Species Distribution Modelling of Crinoids, Bryozoans and Ascidians in the Newfoundland and Labrador Region

J. Guijarro, E. Kenchington, F.J. Murillo, L. Beazley, C. Lirette, V. Wareham,

M. Koen-Alonso

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 3181

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 Fig. 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 Fig. 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 3181

2016

Species Distribution Modelling of Crinoids, Bryozoans and Ascidians in the Newfoundland and Labrador Region

by

J. Guijarro[1], E. Kenchington[1], F.J. Murillo[1], L. Beazley[1], C. Lirette[1], V. Wareham[2],

M. Koen-Alonso[2]

Fisheries and Oceans Canada

[1]Ocean and Ecosystem Sciences Division

Maritimes Region

Bedford Institute of Oceanography

P.O. Box 1006, Dartmouth, N.S.

B2Y 4A2

[2]Newfoundland and Labrador Region

Northwest Atlantic Fisheries Centre

P.O. Box 5667, St. John's, Newfoundland

A1C 5X1

Cat. No. Fs97-6/3181E-PDF ISBN 978-0-660-06544-1 ISSN 1488-5379

Correct citation for this publication:

Guijarro, J., Kenchington, E., Murillo, F.J., Beazley, L., Lirette, C., Wareham, V., Koen-Alonso, M. 2016. Species Distribution Modelling of Crinoids, Bryozoans and Ascidians in the Newfoundland and Labrador Region. Can. Tech. Rep. Fish. Aquat. Sci. 3181: v + 60p.

TABLE OF CONTENTS

ABSTRACT

Guijarro, J., Kenchington, E., Murillo, F.J., Beazley, L., Lirette, C., Wareham, V., and Koen-Alonso, M. 2016. Species Distribution Modelling of Crinoids, Bryozoans and Ascidians in the Newfoundland and Labrador Region. Can. Tech. Rep. Fish. Aquat. Sci. 3181: v + 60p.

Species distribution models were performed using a random forest model for Crinoids (Phylum Echinodermata: Class Crinoidea), Ascidians (Phylum Chordata: Class Ascidiacea) and Bryozoans (Phylum Bryozoa) for the spatial extent of DFO's Placentia Bay-Grand Bank and Newfoundland and Labrador Shelves Large Ocean Management Areas (LOMA). Sixty-six environmental variables derived from various sources and native spatial resolutions were used as predictor variables in the random forest models. Species occurrence was predicted using all presence and absence data (unbalanced model), and equal numbers of presence and absence records (balanced model). The models produced from the unbalanced data were chosen as the better prediction surfaces for all three groups. The unbalanced model appeared to more accurately predict in areas of extrapolation, although none of the models were independently validated. Regression random forest models were employed to predict biomass distribution of the three taxonomic groups, but at best only explained a very small portion of the variation for the Crinoids, and were unreliable for the other groups.

RÉSUMÉ

Guijarro, J., Kenchington, E., Murillo, F.J., Beazley, L., Lirette, C., Wareham, V., and Koen-Alonso, M. 2016. Modélisation de la répartition des espèces de crinoïdes, de bryozoaires et d'ascidies dans la région de Terre-Neuve-et-Labrador. Can. Tech. Rep. Fish. Aquat. Sci. 3181: v+60p.

Les modèles de distribution des espèces ont été réalisés à l'aide d'un modèle de forêts d'arbres décisionnels pour les crinoïdes (phylum *Echinodermata*, classe *Crinoidea*), les ascidies (phylum *Chordata*, classe *Crinoidea*) et les bryozoaires (phylum *Bryozoa*) pour deux zones étendues de gestion des océans de Pêches et Océans Canada (MPO), soit celle de la baie Placentia et des Grands Bancs et celle des plateaux de Terre-Neuve-et-Labrador. Soixante-six variables environnementales provenant de diverses sources et résolutions spatiales natives ont servi de variables prédictives dans les modèles de forêts d'arbres décisionnels. L'occurrence des espèces a été prédite à l'aide de toutes les données de présence et d'absence (modèle déséquilibré) ainsi que d'un nombre égal de rapports de présence et d'absence (modèle équilibré). Les modèles produits à partir des données déséquilibrées ont été retenus, car ils ont permis d'obtenir la meilleure prédiction pour les trois groupes. Le modèle déséquilibré semble donner des prédictions plus précises dans les secteurs d'extrapolation, bien qu'aucun des modèles n'ait été validé de manière indépendante. Des modèles de forêts d'arbres décisionnels de régression ont servi à prédire la répartition de la biomasse des trois groupes taxonomiques. Toutefois, dans le meilleur des cas, ils n'expliquent que dans une très petite mesure la variation de crinoïdes et n'ont pas été fiables pour les autres groupes.

INTRODUCTION

The Canadian Policy for Managing the Impact of Fishing on Sensitive Benthic Areas was developed by Fisheries and Oceans, Canada (DFO) in 2009. Under that policy Significant Benthic Areas are defined in DFO's Ecological Risk Assessment Framework (ERAF) as "significant areas of cold-water corals and sponge dominated communities", where significance is determined "through guidance provided by DFO-lead processes based on current knowledge of such species, communities and ecosystems". This policy has parallels with international policy for the protection of vulnerable marine ecosystems (VMEs) established through the United Nations General Assembly (UNGA) Resolutions pertaining to sustainable fisheries. VMEs can be characterized by other species beyond corals and sponges, and can even be applied to fish species (FAO, 2009).

Responding to the UNGA Resolutions for the high seas, Murillo et al. (2011) reviewed over 500 benthic taxa known to occur in the Northwest Atlantic Fisheries Organization (NAFO) Regulatory Area (NRA) of Flemish Cap and the Nose and Tail of the Grand Bank, against traits for identifying VME indicators (FAO, 2009). Briefly, those traits related to functional significance, fragility, and life-history characteristics that produce a slow recovery to disturbance. The data were derived from research vessel surveys of the area and revealed three faunal groups in addition to the corals (including sea pens) and sponges. Those were Crinoids, Erect Bryozoans, and Large Sea Squirts (Ascidians). Murillo et al. (2011) summarized the literature supporting their VME status and later (Murillo et al., 2016a) found *Boltenia ovifera* (Large Sea Squirt) and *Eucratea loricata* (Erect Bryozoan) to be characteristic of epibenthic assemblages on the edge, and on the medium to fine sands of the inner continental shelf of the Tail of the Grand Bank, respectively. For each group, only dense aggregations (beds/fields), which establish functional significance, are considered to be VMEs. Crinoids, Erect Bryozoans and Large Sea Squirts can also constitute Ecologically or Biologically Significant Areas (EBSAs) under the Convention on Biological Diversity (CBD). Kenchington (2014b) described Crinoid Beds, Erect Bryozoan Turf and Stalked Tunicate Fields as structure-forming biogenic habitats on the Scotian Shelf.

NAFO then conducted kernel density analyses (KDE) on the data for the new VME taxa (NAFO, 2013) to identify dense aggregations. For corals (including sea pens) and sponges, which had been heavily ground-truthed (Kenchington et al., 2014a), NAFO considers the KDE polygons to be VMEs (NAFO, 2014). The KDE for the new VME taxa produced good models for both Erect Bryozoans and Large Sea Squirts. However, there were insufficient data to perform the analyses on the crinoids which are very fragile and are not well sampled by the research vessel surveys. The areas which were identified by the KDE analyses for the other taxa were on the Tail of the Grand Bank. For Erect Bryozoans the area encompassed was quite extensive and NAFO called for further validation of the models, as unlike the coral and sponge KDE polygons the model outputs had not been ground-truthed at the time of modelling. Consequently the KDE polygons for Erect Bryozoans and Large Sea Squirts were only considered to be significant concentrations of VME indicators until further information could be obtained. In 2015, DFO conducted *in situ* photographic surveys of the NAFO KDE polygons for Erect Bryozoans and Large Sea Squirts, and concluded that for most of the area sampled the bottom consisted of soft sediments, and therefore was not conducive to forming large habitat areas consistent with VMEs. The bryozoans

were thought to occupy small patches of hard substrate in the KDE area. Species distribution models (SDM) have not yet been performed to try to improve the knowledge base for the NRA. However Murillo et al. (2016b) performed SDM on stalked tunicates (*Boltenia ovifera*) in the Gulf of St. Lawrence, with good success both for species occurrence and biomass predictions.

Here we present the first species distribution models of crinoids, ascidians (including Large Sea Squirts) and bryozoans (including Erect Bryozoans) for the Newfoundland and Labrador Region generated using random forest, a machine-learning technique. Using the same methodology, random forest SDMs were recently generated on coral and sponge catch data from the Newfoundland and Labrador Region (see Guijarro et al., 2016) and were used to delineate coral and sponge significant benthic areas as part of a national DFO Ecosystems and Oceans Science Sector advisory process, held March 8-10, 2016 in Halifax, N.S.

MATERIAL AND METHODS

Study Area

The combined spatial extent of DFO's Placentia Bay-Grand Bank and Newfoundland and Labrador Shelves Large Ocean Management Areas (LOMA) (termed the 'Newfoundland and Labrador Region' herein) was used as the boundary for species distribution modelling in this report (Figure 1). This extent is delimited by the 200 nautical mile exclusive Economic Zone (EEZ) in the east, and DFO's Maritimes Region and Central and Arctic administrative boundaries in the southwest and north, respectively. A 20-km buffer was placed around all land to avoid its inclusion in the models. The total area covered in the study extent is approximately 1,012,900 km^2 based on a NAD 1983 UTM Zone 21N projection.

Environmental Data Layers

Sixty-six environmental variables derived from various sources and native spatial resolutions were used as predictor variables in the random forest models (Table 1). Variables were chosen based on their availability and assumed relevance to the distribution of benthic fauna. Bathymetry was derived from the Canadian Hydrographic Service (CHS) Atlantic Bathymetry Compilation (ABC). This data was the highest resolution bathymetry available for the entire study area. In the Newfoundland and Labrador Region the data were resolved to 15 arc-seconds which is equivalent to approximately 500 m. Slope in degrees was derived from the depth raster using the 'Slope' tool in ArcMap's Spatial Analyst toolbox, ArcMap version 10.2.2 (ESRI, 2011). All other environmental variables were derived from long-term modelled oceanographic or remote-sensing data and were spatially interpolated across the study area using ordinary kriging in ArcMap. Specific details on the methods used for the spatial interpolation of these variables are documented in a separate technical report (Guijarro et al. in prep., although see Beazley et al., 2016 for information on the same environmental data sources and variables for the Gulf of St. Lawrence). Only variables that were spatially interpolated with reasonable confidence were used in this report, and a number of variables (e.g., dissolved oxygen, silicate) were not considered due to their poor coverage and/or data properties. All predictor layers were displayed in raster format with geographic coordinates using the WGS 1984 datum and a ~0.015º cell size (approximately equal to 1 km horizontal resolution in the Newfoundland and Labrador Region).

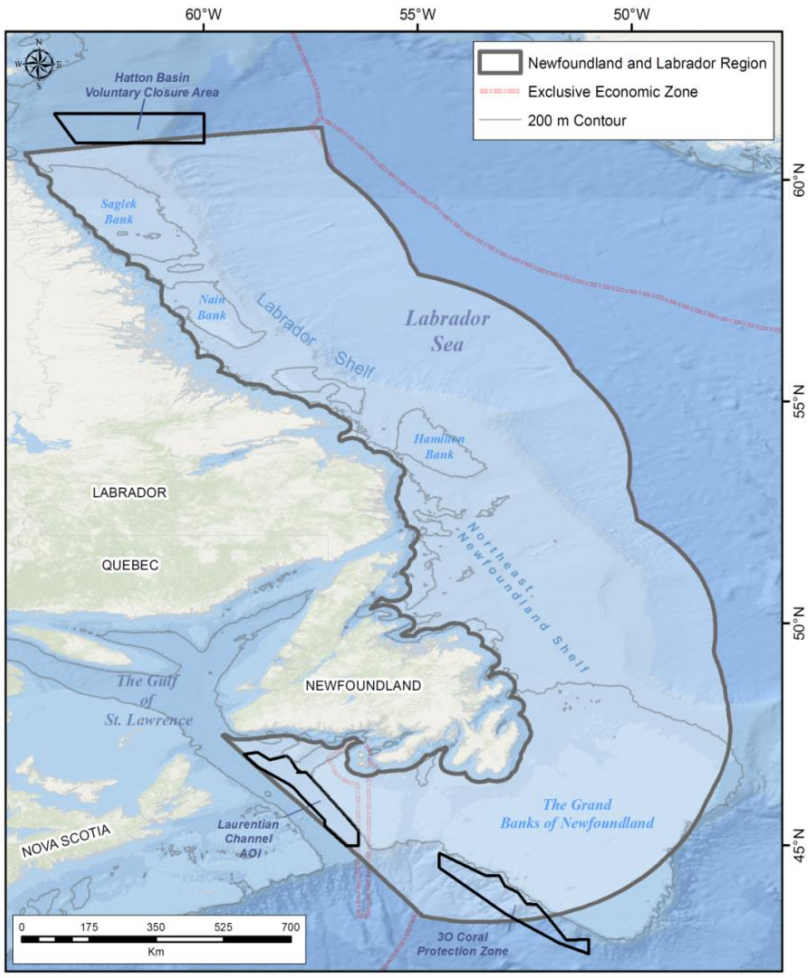

Figure 1. Extent of the boundary used for species distribution modelling (grey polygon) in the Newfoundland and Labrador Region. Place names are indicated on the map.

Response Data

Species composition as determined at sea of the three taxonomic groups modelled in this report is presented in Table 2. These are presented for purposes of re-extracting the data and should not be considered as taxonomically certain. For each group, presence-absence records were derived from catch data from DFO research vessel multispecies trawl surveys conducted on the CCGS *Needler, Teleost, or Templeman.* All tows were conducted following a depth stratified random design using Campelen trawl gear. DFO invertebrate catch data were provided by DFO's Newfoundland and Labrador Region where they are archived and managed. Data were available from 2006 to 2015 for ascidians, from 2010 to 2015 for bryozoans and from 2009 to 2015 for crinoids. Absence records were created from null (zero) catches that occurred in the same surveys.

The presence-absence records used in each random forest model were filtered so that only one presence or absence occurred within a single environmental data raster cell (~1 km). Presence records took precedence over an absence record when both occurred within the same raster cell. Biomass (kg) data associated with the DFO multispecies trawl survey records were averaged across multiple tows occurring within the same environmental raster cell.

Table 1. Summary of the 66 environmental variables used as predictor variables in random forest 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	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Fall Chlorophyll *a* Minimum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Fall Chlorophyll *a* Maximum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Fall Chlorophyll *a* Range	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Spring Chlorophyll *a* Mean	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Spring Chlorophyll *a* Minimum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Spring Chlorophyll *a* Maximum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Spring Chlorophyll *a* Range	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Summer Chlorophyll *a* Mean	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Summer Chlorophyll *a* Minimum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Summer Chlorophyll *a* Maximum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Summer Chlorophyll *a* Range	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Annual Chlorophyll *a* Mean	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Annual Chlorophyll *a* Minimum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Annual Chlorophyll *a* Maximum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Annual Chlorophyll *a* Range	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	mg m^{-3}	9 km
Fall Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Fall Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Fall Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Fall Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Mean	SeaWiFS Level-3 with other input	2006 - 2010	mg C m^{-2} day^{-1}	9 km

	parameters			
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 other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
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

Table 2. Species composition in each of the three taxonomic groups modelled using random forest. Also shown are the species/taxon codes associated with data entry of the DFO multispecies surveys over the time frame analyzed.

Taxon	Species/Taxon	Taxon Code
Crinoidea (Crinoids)	Crinoidea	8261
Ascidiacea (Ascidians)	Ascidiacea	8680
	Ascidia sp.	8742
	Boltenia ovifera	8792
	Dendrodoa sp.	8758
	Enterogona	8681
	Polyclinidae	8690
	Pyuridae	8790
Bryozoa (Bryozoans)	Ectoprocta	2670
	Bryozoa (Ectoprocta or Entoprocta)	9992

Random Forest Modelling

Random forest (Breiman, 2001), a non-parametric, machine learning technique, was used to generate probability of occurrence and biomass models for the three taxonomic groups in this report. Details of this modelling approach and assessment are reported in Guijarro et al. (2016).

RESULTS

Crinoids

Data Sources and Distribution

Crinoid (Phylum Echinodermata: Class Crinoidea) catch data was collected over a span of 7 years from 2009 to 2015 (Table 3) and consisted of 105 presence and 1325 absence records. Both presence and absence records were absent from Saglek Bank, the majority of Grand Bank, and in the deep waters off Newfoundland and Labrador (Figure 3). Presence records were concentrated on the banks off central Labrador. The highest mean biomass records (up to 5.90 kg) occurred on the South of Nain Bank.

Table 3. Number of presence and absence records of crinoid catch recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2009 and 2015.

Year	Total of number of presences	Total of number of absences
2009	2	72
2010	15	249
2011	14	160
2012	26	226
2013	24	190
2014	18	219
2015	6	209

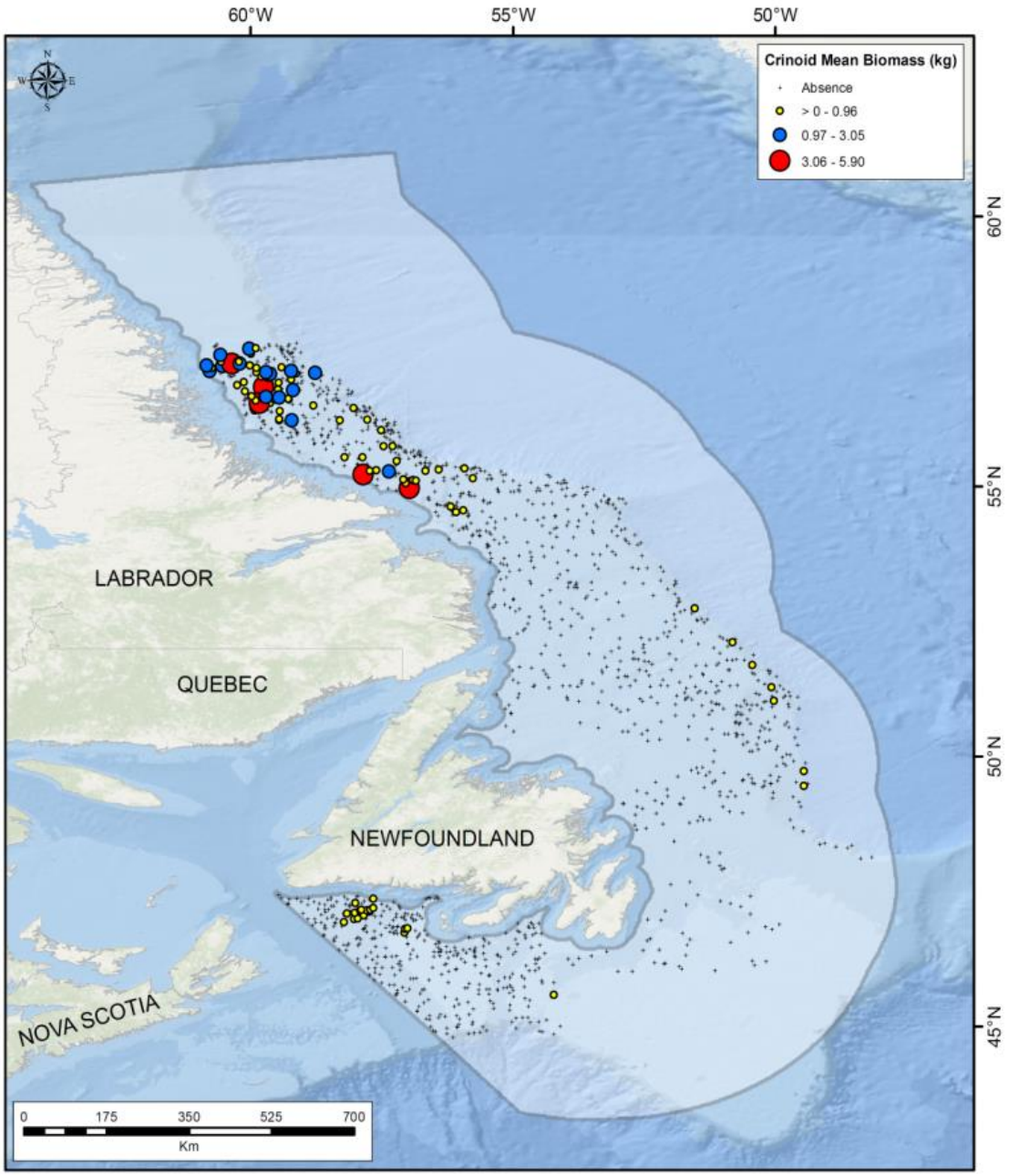

Figure 3. Mean biomass (kg) per grid cell of crinoids recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2009 and 2015. Also shown are absence records from the same surveys.

Model 1 – Balanced species prevalence

Accuracy measures (mean AUC, sensitivity and specificity) for the random forest model on balanced species prevalence (105 presences and 105 absences; Model 1) are presented in Table 4. The average AUC was 0.892, indicating very good model performance. The highest AUC of 0.930 was associated with Model Run 8. The sensitivity and specificity measures of this model were 0.867 and 0.857, respectively. The confusion matrix of this model showed that class errors for both the presence and absence classes were relatively moderate (0.133 and 0.143, respectively; Table 4).

Table 4. Accuracy measures for all 10 model repetitions of 10-fold cross validation from the random forest model of crinoid presence-absence data collected within the Newfoundland and Labrador Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 8) which is considered the optimal model for predicting the presence probability of crinoids.

Model Run	AUC	Sensitivity	Specificity
1	0.859	0.800	0.762
2	0.877	0.819	0.781
3	0.901	0.800	0.829
4	0.915	0.829	0.838
5	0.850	0.771	0.771
6	0.895	0.800	0.829
7	0.906	0.800	0.886
8	**0.930**	**0.867**	**0.857**
9	0.907	0.838	0.829
10	0.882	0.819	0.829
Mean	**0.892**	**0.814**	**0.821**
SD	**0.025**	**0.026**	**0.039**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	90	15	105	0.143
Presence	14	91	105	0.133

The presence probability prediction surface of the crinoids is presented in Figure 4. The highest predictions of presence probability occurred on and south of Nain Bank and a small area south of Newfoundland. These areas of high presence probability corresponded well with the spatial distribution of presence records (see Figure 5). Saglek Bank was also predicted to have a moderate to high presence probability of crinoids, where there are no presence or absence records to support it. The model appeared to greatly extrapolate areas of presence probability beyond the location of presence observations, particularly in deeper waters off the northeast

Newfoundland Shelf. Figure 6 shows the actual presence and absence data observations (105 presences and 105 absences) used in the optimal Model 1. Areas of extrapolation are also shown in Figure 6. The area of high predicted presence probability of crinoids off the Labrador Slope was considered extrapolated area.

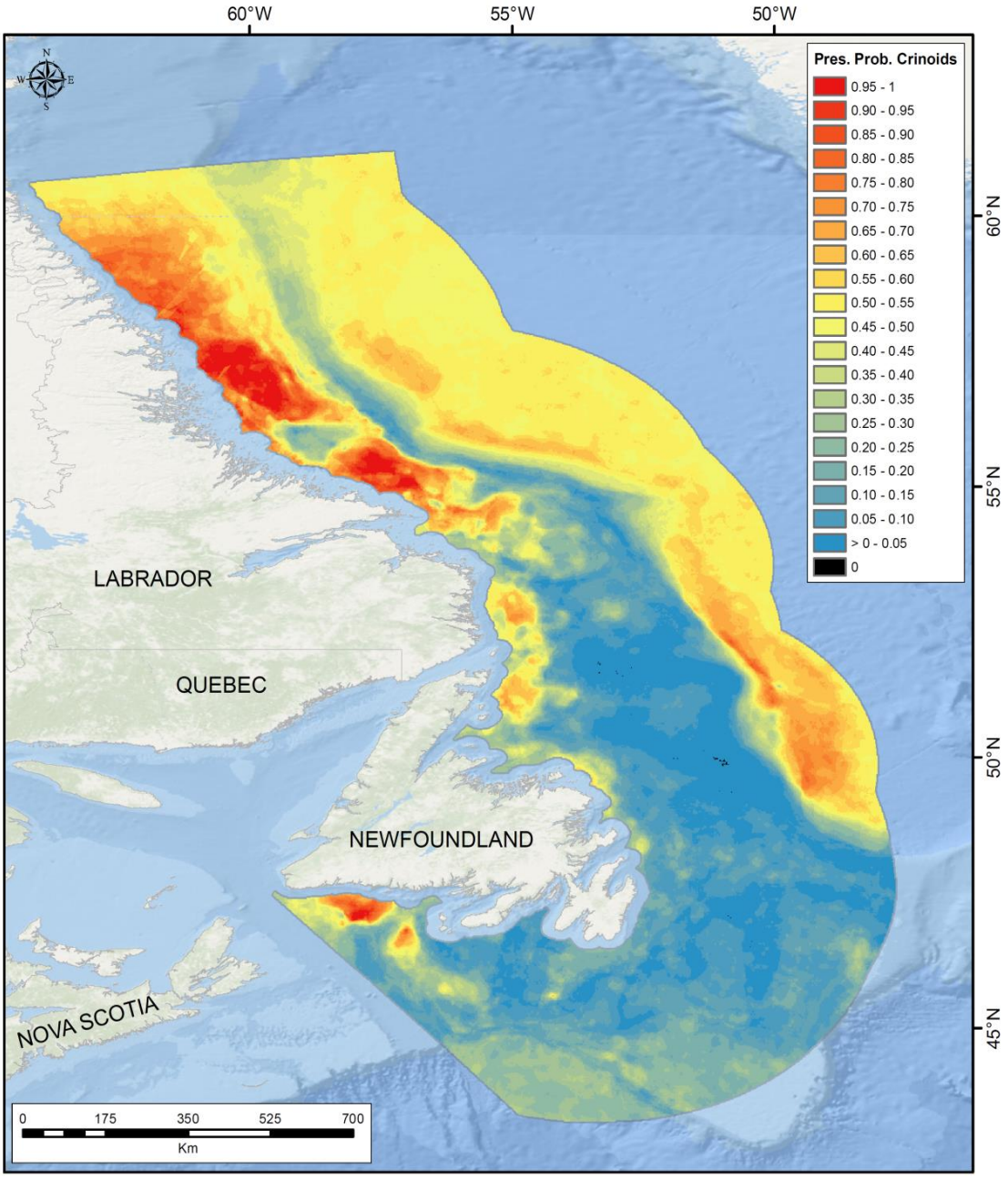

Figure 4. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of crinoid presence and absence collected from DFO multispecies surveys conducted within the Newfoundland and Labrador Region between 2009 and 2015.

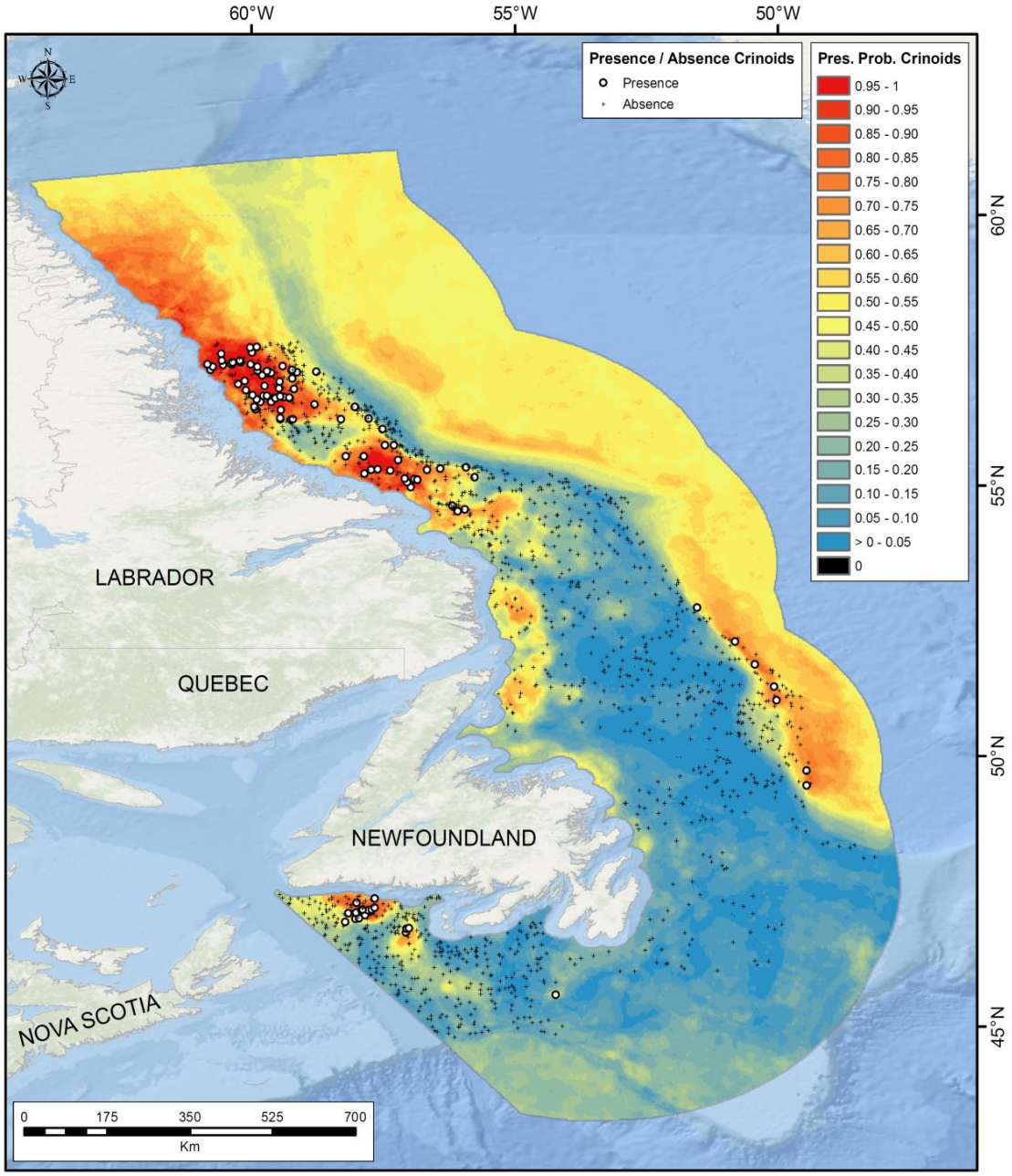

Figure 5. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of crinoid presence and absence data collected from DFO multispecies surveys conducted within the Newfoundland and Labrador Region between 2009 and 2015.

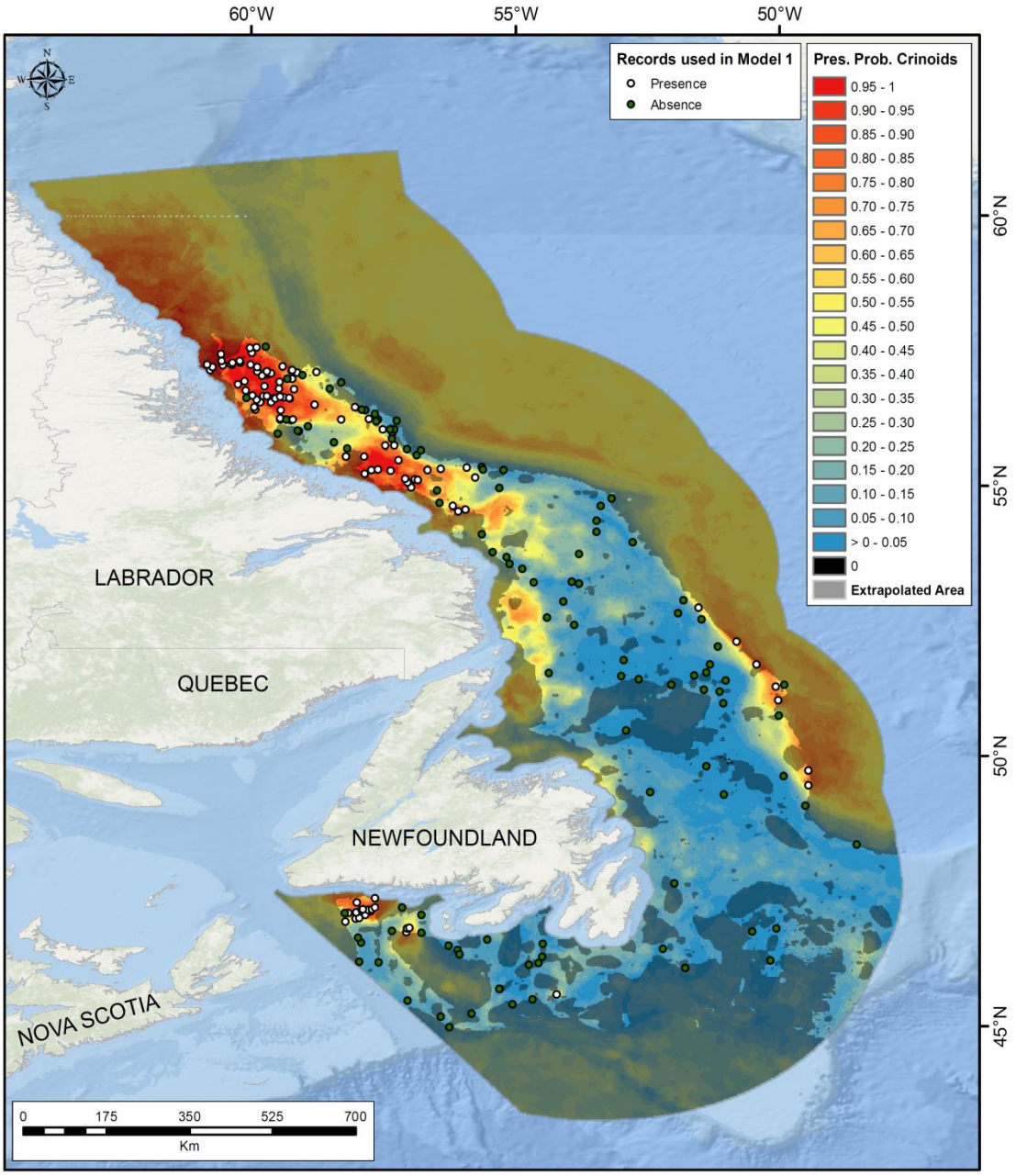

Figure 6. Map of the 210 data observations (105 presences and 105 absences) of crinoids used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of crinoids generated from Model 1 and areas of model extrapolation.

Of all 66 environmental predictor variables used in the model, Bottom Salinity Average Range was the most important for the classification of the crinoid presence-absence data (Figure 7). This variable displayed a right-skewed distribution prior to spatial interpolation (Guijarro et al., in prep). Examination of the Q-Q plot revealed a strong spatial pattern to those data points over- and under-predicted by a normal distribution, with over-predicted points located mainly in the deep waters beyond the Labrador Shelf, and under-predicted points located along the Newfoundland and Labrador slopes. Bottom Salinity Average Range was followed in importance by Spring Primary Production Average Range and Surface Temperature Mean. Partial dependence plots for the top 6 predictor variables are shown in Figure 8. The highest presence probability of crinoids along the gradient in Bottom Salinity Average Range occurred around 0.6. The fit between predicted and observed values for Bottom Salinity Average Range values up to ~0.7 was good, with deviation from the 1:1 reference line occurring for values greater than 0.7.

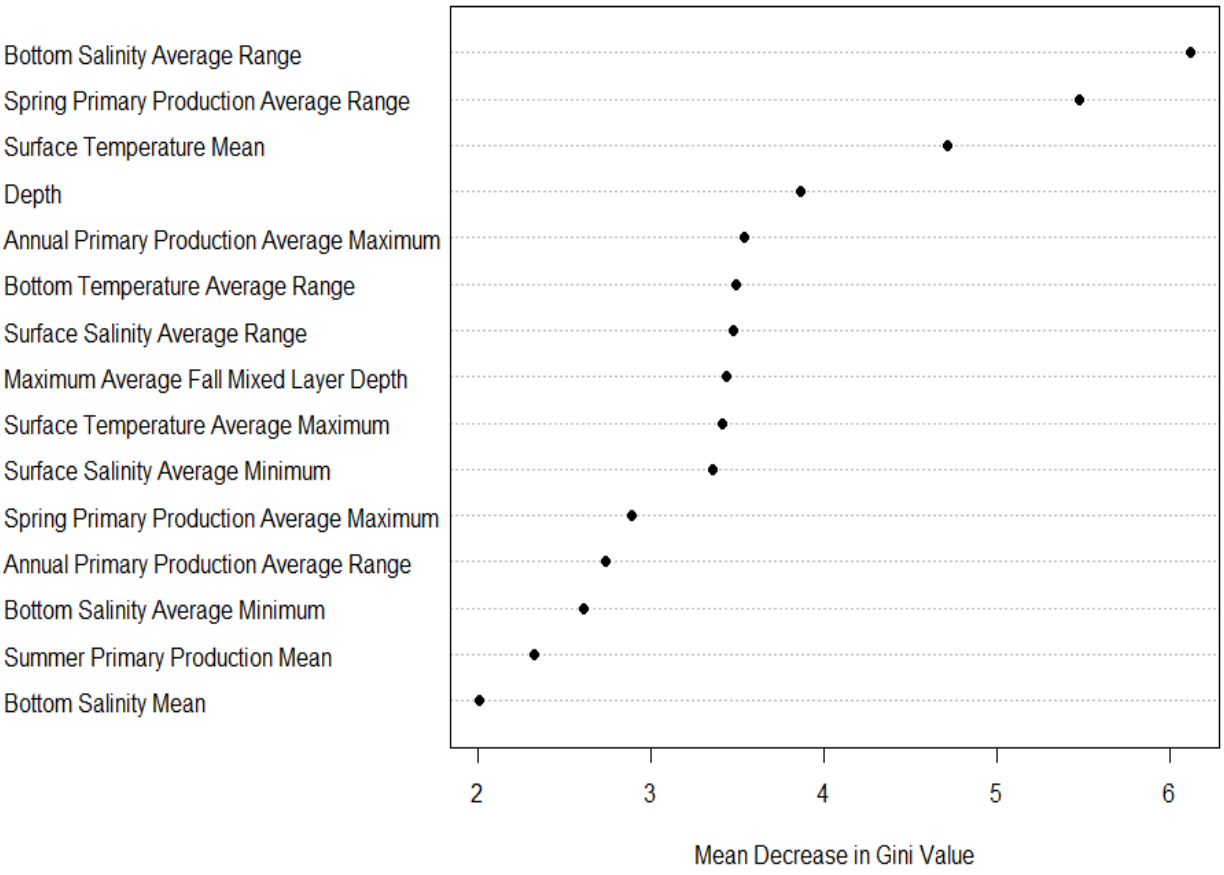

Figure 7. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini Value of the optimal random forest model predicting crinoid presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini value the more important the variable is for predicting the response data.

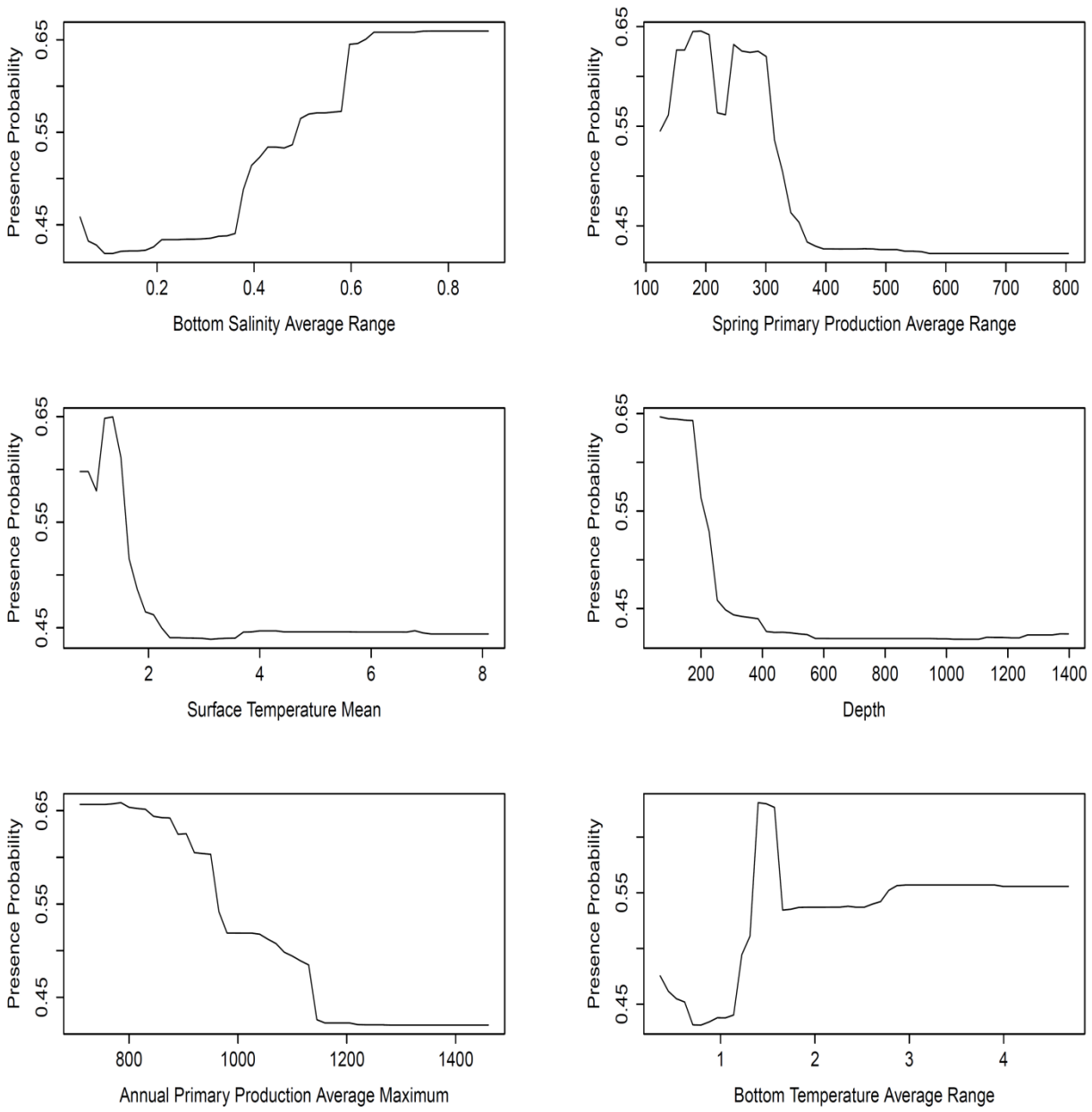

Figure 8. Partial dependence plots of the top 6 predictors from the optimal random forest model of crinoid presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

In order to evaluate the effects of the bias in the variable Bottom Salinity Average Range on Model 1 performance and its prediction surface, we re-ran the analysis excluding that variable. Model performance was little changed; the AUC was 0.892 with all variables and 0.890 with Bottom Salinity Average Range removed. The top variable in the latter analysis became Surface Temperature Average Maximum, followed by Depth, and the prediction surfaces were very similar. Comparison of the two prediction surfaces showed good overall congruence, although areas of extrapolation had higher probability of occurrence when Bottom Salinity Average Range

was removed, likely due to the higher influence of Depth with which it was highly correlated (Spearman's rank correlation = 0.94). The areas of high probability identified in the original analysis near the opening of the Strait of Belle Isle (Figure 5) were still present but with a lower probability after removing the top predictor. Overall we judged the influence of the bias in the environmental variable Bottom Salinity Average Range to have minimal influence on the model output.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 5 shows the accuracy measures for the random forest model using all crinoid presence and absence data (1325 absences and 105 presences; Model 2) and a threshold equal to species prevalence (0.07). The average AUC calculated from this model was slightly higher than that of Model 1 (0.910 compared to 0.892 of Model 1), indicating excellent model performance. Sensitivity and specificity (0.829 and 0.837 respectively) were lower than that of Model 1. Class error of the presence and absence classes was comparable to Model 1.

The predicted crinoid presence probability surface generated from Model 2 is shown in Figure 9. The areas of high predicted presence probability from Model 1 are greatly reduced in this model. The highest crinoid presence probabilities still occurred on the banks off Labrador and in a small pocket south of Newfoundland. However, the model does not appear to extrapolate high probabilities far beyond the location of presence observations (Figure 10), likely due to the inclusion of all absence records in the model. Figure 11 depicts the classification of crinoid presence probability into presence and absence categories based on the prevalence threshold of 0.07. In this map, all presence probability values generated from Model 2 greater than 0.07 were classified as presence, while values less than 0.07 were classed as absence.

Table 5. Accuracy measures and confusion matrix from 10-fold cross validation from random forest modelling of presence and absence of crinoids within the Newfoundland and Labrador Region. Observ. =Observations, Sensit.= Sensitivity, Specif. = Specificity.

Model Fold	**AUC**	**Observ.**	**Predictions**		**Total n**	**Class error**	**Sensit.**	**Specif.**
1	0.971		**Absence**	**Presence**				
2	0.896	**Absence**	1109	216	1325	0.163	0.829	0.837
3	0.800	**Presence**	18	87	105	0.171		
4	0.885							
5	0.908							
6	0.933							
7	0.939							
8	0.945							
9	0.975							
10	0.854							
Mean	**0.910**							
SD	**0.054**							

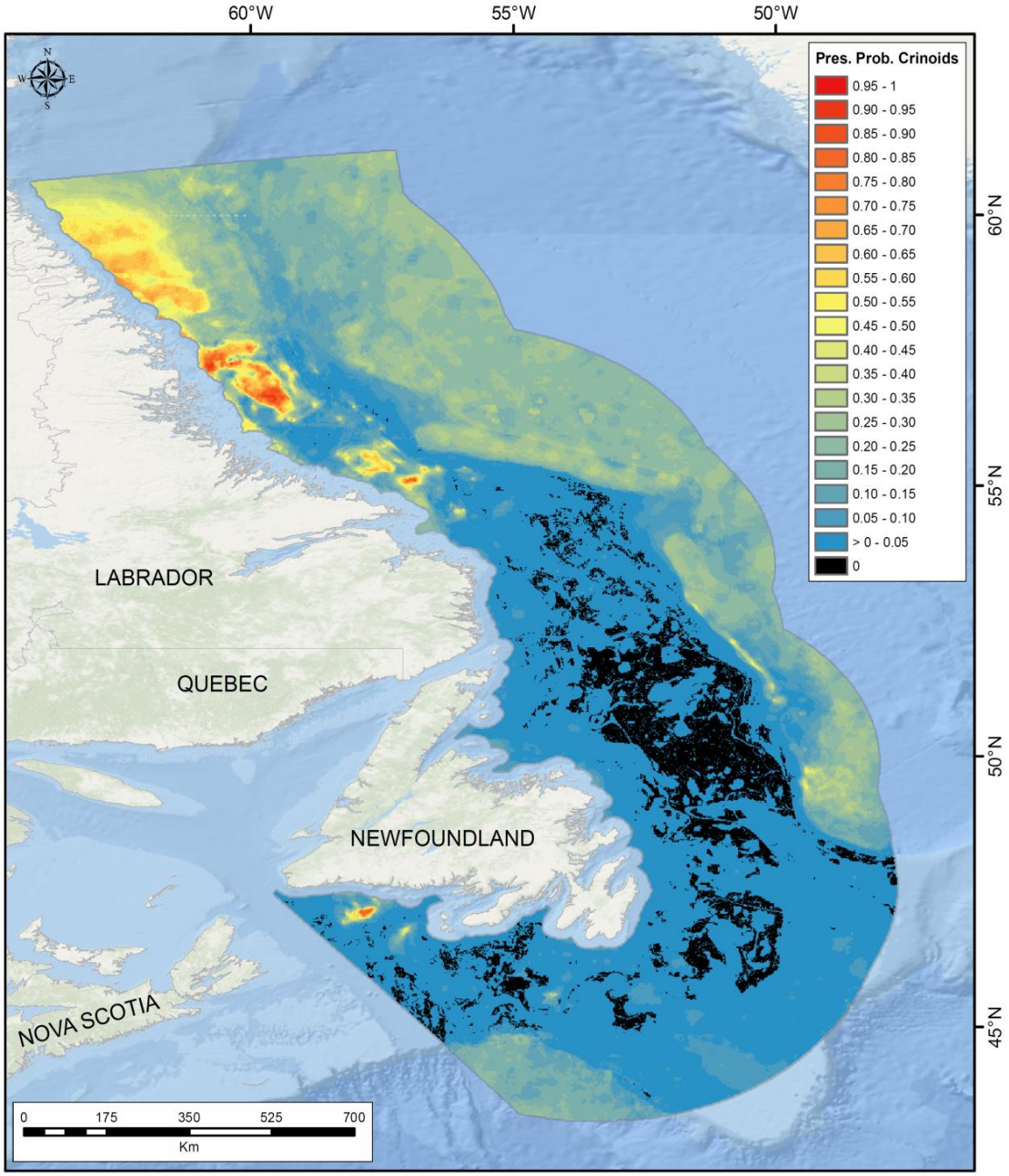

Figure 9. Predictions of presence probability (Pres. Prob.) from the unbalanced random forest model of crinoid presence and absence data collected from DFO multispecies surveys conducted within the Newfoundland and Labrador Region between 2009 and 2015.

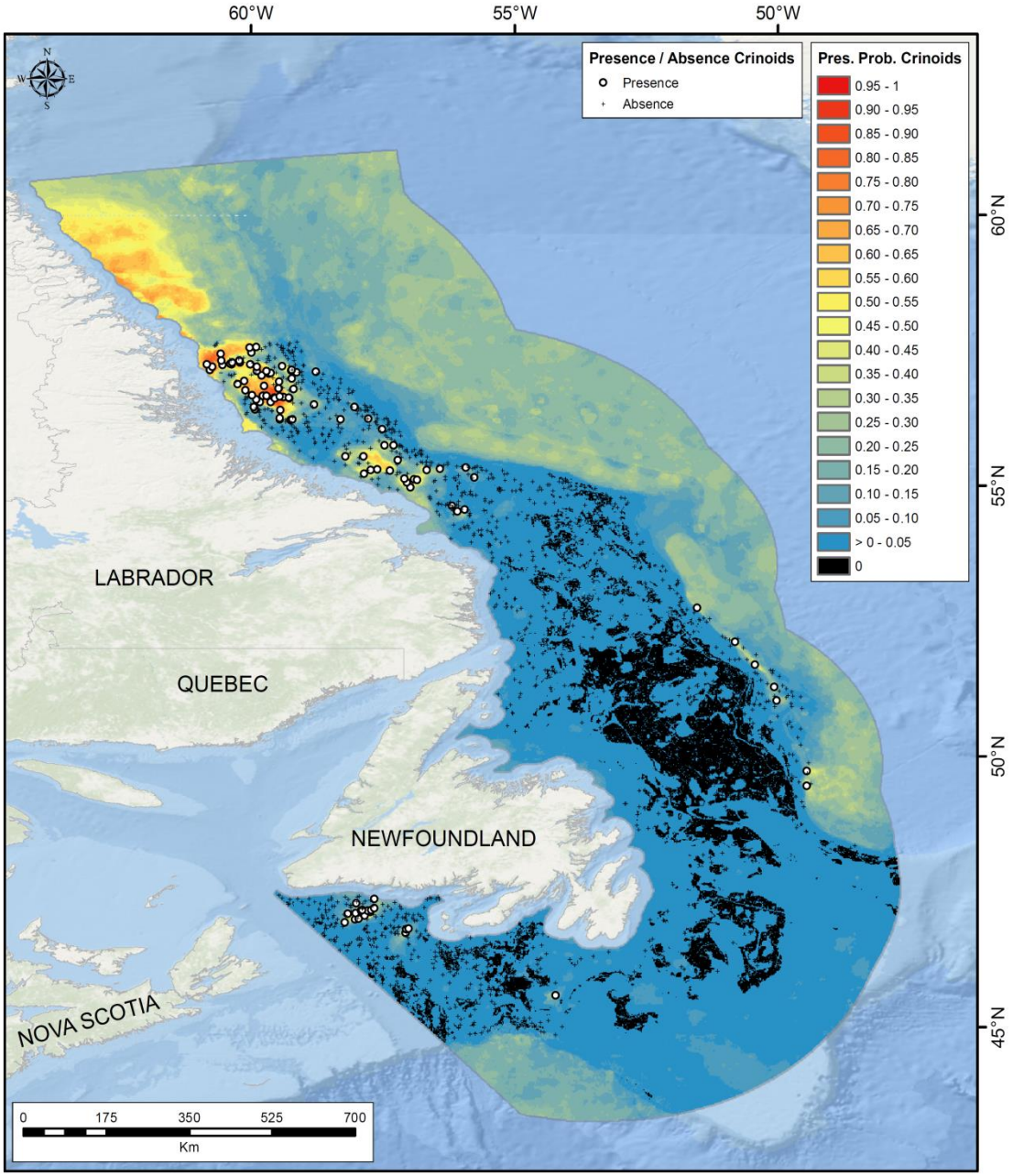

Figure 10. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the unbalanced random forest model of crinoid presence and absence data collected from DFO multispecies surveys conducted within the Newfoundland and Labrador Region between 2009 and 2015.

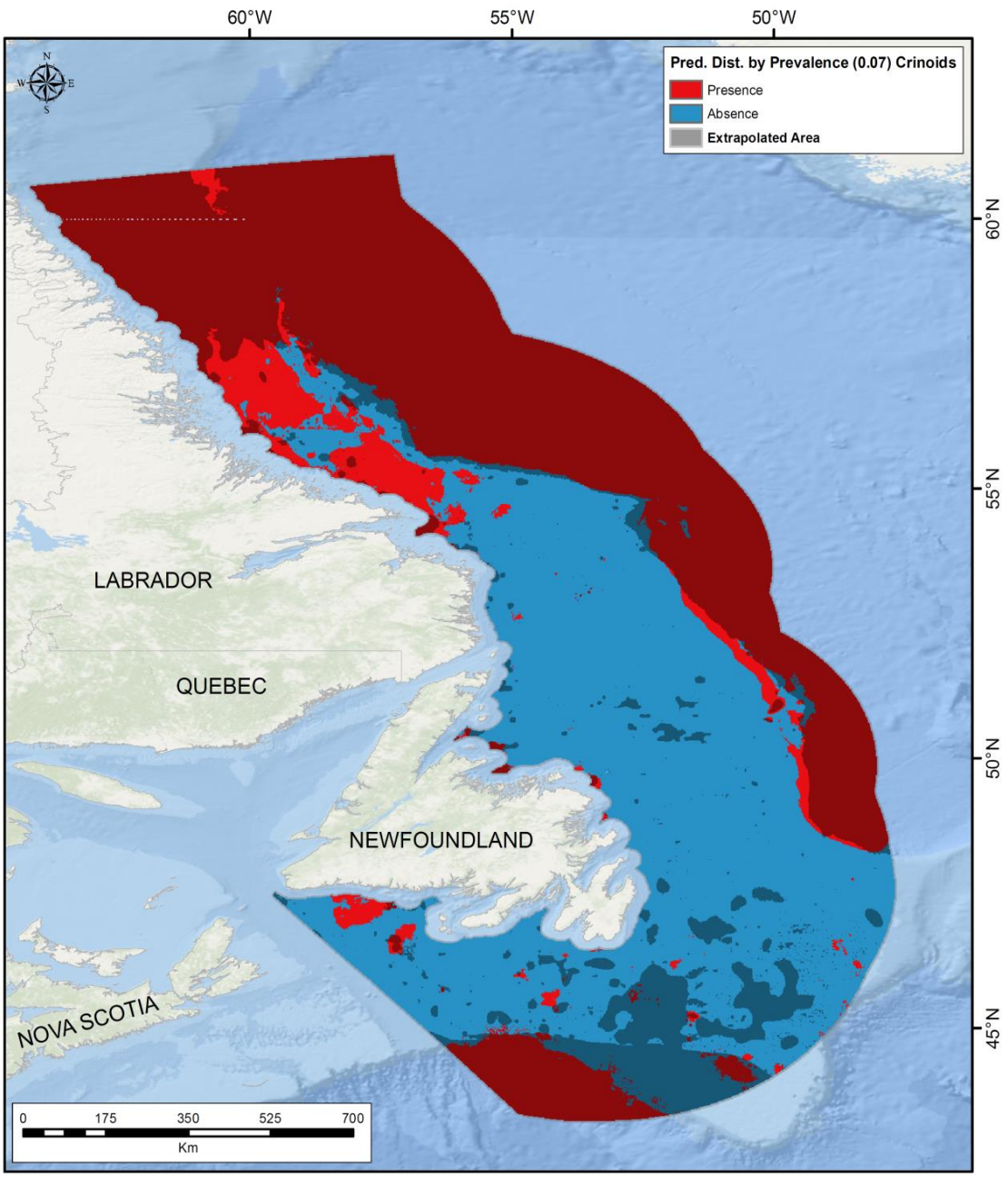

Figure 11. Predicted distribution (Pred. Dist.) of crinoids in the Newfoundland and Labrador Region based on the prevalence threshold of 0.07 of crinoid presence and absence data used in Model 2. Also shown are the areas of model extrapolation. Grey polygon may appear dark red or blue when overlain on the presence-absence surface.

The importance of the environmental predictor variables for predicting the presence probability of crinoids is presented in Figure 12. Surface Temperature Mean was most important for the classification of the crinoid presence and absence data (Figure 12). This variable displayed a near-normal distribution prior to spatial interpolation (Guijarro et al., in prep.). Examination of the Q-Q plot revealed a spatial pattern to those data points over- and under-predicted by a normal

distribution, with over- predicted points located mainly in the northern portion of the study extent on Saglek Bank and in the south of the Grand Banks of Newfoundland, and under-predicted points located on the Grand Banks of Newfoundland and the Northeast Newfoundland Shelf. However, the semivariogram showed weak autocorrelation present in the data and there was an excellent fit between the predicted and measured values. Surface Temperature Mean was followed closely in importance by Depth (non-interpolated variable) and Spring Primary Production Average Range. Partial dependence plots for the top 6 predictor variables are shown in Figure 13. The highest presence probability of crinoids along the gradient in Surface Temperature Mean occurred at ~1.5 °C. Along the Depth gradient, high presence probability occurred in two peaks; one at ~200 m and the other at ~1200 m. This may represent differences in the depth distribution of the different species within this taxonomic group.

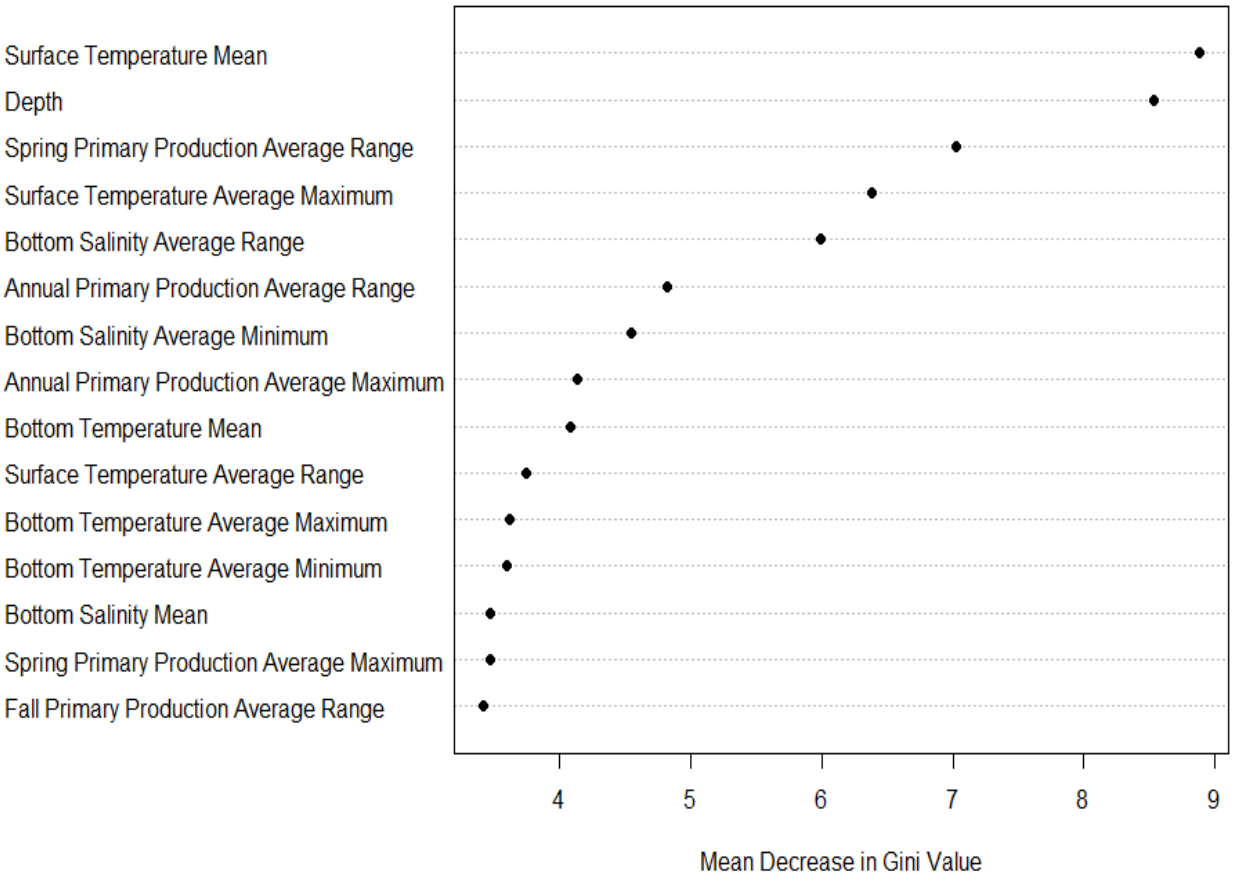

Figure 12. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini Value of the unbalanced random forest model predicting crinoid presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini Value the more important the variable is for predicting the response data.

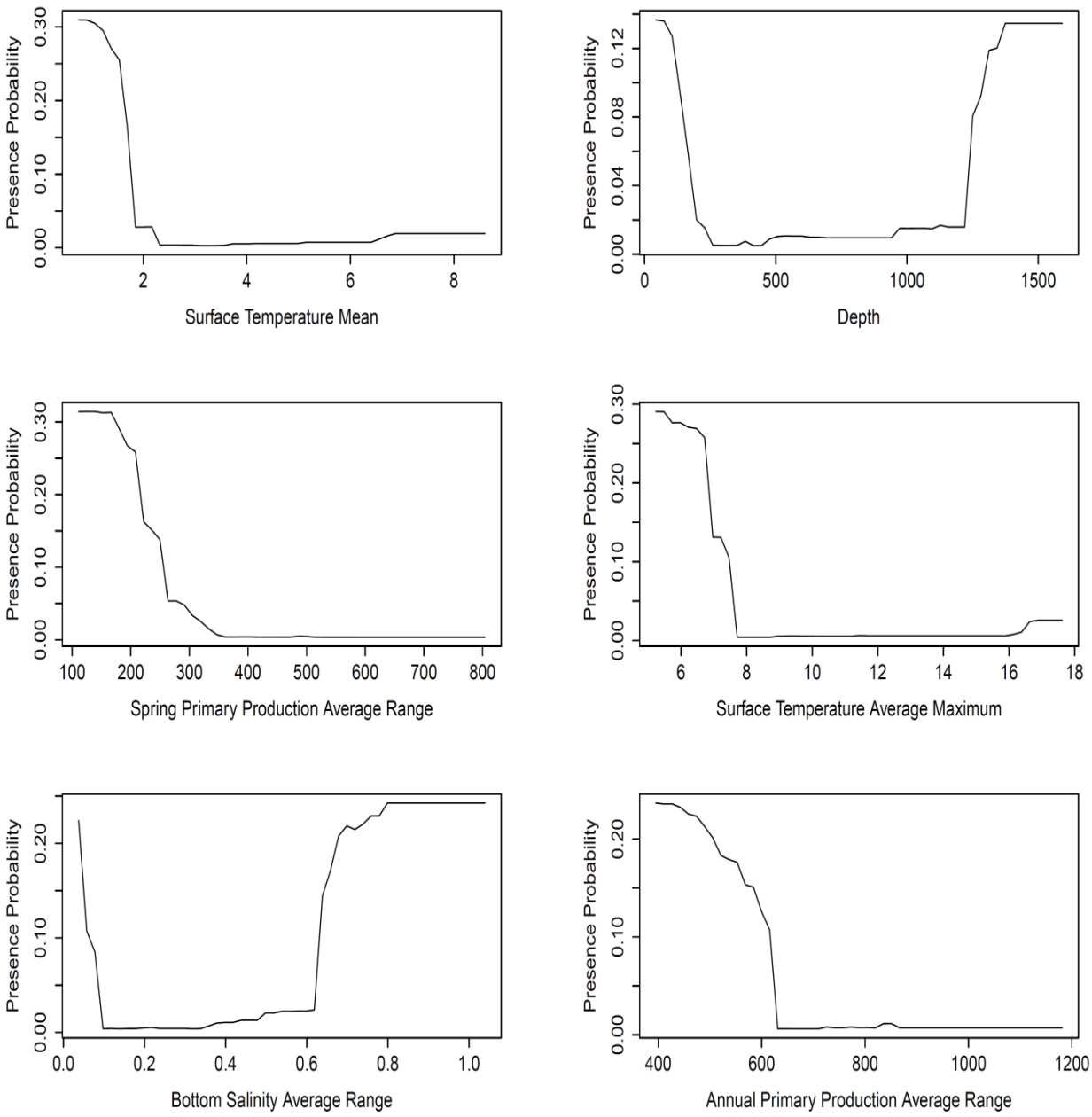

Figure 13. Partial dependence plots of the top 6 predictors from the unbalanced random forest model of crinoid presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using all available crinoid records and an unbalanced species prevalence and threshold equal to 0.07 (Model 2) was chosen as the best predictor of crinoid distribution in the Newfoundland and Labrador Region. The AUC for Model 2 is slightly higher than that for Model 1. Model 1 (balanced species prevalence) was considered a less good predictor of presence probability of crinoid due to its prediction of high presence probability beyond the location of presence data. It is possible that those predictions are accurate but without data to validate the models we feel that the Model 2 output should be used in preference to Model 1 output.

Prediction of Biomass using Random Forest

The accuracy measures of the regression random forest model on mean crinoid biomass per grid cell are presented in Table 6. The highest R^2 value was 0.234, while the average was 0.119 ± 0.080 SD. The average Normalized Root-Mean-Square Error (RMSE) was 0.051 ± 0.026 SD. This model explained a low percentage of variance in the biomass data (average = 2.69% ± 2.62 SD).

Figures 14 and 15 show the crinoid biomass (kg) predictions per grid cell generated by the random forest model. Saglek Bank off northern Labrador was predicted to have the highest biomass of crinoids, despite there being no data observations to support it (see Figure 15). This area is considered an area of extrapolation by the model and should be a priority for validation efforts. This area was also predicted to have high species occurrence probability under Model 1 (Figures 4-6) and under the Model 2 prevalence map (Figures 9-11).

Table 6. Accuracy measures from 10-fold cross validation from random forest modelling of average crinoid biomass (kg) per grid cell recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2009 and 2015. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	0.234	0.163	0.028	2.03
2	0.081	0.281	0.048	2.85
3	0.060	0.282	0.048	3.42
4	0.116	0.485	0.082	2.72
5	0.126	0.461	0.078	-0.14
6	0.019	0.157	0.027	0.24
7	0.124	0.413	0.070	3.01
8	0.208	0.185	0.032	0.53
9	0.212	0.197	0.034	3.03
10	0.009	0.556	0.094	9.18
Mean	**0.119**	**0.318**	**0.051**	**2.69**
SD	**0.080**	**0.149**	**0.026**	**2.62**

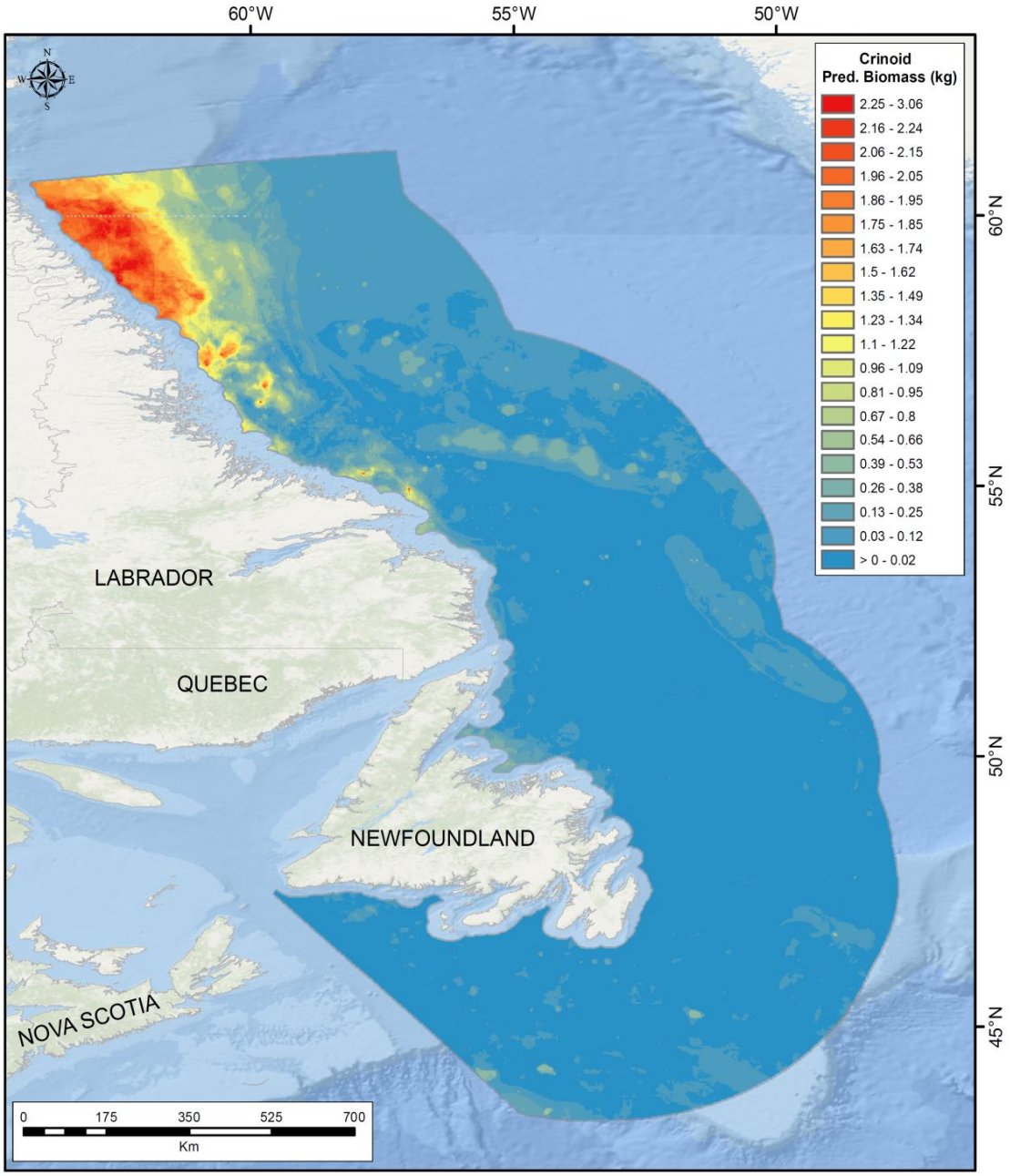

Figure 14. Predictions of biomass (kg) of crinoids from catch data recorded in DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2009 and 2015.

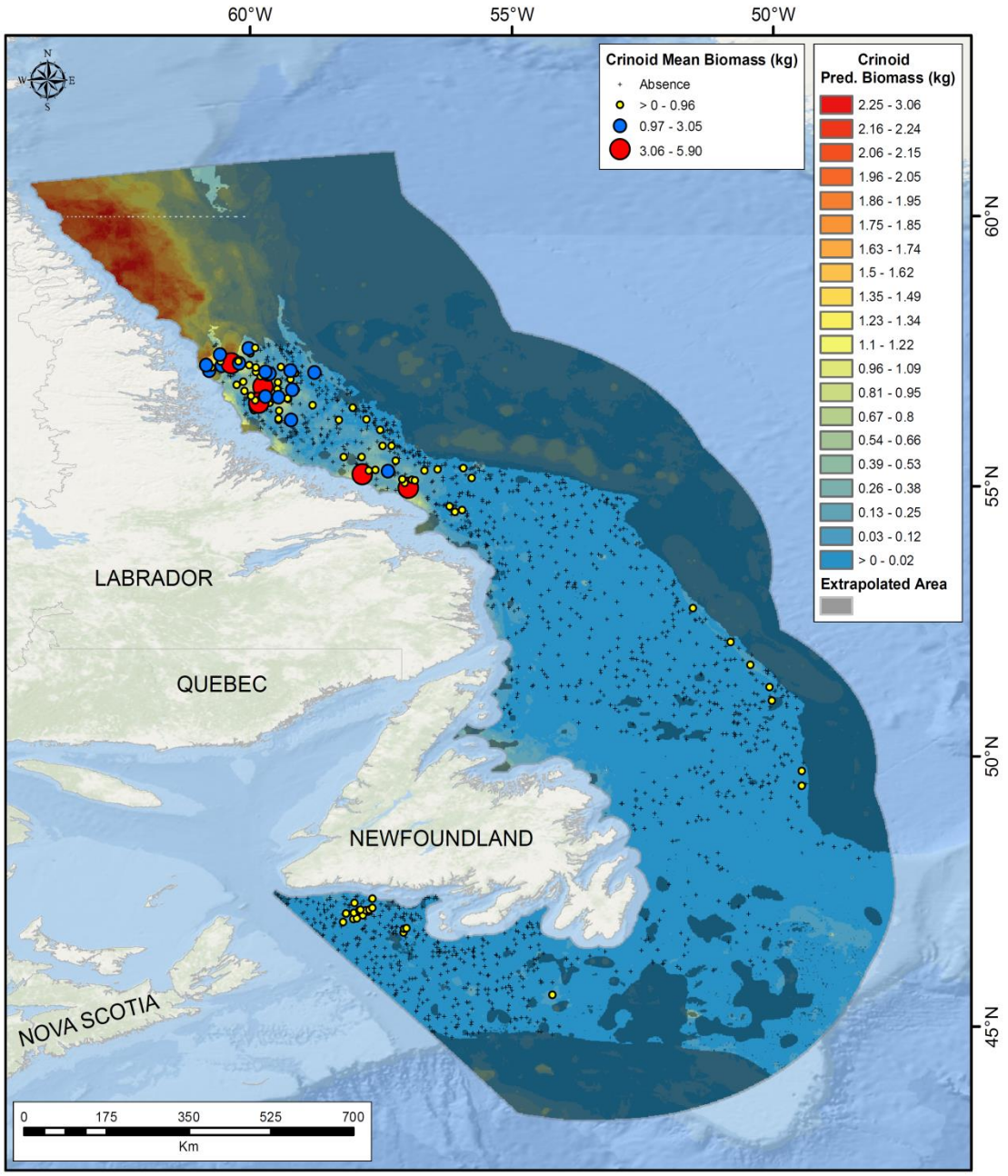

Figure 15. Predictions of biomass (kg) of crinoids from catch data recorded in DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2009 and 2015. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

Like in Model 2 generated on the presence-absence data, Surface Temperature Mean was the most important variable for predicting the biomass distribution of crinoids (Figure 16). The partial dependence of crinoid biomass on the top 6 most important variables is shown in Figure 17. Predicted biomass was highest at the lowest Surface Temperature Mean (< 2 °C). Consequently the over prediction of this variable may have distorted the model prediction surface. Surface Temperature Mean was followed distantly by Bottom Salinity Average Range, and the remaining variables in the model (Figure 16).

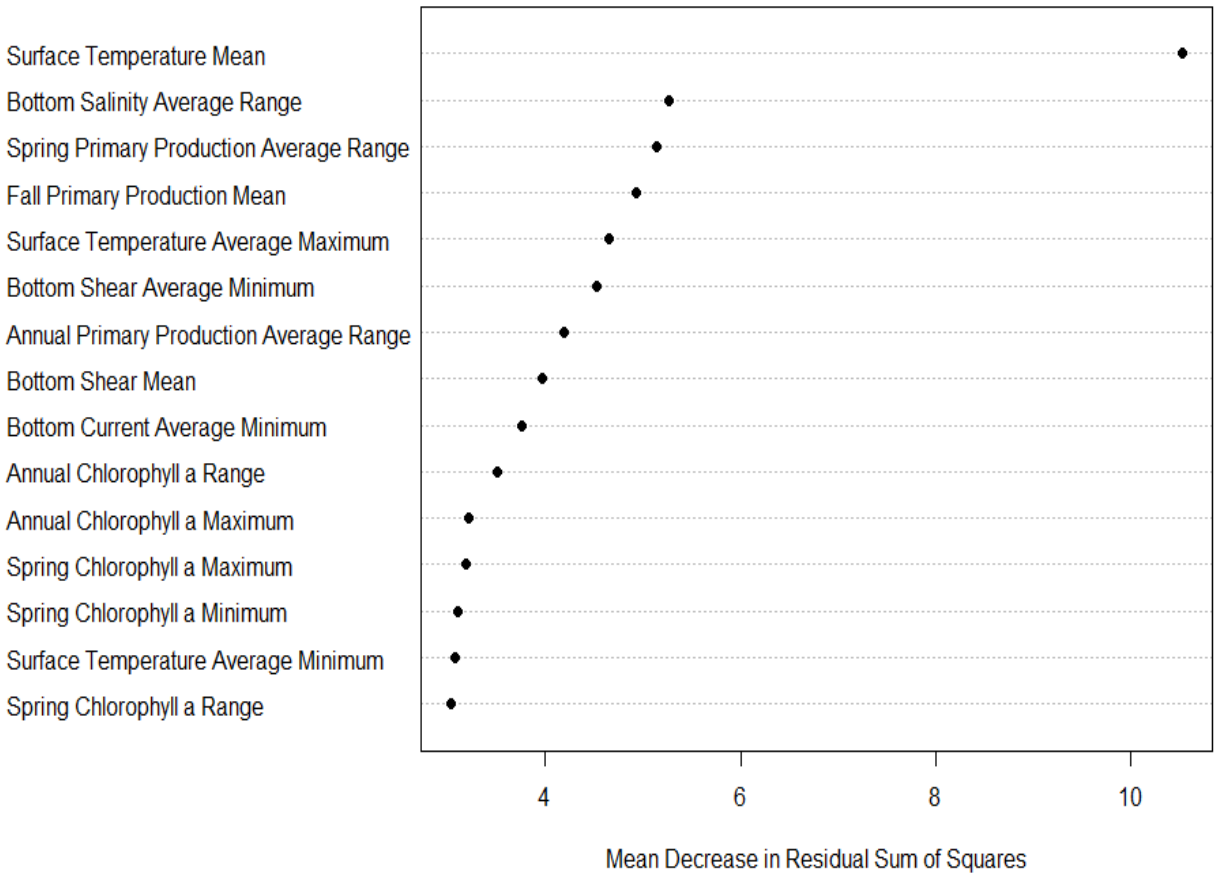

Figure 16. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on crinoid mean biomass data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

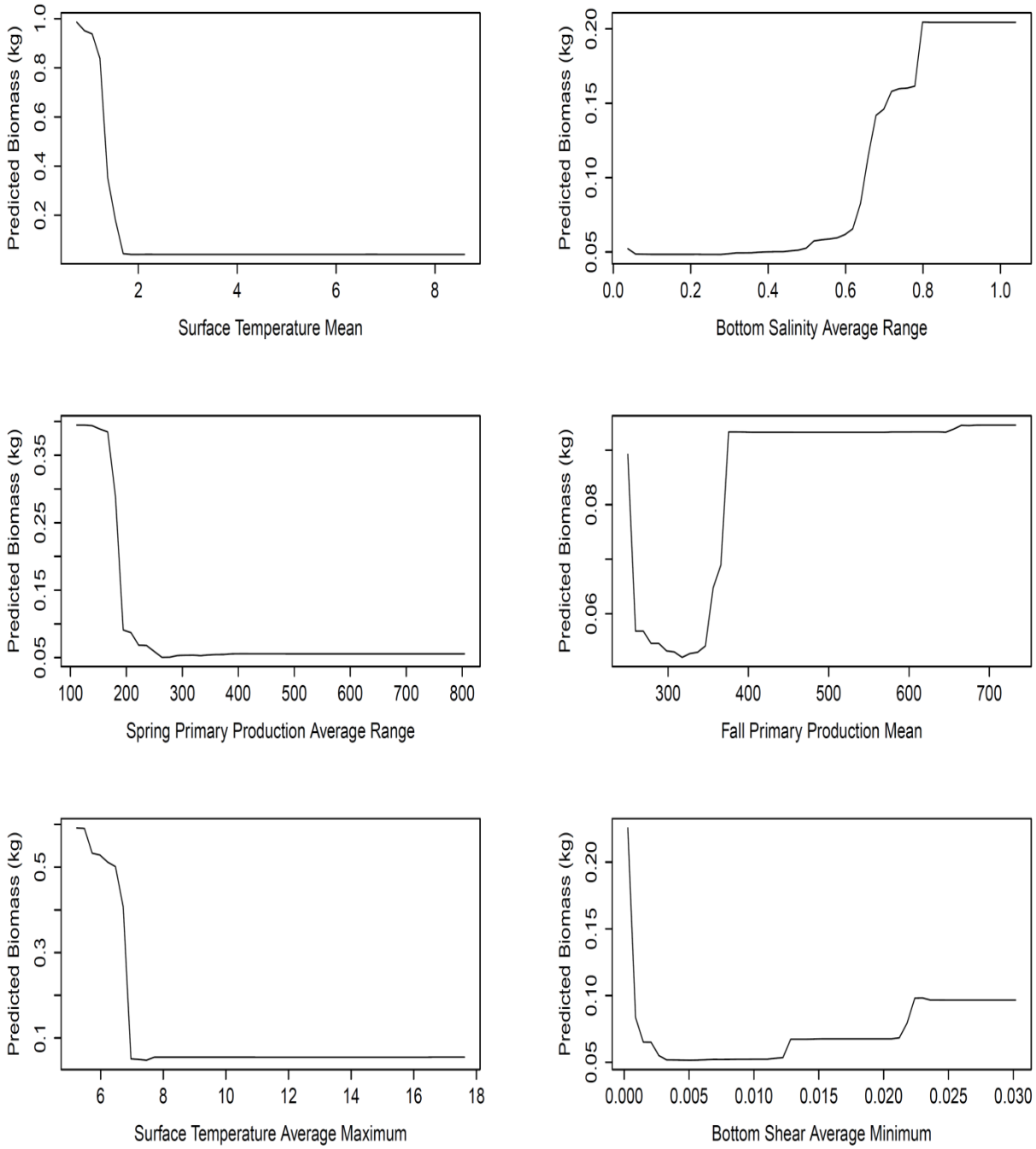

Figure 17. Partial dependence plots of the top six predictors from the random forest model of crinoid biomass data collected within the Newfoundland and Labrador Region between 2009 and 2015, ordered left to right from the top. Predicted biomass is shown on the *y*-axis of each graph.

Ascidians (including Large Sea Squirts)

Data Sources and Distribution

Ascidian (Phylum Chordata: Class Ascidiacea) catch data was collected over a span of 10 years from 2006 to 2015 (Table 7) and consisted of 654 presence and 5592 absence records. Presence and absence records were distributed relatively evenly across the study extent. Both presences and absences were absent from the northern Labrador Shelf and beyond the continental slope. The highest mean biomass records (up to 59.85 kg) were located southwest of Saint-Pierre and Miquelon. A few large catches also occurred southeast of Newfoundland on Grand Bank (Figure 18).

Table 7. Number of presence and absence records of ascidian catch recorded from DFO multispecies surveys conducted between 2006 and 2015 in the Newfoundland and Labrador Region.

Year	Total number of presences	Total number of absences
2006	5	62
2007	31	320
2008	28	437
2009	104	631
2010	101	828
2011	68	687
2012	66	892
2013	104	811
2014	115	606
2015	32	318

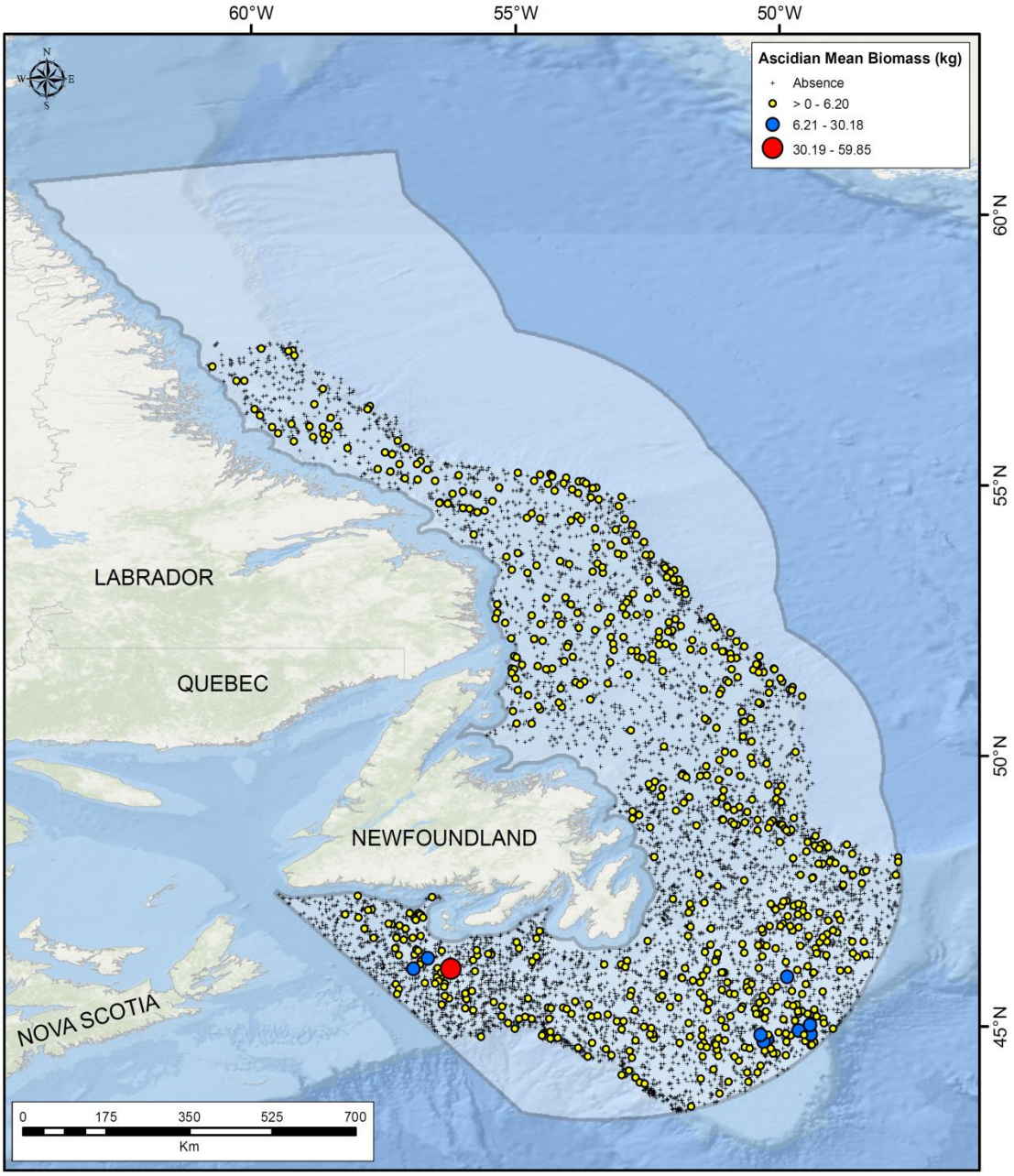

Figure 18. Mean biomass (kg) per grid cell of ascidians recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2006 and 2015. Also, shown are absence records from the same surveys.

Model 1 – Balanced Species Prevalence

Accuracy measures (mean AUC, sensitivity and specificity) for the random forest model on balanced species prevalence (654 presences and 654 absences; Model 1) are presented in Table 8. The average AUC was 0.669, indicating only a fair model performance. The highest AUC of 0.703 was associated with Model Run 3. The sensitivity and specificity measures of this model were 0.648 and 0.636, respectively. The confusion matrix of this model is also presented in Table 8. Class errors for both the presence and absence classes were relatively moderate (0.352 and 0.364, respectively).

Table 8. Accuracy measures for all 10 model repetitions of 10-fold cross validation from the random forest model of ascidian presence-absence data collected within the Newfoundland and Labrador 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 the presence probability of ascidians.

Model Run	AUC	Sensitivity	Specificity
1	0.654	0.622	0.599
2	0.678	0.627	0.639
3	**0.703**	**0.648**	**0.636**
4	0.656	0.648	0.601
5	0.673	0.659	0.627
6	0.676	0.648	0.628
7	0.676	0.661	0.609
8	0.624	0.618	0.564
9	0.651	0.596	0.609
10	0.694	0.639	0.658
Mean	**0.669**	**0.637**	**0.626**
SD	**0.023**	**0.020**	**0.026**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	416	238	654	0.364
Presence	230	424	654	0.352

The presence probability prediction surface of ascidians is presented in Figure 19. The highest predictions of presence probability occurred on Grand Bank, on the northeast Newfoundland Shelf and slopes of Newfoundland and Labrador. These areas of high presence probability correspond well with the spatial distribution of presence records (Figure 20). However, the model extrapolates to large predicted areas of presence probability beyond the location of presence observations, particularly in deeper waters off Labrador. Figure 21 shows the actual presence and absence data observations (654 presences and 654 absences) used in the optimal Model 1. Areas of extrapolation are also shown in this figure. The area of high predicted presence probability of ascidians off the Labrador Slope was considered extrapolated area.

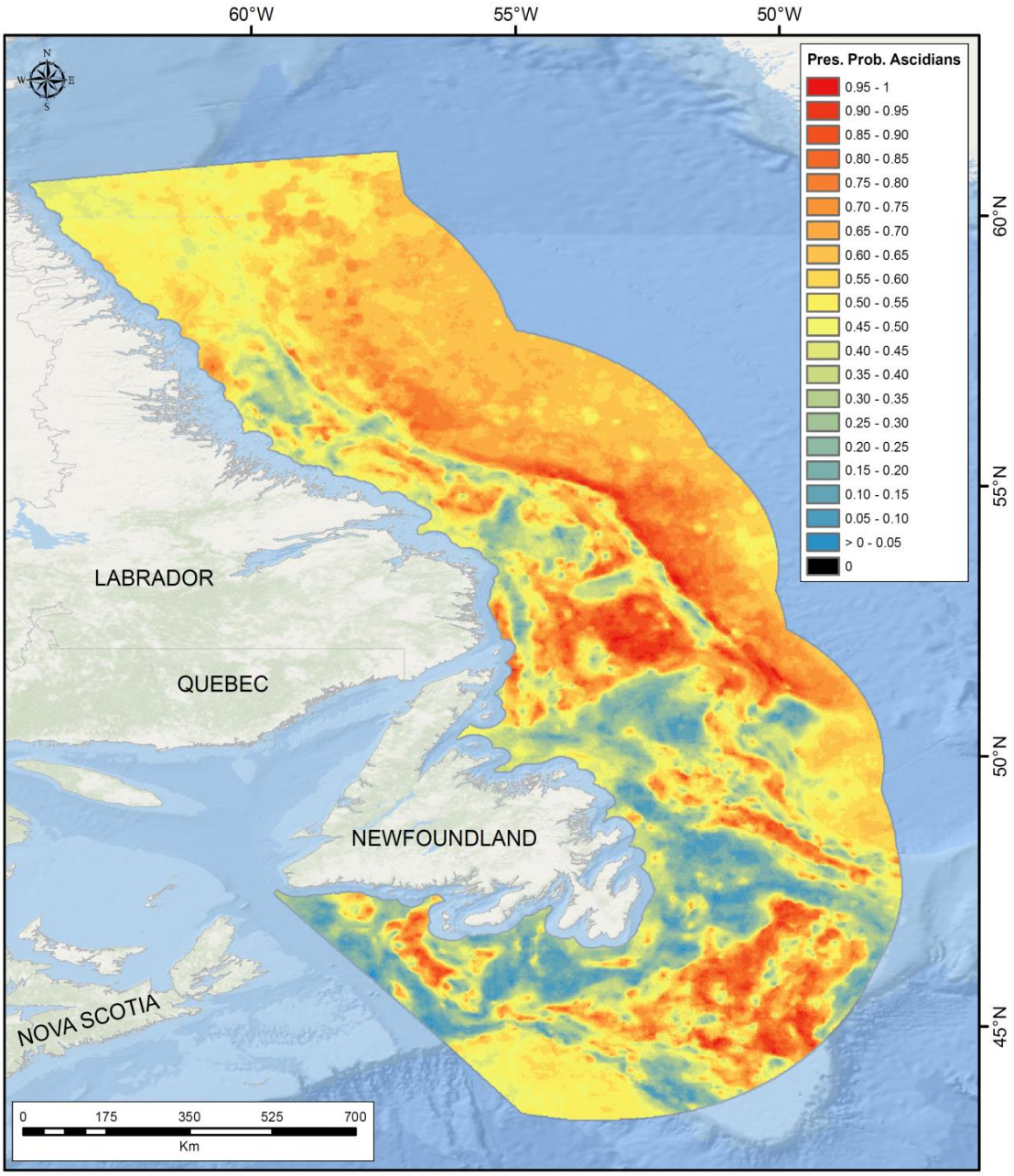

Figure 19. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of ascidian presence and absence data collected from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2006 and 2015.

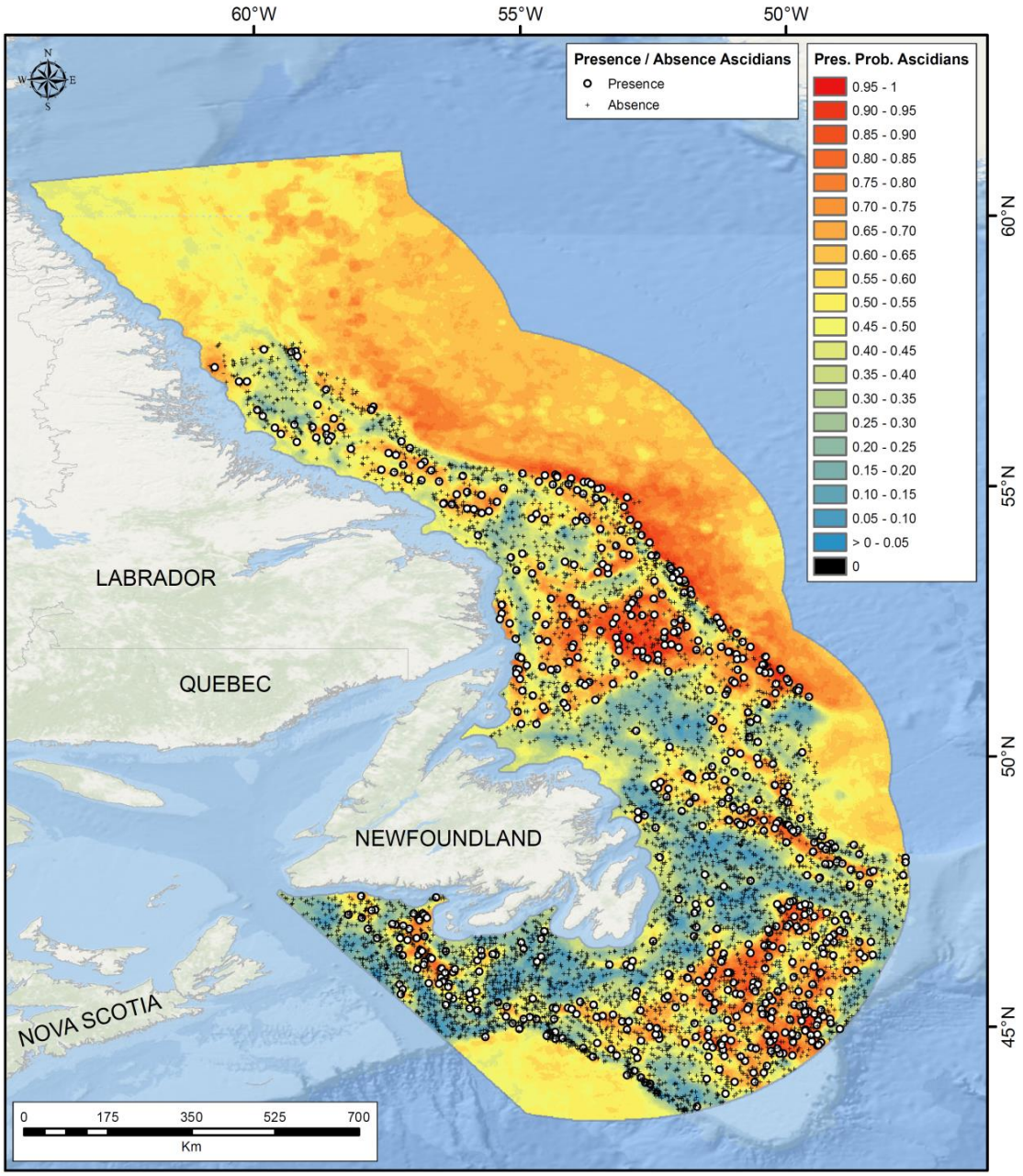

Figure 20. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of ascidian presence and absence data recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2006 and 2015.

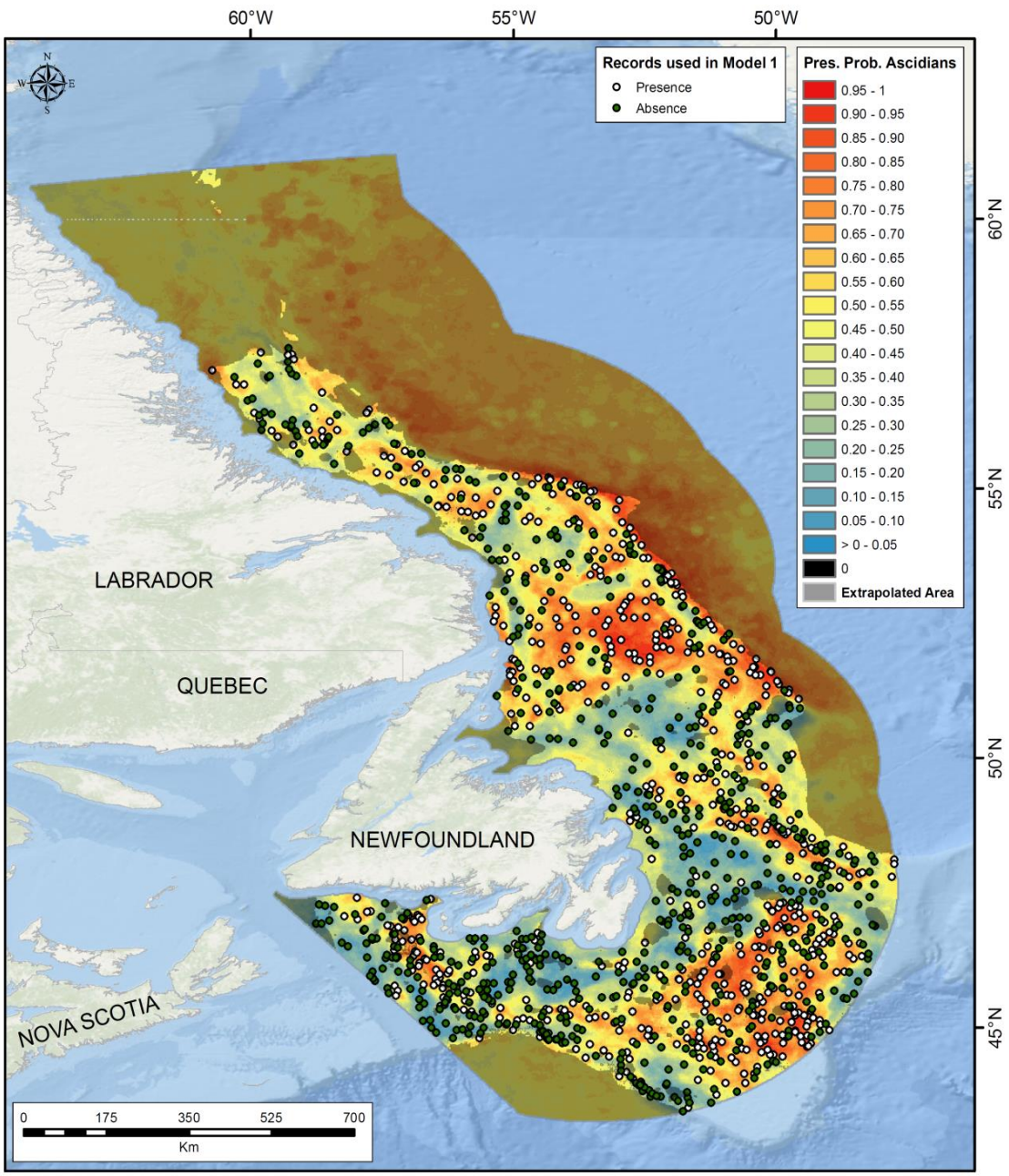

Figure 21. Map of the 1308 data observations (654 presences and 654 absences) of ascidians used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of ascidians generated from Model 1 and areas of model extrapolation.

Of all 66 environmental predictor variables used in the model, Annual Chlorophyll *a* Mean was most important for the classification of the ascidian presence-absence data (Figure 22). This variable displayed a right-skewed distribution prior to spatial interpolation (Guijarro et al., in prep). The data were higher than predicted by a normal distribution at the upper and lower range and lower than predicted at mid values. Examination of the Q-Q plot revealed no strong spatial pattern to the points over- and under-predicted by a normal distribution. This variable was followed by Spring Chlorophyll *a* Mean and Surface Salinity Mean. Partial dependence plots for the top 6 predictor variables are shown in Figure 23. The highest presence probability of ascidians along the gradient in Annual Chlorophyll *a* Mean occurred between 0.8 and 1 mg m^{-3}.

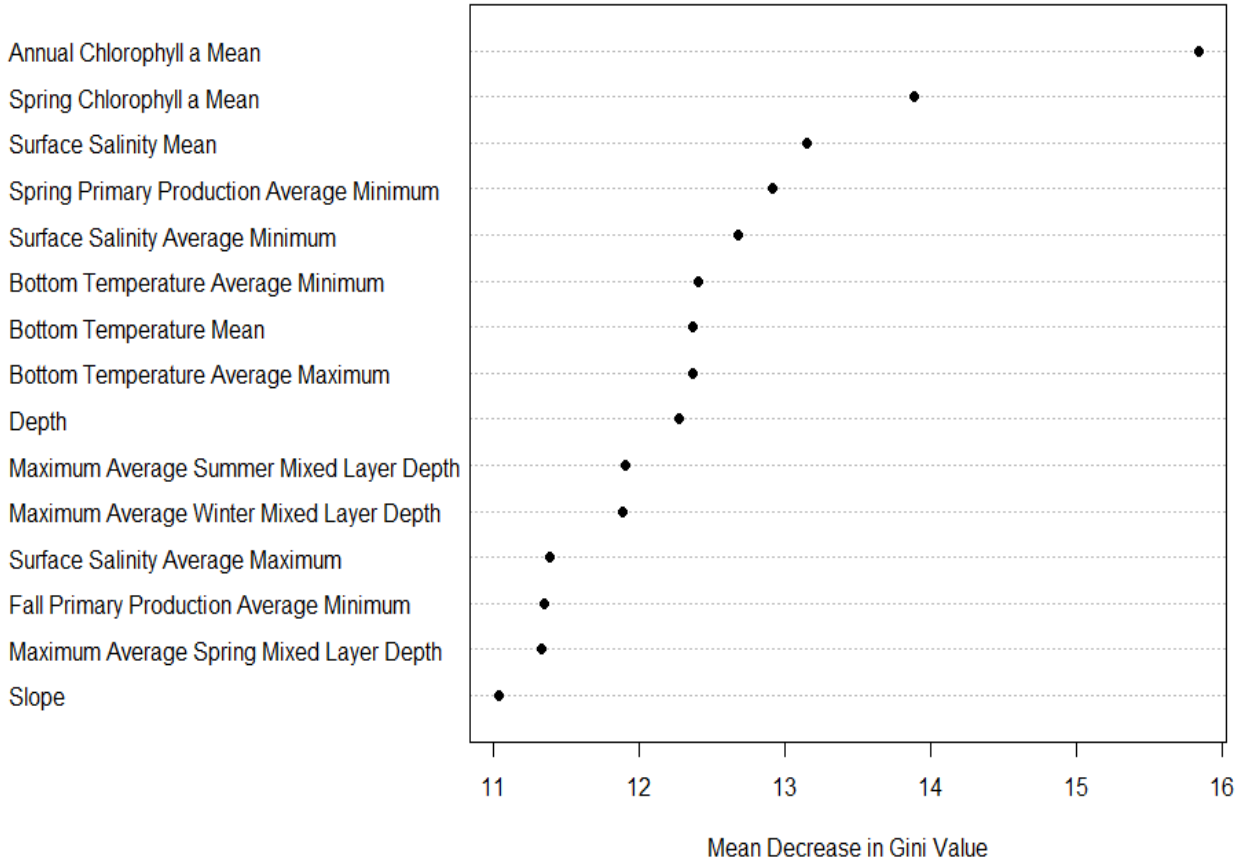

Figure 22. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting ascidian presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini value the more important the variable is for predicting the response data.

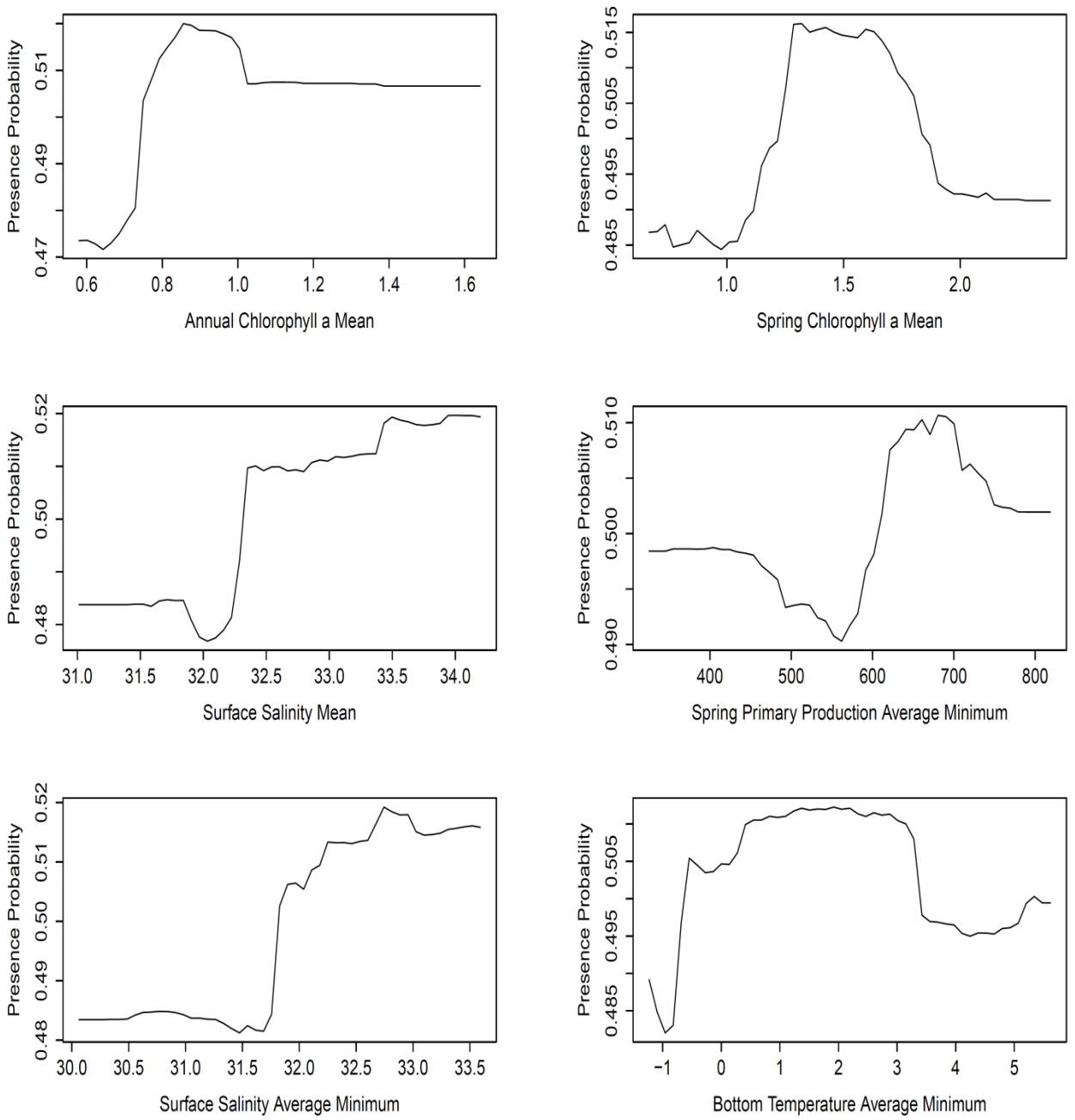

Figure 23. Partial dependence plots of the top 6 predictors from the optimal random forest model of ascidian presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 9 shows the accuracy measures for the random forest model using all ascidian presence and absence data (5592 absences and 654 presences; Model 2) and a threshold equal to species prevalence (0.10). The average AUC calculated from this model was slightly higher than that of Model 1 (0.671 compared to 0.669 of Model 1). Sensitivity was lower than that of Model 1 while specificity was higher. Class error of the presence and absence classes was comparable to Model 1.

The predicted ascidian presence probability surface generated from Model 2 is shown in Figure 24. The areas of high predicted presence probability from Model 1 are greatly reduced in this model. The highest ascidian presence probability occurred in small pockets on the Northeast Newfoundland Shelf and along the slopes of Newfoundland and Labrador. However, the model does not appear to extrapolate high probabilities far beyond the location of presence observations (Figure 25), likely due to the inclusion of all absence records in the model. Figure 26 depicts the classification of ascidian presence probability into presence and absence categories based on the prevalence threshold of 0.10. In this map, all presence probability values generated from Model 2 greater than 0.10 were classified as presence, while values less than 0.10 were classed as absence.

Table 9. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model presence and absence of ascidians within the Newfoundland and Labrador Region. Observ. = Observations, Sensit.= Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.646		**Absence**	**Presence**				
2	0.649	**Absence**	3567	2025	5592	0.362	0.596	0.638
3	0.666	**Presence**	264	390	654	0.404		
4	0.699							
5	0.695							
6	0.617							
7	0.664							
8	0.706							
9	0.691							
10	0.678							
Mean	**0.671**							
SD	**0.028**							

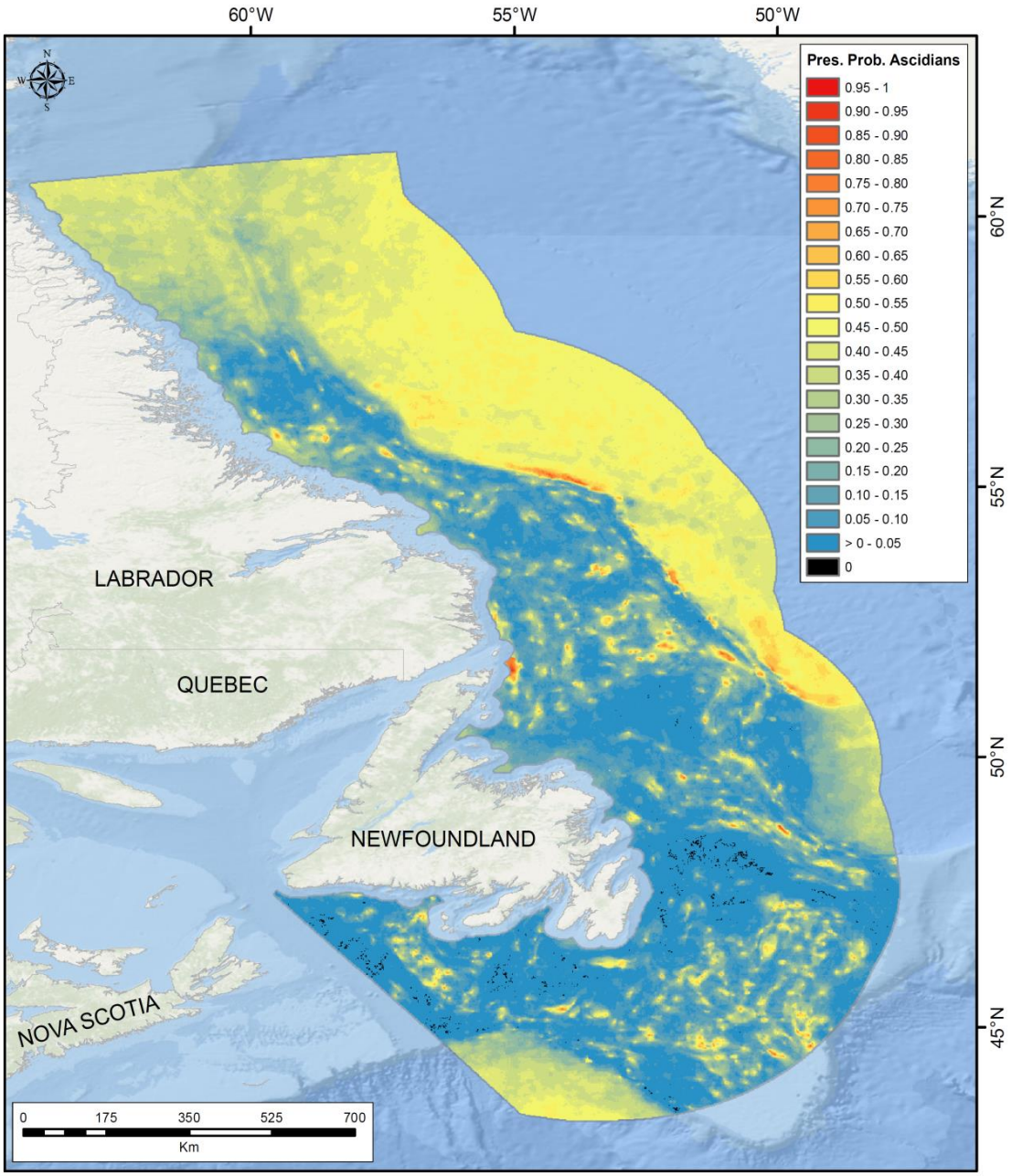

Figure 24. Predictions of presence probability (Pres. Prob.) from the unbalance random forest model of ascidian presence and absence data collected from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2006 and 2015.

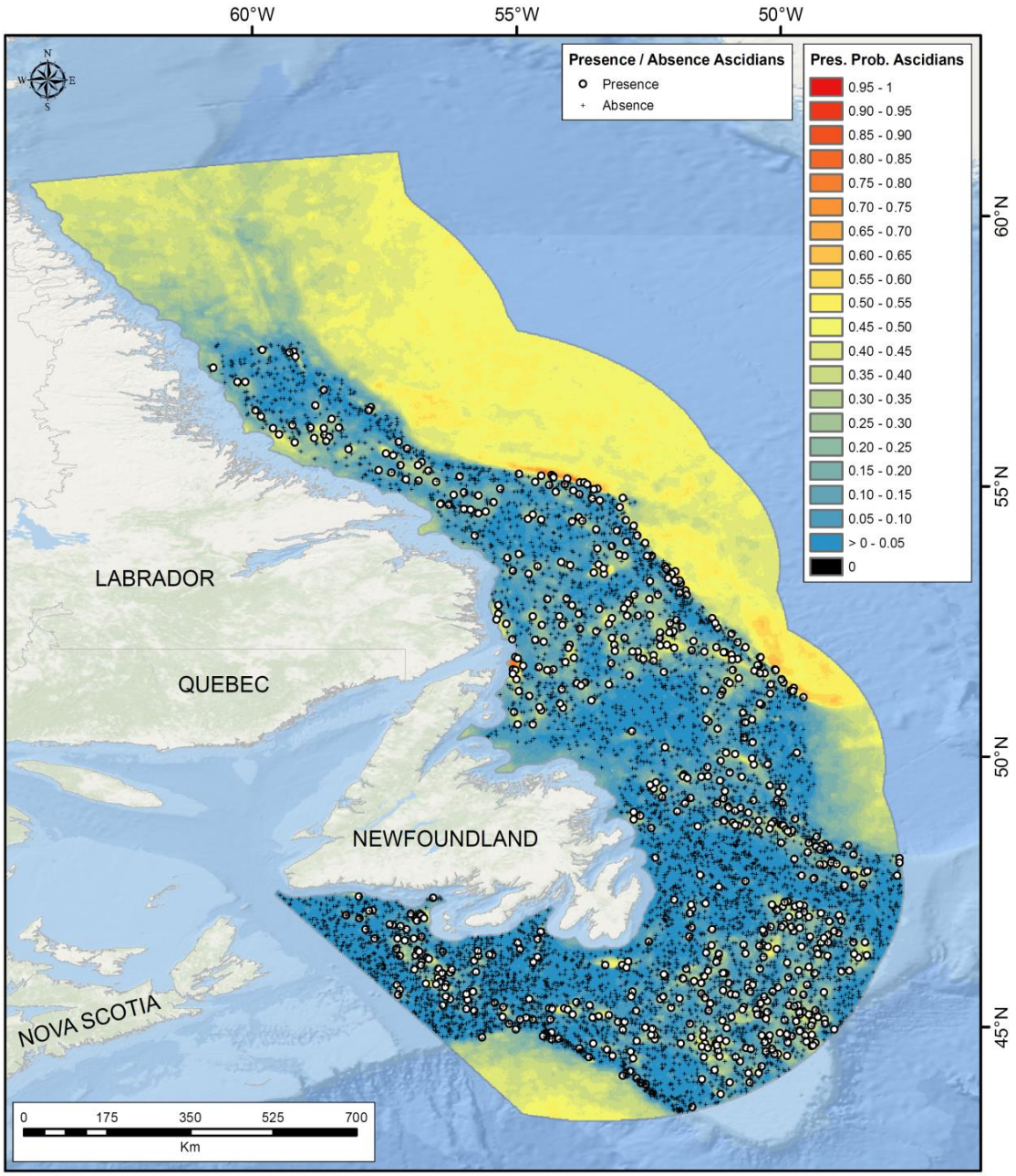

Figure 25. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the unbalanced random forest model of ascidian presence and absence data recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2006 and 2015.

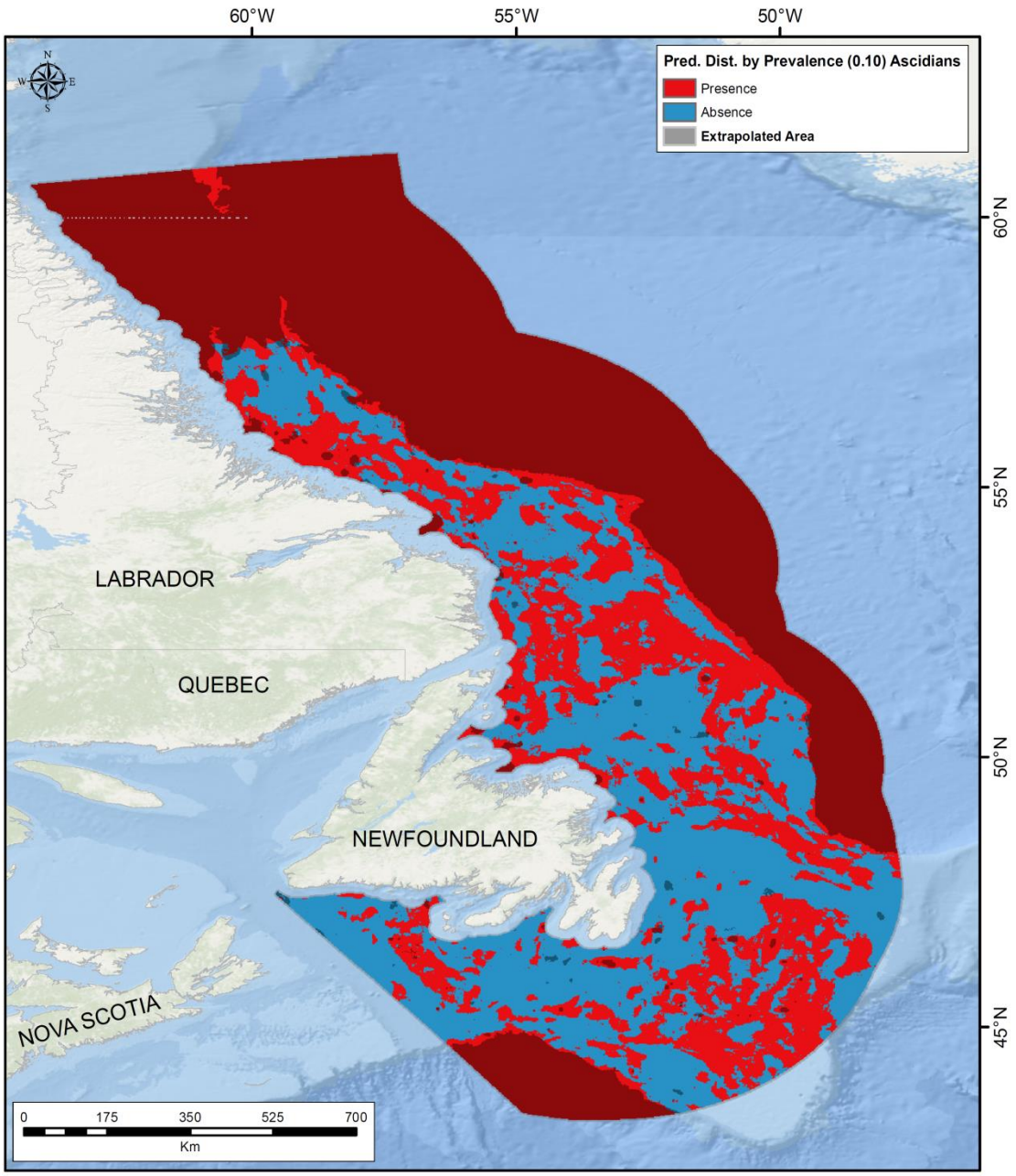

Figure 26. Predicted distribution (Pred. Dist.) of ascidians in the Newfoundland and Labrador Region based on the prevalence threshold of 0.10 of ascidian presence and absence data used in Model 2. Also shown are the areas of model extrapolation. Grey polygon may appear dark red or blue when overlain on the presence-absence surface.

The importance of the environmental predictor variables for predicting the presence probability of ascidians is presented in Figure 27. Depth (non-interpolated variable) was most important for the classification of the ascidian presence-absence data (Figure 27). Depth was followed more distantly in importance by Annual Chlorophyll *a* Minimum and Summer Chlorophyll *a* Mean. Partial dependence plots for the top 6 predictor variables are shown in Figure 28. Presence probability was highest at the shallowest depths.

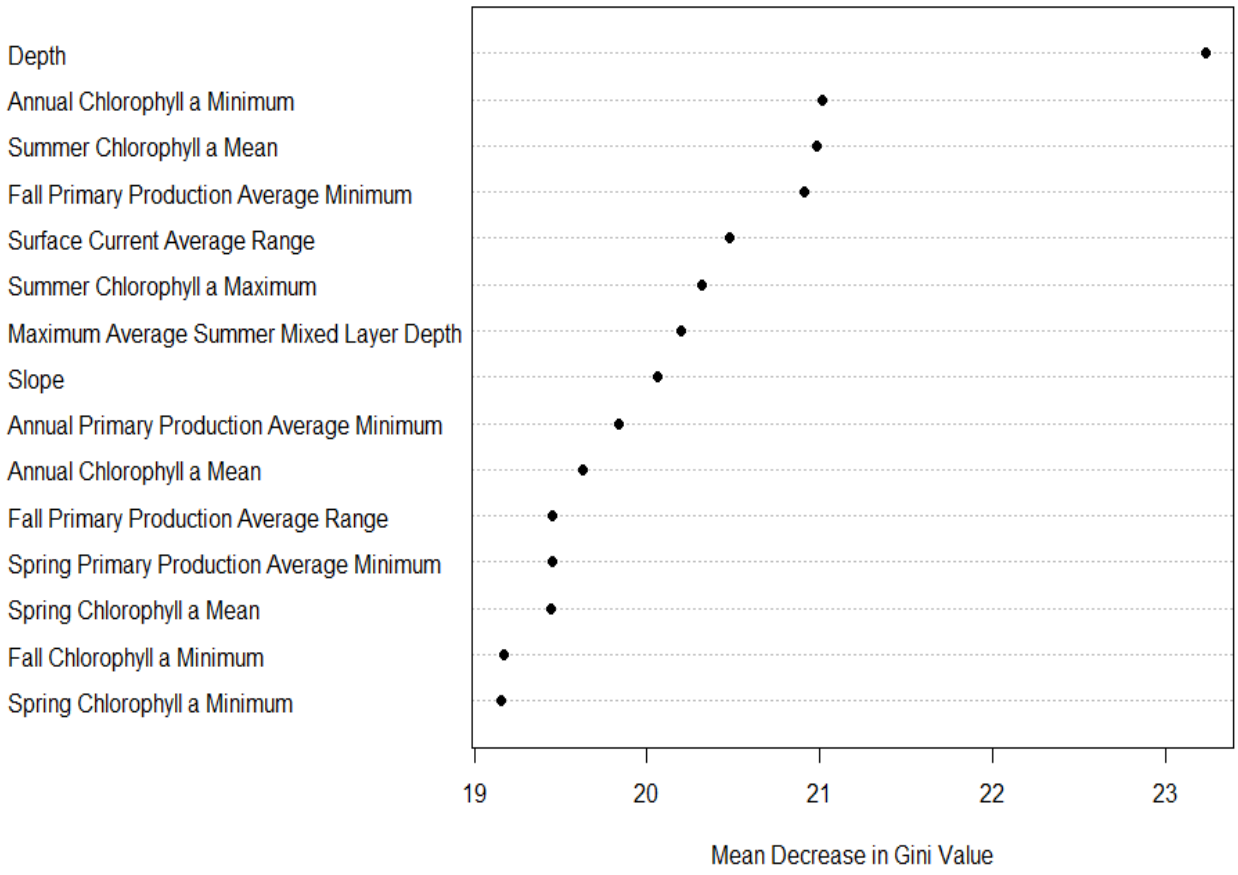

Figure 27. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini Value of the unbalanced random forest model of ascidian presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini Value the more important the variable is for predicting the response data.

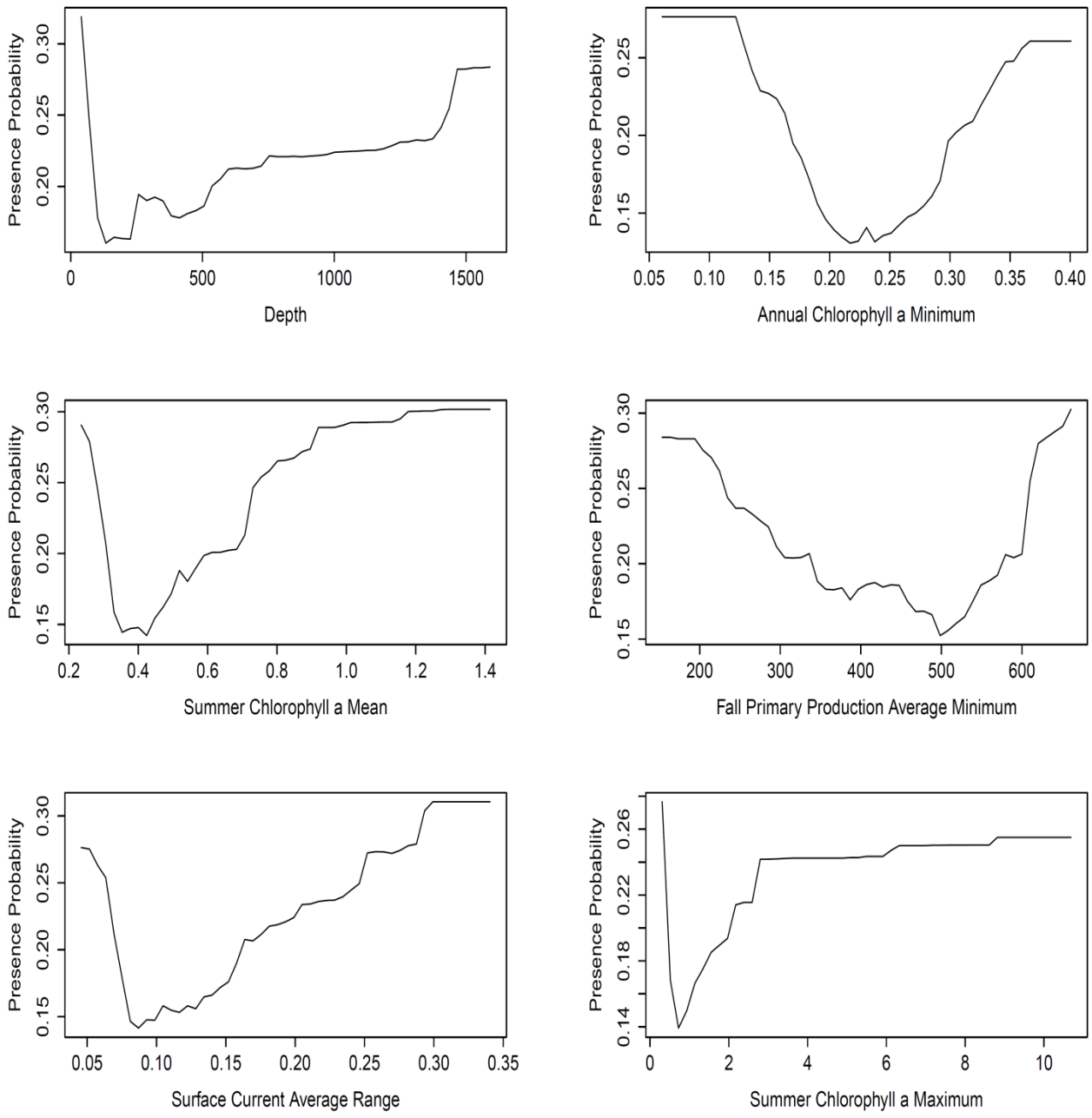

Figure 28. Partial dependence plots of the top 6 predictors from the unbalanced random forest model of ascidian presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using all available ascidian records and an unbalanced species prevalence and threshold equal to 0.10 (Model 2) was chosen as the best predictor of ascidian distribution in the Newfoundland and Labrador Region. Despite comparable AUC values, Model 1 (balanced species prevalence) was considered a poorer predictor of presence probability of ascidians due to its identification of high presence probability beyond the location of presence data in the extrapolated areas, particularly along the slopes and in deeper waters off Newfoundland and Labrador. This phenomenon is likely due to random down-sampling of the absence data.

Prediction of Biomass using Random Forest

The accuracy measures of the regression random forest model on mean ascidian biomass per grid cell are presented in Table 10. The highest R^2 value was 0.051, while the average was 0.013 ± 0.016 SD. The average Normalized Root-Mean-Square Error (RMSE) was 0.015 ± 0.013 SD. The percent variance explained for each fold was negative, indicating that the model had no predictive power. Therefore, the predictive surfaces of this model are not displayed in this report.

Table 10. Accuracy measures from 10-fold cross validation from random forest modelling of average ascidian biomass (kg) per grid cell recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2006 and 2015. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	0.001	0.561	0.011	-6.63
2	0.003	0.478	0.009	-6.24
3	0.019	0.335	0.007	-6.30
4	0.002	0.804	0.016	-5.01
5	0.004	0.468	0.009	-5.24
6	0.006	0.408	0.008	-4.54
7	0.026	0.366	0.007	-4.47
8	0.003	1.322	0.026	-5.46
9	0.051	0.578	0.011	-4.18
10	0.015	2.451	0.047	-5.11
Mean	**0.013**	**0.777**	**0.015**	**-5.32**
SD	**0.016**	**0.656**	**0.013**	**0.84**

Bryozoans (including Erect Bryozoans)

Data Sources and Distribution

Bryozoan (Phylum Bryozoa) catch data was collected over a span of 6 years from 2010 to 2015 and consisted of 83 presence and 1782 absence records (Table 11). Presences and absences records were distributed relatively evenly across the study extent (Figure 29). Both presences and absences were absent from the northern Labrador Shelf and beyond the continental slope. The highest mean catches (up to 3.56 kg) occurred southwest of Saint-Pierre and Miquelon.

Table 11. Number of presence and absence records of bryozoan catch recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2010 and 2015.

Year	Total number of presences	Total number of absences
2010	6	140
2011	7	209
2012	12	231
2013	24	542
2014	14	356
2015	20	304

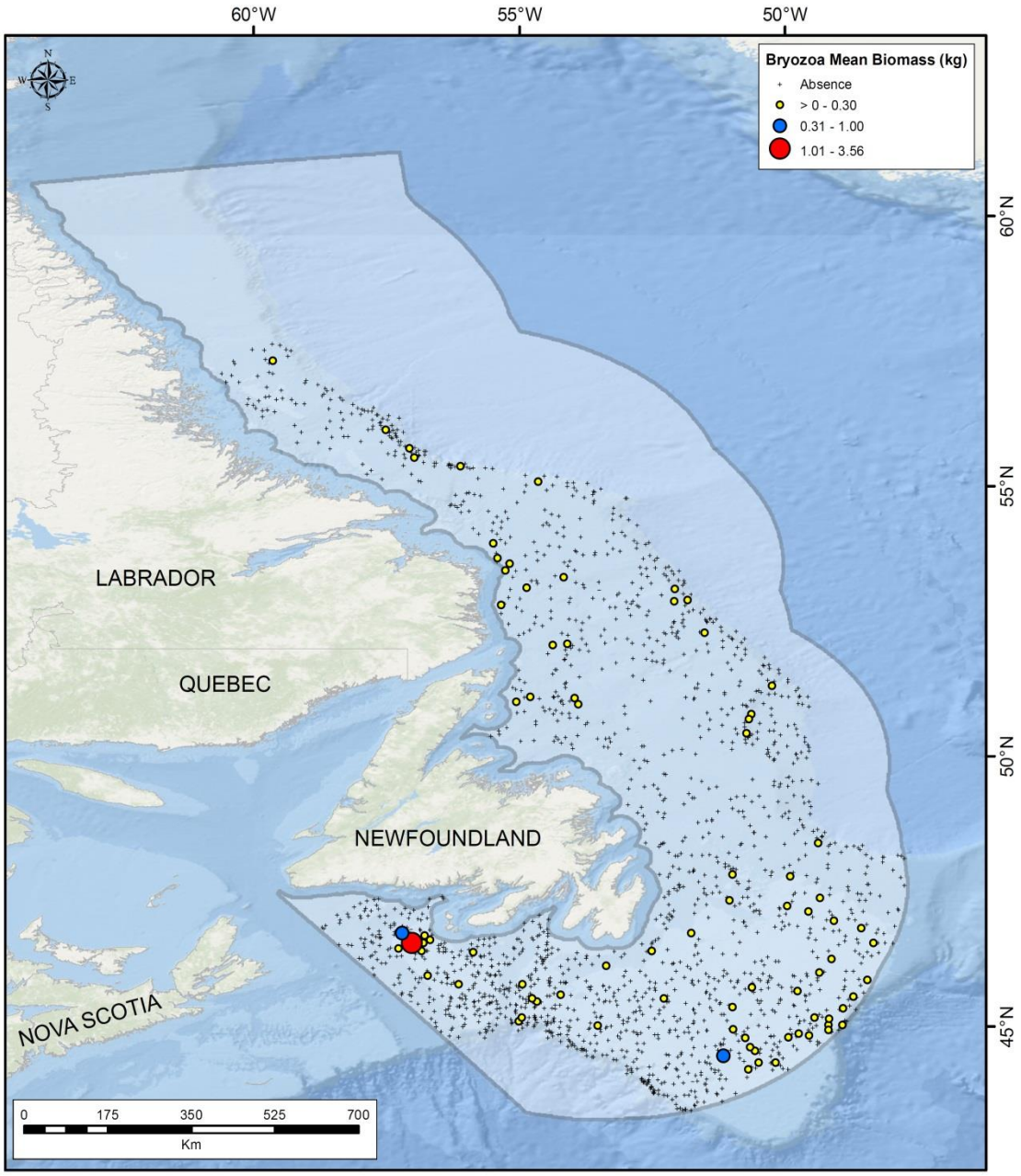

Figure 29. Mean biomass (kg) per grid cell of bryozoan catches recorded from DFO multispecies surveys conducted within the Newfoundland and Labrador Region between 2010 and 2015. Also shown are absence records from the same surveys.

Model 1 – Balanced species prevalence

Accuracy measures (mean AUC, sensitivity and specificity) for the random forest model on balanced species prevalence (83 presences and 83 absences; Model 1) are presented in Table 12. The average AUC was 0.664, indicating fair model performance. The highest AUC of 0.747 was associated with Model Run 3. The sensitivity and specificity measures of this model were 0.602

and 0.639, respectively. The confusion matrix of this model is also presented in Table 12. Class errors for both the presence and absence classes were moderate (0.398 and 0.361, respectively).

Table 12. Accuracy measures for all 10 model repetitions of 10-fold cross validation from the random forest model of bryozoan presence-absence data collected within the Newfoundland and Labrador 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 the presence probability of Bryozoa.

Model Run	AUC	Sensitivity	Specificity
1	0.606	0.615	0.615
2	0.673	0.627	0.578
3	**0.747**	**0.602**	**0.639**
4	0.693	0.675	0.578
5	0.658	0.651	0.627
6	0.662	0.615	0.566
7	0.700	0.615	0.639
8	0.594	0.590	0.506
9	0.623	0.651	0.590
10	0.683	0.675	0.602
Mean	**0.664**	**0.631**	**0.594**
SD	**0.047**	**0.030**	**0.040**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	53	30	83	0.361
Presence	33	50	83	0.398

The presence probability prediction surface of the bryozoans is presented in Figure 30. The highest predictions of presence probability occurred on the Northeast Newfoundland Shelf, Grand Bank, and in a small area southwest of Saint-Pierre and Miquelon. These areas of high presence probability corresponded well with the spatial distribution of presence records (see Figure 31). However, the model appears to have predicted large areas of presence probability beyond the location of presence observations in the areas of extrapolation. Figure 32 shows the actual presence and absence data observations (83 presences and 83 absences) used in the optimal Model 1. Areas of extrapolation are also shown in this figure. Deep water beyond the slope was considered extrapolated area. Smaller pockets of extrapolated area are distributed across the shelf and in coastal areas.

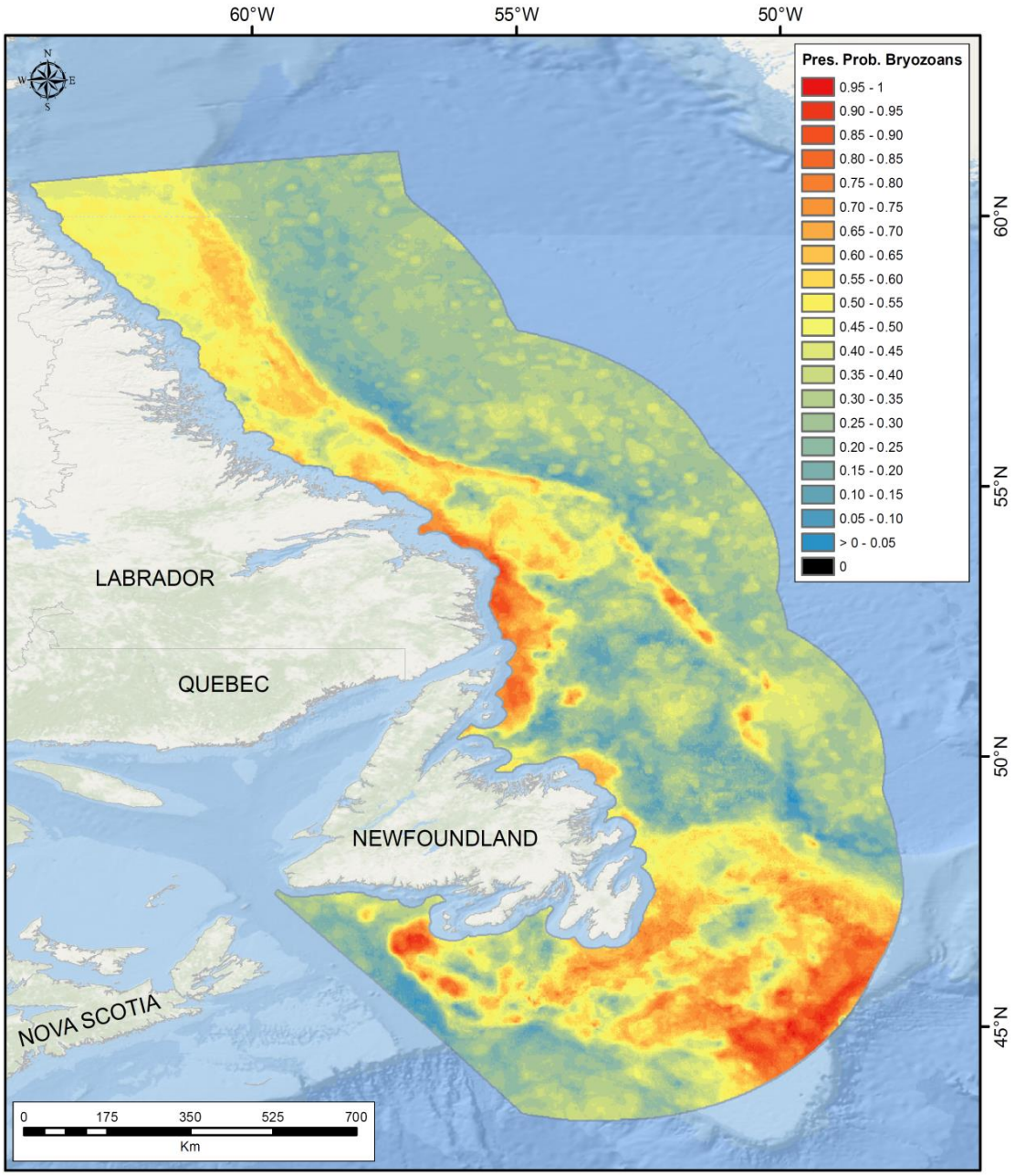

Figure 30. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of bryozoan presence and absence data collected from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2010 and 2015.

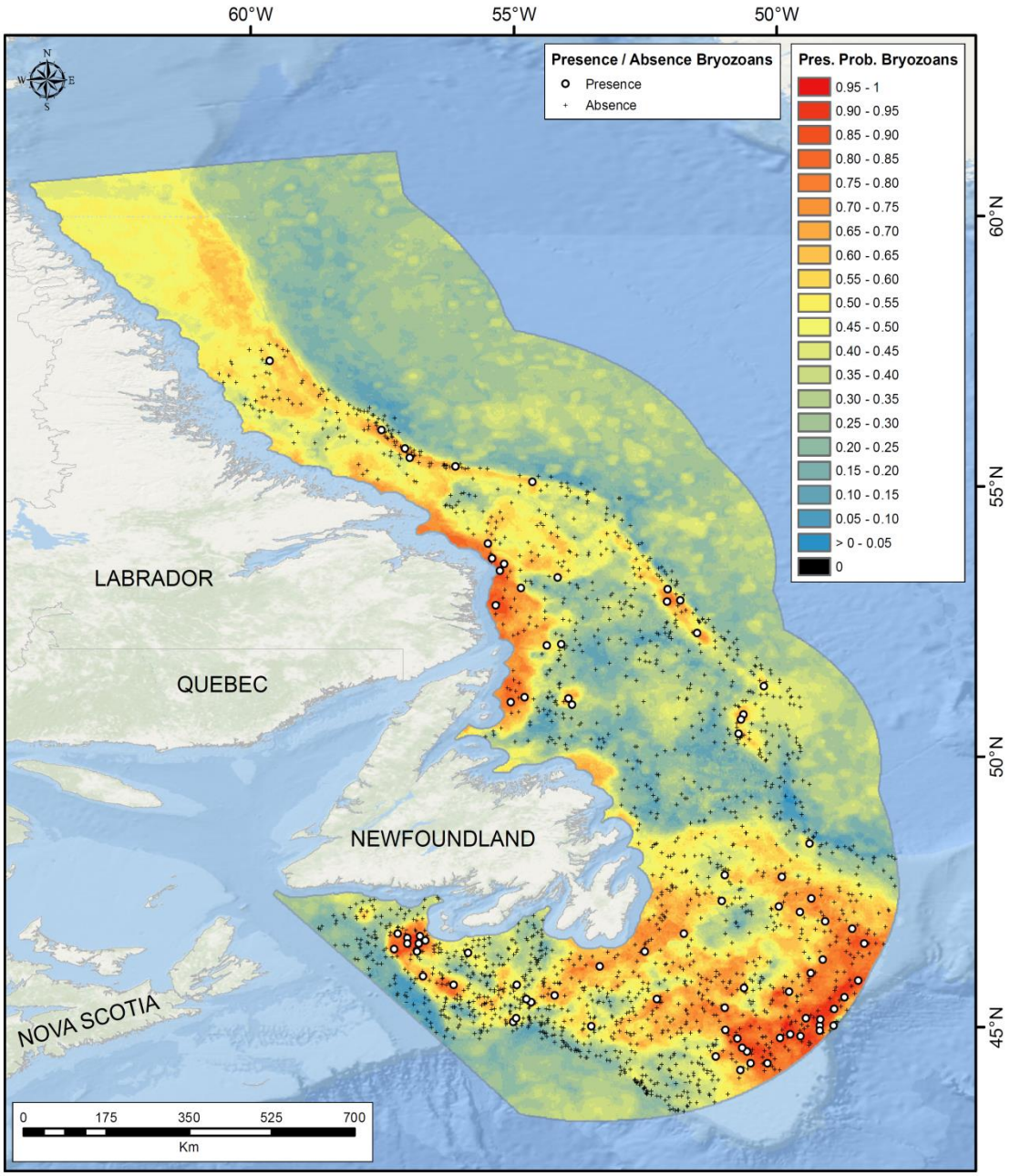

Figure 31. Presence and absence observations and predictions of presence probability (Pres. Prob.) from the optimal random forest model of bryozoan presence and absence data recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2010 and 2015.

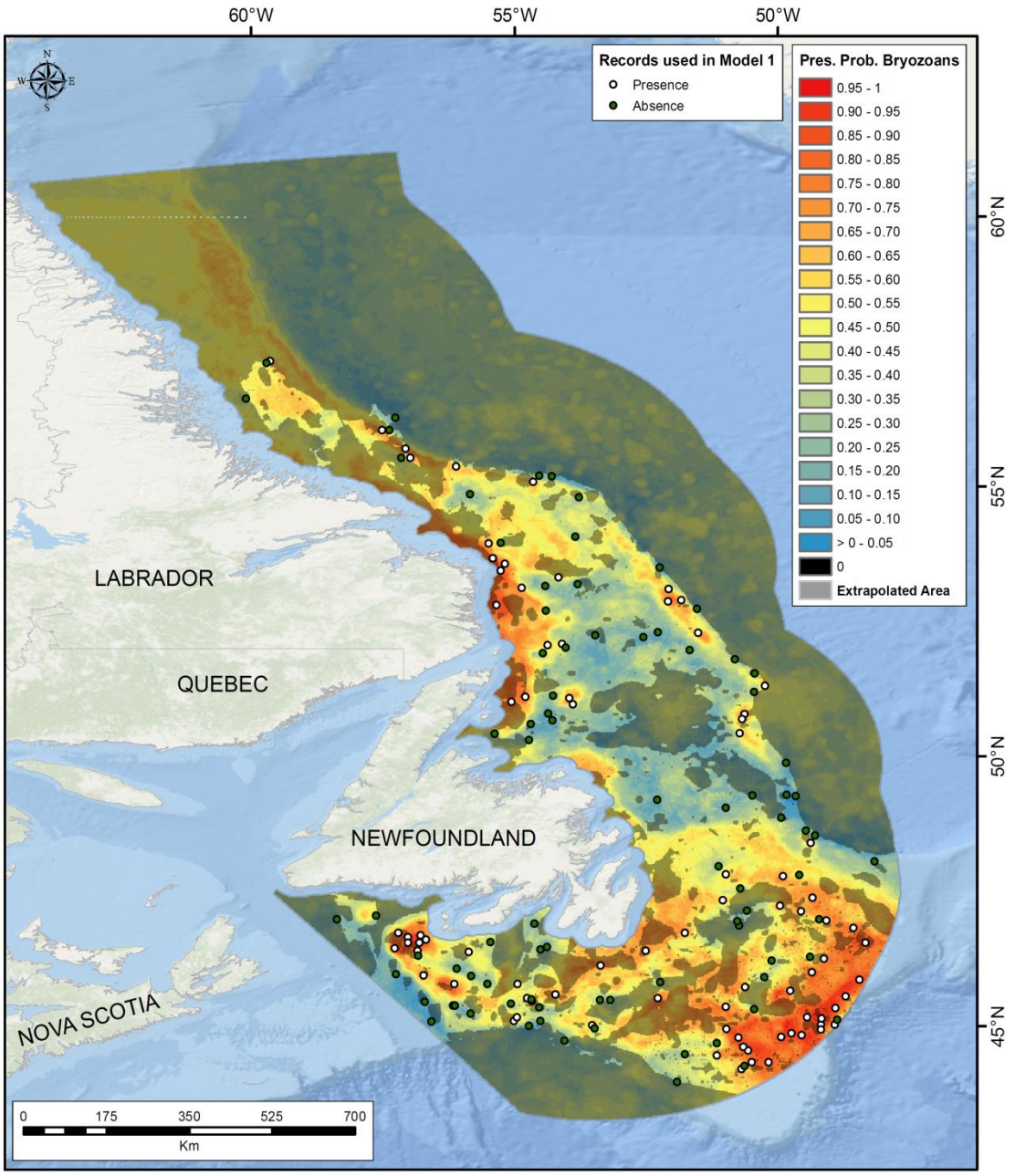

Figure 32. Map of the 166 data observations (83 presences and 83 absences) of bryozoans used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of bryozoans generated from Model 1 and areas of model extrapolation**.**

Of the 66 environmental predictor variables used in the model, Bottom Salinity Average Range was most important for the classification of the bryozoan presence-absence data (Figure 33). This variable displayed a right-skewed distribution prior to spatial interpolation (Guijarro et al., in prep). Examination of the Q-Q plot revealed a strong spatial pattern to those data points over- and under-predicted by a normal distribution, with over-predicted points located mainly in the deep waters beyond the Labrador Shelf, and under-predicted points located along the Newfoundland and Labrador slopes. This variable was followed by Spring Primary Production Average Range and Fall Primary Production Average Range. Partial dependence plots for the top 6 predictor variables are shown in Figure 34. The highest presence probability of bryozoans along the gradient in Bottom Salinity Average Range occurred between 0.6 and 0.8, while along the gradient in Spring Primary Production Average Range occurred in between ~ 500 and 600 mg C m^{-2} day^{-1}.

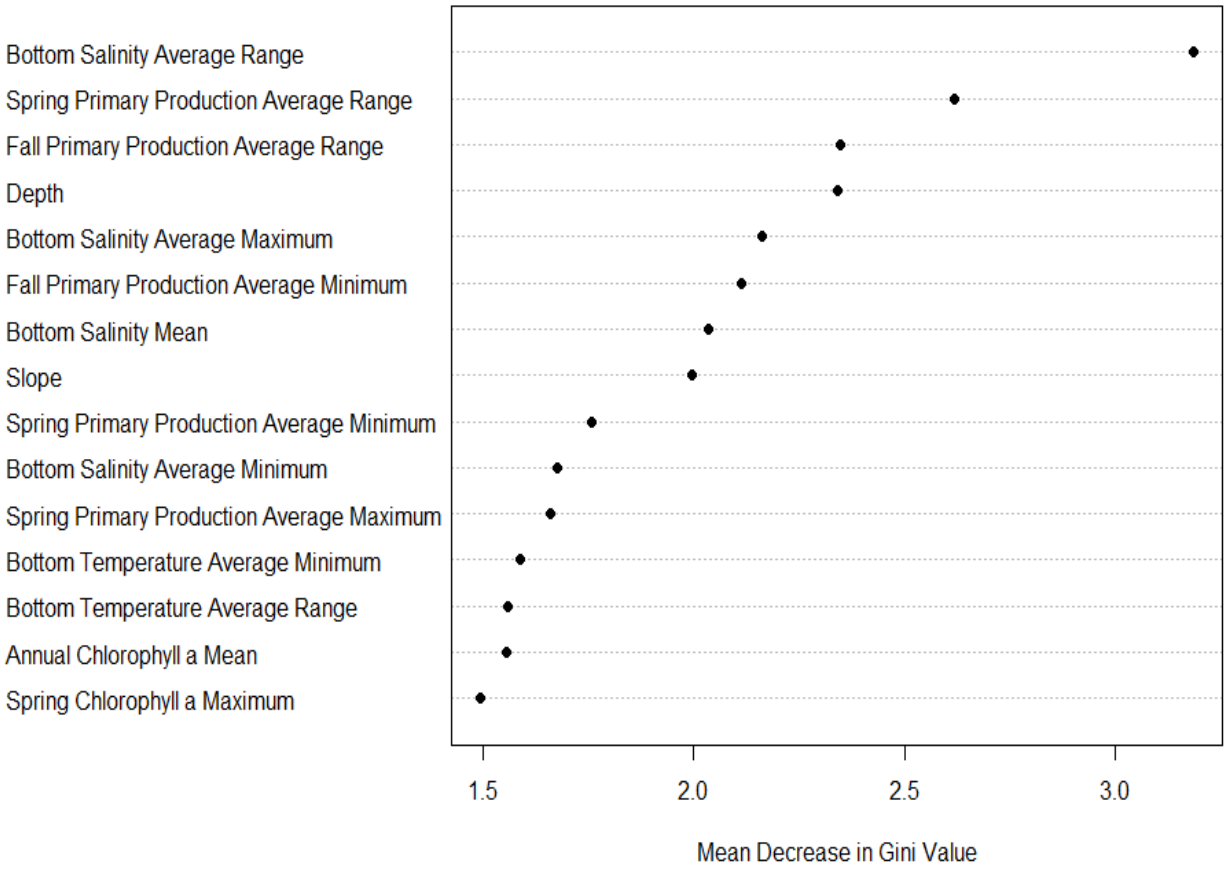

Figure 33. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini Value of the optimal random forest model predicting bryozoan presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini Value the more important the variable is for predicting the response data.

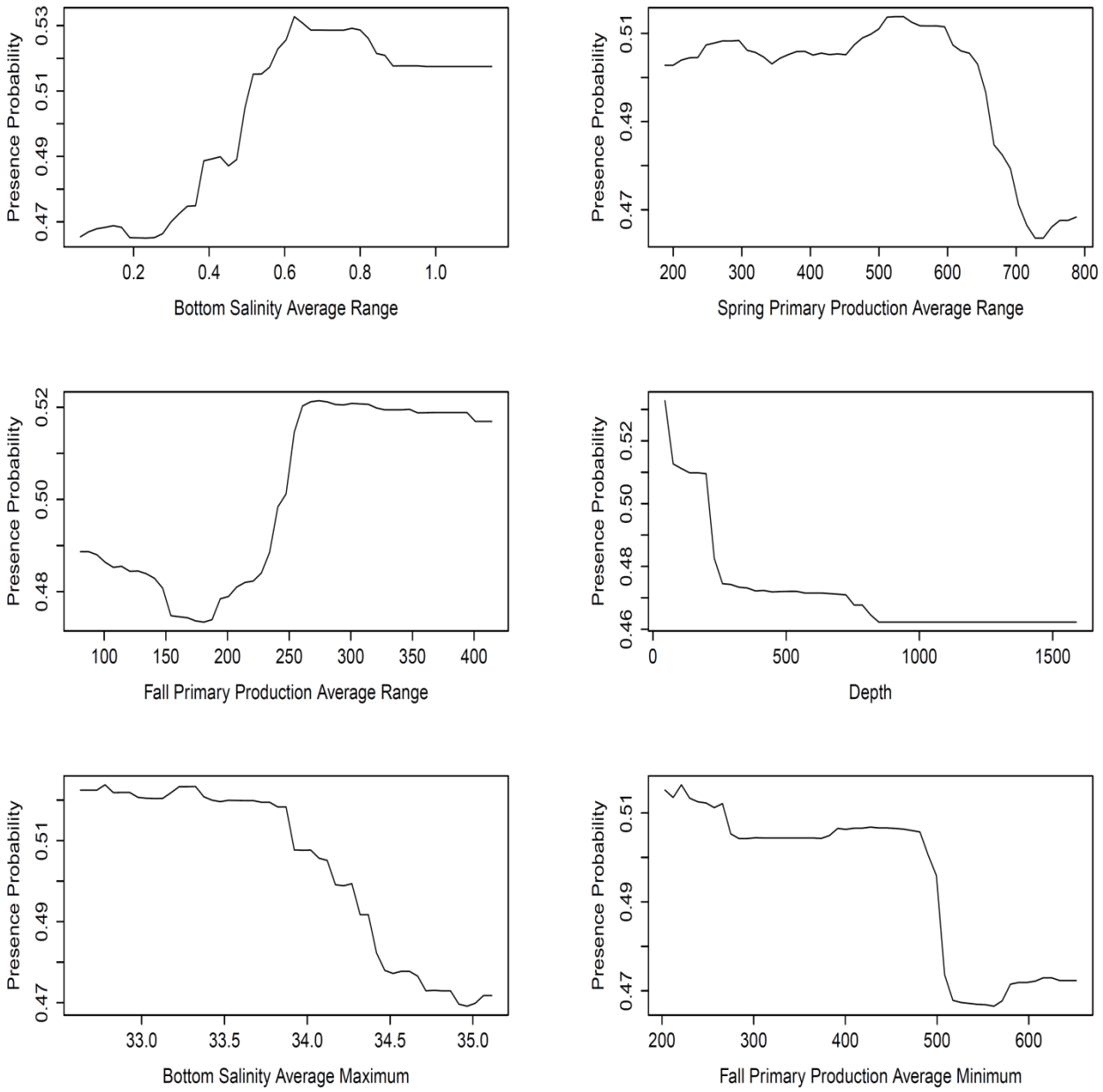

Figure 34. Partial dependence plots of the top 6 predictors from the optimal random forest model of bryozoan presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Predicted presence probability is shown on the *y*-axis of each graph.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 13 shows the accuracy measures for the random forest model using all bryozoan presence and absence data (1782 absences and 83 presences; Model 2) and a threshold equal to species prevalence (0.04). The average AUC calculated from this model was slightly lower than that of Model 1 (0.650 compared to 0.664 of Model 1). Sensitivity was lower than that of Model 1 while specificity was higher than that of Model 1. Class error of the presence and absence classes was comparable to Model 1.

The predicted bryozoan presence probability surface generated from Model 2 is shown in Figure 35. The areas of high predicted presence probability from Model 1 are greatly reduced in this model. The highest bryozoan presence probabilities still occurred on the Tail of the Grand Bank and a small pocket southwest of Saint-Pierre and Miquelon. However, the model does not appear to have greatly extrapolated high probabilities far beyond the location of presence observations (Figure 36), likely due to the inclusion of all absence records in the model. A large area in the northwestern portion of the study extent off Labrador was predicted to have a moderate to high presence probability of bryozoans, despite there being no presence records there. Figure 37 depicts the classification of bryozoan presence probability into presence and absence categories based on the prevalence threshold of 0.04. In this map, all presence probability values generated from Model 2 greater than 0.04 were classified as presence, while values less than 0.04 were classed as absence.

Table 13. Accuracy measures and confusion matrix from 10-fold cross validation from random forest modelling of presence and absence of bryozoans within the Newfoundland and Labrador Region. Observ. =Observations, Sensit.= Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.579		**Absence**	**Presence**				
2	0.520	**Absence**	1154	628	1782	0.352	0.566	0.648
3	0.718	**Presence**	36	47	83	0.434		
4	0.632							
5	0.681							
6	0.608							
7	0.643							
8	0.765							
9	0.632							
10	0.725							
Mean	**0.650**							
SD	**0.074**							

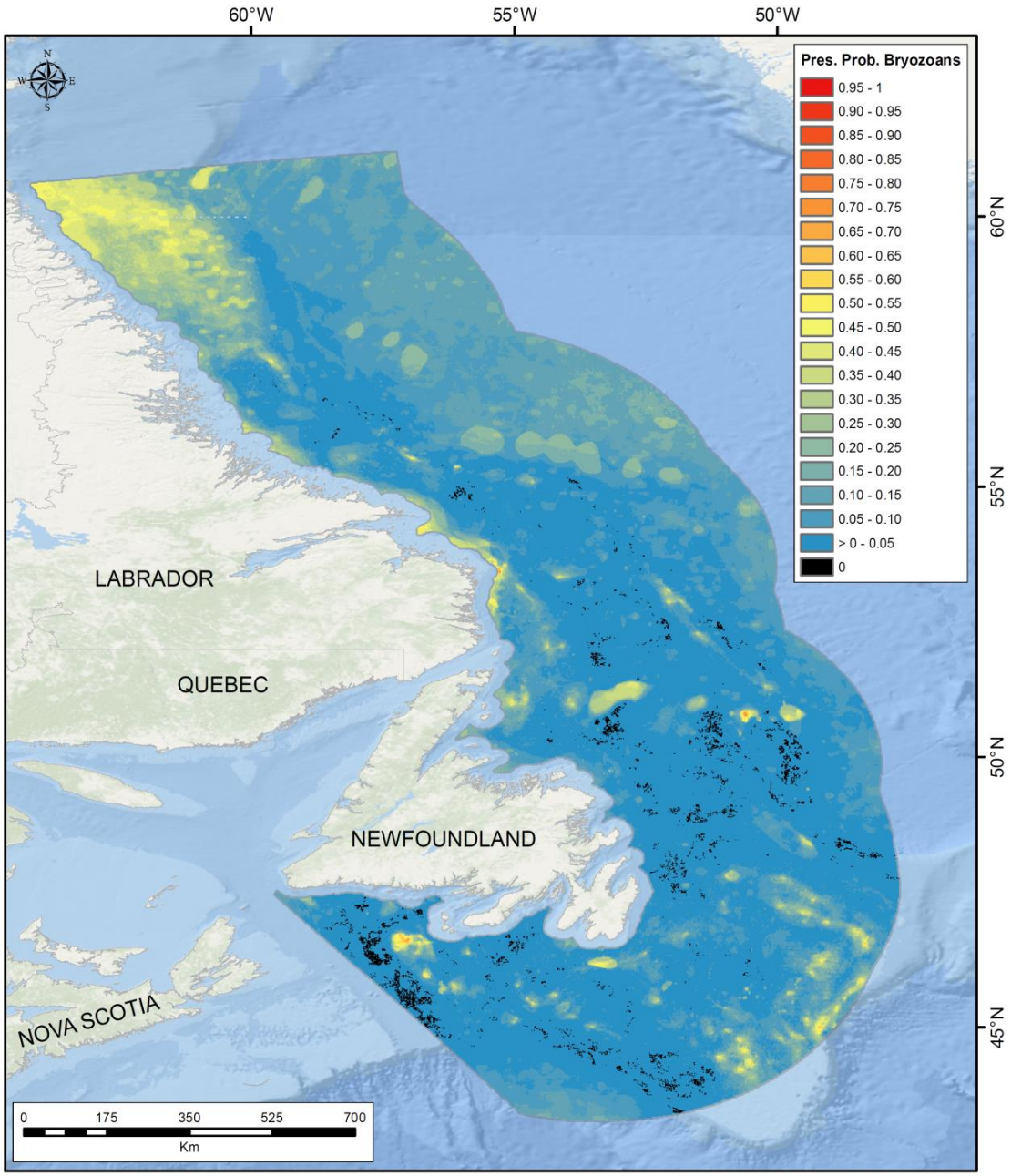

Figure 35. Predictions of presence probability (Pres. Prob.) from the unbalanced random forest model of bryozoan presence and absence data collected from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2010 and 2015.

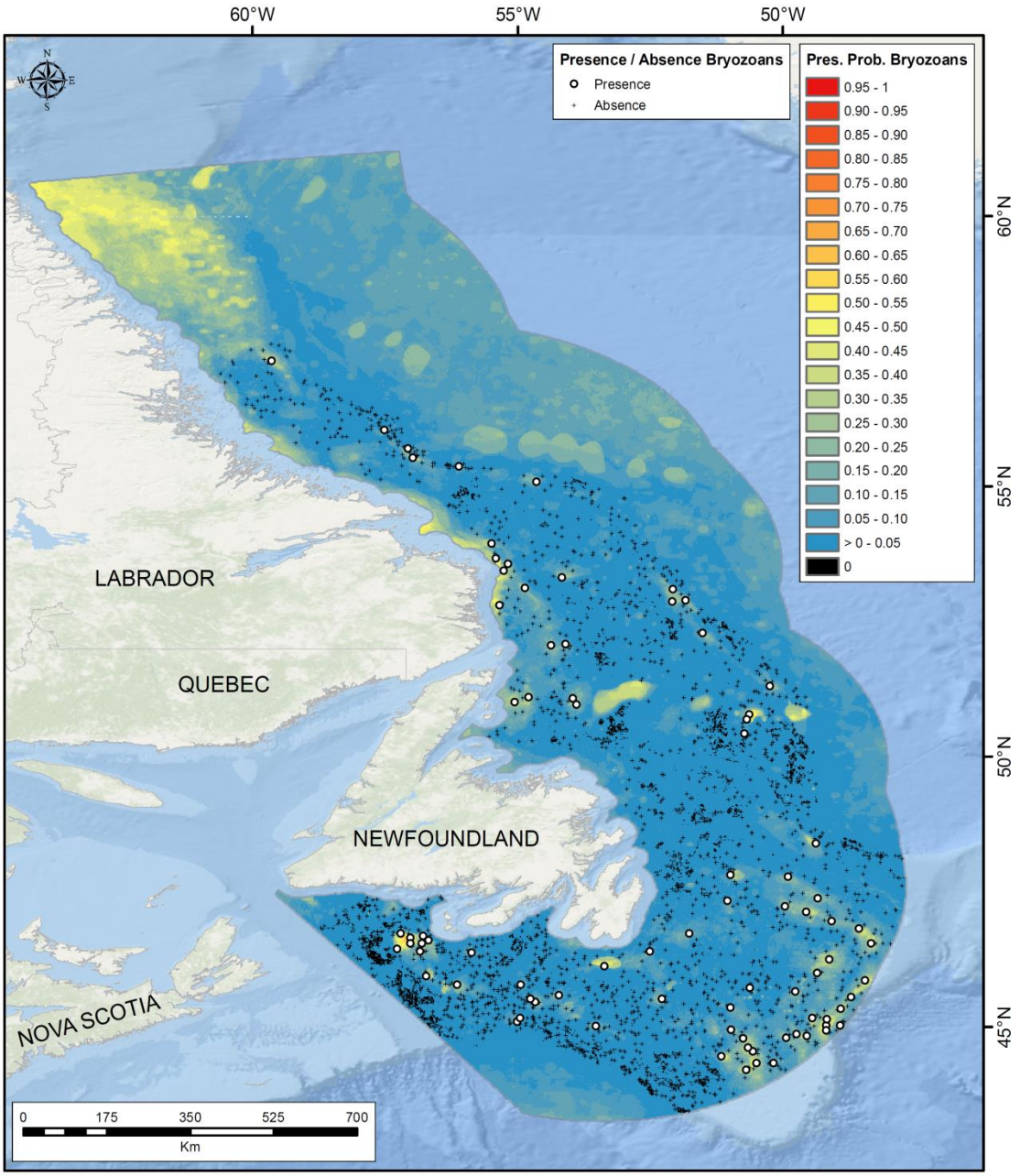

Figure 36. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the unbalanced random forest model of bryozoan presence and absence data recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2010 and 2015.

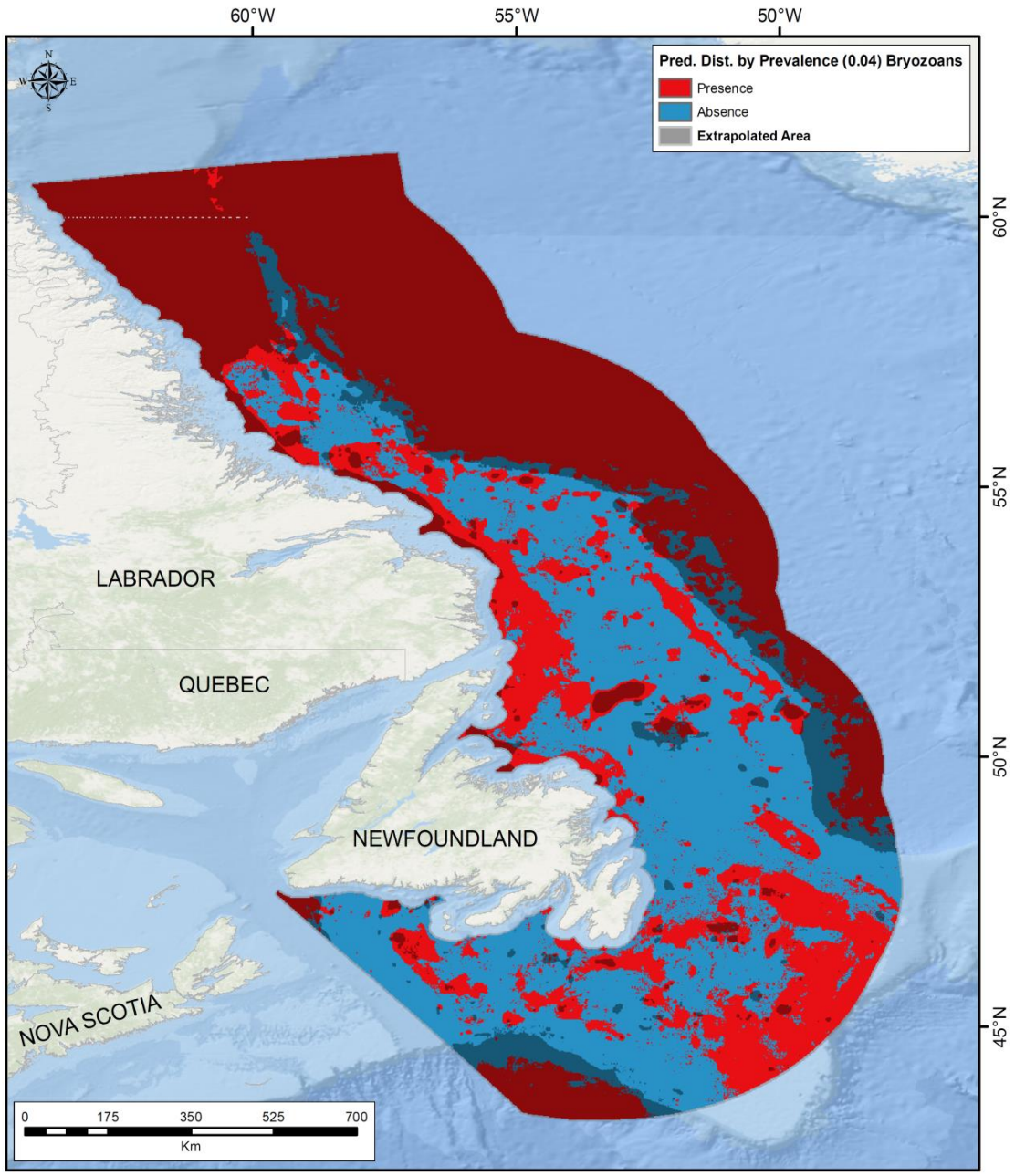

Figure 37. Predicted distribution (Pred. Dist.) of bryozoans in the Newfoundland and Labrador Region based on the prevalence threshold of 0.04 of bryozoan presence and absence data used in Model 2. Also shown are the areas of model extrapolation. Grey polygon may appear dark red or blue when overlain on the presence-absence surface.

The importance of the environmental predictor variables for predicting the presence probability of bryozoans is presented in Figure 38. In this model, Slope (non-interpolated variable) was the most important variable for the classification of the bryozoan presence-absence data. This variable was followed closely by Depth and Fall Primary Production Average Minimum. Partial dependence of the bryozoan presence and absence data on the top 6 predictor variables is shown in Figure 39. The highest presence probability of bryozoans occurred in relatively flat (slope < 0.5˚) areas in shallow water ≤100 m depth.

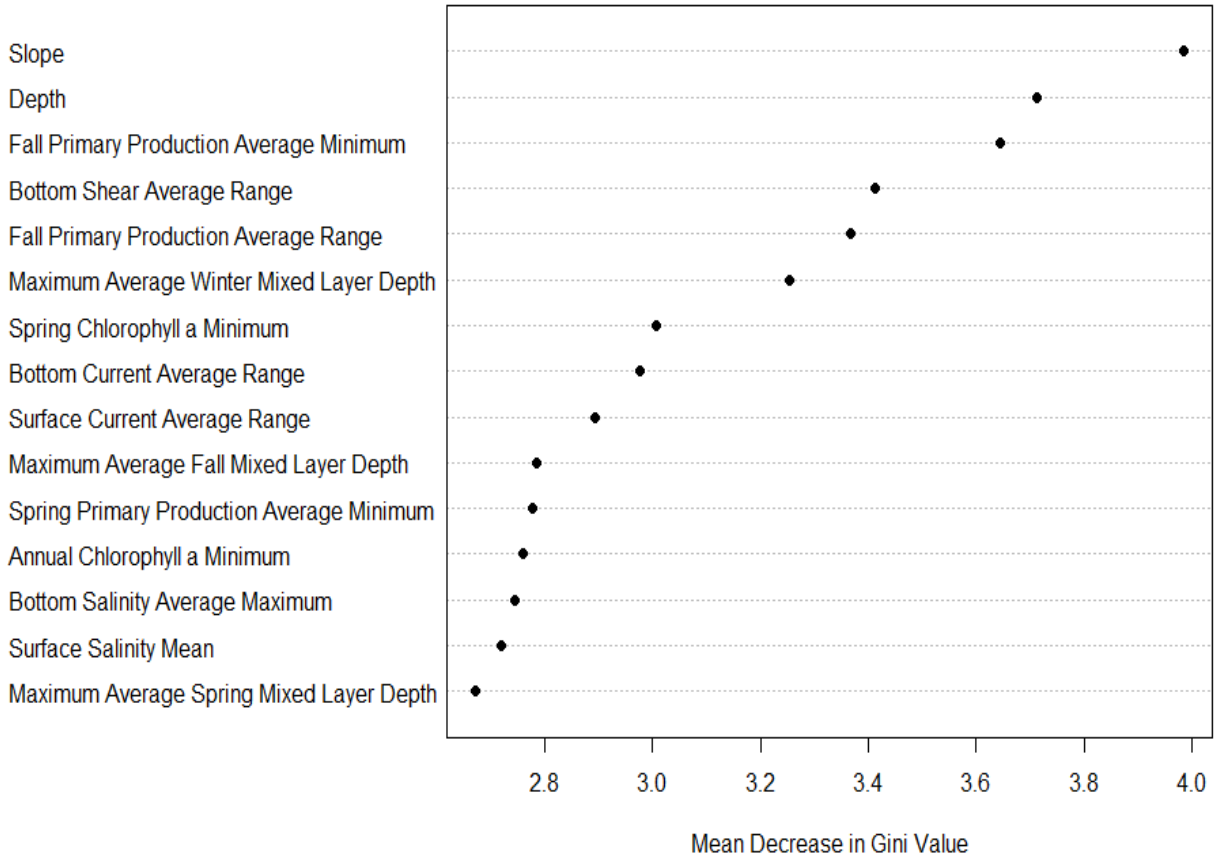

Figure 38. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini Value of the unbalanced random forest model of bryozoan presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini Value the more important the variable is for predicting the response data.

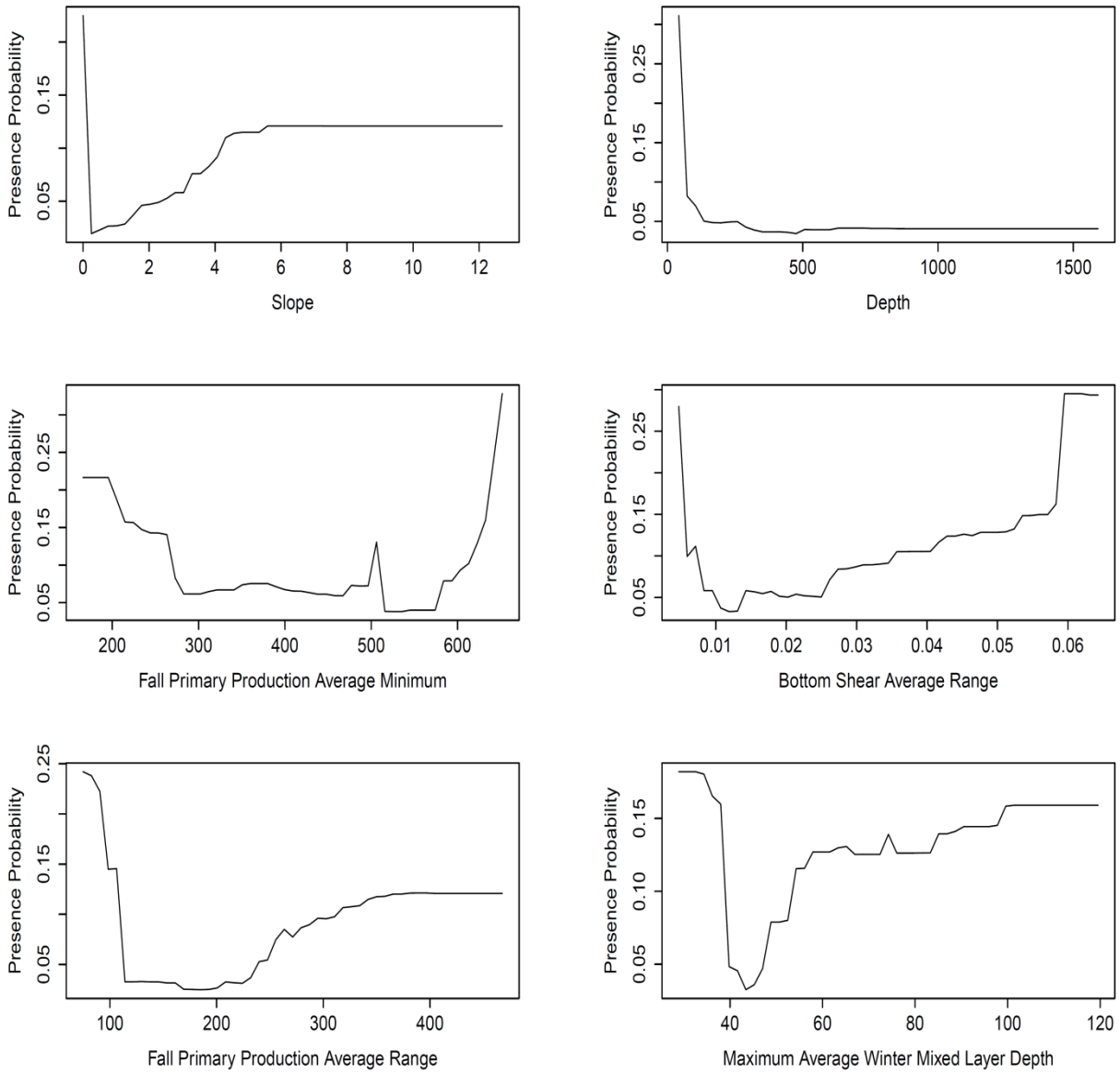

Figure 39. Partial dependence plots of the top 6 predictors from the unbalance random forest model of bryozoan presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model Selection

The random forest model using all available bryozoan records and an unbalanced species prevalence and threshold equal to 0.04 (Model 2) was chosen as the best predictor of bryozoan distribution in the Newfoundland and Labrador Region. Model 1 (balanced species prevalence) was considered a poor predictor of presence probability of bryozoans due to its exaggeration of high presence probability beyond the location of presence data, particularly in the slope off the northeast Newfoundland Shelf. This phenomenon is likely due to random down-sampling of the absence data.

Prediction of Biomass using Random Forest

The accuracy measures of the regression random forest model on mean bryozoan biomass per grid cell are presented in Table 14. The highest R^2 value was 0.092, while the average was 0.017 ± 0.031 SD. The highest Normalized Root-Mean-Square Error (RMSE) was 0.016 ± 0.022 SD. The percent variance explained for each fold was negative, indicating that the model had no predictive power. Therefore, the predictive surfaces of this model are not displayed in this report.

Table 14. Accuracy measures from 10-fold cross validation from random forest modelling of average bryozoan biomass (kg) per grid cell recorded from DFO multispecies surveys conducted in the Newfoundland and Labrador Region between 2010 and 2015. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error (RMSE/range of biomass values for response).

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	5.611×10^{-06}	0.015	0.004	-11.49
2	0.014	0.090	0.025	-15.27
3	7.182×10^{-05}	0.024	0.007	-10.52
4	0.001	0.261	0.074	-2.56
5	1.821×10^{-04}	0.098	0.028	-14.60
6	8.175×10^{-05}	0.013	0.004	-12.10
7	0.009	0.016	0.005	-10.93
8	0.092	0.027	0.008	-12.89
9	0.001	0.007	0.002	-11.85
10	0.050	0.028	0.008	-8.30
Mean	**0.017**	**0.058**	**0.016**	**-11.05**
SD	**0.031**	**0.078**	**0.022**	**3.58**

DISCUSSION

The species distribution models for the crinoids, ascidians and bryozoans in general did not perform well when predicting biomass distribution. The regression random forest model at best only explained a very small portion of the variation for the crinoids, and was unreliable for the other groups. However, the models predicting species presence-absence performed consistently better. In all cases Model 2, using an unbalanced design and prevalence equal to model occurrence was preferred over balanced designs (Model 1) as the prediction surfaces were better matched to the location of known presence data. However, the use of prevalence showed presence in the deep water of the extrapolated areas that was not shown in the Model 1 outputs. We therefore conclude that Model 2 gives the best prediction surface but that the extrapolated area should be discounted until further validation can be provided.

Using random forest, Murillo et al. (2016b) modelled the distribution of stalked tunicates (*Boltenia ovifera*) in the Gulf of St. Lawrence and found that Bottom Temperature Mean and Depth were the top two predictors, followed by other physical variables. In the Newfoundland and Labrador Region, Depth was the most important variable for ascidians (which include *Boltenia ovifera*) followed distantly by a suite of variables that are proxies for food (Chlorophyll *a* and Primary Production). Murillo et al. (2016b) found such variables to be important predictors of biomass but not occurrence. The differences between these two studies could be attributed to differences in species composition between the two regions, differences in the taxonomic resolution of the taxa modelled, or to differences in the regional physical settings, with the Gulf representing a semi-enclosed sea, while the Newfoundland and Labrador Region as assessed herein is a continental shelf system, open to influences from the wider North Atlantic. In both studies presence probability was higher in shallower water but food may be a limiting factor in distribution on the continental shelf.

The SDMs of crinoid occurrence in the Newfoundland and Labrador Region suggest localized areas of suitable habitat largely on the continental shelf north of Cartwright, Labrador and extending through extrapolation to the northern portion of the boundary (Figure 1). High occurrence prediction was also shown on the continental slopes, at depths greater than 1000 m. Surface Temperature Mean and Depth were the top two predictors for this species group with occurrence more probable in colder surface water. The biomass regression model performed well for this taxon, although only a small proportion of the variance was explained by the model. The greatest biomass was also predicted to occur on the continental shelf off Labrador, north of Nain.

The SDMs of bryozoans show their predicted occurrence on the continental shelf, and on the Tail of the Grand Bank. Although different species can be included in this group, likely the highest biomass found corresponded to erect bryozoans and these results could extend the analyses of Murillo et al. (2016a) into Canadian waters. Slope, Depth and Fall Primary Production Average Minimum were the top predictors, combining physical and biological predictor variables as for the ascidians.

The groups modelled in this report include a highly diverse set of species as can be seen by the high level of taxonomic resolution in Table 2. For instance, crinoids and bryozoans were only identified to the taxonomic level of class and phylum, respectively. Ascidians include some

identifications to species or genus level (i.e., *Boltenia ovifera*, *Dendrodoa* sp.), but most of the records (96%) were at the class level (Ascidiacea) where not all the species are considered VME indicators (Murillo et al., 2011) and which could include solitary as well as colonial ascidians that can have different environmental requirements, making it difficult to find a common response to the predictors. In order to improve the performance of the models and to be able to model only the VME indicator taxa it would be necessary to improve the taxonomic resolution of the identifications from the research surveys. However, pending a better taxonomic resolution for some of these groups, the results presented in this report can be used to indicate areas where these taxa, which include VME indicators, are absent, and areas which have a greater probability of presence and higher biomass. Validation of the models should be considered with independent data collected in future years, although at present the Fisheries Observer Program does not collect data on these species so validation would have to come from future DFO NL multispecies surveys. At present, we suggest that these distribution models be used as supporting information for management decisions and as the basis for hypothesis testing for future ground-truthing exercises.

REFERENCES

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

Breiman, L. 2001. Random forests. Machine Learning, 45: 5–32.

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

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

Guijarro, J., Beazley, L., Lirette, C., Kenchington, E., Wareham, V., Gilkinson, K., Koen-Alonso, M. and F.J. Murillo. 2016. Species Distribution Modelling of Corals and Sponges from Research Vessel Survey Data in the Newfoundland and Labrador Region for use in the Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3171: vi+126p.

Guijarro, J., Beazley, L., Lirette, C., Wang, Z., and Kenchington, E. (in prep). Characteristics of Environmental Data Layers for Use in Species Distribution Modelling in the Newfoundland and Labrador Region. Can. Tech. Rep. Fish. Aquat. Sci.

Kenchington, E., Murillo, F.J., Lirette, C., Sacau, M., Koen-Alonso, M., Kenny, A., Ollerhead, N., Wareham, V., and Beazley, L. 2014a. Kernel density surface modelling as a means to identify significant concentrations of vulnerable marine ecosystem indicators. PLoS ONE 10(1): e0117752. doi:10.1371/journal.pone.0117752.

Kenchington, E. 2014b. A General Overview of Benthic Ecological or Biological Significant Areas (EBSAs) in Maritimes Region. Can. Tech. Rep. Fish. Aquat. Sci. 3072: iv + 45p.

Murillo, F.J., Serrano, A., Kenchington E., and Mora, J. 2016a. Epibenthic assemblages of the Tail of the Grand Bank and Flemish Cap (northwest Atlantic) in relation to environmental parameters and trawling intensity. Deep Sea Res. I 109: 99-122.

Murillo, F.J., Kenchington, E., Beazley, L., Lirette, C., Knudby, A., Guijarro, J., Benoît, H., Bourdage, H. and Sainte-Marie, B. 2016b. 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 Sensitive Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3170: vi + 132 p.

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

NAFO. 2013. Report of the 6th meeting of the NAFO Scientific Council Working Group on Ecosystem Science and Assessment (WGESA). NAFO SCS, Doc. 13/24, Serial No. N6277. 208 pp.

NAFO., 2014. Part E: Report of the Scientific Council Meeting, 31 May – 12 June 2014. NAFO Scientific Council Report, 238 p. http://www.nafo.int/publications/frames/sci-reports.html.