Species Distribution Modelling of Corals and Sponges in the Eastern Arctic for Use in the Identification of Significant Benthic Areas

L. Beazley, F.J. Murillo, E. Kenchington, J. Guijarro, C. Lirette, T. Siferd,
M. Treble, E. Baker, M. Bouchard Marmen, G. Tompkins MacDonald

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 3175

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 3175

2016

Species Distribution Modelling of Corals and Sponges in the Eastern Arctic for Use in the Identification of Significant Benthic Areas

by

L. Beazley[1], F.J. Murillo[1], E. Kenchington[1], J. Guijarro[1], C. Lirette[1], T. Siferd[2],
M. Treble[2], E. Baker[1], M. Bouchard Marmen[1], G. Tompkins MacDonald[1]

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]Arctic Stock Assessment and Conservation Research
Central & Arctic Region
501 University Crescent, Winnipeg
Manitoba R3T 2N6

Cat. No. Fs97-6/3175E-PDF ISBN 978-0-660-05892-4 ISSN 1488-5379

Correct citation for this publication:

Beazley, L., Murillo, F.J., Kenchington, E., Guijarro, J., Lirette, C., Siferd, T., Treble, M., Baker, E., Bouchard Marmen, M., Tompkins MacDonald, G. 2016. Species Distribution Modelling of Corals and Sponges in the Eastern Arctic for Use in the Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3175: vii + 210p.

TABLE OF CONTENTS

ABSTRACT vi
RÉSUMÉ vii
INTRODUCTION 1
METHODOLOGY 2
  Study Area 2
  Environmental Data 5
  Response Data 5
  Random Forest Modelling 10
  Model Evaluation 11
    Presence-Absence Response Data – Classification Model 11
    Biomass Response Data – Regression Model 12
  Model Extrapolation 12
  Ecological Interpretation 13
  Alternative Prediction Models 13
RESULTS 14
  Hudson Strait – Ungava Bay Region 14
    Sponges (Porifera) 14
      Data Sources and Distribution 14
      Model 1 – Balanced Species Prevalence 16
      Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence 24
      Model Selection 29
      Prediction of Sponge Biomass Using Random Forest 30
        Cosmos Trawl Gear 30
  Eastern Arctic Region 35
    Sponges (Porifera) 35
      Data Sources and Distribution 35
      Model 1 – Balanced Species Prevalence 38
      Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence 43
      Model Selection 49
      Validation of Selected Model Using Independent Data 50
      Prediction of Sponge Biomass Using Random Forest 53
        Alfredo Trawl Gear 53
        Campelen Trawl Gear 57
        Cosmos Trawl Gear 62
    Sea Pens (Pennatulacea) 67
      Data Sources and Distribution 67
      Model 1 – Balanced Species Prevalence 71
      Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence 77
      Model Selection 83
      Validation of Selected Model Using Independent Data 84
      Prediction of Sea Pen Biomass Using Random Forest 87
        Alfredo Trawl Gear 87
        Campelen Trawl Gear 91
        Cosmos Trawl Gear 96

Large Gorgonian Corals 101
Data Sources and Distribution 101
Model 1 – Balanced Species Prevalence 105
Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence 111
Model Selection 117
Validation of Selected Model Using Independent Data 118
Prediction of Large Gorgonian Coral Biomass Using Random Forest 121
Alfredo Trawl Gear 121
Campelen Trawl Gear 125
Small Gorgonian Corals 130
Data Sources and Distribution 130
Model 1 – Balanced Species Prevalence 134
Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence 140
Model Selection 147
Validation of Selected Model Using Independent Data 148
Prediction of Small Gorgonian Biomass Using Random Forest 151
Alfredo Trawl Gear 151
Campelen Trawl Gear 155
DISCUSSION 161
ACKNOWLEDGMENTS 168
REFERENCES 168
APPENDIX 1 171
Alternative Prediction Models- Generalized Additive Models for Predicting Coral and Sponge Biomass in the Eastern Arctic 171
Hudson Strait – Ungava Bay Region 172
Sponges (Porifera) 172
Cosmos Trawl Gear 172
Eastern Arctic Region 177
Sponges (Porifera) 177
Alfredo Trawl Gear 177
Campelen Trawl Gear 180
Cosmos Trawl Gear 183
Sea Pens (Pennatulacea) 186
Alfredo Trawl Gear 186
Campelen Trawl Gear 190
Cosmos Trawl Gear 194
Large Gorgonian Corals 197
Alfredo Trawl Gear 197
Campelen Trawl Gear 200
Small Gorgonian Corals 205
Alfredo Trawl Gear 205
Campelen Trawl Gear 208

ABSTRACT

Beazley, L., Murillo, F.J., Kenchington, E., Guijarro, J., Lirette, C., Siferd, T., Treble, M., Baker, E., Bouchard Marmen, M., Tompkins MacDonald, G. 2016. Species Distribution Modelling of Corals and Sponges in the Eastern Arctic for Use in the Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3175: vii + 210p.

Species distribution modelling using a random forest (RF) machine learning approach was used to predict the probability of occurrence and biomass of sponges, sea pens, and large and small gorgonian corals in the Hudson Strait portion of Fisheries and Oceans, Canada's (DFO) Hudson Bay Complex Biogeographic Zone (sponges only), and in the eastern extent (Davis Strait and Southern Baffin Bay) of the Eastern Arctic Biogeographic Zone. A suite of 54 environmental predictor variables from different data sources were used. Models utilized catch records from the DFO multispecies trawl surveys and DFO/industry northern shrimp surveys collected between 2006 and 2014. For each taxonomic group in each region, both presence-absence random forest models using data collected across gear types (Alfredo, Campelen, and Cosmos trawls), and biomass random forest models using data collected within gear types were run. Most presence-absence models had good predictive capacity with cross-validated Area Under the Receiver Operating Characteristic Curve (AUC) values ranging from 0.643 to 0.894. The lower AUC was produced from the Hudson Strait sponge model, which also had poor sensitivity and specificity relative to the models performed in the Eastern Arctic Biogeographic Zone. The random forest biomass models performed inconsistently within taxa by gear type, with the models for sponges using data from Alfredo and Campelen trawl surveys perfoming best (R^2 = 0.327 and 0.480 respectively). Generalized additive models (GAMs) were developed to predict the biomass distribution of each taxonomic group and serve as a comparison to the RF models. Aside from providing continuous prediction maps of significant benthic taxa for these regions, our results will be useful in ecosystem management decision-making processes. In particular, good SDM models could be used to refine the outer boundaries of significant concentrations of these organisms identified by kernel density analyses and identify new suitable habitat not sampled by the trawl surveys in areas of extrapolation.

RÉSUMÉ

Beazley, L., Murillo, F.J., Kenchington, E., Guijarro, J., Lirette, C., Siferd, T., Treble, M., Baker, E., Bouchard Marmen, M., Tompkins MacDonald, G. 2016. Modélisation de la répartition des espèces de coraux et d'éponges dans l'est de l'Arctique aux fins d'utilisation dans la détermination des zones benthiques importantes. Can. Tech. Rep. Fish. Aquat. Sci. 3175: vii + 210 p.

La modélisation de la répartition des espèces au moyen d'une approche d'apprentissage machine de forêts aléatoires (RF) a été utilisée pour prédire la probabilité de présence et la biomasse des éponges, des pennatules et des grandes et petites gorgones dans la partie du détroit d'Hudson gérée par Pêches et Océans Canada (MPO), la zone biogéographique du complexe de la baie d'Hudson (éponges seulement) et la partie est (détroit de Davis et sud de la baie de Baffin) de la zone biogéographique de l'est de l'Arctique. Un ensemble de 54 variables environnementales explicatives provenant de différentes sources de données a été utilisé. Les modèles utilisent les registres de pêches tirés des relevés plurispécifiques au chalut du MPO ainsi que les relevés sur la crevette nordique menés par le MPO et l'industrie entre 2006 et 2014. Pour chaque groupe taxonomique, nous avons généré des modèles de forêts aléatoires sur la présence et l'absence des espèces à l'aide des données recueillies en fonction des types d'engins (chaluts Alfredo, Cosmos et Campelen) et des modèles de forêts aléatoires de la biomasse à l'aide des données recueillies en fonction des types d'engins. La plupart des modèles sur la présence et l'absence avaient une bonne efficacité de prévision selon des valeurs contre-validées de l'aire sous la courbe de la fonction d'efficacité du récepteur variant de 0,643 à 0,894. Les valeurs inférieures de l'aire sous la courbe ont été générées à partir du modèle sur les éponges dans le détroit d'Hudson, dont la sensibilité et la spécificité étaient mauvaises par rapport aux modèles de la zone biogéographique de l'est de l'Arctique. Le rendement des modèles de biomasse de forêts aléatoires variait en fonction des taxons par type d'engin; les modèles pour les éponges utilisant les données des relevés au chalut Alfredo et Campelen généraient les meilleurs résultats (R^2 = 0,327 et 0,480 respectivement). Des modèles additifs généralisés ont été élaborés pour prédire la répartition de la biomasse de chaque groupe taxonomique et servent de points de comparaison aux modèles RF. En plus de fournir des cartes de prévision continue des taxons benthiques importants pour ces régions, nos résultats seront utiles dans le cadre des processus décisionnels sur la gestion de l'écosystème. Plus précisément, de bons modèles de répartition de l'espèce pourraient être utilisés pour préciser les limites extérieures des concentrations importantes des organismes désignés par les analyses de noyaux de densité et pour trouver un nouvel habitat convenable qui n'a pas été échantillonné par les relevés au chalut dans les zones d'extrapolation.

INTRODUCTION

The Davis Strait area of the Eastern Arctic joins two oceanic basins, Baffin Bay and the Labrador Sea, and separates western Greenland and Baffin Island, the latter constituting the largest island in the Canadian Arctic Archipelago. It connects to the Arctic Ocean in the north via Baffin Bay and to the Atlantic Ocean in the south via the Labrador Sea. It is considered the world's largest strait and is renowned for exceptionally strong tides, which range from 9 to 18 m, and a complex hydrography (Hamilton and Wu, 2013). The larger region includes the North Water Polynya, one of the Arctic's largest open-water areas and, historically, one of the most biologically productive waters in the Arctic. The shelves extending from both Canada and Greenland include several large shoals or banks typically ranging between 20 and 100 m in depth and traversed by deep troughs. At its narrowest point, a ridge or sill up to approximately 600 m depth extends between Greenland (at Holsteinborg, Sisimiut) and Baffin Island (at Cape Dyer). The slopes at the Labrador Sea flank of the ridge drop to 2500 m or more.

Baffin Bay and Davis Strait have the only large-scale commercial fisheries in Canada's Arctic. These are trawl fisheries for turbot (Greenland halibut) and shrimp both of which are managed under quotas set by the Northwest Atlantic Fisheries Organization (NAFO). These fisheries have undergone considerable expansion in recent decades but nothing is known about their impact on vulnerable marine ecosystems (VMEs) such as corals and sponges which are known to occur in close proximity to the fisheries. Recently, dense bamboo coral forests (*Keratoisis* sp.) have been observed *in situ* in muddy environments in southeast Baffin Bay deeper than 900 m (Neves et al., 2015). These dense aggregations could form habitat for other organisms in the deep and muddy Arctic environment. In the Davis Strait region, corals and sponges constitute the greatest proportion of benthic biomass in some areas, with up to two metric tonnes of sponge being removed during a single research vessel tow and coral catches so heavy as to break the nets (Kenchington et al., 2010; Neves et al., 2015). Yet previous benthic studies have been restricted to shallow-water, soft-bottom macrofaunal communities (e.g., Stewart et al., 1985; Turner,2002), which are more dynamic in nature than those dominated by the long-lived, slow-growing corals and sponges of the deep-water slopes. Consequently the ecosystem function of these taxa both in Davis Strait and more generally, and the physical and biological conditions which sustain them are unknown. However, food availability is likely to be a prime determinant of biomass and distribution.

The waters off west Greenland support intense phytoplankton blooms in spring. The timing of the onset in the ice-free waters upstream (i.e. to the southwest of Davis Strait) is determined by the establishment of water column stability, driven by ice-melt from West Greenland (Frajka-Williams and Rhines, 2010). In Davis Strait itself blooms appear as the seasonal ice-cover retreats northwards. These blooms are characterized by high phytoplankton biomass and a community of grazers dominated by large copepods, i.e. *Calanus* (Huntley et al., 1983; Head et al., 2003). Most higher trophic levels in the Arctic feed directly on *Calanus* (Falk-Petersen et al. 2009), which also play a key ecological role in supplying the benthic communities with high quality food via their production of large and fast-sinking faecal pellets (Juul-Pedersen et al., 2006). The vertical flux of faecal pellets sinking to the sea floor can be an important food source for benthic communities (Turner, 2002). Huntley et al. (1983) suggested a link between zooplankton community development, spring bloom dynamics and hydrography, with initial increases in zooplankton biomass arising from grazing of the spring phytoplankton bloom off west Greenland and proceeding in a counter-clockwise direction reaching Hudson Strait in September to October. Kenchington et al. (2016a) further identified a strong seasonal pattern for the 700 - 900 m depth range on the Greenland slope, associated with a recently discovered coral (*Lophelia pertusa*) reef. Both temperature and salinity reduced to their annual minimal values at the end of March and stayed low for one month with an indication of a second minimum in June, three months later. The occurrence and temporal extent of these minima likely arose through a combination of local convection from the surface and advection of cooled and freshened waters at depth from the Irminger Sea. This occurrence may extend across the sill in the southern Davis Strait, providing another mechanism for food supply to benthic species at depth in that area.

Species distribution modelling in this region should help to not only refine the KDE polygons determined from the analysis of the trawl survey catch (Kenchington et al., 2016b), but also to identify important areas for corals and sponges that are not covered by the surveys and to document the environmental parameters that help to shape those distributions. In this region, changes to the environment are expected to influence ecosystem structure and functioning at large spatial and temporal scales, due to recent decreases in multi-year ice (Ribergaard, 2012; Myers and Ribergaard, 2013). Identification of the parameters that control the distribution of corals and sponges will help to determine their susceptibility to the projected rapid and eminent environmental change in this region.

METHODOLOGY

Study Area

Two DFO Biogeographic Zone (see DFO, 2009) boundaries were modified and used as the spatial extent for species distribution modelling in this report: The Hudson Bay Complex Biogeographic Zone and the Eastern Arctic Biogeographic Zone. In the Hudson Bay Complex Biogeographic Zone, coral and sponge catch data from research vessel surveys are restricted to the Hudson Strait and Ungava Bay in the eastern portion of the zone. Consequently, species distribution models were run only in this area, termed the Hudson Strait - Ungava Bay Region herein (Figure 1). The total area covered in this extent is ~109,573 km^2 based on a NAD 1983 UTM Zone 19N projection. A 20-km land buffer was added around all land points to prevent their inclusion in the models.

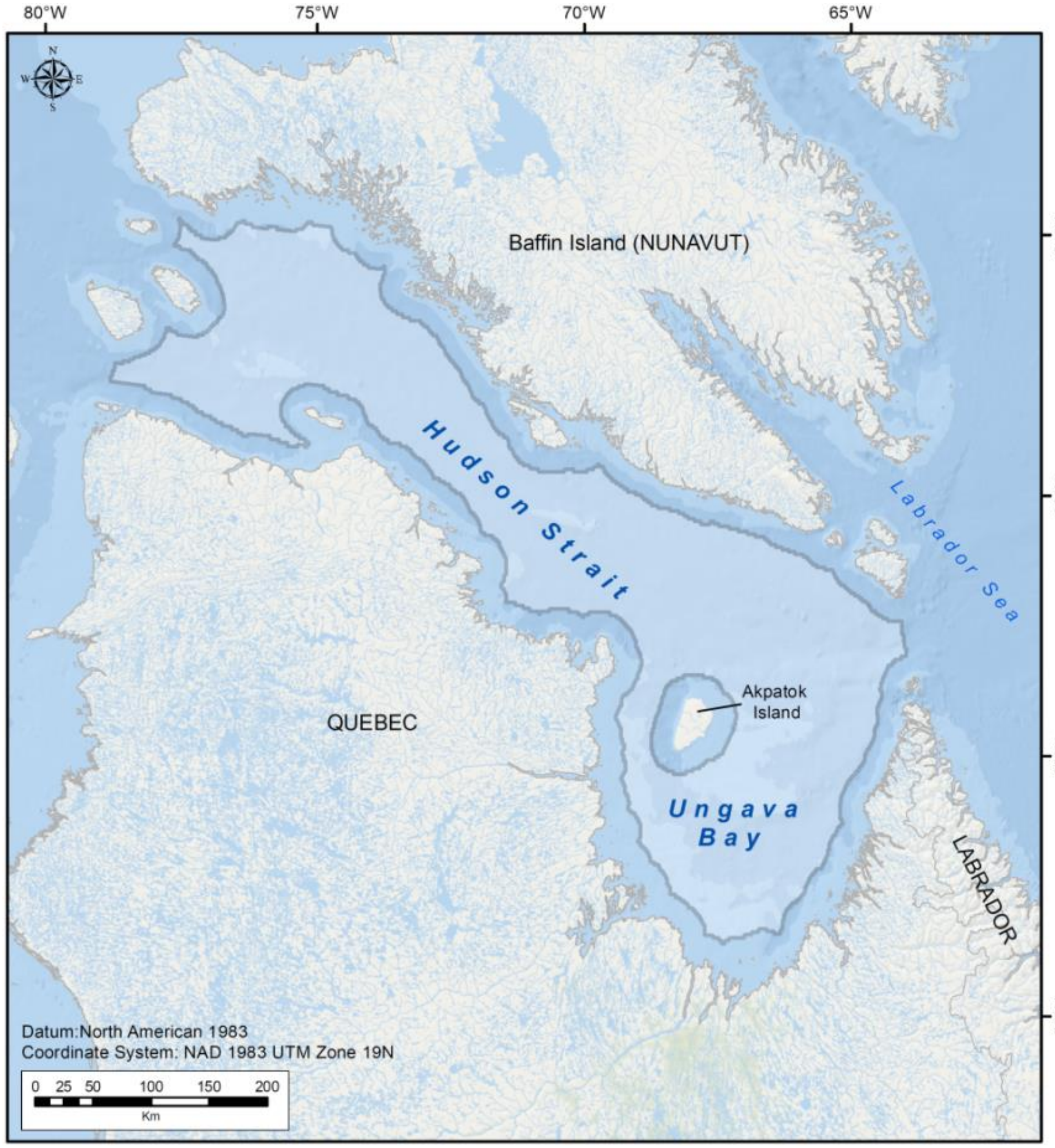

Figure 1. Extent of the Hudson Strait – Ungava Bay Region boundary used for species distribution modelling. Place names and major bodies of water are indicated.

The Eastern Arctic Biogeographic Zone is bounded by the Newfoundland and Labrador Region boundary in the south and the Canadian Exclusive Economic Zone (EEZ) in the east. In the north, the boundary ends at the Nares Strait in northern Baffin Bay and includes Lancaster Sound to the Barrow Strait and the Gulf of Boothia. The 20-km land buffer eliminated much of the study extent in Jones Sound above Devon Island and so this area was excluded from the study extent. The final study extent is referred to as the Eastern Arctic Region herein (Figure 2). There are two closure areas included in this study extent, the Narwhal Over-wintering and Deep-Sea Coral Conservation Area in Baffin Bay, and the Hatton Basin Voluntary Closure Area in Davis Strait. The total area covered in this extent is ~511,173 km^2 based on a NAD 1983 UTM Zone 20N projection.

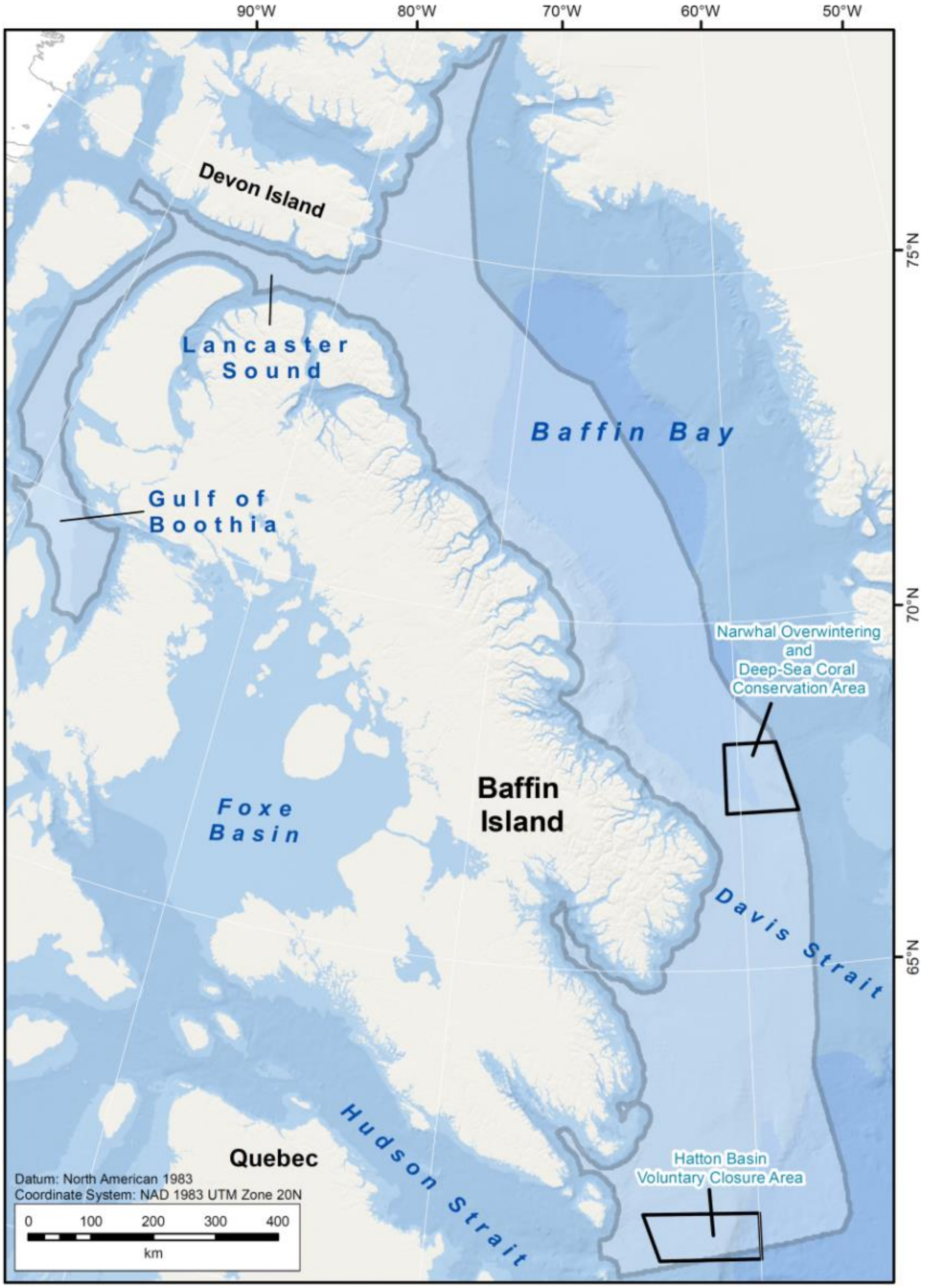
90°W
80°W
70°W
60°W
50°W
75°N
70°N
65°N
Devon Island
Lancaster Sound
Baffin Bay
Gulf of Boothia
Narwhal Overwintering and Deep-Sea Coral Conservation Area
Baffin Island
Foxe Basin
Davis Strait
Hudson Strait
Quebec
Hatton Basin Voluntary Closure Area
Datum: North American 1983
Coordinate System: NAD 1983 UTM Zone 20N
0
100
200
300
400
km

Figure 2. Extent of the Eastern Arctic Region boundary used for species distribution modelling. Place names, major bodies of water, and the location of the Hatton Basin Voluntary Closure Area and the Narwhal Over-wintering and Deep-Sea Coral Conservation Area are indicated.

Environmental Data

Fifty-four environmental variables derived from various sources and native spatial resolutions were used as predictor variables in the random forest models (Table 1). Compared with other species distribution models for Eastern Canada (Beazley et al., 2016a; Murillo et al., 2016; Guijarro et al., 2016), some seasonal productivity variables (Chlorophyll *a*, Primary Production) were not included due to their poor spatial coverage. Variables were chosen based on their availability and assumed relevance to the distribution of benthic fauna. Bathymetry was derived from the General Bathymetric Chart of the Oceans (GEBCO; http://www.gebco.net/data_and_products/gridded_bathymetry_data/). This data is the highest resolution bathymetry available for the entire study area. The data are resolved to 30 arc-seconds which is equivalent to approximately 500 m at 75°N in the eastern Arctic. 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 oceanographic or remote-sensing data and were spatially interpolated across the entire eastern Arctic using ordinary kriging in ArcMap. Specific details on data sources and methodology used for the spatial interpolation of these variables are documented in a separate technical report (in prep; although see Beazley et al., 2016b for information on similar environmental data sources and variables for the Estuary and Gulf of St. Lawrence). Only variables that were spatially interpolated with reasonable confidence were included in this report, and as a result, a number of available data layers were not considered. All predictor layers were displayed in raster format with geographic coordinates using the WGS 1984 datum and a ~0.013º cell size (approximately equal to 1 km horizontal resolution at 75ºN, approximately at the centre of the study extent).

Response Data

Species composition of the four taxonomic groups modelled in this report is presented in Table 2. Note that sponge collections at sea are not identified beyond the phylum (Porifera) level, and are further identified to the species level by specialists from Central and Eastern Arctic. Sea pen and large and small gorgonian coral catch records from the Hudson Strait - Ungava Bay Region are limited and were insufficient for modelling, and so only sponges from this area were modelled. Sponge catch records were collected between 2006 and 2014 and were derived from DFO/industry northern shrimp surveys conducted on fishing vessels *Cape Ballard*, *Aqviq*, or *Kinguk* between 2006 and 2014, and on the *Paamiut* in 2007, 2009, 2011, and 2013. Surveys conducted on the *Cape Ballard*, *Aqviq*, and *Kinguk* used Campelen trawl gear, while Cosmos trawl gear was used on the *Paamiut*. Coral and sponge catch data for the Eastern Arctic Region is derived from several different sources. Several surveys were conducted on the *Paamuit.* Initially targeting only Greenland Halibut, these surveys began targeting shrimp in 2006 and were conducted using both Alfredo and Cosmos trawl gear. Data was also available from the northern shrimp surveys conducted on the *Cape Ballard*, *Aqviq*, and *Kinguk* using Campelen trawl gear. Surveys from both the Hudson Strait – Ungava Bay and Eastern Arctic Regions were stratified-random. Absence records were created from null (zero) catches that occurred in the same surveys.

Table 1. Summary of the 54 environmental variables used as predictor variables in random forest modelling. N/A = Not applicable.

Variable	**Data Source**	**Temporal Range**	**Unit**	**Native Resolution**
Depth	GEBCO	N/A	metres	30 arc-sec.
Slope	GEBCO	N/A	degrees	30 arc-sec.
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 Fall Mixed Layer Depth	GLORYS2V1	1993 - 2011	metres	¼ º
Maximum Average Winter Mixed Layer Depth	GLORYS2V1	1993 - 2011	metres	¼ º
Maximum Average Spring Mixed Layer Depth	GLORYS2V1	1993 - 2011	metres	¼ º
Maximum Average Summer Mixed Layer Depth	GLORYS2V1	1993 - 2011	metres	¼ º
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,	2001 – 2010	$mg\ m^{-3}$	9 km

	NASA's OceanColor			
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
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 four taxonomic groups modelled using random forest. The asterisk (*) was used to indicate species/taxa recorded in both the Eastern Arctic and Hudson Strait – Ungava Bay Regions.

Taxonomic Group	**Species/Taxon**	**Taxon Code**
Sponges (Porifera)	Porifera P.	1101
Sea Pens (Pennatulacea)	Pennatulacea O.	8901
	Anthoptilum grandiflorum	8937
	Halipteris finmarchica	8936
	Pennatula grandis	8935
	Pennatula sp.	8954
	Umbellula sp.*	8972
	Sea pen sp.	8901
Large Gorgonian Corals	*Acanthogorgia armata**	8907
	Keratoisis ornata	8906
	*Paragorgia arborea**	8903
	Paramuricea sp.	8912
	Paramuricea placomus [28S-b]	8940
	*Primnoa resedaeformis**	8902
Small Gorgonian Corals	*Acanella arbuscula*	8909
	Anthothela cf. *grandiflora*	8915
	Radicipes gracilis	8910

The presence-absence records used in each random forest model (see below) 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.

Records from the Fisheries Observer Program (FOP) data (for more details contact V. Wareham, DFO, NWAFC, St. John's, NL; pers. comm.) from the period of 1998 to 2013 were used to validate the presence probability maps for all four taxonomic groups in the Eastern Arctic Region. A total of 4029 sponge, 1345 sea pen, 227 large gorgonian, and 1588 small gorgonian coral records were obtained. An additional 2238 sponge records were obtained from commercial surveys conducted between 1979 and 2001 using bottom otter trawl and shrimp trawl gear (S. Fuller, Ecology Action Centre, Halifax, NS, pers. comm.) and used for model validation. This second data set included data from the first for the overlapping years but as each dataset was separately edited by the data providers we chose to analyze each separately. In both cases only one position was given for each catch location. No validation data was available for the Hudson Strait-Ungava Bay Region.

Biomass (kg) data associated with the DFO trawl survey records were also extracted and used in random forest modelling. In order to avoid the introduction of bias related to differences in catchability between gear types, biomass regression random forest models were run separately on each of the three gear types when applicable. For each taxonomic group, weights were averaged across multiple tows occurring within the same environmental raster cell. Catch weight provided for each coral record from the DFO/industry northern shrimp surveys was used unless sample weight exceeded catch weight, in which case the latter was used.

Random Forest Modelling

Random forest (Breiman, 2001) is a non-parametric machine learning technique, where multiple regression or classification trees (usually $\geq$ 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 covariate relation to the response in comparison to other classical statistical models such as generalized linear models (GLM) or generalized additive models (GAM).

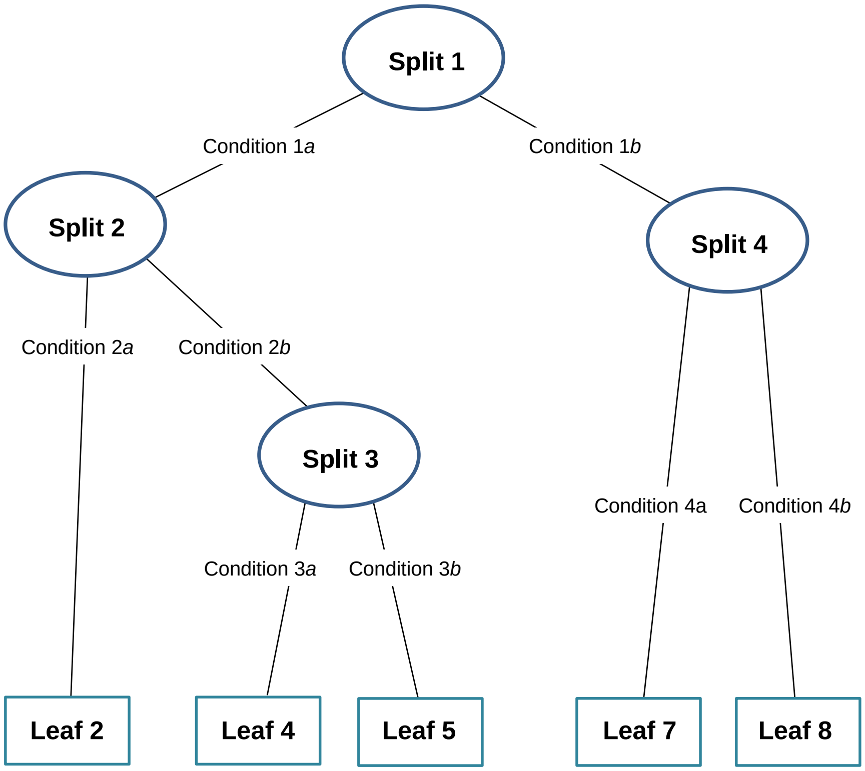

Figure 3. An example of a regression model tree (adapted from Kuhn and Johnson, 2013).

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

Model Evaluation

Presence-Absence Response Data – Classification Model

The catch records for some taxonomic groups are characterized by a higher number of absences relative to presences (i.e. unbalanced species prevalence, where prevalence is the proportion of presences in relation to the total dataset). 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 (Jimenez-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 modeling prevalence (Liu et al., 2005; Hanberry and He, 2013).

Given the numerically and/or spatially biased presence and absence data of most taxonomic groups in this study, we employed two different modelling approaches and evaluated their performance. The first approach 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 set the threshold equal to species prevalence. 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 derived from validated data 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, or Area Under the Receiver Operating Characteristic Curve. 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).

For models generated using a balanced species prevalence and threshold of 0.5, 10 data subsets were created with the same number of presence and absences (balanced data) and the AUC was determined by averaging the 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 folds to give one confusion matrix from which sensitivity and specificity were calculated.

Biomass Response Data – Regression Model

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 (=folds) to better evenly distribute the biomass values across splits. Models were built using each calibrated and validated dataset and accuracy measures were calculated for each of the 10 model runs. The accuracy measures used to validate the models included the goodness-of-fit statistic R^2, the Root-Mean-Square Error (RMSE) and the percentage of variance explained. RMSE was normalized 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. 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.

Model Extrapolation

The spatial distribution of data observations, particularly in the Eastern Arctic Region, is mainly limited to the southern portion of the study extent in Davis Strait and the shelf in Baffin Bay. There are no data observations in the deeper waters off Baffin Island Shelf, in Lancaster Sound, or in the Gulf of Boothia. Extrapolation of model predictions to areas outside of the range of data observations may produce unreliable predictions in those areas (Elith et al., 2010). Random forest models average the decision across regression trees to predict piecewise constant functions, giving a constant value for inputs falling under each leaf. When extrapolating outside the domain of the training data, where different physical conditions from those used to train the model likely exist, random forest models predict the same value as they would for the closest value in the tree for which they had training data (Breiman et al., 1984). For each random forest model, we highlight those areas within the study extent where model predictions are extrapolated. We define areas of extrapolation as those areas where at least one environmental variable has values above or below its sampled range.

Ecological Interpretation

Ecological interpretation of the models was aided by predictor variable importance measures and partial dependence plots. In classification models, 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 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 six most important variables. Partial dependence plots show the relationship between a particular predictor variable and the log-transformed predicted probabilities of presence for classification models or the biomass regression function for regression models. The other predictor variables are held constant at their mean observed value. Partial dependence plots 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 -∞ 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.

Alternative Prediction Models

Generalized additive models (GAMs) were developed to predict the biomass distribution of each taxonomic group. GAMs were developed to compare a regression approach to the machine learning random forest results and to determine whether predictions could be improved for the areas considered as extrapolated by random forest models. Methodology and results for the GAM models are presented in Appendix 1.

RESULTS

Hudson Strait – Ungava Bay Region

Sponges (Porifera)

Data Sources and Distribution

Sponge catch data for the Hudson Strait – Ungava Bay Region was collected between 2006 and 2014 and consisted of 49 presences and 82 absences from the *Cape Ballard*, *Aqviq*, and *Kinguk* surveys using Campelen trawl gear, and 181 presences and 88 absences from the *Paamuit* surveys using Cosmos gear (Figure 4). The combined dataset consisting of 230 presences and 170 absences (see Table 3) had an uneven spatial distribution across the study area (Figure 5). Both presences and absences were concentrated in eastern Hudson Strait and Ungava Bay, with few data points locate south of Akpatok Island. The western portion of Hudson Strait had only a few absence records that were located near the coast off northern Quebec. Presences records in this area had a relatively uniform distribution. The highest mean biomass records (up to 14.12 kg) were found in Ungava Bay, particularly where Hudson Strait meets the Labrador Sea.

Table 3. Number of presence and absence records of sponge catch recorded from DFO surveys using both Campelen and Cosmos trawl gear between 2006 and 2014 in the Hudson Strait – Ungava Bay Region.

	Campelen		**Cosmos**	
Year	**Presences**	**Absences**	**Presences**	**Absences**
2006	3	7	-	-
2007	2	7	55	1
2008	1	5	-	-
2009	-	10	83	17
2010	6	5	-	-
2011	5	6	24	33
2012	2	7	-	-
2013	4	8	19	37
2014	26	27	-	-
TOTAL	**49**	**82**	**181**	**88**

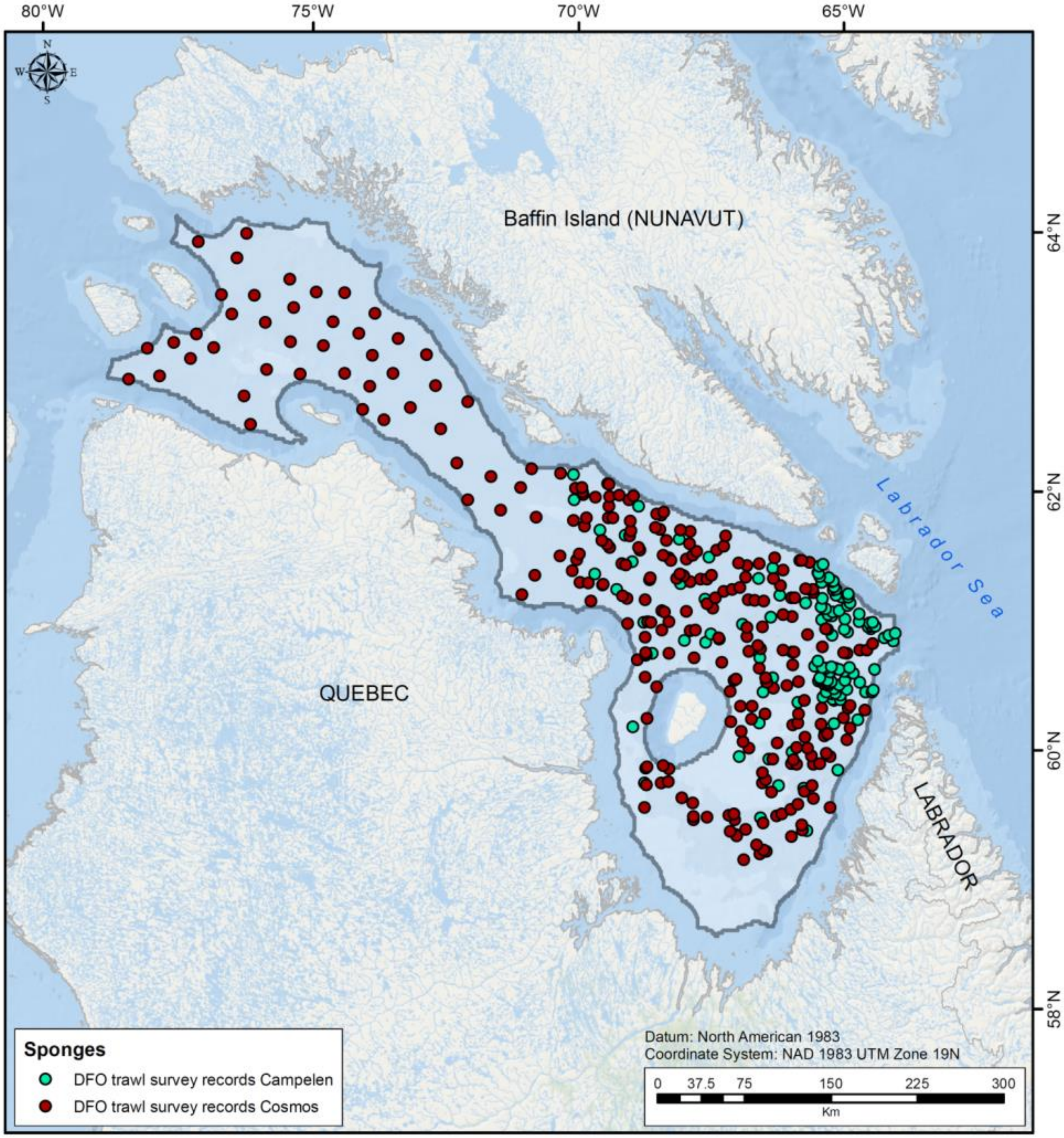
80°W
75°W
70°W
65°W
64°N
62°N
60°N
58°N
N
W
E
S
Baffin Island (NUNAVUT)
Labrador Sea
QUEBEC
LABRADOR
Sponges
● DFO trawl survey records Campelen
● DFO trawl survey records Cosmos
Datum: North American 1983
Coordinate System: NAD 1983 UTM Zone 19N
0 37.5 75 150 225 300
Km

Figure 4. Available sponge presence and absence records by gear type in the Hudson Strait – Ungava Bay Region from DFO trawl surveys conducted between 2006 and 2014.

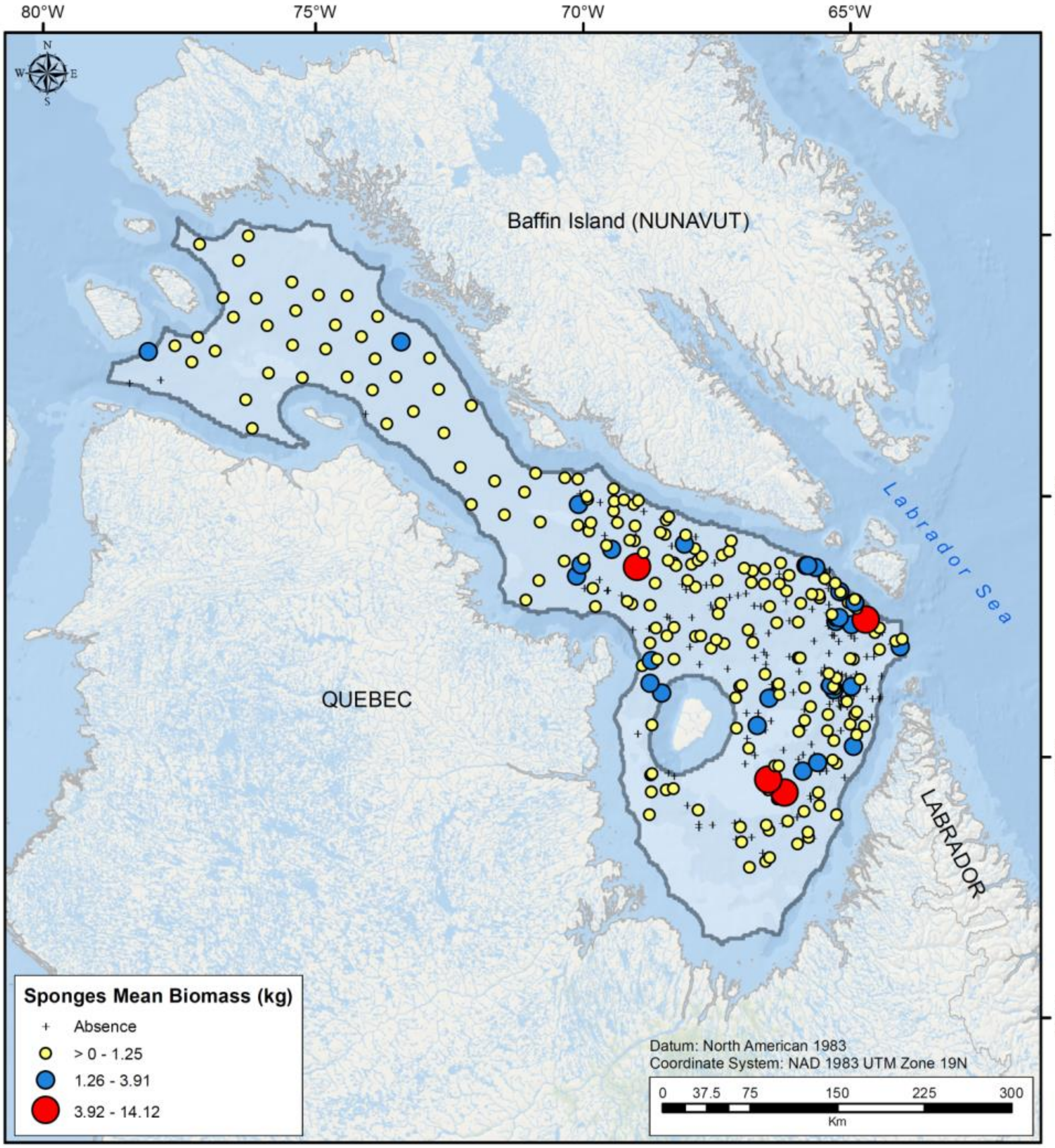

Figure 5. Mean biomass (kg) per grid cell of sponge catch data recorded from DFO trawl surveys conducted in the Hudson Strait – Ungava Bay Region between 2006 and 2014. Also shown are absence records from the same surveys.

Model 1 – Balanced Species Prevalence

Accuracy measures for the random forest model on balanced species prevalence (170 presences and 170 absences; Model 1) are presented in Table 4. The highest mean AUC of 0.677 was associated with Model Run 1 and is therefore considered the optimal model for the prediction of the sponge response data. The sensitivity and specificity measures of this model fold were 0.594 and 0.647, respectively and were slightly higher than the average of all model folds. The confusion matrix of the optimal model is also presented in Table 4. Class error for both the presence and absence classes was moderate.

Table 4. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of sponge presence-absence data collected within the Hudson Strait – Ungava Bay Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 1) which is considered the optimal model for predicting the presence probability of sponge in the region.

Model Run	AUC	Sensitivity	Specificity
1	**0.677**	**0.594**	**0.647**
2	0.671	0.624	0.647
3	0.597	0.559	0.571
4	0.616	0.535	0.641
5	0.665	0.571	0.618
6	0.652	0.635	0.600
7	0.638	0.618	0.606
8	0.641	0.576	0.624
9	0.620	0.576	0.565
10	0.599	0.541	0.624
Mean	**0.638**	**0.583**	**0.614**
SD	**0.029**	**0.034**	**0.029**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	**Presence**		
Absence	110	60	170	0.353
Presence	69	101	170	0.406

The presence probability prediction surface of sponges generated from Model 1 is presented in Figure 6. Western Hudson Strait was predicted to have high and relatively even presence probability of sponges. Pockets of high sponge presence probability were distributed across eastern Hudson Strait and Ungava Bay, with larger areas of high presence probability located northeast of Akpatok Island and south of Baffin Island. Areas of high and low presence probability of sponges corresponded well with the location of presence and absence data points (Figure 7). Figure 8 shows the actual data points used in Model 1. There was extrapolation of high and low sponge presence probability beyond the location of presence and absence data, respectively. Even with the reduction in the number of presence data points used in the model from western Hudson Strait, this area still had a high and even presence probability of sponges. Areas of model extrapolation are also shown in this figure. Southern Ungava Bay and smaller areas in western Hudson Strait are considered areas of model extrapolation.

Of all 54 environmental predictor variables used in the model, Surface Current Mean was the most important for the classification of the sponge presence and absence data (Figure 9). This was followed by Surface Temperature Average Minimum, Summer Chlorophyll a Minimum, and the remaining variables in the model. Partial dependence plots for the top 6 predictor variables are shown in Figure 10. The highest predicted sponge presence probabilities occurred between 0.02 and 0.08 m s^{-1} along the gradient in Surface Current Mean. Sponge presence probability was highest between Surface Temperature Average Minimum values of -1.85 and -1.75ºC.

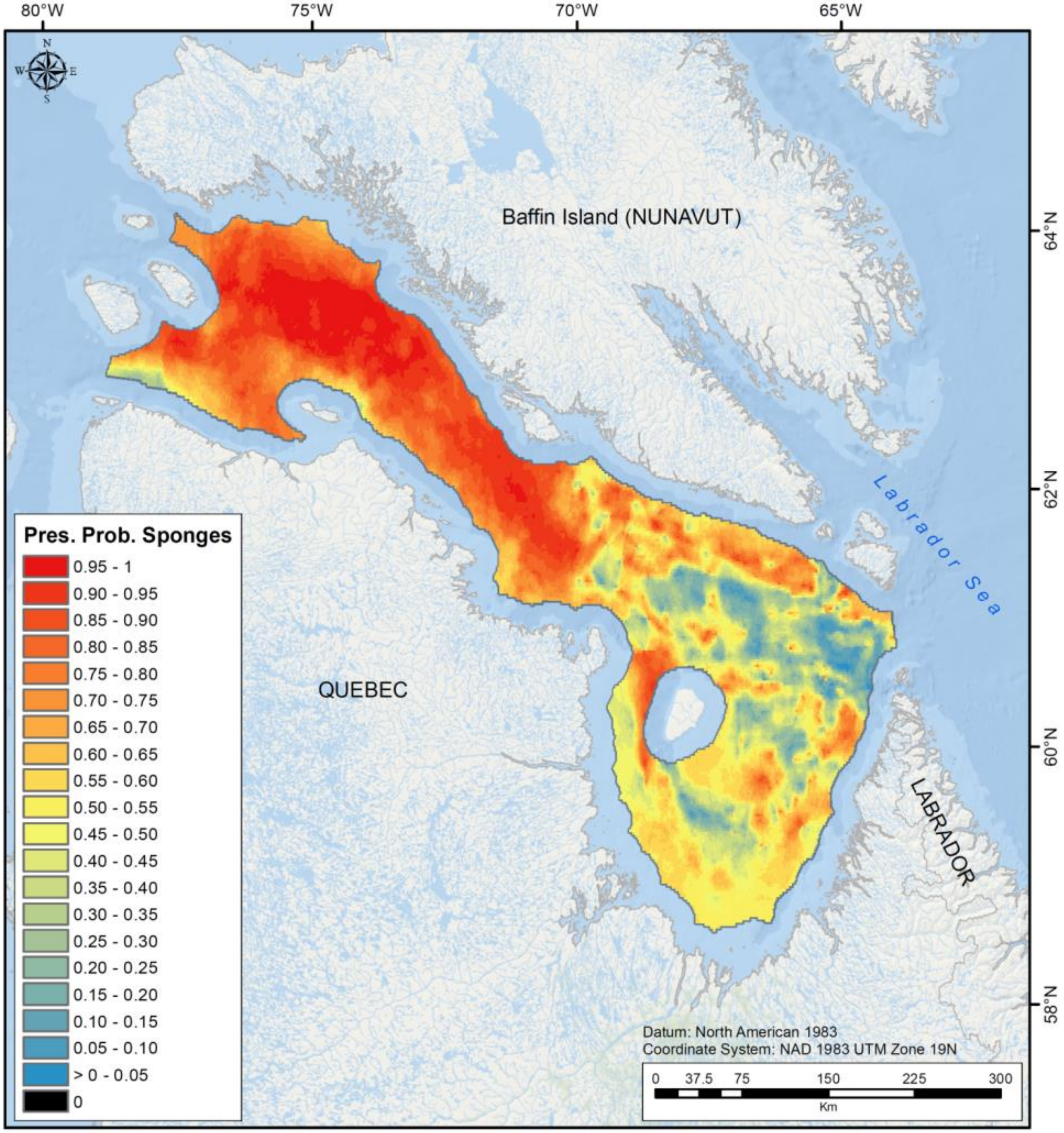

Figure 6. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of sponge presence and absence data collected from DFO trawl surveys conducted in the Hudson Strait – Ungava Bay Region between 2006 and 2014.

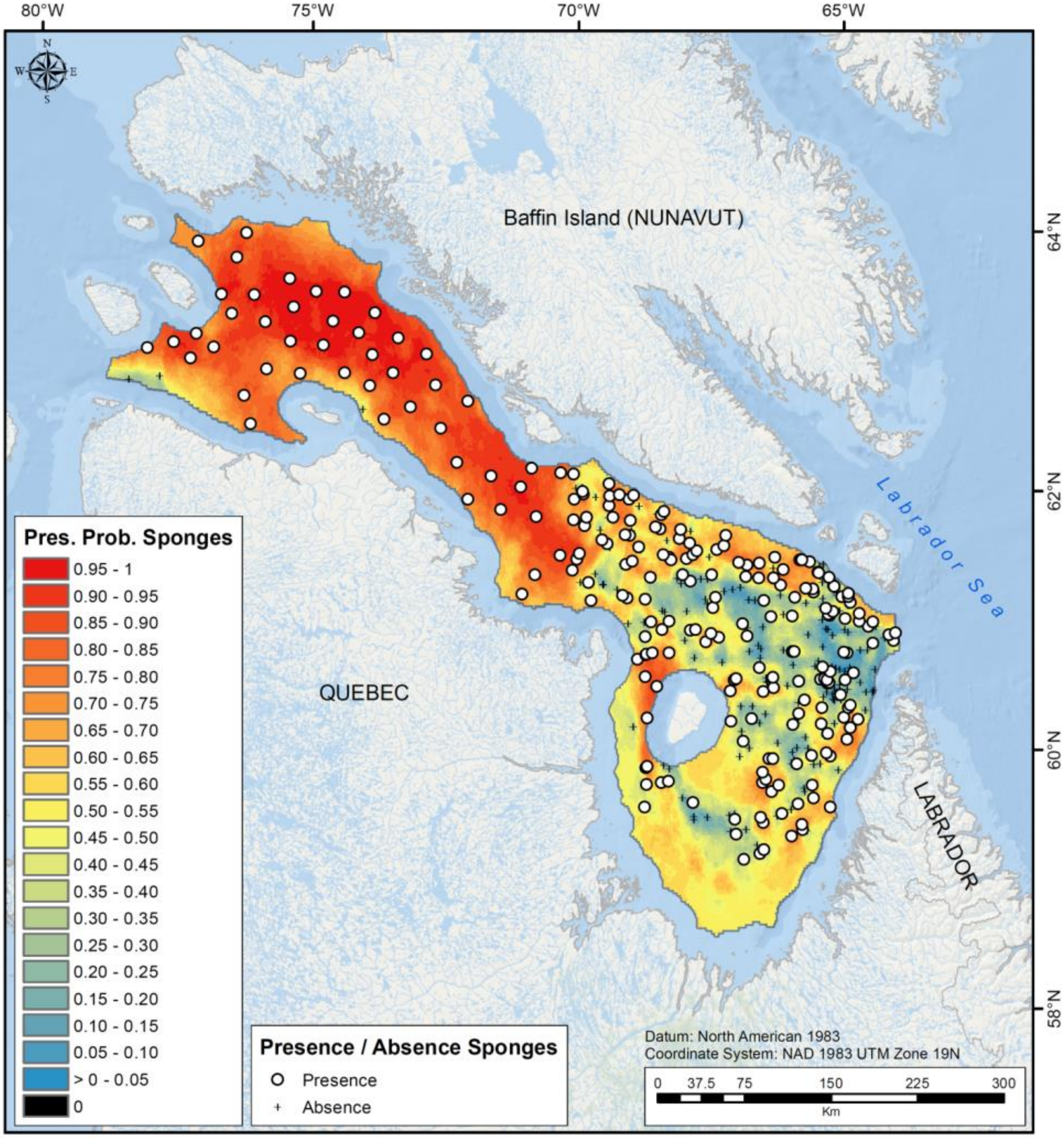

Figure 7. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of sponge presence and absence data recorded from DFO trawl surveys conducted in the Hudson Strait – Ungava Bay Area between 2006 and 2014.

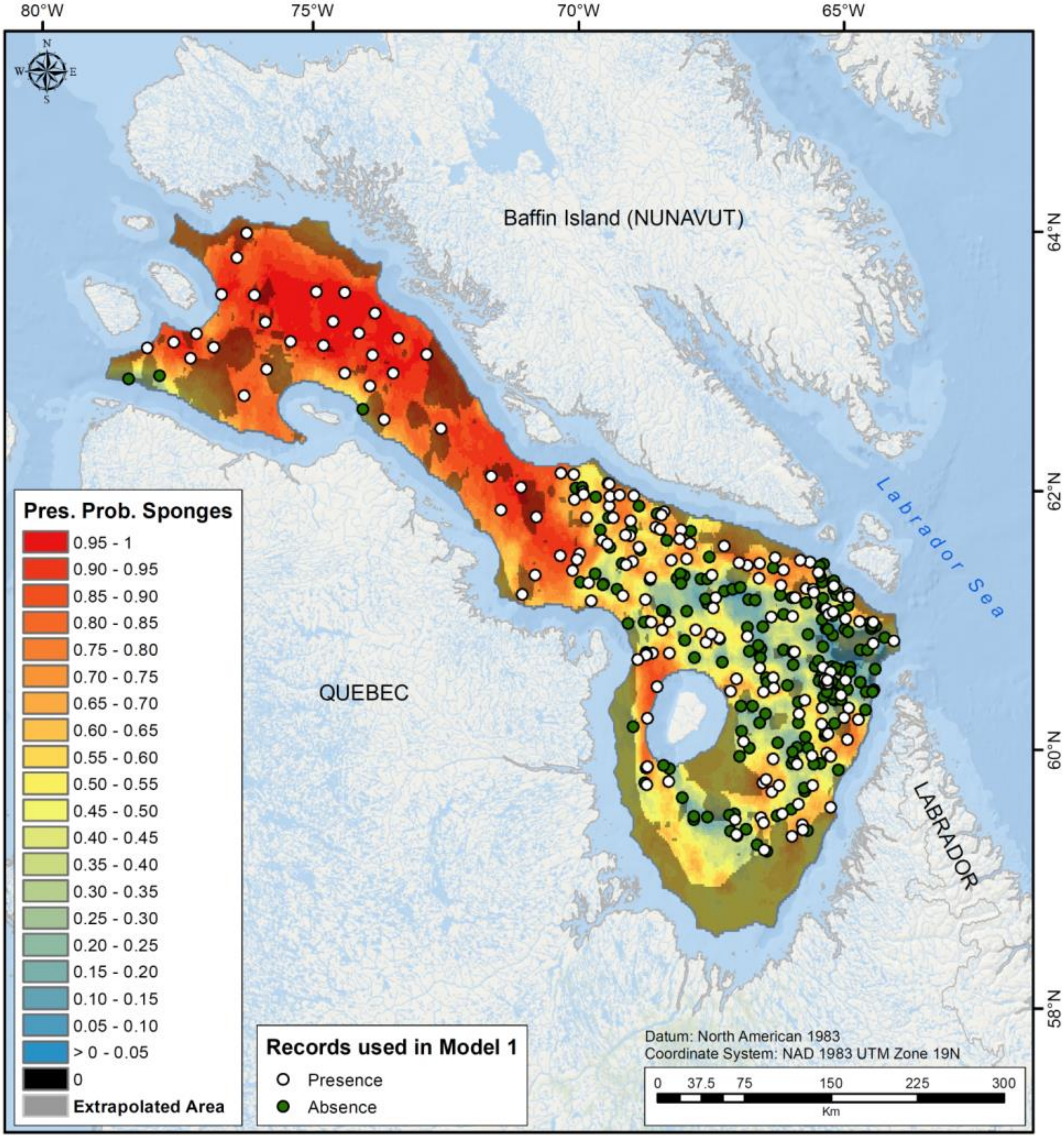

Figure 8. Map of the 340 data observations (170 presences and 170 absences) of sponges used in the optimal random forest Model 1. Also shown is the predicted presence probability (Pres. Prob.) of sponges and the areas of model extrapolation.

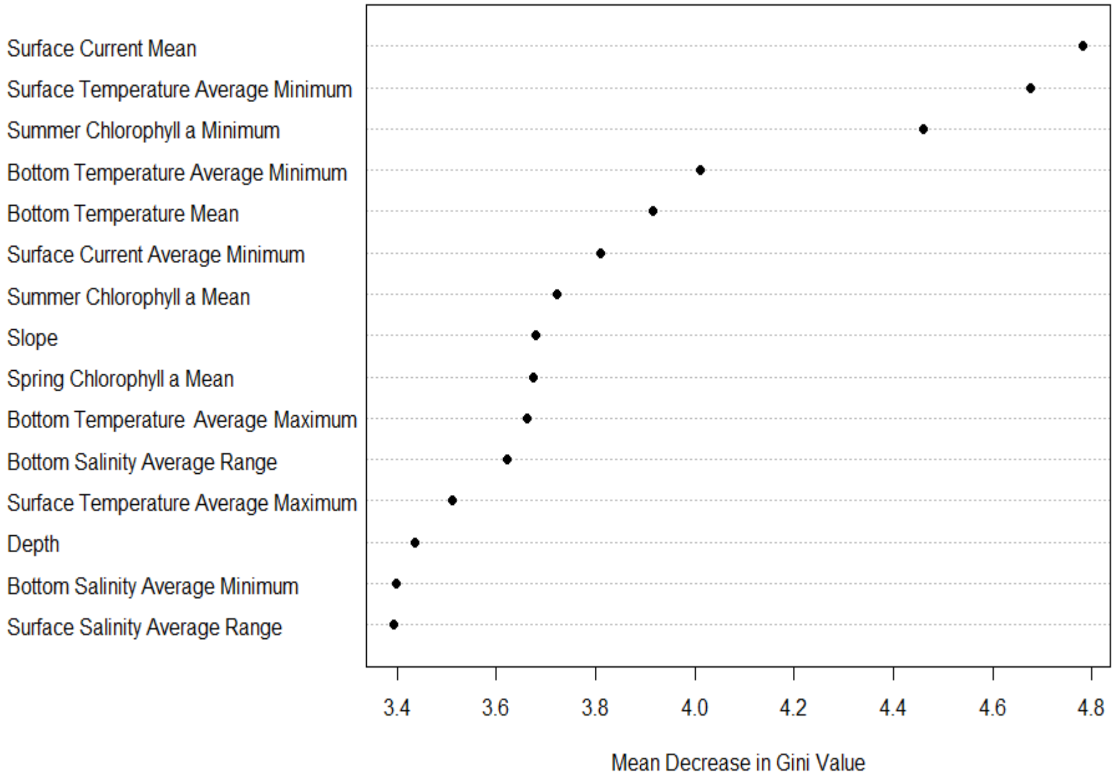

Figure 9. Importance of the top 15 predictor variables measures as the Mean Decrease in Gini value of the optimal random forest model predicting sponge presence and absence data within the Hudson Strait – Ungava Bay Area. The higher the Mean Gini value the more important the variable is for predicting the response data.

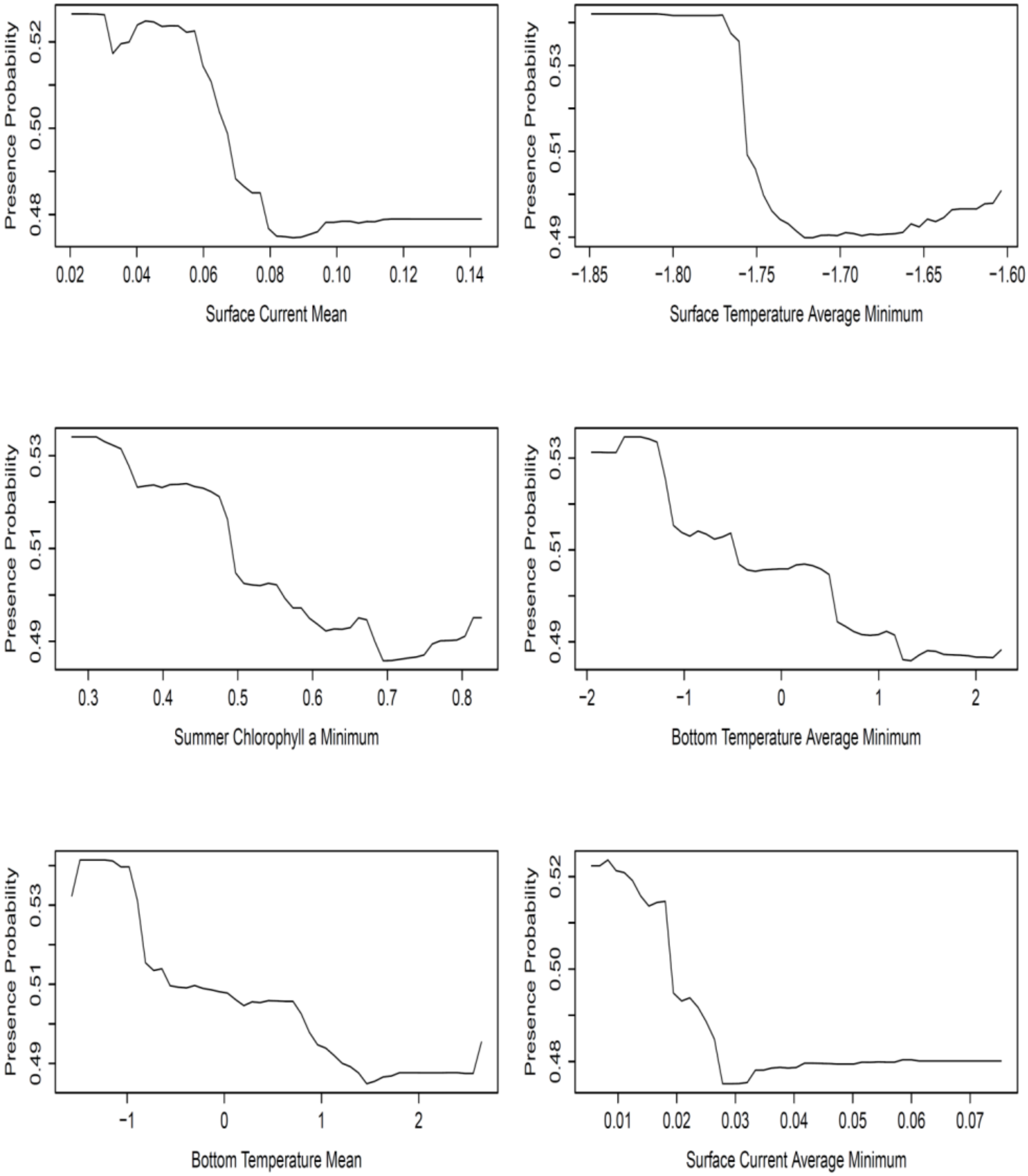

Figure 10. Partial dependence plots of the top six predictors from the optimal random forest model of sponge presence and absence data collected within the Hudson Strait – Ungava Bay Area, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence

Model accuracy measures for the random forest model on unbalanced species prevalence (230 presences and 170 absences; Model 2) and threshold equal to species prevalence (0.58) are presented in Table 5. The average AUC from this model was 0.643, indicating poor model performance. Sensitivity and specificity were also low (0.574 and 0.612, respectively). Class error of both the presence and absence classes was relatively high (0.426 and 0.388, respectively).

The surface of predicted presence probability of sponges generated from Model 2 is presented in Figure 11. Western Hudson Strait was predicted to have high and even presence probability of sponges. Pockets of sponge presence probability were distributed across eastern Hudson Strait and Ungava Bay, with larger areas of high presence probability located northwest of Akpatok Island and south of Baffin Island. Areas of high and low presence probability of sponges corresponded well with the location of presence and absence data points (Figure 12). Figure 13 depicts the classification of sponge presence probability into presence and absence categories based on the prevalence threshold of 0.58. Most of western Hudson Strait was classified as presence of sponges. The northern portion of Ungava Bay east of Aktapok Island was classified as absence of sponges. Areas of model extrapolation are also shown in this figure. The largest area of extrapolation lies in southern Ungava Bay.

Table 5. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model presence and absence of sponges within the Hudson Strait – Ungava Bay Area. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	**AUC**	**Observ.**	**Predictions**		**Total n**	**Class error**	**Sensit.**	**Specif.**
1	0.569		**Absence**	**Presence**				
2	0.732	**Absence**	104	66	170	0.388	0.574	0.612
3	0.692	**Presence**	98	132	230	0.426		
4	0.697							
5	0.638							
6	0.668							
7	0.506							
8	0.724							
9	0.510							
10	0.689							
Mean	**0.643**							
SD	**0.085**							

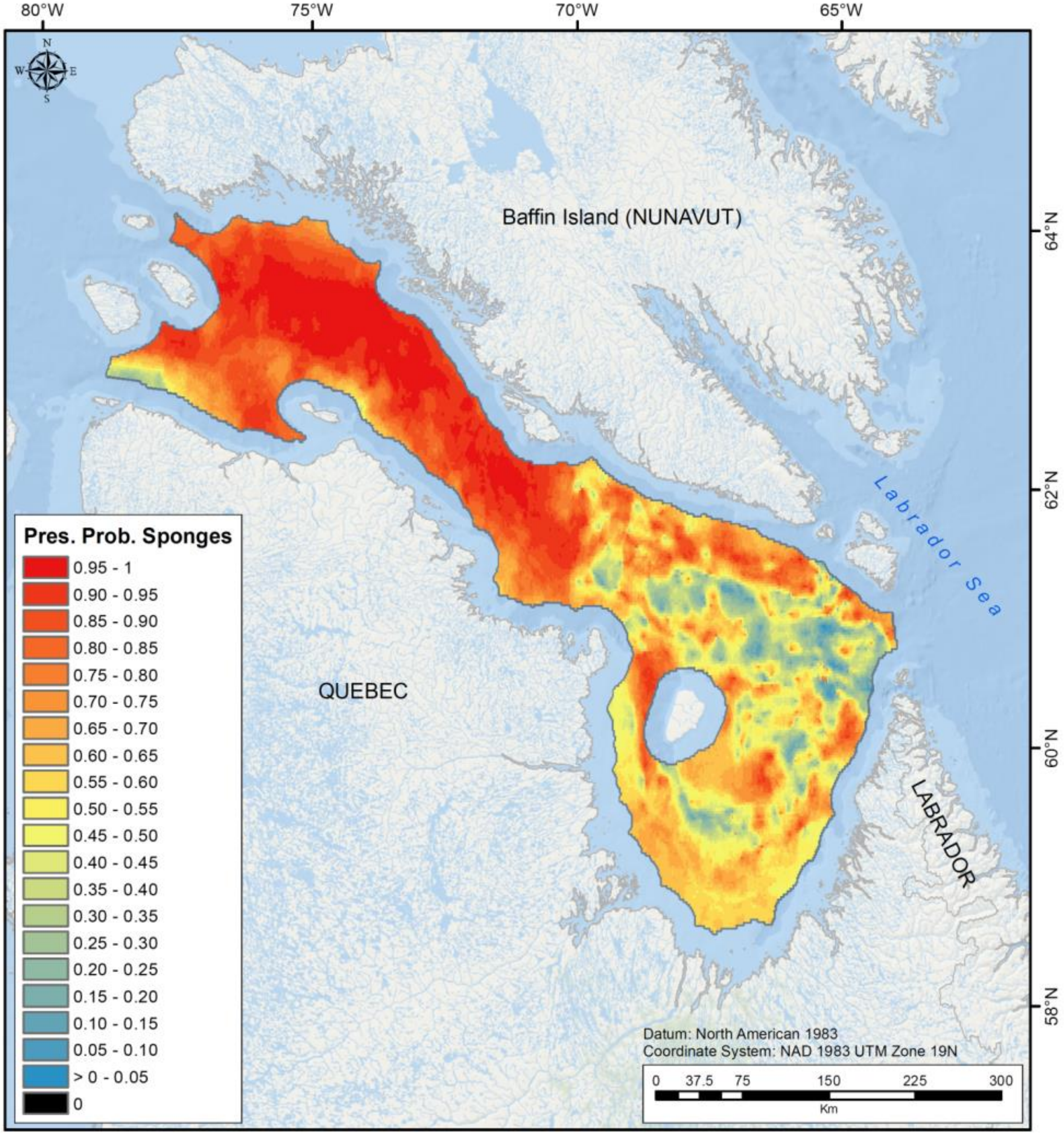

Figure 11. Predictions of presence probability from the unbalanced random forest model of sponge presence and absence data collected from DFO trawl surveys conducted in the Hudson Strait – Ungava Bay Area between 2006 and 2014.

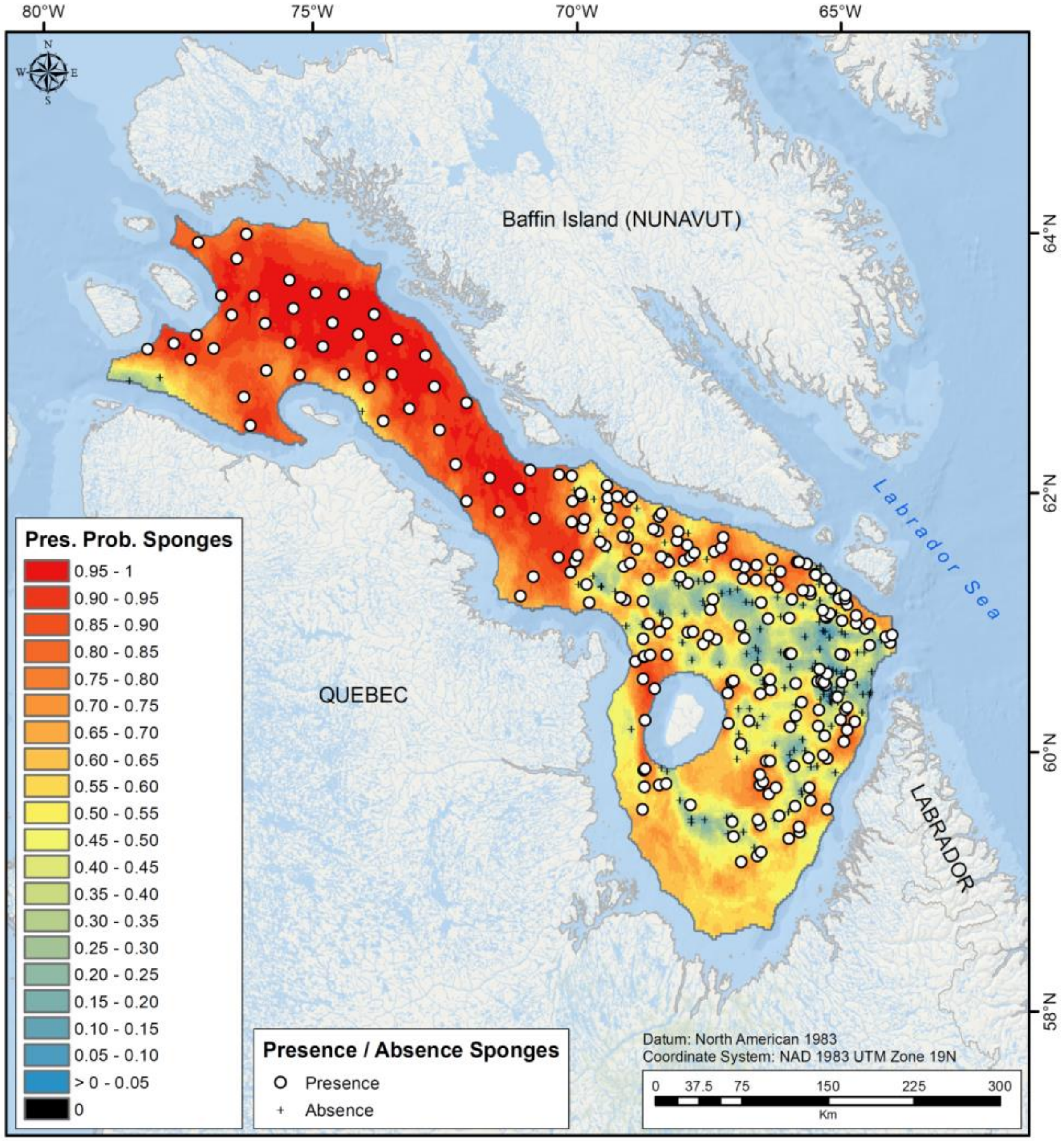

Figure 12. Presence and absence observations and predictions of presence probability of the unbalanced random forest model of sponge presence and absence data recorded from DFO surveys conducted in the Hudson Strait - Ungava Bay Region between 2006 and 2014.

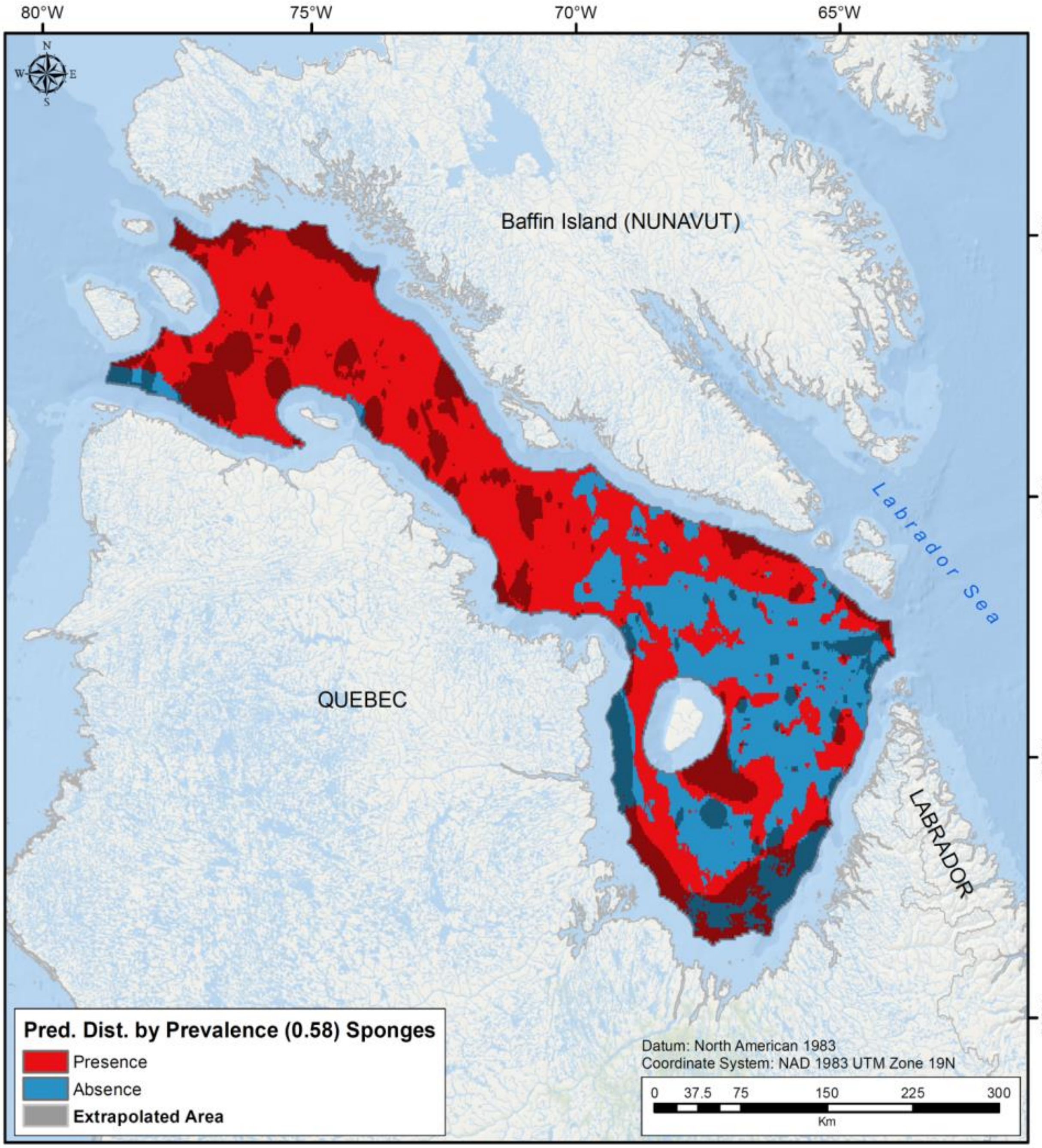

Figure 13. Predicted distribution (Pred. Dist.) of sponges in the Hudson Strait – Ungava Bay Area based on the prevalence threshold of 0.58 of sponge presence and absence data used in Model 2. Also shown are the areas of model extrapolation (grey polygon may appear red or blue).

The order of importance of the environmental predictor variables in the threshold equal to species prevalence Model was shown in Figure 14. Of all 54 environmental predictor variables used in the model, Surface Current Mean was the most important for the classification of the sponge presence and absence data. Surface Current Mean was followed closely by Summer

Chlorophyll *a* Minimum and more distantly by Surface Current Average Minimum and Slope. Partial dependence plots for the top 6 predictor variables are shown in Figure 15. The highest presence probability occurred between 0.02 - 0.08 m s^{-1} along the Surface Current Mean gradient.

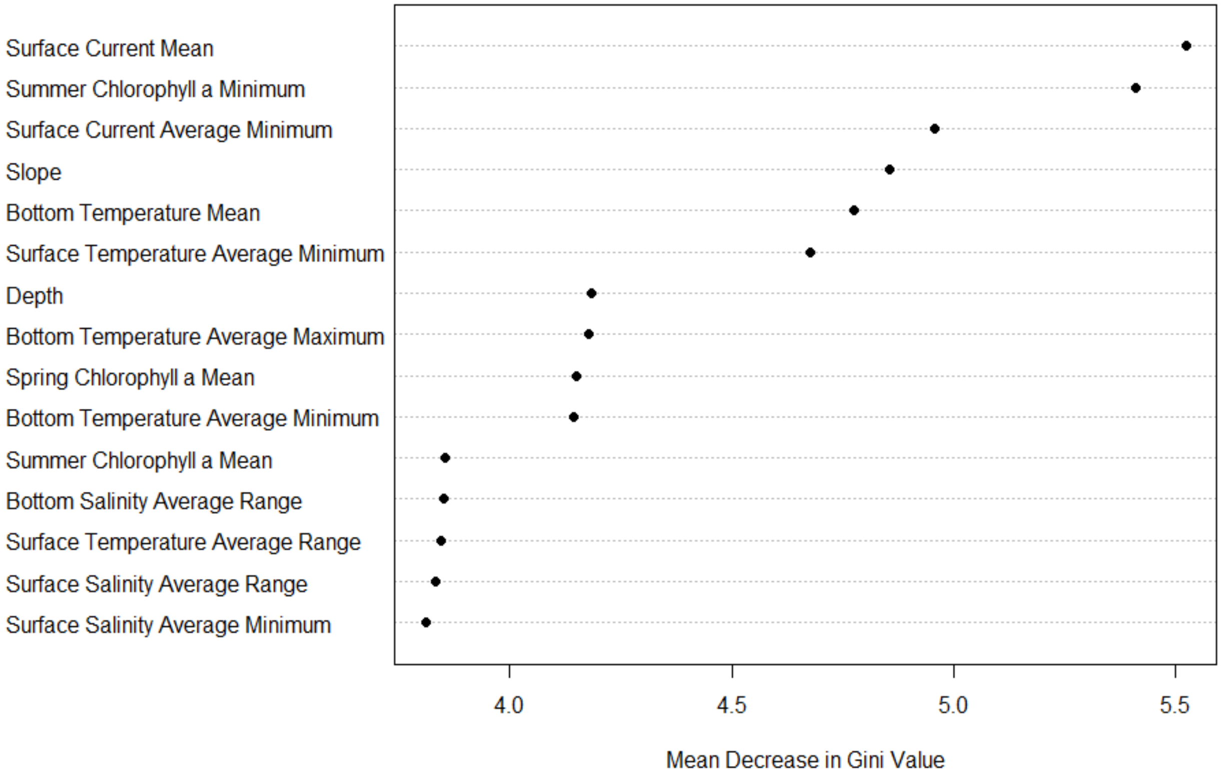

Figure 14. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini Value of the unbalanced random forest model predicting sponge presence and absence data within the eastern portion of the Hudson Bay Complex Region. The higher the Mean Decrease in Gini value the more important the variable is for predicting the response data.

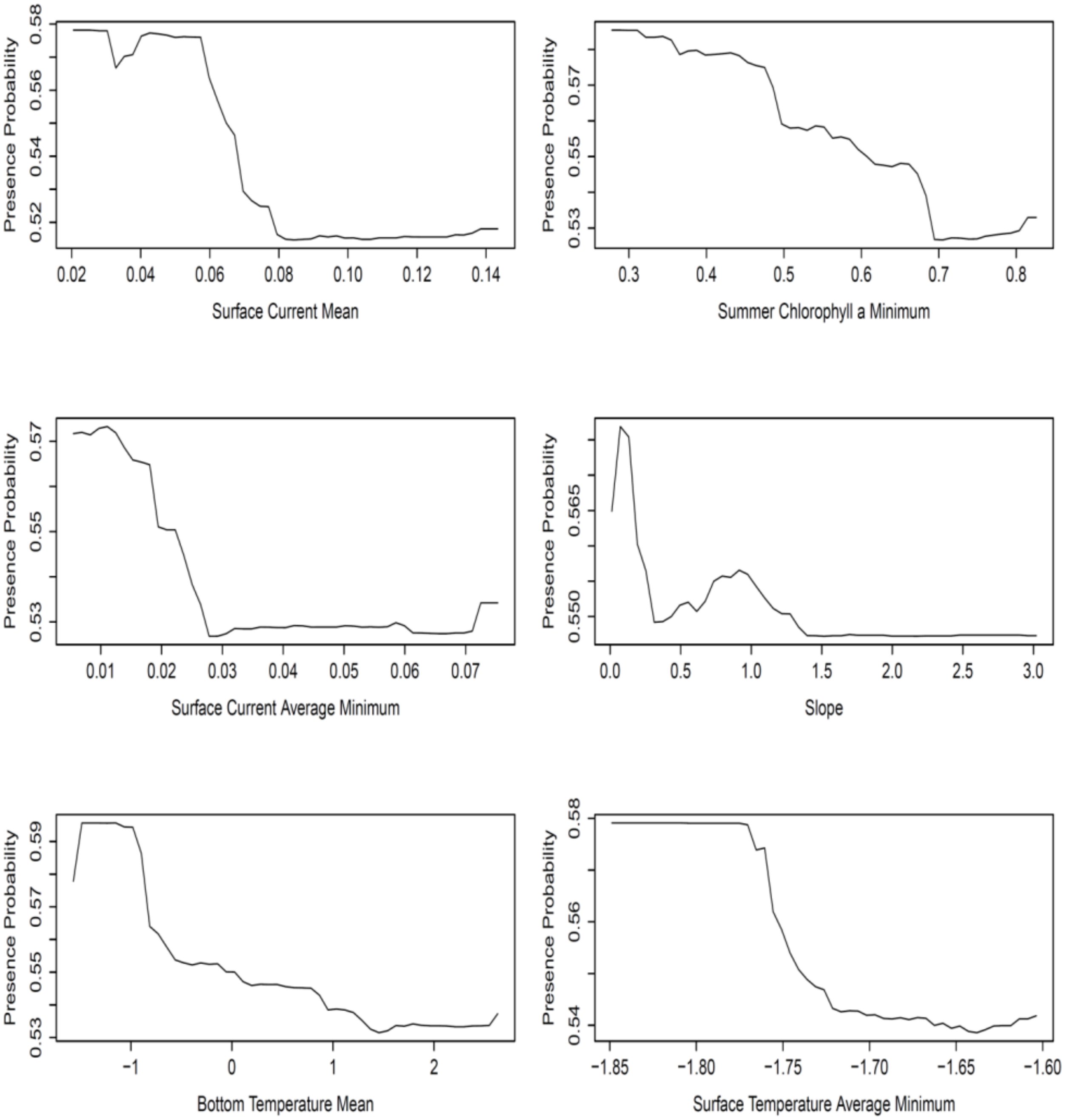

Figure 15. Partial dependence plots of the top 6 predictors from the unbalance random forest model of sponge presence and absence data collected in the Hudson Strait – Ungava Bay Area, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using all available sponge records and an unbalanced species prevalence (Model 2) was selected as the best predictor of sponge distribution in the Hudson

Strait – Ungava Bay Region. In this dataset, the number of presences was higher than the number of absences. Although accuracy measures were comparable between both Models 1 (balanced dataset) and 2, Model 2 allowed for the use of all presence data in the region, providing a more accurate depiction of the distribution of sponges. Note that there was no independent data for use in model validation in the Hudson Strait – Ungava Bay Region.

Prediction of Sponge Biomass Using Random Forest

Cosmos Trawl Gear

Regression random forest was used to model mean sponge biomass per grid cell from the Cosmos surveys only. Accuracy measures are presented in Table 6. The highest R^2 was 0.246, while the average was 0.101 ± 0.086 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.075 ± 0.042 SD. The negative percentage variance explained by this model indicates poor model performance (average = -8.67% ± 2.41 SD).

Figures 16 and 17 show the predicted biomass surface of sponges. The majority of the spatial extent was predicted to have low (0 – 0.14 kg) sponge biomass. The area of highest predicted biomass occurred southeast of Atpatok Island. This area coincided with the location of the highest mean biomass value (Figure 17). A narrow strip of moderate to high predicted biomass occurred around Ungava Bay. These areas of high predicted biomass are considered areas of model extrapolation.

Table 6. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of average sponge biomass (kg) recorded from DFO trawl surveys conducted using Cosmos gear in the Hudson Strait - Ungava Bay Region. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	**R^2**	**RMSE**	**NRMSE**	**Percent (%) variance explained**
1	0.165	0.513	0.059	-5.69
2	0.000	0.776	0.089	-5.55
3	0.095	0.449	0.051	-10.81
4	0.034	1.653	0.189	-7.31
5	0.031	0.441	0.050	-8.73
6	0.228	0.448	0.051	-7.75
7	0.094	0.630	0.072	-9.13
8	0.087	0.644	0.074	-10.85
9	0.246	0.515	0.059	-13.17
10	0.029	0.462	0.053	-7.69
Mean	**0.101**	**0.653**	**0.075**	**-8.67**
SD	**0.086**	**0.368**	**0.042**	**2.41**

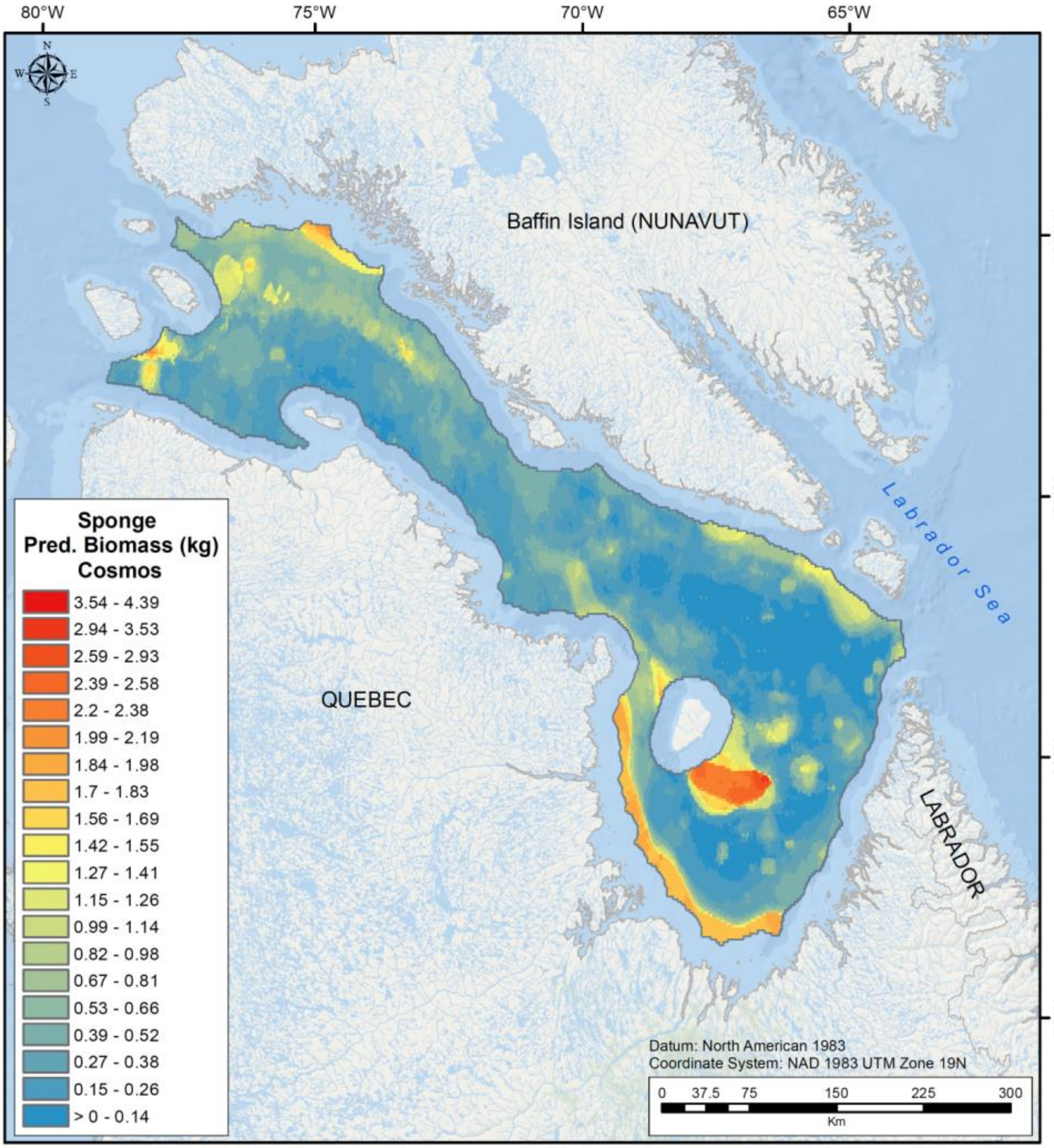

Figure 16. Predictions of biomass (kg) per sponges from catch data recorded in DFO trawl surveys conducted using Cosmos gear in the Hudson Strait - Ungava Bay Region between 2006 and 2014.

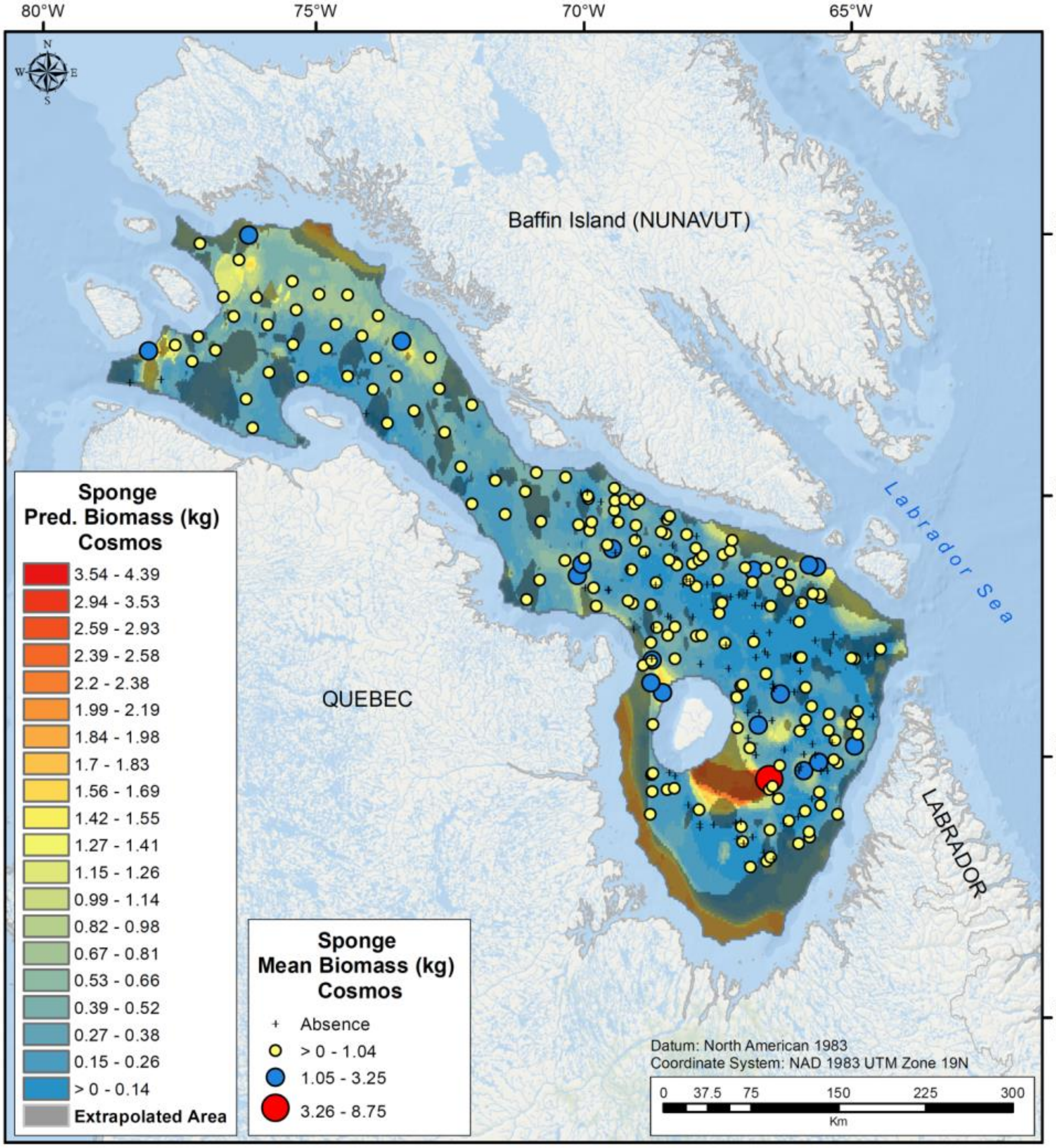

Figure 17. Predictions of biomass (kg) of sponges from catch data recorded in DFO trawl surveys conducted using Cosmos gear in the Hudson Strait - Ungava Bay Region between 2006 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting sponge biomass are shown in Figure 18. Like the presence-absence model on unbalanced data (Model 2), Surface Current Mean was the most important variable in the biomass model. This variable was followed more distantly by Bottom Temperature Average Range, Bottom Salinity Average Maximum, and the other variables in the model displayed a right-skewed distribution with The partial dependence of sponge biomass on the top 6 most important variables is shown in Figure 19. Predicted biomass was highest at the lowest Surface Current Mean values (< 0.03 m s^{-1}) and highest Bottom Temperature Average Range values (> 1.0).

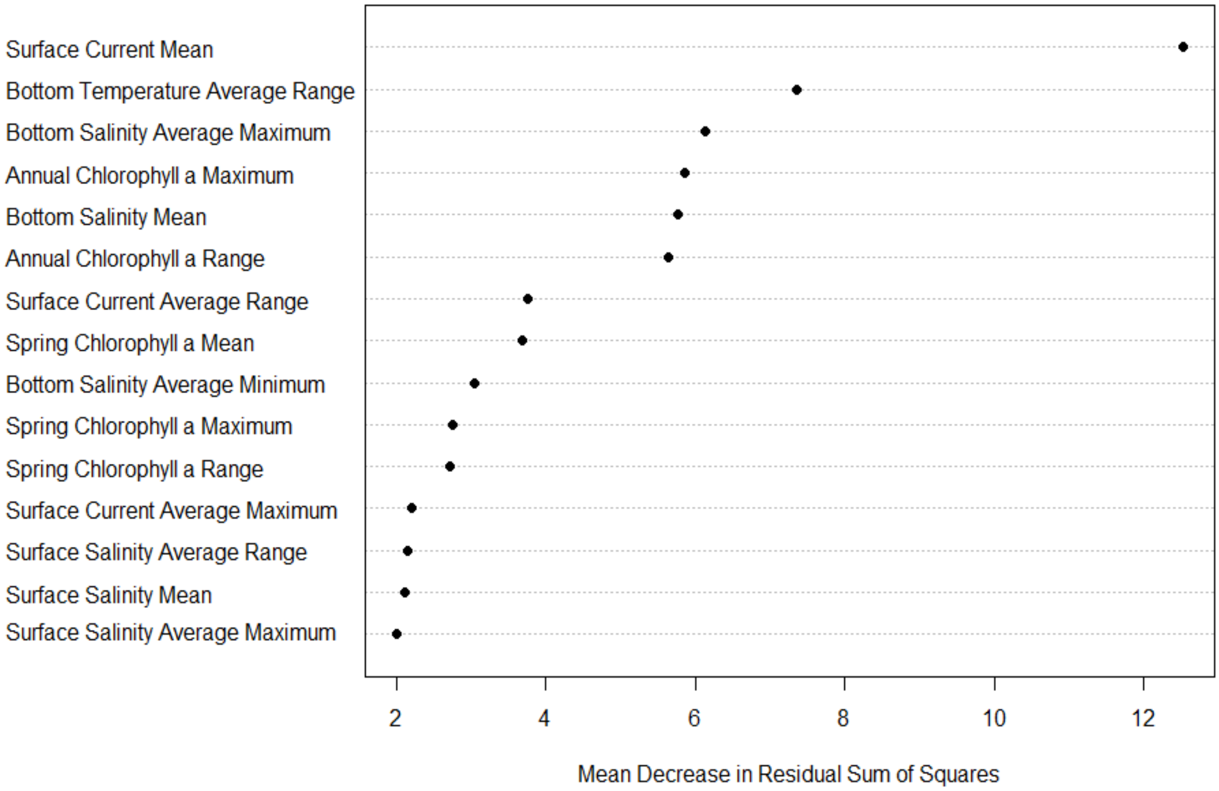

Figure 18. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on sponge mean biomass data collected in the Hudson Strait - Ungava Bay Region. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

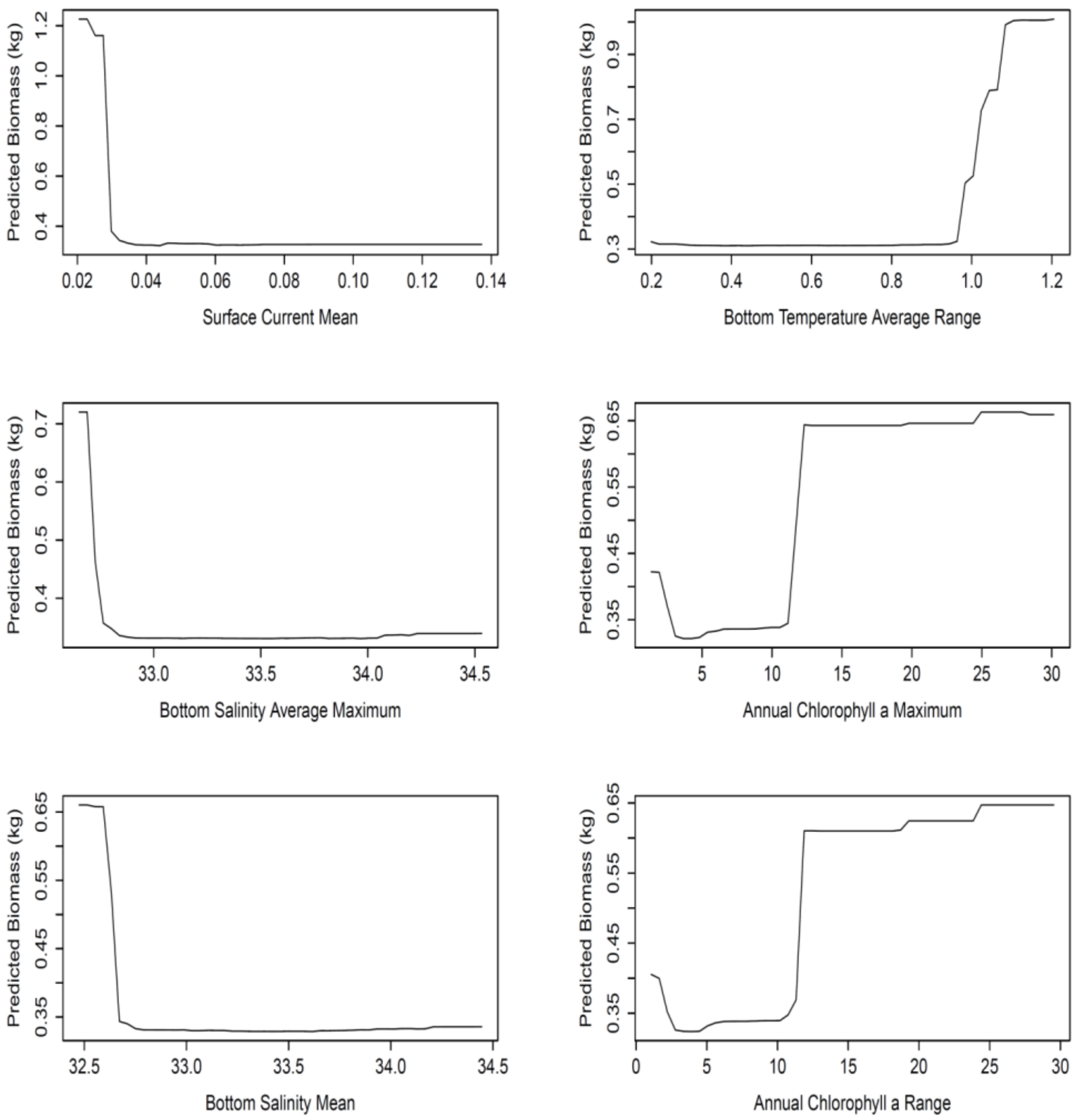

Figure 19. Partial dependence plots of the top six predictors from the random forest model of sponge biomass data collected within the Hudson Strait - Ungava Bay Region, ordered left to right from the top. Predicted biomass is shown on the *y*-axis.

Eastern Arctic Region

Sponges (Porifera)

Data Sources and Distribution

Sponge catch data for the Eastern Arctic Region were collected between 1996 and 2014 and consisted of 635 presences and 168 absences from *Paamuit* surveys conducted using Alfredo trawl gear, 665 presences and 775 absences from the *Cape Ballard*, *Aqviq*, and *Kinguk* surveys conducted using Campelen trawl gear, and 149 presences and 39 absences from *Paamuit* surveys using Cosmos gear (Table 7). DFO Alfredo trawl gear records had the widest spatial distribution in the study extent (Figure 20). Campelen records were restricted to the Davis Strait, while Cosmos records were distributed along the Baffin Island Shelf and in Davis Strait. Several i*n situ* benthic imagery records were distributed off Devon Island, in the Narwhal Over-wintering and Deep-Sea Coral Conservation Area in southern Baffin Bay, and in the Hatton Basin Voluntary Closure Area in Davis Strait. Presence-absence random forest models were generated on the combined dataset consisting of 1449 presences and 982 absences (see Figure 21). The highest mean biomass record (2000 kg) was recorded in the Davis Strait from Campelen gear.

Table 7. Number of presence and absence records of sponge catch by gear type from DFO trawl surveys conducted between 1996 and 2014 in the Eastern Arctic Region.

	Alfredo		**Campelen**		**Cosmos**	
Year	**Presences**	**Absences**	**Presences**	**Absences**	**Presences**	**Absences**
1996	-	-	2	-	-	-
1999	23	-	-	-	-	-
2000	48	-	-	-	-	-
2001	13	-	-	-	-	-
2005	-	-	47	103	-	-
2006	41	18	53	86	51	24
2007	-	-	63	66	11	2
2008	74	9	55	87	62	3
2009	-	1	31	115	-	-
2010	89	31	95	51	15	7
2011	78	6	75	74	-	-
2012	88	69	64	85	10	3
2013	77	8	73	74	-	-
2014	104	26	107	34	-	-
TOTAL	635	168	665	775	149	39

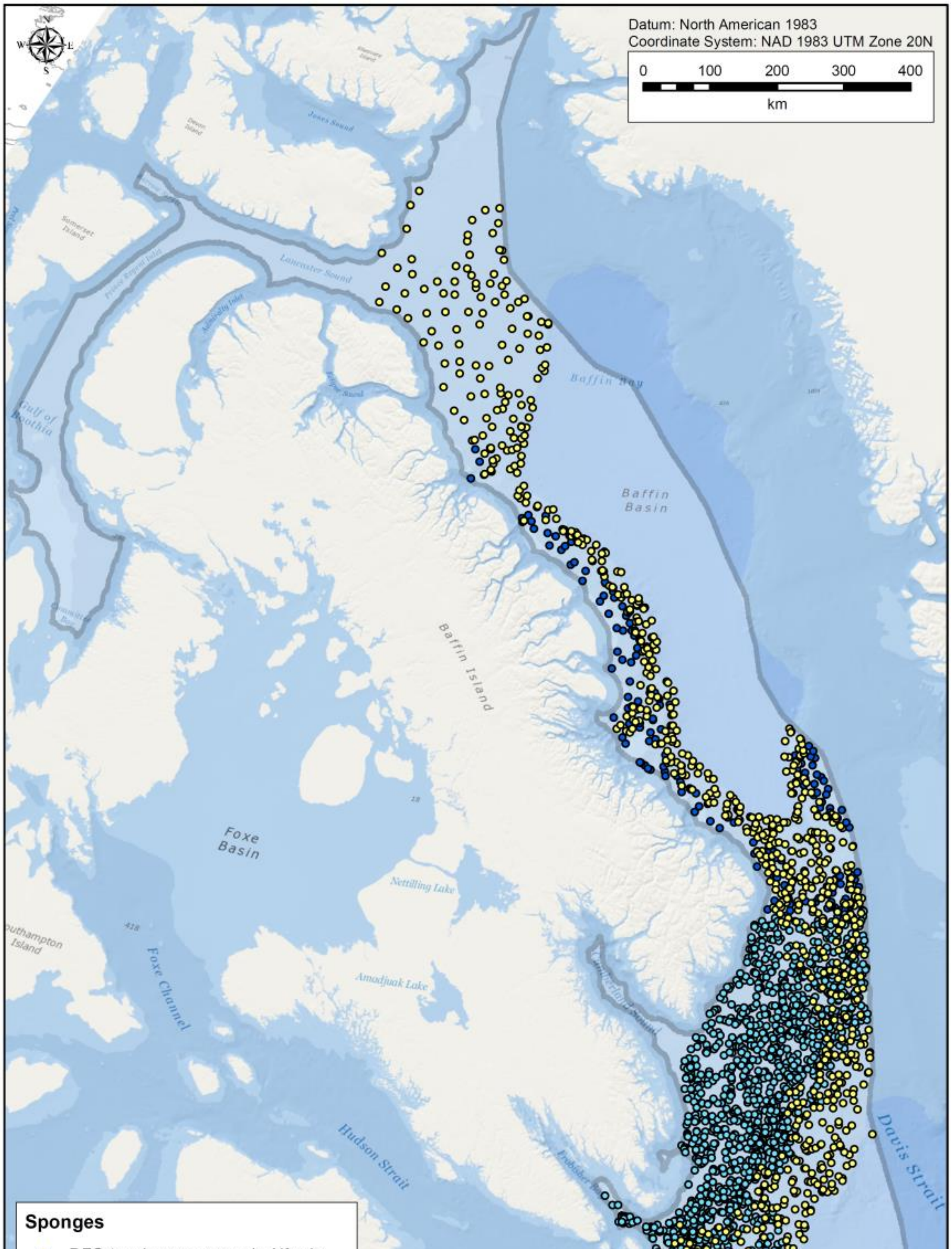

Figure 20. Available sponge presence and absence records in the Eastern Arctic from DFO trawl surveys using Alfredo, Campelen, and Cosmos trawl gear.

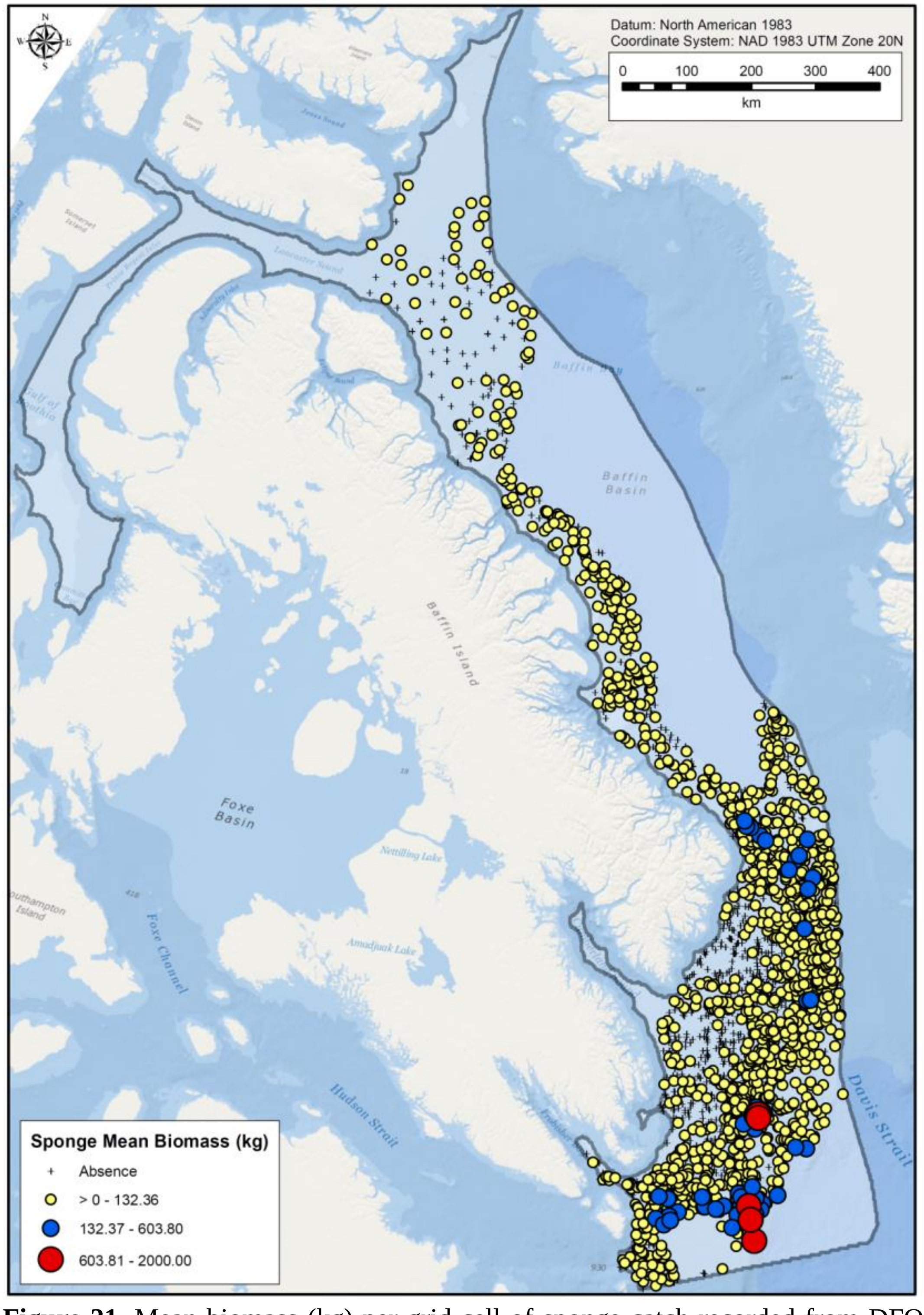

Figure 21. Mean biomass (kg) per grid cell of sponge catch recorded from DFO trawl surveys conducted in the Eastern Arctic between 1996 and 2014.

Model 1 – Balanced Species Prevalence

Accuracy measures for the random forest model on balanced species prevalence (982 presences and 982 absences; Model 1) are presented in Table 8. In this model, the presence data (1449) were randomly down-sampled to match the number of absences. The mean AUC of all model runs was 0.791 ± 0.005, indicating good model performance. The highest mean AUC of 0.799 was associated with Model Fold 5 and is therefore considered the optimal model for the prediction of the small gorgonian coral response data. The sensitivity and specificity measures of this model fold were 0.717 and 0.745, respectively. The confusion matrix of the optimal model is also presented in Table 8. Class error for both the presence and absence classes was moderate.

Table 8. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of sponges collected in the Eastern Arctic Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 5) which is considered the optimal model for predicting the presence probability of sponges in the region.

Model Run	AUC	Sensitivity	Specificity
1	0.792	0.722	0.731
2	0.793	0.729	0.724
3	0.783	0.723	0.726
4	0.791	0.727	0.725
5	**0.799**	**0.717**	**0.745**
6	0.794	0.724	0.734
7	0.791	0.725	0.740
8	0.787	0.713	0.734
9	0.784	0.707	0.715
10	0.797	0.733	0.731
Mean	**0.791**	**0.722**	**0.731**
SD	**0.005**	**0.008**	**0.009**

Confusion matrix of model with highest AUC:

Observations	**Predictions**		**Total n**	**Class error**
	Absence	**Presence**		
Absence	732	250	982	0.255
Presence	278	704	982	0.283

The presence probability prediction surface of sponges is presented in Figure 22. The highest predictions of sponge presence probability occurred on Baffin Island Shelf and in deeper water in the Davis Strait. Lancaster Sound, the Gulf of Boothia, and the deep waters off Baffin Island Shelf were predicted to have moderate presence probability of sponges. Areas of high and low presence probability of sponges corresponded well with the spatial distribution of presence and absence records, respectively (see Figure 23).

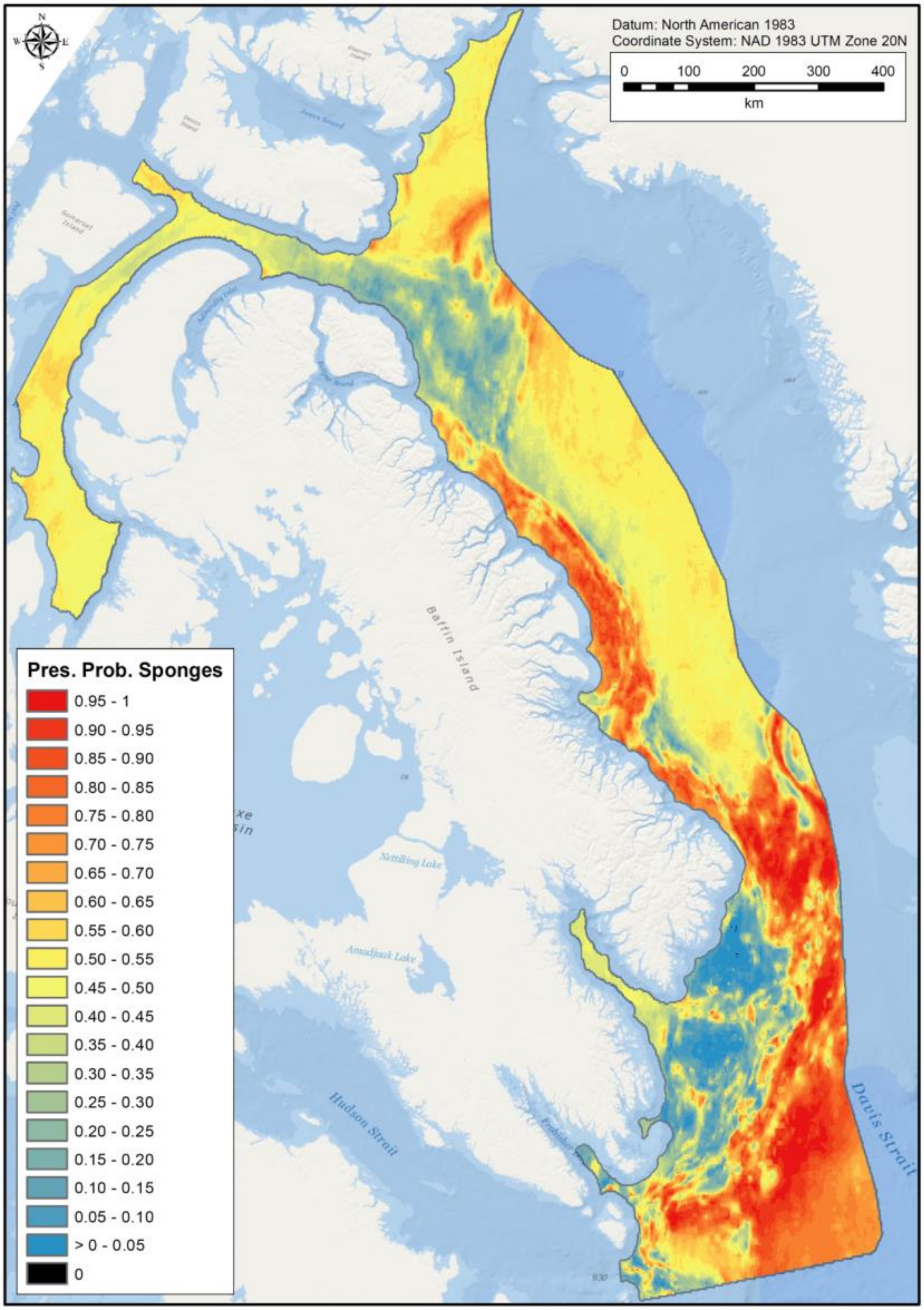

Figure 22. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of sponge presence and absence data collected from DFO trawl surveys conducted in the Eastern Arctic Region between 1996 and 2014.

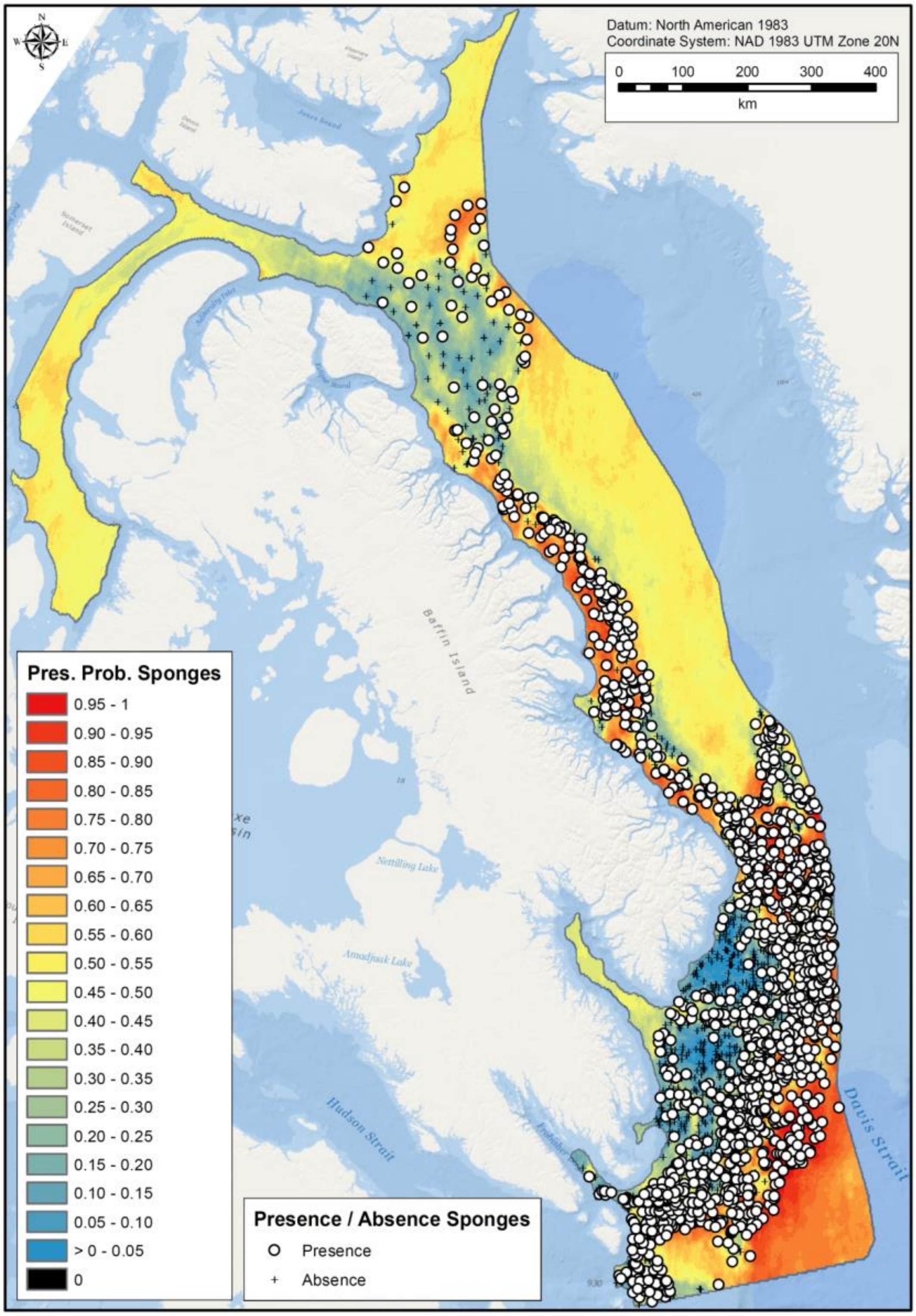

Figure 23. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of sponge presence and absence data collected from DFO trawl surveys in the Eastern Arctic Region between 1996 and 2014.

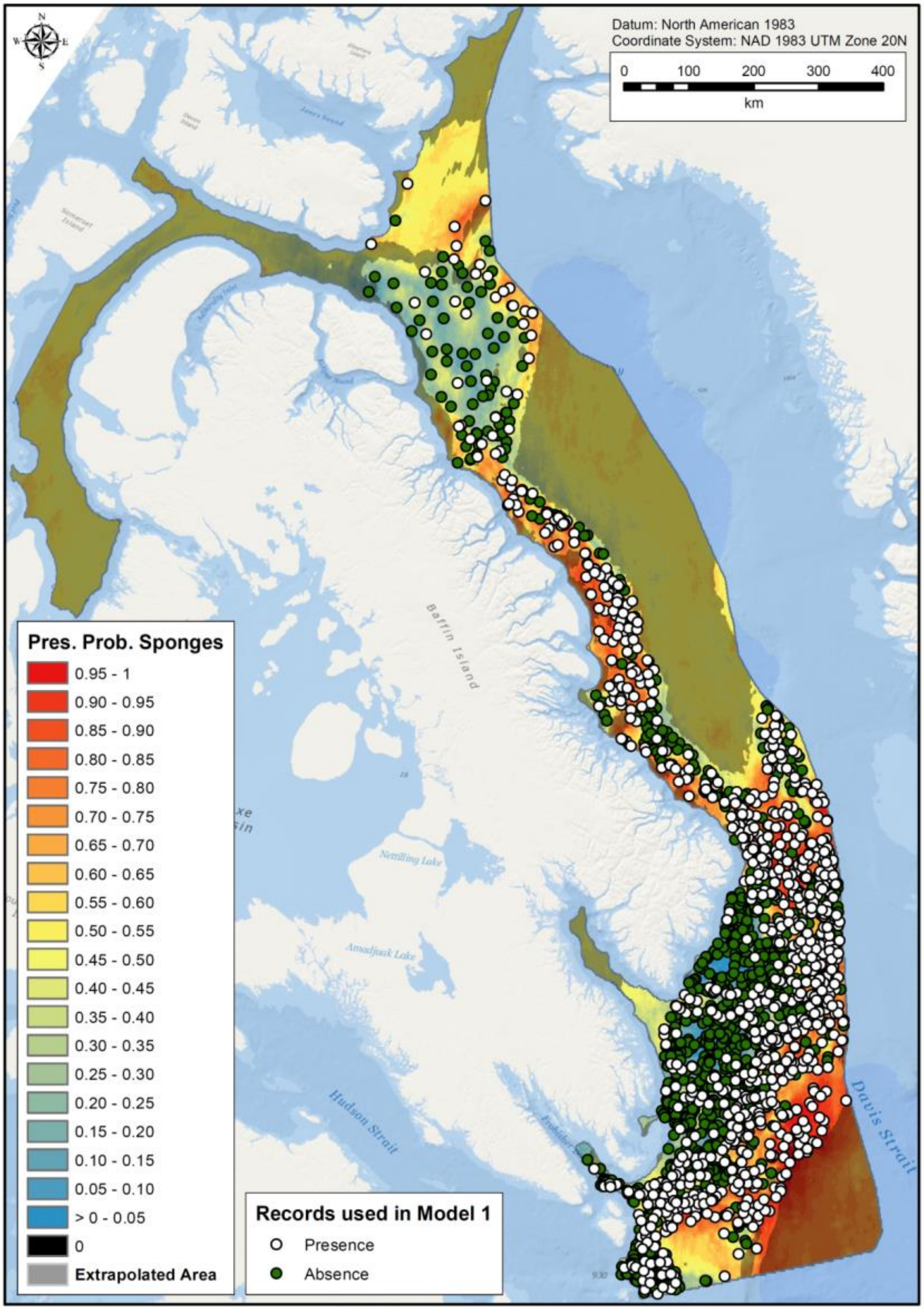

Figure 24. Map of the 1964 data observations (982 presences and 982 absences) of sponges used in the optimal random forest Model 1. Also shown is the predicted presence probability (Pres. Prob.) of sponges and the areas of model extrapolation.

Figure 24 shows the actual presence and absence data used in the optimal model fold of Model 1. There was little spatial bias in the presence records chosen through random down-sampling. Areas where there were no data observations appear to have been predicted with presence probabilities near 50%. Areas of model extrapolation are also shown in this figure. Lancaster Sound, the Gulf of Boothia, the deep waters off Baffin Bay Shelf, and the southeast corner of the spatial extent in Davis Strait was considered extrapolated area. Some coastal areas were also considered extrapolated by the model.

Of all 54 environmental variables used in the model, Depth (a non-interpolated variable) was the most important for the classification of the sponge presence and absence data (Figure 25). Depth was followed by Bottom Salinity Average Range and Bottom Salinity Average Minimum. Both bottom and surface variables ranked high in this model. Partial dependence plots for the top six predictor variables are shown in Figure 26. Presence probability of sponges increased beginning at 500 m and remained high, although it decreased slightly near the upper depth values. Along the gradient in Bottom Salinity Average Range presence probability was highest at ~0.05.

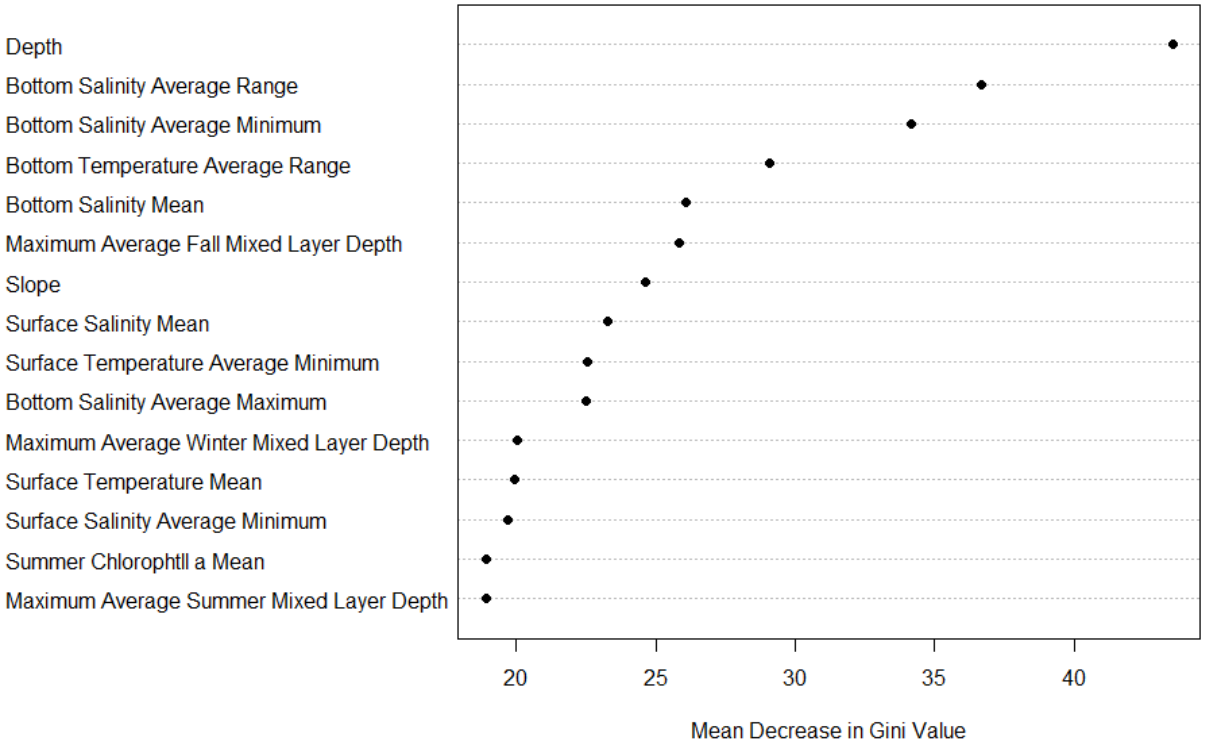

Figure 25. Importance of the top 15 predictor variables measured by the Mean Decrease in Gini Value of the optimal random forest model predicting sponge presence and absence data in the Eastern Arctic Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

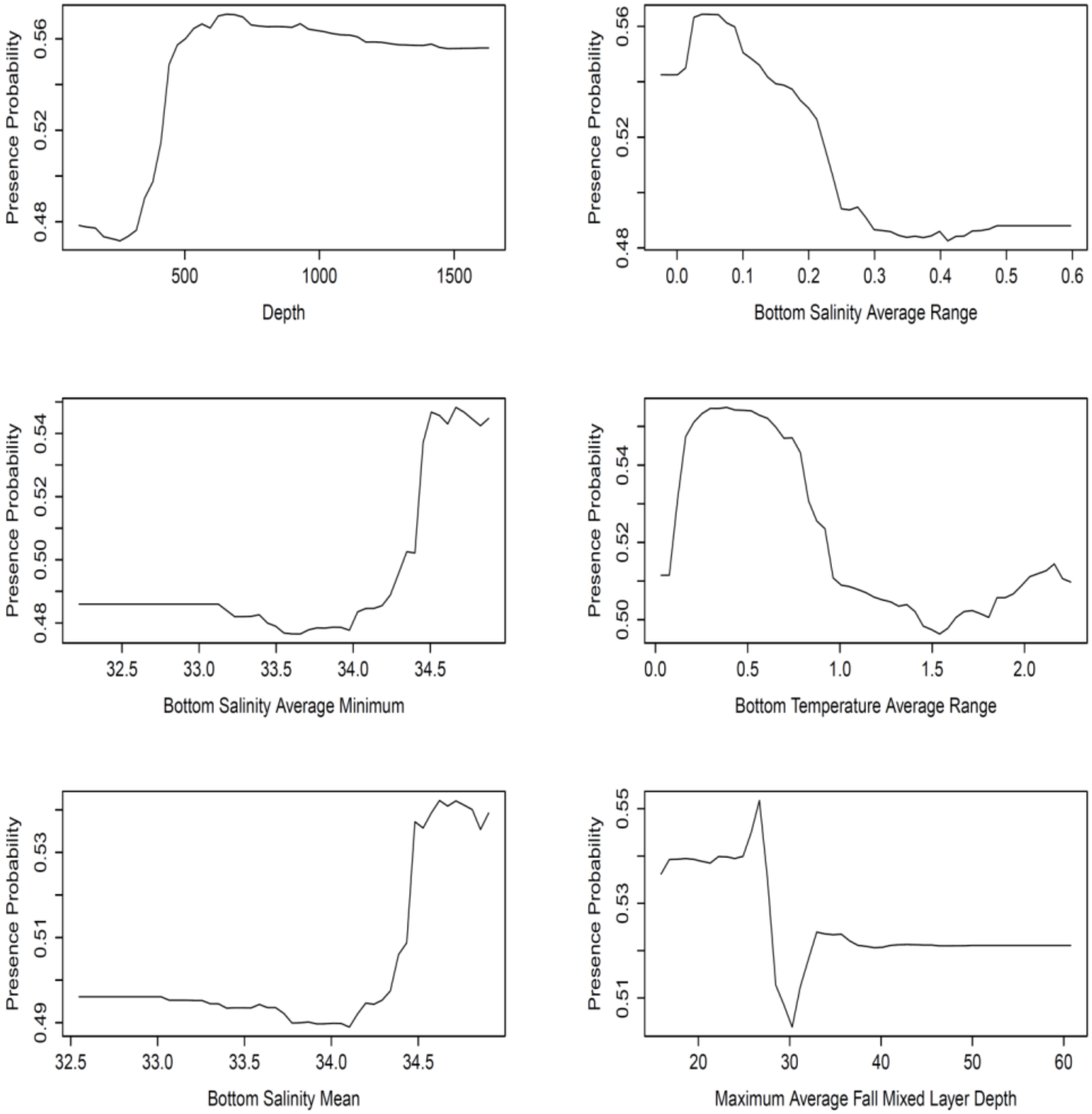

Figure 26. Partial dependence plots of the top six predictor variables from the optimal random forest model of sponge presence and absence data collected within the Eastern Arctic Region, ordered left to right from the top. Predicted presence probability is shown on the *y*-axis.

Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence

Table 9 shows the accuracy measures for the random forest model on all sponge presence and absence data (1449 presences 982 absences; Model 2) and a threshold equal to species prevalence (0.60). The average AUC calculated from this model was 0.791, only slightly higher than that of Model 1 (0.791). Sensitivity and specificity were 0.709 and 0.736, respectively, slightly lower than Model 1. Class error of both the presence and absence classes was slightly higher than Model 1.

The surface of predicted of presence probability of sponges generated from Model 2 is presented in Figure 27. Areas of high presence probability in Baffin Bay along the shelf and in Davis Strait are expanded in this model due to the increased number of presence points (see Figure 28). The southeast corner of the study extent in Davis Strait was predicted to have moderate to high sponge presence probability despite there being no data observations there (Figure 28). Figure 29 depicts the classification of sponge presence probability into presence and absence categories based on the prevalence threshold of 0.60. In this map, all presence probability values greater than 0.60 were classified as presence, while values below 0.60 were classed as absence. Much of the Davis Strait, southeast Baffin Bay and Baffin Island Shelf were predicted as presence of sponges. Areas of extrapolation are also shown in Figure 29. Lancaster Sound, the Gulf of Boothia, and the deeper waters off Baffin Island Shelf are considered extrapolated. The area of moderate to high predicted presence probability of sponges in the Davis Strait is also considered extrapolated area.

Table 9. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model presence and absence of sponges within the Eastern Arctic Biogeographic Region. Observe. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	**AUC**	**Observ.**	**Predictions**		**Total n**	**Class error**	**Sensit.**	**Specif.**
1	0.778		**Absence**	**Presence**				
2	0.777	**Absence**	723	259	982	0.264	0.709	0.736
3	0.806	**Presence**	421	1028	1449	0.291		
4	0.750							
5	0.811							
6	0.809							
7	0.742							
8	0.801							
9	0.804							
10	0.836							
Mean	**0.791**							
SD	**0.029**							

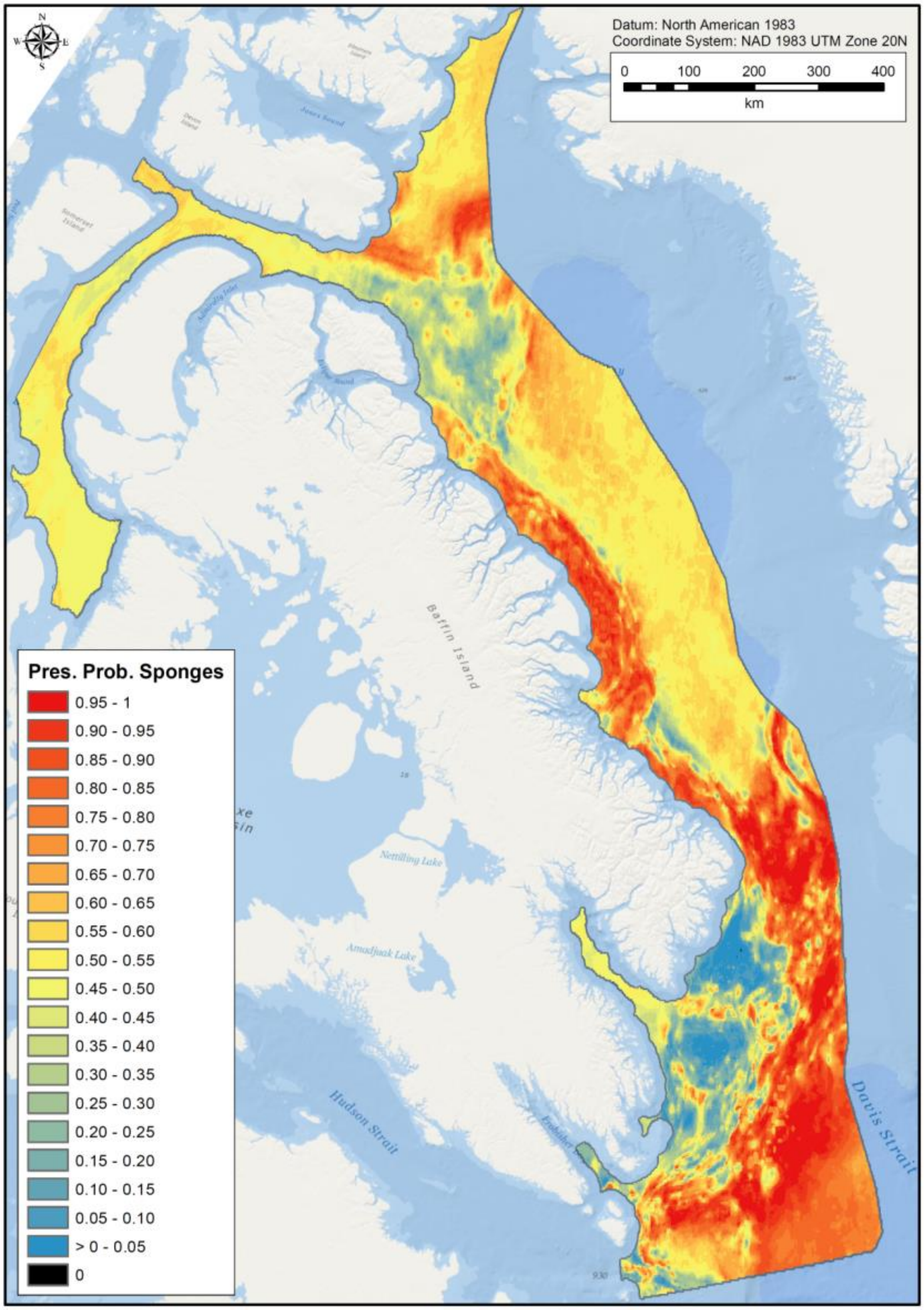

Figure 27. Prediction of presence probability (Pres. Prob.) of sponges based on a random forest model on unbalanced sponge presence and absence catch data collected from DFO trawl surveys conducted in the Eastern Arctic Region between 1996 and 2014.

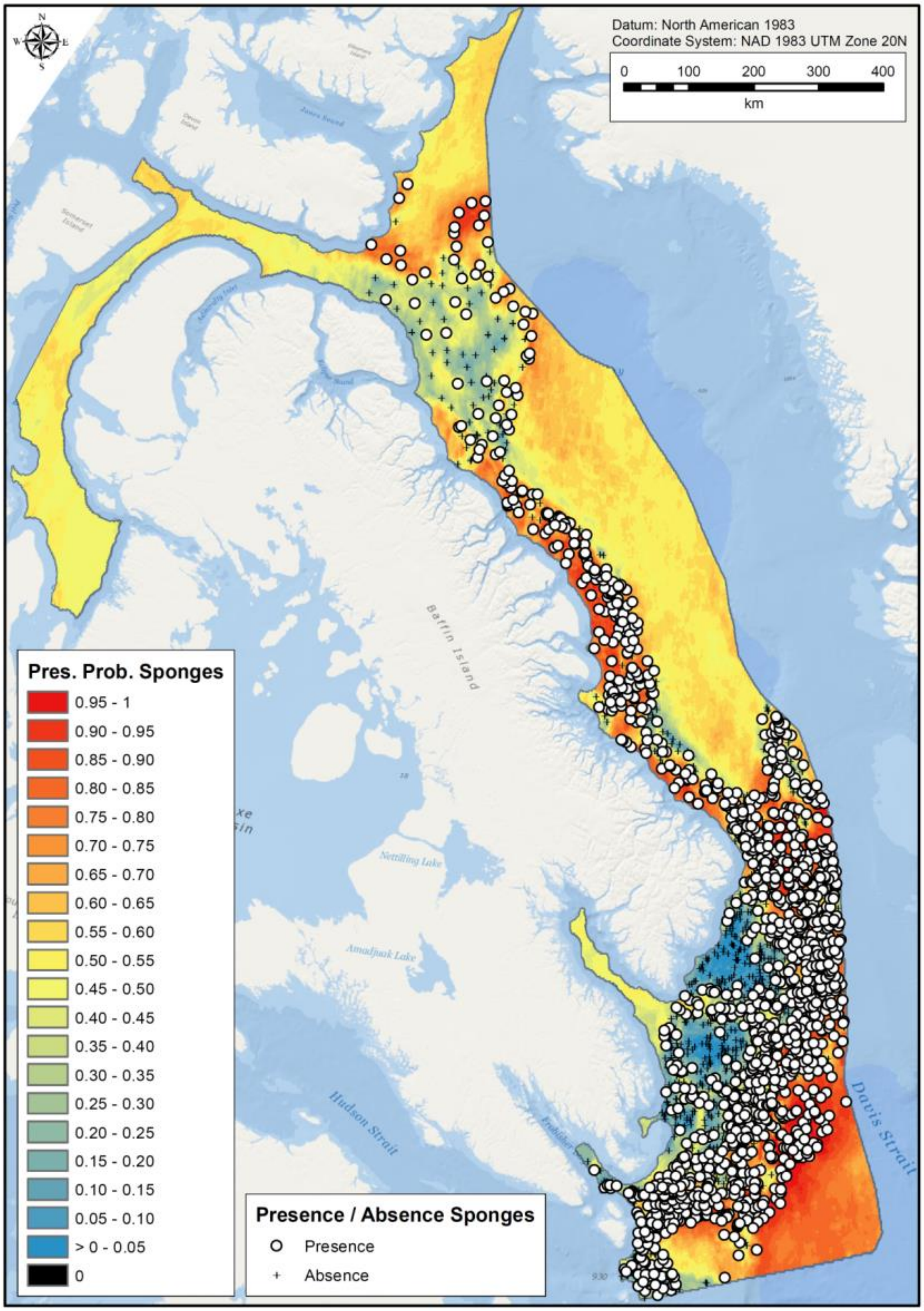

Figure 28. Presence and absence observations and prediction of presence probability (Pres. Prob.) of sponges based on a random forest model on unbalanced sponge presence and absence catch data collected from DFO trawl surveys conducted in the Eastern Arctic Region between 1996 and 2014.

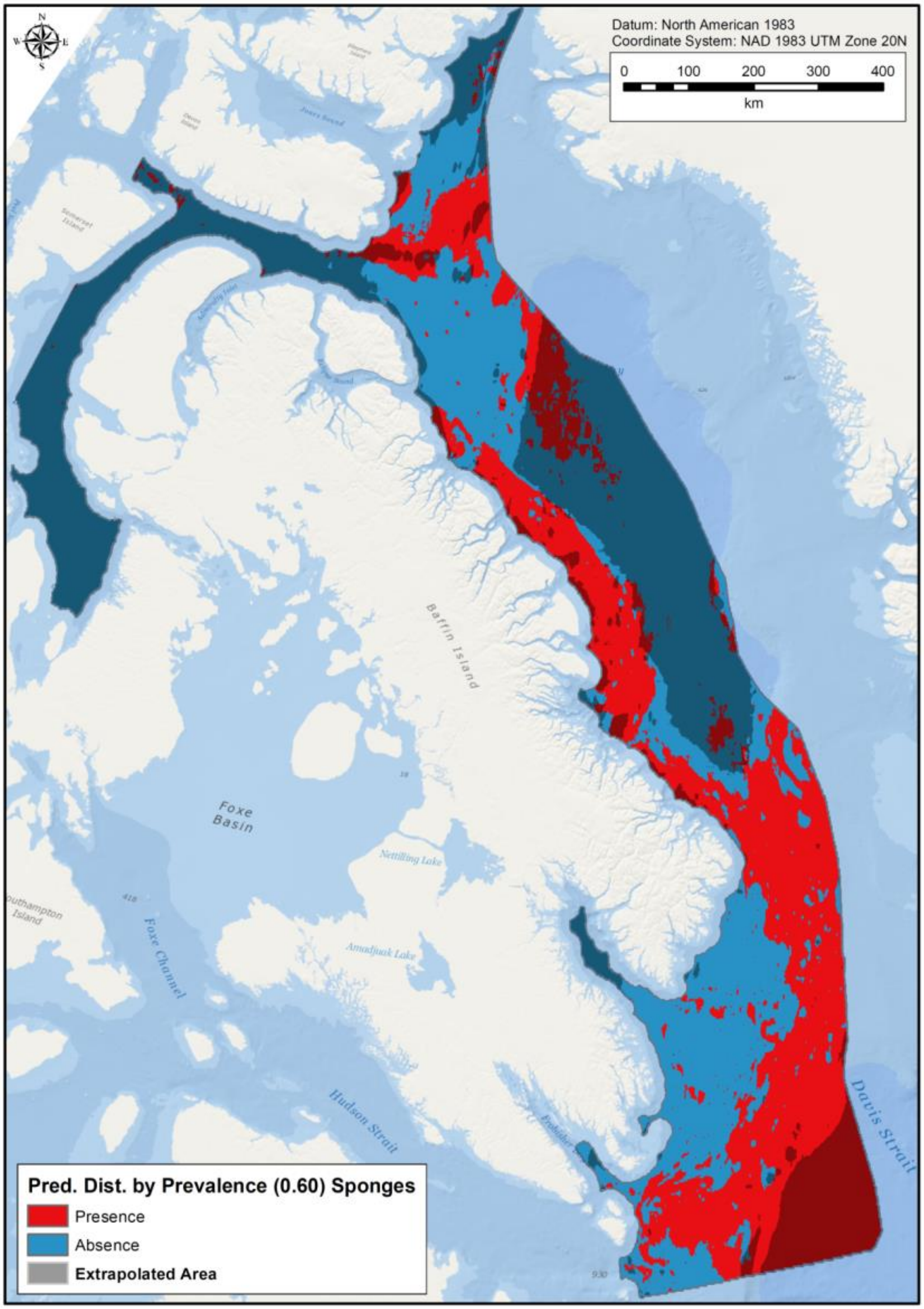

Figure 29. Predicted distribution (Pred. Dist.) of sponges in the Eastern Arctic Region based on the prevalence threshold of 0.60 of sponge presence and absence data used in Model 2. Also shown are the areas of model extrapolation (grey polygon may appear red or blue).

The order of importance of the environmental predictor variables in Model 2 (Figure 30) was slightly different from that of Model 1. Depth was the most important variable in Model 2, and was followed by Bottom Salinity Average Range and Bottom Salinity Average Minimum. Partial dependence plots for the top six predictor variables are shown in Figure 31. Like in Model 1, presence probability was highest between 500 and 1500 m depth, although the decrease at higher values was more prominent in this model.

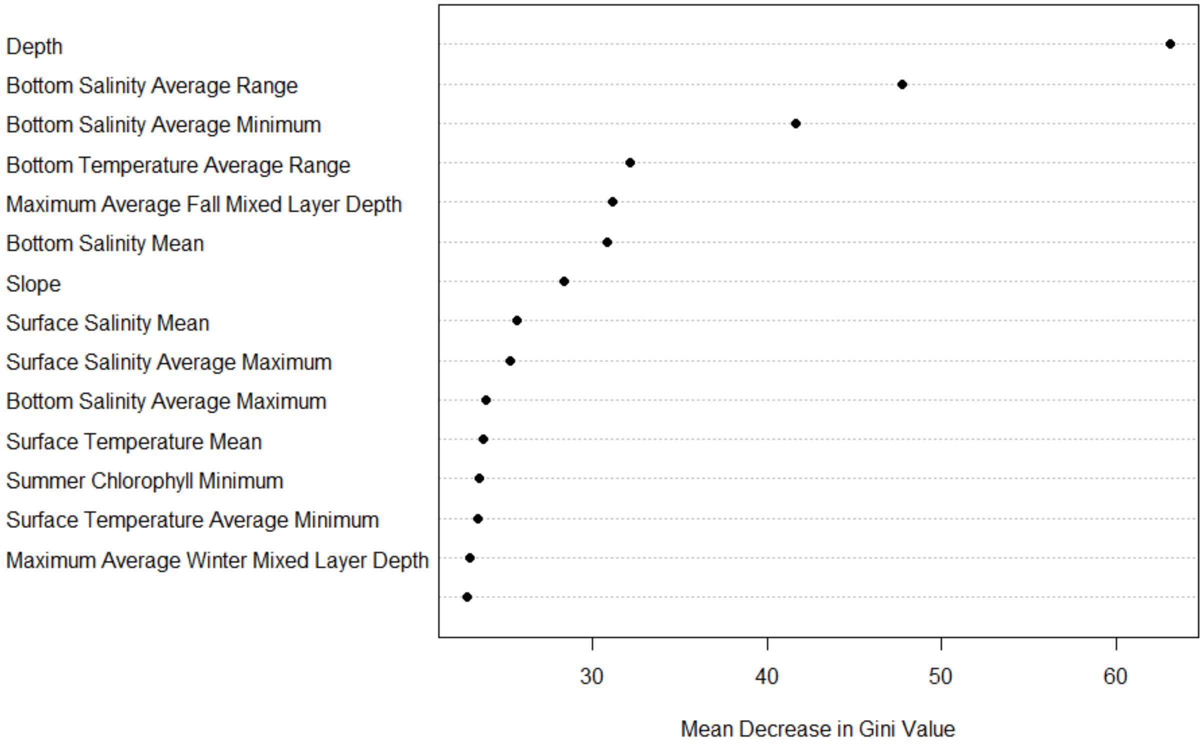

Figure 30. Importance of the top 15 predictor variables measured by the Mean Decrease in Gini Value of the random forest model on unbalanced sponge presence and absence data in the Eastern Arctic Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

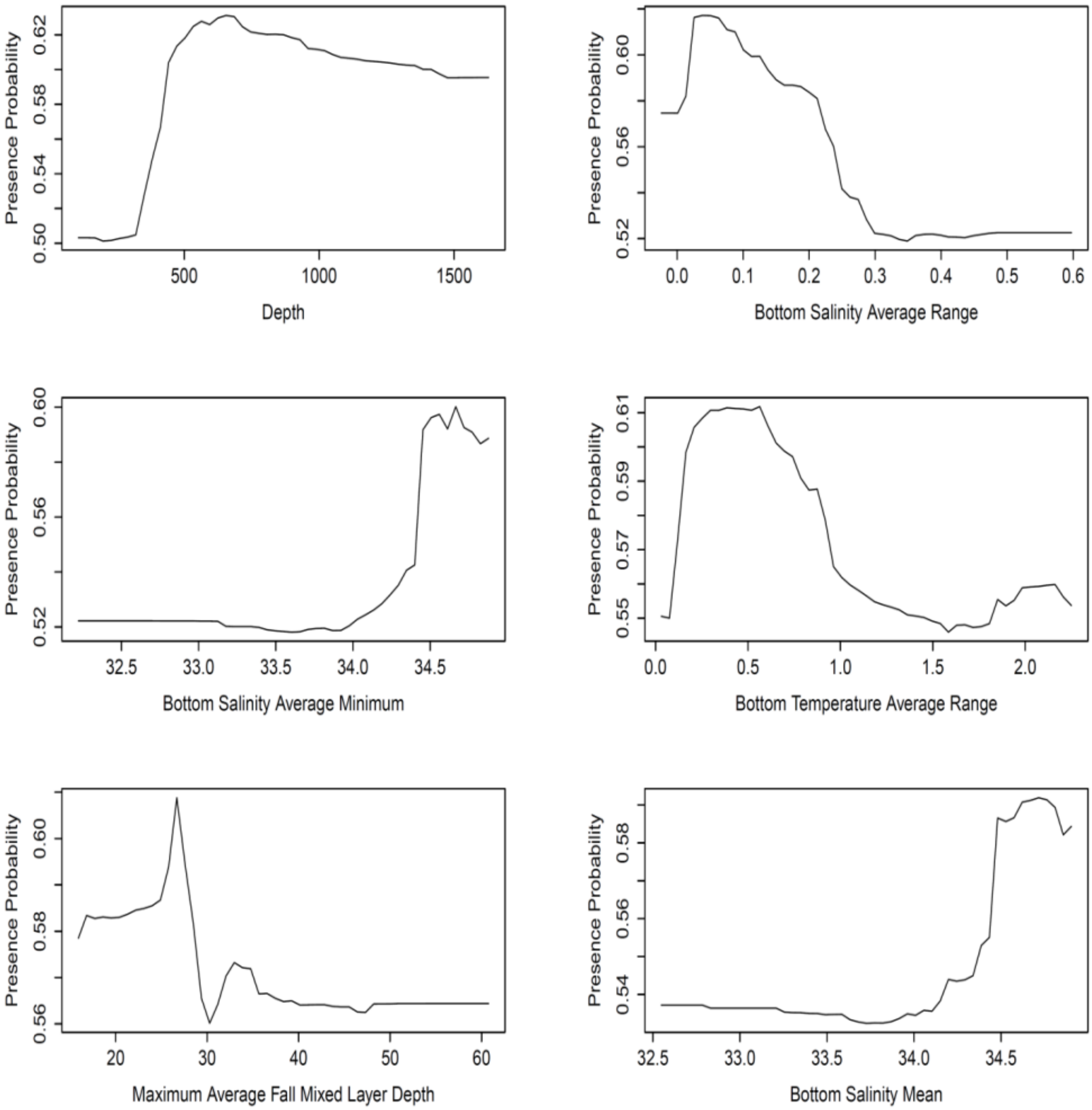

Figure 31. Partial dependence plots of the top six predictors from the random forest model of sponge unbalanced presence and absence data from the Eastern Arctic Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using all the available sponge records and an unbalanced species prevalence (Model 2) was selected as the best predictor of sponge distribution in the Eastern

Arctic Region. In this dataset, the number of presences was higher than the number of absences. Although accuracy measures were comparable between both Models 1 (balanced dataset) and 2, Model 2 allowed for the use of all presence data in the region, providing a more accurate depiction of the distribution of sponges. Sponge absence is predicted in shelf areas where there are major inlets (Figure 29) and in the case of the Cumberland Sound, in an area where a polynya forms (DFO, 2015).

Validation of Selected Model Using Independent Data

Figure 32 shows the predicted presence probability of sponges generated from Model 2 at the location of sponge records from benthic camera observations collected during DFO scientific missions to the Eastern Arctic in 2012 and 2013. Overall there was good spatial congruence between the location of sponge records from the *in situ* surveys and areas of high presence probability predicted by the model. In 2012, several sponges were recorded southeast of Devon Island where the model predicted a relatively high presence probability of sponges. In the two closure areas, sponges were located in areas where the model predicted relatively high presence probability of sponges. Of the 46 sponge records from the 2012 survey, 41 (89%) were predicted as presence based on the prevalence threshold of 0.60. The records predicted as absence by the model were located in the Narwhal Closure.

In 2013, several sponges were recorded in relatively shallow water on Baffin Island Shelf in an area where the model predicted moderate to high presence probability of sponge. A single sponge record was located in the Narwhal Closure area. Several sponges were observed in the Hatton Basin closure in an area where the model predicted moderate to high presence probability of sponges. Of the 18 sponge records from 2013, 17 (94.4%) were predicted as presence based on the prevalence threshold. The single sponge record predicted as absence by the model was located in the Hatton Basin closure area. The positive occurrence of sponges there suggests that this is suitable sponge habitat.

There was good spatial congruence between the location of the FOP data and areas of high presence probability of sponges (Figure 33). Of the 2238 sponge records from 1979 to 2001 (Figure 33, left), 1446 (65%) were predicted as presence by the model. The records predicted as absence were located mainly off the southern tip of Baffin Island and in the deep waters of Baffin Bay. Of the 4029 records collected between 1998 and 2013 (Figure 33, right), 1320 (33%) were predicted as presence by the model, with absences located in the northern portion of the study extent and off the southern tip of Baffin Island. Commercial trawls are often very long compared with the research vessel trawls of approximately 1 km. They can be 10s of km and the exact location of the sponge catch along the trawl track is unknown. For this reason the utility of this type of data to validate the models in areas where there is fine-scale heterogeneity in presence probability is limited. The expectation is for more mismatches arising from presences recorded where the start position indicates an absence, due to the potential for transit over presence areas during the tows. Validation could be improved if the actual trawl track were available (VMS data), however our data only included the start and end positions for each tow.

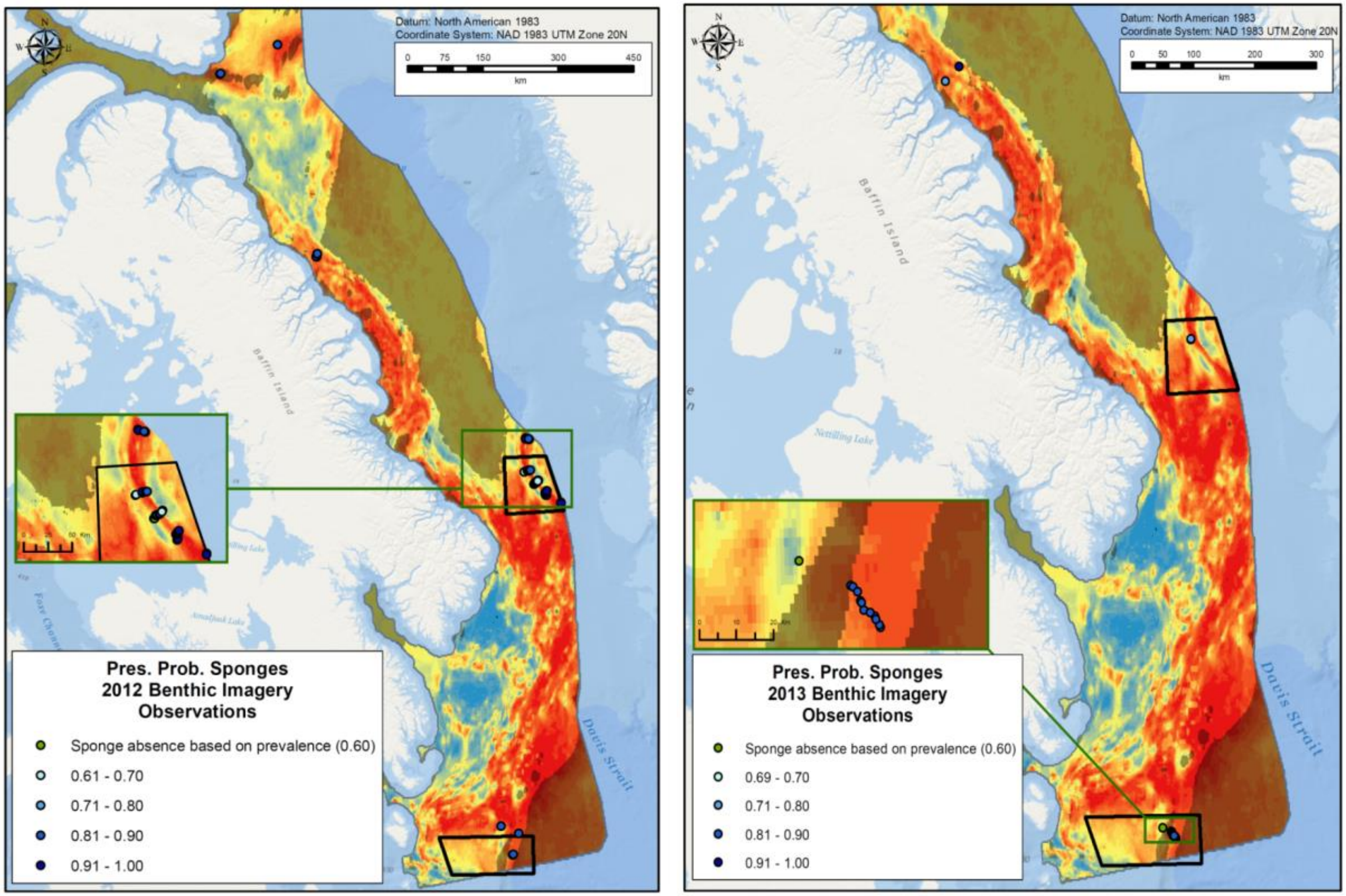

Figure 32. Validation of sponge presence probability from Model 2 using *in situ* camera records of sponges collected during DFO scientific missions conducted in 2012 (left) and 2013 (right). Also shown are the Narwhal Overwintering and Deep-Sea Coral Conservation Area and the Hatton Basin Voluntary Closure Area. Inset maps show the Narwhal (left) and Hatton Basin (left) closures.

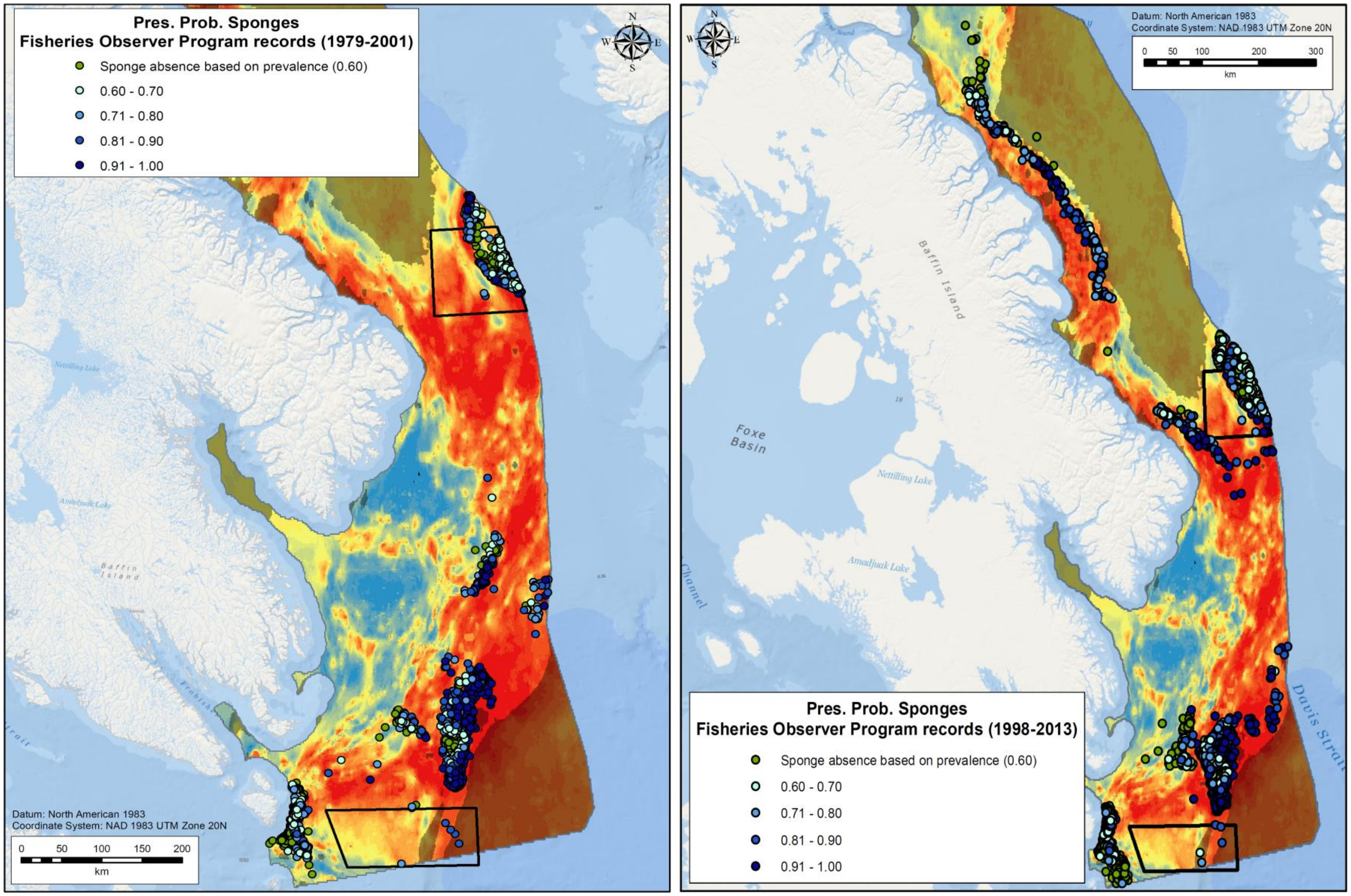
Pres. Prob. Sponges
Fisheries Observer Program records (1979-2001)
● Sponge absence based on prevalence (0.60)
○ 0.60 - 0.70
● 0.71 - 0.80
● 0.81 - 0.90
● 0.91 - 1.00
N
W
E
S
Nettilling Lake
Amadjuak Lake
Baffin Island
Datum: North American 1983
Coordinate System: NAD 1983 UTM Zone 20N
0 50 100 150 200
km
Datum: North American 1983
Coordinate System: NAD 1983 UTM Zone 20N
0 50 100 200 300
km
Baffin Island
Foxe Basin
Nettilling Lake
Amadjuak Lake
Channel
Davis Strait
Pres. Prob. Sponges
Fisheries Observer Program records (1998-2013)
● Sponge absence based on prevalence (0.60)
○ 0.60 - 0.70
● 0.71 - 0.80
● 0.81 - 0.90
● 0.91 - 1.00

Figure 33. Validation of sponge presence probability from Model 2 using sponge records collected by the Fisheries Observer Program between 1979 – 2001 (left) and 1998 – 2013 (right). Also shown are the Narwhal Overwintering and Deep-Sea Coral Conservation Area and the Hatton Basin Voluntary Closure Area.

Prediction of Sponge Biomass Using Random Forest

Alfredo Trawl Gear

The accuracy measures of the regression random forest model on mean sponge biomass per grid cell from Alfredo trawl records are presented in Table 10. This model performed relatively well, and a mean R^2 value of 0.327 ± 0.242 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.042 ± 0.027 SD. This model explained a somewhat moderate percentage of variance in the biomass data (average = 15.44% ± 6.98 SD).

Figures 34 and 35 show the predicted biomass surface of sponges. The majority of the spatial extent was predicted to have low (> 0 to 6.46 kg) sponge biomass, even in areas where low (up to 69.30 catches were recorded (Figure 34). The highest predicted sponge biomass occurred in isolated areas of the Davis Strait, and corresponded to large catches there. Higher biomass values were predicted to occur northeast of Devon Island in northern Baffin Bay. The southeast corner in the Davis Strait was predicted to have moderate to high sponge biomass despite there being no data observations there. This area is considered an area of extrapolation (Figure 35).

Table 10. Accuracy measures for all 10 model repetitions of 10-fold cross validation of random forest model of average of sponge biomass (kg) per grid cell recorded from DFO trawl surveys conducted using Alfredo trawl gear the Eastern Arctic Region. NRMSE= Normalized Root-Mean-Square Error, RMSE= Root-Mean-Square Error.

Model Fold	**R^2**	**RMSE**	**NRMSE**	**Percent (%) variance explained**
1	0.093	117.856	0.108	33.16
2	0.639	20.453	0.019	12.31
3	0.132	34.595	0.032	14.18
4	0.371	27.047	0.025	13.95
5	0.157	42.140	0.039	15.37
6	0.606	17.139	0.016	14.83
7	0.464	50.791	0.047	7.77
8	0.628	65.8267	0.060	9.37
9	0.132	40.209	0.037	14.29
10	0.044	38.527	0.035	19.20
Mean	**0.327**	**45.458**	**0.042**	**15.44**
SD	**0.242**	**29.166**	**0.027**	**6.98**

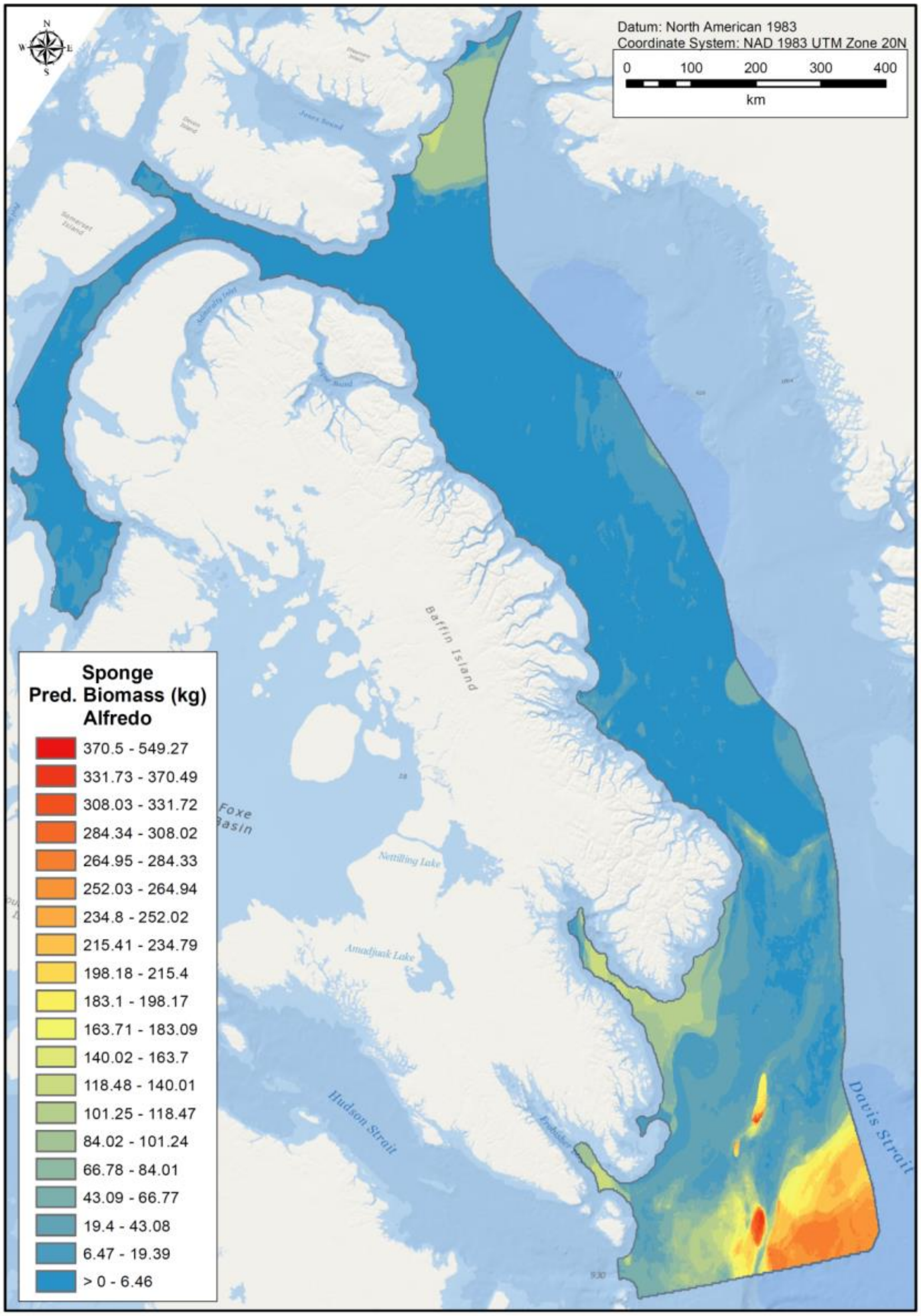

Figure 34. Predictions of biomass (kg) of sponges from catch data recorded in DFO trawl surveys conducted using Alfredo trawl gear in the Eastern Arctic Region between 1999 and 2014.

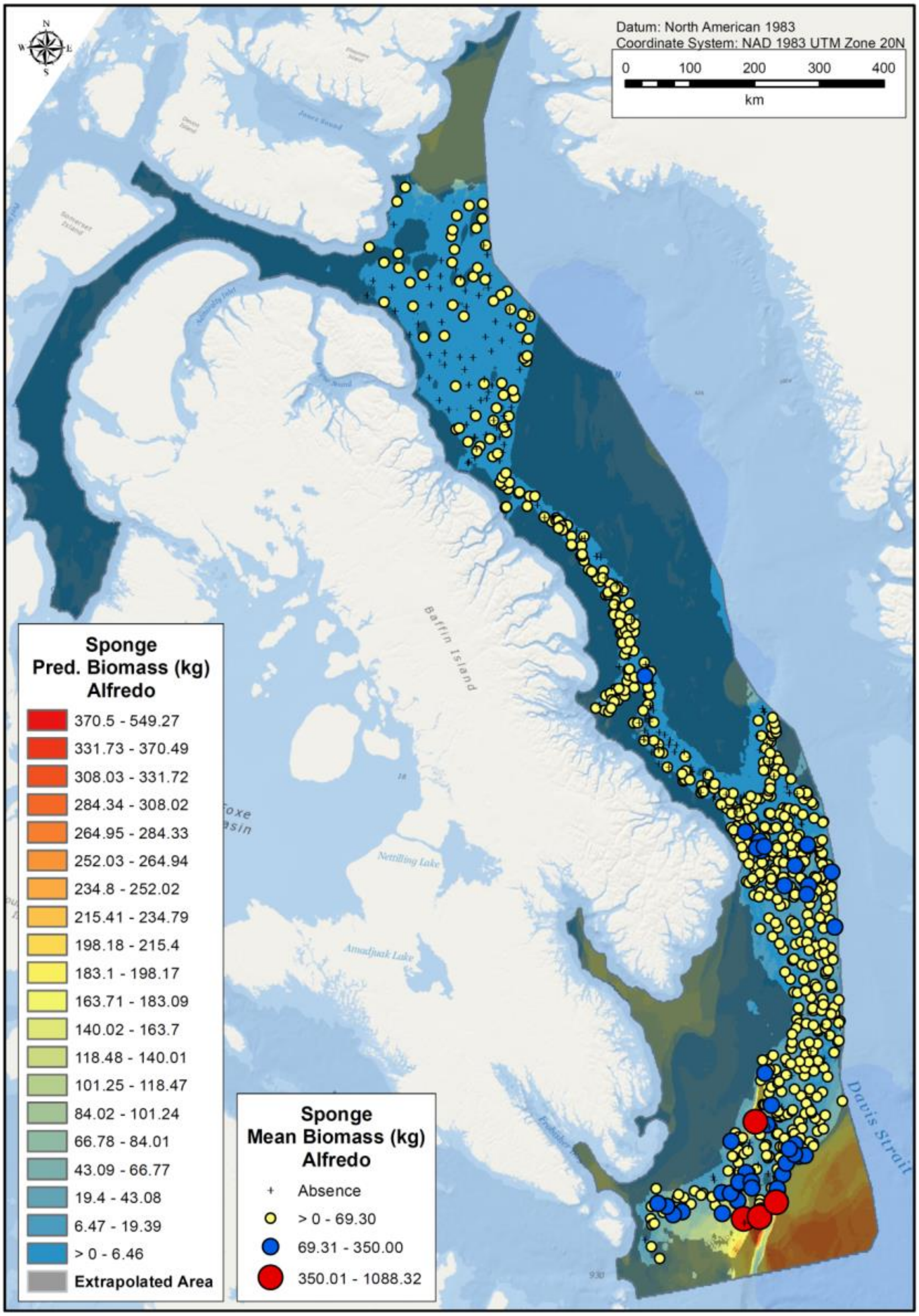

Figure 35. Predictions of biomass (kg) of sponges from catch data recorded in DFO trawl surveys conducted using Alfredo trawl gear in the Eastern Arctic Region between 1999 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting sponge biomass are shown in Figure 36. Unlike the sponge presence-absence models, bottom temperature variables were most important for the prediction of sponge biomass in the Eastern Arctic. Bottom Temperature Average Maximum was the most important variable, followed by Bottom Temperature Average Minimum, Bottom Temperature Mean, and the remaining variables in the model. The partial dependence plots of the top six environmental predictor variables are shown in Figure 37. In general, predicted biomass was highest at bottom temperature values greater than 3ºC.

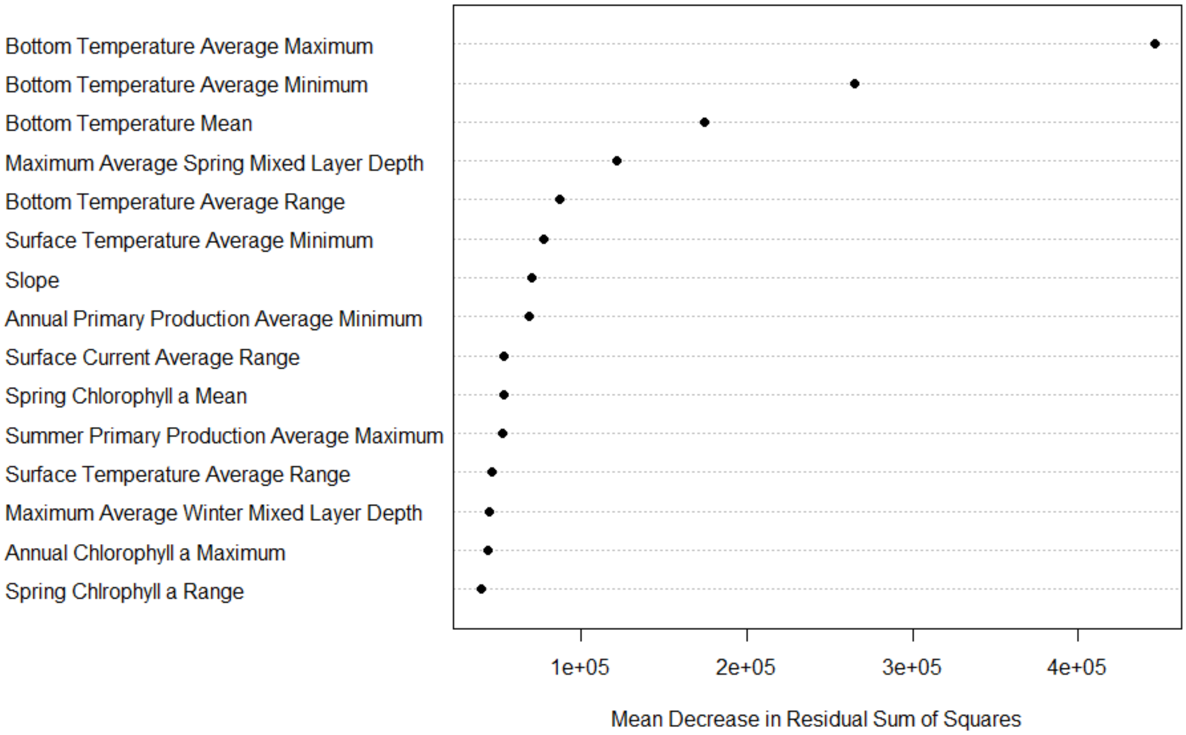

Figure 36. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on sponge biomass data collected from DFO trawl surveys conducted using Alfredo trawl gear. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

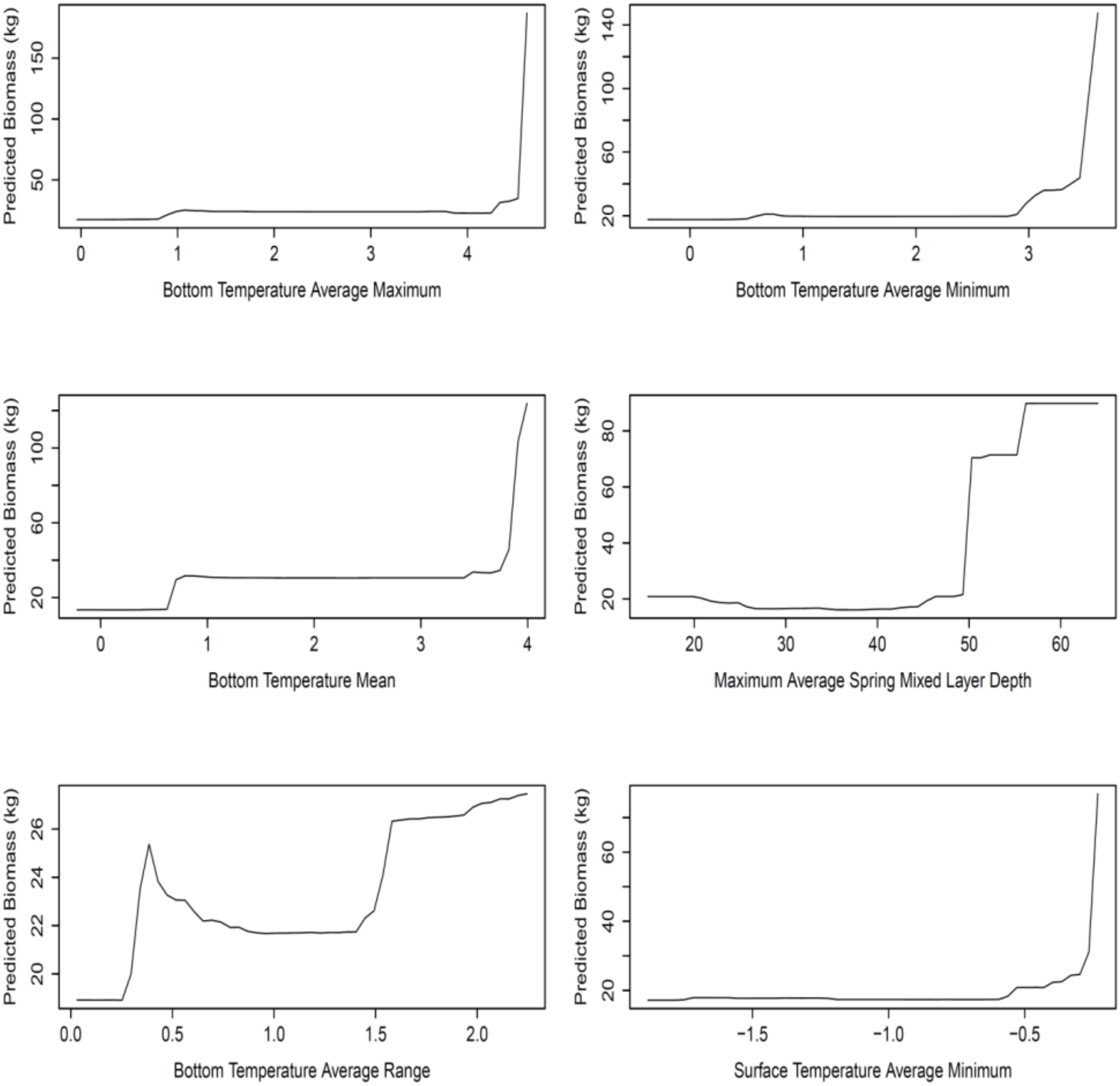

Figure 37. Partial dependence plots of the top six predictors from the random forest model of sponge biomass data collected from DFO trawl surveys using Alfredo trawl gear in the Eastern Arctic Region, ordered from left to right from the top.

Campelen Trawl Gear

Accuracy measures from the regression random forest model on mean sponge biomass records from trawl surveys using Campelen trawl gear are presented in Table 11. The highest R^2 value was 0.803 while the average was 0.480 ± 0.174 SD, indicating good model performance. The average Normalized Root-Mean-Square-Error (NRMSE) was 0.032 ± 0.018 SD. The average percent variance explained by the model was 31.91 ± 4.82 SD.

Figures 38 and 39 show the predicted biomass surface of sponges using mean biomass data from Campelen trawl surveys. Most of Baffin Bay, Lancaster Sound, and the Gulf of Boothia where there are no data observations were predicted to have relatively low sponge biomass (<164.88 kg). These areas were considered areas of extrapolation by the model (Figure 39). However, a large area of the Davis Strait where presence and absence observation data are concentrated was predicted to have the lowest biomass of the entire region, likely due to the inclusion of zero catches (absences) from that area. Predicted biomass was highest in the southeast corner of the study extent in Davis Strait. This area was also considered extrapolated area by the model.

Table 11. Accuracy measures for all 10 model repetitions of 10-fold cross validation of the random forest model of average of sponge biomass (kg) per grid cell recorded from DFO trawl surveys conducted using Campelen trawl gear the Eastern Arctic Region. RMSE= Root-Mean-Square Error, NRMSE= Normalized Root-Mean-Square Error.

Model Fold	**R^2**	**RMSE**	**NRMSE**	**Percent (%) variance explained**
1	0.393	38.978	0.019	35.08
2	0.643	130.925	0.065	33.39
3	0.494	25.931	0.013	32.88
4	0.555	89.780	0.045	26.12
5	0.214	76.343	0.038	38.84
6	0.803	31.098	0.016	26.30
7	0.323	52.439	0.026	33.18
8	0.364	76.108	0.038	34.04
9	0.604	34.705	0.017	35.40
10	0.407	110.839	0.055	23.89
Mean	**0.480**	**66.715**	**0.032**	**31.91**
SD	**0.174**	**36.075**	**0.018**	**4.82**

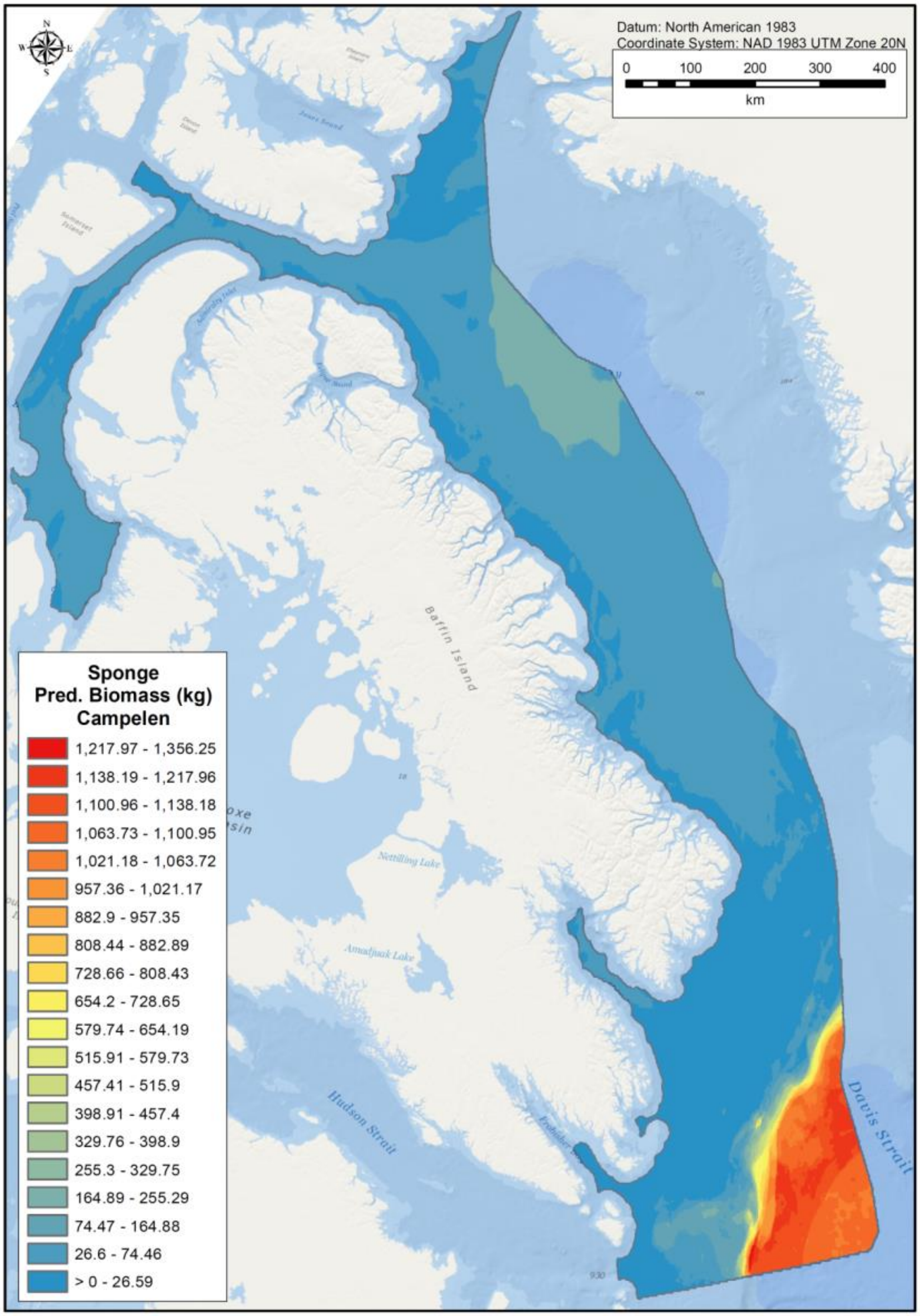

Figure 38. Predictions of biomass (kg) of sponges from catch data recorded in DFO multispecies trawl surveys conducted using Campelen trawl gear in the Eastern Arctic Region between 1996 and 2014.

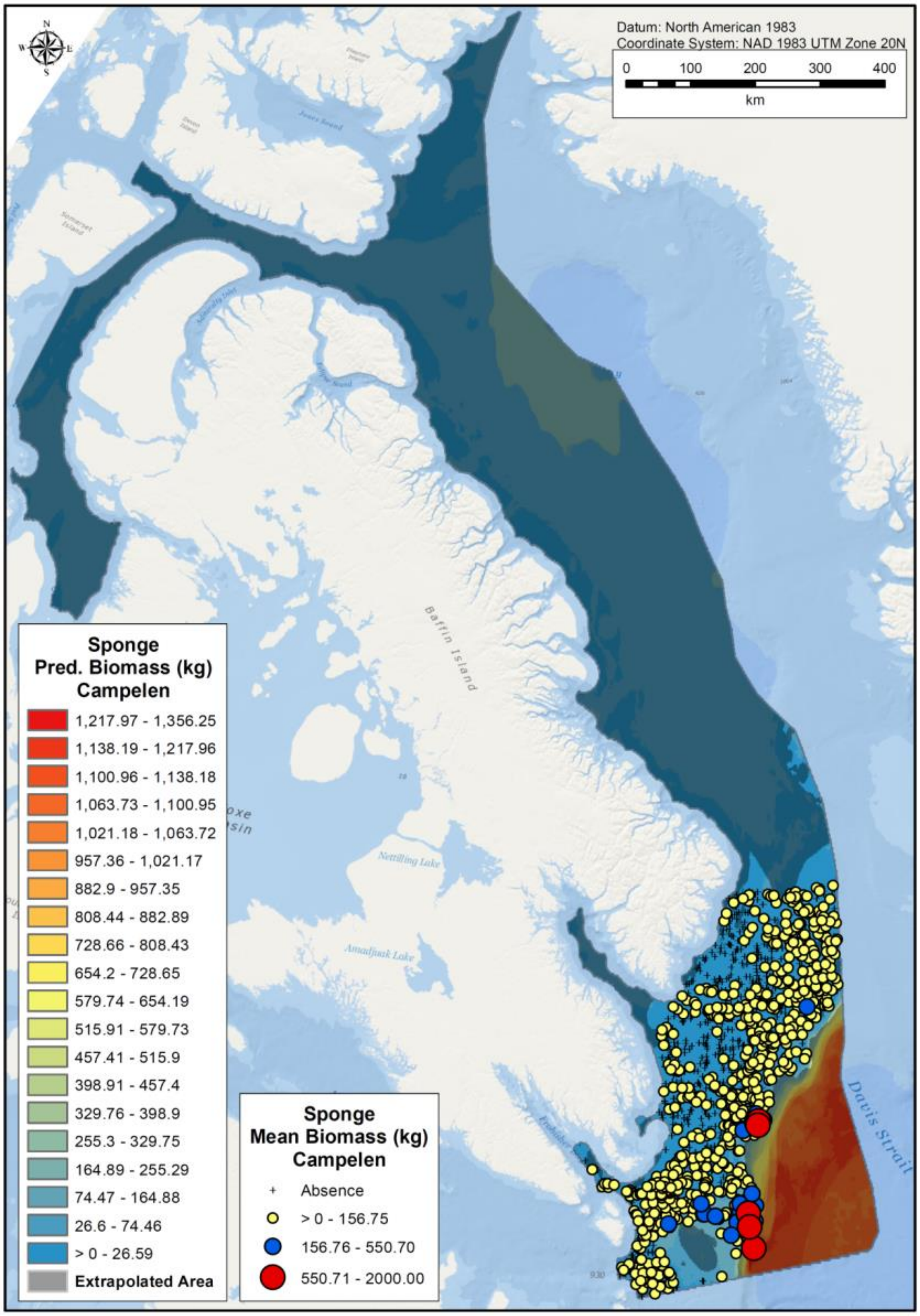

Figure 39. Predictions of biomass (kg) of sponges from catch data recorded in DFO multispecies trawl surveys conducted using Campelen trawl gear in the Eastern Arctic Region between 1996 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting sponge biomass are shown in Figure 40. In this model, Surface Salinity Average Minimum was the most important variable, followed closely by Bottom Temperature Average Minimum. Surface variables ranked higher in this model compared to the previous model using sponge records from Alfredo trawl gear. Partial dependence plots of the top six environmental predictor variables are shown in Figure 41. Predicted biomass was highest at Surface Salinity Average Minimum values between 31.5 and 32.

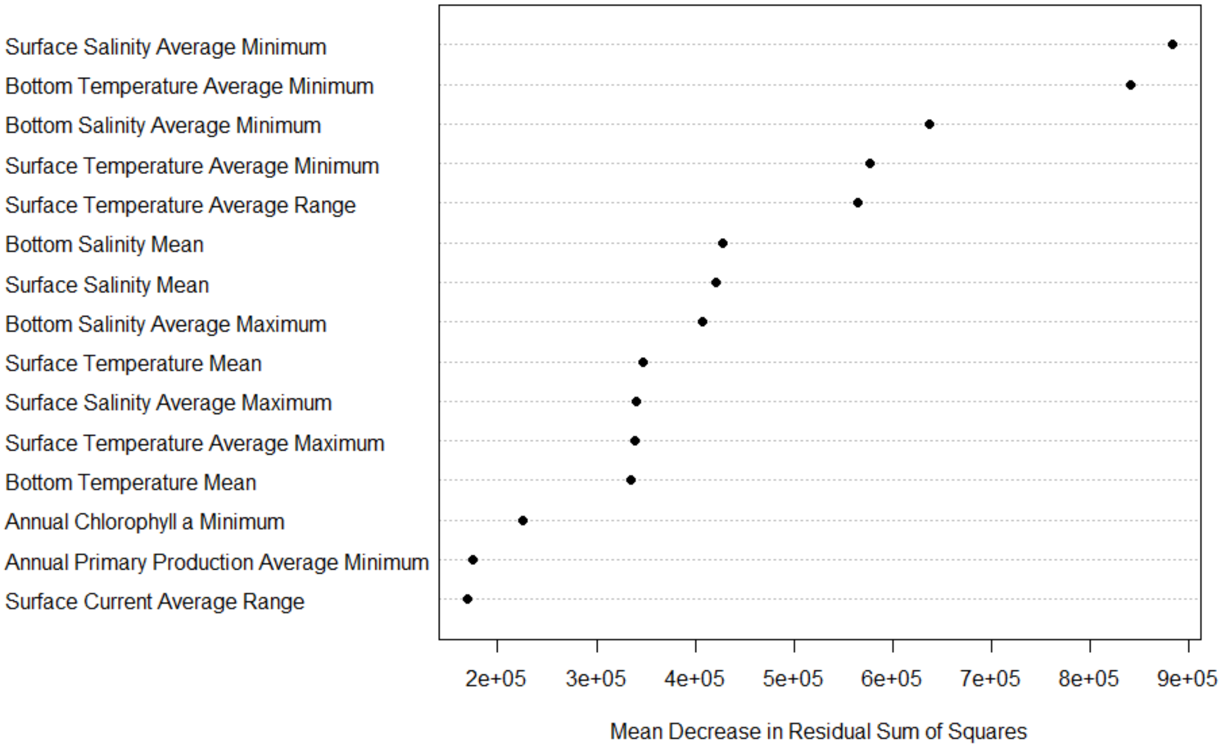

Figure 40. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on sponge biomass data collected from DFO trawl surveys conducted using Campelen trawl gear. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

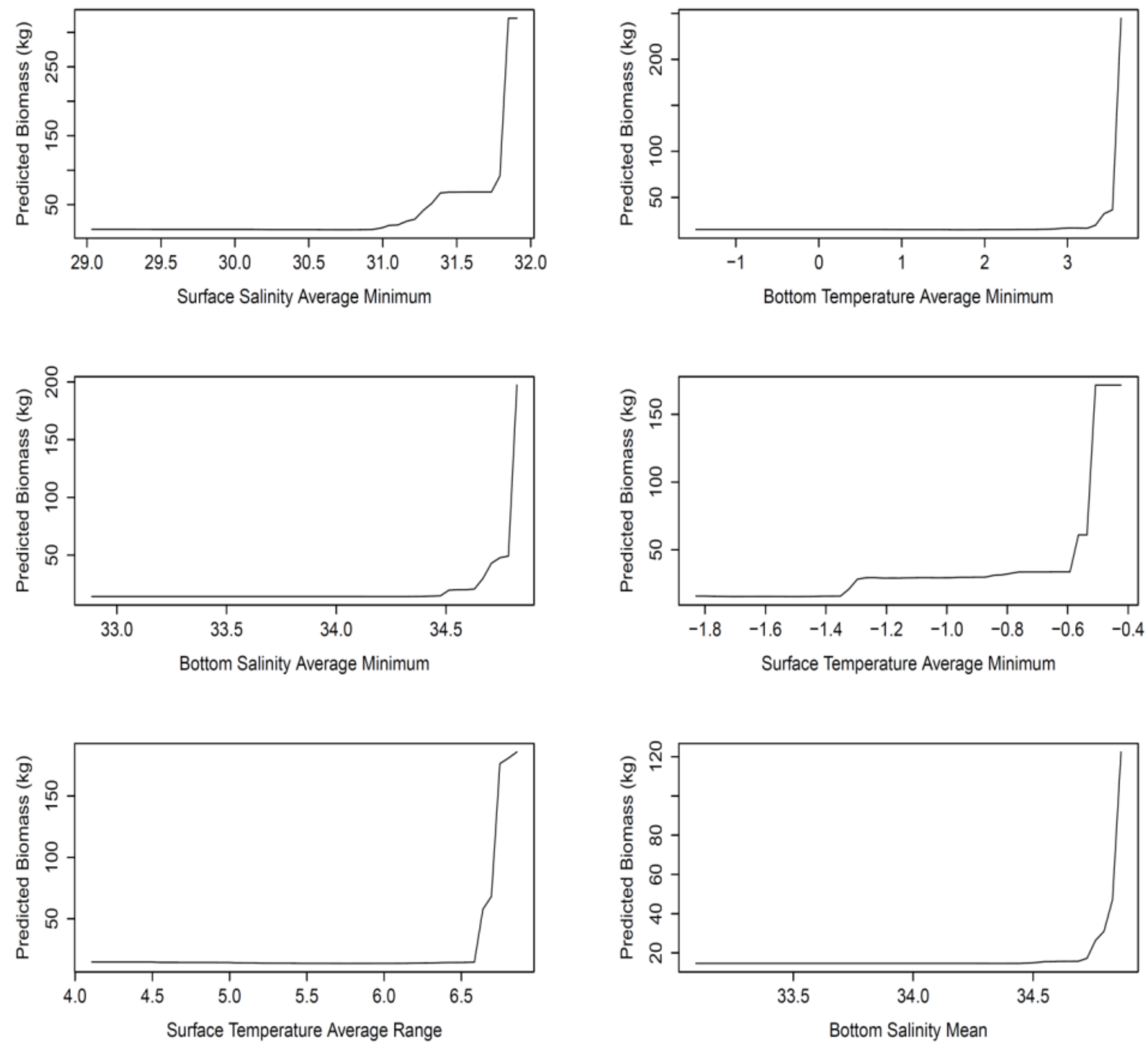

Figure 41. Partial dependence plots of the top six predictors from the random forest model of sponge biomass data collected from DFO trawl surveys using Campelen trawl gear in the Eastern Arctic Region, ordered from left to right from the top.

Cosmos Trawl Gear

Accuracy measures from the regression random forest model on mean sponge biomass records from trawl surveys using Cosmos trawl gear are presented in Table 12. The highest R^2 value was 0.656 while the average was 0.295 ± 0.208 SD. The average Normalized Root-Mean-Square-Error (NRMSE) was 0.031 ± 0.033 SD. The standard deviation was higher than the mean, indicating high variability between model folds. The average percent variance explained by the model was -14.81% ± 11.93 SD.

Figures 42 and 43 show the predicted biomass surface of sponges using mean biomass data from Cosmos trawl surveys. The deeper waters in the Davis Strait and northern Baffin Bay in Baffin Basin were predicted to have moderate to high biomass of sponge. These areas contained no data observations and were considered extrapolated area by the model (Figure 43). Coastal areas where the majority of the positive biomass catches reside were predicted to have low biomass of sponge.

Table 12. Accuracy measures for all 10 model repetitions of 10-fold cross validation of the random forest model of average of sponge biomass (kg) per grid cell recorded from DFO trawl surveys conducted using Cosmos trawl gear the Eastern Arctic Region. RMSE= Root-Mean-Square Error, NRMSE= Normalized Root-Mean-Square Error.

Model Fold	**R^2**	**RMSE**	**NRMSE**	**Percent (%) variance explained**
1	0.002	19.353	0.018	-15.86
2	0.275	6.567	0.006	-16.22
3	0.656	20.051	0.018	-19.42
4	0.398	6.792	0.006	-17.35
5	0.097	131.495	0.121	18.10
6	0.234	34.554	0.032	-22.47
7	0.384	18.975	0.017	-16.79
8	0.506	35.459	0.033	-22.30
9	0.359	36.251	0.033	-21.67
10	0.042	31.264	0.029	-14.14
Mean	**0.295**	**34.076**	**0.031**	**-14.81**
SD	**0.208**	**35.975**	**0.033**	**11.93**

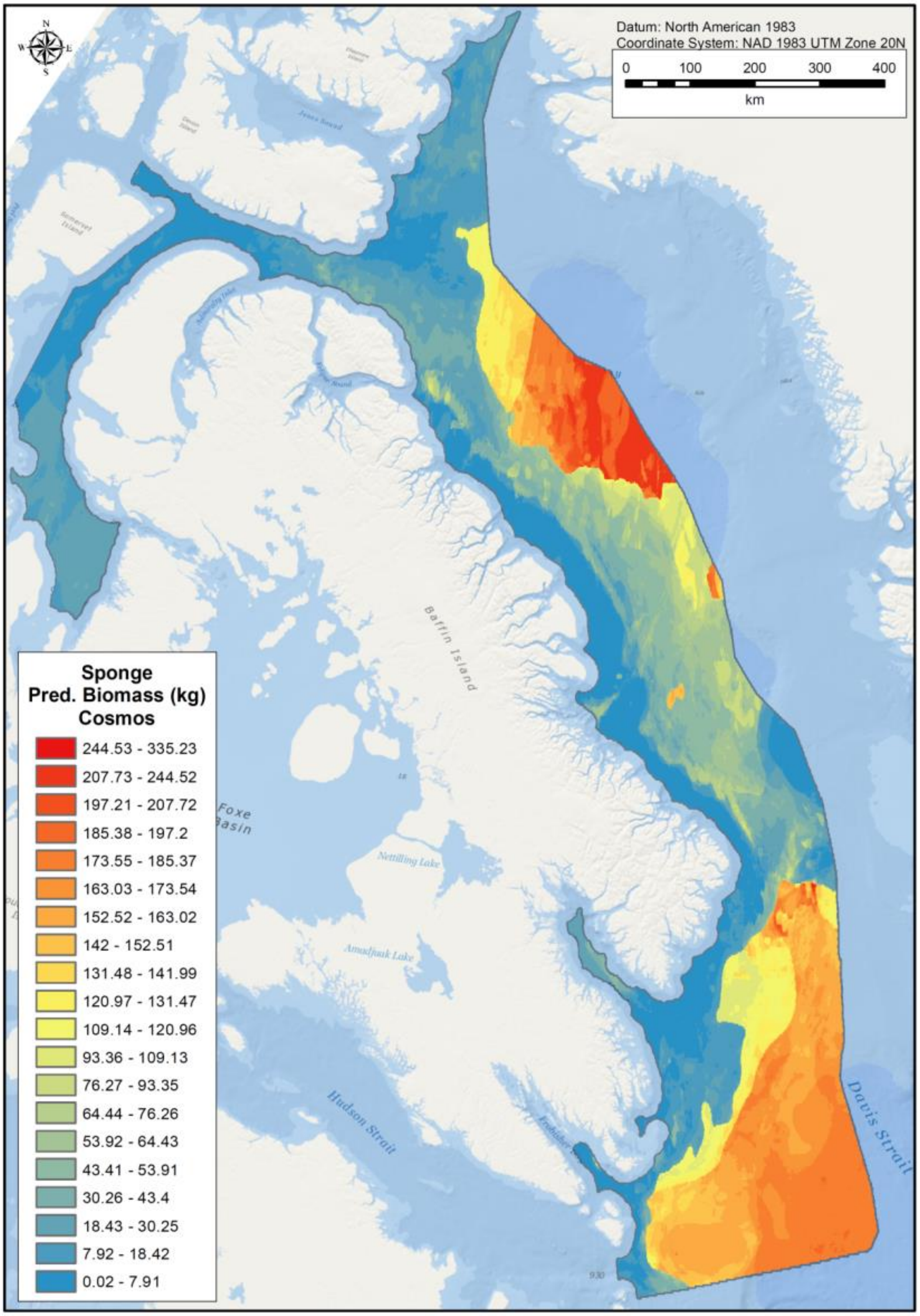

Figure 42. Predictions of biomass (kg) of sponges from catch data recorded in DFO trawl surveys conducted using Cosmos trawl gear in the Eastern Arctic Region between 2006 and 2012.

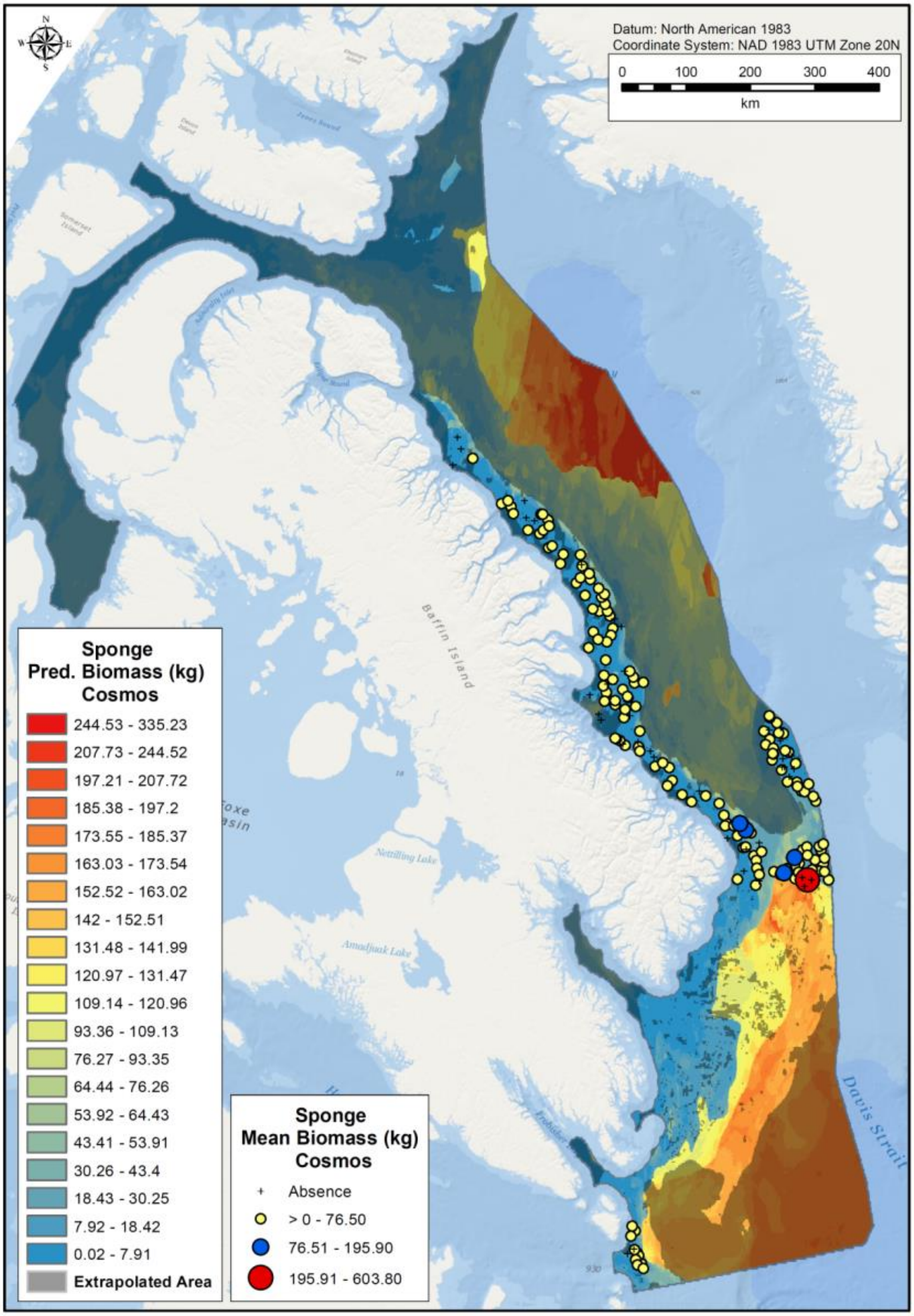

Figure 43. Predictions of biomass (kg) of sponges from catch data recorded in DFO trawl surveys conducted using Cosmos trawl gear in the Eastern Arctic Region between 2006 and 2012. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

Of the 54 variables included in the model, Maximum Average Summer Mixed Layer Depth was most important for predicting sponge biomass from the Cosmos trawl gear records (Figure 44). This variable was followed very distantly by Bottom Salinity Average Minimum, Spring Chlorophyll *a* Minimum, and the remaining variables in the model. The partial dependence plots of the top six environmental predictor variables are shown in Figure45. Predicted biomass was highest at the highest Maximum Average Summer Mixed Layer Depth (> 13.5 m) and Bottom Salinity Average Minimum (> 34.5) values.

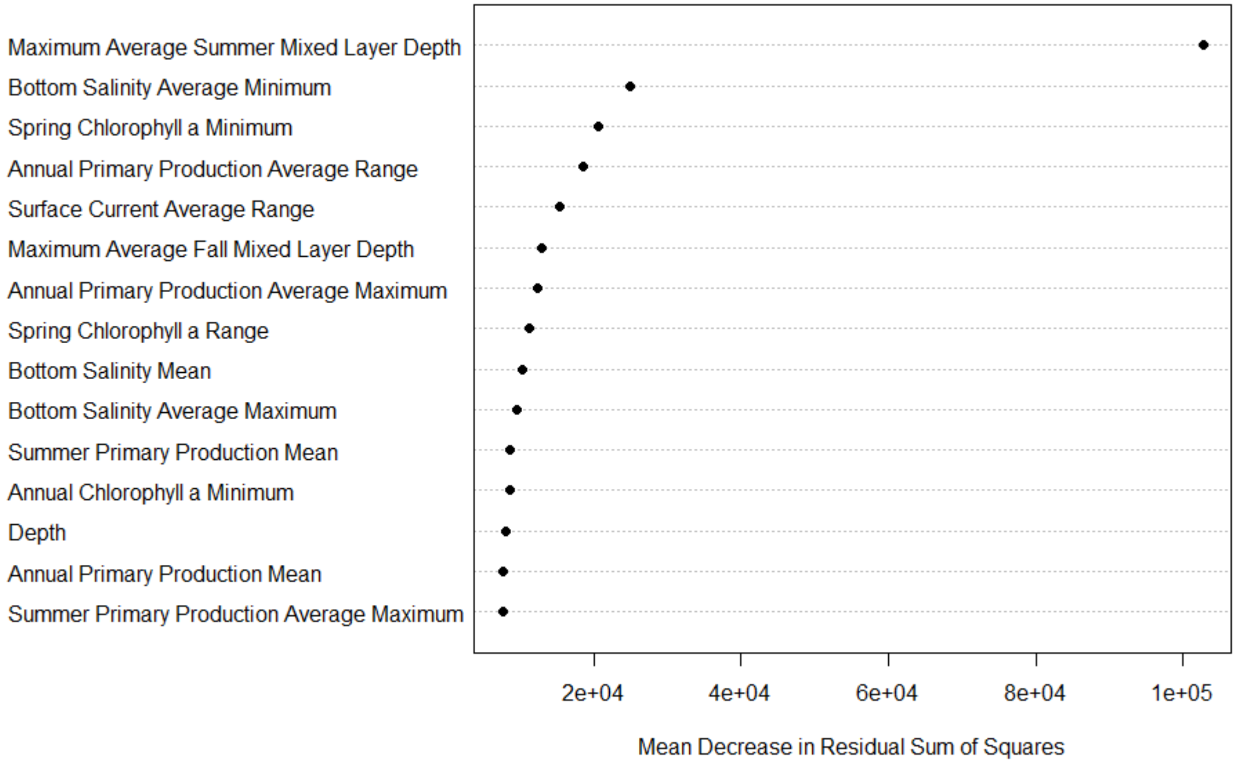

Figure 44. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on sponge biomass data collected from DFO trawl surveys conducted using Cosmos trawl gear. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

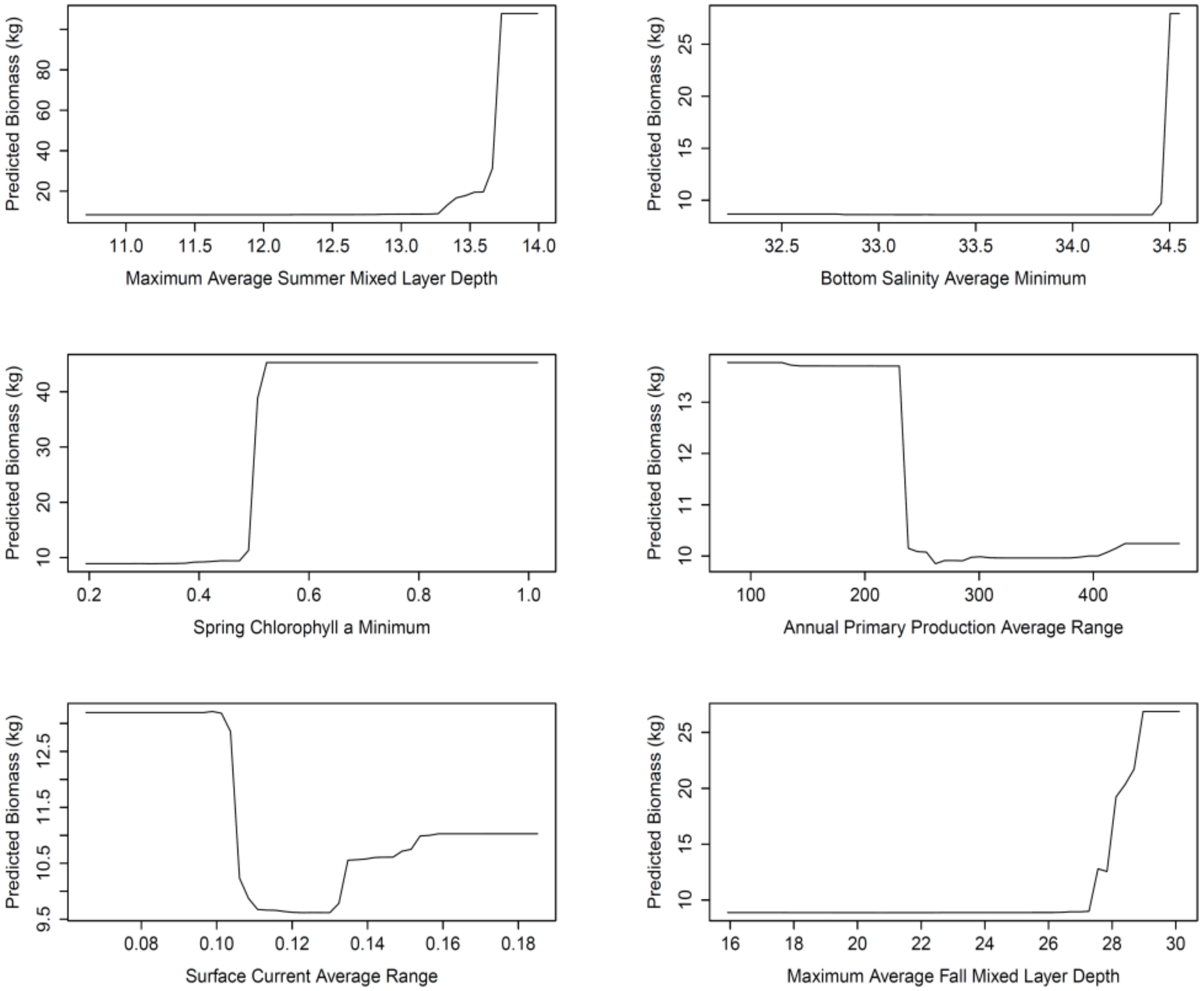

Figure 45. Partial dependence plots of the top six predictors from the random forest model of sponge biomass data collected from DFO trawl surveys using Cosmos trawl gear in the Eastern Arctic Region, ordered from left to right from the top.

Sea Pens (Pennatulacea)

Data Sources and Distribution

Sea pen catch data for the Eastern Arctic Region was collected between 1999 and 2014 and consisted of 302 presences and 451 absences from *Paamuit* surveys conducted using Alfredo trawl gear, 66 presences and 1375 absences from the *Cape Ballard*, *Aqviq*, and *Kinguk* surveys conducted using Campelen trawl gear, and 52 presences and 135 absences from *Paamuit* surveys using Cosmos gear (Table 13). DFO trawl survey records using Alfredo gear had the widest spatial distribution in the study extent (Figure 46). Campelen records were restricted to the Davis Strait, while Cosmos records were distributed along the Baffin Island Shelf and in Davis Strait.

Several i*n situ* benthic imagery records were distributed off Devon Island, in the Narwhal Overwintering and Deep-Sea Coral Conservation Area in southern Baffin Bay, and in the Hatton Basin Voluntary Closure Area in Davis Strait.

Presence-absence random forest models were generated on the combined dataset consisting of 420 presences and 1961 absences (see Figure 47). The highest mean biomass record (5 kg) was recorded in the Davis Strait from a Campelen survey. Another cluster of high mean biomass records occurred in northern Baffin Bay east of Devon Island.

Table 13. Number of presence and absence records of sea pen catch recorded from DFO trawl surveys conducted between 1999 and 2014 in the Eastern Arctic Region.

	Alfredo		**Campelen**		**Cosmos**	
Year	**Presences**	**Absences**	**Presences**	**Absences**	**Presences**	**Absences**
1999	6	-	-	-	-	-
2000	6	-	-	-	-	-
2005	-	-	8	142	-	-
2006	31	30	12	130	15	60
2007	-	-	4	129	0	13
2008	38	45	9	134	28	38
2009	0	17	7	138	-	-
2010	73	48	5	139	8	14
2011	23	61	5	145	-	-
2012	71	87	4	147	1	10
2013	19	67	9	141	-	-
2014	35	96	3	130	-	-
TOTAL	**302**	**451**	**66**	**1375**	**52**	**135**

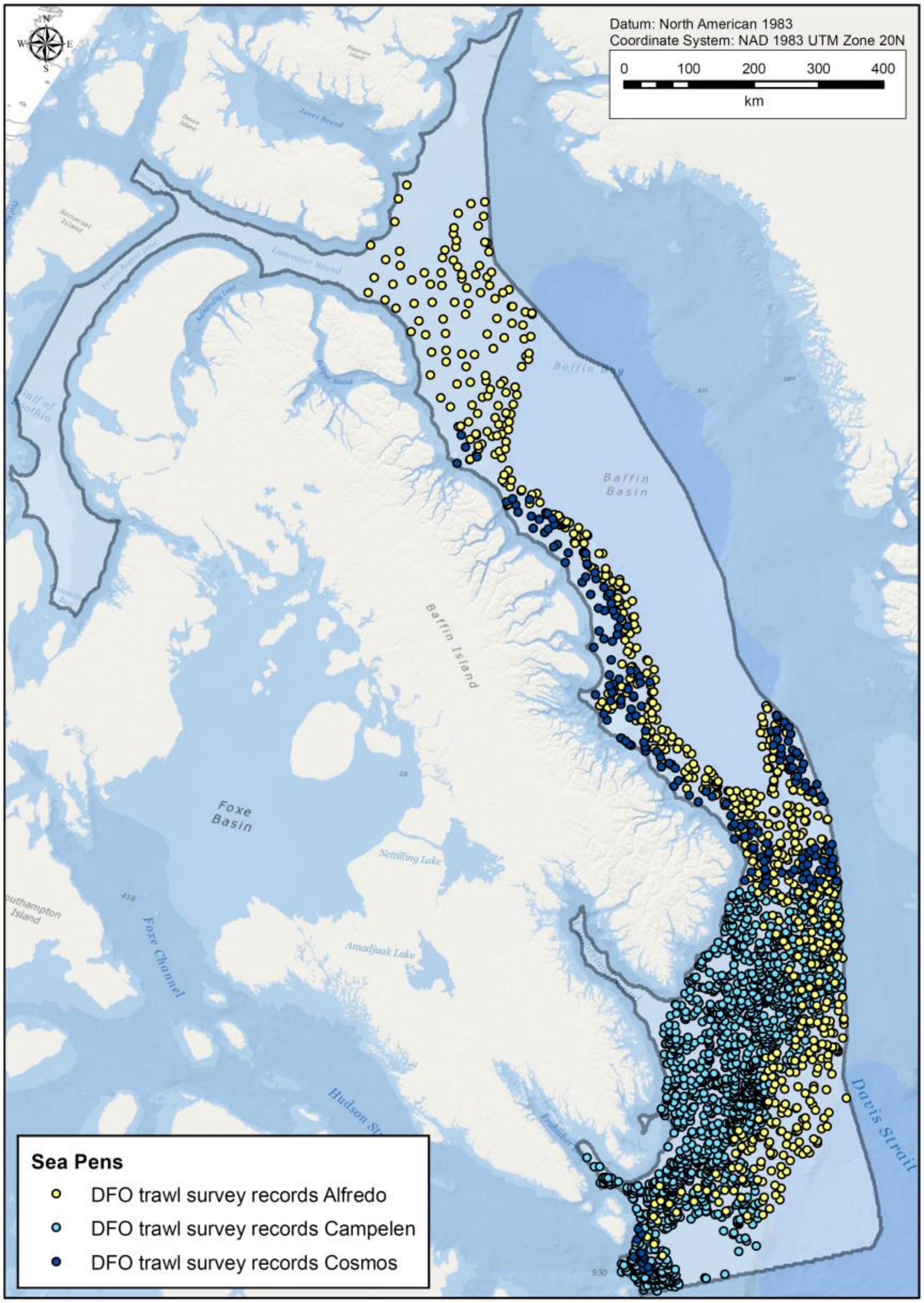

Figure 46. Available sea pen presence data in the Eastern Arctic Region from DFO trawl surveys conducted between 1999 and 2014 and *in situ* benthic imagery observations from a DFO scientific mission conducted in 2012.

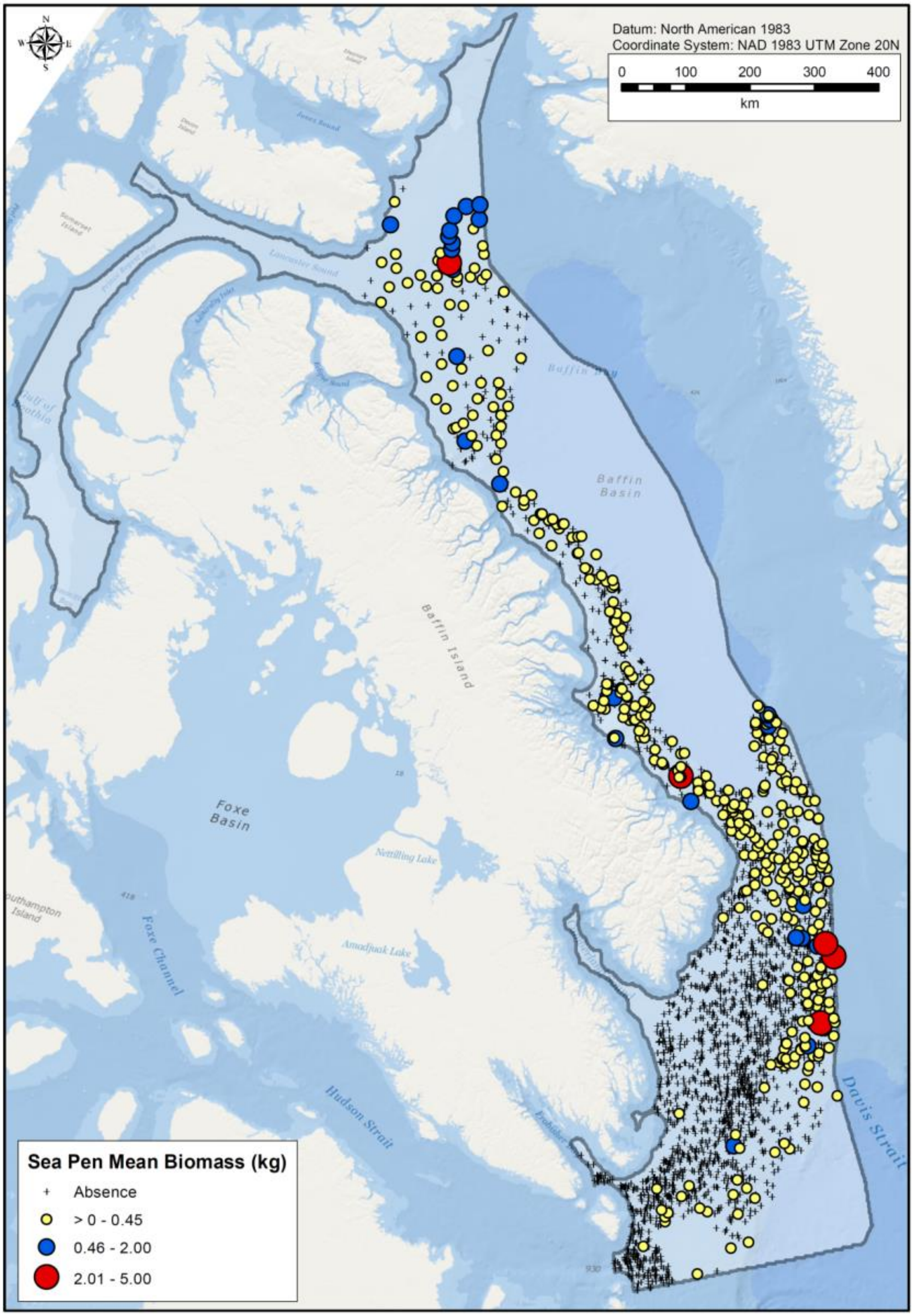

Figure 47. Mean biomass (kg) per grid cell of sea pen catch recorded from DFO trawl surveys conducted in the Eastern Arctic Region between 1999 and 2014.

Model 1 – Balanced Species Prevalence

Accuracy measures for the random forest model on balanced species prevalence (420 presences and 420 absences; Model 1) are presented in Table 14. The mean AUC value was 0.839 ± 0.014 SD, indicating very good model performance. The highest mean AUC of 0.860 was associated with Model Run 9. This model run was therefore considered the optimal model for the prediction of the sea pen response data. The sensitivity and specificity of this model run were 0.826 and 0.743, respectively. The confusion matrix of the optimal model is also presented in Table 14. Class error was higher for the absence class (0.174 for the presence class versus 0.257 for the absence class).

Table 14. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of sea pens within the Eastern Arctic Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 9) which is considered the optimal model for predicting the presence probability of sea pens in the region.

Model Run	**AUC**	**Sensitivity**	**Specificity**
1	0.840	0.800	0.698
2	0.831	0.798	0.729
3	0.844	0.812	0.714
4	0.824	0.810	0.707
5	0.840	0.826	0.726
6	0.813	0.800	0.669
7	0.835	0.833	0.743
8	0.854	0.819	0.745
9	**0.860**	**0.826**	**0.743**
10	0.844	0.838	0.700
Mean	**0.839**	**0.816**	**0.717**
SD	**0.014**	**0.015**	**0.025**

Confusion matrix of model with highest AUC:

Observations	**Predictions**		**Total n**	**Class error**
	Absence	**Presence**		
Absence	312	108	420	0.257
Presence	73	347	420	0.174

The presence probability prediction surface of sea pens is presented in Figure 48. Most of Baffin Bay and Lancaster Sound had moderate to high predicted presence probability of sea pens. The highest predictions of sea pen presence occurred in northern Baffin Bay southeast of Devon Island. Southern Davis Strait was mostly predicted to have low probability of occurrence of sea pens except for an area in deeper water southeast of Hall Peninsula. Areas of high and low predicted presence probability of sea pens corresponded well with the location of presence and absence records (Figure 49), although extrapolation of moderate to high presence probability to the deeper waters off Baffin Island Shelf where there are no presence observations has occurred.

The actual presence and absence records selected for use in the optimal model fold of Model 1 (420 presences and 420 absences; Figure 50) were slightly spatially biased across the study area. Despite there being absence records in northern Baffin Bay, very few were selected during random down-sampling of the data prior to modelling. This likely caused the over-extension of high predicted presence probability in this area. Areas of model extrapolation are also shown in Figure 50. All deep water beyond the Baffin Island Shelf is considered extrapolated area. Lancaster Sound, the Gulf of Boothia, and the southeast Davis Strait are also considered extrapolated area.

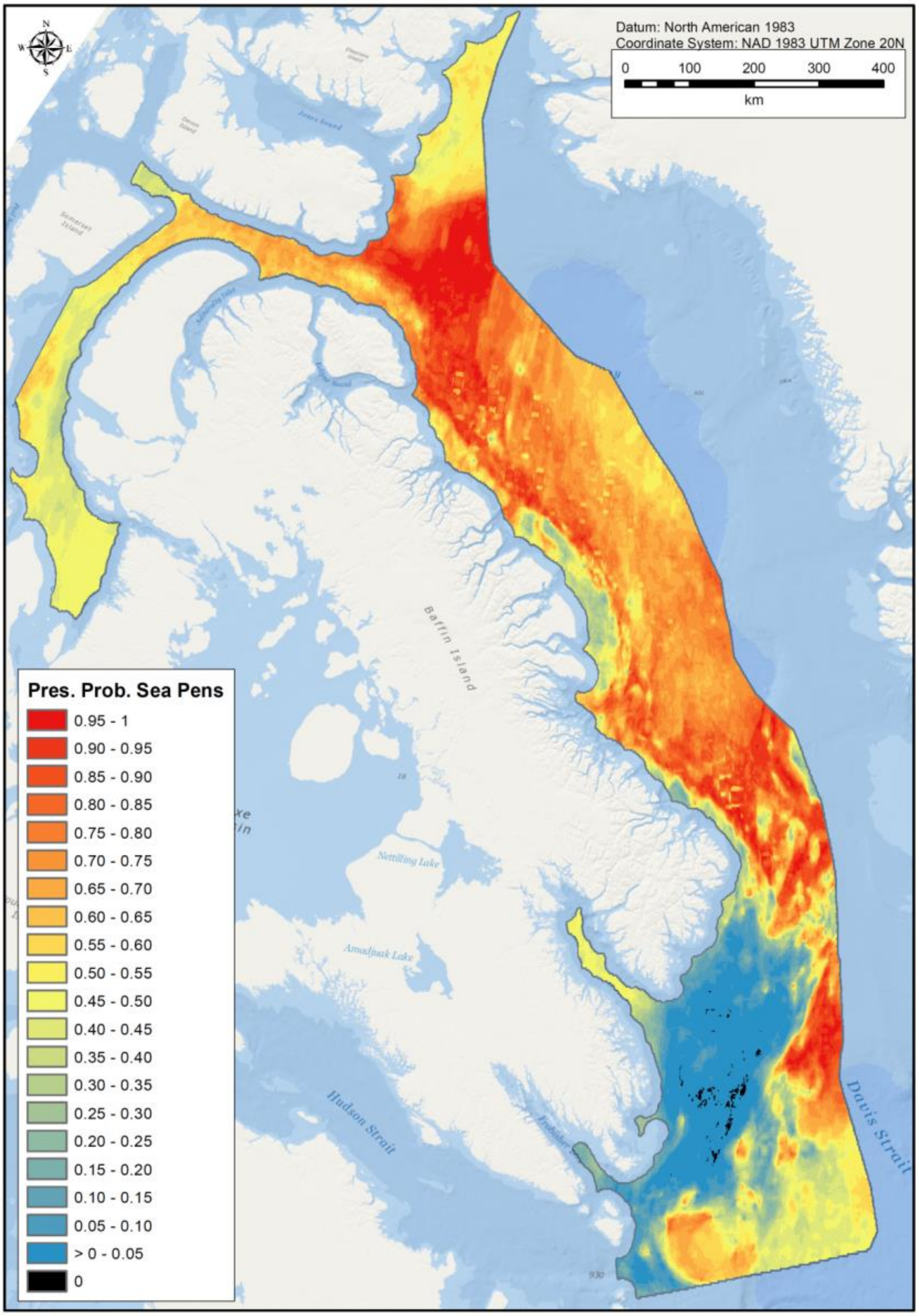

Figure 48. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of sea pen presence and absence data collected from DFO trawl surveys in the Eastern Arctic Region between 1999 and 2014.

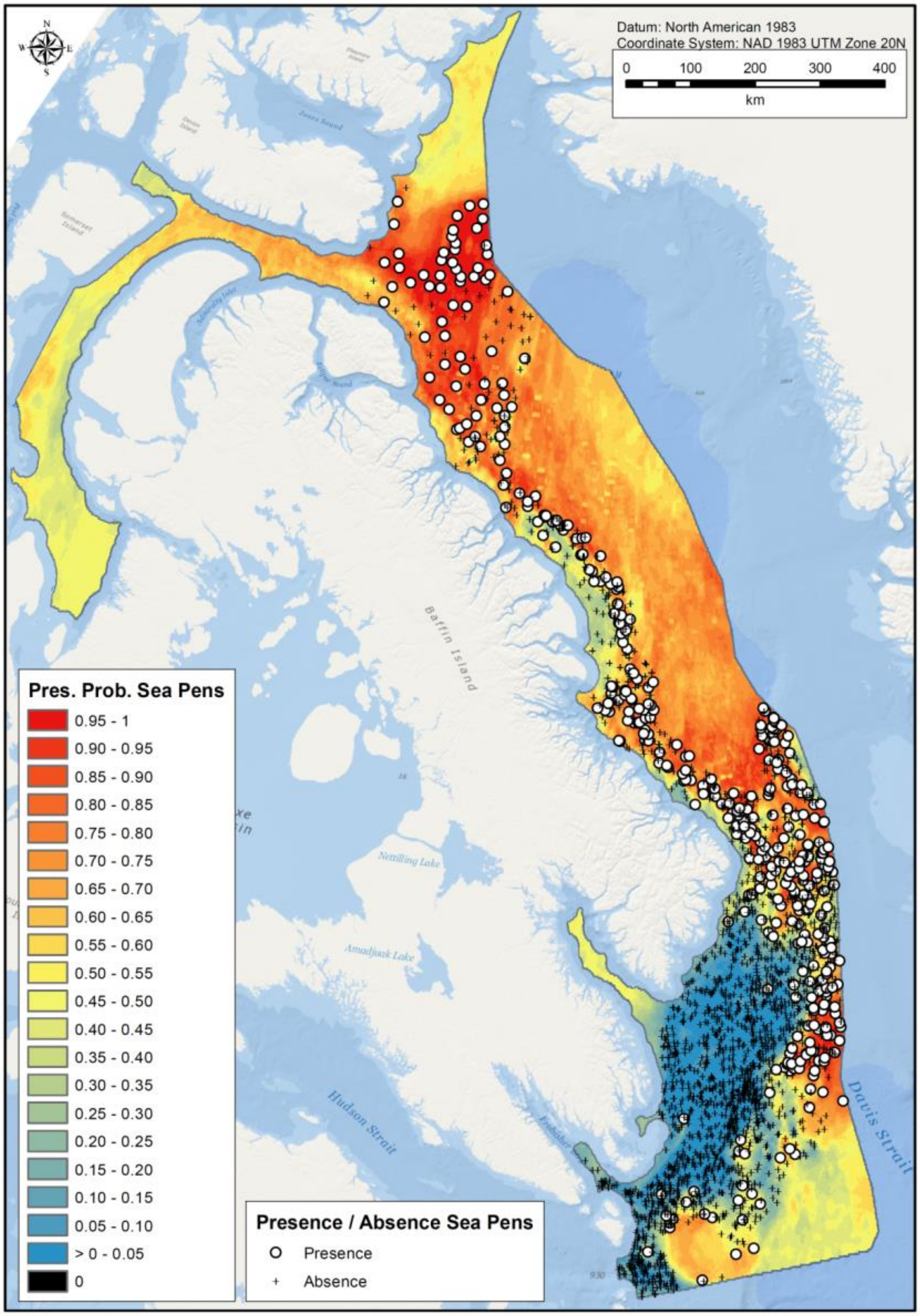

Figure 49. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of sea pen presence and absence data collected from DFO trawl surveys in the Eastern Arctic Region between 1999 and 2014.

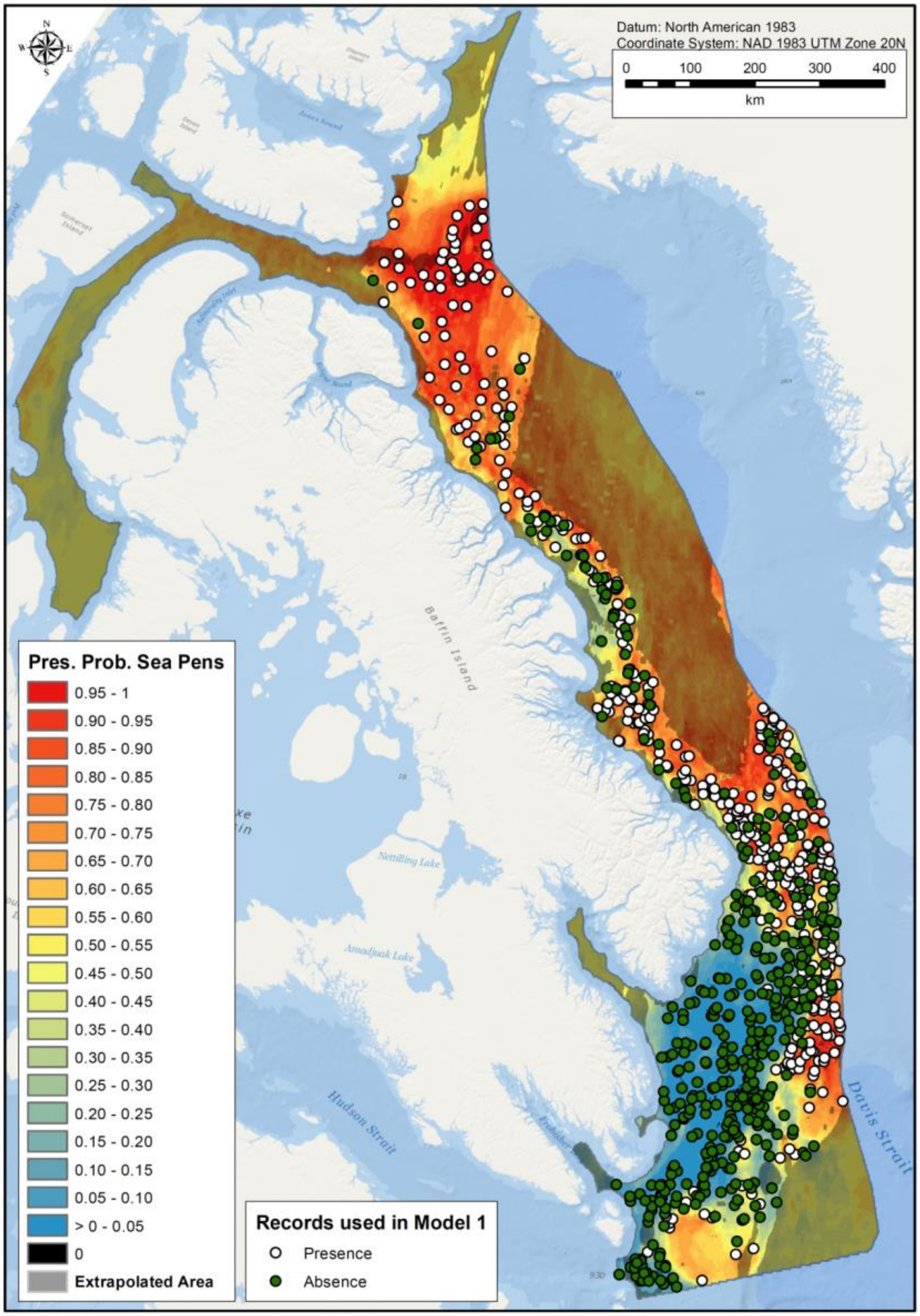

Figure 50. Map of the 840 data observations (420 presences and 420 absences) of sea pens used in the optimal random forest model on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of sea pens and areas of model extrapolation.

Of the 54 environmental predictor variables used in the model, Depth (a non-interpolated variable) was the most important for the classification of the sea pen presence and absence data

(Figure 51). Depth was followed somewhat distantly by Bottom Salinity Average Range and Bottom Temperature Average Range. Surface and bottom temperature variables ranked high in this model. Slope was the 15th most important variable. Partial dependence plots for the top 6 predictor variables are shown in Figure 52. Presence probability of sea pens was highest at Depth values between 500 and 1000 m. Presence probability was highest at lower Bottom Salinity Average Range and Bottom Temperature Average Range values.

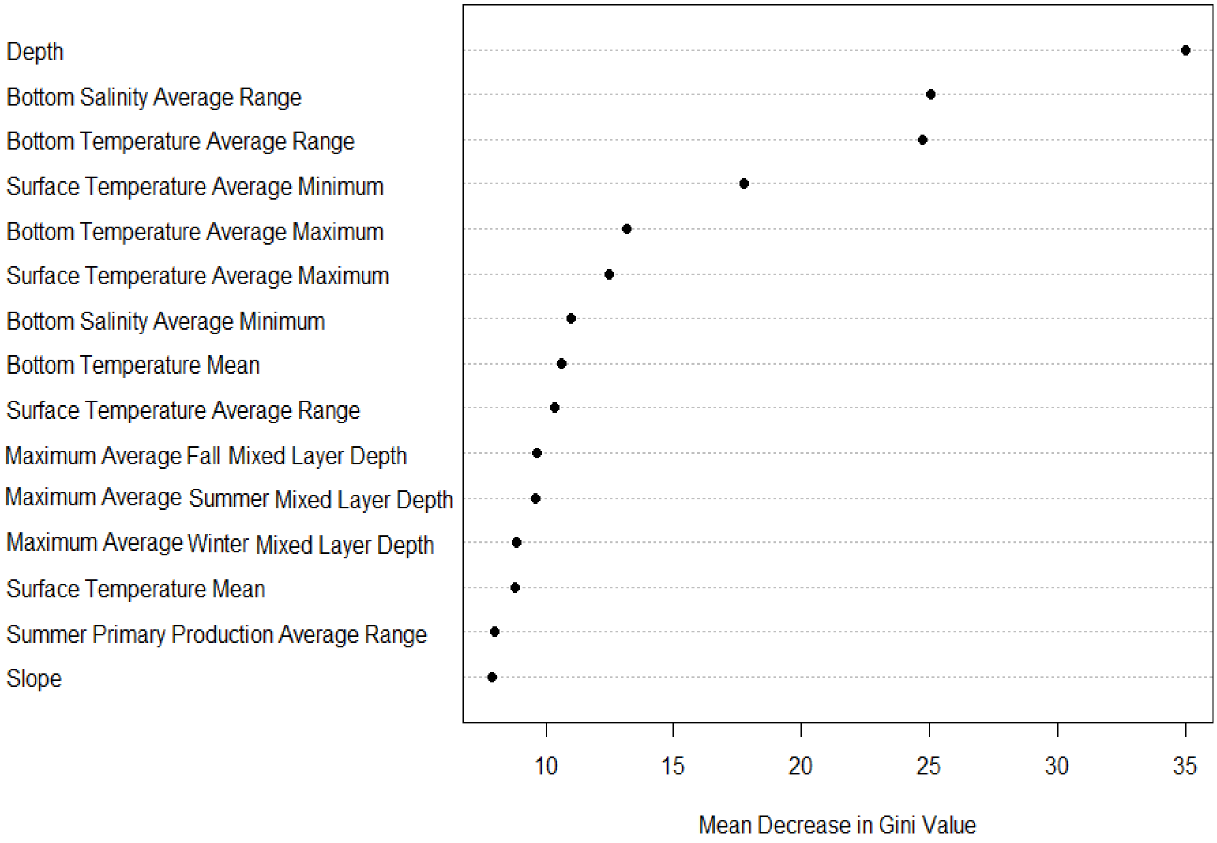

Figure 51. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model of sea pen presence and absence data collected from the Eastern Arctic Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

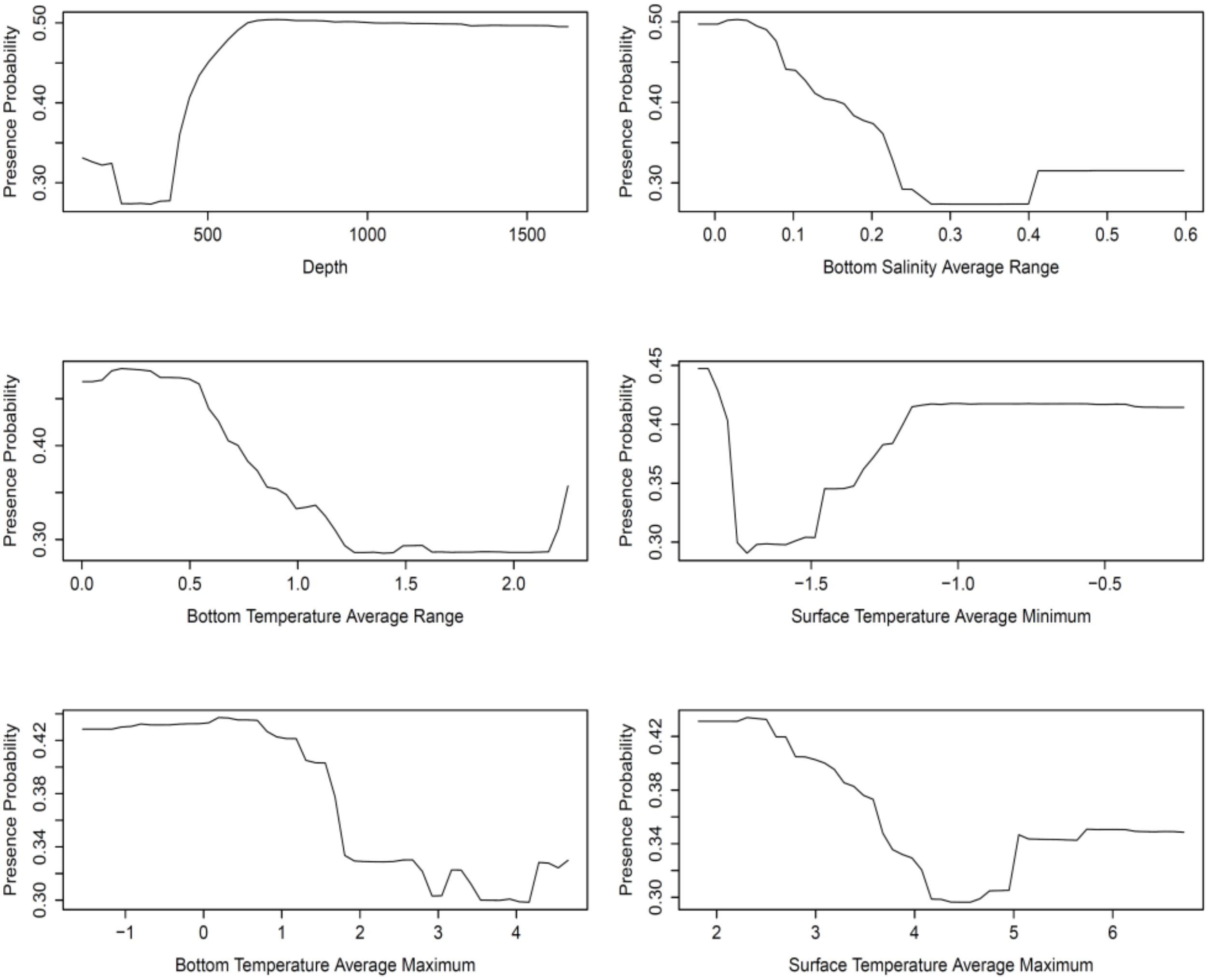

Figure 52. Partial dependence plots of the top six predictors from the optimal random forest model of sea pen presence and absence data collected within the Eastern Arctic Region, ordered from left to right from the top. Predicted presence probability is shown on the *y*-axis of each graph.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 15 shows the accuracy measures for the random forest model on all sea pen presence and absence data (420 presences and 1961 absences; Model 2) and a threshold equal to species prevalence (0.18). The mean AUC was nearly identical to that of Model 1 (0.839) and indicates very good model performance. Sensitivity was slightly lower than that of Model 1 (0.814 versus 0.816 of Model 1), while specificity was slightly higher (0.721 versus 0.717 of Model 1). Class error for both the presence and absence classes was slightly higher than that of Model 1.

The predicted presence probability surface of sea pens generated from Model 2 is shown in Figure 53. Overall, areas of high presence probability are much reduced in this model compared to model 1. Like Model 1, the largest area of high presence probability occurred in northern Baffin Bay southeast of Devon Island. The edge of the Baffin Island Shelf also had smaller pockets of high sea pen presence probability. Much of the Davis Strait was predicted to have zero or low presence probability of sea pens. Figure 54 shows the spatial distribution of the sea pen presence and absence data used in Model 2 over the predicted presence probability surface. Predicted presence probability was low in locations where a high number of presence observations occurred, particularly along the shelf break of Baffin Island and in Davis Strait. This could be due to the high overlap between presence and absence data points in those areas and the inclusion of all absence data in the model.

Figure 55 shows the classification of sea pen presence probability into presence and absence categories based on the prevalence threshold of 0.18. In this map, presence probabilities greater than 0.18 were classed as presence, while probabilities lower than 0.18 were classed as absence. Most of the study extent was predicted as presence of sea pens. The largest area predicted as absence of sea pens occurred in the southern portion of the study extent in Davis Strait. Smaller pockets of sea pen absence were distributed on Baffin Island Shelf. Areas of extrapolation in this model (Figure 55) are similar to that of Model 1 with areas of model extrapolation occurring in Lancaster Sound, the Gulf of Boothia, in the deep water off Baffin Island Shelf, and in the southeast corner of the spatial extent in Davis Strait.

Table 15. Accuracy measures for unrepeated 10-fold cross validation of a random forest model of presence and absence of sea pens within the Eastern Arctic Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.842		**Absence**	**Presence**				
2	0.838	**Absence**	1413	548	1961	0.279	0.814	0.721
3	0.836	**Presence**	78	342	420	0.186		
4	0.820							
5	0.829							
6	0.843							
7	0.838							
8	0.848							
9	0.822							
10	0.868							
Mean	**0.838**							
SD	**0.014**							

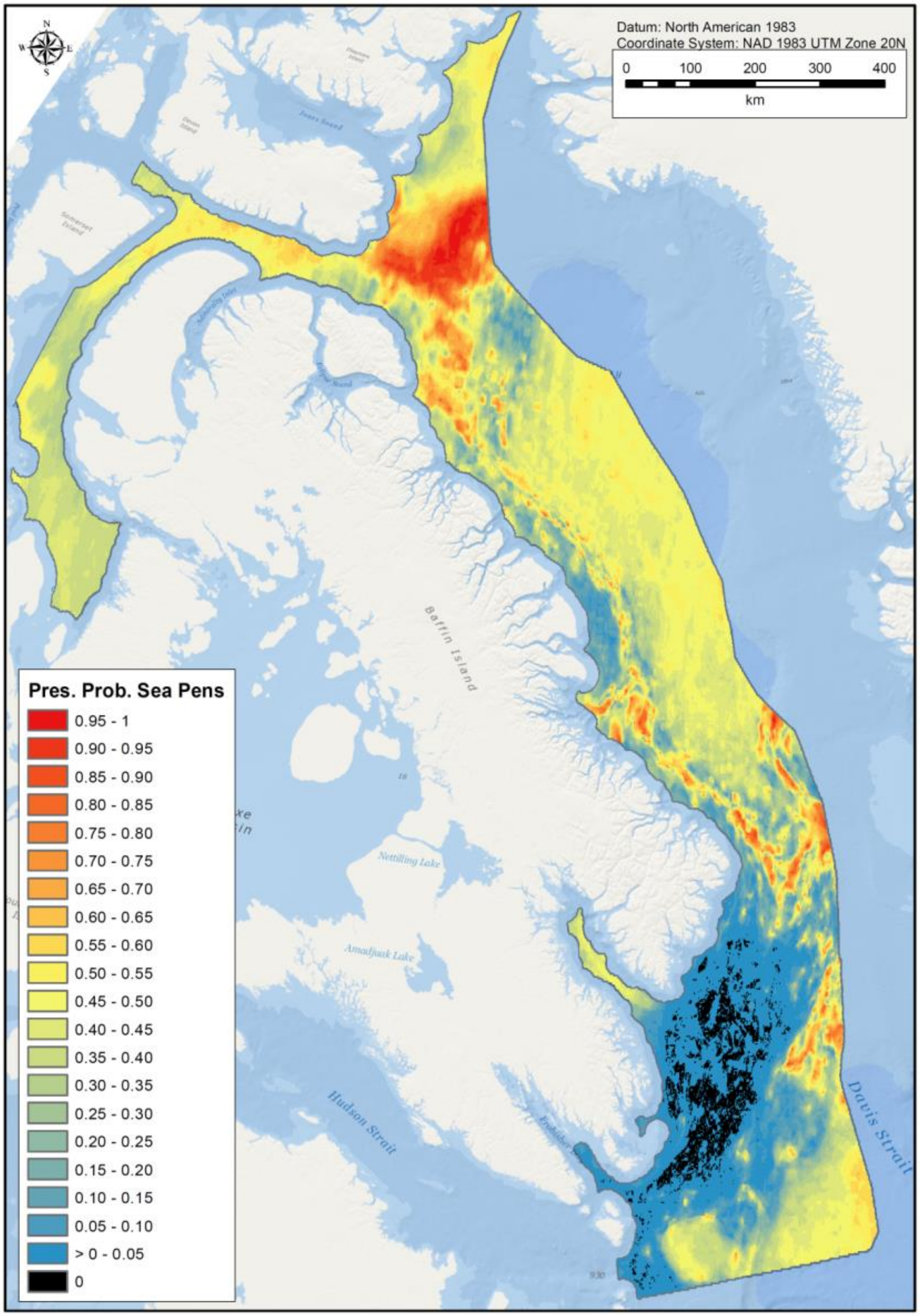

Figure 53. Prediction of presence probability (Pres. Prob.) of sea pens based on a random forest model on unbalanced presence and absence sea pen catch data collected from DFO trawl surveys conducted within the Eastern Arctic Region between 1999 and 2014.

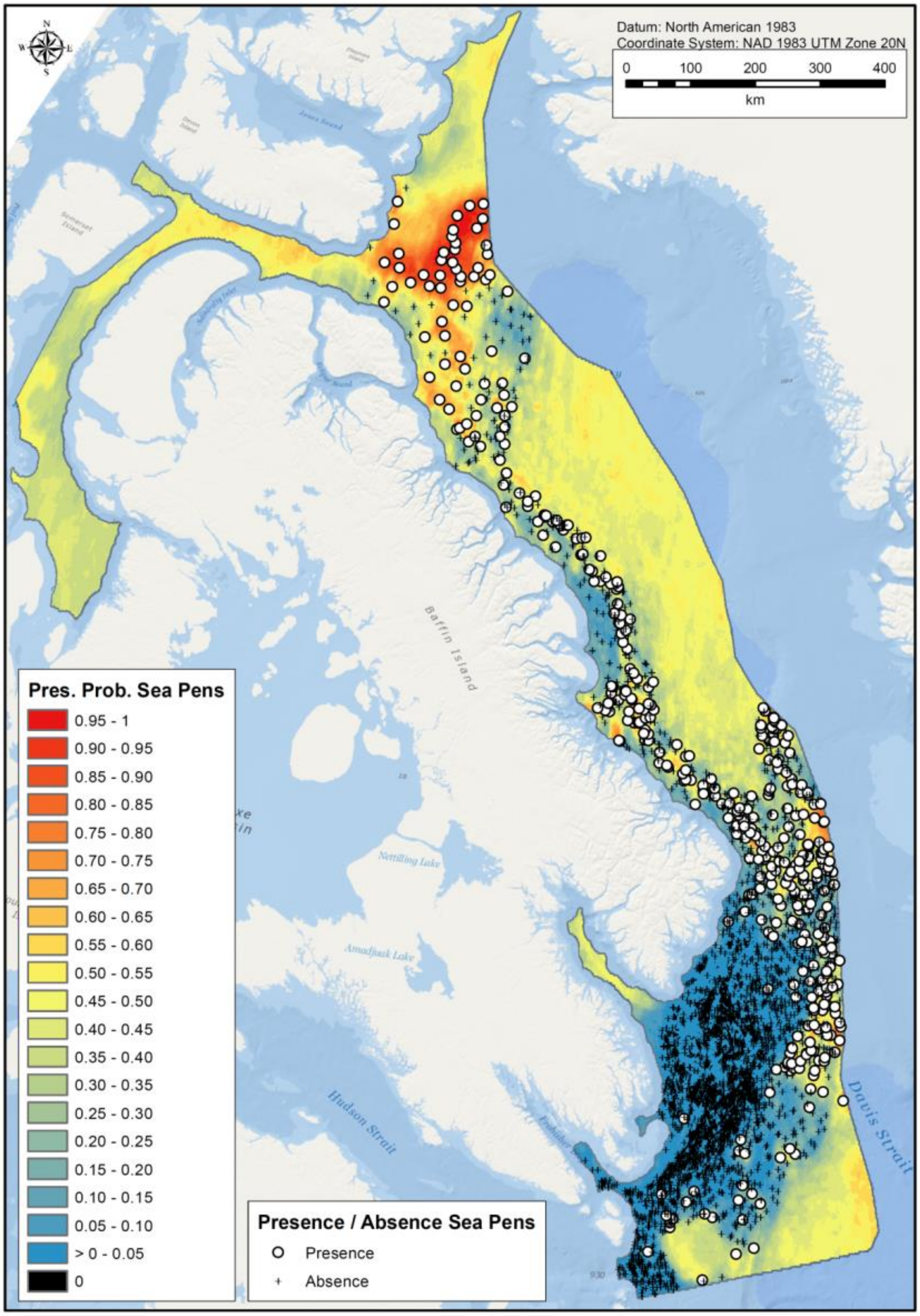

Figure 54. Presence and absence observations and prediction of presence probability (Pres. Prob.) of sea pens based on a random forest model on unbalanced presence and absence sea pen catch data collected from DFO trawl surveys conducted within the Eastern Arctic Region between 1999 and 2014.

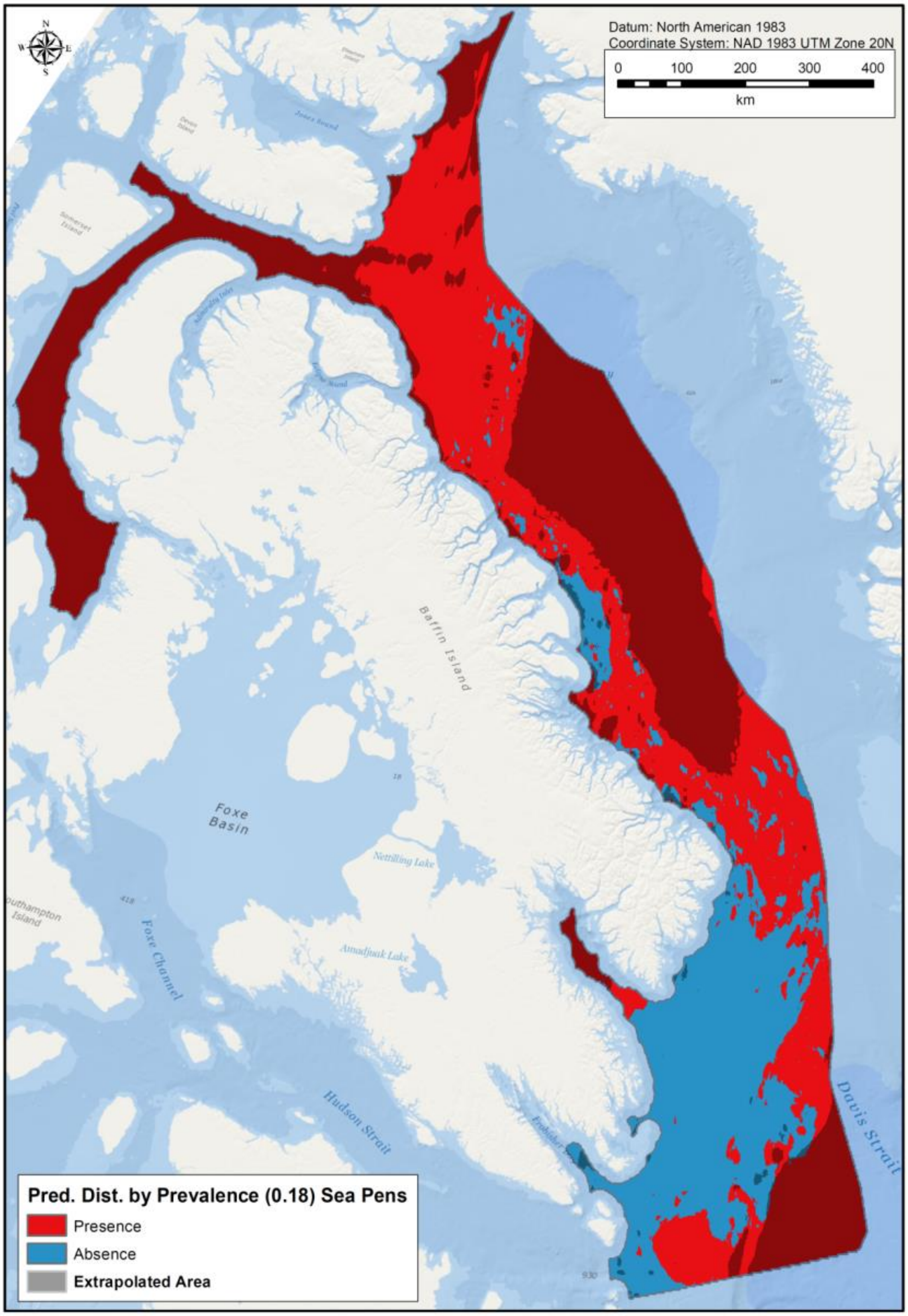

Figure 55. Predicted distribution (Pred. Dist.) of sea pens in the Eastern Arctic Region based on the prevalence threshold of 0.18 of sea pen presence and absence data used in Model 2. Also shown are areas of model extrapolation (grey polygon may appear red or blue).

The order of importance of environmental predictor variables in Model 2 (Figure 56) was similar to that of Model 1. Bottom Salinity Average Range was the most important variable in Model 2, compared to Depth in Model 1. Depth and Bottom Temperature Average Range were the second and third-most important variables in this model. Partial dependence plots for the top 6 predictor variables are shown in Figure 57. Presence probability of sea pens was highest at the lowest Bottom Salinity Average Range values and decreased to near-zero at salinity values of 0.25. Like in Model 1, sea pen presence probability was highest between 500 and 1000 m along the Depth gradient.

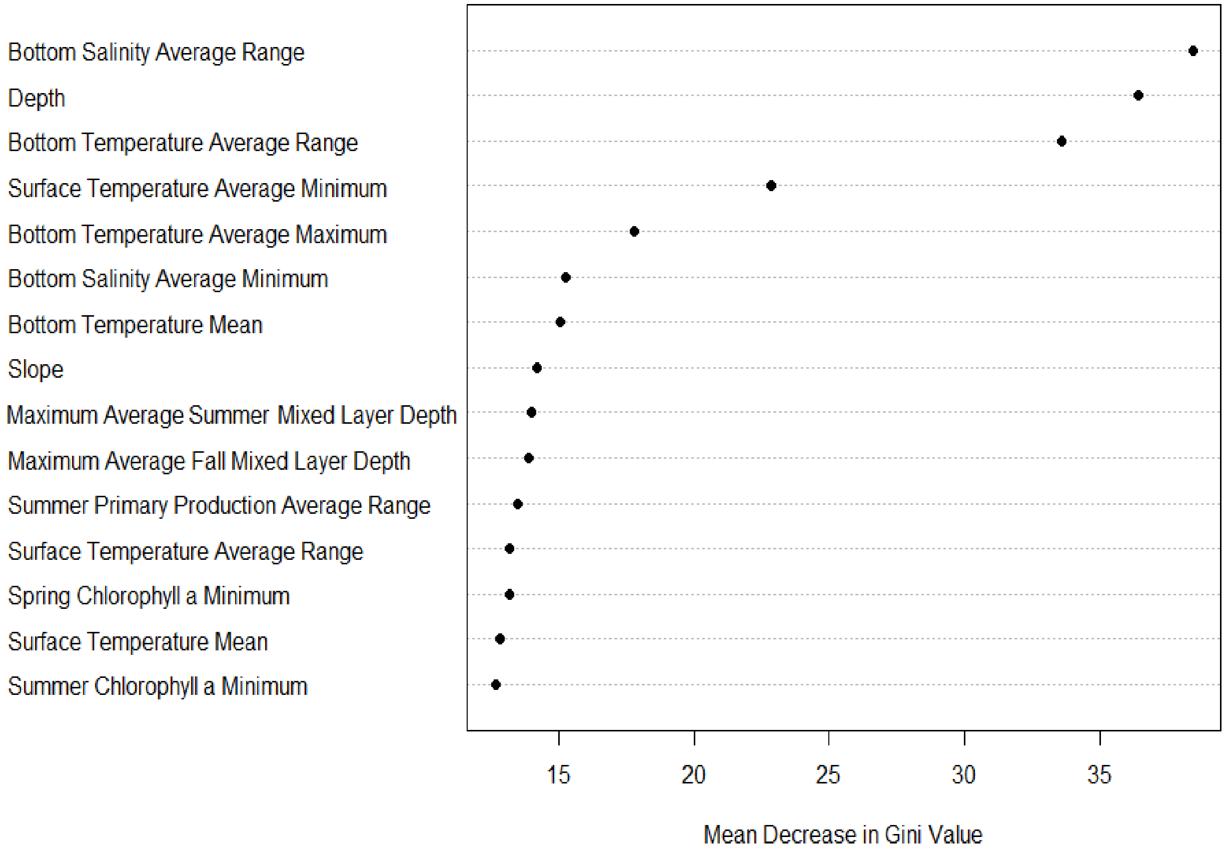

Figure 56. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced sea pen presence and absence data from the Eastern Arctic Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

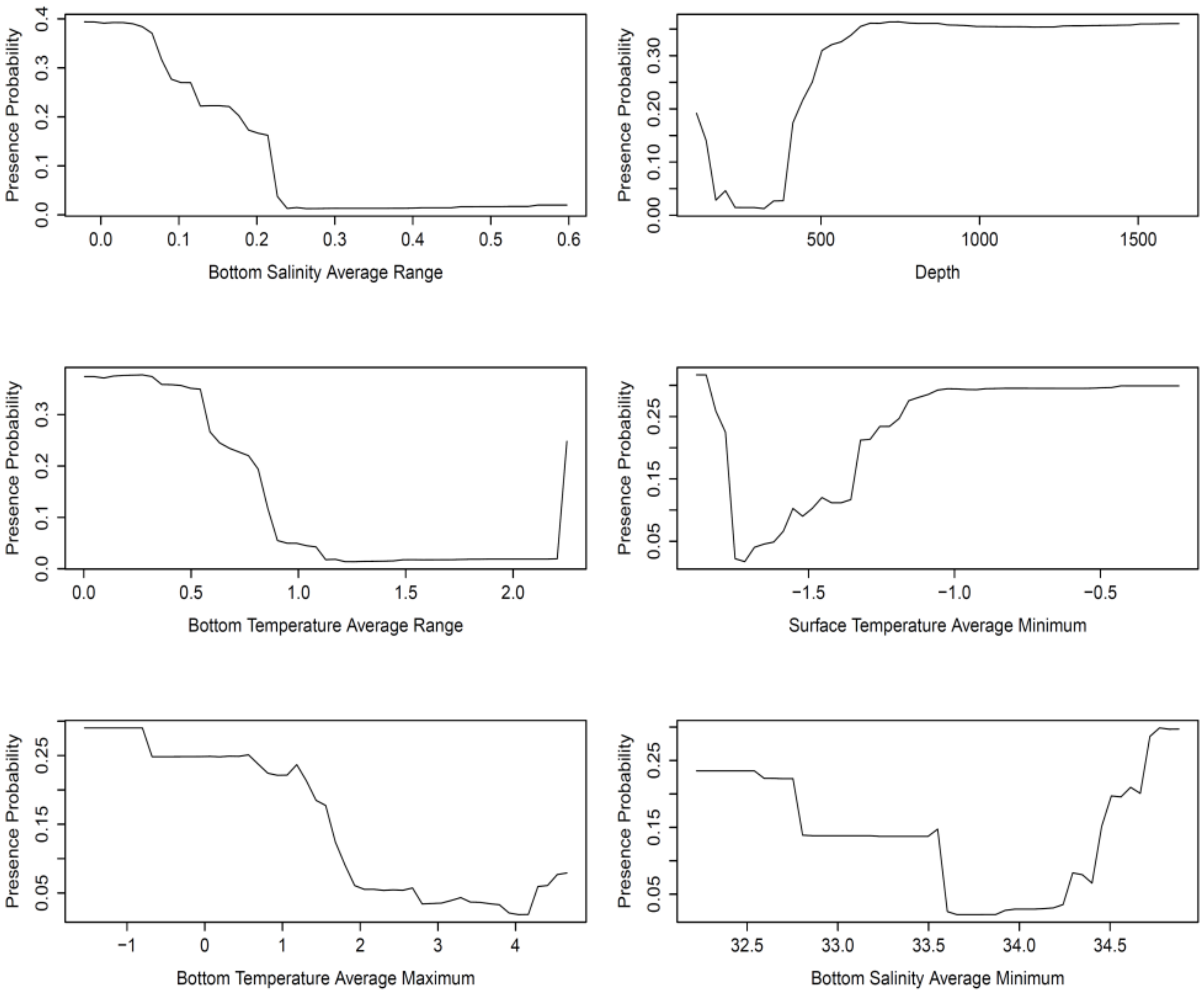

Figure 57. Partial dependence plots of the top six predictors from the random forest model of sea pen unbalanced presence and absence data collected within the Eastern Arctic Region, ordered from left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using all sea pen records and an unbalanced species prevalence (Model 2) was selected as the best predictor of sea pen distribution in the Eastern Arctic Region (Figure 53). Although AUC values were identical between models (0.838) and sensitivity and specificity were comparable (0.816 sensitivity and 0.717 specificity of Model 1, compared to 0.814 and 0.721 sensitivity and specificity of Model 2), Model 1 was considered a poorer predictor of presence probability of sea pens due to its exaggeration of high presence probability, particularly in the deep waters off Baffin Island where there were no data observations. This phenomenon was due to random down-sampling of the absence data. Model 2 produced a more realistic depiction of the potential distribution of sea pens in the Eastern Arctic Region with little extrapolation of high presence probability beyond presence points.

Validation of Selected Model Using Independent Data

Figure 58 shows the predicted presence probability of sea pens generated from Model 2 at the location of sea pen records from benthic camera observations collected during DFO scientific missions to the Eastern Arctic in 2012 and 2013. In 2012, several sea pen records occurred southeast of Devon Island where the model predicted high presence probability of sea pens. In the Narwhal Over-wintering and Deep-Sea Coral Conservation Area, there was relatively good spatial congruence between the location of sea pens from the *in situ* surveys and areas of high predicted presence probability from the model. There were only two sea pen records recorded in the Hatton Basin closure in Davis Strait, in an area where the model predicted low (0.2 to 0.5) presence probability of sea pens. Of the 39 sea pen records from the 2012 survey, all (100%) were predicted as presence based on the prevalence threshold of 0.18.

In 2013, several sea pens were recorded in relatively shallow water on Baffin Island Shelf in an area where the model predicted low presence probability. A single sea pen was recorded in the Narwhal Closure area. Several sea pens were observed in the Hatton Basin closure in an area where the model predicted low presence probability of sea pens. Of the 15 sea pen records from 2013, 14 (93.3%) were predicted as presence based on the prevalence threshold. The single sea pen record predicted as absence by the model was located in the Hatton Basin closure area. The positive occurrence of sea pens there suggests that this is suitable habitat for these organisms.

The spatial congruence between the FOP sea pen records was good in the deeper waters of Baffin Bay (Figure 59). However, there were several sea pen records located in shallower waters along Baffin Island that were predicted with a lower presence probability by the model. Of the 1345 records, 772 (57%) were predicted as presence, with the majority of the absences being located in Davis Strait. As for the sponges, FOP records will tend to produce mismatches arising from presences recorded where the start position indicates an absence in areas of prediction spatial heterogeneity, due to the potential for transit over presence areas during the very long commercial tows.

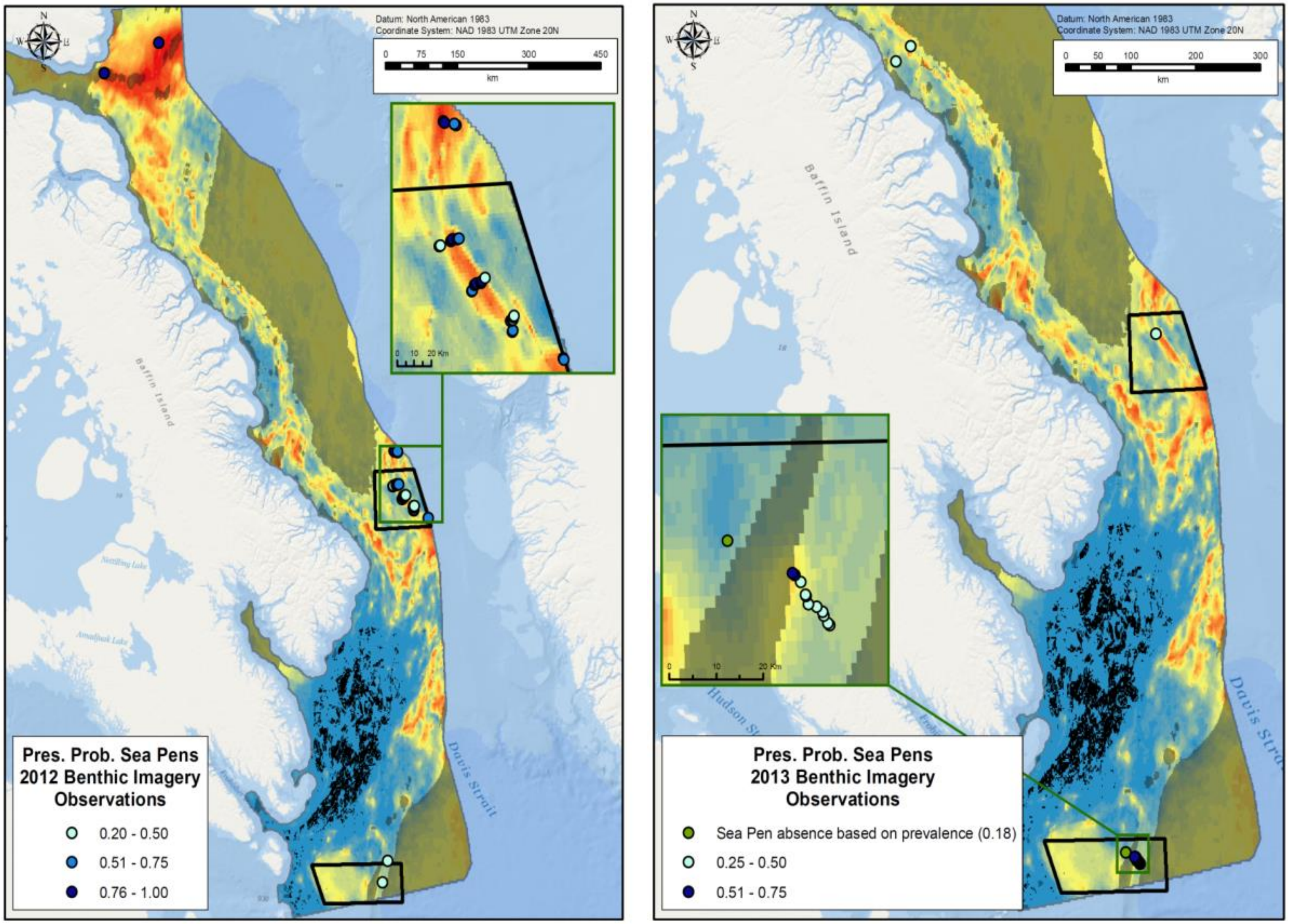

Figure 58. Validation of sea pen presence probability from Model 2 using *in situ* camera records of sea pens collected during DFO scientific missions conducted in 2012 (left) and 2013 (right). Also shown are the Narwhal Overwintering and Deep-Sea Coral Conservation Area and the Hatton Basin Voluntary Closure Area. Inset maps show the Narwhal (left) and Hatton Basin (left) closures.

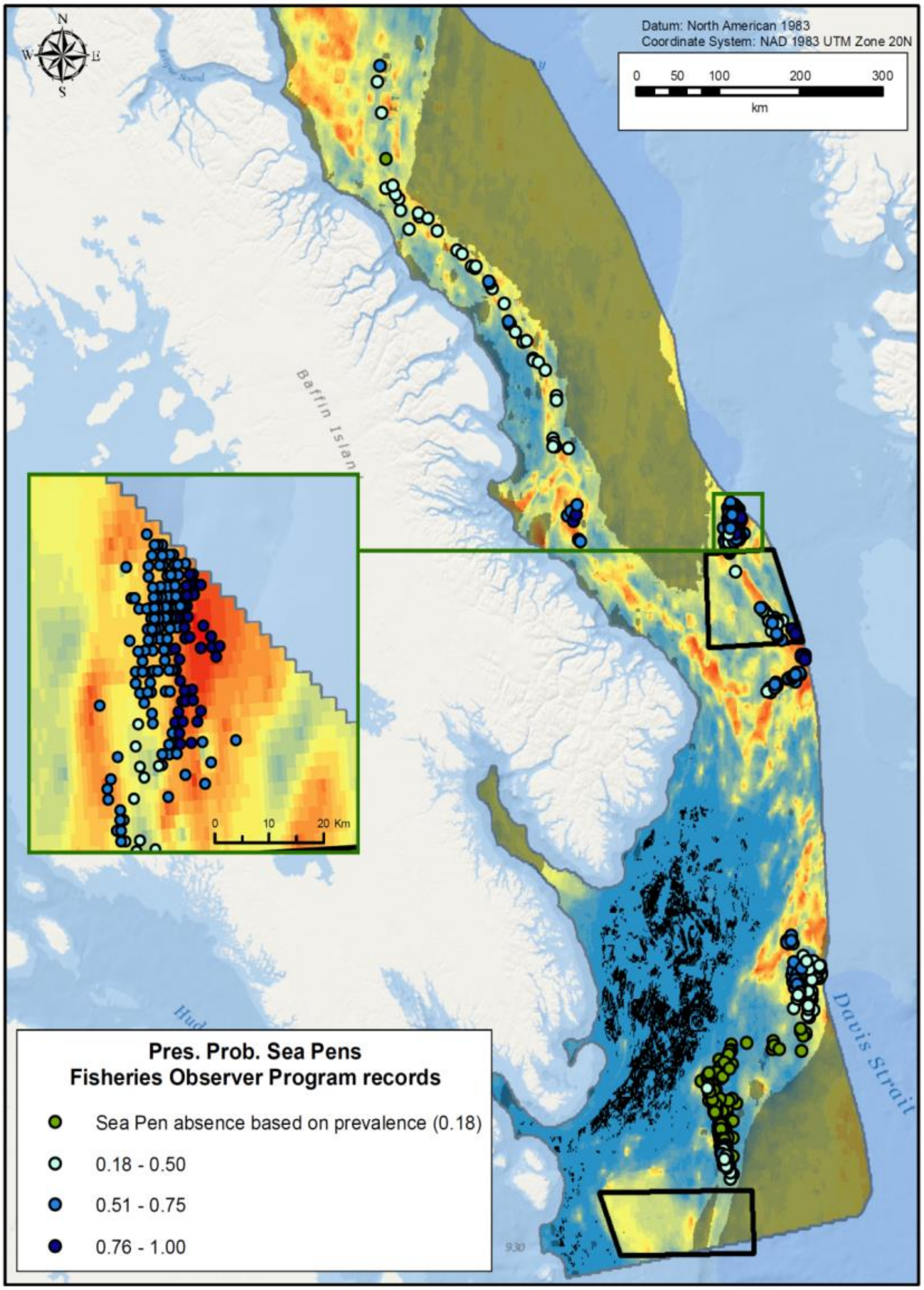
N
S
E
W
Datum: North American 1983
Coordinate System: NAD 1983 UTM Zone 20N
0
50
100
200
300
km
Baffin Islan
0
10
20 Km
Davis Strait
Hud
Pres. Prob. Sea Pens
Fisheries Observer Program records
● Sea Pen absence based on prevalence (0.18)
● 0.18 - 0.50
● 0.51 - 0.75
● 0.76 - 1.00

Figure 59. Validation of sea pen presence probability from Model 2 using sea pen records collected by the Fisheries Observer Program between 1998 – 2013. Also shown are the Narwhal Overwintering and Deep-Sea Coral Conservation Area and the Hatton Basin Voluntary Closure Area.

Prediction of Sea Pen Biomass Using Random Forest

Alfredo Trawl Gear

Accuracy measures from the regression random forest model on mean sea pen biomass records from trawl surveys using Alfredo trawl gear are presented in Table 16. The highest R^2 value was 0.2017 while the average was 0.089 ± 0.069 SD, indicating poor model performance. The average Normalized Root-Mean-Square-Error (NRMSE) was 0.062 ± 0.034 SD. The average percent variance explained by the model was -3.03 ± 2.41 SD.

Figures 60 and 61 show the predicted biomass surface of sea pens using mean biomass data from Alfredo trawl surveys. The majority of the spatial extent was predicted to have low (< 0.106 kg) biomass of sea pens. The highest predicted biomass occurred in the southeast corner of the study extent in Davis Strait where there are no data observations (Figure 61). Northern Baffin Bay, Lancaster Sound, and the Gulf of Boothia were predicted to have moderate sea pen biomass, as well as Cumberland Sound and Frobisher Bay. Areas of moderate to high predicted biomass were considered extrapolated by the model (Figure 61).

Table 16. Accuracy measures for all 10 model repetitions of 10-fold cross validation of the random forest model of average of sea pen biomass (kg) per grid cell recorded from DFO trawl surveys conducted using Alfredo trawl gear the Eastern Arctic Region. RMSE= Root-Mean-Square Error, NRMSE= Normalized Root-Mean-Square Error.

Model Fold	**R^2**	**RMSE**	**NRMSE**	**Percent (%) variance explained**
1	0.011	0.181	0.036	-3.48
2	0.004	0.414	0.083	-0.05
3	0.155	0.265	0.053	-4.57
4	0.001	0.676	0.135	1.66
5	0.098	0.228	0.046	-1.93
6	0.138	0.189	0.038	-3.56
7	0.126	0.129	0.026	-3.97
8	0.202	0.360	0.072	-6.99
9	0.066	0.462	0.092	-3.73
10	0.088	0.188	0.038	-3.64
Mean	**0.089**	**0.309**	**0.062**	**-3.03**
SD	**0.069**	**0.169**	**0.034**	**2.41**

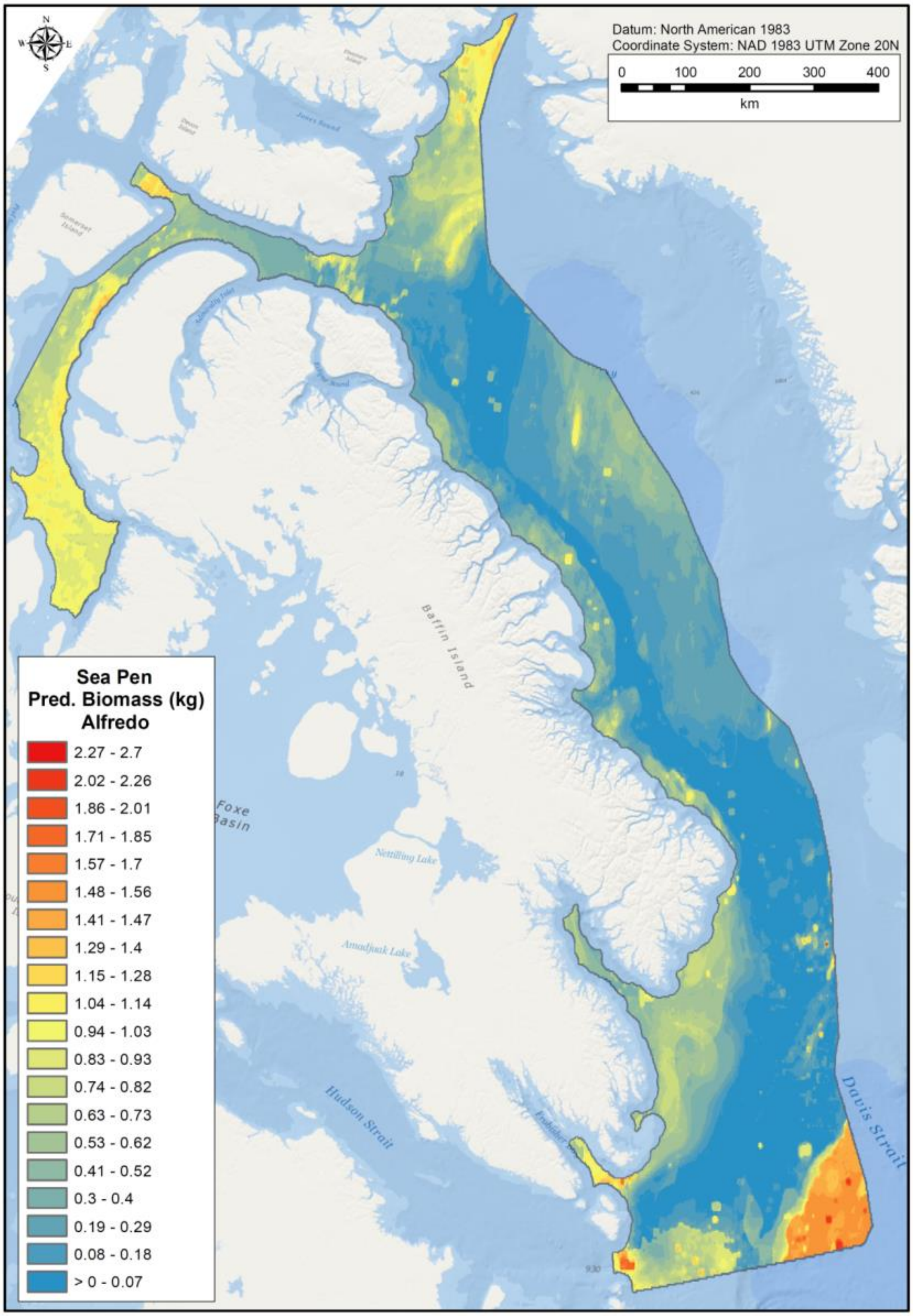

Figure 60. Predictions of biomass (kg) of sea pens from catch data recorded in DFO multispecies trawl surveys conducted using Alfredo trawl gear in the Eastern Arctic Region between 1999 and 2014.

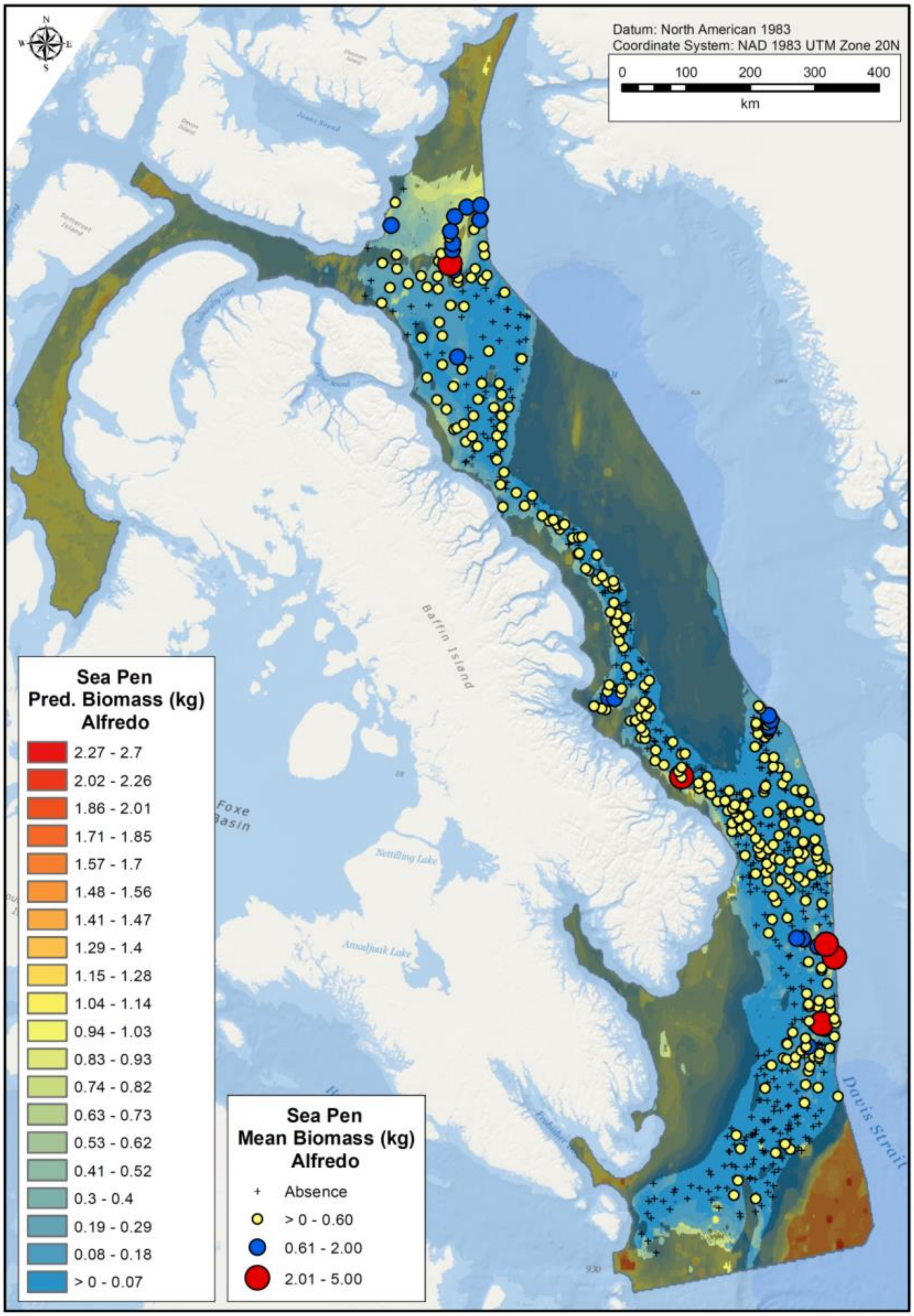

Figure 61. Predictions of biomass (kg) of sea pens from catch data recorded in DFO multispecies trawl surveys conducted using Alfredo trawl gear in the Eastern Arctic Region between 1999 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting sea pen biomass are shown in Figure 62. Summer Chlorophyll *a* Mean was the most important variable for predicting the biomass of the sea pen catch data from Alfredo trawl gear in the Eastern Arctic. This variable was followed more distantly by Summer Chlorophyll *a* Range and the remaining variables in the model. The partial dependence plots of the random forest model on sea pen catch records from Alfredo trawl gear are shown in Figure 63. Predicted biomass was highest at Summer Chlorophyll *a* Mean values greater than 0.8 mg m^{-3}.

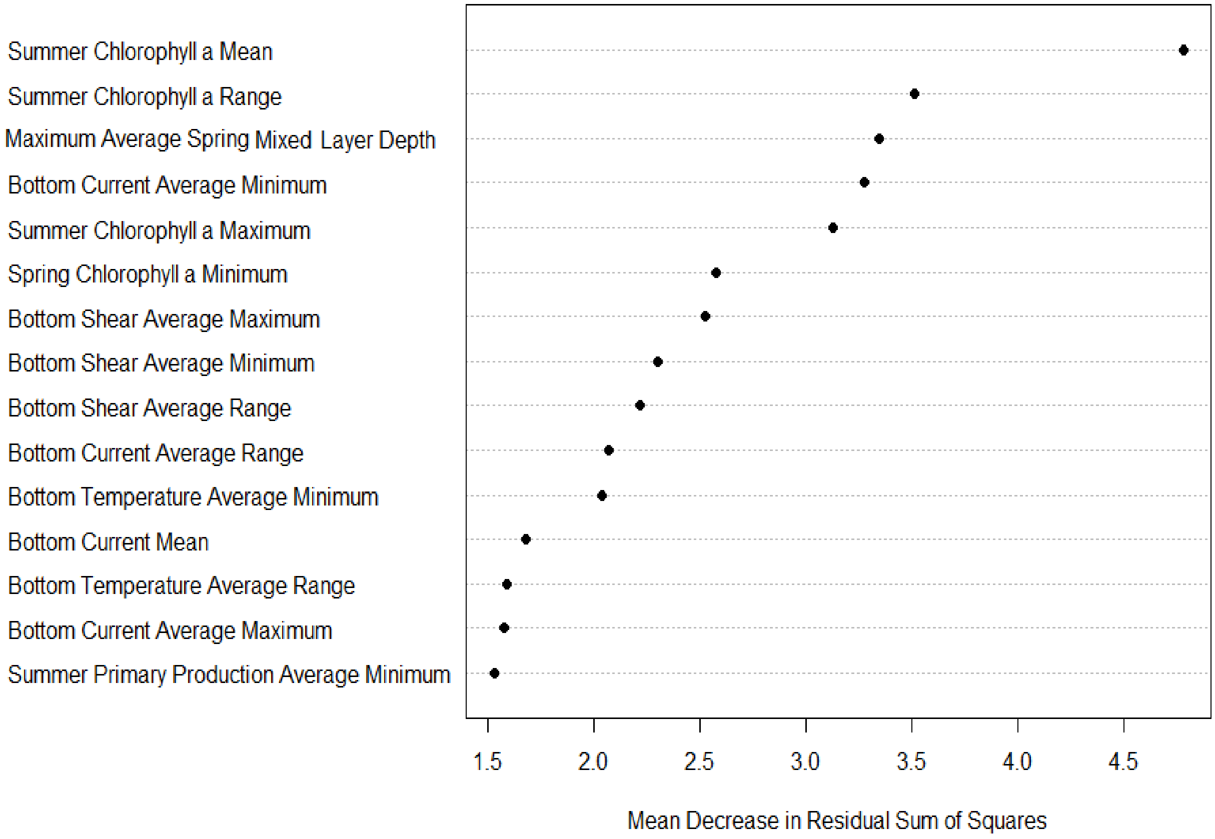

Figure 62. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on sea pen biomass data collected from DFO trawl surveys conducted using Alfredo trawl gear. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

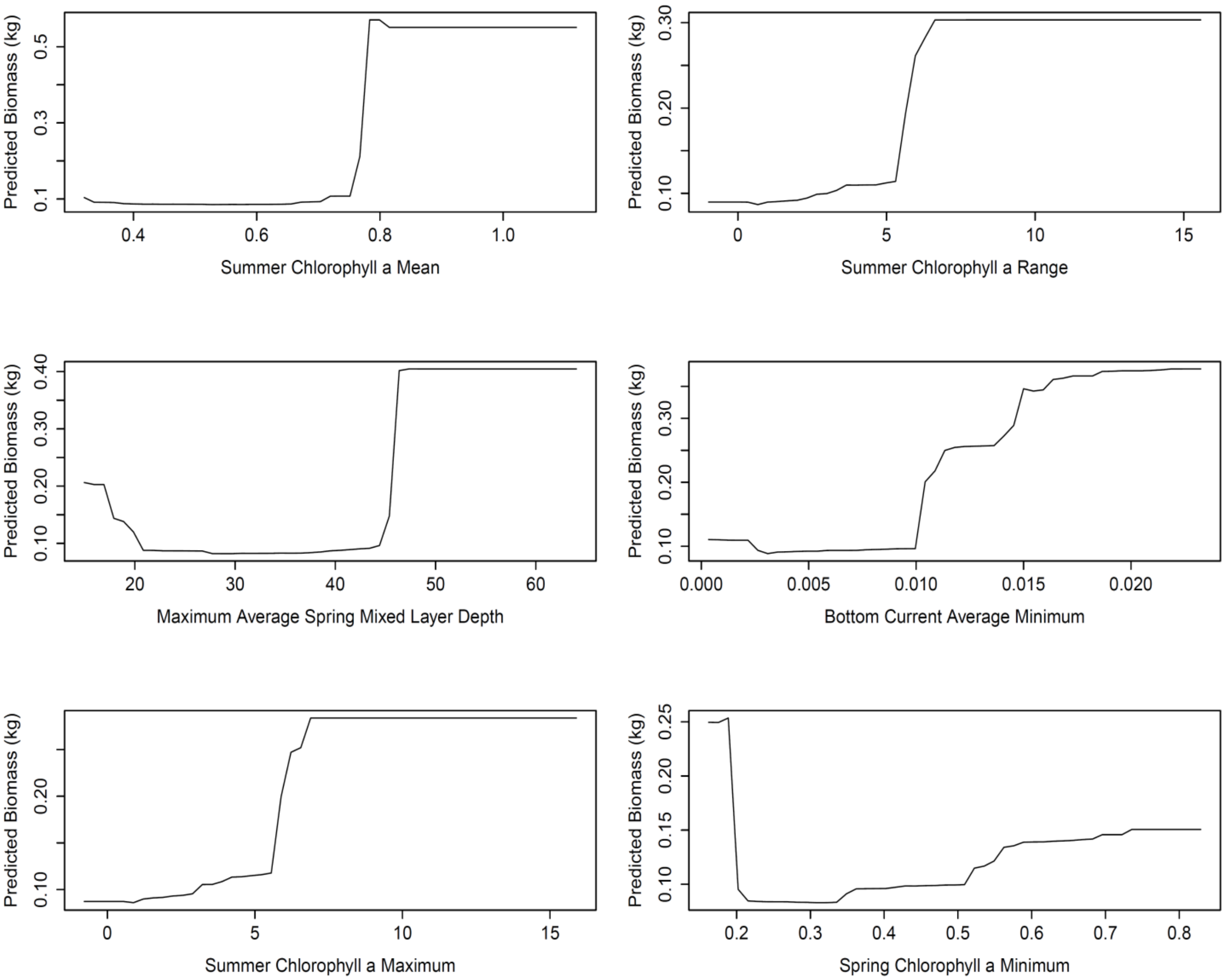

Figure 63. Partial dependence plots of the top six predictors from the random forest model of sea pen biomass data collected from DFO trawl surveys using Alfredo trawl gear in the Eastern Arctic Region, ordered from left to right from the top.

Campelen Trawl Gear

Accuracy measures from the regression random forest model on mean sea pen biomass records from trawl surveys using Campelen trawl gear are presented in Table 17. The highest R^2 value was 0.202 while the average was 0.041 ± 0.062 SD, indicating poor model performance. The average Normalized Root-Mean-Square-Error (NRMSE) was 0.042 ± 0.025 SD. The average percent variance explained by the model was -8.99 ± 3.96 SD.

Figures 64 and 65 show the predicted biomass surface of sea pens using mean biomass data from Campelen trawl surveys. The majority of the spatial extent was predicted to have low (< 0.039 kg) biomass of sea pens. The highest predicted biomass occurred in a very small area in the Davis Strait where there are no data observations (Figure 65). In general, the deep waters off Baffin Island Shelf were predicted to have moderate to high biomass of sea pens. This area, including Lancaster Sound and the Gulf of Boothia, was considered extrapolated area (Figure 65). The southeast corner of the study extent in Davis Strait was also considered extrapolated area.

Table 17. Accuracy measures for all 10 model repetitions of 10-fold cross validation of the random forest model of average of sea pen biomass (kg) per grid cell recorded from DFO trawl surveys conducted using Campelen trawl gear the Eastern Arctic Region. RMSE= Root-Mean-Square Error, NRMSE= Normalized Root-Mean-Square Error.

Model Fold	**R^2**	**RMSE**	**NRMSE**	**Percent (%) variance explained**
1	0.074	0.017	0.020	-11.44
2	0.202	0.022	0.026	-10.20
3	0.005	0.071	0.084	-4.38
4	0.001	0.043	0.052	-2.36
5	0.025	0.018	0.022	-12.18
6	0.010	0.072	0.085	-12.63
7	0.029	0.040	0.048	-14.08
8	0.058	0.022	0.026	-9.79
9	0.003	0.029	0.035	-4.76
10	4.723×10^{-4}	0.017	0.020	-8.06
Mean	**0.041**	**0.035**	**0.042**	**-8.99**
SD	**0.062**	**0.021**	**0.025**	**3.96**

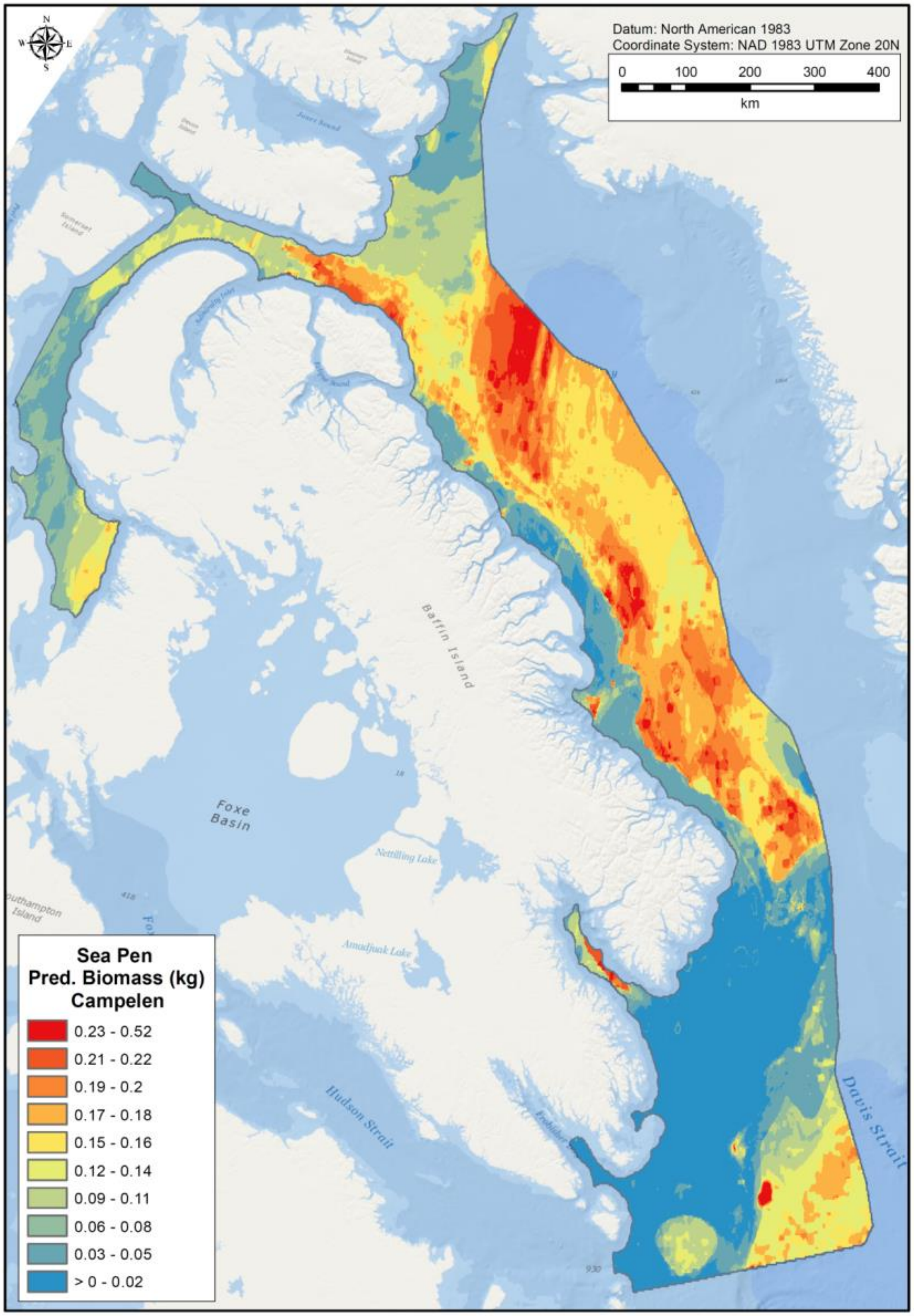

Figure 64. Predictions of biomass (kg) of sea pens from catch data recorded in DFO multispecies trawl surveys conducted using Campelen trawl gear in the Eastern Arctic Region between 2005 and 2014.

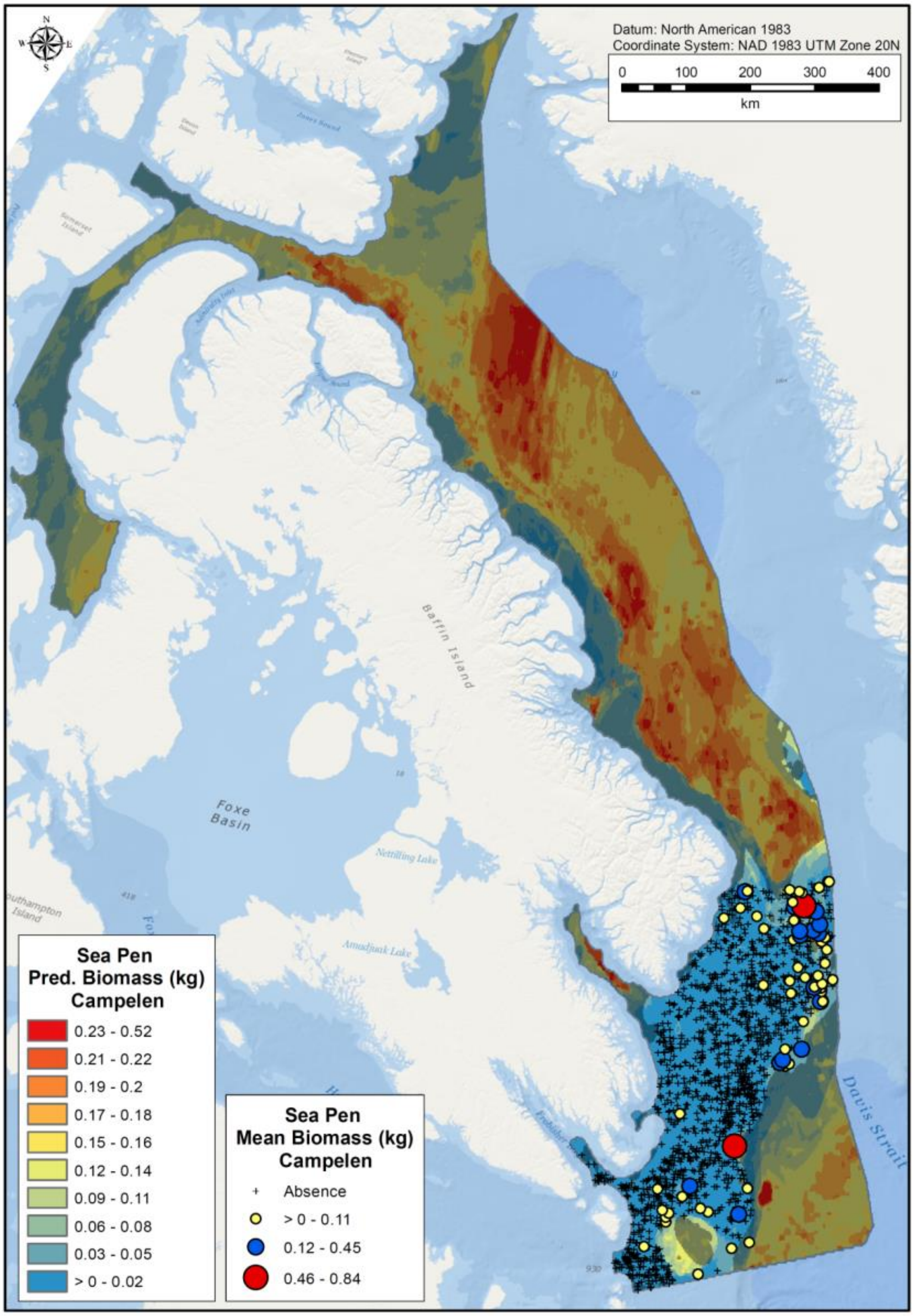

Figure 65. Predictions of biomass (kg) of sea pens from catch data recorded in DFO multispecies trawl surveys conducted using Campelen trawl gear in the Eastern Arctic Region between 2005 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting sea pen biomass are shown in Figure 66. Unlike the model on Alfredo trawl records, Depth was the most important variable in this model using Campelen trawl records. This was followed by Spring Chlorophyll *a* Minimum and Summer Primary Production Average Minimum. The partial dependence plots of the top six predictor variables are shown in Figure 67. Predicted biomass was highest between 700 and 800 m depth. In general, predicted biomass was highest at the lowest spring chlorophyll *a* and summer primary production values.

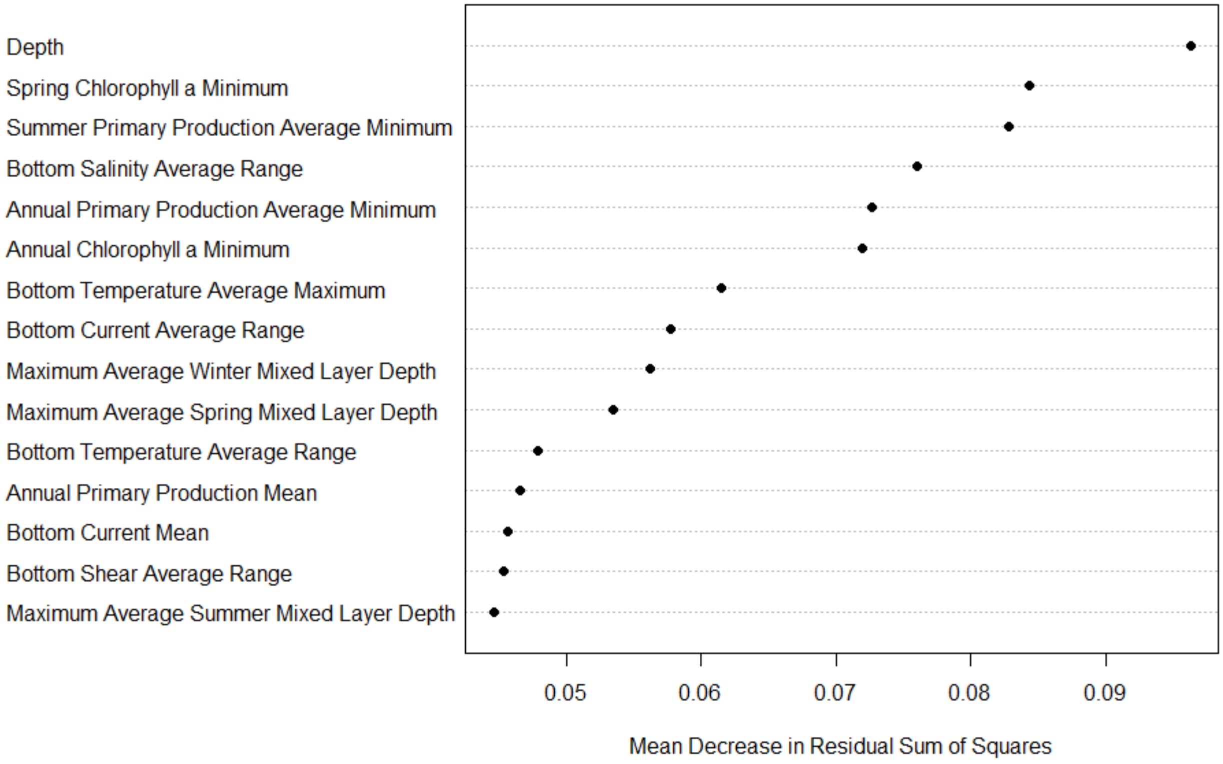

Figure 66. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on sea pen biomass data collected from DFO trawl surveys conducted using Campelen trawl gear. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

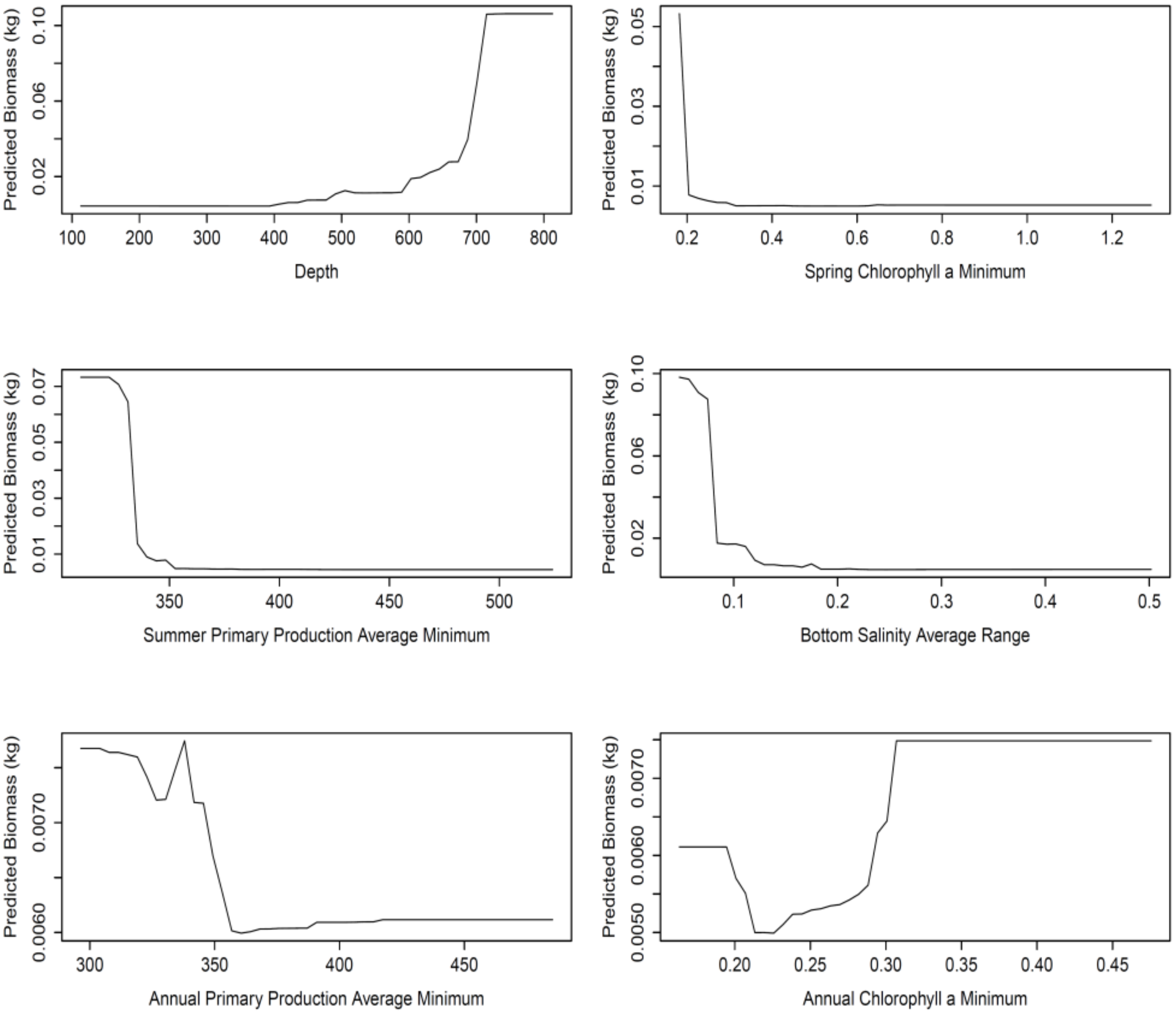

Figure 67. Partial dependence plots of the top six predictors from the random forest model of sea pen biomass data collected from DFO trawl surveys using Campelen trawl gear in the Eastern Arctic Region, ordered from left to right from the top.

Cosmos Trawl Gear

Accuracy measures from the regression random forest model on mean sea pen biomass records from trawl surveys using Cosmos trawl gear are presented in Table 18. The highest R^2 value was 0.564 while the average was 0.087 ± 0.176 SD. The standard deviation is higher than the mean, indicating high variability between model folds. The average Normalized Root-Mean-Square-Error (NRMSE) was 0.101 ± 0.064 SD. The average percent variance explained by the model was -12.47 ± 3.43 SD.

Figures 68 and 69 show the predicted biomass surface of sea pens using mean biomass data from Cosmos trawl surveys. Most of Baffin Bay and Davis Strait were predicted to have low (< 0.03

kg) biomass of sea pens. Small pockets of high predicted biomass occurred in the Gulf of Boothia and The Prince Regent Strait, which connects the Gulf of Boothia to Lancaster Sound. These areas contained no data observations and were considered extrapolated area by the model (Figure 69). Small pockets of biomass were predicted off Hall Peninsula where the highest mean biomass catches of sea pens were recorded. Moderate biomass of sea pens was predicted to occur in the northern Baffin Bay and southeast Davis Strait. Given the poor spatial distribution of data observations from Cosmos gear, most of the study extent was considered extrapolated area by the model.

Table 18. Accuracy measures for all 10 model repetitions of 10-fold cross validation of the random forest model of average of sea pen biomass (kg) per grid cell recorded from DFO trawl surveys conducted using Cosmos trawl gear the Eastern Arctic Region. RMSE= Root-Mean-Square Error, NRMSE= Normalized Root-Mean-Square Error.

Model Fold	**R^2**	**RMSE**	**NRMSE**	**Percent (%) variance explained**
1	0.017	0.051	0.065	-11.27
2	0.040	0.064	0.081	-10.27
3	8.777×10^{-4}	0.086	0.111	-12.79
4	0.005	0.184	0.236	-4.69
5	0.006	0.035	0.045	-14.64
6	5.000×10^{-4}	0.049	0.063	-11.43
7	1.367×10^{-4}	0.032	0.041	-14.38
8	0.168	0.113	0.145	-13.20
9	0.066	0.039	0.050	-14.65
10	0.564	0.131	0.168	-17.39
Mean	**0.087**	**0.078**	**0.101**	**-12.47**
SD	**0.176**	**0.050**	**0.064**	**3.43**

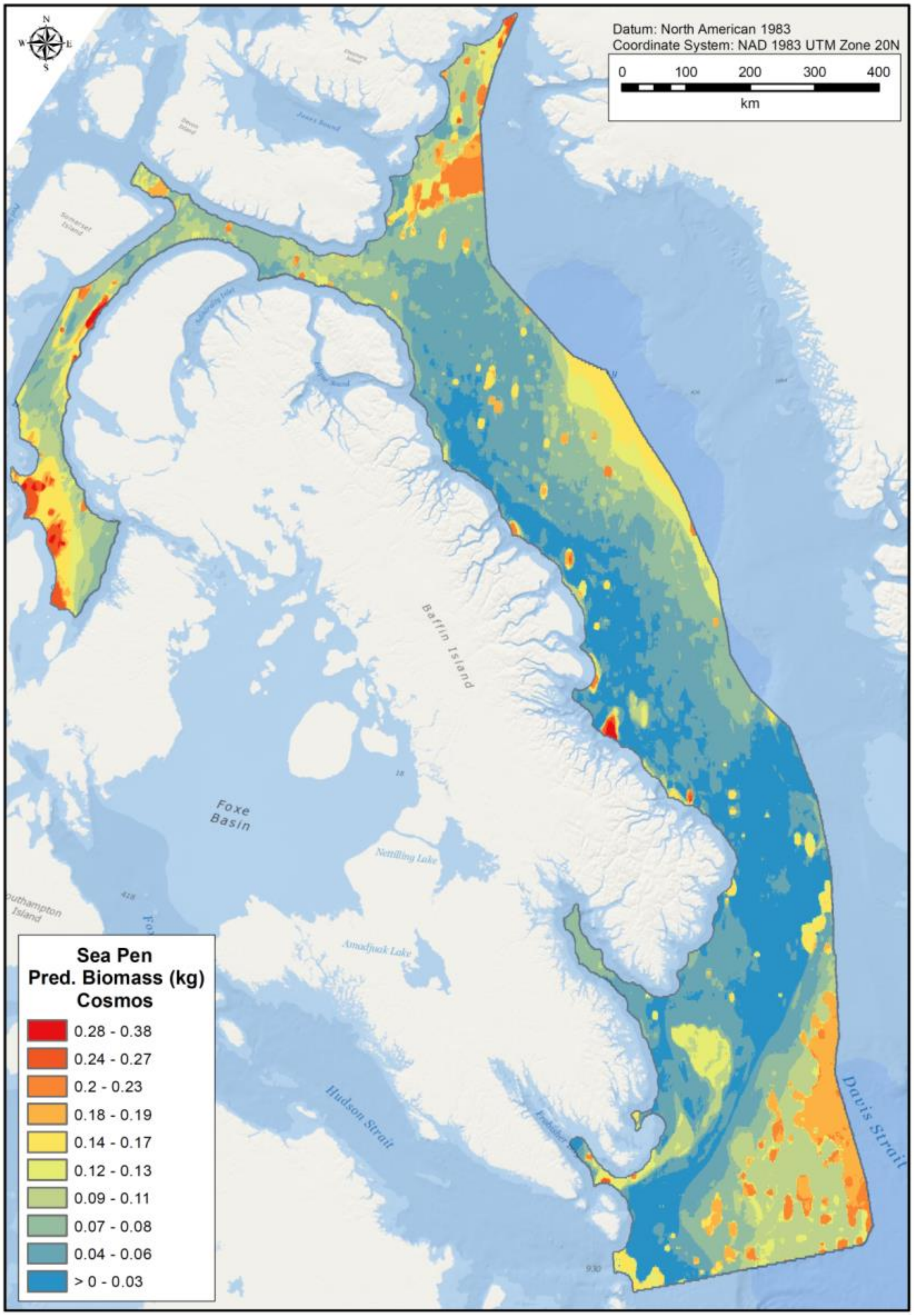

Figure 68. Predictions of biomass (kg) of sea pens from catch data recorded in DFO multispecies trawl surveys conducted using Cosmos trawl gear in the Eastern Arctic Region between 2006 and 2012.

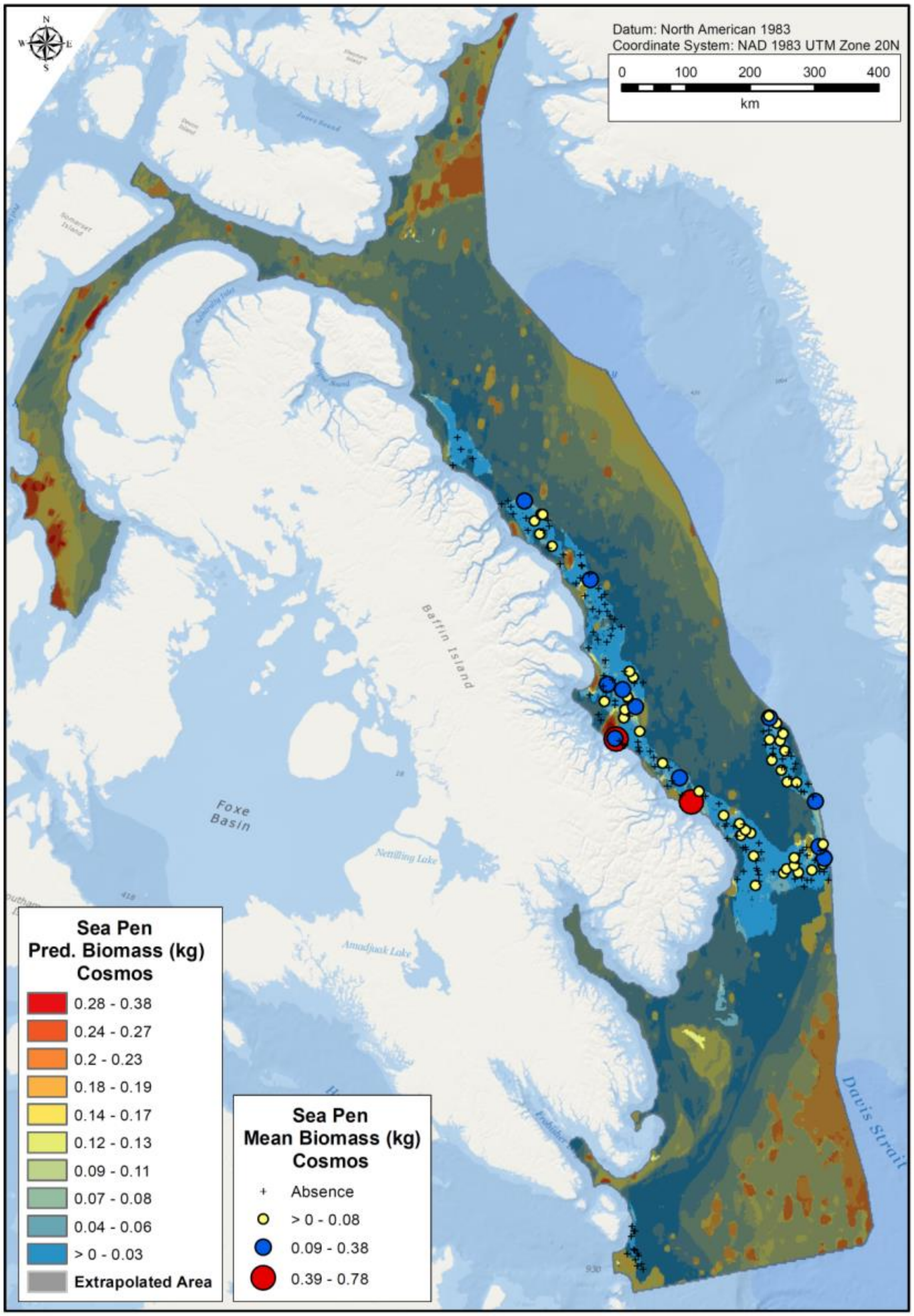

Figure 69. Predictions of biomass (kg) of sea pens from catch data recorded in DFO multispecies trawl surveys conducted using Cosmos trawl gear in the Eastern Arctic Region between 2006 and 2012. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting sea pen biomass are shown in Figure 70. Annual Chlorophyll *a* Range was the most important variable in this model. This variable was followed distantly by Bottom Salinity Average Range, Annual Chlorophyll *a* Maximum, Depth, and the remaining variables in the model. The partial dependence plots of the top six predictor variables are shown in Figure 71. Along the gradient in Annual Chlorophyll *a* Range, predicted biomass was highest between 6 and 10 mg m^{-3}. Along the Depth gradient, predicted biomass was high at shallow depths, likely coinciding with the Baffin Island Shelf, decreased and then increased again between 600 to 800 m, coinciding with deeper waters in Baffin Bay and Davis Strait.

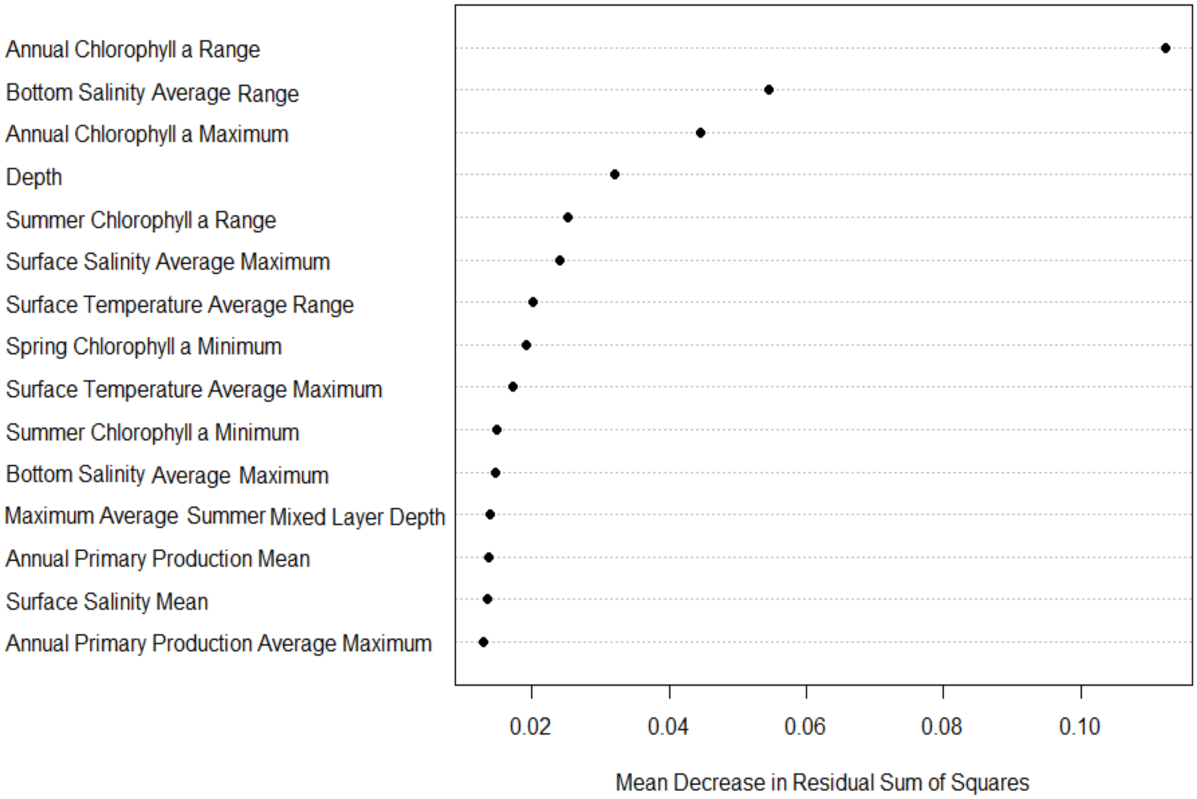

Figure 70. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on sea pen biomass data collected from DFO trawl surveys conducted using Cosmos trawl gear. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

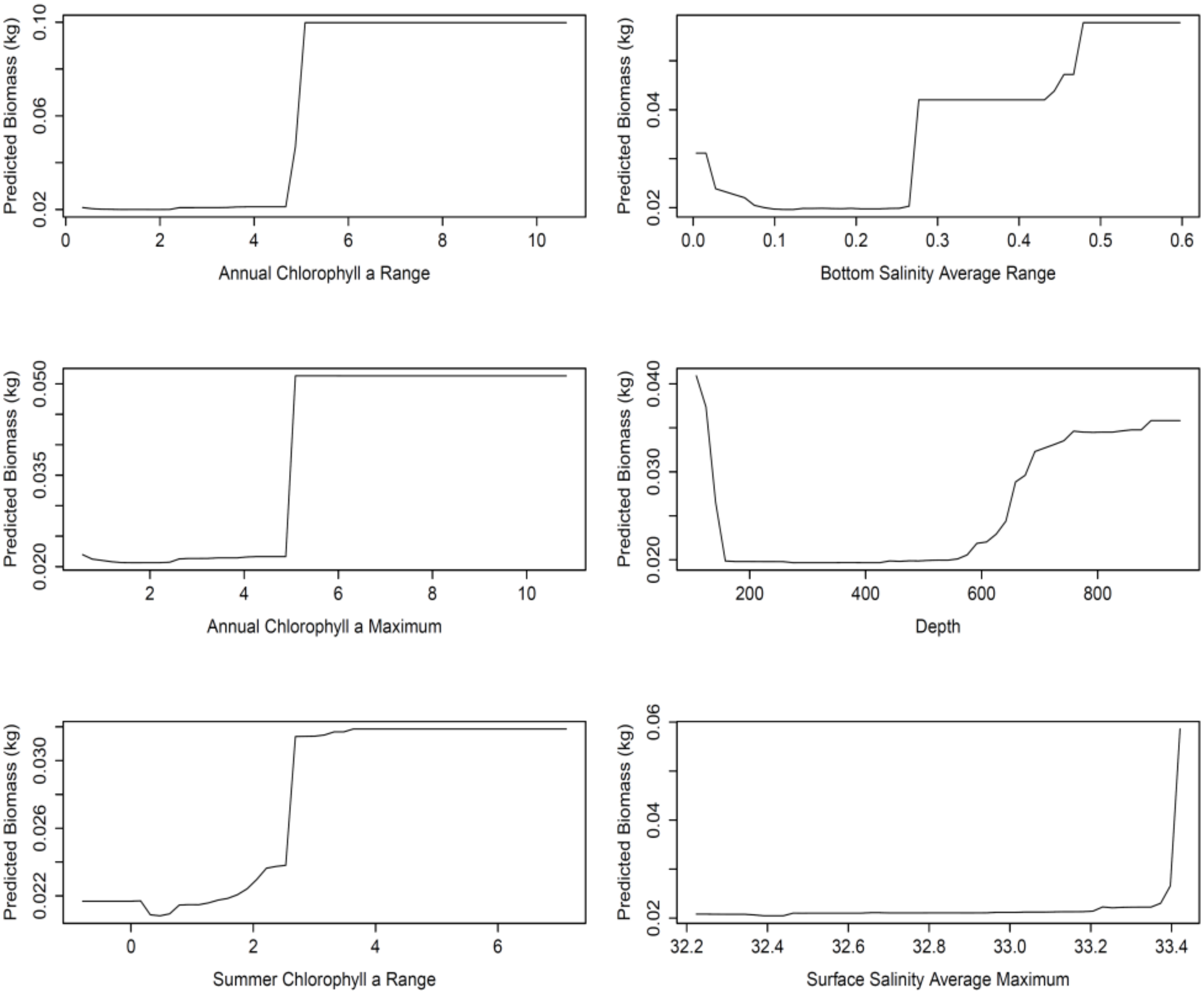

Figure 71. Partial dependence plots of the top six predictors from the random forest model of sea pen biomass data collected from DFO trawl surveys using Cosmos trawl gear in the Eastern Arctic Region, ordered from left to right from the top.

Large Gorgonian Corals

Data Sources and Distribution

Large gorgonian coral catch data for the Eastern Arctic Region was collected between 1999 and 2014 and consisted of 39 presences and 698 absences from Paamuit surveys conducted using Alfredo trawl gear, 115 presences and 1329 absences from the *Cape Ballard*, *Aqviq*, and *Kinguk* surveys conducted using Campelen gear, and 1 presence and 185 absences from *Paamuit* surveys using Cosmos gear (Table 19). Alfredo trawl records had the widest spatial distribution across the study extent (Figure 72). Campelen records were restricted to the Davis Strait, while Cosmos records were distributed along the coast of Baffin Island and on the eastern edge of the study extent in Davis Strait. Presence-absence random forest models were generated on the combined dataset consisting of 155 presences and 2212 absences (see Figure 73). The highest mean biomass record (up to 2000 kg) was recorded in the Narwhal Over-wintering and Deep-Sea Coral Conservation Area.

Table 19. Number of presence and absence records of large gorgonian coral catch by gear type recorded from DFO trawl surveys conducted between 1999 and 2014 in the Eastern Arctic.

	Alfredo		**Campelen**		**Cosmos**	
Year	**Presences**	**Absences**	**Presences**	**Absences**	**Presences**	**Absences**
1999	1	-	-	-	-	-
2000	-	-	-	-	-	-
2001	-	-	-	-	-	-
2005	-	-	6	145	-	-
2006	-	60	27	115	-	75
2007	-	-	5	128	-	13
2008	2	77	7	136	-	65
2009	-	17	4	140	-	-
2010	5	111	11	134	-	22
2011	6	76	20	130	-	-
2012	1	160	21	129	1	10
2013	18	67	11	140	-	-
2014	6	130	3	132	-	-
TOTAL	39	698	115	1329	1	185

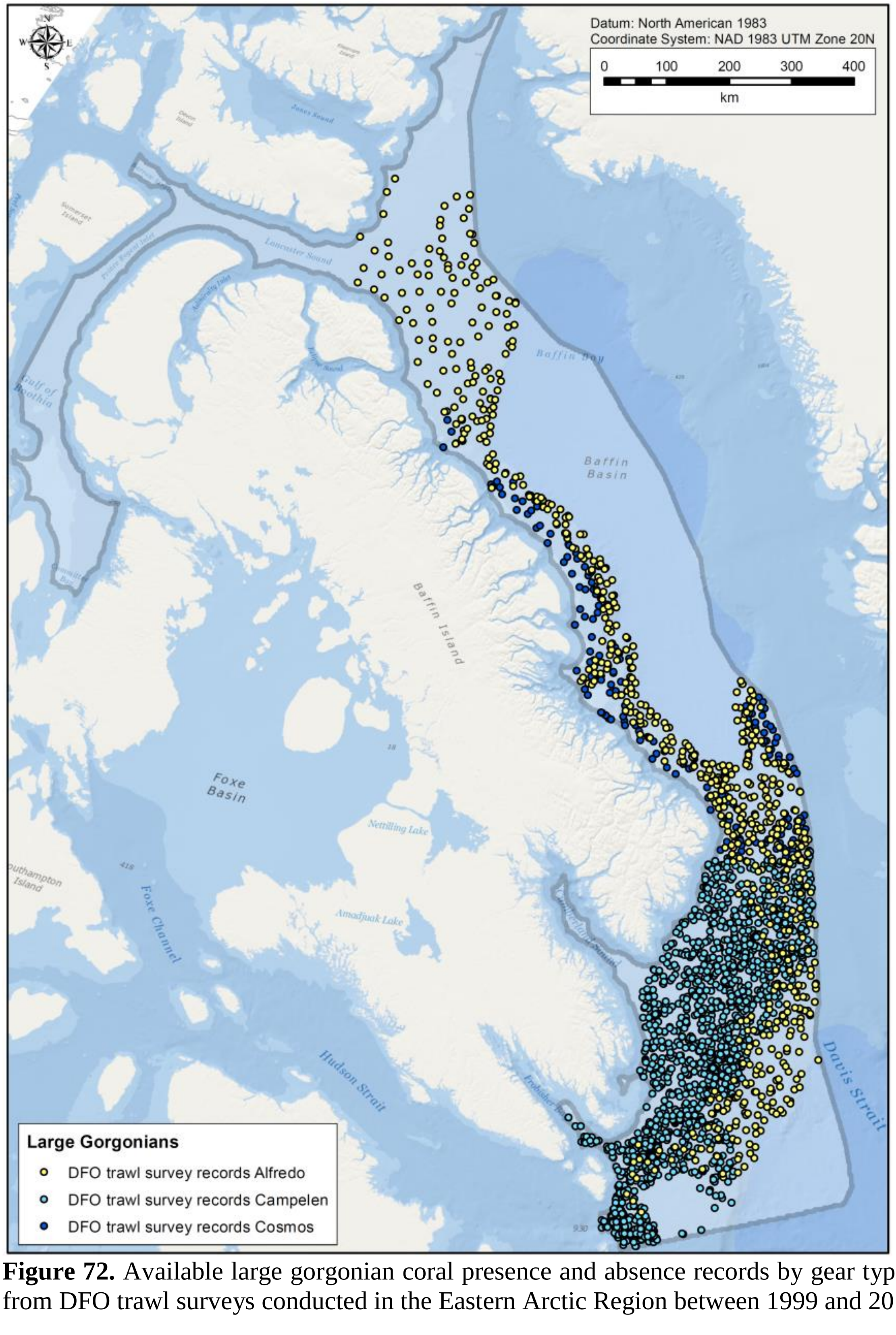

Figure 72. Available large gorgonian coral presence and absence records by gear type collected from DFO trawl surveys conducted in the Eastern Arctic Region between 1999 and 2014.

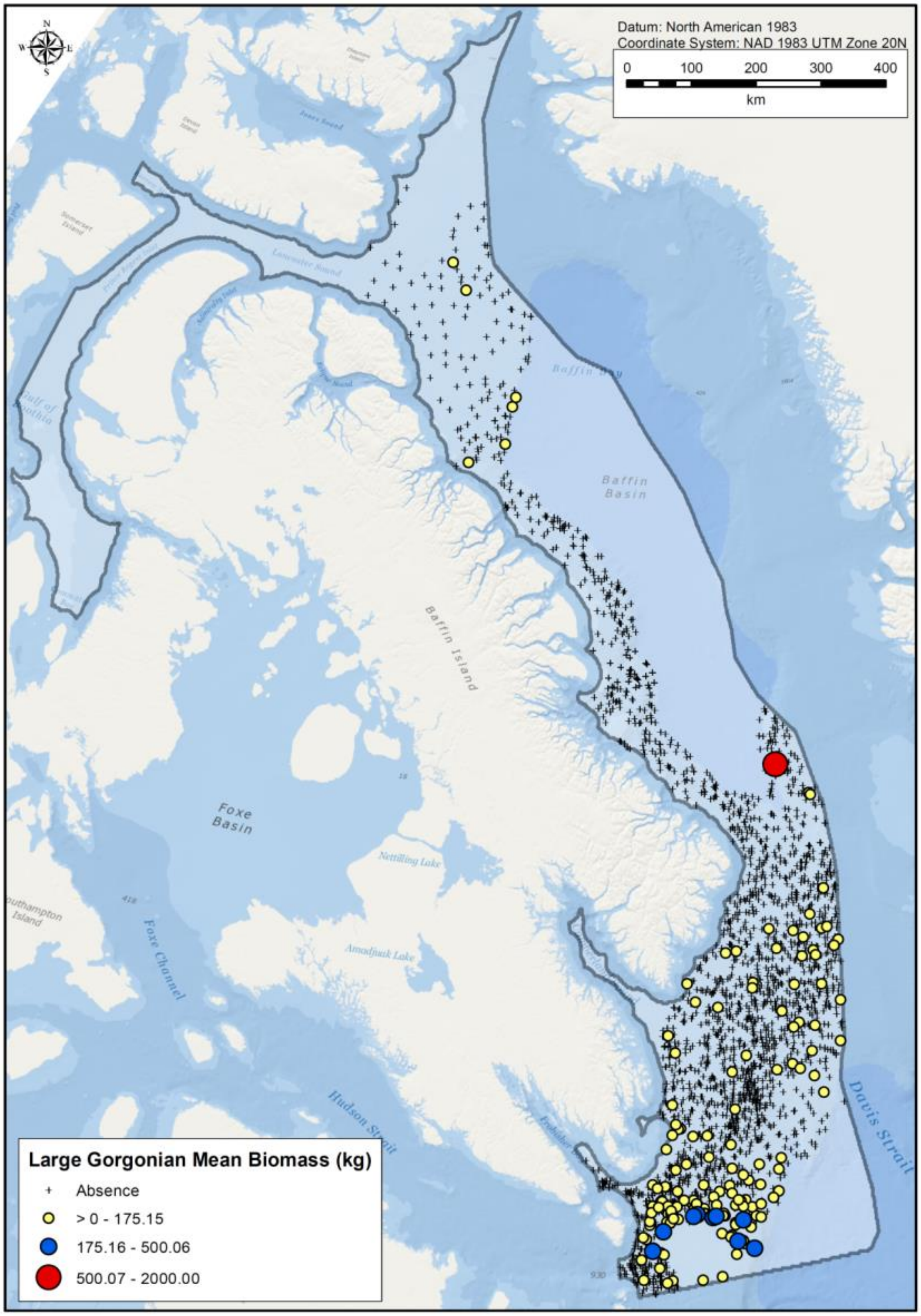

Figure 73. Mean biomass (kg) per grid cell of large gorgonian coral catch recorded from DFO trawl surveys conducted in the Eastern Arctic Region between 1999 and 2014.

Model 1 – Balanced Species Prevalence

Accuracy measures for the random forest model on balanced species prevalence (179 presences and 179 absences; Model 1) are presented in Table 20. The highest mean AUC of 0.801 was associated with Model Run 4 and is therefore considered the optimal model for the prediction of the large gorgonian coral response data. The sensitivity and specificity measures of this model fold were 0.677 and 0.787, respectively. In general, sensitivity was much lower than specificity for all model folds. The confusion matrix of the optimal model is also presented in Table 20. Class error for both the presence and absence classes was moderate.

Table 20. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of large gorgonian corals within the Eastern Arctic Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 4) which is considered the optimal model for predicting the presence probability of large gorgonian corals in the region.

Model Run	**AUC**	**Sensitivity**	**Specificity**
1	0.768	0.652	0.768
2	0.784	0.652	0.761
3	0.721	0.632	0.716
4	**0.801**	**0.677**	**0.787**
5	0.770	0.632	0.729
6	0.775	0.671	0.787
7	0.775	0.652	0.742
8	0.786	0.684	0.703
9	0.787	0.632	0.755
10	0.722	0.581	0.690
Mean	**0.769**	**0.646**	**0.744**
SD	**0.027**	**0.030**	**0.034**

Confusion matrix of model with highest AUC:

Observations	**Predictions**		**Total n**	**Class error**
	Absence	**Presence**		
Absence	122	33	155	0.213
Presence	50	105	155	0.323

The presence probability prediction surface of large gorgonian corals is presented in Figure 74. Shallower waters in Baffin Bay were predicted as absence of large gorgonian corals. The largest area of high presence probability occurred in southwest Davis Strait and along a narrow band running southwest to northeast in Davis Strait. These areas of high presence probability corresponded well with the distribution of presence observations (Figure 75), with little extrapolation beyond these locations except in southwest Davis Strait.

The actual presence and absence records selected for use in the optimal model fold of Model 1 (155 presences and 155 absences; Figure 76) shows some slight spatial bias across the study area. Random down-sampling of the absence records in Davis Strait likely contributed to the higher predictions of presence probability there. Also shown in this figure are areas of model extrapolation. Northern Baffin Bay, Lancaster Sound, the Gulf of Boothia, deep waters off Baffin Island Shelf, and the southeast corner of the study extent were all considered areas of extrapolation by the model.

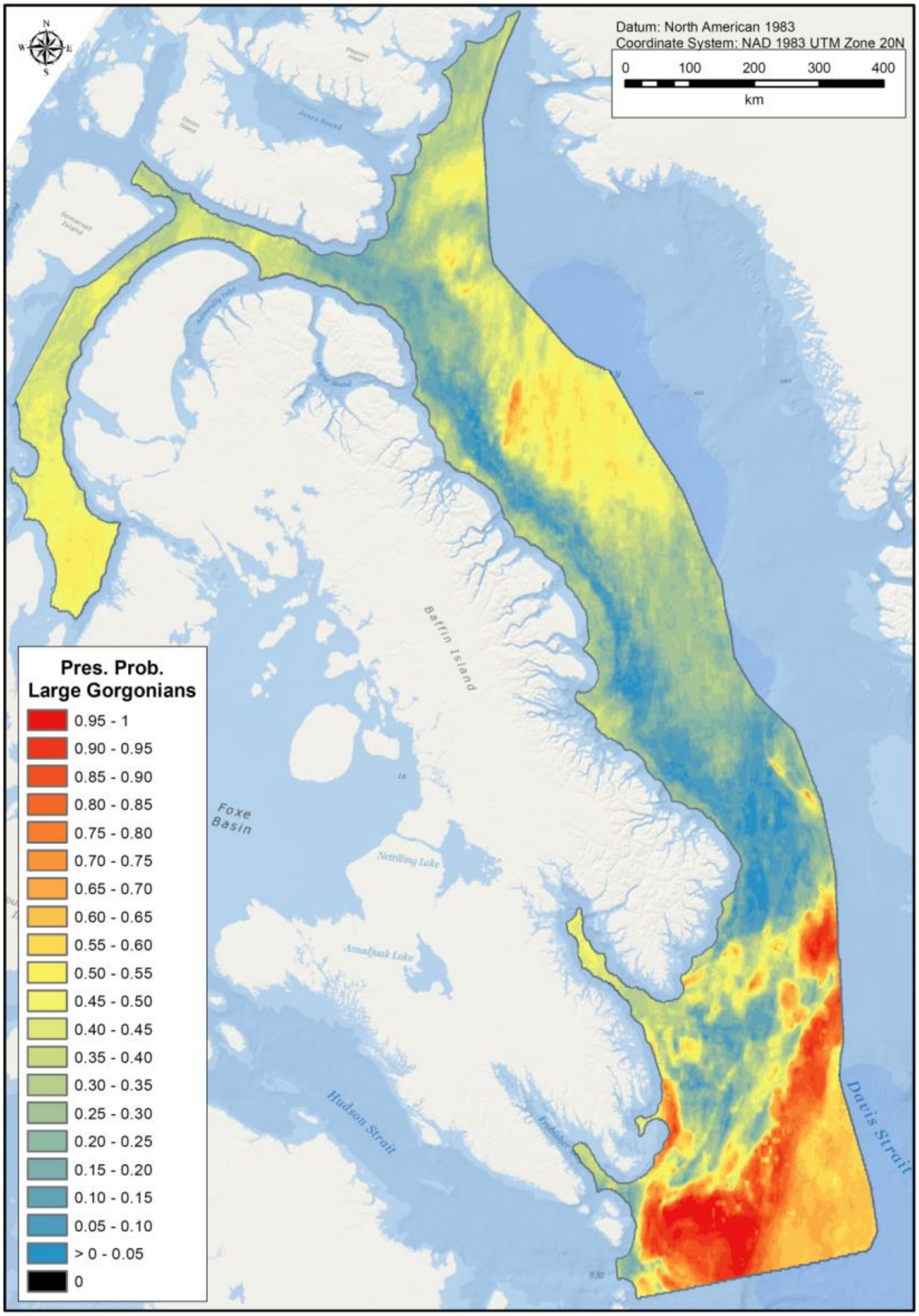

Figure 74. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of large gorgonian coral presence and absence data collected from DFO trawl surveys in the Eastern Arctic Region between 1999 and 2014.

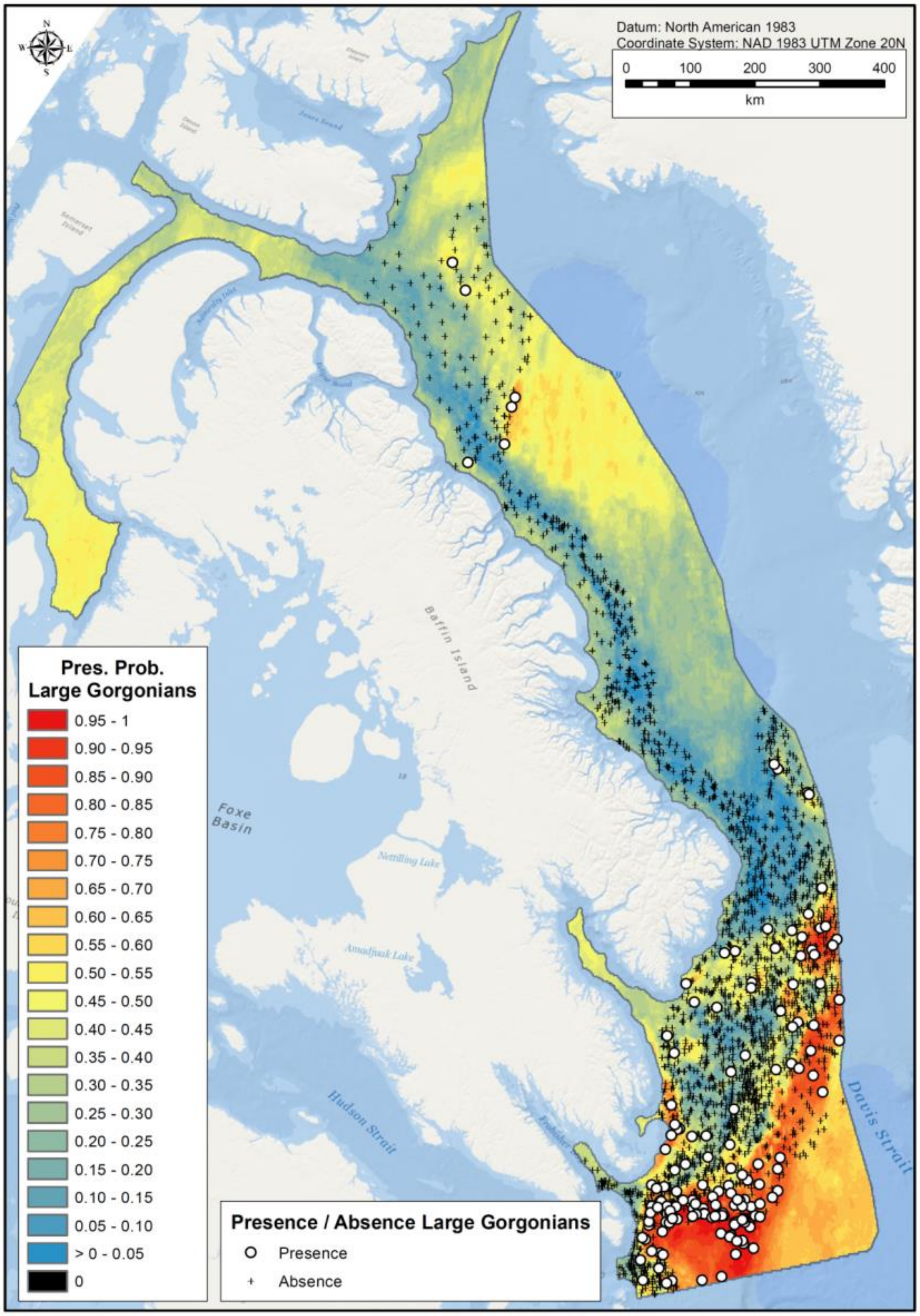

Figure 75. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of large gorgonian coral presence and absence data collected from DFO trawl surveys in the Eastern Arctic Region between 1999 and 2014.

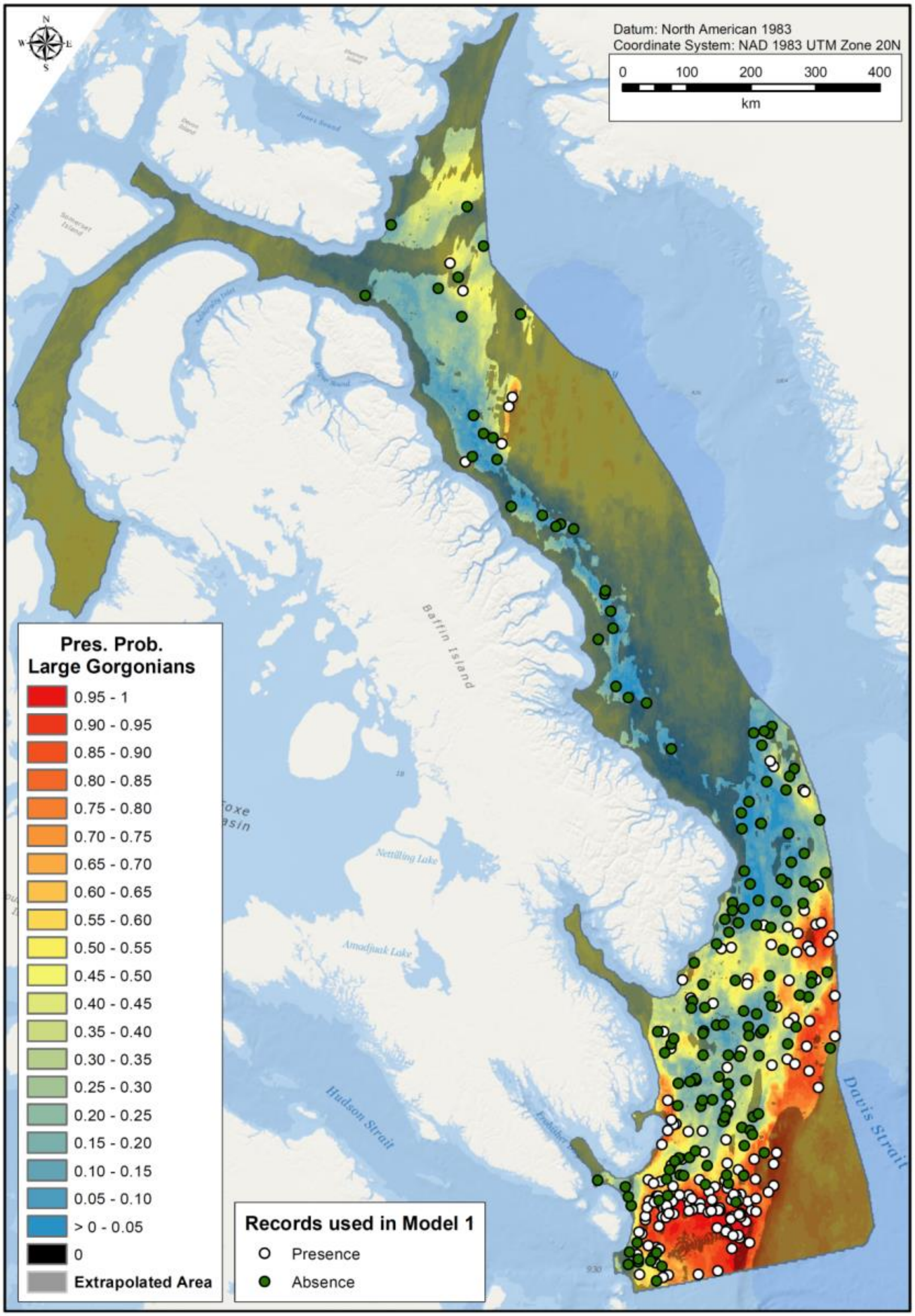

Figure 76. Map of the 310 data observations (155 presences and 155 absences) of large gorgonian corals used in the optimal random forest model on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of large gorgonian corals and areas of model extrapolation.

Of the 54 environmental predictor variables used in the model, Bottom Salinity Mean was the most important for the classification of the large gorgonian presence and absence data in the Eastern Arctic Region (Figure 77). This variable was followed very closely by Annual Chlorophyll a Mean. The partial dependence plots for the top six predictor variables are shown in Figure 78. Presence probability of large gorgonians was highest at high values (> 34.5) along the gradient in Bottom Salinity Mean. Along the Annual Chlorophyll *a* Mean gradient, presence probability was highest at ~0.75 mg m^{-3}.

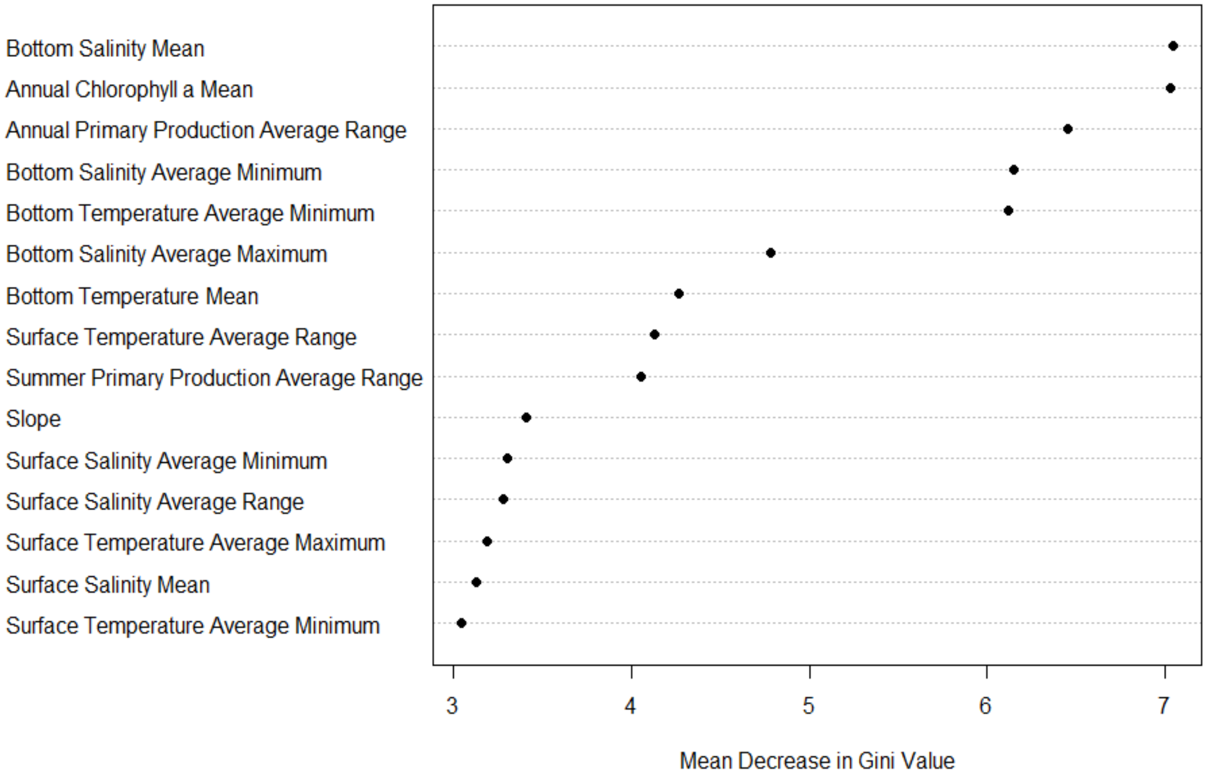

Figure 77. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model of large gorgonian coral presence and absence data collected from the Eastern Arctic Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

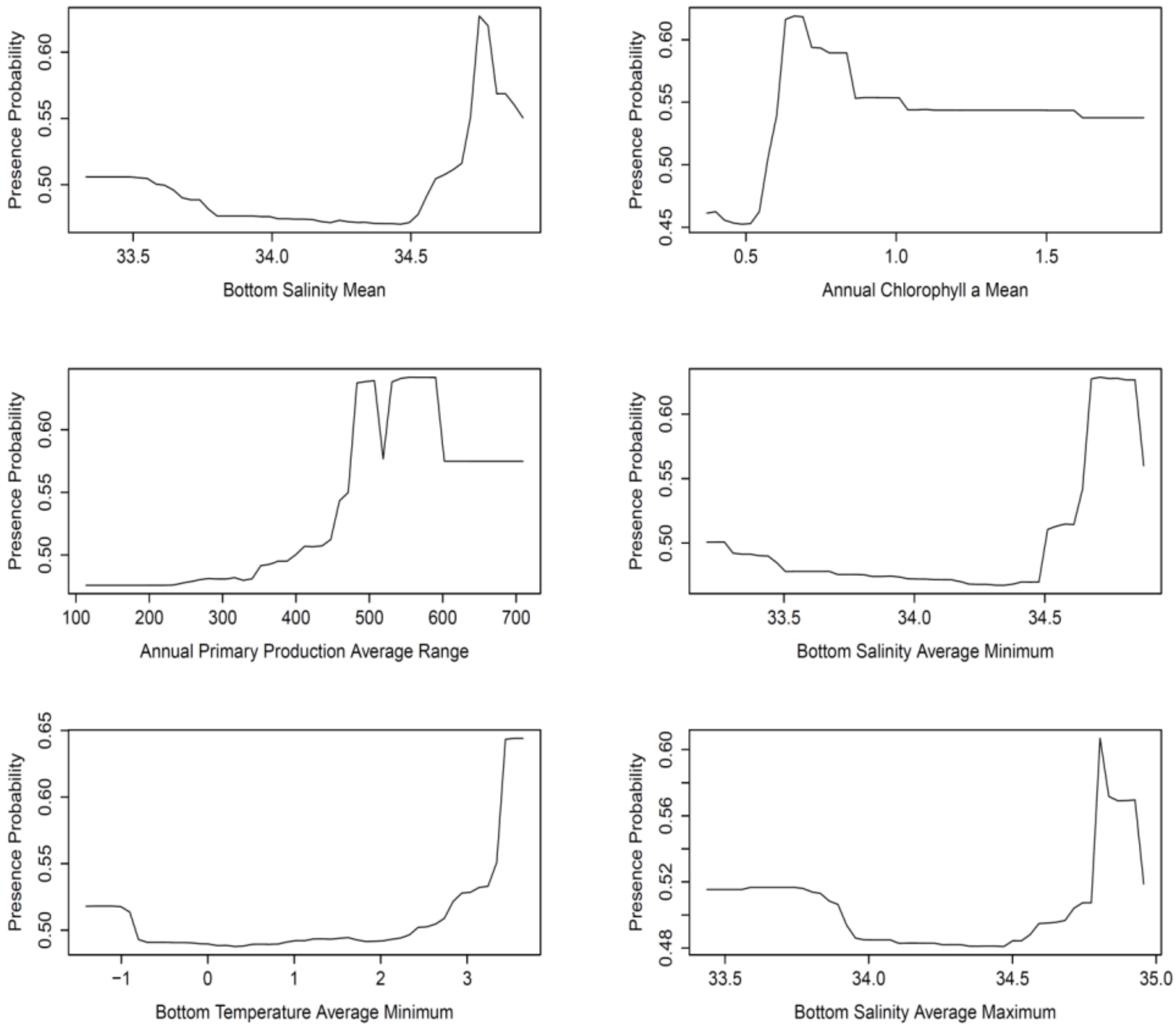

Figure 78. Partial dependence plots of the top six predictors from the optimal random forest model of large gorgonian coral presence and absence data collected in the Eastern Arctic Region, ordered from left to right from the top. Predicted presence probability is shown on the *y*-axis of each graph.

Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence

Table 21 shows the accuracy measures for the random forest model on all large gorgonian coral presence and absence data (2212 presences 155 absences) and a threshold equal to species prevalence (0.07). The mean AUC was slightly lower than that of Model 1 (0.769) and indicates good model performance. Sensitivity was slightly lower than that of Model 1 (0.626 versus 0.646 of Model 1), while specificity was slightly higher (0.786 versus 0.744 of Model 1). Class error for the presence class was moderate in this model.

The surface of predicted of presence probability of large gorgonian corals generated from Model 2 is presented in the Figure 79. Overall, areas of high presence probability are much reduced in this model compared to Model 1. Like Model 1, the largest area of high presence probability occurred in southwest portion of the extent in Davis Strait. The southeast corner of the extent, and the deeper waters off Baffin Island Shelf were predicted to have moderate presence probability of large gorgonian corals.

Figure 80 shows the spatial distribution of the large gorgonian coral presence and absence data used in Model 2 over the predicted presence probability surface. Predicted presence probability was low in locations where a high number of presence observations occurred, particularly in Davis Strait. This could be due to the high overlap between presence and absence data points in those areas and the inclusion of all absence data in the model.

Figure 81 shows the classification of large gorgonian coral presence probability into presence and absence categories based on the prevalence threshold of 0.07. In this map, presence probabilities greater than 0.07 were classed as presence, while probabilities lower than 0.07 were classed as absence. With the exception of Lancaster Sound and the Gulf of Boothia, much of the shallow portion of the study extent in Baffin Bay and Davis Strait were classified as absence of large gorgonian corals. The deep waters in Baffin Basin and Davis Strait were predicted as presence of large gorgonian corals.

Table 21. Accuracy measures for unrepeated 10-fold cross validation of a random forest model of presence and absence of large gorgonian corals within the Eastern Arctic Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	**AUC**	**Observ.**	**Predictions**		**Total n**	**Class error**	**Sensit.**	**Specif.**
1	0.814		**Absence**	**Presence**				
2	0.879	**Absence**	1738	474	2212	0.214	0.626	0.786
3	0.690	**Presence**	58	97	155	0.374		
4	0.681							
5	0.726							
6	0.764							
7	0.572							
8	0.752							
9	0.790							
10	0.849							
Mean	**0.752**							
SD	**0.090**							

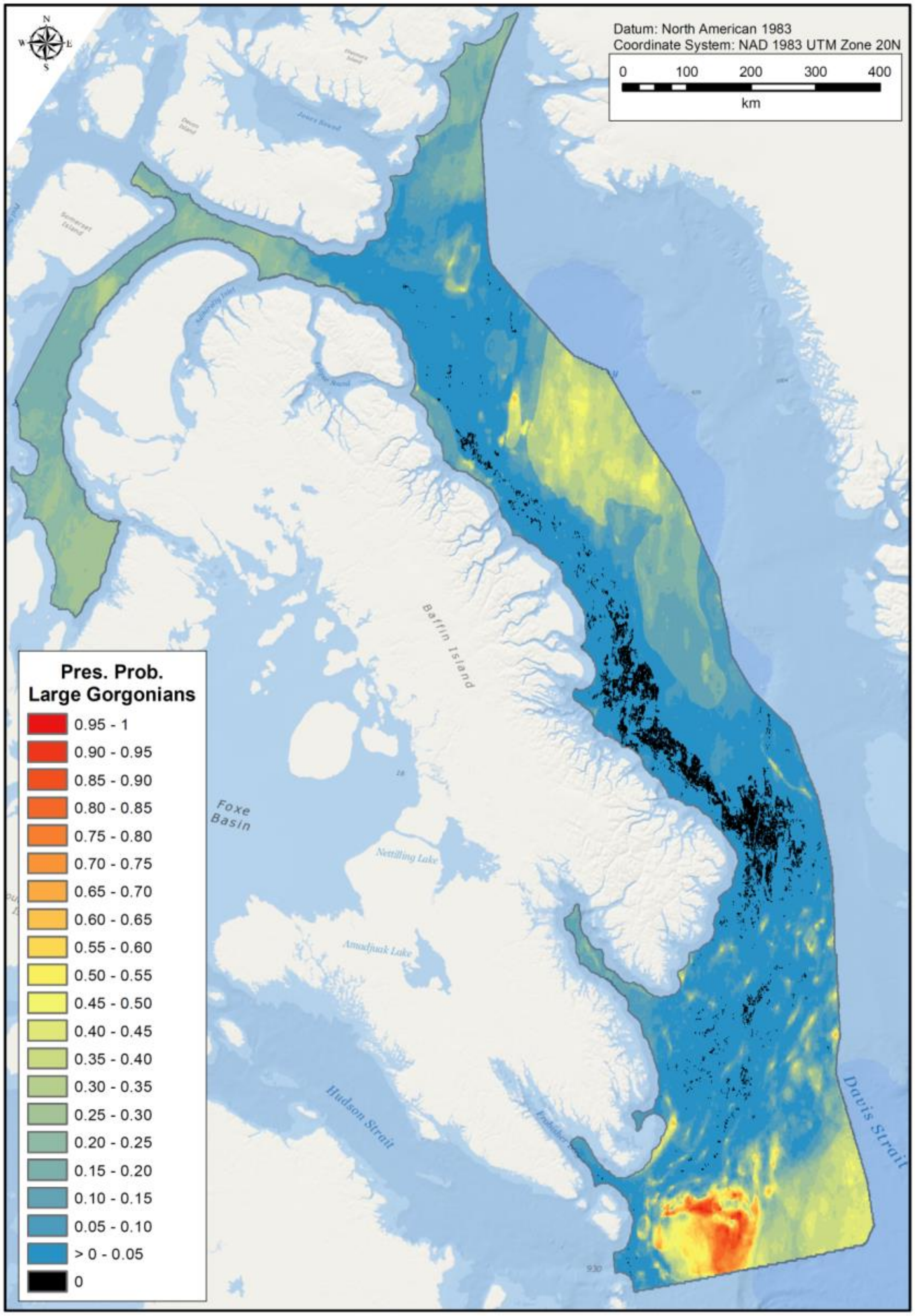

Figure 79. Prediction of presence probability (Pres. Prob.) of large gorgonian corals based on a random forest model on unbalanced presence and absence large gorgonian catch data collected from DFO trawl surveys conducted within the Eastern Arctic Region between 1999 and 2014.

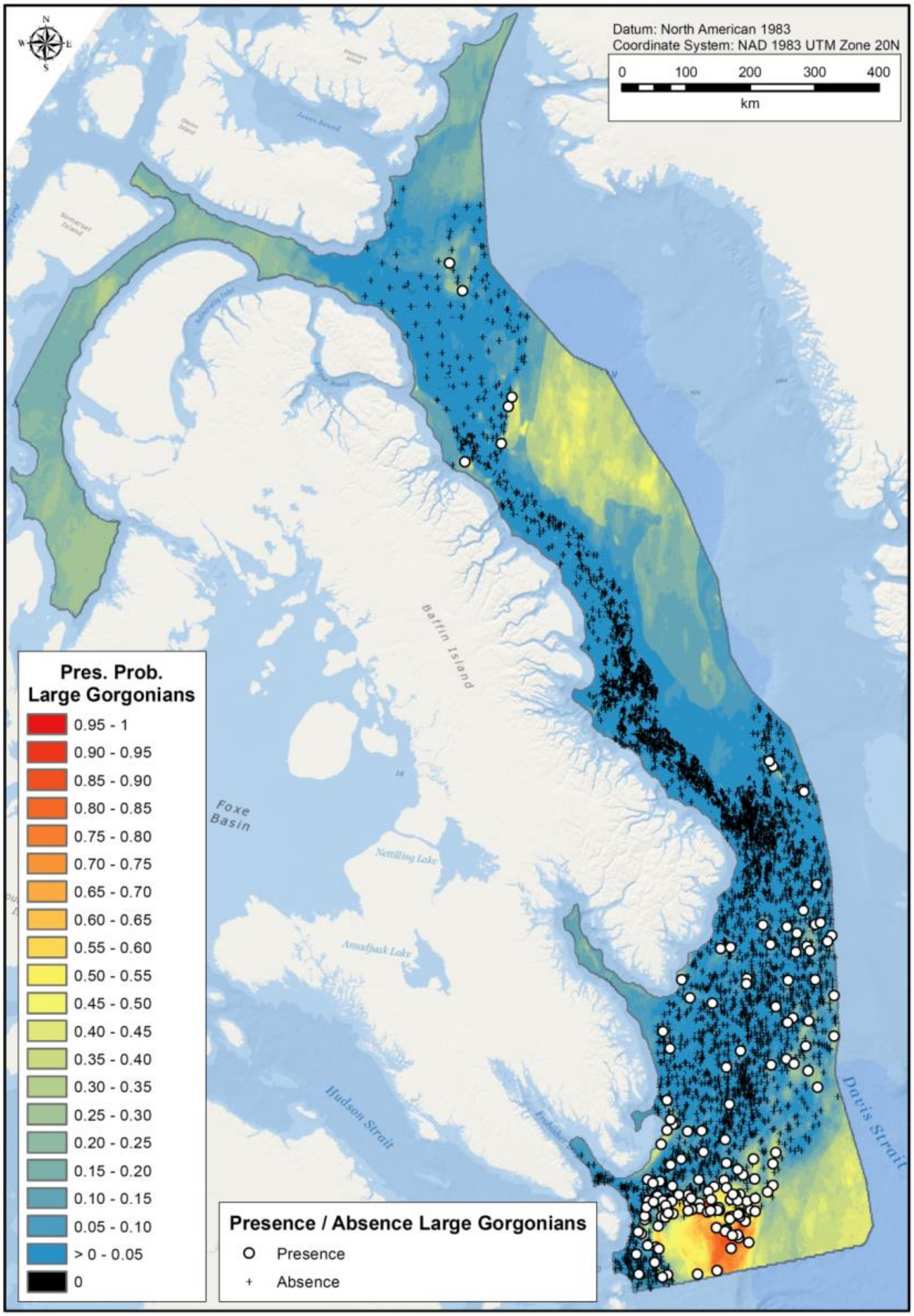

Figure 80. Presence and absence observations and prediction of presence probability (Pres. Prob.) of large gorgonian corals based on a random forest model on unbalanced presence and absence large gorgonian catch data collected from DFO trawl surveys conducted within the Eastern Arctic Region between 1999 and 2014.

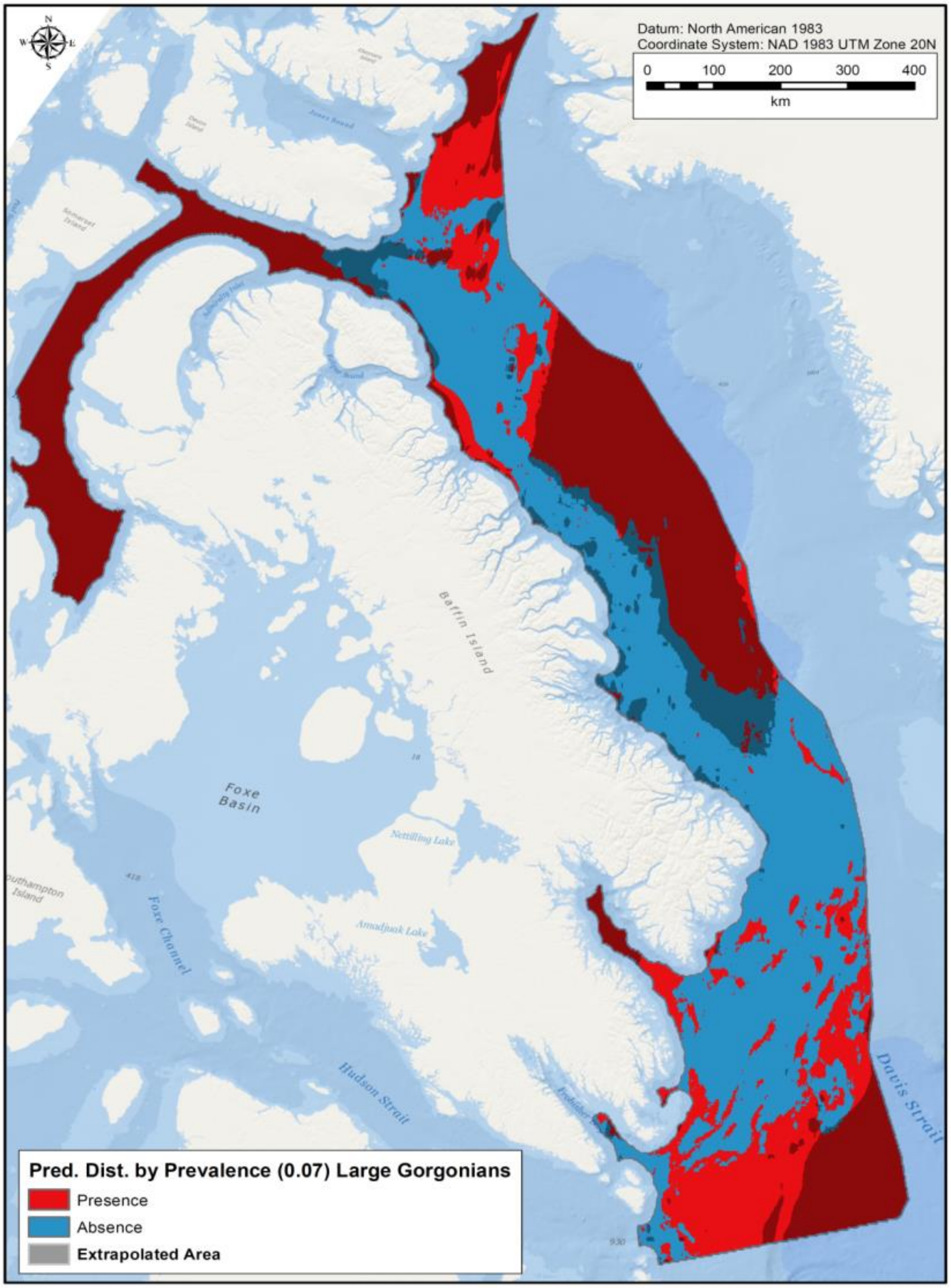

Figure 81. Predicted distribution (Pred. Dist.) of large gorgonian corals in the Eastern Arctic Region based on the prevalence threshold of 0.07 of large gorgonian presence and absence data used in Model 2. Also shown are areas of model extrapolation (grey polygon may appear red or blue).

The order of importance of environmental predictor variables in Model 2 (Figure 82) was slightly different from that of Model 1. In this model, Bottom Temperature Average Minimum was the most important predictor, followed distantly by Bottom Temperature Mean, Bottom Salinity Average Minimum, and the remaining variables in the model. The partial dependence plots for the top six environmental predictor variables are shown in Figure 83. Along the gradient in Bottom Temperature Average Minimum, presence probability was moderate at low (< -1ºC) temperature values, decreased, and then increased beginning at temperature values of ~2ºC.

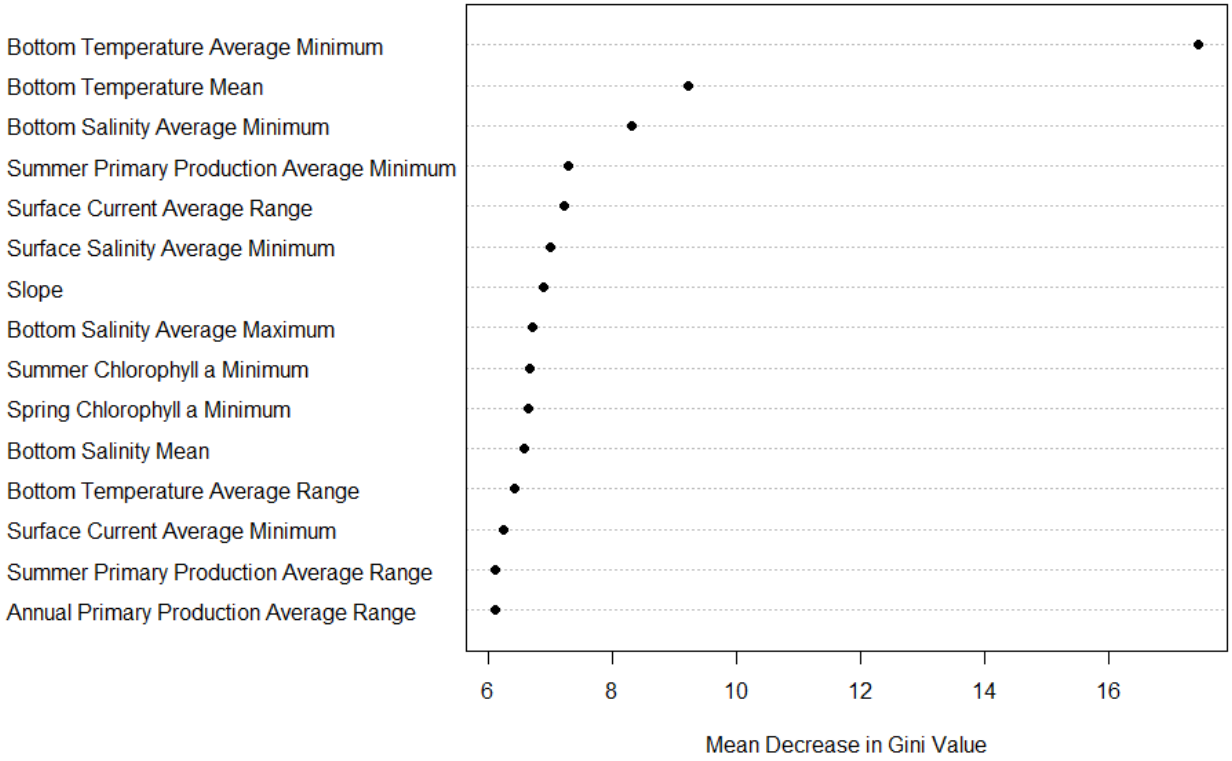

Figure 82. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced large gorgonian coral presence and absence data from the Eastern Arctic Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

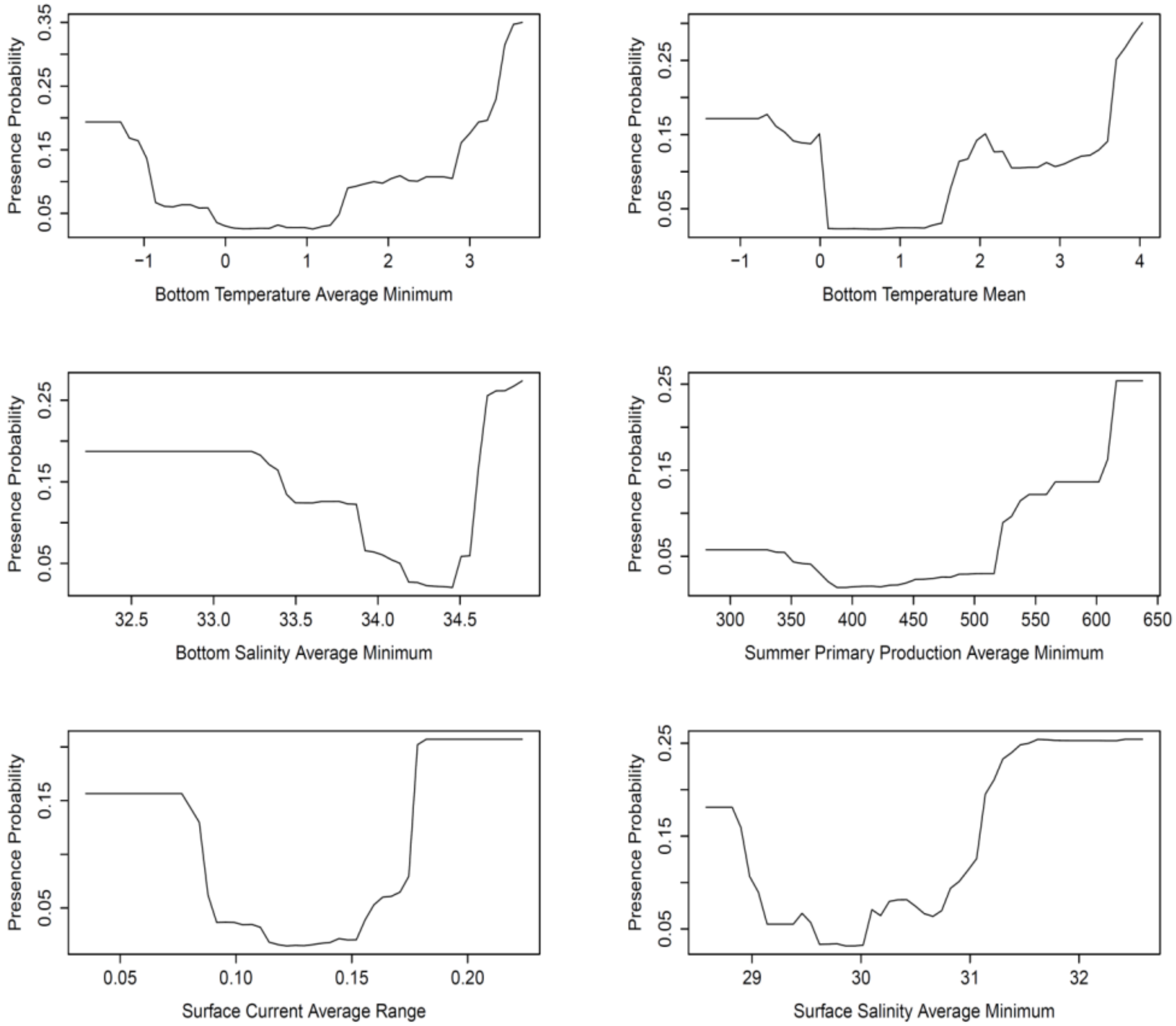

Figure 83. Partial dependence plots of the top six predictors from the random forest model of large gorgonian coral unbalanced presence and absence data collected within the Eastern Arctic Region, ordered from left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using all available large gorgonian coral data and an unbalanced species prevalence (Model 2) was selected as the best predictor of large gorgonian coral distribution in the Eastern Arctic Region. Although accuracy measures were slightly better for Model 1 (balanced dataset), this model was considered a poorer predictor of large gorgonian coral presence probability due to its excessive exaggeration of high presence probabilities mainly in the deep waters of Davis Strait. This phenomenon was due to random down-sampling of the absence data.

Validation of Selected Model Using Independent Data

Figure 84 shows the predicted presence probability of large gorgonian corals generated from Model 2 at the location of large gorgonian coral records collected during two DFO scientific missions to the Eastern Arctic in 2012 and 2013. Records from both years were combined for display as there was little overlap between them. From 2012, there were two records of large gorgonians, located southeast of Devon Island in northern Baffin Bay. These records were predicted as absence by the model. From 2013, there were a total of 10 large gorgonian corals, of which all (100%) were predicted as presence by the model. These records were located mainly in the Hatton Basin closure area and were predicted with moderate (0.24 to 0.84) presence probability of large gorgonians by the model.

Figure 85 shows the spatial congruence between the FOP large gorgonian coral records and predicted presence probability from Model 2. Of the 227 records, 200 (88%) were predicted as presence by the model. The absences were located off the southern coast of Baffin Island and in Davis Strait. Several FOP records were located in the Narwhal Overwintering and Deep-Sea Coral Conservation Area. These were predicted with a higher presence probability. The expectation is for more mismatches arising from presences recorded where the start position indicates an absence, due to the potential for transit over presence areas during the tows.

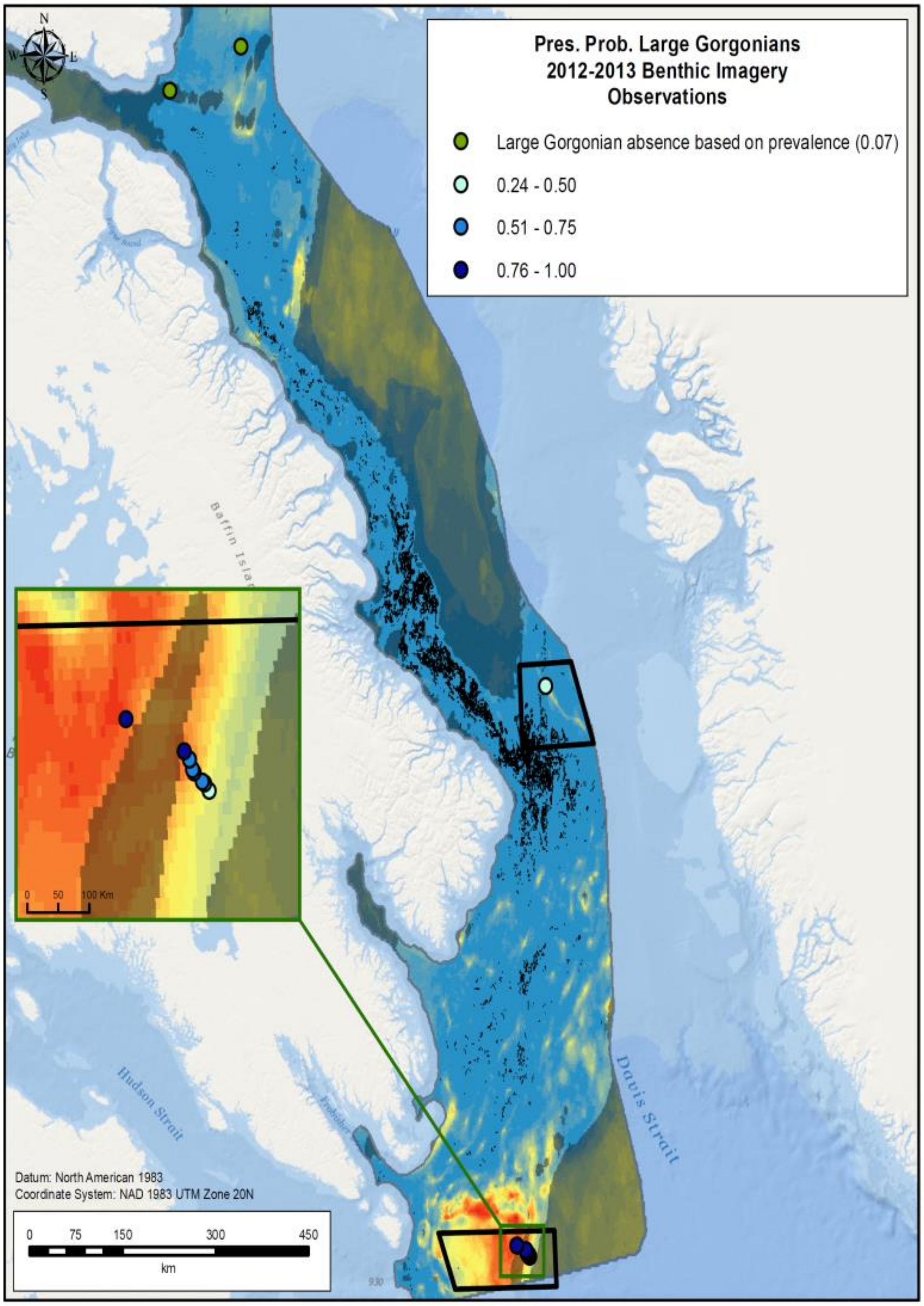
Pres. Prob. Large Gorgonians
2012-2013 Benthic Imagery
Observations
Large Gorgonian absence based on prevalence (0.07)
0.24 - 0.50
0.51 - 0.75
0.76 - 1.00
N
W
E
S
Baffin Isla
Davis Strait
Hudson Strait
0
50
100 Km
Datum: North American 1983
Coordinate System: NAD 1983 UTM Zone 20N
0
75
150
300
450
km

Figure 84. Validation of large gorgonian coral presence probability from Model 2 using in situ camera records of large gorgonians collected during DFO scientific missions conducted in 2012 and 2013 (records were combined for display). Also shown are the Narwhal Overwintering and Deep-Sea Coral Conservation Area and the Hatton Basin Voluntary Closure Area. Inset map shows the Hatton Basin (bottom) closures.

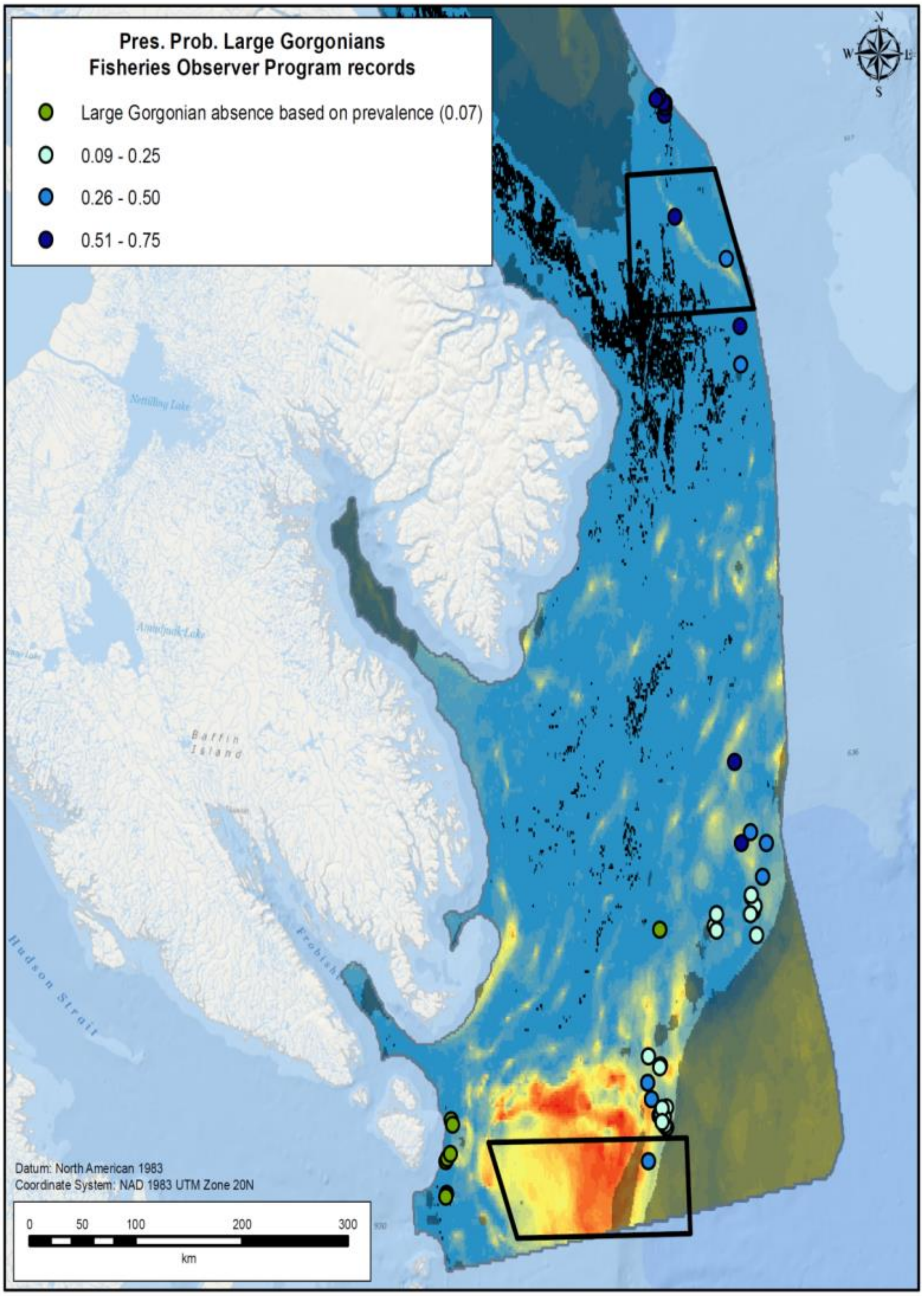
Pres. Prob. Large Gorgonians
Fisheries Observer Program records
Large Gorgonian absence based on prevalence (0.07)
0.09 - 0.25
0.26 - 0.50
0.51 - 0.75
N
W
E
S
Nettilling Lake
Amadjuak Lake
Baffin Island
Hudson Strait
Frobisher
Datum: North American 1983
Coordinate System: NAD 1983 UTM Zone 20N
0
50
100
200
300
km

Figure 85. Validation of large gorgonian coral presence probability from Model 2 using large gorgonian coral records collected by the Fisheries Observer Program between 2004 and 2011. Also shown are the Narwhal Overwintering and Deep-Sea Coral Conservation Area and the Hatton Basin Voluntary Closure Area.

Prediction of Large Gorgonian Coral Biomass Using Random Forest

Alfredo Trawl Gear

Accuracy measures from the regression random forest model on mean large gorgonian coral biomass records from trawl surveys using Alfredo trawl gear are presented in Table 22. The highest R^2 value was 0.033 while the average was 0.006 ± 0.011 SD, indicating poor model performance. The average Normalized Root-Mean-Square Error (NRMSE) was 0.021 ± 0.035 SD. The standard deviation is higher than the mean, indicating high variability between model folds. The average percent variance explained by the model was -6.54% ± 9.53 SD.

Figures 86 and 87 show the predicted biomass surface of large gorgonian corals using mean biomass data from Alfredo trawl surveys. The majority of the spatial extent was predicted to have low (> 0 – 4.13 kg) biomass of sea pens. The highest predicted biomass occurred in the southern Baffin Bay in the Narwhal Over-wintering and Deep-Sea Coral Conservation Area, corresponding to a large mean catch in that location (Figure 87). Northern Baffin Bay and southwest Davis Strait were predicted to have somewhat moderate biomass of large gorgonian corals.

Table 22. Accuracy measures for all 10 model repetitions of 10-fold cross validation of random forest model of average large gorgonian coral biomass (kg) per grid cell recorded from DFO trawl surveys within the Eastern Arctic Region. RMSE= Root-Mean-Square Error, NRMSE= Normalized Root-Mean-Square Error.

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	9.480×10^{-5}	15.437	0.008	-11.86
2	0.001	30.689	0.015	-9.14
3	0.001	14.502	0.007	-13.67
4	2.293×10^{-4}	10.296	0.005	-9.19
5	1.91×10^{-7}	59.375	0.030	-3.34
6	0.017	22.226	0.011	-12.32
7	0.001	234.083	0.117	19.32
8	0.003	4.094	0.002	-9.13
9	2.702×10^{-4}	27.018	0.014	-7.41
10	0.033	6.042	0.003	-8.70
Mean	**0.006**	**42.376**	**0.021**	**-6.54**
SD	**0.011**	**69.231**	**0.035**	**9.53**

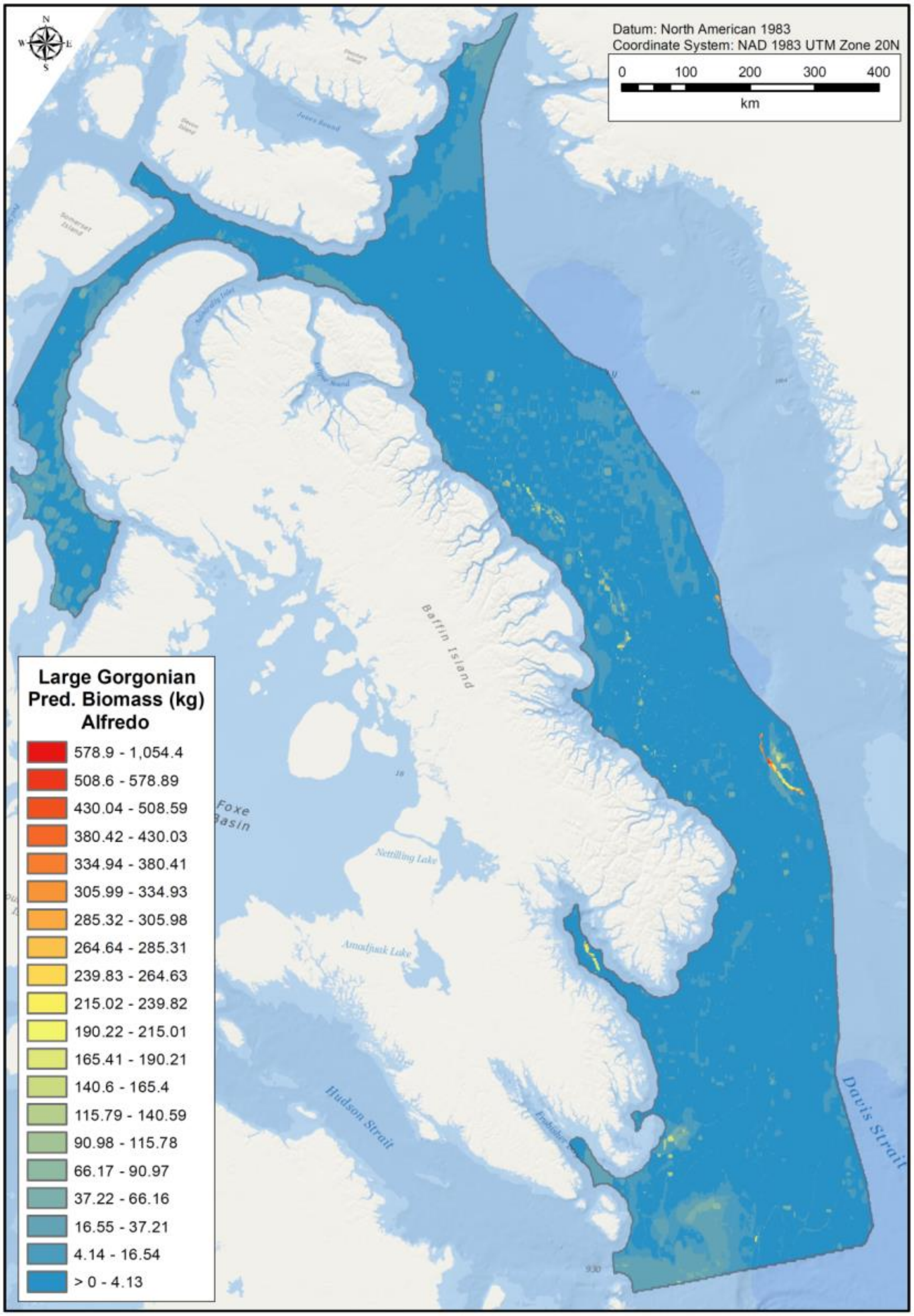

Figure 86. Predictions of biomass (kg) of large gorgonian corals from catch data recorded in DFO trawl surveys conducted using Alfredo trawl gear in the Eastern Arctic Region between 1999 and 2014.

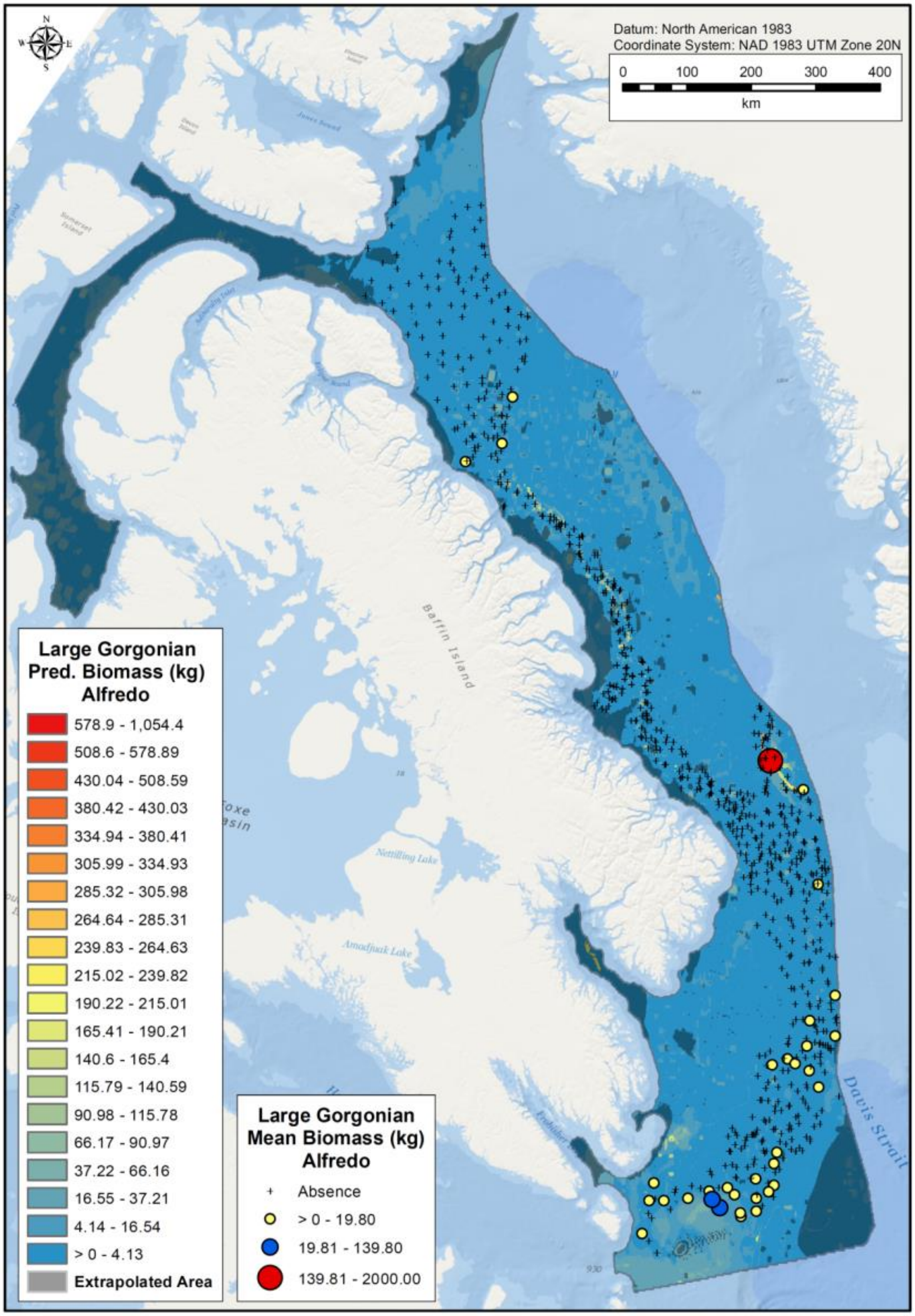

Figure 87. Predictions of biomass (kg) of large gorgonian corals from catch data recorded in DFO trawl surveys conducted using Alfredo trawl gear in the Eastern Arctic Region between 1999 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

Of the 54 environmental variables used in the model, Slope was the most important for predicting biomass of large gorgonian corals (Figure 88). This variable was followed by Surface Current Average Minimum. The partial dependence plots of the top six environmental predictor variables are shown in Figure 89. Along the Slope gradient, predicted biomass was highest between 4 and 7º.

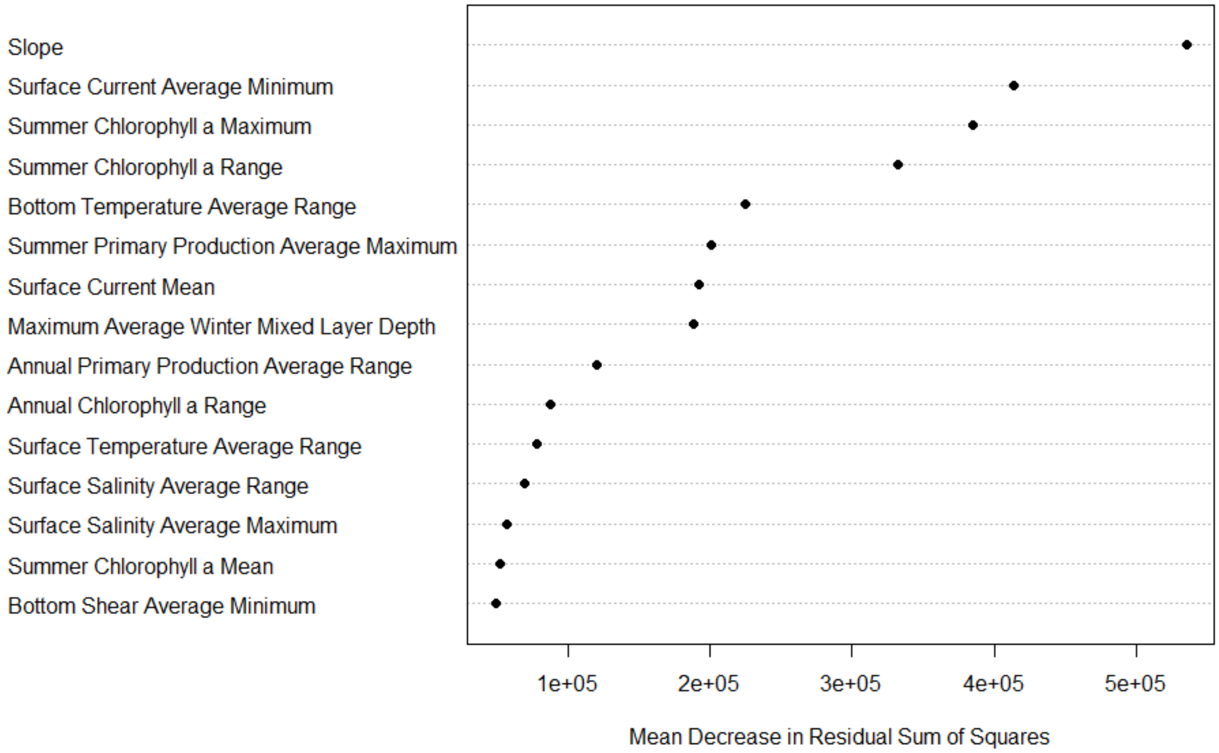

Figure 88. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on large gorgonian coral biomass data collected from DFO trawl surveys conducted using Alfredo trawl gear. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

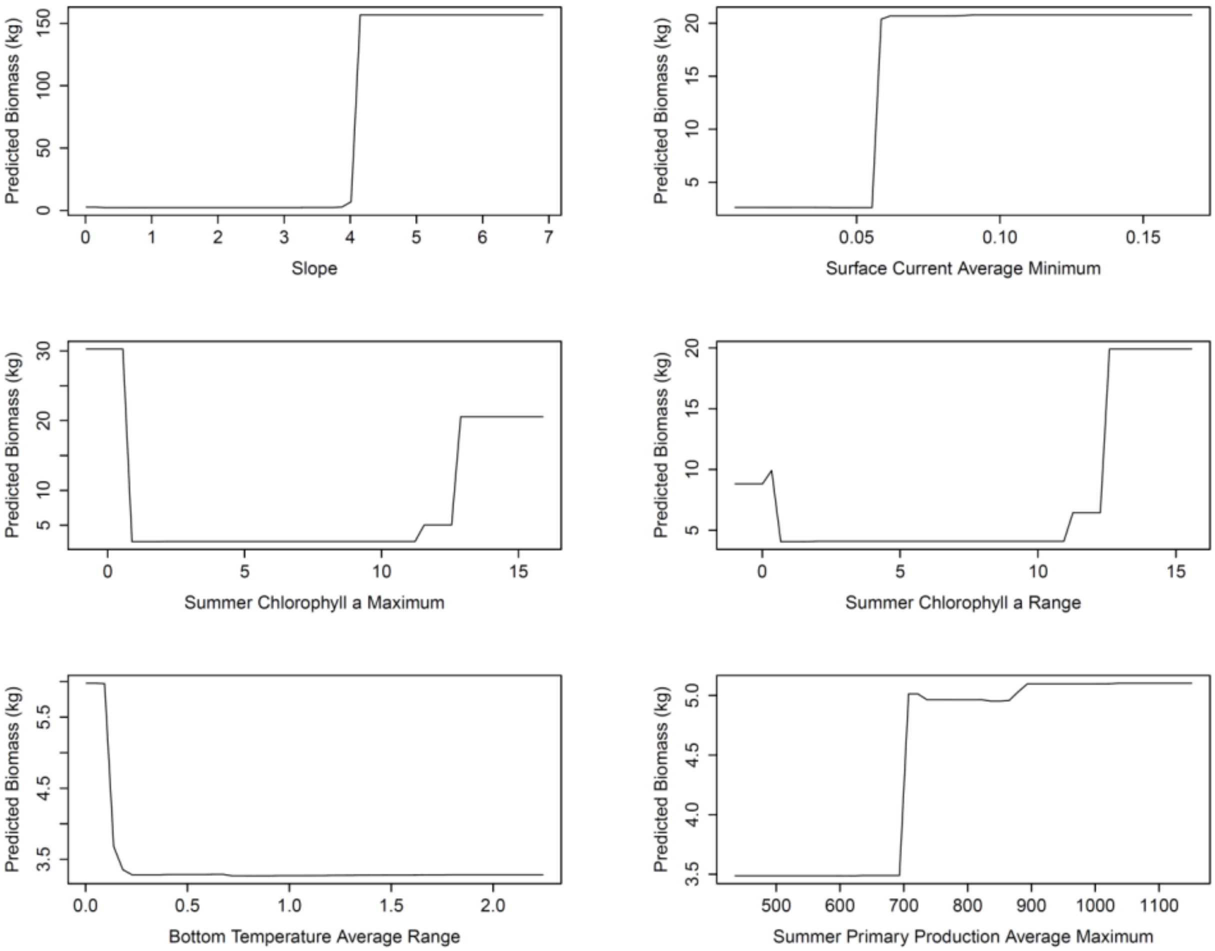

Figure 89. Partial dependence plots of the top six predictors from the random forest model of large gorgonian coral biomass data collected from DFO trawl surveys using Alfredo trawl gear in the Eastern Arctic Region, ordered from left to right from the top.

Campelen Trawl Gear

The accuracy measures of the regression random forest model on mean large gorgonian coral biomass per grid cell from DFO trawl surveys with Campelen are presented in Table 23. The highest R^2 value was 0.470, while the average was 0.186 ± 0.160 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.013 ± 0.007 SD. This model explained a somewhat moderate percentage of variance in the biomass data (average = 16.86% ± 4.99 SD).

Figures 90 and 91 show the predicted biomass surface of large gorgonians. The majority of the spatial extent was predicted to have low (< 18.9 kg) large gorgonian coral biomass. The highest predicted biomass (up to 322 kg) occurred north of the voluntary Closure Area in Hatton Basin in Davis Strait and coincided with a cluster of large mean catches (Figure 91). The southeast corner of the study extent in Davis Strait was predicted to have moderate biomass of large gorgonian corals, however, this area was considered an area of model extrapolation (Figure 91).

Table 23. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of average large gorgonian coral biomass (kg) per grid cell recorded from DFO groundfish trawl surveys with Campelen gear conducted in the Eastern Arctic Region. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	**R^2**	**RMSE**	**NRMSE**	**Percent (%) variance explained**
1	0.074	14.210	0.007	20.39
2	0.376	32.945	0.016	9.92
3	0.179	17.858	0.009	20.32
4	0.337	14.518	0.007	16.17
5	0.470	58.188	0.029	6.53
6	0.165	25.158	0.013	18.14
7	0.033	17.799	0.009	21.41
8	0.172	18.064	0.009	15.78
9	0.039	42.903	0.022	19.03
10	0.009	19.904	0.010	20.88
Mean	**0.186**	**26.155**	**0.013**	**16.86**
SD	**0.160**	**14.434**	**0.007**	**4.99**

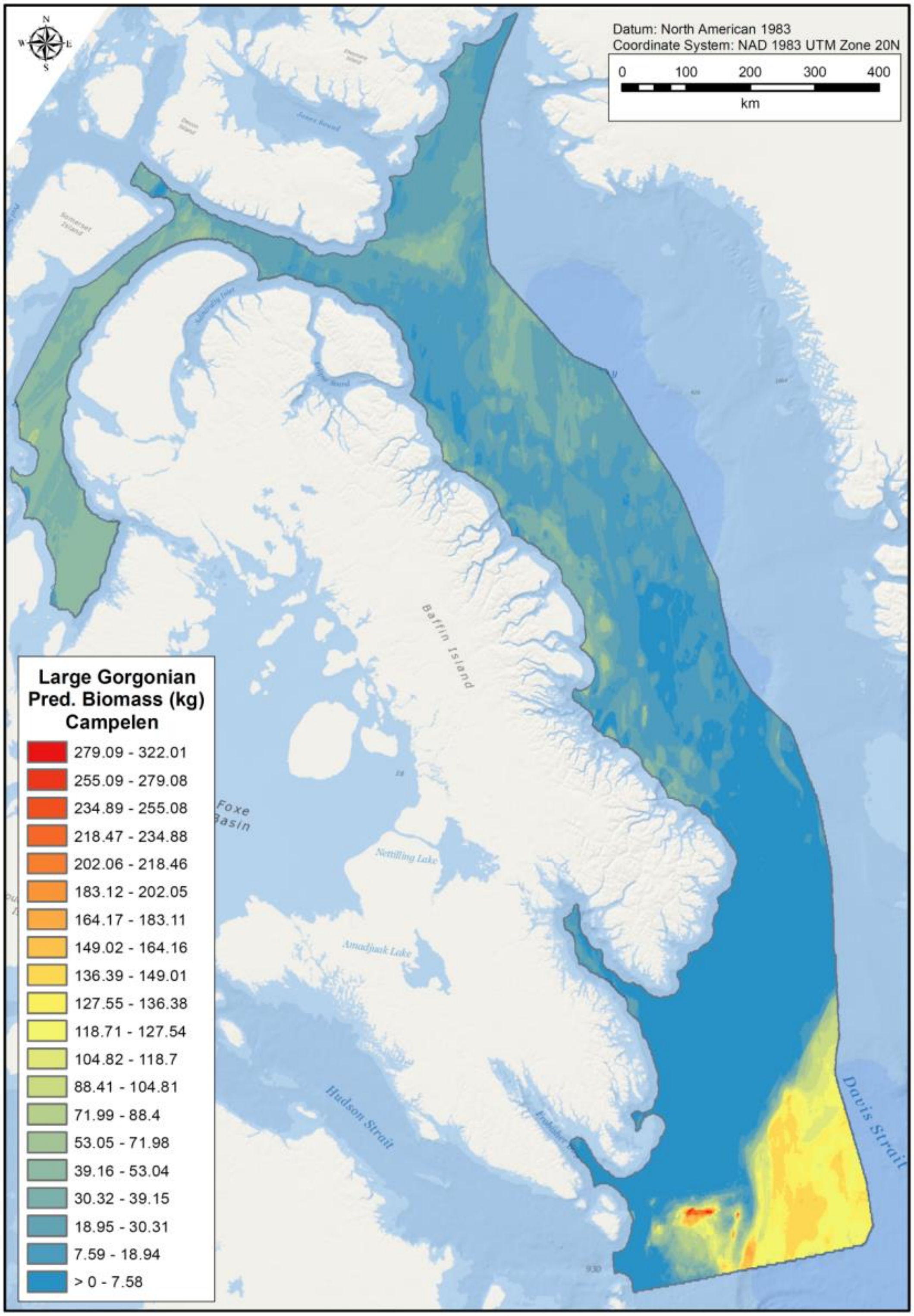

Figure 90. Predictions of biomass (kg) per grid cell of large gorgonian corals from catch data recorded in DFO trawl surveys with Campelen gear conducted in the Eastern Arctic Region between 2005 and 2014.

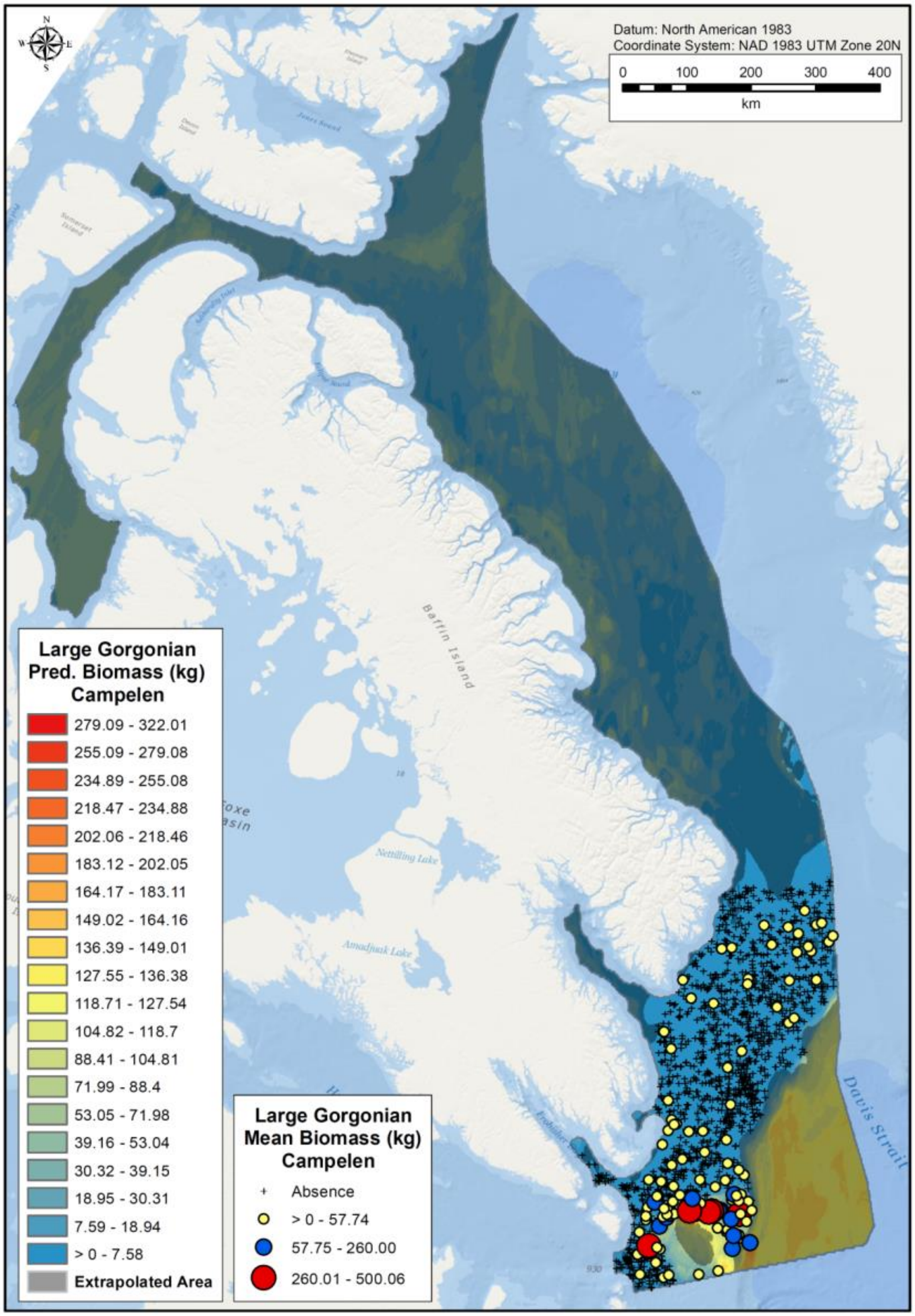

Figure 91. Predictions of biomass (kg) per grid cell of large gorgonian corals from catch data recorded in DFO trawl surveys with Campelen gear conducted in the Eastern Arctic Region between 2005 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting large gorgonian biomass are shown in Figure 92. Bottom Temperature Average Minimum was the most important variable in the model, followed by Annual Primary Production Average Minimum, Bottom Temperature Average Range, and Surface Salinity Average Minimum. The partial dependence of small gorgonian coral biomass on the top 6 most important variables is shown in Figure 93. Predicted biomass was highest at Bottom Temperature Average Minimum values of 0.007 m s^{-1} and greater.

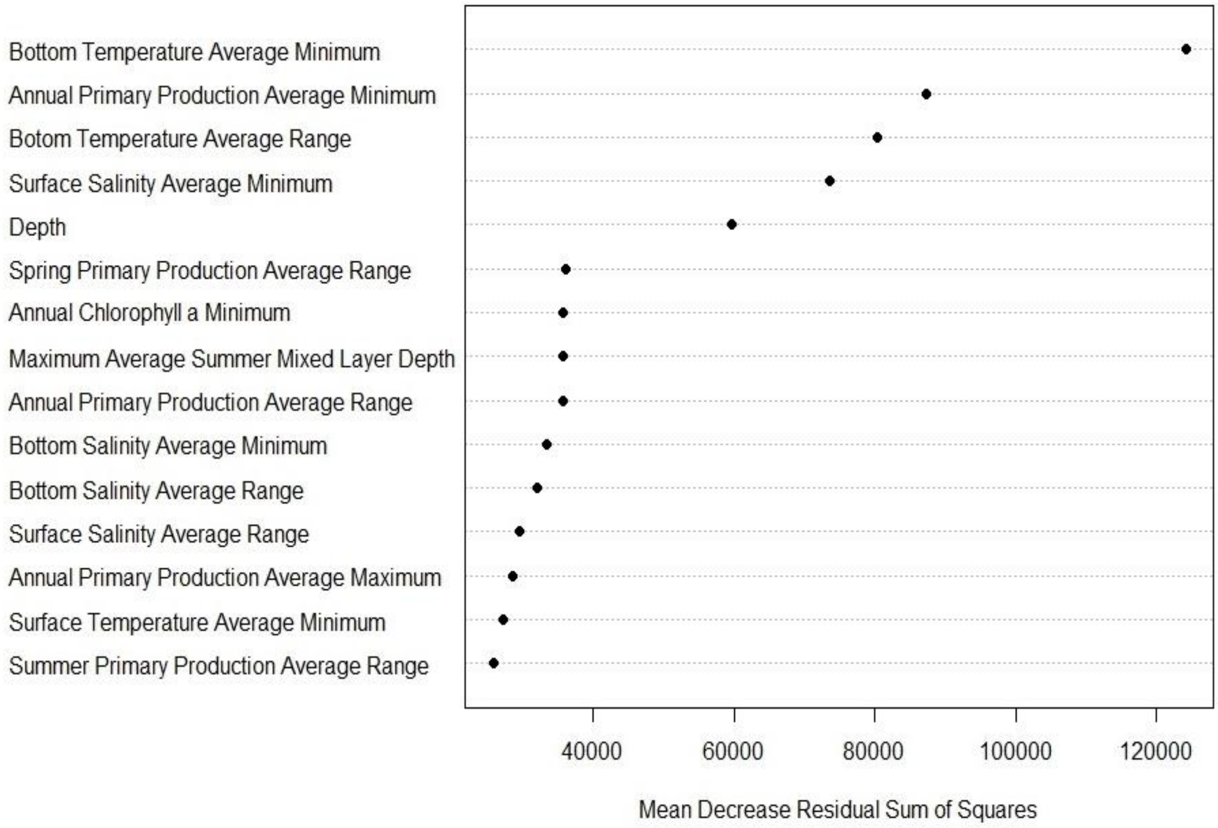

Figure 92. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on large gorgonian coral mean biomass data averaged per grid cell will Campelen gear. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

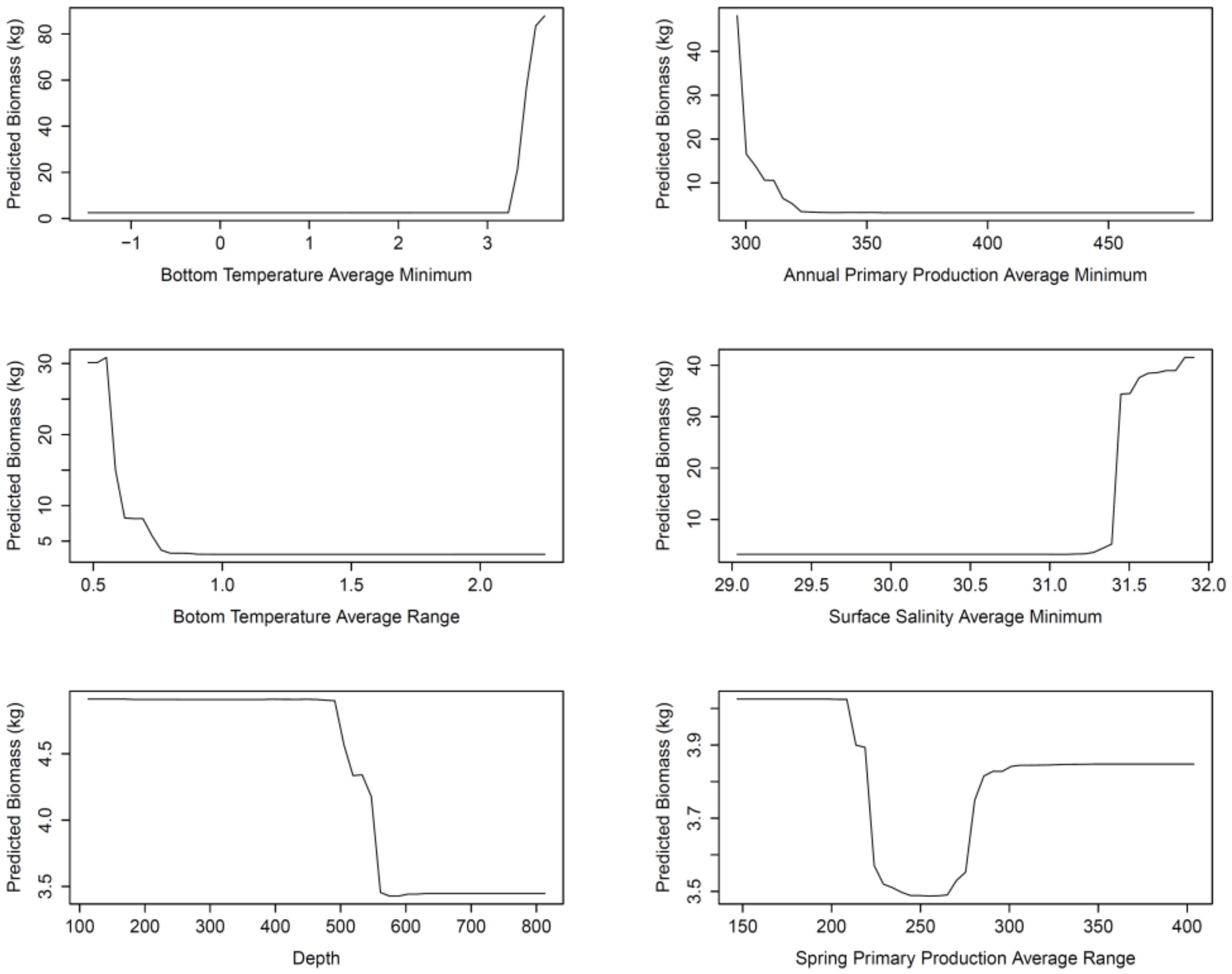

Figure 93. Partial dependence plots of the top six predictors from the random forest model of large gorgonian coral biomass data collected with Campelen gear within the Eastern Arctic Region, ordered left to right from the top.

Small Gorgonian Corals

Data Sources and Distribution

Small gorgonian coral catch data for the Eastern Arctic was collected between 2005 and 2014, and consisted of 85 presences and 655 absences from *Paamuit* surveys conducted using Alfredo trawl gear, 90 presences and 1354 absences from the *Cape Ballard*, *Aqviq*, and *Kinguk* surveys conducted using Campelen trawl gear, and 4 presences and 185 absences from *Paamuit* surveys using Cosmos gear (Table 24). The majority of the presence records were distributed in the southern Baffin Bay and Davis Strait. There was relatively good congruence in the spatial distribution of presence records originating from the different data sources, particularly of the Alfredo and Campelen records. Alfredo trawl records had the widest spatial distribution in the study extent (Figure 94). Several i*n situ* benthic imagery records were distributed in the Narwhal

Over-wintering and Deep-Sea Coral Conservation Area in southern Baffin Bay, and in the Hatton Basin Voluntary Closure Area in Davis Strait.

Presence-absence random forest models were generated on the combined dataset consisting of 179 presences and 2187 absences (see Figure 95). The highest mean biomass record was 1.50 kg and was distributed in the Davis Strait.

Table 24. Number of presence and absence records of small gorgonian catch by gear type recorded from DFO trawl surveys conducted between 2005 and 2014 in the Eastern Arctic Region.

	Alfredo		**Campelen**		**Cosmos**	
Year	**Presences**	**Absences**	**Presences**	**Absences**	**Presences**	**Absences**
2005	-	-	7	144	-	-
2006	2	58	16	126	2	73
2007	-	-	15	118	0	14
2008	0	79	16	127	2	65
2009	1	16	4	139	-	-
2010	1	116	10	134	0	22
2011	32	51	4	148	-	-
2012	0	161	8	143	0	11
2013	25	61	10	141	-	-
2014	24	113	0	134	-	-
TOTAL	**85**	**655**	**90**	**1354**	**4**	**185**

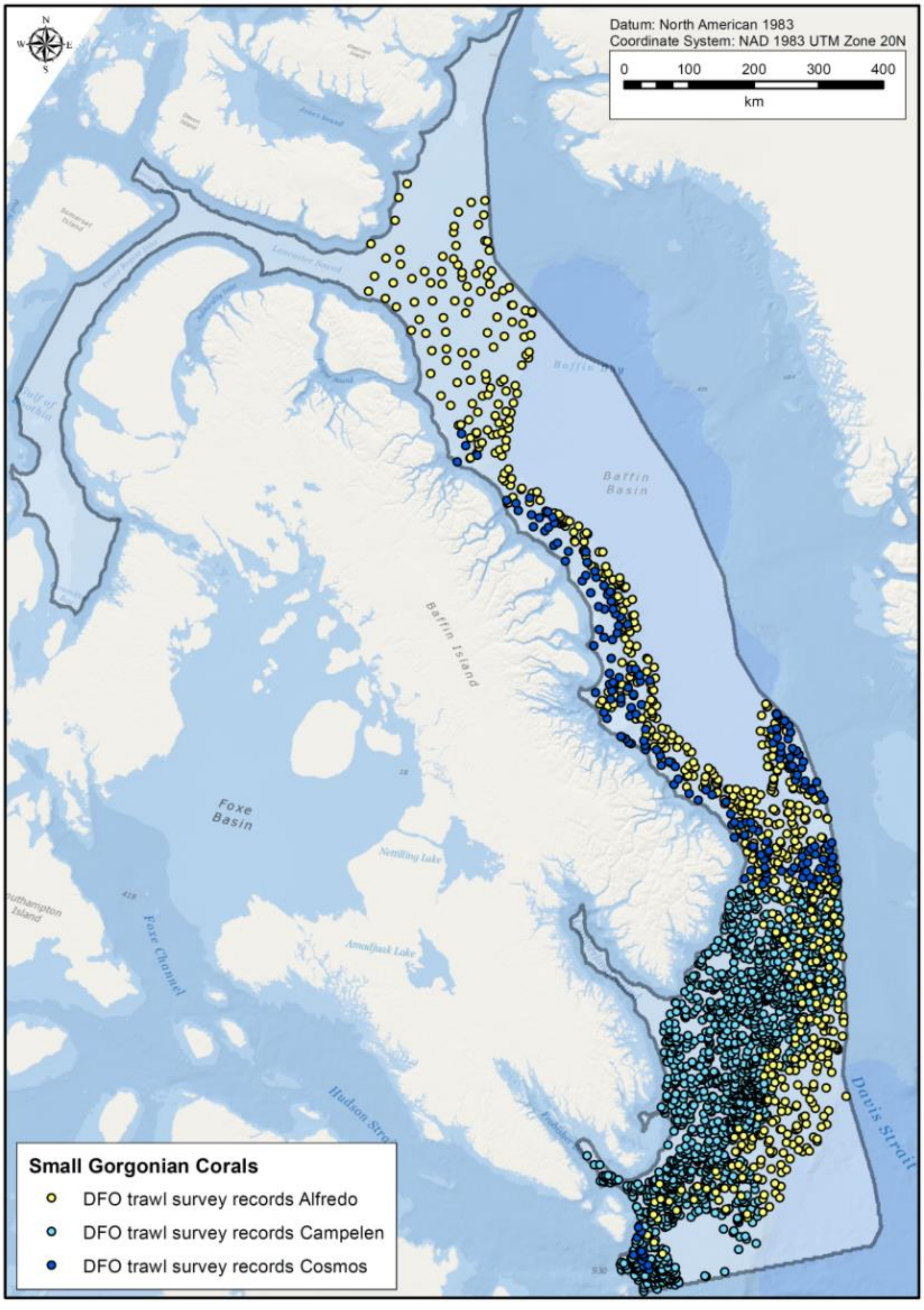

Figure 94. Available small gorgonian coral data in the Eastern Arctic Region from DFO trawl surveys conducted between 2005 and 2014.

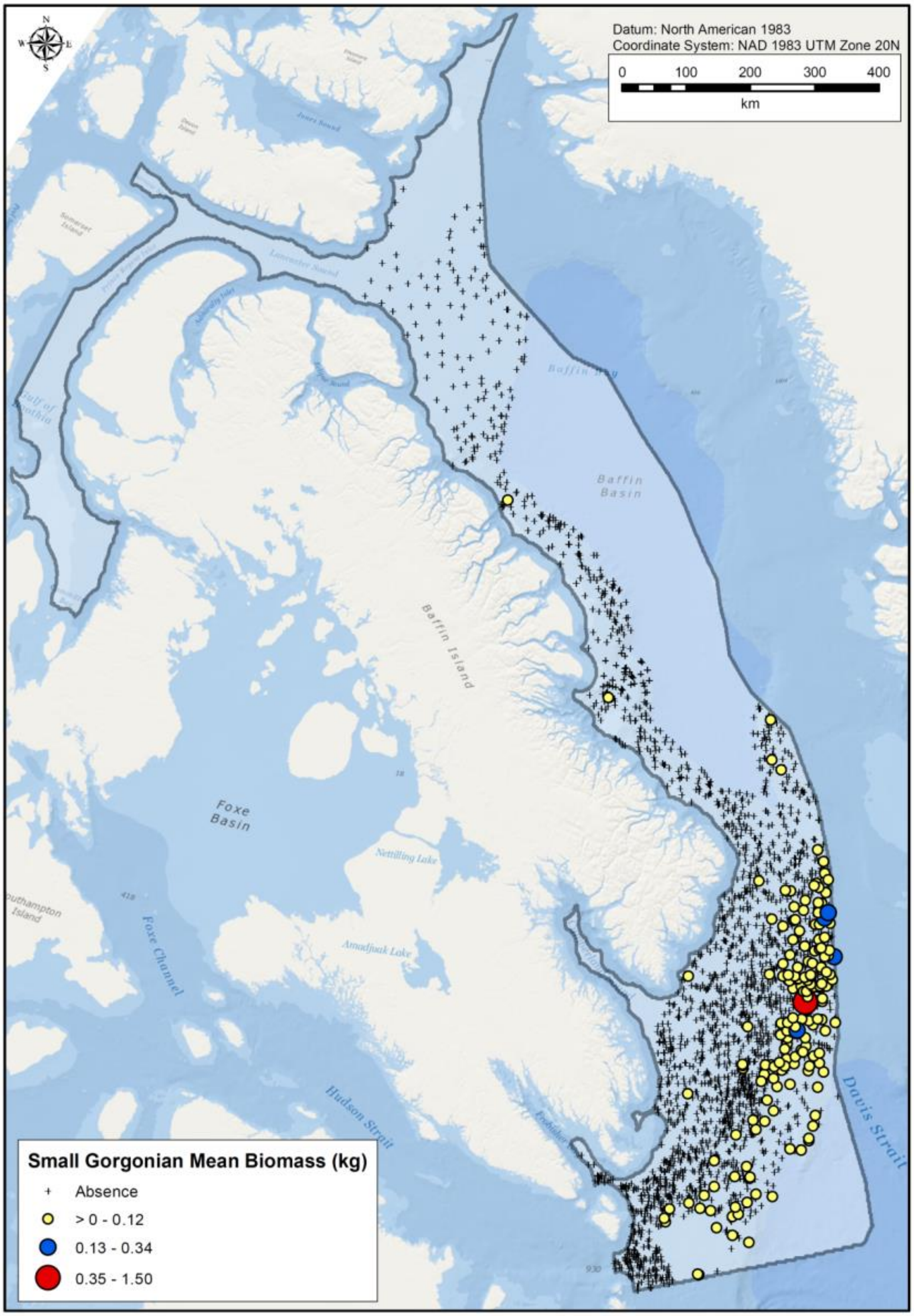

Figure 95. Mean biomass (kg) per grid cell of small gorgonian coral catch recorded in the Eastern Arctic Region from DFO trawl surveys conducted between 2005 and 2014.

Model 1 – Balanced Species Prevalence

Accuracy measures for the random forest model on balanced species prevalence (179 presences and 179 absences; Model 1) are presented in Table 25. The average AUC was 0.882 ± 0.017, indicating very good model performance. The highest AUC of 0.902 was associated with Model Run 7 and is therefore considered the optimal model for the prediction of the small gorgonian coral response data. The sensitivity and specificity measures of this model fold were 0.849 and 0.827, respectively. The confusion matrix of the optimal model is also presented in Table 25. Class error for the presence and absences classes was relatively low (0.151 and 0.173, respectively).

Table 25. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of small gorgonian corals within the Eastern Arctic 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 small gorgonian corals in the region.

Model Run	AUC	Sensitivity	Specificity
1	0.882	0.838	0.816
2	0.892	0.838	0.810
3	0.855	0.827	0.793
4	0.892	0.821	0.832
5	0.859	0.827	0.793
6	0.869	0.827	0.805
7	**0.902**	**0.849**	**0.827**
8	0.879	0.832	0.793
9	0.892	0.855	0.816
10	0.902	0.844	0.849
Mean	**0.882**	**0.836**	**0.813**
SD	**0.017**	**0.011**	**0.019**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	**Presence**		
Absence	148	31	179	0.173
Presence	27	152	179	0.151

The presence probability prediction surface of small gorgonian corals is presented in Figure 96. The highest predictions of small gorgonian coral presence probability occurred in Davis Strait in deep water southeast of Hall Peninsula on Baffin Island. These areas of high presence probability corresponded well with the location of presence points, although there was extrapolation of moderate/high presence probability to the southeast corner of the study extent in Davis Strait where there are no data observations (Figure 97). The deep waters off the Baffin Island Shelf were also predicted to have moderate presence probability of small gorgonian corals where there are no presence observations.

The actual presence and absence data observations (179 presences and 179 absences) used in the optimal fold of Model 1 (Figure 98). The absences in Davis Strait were greatly reduced by the random down-sampling, although some absence points in the vicinity of the presence data were selected. This figure also shows the areas of model extrapolation. Most of Baffin Bay is considered extrapolated area. Predictions in Lancaster Sound and the Gulf of Boothia were extrapolated by the model.

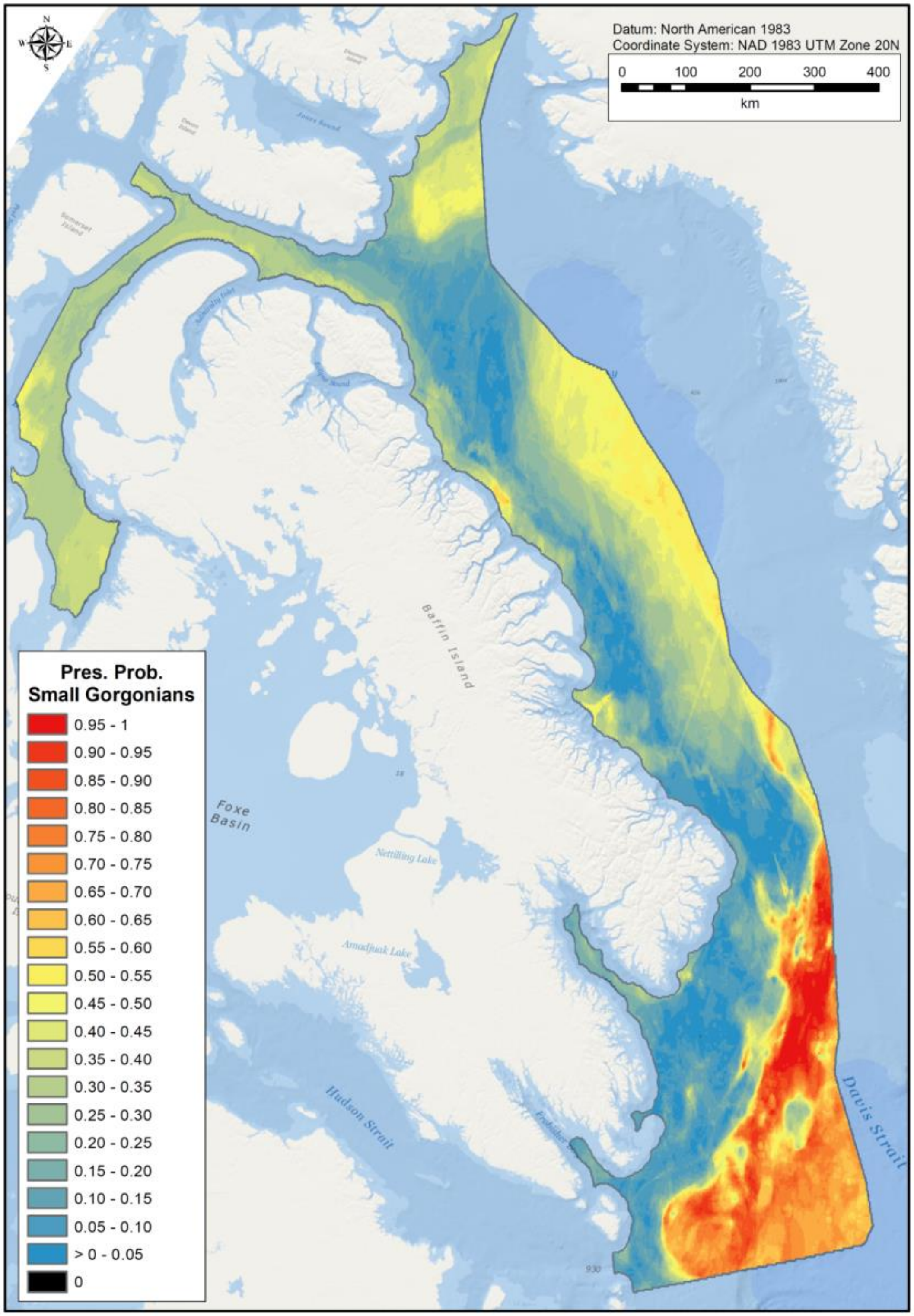

Figure 96. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of small gorgonian coral presence and absence data collected from DFO trawl surveys in the Eastern Arctic Region between 2005 and 2014.

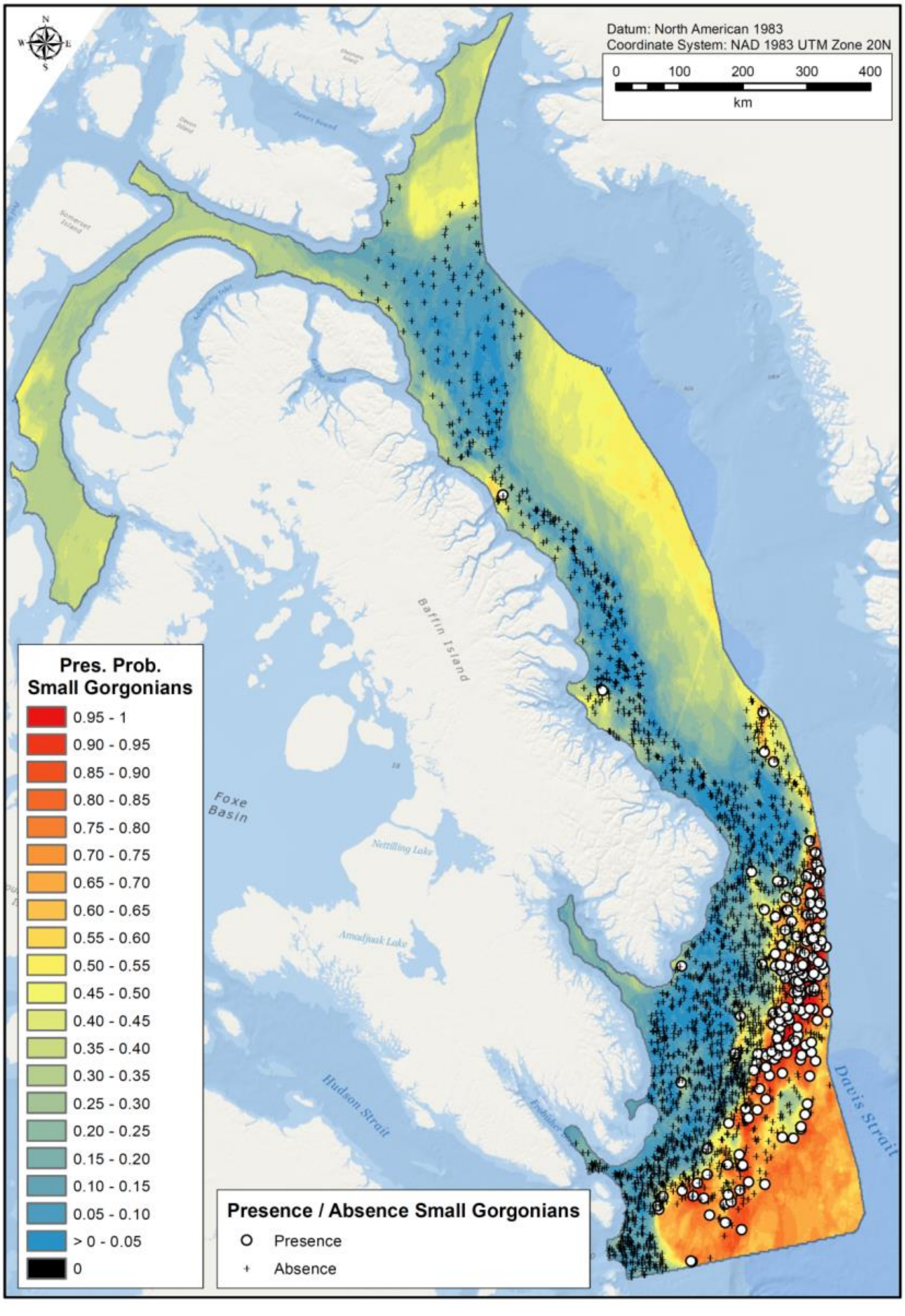

Figure 97. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of small gorgonian presence and absence data recorded from DFO trawl surveys in the Eastern Arctic Region between 2005 and 2014.

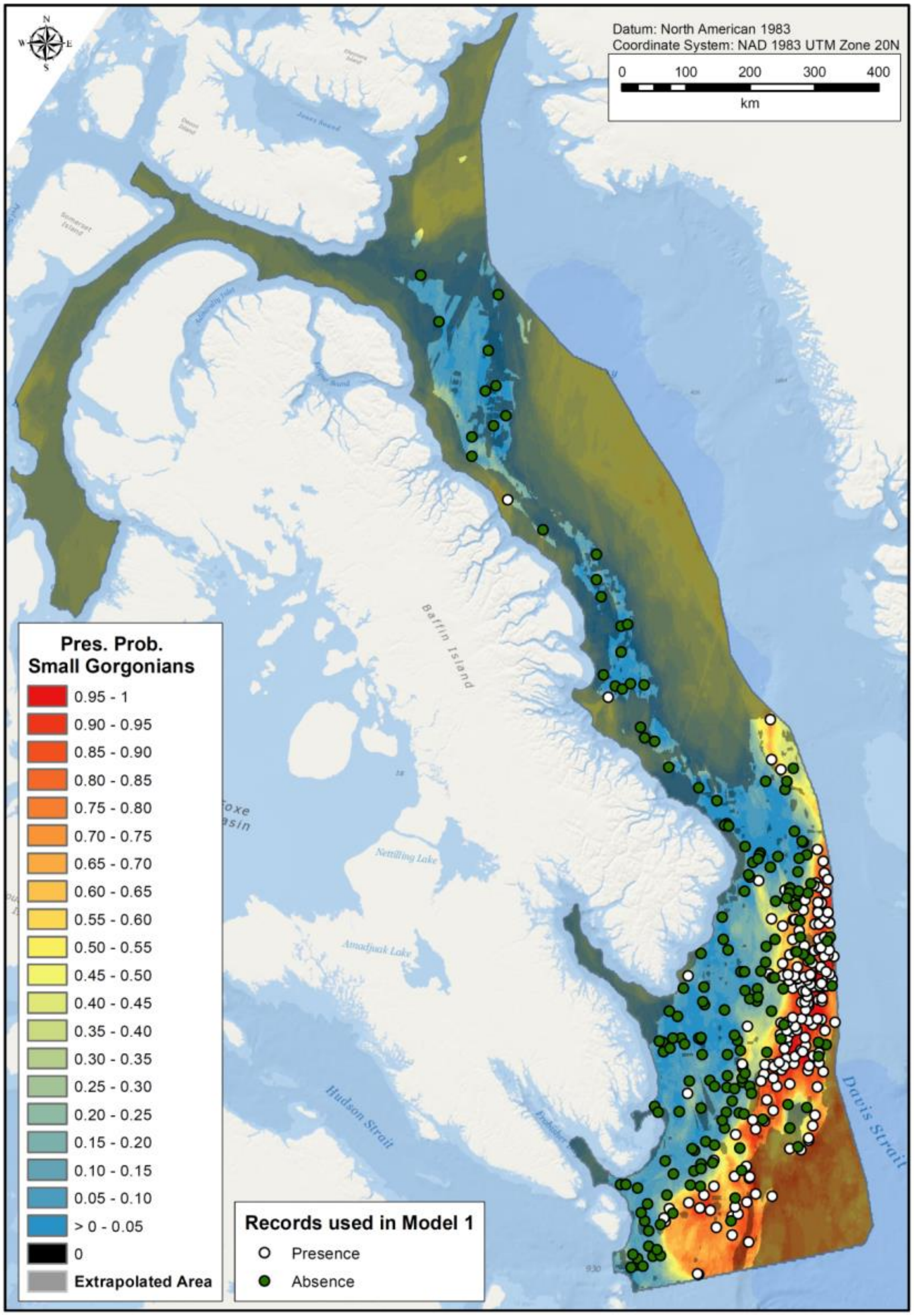

Figure 98. Map of the 358 data observations (179 presences and 179 absences) of small gorgonian corals used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of small gorgonian corals generated from Model 1 and areas of model extrapolation.

Of all 54 environmental predictor variables used in the model, Surface Salinity Mean was the most important for the classification of the small gorgonian presence and absence data (Figure 99). This variable was followed more distantly in terms of its Mean Decrease in Gini Value by Surface Salinity Average Maximum, Surface Salinity Average Minimum, and the remaining variables in the model. The partial dependence plots for the top 6 most important predictors are shown in Figure 100. Presence probability of small gorgonians was highest at Surface Salinity Mean values of 32.5 and higher.

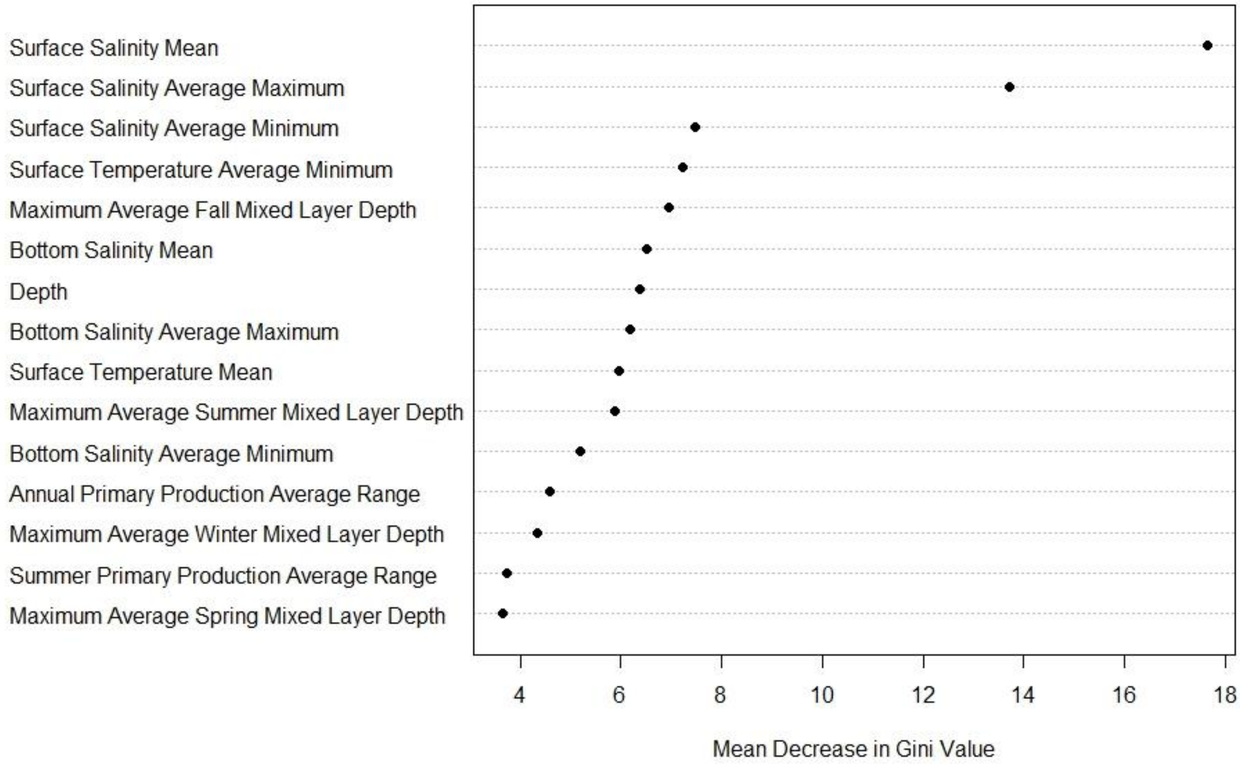

Figure 99. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting small gorgonian coral presence and absence data within the Eastern Arctic Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

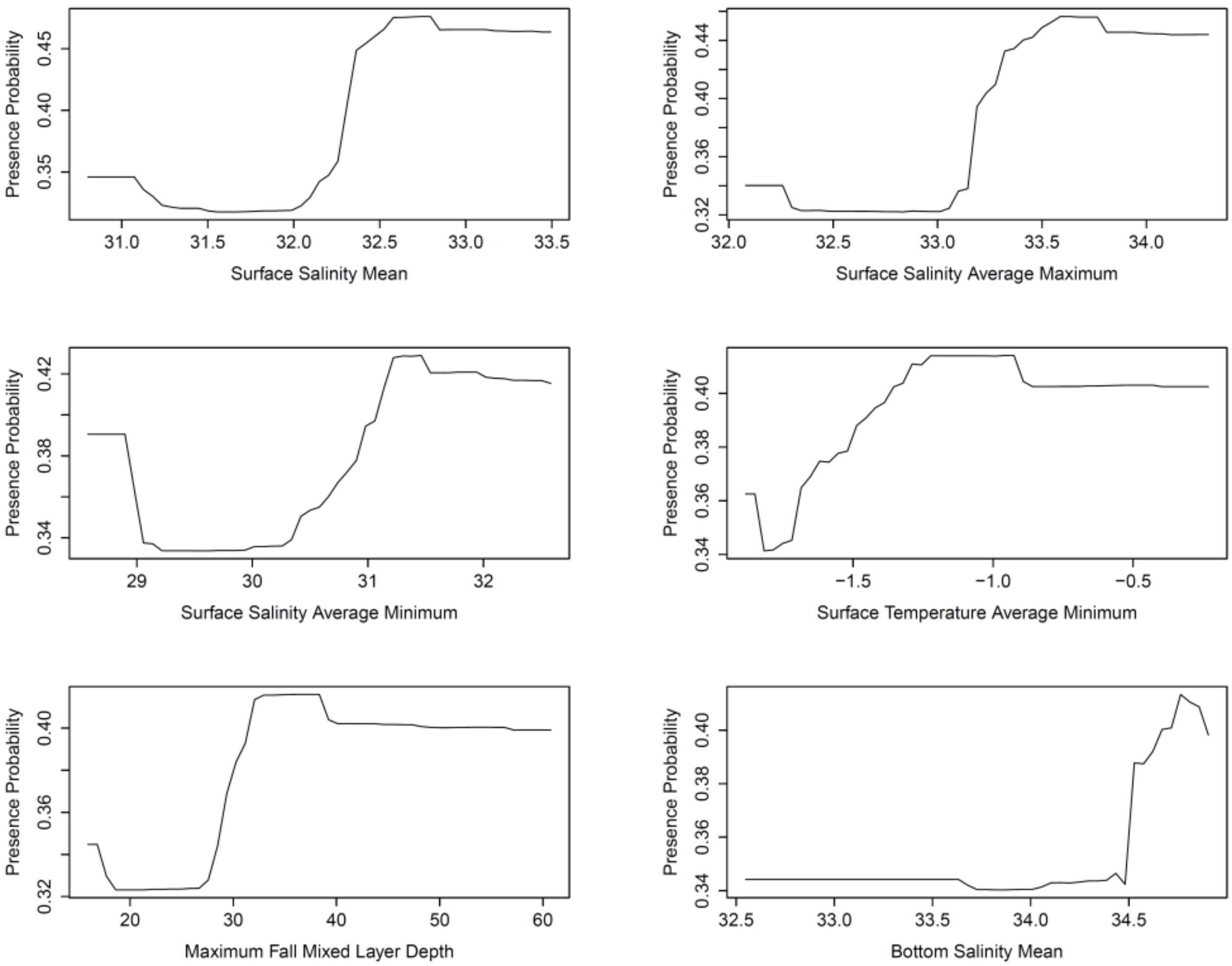

Figure 100. Partial dependence plots of the top six predictors from the optimal random forest model of small gorgonian presence and absence data collected within the Eastern Arctic Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 26 shows the accuracy measures for the random forest model on all small gorgonian coral presence and absence data (179 presences and 2189 absences; Model 2) and a threshold equal to species prevalence (0.08). The average AUC calculated was 0.894, slightly higher than that of Model 1 (0.882 in Model 1). Sensitivity and specificity were comparable to Model 1. Class error for the presence class was slightly higher in Model 2 (0.179 versus 0.151 for Model 1).

The predicted presence probability surface of small gorgonian corals generated from Model 2 is shown in Figure 101. The area of high presence probability in the Davis Strait was much reduced in this model. At the location of some presence points predicted presence probability was not high due to the high overlap of presence and absence points (Figure 102). Figure 103 depicts the classification of small gorgonian coral presence probability into presence and absence categories based on the prevalence threshold of 0.08. In this map, all presence probability values generated from Model 2 greater than 0.08 were classified as presence, while values less than 0.09 were classed as absence. The majority of Baffin Bay, Lancaster Sound, and the Gulf of Boothia were classified as absence of small gorgonian corals. The deep waters off Baffin Island Shelf, northern Baffin Bay, and the southeast Davis Strait were classified as presence of small gorgonian corals. Areas of model extrapolation are also shown in this figure. Much of Lancaster Sound, Gulf of Boothia, and the deep water off Baffin Island Shelf was considered extrapolated area.

Table 26. Accuracy measures for unrepeated 10-fold cross validation of a random forest model of presence and absence of small gorgonians within the Eastern Arctic Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.858		**Absence**	**Presence**				
2	0.936	**Absence**	1800	387	2187	0.177	0.821	0.823
3	0.895	**Presence**	32	147	179	0.179		
4	0.876							
5	0.886							
6	0.806							
7	0.945							
8	0.916							
9	0.930							
10	0.889							
Mean	**0.894**							
SD	**0.042**							

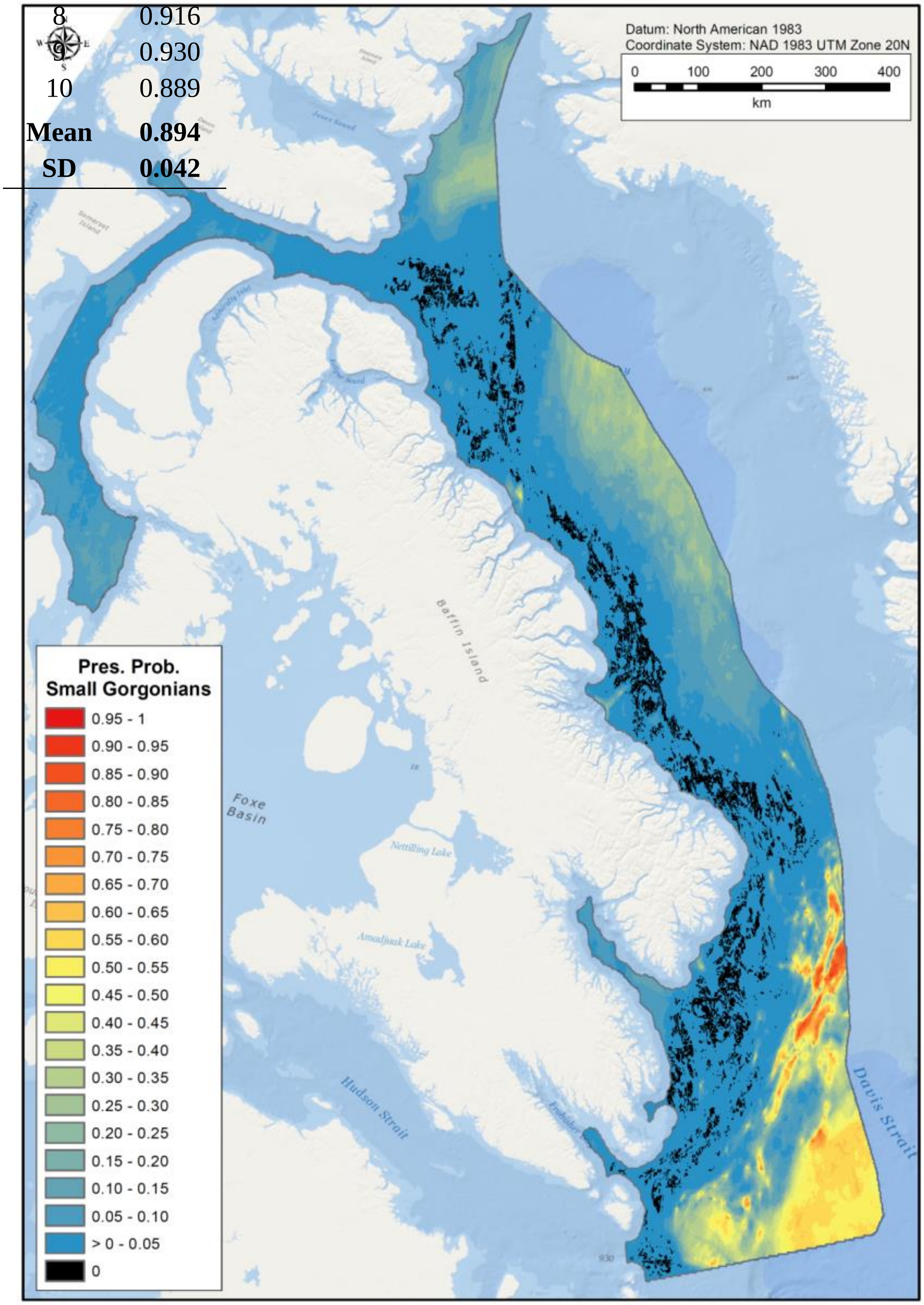

Figure 101. Predictions of presence probability (Pres. Prob.) of small gorgonian corals based on a random forest model on unbalanced presence and absence small gorgonian coral catch data collected from DFO trawl surveys conducted within the Eastern Arctic Region between 2005 and 2014.

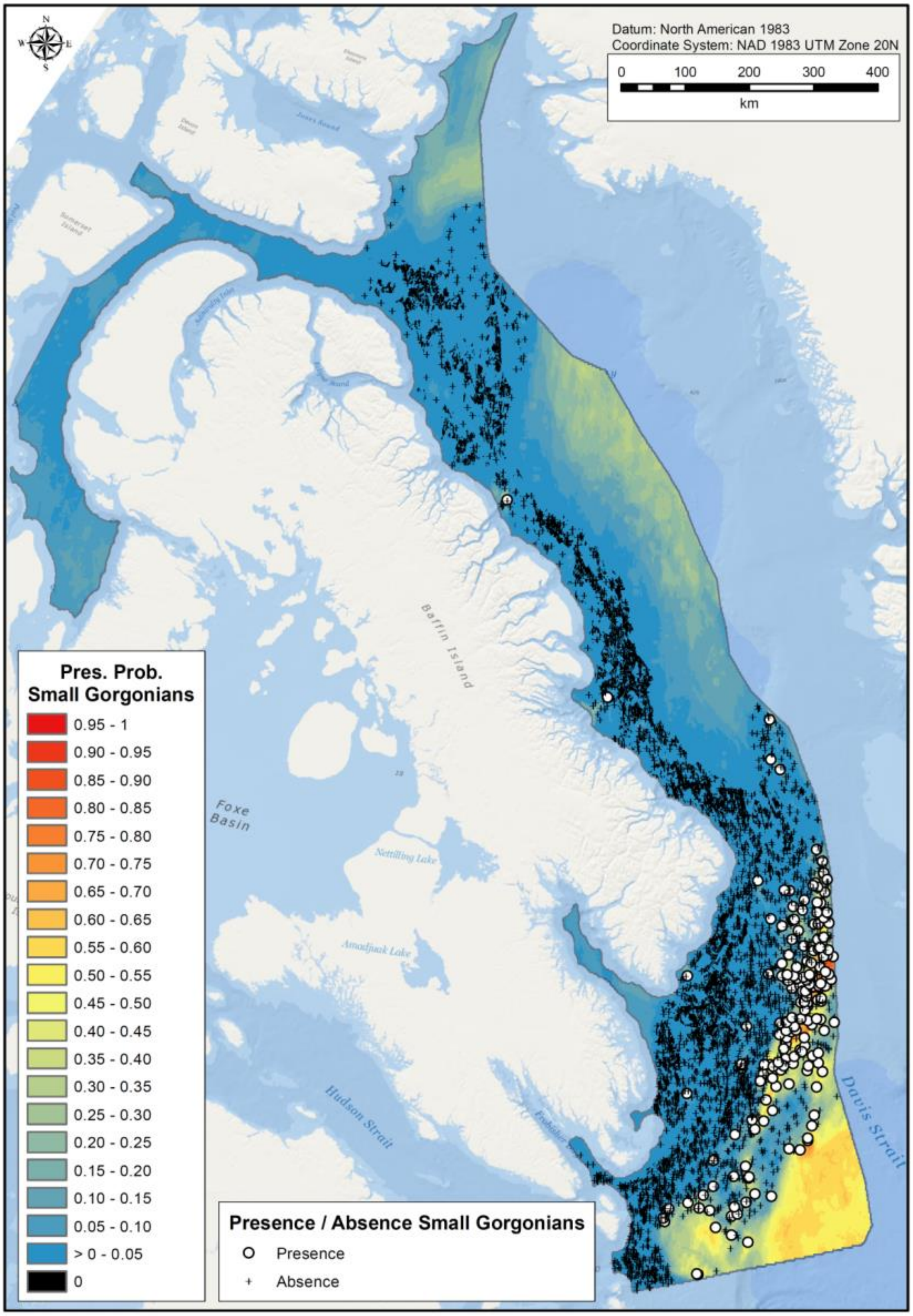

Figure 102. Presence and absence observations and predictions of presence probability (Pres. Prob.) of small gorgonian corals based on a random forest model on unbalanced presence and absence small gorgonian coral catch data collected from DFO trawl surveys conducted within the Eastern Arctic Region between 2005 and 2014.

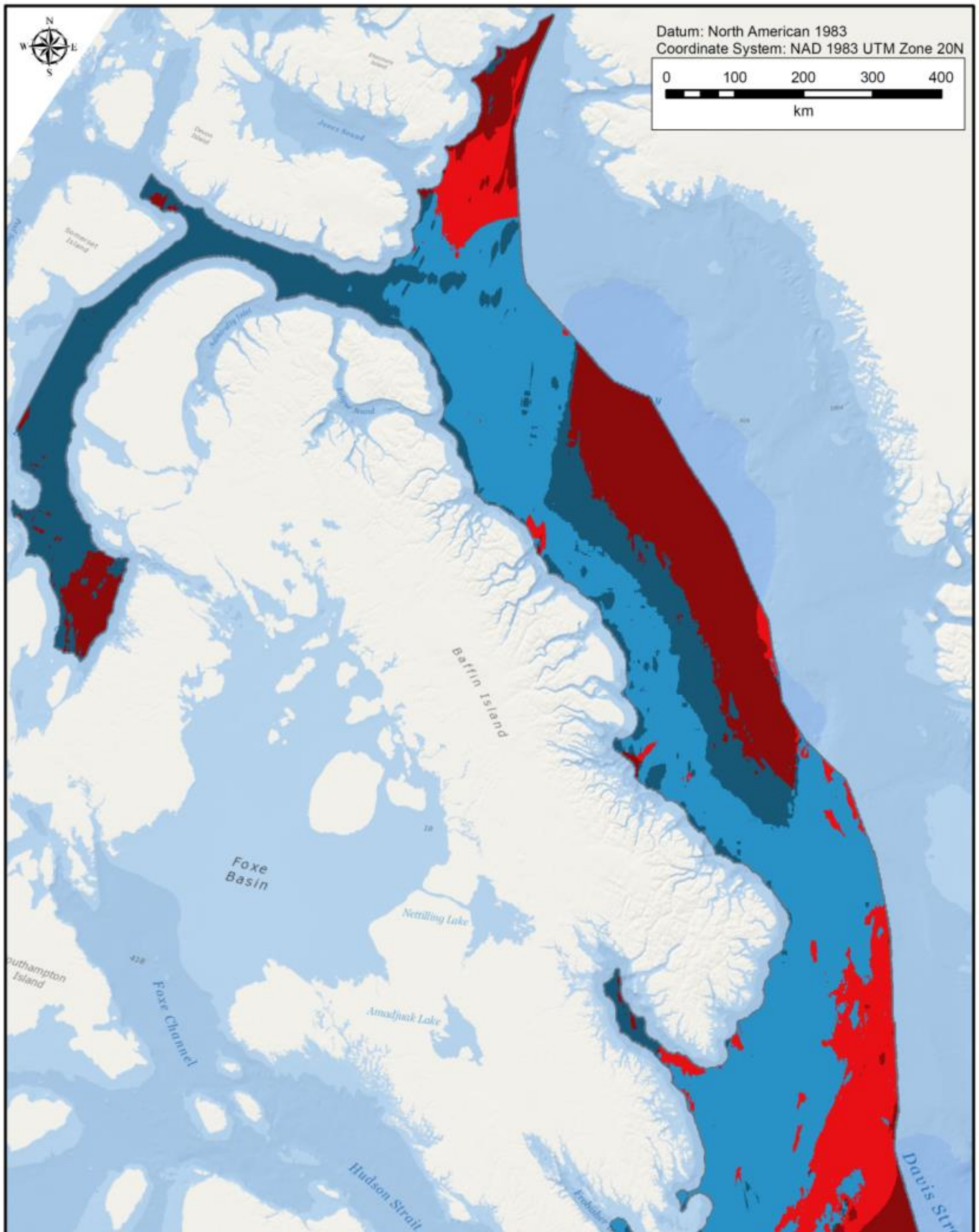

Figure 103. Predicted distribution (Pred. Dist.) of small gorgonian corals in the Eastern Arctic Region based on the prevalence threshold of 0.08 of small gorgonian coral presence and absence data used in Model 2. Also shown are the areas of model extrapolation (grey polygon may appear red or blue).

The order of importance of the environmental predictor variables in Model 2 was slightly different from that of Model 1 (Figure 104). Like Model 1, Surface Salinity Mean was the most important variable in Model 2. This was followed very distantly by Surface Temperature Average Maximum, Maximum Average Summer Mixed Layer Depth, and the remaining variables in the model. Partial dependence of small gorgonian presence and absence data on the top 6 predictor variables is shown in Figure 105. Presence probability of small gorgonians was highest at Surface Salinity Mean values of 32.5 and higher.

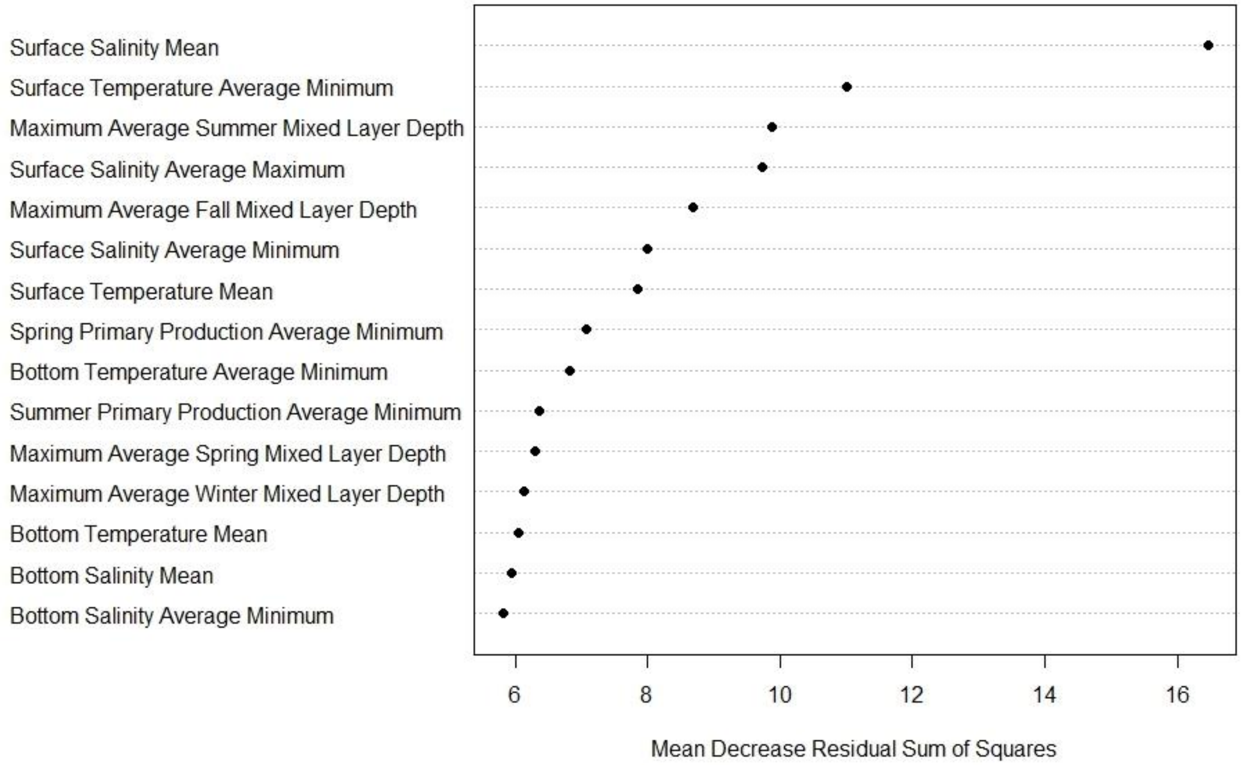

Figure 104. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced small gorgonian coral presence and absence data within the Eastern Arctic Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

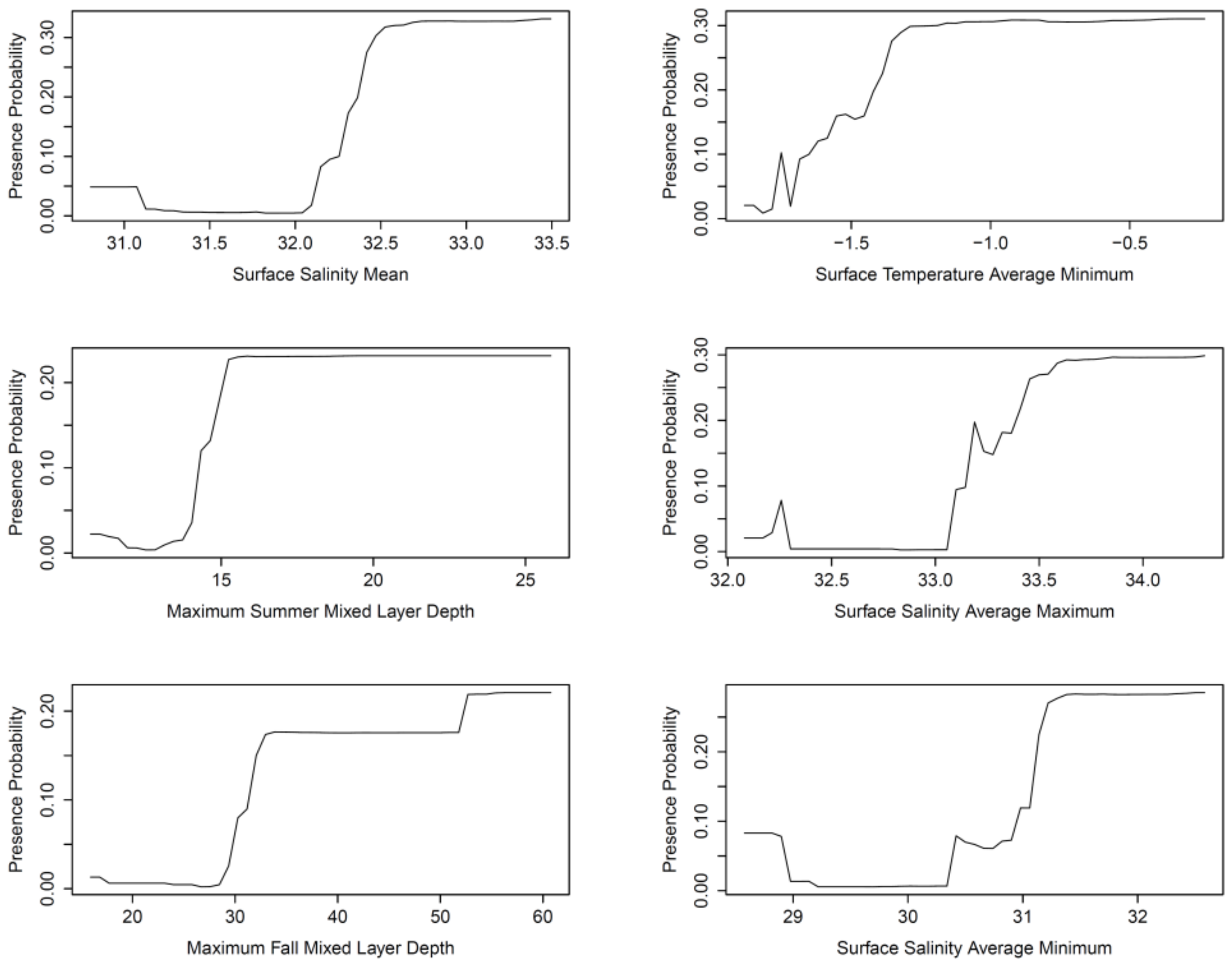

Figure 105. Partial dependence plots of the top six predictors from the random forest model of small gorgonian coral unbalanced presence and absence data collected within the Eastern Arctic Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model Selection

The random forest model using all small gorgonian coral records and unbalanced species prevalence (Model 2) was selected as the best predictor of small gorgonian coral distribution in the Eastern Arctic Region. Accuracy measures were very good and similar for both models, with AUC values around 0.9 in both cases (0.90 for Model 1 and 0.89 for Model 2). Sensitivity and Specificity was higher than 0.8 in both models. However, Model 1 (balanced species prevalence) was considered a worse predictor of presence probability of this group than Model 2 due to its exaggeration of high presence probability in some areas, mainly in the deep waters along the Davis Strait. This phenomenon was due to random down-sampling of the absence data in areas with no data that were predicted as high probability areas. This model also increases the extrapolated areas. Model 2, which was generated using the same presence-absence dataset but using all absence data, produced a more realistic presence probability surface with no exaggeration beyond the location of presence points.

Validation of Selected Model Using Independent Data

Figure 106 shows the predicted presence probability of small gorgonian corals generated from Model 2 at the location of small gorgonian coral records collected during two DFO scientific missions to the Eastern Arctic in 2012 and 2013. Records from both years were combined for display as there was little overlap between them. Small gorgonian coral records from these surveys were concentrated in the Narwhal Over-wintering and Deep-Sea Coral Conservation Area and the Hatton Basin Voluntary Closure in Davis Strait. Of the 19 small gorgonian coral records from the 2012 survey, 10 (52%) were predicted as presence based on the prevalence threshold of 0.08. Of the 12 records from the 2013 survey, 100% were predicted as presence. The records predicted as absence by the model were located in the Narwhal Closure Area in Baffin Bay, suggesting that the model is under-predicting small gorgonian presence in this area. In the Hatton Basin closure, presence probability at the location of these sea pen records was low.

There were a large number of records (1588) of small gorgonian corals from the Fisheries Observer Program used for model validation (Figure 107). Of these, 1179 (74%) were predicted as presence by Model 2. The absences were located in the Davis Strait. Records predicted with a high presence probability were located north and south of the Narwhal Overwintering and Deep-Sea Coral Conservation Area. As explained for the other taxa, the expectation is for more mismatches arising from presences recorded where the start position indicates an absence, due to the potential for transit over presence areas during the tows.

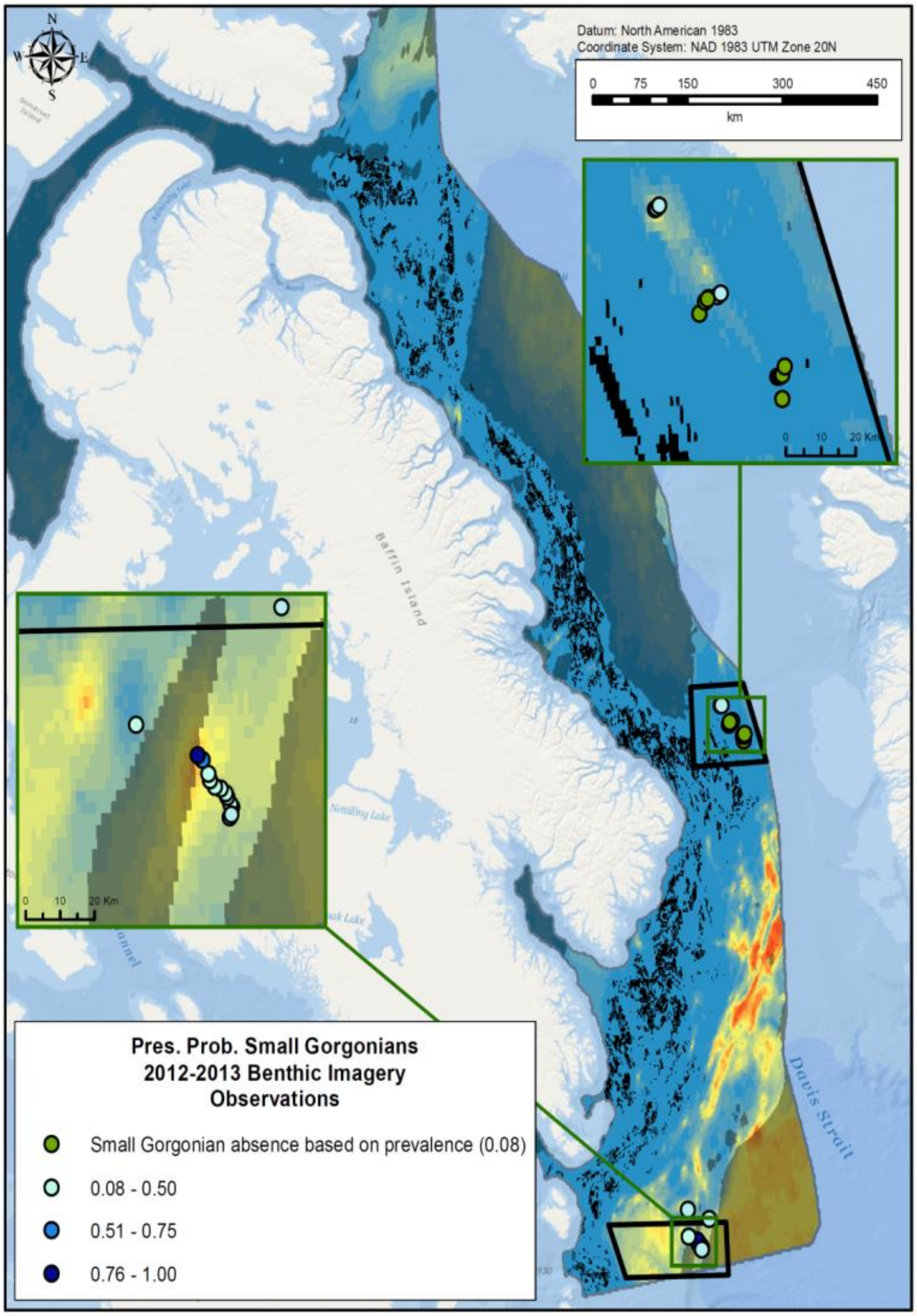
N
S
E
W
Datum: North American 1983
Coordinate System: NAD 1983 UTM Zone 20N
0 75 150 300 450
km
0 10 20 Km
0 10 20 Km
Baffin Island
Nettilling Lake
Davis Strait
Pres. Prob. Small Gorgonians
2012-2013 Benthic Imagery
Observations
● Small Gorgonian absence based on prevalence (0.08)
○ 0.08 - 0.50
● 0.51 - 0.75
● 0.76 - 1.00

Figure 106. Validation of small gorgonian coral presence probability from Model 2 using *in situ* camera records of small gorgonians collected during DFO scientific missions conducted in 2012 and 2013 (records were combined for display). Also shown are the Narwhal Overwintering and Deep-Sea Coral Conservation Area and the Hatton Basin Voluntary Closure Area. Inset maps show the Narwhal (top) and Hatton Basin (bottom) closures.

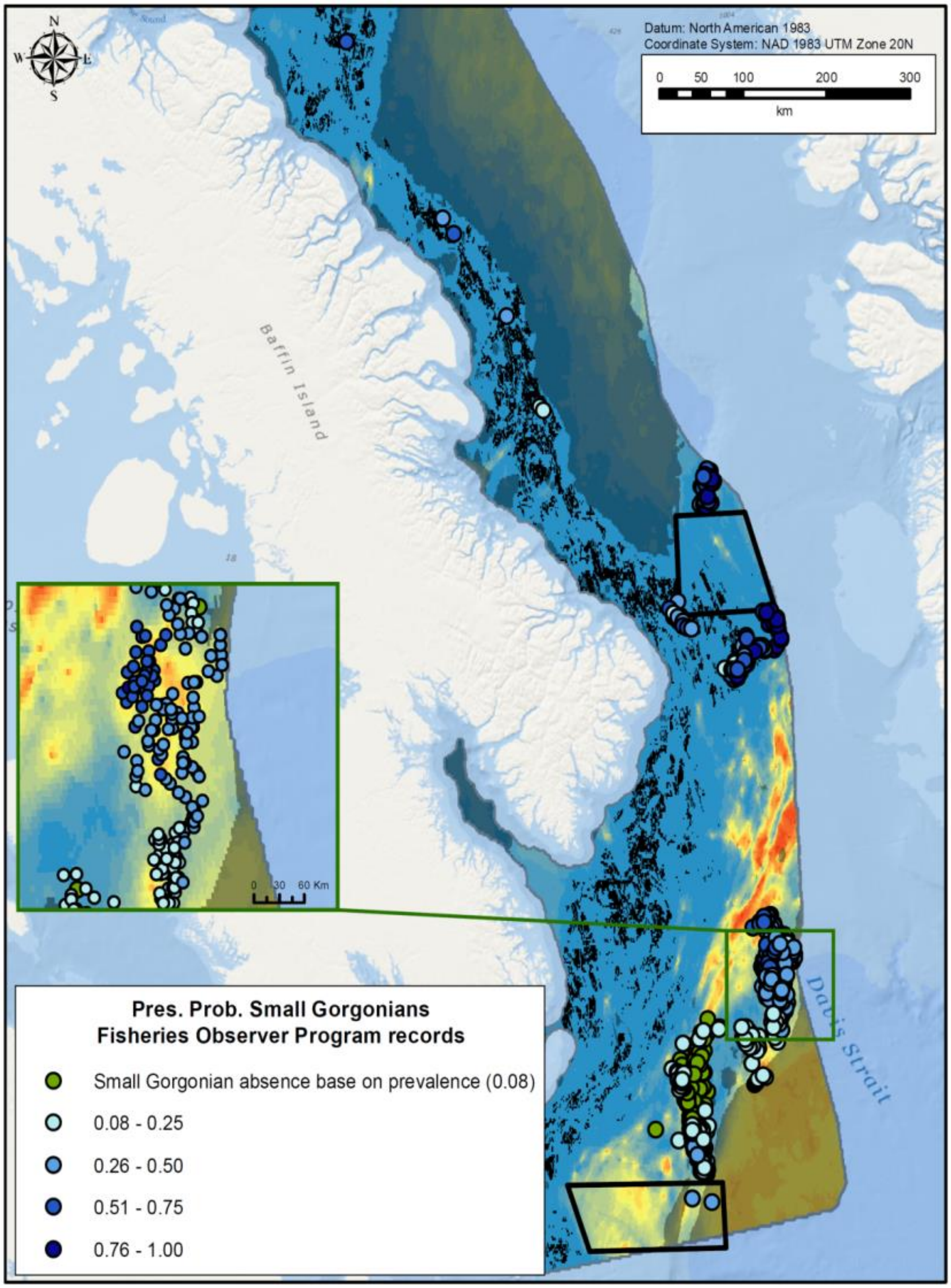
N
W
E
S
Datum: North American 1983
Coordinate System: NAD 1983 UTM Zone 20N
0 50 100 200 300
km
Baffin Island
Davis Strait
0 30 60 Km
Pres. Prob. Small Gorgonians
Fisheries Observer Program records
Small Gorgonian absence base on prevalence (0.08)
0.08 - 0.25
0.26 - 0.50
0.51 - 0.75
0.76 - 1.00

Figure 107. Validation of small gorgonian coral presence probability from Model 2 using small gorgonian coral records collected by the Fisheries Observer Program between 2004 and 2013. Also shown are the Narwhal Overwintering and Deep-Sea Coral Conservation Area and the Hatton Basin Voluntary Closure Area. Inset maps show the Narwhal (top) and Hatton Basin (bottom) closures.

Prediction of Small Gorgonian Biomass Using Random Forest

Alfredo Trawl Gear

The accuracy measures of the regression random forest model on mean small gorgonian coral biomass per grid cell from DFO trawl surveys with Alfredo gear are presented in Table 27. The highest R^2 value was 0.677, while the average was 0.292 ± 0.213 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.044 ± 0.026 SD. The percent variance explained ranged from 6.91 to 27.97 with an average of 11.78 ± 6.32 SD.

Figures 108 and 109 show the predicted biomass surface of small gorgonian corals from Alfredo trawl gear records. The majority of the spatial extent was predicted to have low (0 – 0.006 kg) small gorgonian biomass, even in areas where there are moderate biomass records (Figure 109). The southeast corner of the study extent in Davis Strait was predicted to have the highest biomass of small gorgonian corals. This area is considered extrapolated by the model (Figure 109). Smaller pockets of moderate to high biomass were predicted to occur in Baffin Basin and in Lancaster Sound.

Table 27. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of average small gorgonian coral biomass (kg) per grid cell recorded from DFO trawl surveys with Alfredo gear conducted in the Eastern Arctic Region. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	**R^2**	**RMSE**	**NRMSE**	**Percent (%) variance explained**
1	0.139	0.010	0.040	13.67
2	0.086	0.009	0.039	12.89
3	0.451	0.010	0.040	6.91
4	0.677	0.003	0.011	9.01
5	0.185	0.012	0.051	14.31
6	0.024	0.008	0.034	7.93
7	0.448	0.008	0.033	10.17
8	0.314	0.006	0.024	7.05
9	0.128	0.014	0.058	7.89
10	0.471	0.025	0.106	27.97
Mean	**0.292**	**0.010**	**0.044**	**11.78**
SD	**0.213**	**0.006**	**0.026**	**6.32**

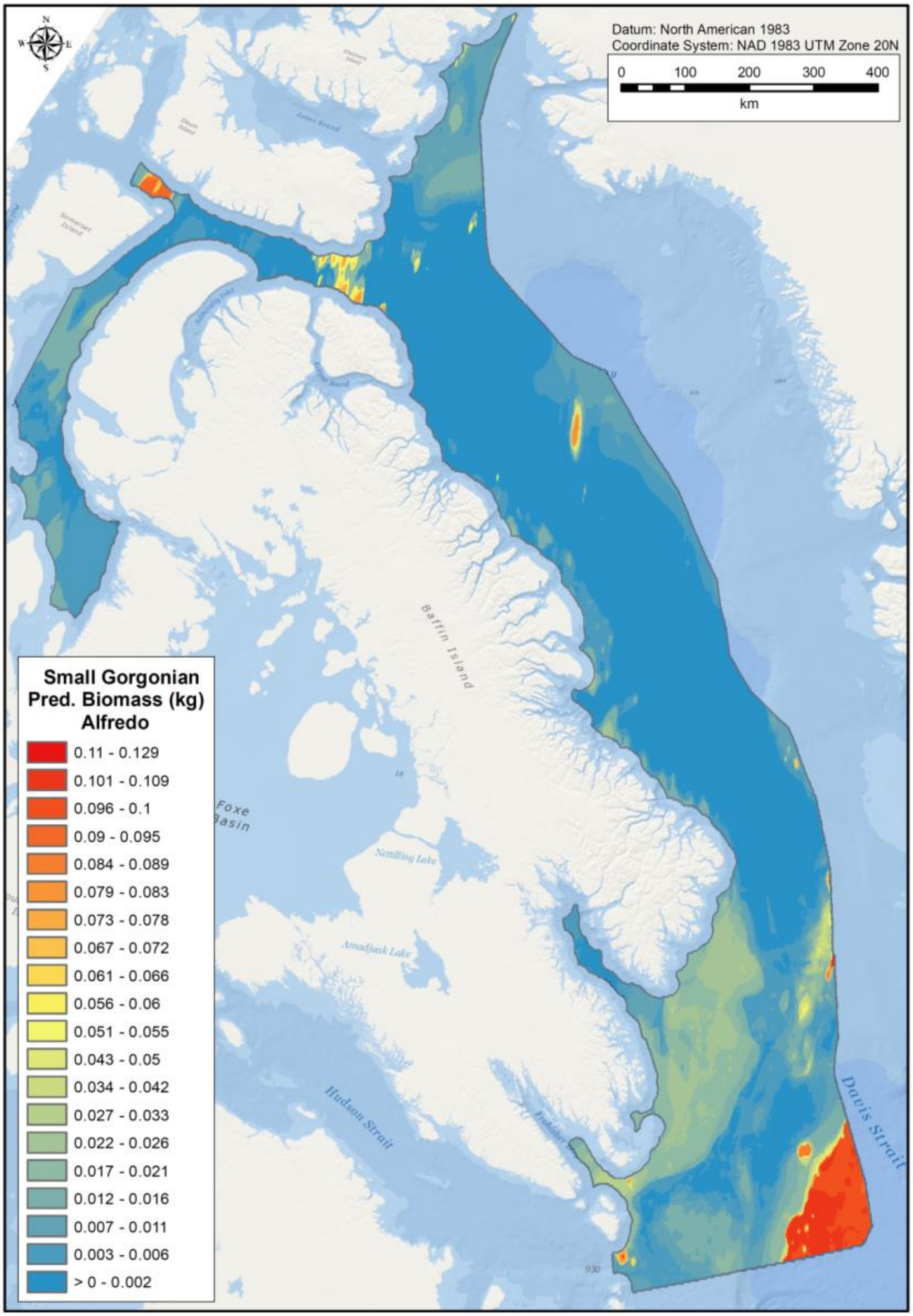

Figure 108. Predictions of biomass (kg) per grid cell of small gorgonian corals from catch data recorded in DFO trawl surveys with Alfredo gear conducted in the Eastern Arctic Region between 2006 and 2014.

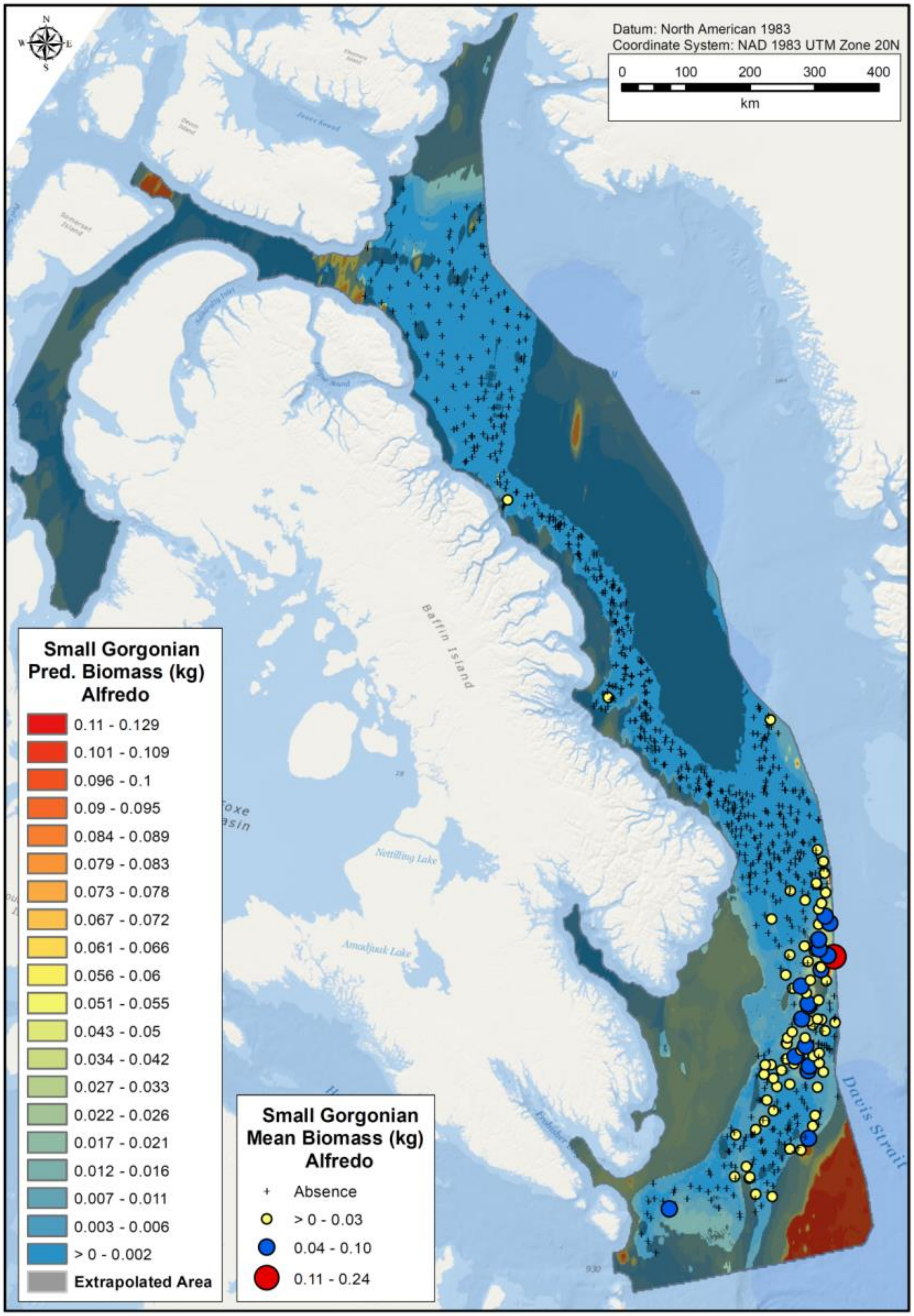

Figure 109. Predictions of biomass (kg) per grid cell of small gorgonian corals from catch data recorded in DFO trawl surveys with Alfredo gear conducted in the Eastern Arctic Region between 2006 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting small gorgonian biomass are shown in Figure 110. Bottom Shear Average Maximum was the most important variable followed distantly by Bottom Shear Average Range, Bottom Current Average Maximum, and the remaining variables in the model. The partial dependence of small gorgonian coral biomass on the top 6 most important variables is shown in Figure 111. Predicted biomass was highest at Bottom Shear Average Maximum values greater than ~0.027 m s^{-1}.

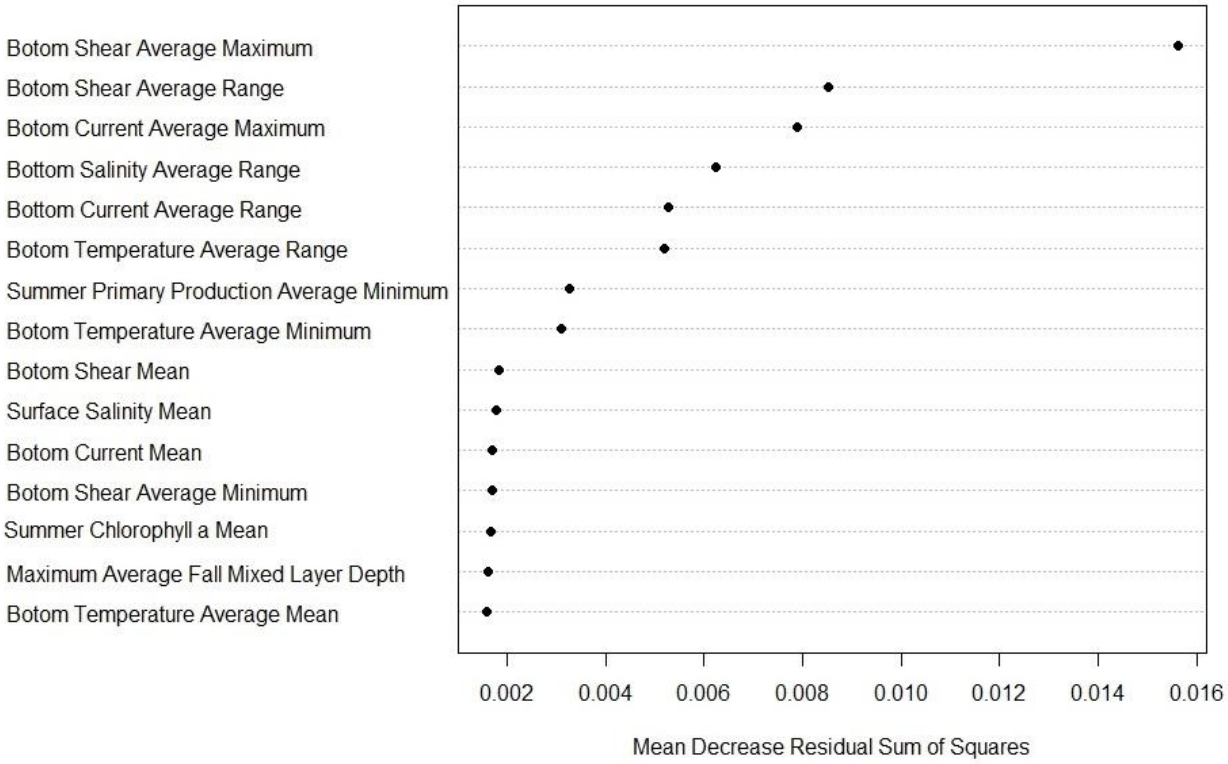

Figure 110. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on small gorgonian coral mean biomass data averaged per grid cell will Alfredo gear. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

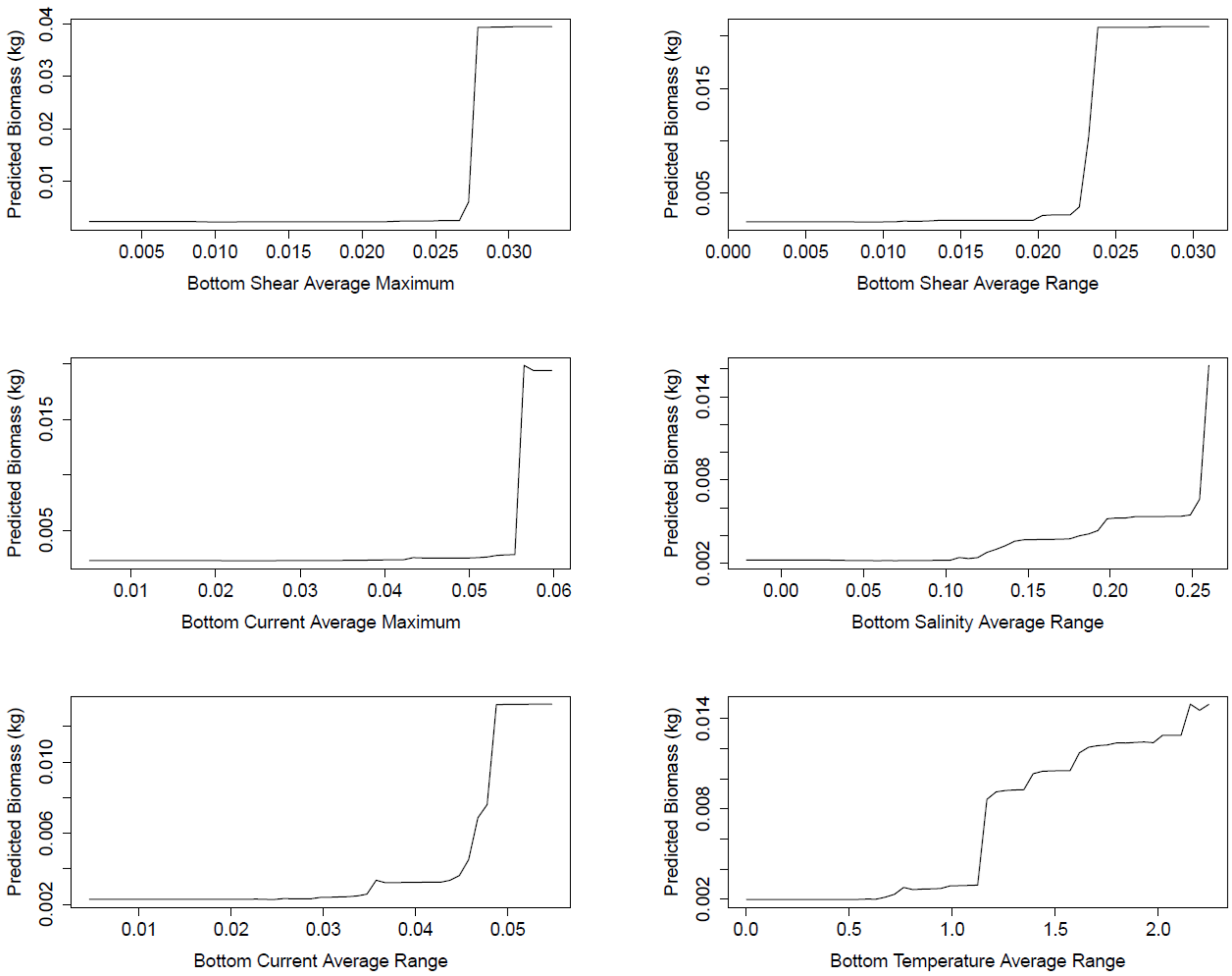

Figure 111. Partial dependence plots of the top six predictors from the random forest model of small gorgonian coral biomass data collected with Alfredo gear within the Eastern Arctic Region, ordered left to right from the top.

Campelen Trawl Gear

The accuracy measures of the regression random forest model on mean small gorgonian coral biomass per grid cell from DFO trawl surveys conducted with Campelen gear are presented in Table 28. The highest R^2 value was 0.483, while the average was 0.100 ± 0.194 SD, indicating poor model performance. The average Normalized Root-Mean-Square Error (NRMSE) was 0.018 ± 0.026 SD. The percent variance explained for each fold was negative.

Figures 112 and 113 show the predicted biomass surface of small gorgonian corals. The majority of the spatial extent was predicted to have low (< 0.05 kg) small gorgonian biomass. The highest predicted biomass (up to 0.91 kg) occurred in a small area in Davis Strait and was associated with a higher catch value (Figure 113). A large area covering the Baffin Shelf and Basin was predicted to have moderate small gorgonian coral biomass. Like the model using Alfredo trawl gear records, this model predicted higher biomass in the southeast corner of the extent in Davis Strait.

Table 28. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of average small gorgonian coral biomass (kg) per grid cell recorded from DFO trawl surveys with Campelen gear conducted in the Eastern Arctic Region. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	**R^2**	**RMSE**	**NRMSE**	**Percent (%) variance explained**
1	0.003	0.004	0.003	-16.66
2	0.018	0.007	0.005	-16.37
3	2.149×10^{-4}	0.019	0.013	-13.50
4	3.637×10^{-4}	0.011	0.007	-17.58
5	0.046	0.009	0.006	-11.78
6	1.156×10^{-6}	0.006	0.004	-14.19
7	0.451	0.128	0.085	-15.47
8	1.266×10^{-4}	0.019	0.013	-14.91
9	0.483	0.052	0.035	-16.19
10	4.688×10^{-4}	0.011	0.008	-14.75
Mean	**0.100**	**0.027**	**0.018**	**-15.14**
SD	**0.194**	**0.038**	**0.026**	**1.70**

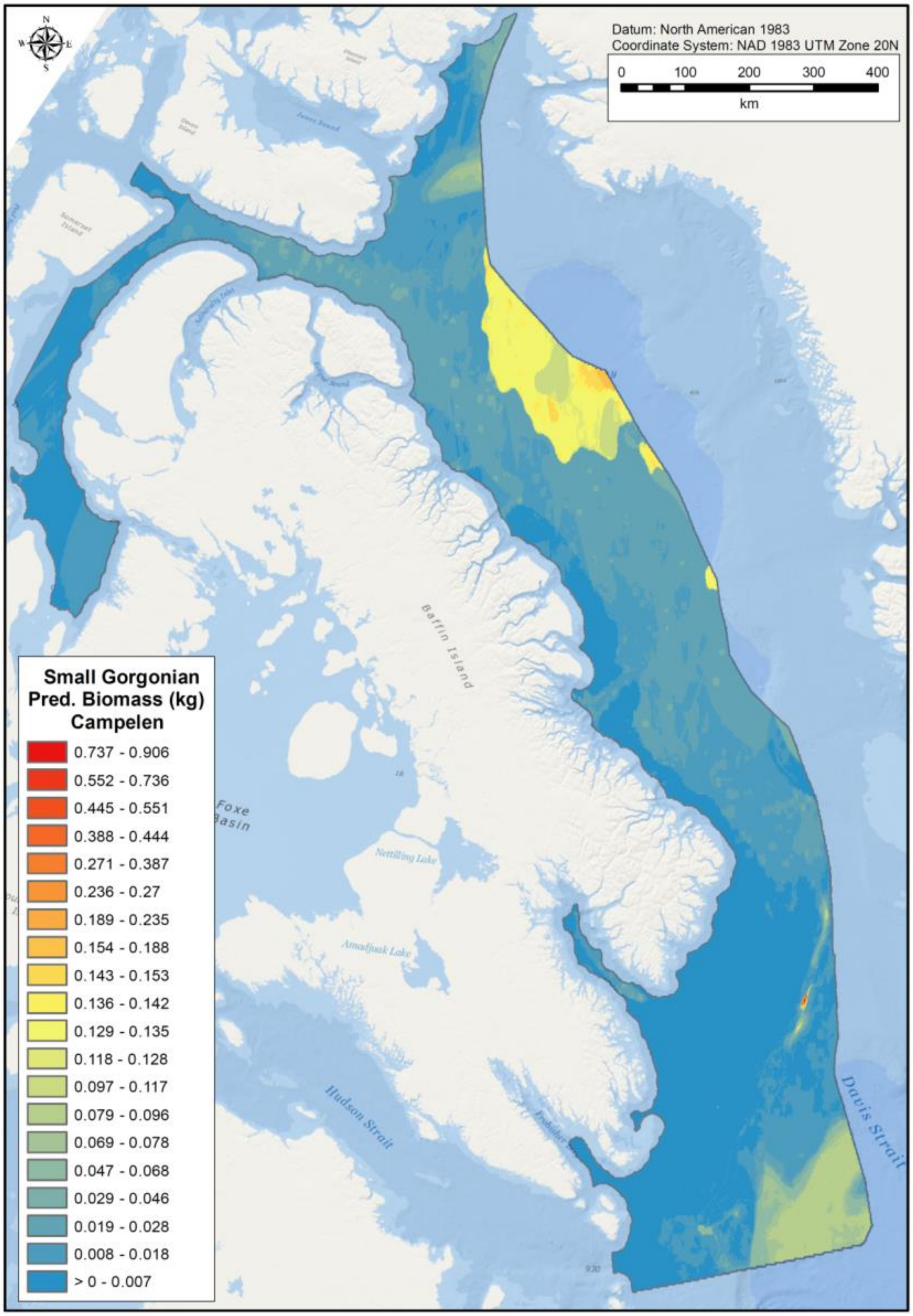

Figure 112. Predictions of biomass (kg) of small gorgonian corals from catch data recorded in DFO trawl surveys with Alfredo gear conducted in the Eastern Arctic Region between 2006 and 2014.

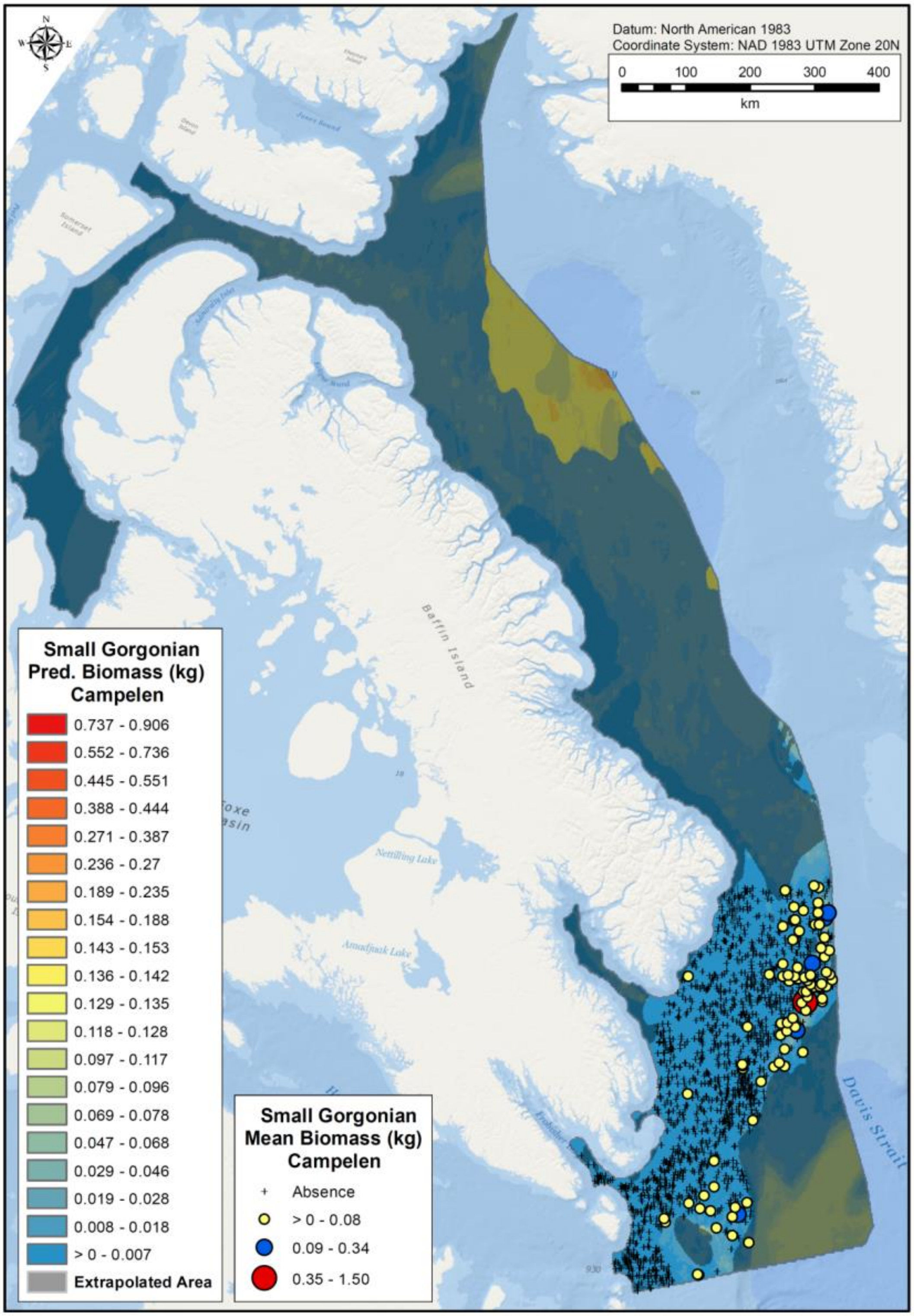

Figure 113. Predictions of biomass (kg) of small gorgonians from catch data recorded in DFO trawl surveys with Campelen gear conducted in the Eastern Arctic Region between 2005 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting small gorgonian biomass are shown in Figure 114. Bottom Temperature Average Range was the most important variable followed very distantly by Bottom Salinity Average Range, and the remaining variables in the model. The partial dependence of small gorgonian coral biomass on the top 6 most important variables is shown in Figure 115. Predicted biomass was highest at Bottom Temperature Average Range values greater than 2.0 °C.

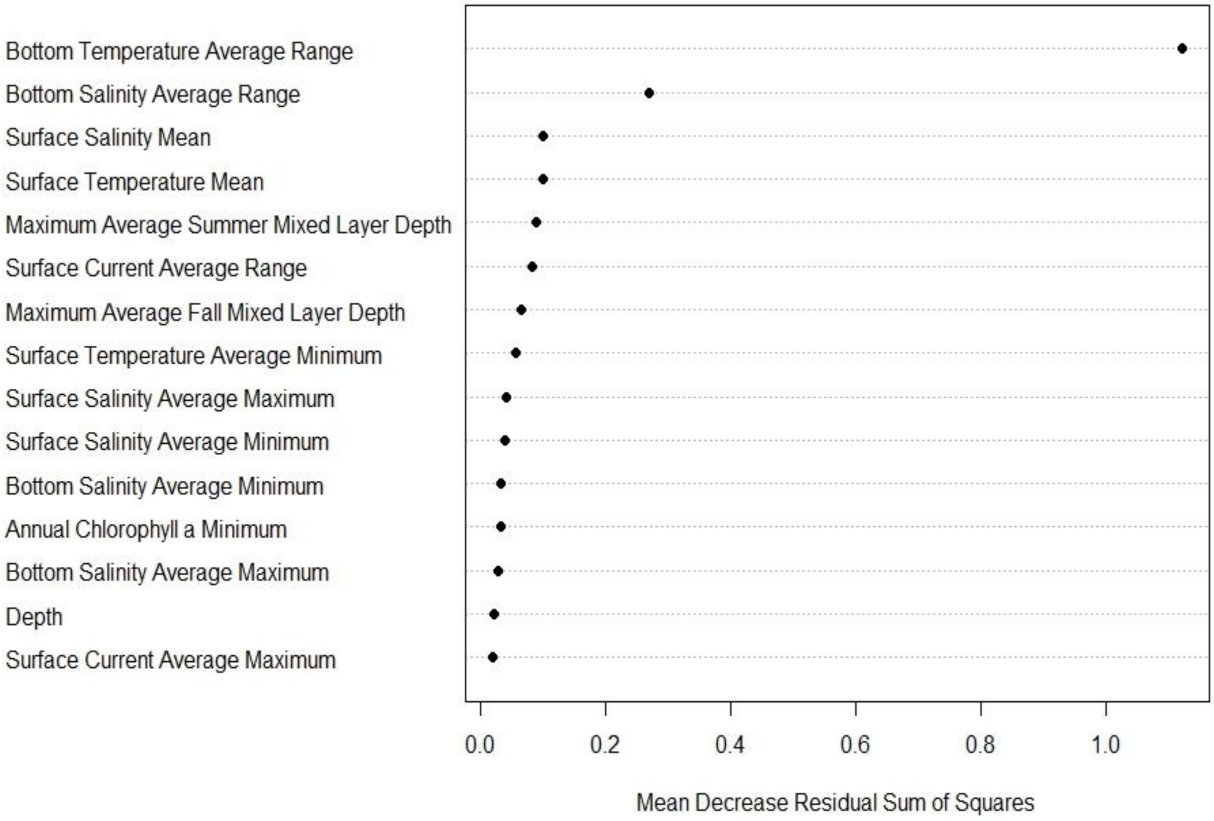

Figure 114. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on small gorgonian coral mean biomass data averaged per grid cell will Campelen gear. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

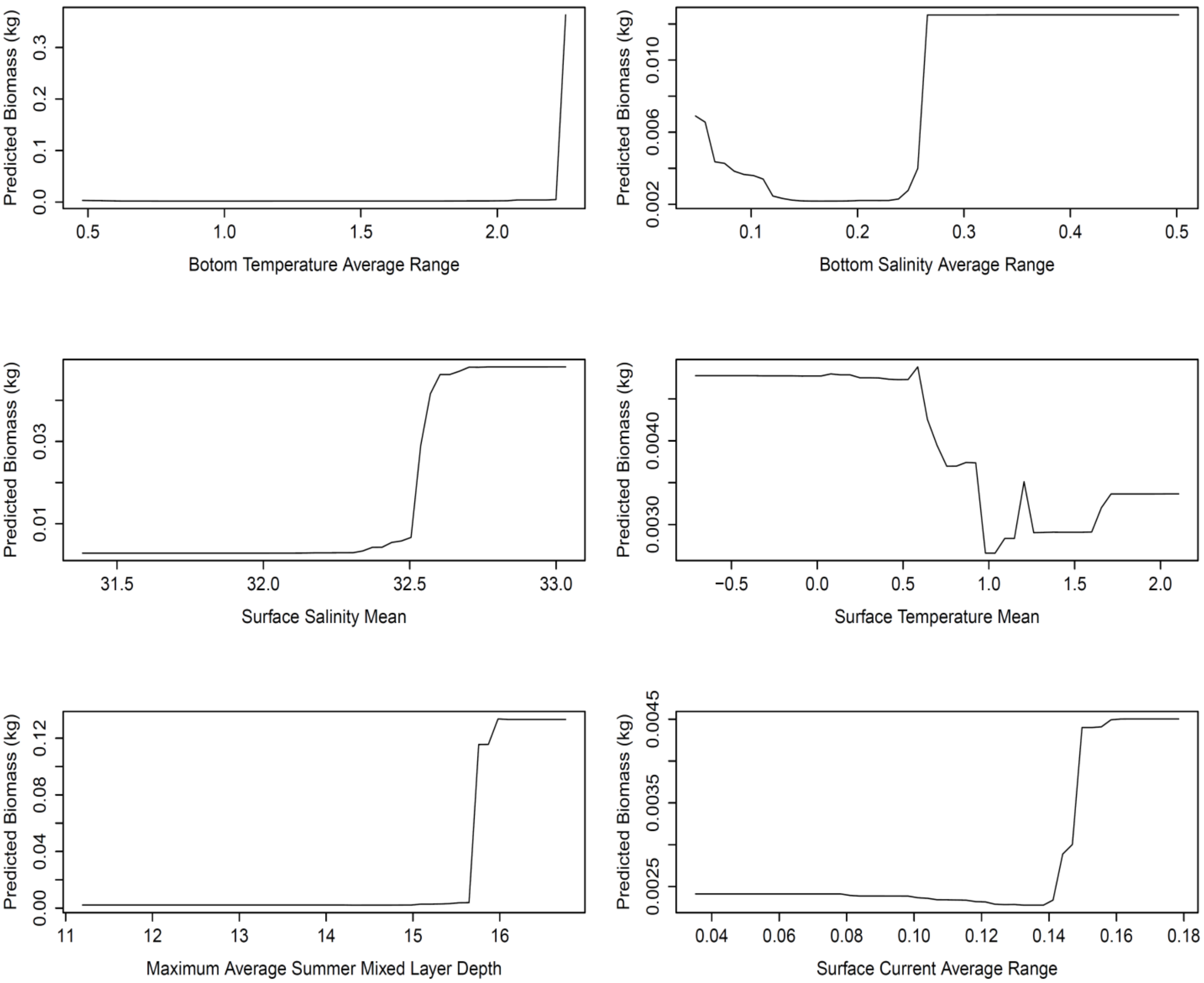

Figure 115. Partial dependence plots of the top six predictors from the random forest model of small gorgonian coral biomass data collected with Campelen gear within the Eastern Arctic Region, ordered left to right from the top.

DISCUSSION

The species distribution model results for the Hudson Strait-Ungava Bay and Eastern Arctic Regions were variable. Cross-validated AUCs ranged from 0.643 to 0.894 for the presence-absence models (Table 28). Sensitivity and Specificity were also variable and both were poor for the sponges in the Hudson Strait-Ungava Bay Region. Overall the presence-absence models performed best with the small gorgonian corals in the Eastern Arctic Region, with the sea pen model also having a very good AUC and Sensitivity but slightly poorer Specificity (Table 28). The reasons for these results are not clear. The different sea pen species included in the Sea Pen (Pennatulacea) taxon modelled in this study differ in their depth and spatial distribution (see Figure 116. To determine whether taxonomic resolution was a factor, we performed an additional random forest model generated with just the *Umbellula* sp. records from Alfredo trawl gear. The result was a poorer performance of this model (AUC = 0.761, Sensitivity = 0.636, Specificity = 0.697 vs. Table 28) compared to the model run on all sea pen taxa combined. Hui et al. (2013) compared the predictive performance of species groups against separate SDMs using a number of multi-species data sets. They found that using groups of species with similar environmental requirements improved model accuracy and discriminatory capacity compared to separate SDMs. Further, the approach was endorsed for rare species. Consequently, we do not believe that taxonomic resolution explains the lower model performance compared with other regions (e.g., Beazley et al., 2016a). The Eastern Arctic Region is governed by a complex physical environment, with several interacting environmental conditions likely controlling the distribution of the benthos and influencing SDM performance.

Table 28. Summary of the mean accuracy measures for selected presence-absence models for each of the four taxonomic groups.

	AUC	**Sensitivity**	**Specificity**	**Top Predictor Variable**
Hudson Strait – Ungava Bay Region				
Sponges (Porifera)	0.643	0.574	0.612	Surface Current Mean
Eastern Arctic Region				
Sponges (Porifera)	0.791	0.709	0.736	Depth
Sea Pens (Pennatulacea)	0.838	0.814	0.721	Bottom Salinity Avg. Range
Large Gorgonian Corals	0.752	0.626	0.786	Bottom Temp. Avg. Minimum
Small Gorgonian Corals	0.894	0.8212	0.823	Surface Salinity Mean

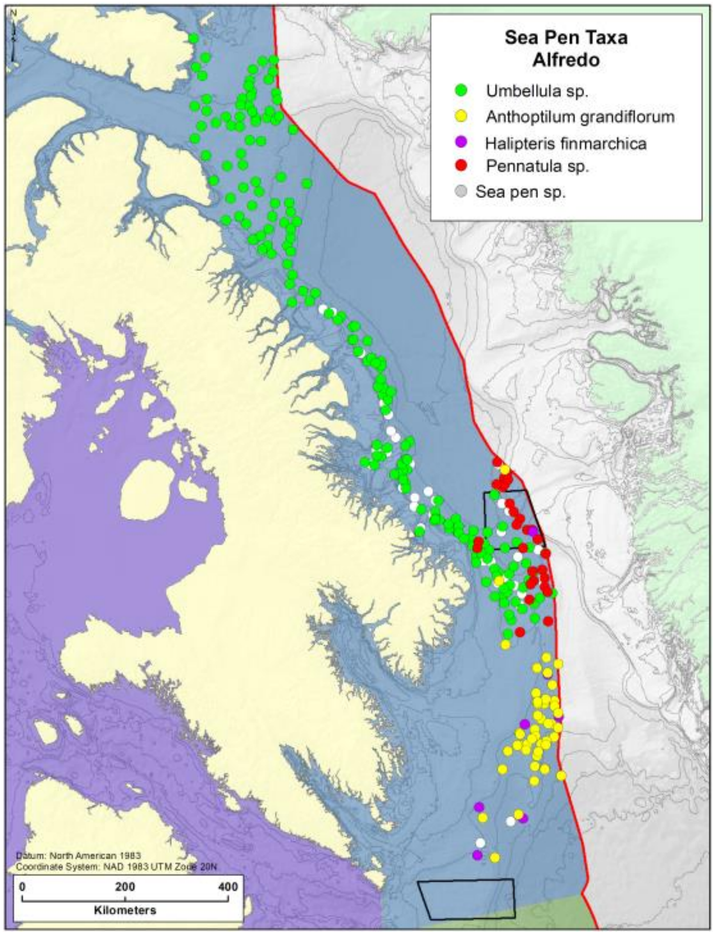

Figure 116. Distribution of individual sea pen taxa comprising the higher-level Sea Pen (Pennatulacea) group modelled in this report.

Knudby et al. (2013a) used random forest to predict the distribution of several sponge species and sponge grounds in the northwest Atlantic, including the Eastern Arctic Region. In the Eastern Arctic, models were run on two subareas, the Hudson Strait subarea and the Baffin Bay subarea, designated based on the available data and the oceanographic conditions in the area. The boundary between these two subareas lies at the sill separating the Labrador and Baffin Basins (at 65ºN). Knudby et al. (2013a) found models trained on data from the Baffin Bay subarea were poor predictors of sponge distribution elsewhere in the northwest Atlantic, and noted that correlations between environmental variables north and south of the sill at 65ºN greatly differed, likely resulting from the different oceanographic conditions in Baffin Bay compared to the Hudson Strait subarea. We generated an additional model on all sponge presence-absence records north and south of the sill at 65ºN to determine whether model performance could be improved over models generated on sponge data over the full extent (Table 29, Figure 117). The cross-validated AUC, sensitivity, and specificity of models generated north and south of 65ºN were very comparable to one another and to the model generated on data over the entire spatial extent (see Table 29), although the model south of 65ºN performed slightly worse. The predicted presence probability surfaces were nearly identical between that of the full extent (left panel of Figure 117) and those above and below the sill (right panel of Figure 117). Examination of the top environmental variables for the presence-absence models of all four taxonomic groups in the

Eastern Arctic Region (see Figure 118) revealed no strong spatial differences in Depth and Bottom Salinity Average Range between Baffin Bay and Davis Strait. Bottom Temperature Average Range and Surface Salinity Mean showed higher values in the Davis Strait compared to Baffin Bay. These areas of higher bottom temperature and surface salinity may correspond to the flow of the Irminger Current, which circulates cyclonically around the northern Labrador Sea and along the southern edge of Davis Strait, where it then turns southwest towards the Labrador Sea (Hamilton and Wu, 2013). The Polar Water is found close to the coast and the Atlantic Water is found as a 500–800 m thick layer over the continental slope with a core at about 200–300 m depth (Yashayaev, 2007). Due to this circulation and water mass distribution, the southeast corner of the spatial extent in Davis Strait and the deep waters off the Baffin Island Shelf have a similar depth range but different temperature and salinity regimes (Figure 118). Interestingly, these areas were commonly predicted with high presence probability or biomass by the random forest models despite there being no data observations from there, and were considered extrapolated area by the models. These results highlight the complexity of the physical oceanographic conditions in the Eastern Arctic. We note that sponge absence is predicted in shelf areas where there are major inlets (Figure 29) and in the case of the Cumberland Sound, in an area where a polynya forms (DFO, 2015). Given the importance of the depth and salinity variables in determining sponge presence-absence, it is possible that freshwater input due to ice and snow melt is an unrecorded determinant. Future models could include distance from shore as a variable. That variable was found to improve model performance for lobsters in the Gulf of Maine (K. Tanaka, University of Maine, Orono, Maine, USA, pers. comm.)

Table 29. Summary of the accuracy measures for sponge presence-absence models generated on the full Eastern Arctic extent, and north and south of 65ºN.

Spatial Extent	**AUC**	**Sensitivity**	**Specificity**
Full Extent	0.791	0.709	0.736
North of 65ºN	0.791	0.743	0.689
South of 65ºN	0.778	0.681	0.740

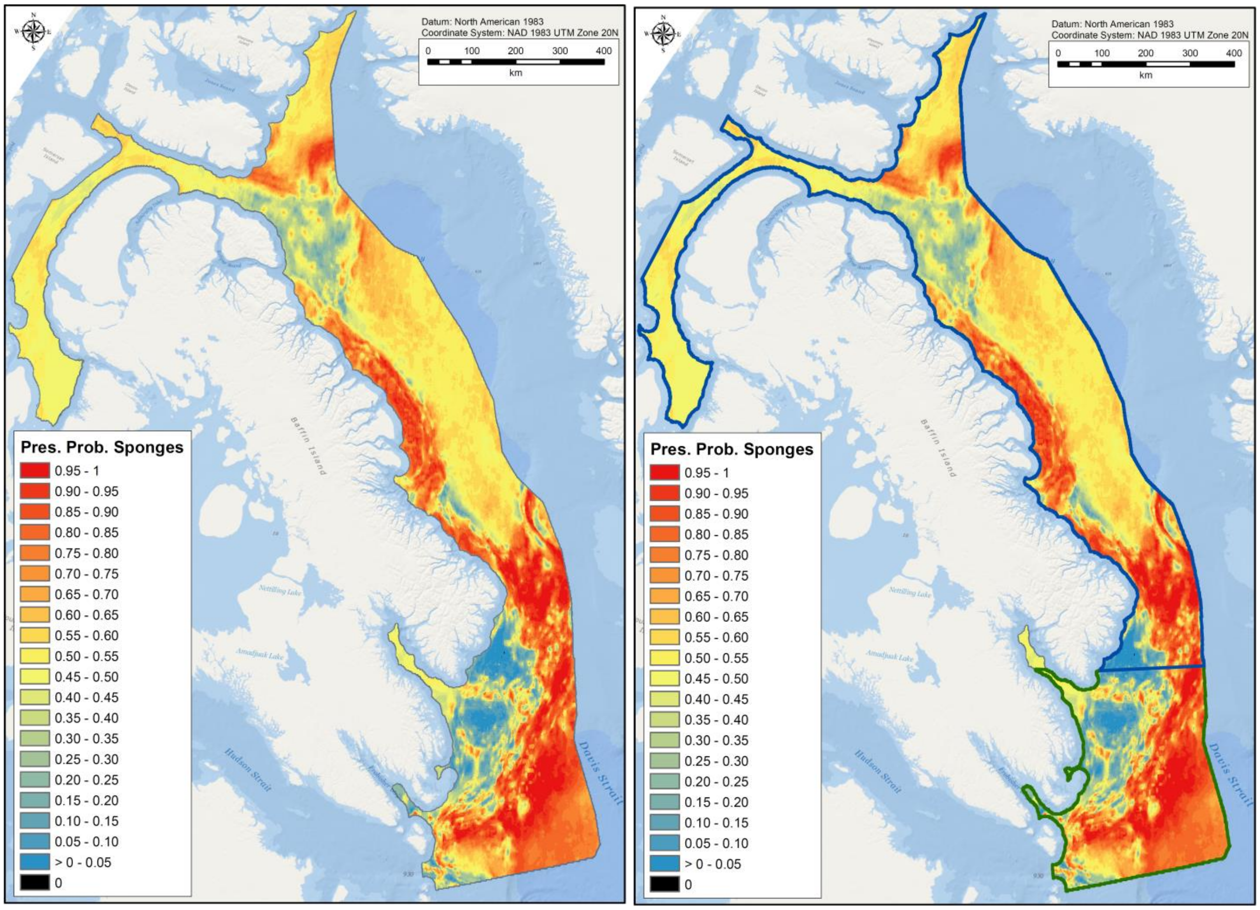

Figure 117. Spatial concordance between left) the predicted presence probability of sponges from Model 2 of this study on the full spatial extent of the Eastern Arctic Region, and right) the predicted presence probability of sponge records above (blue polygon) and below (green polygon) the sill separating the Labrador and Baffin Basins at 65ºN.

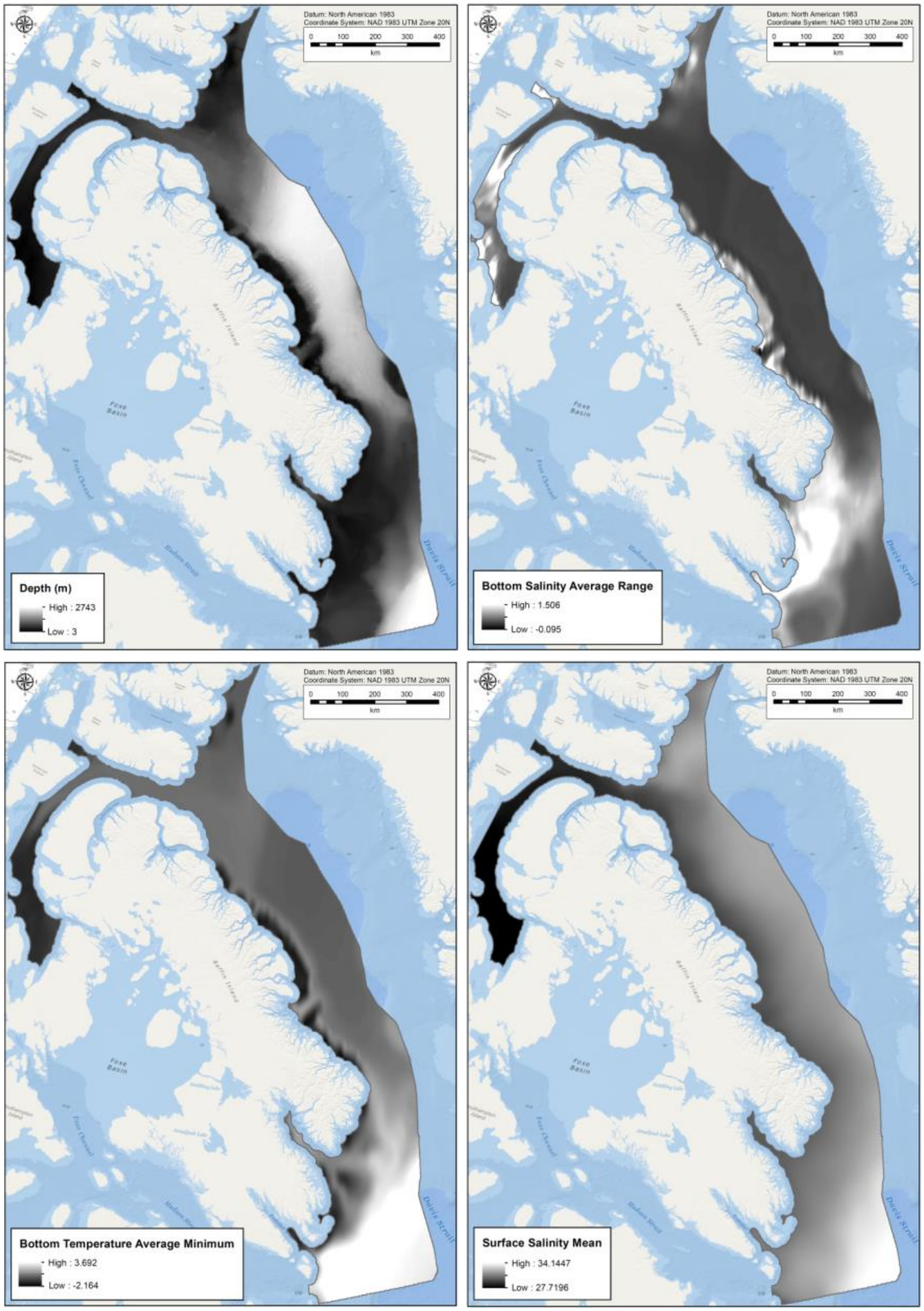

Figure 118. Spatial distribution of the top environmental predictor variables for presence-absence models of each taxonomic group modelled in the Eastern Arctic Region. Upper left: Depth (Sponges); Upper right: Bottom Salinity Average Range (Sea Pens); Lower left: Bottom Temperature Average Minimum (Large Gorgonian Corals); Lower right: Surface Salinity Mean (Small Gorgonian Corals).

The random forest biomass models performed inconsistently within taxa by gear type (see Table 29 for summary). The spatial and depth distribution of records from each gear type differed, with Alfredo records covering the greatest portion of the study extent in Baffin Bay and Davis Strait, Campelen records restricted mainly to the Davis Strait and southern Baffin Bay, and Cosmos records distributed on the slope off Baffin Island Shelf and on the eastern of the extent in Davis Strait. Despite the greater distribution of the Alfredo surveys, only two biomass random forest models using these records performed well (see results for Sponges and Small Gorgonian Corals in Table 29). For sponges, the differences between the Alfredo and Campelen biomass models are illustrated in Figure 119. The greatest differences in prediction between the gear types occurred in the southeast corner of the study extent in Davis Strait. However, the raw biomass predicted surfaces show that both models predicted high biomass to occur there, as well as in the deeper waters of Baffin Basin. Similar to the presence-absence models, both sponge biomass models predicted high biomass. These results are consistent with the presence-absence models which have also predicted high presence probability to occur there. Both areas are considered areas of model extrapolation and require validation.

Table 29. Summary of the mean accuracy measures for biomass random forest models for each of the four taxonomic groups. NRMSE = Normalized Root-Mean-Square Error; % Var. Exp. = Percentage Variance Explained.

	R^2	**NRMSE**	**% Var. Exp.**
Hudson Strait – Ungava Bay Region			
Sponges (Porifera)			
Cosmos Trawl Gear	0.101	0.0746	-8.67
Eastern Arctic Region			
Sponges (Porifera)			
Alfredo Trawl Gear	0.327	0.042	15.44
Campelen Trawl Gear	0.480	0.032	31.91
Cosmos Trawl Gear	0.295	0.031	-14.81
Sea Pens (Pennatulacea)			
Alfredo Trawl Gear	0.089	0.062	-3.03
Campelen Trawl Gear	0.041	0.042	-8.99
Cosmos Trawl Gear	0.0868	0.101	-12.47
Large Gorgonian Corals			
Alfredo Trawl Gear	0.006	0.021	-6.54
Campelen Trawl Gear	0.186	0.013	16.86
Small Gorgonian Corals			
Alfredo Trawl Gear	0.292	0.044	11.78
Campelen Trawl Gear	0.100	0.018	-15.14

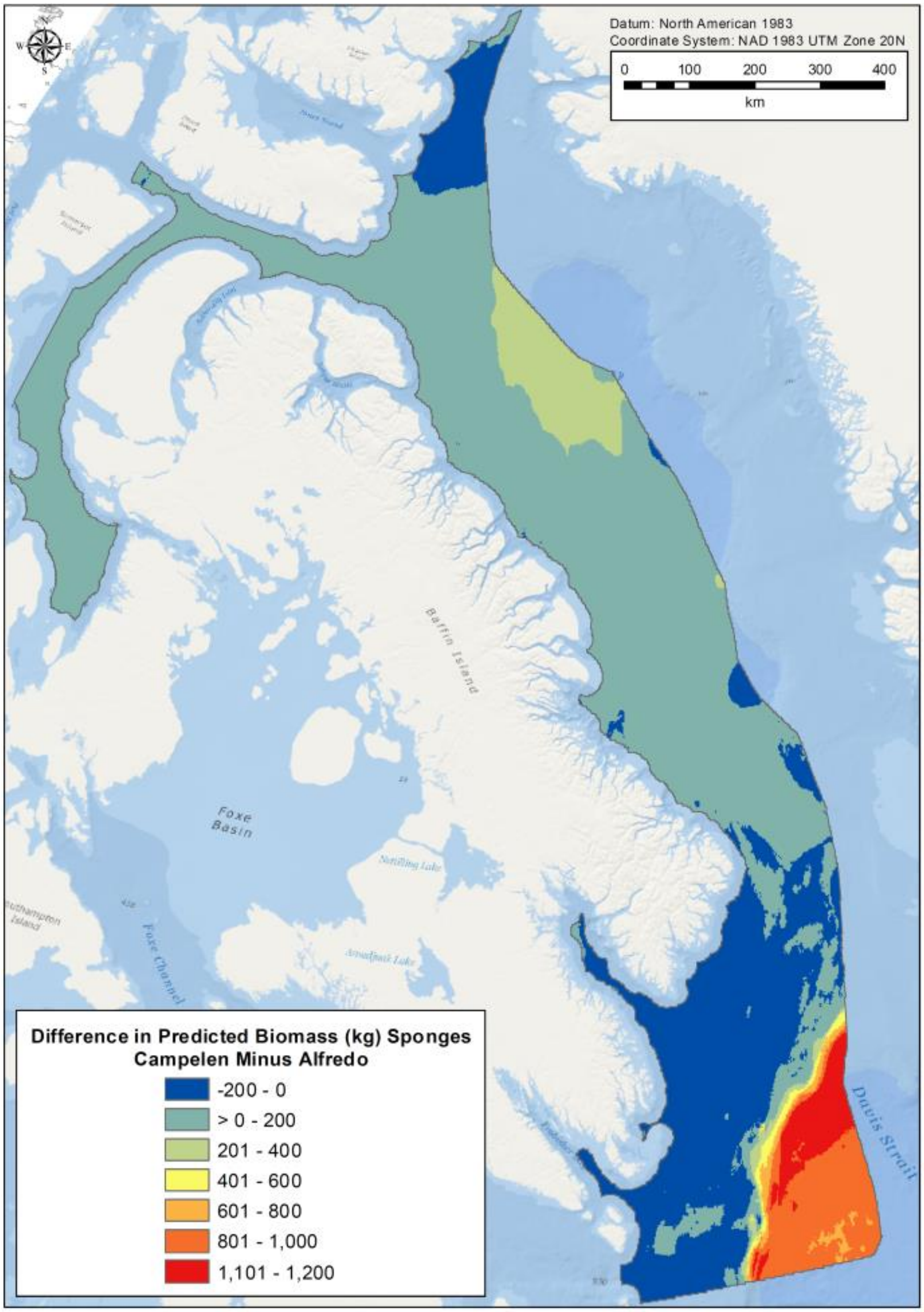

Figure 119. Difference in predicted biomass (kg) surfaces of sponges from Alfredo and Campelen trawl gear. The predicted biomass surface from Alfredo gear was subtracted from that of the Campelen model. Areas classified as negative biomass (-200 to 0 kg) are where the Alfredo predicted biomass was higher than the Campelen.

In conclusion, the probability of occurrence species distribution models for the Eastern Arctic performed well overall, with very good to excellent model performance for the sea pens and small gorgonian corals. The sponges had good model performance and efforts to run the models based on oceanographic provinces did not greatly improve the results. Although it was not the case for sea pens, improvement in the sponge models in this region may be seen through further taxonomic breakdown of the catch records. The relatively poorer model performance for the sponges in Hudson Strait may be due to the fewer response data records there. The top predictor, Surface Current Mean did not show any unusual patterns in the absolute values or error distribution which would explain the results. Re-analysis in the future when more survey records are available could improve the model performance. The GAM models were supportive of the RF models, and along with the fisheries observer records used to validate, infer robustness to the RF predictions. Our models can be used in support of decision-making under 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) and an International Governance Strategy (IGS) project to EK. 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. We also thank Vonda Wareham (DFO-NWAFC, St. John’s, NL) and Dr. Susanna Fuller (Ecology Action Centre, Halifax, Nova Scotia) for their kind provision of the fishery observer data that they had in their possession.

REFERENCES

Beazley, L., Kenchington, E., Murillo, J., Lirette, C., Guijarro, J., McMillan, A., and Knudby, A. 2016a. Species Distribution Modelling of Corals and Sponges in the Maritimes Region for Use in the Identification of Sensitive Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3172: vi + 188p.

Beazley, L., Lirette, C., Sabaniel, J., Wang, Z., Knudby, A., and Kenchington, E. 2016b. 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.

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

DFO. 2015. Ecologically and Biologically Significant Areas in Canada’s Eastern Arctic Biogeographic Region, 2015. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2015/049.

DFO. 2009. Development of a Framework and Principles for the Biogeographic Classification of Canadian Marine Areas. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2009/056.

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

Elith, J., Kearney, M., and Phillips, S. 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1: 330-342.

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. Eds: Drew, C.A., Wiersma, Y.F., and Huettmann, F. Springer, NY.

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

Falk-Petersen, S., Mayzaud, P., Kattner, G., and Sargent, J. 2009. Lipids and life strategy of Arctic *Calanus*. Mar. Biol. Res. 5: 18-39.

FAO. 2009. Report of the Technical Consultation on International Guidelines for the Management of Deep-sea Fisheries in the High Seas. Rome, 4–8 February and 25–29 August 2008. ftp://ftp.fao.org/docrep/fao/011/i0605t/i0605t00.pdf

Frajka-Williams, E., and Rhines, P. 2010. Physical controls and interannual variability of the Labrador Sea spring phytoplankton bloom in distinct regions. Deep-Sea Res. I. 57: 541-552.

Guijarro, J., Beazley, L., Lirette, C., Kenchington, E., Wareham, V., Gilkinson, K., Koen-Alonso, M., Murillo, F.J. 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 + 125p.

Hamilton, J.M., and Wu, Y. 2013. Synopsis and trends in the physical environment of Baffin Bay and Davis Strait. Can. Tech. Rep. Hydrogr. Ocean Sci. 282: vi + 39 p.

Hanberry, B.B., and He, H.S. 2013. Prevalence, statistical thresholds, and accuracy assessment for species distribution models. Web Ecol. 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.

Head, E.J.H., Harris, L.R., and Yashayaev, I. 2003. Distributions of *Calanus* spp. and other mesozooplankton in the Labrador Sea in relation to hydrography in spring and early summer (1995-2000). Prog. Oceanogr. 59: 1-30.

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.

Hui, F.K., Warton, D.I., Foster, S.D., and Dunstan, P.K. 2013. To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. Ecology 94: 1913-1919.

Huntley, M., Strong, K.W., and Dengler, A.T. 1983. Dynamics and community structure of zooplankton in the Davis Strait and northern Labrador Sea. Arctic 36(2):143-161.

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

Juul-Pedersen, T., Gissel Nielsen, T., Michel, C., Friis Møller, E., Tiselius, P., Thor, P., Olesen, M., Selander, E., and Gooding, S. 2006. Sedimentation following the spring bloom in Disko Bay, West Greenland, with special emphasis on the role of copepods. Mar. Ecol. Prog. Ser. 314: 239-255.

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., Yashayaev, I., Tendal, O.S., and Jorgensbye, H. 2016a. Water mass characteristics and associated fauna of a recently discovered *Lophelia pertusa* (Scleractinian: Anthozoa) reef in Greenlandic waters. Polar Biol. (published online May 2016 doi: 10.1007/s00300-016-1957-3).

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. 2016b. 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.

Kuhn, M., and Johnson, K. 2013. Applied Predictive Modeling. New York: Springer Science + Business Media.

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.

Marra, G., and Wood, S.N. 2011. Practical Variable Selection for Generalized Additive Models. Comput. Stat. Data. An. 55: 2372-2387.

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.

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

Myers, P.G., and Ribergaard, M.H. 2013. Warming of the Polar Water Layer in Disko Bay and potential impact on Jakobshavn Isbrae. J. Phys. Oceanogr. 43: 2629–2640

Neves, B.M., Edinger, E., Hillaire-Marcel, C., Saucier, E.H., France, S.C., Treble, M., and Wareham, V. 2015. Deep-water bamboo coral forests in a muddy Arctic environment. Mar. Biodiv. 45: 867-871.

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

Ribergaard, M.H. 2012. Oceanographic Investigations off West Greenland 2011. NAFO Scientific Council Documents, 12/002

Shono, H. 2008. Application of the Tweedie Ddistribution of Zero-catch Data in CPUE Analysis. Fish. Res. 93: 154–162.

Stewart, P.L., Pocklington, P., and Cunjak, R.A. 1985. Distribution, abundance and diversity of benthic macroinvertebrates on the Canadian continental shelf and slope of southern Davis Strait and Ungava Bay. Arctic 38: 281-291.

Turner, J.T. 2002. Zooplankton fecal pellets, marine snow and sinking phytoplankton blooms. Aquat. Microb. Eco. 27: 57-102.

Wood, S.N. 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC Press, Boca Raton, FL.

Yashayaev, I. 2007. Hydrographic changes in the Labrador Sea, 1960–2005. Progr Oceanogr 73: 242–276.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., and Smith, G.M. 2009. Mixed-Effects Models and Extensions in Ecology with R. Springer, New York.

APPENDIX 1

Alternative Prediction Models- Generalized Additive Models for Predicting Coral and Sponge Biomass in the Eastern Arctic

Generalized additive models (GAMs) were performed on the biomass data from the research vessel surveys for each taxonomic group (Sponges, Sea Pens, Large and Small Gorgonian Corals). GAMs use regressions to make predictions and so represent a fundamentally different approach to the machine learning random forest (RF) models. We wished to cross compare biomass prediction surfaces using the two approaches and were particularly interested to see whether the GAMs were better able to model distribution in the extrapolated areas of the RF surfaces.

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:

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

where an exponential family distribution is specified for Y along with a link function g. The functions fj(xj) 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.

The mgcv package in R (Wood, 2006) was used to construct GAM models to predict the biomass of the taxa considered in order to compare with the RF models. We performed the GAMs using two methods for selecting the environmental variables. In our first approach we used the top variables from the equivalent RF biomass regression models, evaluating inclusion using the decrease in Gini value, where a natural break in the Mean Decrease in Sum of Squares was used to select the environmental variables for GAM modelling. This model was termed “GAM RF Variables”. In the second approach we used only variables that were correlated at less than 0.07. This second approach was termed “GAM 0.07 Variables”. This approach was also used for the Maritimes and Newfoundland and Labrador Regions (Beazley et al., 2016a; Guijarro et al., 2016) but differed slightly for the Gulf of St. Lawrence (Murillo et al., 2016).

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)). The full model followed 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 indicated a thin plate regression spline smoothing function. In addition, as for the Maritimes and Newfoundland and Labrador Regions, (see Beazley et al., 2016a; Guijarro et al., 2016), shrinkage smoothers (Zuur et al., 2009; Marra and Wood, 2011) were evaluated. A Tweedie model is an expansion of a compound Poisson model derived from a 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). The Tweedie factor was estimated inside the model.

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, randomized quantile residuals (Dunn and Smyth, 1996) were generated using the rqgam.check function of the dsm package in R (Miller et al., 2015). Randomized quantile residuals transform the residuals to be exactly normally distributed making the residuals vs. linear predictor plot much easier to interpret as it does not include the artifacts generated by the link function.

The goodness-of-fit statistic R^2 and the percentage of variance explained were used to evaluate the performance of the models as well as the prediction map derivate of the model in comparison to the real data. The Akaike information criterion (AIC) was used to evaluate the relative quality of the models for each set of data. An alternative method of model selection, the Bayesian information criterion (BIC) was also used.

Hudson Strait – Ungava Bay Region

Sponges (Porifera)

Cosmos Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sponge biomass using Cosmos trawl gear are presented in Table A1.1. The R^2 was fair to moderate for both models, and slightly higher for the GAM model using the RF-selected variables than for the model using the variables correlated at less than 0.7. The deviance explained was similar in both models. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.2 and A1.3, respectively.

Figure A1.1 shows the graphical diagnostics for both models. Both models showed fairly normal residuals and only small patterns in the residuals vs. linear predictor plots. The response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

Figure A1.2 shows the biomass surface of sponges generated from the GAM 0.7 Variables model. The areas of highest biomass of sponges was predicted to occur southeast of Atpatok Island and in the West of the Baffin Island. It is consistent with the results of the RF model (see Figures 16 and 17), and it predicted high biomass in the areas of model extrapolation as well. These areas of higher biomass are consistent with the distribution of sponge catches from the RV surveys (Figure A1.2, right panel). Biomass surface of sponges generated from the GAM RF Variables model showed erroneously high predicted biomass values that were not alleviated with the inclusion of latitude and longitude and therefore this predicted surface is not presented.

Table A1.1. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sponges using Cosmos trawl gear in the Hudson Strait-Ungava Bay Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.176	0.143
Deviance explained	23.8%	24.5%
AIC	365.583	361.305
BIC	432.313	425.521

Table A1.2. Results of the GAM RF Variables model built to predict the biomass of sponges using Cosmos trawl gear in the Hudson Strait-Ungava Bay Region.The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Surface Current Mean	1.175	3.769	0.035*
Bottom Temperature Average Range	2.951	4.384	2.890×10^{-3}*
Bottom Salinity Average Maximum	1.348	0.393	0.640
Annual Chlorophyll *a* Maximum	9.887×10^{-5}	0.000	1.000
Bottom Salinity Mean	1.943	0.996	0.377
Annual Chlorophyll *a* Range	2.887	2.933	0.029*
Surface Current Average Range	7.923×10^{-5}	0.075	0.997
Spring Chlorophyll *a* Mean	2.059	1.763	0.159
Bottom Salinity Average Minimum	1.291×10^{-4}	0.007	0.999
Spring Chlorophyll *a* Maximum	1.228×10^{-1}	0.299	0.800
Spring Chlorophyll *a* Range	2.946×10^{-4}	0.236	0.991

Table A1.3. Results of the GAM 0.7 Variables model built to predict the biomass of sponges Cosmos trawl gear in the Hudson Strait-Ungava Bay Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Current Average Maximum	9.104×10^{-1}	2.648	0.095
Bottom Salinity Average Range	7.809×10^{-1}	1.668	0.193
Bottom Temperature Average Maximum	1.033×10^{-5}	0.000	1.000
Bottom Temperature Average Range	1.288	4.069	0.026*
Spring Chlorophyll *a* Maximum	2.921	3.299	0.017*
Spring Chlorophyll *a* Minimum	2.035×10^{-5}	0.001	1.000
Summer Chlorophyll *a* Maximum	1.57×10^{-1}	0.412	0.729
Summer Chlorophyll *a* Minimum	8.869×10^{-2}	0.431	0.793
Depth	1.505	8.317	3.890×10^{-4}*
Maximum Average Fall Mixed Layer Depth	1.014×10^{-5}	0.316	0.998
Maximum Average Spring Mixed Layer Depth	1.305×10^{-5}	0.000	1.000
Annual Primary Production Average Range	3.115×10^{-5}	0.081	0.998
Summer Primary Production Average Maximum	2.545×10^{-1}	0.750	0.566
Summer Primary Production Average Minimum	4.532×10^{-5}	0.392	0.995
Surface Current Average Maximum	3.043	4.278	3.012×10^{-3}*
Surface Current Average Minimum	1.044×10^{-5}	0.001	1.000
Surface Temperature Average Maximum	1.829×10^{-5}	0.003	1.000
Surface Temperature Average Minimum	2.723×10^{-5}	0.423	0.996
Slope	5.276×10^{-1}	0.998	0.359

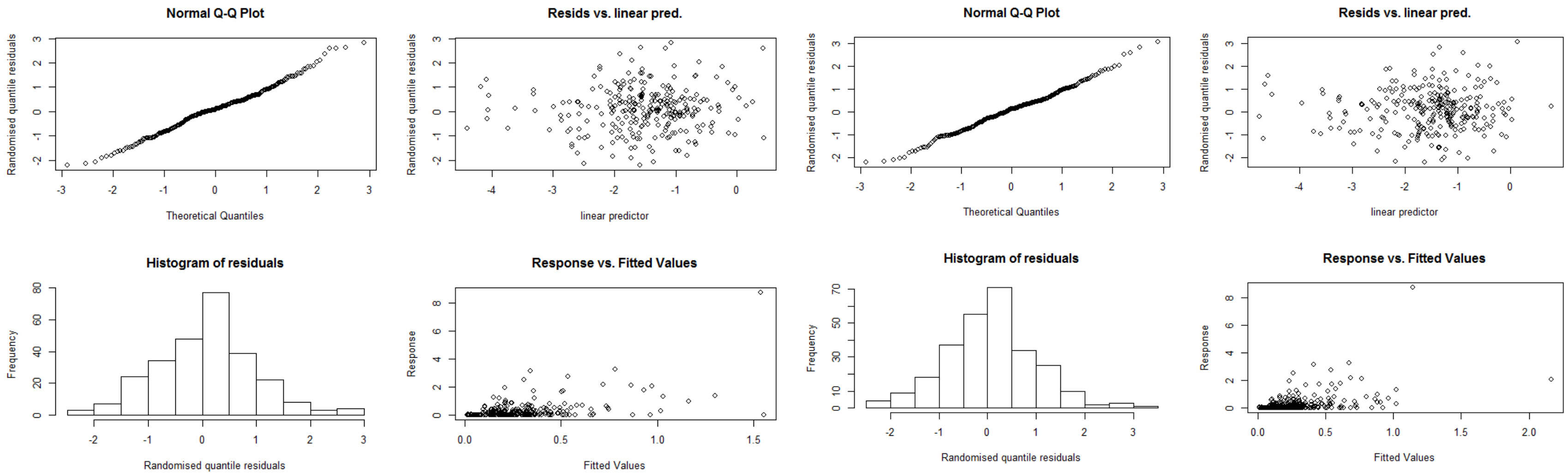

Figure A1.1. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of sponges biomass using Cosmos trawl gear in the Hudson Strait-Ungava Bay Region.

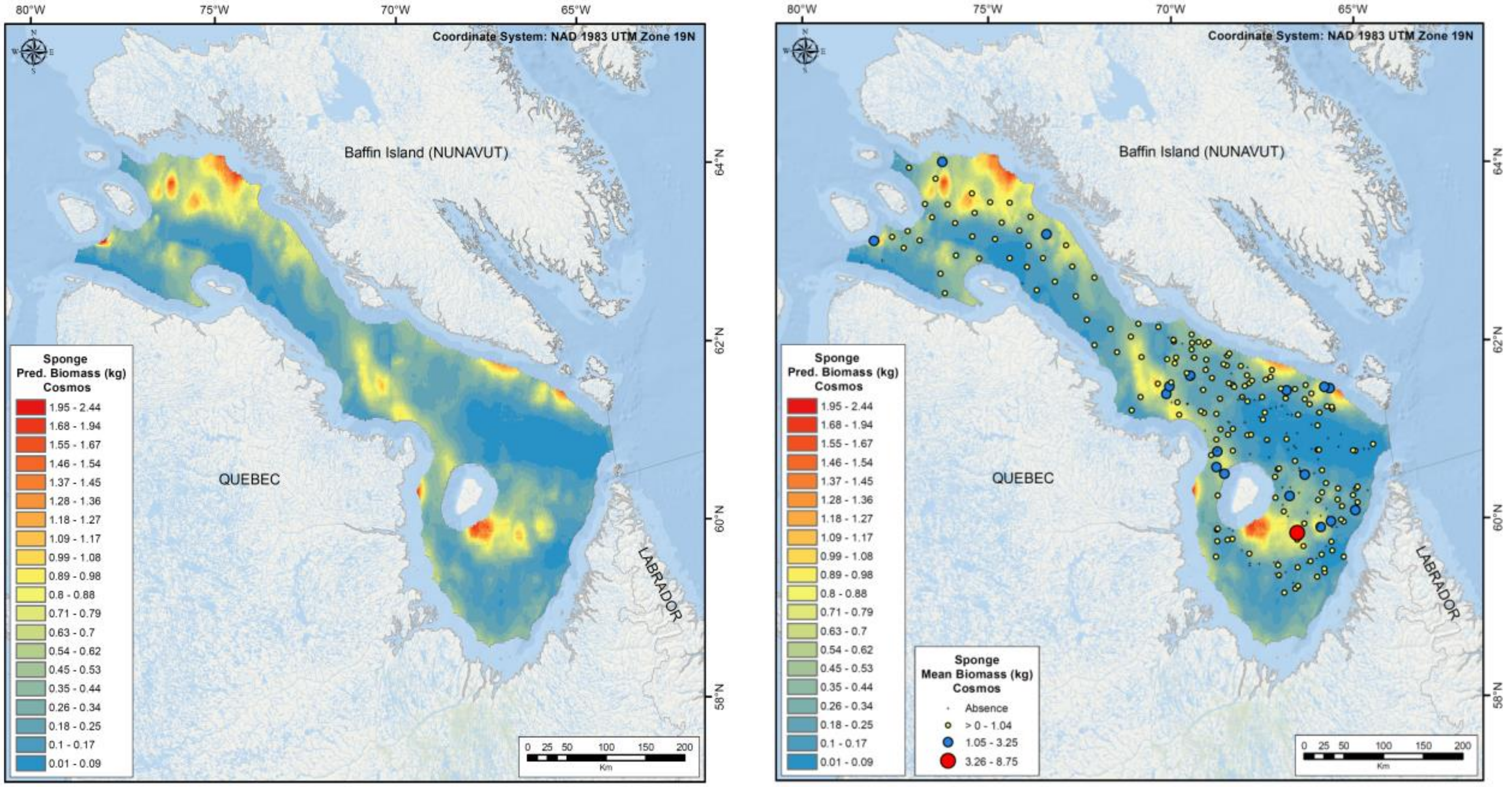

Figure A1.2. Prediction of sponge biomass (kg) using Cosmos trawl gear from the GAM 0.7 Variables model in the Hudson Strait-Ungava Bay Region. Right map shows the sponges mean biomass observations overlain.

Eastern Arctic Region

Sponges (Porifera)

Alfredo Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sponge biomass using Alfredo trawl gear are presented in Table A1.4. The R^2 was fair for both models, and slightly higher for the GAM model using the variables correlated at less than 0.7 than for the model using the RF-selected variables. The deviance explained was higher for the GAM 0.7 Variables model as well. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.5 and A1.6, respectively.

Figure A1.3 shows the graphical diagnostics for both models. Residuals from model using the variables correlated at less than 0.7 were closer to normality than residuals from the model using the RF-selected variables. The response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

When predicted to the entire extent of the study area, the models showed erroneously high predicted biomass values. High predictions of biomass were not alleviated with the inclusion of latitude and longitude or with modifications to the k value for each predictor. Predicted surfaces are therefore not presented for this taxonomic group.

Table A1.4. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sponges using Alfredo trawl gear in the Eastern Arctic Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.219	0.338
Deviance explained	58.6%	63.7%
AIC	3661.189	3600.759
BIC	3876.421	3944.498

Table A1.5. Results of the GAM RF Variables model built to predict the biomass of sponges using Alfredo trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Temperature Average Maximum	6.344	3.803	3.550 x 10^{-4}*
Bottom Temperature Average Minimum	5.910	3.233	2.251 x 10^{-3}*
Bottom Temperature Mean	6.972	10.064	6.440 x 10^{-13}*
Maximum Average Spring Mixed Layer Depth	3.293 x 10^{-3}	0.248	0.968
Bottom Temperature Average Range	8.224	11.157	7.750 x 10^{-16}*
Surface Temperature Average Minimum	3.521	7.534	6.710 x 10^{-6}*
Bottom Salinity Average Range	1.488	2.088	0.120
Slope	5.333	4.390	2.130 x 10^{-4}*

Table A1.6. Results of the GAM 0.7 Variables model built to predict the biomass of sponges using Alfredo trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Current Average Maximum	7.423	9.353	1.500 x 10^{-12}*
Bottom Temperature Average Maximum	6.258	14.144	< 2 x 10^{-16}*
Annual Chlorophyll *a* Maximum	0.914	0.857	0.366
Annual Chlorophyll *a* Minimum	0.763	1.504	0.216
Spring Chlorophyll *a* Maximum	2.713 x 10-3	0.244	0.973
Spring Chlorophyll *a* Minimum	6.278	8.592	2.660 x 10^{-10}*
Summer Chlorophyll *a* Maximum	3.710	3.802	3.838 x 10^{-3}*
Summer Chlorophyll *a* Minimum	1.409	7.252	1.378 x 10^{-3}*
Depth	1.568	2.140	0.116
Maximum Average Spring Mixed Layer Depth	6.724	4.893	1.030 x 10^{-5}*
Maximum Average Summer Mixed Layer Depth	4.2168	18.558	< 2 x 10^{-16}*
Summer Primary Production Average Maximum	6.369	6.231	2.820 x 10^{-7}*
Summer Primary Production Average Minimum	6.147	8.414	4.180 x 10^{-10}*
Summer Primary Production Average Range	3.000	1.886	0,108
Surface Current Average Maximum	6.633 x 10-4	0.002	0.999
Surface Temperature Average Minimum	3.861	14.501	9.870 x 10^{-12}*
Slope	1.431	8.666	3.960 x 10^{-4}*

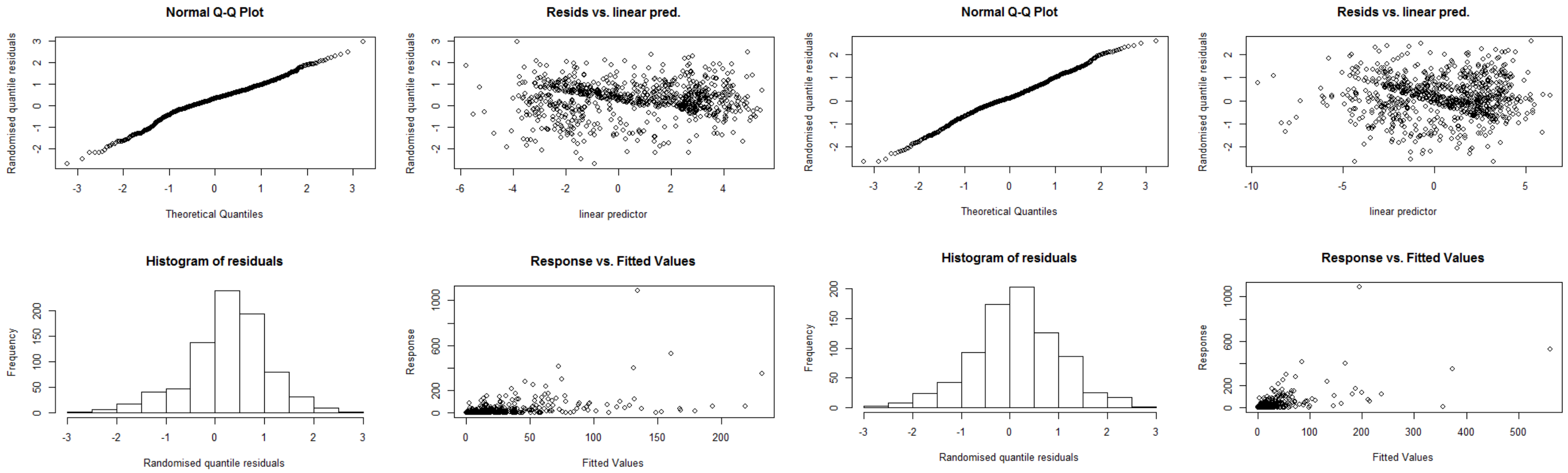

Figure A1.3. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of sponge biomass using Alfredo trawl gear in the Eastern Arctic Region.

Campelen Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sponge biomass using Campelen trawl gear are presented in Table A1.7. The R^2 was good for both models, and slightly higher for the GAM model using the variables correlated at less than 0.7 than for the model using the RF-selected variables. The deviance explained was slightly higher for the GAM RF Variables model. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.8 and A1.9, respectively.

Figure A1.4 shows the graphical diagnostics for both models. Both models showed fairly normal residuals and only small patterns in the residuals vs. linear predictor plots. However, the response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

When predicted to the entire extent of the study area, the models showed erroneously high predicted biomass values. High predictions of biomass were not alleviated with the inclusion of latitude and longitude or with modifications to the k value for each predictor. Predicted surfaces are therefore not presented for this taxonomic group.

Table A1.7. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sponges using Campelen trawl gear in the Eastern Arctic Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.459	0.606
Deviance explained	68.9%	68.5%
AIC	5534.356	5554.637
BIC	6876.936	5892.677

Table A1.8. Results of the GAM RF Variables model built to predict the biomass of sponges using Campelen trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Surface Salinity Average Minimum	3.072	1.917	0.106
Bottom Temperature Average Minimum	5.087	4.630	8.470×10^{-5}*
Bottom Salinity Average Minimum	4.115	3.019	0.010*
Surface Temperature Average Minimum	2.404	3.160	0.024*
Surface Temperature Average Range	3.024	1.336	0.254
Bottom Salinity Mean	3.621	1.318	0.255
Surface Salinity Mean	5.349	2.986	5.340×10^{-3}*
Bottom Salinity Average Maximum	2.265×10^{-3}	0.154	0.981
Surface Temperature Mean	4.371	1.559	0.161
Surface Salinity Average Maximum	7.339	7.129	2.270×10^{-9}*
Surface Temperature Average Maximum	5.894	4.566	5.540×10^{-5}*
Bottom Temperature Mean	6.776	5.763	4.970×10^{-7}*
Annual Chlorophyll *a* Mean	1.043	0.645	0.468

Table A1.9. Results of the GAM 0.7 Variables model built to predict the biomass of sponges using Campelen trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Current Average Maximum	4.762	10.130	2.420×10^{-10}*
Bottom Temperature Average Maximum	0.800	1.332	0.245
Annual Chlorophyll *a* Maximum	3.214×10^{-3}	0.255	0.970
Annual Chlorophyll *a* Minimum	2.687	4.267	3.958×10^{-3}*
Spring Chlorophyll *a* Maximum	0.846	1.695	0.186
Spring Chlorophyll *a* Minimum	0.794	1.196	0.273
Summer Chlorophyll *a* Maximum	1.696×10^{-3}	0.144	0.983
Summer Chlorophyll *a* Minimum	2.854	10.860	1.470×10^{-7}*
Depth	5.185	28.374	$< 2 \times 10^{-16}$*
Maximum Average Spring Mixed Layer Depth	5.711	10.179	1.150×10^{-11}*
Maximum Average Summer Mixed Layer Depth	4.350	3.827	1.582×10^{-3}*
Summer Primary Production Average Maximum	4.741	4.142	5.460×10^{-4}*
Summer Primary Production Average Minimum	6.459	4.525	3.640×10^{-5}*
Summer Primary Production Average Range	1.532	0.615	0.562
Surface Current Average Maximum	0.375	0.875	0.472
Surface Temperature Average Minimum	2.553	3.848	8.850×10^{-3}*
Slope	6.980	7.419	1.940×10^{-9}*

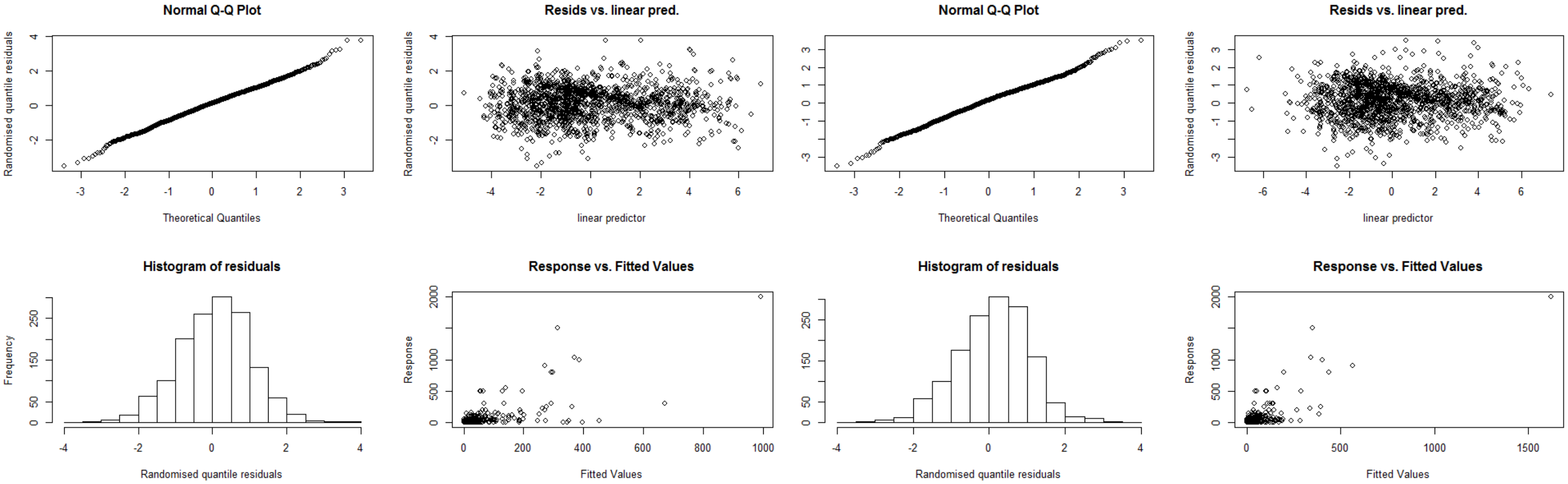

Figure A1.4. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of sponge biomass using Campelen trawl gear in the Eastern Arctic Region.

Cosmos Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sponge biomass using Cosmos trawl gear are presented in Table A1.10. The R^2 was fair for the GAM RF Variables model and negative for the GAM 0.7 Variables models. The deviance explained was slightly higher for the GAM RF Variables model. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.11 and A1.12, respectively.

Figure A1.5 shows the graphical diagnostics for both models. Residuals from the model using the variables correlated at less than 0.7 were closer to normality than residuals from the model using the RF-selected variables. The response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

When predicted to the entire spatial extent of the study area, the models showed erroneously high predicted biomass values. High predictions of biomass were not alleviated with the inclusion of latitude and longitude or with modifications to the k-value for each predictor. Predicted surfaces are therefore not presented for this taxonomic group/gear combination.

Table A1.10. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sponges using Cosmos trawl gear in the Eastern Arctic Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.156	-0.026
Deviance explained	61.6%	59.3%
AIC	459.808	486.679
BIC	529.906	578.911

Table A1.11. Results of the GAM RF Variables model built to predict the biomass of sponges using Cosmos trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Maximum Average Summer Mixed Layer Depth	9.560×10^{-1}	2.327	0.117
Bottom Salinity Average Minimum	4.606	10.772	2.860×10^{-9}*
Spring Chlorophyll *a* Minimum	4.921×10^{-6}	0.007	9.998×10^{-1}
Annual Primary Production Average Range	1.026	2.617	0.093
Surface Current Average Range	7.190	8.016	2.200×10^{-9}*
Maximum Average Fall Mixed Layer Depth	1.342	5.724	6.560×10^{-3}*
Annual Primary Production Average Maximum	3.984×10^{-6}	0.127	9.993×10^{-1}
Spring Chlorophyll *a* Range	3.608×10^{-6}	0.025	9.997×10^{-1}
Bottom Salinity Mean	1.676×10^{-4}	0.103	9.954×10^{-1}
Bottom Salinity Average Maximum	3.287×10^{-4}	0.172	9.918×10^{-1}
Summer Primary Production Mean	6.709×10^{-6}	0.061	9.994×10^{-1}
Annual Chlorophyll *a* Minimum	5.015×10^{-6}	0.006	9.998×10^{-1}
Depth	1.382×10^{-4}	0.072	9.965×10^{-1}
Annual Primary Production Mean	5.664×10^{-6}	0.013	9.997×10^{-1}
Summer Primary Production Average Maximum	8.033×10^{-1}	1.358	0.243
Surface Salinity Average Maximum	5.489×10^{-6}	0.001	9.999×10^{-1}
Surface Temperature Average Minimum	4.478×10^{-6}	0.139	9.992×10^{-1}

Table A1.12. Results of the GAM 0.7 Variables model built to predict the biomass of sponges using Cosmos trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Current Average Maximum	1.416	5.931	4.873×10^{-3}*
Bottom Temperature Average Maximum	4.125	6.077	3.190×10^{-5}*
Annual Chlorophyll *a* Maximum	2.567×10^{-5}	0.251	0.997
Annual Chlorophyll *a* Minimum	1.612×10^{-5}	0.021	0.999
Spring Chlorophyll *a* Maximum	2.384×10^{-4}	0.248	0.991
Spring Chlorophyll *a* Minimum	3.530×10^{-5}	0.146	0.997
Summer Chlorophyll *a* Maximum	2.068×10^{-4}	0.140	0.994
Summer Chlorophyll *a* Minimum	1.336	2.240	0.113
Depth	1.437	1.952	0.145
Maximum Average Spring Mixed Layer Depth	3.387	8.187	3.380×10^{-6}*
Maximum Average Summer Mixed Layer Depth	1.848	8.100	2.240×10^{-4}*
Summer Primary Production Average Maximum	3.838×10^{-5}	0.233	0.997
Summer Primary Production Average Minimum	1.377	3.297	0.044*
Summer Primary Production Average Range	8.523×10^{-1}	2.148	0.137
Surface Current Average Maximum	5.041×10^{-1}	1.009	0.380
Surface Temperature Average Minimum	2.666	4.468	4.100×10^{-3}*
Slope	1.098	0.932	0.354

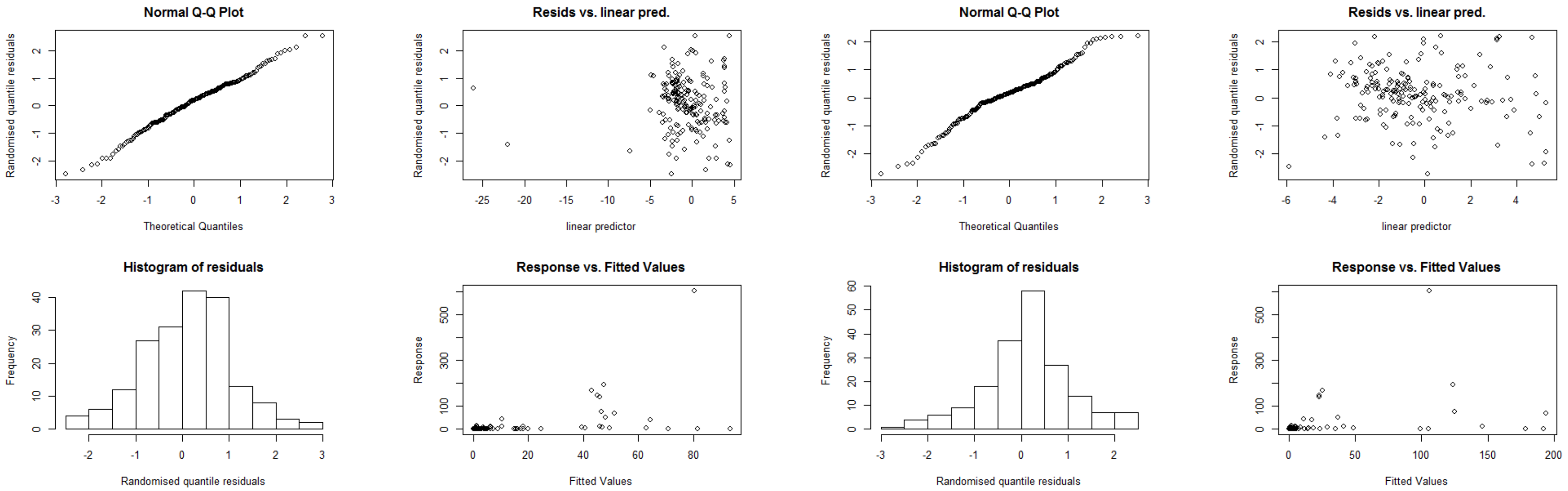

Figure A1.5. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of sponge biomass using Cosmos trawl gear in the Eastern Arctic Region.

Sea Pens (Pennatulacea)

Alfredo Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sea pen biomass using Alfredo trawl gear are presented in Table A1.13. The R^2 was fair for both models, and higher for the GAM model using the variables correlated at less than 0.7 than for the model using the RF-selected variables. The deviance explained was higher for the GAM 0.7 Variables model as well. The AIC/BIC was also lower for the GAM 0.7 Variables model. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.14 and A1.15, respectively.

Figure A1.6 shows the graphical diagnostics for both models. Residuals from the model using the variables correlated at less than 0.7 were closer to normality than residuals from the model using the RF-selected variables. The response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

The GAM RF Variables model predicted erroneously-high biomass values when applied to the Easter Arctic study spatial extent, therefore the predicted surface is not presented here. Although performance measures (R^2 and percent deviance explained) were slightly improved, erroneous biomass values were still predicted to occur by a model including latitude and longitude, therefore these results were not considered further.

Figure A1.7 shows the predicted biomass surface of sea pens from the GAM 0.7 Variables model. Higher sea pen biomass was predicted to occur in the southeast of Baffin Island although there are nt data observations in this area to validate this prediction. The Random Forest model predicted medium biomass in this area. Another area predicted to have high biomass was off the eastern coast of Devon Island, in the north (Figure A1.7). This area was supported with average biomass data observations. However, the prediction biomass is not consistent with the high sea pens catches recorded in the area which was predicted to have medium biomass (Figure A1.7, right panel). Similar results were found with the RF model (see Figures 60 and 61), although this predicted higher biomass values in the area of model extrapolation.

Table A1.13. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sea pens using Alfredo trawl gear in the Eastern Arctic Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.135	0.236
Deviance explained	29.2%	44.4%
AIC	248.137	145.697
BIC	377.9094	340.461

Table A1.14. Results of the GAM RF Variables model built to predict the biomass of sea pens using Alfredo trawl gear in the Eastern Arctic Region.The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Summer Chlorophyll *a* Mean	1.000	6.821	0.009*
Summer Chlorophyll *a* Range	1.000	1.666	0.197
Maximum Average Spring Mixed Layer Depth	1.000	8.211	0.004*
Bottom Current Average Minimum	3.019	9.034	1.84×10^{-6}*
Summer Chlorophyll *a* Maximum	1.000	2.022	0.155
Spring Chlorophyll *a* Minimum	3.113	2.681	0.029*
Bottom Shear Average Maximum	3.081	4.579	0.002*
Bottom Shear Average Minimum	1.000	16.076	6.69×10^{-5}*
Bottom Shear Average Range	3.131	4.867	0.002*
Bottom Current Average Range	1.001	17.788	2.77×10^{-5}*
Bottom Temperature Average Minimum	3.179	13.475	6.76×10^{-10}*

Table A1.15. Results of the GAM 0.7 Variables model built to predict the biomass of sea pens using Alfredo trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Maximum	2.859	5.488	5.85×10^{-13}*
Bottom Temperature Average Range	5.211	11.071	$< 2 \times 10^{-16}$*
Annual Chlorophyll *a* Maximum	1.532	0.937	0.002*
Annual Chlorophyll *a* Minimum	2.473×10^{-5}	0.000	0.543
Spring Chlorophyll *a* Maximum	3.666	2.026	9.71×10^{-5}*
Spring Chlorophyll *a* Minimum	1.601	1.305	< 0.001*
Summer Chlorophyll *a* Maximum	1.629	1.929	7.05×10^{-6}*
Summer Chlorophyll *a* Minimum	1.247×10^{-5}	0.000	1.000
Depth	4.363	4.470	1.19×10^{-9}*
Maximum Average Spring Mixed Layer Depth	7.321×10^{-1}	0.234	0.061
Maximum Average Summer Mixed Layer Depth	7.427×10^{-6}	0.000	0.629
Summer Primary Production Average Maximum	5.499×10^{-5}	0.000	0.188
Summer Primary Production Average Minimum	1.487	0.577	0.020*
Summer Primary Production Average Range	3.471	1.522	0.001*
Surface Current Average Maximum	3.780×10^{-1}	0.054	0.236
Surface Salinity Average Range	1.585	1.019	0.001*
Surface Temperature Average Minimum	3.402	3.083	4.23×10^{-7}*
Slope	7.347×10^{-6}	0.000	0.718

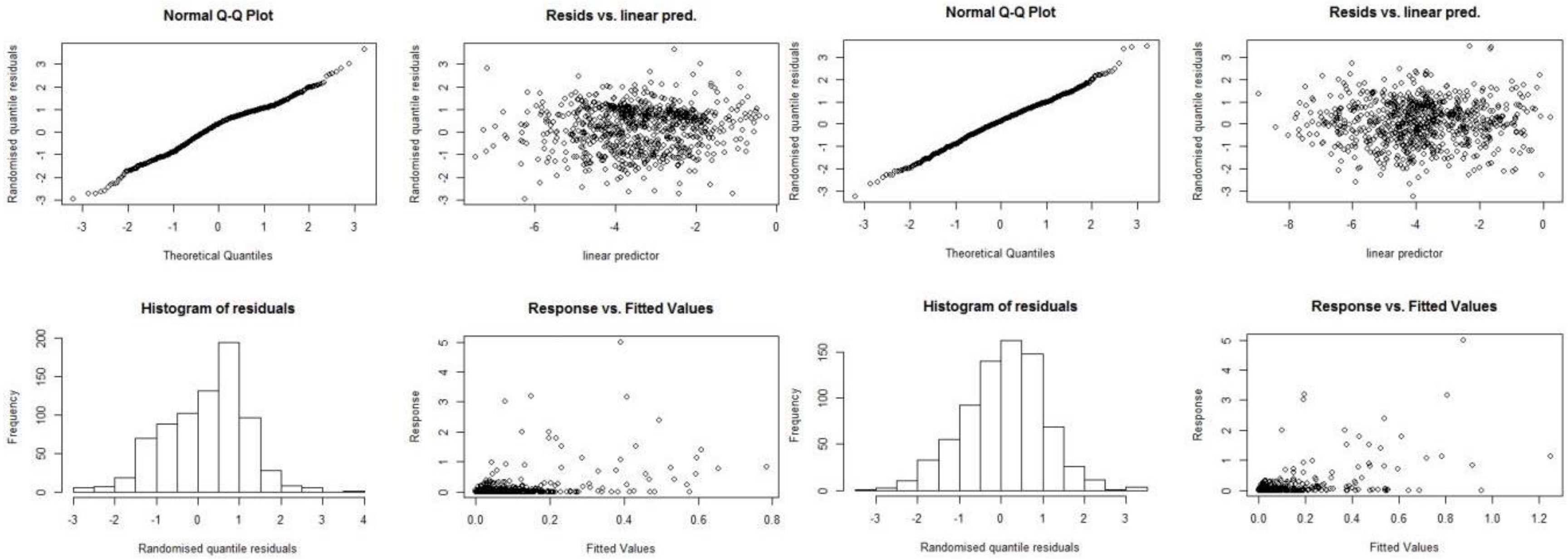

Figure A1.6. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of sea pen biomass using Alfredo trawl gear in the Eastern Arctic Region.

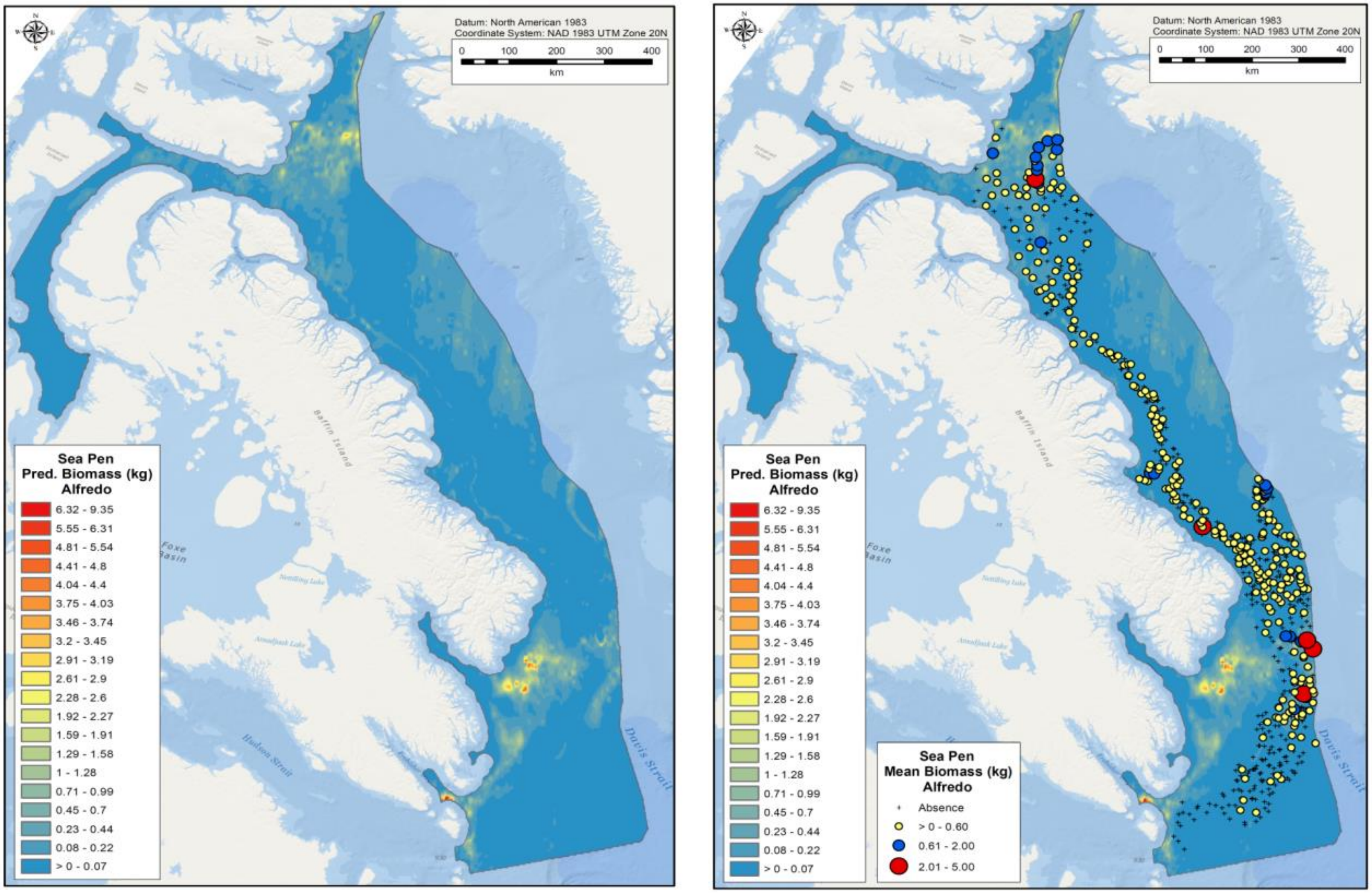

Figure A1.7. Prediction of sea pen biomass (kg) using Alfredo trawl gear from the GAM 0.7 Variables model in the Eastern Arctic Region. Right map shows the sea pen mean biomass observations overlain.

Campelen Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sea pen biomass using Campelen trawl gear are presented in Table A1.16. The R^2 was fair for both models, and slightly higher for the GAM model using the variables correlated at less than 0.7 than for the model using the RF-selected variables. The deviance explained was similar in both models. The AIC/BIC was comparable between the two models. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.17 and A1.18, respectively.

Figure A1.8 shows the graphical diagnostics for both models. Both models showed fairly normal residuals and only small patterns in the residuals vs. linear predictor plots. However, the response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

When predicted to the entire extent of the study area, the models showed erroneously very high predicted biomass values in the deeper areas. High predictions of biomass were not alleviated with the inclusion of latitude and longitude in the model. The predicted surfaces therefore are not presented in this report.

Table A1.16. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sea pens using Campelen trawl gear in the Eastern Arctic Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.15	0.193
Deviance explained	55.9%	57.3%
AIC	1196.001	1195.611
BIC	1293.353	1302.531

Table A1.17. Results of the GAM RF Variables model built to predict the biomass of sea pens using Campelen trawl gear in the Eastern Arctic Region.The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Depth	3.015	7.891	$< 2 \times 10^{-16}$*
Spring Chlorophyll *a* Minimum	1.299	0.991	0.001*
Summer Primary Production Average Minimum	2.920	3.243	7.21×10^{-8}*
Bottom Salinity Average Range	3.319×10^{-1}	0.037	0.269
Annual Primary Production Average Minimum	5.314×10^{-6}	0.000	0.677
Annual Chlorophyll *a* Minimum	1.109	0.709	0.005*
Bottom Temperature Average Maximum	1.594×10^{-1}	0.017	0.279
Bottom Current Average Range	9.718×10^{-1}	0.489	0.016*
Maximum Average Winter Mixed Layer Depth	1.302×10^{-5}	0.000	0.774
Maximum Average Spring Mixed Layer Depth	9.942×10^{-6}	0.000	1.000
Bottom Temperature Average Range	2.210	0.793	0.007*
Annual Primary Production Mean	8.363×10^{-6}	0.000	0.877
Bottom Current Mean	2.214×10^{-5}	0.000	0.309
Bottom Shear Average Range	1.272×10^{-5}	0.000	0.307
Maximum Average Summer Mixed Layer Depth	4.773×10^{-6}	0.000	0.785
Surface Salinity Mean	1.879×10^{-4}	0.000	0.458
Bottom Shear Mean	7.418×10^{-6}	0.000	0.327
Bottom Shear Average Minimum	6.675×10^{-5}	0.000	0.427

Table A1.18. Results of the GAM 0.7 Variables model built to predict the biomass of sea pens using Campelen trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Current Average Maximum	8.442×10^{-1}	0.300	0.045*
Bottom Temperature Average Range	2.362	1.361	0.001*
Annual Chlorophyll *a* Maximum	9.837×10^{-6}	0.000	0.620
Annual Chlorophyll *a* Minimum	1.041	0.577	0.010*
Spring Chlorophyll *a* Maximum	6.869×10^{-6}	0.000	0.616
Spring Chlorophyll *a* Minimum	1.238	0.861	0.002*
Summer Chlorophyll *a* Maximum	1.895×10^{-4}	0.000	0.463
Summer Chlorophyll *a* Minimum	7.306×10^{-1}	0.166	0.122
Depth	2.610	11.008	$< 2 \times 10^{-16}$*
Maximum Average Spring Mixed Layer Depth	2.636×10^{-6}	0.000	1.000
Maximum Average Summer Mixed Layer Depth	1.708×10^{-6}	0.000	0.705
Summer Primary Production Average Maximum	1.756×10^{-6}	0.000	1.000
Summer Primary Production Average Minimum	2.155	2.618	6.29×10^{-7}*
Summer Primary Production Average Range	1.535×10^{-6}	0.000	1.000
Surface Current Average Maximum	9.734×10^{-5}	0.000	0.566
Surface Salinity Average Range	3.941×10^{-6}	0.000	1.000
Surface Temperature Average Minimum	3.646×10^{-5}	0.000	0.431
Slope	2.469	0.956	0.012*

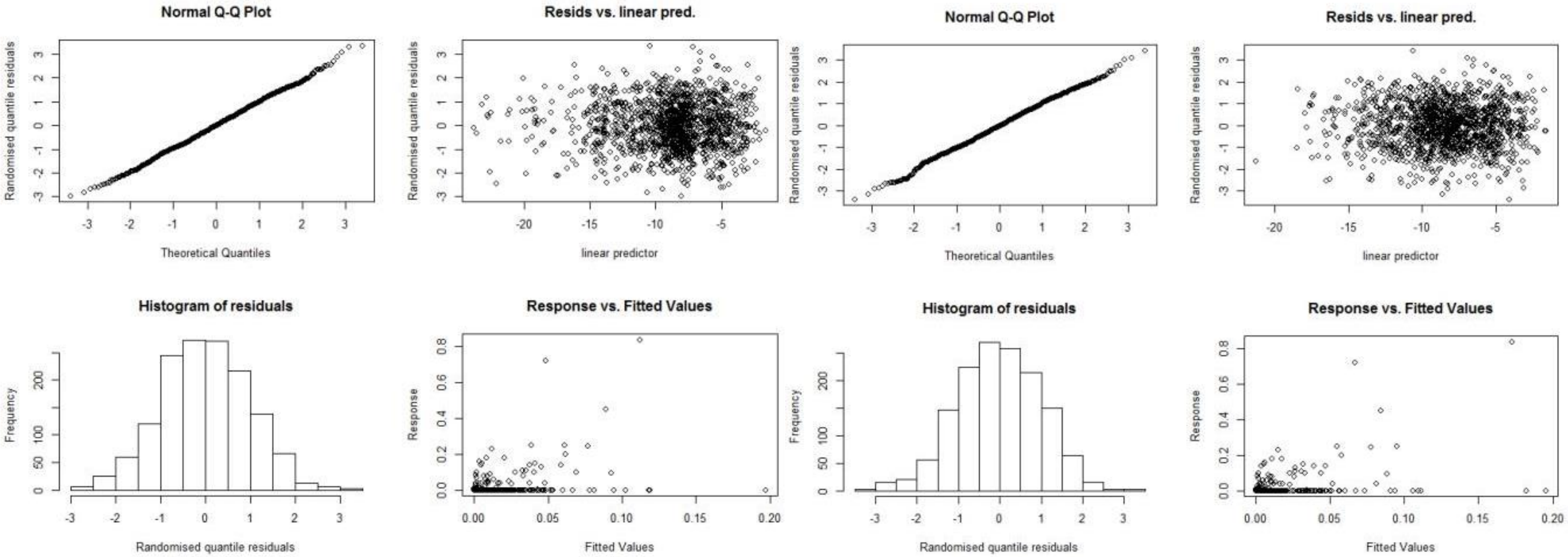

Figure A1.8. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of sea pen biomass using Campelen trawl gear in the Eastern Arctic Region.

Cosmos Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sea pen biomass using Cosmos trawl gear are presented in Table A1.19. The R^2 was fair for both models, and slightly higher for the GAM model using RF-selected variables than for the model using the variables correlated at less than 0.7. However, the deviance was higher in the GAM model using the variables correlated at less than 0.7. The AIC was higher in the GAM model using RF-selected variables than for the model using the variables correlated at less than 0.7 but the BIC lower. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.20 and A1.21, respectively.

Figure A1.9 shows the graphical diagnostics for both models. Both models showed fairly normal residuals and only small patterns in the residuals vs. linear predictor plots. However, the response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

When predicted to the entire extent of the study area, the models showed erroneously very high predicted biomass values in the deeper areas. High predictions of biomass were not alleviated with the inclusion of latitude and longitude in the model. The predicted surfaces therefore are not presented in this report.

Table A1.19. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sea pens using Cosmos trawl gear in the Eastern Arctic Region.

	GAM RF Variables	GAM 0.7 Variables
R^2	0.12	0.099
Deviance explained	16.2%	30.1%
AIC	64.647	55.525
BIC	97.107	105.286

Table A1.20. Results of the GAM RF Variables model built to predict the biomass of sea pens using Cosmos trawl gear in the Eastern Arctic Region.The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Annual Chlorophyll *a* Range	4.523×10^{-5}	0.000	0.2958
Bottom Salinity Average Range	1.308	0.725	0.005*
Annual Chlorophyll *a* Maximum	3.018×10^{-5}	0.000	0.343
Depth	5.474×10^{-1}	0.112	0.122
Summer Chlorophyll *a* Range	7.482×10^{-1}	0.171	0.128
Surface Salinity Average Maximum	1.425	0.434	0.038*
Surface Temperature Average Range	1.849×10^{-6}	0.000	0.207
Spring Chlorophyll *a* Minimum	3.079×10^{-5}	0.000	0.545
Surface Temperature Average Maximum	1.227	0.559	0.013*

Table A1.21. Results of the GAM 0.7 Variables model built to predict the biomass of sea pens using Cosmos trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Current Average Maximum	2.039	1.050	0.003*
Bottom Temperature Average Range	3.779×10^{-6}	0.000	0.783
Annual Chlorophyll *a* Maximum	6.070×10^{-3}	0.001	0.201
Annual Chlorophyll *a* Minimum	4.404×10^{-6}	0.000	0.907
Spring Chlorophyll *a* Maximum	1.410	0.683	0.013*
Spring Chlorophyll *a* Minimum	1.204	0.759	0.004*
Summer Chlorophyll *a* Maximum	1.914×10^{-5}	0.000	0.404
Summer Chlorophyll *a* Minimum	1.343	0.763	0.005*
Depth	8.606×10^{-1}	0.288	0.060
Maximum Average Spring Mixed Layer Depth	5.063×10^{-6}	0.000	0.751
Maximum Average Summer Mixed Layer Depth	1.111	0.448	0.019*
Summer Primary Production Average Maximum	2.485×10^{-6}	0.000	0.832
Summer Primary Production Average Minimum	6.635×10^{-1}	0.139	0.124
Summer Primary Production Average Range	3.516×10^{-6}	0.000	0.693
Surface Current Average Maximum	6.839×10^{-1}	0.171	0.114
Surface Salinity Average Range	1.782×10^{-6}	0.000	0.878
Surface Temperature Average Minimum	3.848×10^{-5}	0.000	0.469
Slope	1.567×10^{-5}	0.000	0.389

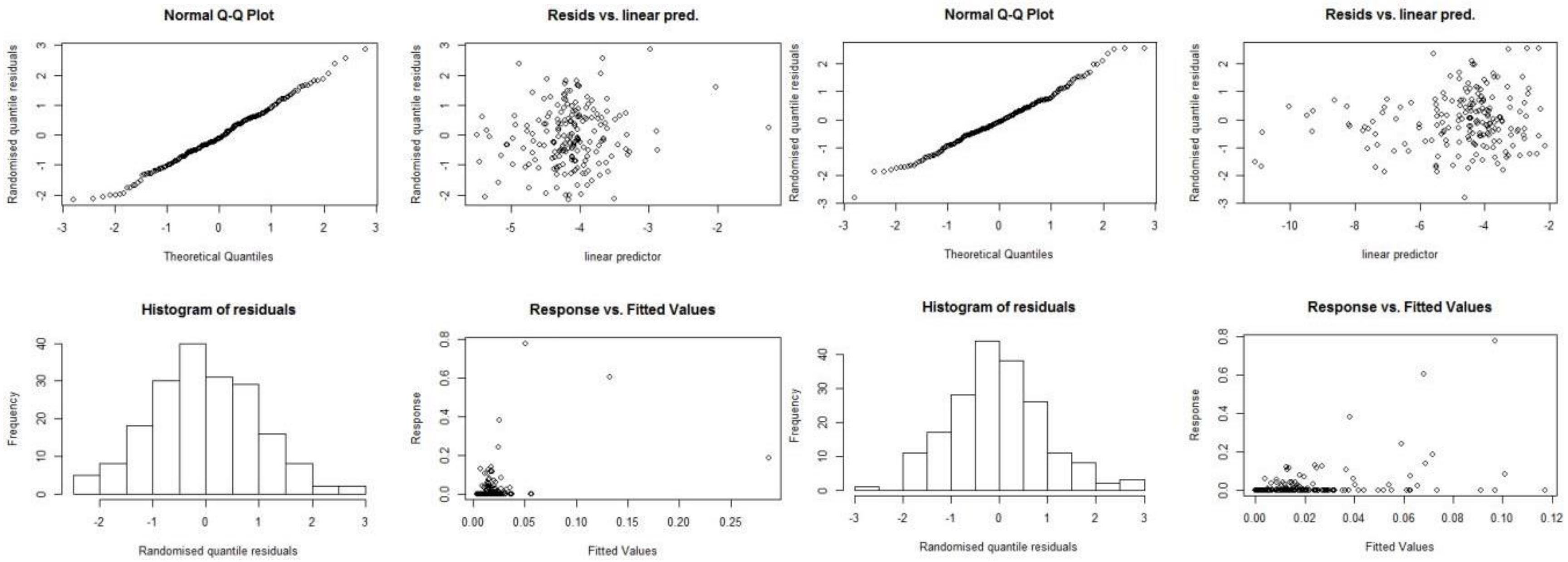

Figure A1.9. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of sea pen biomass using Cosmos trawl gear in the Eastern Arctic Region.

Large Gorgonian Corals

Alfredo Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean large gorgonian coral biomass using Alfredo trawl gear are presented in Table A1.22. The R^2 was very low for both models, and slightly higher for the GAM model using the RF-selected variables than the model using the variables correlated at less than 0.7. The deviance explained was similar for both models. The AIC/BIC was comparable between the two models. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.23 and A1.24, respectively.

Figure A1.10 shows the graphical diagnostics for both models. Both models showed fairly normal residuals. However, the response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

When predicted to the entire extent of the study area, the models showed erroneously very high predicted biomass values in the deeper areas. High predictions of biomass were not alleviated with the inclusion of latitude and longitude in the model. The predicted surfaces therefore are not presented in this report.

Table A1.22. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of large gorgonian corals using Alfredo trawl gear in the Eastern Arctic Region.

	GAM RF Variables	GAM 0.7 Variables
R^2	0.078	0.051
Deviance explained	80.6%	81.6
AIC	860.425	858.754
BIC	942.583	960.293

Table A1.23. Results of the GAM RF Variables model built to predict the biomass of large gorgonian corals using Alfredo trawl gear in the Eastern Arctic Region.The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Slope	3.724	8.074	4.0 x 10^{-16}*
Surface Current Average Minimum	8.964 x 10^{-5}	0.000	0.407
Summer Chlorophyll *a* Maximum	2.116	1.764	< 0.001*
Summer Chlorophyll *a* Range	1.169 x 10^{-3}	0.000	0.007*
Bottom Temperature Average Range	9.445 x 10^{-1}	0.332	0.046*
Summer Primary Production Average Maximum	1.590	0.465	0.049*
Surface Current Mean	8.170 x 10^{-1}	0.231	0.087
Maximum Average Winter Mixed Layer Depth	2.886 x 10^{-5}	0.000	1.000
Annual Primary Production Average Range	2.938	1.759	7.9 x 10^{-5}*

Table A1.24. Results of the GAM 0.7 Variables model built to predict the biomass of large gorgonian corals using Alfredo trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Current Average Maximum	8.933 x 10^{-6}	0.000	1.000
Bottom Temperature Average Range	3.627 x 10^{-5}	0.000	0.478
Annual Chlorophyll *a* Maximum	9.122 x 10^{-4}	0.000	0.554
Annual Chlorophyll *a* Minimum	1.409 x 10^{-5}	0.000	0.764
Spring Chlorophyll *a* Maximum	1.763 x 10^{-5}	0.000	1.000
Spring Chlorophyll *a* Minimum	1.529 x 10^{-5}	0.000	1.000
Summer Chlorophyll *a* Maximum	1.651 x 10^{-1}	0.050	0.084.
Summer Chlorophyll *a* Minimum	1.985 x 10^{-5}	0.000	0.752
Depth	4.892 x 10^{-5}	0.000	0.541
Maximum Average Spring Mixed Layer Depth	5.624	3.258	7.31 x 10^{-6}*
Maximum Average Summer Mixed Layer Depth	2.174	0.602	0.037*
Summer Primary Production Average Maximum	2.586	1.604	< 0.001*
Summer Primary Production Average Minimum	1.879 x 10^{-5}	0.000	1.000
Summer Primary Production Average Range	1.599 x 10^{-5}	0.000	1.000
Surface Current Average Maximum	7.900 x 10^{-6}	0.000	0.655
Surface Salinity Average Range	1.495 x 10^{-5}	0.000	0.632
Surface Temperature Average Minimum	1.643	0.589	0.017*
Slope	3.409	8.843	< 2 x 10^{-6}*

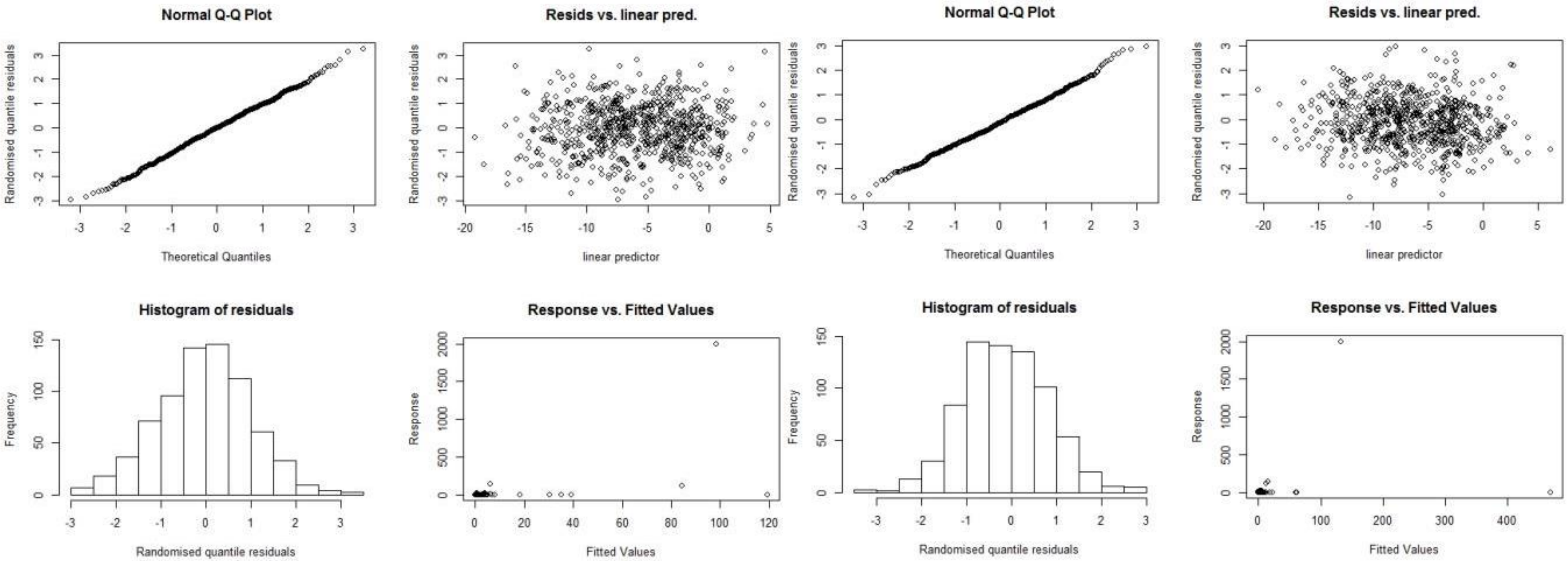

Figure A1.10. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of large gorgonian coral biomass using Alfredo trawl gear in the Eastern Arctic Region.

Campelen Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean large gorgonian coral biomass using Campelen trawl gear are presented in Table A1.25. The R^2 was fair moderate for both models, and slightly higher for the GAM model using the variables correlated at less than 0.7 than for the model using the RF-selected variables. The deviance explained was similar in both models. The AIC/BIC was lower in the GAM model using RF-selected variables than for the model using the variables correlated at less than 0.7. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.26 and A1.27, respectively.

Figure A1.11 shows the graphical diagnostics for both models. Both models showed fairly normal residuals. Variable model showed patterns in the residuals vs. linear predictor plot indicative of heteroskedasticity. The response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

Figures A1.12 and A1.13 show the biomass surface of large gorgonian corals generated from the GAM RF Variables and GAM 0.7 Variables models, respectively. For the GAM RF Variables model, the majority of the spatial extent was predicted to have low (< 13.31 kg) large gorgonian coral biomass. High biomass of large gorgonians was predicted to occur along the voluntary Closure Area in Hatton Basin in Davis Strait, although the highest biomass predicted was to the south, where no data were available (Figure A1.12, right panel) and differed with the results of the RF model (see Figures 87 and 88). However, this model did not predict high biomass in the areas of model extrapolation. The GAM 0.7 Variables model predicted similar results (Figure A1.13), however the area of high biomass was localized further south than in the GAM RF Variables Model and data were not available to validate this area of high biomass.

The majority of the spatial extent was predicted to have low (< 18.9 kg) large gorgonian coral biomass. The highest predicted biomass (up to 322 kg) occurred north of the voluntary Closure Area in Hatton Basin in Davis Strait and coincided with a cluster of large mean catches.

Table A1.25. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of large gorgonian corals using Campelen trawl gear in the Eastern Arctic Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.296	0.332
Deviance explained	78.7%	81.5%
AIC	1922.226	1926.832
BIC	2013.037	2095.893

Table A1.26. Results of the GAM RF Variables model built to predict the biomass of large gorgonian corals using Campelen trawl gear in the Eastern Arctic Region.The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Temperature Average Minimum	6.436	20.355	$< 2 \times 10^{-16}$*
Annual Primary Production Average Minimum	1.687	3.164	2.89×10^{-8}*
Bottom Temperature Average Range	3.671×10^{-4}	0.000	0.730
Surface Salinity Average Minimum	5.782×10^{-3}	0.000	0.399
Depth	3.548	2.099	9.61×10^{-5}*

Table A1.27. Results of the GAM 0.7 Variables model built to predict the biomass of large gorgonian corals Campelen trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Current Average Maximum	3.767×10^{-5}	0.000	0.582
Bottom Temperature Average Range	2.711	5.436	4.84×10^{-14}*
Annual Chlorophyll *a* Maximum	8.492×10^{-6}	0.000	1.000
Annual Chlorophyll *a* Minimum	5.967×10^{-1}	0.123	0.136
Spring Chlorophyll *a* Maximum	8.815×10^{-6}	0.000	0.840
Spring Chlorophyll *a* Minimum	7.865×10^{-1}	0.199	0.089
Summer Chlorophyll *a* Maximum	1.390×10^{-5}	0.000	1.000
Summer Chlorophyll *a* Minimum	1.128×10^{-5}	0.000	1.000
Depth	4.219	2.633	9.62×10^{-6}*
Maximum Average Spring Mixed Layer Depth	1.929×10^{-2}	0.001	0.460
Maximum Average Summer Mixed Layer Depth	1.299	1.188	< 0.001*
Summer Primary Production Average Maximum	3.018	1.330	0.002*
Summer Primary Production Average Minimum	4.913	2.028	0.001*
Summer Primary Production Average Range	8.047×10^{-6}	0.000	1.000
Surface Current Average Maximum	1.210×10^{-3}	0.000	0.495
Surface Salinity Average Range	1.807	1.887	1.48×10^{-5}*
Surface Temperature Average Minimum	2.307×10^{-1}	0.024	0.252
Slope	3.495	1.771	0.001*

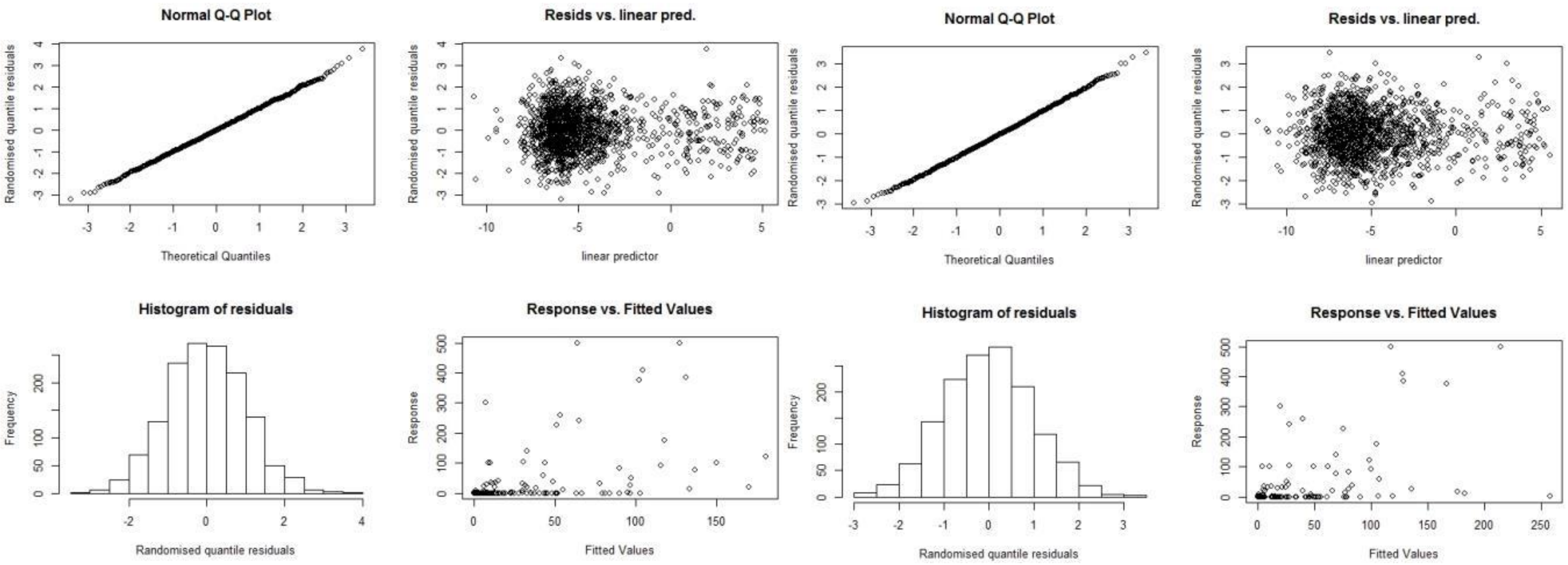

Figure A1.11. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of large gorgonian coral biomass using Campelen trawl gear in the Eastern Arctic Region.

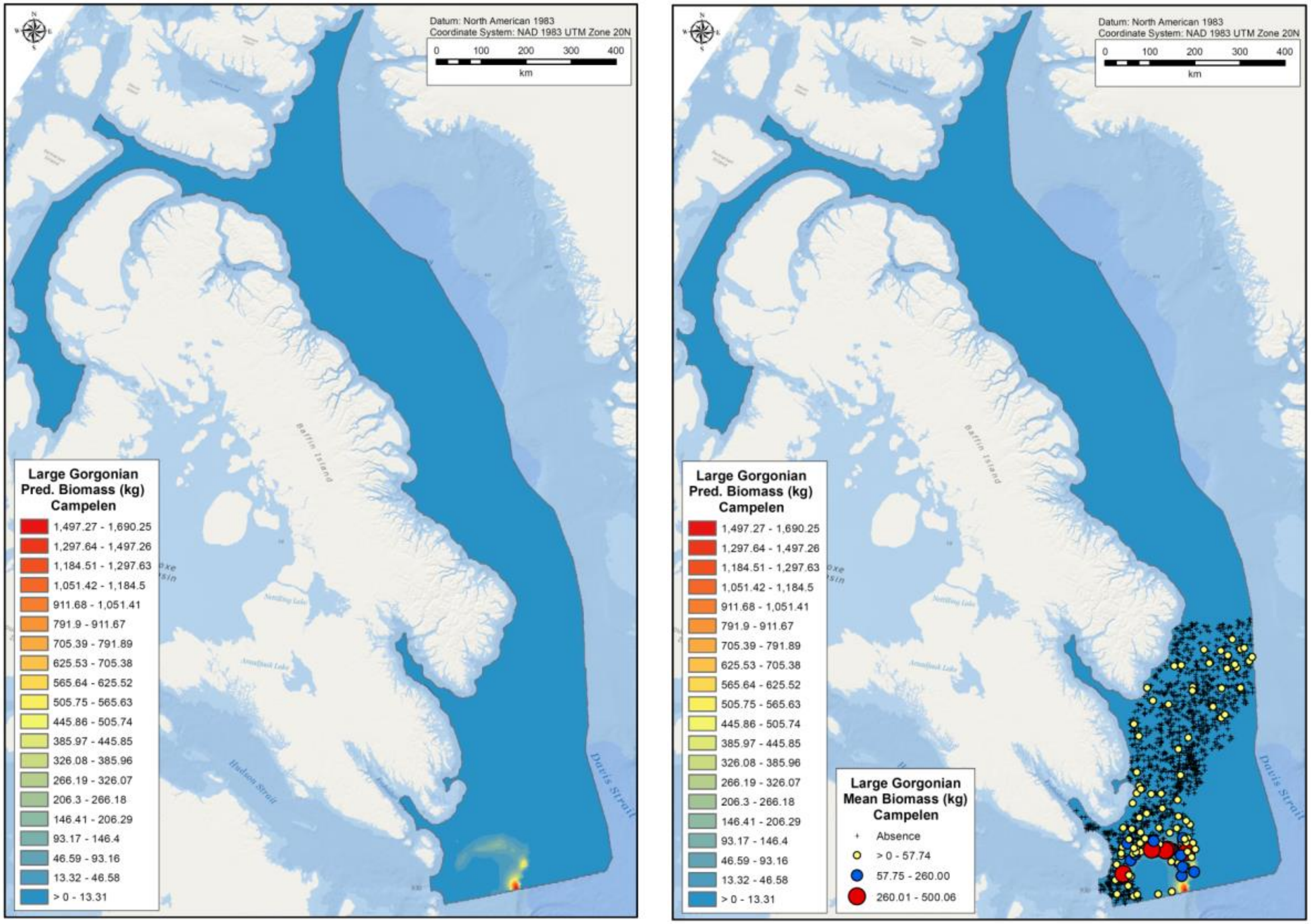

Figure A1.12. Prediction of large gorgonian coral biomass (kg) using Campelen trawl gear from the GAM RF Variables model in the Eastern Arctic Region. Right map shows the sea pen mean biomass observations overlain.

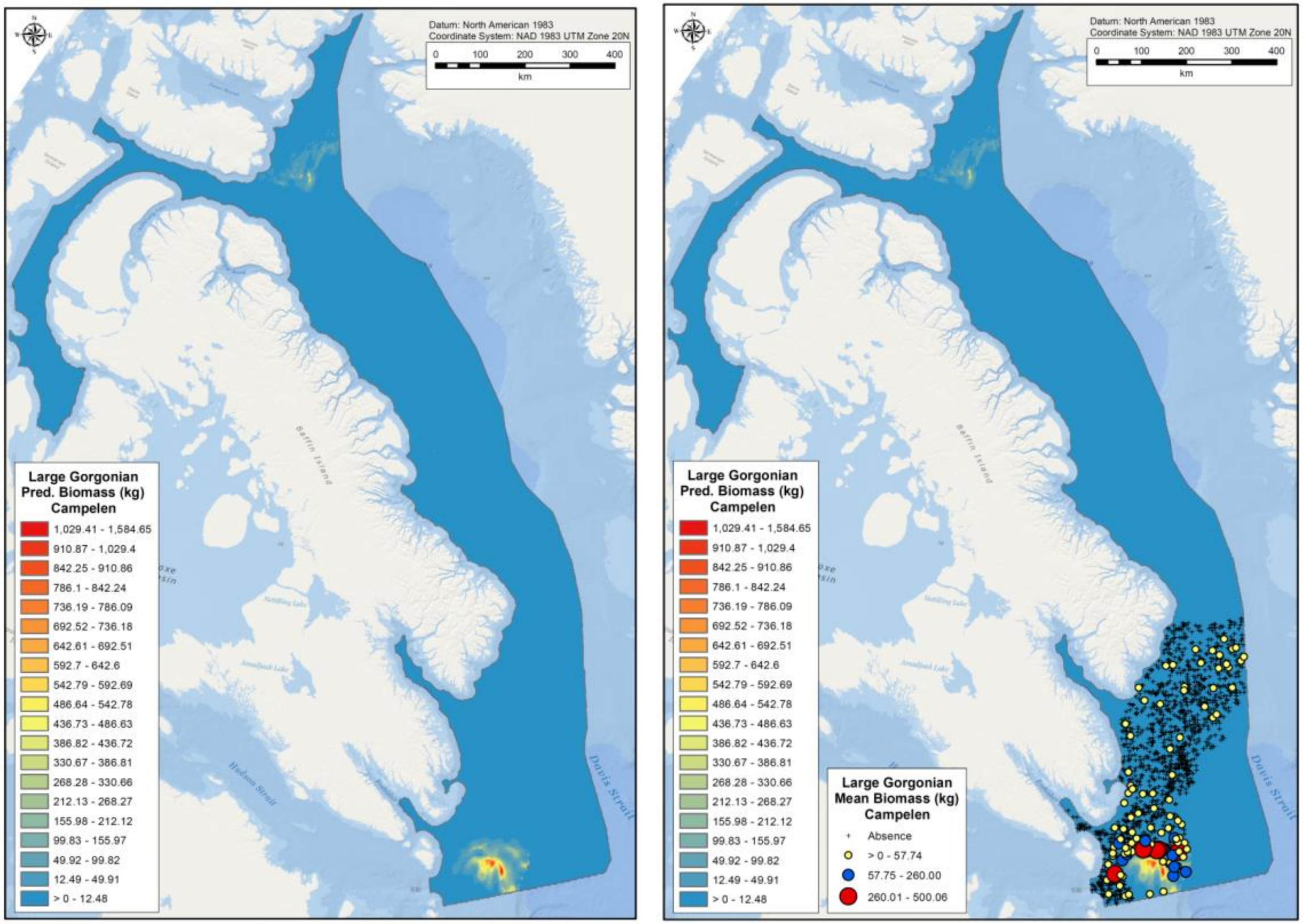

Figure A1.13. Prediction of large gorgonian coral biomass (kg) using Campelen trawl gear from the GAM 0.7 Variables model in the Eastern Arctic Region. Right map shows the sea pen mean biomass observations overlain.

Small Gorgonian Corals

Alfredo Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean small gorgonian coral biomass using Alfredo trawl gear are presented in Table A1.28. The R^2 was good for both models, and higher for the GAM model using the variables correlated at less than 0.7. The deviance explained was high for both models. The AIC was higher in the GAM model using RF-selected variables than for the model using the variables correlated at less than 0.7 but the BIC was lower. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.29 and A1.30, respectively.

Figure A1.14 shows the graphical diagnostics for both models. Both models showed fairly normal residuals and only small patterns in the residuals vs. linear predictor plots. However, the response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

When predicted to the entire extent of the study area, the models showed erroneously very high predicted biomass values in the deeper areas. High predictions of biomass were not alleviated with the inclusion of latitude and longitude in the model. The predicted surfaces therefore are not presented in this report.

Table A1.28. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of small gorgonian corals using Alfredo trawl gear in the Eastern Arctic Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.415	0.641
Deviance explained	62.3%	77.2%
AIC	144.075	134.230
BIC	215.168	332.851

Table A1.29. Results of the GAM RF Variables model built to predict the biomass of small gorgonian corals using Alfredo trawl gear in the Eastern Arctic Region.The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Shear Average Maximum	6.668×10^{-5}	0.065	0.998
Bottom Shear Average Range	1.005	1.467	0.220
Bottom Current Average Maximum	7.089×10^{-5}	0.332	0.995
Bottom Salinity Average Range	4.910×10^{-6}	0.164	0.999
Bottom Current Average Range	3.447	5.203	3.600×10^{-4}*
Bottom Temperature Average Range	1.357	3.464	0.039*
Summer Primary Production Average Minimum	9.018×10^{-1}	2.890	0.082
Bottom Temperature Average Minimum	3.381	3.948	3.370×10^{-3}*

Table A1.30. Results of the GAM 0.7 Variables model built to predict the biomass of small gorgonian corals Alfredo trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	**edf**	***F***	***p*-value**
Bottom Current Average Maximum	2.631	9.201	3.690×10^{-6}*
Bottom Temperature Average Range	5.115×10^{-6}	0.221	0.999
Annual Chlorophyll *a* Maximum	3.871×10^{-5}	0.332	0.996
Annual Chlorophyll *a* Minimum	4.227	2.406	0.035*
Spring Chlorophyll *a* Maximum	3.102	4.085	3.890×10^{-3}*
Spring Chlorophyll *a* Minimum	4.030×10^{-1}	0.718	0.490
Summer Chlorophyll *a* Maximum	1.849×10^{-4}	0.263	0.992
Summer Chlorophyll *a* Minimum	3.125×10^{-6}	0.031	1.000
Depth	1.106×10^{-5}	0.101	0.999
Maximum Average Spring Mixed Layer Depth	1.922	5.345	3.130×10^{-3}*
Maximum Average Summer Mixed Layer Depth	1.887	1.635	0.185
Summer Primary Production Average Maximum	8.408×10^{-6}	0.075	0.999
Summer Primary Production Average Minimum	2.060	8.012	9.860×10^{-5}*
Summer Primary Production Average Range	4.974	3.266	4.320×10^{-3}*
Surface Current Average Maximum	2.140	1.446	0.226
Surface Salinity Average Range	5.166	3.687	1.240×10^{-3}*
Surface Temperature Average Minimum	1.931	2.254	0.092
Slope	2.287	2.891	0.037*

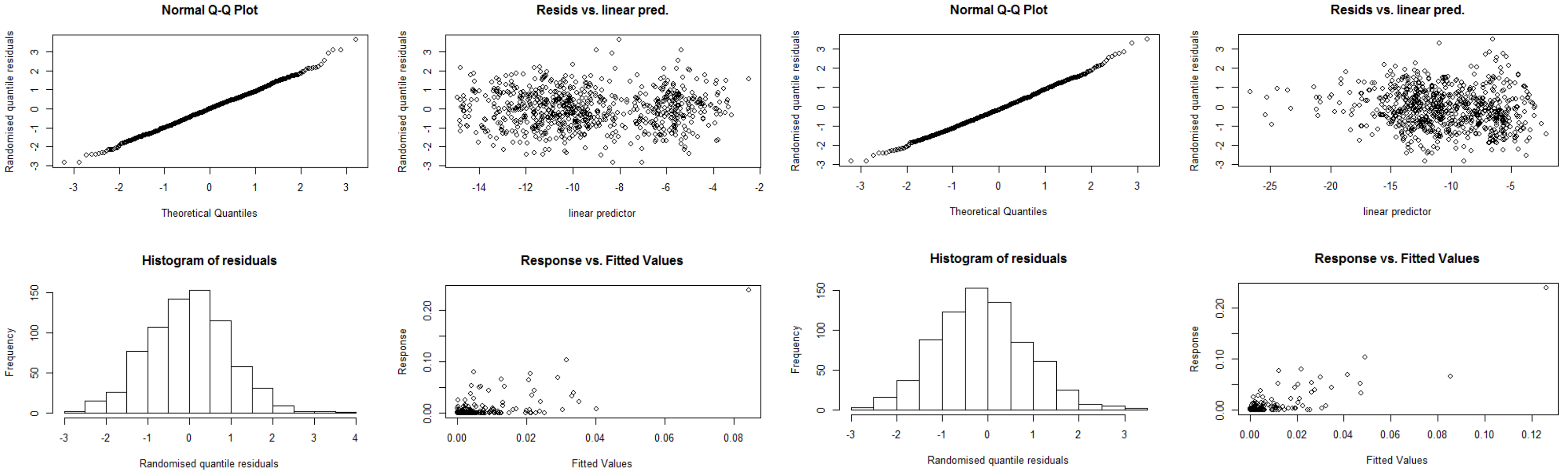

Figure A1.14. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of small gorgonian coral biomass using Alfredo trawl gear in the Eastern Arctic Region.

Campelen Trawl Gear

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean small gorgonian coral biomass using Campelen trawl gear are presented in Table A1.31. The R^2 was fair for the GAM model using the variables correlated at less than 0.7 and very low for the model using the RF-selected variables. The deviance explained was similar in both models. The AIC was higher in the GAM model using RF-selected variables than for the model using the variables correlated at less than 0.7 but the BIC lower The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.32 and A1.33, respectively.

Figure A1.15 shows the graphical diagnostics for both models. Both models showed fairly normal residuals and small patterns in the residuals vs. linear predictor plots. The response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

When predicted to the entire spatial extent of the study area, the models showed erroneously very high predicted biomass values in the deeper areas. High predictions of biomass were not alleviated with the inclusion of latitude and longitude in the model. The predicted surfaces therefore are not presented in this report.

Table A1.31. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of small gorgonian corals using Campelen trawl gear in the Eastern Arctic Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.083	0.128
Deviance explained	64.9%	68%
AIC	781.622	778.393
BIC	893.977	923.773

Table A1.32. Results of the GAM RF Variables model built to predict the biomass of small gorgonian corals using Campelen trawl gear in the Eastern Arctic Region.The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Temperature Average Range	2.848	4.255	3.115 x 10^{-3}*
Bottom Salinity Average Range	5.040	2.939	7.236 x 10^{-3}*
Surface Salinity Mean	2.371	7.980	2.520 x 10^{-5}*
Surface Temperature Mean	1.342 x 10^{-6}	0.029	1.000
Maximum Average Summer Mixed Layer Depth	5.117 x 10^{-7}	0.000	1.000
Surface Current Average Range	5.585 x 10^{-5}	0.079	0.998
Maximum Average Fall Mixed Layer Depth	2.438	10.146	1.220 x 10^{-6}*
Surface Temperature Average Minimum	2.015	7.953	1.420 x 10^{-4}*

Table A1.33. Results of the GAM 0.7 Variables model built to predict the biomass of small gorgonian corals Campelen trawl gear in the Eastern Arctic Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the α= 0.05 level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Maximum	7.145 x 10^{-6}	0.057	0.999
Bottom Temperature Average Range	1.756 x 10^{-5}	0.320	0.997
Annual Chlorophyll *a* Maximum	9.565 x 10^{-6}	0.000	1.000
Annual Chlorophyll *a* Minimum	1.822 x 10^{-1}	0.360	0.734
Spring Chlorophyll *a* Maximum	3.306	4.547	1.310 x 10^{-3}*
Spring Chlorophyll *a* Minimum	7.250 x 10^{-5}	0.246	0.995
Summer Chlorophyll *a* Maximum	1.906	1.447	0.228
Summer Chlorophyll *a* Minimum	2.320 x 10^{-5}	0.147	0.998
Depth	4.238	14.925	1.50 x 10^{-14}*
Maximum Average Spring Mixed Layer Depth	7.853 x 10^{-6}	0.146	0.999
Maximum Average Summer Mixed Layer Depth	8.303 x 10^{-1}	2.844	0.089
Summer Primary Production Average Maximum	6.170 x 10^{-6}	0.010	1.000
Summer Primary Production Average Minimum	1.402	8.686	4.11 x 10^{-4}*
Summer Primary Production Average Range	5.879 x 10^{-6}	0.002	1.000
Surface Current Average Maximum	3.521	3.888	2.831 x 10^{-3}*
Surface Salinity Average Range	4.496	6.108	7.030 x 10^{-6}*
Surface Temperature Average Minimum	1.478 x 10^{-5}	0.297	0.998
Slope	7.088 x 10^{-6}	0.006	1.000

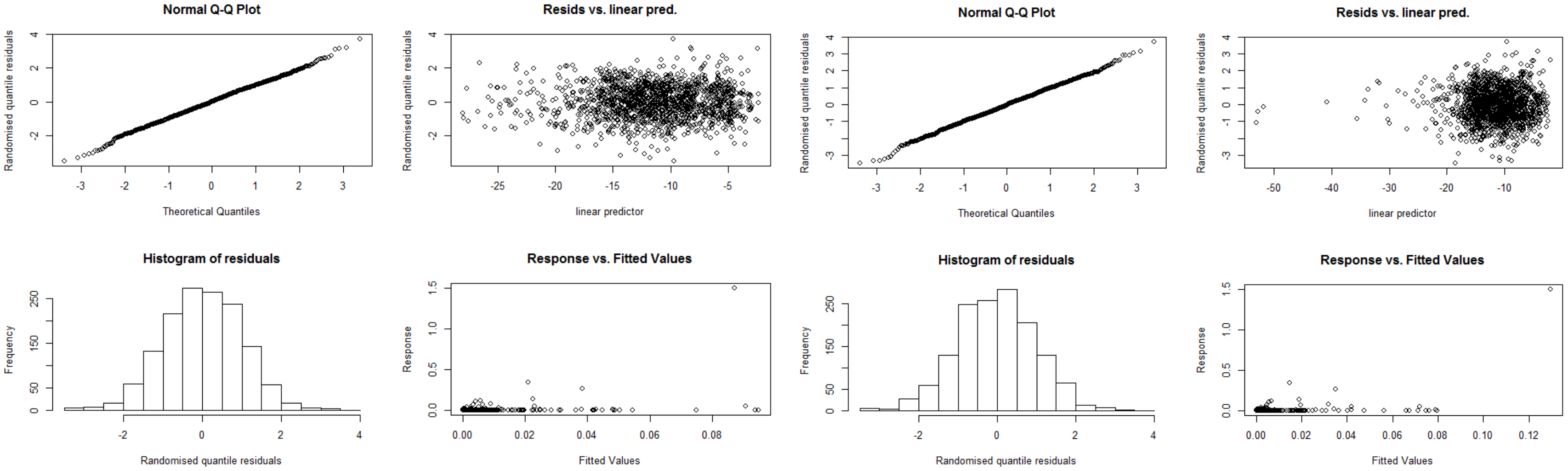

Figure A1.15. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of small gorgonian coral biomass using Campelen trawl gear in the Eastern Arctic Region.