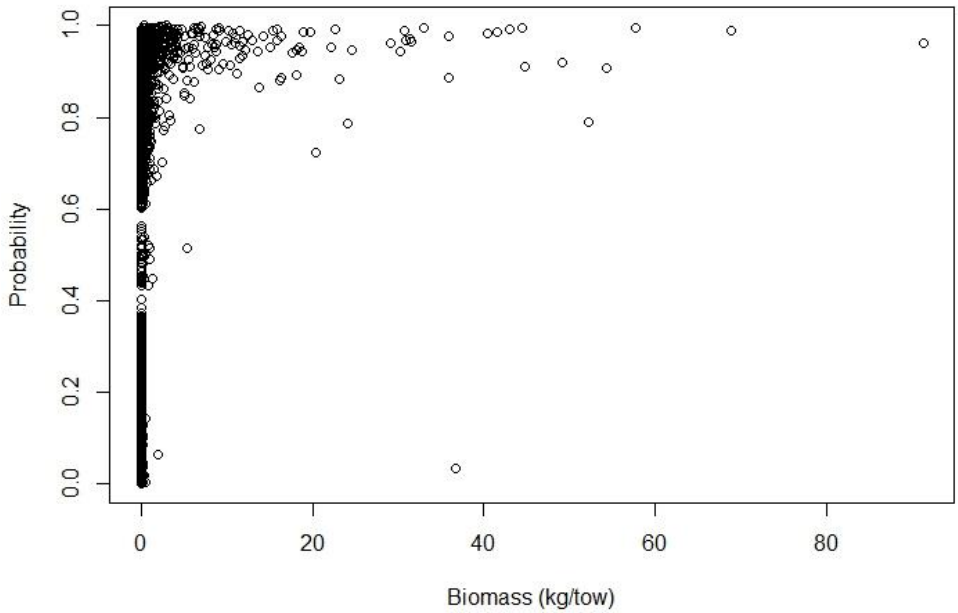

Figure 74. Scatterplot between the presence probabilities of stalked tunicates based on balanced data with the mean biomass (kg) per grid cell of stalked tunicates.

Model Selection

Model accuracy measures were very similar for both models, although Model 1 presented slightly higher AUC and sensitivity and specificity. The selection of Model 1, based on balanced data, allows for the use of the presence probabilities considering values higher than 0.5 as presences and lower as absences to maintain fairly constant error rates.

Prediction of Stalked Tunicates Biomass Using Random Forest

c) Southern Gulf of St. Lawrence

The accuracy measures of the regression random forest model on mean stalked tunicate biomass per grid cell for the southern Gulf of St. Lawrence are presented in Table 17. The average R-square (R^2) value was 0.179 ± 0.080, the average Root-Mean-Square Error (RMSE) was 4.347 ± 1.495, and the average percent variance explained was 11.98 ± 1.61.

Table 17. Accuracy measures from 10-fold cross validation from random forest modelling of biomass of stalked tunicates within the southern Gulf of St. Lawrence. NRMSE, is the normalized root-mean square deviation (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	%variance explained
1	0.196	3.349	0.037	11.14
2	0.308	4.042	0.044	9.11
3	0.200	5.193	0.057	10.56
4	0.084	4.760	0.052	12.76
5	0.184	6.886	0.075	13.78
6	0.085	6.031	0.066	13.67
7	0.080	3.100	0.034	13.13
8	0.170	5.017	0.055	12.92
9	0.288	2.963	0.032	10.19
10	0.199	2.125	0.023	12.54
Mean	**0.179**	**4.347**	**0.048**	**11.98**
SD	**0.080**	**1.495**	**0.016**	**1.61**

Figures 76 and 77 show the surface of stalked tunicate biomass (kg) predictions per grid cell generated from the random forest model in the southern Gulf of St. Lawrence. The maximum biomass was predicted in the northeast of Prince Edward Island.

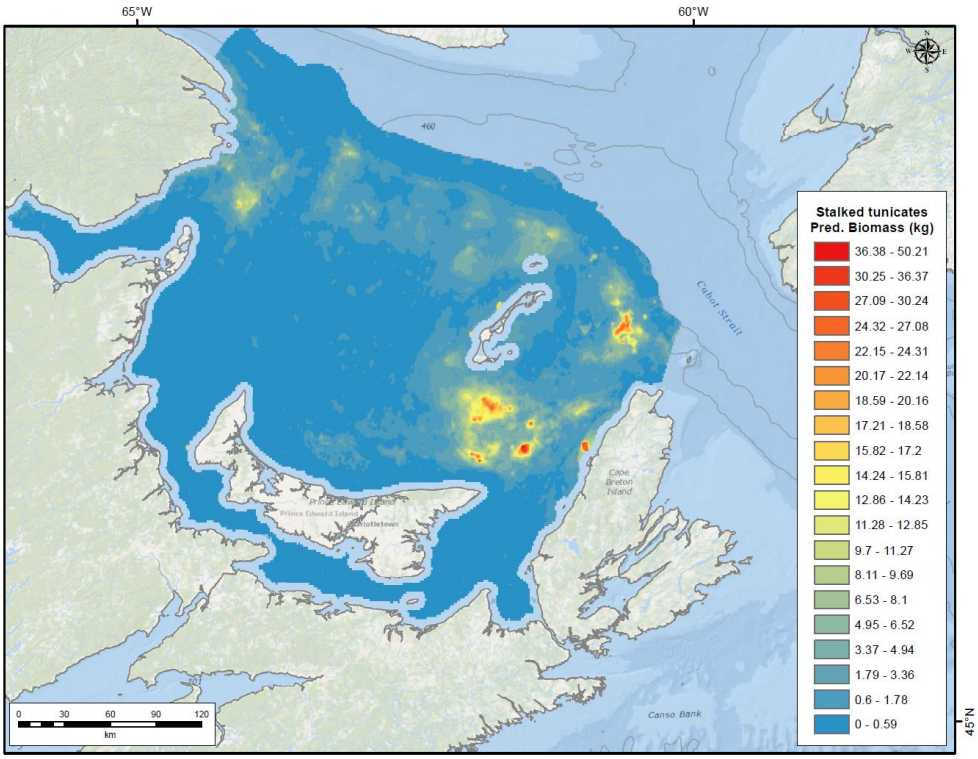

Figure 76. Predictions of stalked tunicate biomass probability from the random forest model (see Table 17) within the southern Gulf of St. Lawrence.

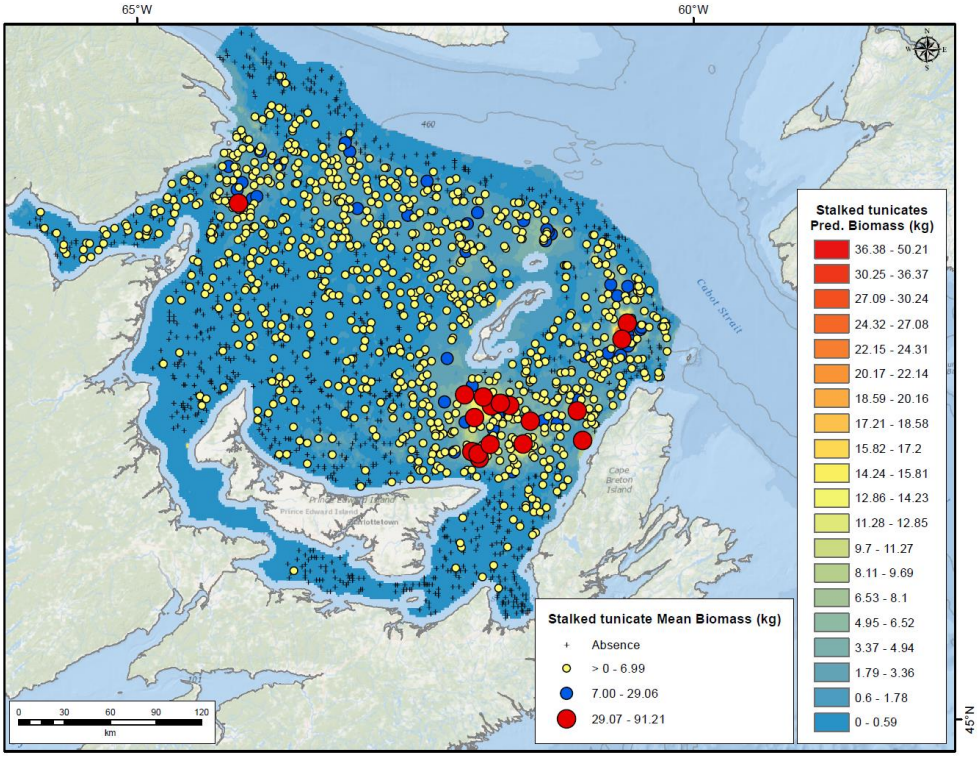

Figure 77. Stalked tunicate biomass catches overlaid to the stalked tunicate biomass probability from the random forest model (see Table 17) within the southern Gulf of St. Lawrence.

The top 15 most important environmental variables for predicting stalked tunicate biomass in the southern Gulf of St. Lawrence are shown in Figure 78. MERIS Case II Fall Chlorophyll *a* Minimum was the most important variable in the model, followed more distantly by Bottom Current Average Range and Annual Primary Production Average Range. The partial dependence of stalked tunicate biomass on the top 6 most important variables is shown in Figure 79. Predicted biomass (kg) was maximum for MERIS Case II Fall Chlorophyll *a* Minimum values lower than 0.5 mg m^{-3}.

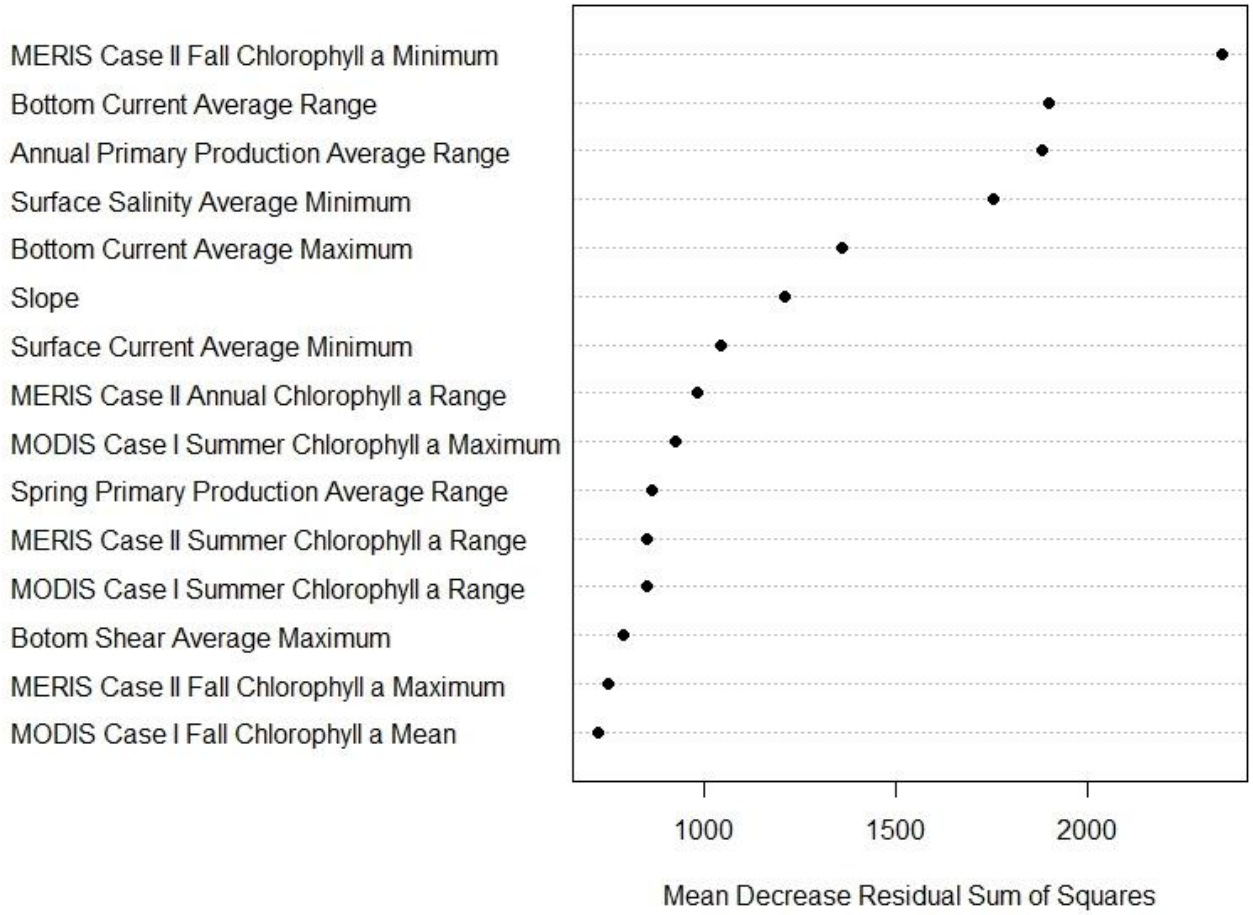

Figure 78. Importance of the top 15 predictor variables measured as the Mean Decrease Residual Sum of Squares of the random forest model (see Table 17) of stalked tunicate biomass data within the southern Gulf of St. Lawrence.

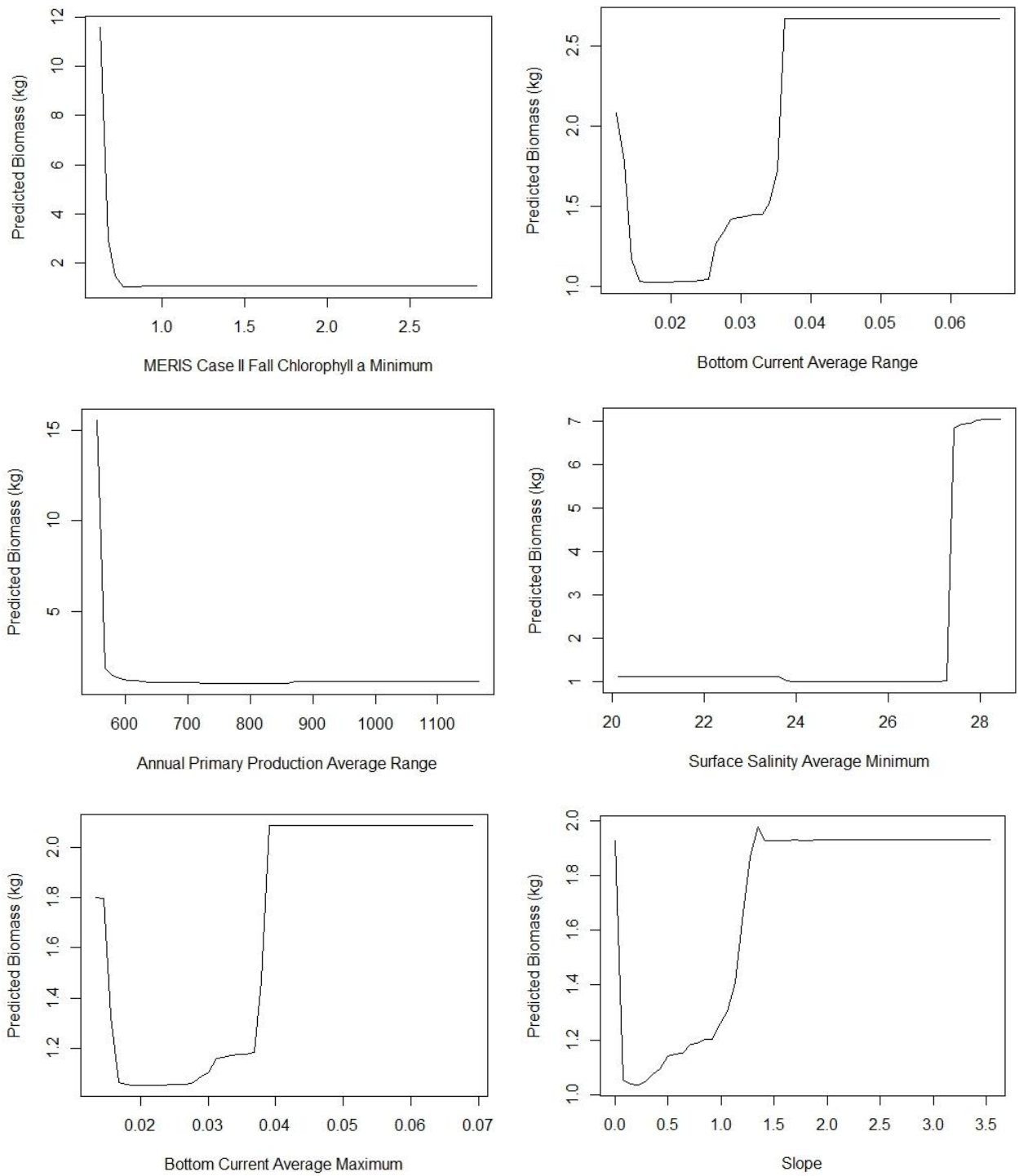

Figure 79. Partial dependence plots of the top six predictors of the random forest model (see Table 17) of stalked tunicate biomass data within the southern Gulf of St. Lawrence, ordered left to right from the top.

Model 4 – Northern Gulf Region

The accuracy measures of the regression random forest model on mean stalked tunicate biomass per grid cell for the northern Gulf of St. Lawrence are presented in Table 18. The average R-square (R^2) value was 0.409 ± 0.215, the average Root-Mean-Square Error (RMSE) was 1.082 ± 0.560, and the average percent variance explained was 28.17 ± 9.87. The mean accuracy metrics were better than for the southern Gulf of St. Lawrence.

Table 18. Accuracy measures from 10-fold cross validation from random forest modelling of biomass of stalked tunicates within the northern Gulf of St. Lawrence. NRMSE, is the normalized root-mean square deviation (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	%variance explained
1	0.572	0.854	0.024	29.62
2	0.433	1.632	0.045	39.25
3	0.225	1.106	0.031	28.96
4	0.824	1.175	0.033	10.33
5	0.199	0.622	0.017	34.42
6	0.495	0.506	0.014	33.02
7	0.252	0.535	0.015	30.05
8	0.243	1.530	0.043	30.84
9	0.228	0.665	0.018	34.82
10	0.622	2.196	0.061	10.39
Mean	**0.409**	**1.082**	**0.030**	**28.17**
SD	**0.215**	**0.560**	**0.016**	**9.87**

Figures 80 and 81 show the surface of stalked tunicate biomass (kg) predictions per grid cell generated from the random forest model in the northern Gulf of St. Lawrence. The maximum biomass was predicted in the northwest of Anticosti Island and in the Strait of Belle Isle.

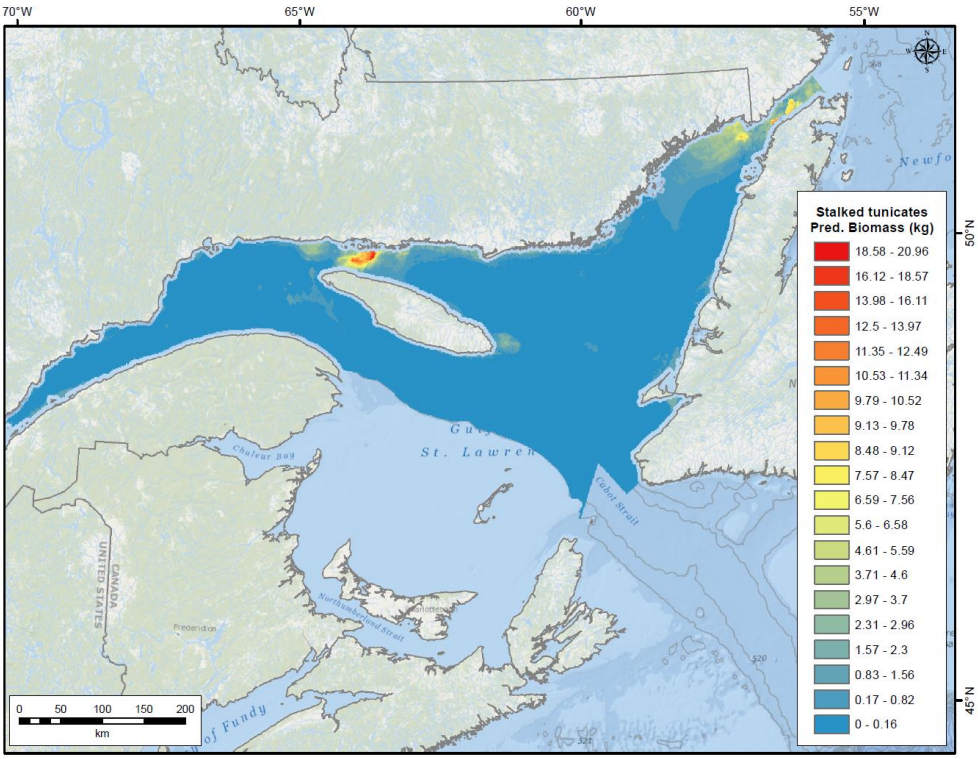

Figure 80. Predictions of stalked tunicate biomass probability from the random forest model (see Table 18) within the northern Gulf of St. Lawrence.

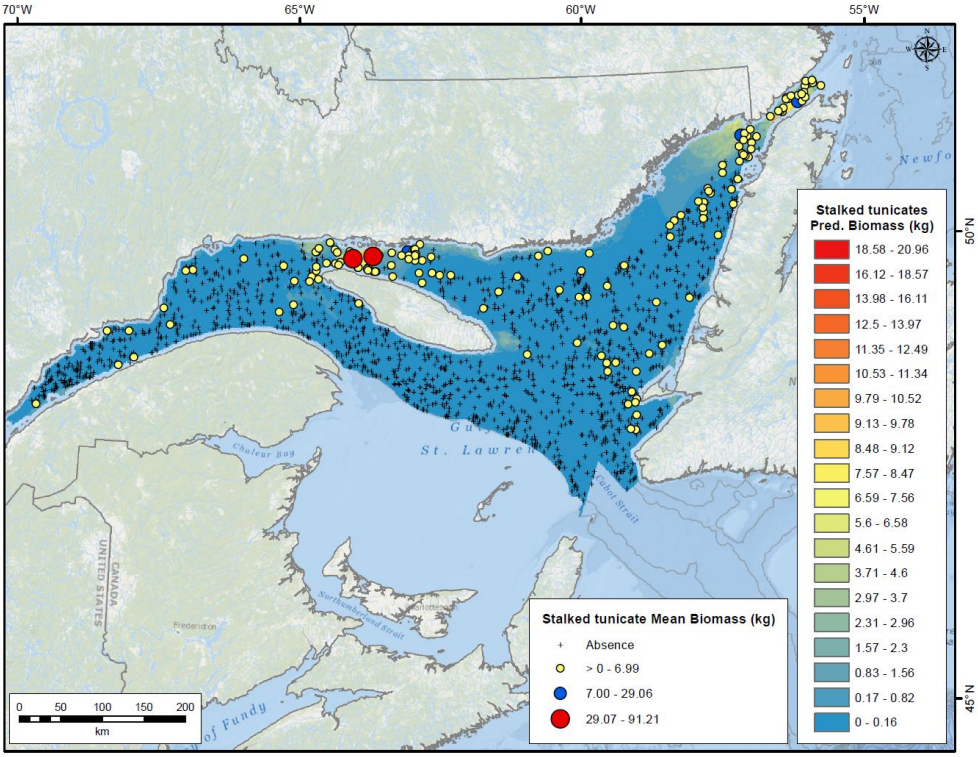

Figure 81. Stalked tunicate biomass catches overlaid to the stalked tunicate biomass probability from the random forest model (see Table 18) within the northern Gulf of St. Lawrence.

The top 15 most important environmental variables for predicting stalked tunicate biomass in the northern Gulf of St. Lawrence are shown in Figure 82. Summer Primary Production Mean was the most important variable in the model, followed distantly by the other variables in the model. The partial dependence of stalked tunicate biomass on the top 6 most important variables is shown in Figure 83. Summer Primary Production Mean was the most important variable for predicting stalked tunicate biomass. Predicted biomass was highest at primary production values lower than 850 mg C m^{-2} day^{-1}.

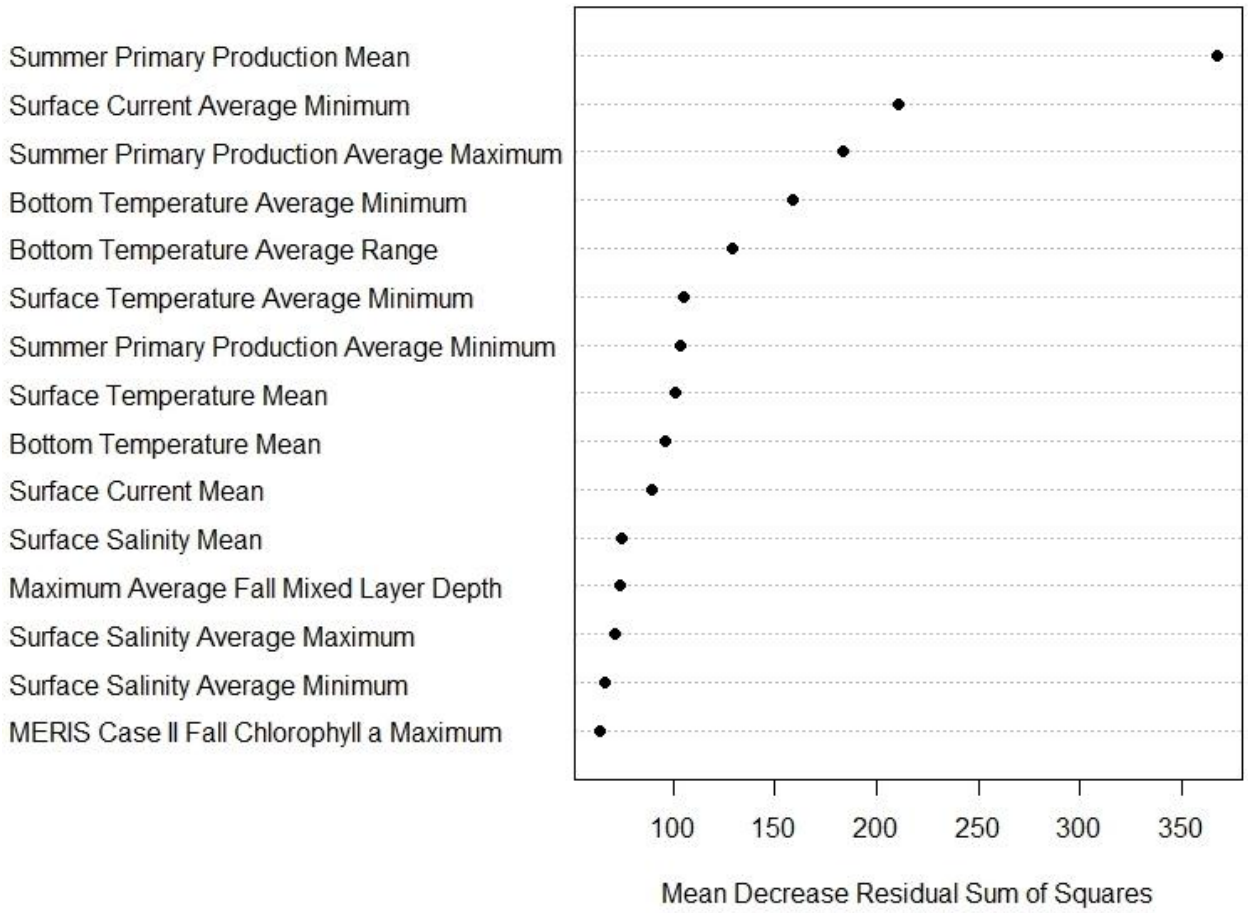

Figure 82 Importance of the top 15 predictor variables measured as the Mean Decrease Residual Sum of Squares of the random forest model (see Table 18) of stalked tunicate biomass data within the northern Gulf of St. Lawrence.

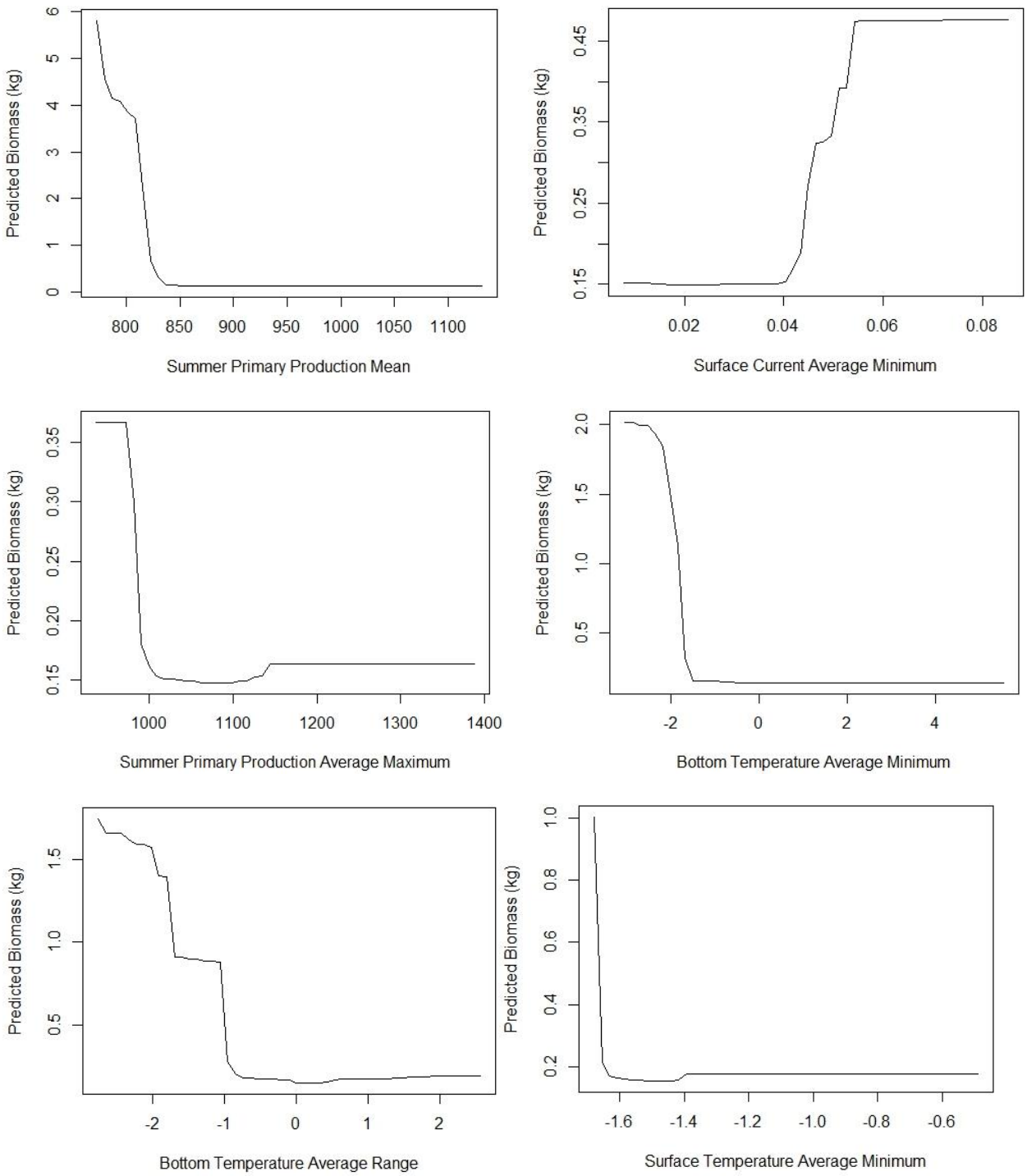

Figure 83. Partial dependence plots of the top six predictors of the random forest model (see Table 18) of stalked tunicate biomass data within the northern Gulf of St. Lawrence, ordered left to right from the top.

Figures 84 and 85 show the surface of stalked tunicate biomass (kg) predictions per grid cell generated from the random forest model in the entire Gulf of St. Lawrence.

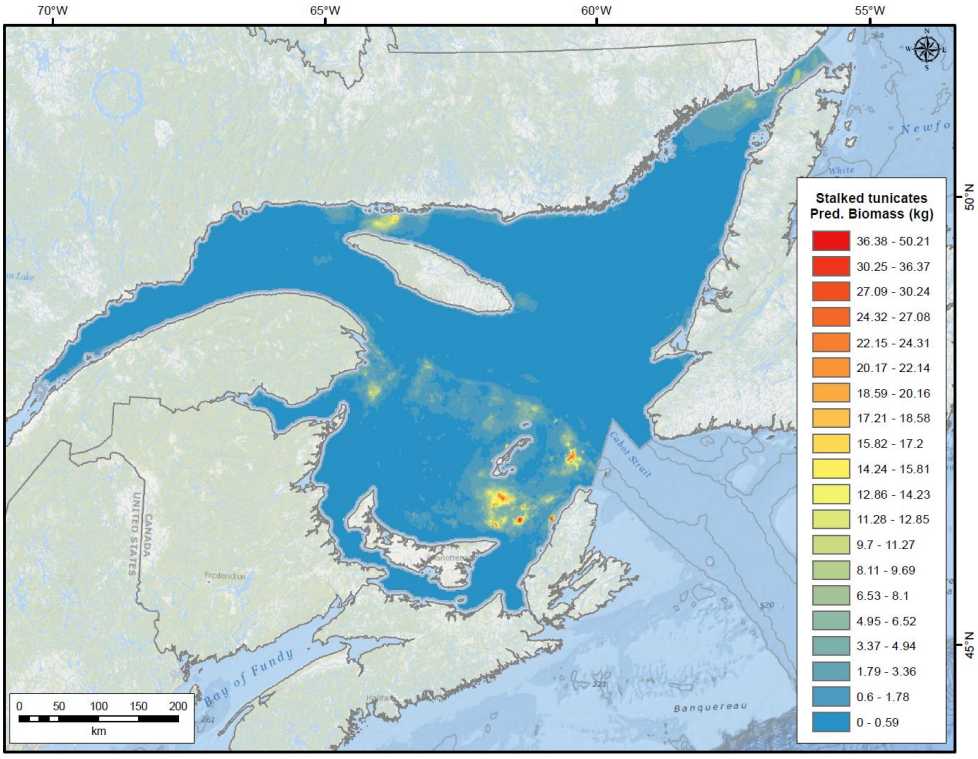

Figure 84. Predictions of stalked tunicate biomass probability from the random forest model within the Gulf Region.

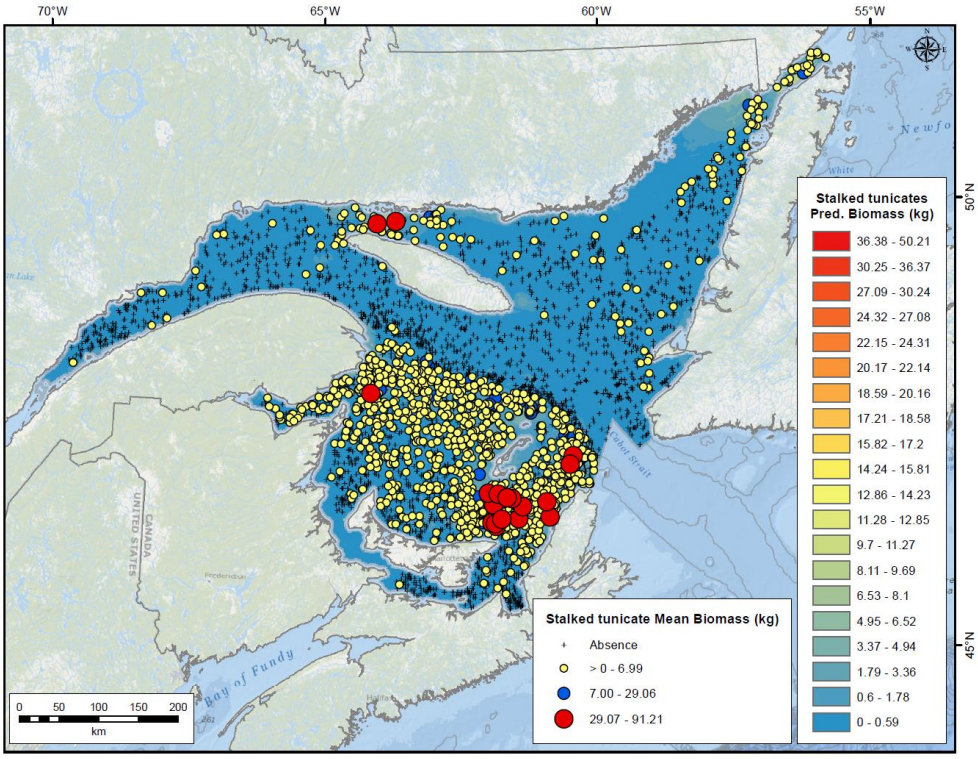

Figure 85. Stalked tunicate biomass catches overlaid to the stalked tunicate biomass probability from the random forest model within the Gulf Region.

Soft corals (Gersemia rubiformis)

Data Sources and Distribution

The soft coral *G. rubiformis*, known also as sea cauliflower, by-catch data available for species distribution modelling was collected over a period of 13 years from 2003 to 2015 (Table 19). In the surveys from the southern Gulf of St. Lawrence one catch of 41.66 kg was observed in an area with no other high catches around it. The next largest by-catch in term of biomass was 6.60 kg. As it was not possible to confirm this data as a valid catch or as an annotation error, it was considered for the presence-absence models but not for the biomass model. The data were filtered so that only one presence or absence record occurred within the same environmental data raster cell (approximately 1 km^2; with presence records taking precedence over an absence record when both occurred in the same cell), giving a total of 730 presence and 1118 absence records for the southern Gulf Region and 326 presence and 1342 absence records for the northern Gulf Region. Biomass (kg) of *G. rubiformis* was averaged across each catch occurring within the same cell. *G. rubiformis* presence and absence records were distributed unevenly across the Gulf Region (Figure 86). Presences were concentrated mainly in the southern Region in shallow waters and scattered along the northern Region, while absences were concentrated on the deeper areas. The highest biomass records occurred in the southern Region.

Table 19. Data date range, gear type, area, and the total number of presences and absences of *G. rubiformis* catch data recorded from DFO trawl surveys.

	Southern Gulf		**Northern Gulf**	
Year	P	A	P	A
2003	52	30	-	-
2004	69	150	-	-
2005	89	135	-	-
2006	75	90	53	128
2007	58	109	26	152
2008	57	119	47	144
2009	46	101	49	122
2010	57	73	52	86
2011	42	77	27	127
2012	52	87	13	157
2013	63	61	14	126
2014	70	86	26	142
2015	-	-	19	158
TOTAL	**730**	**1118**	**326**	**1342**

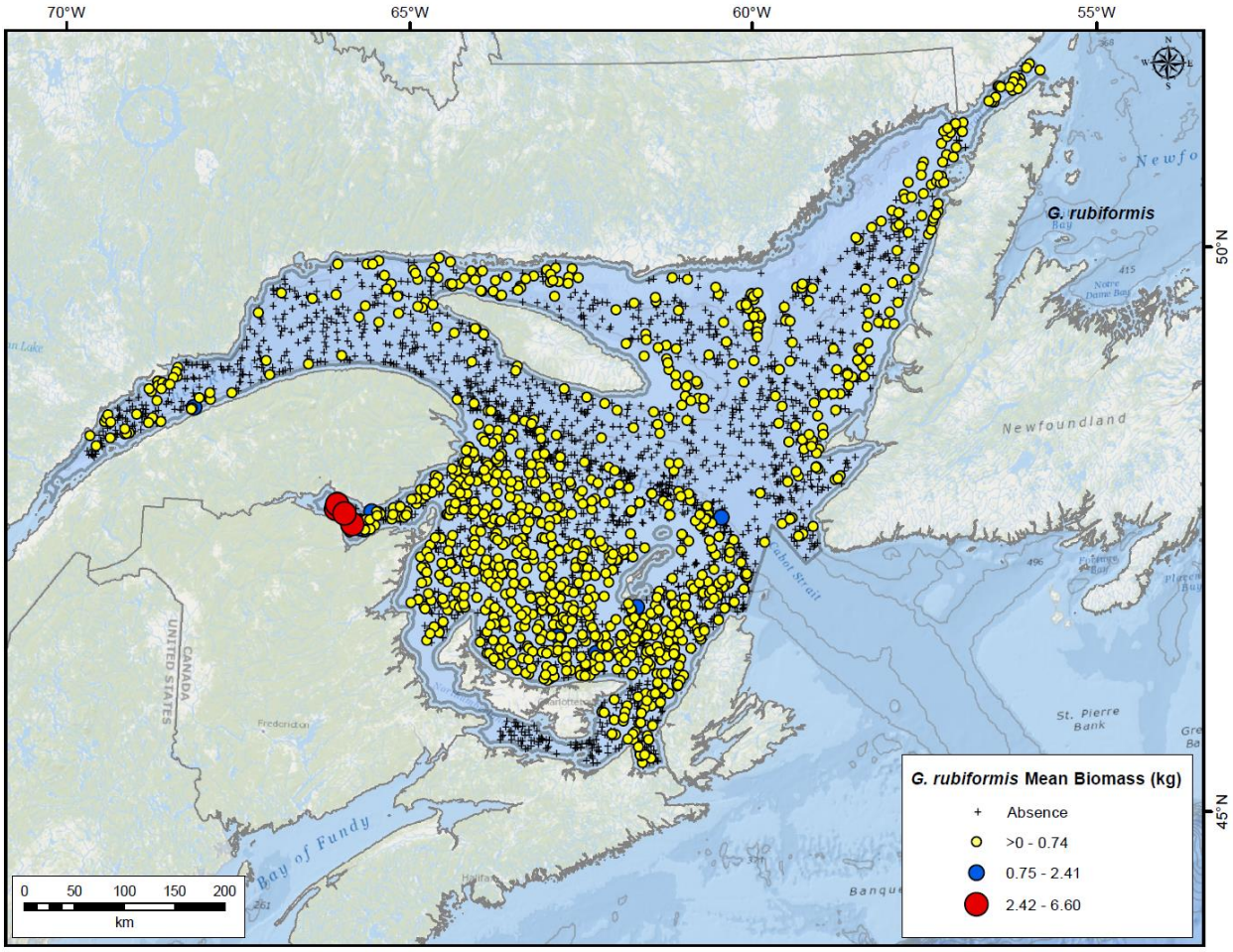

Figure 86. Mean biomass (kg) per grid cell of *G. rubiformis* by-catches recorded from DFO groundfish trawl surveys from 2003 to 2015 within the Gulf Region.

Model 1 – Balanced species prevalence

Accuracy measures (mean AUC, sensitivity and specificity) for the random forest model on balanced species prevalence data are presented in Table 20. The highest AUC was 0.753 associated with Model Run 7 which is considered the optimal model for the prediction of *G. rubiformis* response data. The sensitivity and specificity measures of this model were 0.722 and 0.674, respectively. The confusion matrix of the optimal model is also presented in Table 20. Class error for both classes was above 0.25.

Table 20. Accuracy measures for all 10 model repetitions from 10-fold cross validation from random forest modelling of presence and absence of *G. rubiformis* within the Gulf Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 7) which is considered the optimal model for predicting *G. rubiformis* presence probability.

Model Run	AUC	Sensitivity	Specificity
1	0.739	0.706	0.666
2	0.743	0.729	0.659
3	0.738	0.704	0.652
4	0.730	0.699	0.651
5	0.752	0.729	0.670
6	0.738	0.713	0.662
7	**0.753**	**0.722**	**0.674**
8	0.737	0.709	0.632
9	0.749	0.728	0.670
10	0.733	0.702	0.649
Mean	**0.741**	**0.714**	**0.659**
SD	**0.008**	**0.012**	**0.013**

Confusion matrix of model with maximum AUC:

Observations	Predictions		Total n	Class error
	Absence	**Presence**		
Absence	712	344	1056	0.326
Presence	294	762	1056	0.278

The presence probability prediction surface of *G. rubiformis* is presented in Figure 87. The highest predictions of presence probability occurred in the shallower areas of the Region. Areas of higher presence probability corresponded well with the spatial distribution of presence records (see Figure 88).

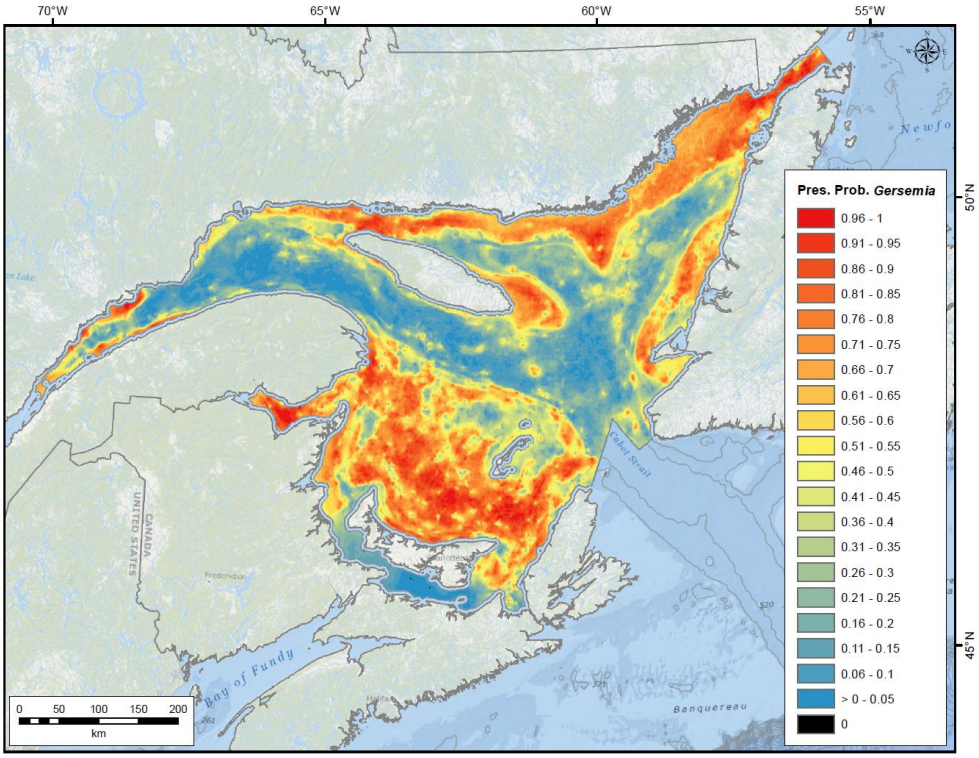

Figure 87. Predictions of presence probability (pres. prob.) of *G. rubiformis* from the optimal random forest model (see Table 20) based on balanced data.

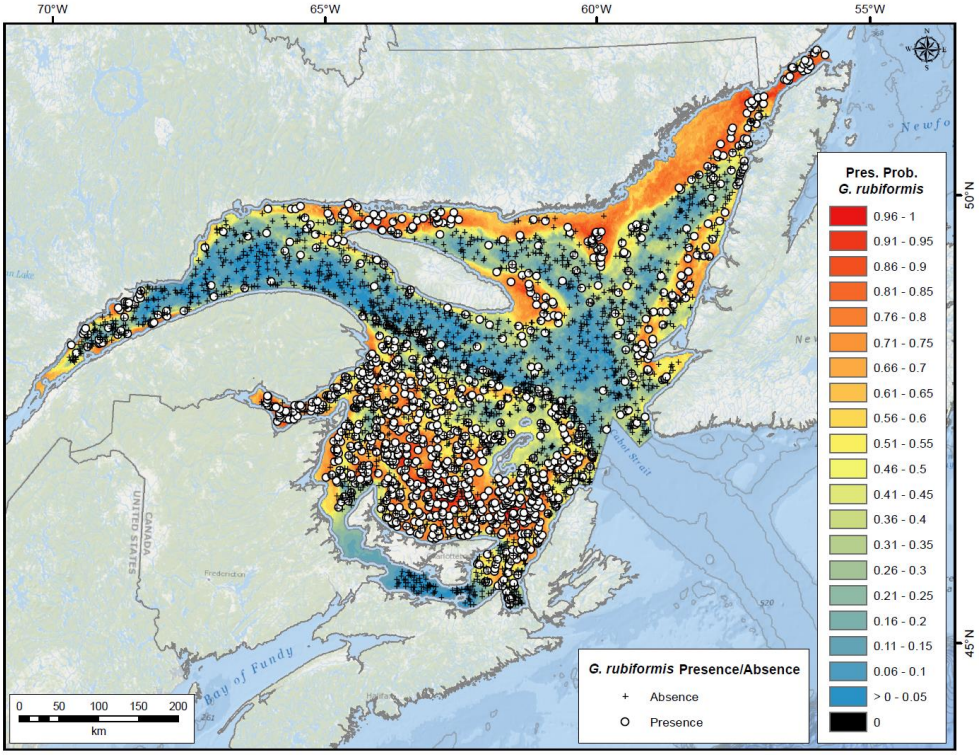

Figure 88. Total presence and absence observations and predictions of presence probability (pres. prob.) of *G. rubiformis* from the optimal random forest model (see Table 20) based on balanced data.

The actual presence and absence data observations (1056 presences and 1056 absences) used in the optimal model did not show important differences (Figure 89) with respect of the distribution of all the absence records across the study area (Figure 88). In this figure are also showed the areas of model extrapolation (i.e. areas where at least one environmental variable has values beyond its sampled range). The south of Prince Edward Island and the shallow portion in the northeast, before the Strait of Belle Isle are considered extrapolated area, as well as smaller areas scattered across the study area.

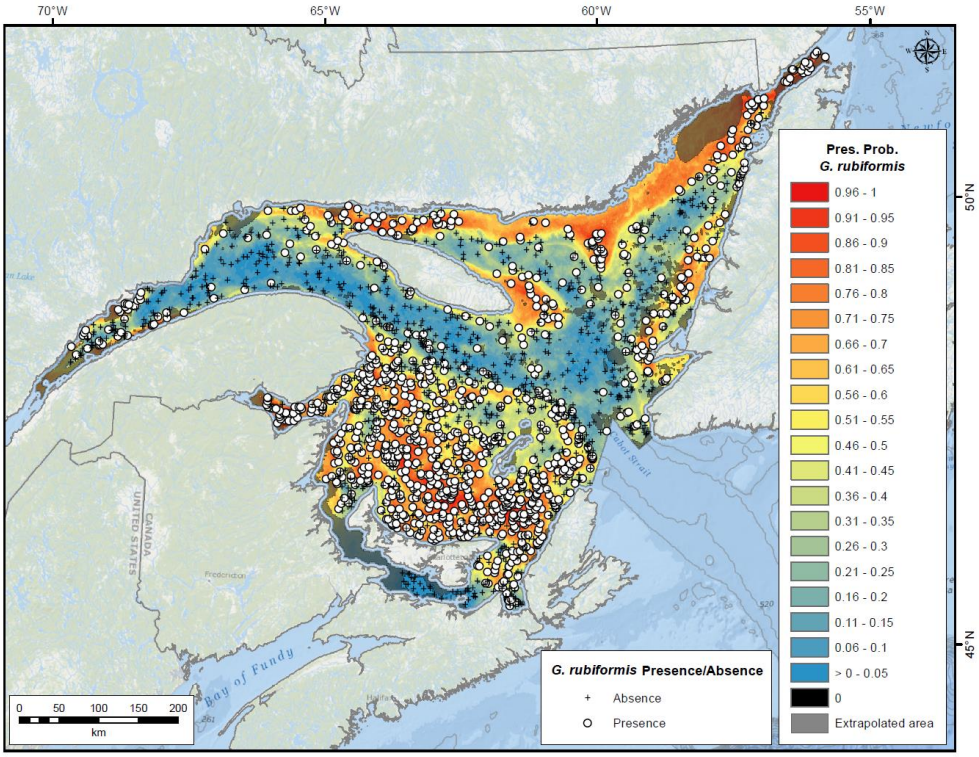

Figure 89. Presence and balanced absence observations and predictions of presence probability (pres. prob.) of *G. rubiformis* from the optimal random forest model (see Table 20) based on balanced data. The grey areas of extrapolation are also shown.

Of all 78 environmental predictor variables used in the model, Depth was the most important for the classification of the *G. rubiformis* presence and absence data (Figure 90). This variable was followed more distantly in terms of its Mean Decrease in Gini Value by Bottom Temperature Mean and other variables. The partial dependence plots for the top 6 most important predictors are shown in Figure 91. Presence probabilities of *G. rubiformis* were higher at depths values lower than 200 m.

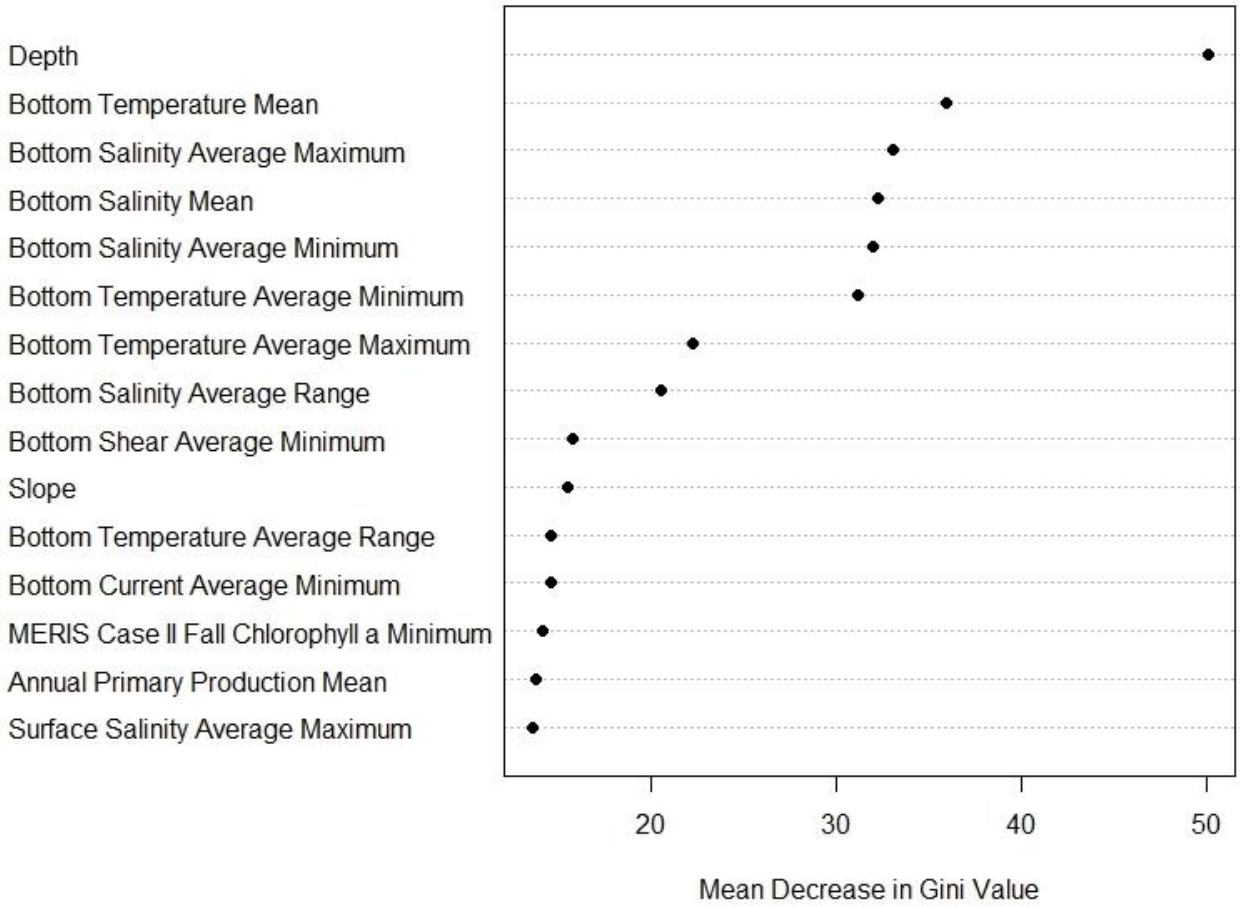

Figure 90. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model of *G. rubiformis* presence and absence data within the Gulf Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

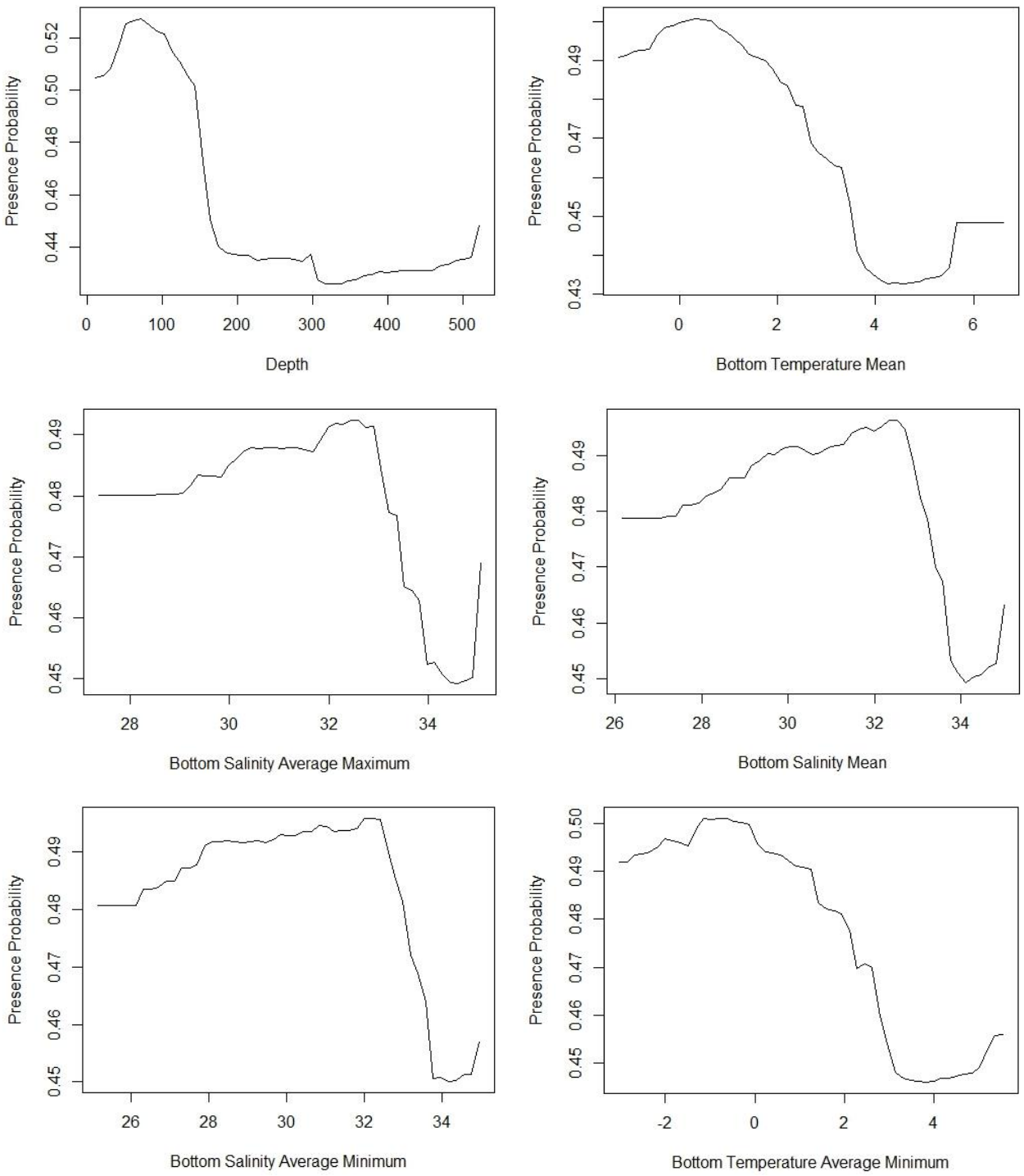

Figure 91. Partial dependence plots of the top six predictors of from the optimal random forest model of *G. rubiformis* presence and absence data collected within the Gulf Region, ordered left to right from the top.

Model 2 – Unbalanced species prevalence

Table 22 shows the accuracy measures for the random forest model on *G. rubiformis* presence and absence data (1056 presences and 2460 absences) and a threshold equal to species prevalence (0.3). The AUC calculated from Model 2 was 0.739, slightly lower than that of Model 1 (0.741). Class error of the presence class was higher than that of Model 1 although class error for the absence class was lower. Sensitivity was lower for Model 2, although specificity was higher.

Table 22. Accuracy measures and confusion matrix from 10-fold cross validation from random forest modelling of presence and absence of *G. rubiformis* within the Gulf Region with a cut-off threshold of 0.3.

Observations	Predictions		Total n	Class error	AUC	Sensitivity	Specificity
	Absence	**Presence**					
Absence	1754	706	2460	0.287	0.739	0.627	0.713
Presence	394	662	1056	0.373			

The predicted presence probability surface of *G. rubiformis* generated from Model 2 is shown in Figure 92. Similarly to Model 1, the highest predictions of presence probability occurred in the shallower areas of the Region. Areas of higher presence probability corresponded well with the spatial distribution of presence records (see Figure 93).

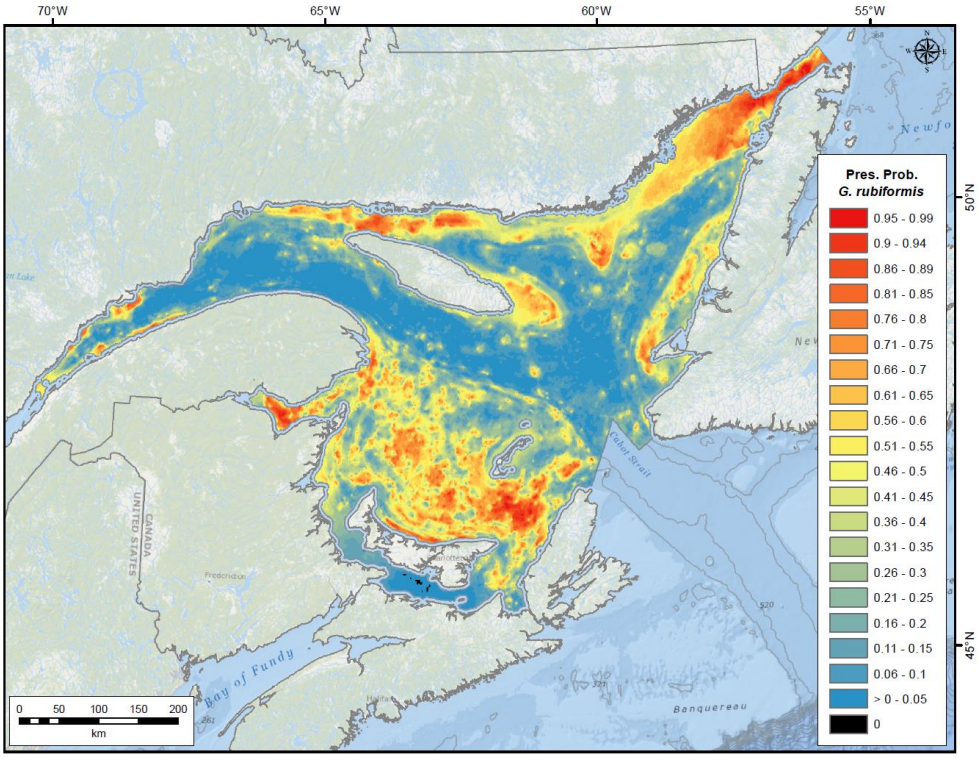

Figure 92. Predictions of presence probability (pres. prob.) of *G. rubiformis* from the unbalanced random forest model (see Table 22).

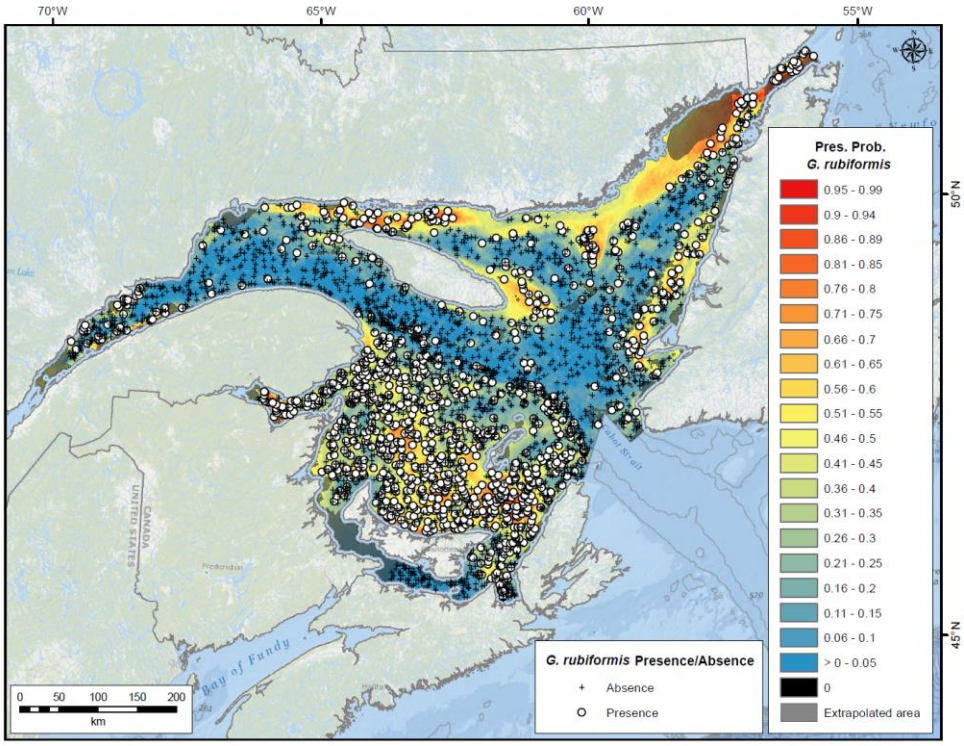

Figure 93. Presence and absence observations and predictions of presence probability (pres. prob.) of *G. rubiformis* from the unbalanced random forest model (see Table 22). The grey areas of extrapolation are also shown.

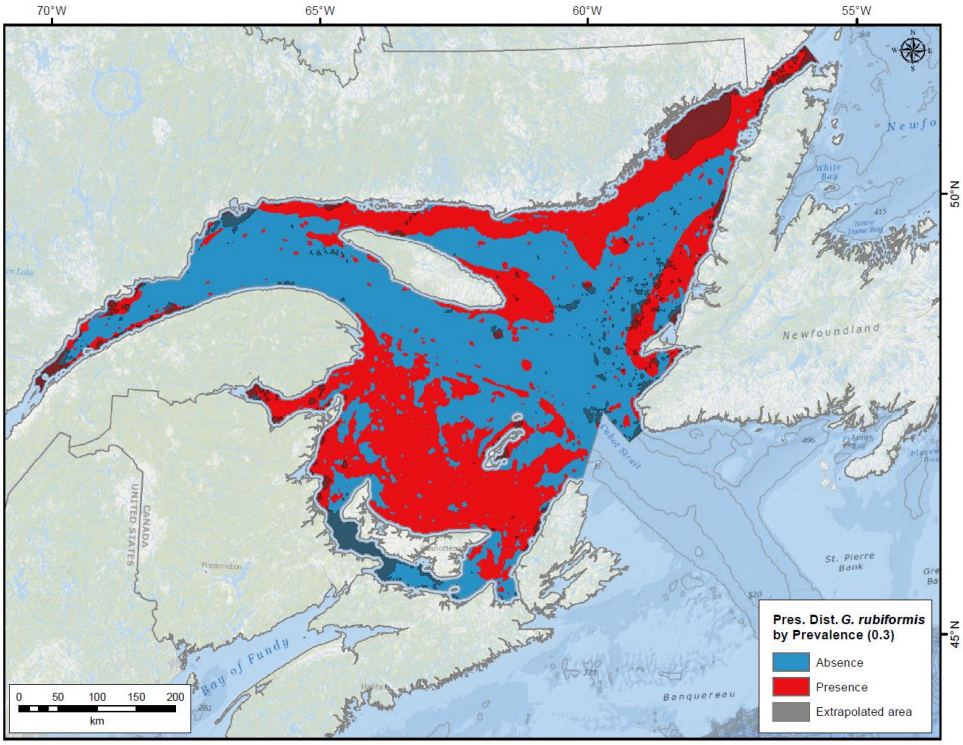

Figure 94. Predicted distribution of *G. rubiformis* from the unbalanced random forest model (see Table 22) using the prevalence cut-off threshold of 0.3. The grey areas of extrapolation are also shown, they appear dark red when overlain on the red presence surface.

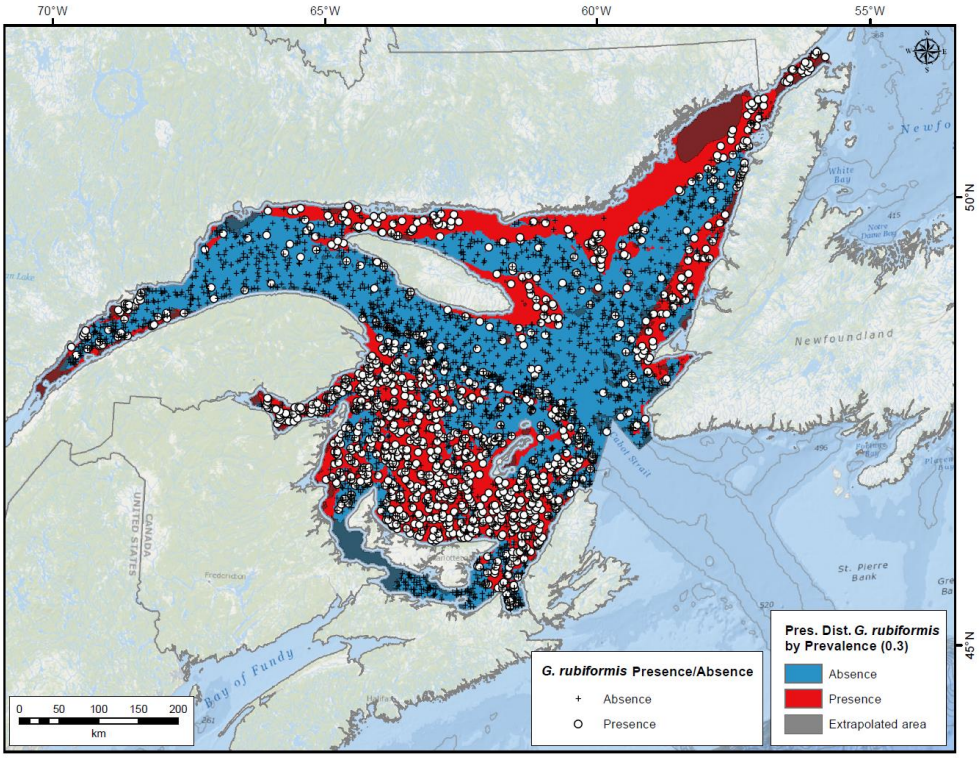

Figure 95. Presence and absence observations and predicted distribution of *G. rubiformis* from the unbalanced random forest model (see Table 22) using the prevalence cut-off threshold of 0.3. The grey areas of extrapolation are also shown, they appear dark red when overlain on the red presence surface.

The six most important environmental predictor variables in Model 2 were the same as that of Model 1 (Figure 96), and Depth and Bottom Temperature Mean were also the most important variables for the classification of *G. rubiformis* in Model 2. Partial dependence of *G. rubiformis* presence and absence data on the top 6 predictor variables is shown in Figure 97 and show similar results to those for Model 1.

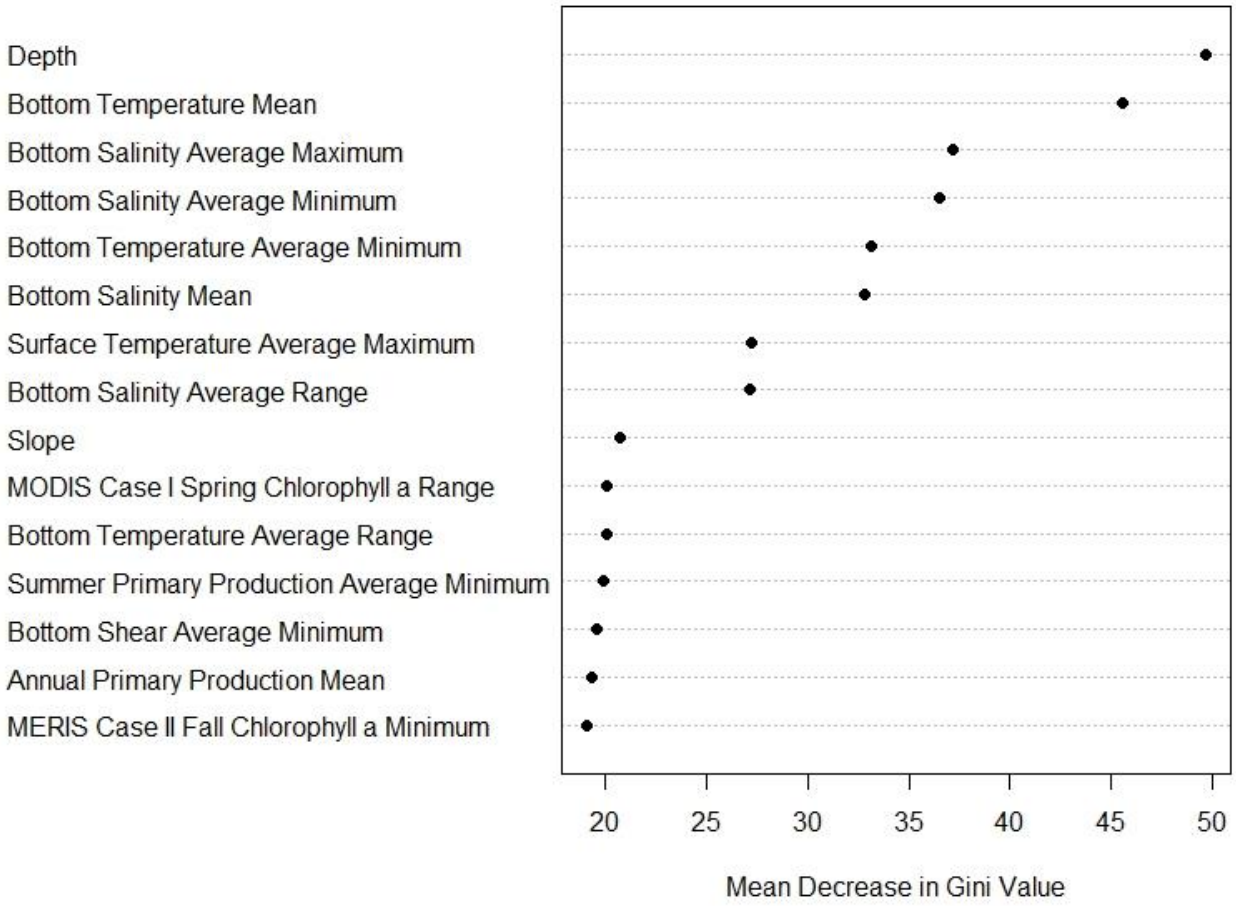

Figure 96. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the unbalanced random forest model of *G. rubiformis* presence and absence data within the Gulf Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

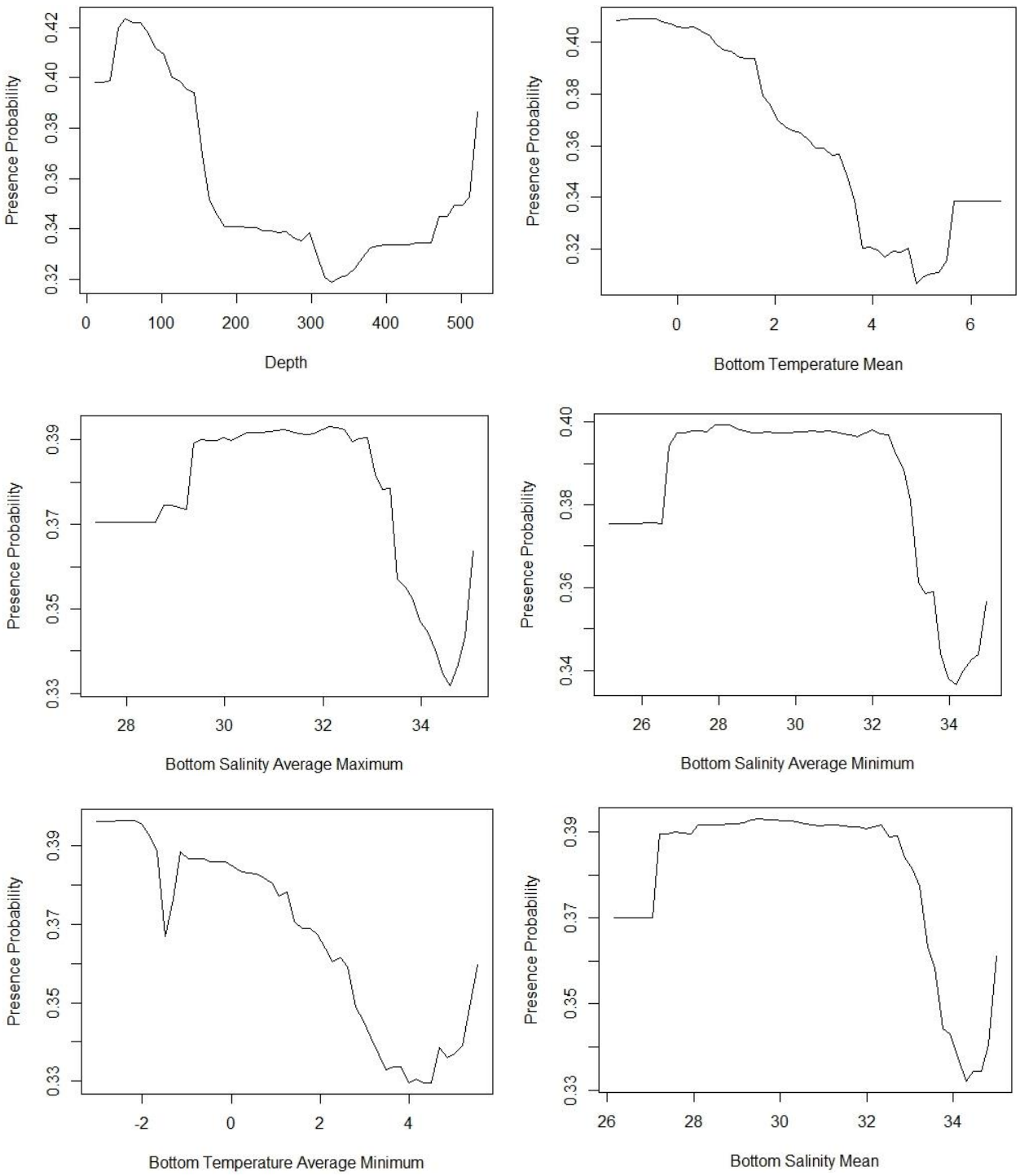

Figure 97. Partial dependence plots of the top six predictors of from the unbalanced random forest model of *G. rubiformis* presence and absence data collected within the Gulf Region, ordered left to right from the top.

Model Selection

Model accuracy measures were very similar for both models, although Model 1 presented slightly higher AUC and sensitivity and lower specificity. The selection of Model 1, based on balanced data, allows for the use of the presence probabilities considering values higher than 0.5 as presences and lower as absences to maintain fairly constant error rates.

Prediction of *G. rubiformis* Biomass Using Random Forest

a) Southern Gulf of St. Lawrence

The accuracy measures of the regression random forest model on mean *G. rubiformis* biomass per grid cell for the southern Gulf of St. Lawrence are presented in Table 23. The average R-square (R^2) value was 0.420 ± 0.287, the average Root-Mean-Square Error (RMSE) was 0.209 ± 0.098, and the average percent variance explained was 18.26 ± 6.28. The high variability around the mean for R^2 indicates high variability of the model fit.

Table 23. Accuracy measures from 10-fold cross validation from random forest modelling of biomass of *G. rubiformis* within the southern Gulf of St. Lawrence. NRMSE, is the normalized root-mean square deviation (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	%variance explained
1	0.373	0.294	0.010	13.90
2	0.003	0.210	0.008	26.25
3	0.110	0.264	0.007	21.67
4	0.547	0.165	0.074	18.16
5	0.703	0.101	0.009	12.84
6	0.403	0.174	0.004	14.49
7	0.037	0.179	0.006	20.20
8	0.596	0.407	0.003	29.91
9	0.611	0.227	0.007	10.04
10	0.820	0.066	0.007	15.16
Mean	**0.420**	**0.209**	**0.005**	**18.26**
SD	**0.287**	**0.098**	**0.002**	**6.28**

Figures 98 and 99 show the surface of *G. rubiformis* biomass (kg) predictions per grid cell generated from the random forest model in the southern Gulf of St. Lawrence. The maximum biomass was predicted in the coastal areas in the southern part.

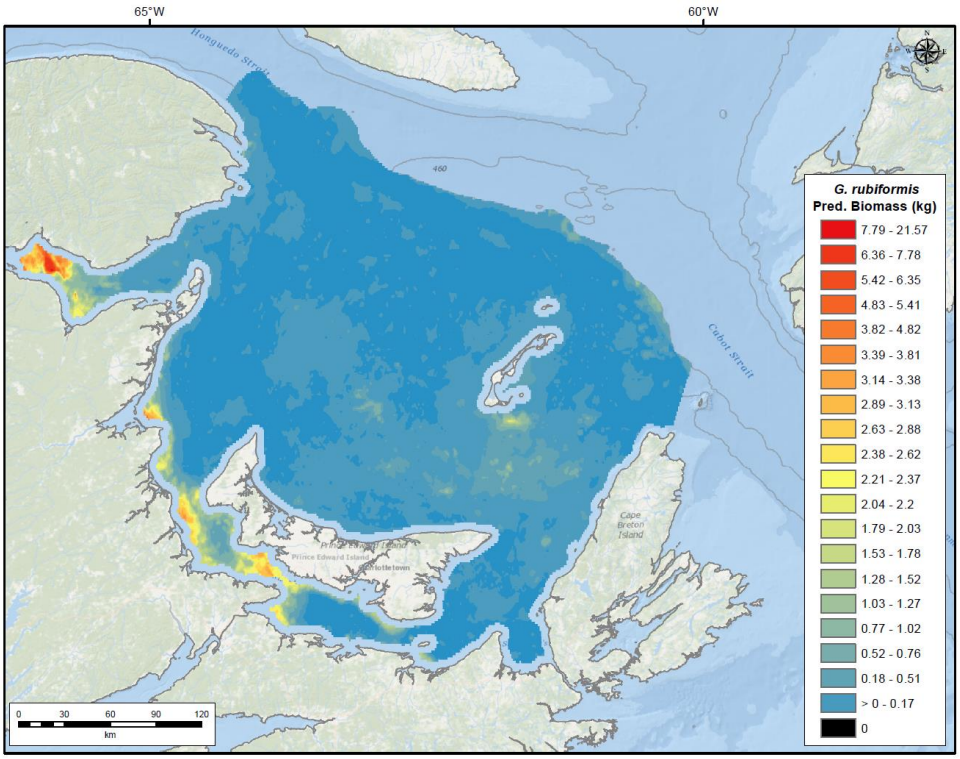

Figure 98. Prediction of *G. rubiformis* biomass (kg) from the random forest model (see Table 23) within the southern Gulf of St. Lawrence.

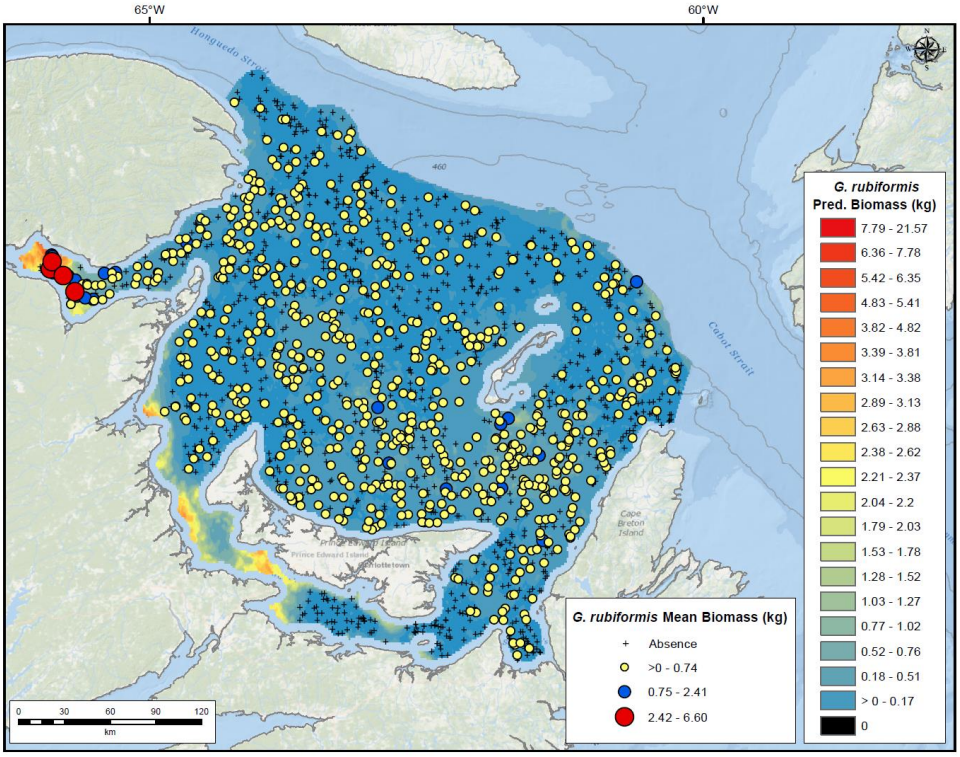

Figure 99. *G. rubiformis* biomass catches overlaid to the *G. rubiformis* biomass (kg) predicted from the random forest model (see Table 23) within the southern Gulf of St. Lawrence.

The top 15 most important environmental variables for predicting *G. rubiformis* biomass in the southern Gulf of St. Lawrence are shown in Figure 100. Maximum Average Spring Mixed Layer Depth was the most important variable in the model, followed by Maximum Average Winter Mixed Layer Depth. The partial dependence of *G. rubiformis* biomass on the top 6 most important variables is shown in Figure 101. Maximum Average Spring Mixed Layer Depth predicted highest biomass at values around 11 m depth.

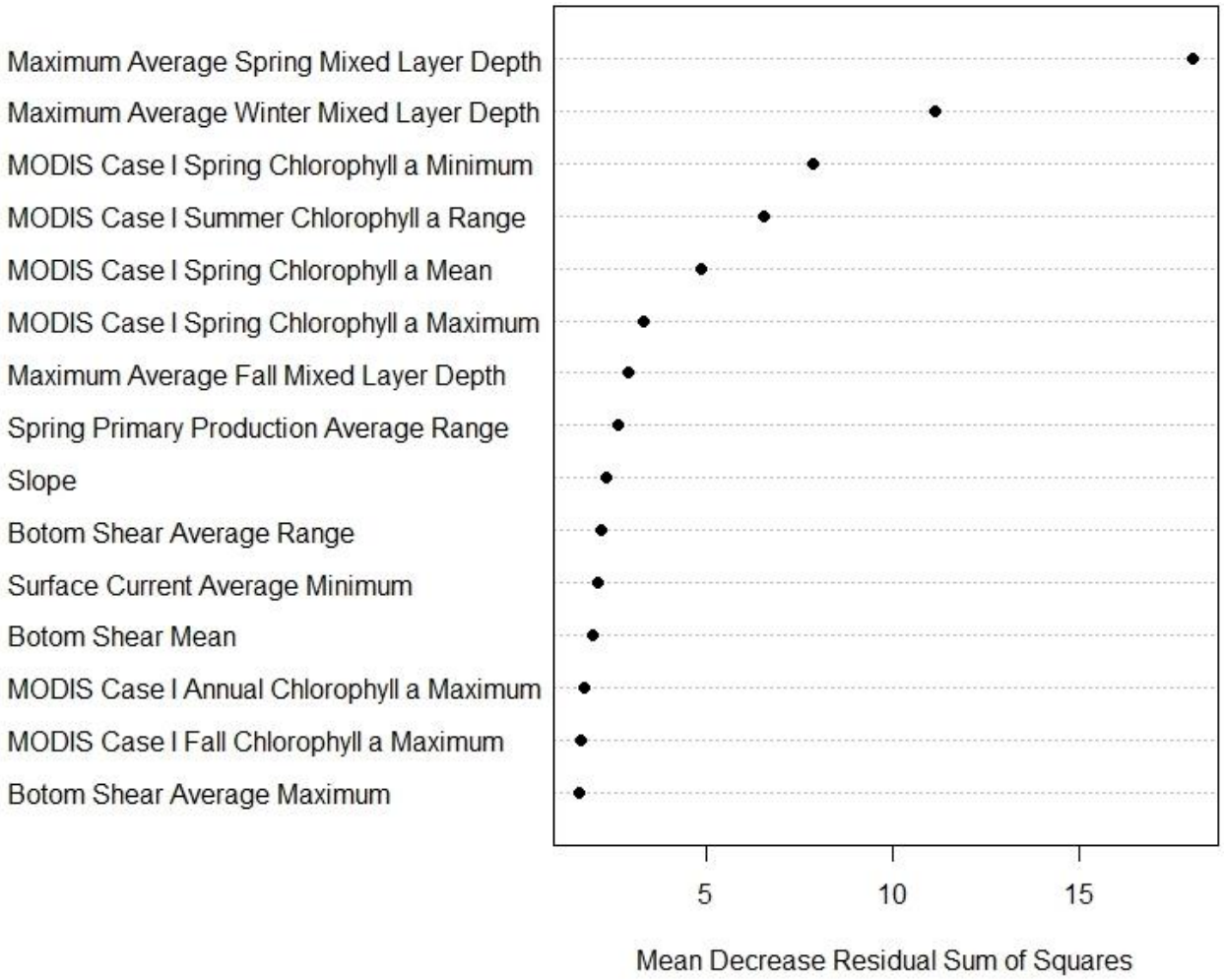

Figure 100. Importance of the top 15 predictor variables measured as the Mean Decrease Residual Sum of Squares of the random forest model (see Table 23) of *G. rubiformis* biomass data within the southern Gulf of St. Lawrence.

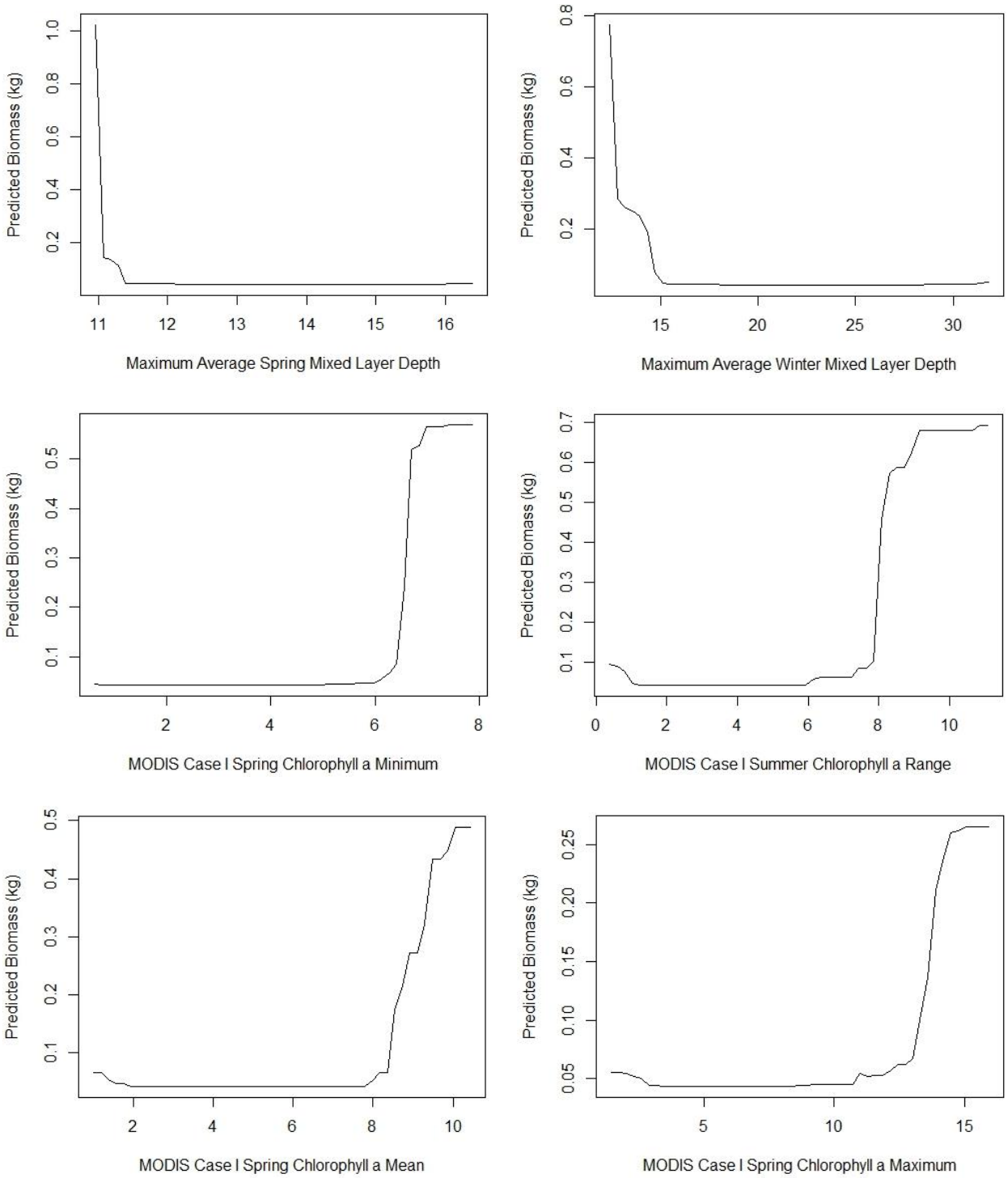

Figure 101. Partial dependence plots of the top six predictors of the random forest model (see Table 23) of *G. rubiformis* biomass data within the southern Gulf of St. Lawrence, ordered left to right from the top.

Model 4 – Northern Gulf Region

The accuracy measures of the regression random forest model on mean *G. rubiformis* biomass per grid cell for the northern Gulf of St. Lawrence are presented in Table 24. The average R-square (R^2) value was 0.192 ± 0.274, the average Root-Mean-Square Error (RMSE) was 0.042 ± 0.038, and the average percent variance explained was 7.002 ± 4.190. The high variability around the mean for both of these metrics indicates high variability of the model fit.

Table 24. Accuracy measures from 10-fold cross validation from random forest modelling of biomass of *G. rubiformis* within the northern Gulf of St. Lawrence. NRMSE, is the normalized root-mean square deviation (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	%variance explained
1	0.003	0.063	0.033	15.500
2	0.542	0.021	0.011	2.820
3	0.287	0.142	0.075	2.920
4	0.024	0.023	0.012	5.660
5	0.029	0.043	0.023	7.180
6	0.004	0.014	0.007	7.790
7	0.050	0.018	0.009	7.150
8	0.038	0.032	0.017	11.880
9	0.800	0.020	0.011	1.970
10	0.137	0.045	0.024	7.150
Mean	**0.192**	**0.042**	**0.022**	**7.002**
SD	**0.274**	**0.038**	**0.020**	**4.190**

Figures 102 and 103 show the surface of *G. rubiformis* biomass (kg) predictions per grid cell generated from the random forest model in the northern Gulf of St. Lawrence. The biomass predicted was low in most of the area except in some shallow areas in the estuary of the St. Lawrence River.

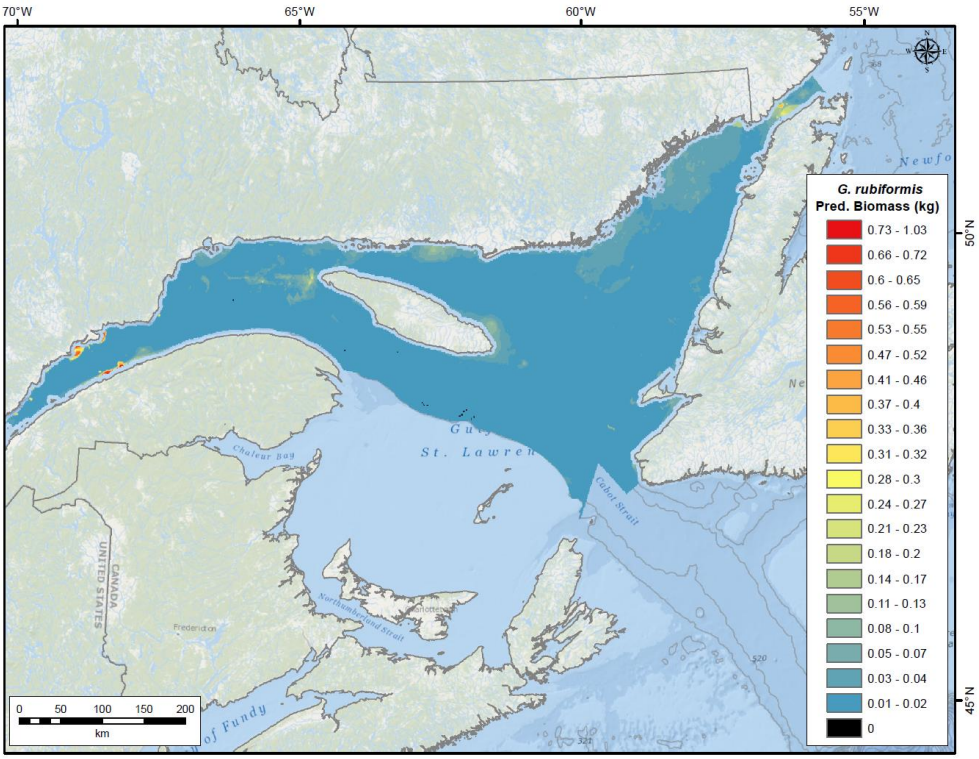

Figure 102. Predictions of *G. rubiformis* biomass probability from the random forest model (see Table 24) within the northern Gulf of St. Lawrence.

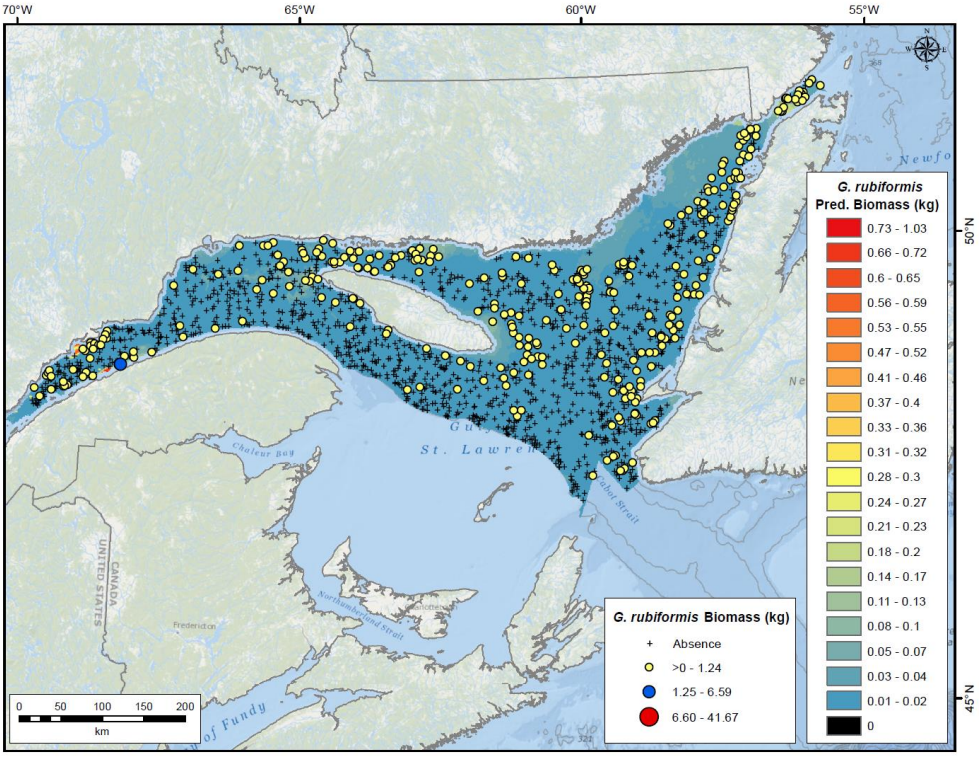

Figure 103. *G. rubiformis* biomass catches overlaid to the *G. rubiformis* biomass probability from the random forest model (see Table 24) within the northern Gulf of St. Lawrence.

The top 15 most important environmental variables for predicting *G. rubiformis* biomass in the northern Gulf of St. Lawrence are shown in Figure 104. The first five top environmental variables are related to MODIS Case I Chlorophyll *a*, being the MODIS Case I Chlorophyll *a* Range the most important variable in the model. The partial dependence of *G. rubiformis* biomass on the top 6 most important variables is shown in Figure 105. Predicted biomass of *G. rubiformis* increased when MODIS Case I Chlorophyll *a* Range was higher than 9 mg m^{-3}.

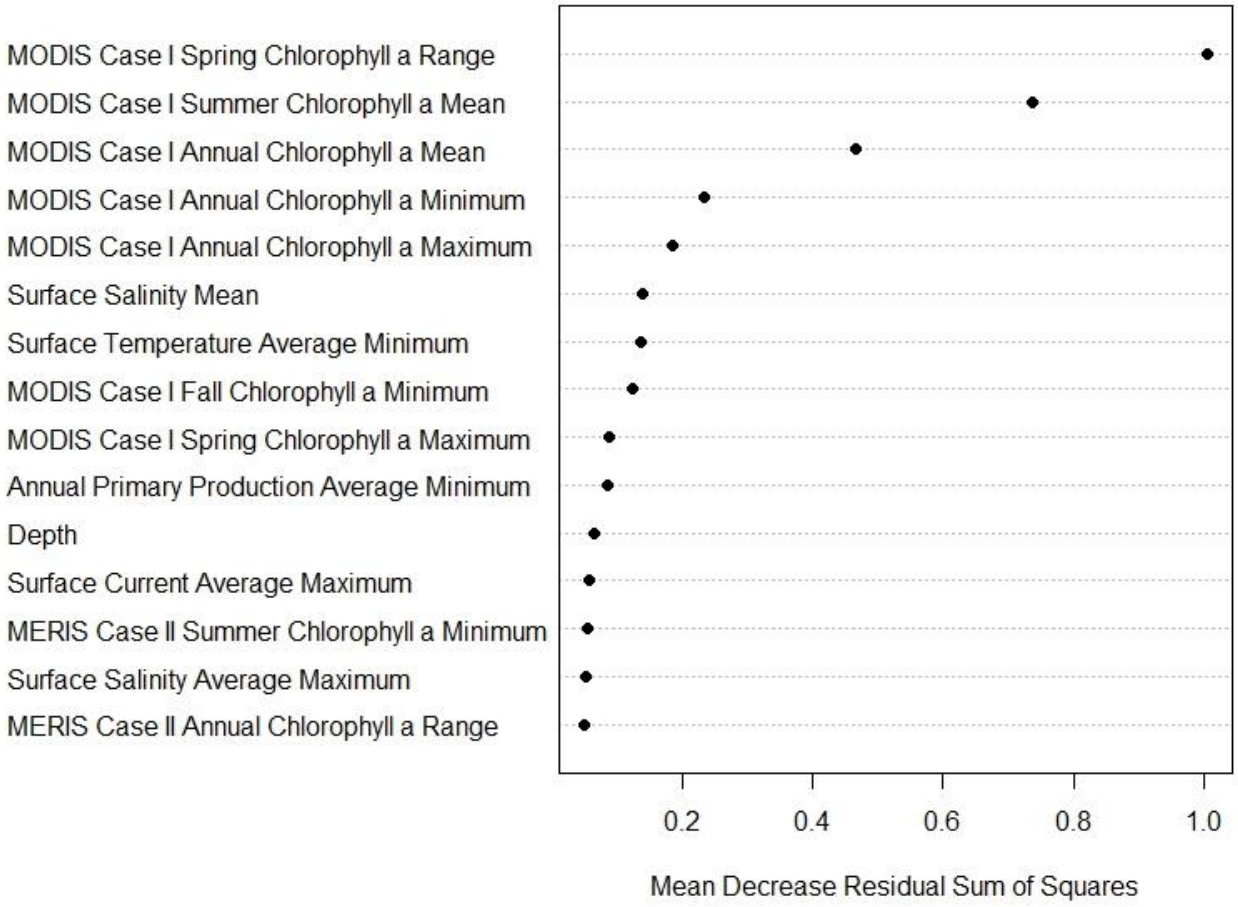

Figure 104 Importance of the top 15 predictor variables measured as the Mean Decrease Residual Sum of Squares of the random forest model (see Table 24) of *G. rubiformis* biomass data within the northern Gulf of St. Lawrence.

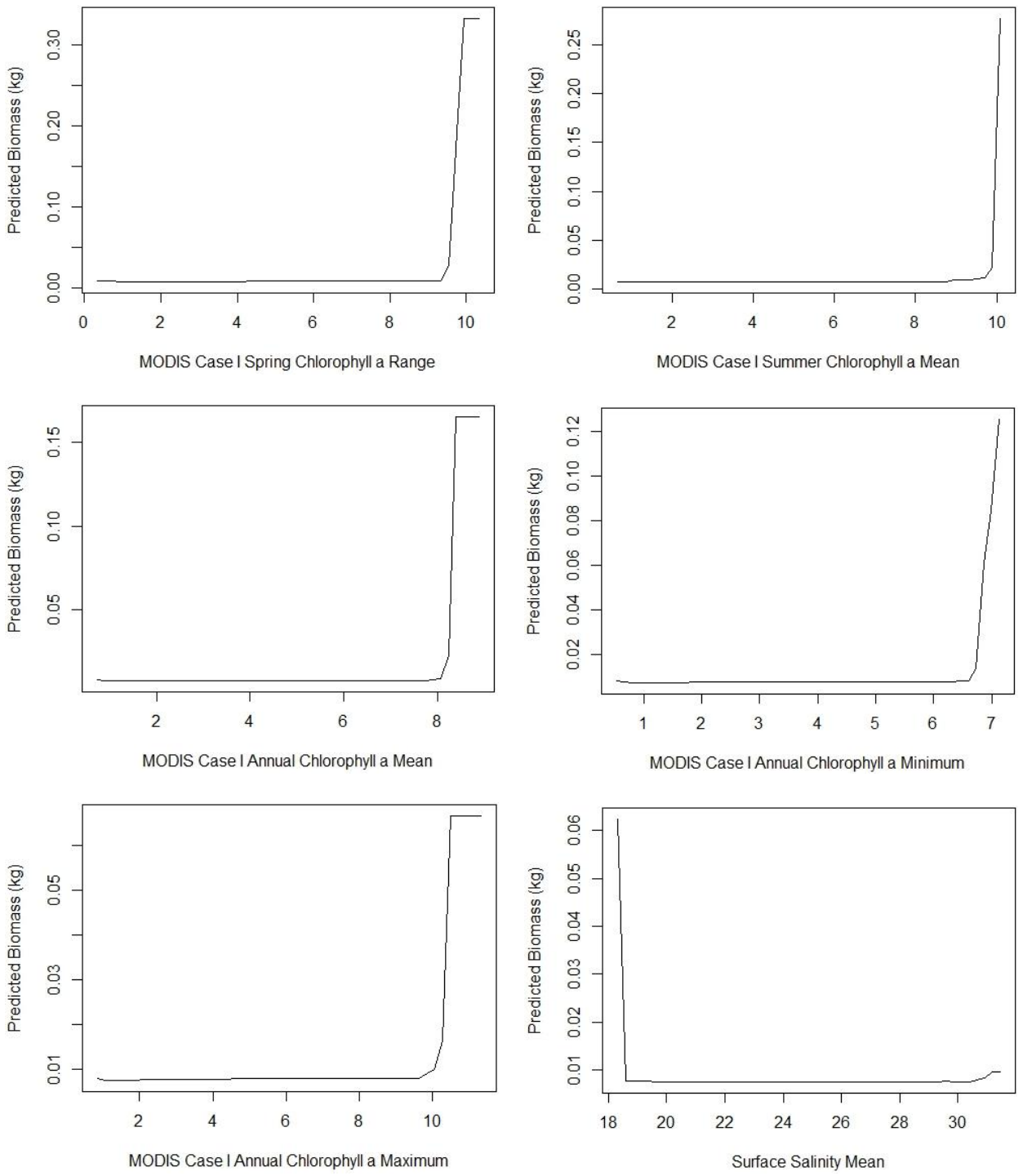

Figure 105. Partial dependence plots of the top six predictors of the random forest model (see Table 24) of *G. rubiformis* biomass data within the northern Gulf of St. Lawrence, ordered left to right from the top.

Figures 106 and 107 show the surface of *G. rubiformis* biomass (kg) predictions per grid cell generated from the random forest model in the entire Gulf of St. Lawrence.

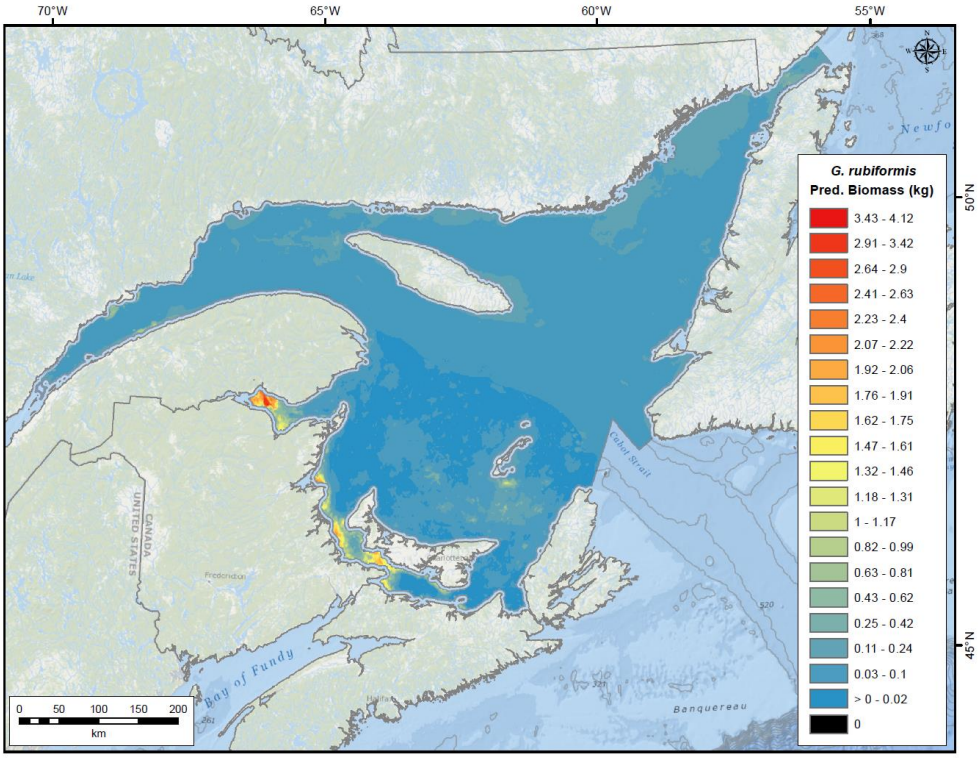

Figure 106. Predictions of *G. rubiformis* biomass from the random forest model within the Gulf of St. Lawrence.

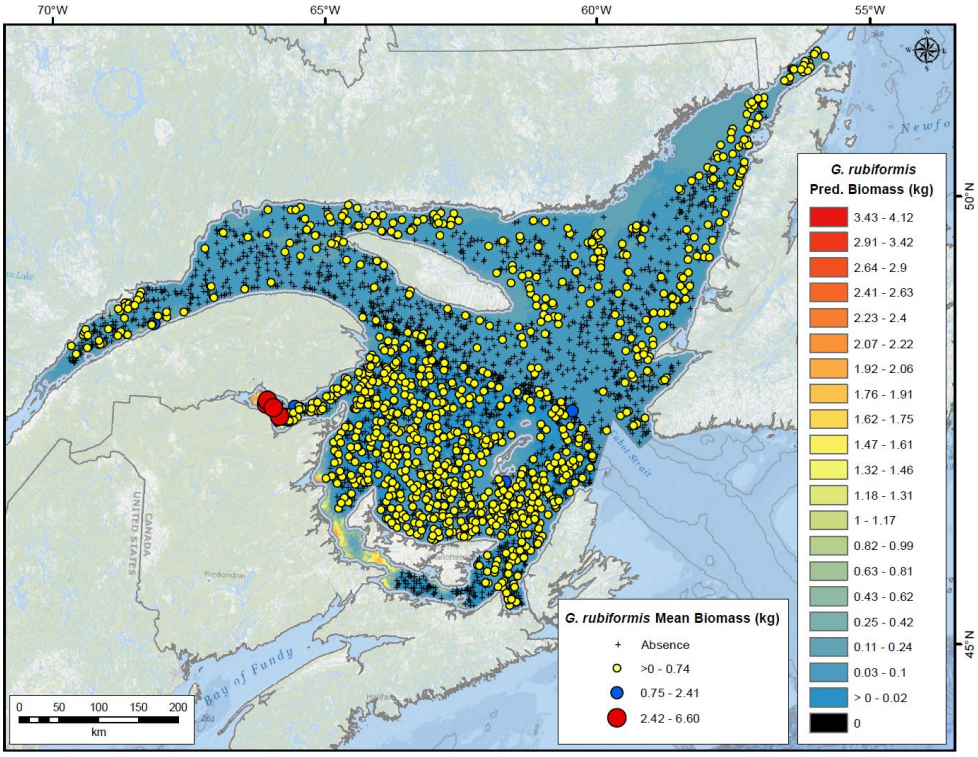

Figure 107. *G. rubiformis* biomass catches overlaid to the *G. rubiformis* biomass predicted from the random forest model within the Gulf of St. Lawrence.

DISCUSSION

Random forest-generated species distribution models have been produced for sea pens (Pennatulacea), sponges (Porifera), stalked tunicates (*Boltenia ovifera*) and soft corals (*Gersemia rubiformis*). Models for presence-absence response data performed very well for sea pens and stalked tunicates and performed well for soft corals and sponges (Table 25). Prediction surfaces for the four groups produced clearly-defined areas of high occurrence probability. For the models based on biomass, soft corals and sea pens had the highest R^2 values in the southern Gulf of St. Lawrence and stalked tunicates and sea pens in the north. Sponges had R^2 values less than 0.1 in both areas indicating very bad model performance. The group "sponges" includes a highly diverse set of species as we can see in Table 1 that can have different environmental requirements making it more difficult to find a common response to the predictors than for the other taxa modelled in this study. In the case of sea pens, different species are also included (see Table 1) but all of them have similar environmental requirements (Williams, 1995) and they constitute extensive aggregations (sea pens fields) in suitable conditions (Fuller et al., 2008).

Table 25. Accuracy measures for the models selected for each group based on presence-absence and on biomass data.

	Presence-Absence			**Biomass**			
				Southern Gulf		**Northern Gulf**	
Taxon	**AUC**	**Sensitivity**	**Specificity**	**R^2**	**NRMSE**	**R^2**	**NRMSE**
Sea pens	0.91	0.840	0.822	0.37	0.04	0.27	0.04
Sponges	0.71	0.646	0.662	0.07	0.02	0.03	0.04
Stalked tunicates	0.89	0.837	0.805	0.18	0.05	0.41	0.03
Soft corals	0.75	0.722	0.674	0.42	0.01	0.19	0.02

Recently, Moritz et al. (2013) have modelled the spatial distribution of epibenthic communities in the Lower Estuary and Northern Gulf of St. Lawrence using generalized linear models. Predicted high and low suitability zones for six benthic communities were mapped in relation to two sets of environmental variables. Most of the models had a good performance with AUC values higher than 0.8. Of the taxonomic groups modelled in this report, only sea pens (Pennatulacea) and sponges (Porifera) were present as dominant taxa in some of the benthic communities groups identified by Moritz et al. (2013). Sea pens were included in the five dominant taxa in 5 of the 6 communities, and in the group B (estuarine community) they were found with the highest frequency of occurrence (91%). This group was characterized by low salinity, poor oxygen saturation and turbidity. This distribution is similar to the distribution found in the presence probability maps (Figures 10 and 15) where the Bottom Salinity Average Minimum was the main variable in importance in the balanced model and Depth in the unbalanced model. Sponges were included as one of the five dominant taxa in 3 of the 6 communities, and the highest frequency was 67% in group A. Sponges are found in almost all of the area in our study and due to the poor resolution of this group no further comparisons can be done.

Large numbers of collinear predictors can create different problems (such as inaccurate model parameterization or exclusion of significant predictors during model creation) and make models

more difficult to interpret (Graham, 2003). In the case of RF, the model is obtained by combining base models trained on different bootstrap replicate samples of the data together and only a random subset of the available variables is used for the candidate splitting variables at each node, alleviating the problem of correlated variables (Breiman, 1996). This has allowed us to use all of the environmental variables in the RF model and the performance obtained has produced better results than when a subset of predictors was considered (i.e. correlated less than a threshold). Nevertheless, the interpretability of the model can still be difficult. From the partial plots derived from the RF models it can be observed that the environmental variable values from which the presence probability (or biomass in regression models) increases towards the maximum on the *x* axis. These values may indicate an environmental barrier or environmental ranges in the distribution of the species. However, we caution that when such values are observed, the interpolated prediction surface of the environmental variable and the surface of the prediction standard error along with the Q-Q plots generated (Beazley et al., 2016) should be considered together to interpret the results. In the process of interpolation some of the areas can be over- or under-estimated which could affect the posterior results during the modelling process.

Although it was not our goal to identify the specific niche requirements of each taxon considered, some preliminary conclusions can be obtained from the top predictor of each model (Table 26). Predictors relating to the physical environment, such as bottom temperature and salinity, were the most important predictors in presence-absence models, whereas those relating to biological factors, such as primary production or chlorophyll *a*, where the main predictors in the biomass models. This could be explained by the physical environment being a limiting factor for the presence of a taxon, but its development or growth could be more related to biological factors controlling food availability which is consistent with the biology of these groups of sessile filter feeders.

Table 26. Top important environmental variable from each RF model. For presence-absence only the model selected between balanced and unbalanced data is considered.

	Presence-Absence	**Biomass**	
Taxon		**Southern Gulf**	**Northern Gulf**
Sea Pens	Bottom Salinity Average Minimum	Spring Primary Production Average Minimum	Surface Temperature Average Maximum
Sponges	Depth	MERIS Case II Spring Chlorophyll *a* Range	Maximum Average Summer Mixed Layer Depth
Stalked tunicates	Bottom Temperature Mean	MERIS Case II Fall Chlorophyll *a* Minimum	Surface Primary Production Mean
Soft corals	Depth	Maximum Average Spring Mixed Layer Depth	MODIS Case I Spring Chlorophyll *a* Range

The number of available methods for predicting species' distribution has increased in response to the demand for predictive modelling in recent years. According to Knudby et al. (2010), RF is superior in predicting species richness and diversity, whereas GAM is superior in predicting biomass. Some authors suggest the use of an ensemble approach where different methods are applied (Bučas et al., 2013). We have compared the results obtained from RF and GAM in the case of sea pens and additional GAM models for the other taxa are presented in Appendix 1. RF

biomass models performed similar or better than GAM models when latitude and longitude were not included as predictor variables. However, including latitude and longitude as a tensor product in the GAM models increased the R^2 but the prediction capacity outside the range of data observations decreased. Depending on the objective of the study, it can be useful to use latitude and longitude as predictors. In the case of refining areas already identified by other means, such as Kernel Density Estimation (Kenchington et al., 2016), it seems that either the RF or the GAM models considering geographic distance give accurate results and are able to map sea pen habitat when a threshold is defined by the KDE analysis in advance. However, if the objective is to predict biomass in new areas where no data are available, the GAM without latitude and longitude seemed to generalize better than RF. RF is a robust statistical method that requires very little input from the user and is able to learn patterns in the data automatically, whereas in GAM models several parameters need to be defined, such as the distribution type or the degrees of freedom of the smothers. In addition, when latitude and longitude are considered in a GAM model as a tensor product the model becomes computationally costly and most of the explanatory power related to the environmental variables is collected by these predictors, making interpretability of the model difficult. In the Gulf of St. Lawrence where good coverage of the response data was available, the main objective was to present SDM-maps that can be used to refine the significant polygons (Figure 108). For this objective it seems that RF and GAM models provide similar results and are both good options, although the fewer assumptions required for RF makes this method more convenient.

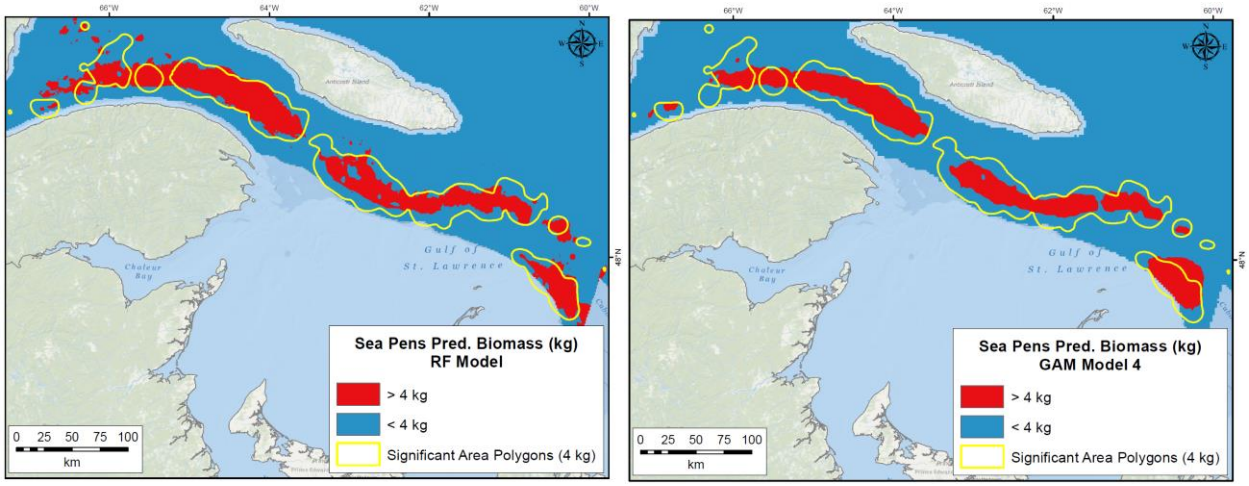

Figure 108. Sea pen biomass predicted (higher and lower than 4 kg) overlaid with the significant area polygons from Kenchington et al. (2016) in the northern Gulf of St. Lawrence. Left panel (RF Model output); Right panel (GAM Model 4 output).

The models provided in this study do not consider the effect of disturbance by human activities. The Gulf of St. Lawrence has been subjected to the effects of fisheries for centuries (Perley, 1849). Predicted biomass and to a lesser degree distribution can therefore be confounded by fishing activities, and areas that are physically suitable but are predicted to have low occurrence or biomass may not necessarily indicate bad model performance. Three of the taxa considered in this report (sea pens, sponges, and stalked tunicates) are vulnerable marine ecosystems (VME) indicators (Murillo et al., 2011) and are highly aggregating, structure-forming megafaunal groups

that can be found in ‘significant concentrations’ constituting VME (Kenchington et al., 2014). The life-history traits of these species, such as slow growth rates, late age of maturity, or their structural complexity make them very vulnerable to fishing activities (FAO, 2009). In order to consider how anthropogenic pressure has influenced these ecosystems, a measure of this, such as fishing intensity, should be included as a predictor variable in the RF models and the effects of changes in the pressure explored (Bergstrom et al. 2013). This kind of analysis would point out potential species distribution and could indicate areas for future restoration initiatives helping the implementation of the Policy for Managing the Impact of Fishing on Sensitive Benthic Areas developed by DFO in 2009 to ensure Canadian fisheries are conducted in a manner that supports marine conservation and sustainable resource use within and outside Canada's 200 nautical mile exclusive economic zone.

ACKNOWLEDGMENTS

This project was funded in part by a one year project under DFO’s Strategic Program for Ecosystem-Based Research and Advice (SPERA) to EK and through financial support by DFO’s Oceans Management Program, Gulf Region. We thank Ray MacIsaac for the latter contribution and guidance in the preparation of this report. We thank C. Rooper (NMFS - RACE Division, Washington, USA) and K. Tanaka (U of Maine, Maine, USA) for their constructive comments on SDM and GAMs.

REFERENCES

Beazley, L., Lirette, C., Sabaniel, J., Wang, Z., Knudby, A., and Kenchington, E. 2016. Characteristics and Environmental Data Layers for Use in Species Distribution Modelling in the Gulf of St. Lawrence. Can. Tech. Rep. Fish. Aquat. Sci. 3154: viii + 357 p.

Bergström, U., Sundblad, G., Downie, A.-L., Snickars, M., Bostöm, C., and Lindegarth, M. 2013. Evaluating eutrophication management scenarios in the Baltic Sea using species distribution modelling. J. Appl. Ecol. 50: 680–690.

Bučas, M., Bergström, U., Downie, A-L., Sundblad, G., Gullström, M., von Numers, M., Šiaulys, A., and Lindegarth, M. 2013. Empirical modelling of benthic species distribution, abundance, and diversity in the Baltic Sea: evaluating the scope for predictive mapping using different modelling approaches. ICES J. Mar. Sci. doi: 10.1093/icesjms/fst036.

Breiman, L. 1996. Bagging Predictors. Mach. Learn. 24: 123–140.

Breiman, L. 2001. Random Forests. Mach. Learn. 45: 5–32.

Chen, C., Liaw, A., and Breiman, L. 2004. Using random forest to learn imbalanced data. University of California, Berkeley. 12 p.

Cutler, D.R., Edwards Jr., T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., and Lawler, J.J. 2007. Random Forest for classification in Ecology. Ecology 88: 2783–2792.

Davies, A.J., Wisshak, M., Orr, J.C., and Roberts, J.M. 2008. Predicting suitable habitat for the cold-water coral *Lophelia pertusa* (Scleractinia). Deep-Sea Res. I 55: 1048–1062.

Davies, A.J., and Guinotte, J.M. 2011. Global habitat suitability for framework-forming coldwater corals. PLoS ONE 6(4): e18483. doi:10.1371/journal.pone.0018483

DFO. 2007. Assessment of cod stock in the northern gulf of St. Lawrence (3Pn, 4RS) in 2006. DFO Canadian Science Advisory Secretariat Science Advisory Report 2007/003. 14 p.

Dolan, M.F.J., Grehan, A.J., Guinan, J.C., and Brown, C. 2008. Modelling the local distribution of cold-water corals in relation to bathymetric variables: Adding spatial context to deep-sea video data. Deep-Sea Res. I 55: 1564–1579.

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.

Dunn, P. K., and Smyth, G. K. 1996. Randomized quantile residuals. J. Comput. Graph. Stat. 5: 236–244.

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.

ESRI, 2011. ArcGIS Desktop: Release 10. Environmental Systems Research Institute, Redlands, CA.

Evans J.S., Murphy, M.A., Holden, Z.A., and Cushman, S.A. 2011. Modeling Species Distribution and change Using Random Forests. *In* Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications. Edited by C.A Drew, Y.F. Wiersma, and F. Huettmann. Springer, New York. pp. 139–159.

FAO. 2009. International Guidelines for the Management of Deep-sea Fisheries in the High Seas. FAO, Rome. 73p.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recog. Lett. 27: 861–874.

Franklin, J. 2009. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press, Cambridge, U.K. 338 p.

Fuller, S.D., Murillo Perez, F.J., Wareham, V., and Kenchington, E. 2008. Vulnerable Marine Ecosystems Dominated by Deep-Water Corals and Sponges in the NAFO Convention Area. Ser No N5524. NAFO SCR Doc 08/22. 24 p.

Graham, M. 2003. Confronting multicollinearity in ecological multiple regression. Ecology 84: 2809-2815. doi:10.1890/02-3114.

Guinan, J., Grehan, A.J., Dolan, M.F.J., and Brown, C. 2009. Quantifying relationships between video observations of cold-water coral cover and seafloor features in Rockall Trough, West of Ireland. Mar. Ecol. Prog. Ser. 375: 125–138.

Guisan, A., and Zimmerman, N.E. 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135: 147–186.

Hanberry, B.B., and He, H.S. 2013. Prevalence, statistical thresholds and accuracy assessment for species distribution models. Web Ecology 13: 13–19.

Hastie, T., and Tibshirani, R. 1986. Generalized additive models. Stat. Sci. 1: 297–318.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. 2005. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Second Edition. Springer+Verlag.

Herrick, K.K., Huettmann, F., and Lindgren, M.A. 2013. A global model of avian influenza prediction in wild birds: the importance of northern region. Vet. Res. 44: 42.

Jiménez-Valverde, A. and Lobo, J. M. 2006. The ghost of unbalanced species distribution data in geographical model predictions. Divers. Distrib. 12: 521–524.

Kenchington, E., Lirette, C., Cogswell, A., Archambault, D., Archambault, P., Benoit, H., Bernier, D., Brodie, B., Fuller, S., Gilkinson, K., Lévesque, M., Power, D., Siferd, T., Treble, M., and Wareham, V. 2010. Delineating coral and sponge concentrations in the biogeographic regions of the east coast of Canada using spatial analyses. Can. Sci. Advis. Sec. Res. Doc. 2010/041: vi + 202 p.

Kenchington, E., Murillo, F.J., Lirette, C., Sacau, M., Koen-Alonso, M., Kenny, A., Ollerhead, N., Wareham, V., and Beazley, L. 2014. Kernel density surface modelling as a means to identify significant concentrations of vulnerable marine ecosystem indicators. *PLoS ONE* 9(10): e109365. doi:10.1371/journal.pone.0109365

Kenchington, E., Lirette, C., Murillo, F.J., Beazley, L., Guijarro, J., Wareham, V., Gilkinson, K., Koen Alonso, M., Benoît, H., Bourdages, H., Sainte-Marie, B., Treble, M., and Siferd, T. 2016. Kernel Density Analyses of Coral and Sponge Catches from Research Vessel Survey Data for Use in Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3167: viii + 207 p.

Knudby, A., Brenning, A., and LeDrew, E. 2010. New approaches to modelling fish-habitat relationships. Ecol. Model. 221: 503–511.

Knudby, A., Kenchington, E., Cogswell, A.T., Lirette, C.G., and Murillo, F.J. 2013a. Distribution modelling for sponges and sponge grounds in the northwest Atlantic Ocean. Can. Tech. Rep. Fish. Aquat. Sci. 3055: v + 73 p.

Knudby, A., Kenchington, E., and Murillo, F.J. 2013b. Modeling the distribution of *Geodia* sponges and sponge grounds in the northwest Atlantic Ocean. PLoS ONE 8, e82306. http://dx.doi.org/10.1371/journal.pone.0082306.

Knudby, A., Lirette, C., Kenchington, E., and Murillo, F.J. 2013c. Species Distribution Models of Black Corals, Large Gorgonian Corals and Sea Pens in the NAFO Regulatory Area. Ser No N6276. NAFO SCR Doc 13/78. 17p.

Kuhn, M. and Johnson, K. 2013. Applied Predictive Modeling. Springer. New York. 574 p.

Kulka, D., Swain. D., Simpson, M.R., Miri, C.M., Simon, J., Gauthier, J., McPhie, R., Sulikowski, J. and Hamilton, J. 2006. Distribution, abundance, and life history of *Malacoraja senta* (smooth skate) in Canadian Atlantic waters with reference to its global distribution. DFO Can. Sci. Advis. Sec. Res. Doc. 2006/093: 140 p.

Lawler, J.J., Shafer, S.L., White, D., Kareiva, P., Maurer, E.P., Blaustein, A.R., and Bartlein, P.J. 2009. Projected climate-induced faunal change in the Western Hemisphere. Ecology 90: 588–597.

Liaw, A., and Wiener, M. 2002. Classification and regression by randomForest. R News 2: 18–22.

Liu, C., Berry, P.M., Dawson, T.P., and Pearson, R.G. 2005. Selecting thresholds of occurrence in prediction of species distribution. Ecography 28: 385–393.

McPherson, J.M., Jetz, W., and Rogers, D.J. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artifact? J. Appl. Ecol. 41: 811–823.

Miller, D.L., Rexstad, E., Burt, L., Bravington, M.V., and Hedley, S. 2015. Package 'dsm'. 26 p.

Moritz, C., Lévesque, M., Gravel, D., Vaz, S., Archambault, D., and Archambault, P. 2013. Modelling spatial distribution of epibenthic communities in the Gulf of St. Lawrence (Canada). J. Sea Res. 78: 75–84.

Murillo, F.J., Kenchington, E., Sacau, M., Piper, D.J.W., Wareham V, and Muñoz, A. 2011. New VME indicator species (excluding corals and sponges) and some potential VME elements of the NAFO Regulatory Area. Ser No N6003. NAFO SCR Doc 11/73. 20 p.

NAFO. 2013. Conservation and Enforcement Measures. NAFO/FC, Doc. 13/1, Serial No. N6131. 103 p.

NAFO. 2015. Report of the 8th meeting of the NAFO Scientific Council Working Group on Ecosystem Science and Assessment (WGESA). NAFO SCS, in press.

Perley, M. H. 1849. Report on the fisheries of the Gulf of Saint Lawrence.

Peterson, A. 2003. Predicting the geography of species' invasions via ecological niche modeling. Q. Rev. Biol. 78: 419–433.

Peterson, A., and Robins, C. 2003. Using ecological-niche modeling to predict barred owl invasions with implications for spotted owl conservation. Conserv. Biol. 17: 1161–1165.

Porfirio, L.L., Harris, R.M.B., Lefroy, E.C., Hugh, S., Gould, S.F., Lee, G., Bindoff, N.L., and Mackey, B. 2014. Improving the Use of Species Distribution Models in Conservation Planning and Management under Climate Change. PLoS ONE 9(11): e113749. doi:10.1371/journal.pone.0113749.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rooper, C.N., Zimmermann, M., Prescott, M.M., and Hermann, A.J. 2014. Predictive models of coral and sponge distribution, abundance and diversity in bottom trawl surveys of the Aleutian Islands, Alaska. Mar. Ecol. Prog. Ser. 503: 157–176.

Shono, H. 2008. Application of the Tweedie distribution of zero-catch data in CPUE analysis. Fish. Res. 93: 154–162.

Tittensor, D.P., Baco, A.R., Brewin, P.E., Clark, M.R., Consalvey, M., Hall-Spencer, J., Rowden, A.A., Schlacher, T., Stocks, K.I., and Rogers, A.D. 2009. Predicting global habitat suitability for stony corals on seamounts. J. Biogeogr. 36: 1111–1128.

Williams, G. C. 1995. Living genera of sea pens (Coelenterata: Octocorallia: Pennatulacea): illustrated key and synopses. Zool. J. Linn. Soc.- Lond. 113: 93–140.

Wood, S.N. 2006. Generalized additive models: an introduction with R. Chapman & Hall/CRC Press, Boca Raton, FL.

Yesson, C., Taylor, M.L., Tittensor, D.P., Davies, A.J., Guinotte, J., Baco, A., Black, J., Hall-Spencer, J.M., and Rogers, A.D. 2012. Global habitat suitability of cold-water octocorals. J. Biogeogr. 39: 1278–1292.

APPENDIX 1

Prediction of Sponges, Stalked Tunicates and Soft Corals Biomass Using GAM

Sponges (Porifera)

a) Southern Gulf of St. Lawrence

The GAM results on mean sponge biomass per grid cell for the southern Gulf of St. Lawrence are presented in Table A1.

Table A1. GAM results. Significant variables ($p < 0.05$) are listed in order of their significance (highest to lowest) in the model, the deviance explained and the estimated degrees of freedom (edf) and the R^2 are also provided. Variables are Summer Primary Production Average Range (pp_sum_ran_avg), Spring Primary Production Average Range (pp_spr_ran_avg), Depth, Surface Temperature Average Minimum (s_tmp_min_avg), Summer Primary Production Average Minimum (pp_sum_min_avg), MERIS Case II Fall Chlorophyll *a* Minimum (M_chl_fall_min), Maximum Summer Mixed Layer Depth (mld_sum_max_avg), Summer Primary Production Average Maximum (pp_sum_max_avg), MODIS Case I Fall Chrolophyll *a* Range (chl_fall_ran), MODIS Case I Spring Chrolophyll *a* Range (chl_spr_ran), Surface Current Average Range (s_cur_ran_avg), Slope, location (lat, long), Surface Current Average Minimum (s_cur_min_avg).

Model	Significant variables	Deviance explained	edf	R^2
1	s(pp_sum_ran_avg) + s(pp_spr_ran_avg) + s(depth) + s(s_tmp_min_avg) + s(pp_sum_min_avg) + s(M_chl_fall_min) + s(mld_sum_max_avg) +s(pp_sum_max_avg) +s(chl_fall_ran) + s(chl_spr_ran) + s(s_cur_ran_avg) + s(slope)	27.1%	3.5; 4.3; 1; 4.7; 3.9; 1.5; 3.9; 1; 4.4; 3.6; 4.2; 1.9	0.05
2	te(lat, long) + s(depth) + s(pp_sum_ran_avg) + s(s_cur_min_avg) + s(M_chl_fall_min) + s(mld_sum_max_avg)	55%	194; 1; 1; 3.3; 1.5; 3.1	0.26

Figures A1 and A2 show the surface of sponge biomass (kg) predictions per grid cell generated from GAM Models 1 and 2 in the southern Gulf of St. Lawrence. The biomass predicted was higher around the maximum sponge catch, but in Model 2 (Figure A2) deviance explained and R^2 were higher, and very large predictions (> 4.000 kg) were obtained in two small areas at the northeast of Prince Edward Island in areas where the maximum biomass sponge recorded was less than 10 kg (Figure A3).

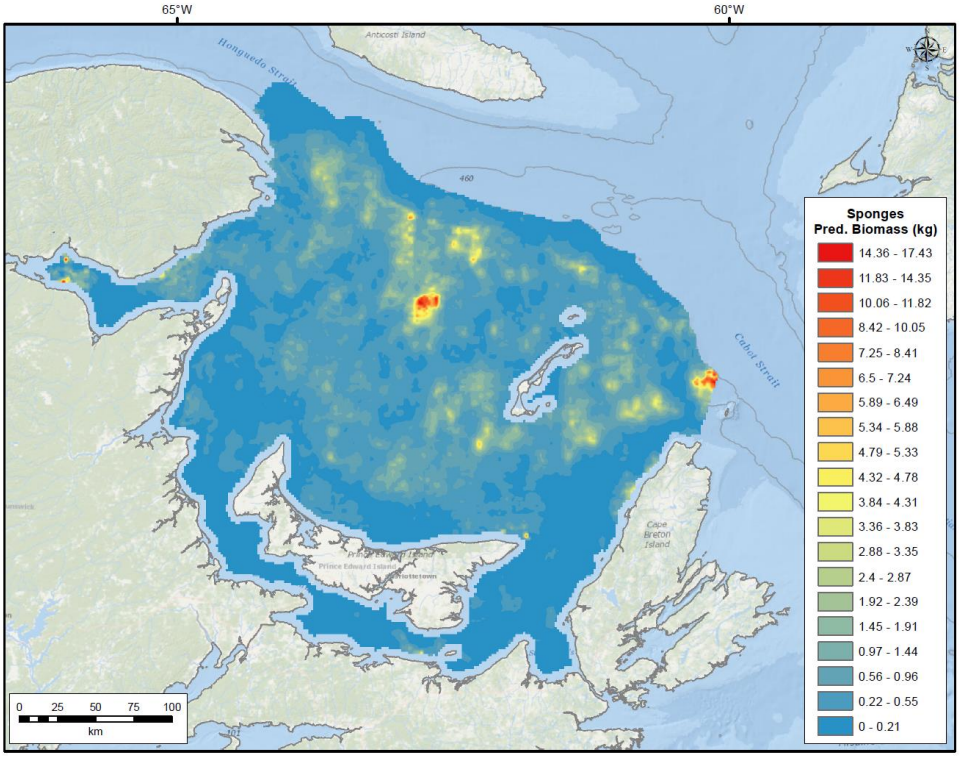

Figure A1. Prediction of sponge biomass (kg) from the GAM Model 1 (see Table A1) within the southern Gulf of St. Lawrence.

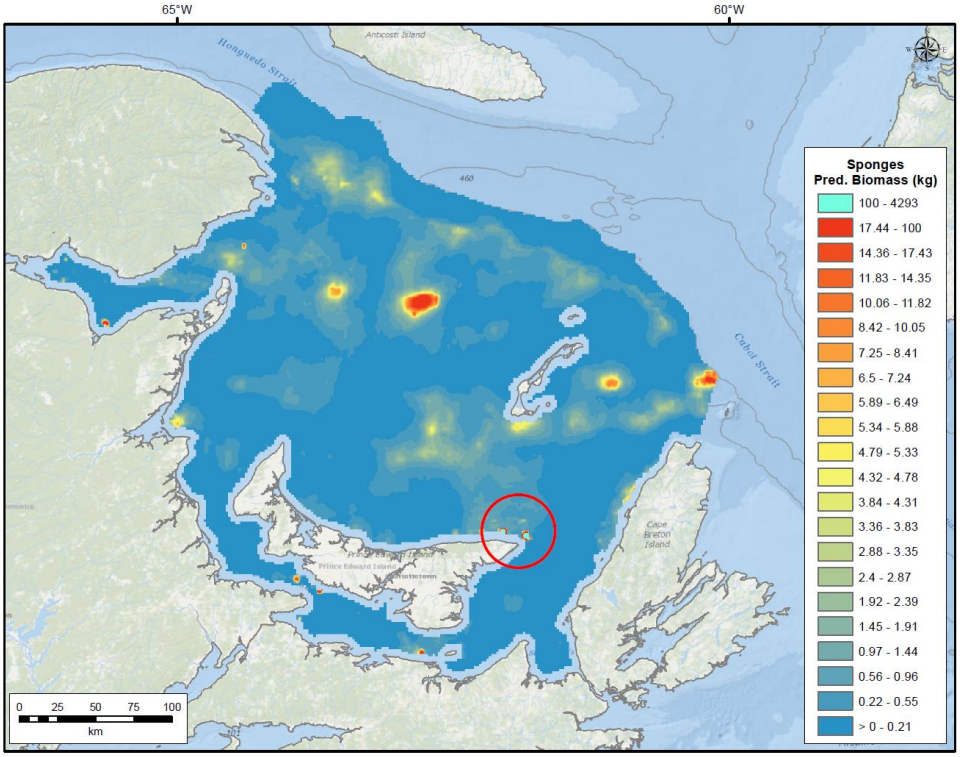

Figure A2. Prediction of sponge biomass (kg) from the GAM Model 2 (see Table A1) within the southern Gulf of St. Lawrence. Red circle highlights an area of high predicted biomass that was not as prominent in the Model 1 analysis (Figure A1).

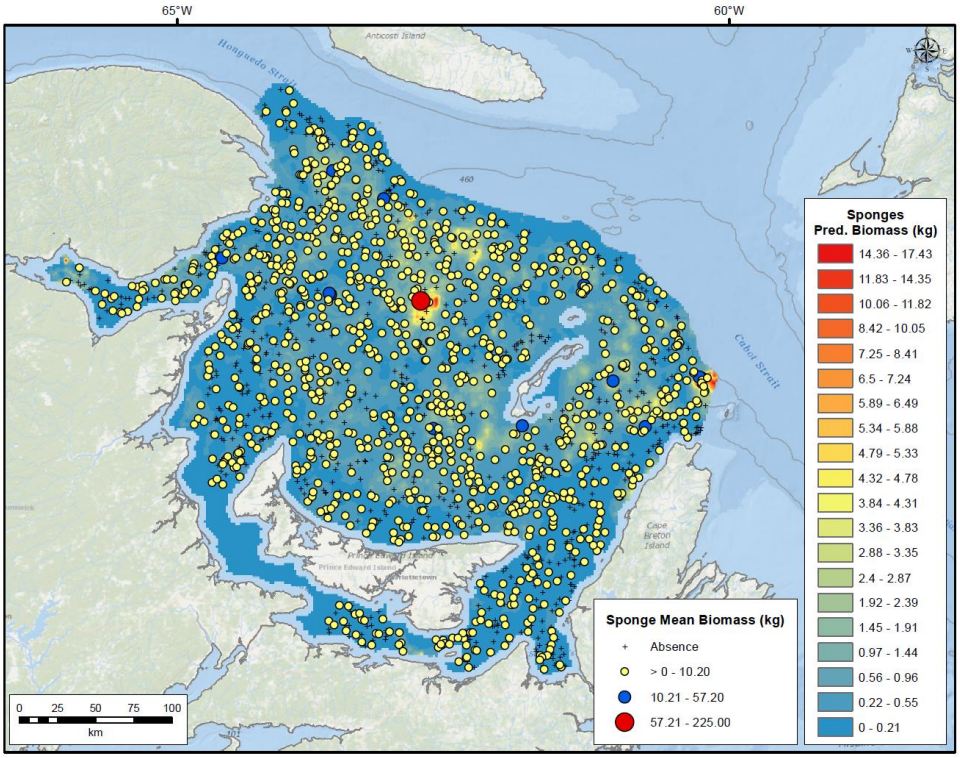

Figure A3. Sponge biomass catches overlaid on the sponge biomass (kg) predicted from the GAM Model 1 (see Table A1) within the southern Gulf of St. Lawrence.

b) Northern Gulf of St. Lawrence

The GAM results on mean sponge biomass per grid cell for the northern Gulf of St. Lawrence are presented in Table A2.

Table A2. GAM results. Significant variables ($p < 0.05$) are listed in order of their significance (highest to lowest) in the model, the deviance explained and the estimated degrees of freedom (edf) and the R^2 are also provided. Variables are Maximum Summer Mixed Layer Depth (mld_sum_max_avg), Surface Current Average Minimum (s_cur_min_avg), Depth, Surface Current Average Range (s_cur_ran_avg), MERIS Case II Fall Chlorophyll *a* Minimum (M_chl_fall_min), MODIS Case I Spring Chrolophyll *a* Range (chl_spr_ran), Bottom current Average Range (b_cur_ran_avg), Summer Primary Production Average Minimum (pp_sum_min_avg), Summer Primary Production Average Maximum (pp_sum_max_avg), Summer Primary Production Average Range (pp_sum_ran_avg), Bottom Temperature Average Range (b_tmp_ran_avg), location (lat, long), Slope.

Model	Significant variables	Deviance explained	edf	R^2
1	s(mld_sum_max_avg) + s(s_cur_min_avg) + s(depth) + s(s_cur_ran_avg) + s(M_chl_fall_min) + s(chl_spr_ran) + s(b_cur_ran_avg) + s(pp_spr_min_avg) + s(pp_spr_ran_avg) + s(pp_sum_min_avg) + s(pp_sum_max_avg) + s(pp_sum_ran_avg) + s(b_tmp_ran_avg)	22%	4.5; 4.8; 1.8; 4.8; 1; 1; 1.7; 3.5; 4.1; 3.6; 1.8; 1; 1	0.03
2	te(lat, long) + s(pp_sum_ran_avg) + s(slope) + s(M_chl_fall_min) + s(depth) + s(s_cur_ran_avg) + s(pp_spr_ran_avg) + s(pp_sum_max_avg) + s(mld_sum_max_avg) + s(b_cur_ran_avg)	45.2%	138; 1; 1.8; 1.5; 1; 1; 4.1; 1.5; 4.2; 1	0.136

Figures A4 and A5 show the surface of sponge biomass (kg) predictions per grid cell generated from GAM Models 1 and 2 in the northern Gulf of St. Lawrence. Both surfaces showed very similar predictions with the maximum predicted biomass west and north of Anticosti Island. The biomass records overlaid on the sponge biomass predicted from the GAM Model 2 can be observed in Figure A6.

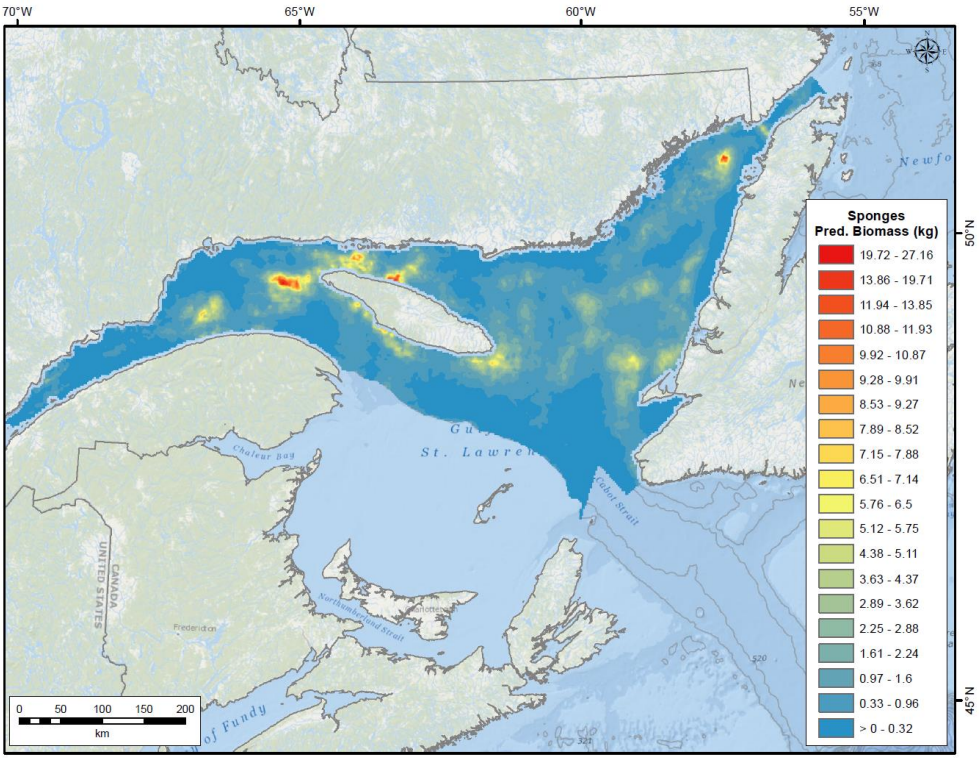

Figure A4. Prediction of sponge biomass (kg) from the GAM Model 1 (see Table A2) within the northern Gulf of St. Lawrence.

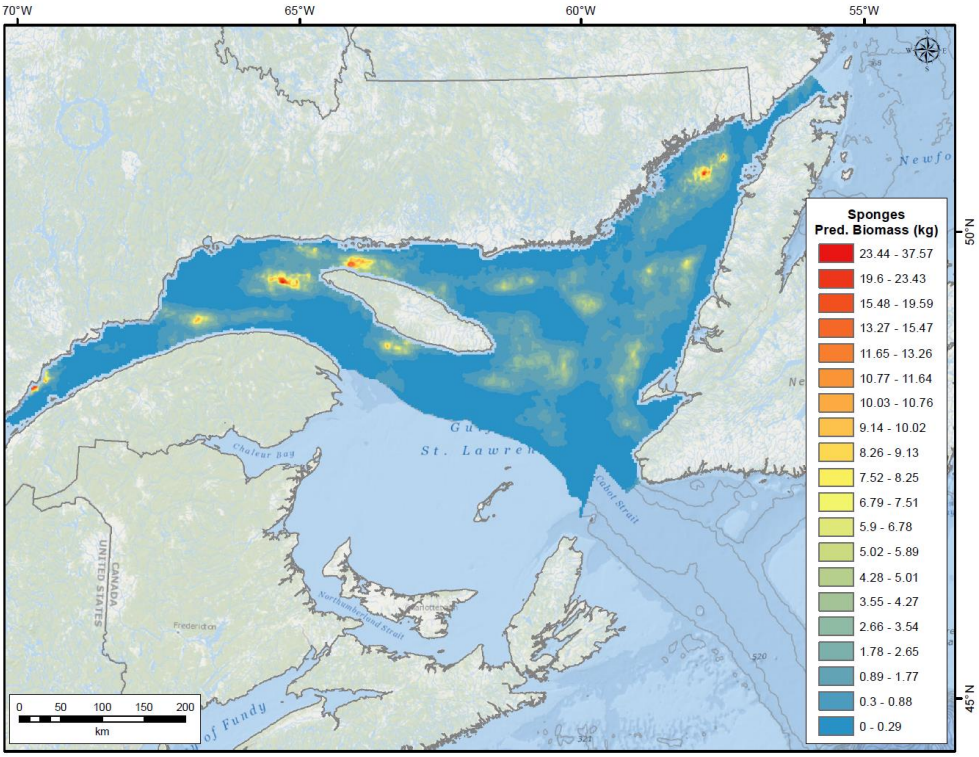

Figure A5. Prediction of sponge biomass (kg) from the GAM Model 2 (see Table A2) within the northern Gulf of St. Lawrence.

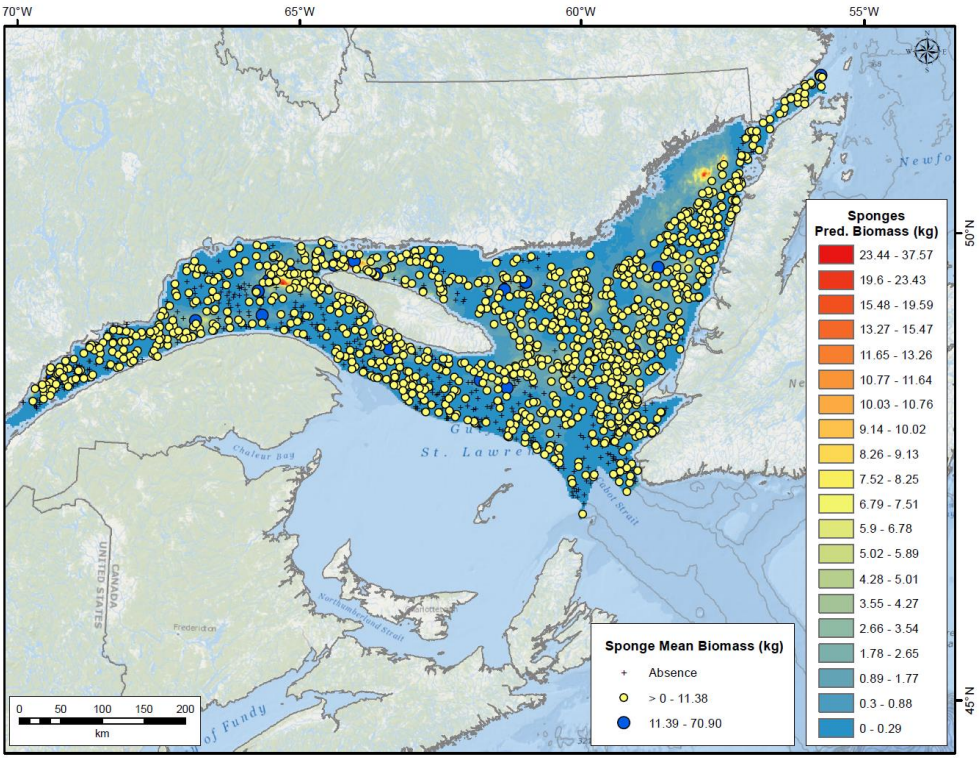

Figure A6. Sponge biomass catches overlaid on the sponge biomass (kg) predicted from the GAM Model 2 (see Table A2) within the northern Gulf of St. Lawrence.

Stalked tunicates (*Boltenia ovifera*)

a) Southern Gulf of St. Lawrence

The GAM results on mean stalked tunicates biomass per grid cell for the southern Gulf of St. Lawrence are presented in Table A3.

Table A3. GAM results. Significant variables ($p < 0.05$) are listed in order of their significance (highest to lowest) in the model, the deviance explained and the estimated degrees of freedom (edf) and the R^2 are also provided. Variables are Depth, MERIS Case II Fall Chlorophyll *a* Minimum (M_chl_fall_min), Slope, Maximum Summer Mixed Layer Depth (mld_sum_max_avg), Spring Primary Production Average Range (pp_spr_ran_avg), Bottom Temperature Average Maximum (b_tmp_max_avg), Surface Current Average Minimum (s_cur_min_avg), MERIS Case II Summer Chlorophyll *a* Range (M_chl_sum_ran), MODIS Case I Spring Chrolophyll *a* Range (chl_spr_ran), Spring Primary Production Average Minimum (pp_spr_min_avg), Annual Primary Production Average Minimum (pp_ann_min_avg), Bottom Temperature Average Range (b_tmp_ran_avg), Bottom Current Average Range (b_cur_ran_avg), Surface Current Average Range (s_cur_ran_avg), MODIS Case I Fall Chrolophyll *a* Range (chl_fall_ran), Surface Temperature Average Minimum (s_tmp_min_avg), Summer Primary Production Average Maximum (pp_sum_max_avg), Summer Primary Production Average Range (pp_sum_ran_avg), location (lat, long).

Model	Significant variables	Deviance explained	edf	R^2
1	s(depth) + s(M_chl_fall_min) + s(slope) + s(mld_sum_max_avg) + s(pp_spr_ran_avg) + s(b_tmp_max_avg) + s(s_cur_min_avg) + s(M_chl_sum_ran) + s(chl_spr_ran) + s(pp_spr_min_avg) + s(pp_ann_min_avg) + s(b_tmp_ran_avg) + s(b_cur_ran_avg) + s(s_cur_ran_avg) + s(chl_fall_ran) + s(s_tmp_min_avg) + s(pp_sum_max_avg) + s(pp_sum_ran_avg)	52.8%	1; 2; 1.9; 4.8; 4.5; 1; 2.8; 4.3; 4.8; 4.5; 1.8; 1; 1; 1; 2.7; 4.7; 1.8; 2.9	0.18
2	te(lat, long) + s(depth) + s(slope) + s(chl_spr_ran) + s(M_chl_fall_min)	71.3%	170.3; 1.4; 1; 4.4; 1	0.36

Figures A7 and A8 show the surface of stalked tunicate biomass (kg) predictions per grid cell generated from GAM Models 1 and 2 in the southern Gulf of St. Lawrence. GAM Model 2 had higher deviance explained and R^2 than Model 1 and was able to predict the large stalked tunicate biomass between Prince Edward Island and Magdalen Islands. However, both of them predicted high biomass at the northwest of the study region, where no large stalked tunicate biomass catches have been found to date (Figure A9).

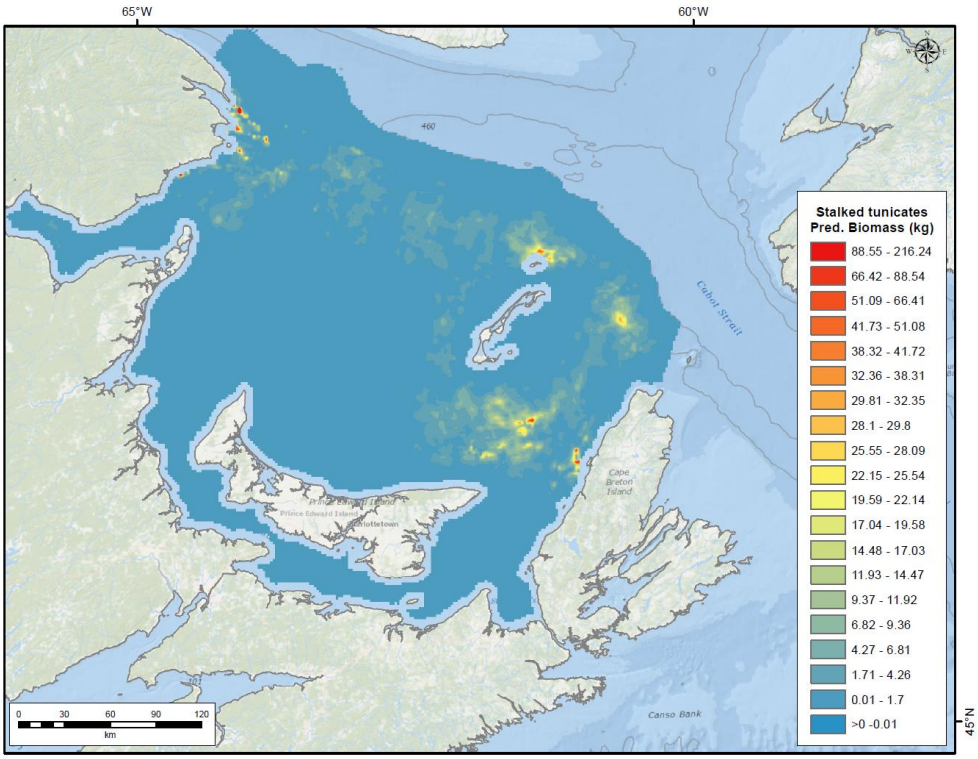

Figure A7. Prediction of stalked tunicate biomass (kg) from the GAM Model 1 (see Table A3) within the southern Gulf of St. Lawrence.

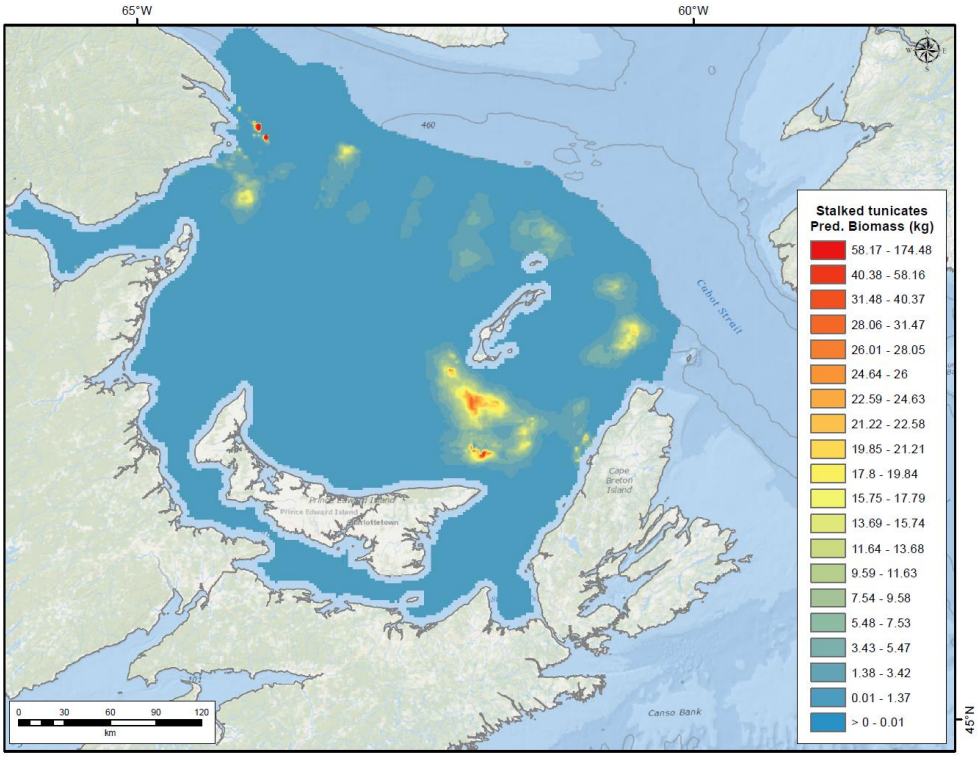

Figure A8. Prediction of stalked tunicate biomass (kg) from the GAM Model 2 (see Table A3) within the southern Gulf of St. Lawrence.

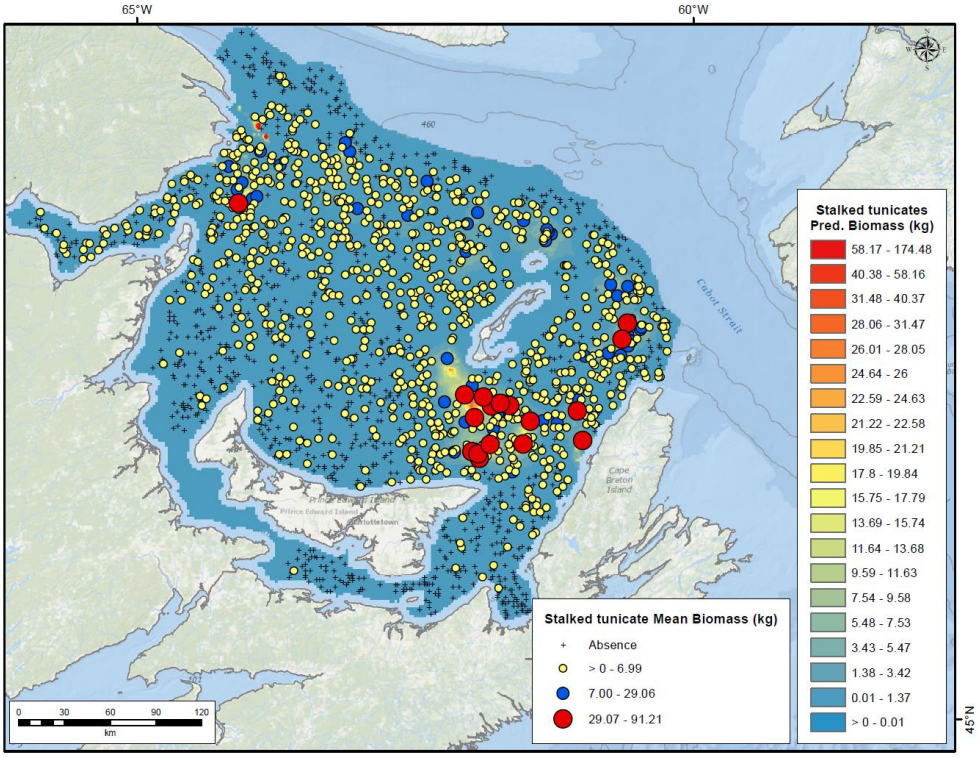

Figure A9. Stalked tunicate biomass catches overlaid to the stalked tunicate biomass (kg) predicted from the GAM Model 2 (see Table A3) within the southern Gulf of St. Lawrence.

b) Northern Gulf of St. Lawrence

The GAM results on mean stalked tunicates biomass per grid cell for the northern Gulf of St. Lawrence are presented in Table A4.

Table A4. GAM results. Significant variables ($p < 0.05$) are listed in order of their significance (highest to lowest) in the model, the deviance explained and the estimated degrees of freedom (edf) and the R^2 are also provided. Variables are Surface Current Average Minimum (s_cur_min_avg), Depth, Surface Current Average Range (s_cur_ran_avg), Bottom Current Average Range (b_cur_ran_avg), Summer Primary Production Average Maximum (pp_sum_max_avg), Surface Temperature Average Minimum (s_tmp_min_avg), Maximum Summer Mixed Layer Depth (mld_sum_max_avg), Spring Primary Production Average Minimum (pp_spr_min_avg), Summer Primary Production Average Range (pp_sum_ran_avg), location (lat, long), Spring Primary Production Average Range (pp_spr_ran_avg), MERIS Case II Fall Chlorophyll *a* Minimum (M_chl_fall_min), Slope, Bottom Temperature Average Range (b_tmp_ran_avg).

Model	Significant variables	Deviance explained	edf	R^2
1	s(s_cur_min_avg) + s(depth) + s(s_cur_ran_avg) + s(b_cur_ran_avg) + s(pp_sum_max_avg) + s(s_tmp_min_avg) + s(mld_sum_max_avg) + s(pp_spr_min_avg) + s(pp_sum_ran_avg)	82.7%	4.7; 1.9; 4.1; 1; 1; 3.3; 3.7; 1; 1	0.42
2	te(lat, long) + s(pp_spr_ran_avg) + s(s_tmp_min_avg) + s(M_chl_fall_min) + s(depth) + s(slope) + s(pp_sum_max_avg) + s(s_cur_ran_avg) + s(b_cur_ran_avg) + s(b_tmp_ran_avg)	85.1%	19.2; 3.2; 1; 1.8; 1; 1; 1; 1.9; 1; 1.8	0.44

Figures A10 and A11 show the surface of stalked tunicate biomass (kg) predictions per grid cell generated from GAM Models 1 and 2 in the northern Gulf of St. Lawrence. GAM Model 2 had slightly higher deviance explained and R^2 than Model 1 and both model predicted large stalked tunicate biomass north of Anticosti island, where the higher biomass of stalked tunicates was found (Figure A12). However, GAM Model 2 predicted very high biomass at the mouth of the St. Lawrence Estuary where low stalked tunicate biomass was recorded (Figure A11).

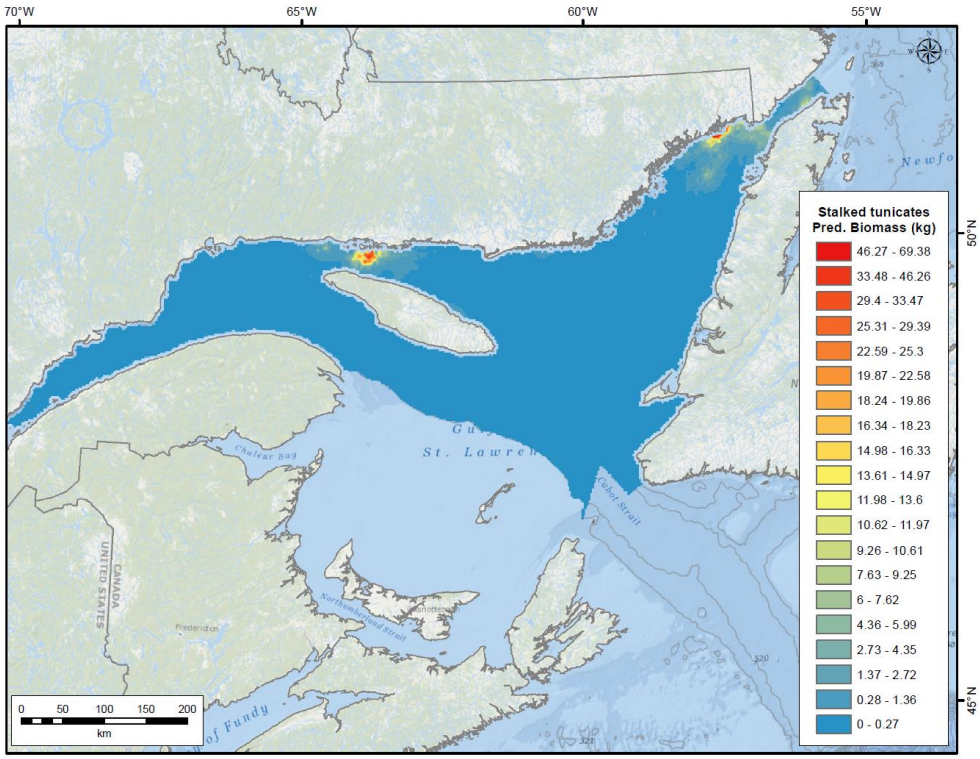

Figure A10. Prediction of stalked tunicate biomass (kg) from the GAM Model 1 (see Table A4) within the northern Gulf of St. Lawrence.

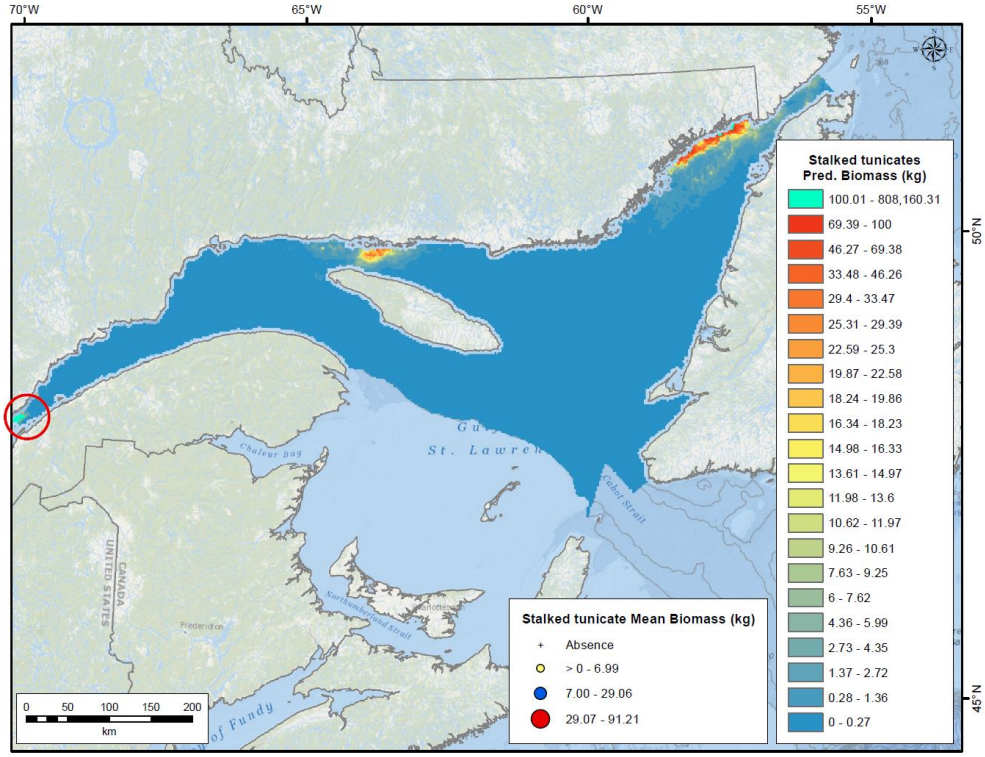

Figure A11. Prediction of stalked tunicate biomass (kg) from the GAM Model 2 (see Table A4) within the northern Gulf of St. Lawrence. Red circle highlights an area of high predicted biomass that was not as prominent in the Model 1 analysis (Figure A10).

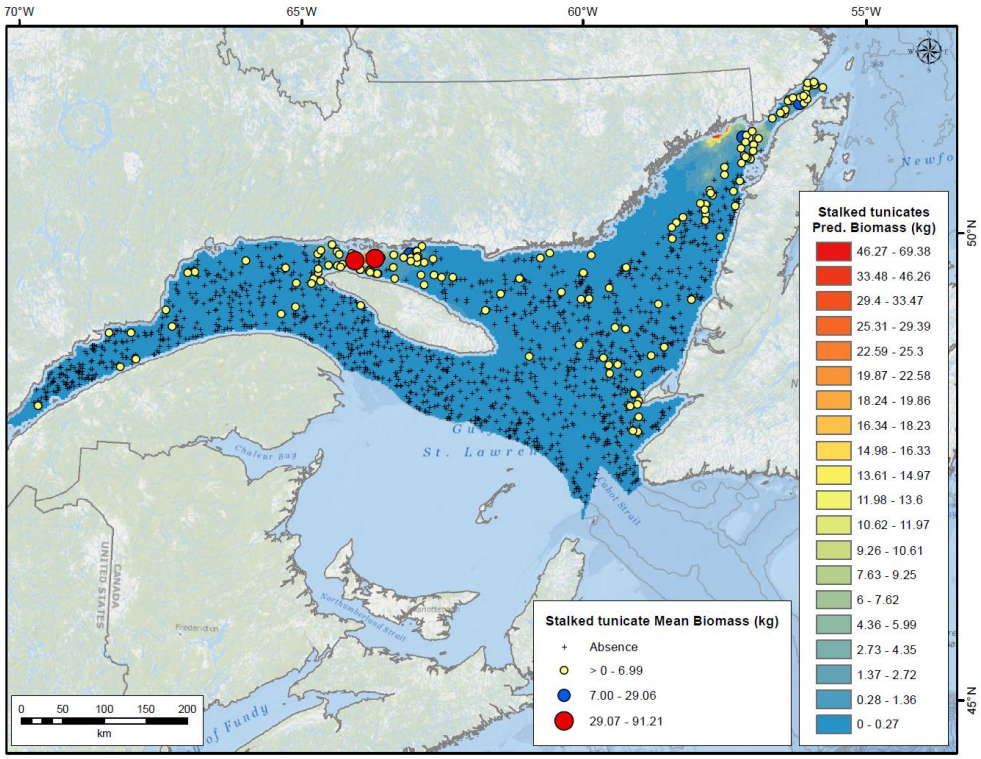

Figure A12. Stalked tunicate biomass catches overlaid to the stalked tunicate biomass (kg) predicted from the GAM Model 2 (see Table A4) within the northern Gulf of St. Lawrence.

Soft corals (*Gersemia rubiformis*)

a) Southern Gulf of St. Lawrence

The GAM results on mean stalked tunicates biomass per grid cell for the southern Gulf of St. Lawrence are presented in Table A5.

Table A5. GAM results. Significant variables ($p < 0.05$) are listed in order of their significance (highest to lowest) in the model, the deviance explained and the estimated degrees of freedom (edf) and the R^2 are also provided. Variables are Spring Primary Production Average Range (pp_spr_ran_avg), Surface Temperature Average Minimum (s_tmp_min_avg), Summer Primary Production Average Minimum (pp_sum_min_avg), Bottom Current Average Range (b_cur_ran_avg), MERIS Case II Fall Chlorophyll *a* Minimum (M_chl_fall_min), Spring Primary Production Average Minimum (pp_spr_min_avg), MERIS Case II Summer Chlorophyll *a* Range (M_chl_sum_ran), Surface Current Average Range (s_cur_ran_avg), Bottom Temperature Average Range (b_tmp_ran_avg), location (lat, long).

Model	Significant variables	Deviance explained	edf	R^2
1	s(pp_spr_ran_avg) + s(s_tmp_min_avg) + s(pp_sum_min_avg) + s(b_cur_ran_avg) + s(M_chl_fall_min) + s(pp_spr_min_avg) + s(M_chl_sum_ran) + s(s_cur_ran_avg) + s(b_tmp_ran_avg)	41.5%	4.4; 4.6; 4.5; 1.8; 1.2; 3.8; 1; 2.6; 1	0.34
2	te(lat, long) + s(M_chl_fall_min) + s(pp_sum_min_avg)	48%	54; 1; 3.7	0.36

Figures A13 and A14 show the surface of *G. rubiformis* biomass (kg) predictions per grid cell generated from GAM Models 1 and 2 in the southern Gulf of St. Lawrence. GAM Model 2 had slightly higher deviance explained and R^2 than Model 1 although both surfaces were very similar showing a good fit to the *G. rubiformis* biomass catches (Figure A15). Biomass is predicted to be highest in the Baie des Chaleurs under both models.

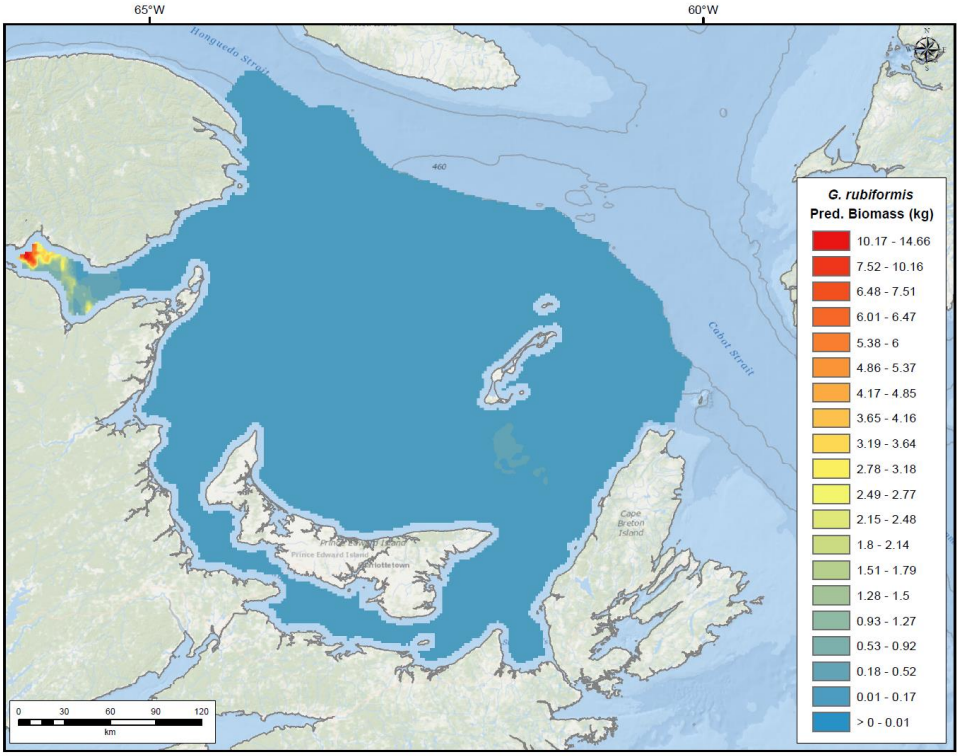

Figure A13. Prediction of *G. rubiformis* biomass (kg) from the GAM Model 1 (see Table A5) within the southern Gulf of St. Lawrence.

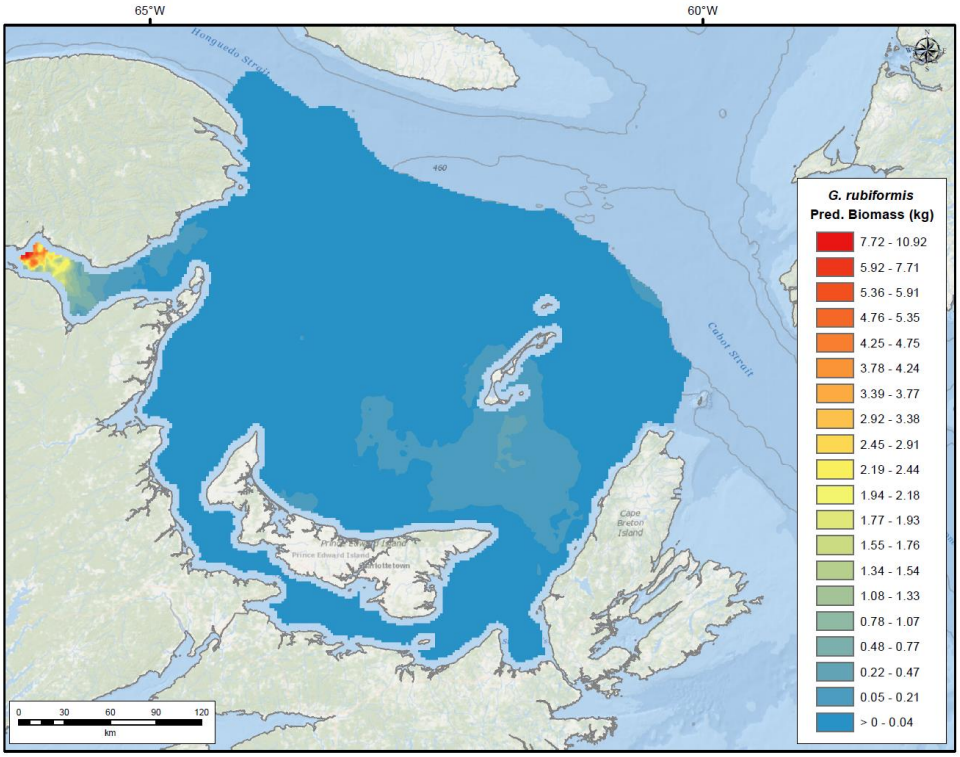

Figure A14. Prediction of *G. rubiformis* biomass (kg) from the GAM Model 2 (see Table A5) within the southern Gulf of St. Lawrence.

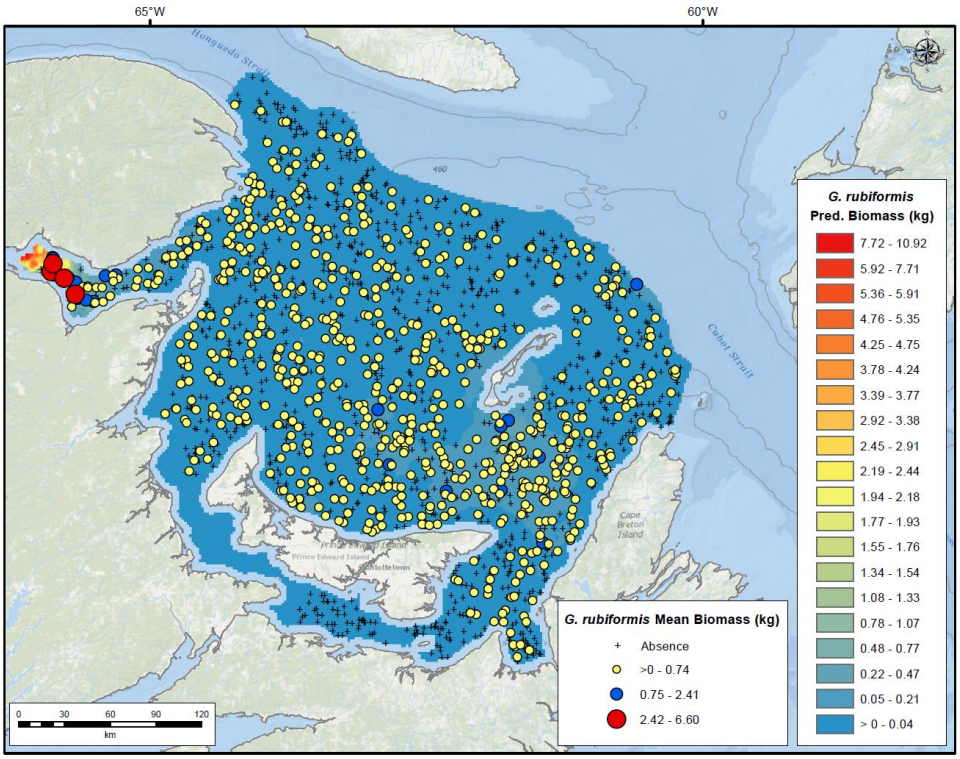

Figure A15. *G. rubiformis* biomass catches overlaid to the *G. rubiformis* biomass (kg) predicted from the GAM Model 2 (see Table A5) within the southern Gulf of St. Lawrence.

b) Northern Gulf of St. Lawrence

The GAM results on mean *G. rubiformis* biomass per grid cell for the northern Gulf of St. Lawrence are presented in Table A6.

Table A6. GAM results. Significant variables ($p < 0.05$) are listed in order of their significance (highest to lowest) in the model, the deviance explained and the estimated degrees of freedom (edf) and the R^2 are also provided. Variables are Depth, Surface Current Average Range (s_cur_ran_avg), Bottom Temperature Average Maximum (b_tmp_max_avg), Spring Primary Production Average Range (pp_spr_ran_avg), Maximum Summer Mixed Layer Depth (mld_sum_max_avg), Annual Primary Production Average Minimum (pp_ann_min_avg), Summer Primary Production Average Maximum (pp_sum_max_avg), Summer Primary Production Average Minimum (pp_sum_min_avg), MODIS Case I Fall Chrolophyll *a* Range (chl_fall_ran), Spring Primary Production Average Minimum (pp_spr_min_avg), Bottom Temperature Average Range (b_tmp_ran_avg), Surface Temperature Average Minimum (s_tmp_min_avg), MODIS Case I Spring Chrolophyll *a* Range (chl_spr_ran), Surface Current Average Minimum (s_cur_min_avg), Summer Primary Production Average Range (pp_sum_ran_avg), Bottom Current Average Range (b_cur_ran_avg), Slope, location (lat, long).

Model	Significant variables	Deviance explained	edf	R^2
1	s(depth) + s(s_cur_ran_avg) + s(b_tmp_max_avg) + s(pp_spr_ran_avg) + s(mld_sum_max_avg) + s(pp_ann_min_avg) + s(pp_sum_max_avg) + s(pp_sum_min_avg) + s(chl_fall_ran) + s(pp_spr_min_avg) + s(b_tmp_ran_avg) + s(s_tmp_min_avg) + s(chl_spr_ran) + s(s_cur_min_avg) + s(pp_sum_ran_avg) + s(b_cur_ran_avg) + s(slope)	38.1%	1.8; 1; 1.7; 1; 1.4; 2; 1.9; 2.5; 1; 3.5; 1; 4.3; 1.6; 1.2; 1; 1.4; 1	0.34
2	te(lat, long) + s(depth) + s(mld_sum_max_avg) + s(b_tmp_max_avg) + s(s_cur_ran_avg) + s(pp_spr_ran_avg) + s(pp_sum_max_avg) + s(chl_spr_ran) + s(b_cur_ran_avg)	46.6%	34.4; 1.9; 1; 1.8; 1; 1; 1.9; 1; 1	0.47

Figures A16 and A17 show the surface of *G. rubiformis* biomass (kg) predictions per grid cell generated from GAM Models 1 and 2 in the northern Gulf of St. Lawrence. GAM Model 2 had higher deviance explained and R^2 than Model 1 and both models predicted similar biomass surfaces, however biomass predicted from GAM Model 2 was closer to the *G. rubiformis* biomass catches (Figure A18).

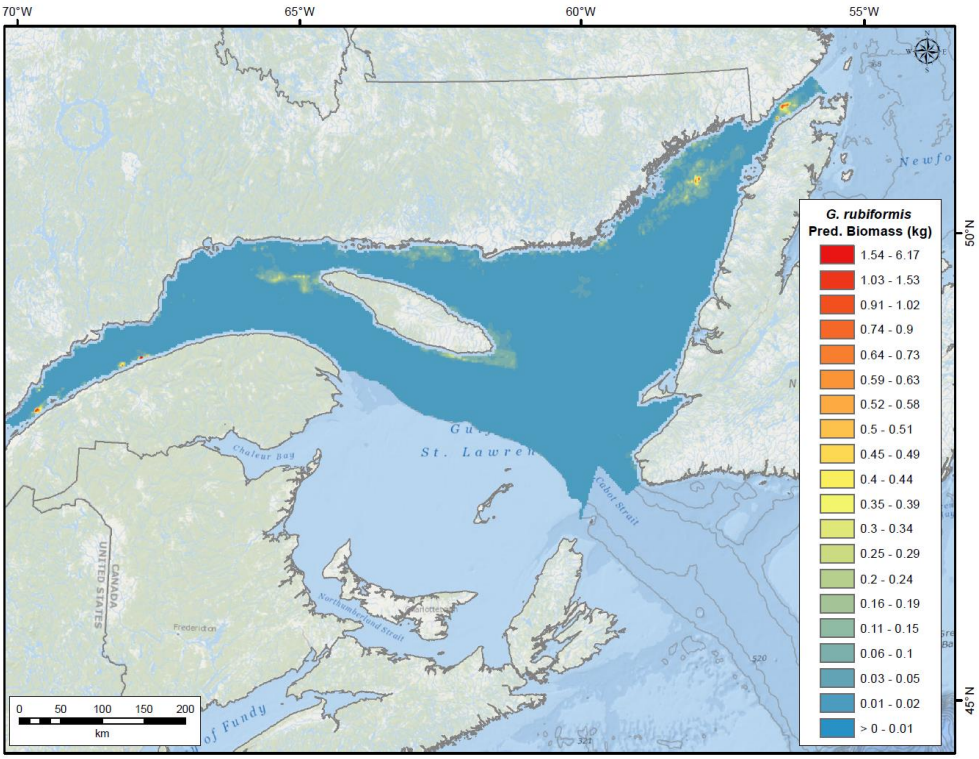

Figure A16. Prediction of *G. rubiformis* biomass (kg) from the GAM Model 1 (see Table A6) within the northern Gulf of St. Lawrence.

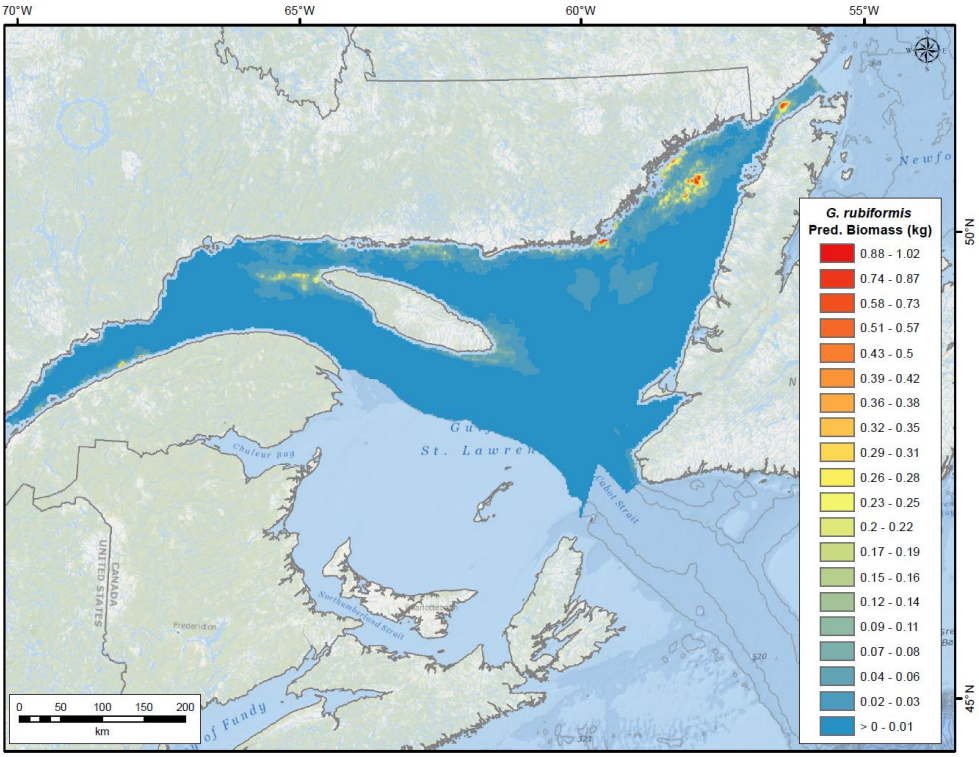

Figure A17. Prediction of *G. rubiformis* biomass (kg) from the GAM Model 2 (see Table A6) within the northern Gulf of St. Lawrence.

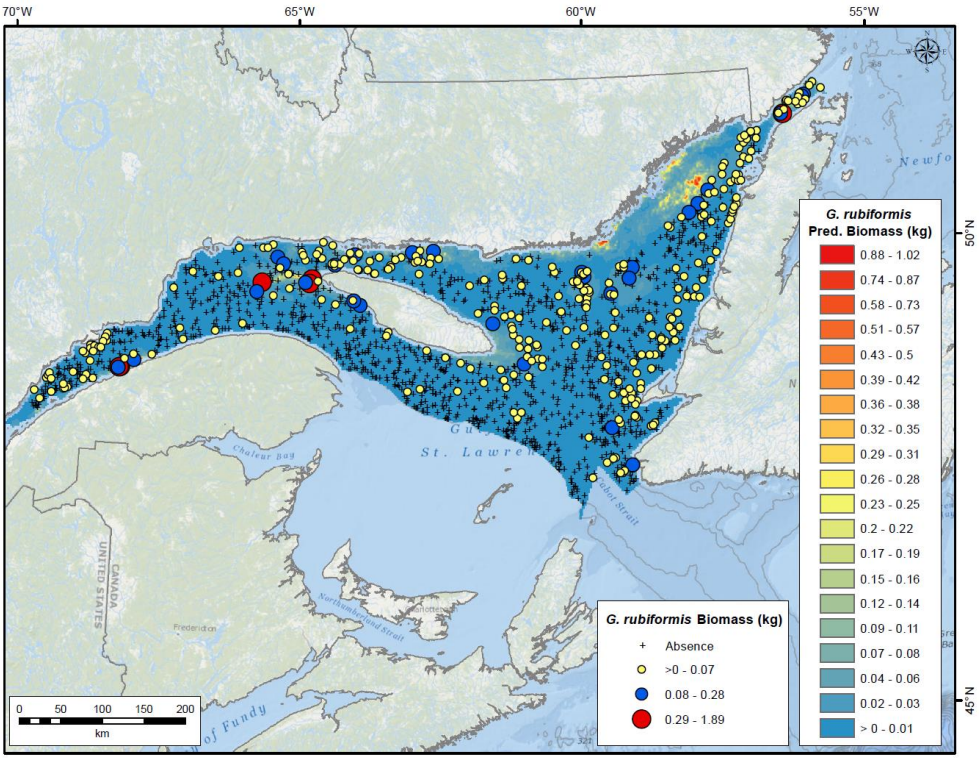

Figure A18. *G. rubiformis* biomass catches overlaid to the *G. rubiformis* biomass (kg) predicted from the GAM Model 2 (see Table A6) within the northern Gulf of St. Lawrence.

APPENDIX 2

Distribution of Solitary Cup Corals (*Flabellum alabastrum*)

Solitary cup corals are not considered to be VME indicator taxa by NAFO (NAFO, 2013) however they are the only representatives of the stony corals in the Gulf of St. Lawrence. Their known distribution from DFO research vessel survey data has been plotted (Figure A19) in order to complement the work presented herein. To date only *Flabellum alabastrum* has been reported and only from surveys in the northern Gulf of St. Lawrence.

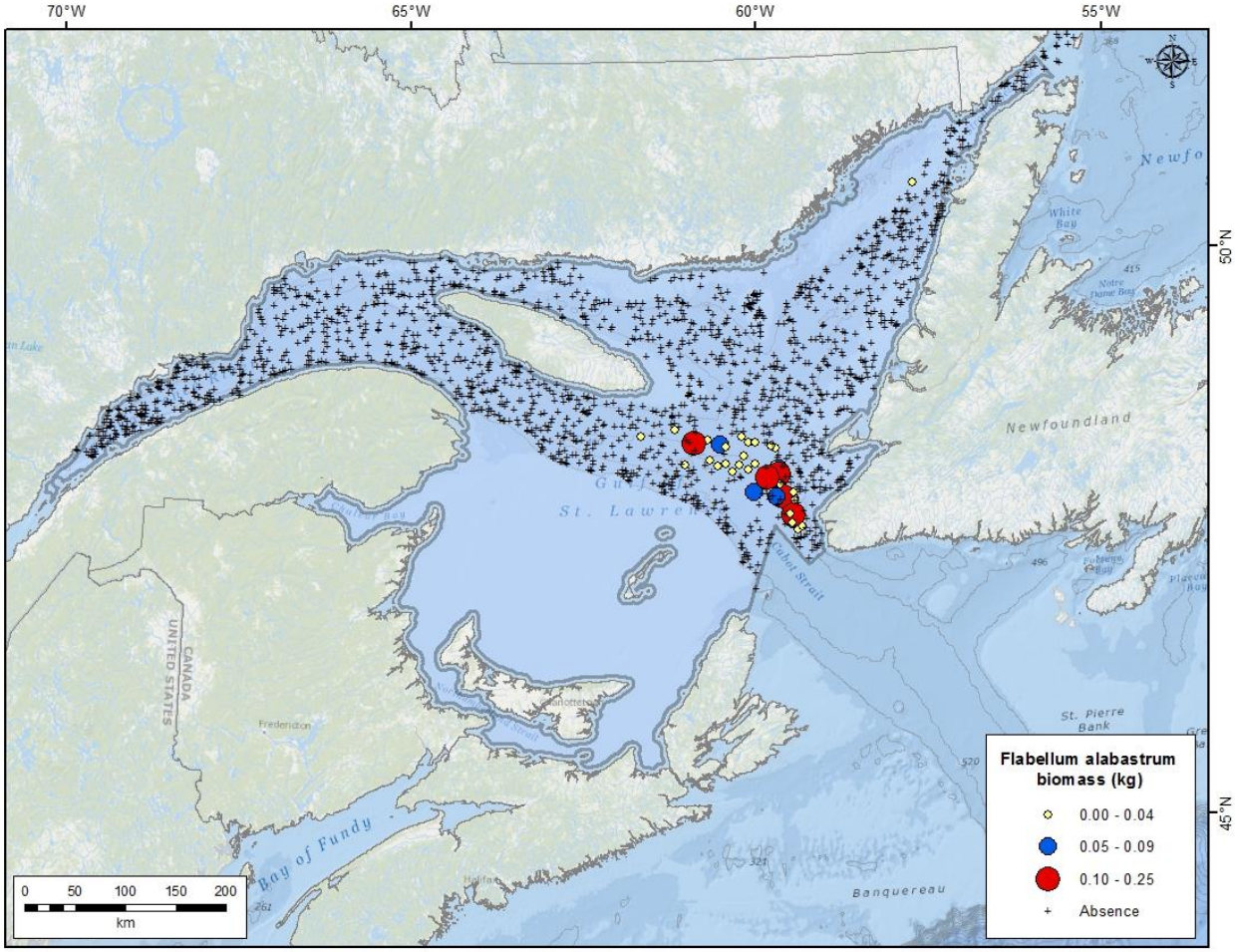

Figure A19. *Flabellum alabastrum* (stony cup coral) biomass catches within the northern Gulf of St. Lawrence. No records were reported from the southern Gulf region.