



NOT TO BE CITED WITHOUT PRIOR  
REFERENCE TO THE AUTHOR(S)

Serial No. N7601

NAFO SCR Doc. 24/063

### **SCIENTIFIC COUNCIL MEETING – NOVEMBER 2024**

## **Vulnerable Marine Ecosystems in the NAFO Regulatory Area: Updated Species Distribution Models of Selected Vulnerable Marine Ecosystem Indicators (Large-Sized Sponges, Sea Pens and Black Corals)**

by

F.J. Murillo<sup>1</sup>, A.-L. Downie<sup>2</sup>, S. Abalo Morla<sup>3</sup>, C. Lirette<sup>1</sup>, N. Paulin<sup>1</sup>, Z. Wang<sup>1</sup>, E. Devred<sup>1</sup>,  
S. Clay<sup>1</sup>, M. Sacau<sup>3</sup>, C. Nozères<sup>1</sup>, M. Koen-Alonso<sup>4</sup>, L. Gullage<sup>4</sup>, and E. Kenchington<sup>1</sup>

<sup>1</sup>Department of Fisheries and Oceans, Bedford Institute of Oceanography, Dartmouth, NS, Canada.

<sup>2</sup>Centre for Environment, Fisheries and Aquaculture Science, Lowestoft, UK.

<sup>3</sup>Instituto Español de Oceanografía (COV-IEO), CSIC, Vigo, Spain.

<sup>4</sup>Department of Fisheries and Oceans, Northwest Atlantic Fisheries Centre, St. John's, NL, Canada.

### **Abstract**

The Northwest Atlantic Fisheries Organization (NAFO) Commission has called for a reassessment of the vulnerable marine ecosystems (VMEs) and impact of bottom fisheries on VMEs for 2027. Species distribution models (SDMs) help to inform on the closed area boundaries and have been used to modify the areas of significant concentrations of Large-Sized Sponges and Large Gorgonian Corals produced through kernel density analyses (KDE) in the previous review. Here we provide Random Forest SDMs for the Large-Sized Sponges, Sea Pens, and Black Corals. For the first time, we provide maps of uncertainty associated with the areas of predicted presence and absence. For the Large-Sized Sponges and Sea Pens, we had sufficient data to model the distributions of subsets of data for each, viz. the sponges of the sub-order Astrospora, the families Tetillidae and Polymastiidae, and sponge grounds (catches above a weight threshold from the KDE analyses), and for Sea Pens, the genera *Balticina*, *Funiculina*, *Anthoptilum* and *Pennatula*. Predictive models for the Large and Small Gorgonian Corals, Erect Bryozoans and Sea Squirrels will be separately presented for the 2025 meeting of the Working Group on Ecosystem Science and Assessment (WG-ESA), following the workflow presented herein.

### **Introduction**

Species distribution models (SDMs) predict the presence, absence, or abundance/biomass of a species or habitat (the response variable) from environmental variables (the predictor variables) thought to influence it. Potential uses of SDMs include 1) explanation, 2) mapping, and 3) transfer (Zurrell et al., 2020), with the first focused on identifying the main factors driving the species distributions, the second on producing maps of the distribution, and the third on forecasting or projecting the distributions into a different geographic region or time period. The primary objective of the SDMs presented here is for 'Mapping' (Zurrell et al., 2020). These models are particularly valuable in areas where survey vessels do not sample (e.g., rough bottom, cliffs) and for non-aggregating taxa such as the black corals that are present in low frequency. The maps will also be used to evaluate the area between trawl sets to determine if the full vulnerable marine ecosystem (VME) polygon (derived from kernel density analyses (KDE), which does not consider environmental variables (Kenchington et al., 2019)) is potential habitat, and to modify the boundaries of the VME polygons if they include areas of

predicted species absence. The latter was previously done to modify the VME polygons for Large-sized Sponges and Large Gorgonian Corals (NAFO, 2019).

SDMs for sponge grounds, a habitat dominated by massive structure forming sponges (Knudby et al., 2013 a,b), the glass sponge *Asconema foliata* (NAFO, 2019), black corals, large gorgonian corals and sea pen corals (Knudby et al., 2013c), erect bryozoans and sea squirts (*Boltenia ovifera*) (Kenchington et al., 2019) have previously been incorporated into the NAFO assessment of VMEs (NAFO, 2019). In support of the 2024 NAFO Commission Request#6, b: *Work towards the reassessment of VMEs and impact of bottom fisheries on VMEs for 2027, and c: Develop materials on the potential of submitting NAFO coral bottom fishing closed areas as OECMs for discussion at the 2025 WG-EAFFM meeting*, SDMs have been created using a common set of environmental predictors and response variables updated to include data through to 2023. For some groups like the sponges, an additional ten years or more of data were considered. We followed the Overview/Conceptualisation, Data, Model fitting, Assessment and Prediction (ODMAP) steps recommended by Zurrell et al. (2020) for standard reporting of SDMs, complemented by the recommendations of Sofaer et al. (2019) for the use of SDMs in decision-making.

Here we present SDMs for the VME functional groups Large-sized Sponges, Sea Pens, and Black Corals (mostly a single species *Stauropathes arctica*). Within the Large-sized Sponges we also present models for the sponge grounds (as in Knudby et al., 2013b), the sponge families Tetillidae and Polymastiidae (excluding genera *Radiella* and *Tentorium* as they are not VME indicator taxa (NAFO, 2024)), as well as for the Astrophorina, a suborder of massive sponges in the class Demospongiae. We were not able to construct a new model for the glass sponge *Asconema foliata* as the published data on this species (Murillo et al., 2016; NAFO, 2019) were collected in one year (2007) and were not provided for the analyses herein (see below). Within the Sea Pens, we also present models for the genera *Funiculina*, *Balticina*, *Anthoptilum* and *Pennatula* (including *P. grandis* which has been reassigned to the genus *Ptilella*). These additional models will be used to compare the results of the predicted distributions of individual taxa versus that of their functional group. The latter formed the original response data to earlier models as there were insufficient records for individual species, or a lack of confidence in the identifications of early records. The models for the subgroups will also be used to examine differential impacts of bottom fisheries and evaluate the proportional protection afforded to these subgroups by the existing closed areas. We discuss recommendations for which models to use in response to Commission Request#6b.

## Methods

### *Environmental data*

All layers were displayed using a NAD83 UTM 23N projection and the resolution of the final raster surfaces was 1 km. The spatial extent of the modelled area is bounded by the Canadian Exclusive Economic Zone (EEZ) to the west, and to the north, south and east by the 2500 m depth contour (derived from GEBCO 2024, see below). This area is referred to as the NAFO Regulatory Area (NRA) and includes Flemish Cap and the Nose and Tail of Grand Bank.

### *Water column variables*

Environmental layers representing water column properties were the physical oceanographic variables bottom temperature, bottom salinity, bottom current speed, bottom stress, mixed layer depth, surface temperature, surface salinity, and the biological oceanographic variables chlorophyll a, and primary productivity (Table 1). Monthly temperature, salinity, current speed, bottom stress, and mixed layer depth were extracted from the Bedford Institute of Oceanography North Atlantic Model (BNAM; Wang et al., 2018) for the period 1990-2023. Mean, maximum, minimum and range values derived from BNAM were calculated for all months within a year and averaged across all years. Bottom stress ( $\tau_b$ ) was calculated as  $\tau_b = 3.5 \times 10^{-3} \times \rho \times U_b^2$  where,  $\rho$  is the density of seawater [kg m<sup>-3</sup>] derived from BNAM, and mean bottom current velocity ( $U_b$ ) was calculated from eastward seawater velocity (U) and northward seawater velocity (V), with the following formula:  $U_b = \sqrt{U^2 + V^2}$ . For mixed layer depth, only maximum values were calculated as above, averaged across the time period and for seasonal time periods (Winter: January – March; Spring: April - June; Summer: July - September; Fall: October – December). Using ArcGIS Pro's Geostatistical

Wizard, BNAM (and BNAM-derived) point data were interpolated using ordinary kriging, and the resulting geostatistical layers were exported to the final raster surfaces.

Daily photosynthetically active radiation (PAR) and surface chlorophyll *a* concentration data collected by the MODIS sensor onboard the Aqua satellite from 2003-2023 were projected onto a 4.64-km resolution equal-area grid by NASA's Ocean Biology Processing Group (OBPG). The data were downloaded from the OBPG website and averaged into 8-day composites to minimize the effect of missing data (e.g., cloud cover, low solar angles), and simplify processing while still retaining sufficient detail in the time scale of the chlorophyll *a* fields to capture short-lived phytoplankton blooms. The Data Interpolating Empirical Orthogonal Functions (DINEOF) method was used to fill the remaining spatial gaps. Primary production values were derived from chlorophyll *a* concentration, PAR, and photosynthetic parameters that describe the rate of production as a function of available light (Platt and Sathyendranath, 2008). The photosynthetic parameters and their seasonal variation were derived from a database of ship-based incubation experiments (i.e., production-irradiance curve) carried out between 1977 and 2011 in the modelled area (NRA), and smoothed into 8-day climatologies to capture the seasonal phenology of the parameters. Averaged across years and seasonal periods, the mean, max, min and the range values were derived from the 8-day chlorophyll *a* and primary production composites. Seasons were delimited in the following manner: Winter: Jan 01 to Mar 29; Spring: Mar 30 to Jul 03; Summer: Jul 04 to Sep 29; Fall: Sep 30 to Dec 31. As with the BNAM data the resulting statistical layers were interpolated using ordinary kriging and the geostatistical layers were exported to the final raster surfaces.

**Table 1.** Water column variables used in the Random Forest models (Max: maximum; Min: minimum; MLD: Mixed Layer Depth; Chl: Chlorophyll; PP: Primary Production; BNAM: Bedford Institute of Oceanography North Atlantic model (Wang et al., 2018); SOPhyE: Satellite Ocean Colour and Phytoplankton Ecology Group at the Bedford Institute of Oceanography).

Variable	Metric	Unit	Native Resolution	Source
Bottom Salinity	Mean, Max, Min, Range	N/A	1/12 <sup>o</sup> lat/long	BNAM
Bottom Temperature	Mean, Max, Min, Range	°C	1/12 <sup>o</sup> lat/long	BNAM
Bottom Current Speed	Mean, Max, Min, Range	m s <sup>-1</sup>	1/12 <sup>o</sup> lat/long	BNAM
Bottom Stress	Mean, Max, Min, Range	m s <sup>-1</sup>	1/12 <sup>o</sup> lat/long	BNAM
Surface Salinity	Mean, Max, Min, Range	N/A	1/12 <sup>o</sup> lat/long	BNAM
Surface Temperature	Mean, Max, Min, Range	°C	1/12 <sup>o</sup> lat/long	BNAM
Surface Current Speed	Mean, Max, Min, Range	m s <sup>-1</sup>	1/12 <sup>o</sup> lat/long	BNAM
Annual MLD	Max	m	1/12 <sup>o</sup> lat/long	BNAM
Summer MLD	Max	m	1/12 <sup>o</sup> lat/long	BNAM
Fall MLD	Max	m	1/12 <sup>o</sup> lat/long	BNAM
Winter MLD	Max	m	1/12 <sup>o</sup> lat/long	BNAM
Spring MLD	Max	m	1/12 <sup>o</sup> lat/long	BNAM
Annual Chl <i>a</i>	Max, Mean, Min, Range	mg m <sup>-3</sup>	4 km	SOPhyE
Spring Chl <i>a</i>	Max, Mean, Min, Range	mg m <sup>-3</sup>	4 km	SOPhyE
Fall Chl <i>a</i>	Max, Mean, Min, Range	mg m <sup>-3</sup>	4 km	SOPhyE
Winter Chl <i>a</i>	Max, Mean, Min, Range	mg m <sup>-3</sup>	4 km	SOPhyE
Summer Chl <i>a</i>	Max, Mean, Min, Range	mg m <sup>-3</sup>	4 km	SOPhyE
Fall PP	Max, Mean, Min, Range	mg C m <sup>-2</sup> day <sup>-1</sup>	4 km	SOPhyE
Winter PP	Max, Mean, Min, Range	mg C m <sup>-2</sup> day <sup>-1</sup>	4 km	SOPhyE
Summer PP	Max, Mean, Min, Range	mg C m <sup>-2</sup> day <sup>-1</sup>	4 km	SOPhyE
Spring PP	Max, Mean, Min, Range	mg C m <sup>-2</sup> day <sup>-1</sup>	4 km	SOPhyE
Annual PP	Max, Mean, Min, Range	mg C m <sup>-2</sup> day <sup>-1</sup>	4 km	SOPhyE

### *Terrain variables*

GIS tools from the R package MultiscaleDTM (Ilich et al., 2023) and the System for Automated Geoscientific Analyses (SAGA) (v. 8.4.1; Conrad et al., 2015) accessed with the R package RSAGA (Brenning et al., 2022) were used to calculate terrain variables (Table 2) in the free statistical computing software R (v. 4.3.2, R Development Core Team, 2023). Terrain variables were derived from a digital elevation model (DEM) produced from the 15 arc-second gridded General Bathymetric Chart of the Oceans (GEBCO) 2024 (GEBCO Compilation Group, 2024) covering the modelled area (NRA). The bathymetric horizontal resolution corresponds to approximately 388 m at the study area's latitude. The GEBCO bathymetry data layer was then projected onto NAD83 UTM23N using the terra R package's "project" function using EPSG 26923 (Hijmans, 2024). The SAGA 'Fill sinks' tool (Wang and Liu, 2006) with a slope threshold of 0.005 was used to smooth out artefacts in the GEBCO DEM before calculating the derivative terrain layers (Wang and Liu, 2006).

SAGA was used to calculate slope, eastness, northness, ruggedness, channel network base level and distance, valley depth, relative slope position, LS-factor, positive/negative openness, and wind exposition index (here interpreted as current exposition). MultiscaleDTM was used to calculate fine- and broad-scale bathymetric position index (BPI). The topographic layers, their units of measurement, and the tools and function arguments used to produce them are summarised in Table 2. Default arguments for each tool or function were used unless otherwise stated.

Eastness and northness values were calculated using the sin and cosine, respectively, of the aspect values calculated by SAGA. The bathymetric position index (BPI) is a modified version of the topographic position index (Weiss, 2001) and measures the difference between the value of a focal cell and the mean value of neighbouring cells in an annulus around the focal cell (Lundblad et al., 2006). Fine-scale BPI was calculated using an annulus with an inner radius of 4 cells and an outer radius of 8 cells, while broad-scale BPI was calculated using an annulus with an inner radius of 4 cells and an outer radius of 64 cells. Ruggedness is a measure of seabed complexity measured as a function of the variability in elevation at a selected scale and was calculated using the vector ruggedness measure (VRM) in SAGA (Sappington et al., 2007).

The channel network base level, channel network distance, relative slope position (RSP), and valley depth are layers derived from two channel network layers. Channels for these layers were generated using Strahler order thresholds of 3 and 5, respectively. The lower order channel network retains smaller channels and delineates finer topographic features and branching. Channel network base level layers were calculated from the channel network layers and denote topographic highs and lows. Channel network distance layers were then calculated using the vertical distance between the base DEM and the channel network base level. Valley depth measures the distance between the DEM and interpolated ridge level defined by the Strahler order. For this variable, Strahler order thresholds of 3 and 5 produced identical results and so only that obtained with a threshold of 3 was retained. RSP (Böhner and Selige, 2006) is the location along the entire length of a slope on a scale from 0 (bottom) to 1 (top). Some errors were introduced in the values for channel network distance, valley depth, and RSP if the interpolated channel network base level was higher than the base DEM elevation in some cells. These errors produced values outside the valid range for these variables. Invalid values were snapped to the closest valid value for each layer: 0 for negative values for channel network distance, valley depth, and RSP, and 1 for values over 1 for RSP.

The LS-factor, a combination of slope length and steepness (gradient over the length), predicts erosion potential in the terrestrial environment (Desmet and Govers, 1996) and can also be applied in the marine context to reflect the potential stability of sediment deposits and hence the likelihood of exposed hard substrata.

Positive and negative topographic openness (Yokoyama et al., 2002) provide information on how prominent or sheltered an area is in relation to surrounding topography. Similarly, the wind exposition index represents how exposed an area is (Böhner and Antonić, 2009) to wind (or currents in the marine environment), where values below 1 are sheltered, and values above 1 are exposed.

All resulting terrain variable layers were then transformed to match the 1-km resolution and origin of other environmental data raster layers with the 'resample' function from the raster R package (Hijmans, 2023) using a bilinear interpolation method. Layers were then cropped and masked to the study area extent.

**Table 2.** Description of terrain variable layers calculated from GEBCO 2024 bathymetry data.

<b>Variable</b>	<b>Short name</b>	<b>Unit</b>	<b>R package</b>	<b>RSAGA library</b>	<b>SAGA module/MultiscaleDTM function</b>	<b>Arguments</b>
Fill-sink bathymetry*	FS005	m	RSAGA	ta_preprocessor	Fill Sinks (Wang & Liu, 2006)	MINSLOPE = 0.005
Slope	SLOPE	degrees	RSAGA	ta_morphometry	Slope, Aspect, Curvature	UNIT_SLOPE = 1
Bathymetric Position Index (fine-scale)	BPIF	index	MultiscaleDTM	N/A	BPI	w = c(4, 8)
Bathymetric Position Index (broad-scale)	BPIB	index	MultiscaleDTM	N/A	BPI	w = c(4,64)
Ruggedness	VRM	index	RSAGA	ta_morphometry	Vector Ruggedness Measure (VRM)	MODE = 0, RADIUS = 3
Eastness (aspect)	EAST	radians	RSAGA	ta_compound	Basic Terrain Analysis	
Northness (aspect)	NORTH	radians	RSAGA	ta_compound	Basic Terrain Analysis	
Channel Network Base Level (3 & 5)	CHNETBL3/5	m	RSAGA	ta_compound	Basic Terrain Analysis	THRESHOLD = 3 & 5
Channel Network Distance (3 & 5)	CHNETD3/5	m	RSAGA	ta_compound	Basic Terrain Analysis	THRESHOLD = 3 & 5
Valley Depth (3)	VALD	m	RSAGA	ta_compound	Basic Terrain Analysis	THRESHOLD = 3
Relative Slope Position (3 & 5)	RSP3/5	index	RSAGA	ta_compound	Basic Terrain Analysis	THRESHOLD = 3 & 5
LS-Factor	LSF	index	RSAGA	ta_compound	Basic Terrain Analysis	
Positive and Negative Openness	POP/NOP	radians	RSAGA	ta_lighting	Topographic Openness	
Wind Exposition Index	WEI	index	RSAGA	ta_morphometry	Wind Exposition Index	

\*Used as the digital elevation model (DEM) for all the other variables requiring a DEM input layer.

### ***Fishing effort variables***

Two methods were used to produce fishing effort layers (NAFO, 2019). The first method used both bottom trawling and bottom longline effort data resolved at a native resolution of 0.05 degrees (approximately 3.8 x 5.6 km<sup>2</sup>) and represented effort as hours fished per grid cell. The second method used only bottom trawling data, was resolved at a native resolution of 1 km<sup>2</sup>, and represented effort as km trawled per km<sup>2</sup> per year. Both fishing effort layers were used in the analyses as they capture different aspects of fishing pressure at different scales of resolution.

#### *Bottom Trawling and Bottom Longline Effort 0.05 x 0.05 Degree Native Resolution*

The method used to estimate cumulative bottom fishing effort (bottom trawling and bottom longline) was based on data from the Vessel Monitoring System (VMS), logbook records (haul-by-haul data), and the IEO Scientific Observer Program, from 2016 to 2022. This work was conducted as part of the "NAFO Potential Vulnerable Marine Ecosystem-Impacts of Deep-sea Fisheries" NEREIDA project, which was funded by the European Union (EU) through NAFO. The analysis was carried out using the improved methodology for "coupling VMS and logbook data", first described by Sacau et al. (2020) and later by Garrido et al. (2023). Recognizing the potential for errors in both data sources, a subset of records from the merged VMS and logbook database was selected for vessels with a Spanish scientific observer on board. The purpose of this selection was to evaluate the extent and nature of errors in each data source, based on the assumption that the actual fishing effort for these specific hauls was accurately reported by the scientific observers on board. The core principle of this method is that haul-by-haul catch data (logbook) and VMS are complementary data sources. By using haul-by-haul data, VMS pings can be classified as "fishing" or "non-fishing" depending on whether they fall within the fishing time intervals reported in the haul-by-haul catch data. This approach allows the start and end time stamps of fishing events from the logbooks to be used to extract relevant VMS points, which are then mapped spatially to represent fishing effort. Since these VMS points occur directly within the reported fishing time interval, they are considered to be associated with fishing activity. The coupling of the two datasets has already proven to be highly effective in describing the spatial distribution of fishing activity with much finer resolution (NAFO, 2017, 2018, 2019). Effort was represented by VMS ping time (i.e. the time interval between consecutive fishing pings), which was summed to produce hours fished and applied to a 0.05 x 0.05 degree grid. The cumulative bottom fishing effort obtained by Garrido et al. (2023) using this methodology was interpolated to a 1x1 km<sup>2</sup> grid employing the nearest neighbor approach with the terra package in R version 4.3.2. Lastly, the raster was cropped and masked to the bottom fishing footprint extent.

#### *Bottom Trawling Effort 1 km<sup>2</sup> Native Resolution*

Fishing effort for bottom trawls was defined as kilometers of trawl track travelled per km<sup>2</sup> per year. This is a departure from the previously used effort unit of hours fished per km<sup>2</sup> per year, which was calculated from the accumulation of the hourly VMS pings. Line features representing the tracks of fishing vessels corresponding to individual fishing events (trawl tows) were initially created by the NAFO Secretariat using VMS data between 2010 and 2018 (NAFO, 2019) and later updated by WG-ESA to cover between 2010 and 2021 (NAFO, 2022). Each track represents the movement of individual fishing vessels which are filtered based on known fishing speeds (0.5 – 5 knots) derived from logbook data (2016-2021). Each line was also attributed with the type of fishing gear used by the vessel. Using VMS tracks instead of raw VMS pings accounts for vessel trajectory, which is relevant for fisheries that follow depth contours, and eliminates the need to associate fishing effort at the grid scale used to collate VMS pings (0.05 degrees / ~ 5 km). Considering that the distance travelled is clearly related to bottom impact for trawlers, as the trawl travels on the sea floor, the fishing effort layer was produced using a moving window approach (NAFO, 2019; NAFO, 2020). The total length of VMS track within a specified neighbourhood was calculated in meters using the ArcGIS Spatial Analyst 'Line Statistics' tool (ArcGIS 10.5). The radius of the circular neighbourhood was set at 565 m (area = 1 km<sup>2</sup>). The resolution of the output raster layer was 1 km. The output was converted to the unit of km/km<sup>2</sup>/year by first converting meters into kilometers and dividing the line length by the number of years of data included in the VMS tracks line feature. Lastly, the raster was cropped and masked to the bottom fishing footprint extent.

### **Biological data**

The data records used for the response data in the SDMs were drawn from the research vessel trawl surveys conducted by the NAFO contracting parties from trawl sets in the NAFO Regulatory Area on Flemish Cap and the Nose and Tail of Grand Bank to 2500 m (Table 3). Over time, the at-sea identification and coding of the VME Indicator taxa has evolved and different identifications were reviewed and consolidated for each modeled taxon (Appendix Table A1). The time frame for the identification of the lower level taxa differs among surveys and is indicated in the descriptions for each taxon below and in the Appendices (Appendix Tables A2, A3). Initially, data from Canada and Spain were identified only by the functional group attribution and not the taxon name or species code, even if such data were recorded at sea. That was because the data were used for the KDE analysis and identification of VME polygons at the functional group level (Kenchington et al., 2014), and finer taxonomic resolution was not needed. Those earlier records could be reviewed to ensure that the taxon names were consistent with the functional groups used today, however, as there are sufficient records that have been validated with taxon names (Appendix Table A1) for SDMs, those earlier records were evaluated on a case-by-case basis for inclusion in the models (see below for details for each functional group).

Survey data were used to record both species biomass (kg), and presence or absence. Absence data at the functional group level (i.e., Large-Sized Sponges, Sea Pens and Black Corals) were determined on a tow-by-tow basis for each mission that recorded the presence of the functional group amongst their trawl sets. The assumption was that if the functional group was recognized and recorded on the survey mission, its absence was not likely due to identification issues. For some functional groups where presence was not consistently recorded in the earlier years the associated absence (null) data were excluded from the SDM. The same procedure was used to identify null data for the SDMs of the lower-level taxa within each functional group. All nulls for the functional group were used, in addition to null data where the subgroups were not observed. This was necessary to fill gaps within functional group distributions where particular taxa may not occur. Once this data set was produced, subgroup-specific nulls were extracted. A pivot table was created for each set and nulls calculated for each subgroup, so the number of nulls will differ by taxon. As an additional check, the number of presences by year was examined to ensure that there were no trends in recording the taxon prior to accepting subgroup null data. This was used to reduce the number of observations for the black corals (Appendix Table A2) as they were not consistently recorded in the earlier years compared with later observations from the same survey area.

**Table 3.** Research Vessel Survey Data from NAFO Contracting Parties (EU and Canada); EU, European Union; DFO, Department of Fisheries and Oceans; NL, Newfoundland and Labrador; IEO, Instituto Español de Oceanografía; IIM, Instituto de Investigaciones Marinas; IPMA, Instituto Português do Mar e da Atmosfera.

<b>Data Source</b>	<b>Period</b>	<b>NAFO Division</b>	<b>Gear</b>	<b>Mesh Size in Codend Liner (mm)</b>	<b>Trawl Duration (min)</b>	<b>Average Wingspread (m)</b>
Spanish 3NO Survey (IEO)	2002 - 2023	3NO	Campelen 1800	20	30	24.2 - 31.9
EU Flemish Cap Survey (IEO, IIM, IPIMAR)	2003 - 2023	3M	Lofoten	35	30	13.89
Spanish 3L Survey (IEO)	2003 - 2023	3L	Campelen 1800	20	30	24.2 - 31.9
DFO NL Multi-species Surveys (DFO)	1995 - 2022	3LNO	Campelen 1800	12.7	15	15 - 20

#### *Large-Sized Sponges*

The available data for the SDM models for the VME functional group Large-Sized Sponges included 7809 validated presence records (Appendix Table A1) and 4907 null records obtained from the surveys shown in Table 3. Only two records from 2013 had biomass values but no taxon name. Those were included in the response data set. Recording of sponges in the catch data appears to have been more consistent after 2006 (Appendix Table A2). Most of the 7809 records (62%) were listed as Porifera (Canadian surveys) or ESPONJAS (EU surveys). While the Canadian surveys presently record sponges under this taxon code only, the EU surveys

have regularly provided information on individual sponge taxa with their data requests since 2011. This has facilitated the modeling of the sponge groups Tetillidae, Polymastiidae (excluding species formerly considered in the genera *Radiella* [e.g., *Radiella hemisphaerica* currently accepted as *Polymastia hemisphaerica*] and *Tentorium* as they are not VME indicator taxa (NAFO, 2024)), and Astrophorina. At present, the number of records for the VME Indicator taxa *Mycale* (N=90) and Axinellidae (N=127) were deemed insufficient for generating separate SDMs and they were only included in the SDM for the functional group. For the genus *Asconema*, which had previously been modeled using data collected in 2007 (Murillo et al., 2016; NAFO, 2019) but not provided for this analysis, 388 records were available under the at-sea identifications 'Asconema' and 'ASCONEMA SP'. Although this represented a sufficient number of records to support model generation, closer examination of their spatial distribution revealed that no data with these codes had been collected in the EU Spanish surveys of Flemish Cap (Appendix Table A3), where the species was previously found to have the highest probability of occurrence (Murillo et al., 2016; NAFO, 2019 Fig. 12.27). Consequently, this taxon could not be modeled with the available data.

To ensure SDMs of sponge groups Tetillidae, Polymastiidae (with the exceptions noted above), and Astrophorina could be directly compared with SDMs of the Large-Sized Sponge functional group and the Sponge Grounds, the response data compiled for Large-Sized Sponges comprised only of EU records from 2011-2023.

With much of the data recorded at a level of taxonomic resolution that could not ensure the exclusion of non-VME taxa, the model of the Large-Sized Sponge functional group omitted data that were listed as '*Radiella* sp.', '*Tentorium* sp.', '*Rhizaxinella* spp.', '*Stylocordyla* sp.', and Sycettidae which are not VME indicator taxa (NAFO, 2024) as well as the 'DEMOSPONGIDAE', 'Porifera' and 'ESPONJAS (PORIFERA)'. Records were included for taxa which meet the FAO (2009) guidelines for VME indicator species even if they are not presently included in the VME indicator taxa list (NAFO, 2024) as they will be recommended for inclusion in the next revision of the taxa in 2027. These taxa include 'Pheronematidae' and '*Poecillastra compressa*'. Records of '*Isops* spp.' were included as '*Isops phlegraei*', formerly *Geodia phlegraei*. Higher order taxa for 'Astrophorida', 'Astrophorina' and 'ASTROPHORINA (ASTROPHORIDA)' were included, as they include the VME geodiids and other taxa listed in the VME indicator list (NAFO, 2024), as were 'Ancorinidae', which is the family of *Stelletta* spp. and *Stryphnus* spp. and are also included in the VME indicator list.

Lastly, to compare results with those of Knudby et al. (2013a,b), the biomass threshold used to identify significant concentrations of sponges (i.e., >100 kg/tow) (Kenchington et al., 2019), was used to select a subset of sponge observations and generate a presence/absence Random Forest model for Sponge Grounds.

The final biological data used for the response data in the Large-Sized Sponges SDMs included 1182 presence/biomass records from 2011-2023 (excluding 2014 where species identification was not consistent, Appendix Table A3), and 1716 associated null data. For Sponge Grounds there were 67 presence/biomass records from 2011-2023, and 2831 associated null data. Response variables for the SDM of Tetillidae included 211 records recorded as '*Craniella*', '*CRANIELLA* SP', '*Craniella* spp' and 'Tetillidae', and 4096 null records. Response variables for the SDM of Polymastiidae included 621 records listed as 'Polymastiidae' (Appendix Table A1) and 3686 null records. The models for the sponge suborder Astrophorina used 401 records, from the at-sea identifications for 'Ancorinidae', 'Astrophorida', 'Astrophorina', 'ASTROPHORINA (ASTROPHORIDA)', 'Geodia', 'GEODIA SP.', 'Geodia spp', 'Geodiidae', 'Isops spp.', 'Poecillastra compressa', 'STELLETA SP', 'STELLETA SPP', 'Stelletta', 'Stryphnus', 'Stryphnus sp.', 'STRYPHNUS SPP', 'Thenea', 'Thenea levis', 'THENEA MURICATA', 'THENEA SP', and 'Thenea spp.', and 3906 null records (Table 4). This data set has been archived on the NAFO Sharepoint site.

### Sea Pens

The available data for the VME functional group Sea Pens included 4017 presence records (Appendix Table A1) and 5786 null records obtained from the surveys shown in Table 3. Records for the genera *Anthoptilum* and *Balticina* (formerly *Halipteris*) first appeared in 2005 from the Canadian surveys (Table 3) when survey identification codes for sea pens were first introduced, while *Funiculina* was not recorded until 2006 and *Pennatula* until 2009. The EU surveys also began recording sea pens in 2005, and details for individual sea pen taxa became available in 2011. This chronology is reflected in the numbers of records with taxon names reported in each year (Appendix Table A2).



The modelled genera are well represented in the data records (Appendix Table A1, Appendix Table A2). For *Anthoptilum*, 1292 records were obtained under the identifications ‘Anthoptilum’, ‘Anthoptilum grandiflorum’, ‘ANTHOPTILUM GRANDIFLORUM’, ‘ANTHOPTILUM MURRAYI’, ‘Anthoptilum murrayi’, ‘ANTHOPTILUM SP’, ‘Anthoptilum sp.’, and ‘Anthoptilum spp’, with 69% of the records listed as ‘Anthoptilum’. For *Balticina* (formerly *Halipterus*), there were 688 records compiled from the following at-sea identifications: ‘Balticina finmarchica (=Halipterus)’, ‘Halipteridae’, ‘Halipterus cf. christii’, ‘Halipterus christii’, ‘HALIPTERIS FINMARCHICA’, and ‘Halipterus finmarchica’, with 86% of those recorded as ‘Halipterus finmarchica’ and ‘HALIPTERIS FINMARCHICA’. *H. christii* and *H. finmarchica* have different distributions, with the former found in the shallower waters on Flemish Cap. For *Funiculina*, of the 423 records, 98% of them were ‘F. quadrangularis’ (including ‘FUNICULINA QUADRANGULARIS’), with 8 records appearing as ‘Funiculina’. It is highly likely that the SDM for this genus is representative of *F. quadrangularis*. For *Pennatula* 524 records were used as response data in the SDM, including records for ‘Pennatula’, ‘Pennatula aculeata’, ‘PENNATULA ACULEATA/PHOSPHOREA’, ‘Pennatula grandis’, ‘PENNATULA GRANDIS’, ‘Ptilella grandis (=Pennatula)’, ‘Pennatula phosphorea’, and ‘Pennatula sp.’. For 747 records no taxon name or biomass was provided. These were collected mostly from 2005-2010, although one record a year exists for 2012, 2013 and 2014 which were excluded from the models. A breakdown of the number of records in each taxon group by year is provided in Appendix Table A4.

To ensure SDMs of *Anthoptilum* spp., *Balticina* spp., *Funiculina* spp. and *Pennatula* spp. could be directly compared with those of the Sea Pen functional group, the response data compiled for Sea Pens consisted only of records from 2011-2023 (Appendix Table A4).

The final biological data used for the response data in the Sea Pen SDMs included 1721 presence/biomass records from 2011-2023, and 3988 associated null data. Presence/biomass (absence) records for named taxa from 2011-2023 included 1200 (4509) records for *Anthoptilum* spp., 642 (5067) records for *Balticina* spp., 391 (5318) records for *Funiculina* spp., and 441 (5268) records of *Pennatula* spp. These data came from 5709 trawl sets (Table 4). This data set has been archived on the NAFO Sharepoint site.

#### *Black Corals*

The biological data for the SDM models for the VME functional group Black Corals included 365 data presence/biomass records (Appendix Table A1) and 6744 null records obtained from the surveys shown in Table 3. These presence/biomass records were compiled from those identified at sea as ‘Antipatharia’, ‘Antipatharia sp. (ORDER)’, ‘Stauropathes arctica’, ‘STAUROPATHES ARCTICA’, and ‘Leiopathes cf. expansa’, and also included records for the functional group but with no taxon name or biomass provided (N=111). The majority of records with taxon names (N=254) were of *Stauropathes arctica* (74%).

Of those records with taxon names, the earliest were from the Canadian surveys with gaps in reporting years (Appendix Table A2), which raised questions surrounding the consistency of recording and the validity of these data. Consequently, presence/biomass records with taxon names (N=14) and without taxon names (N=111) collected from 2002-2010 and associated absence records (N=1971), were excluded from the models.

The final biological data used for the response data in the Black Coral SDMs included 240 presence/biomass records from 2011-2023, and 4776 associated null data records (Table 4). This data set has been archived on the NAFO SharePoint site.

**Table 4.** Summary of the Response Data Inputs to the Random Forest Species Distribution Models.

Response Group	Period	No. Presences	No. Absences
Large-Sized Sponges	2011 - 2023	1182	1716
Sponge Grounds	2011 - 2023	67	2831
Tetillidae	2011 - 2023	211	4096
Polymastiidae	2011 - 2023	621	3686
Astrophorina	2011 - 2023	401	3906
Sea Pens	2011 - 2023	1721	3988
<i>Anthoptilum</i>	2011 - 2023	1200	4509
<i>Balticina</i>	2011 - 2023	642	5067
<i>Funiculina</i>	2011 - 2023	391	5318
<i>Pennatula</i>	2011 - 2023	441	5268
Black Coral	2011 - 2023	240	4776

### Variable reduction

Preliminary SDMs were generated for each of the modelled taxa using the full suite of predictor variables to rank variable importance (Appendix Table A5). Following this, an iterative approach was used to conduct model specific variable selection. First, Spearman correlations were calculated for variable pairs and for those with correlation scores > 0.70, the least important variable was removed. Subsequently, the variable inflation factor (VIF), which measures the amount of inflation in the variance of a regression coefficient due to multicollinearity, was evaluated for the remaining uncorrelated variables. If VIFs > 10 were observed, the Spearman correlation scores were recomputed with progressively lower thresholds (decreased by increments of 0.05) until all remaining predictor variables achieved a VIF < 10.

### Model fitting

Models predicting the probability of presence for each taxon were built using classification Random Forest models. Random Forest is an ensemble method, where a large number of decision trees (typically 500-1000) are built using random subsets of the data (Breiman, 2001; Cutler et al., 2007). The models were built in the free statistical computing software R (v.3.5.1, R Development Core Team, 2018) using the 'randomForest' package (Liaw and Wiener, 2002) modified to output desired maps and tables (Appendix Table A5). The models were run using the default settings of the randomForest function, using 500 trees.

Predictor importance was investigated for each model using the decrease in end node impurity, measured by the Gini index for presence/absence. Partial response plots were used to visualize the relationship between each predictor variable and the response variables in turn, while accounting for the average effect of the other predictors in the model.

Models were validated using a bootstrap k-fold cross-validation procedure. For each response variable, the data was randomly subsampled into 10 folds and train sets constructed leaving each fold out in turn, to be used as test data (resulting in a 90/10 split, keeping balance of classes equal). Models were built using each train set, and validation statistics calculated for each corresponding test dataset. A cross-validation approach, such as this, gives an average cross-validation score, but also an estimate of variability around the mean. The variability can be used as an indicator of the stability of the model fit, and to check for the arbitrary effects caused by subsetting data to train and test a model. Accuracy measures used to validate the models included Sensitivity, Specificity, Kappa, True Skill Statistic (TSS, Allouche et al., 2006) and Balanced Accuracy, with the mean and standard deviation calculated across model runs (N=10).

Sensitivity, also referred to as the True Positive Rate, corresponds to the proportion of observed presences correctly predicted as such. Conversely, Specificity, or True Negative Rate, is the proportion of absences correctly predicted. These can be used to judge how likely a model is to detect presence and how specific the

predictions are to the correct class. High sensitivity with a low specificity indicates a model that is overpredicting, whilst an underpredicting model shows high specificity and low sensitivity. The overall accuracy was additionally investigated using the Kappa statistic, a measure of performance which takes account of class imbalance. Also computed were the TSS (Sensitivity + Specificity - 1) and Balanced Accuracy (average of Sensitivity and Specificity) which, unlike Kappa, are both independent of prevalence and can give a much better estimate of overall model performance where the classes are unbalanced.

Binary presence/absence maps were created by using two thresholds, the prevalence of the data and a threshold optimised to ensure that resulting Sensitivity and Specificity are afforded equal weight (Sensitivity=Specificity). The former was used in previous work (Kenchington et al., 2019), as a threshold to account for the class imbalance in data. However, in this model iteration the Sensitivity=Specificity threshold was used in preference, as it affords equal weight to detection of presence and absence, minimising both false positives and false negatives.

The final model output was plotted as the class (presence/absence) with the majority vote of all 10 model runs. Two confidence map layers were also produced consisting of: 1) the frequency of the most common class (N/10), and 2) the average probability over all 10 model runs of the majority vote class.

## Results

### *Assessment and Prediction of Large-Sized Sponges*

Random Forest models predicting the probability of the presence of Large-Sized Sponges generally had high accuracy scores across the validation statistics (Balanced Accuracy, Sensitivity, and Specificity all > 0.7; Table 5). However, Kappa, which measures the extent to which the agreement between observed and predicted is higher than that expected by chance alone, was of 'moderate' (> 0.5) performance for Large-Sized Sponges functional group and Sponge Grounds, and 'fair' (> 0.3) for individual sponge taxa (Tetillidae, Polymastiidae, Astrophorina). The TSS, defined as the average of the net prediction success rate for present sites and that for absent sites was 0.86 for Sponge Grounds which indicates high model performance, 0.65 for Large-Sized Sponges functional group and 0.61 for Tetillidae, which indicates a good model performance, and a fair model performance for Polymastiidae and Astrophorina.

**Table 5.** Model Validation Results for the Presence/Absence Random Forest Model for the Large-Sized Sponges VME Functional Group, Sponge Grounds, and Subgroups. TSS=True Skill Statistic (Sensitivity + Specificity - 1).

	<b>Large-Sized Sponges Functional group</b>	<b>Sponge Grounds</b>	<b>Tetillidae</b>	<b>Polymastiidae</b>	<b>Astrophorina</b>
<b>Accuracy Measure</b>	<b>Mean ± SD</b>	<b>Mean ± SD</b>	<b>Mean ± SD</b>	<b>Mean ± SD</b>	<b>Mean ± SD</b>
Sensitivity	0.83 ± 0.03	0.94 ± 0.08	0.81 ± 0.05	0.78 ± 0.02	0.79 ± 0.03
Specificity	0.82 ± 0.03	0.92 ± 0.04	0.79 ± 0.05	0.77 ± 0.02	0.79 ± 0.03
Kappa	0.64 ± 0.05	0.53 ± 0.16	0.22 ± 0.06	0.38 ± 0.04	0.32 ± 0.05
Balanced Accuracy	0.82 ± 0.03	0.93 ± 0.06	0.80 ± 0.05	0.78 ± 0.02	0.79 ± 0.03
TSS	0.65 ± 0.05	0.86 ± 0.11	0.61 ± 0.09	0.55 ± 0.05	0.58 ± 0.06

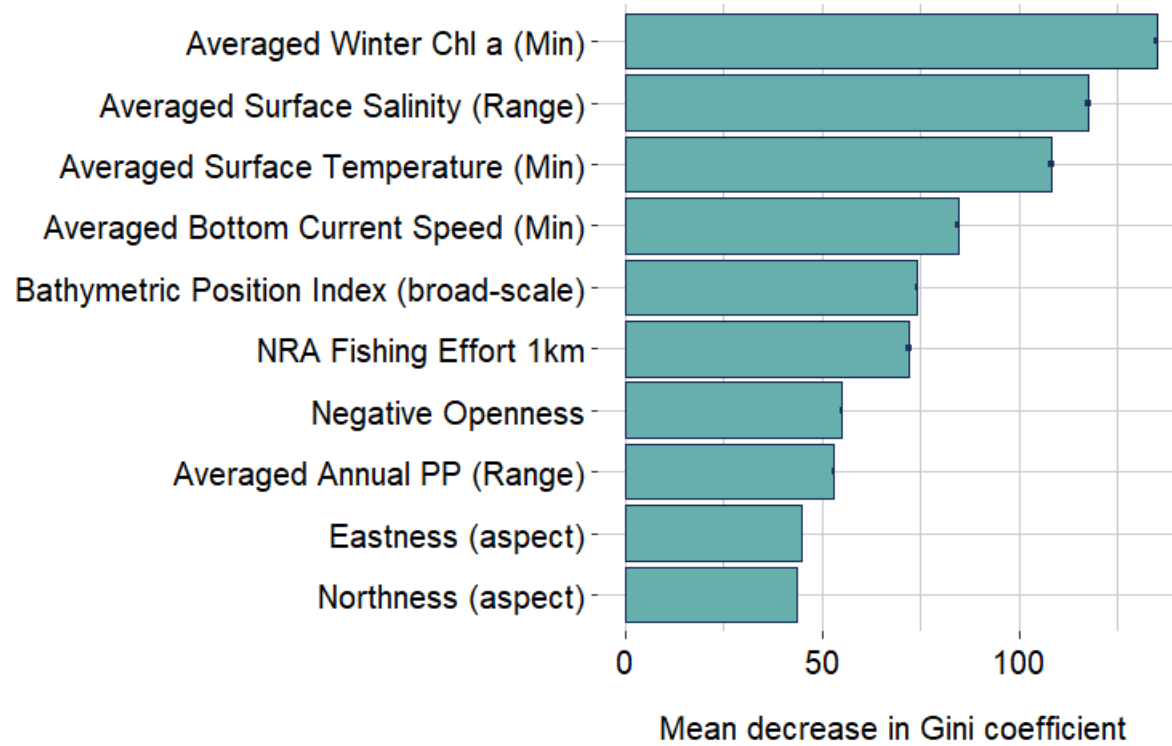
### *Large-Sized Sponges Functional Group*

The five most important variables for the Large-Sized Sponges functional group were the averaged winter mean value of chlorophyll *a*, followed by the averaged range of surface salinity, the averaged minimum value of sea surface temperature, the bottom current speed, the broad-scale of the bathymetric position index, and the bottom trawl fishing effort in the NRA (1 km resolution) (Figure 1). The models indicate that the Large-Sized Sponges are found in depressed or elevated areas with a winter mean value of chlorophyll *a* < 0.3 mg m<sup>-3</sup>, ranges of surface salinity > 1.3‰, and with low bottom trawl fishing effort (< 10 km/km<sup>2</sup>/year) (Figure 2).

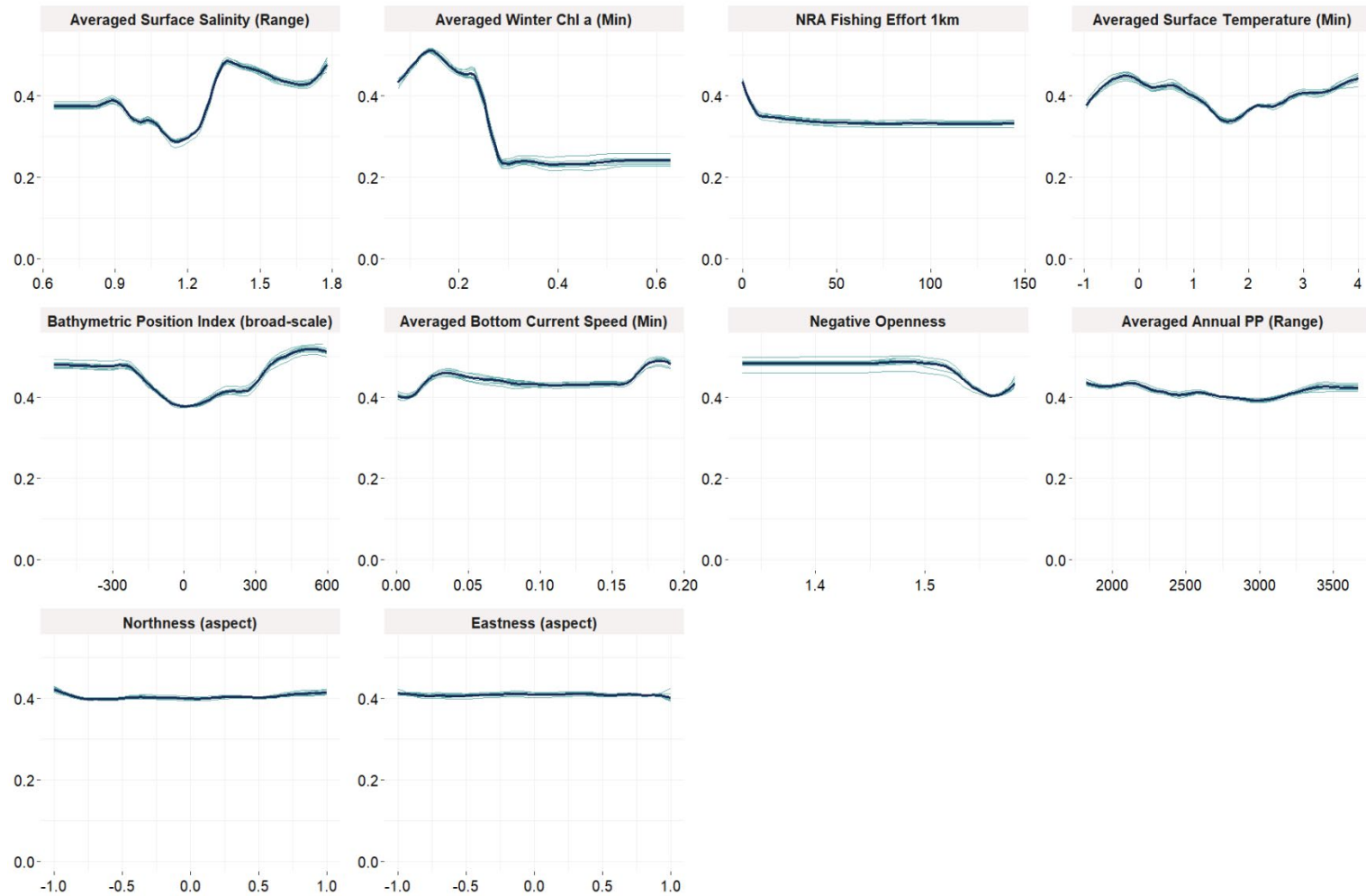
The predicted distribution maps are shown in Figure 3 as binary plots of presence/absence based on two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The data distribution is

shown overlain on the binary map of presence/absence based on Sensitivity=Specificity (Figure 4). Outside the areas extrapolated by the model, the Large-Sized Sponges are distributed across Flemish Cap, Flemish Pass, and the flanks of the Grand Bank of Newfoundland.

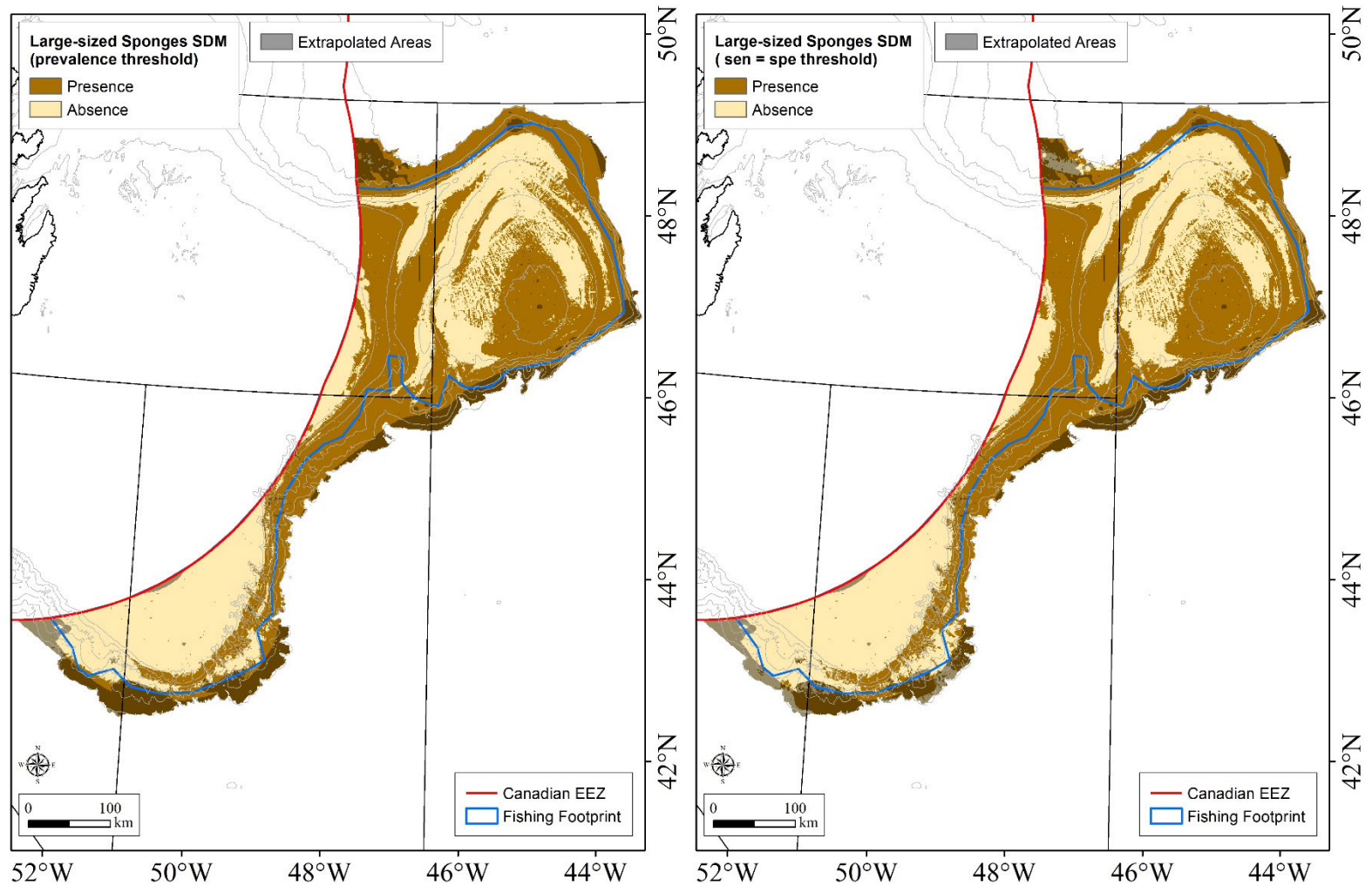
The uncertainty expressed as the frequency of presences and absences for the 10 cross-validation runs (Figure 4), the areas of extrapolation (Figures 3-5) and the average probability of the maximum frequency class (Figure 5) indicated high certainty within the fishing footprint for both presence and absence predictions.



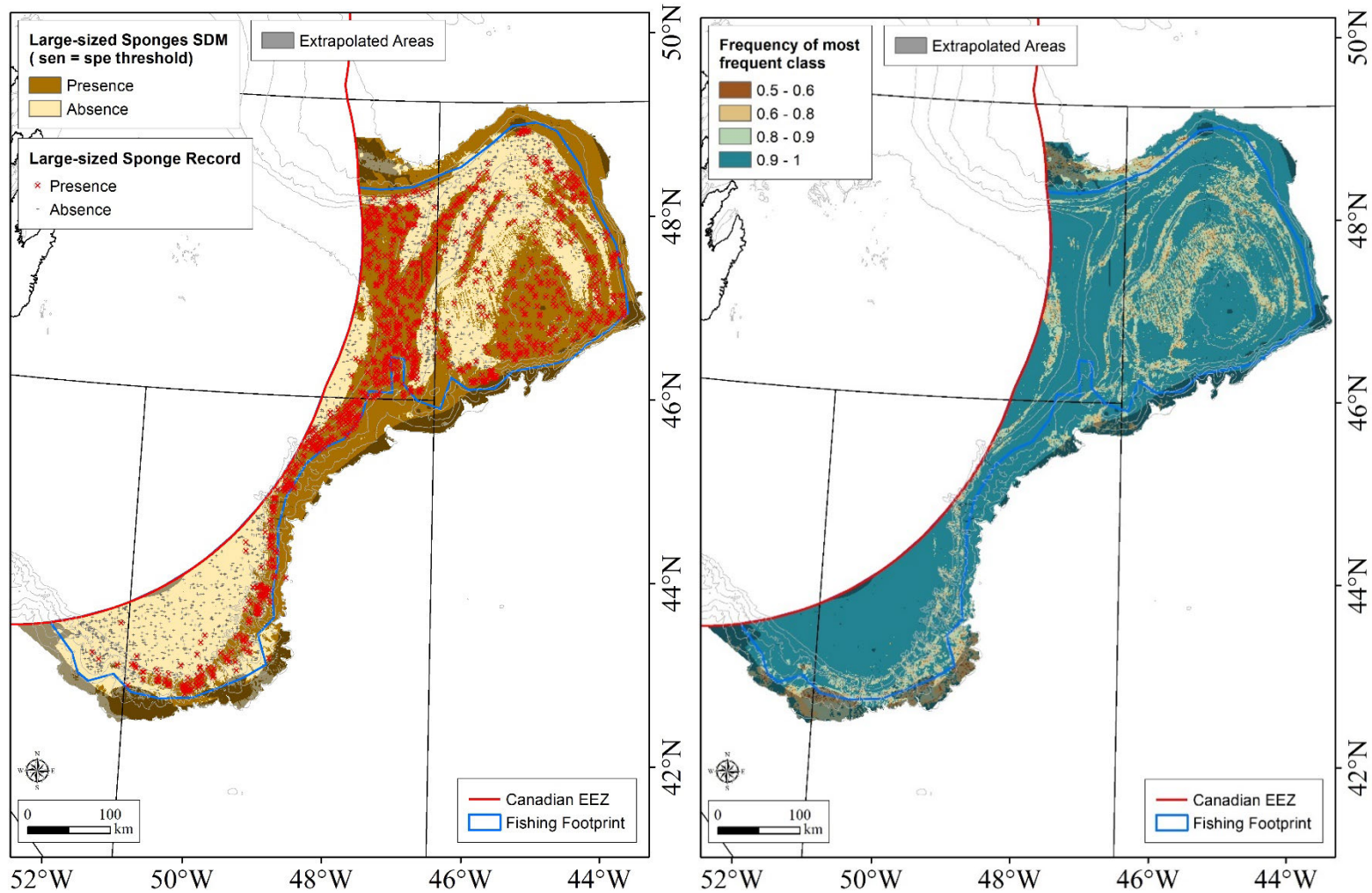
**Figure 1.** Plot of mean decrease and standard deviation in Gini Value for the 10 predictor variables in the Random Forest model for the Large-Sized Sponge functional group, indicating their relative importance and variation across 10 model folds.



**Figure 2.** Response curves showing the partial dependence of the probability of presence on the predictors (Figure 1) identified in the Random Forest model for the Large-Sized Sponges functional group. For each variable, the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.

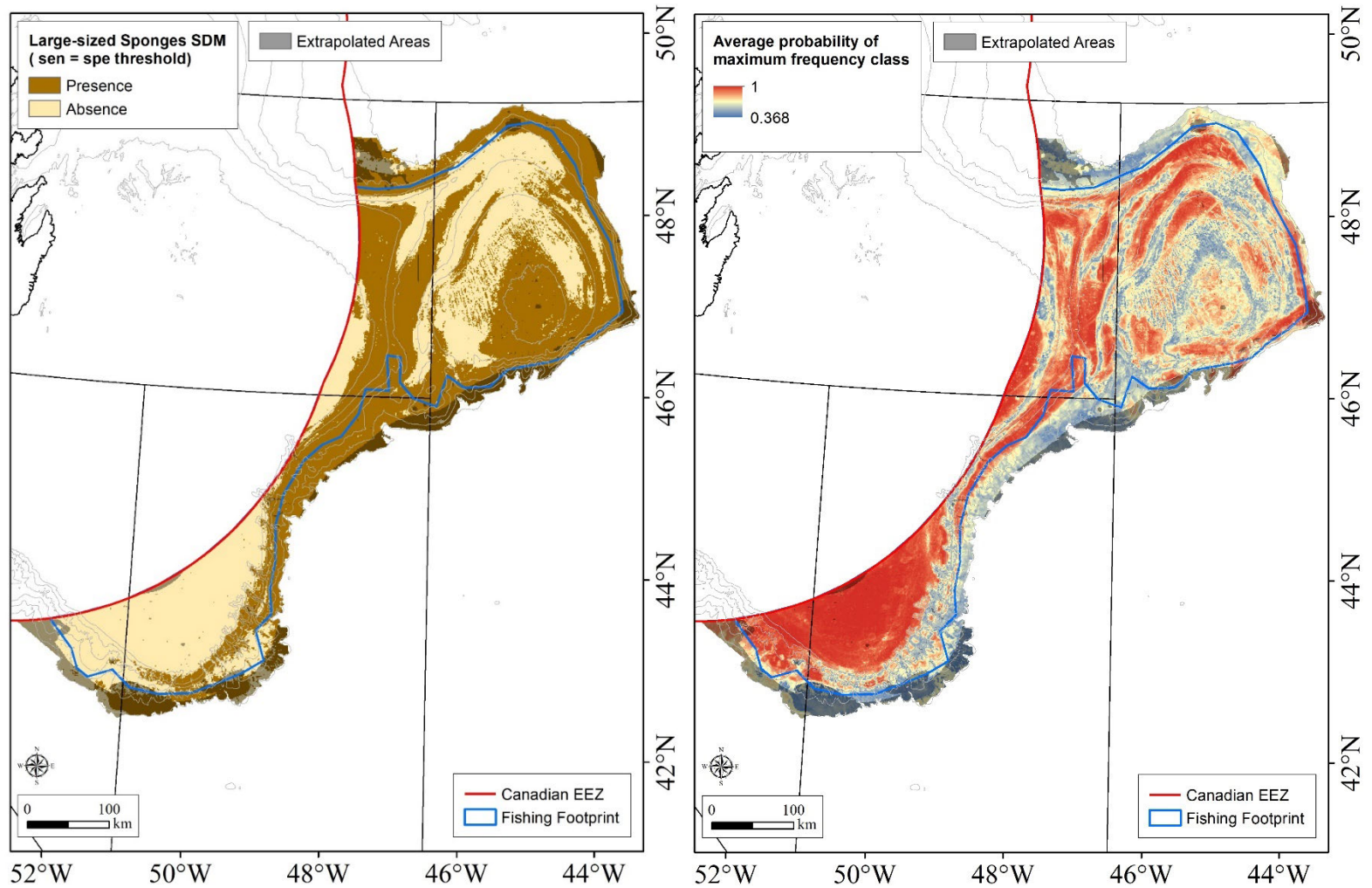


**Figure 3.** Random Forest species distribution model for the VME functional group Large-Sized Sponges showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 4.** Random Forest species distribution model for the VME functional group Large-Sized Sponges showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.





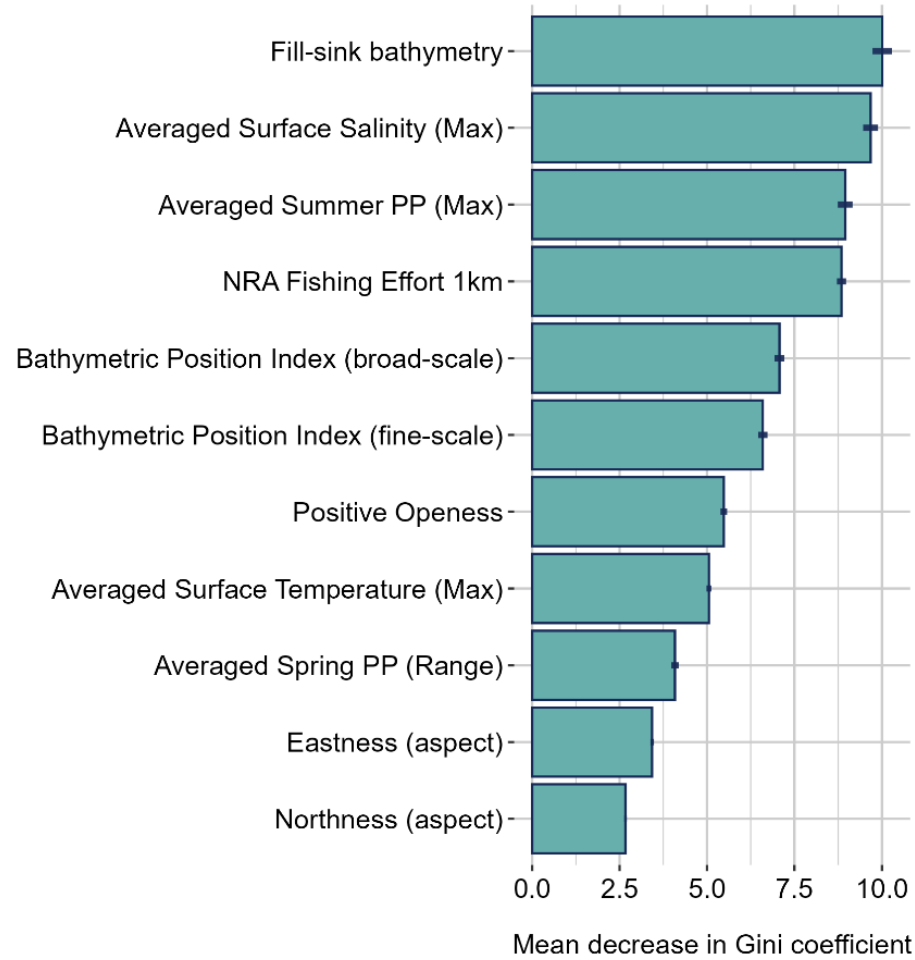
**Figure 5.** Random Forest species distribution model for the functional group Large-Sized Sponges showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated as the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.

### ***Sponge Grounds***

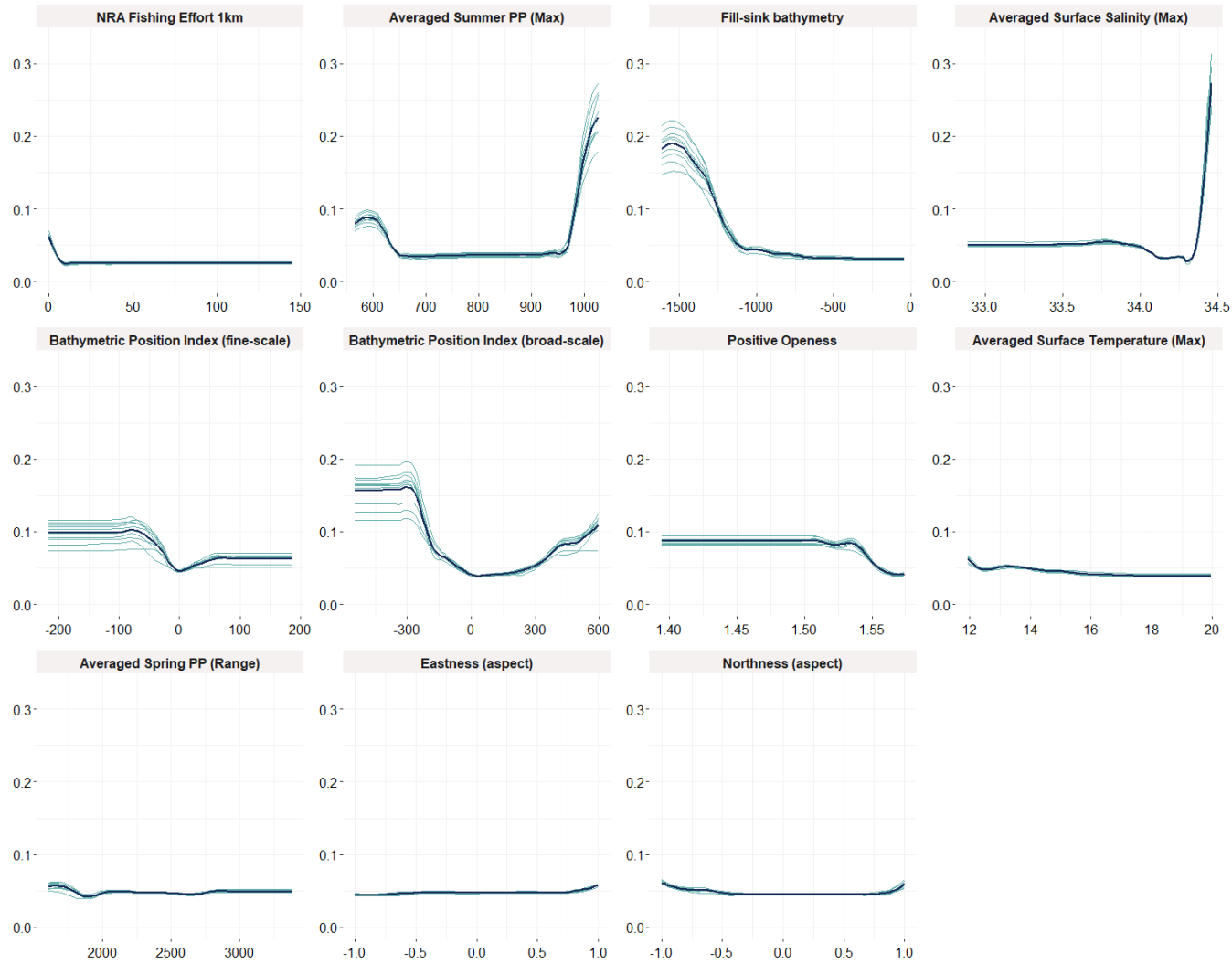
The most important variables for the Sponge Grounds were the fill-sink bathymetry (Depth), averaged maximum value of surface salinity, the averaged maximum value of summer primary productivity, the bottom trawl fishing effort in the NRA (1 km resolution), and the broad-scale bathymetric position index (Figure 6). The models indicate that the Sponge Grounds are typically located in depressed areas at depths > 1000 m, with maximum values of surface salinity > 34.3 ‰, maximum values of summer primary productivity > 950 mg C m<sup>-2</sup> day<sup>-1</sup>, and low bottom trawl fishing effort (< 10 km<sup>2</sup>/km<sup>2</sup>/year) (Figure 7).

The predicted distribution maps are presented in Figure 8 as binary plots indicating presence/absence based on two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The data distribution is shown overlain on the binary map of presence/absence based on Sensitivity=Specificity in Figure 9. Outside the model extrapolation areas, the Sponge Grounds are distributed on East Flemish Cap, the southern part of Flemish Pass, and the Tail and canyons of the Grand Bank of Newfoundland.

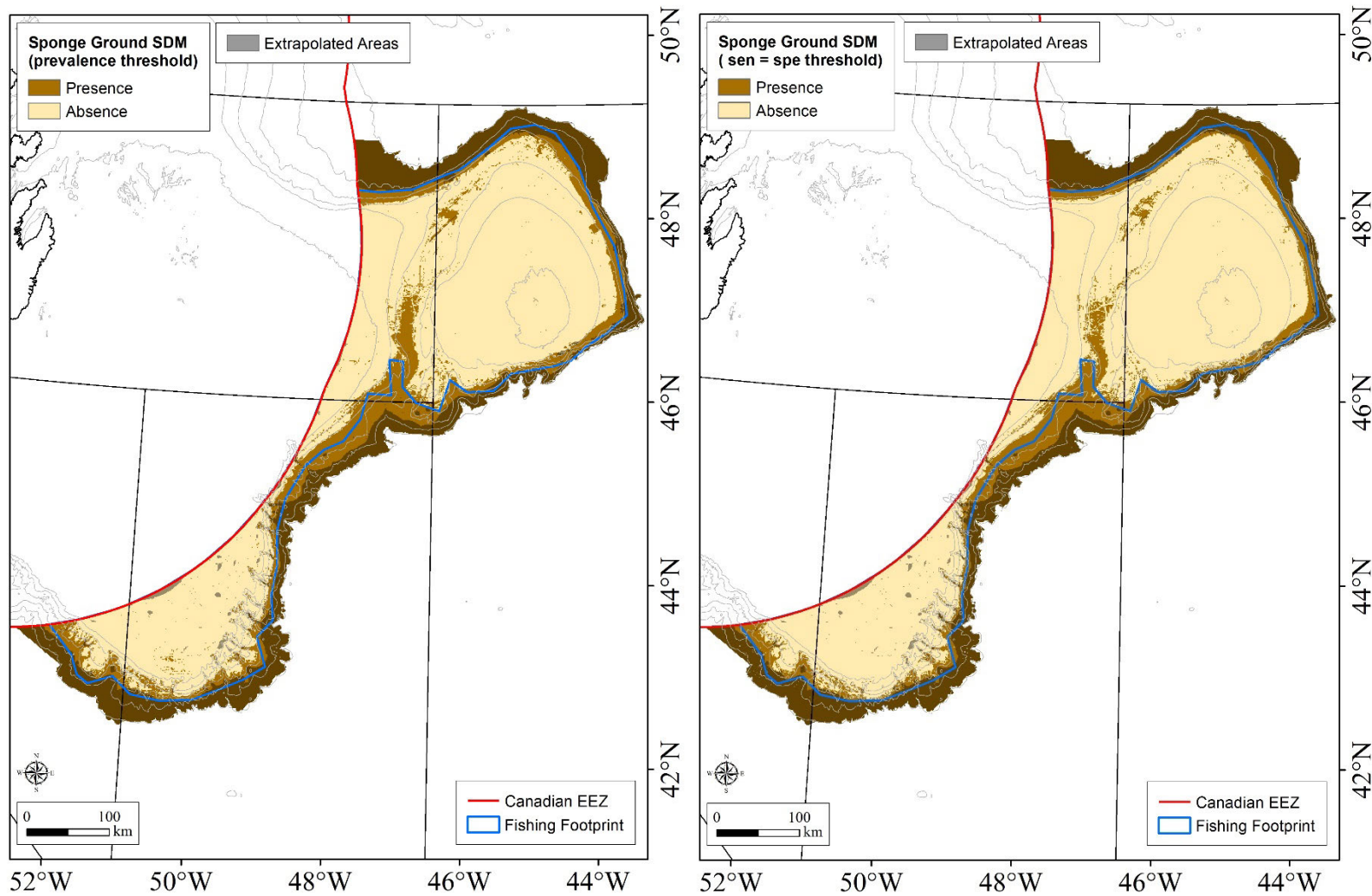
The uncertainty expressed as the frequency of P/A from the 10 cross-validation runs (Figure 9), the areas of extrapolation (Figures 8-10) and the average probability of the maximum frequency class (Figure 10) indicated high certainty within the fishing footprint for both presence and absence predictions. However, there was increased uncertainty in the deeper slope waters both in areas of interpolation and extrapolation.



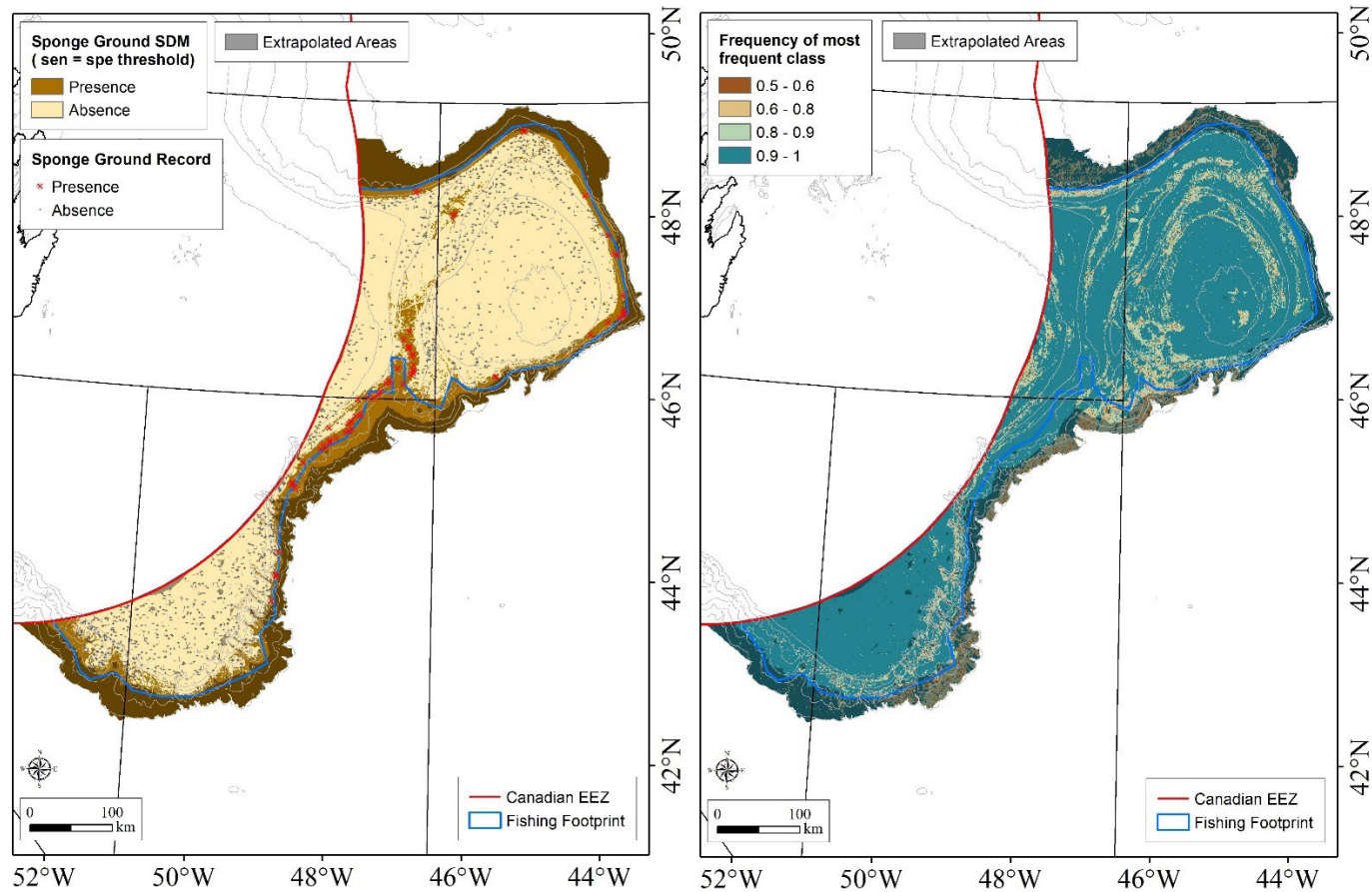
**Figure 6.** Plot of mean decrease and standard deviation in Gini Value for the 11 predictor variables in the Random Forest model for the Sponge Grounds, indicating their relative importance and variation across 10 model folds.



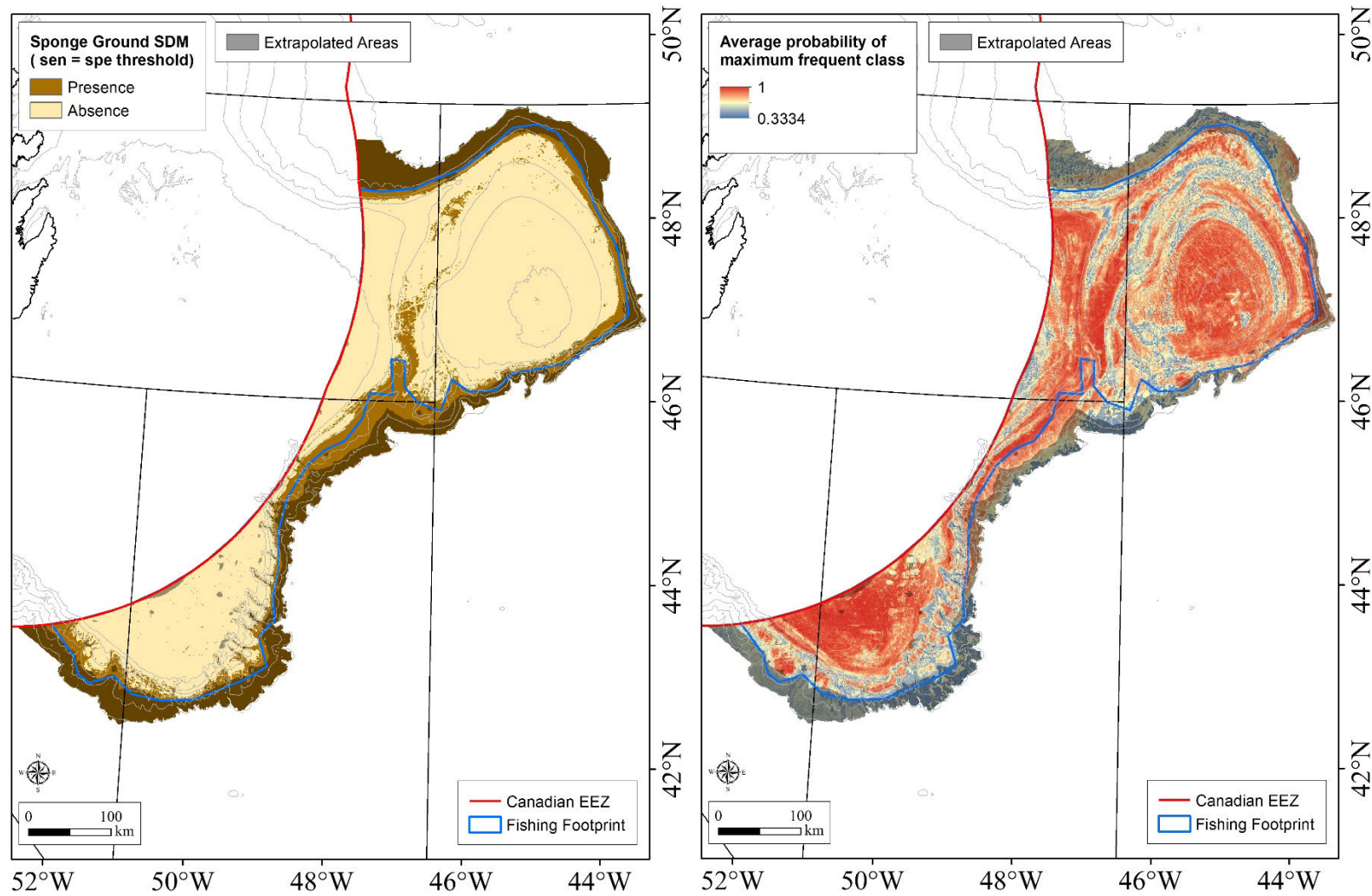
**Figure 7.** Response curves showing the partial dependence of the probability of presence on the predictors (Figure 6) identified in the Random Forest model for the Sponge Grounds. For each variable the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.



**Figure 8.** Random Forest species distribution model for the Sponge Grounds showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). The areas of extrapolation show where model predictions extend into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 9.** Random Forest species distribution model for the Sponge Grounds showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 10.** Random Forest species distribution model for the Sponge Grounds showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.

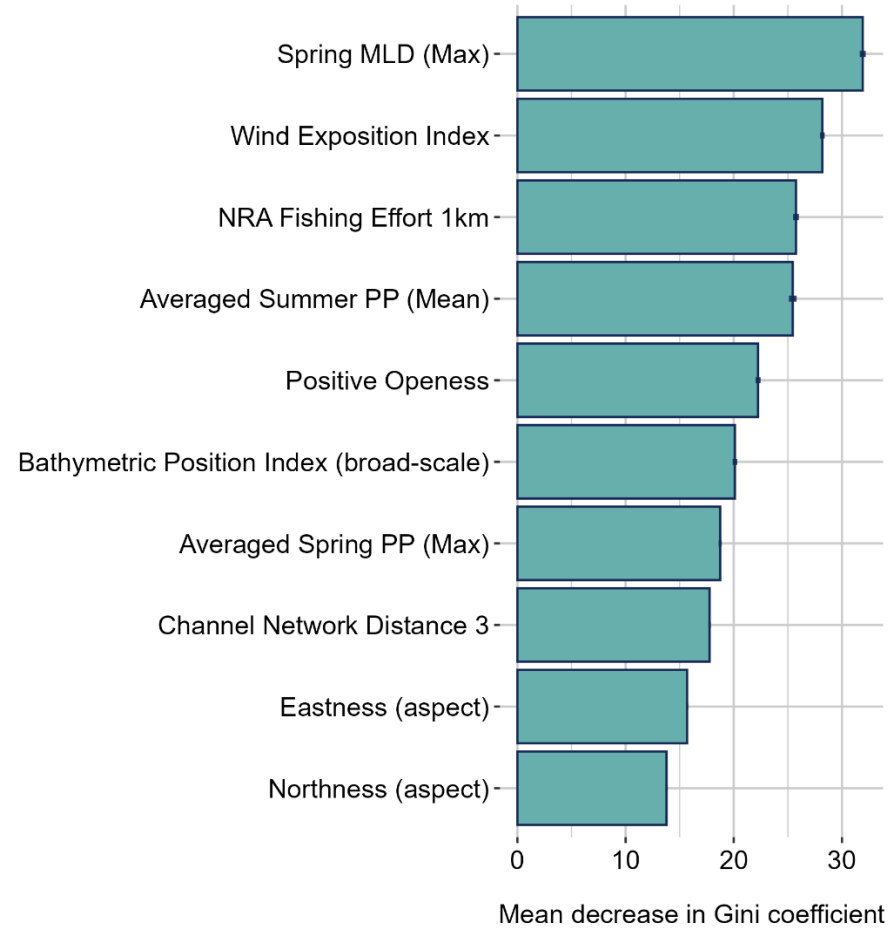
### ***Tetillidae***

The most important variables for the Tetillidae were the maximum of the mixed layer depth in the spring, the wind exposition index, the bottom trawl fishing effort in the NRA (1 km resolution), the averaged mean value of summer primary productivity, and the positive openness (Figure 11). The models indicate that the Tetillidae taxa are found in moderately sheltered areas, and with maximum mixed layer depth in the spring < 17 m (Figure 12).

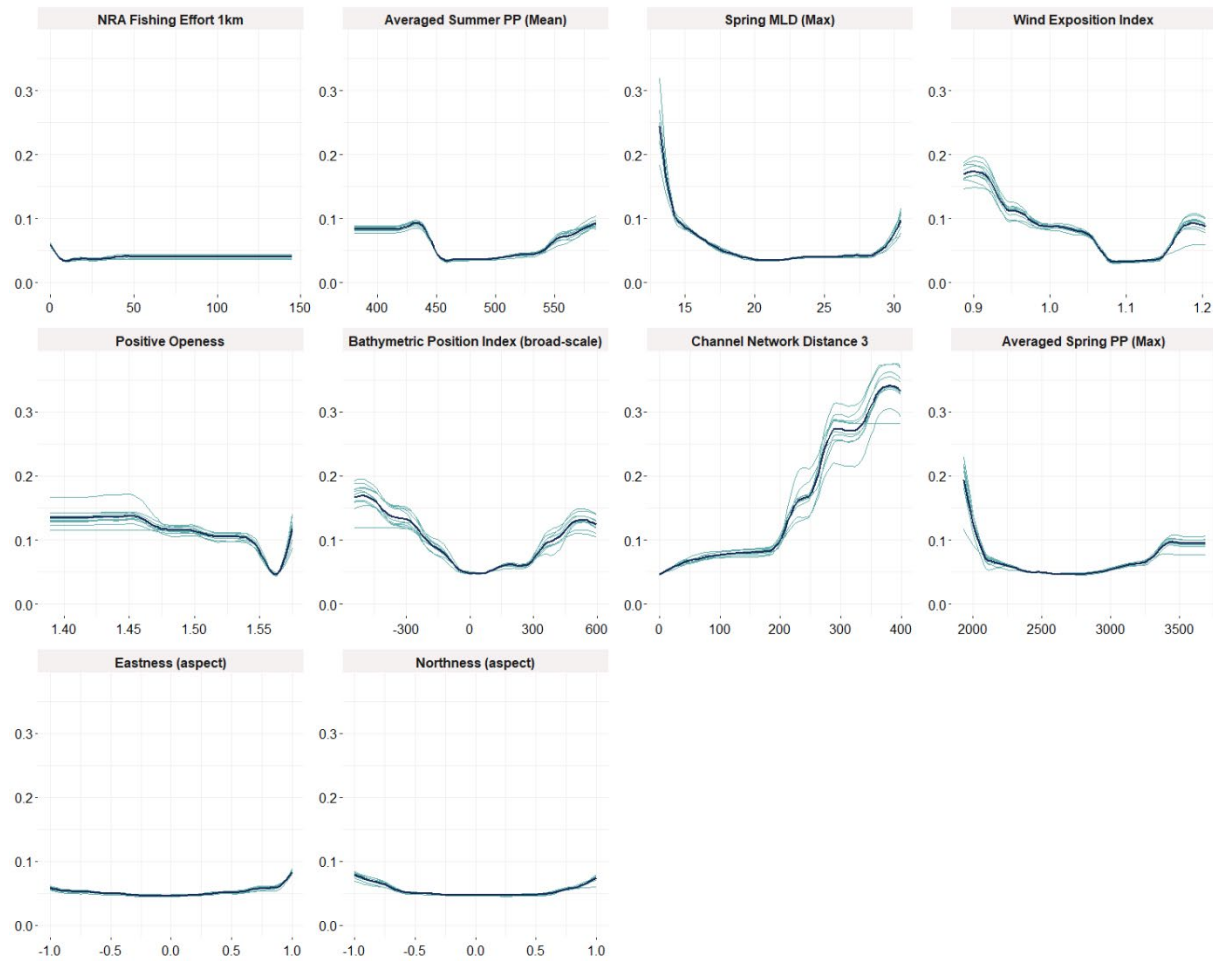
The predicted distribution maps are presented in Figure 13 as binary plots indicating presence/absence based on two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The data distribution is shown overlain on the binary map of presence/absence based on Sensitivity=Specificity in Figure 14. Outside the model extrapolation areas, the Tetillidae group is distributed on the South Flemish Cap, on Flemish Pass, and the canyons of Grand Bank of Newfoundland.

The uncertainty expressed as the frequency of P/A from the 10 cross-validation runs (Figure 14), the areas of extrapolation (Figures 13-15) and the average probability of the maximum frequency class (Figure 15) indicated increased uncertainty in the deeper slope waters and in areas of transition between the presence and absence classes. The average probability of the maximum frequency class was lower over areas of predicted presence in some areas (Figure 15).

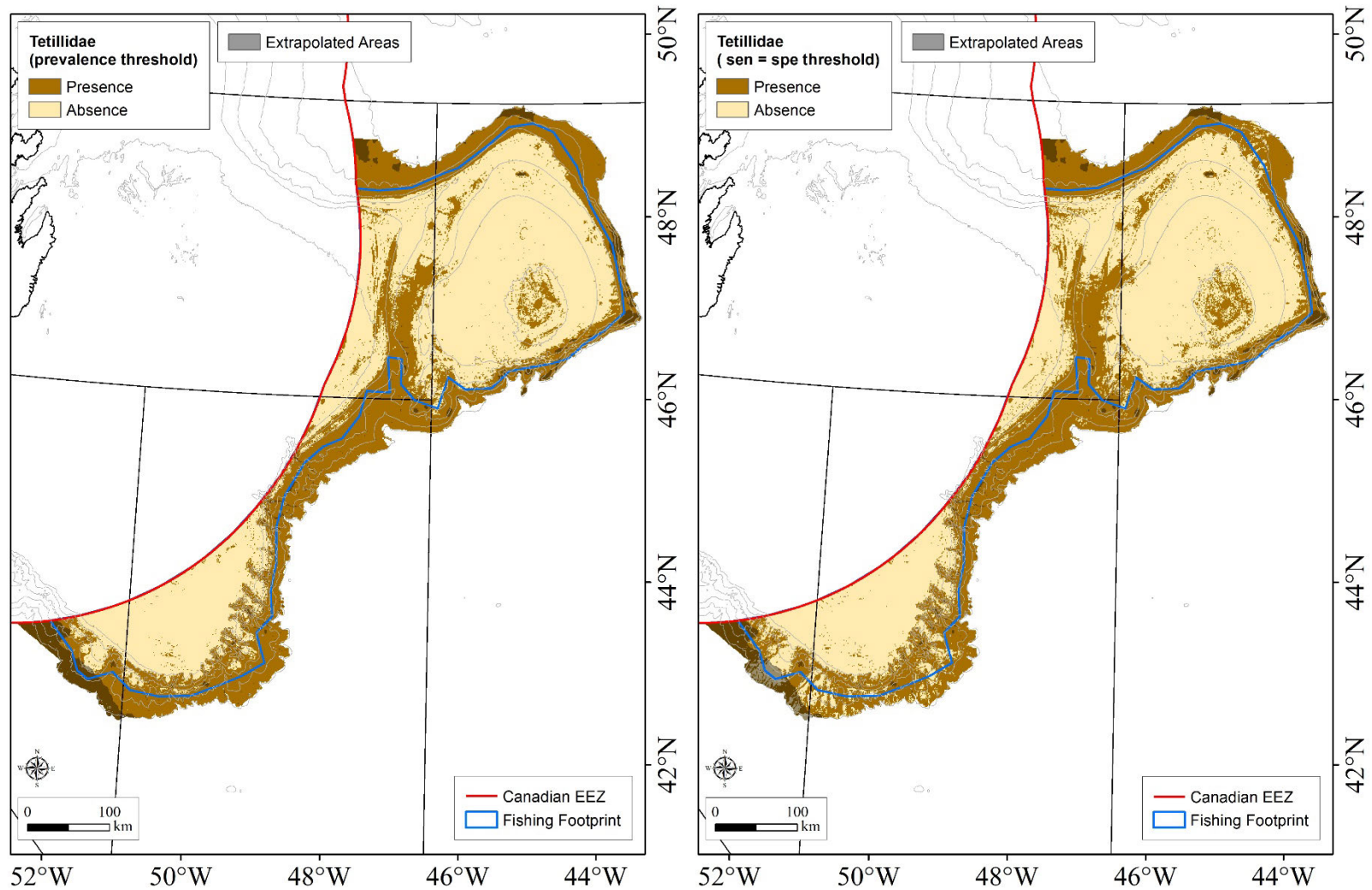




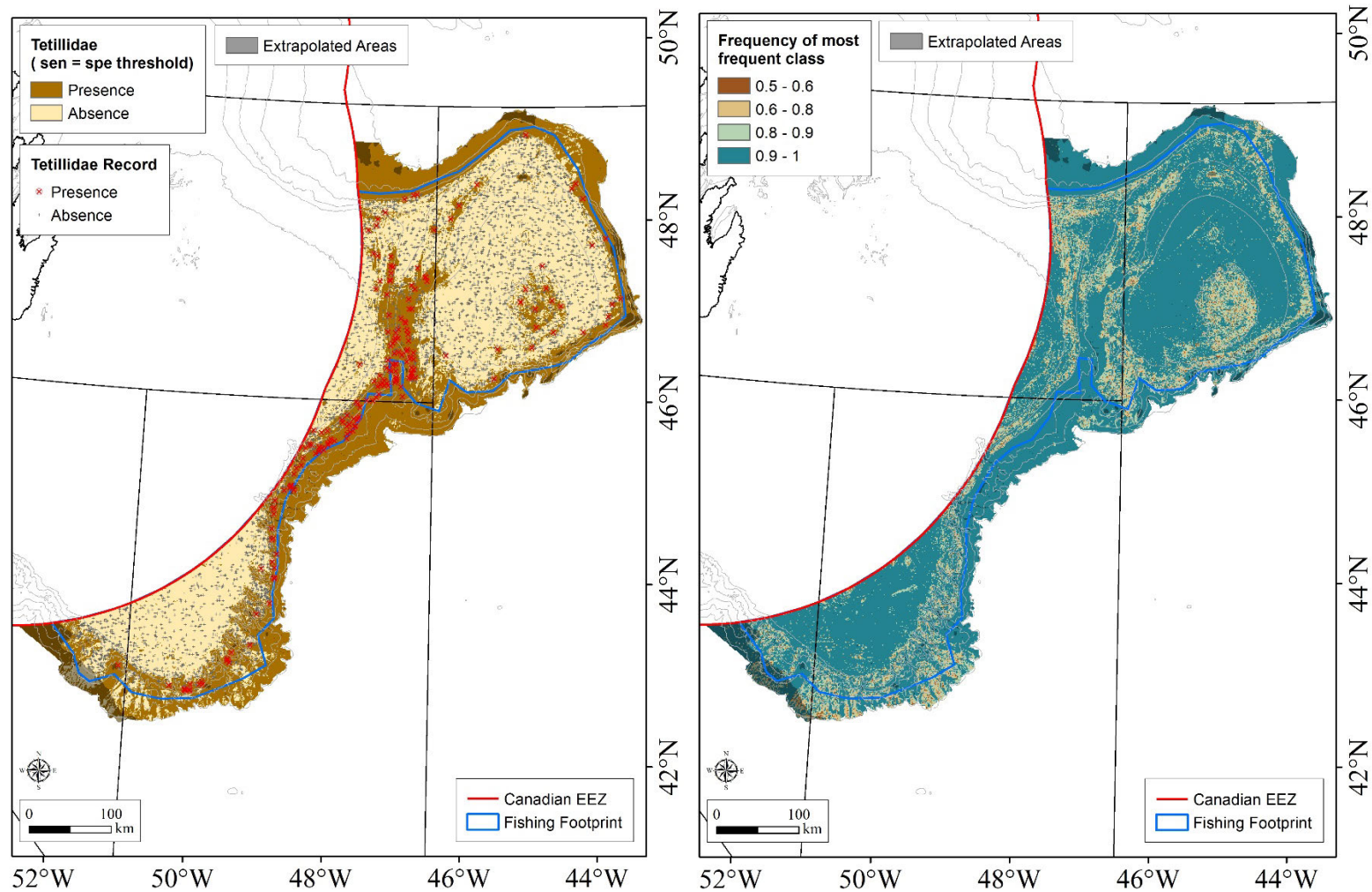
**Figure 11.** Plot of mean decrease and standard deviation in Gini Value for the 10 variables in the Random Forest model for the Tetillidae, indicating their relative importance and variation across 10 data folds.



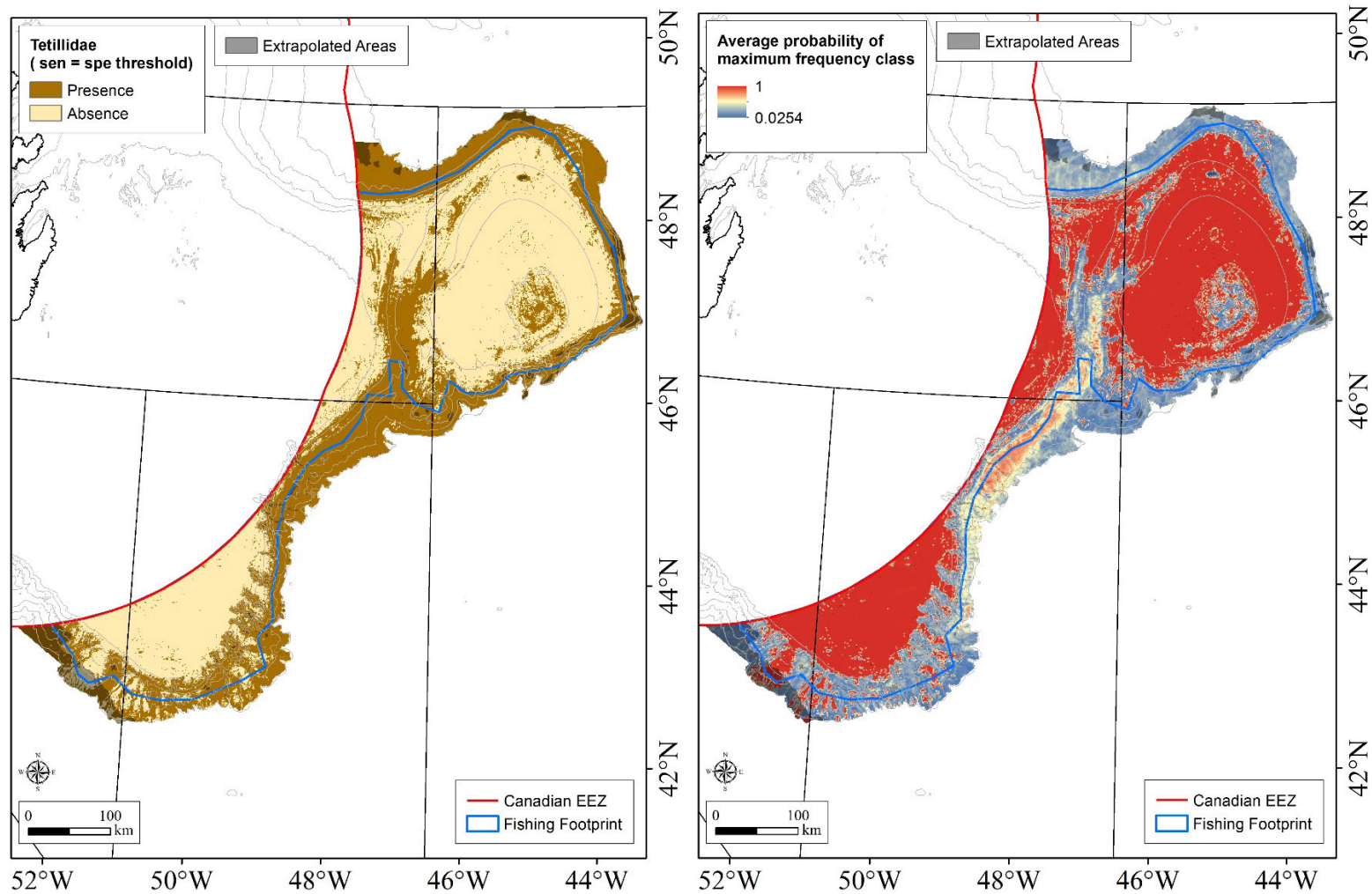
**Figure 12.** Response curves showing the partial dependence of the probability of presence on the predictors (Figure 11) identified in the Random Forest model for the Tetillidae. For each variable, the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.



**Figure 13.** Random Forest species distribution model for the Tetillidae showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 14.** Random Forest species distribution model for the Tetillidae showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



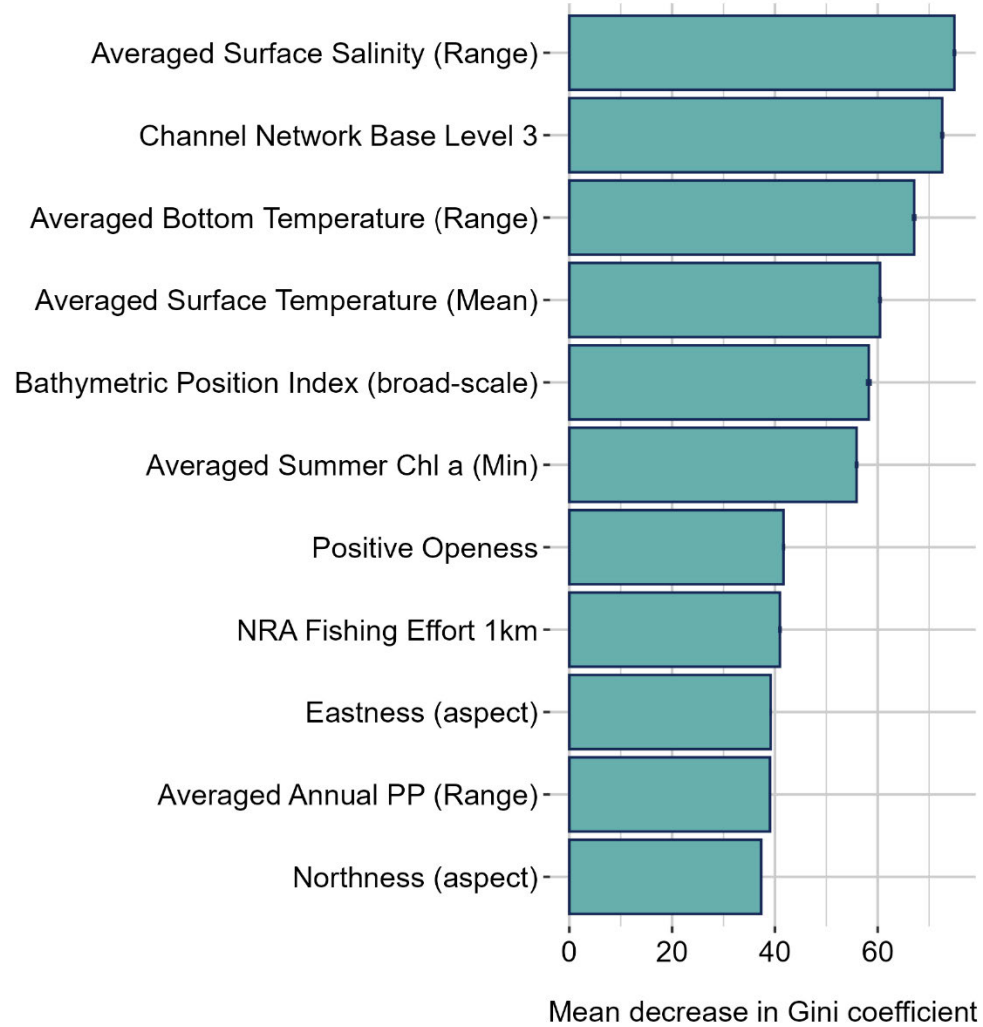
**Figure 15.** Random Forest species distribution model for the VME functional group Tetillidae showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.

***Polymastiidae***

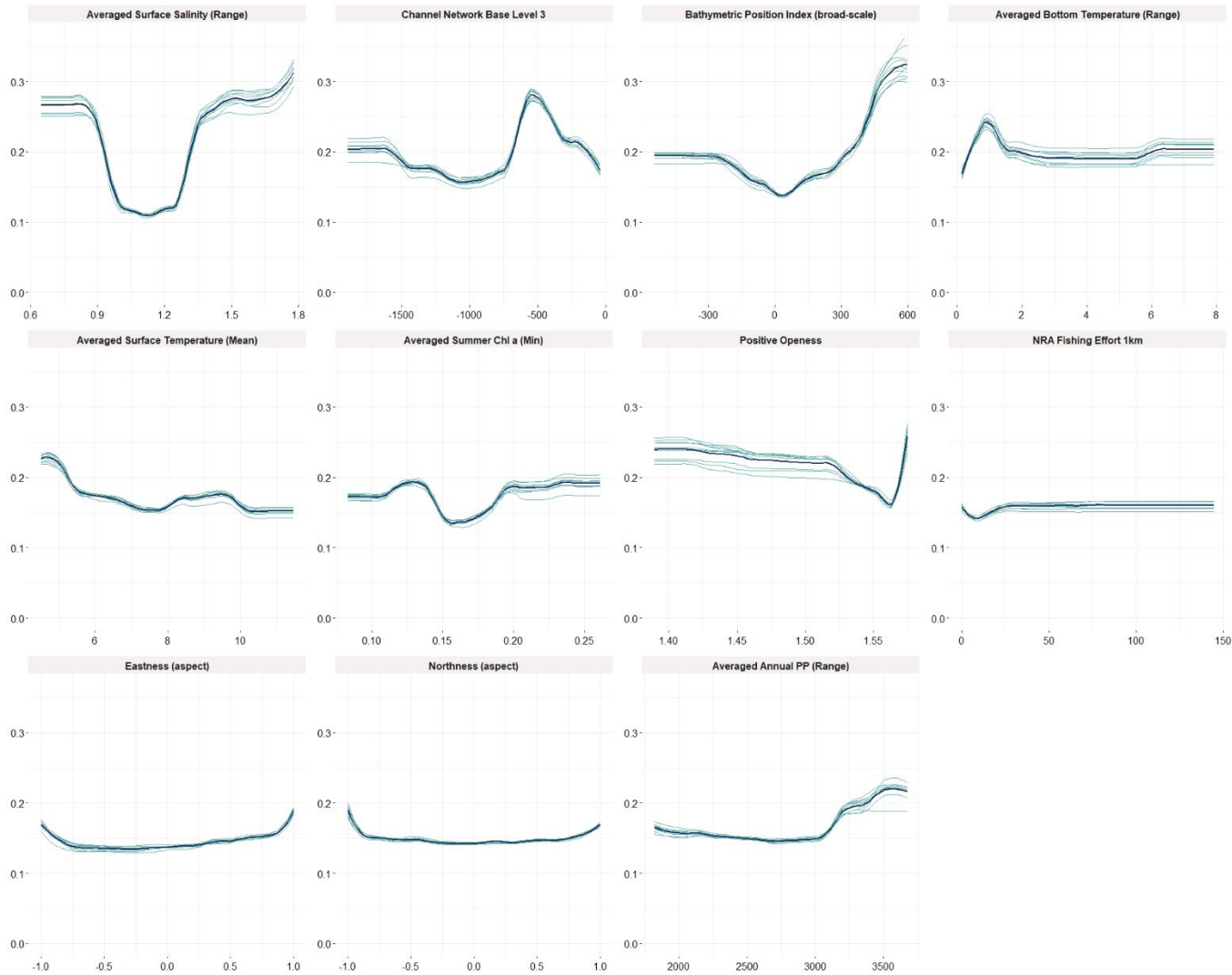
The most important variables for the *Polymastiidae* group were the range of the surface salinity, the Channel Network Base Level 3, the range of the bottom temperature, the mean value of surface temperature, and the broad-scale bathymetric position index (Figure 16). The models indicate that the *Polymastiidae* group are found in elevated areas with moderate changes of surface salinity, with mean surface temperatures < 6°C and relatively stable bottom temperatures (Figure 17).

The predicted distribution maps are presented in Figure 18 as binary plots indicating presence/absence based on two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The data distribution is shown overlain on the binary map of presence/absence based on Sensitivity=Specificity in Figure 19. Outside the model extrapolation areas, the *Polymastiidae* group are distributed on the Flemish Cap, Flemish Pass, and the southeastern region of the Grand Bank of Newfoundland.

The uncertainty expressed as the frequency of P/A from the 10 cross-validation runs (Figure 19), the areas of extrapolation (Figures 18-20) and the average probability of the maximum frequency class (Figure 20) indicated uncertainty for much of the area of predicted presence in the latter indicator (Figure 20), but not the former (Figure 19).

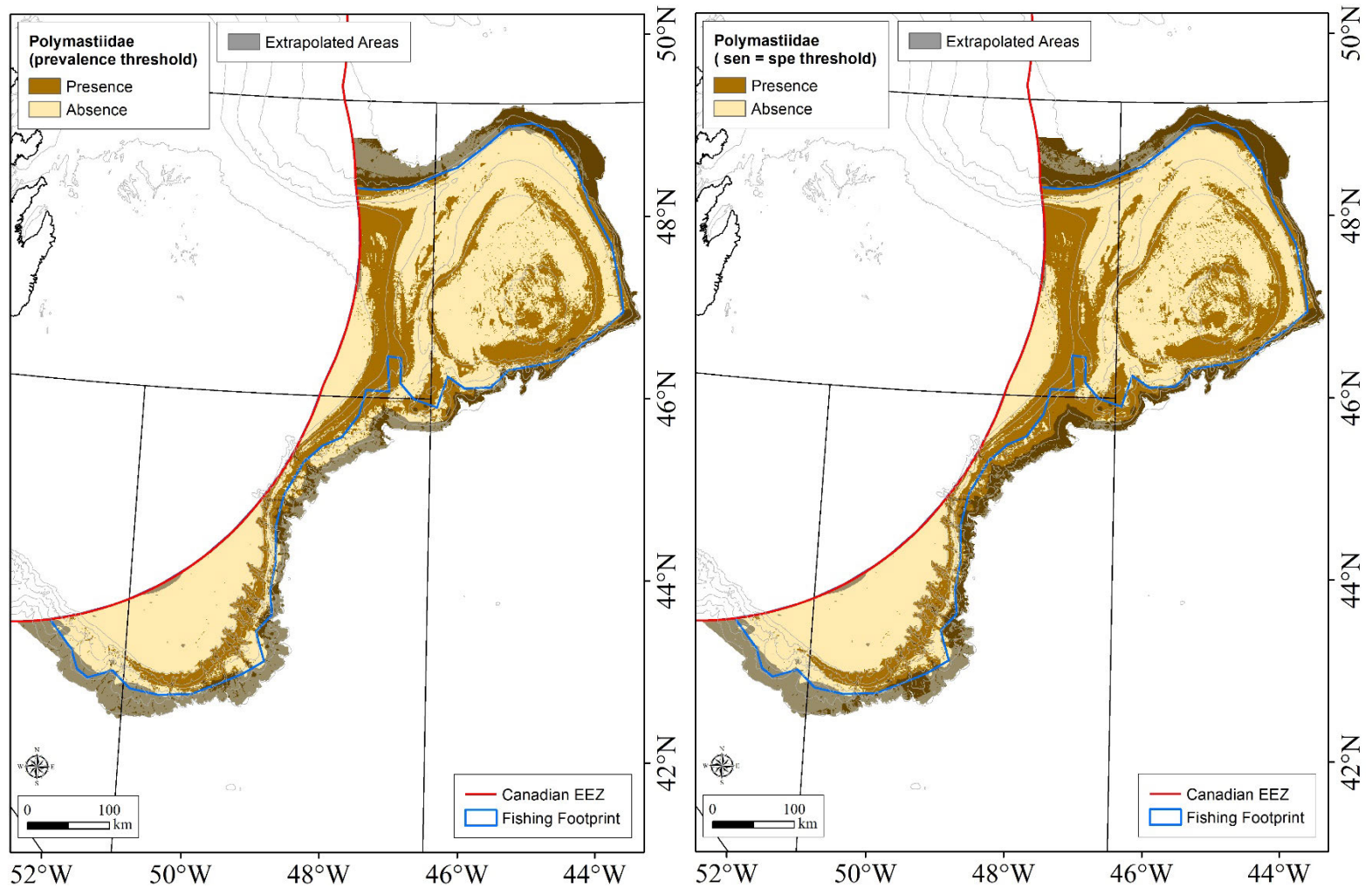


**Figure 16.** Plot of mean decrease and standard deviation in Gini Value for the 11 variables in the Random Forest model for the Polymastiidae, indicating their relative importance and variation across 10 data folds.

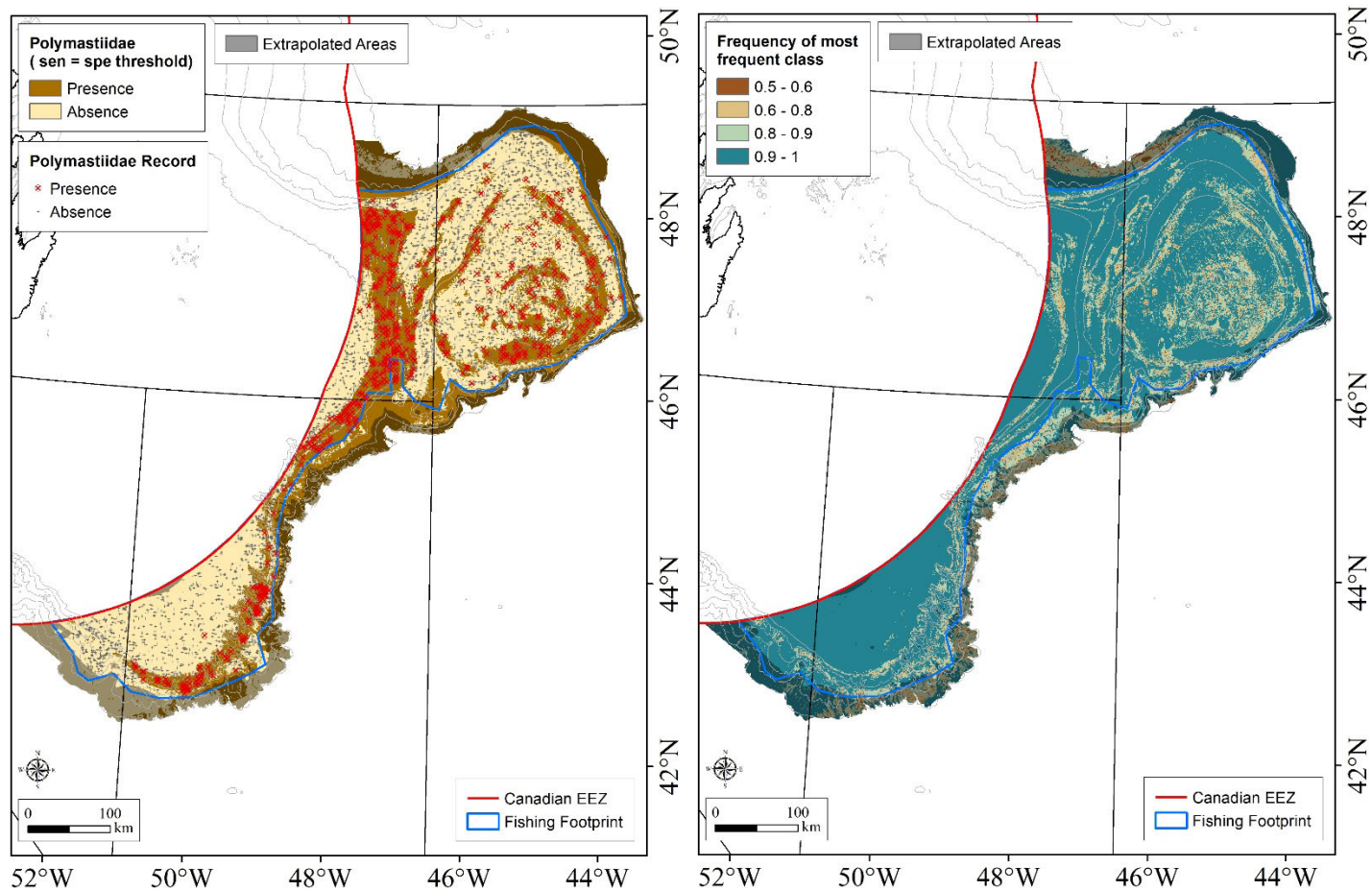


**Figure 17.** Response curves showing the partial dependence of the probability of presence on the predictors (Figure 16) identified in the Random Forest model for the Polymastiidae. For each variable, the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.

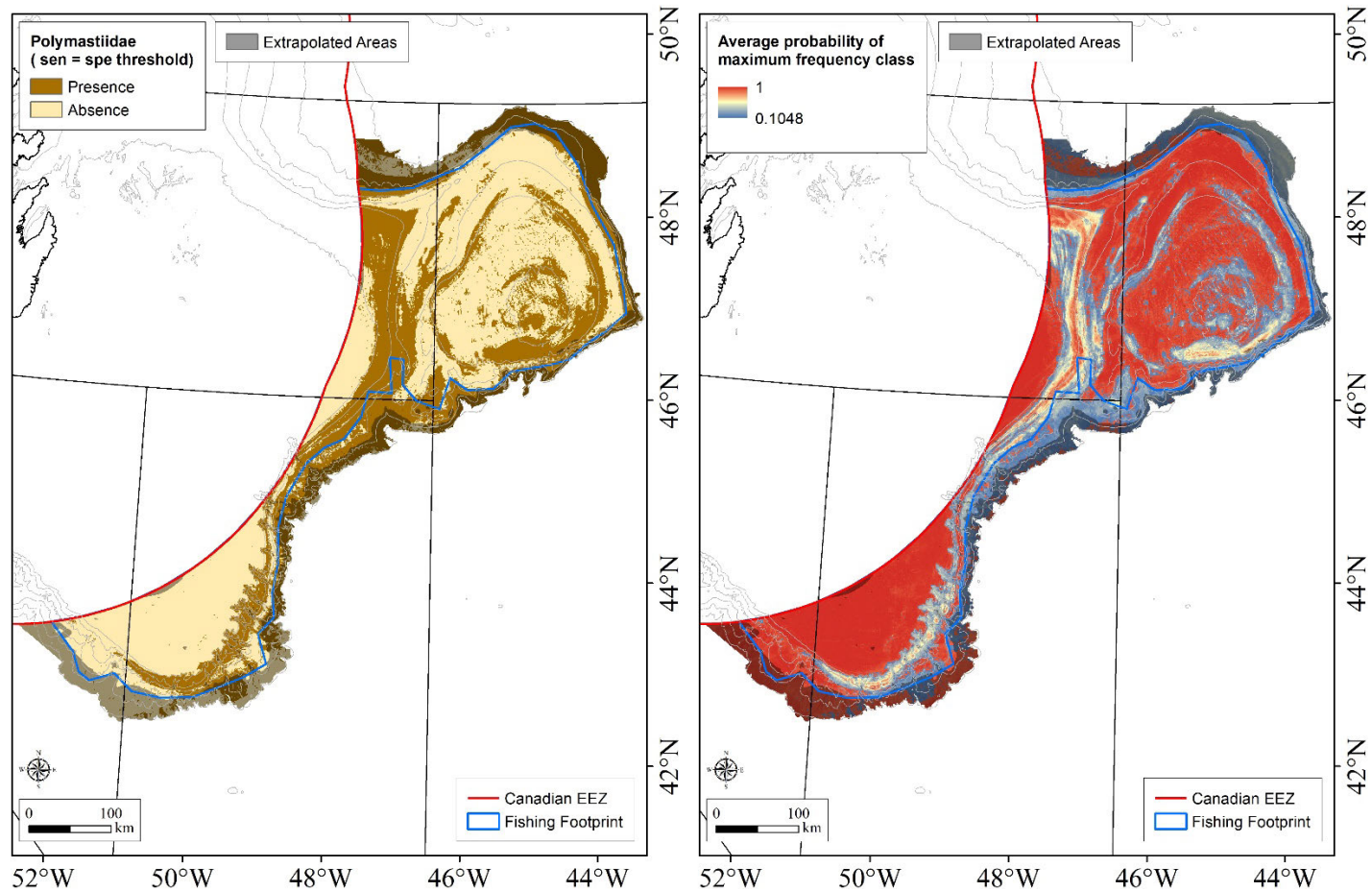




**Figure 18.** Random Forest species distribution model for the Polymastiidae showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 19.** Random Forest species distribution model for the Polymastiidae showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



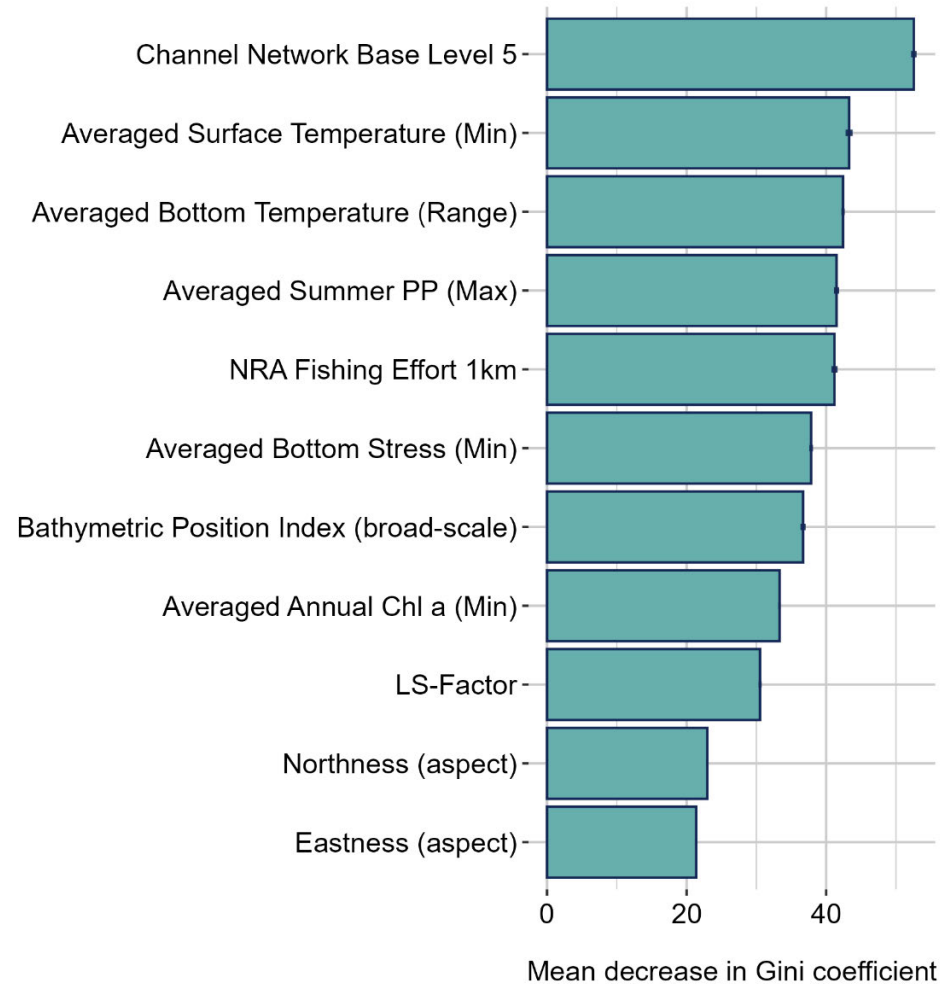
**Figure 20.** Random Forest species distribution model for the VME functional group *Polymastiidae* showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.

***Astrophorina***

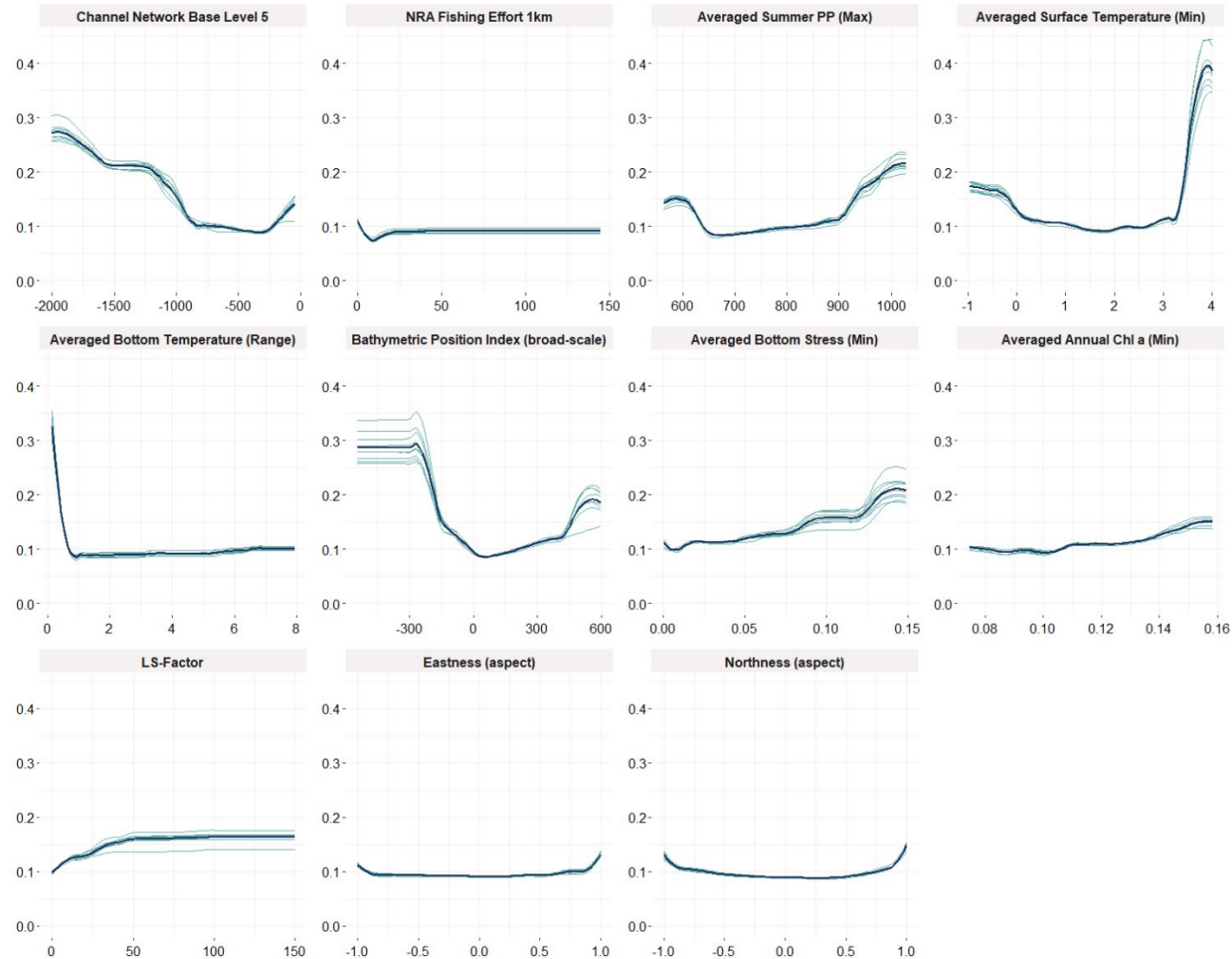
The most important variables for the *Astrophorina* group were the Channel Network Base Level 5, the minimum value of average surface temperature, the range of bottom temperature, the maximum value of the primary productivity in summer, and the bottom trawl fishing effort in the NRA (1 km resolution) (Figure 21). The models indicate that the *Astrophorina* group are found in depressed areas, with maximum values of primary productivity in summer  $> 900 \text{ mg C m}^{-2} \text{ day}^{-1}$ , minimum surface temperature  $> 4^{\circ}\text{C}$ , and stable environment of bottom temperatures (Figure 22).

The predicted distribution maps are presented in Figure 23 as binary plots indicating presence/absence based on two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The data distribution is shown overlain on the binary map of presence/absence based on Sensitivity=Specificity (Figure 24). Outside the model extrapolation areas, the *Astrophorina* group are distributed on the Flemish Cap, Flemish Pass, and the flanks of the Grand Bank of Newfoundland.

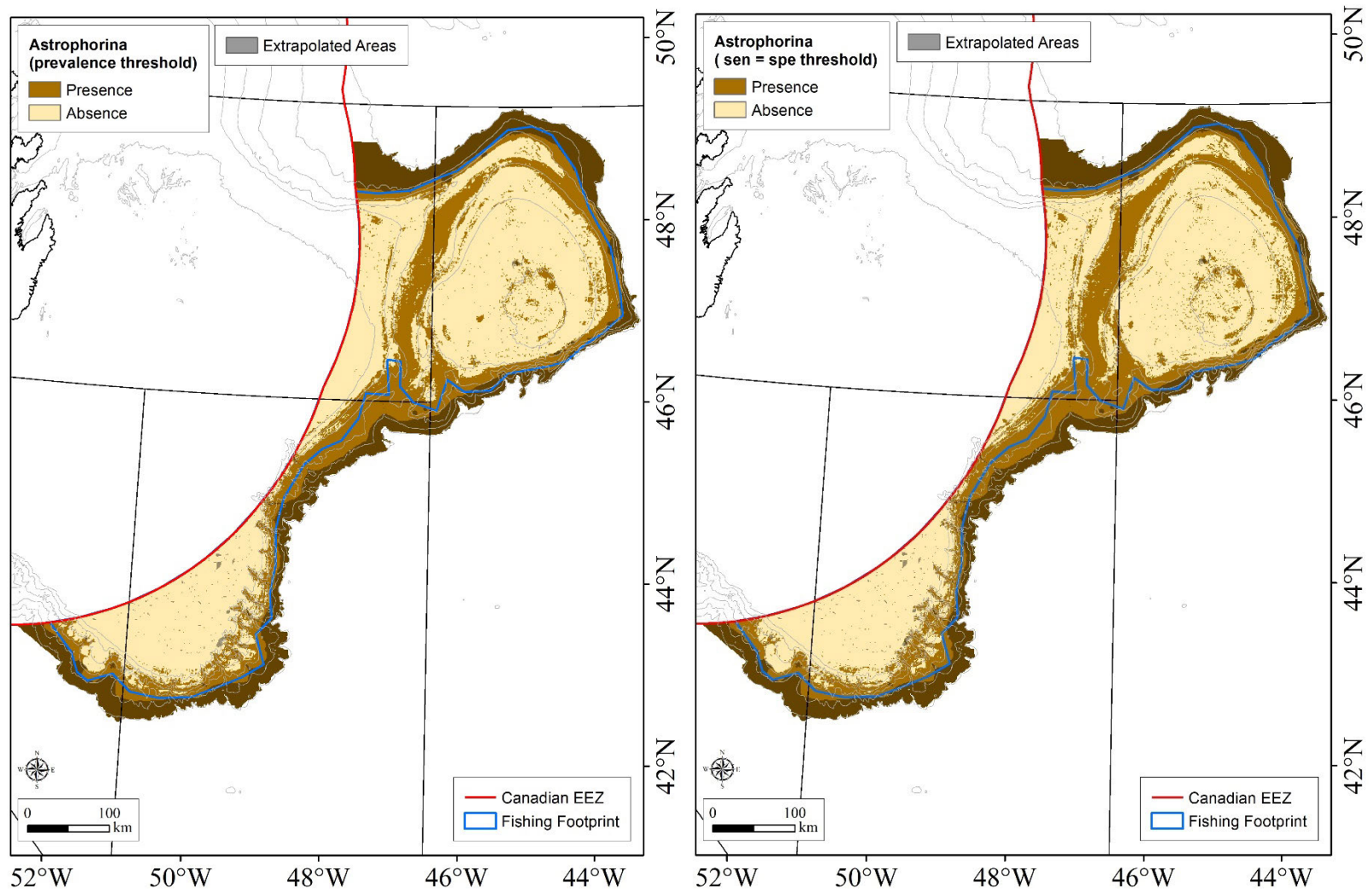
The uncertainty expressed as the frequency of P/A from the 10 cross-validation runs (Figure 24), the areas of extrapolation (Figures 23-25) and the average probability of the maximum frequency class (Figure 25) indicated high certainty within the fishing footprint for both presence and absence predictions. However, there was increased uncertainty in the deeper slope waters both in areas of interpolation and extrapolation.



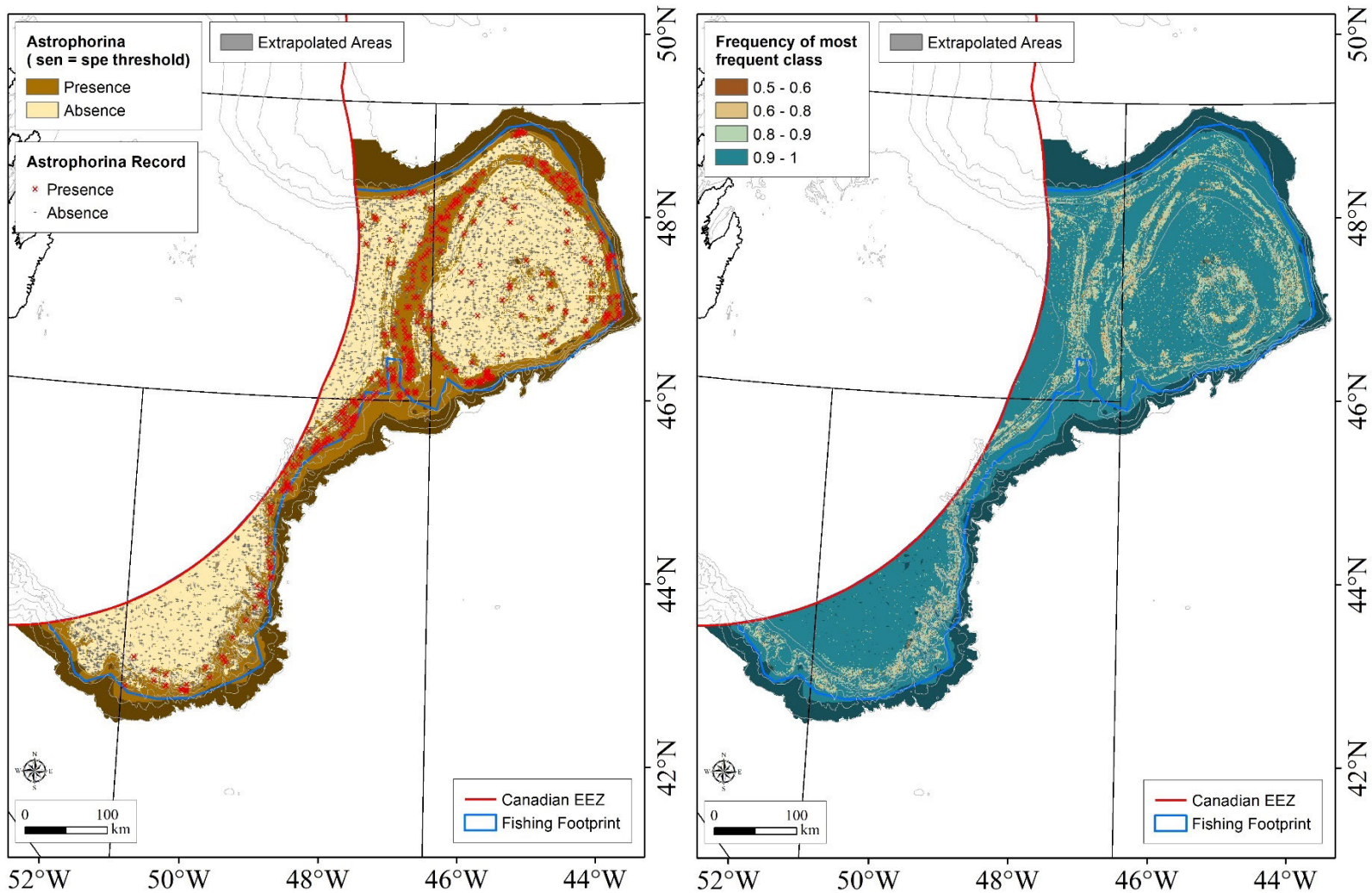
**Figure 21.** Plot of mean decrease and standard deviation in Gini Value for the 11 variables in the Random Forest model for the Astrophorina, indicating their relative importance and variation across 10 model folds.



**Figure 22.** Response curves showing the partial dependence of the probability of presence on the predictors (Figure 21) identified in the Random Forest model for the Astrophorina. For each variable, the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.

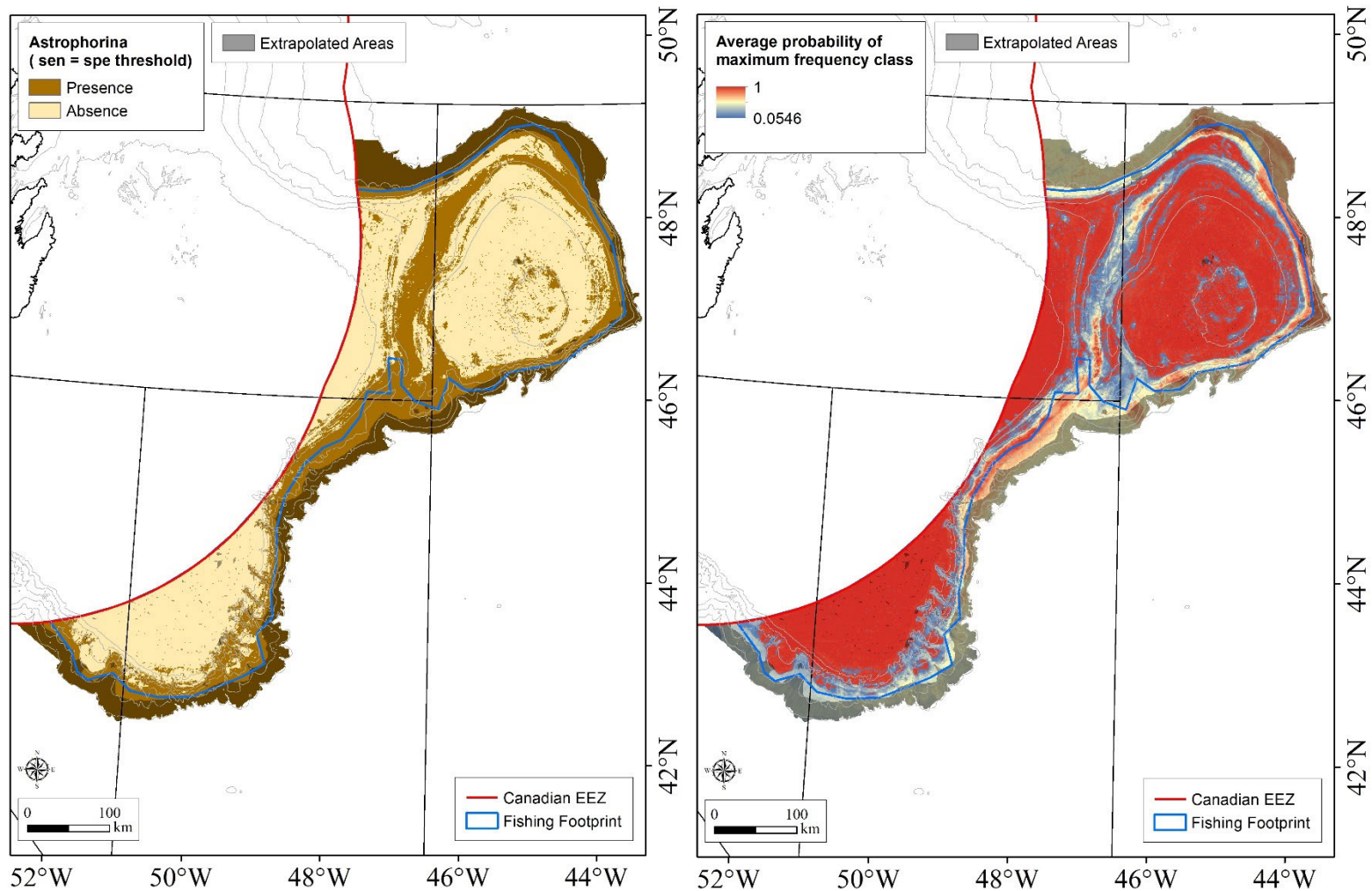


**Figure 23.** Random Forest species distribution model for the *Astrophorina* showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 24.** Random Forest species distribution model for the *Astrophorina* showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.





**Figure 25.** Random Forest species distribution model for the *Astrophorina* showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.

### Assessment and Prediction of Sea Pens

Random Forest models predicting the probability of the presence of Sea Pens generally had high accuracy scores across the validation statistics (Balanced Accuracy, Sensitivity and Specificity all  $\geq 0.79$ ; Table 6). Kappa, which measures the extent to which the agreement between observed and predicted is higher than that expected by chance alone, was 'high' ( $> 0.71$ ) for the Sea Pens functional group, 'moderate' ( $> 0.54$ ) for *Anthoptilum* spp. and 'fair' ( $> 0.31$ ) for *Balticina* spp., *Funiculina* spp. and *Pennatula* spp. The disparity in Kappa values reflects its dependence on prevalence, which for the Sea Pens functional group was close to equal (0.43), for *Anthoptilum* spp. 0.21, and for *Balticina* spp., *Funiculina* sp. and *Pennatula* spp. 0.11, 0.07, and 0.08, respectively. The TSS, defined as the average of the net prediction success rate for present sites and that for absent sites was highest at 0.72 for the Sea Pens functional group and lowest at 0.58 for *Balticina* spp., with other taxa ranging from 0.60-0.64.

**Table 6.** Model Validation Results for the Presence/Absence Random Forest Model for the Sea Pens VME Functional Group and Subgroups. TSS=True Skill Statistic (Sensitivity + Specificity - 1).

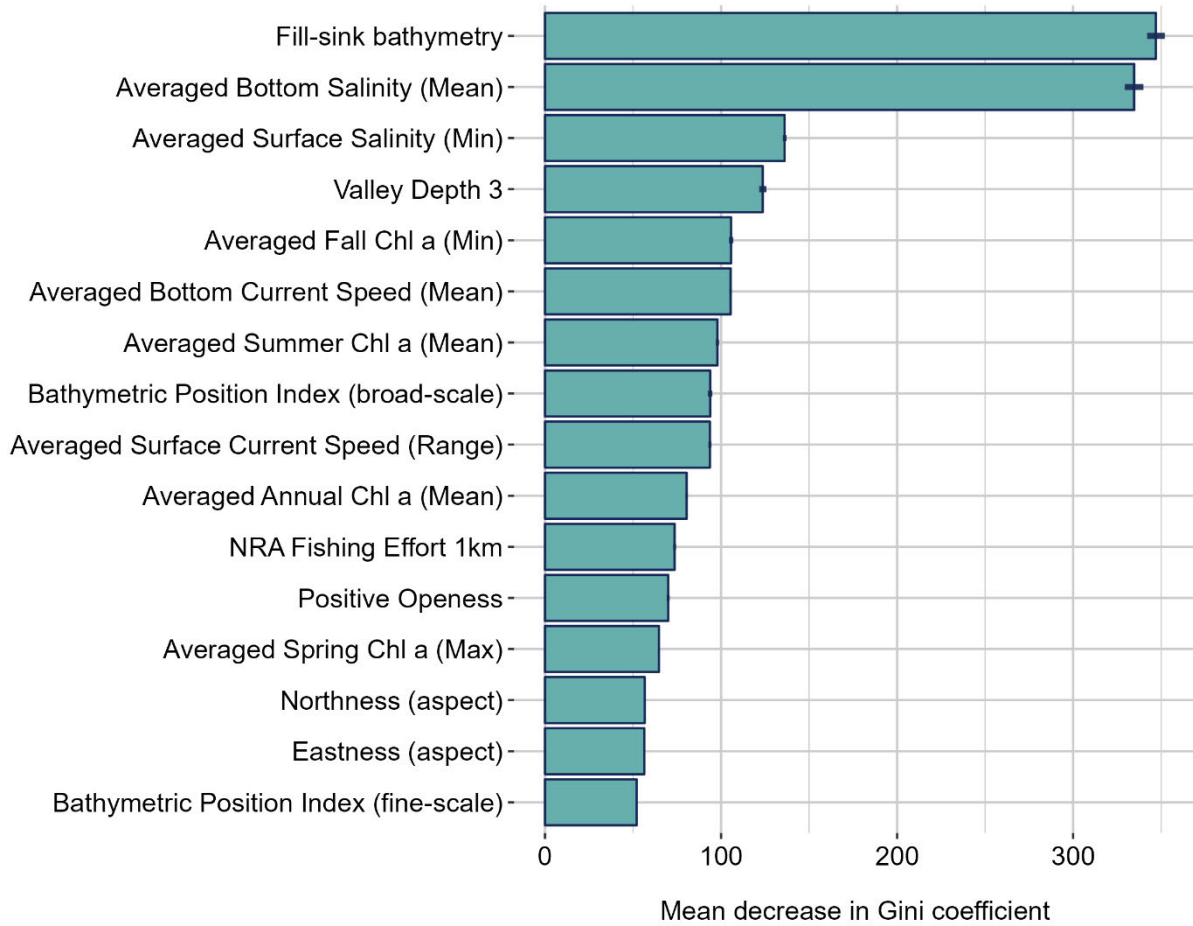
	Sea Pen Functional group	<i>Anthoptilum</i> spp.	<i>Balticina</i> spp.	<i>Funiculina</i> spp.	<i>Pennatula</i> spp.
Accuracy Measure	Mean $\pm$ SD	Mean $\pm$ SD	Mean $\pm$ SD	Mean $\pm$ SD	Mean $\pm$ SD
Sensitivity	0.86 $\pm$ 0.02	0.82 $\pm$ 0.02	0.79 $\pm$ 0.02	0.82 $\pm$ 0.04	0.80 $\pm$ 0.03
Specificity	0.86 $\pm$ 0.02	0.82 $\pm$ 0.02	0.79 $\pm$ 0.02	0.82 $\pm$ 0.03	0.80 $\pm$ 0.04
Kappa	0.71 $\pm$ 0.03	0.54 $\pm$ 0.04	0.35 $\pm$ 0.04	0.31 $\pm$ 0.05	0.31 $\pm$ 0.07
Balanced Accuracy	0.86 $\pm$ 0.02	0.82 $\pm$ 0.02	0.79 $\pm$ 0.02	0.82 $\pm$ 0.03	0.80 $\pm$ 0.04
TSS	0.72 $\pm$ 0.03	0.64 $\pm$ 0.03	0.58 $\pm$ 0.05	0.64 $\pm$ 0.06	0.60 $\pm$ 0.07

### Sea Pen Functional Group

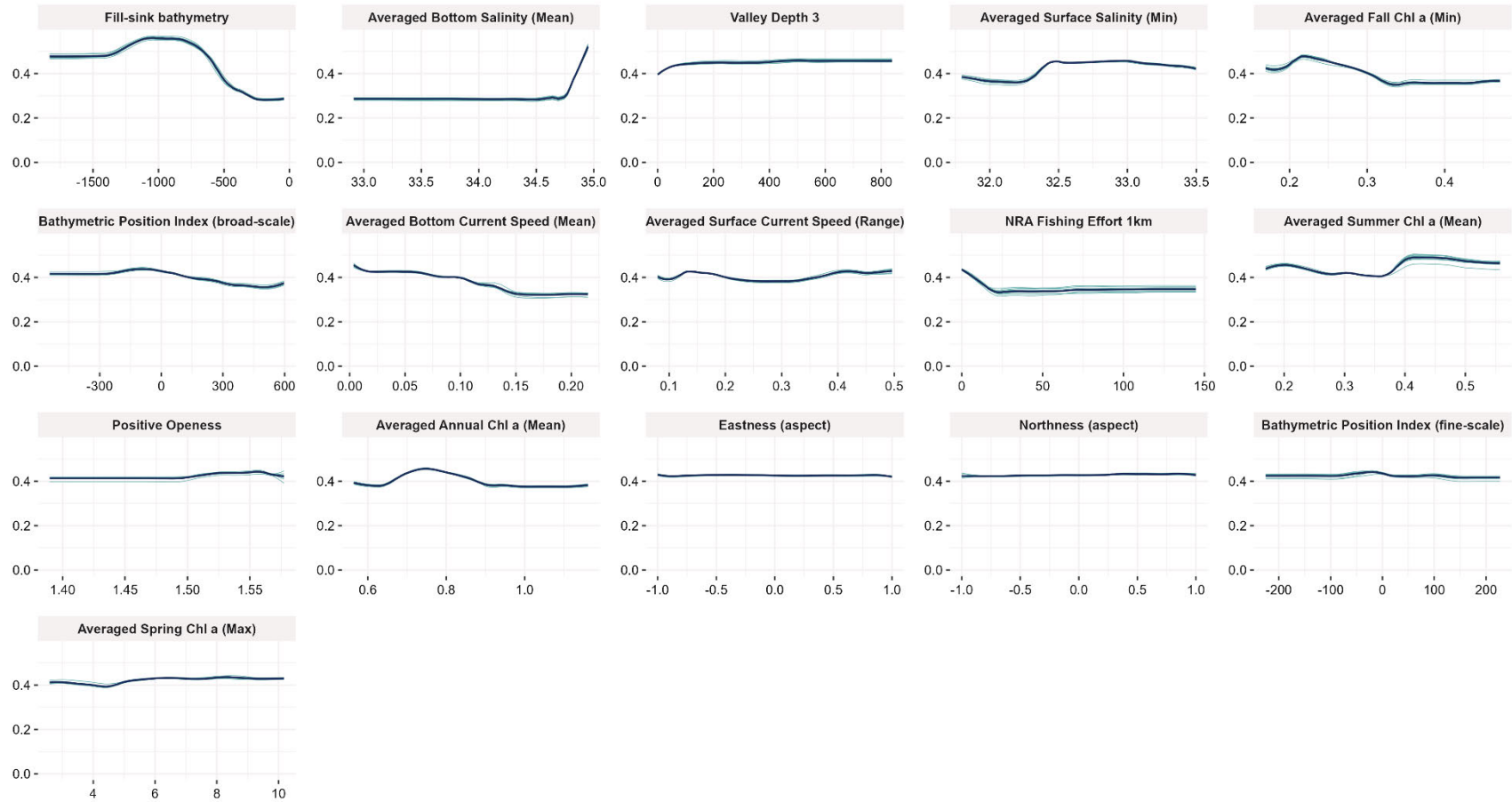
The most important variables for the Sea Pen functional group were fill-sink bathymetry, averaged mean bottom salinity, averaged minimum surface salinity and valley depth, averaged minimum fall chlorophyll concentration and averaged mean bottom current speed (Figure 26). The models indicate that the Sea Pens are typically located in depressed areas at depths less than 500 meters, with a optimum depth band around 700-1250 m depth, low bottom current speeds, mean bottom salinity  $> 34.7\text{‰}$ , average minimum surface salinity of  $32.3\text{‰}$ , and low fall chlorophyll *a* concentrations. Whilst the effect of bottom trawling effort is of lower importance in the model in comparison to the environmental conditions the probability of presence increases at lower fishing effort (Figure 27).

The predicted distribution maps are presented in Figure 28 as binary plots indicating presence/absence based on two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The data distribution is shown overlain on the binary map of presence/absence based on Sensitivity=Specificity in Figure 29. Outside the model extrapolation areas, the Sea Pen functional group forms a band around the West, North and East of the Flemish Cap, and the edge of the Tail of the Grand Bank.

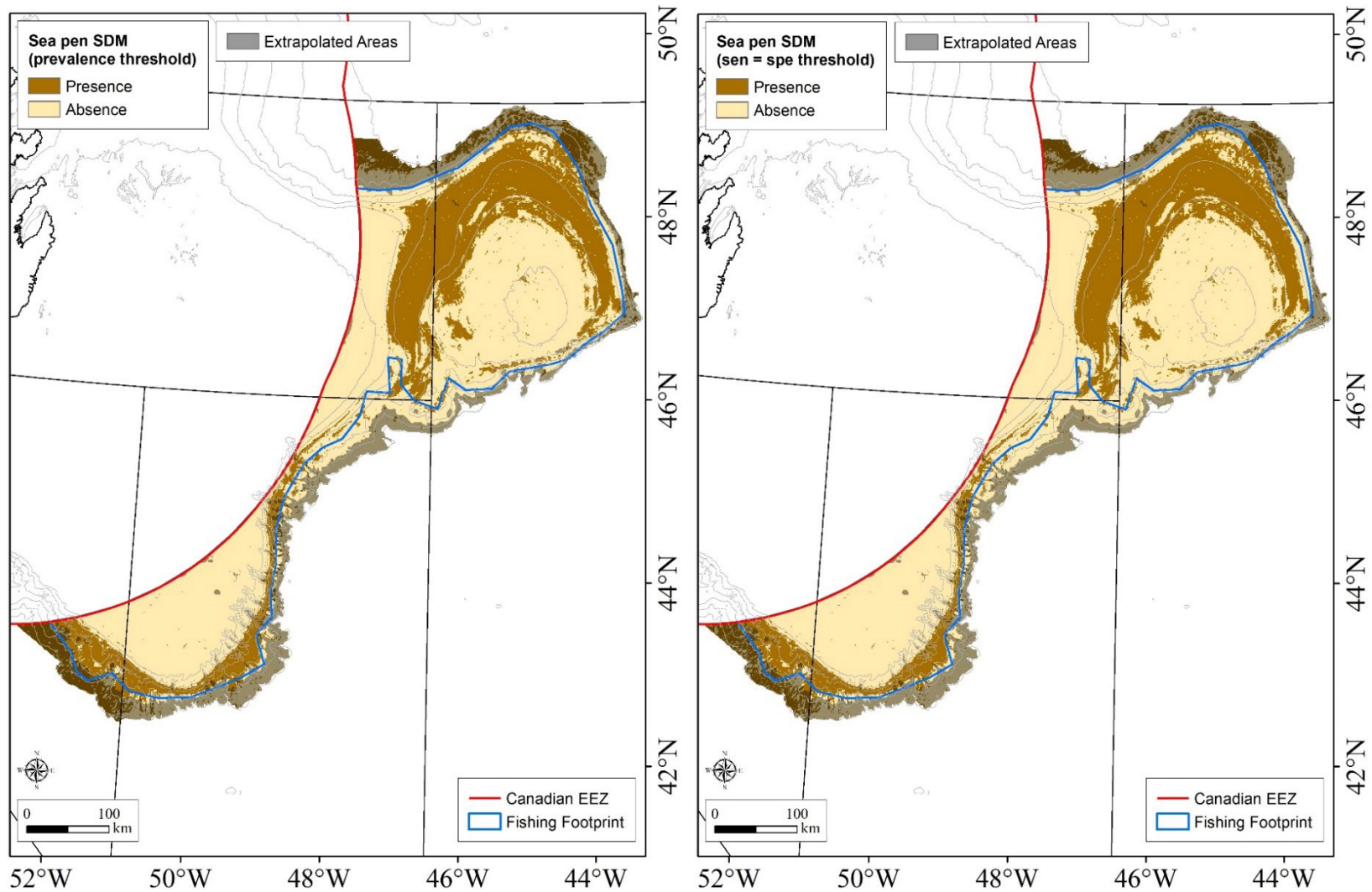
The uncertainty expressed as the frequency of P/A from the 10 cross-validation runs (Figure 29), the areas of extrapolation (Figures 28-30) and the average probability of the maximum frequency class (Figure 30) indicated high certainty within the fishing footprint for both presence and absence predictions. However, there was increased uncertainty in the deeper slope waters both in areas of interpolation and extrapolation (Figure 30).



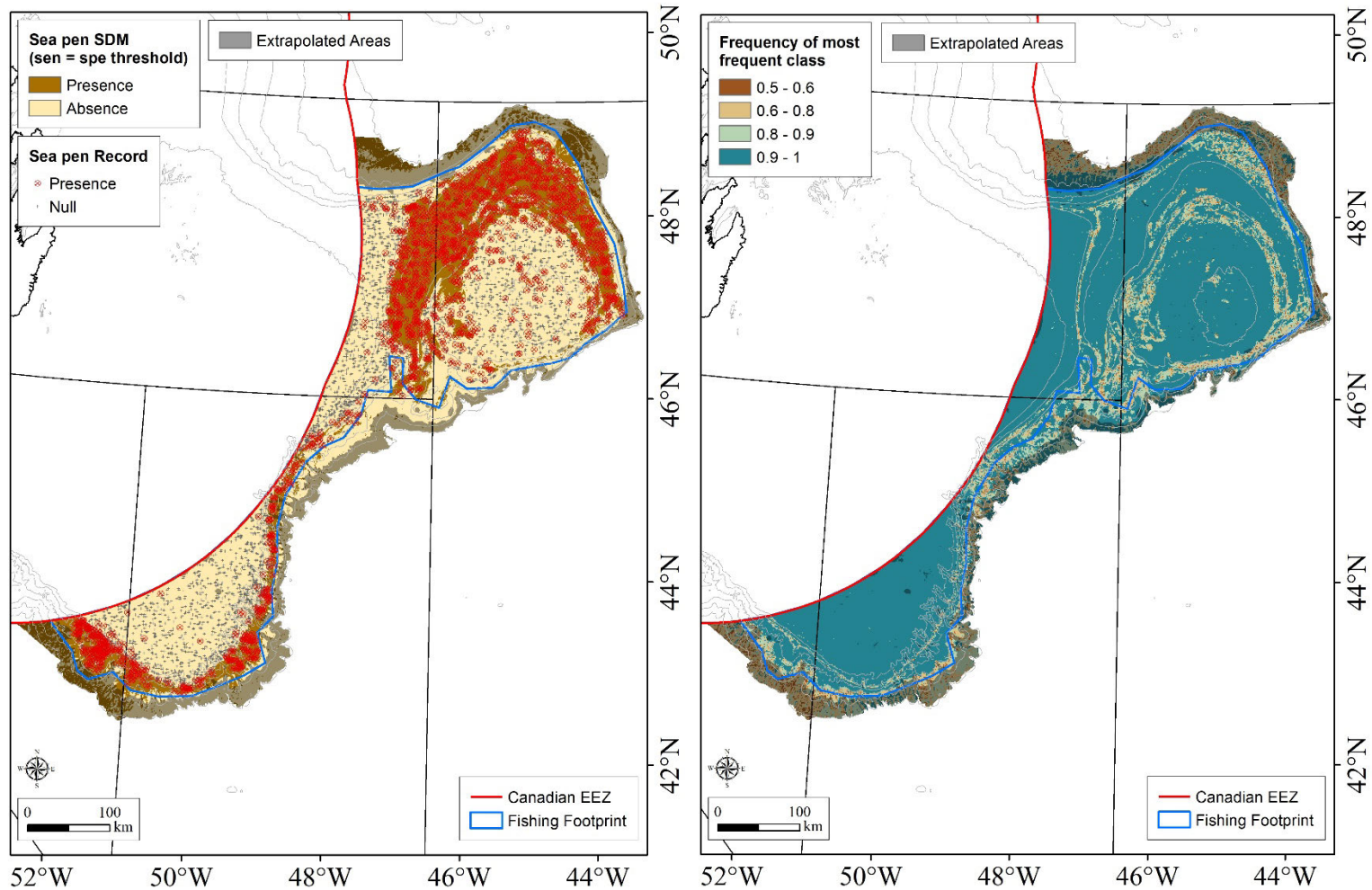
**Figure 26.** Plot of mean and standard deviation showing decrease in Gini Value for the variables in the Random Forest model for the Sea Pen VME functional group, indicating their relative importance and variation across 10 data folds.



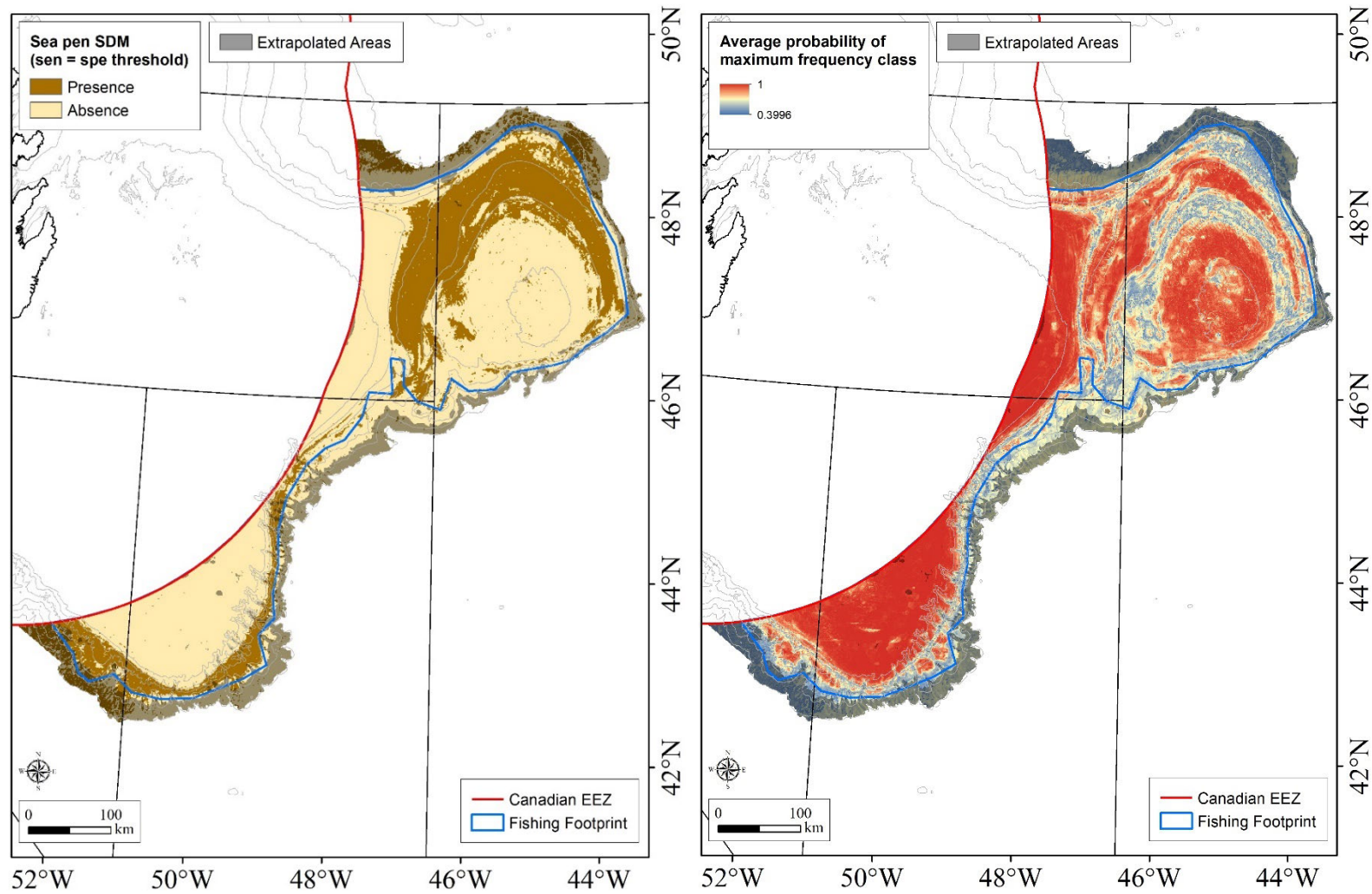
**Figure 27.** Response curves showing the partial dependence of the probability of presence on the predictors (Figure 26) identified in the Random Forest model for the Sea Pen VME functional group. For each variable, the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.



**Figure 28.** Random Forest species distribution model for the Sea Pen VME functional group showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). Areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 29.** Random Forest species distribution model for the Sea Pen VME functional group showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 30.** Random Forest species distribution model for the Sea Pen VME functional group showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.

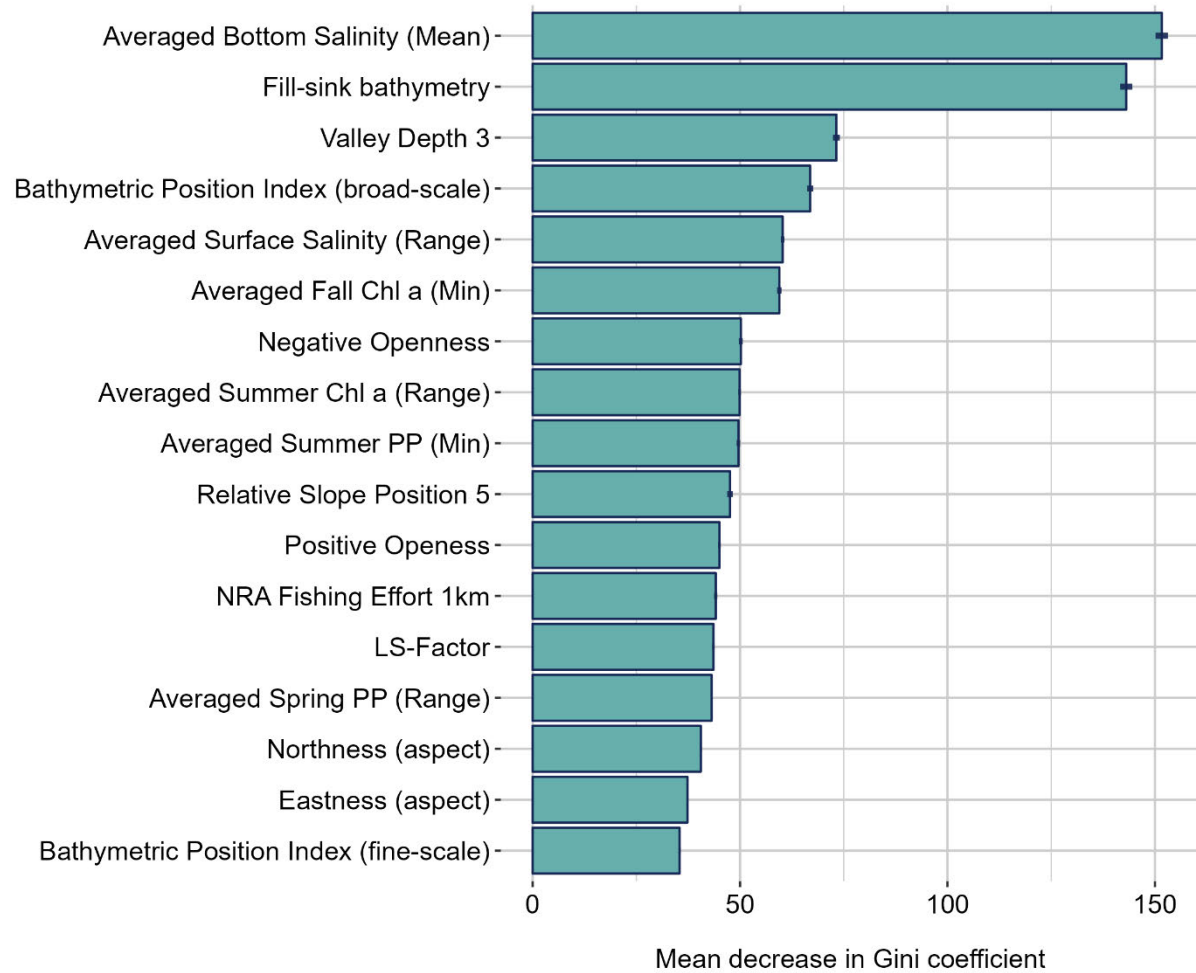
***Anthoptilum spp.***

The most important variables for *Anthoptilum spp.* were averaged mean bottom salinity, bathymetry, valley depth, broad scale bathymetric position index, and averaged minimum surface salinity (Figure 31). The models indicate that *Anthoptilum spp.* are typically located in depressed areas at depths more than 500 meters, with an optimum depth band around 700-1250 m depth, mean bottom salinity > 34.7‰, valley depth > 400 m, and low fall chlorophyll *a* concentration. Whilst the effect of bottom trawling effort is of lower importance in the model in comparison to the environmental conditions the probability of presence increases at lower fishing effort (Figure 32).

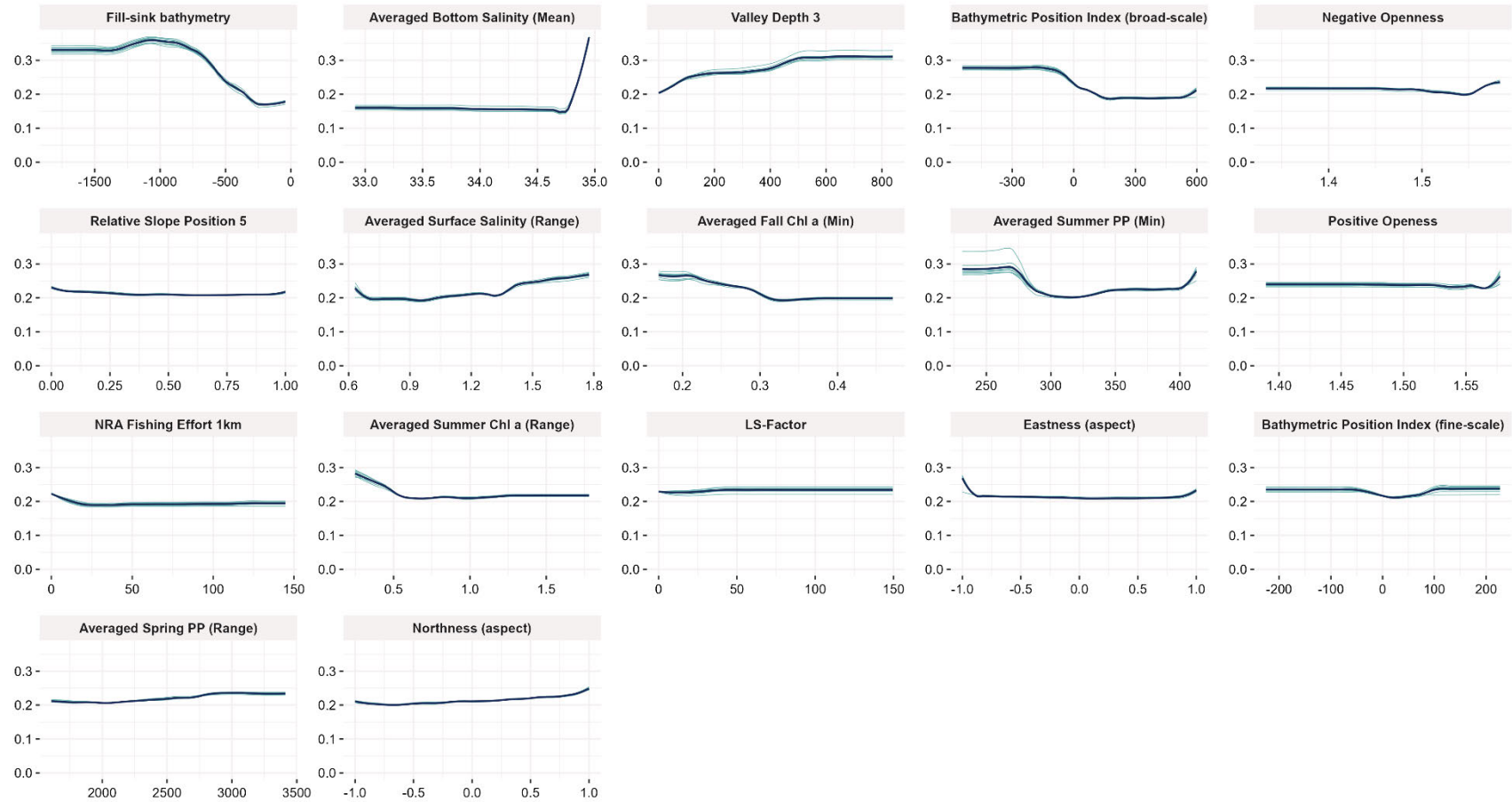
The predicted distribution maps are presented in Figure 33 as binary plots indicating presence/absence based on two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The data distribution is shown overlain on the binary map of presence/absence based on Sensitivity=Specificity (Figure 34). Outside the model extrapolation areas, *Anthoptilum spp.* forms a band around the West, North and East of the Flemish Cap, and the edge of the Tail of the Grand Bank.

The uncertainty expressed as the frequency of P/A from the 10 cross-validation runs (Figure 34), the areas of extrapolation (Figures 33-35) and the average probability of the maximum frequency class (Figure 35) indicated high certainty within the fishing footprint for both presence and absence predictions. However, there was increased uncertainty in the deeper slope waters both in areas of interpolation and extrapolation (Figure 35).

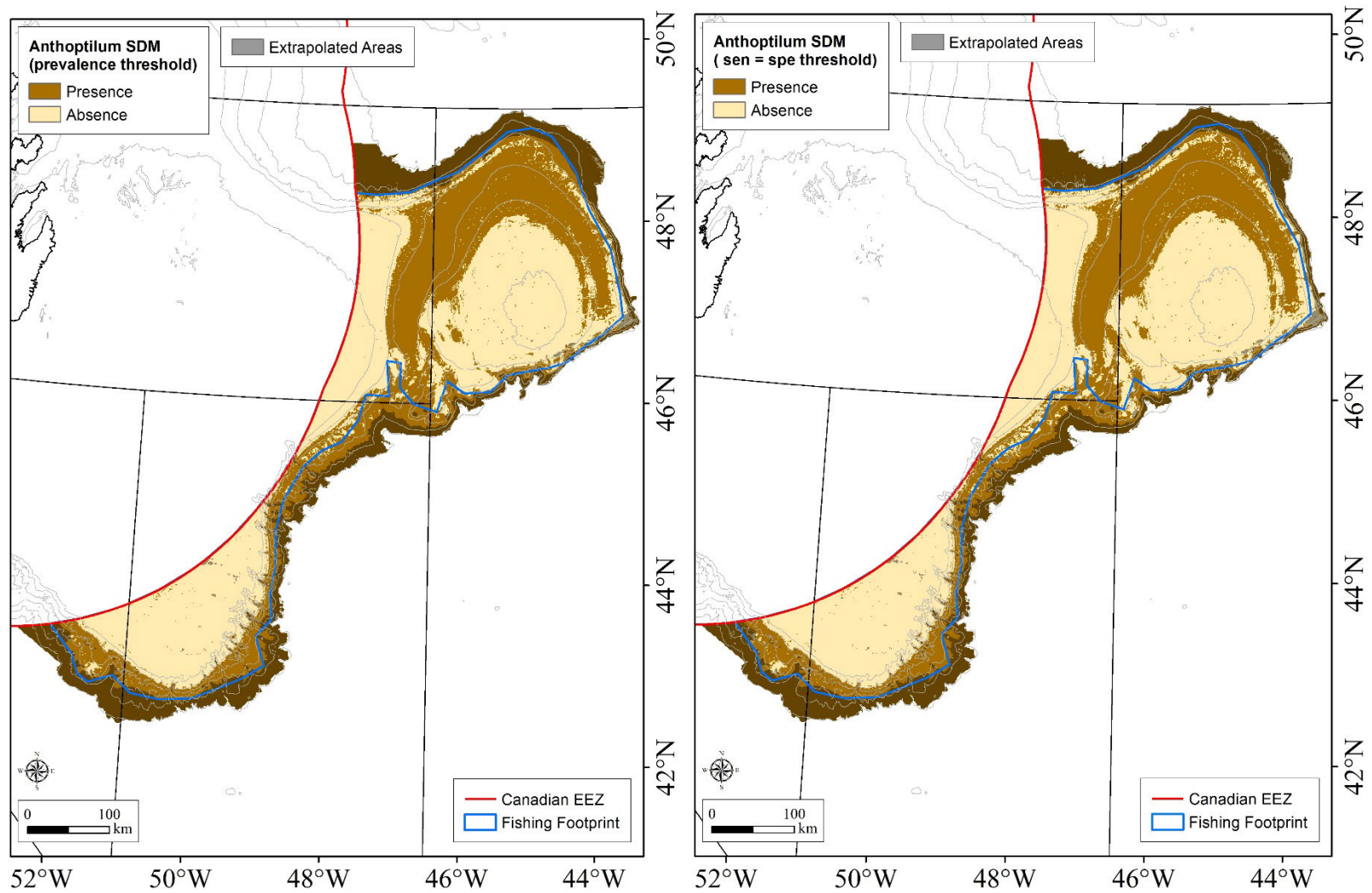




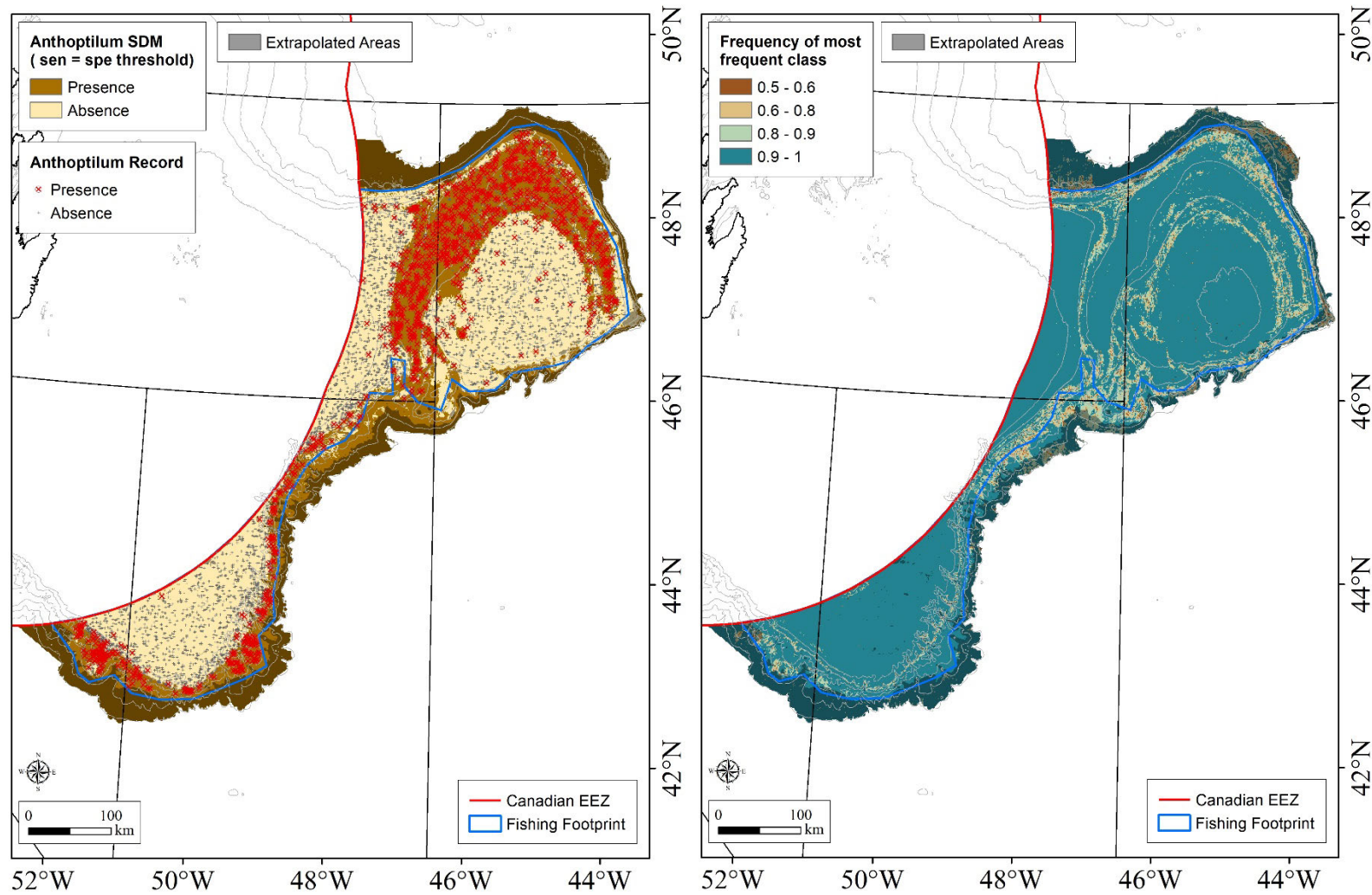
**Figure 31.** Plot of mean and standard deviation showing decrease in Gini Value for the variables in the Random Forest model for *Anthoptilum* spp., indicating their relative importance and variation across 10 data folds.



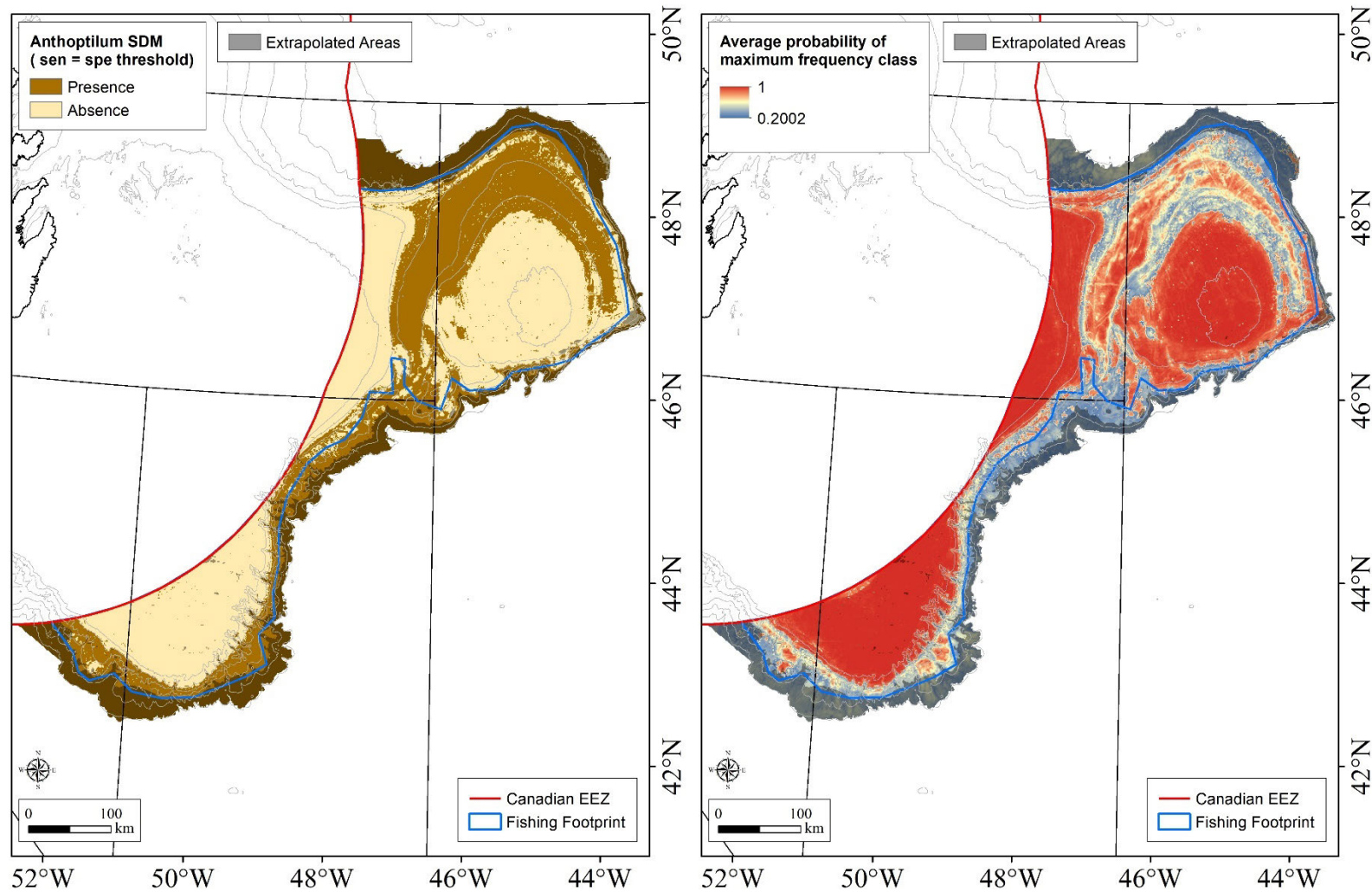
**Figure 32.** Response curves showing the partial dependence of the probability of presence on the predictors (Figure 31) identified in the Random Forest model for *Anthoptilum* spp. For each variable, the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.



**Figure 33.** Random Forest species distribution model for *Anthoptilum* spp. showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 34.** Random Forest species distribution model for *Anthoptilum* spp. showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records.



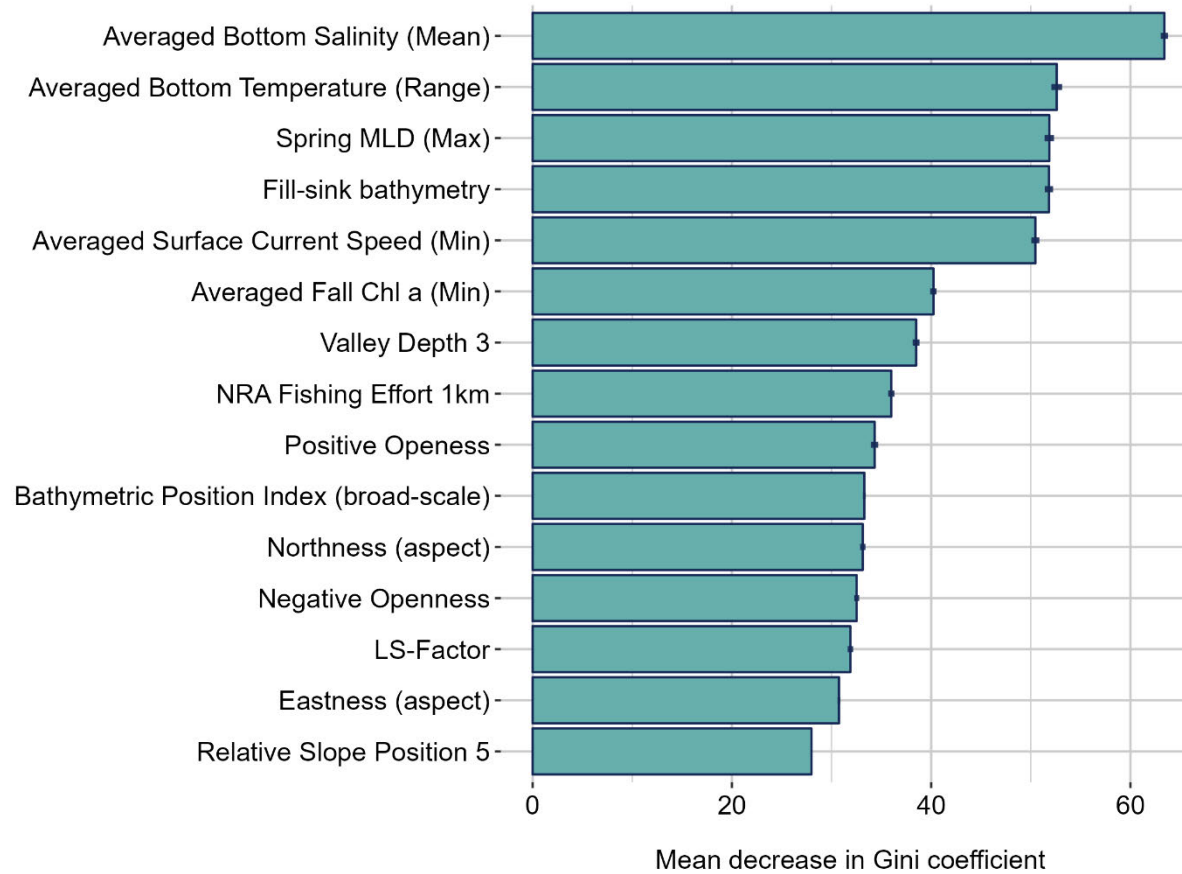
**Figure 35.** Random Forest species distribution model for the Sea Pen VME functional group showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records.

***Balticina spp.***

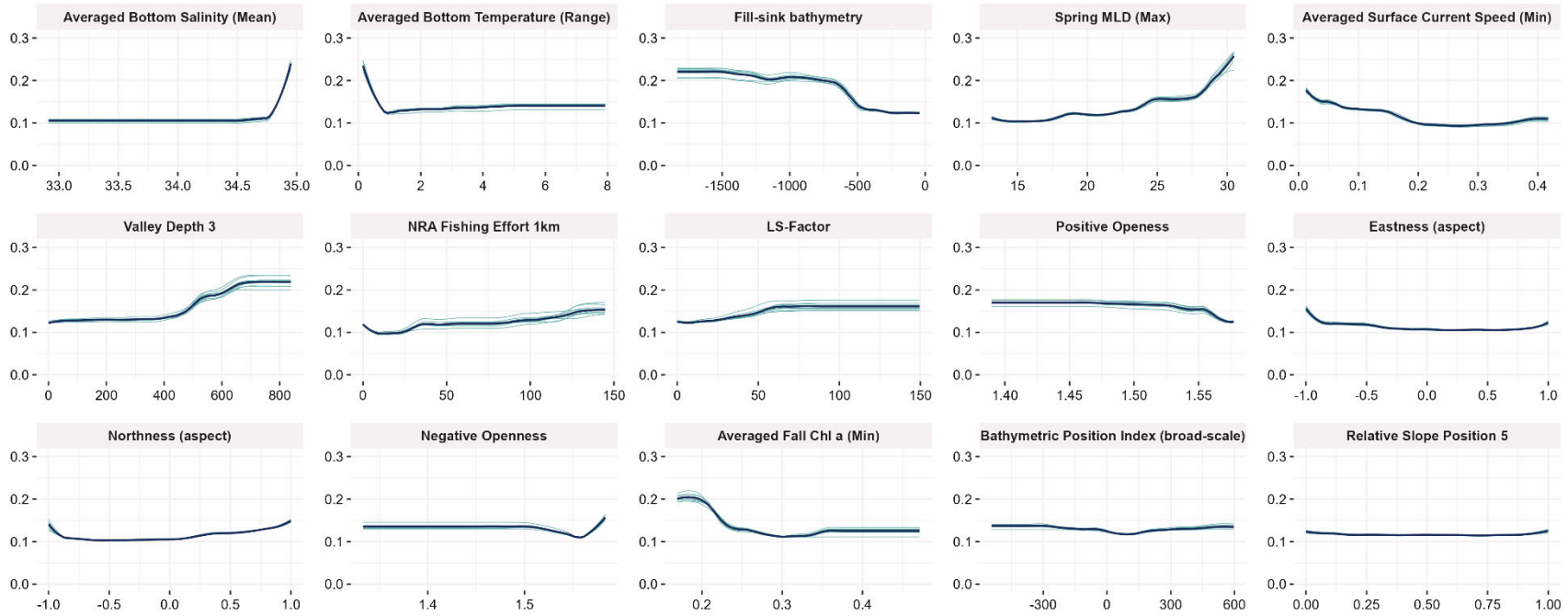
The most important variables for *Balticina* spp. were averaged mean bottom salinity, averaged bottom temperature range, maximum mixed layer depth in spring, bathymetry, averaged minimum surface current speed and averaged minimum fall chlorophyll concentration (Figure 36). The models indicate that *Balticina* spp. are typically located in areas with mean bottom salinity > 34.7‰, low temperature variability with average bottom temperature range < 1°, high spring mixed layer depth (> 22 m), low surface current speeds and low fall chlorophyll *a* concentration (Figure 37). Depth is a less important predictor variable than for the Sea Pens in general, with optimum depth > 500 m. The effect of bottom trawling effort is of low importance in the model but in contrast to the Sea Pen functional group, shows the probability of presence slightly increasing with increased fishing effort.

The predicted distribution maps are presented in Figure 38 as binary plots indicating presence/absence based on two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The distribution of the response data is shown overlain on the binary map of presence/absence based on Sensitivity=Specificity (Figure 39). Outside the model extrapolation areas, *Balticina* spp. is distributed around the Flemish Cap and edge of the Grand Bank seen with the Sea Pens functional group and other sea pen taxa. However, it has the shallowest predicted distribution of all of the individual sea pen taxa modelled, and is also predicted to be present to the South of the Flemish Cap.

The uncertainty expressed as the frequency of P/A from the 10 cross-validation runs (Figure 39), the areas of extrapolation (Figures 38-40) and the average probability of the maximum frequency class (Figure 40) indicated high certainty within the fishing footprint for both presence and absence predictions, although the shallower portions of the predicted presence on Flemish Cap had a lower average probability (Figure 40). Increased uncertainty was also identified in the deeper slope waters (Figure 40).

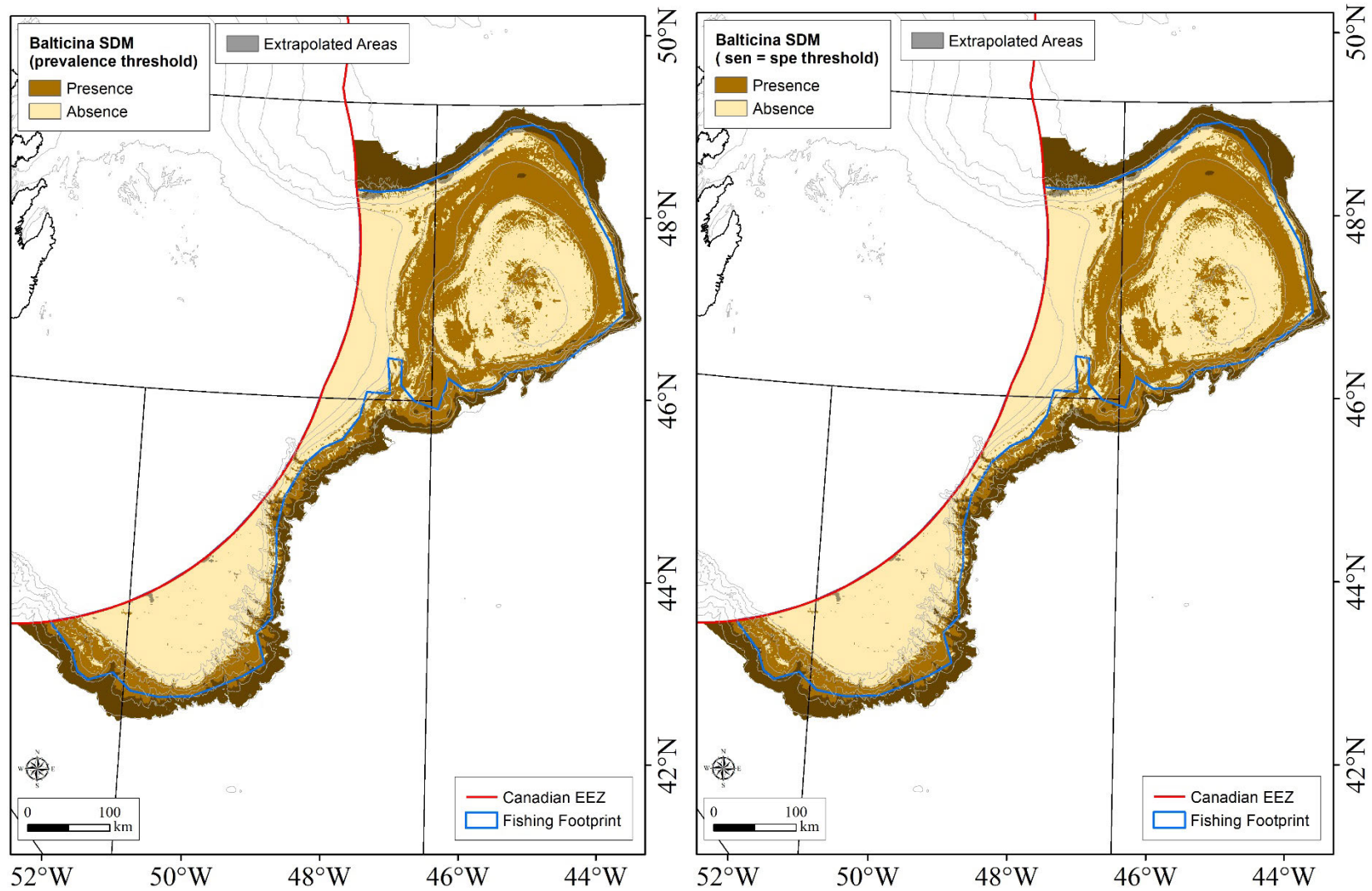


**Figure 36.** Plot of mean and standard deviation showing decrease in Gini Value for the variables in the Random Forest model for *Balticina* spp., indicating their relative importance and variation across 10 data folds.

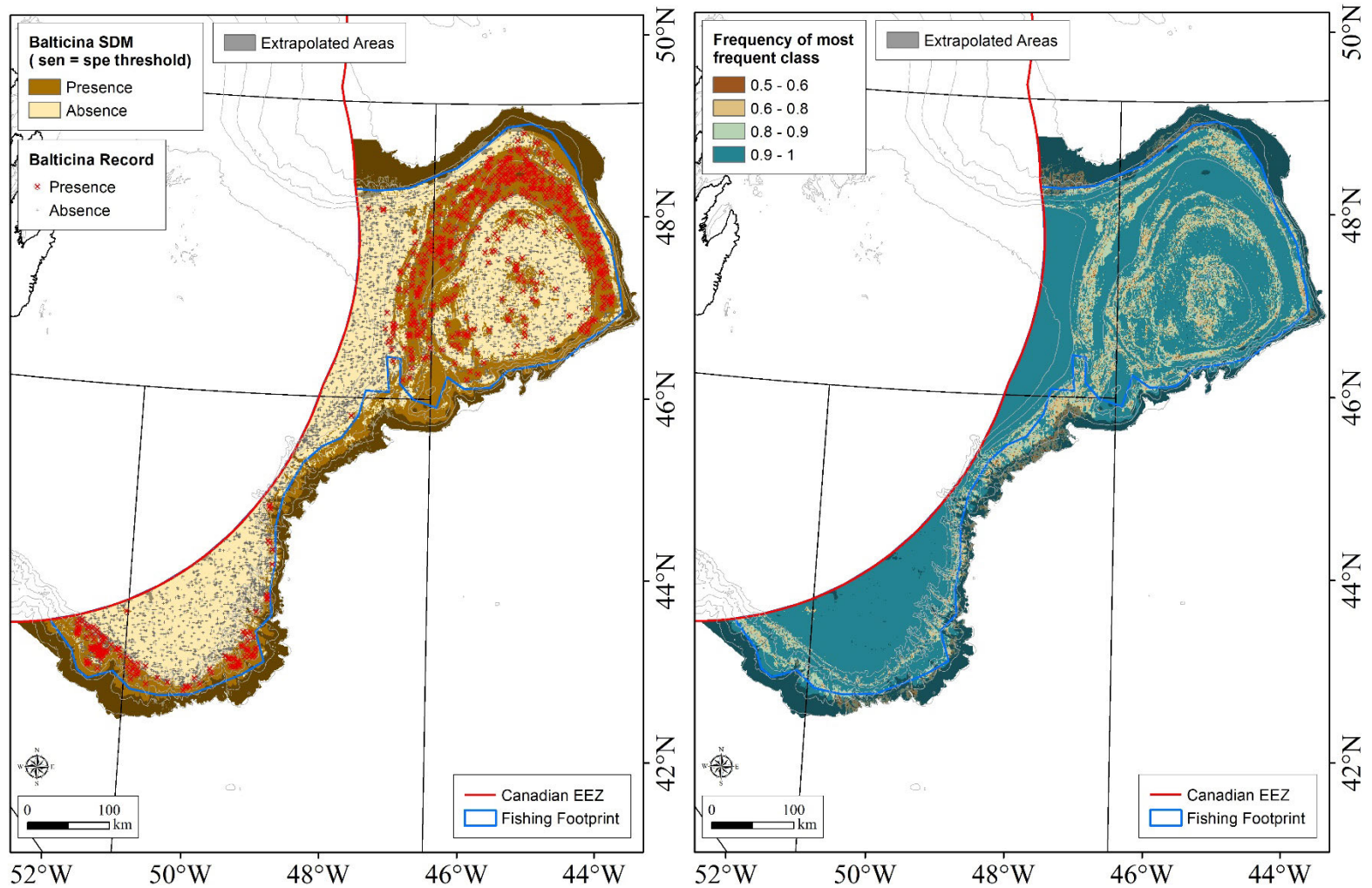


**Figure 37.** Response curves showing the partial dependence of the probability of presence on the predictors (Figure 36) identified in the Random Forest model for *Balticina* spp. For each variable, the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.

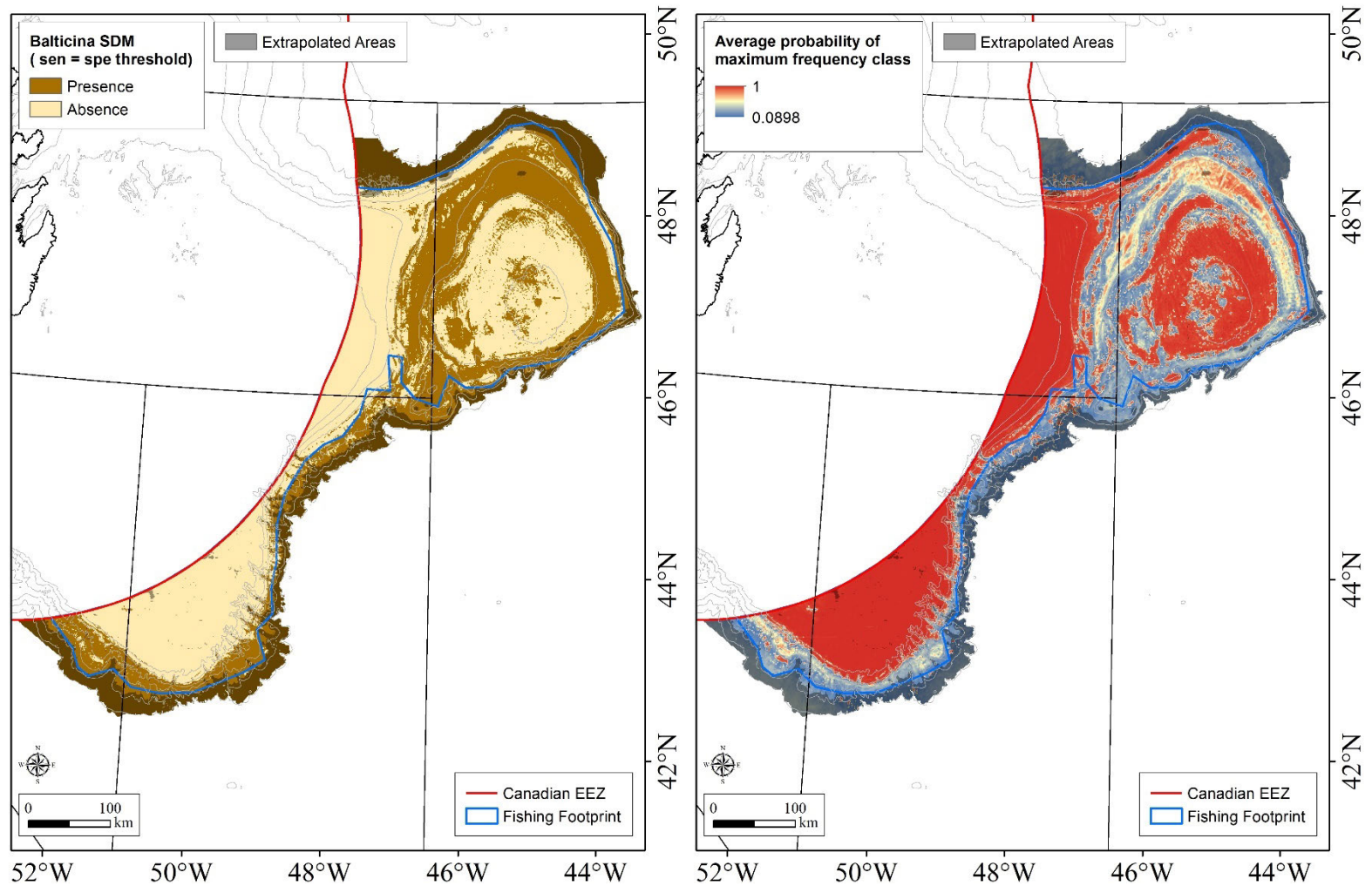




**Figure 38.** Random Forest species distribution model for *Balticina* spp. showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 39.** Random Forest species distribution model for *Balticina* spp. showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



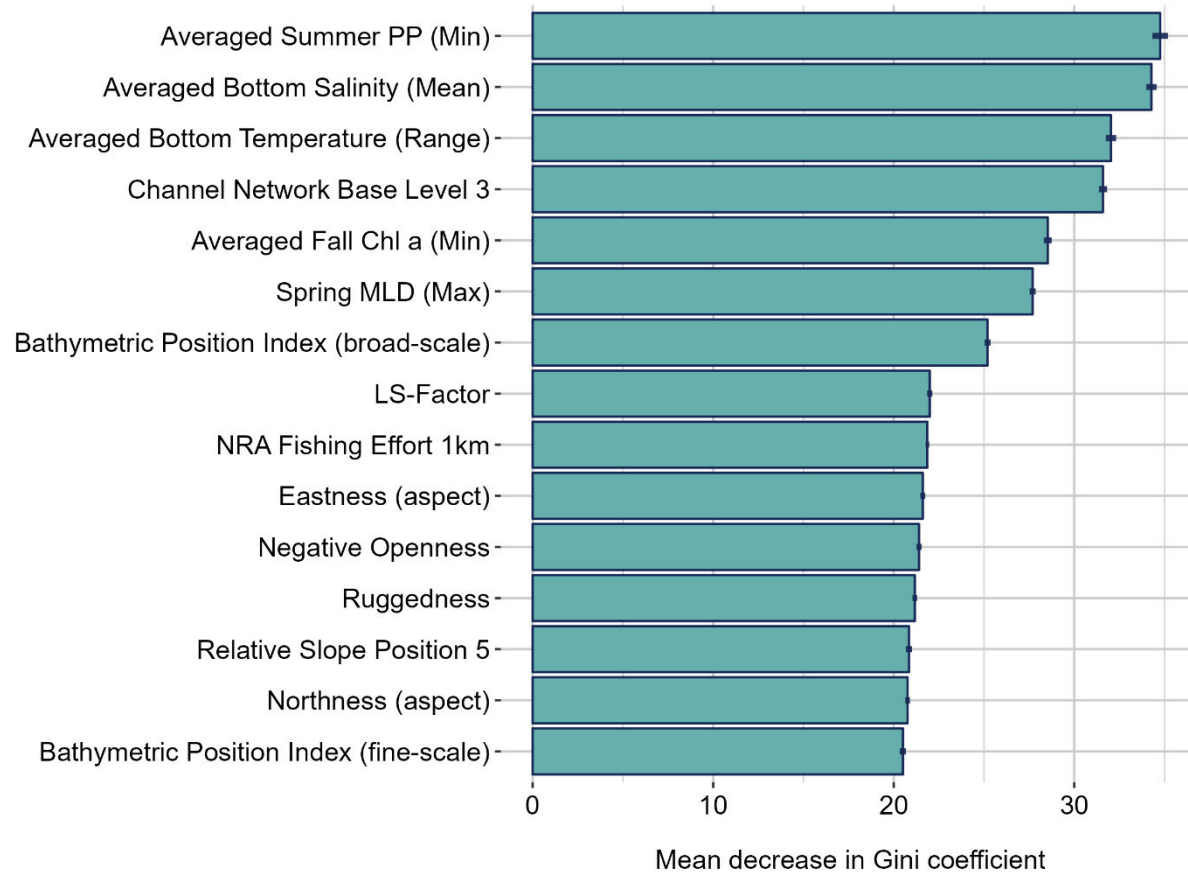
**Figure 40.** Random Forest species distribution model for *Balticina* spp. showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.

***Funiculina* spp.**

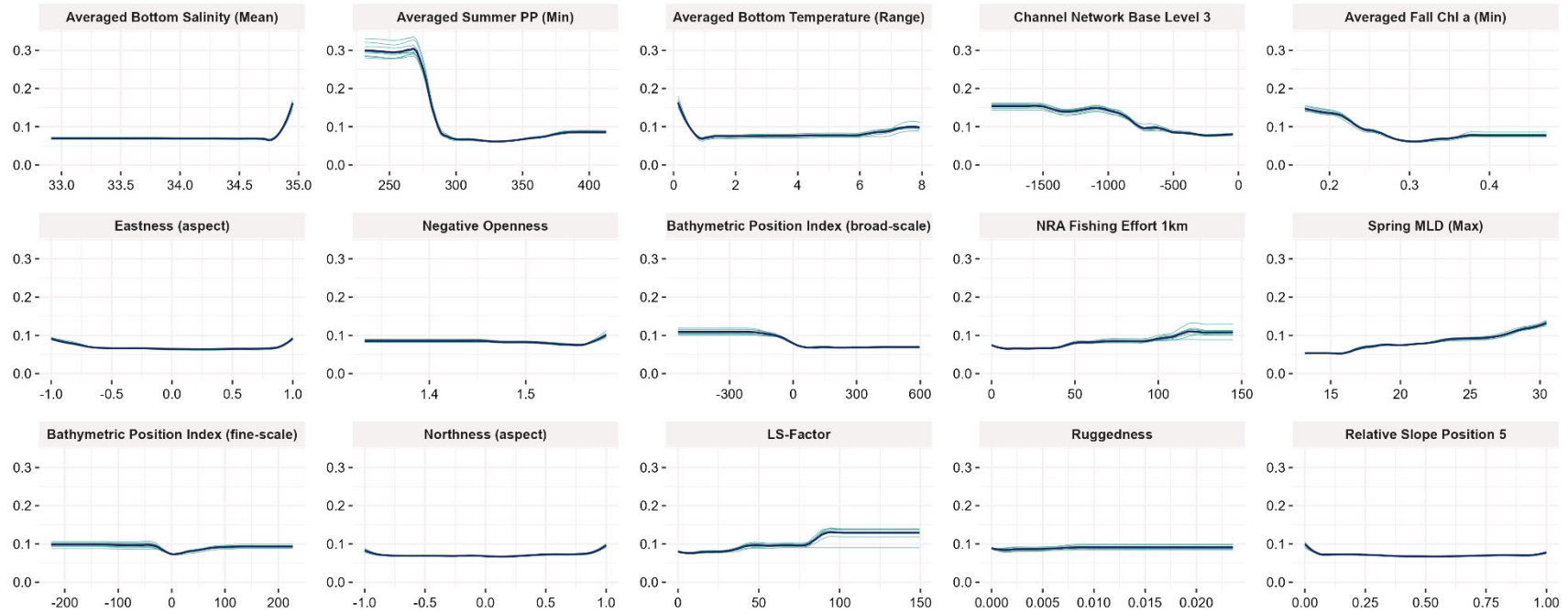
The most important variables for *Funiculina* spp. were the averaged minimum summer primary productivity, averaged mean bottom salinity, averaged bottom temperature range, channel network base level (effectively a coarse scale representation of bathymetry), averaged minimum fall chlorophyll a concentration, and maximum mixed layer depth in spring (Figure 41). The models indicate that *Funiculina* spp. are typically located in areas with low summer primary productivity, mean bottom salinity > 34.7‰, low temperature variability with average bottom temperature range < 1°, base depths > 500 m, and low fall chlorophyll a concentration (Figure 42). The likelihood of occurrence increases with higher spring mixed layer depths. The effect of bottom trawling effort is of low importance in the model but shows the probability of presence slightly increasing with increased fishing effort.

The predicted distribution maps are presented in Figure 43 as binary plots indicating presence/absence based on two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The data distribution is shown overlain on the binary map of presence/absence based on Sensitivity=Specificity (Figure 44). Outside the model extrapolation areas, *Funiculina* spp. generally follows the distribution around the Flemish Cap and edge of the Grand Bank seen with the Sea Pens functional group and other sea pen taxa. However, along with *Balticina* spp., it is also predicted to be present to the South of the Flemish Cap.

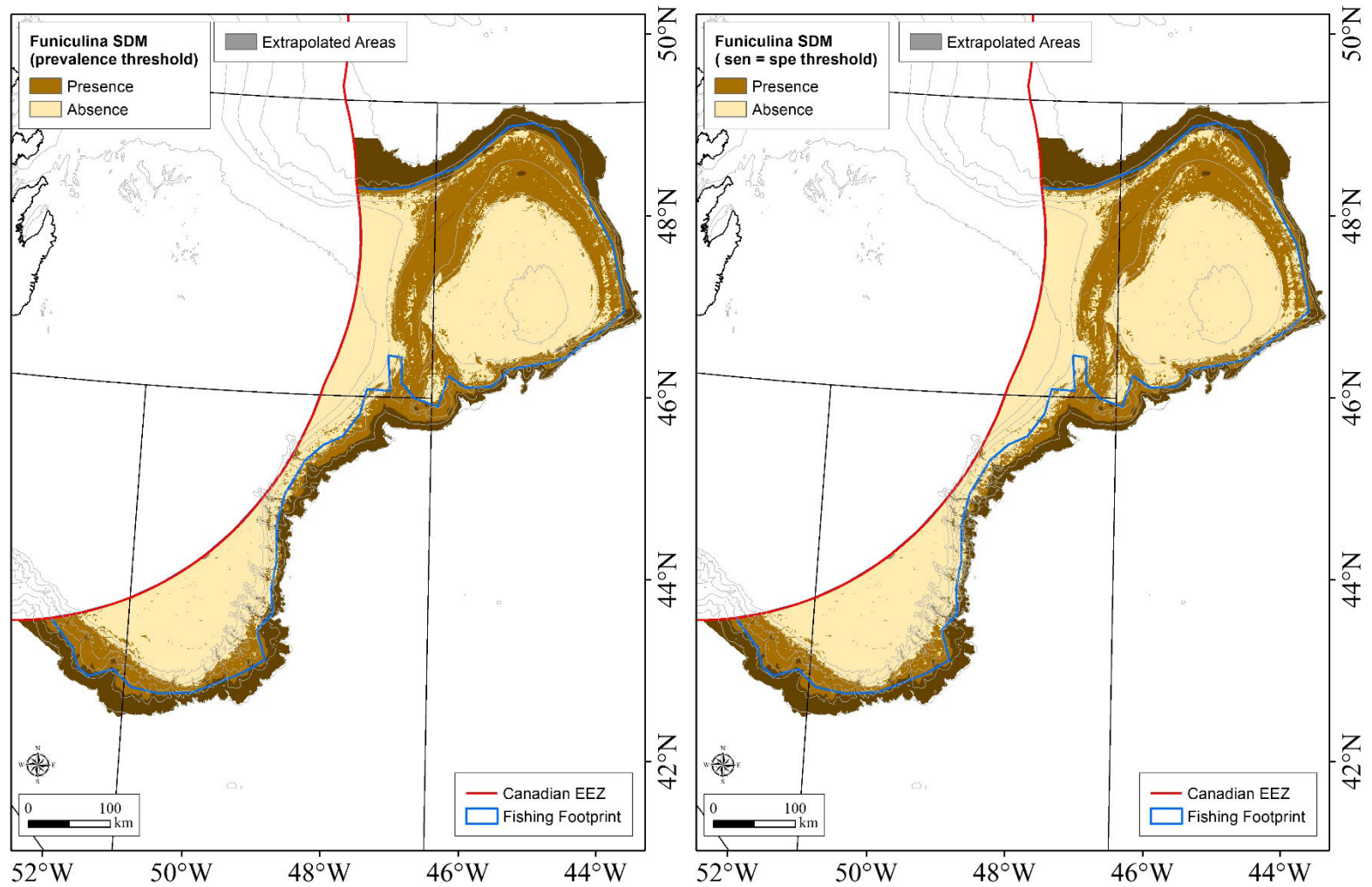
The uncertainty expressed as the frequency of P/A from the 10 cross-validation runs (Figure 44), the areas of extrapolation (Figures 43-45) and the average probability of the maximum frequency class (Figure 45) indicated high certainty across most of the area except for the boundary areas between the predicted presence and absence classes and in the deeper slope waters.



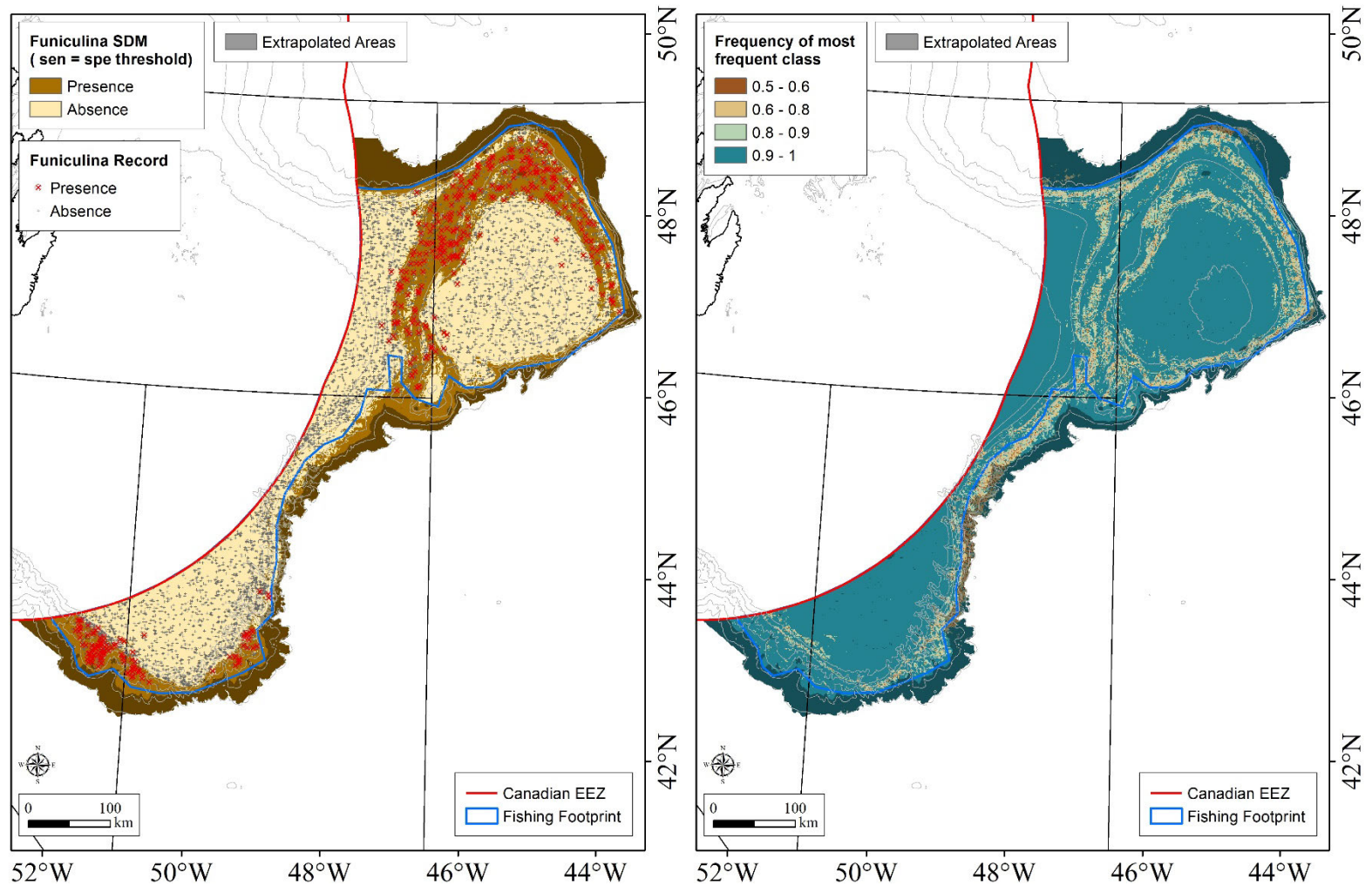
**Figure 41.** Plot of mean and standard deviation showing decrease in Gini Value for the variables in the Random Forest model for *Funiculina* spp., indicating their relative importance and variation across 10 model folds.



**Figure 42.** Response curves showing the partial dependence of the probability of presence on the predictors (Figure 41) identified in the Random Forest model for *Funiculina* spp. For each variable the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.

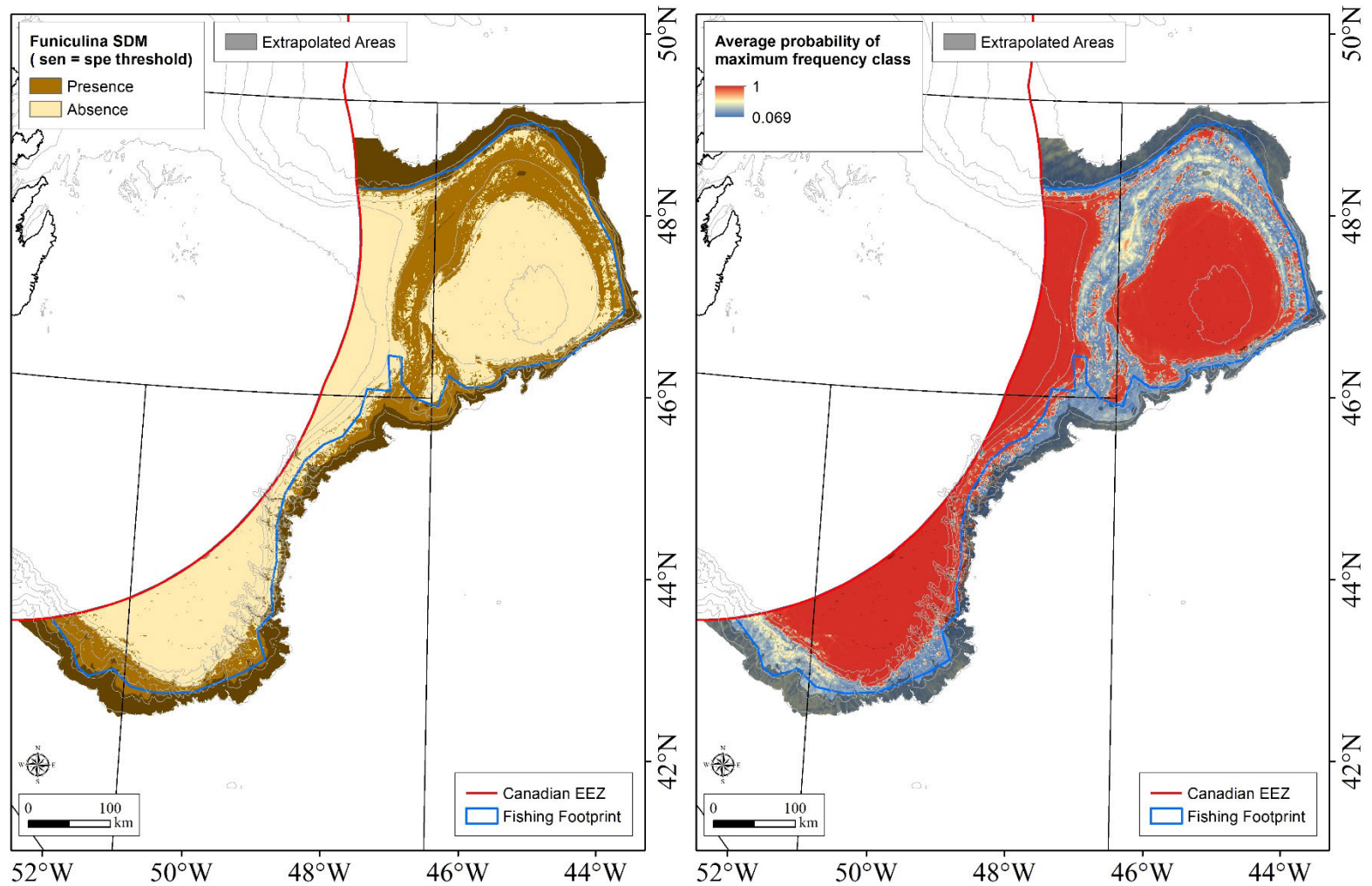


**Figure 43.** Random Forest species distribution model for *Funiculina* spp. showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 44.** Random Forest species distribution model for *Funiculina* spp. showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.





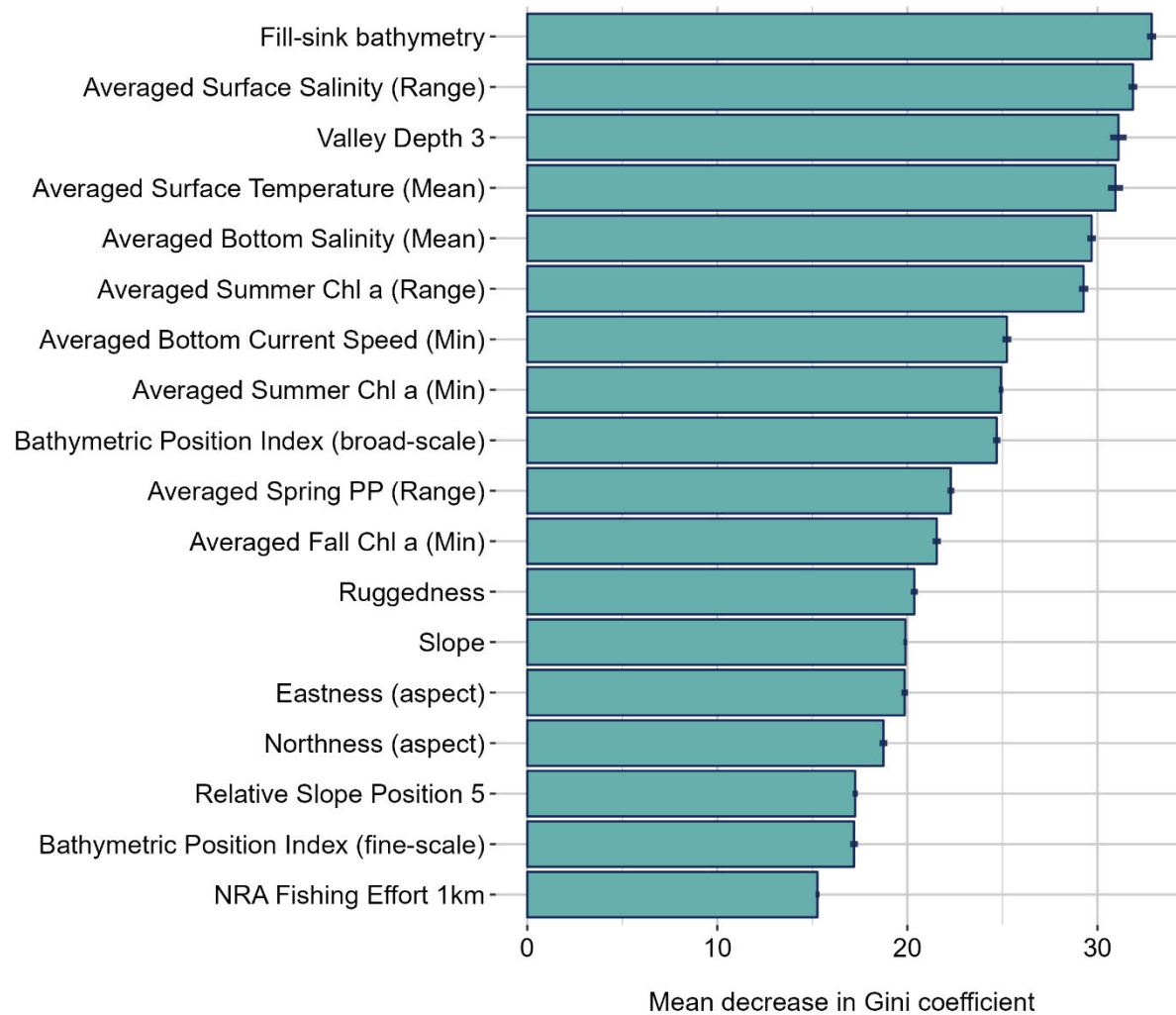
**Figure 45.** Random Forest species distribution model for *Funiculina* spp. showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.

***Pennatula spp.***

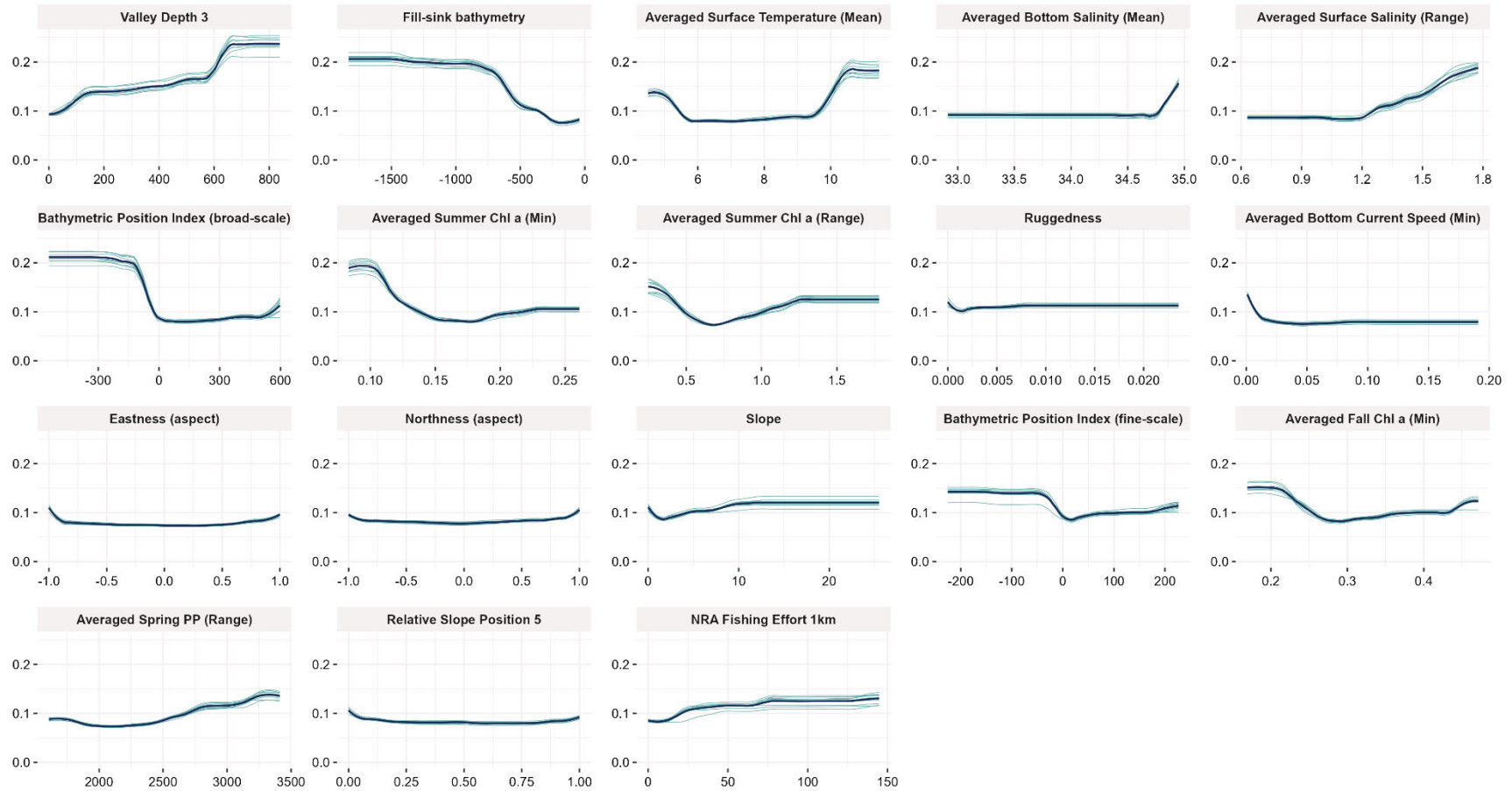
The most important variables for *Pennatula* spp. were bathymetry, averaged surface salinity range, valley depth, averaged mean surface temperature, averaged mean bottom salinity, and averaged range of summer chlorophyll *a* concentration (Figure 46). The models indicate that *Pennatula* spp. are typically located in depressed areas at depths > 500 meters, with optimum depths occurring at > 700 m. Likelihood of presence increased in areas with variable surface salinity, mean bottom salinity > 34.7‰, and low summer chlorophyll *a* concentration (Figure 47). The effect of bottom trawling effort is of low importance in the model but shows the probability of presence slightly increasing with increased fishing effort.

The predicted distribution maps are presented in Figure 48 as binary plots indicating presence/absence based on two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The data distribution is shown overlain on the binary map of presence/absence based on Sensitivity=Specificity (Figure 49). The distribution of *Pennatula* spp. outside the model extrapolation areas differs from the sea pens in general. Unlike the other taxa *Pennatula* spp. occurs on the Sackville Spur and has a wider distribution in the Flemish Pass and on the Nose of the Grand Bank. In contrast, the predicted distribution excludes most of the Northern and Eastern flanks of the Flemish Cap. The distribution along the edge of the Tail of Grand Bank follows that of the other sea pen taxa.

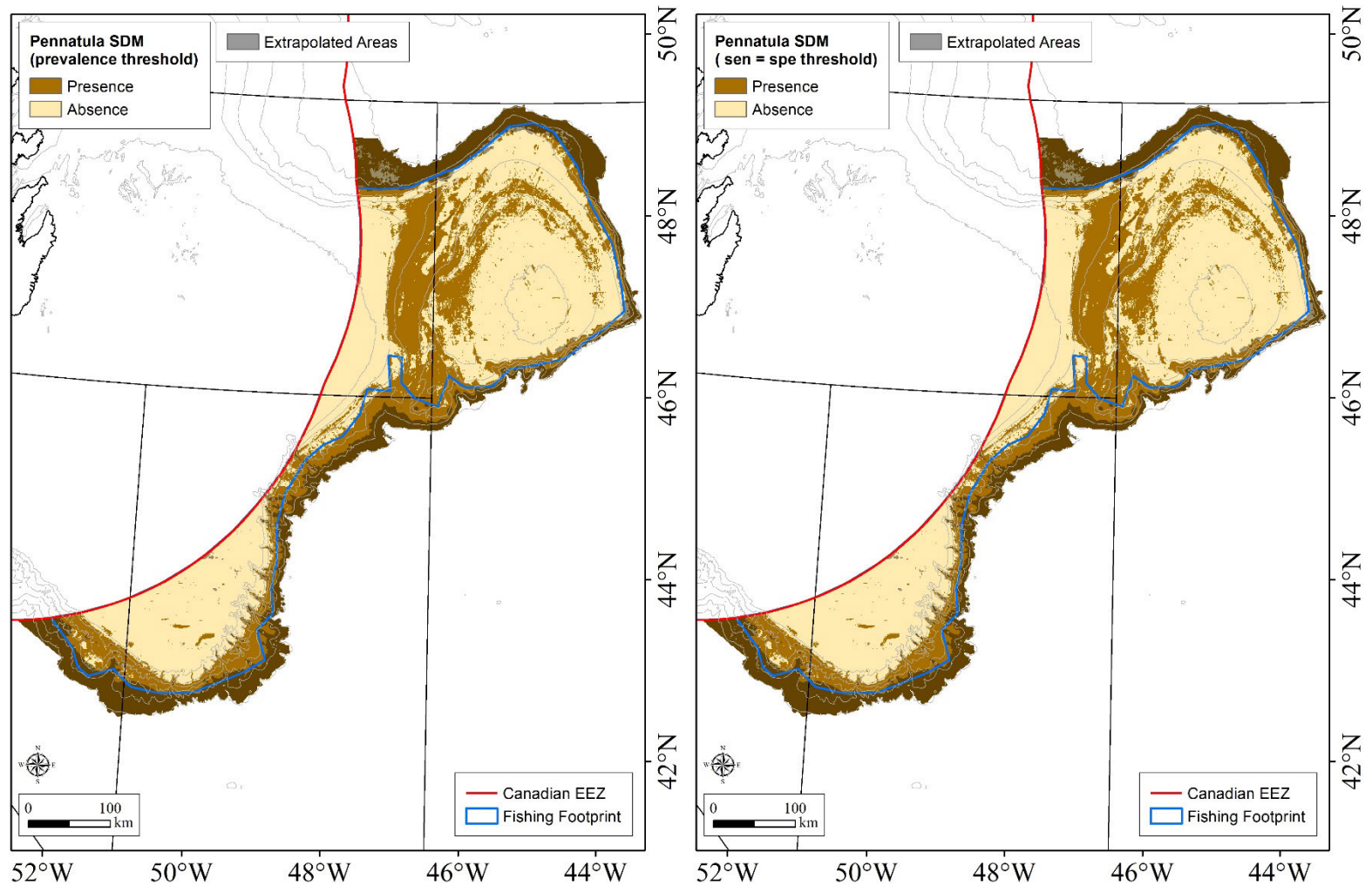
The uncertainty expressed as the frequency of P/A from the 10 cross-validation runs (Figure 49), the areas of extrapolation (Figures 48-50) and the average probability of the maximum frequency class (Figure 50) indicated high certainty within the fishing footprint for both presence and absence predictions, although the shallower portions of the predicted presence on Flemish Cap had a lower average probability (Figure 50). Increased uncertainty was also identified in the deeper slope waters (Figure 50).



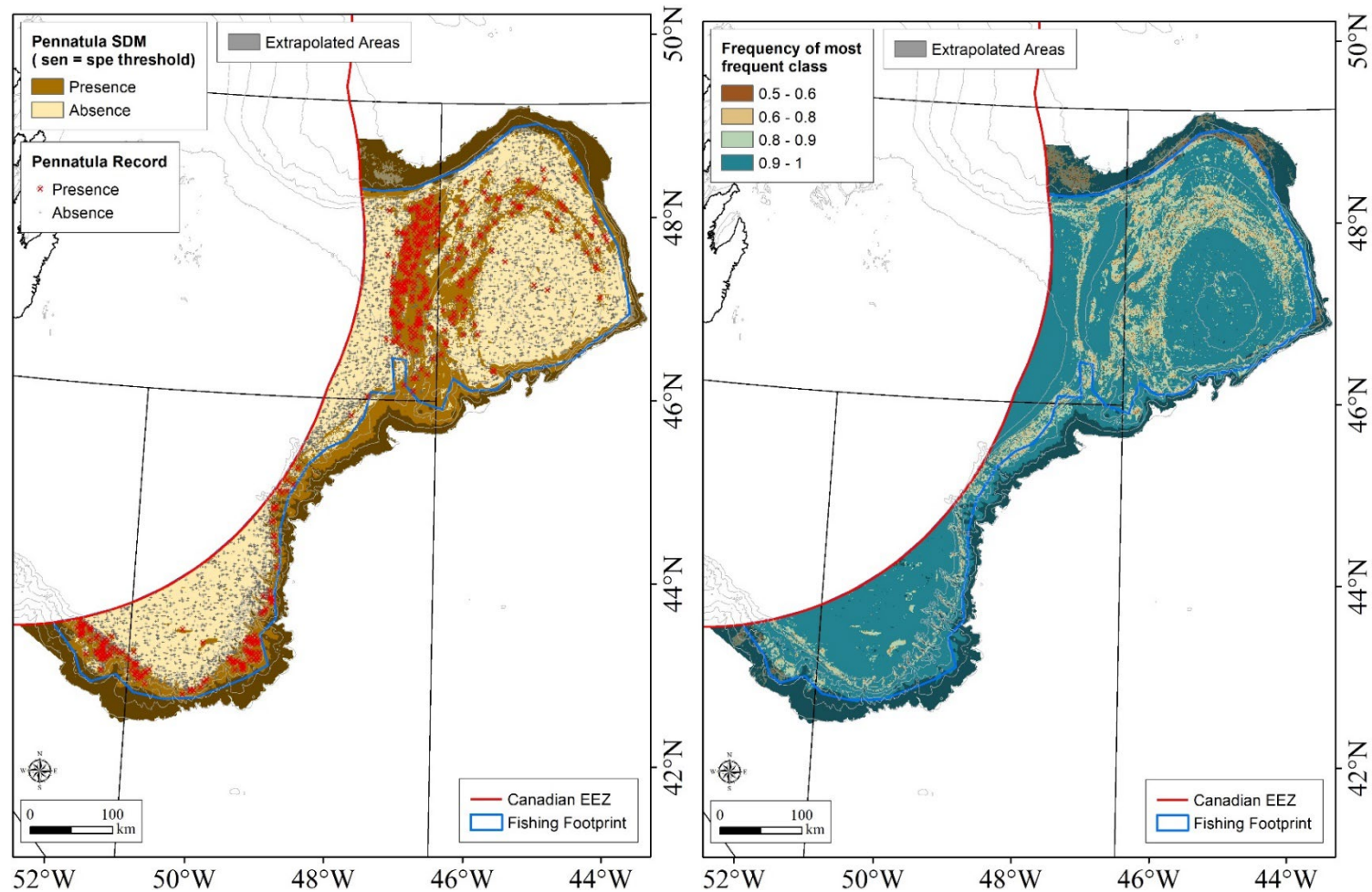
**Figure 46.** Plot of mean and standard deviation showing decrease in Gini Value for the variables in the Random Forest model for *Pennatula* spp., indicating their relative importance and variation across 10 data folds.



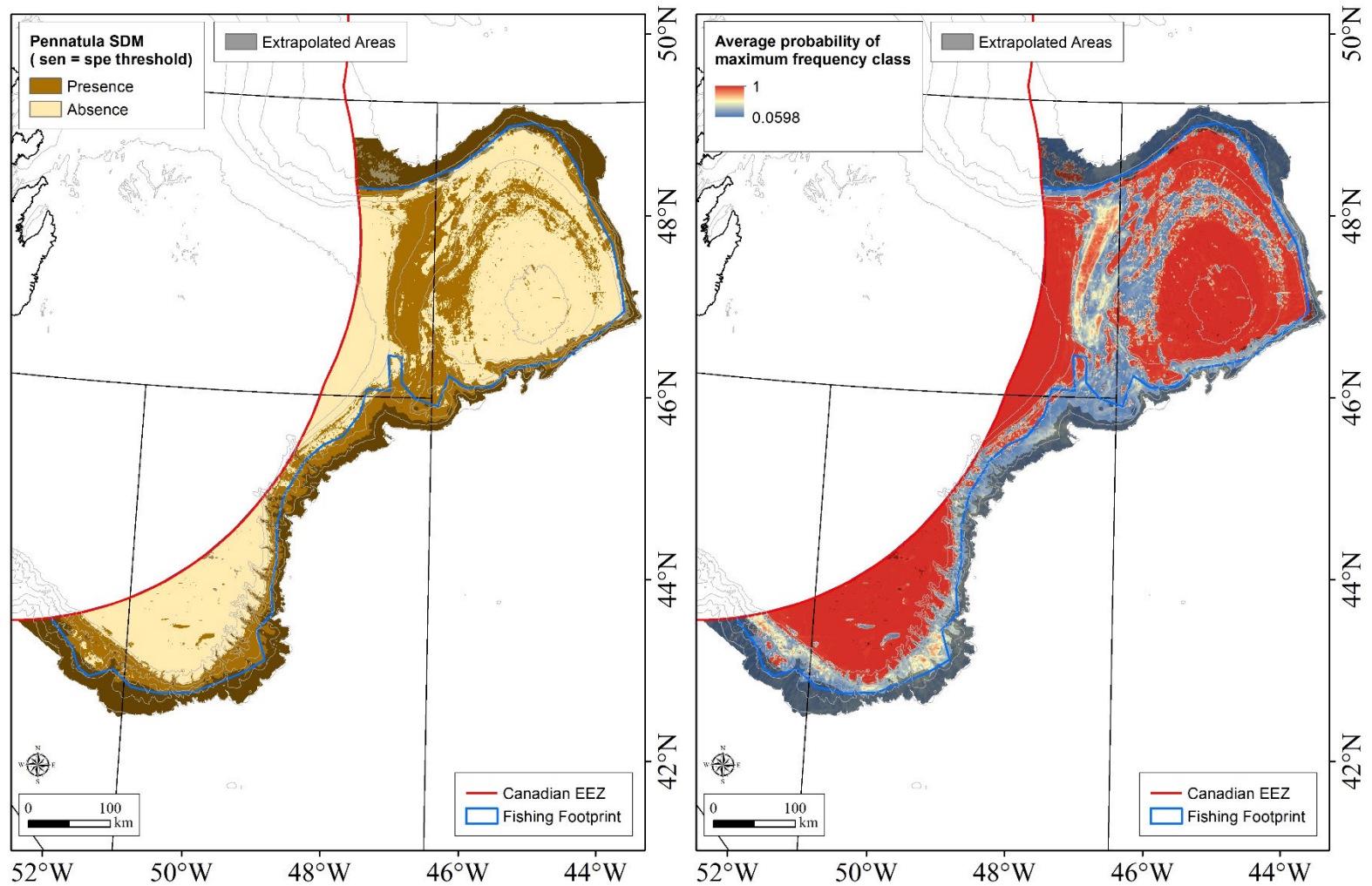
**Figure 47.** Response curves showing the partial dependence of the probability of presence on the predictors (Figure 46) identified in the Random Forest model for *Pennatula* spp. For each variable, the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.



**Figure 48.** Random Forest species distribution model for *Pennatula* spp. showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 49.** Random Forest species distribution model for *Pennatula* spp. showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 50.** Random Forest species distribution model for *Pennatula* spp. showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.

### ***Assessment and Prediction of Black Corals***

Random Forest models predicting the probability of the presence of the Black Coral VME functional group generally scored high accuracy across the validation statistics (Balanced Accuracy, Sensitivity and Specificity all > 0.8; Table 7). However, Kappa, which measures the extent to which the agreement between observed and predicted is higher than that expected by chance alone, was 0.25 which is considered 'fair' performance. The TSS, defined as the average of the net prediction success rate for presence sites and that for absence sites was 0.66 which indicates good model performance.

**Table 7.** Model Validation Results for the Presence/Absence Random Forest Model for the Black Coral VME Functional Group. TSS=True Skill Statistic (Sensitivity + Specificity - 1 ).

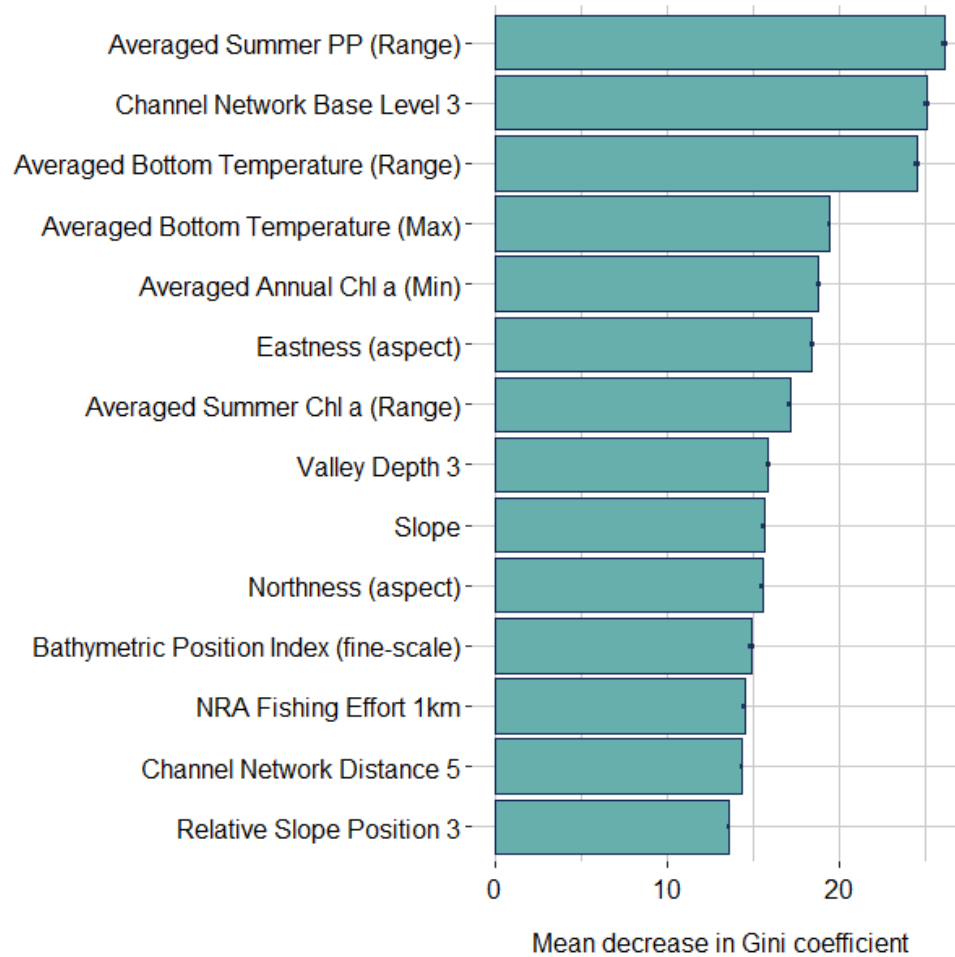
<b>Accuracy Measure</b>	<b>Mean ± SD</b>
Sensitivity	0.85 ± 0.04
Specificity	0.81 ± 0.04
Kappa	0.25 ± 0.06
Balanced Accuracy	0.83 ± 0.04
TSS	0.66 ± 0.07

The most important variables were the averaged range of summer Primary Productivity, followed by the terrain variable Channel Network Base Level (3), the range of the mean bottom temperature, and the mean maximum bottom temperature (Figure 51). The models indicate that the Black Corals are found in areas with a summer primary productivity range > 500 mg C m<sup>-2</sup> day<sup>-1</sup>, stable temperature environments with bottom temperature ranges < 1°C, and topographic complexity (Figure 52).

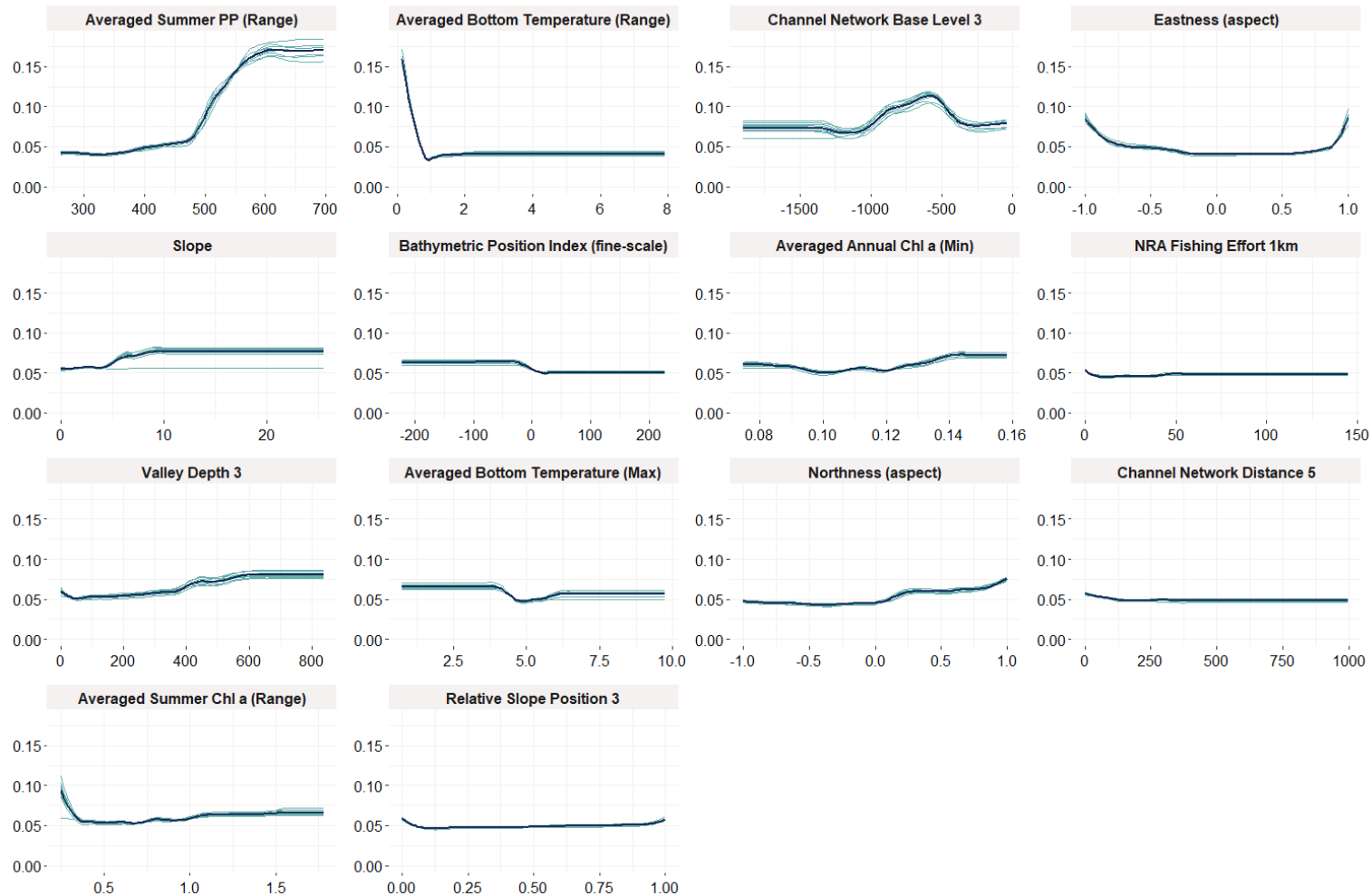
The predicted distribution maps are shown in Figure 53, shown as binary plots of presence/absence based on the two thresholds (Prevalence and Sensitivity=Specificity). These two plots are very similar. The distribution of the data is shown overlain on the binary map of presence/absence based on Sensitivity=Specificity (Figure 54). Outside areas of model extrapolation, the black corals are distributed around the Flemish Cap between 500 and 1000 m depth and south of Flemish Pass.

The uncertainty expressed as the frequency of P/A from the 10 cross-validation runs (Figure 53), the areas of extrapolation (Figures 53-55), and the average probability of the maximum frequency class (Figure 55) indicated high certainty within the fishing footprint for both presence and absence predictions. However, there was increased uncertainty in the deeper slope waters (Figure 55) and in areas of transition between the presence and absence classes (Figures 54-55).

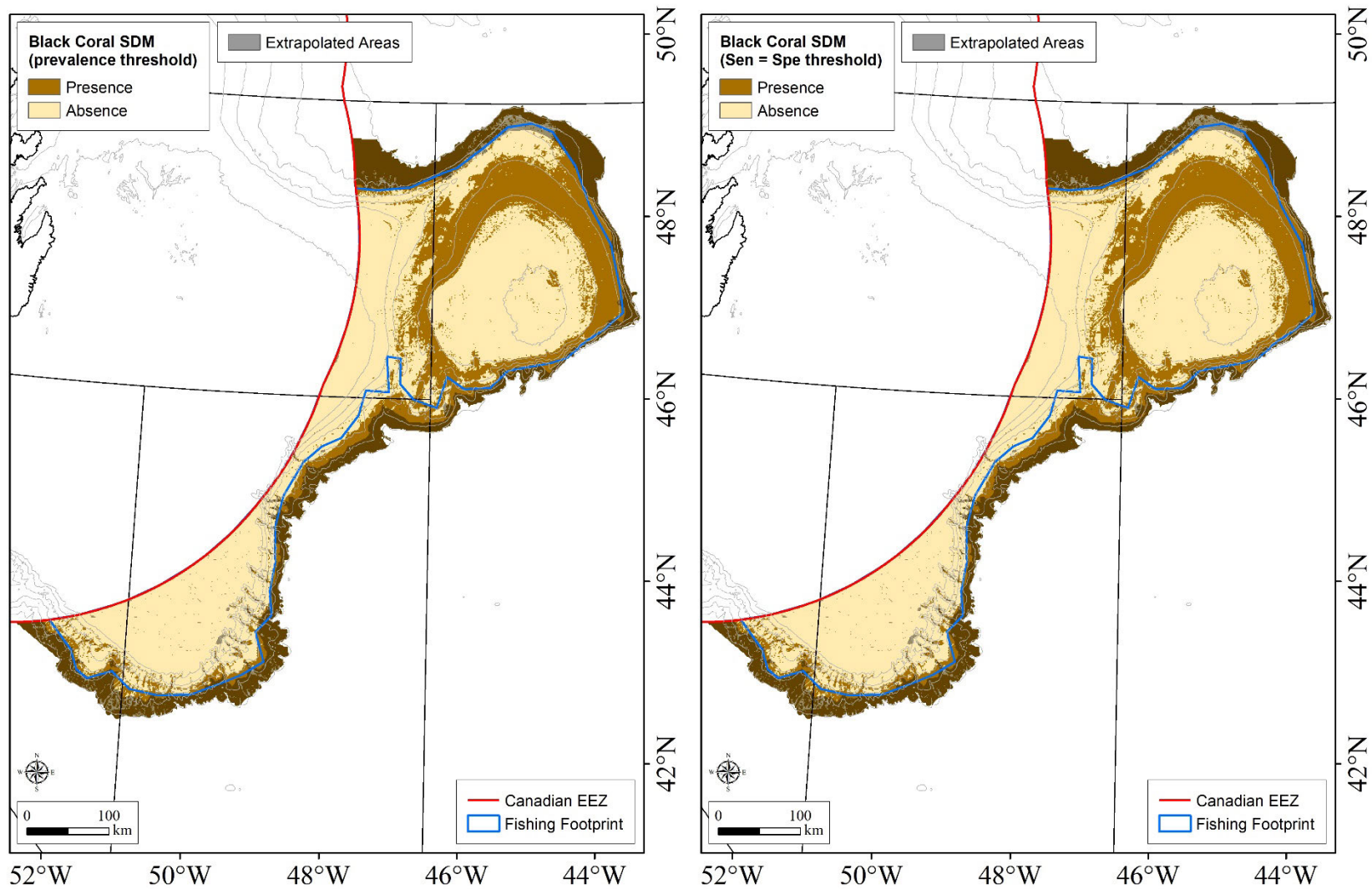




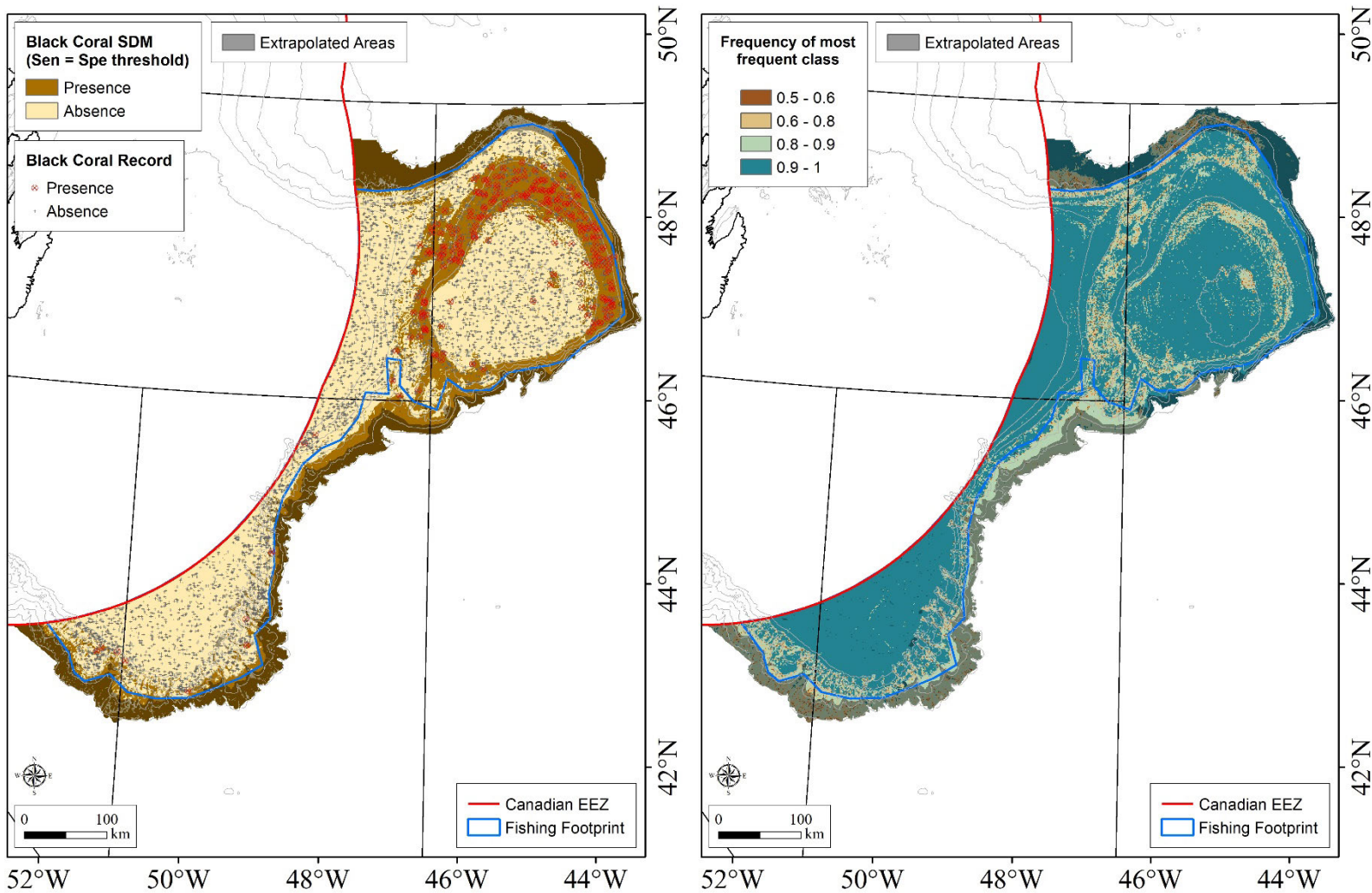
**Figure 51.** Plot of mean and standard deviation showing decrease in Gini Value for the 14 variables in the Random Forest model for the Black Coral VME functional group, indicating their relative importance and variation across 10 data folds.



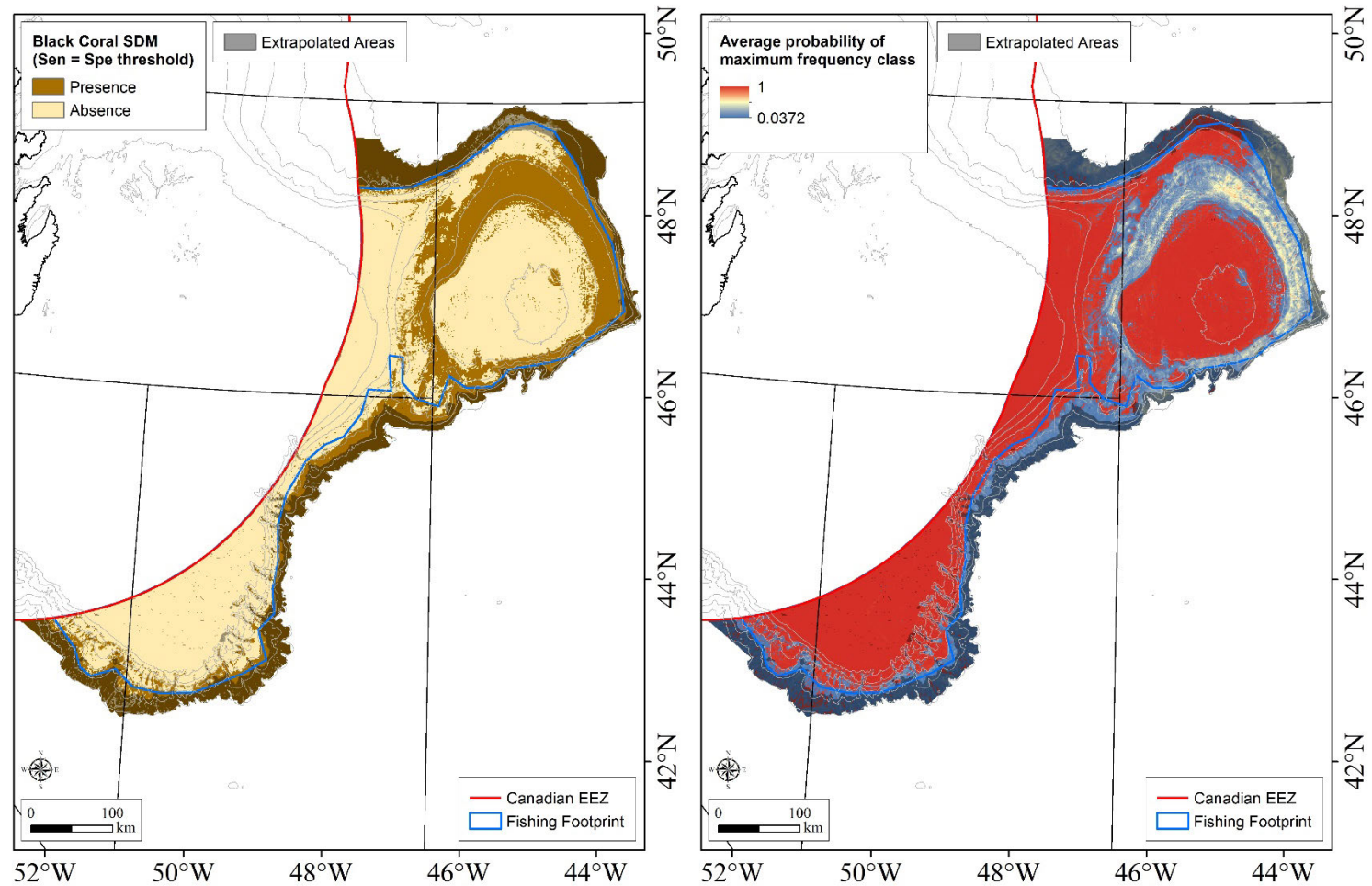
**Figure 52.** Response curves showing the partial dependence of the probability of presence on the 14 predictor variables (Figure 51) identified in the Random Forest model for the Black Coral VME functional group. For each variable, the mean response and curves for each of the model folds are plotted. The plots show the predicted response to each predictor variable in turn, whilst other variables are held at their mean value.



**Figure 53.** Random Forest species distribution model for the VME functional group Black Coral showing binary maps of VME presence thresholded using data prevalence (left panel) and a Sensitivity=Specificity threshold (right panel). Areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



**Figure 54.** Random Forest species distribution model for the VME functional group Black Coral showing the distribution of the presence and absence data overlain on a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the frequency of P/A from the 10 cross-validation runs (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.



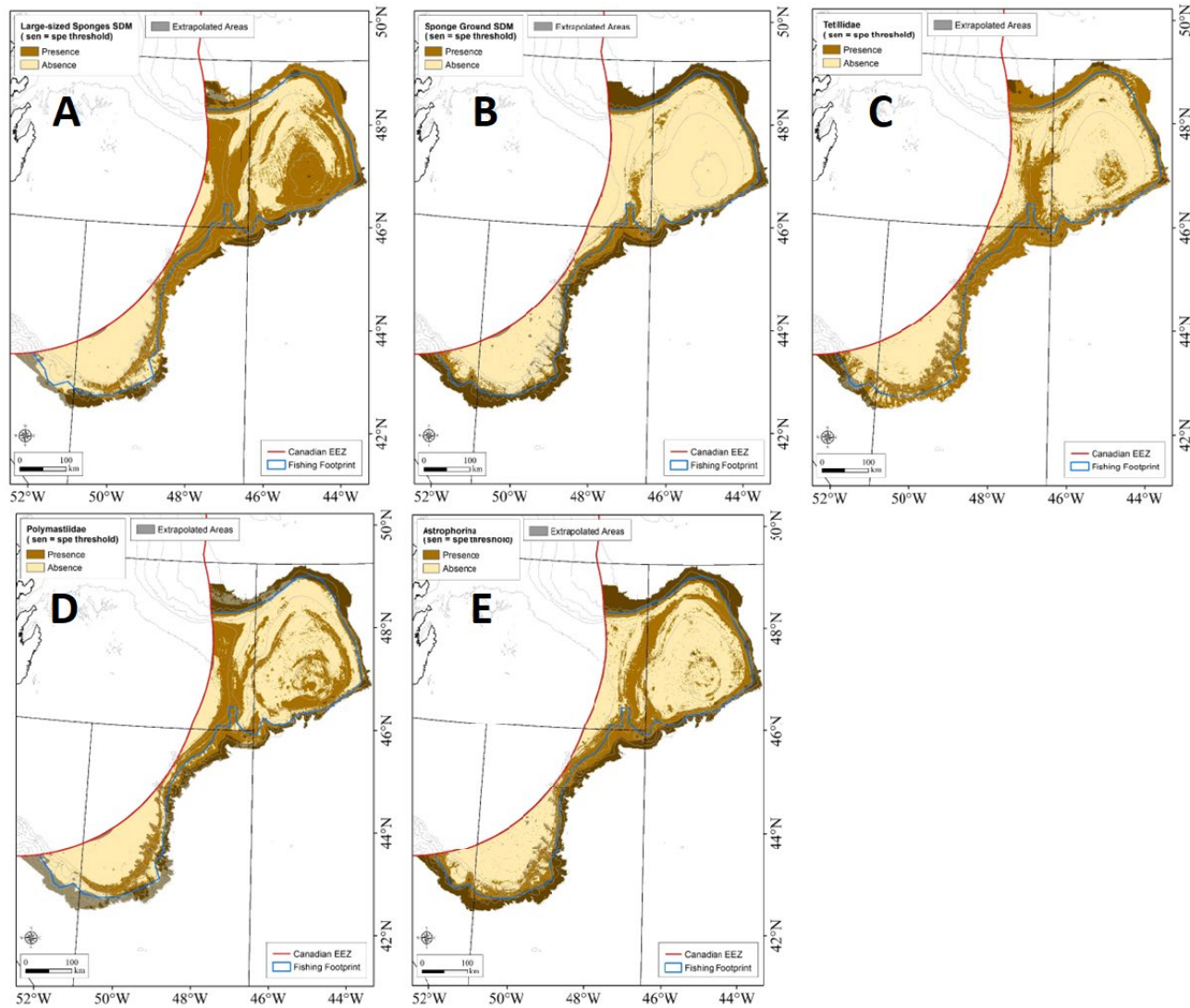
**Figure 55.** Random Forest species distribution model for the VME functional group Black Coral showing a binary map thresholded using a Sensitivity=Specificity threshold (left panel). Model uncertainty is illustrated by showing the average probability of the maximum frequency class (right panel). The areas of extrapolation show where the model has predicted into areas outside of the environment for the presence and absence records. The perimeter of the fishing footprint is shown on both maps.

## Discussion

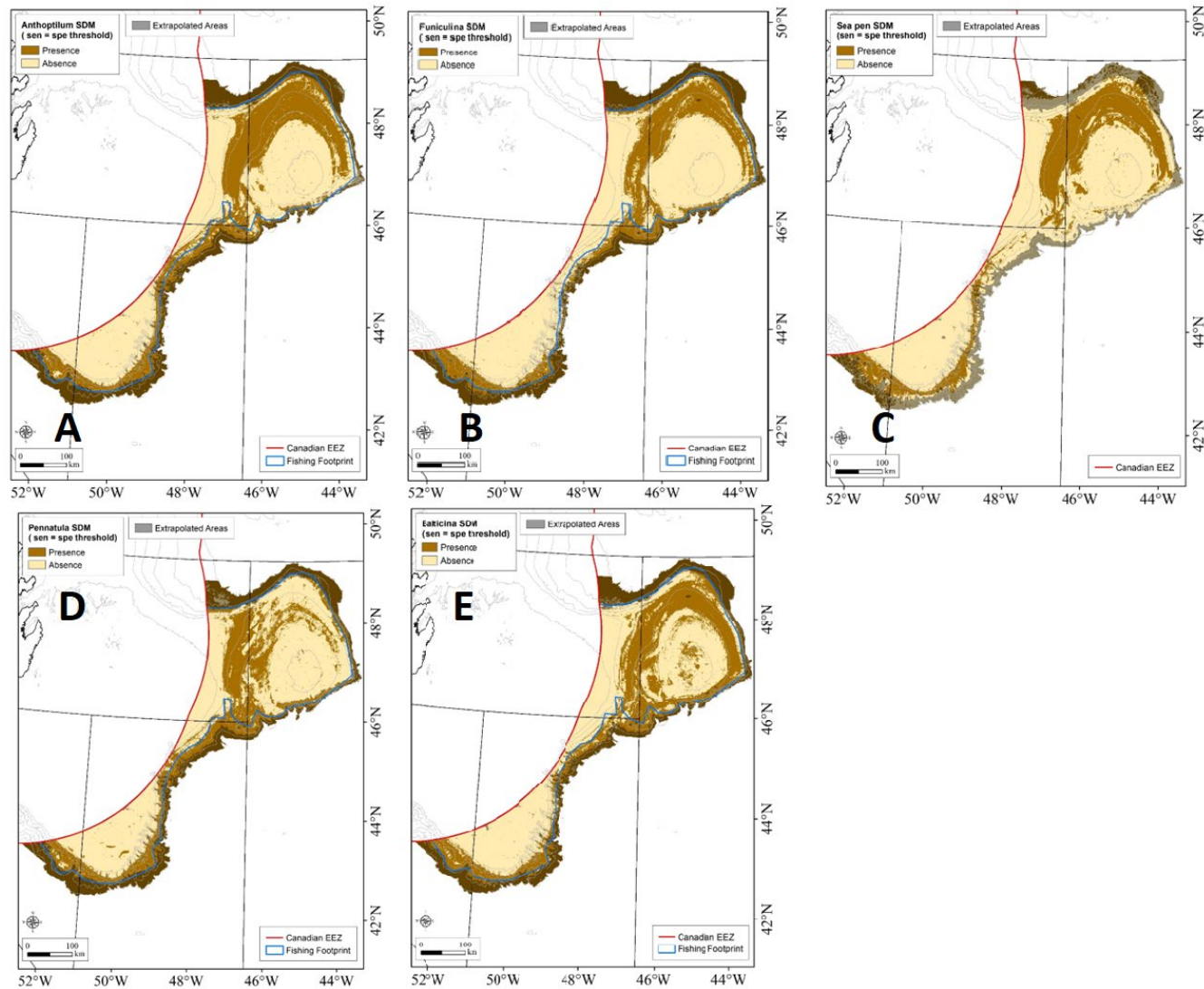
All models generally scored high accuracy across the validation statistics. The binary Presences/Absences maps are based on a threshold of Sensitivity=Specificity, which is the threshold where the chance of correctly predicting a positive or negative observation is the same. Previously, Prevalence (the ratio of Presences/Absences) was used which produced very similar outputs. However, a threshold of Sensitivity=Specificity will be used for the 2027 review of the closed areas.

Another advancement over previous work is the presentation of uncertainty associated with the distributions. This is shown in three distinct ways: 1) inclusion of areas of model extrapolation (predictions occurring outside of the range of environmental conditions encountered by response variables) on all maps; 2) maps showing the frequency of Presences/Absences from the 10 cross-validation runs (values close to 1 give confidence in the Presences/Absences identified in the binary maps); and 3) maps of the average probability of the maximum frequency class (e.g. presence or absence) from the 10 cross-validation runs (areas with lower average probability within the same class can be associated with areas of uncertainty). For each model we provided maps of the predicted Presences/Absences based on a threshold of Sensitivity=Specificity, showing the areas of extrapolation and uncertainty from the 10 cross-validation runs. For all of the models there were areas of uncertainty at the border of the Presence/Absence prediction (e.g., Figure 39 for *Balticina* spp.). This occurrence is not unexpected as at these boundaries some of the model runs are likely to deviate in their predictions. More interesting are the uncertainties expressed within the areas of predicted presence. For example, in the Black Coral functional group model (Figure 55) within the area of predicted presence on Flemish Cap, the uncertainty shown as the average probability increases with depth. This is seen in other models such as in the sea pen *Funiculina* spp. (Figure 45), except uncertainty increases in both deeper and shallower water.

Models of the subgroups for the sponges and sea pens illustrated that there is potential for unequal protection of the VME Indicator taxa. For the sponges (Figure 56), the greatest area of predicted presence is seen in the Large-Sized Sponge functional group, as expected, and indicates that some of the VME indicators in this taxon that were not numerous enough to model independently have an influence on the distribution, particularly in the shallower areas of Flemish Cap. The model of the sponge grounds, which selected catches above the biomass threshold used for the KDE analyses (Kenchington et al., 2019), restricts the distribution to the slopes, some areas of which are predicted with high confidence, but most of which lie in areas of extrapolation. Some of those slope areas were previously validated with underwater camera observations (Kenchington et al., 2019). Comparison with previous sponge grounds models (Knudby et al., 2013a,b) shows very similar areas of predicted presence, despite the different data sets used to construct the models. This gives further confidence in the models themselves and in the validity of the sponge ground model presented here.



**Figure 56.** Random Forest species distribution model for the different sponge data sets showing binary maps of VME presence thresholded using Sensitivity=Specificity. A: Large-Sized Sponges functional group (see also Figure 3); B: Sponge grounds (see also Figure 8); C: Tetillidae (see also Figure 13); D: Polymastiidae (see also Figure 18); E: Astrophorina (see also Figure 23).



**Figure 57.** Random Forest species distribution model for the different sea pen data sets showing binary maps of VME presence thresholded using Sensitivity=Specificity. A: *Anthoptilum* spp. (see also Figure 33); B: *Funiculina* spp. (see also Figure 43); C: Sea Pen functional group (see also Figure 28); D: *Pennatula* spp. (see also Figure 48); E: *Balticina* spp. (see also Figure 38).



Compared to the sponges, there was more similarity among the individual models of sea pens (Figure 57), although there are differences among the genera. *Balticina* spp. extending into shallower waters, especially on the top of Flemish Cap, while *Pennatula* spp. appear in the Flemish Pass and along the slopes of the Tail of Grand Bank. Closer examination of this taxon with respect to the closed areas may show that it is not as well protected as the other genera which have wide bands of predicted presence in the region of the closed areas on Flemish Cap. Overlying these distributions may be an informative approach to discerning the relative protection afforded to the different genera. As they all have very different morphologies, this unequal protection may require different evaluations of significant adverse impacts (SAI) to be conducted.

For all VME functional groups and subgroups, records where there was uncertainty in the accuracy of identification were excluded from the modeling process. In future, those presence records could be used to independently validate model performance by evaluating their position relative to the binary Presence/Absence predictive surfaces. It would also be useful to superimpose the uncertainty predictions directly onto the binary predictive surfaces so that areas of higher uncertainty could be directly evaluated. This will be explored for presentation at the WG-ESA 2025 meeting. To that end, it was agreed that an intersessional meeting of those involved in the reassessment of the areas closed to bottom fishing be held in September 2025 in advance of the WG-ESA meeting in order to agree on the best way to present these results so that appropriate maps can be prepared in advance.

The model of the Astrophorina is very similar to that of the Sponge Grounds, that sub-order being the main constituent of the latter, however it also includes lower catch weights and so the distribution along the mid-slope of Flemish Cap is stronger compared to that of the sponge ground model. The last two groups, the families Tetillidae and Polymastiidae show differing areas of predicted presence with the former not predicted to be present in the northern part of Flemish Pass and for large areas of Flemish Cap in contrast with the Polymastiidae and with different variables influencing the predictions (Figures 11 and 16).

### Acknowledgements

The data collection of the EU Groundfish Surveys used in this paper has been funded by the EU through the European Maritime, Fisheries and Aquaculture Fund (EMFAF) within the Spanish Work Plan for the collection of data in the fisheries and aquaculture sectors in relation to the Common Fisheries Policy. The NEREIDA project was funded by the European Union through the NAFO Secretariat. We also thank Dr. Bárbara Neves and Vonda Wareham, Department of Fisheries and Oceans, Canada based at the Northwest Atlantic Fisheries Centre in St. John's, NL, for provision of the data from the Canadian surveys. This report was prepared by the Species Distribution Modeling Subgroup of WG-ESA. We thank all participants in the 2024 WG-ESA meeting for the valuable comments and contributions to this work.

### References

- Allouche, O., Tsoar, A. and Kadmon, R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* 43: 1223-1232.
- Böhner, J., and Antonić, O. 2009. Land-surface parameters specific to topo-climatology. In: Hengl, T., Reuter, H. [Eds.]: *Geomorphometry - Concepts, Software, Applications*. *Developments in Soil Science* 33: 195-226.
- Böhner, J., and Selige, T. 2006. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In: Boehner, J., McCloy, K.R., Strobl, J. (Eds.). *SAGA - Analysis and Modelling Applications*, Goettinger Geographische Abhandlungen, Goettingen, pp. 13-28.
- Breiman, L. 2001. Random Forests. *Mach. Learn.* 45: 5-32.
- Brenning, A., Bangs, D., and Becker, M. 2022. RSAGA: SAGA geoprocessing and terrain analysis. R package v. 1.4.0. <https://CRAN.R-project.org/package=RSAGA>.
- Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J. 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geosci. Model Dev.* 8: 1991-2007.
- Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J.J. 2007. Random Forests for classification in ecology. *Ecology* 88: 2783-2792.

- Desmet, P.J.J., and Govers, G. 1996. A GIS procedure for automatically calculating the USLE LS factor on topographically complex landscape units. *J. Soil Water Conserv.* 51(5): 427-433.
- Garrido, I., Sacau, M., Durán-Muñoz, P., Baldó, F., González-Costas, F., and González-Troncoso, D. 2023. Update on the analysis of VMS and Logbook data to study the bottom fishing footprint in the NAFO Regulatory Area: NEREIDA project. NAFO SCR Doc. 23/056. Serial No. N7486. <https://www.nafo.int/Portals/0/PDFs/sc/2023/scr23-056.pdf>.
- GEBCO Compilation Group (2024) GEBCO 2024 Grid. doi:10.5285/1c44ce99-0a0d-5f4f-e063-7086abc0ea0f.
- Hijmans, R. 2024. terra: Spatial Data Analysis. R package version 1.7-83. <https://CRAN.R-project.org/package=terra>.
- Hijmans, R. 2023. raster: Geographic data analysis and modelling. R package v. 3.6-26. <https://CRAN.R-project.org/package=raster>.
- Ilich, A. R., Misiuk, B., Lecours, V., and Murawski, S. A. 2023. MultiscaleDTM: An open-source R package for multiscale geomorphometric analysis. *Trans. GIS.* 27(4). <https://doi.org/10.1111/tgis.13067>.
- Kenchington, E., Lirette, C., Murillo, F.J., Beazley, L., and Downie, A. L. 2019. Vulnerable Marine Ecosystems in the NAFO Regulatory Area: Updated Kernel Density Analyses of Vulnerable Marine Ecosystem Indicators. NAFO SCR Doc. 19/058, Serial No. N7030. 68 pp.
- Kenchington, E., Murillo, F.J., Lirette, C., Sacau, M., Koen-Alonso, M., Kenny, A., Ollerhead, N., Wareham, V., and Beazley, L. 2014. Kernel density surface modelling as a means to identify significant concentrations of vulnerable marine ecosystem indicators. *PLoS ONE* 9(10): e109365.
- Knudby, A., Kenchington, E., Cogswell, A. T., Lirette, C.G., and Murillo, F.J. 2013a. Distribution modelling for sponges and sponge grounds in the northwest Atlantic Ocean. *Can. Tech. Rep. Fish. Aquat. Sci.* 3055: v + 73 p.
- Knudby, A., Kenchington, E., and Murillo, F.J. 2013b. Modeling the Distribution of *Geodia* Sponges and Sponge Grounds in the Northwest Atlantic. *PLoS ONE* 8: e82306. doi:10.1371/journal.pone.0082306.
- Knudby, A., Lirette, C., Kenchington, E., and Murillo, F.J. 2013c. Species Distribution Models of Black Corals, Large Gorgonian Corals and Sea Pens in the NAFO Regulatory Area. NAFO SCR Doc. 13/78, Serial No. N6276, 17 pp.
- Liaw, A., and Wiener, M. 2002. Classification and Regression by randomForest. *R News* 2: 18–22.
- Lundblad, E.R., Wright, D.J., Miller, J., Larkin, E.M., Rinehart, R., Naar, D.F., Donahue, B. T., Anderson, S.M., and Battista, T. 2006. A benthic terrain classification scheme for American Samoa. *Mar. Geod.* 29: 89-111.
- Murillo, F.J., Serrano, A., Kenchington, E. and Mora, J. 2016. Epibenthic assemblages of the Tail of the Grand Bank and Flemish Cap (northwest Atlantic) in relation to environmental parameters and trawling intensity. *Deep Sea Res. I* 109: 99-122.
- NAFO. 2024. Northwest Atlantic Fisheries Organization Conservation and Enforcement Measures 2024. NAFO/COM Doc. 24-01. Serial No. N7490. <https://www.nafo.int/Portals/0/PDFs/com/2024/comdoc24-01.pdf>
- NAFO. 2022. Report of the Scientific Council Working Group on Ecosystem Science and Assessment, 15 - 24 November 2022, Dartmouth, Nova Scotia, Canada. NAFO SCS Doc. 22/25.
- NAFO. 2020. Report of the 13th Meeting of the NAFO Scientific Council Working Group on Ecosystem Science and Assessment (WG-ESA). By WebEx 17-26 November 2020. NAFO SCS Doc. 20/23. Serial No. 7148. <https://www.nafo.int/Portals/0/PDFs/sc/2020/scs20-23.pdf>
- NAFO. 2019. Report of the 12<sup>th</sup> Meeting of the NAFO SC Working Group on Ecosystem Science and Assessment (WGESA) – November 2019. NAFO SCS Doc. 19/25. Serial No. N7027. <https://www.nafo.int/Portals/0/PDFs/sc/2019/scs19-25.pdf>

- NAFO. 2018. Report of the 11<sup>th</sup> Meeting of the NAFO Scientific Council Working Group on Ecosystem Science and Assessment (WGESA). NAFO SCS Doc. 18/23. Serial No. N6900.  
<https://www.nafo.int/Portals/0/PDFs/sc/2018/scs18-23.pdf>
- NAFO. 2017. Report of the 10<sup>th</sup> Meeting of the NAFO Scientific Council Working Group on Ecosystem Science and Assessment (WGESA). NAFO SCS Doc. 17/21. Serial No. N6774.  
<https://www.nafo.int/Portals/0/PDFs/sc/2017/scs17-21.pdf>
- Platt, T., and Sathyendranath, S. 1988. Oceanic Primary Production: Estimation by remote sensing at local and regional scales. *Science* 241: 1613-1620.
- R Development Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.
- R Development Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.
- Sacau, M., Durán-Muñoz, P., Garrido, I., and Baldó, F. 2020. Improvements in the methodology to study the bottom fishing footprint in the NRA using VMS and logbook data. NAFO SCS Doc. 20/069. Serial No. N7145.
- Sappington, J. M., Longshore, K. M., and Thompson, D. B. 2007. Quantifying landscape ruggedness for animal habitat analysis: A case study using bighorn sheep in the Mojave desert. *J. Wildl. Manage.* 71(5): 1419-1426.
- Sofaer, H.R., Jarnevich, C.S., Pearse, I.S., Smyth, R.L., Auer, S., Cook, G.L., Edwards, T.C. Jr., Guala, G.F., Howard, T.G., Morisette, J. T., and Hamilton, H. 2019. Development and delivery of species distribution models to inform decision-making. *BioScience* 69: 544-557.
- Wang, L., and Liu, H. 2006. An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. *Int. J. Geogr. Inf. Sci.* 20: 193-213.
- Wang, Z., Lu, Y., Greenan, B., Brickman, D., and DeTracey, B. 2018. An eddy-resolving North Atlantic model (BNAM) to support ocean monitoring. *Can. Tech. Rep. Hydrogr. Ocean Sci.* 327: vii + 18 pp.
- Weiss, A. 2001. Topographic Position and Landforms Analysis. Presented at the ESRI user conference, San Diego, CA.
- Yokoyama, R., Shirasawa, M., and Pike, R.J. 2002. Visualizing topography by openness: A new application of image processing to digital elevation models. *Photogramm. Eng. Rem. S.* 68: 251-266.
- Breiman, L. 2001. Random Forests. *Mach. Learn.* 45: 5-32.
- Zurrell, D., Franklin, J., Konig, C., Bouchet, P.J., Dormann, C.F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Fuisan, A., Lahoz-Monfort, J.J., Leitao, P.J., Park, D.S., Townsend Peterson, A., Rapacciuolo, G., Schmatz, D.R., Schroder, B., Serra-Diaz, J.M., Thuiller, W., Yates, K.L., Zimmermann, N.E., and Merow, C. 2020. A standard protocol for reporting species distribution models. *Ecography (Cop.)* 43: 1261-1277.

## Appendix

**Table A1.** At-sea Identification Nomenclature and Corresponding Number of Records for Each of Large-Sized Sponges, Sea Pens, and Black Corals Considered for the Response Data in the Species Distribution Models. \*Indicates taxon from the records of the Canadian DFO NL Multi-species Surveys (Table 3); All other taxa are as recorded from the EU Surveys undertaken by Spain and Portugal (Table 3).

At-Sea Identification for Large-Sized Sponges	Number of Records	At-Sea Identification for Sea Pens	Number of Records
Ancorinidae	4	Anthoptilum	889
Asconema	286	Anthoptilum grandiflorum*	135
ASCONEMA SP	102	ANTHOPTILUM GRANDIFLORUM*	1
Astrophorida	180	ANTHOPTILUM MURRAYI	1
Astrophorina	26	Anthoptilum murrayi*	1
ASTROPHORINA (ASTROPHORIDA)	15	ANTHOPTILUM SP	64
AXINELLIDAE	116	Anthoptilum sp.	172
Chondrocladia	21	Anthoptilum sp.*	1
Craniella	36	Anthoptilum spp	28
CRANIELLA SP	10	Balticina finmarchica (=Halipteris)	63
Craniella spp	3	Distichoptilum	3
DEMOSPONGIDAE	50	Distichoptilum gracile	59
ESPONJAS (PORIFERA)	99	Distichoptilum gracile*	1
Euplectellidae	2	DISTICHOPTILUM GRACILE	8
Forcepia sp.	6	Funiculina	8
Geodia	43	Funiculina quadrangularis	364
GEODIA SP.	6	Funiculina quadrangularis*	37
Geodia spp	3	FUNICULINIA QUADRANGULARIS	14
Geodiidae	172	Halipteridae	1
Isodictya palmata	1	Halipteris cf. christii	10
Isops spp	3	Halipteris christii	24
Mycale	50	Halipteris finmarchica	561
MYCALE SP	40	HALIPTERIS FINMARCHICA	29
Phakellia	2	Kophobelemnnon stelliferum	8
PHAKELLIA SP.	5	Pennatula	227
Phakellia spp	4	Pennatula aculeata	14
Pheronematidae	1	Pennatula aculeata*	25
Poecillastra compressa	1	PENNATULA ACULEATA/PHOSPHOREA Pennatuloidea sp. (SUPERFAMILY) formerly PENNATULACEA SPP.	26
Polymastiidae	668	(ORDER)*	138
Porifera	3726	Pennatula grandis	163
Porifera*	1074	PENNATULA GRANDIS	10
Radiella	45	Ptilella grandis (=Pennatula)*	39
RADIELLA (TRICHOSTEMMA) HEMISPHERICA	7	Pennatula phosphorea	7

<b>At-Sea Identification for Large-Sized Sponges</b>	<b>Number of Records</b>	<b>At-Sea Identification for Sea Pens</b>	<b>Number of Records</b>
Radiella hemisphaerica	207	Pennatula sp.	13
RADIELLA SP.	1	Pennatulacea	14
Rhizaxinella	10	Umbellula	105
RHIZAXINELLA SPP	1	Umbellula spp	2
STELLETA SP	1	Umbellula sp	5
STELLETA SPP	3	Taxon Name Not Provided	747
		<b>At-Sea Identification for Black Corals</b>	<b>Number of Records</b>
Stelletta	21	Antipatharia	57
Stryphnus	2	Antipatharia sp. (ORDER)*	7
Stryphnus sp.	24	Stauropathes arctica	174
STRYPHNUS SPP	13	STAUROPATHES ARCTICA	5
Stylocordyla	41	Stauropathes arctica*	10
Stylocordyla sp.	1	Leiopathes cf. expansa*	1
Sycettidae	9	Taxon Name Not Provided	111
Tentorium	28		
Tentorium semisuberites	350		
Tentorium sp.	4		
Tetillidae	183		
Thenea	75		
Thenea levis	3		
THENEA MURICATA	1		
THENEA SP	17		
Thenea spp	5		
Taxon Name Not Provided	2		

**Table A.2.** The Number of Records with Taxon Name Provided by VME Functional Group (Large-Sized Sponges, Sea Pens, and Black Corals) by Year.

<b>Year</b>	<b>Large-Sized Sponges</b>	<b>Sea Pens</b>	<b>Black Corals</b>
2002	34	0	1
2003	20	0	0
2004	38	0	0
2005	81	18	0
2006	244	54	4
2007	335	106	7
2008	312	9	0
2009	326	48	0
2010	302	49	2
2011	423	200	8
2012	464	225	6
2013	569	297	21
2014	444	314	29
2015	589	310	21
2016	583	243	12
2017	540	263	26
2018	546	248	14
2019	502	238	20
2020	214	95	19
2021	344	149	16
2022	320	183	19
2023	484	221	15

**Table A.3.** The Number of Records with Taxon Name Provided for the Large-Sized Sponge Subgroups by Mission. Values in red highlight 0 records.

	<b>Astrophorina</b>	<b>Tetillidae</b>	<b>Polymastiidae</b>	<b>Asconema</b>	<b>N Sets</b>
CAFC11	11	1	19	0	139
CAFC12	17	3	12	0	175
CAFC13	18	7	24	0	183
CAFC14	9	1	20	0	181
CAFC15	12	2	16	0	182
CAFC16	18	4	14	0	181
CAFC17	12	1	29	0	184
CAFC18	13	1	13	0	182
CAFC19	14	1	10	0	180
CAFC20	22	2	12	0	180
CAFC21	27	2	24	0	181
CAFC22	20	4	24	0	183
CAFC23	5	1	4	0	184
FN3L11	5	6	13	0	89
FN3L12	6	5	15	11	98
FN3L13	8	12	27	10	101
FN3L14	10	15	27	29	99
FN3L15	13	16	39	38	104
FN3L16	24	19	34	39	105
FN3L17	19	14	35	21	99
FN3L18	5	13	25	29	100
FN3L19	4	3	30	36	96
FN3L23	21	11	34	37	100
PLA11	13	7	14	0	122
PLA12	7	11	20	0	123
PLA13	10	7	17	8	124
PLA14	5	0	0	0	114
PLA15	12	12	12	11	122
PLA16	9	7	12	14	115
PLA17	12	8	13	8	112
PLA18	9	5	20	18	115
PLA19	6	5	10	14	115
PLA21	15	7	18	27	113
PLA22	5	6	12	22	114
PLA23	9	8	20	16	106

**Table A.4.** The Number of Records of Sea Pens for Different Subsets of the Data by Year. Shading of the column 2011 indicates the first year selected for response data in the SDMs.

TAXON	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total
<b>TOTAL SETS</b>	187	490	631	501	541	476	533	598	633	509	669	641	609	612	643	277	350	425	478	9803
Anthoptilum							89	82	109	115	110	99	104	99	82					889
Anthoptilum grandiflorum	3	23	38	3	14	8	4	2	4	4	5	5	2	2	9	1		9		136
Anthoptilum murrayi			1															1		2
ANTHOPTILUM SP																	15	23	26	64
Anthoptilum sp.											1					44	42	45	41	173
Anthoptilum spp																			28	28
<b>TOTAL ANTHOPTILUM</b>	<b>3</b>	<b>23</b>	<b>39</b>	<b>3</b>	<b>14</b>	<b>8</b>	<b>93</b>	<b>84</b>	<b>113</b>	<b>119</b>	<b>115</b>	<b>105</b>	<b>106</b>	<b>101</b>	<b>91</b>	<b>45</b>	<b>57</b>	<b>78</b>	<b>95</b>	<b>1292</b>
Balticina finmarchica (=Halipteris)	1	10	16	2	6	7	1	2	2	1	3	2	2	1	5			2		63
Halipteridae							1													1
Halipteris cf. christii																3	6		1	10
Halipteris christii							1		4	6	5	3	2	1	2					24
Halipteris finmarchica							27	39	42	55	48	52	57	62	49	22	41	42	54	590
<b>TOTAL HALIPTERIS</b>	<b>1</b>	<b>10</b>	<b>16</b>	<b>2</b>	<b>6</b>	<b>7</b>	<b>30</b>	<b>41</b>	<b>48</b>	<b>62</b>	<b>56</b>	<b>57</b>	<b>61</b>	<b>64</b>	<b>56</b>	<b>25</b>	<b>47</b>	<b>44</b>	<b>55</b>	<b>688</b>
Funiculina							6				2									8
Funiculina quadrangularis		4	19	2	5	2	20	23	49	56	26	20	29	25	32	13	26	37	21	409
FUNICULINIA QUADRANGULARIS																			6	6
<b>TOTAL FUNICULINA</b>	<b>0</b>	<b>4</b>	<b>19</b>	<b>2</b>	<b>5</b>	<b>2</b>	<b>26</b>	<b>23</b>	<b>49</b>	<b>56</b>	<b>28</b>	<b>20</b>	<b>29</b>	<b>25</b>	<b>32</b>	<b>13</b>	<b>26</b>	<b>37</b>	<b>27</b>	<b>423</b>
Pennatula							16	25	41	16	32	13	24	26	34					227
Pennatula aculeata					6	1	2	2	3	9	8	1	2		2			1	2	39
PENNATULA ACULEATA/PHOSPHOREA																	12	1	13	26
Pennatula grandis							13	19	10	25	28	18	14	15	10		3	7	11	173
Pennatula phosphorea																			7	7



TAXON	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total
Pennatula sp.																5	4	3	1	13
Ptilella grandis (=Pennatula)	3	4	15	2	5	1		1			1	2	1		2			2		39
<b>TOTAL PENNATULA</b>	<b>3</b>	<b>4</b>	<b>15</b>	<b>2</b>	<b>11</b>	<b>2</b>	<b>31</b>	<b>47</b>	<b>54</b>	<b>50</b>	<b>69</b>	<b>34</b>	<b>41</b>	<b>41</b>	<b>48</b>	<b>5</b>	<b>19</b>	<b>14</b>	<b>34</b>	<b>524</b>
Distichoptilum							1					2								3
Distichoptilum gracile			1				1	3	6	5	11	7	14	3	3	1		5	8	68
Kophobelemnon stelliferum									3	5										8
Umbellula							13	17	20	15	13	11	10	2	4					105
Umbellula sp		3	1		1															5
Umbellula spp																			2	2
<b>TOTAL OTHERS</b>	<b>0</b>	<b>3</b>	<b>2</b>	<b>0</b>	<b>1</b>	<b>0</b>	<b>15</b>	<b>20</b>	<b>29</b>	<b>25</b>	<b>24</b>	<b>20</b>	<b>24</b>	<b>5</b>	<b>7</b>	<b>1</b>	<b>0</b>	<b>5</b>	<b>10</b>	<b>191</b>
Pennatulacea							1	6	2	2			1	1		1				14
Pennatuloidea sp. (SUPERFAMILY) formerly PENNATULACEA SPP. (ORDER)	11	10	15		11	30	4	4	2		18	7	1	11	4	5		5		138
<b>SEA PENS WITH TAXA NAME</b>	<b>18</b>	<b>54</b>	<b>106</b>	<b>9</b>	<b>48</b>	<b>49</b>	<b>200</b>	<b>225</b>	<b>297</b>	<b>314</b>	<b>310</b>	<b>243</b>	<b>263</b>	<b>248</b>	<b>238</b>	<b>95</b>	<b>149</b>	<b>183</b>	<b>221</b>	<b>3270</b>
<b>NULL FOR TAXA</b>	<b>169</b>	<b>436</b>	<b>525</b>	<b>492</b>	<b>493</b>	<b>427</b>	<b>333</b>	<b>373</b>	<b>336</b>	<b>195</b>	<b>359</b>	<b>398</b>	<b>346</b>	<b>364</b>	<b>405</b>	<b>182</b>	<b>201</b>	<b>242</b>	<b>257</b>	<b>6533</b>
<b>SEA PENS FUNCTIONAL GROUP</b>	<b>41</b>	<b>194</b>	<b>260</b>	<b>153</b>	<b>200</b>	<b>180</b>	<b>200</b>	<b>226</b>	<b>298</b>	<b>315</b>	<b>310</b>	<b>243</b>	<b>263</b>	<b>248</b>	<b>238</b>	<b>95</b>	<b>149</b>	<b>183</b>	<b>221</b>	<b>4017</b>
<b>NULL FOR SEA PEN FUNCTIONAL GROUP</b>	<b>146</b>	<b>296</b>	<b>371</b>	<b>348</b>	<b>341</b>	<b>296</b>	<b>333</b>	<b>372</b>	<b>335</b>	<b>194</b>	<b>359</b>	<b>398</b>	<b>346</b>	<b>364</b>	<b>405</b>	<b>182</b>	<b>201</b>	<b>242</b>	<b>257</b>	<b>5786</b>

**Table A5.** R Code Used to Run the Species Distribution Models.

```
#####
#####      NAFO Sea pen distribution models 2024      #####
#####

#Load packages and library files
library(raster)
library(maptools)
library(randomForest)
library(ranger)
library(dplyr)
library(vtable)
library(pdp)
library(sf)
library(data. Table)
library(ggcorrplot)
library(patchwork)
library(caret)
library(dsmextra)

## Colour palette
cpl <- c('#d4ebe7','#cbbcbb','#f5f1f1','#172957','#66afad')
names(cpl) <- c('lt','dbe','lbe','dbl','dt')

OneB_theme <-
  ggplot2::theme(axis. Title.y = element_text(vjust=4, size=12,colour="black"),
    axis. Text.y = element_text(vjust=0.5, size=12,colour="black"),
    axis. Text.x = element_text(vjust=0.5, size=12,colour="black"),
    axis. Title.x = element_text(vjust=-4, size=12,colour="black"),
    strip.background = element_rect(fill=cpl['lbe']),
    strip. Text.x = element_text(size=12, face="bold"),
    panel.grid.major = element_line(colour=cpl['lbe']),
    panel.grid.minor = element_line(colour=cpl['lbe']),
    panel.background = element_rect(fill="white"),
    plot.margin = ggplot2::margin(0.5, 0.5, 1, 0.5, "cm"))

#####
##### STEP 1: IMPORT THE RESPONSE AND ENVIRONMENTAL DATA #####
#####

# Set working directory
wdir = ("C:/Users/AD06/OneDrive - CEFAS/VME/NAFO2024")
setwd(wdir)

# Set response variable title for model outputs
rvar = 'SeaPens'
```

```

### Environmental data -----

# Directory containing environmental rasters
rasterdir = "/ENVDATA/FINAL"

# List of raster files
predictorfiles = list.files(path = paste(wdir, rasterdir, sep=""), pattern = "\\.*Tif$", full.names=T)
# Now read the raster data (create a raster stack)
predictors = stack(predictorfiles,RAT=F)
# Confirm raster stack with all raster layers present
predictors
names(predictors)
names(predictors)[72:73] <- c("NRA_fishing_effort_1km","NRA_fishing_effort_5km")
# Plot raster files
plot(predictors)

# coordinate system for rasters
rprj = st_crs(predictors) # or if is not included in data set manually e.g. CRS('+proj=longlat +datum=WGS84
+no_defs')

#####

### Response data ---

# Directory containing response data
respdir = ("BIODATA/")

# Read and investigate csv
responsefile = read.csv(paste(respdir,"sea_pens.csv", sep=""), header=TRUE)
head(responsefile)
dim(responsefile)

# Response variable name
respvar = "VME_P_A"
# Coordinate variable names
xyvars = c("Start_Long_DD","Start_Lat_DD")

# Select response and coordinate columns
responsedata = data.frame(pa=responsefile[,respvar],x=responsefile[,xyvars[1]],
y=responsefile[,xyvars[2]])
head(responsedata)
dim(responsedata)

# Check response column is a factor
if (!is.factor(responsedata$pa)) {
  responsedata$pa <- as.factor(responsedata$pa)
}

```

```

# Check factor levels
levels(responsetdata$pa)

# Final data removing NAs
response = responsetdata[complete.cases(responsetdata),]

## Convert to spatial --
# Define coordinate system
pprj = CRS('+proj=longlat +datum=WGS84 +no_defs')
# Convert to sf
response_sp = st_as_sf(response,coords=c('x','y'),crs=pprj)
response_sp

# Check response and environmental variables are in the same coordinate system
if (pprj != rprj) {
  response_sp <- st_transform(response_sp, rprj)
}

st_write(response_sp,dsn = paste0('BIODATA/',rvar,'_Data.shp'),append = FALSE)

#####
#### STEP 2. EXTRACT THE VALUES OF PREDICTORS AT RESPONSE LOCATIONS ####
#####

### Extract the values from each predictor for each location in the response data, and put results in a new
dataframe
p.data = extract(predictors, response_sp)
sdata = data.frame(response, p.data) #adds all the extracted values to the existing dataframe
head(sdata)
prnames = colnames(p.data) #get names from new columns
str(sdata)

### Labels to use for environmental variables
# names of predictor columns
prnames
# Read csv file with two columns 'variable' with predictor column names and 'label' with labels to use for
plotting
varlabs <- read.csv("Models/varnames.csv")
# Convert to named vector
envlab <- varlabs$label
names(envlab) <- varlabs$variable
envlab

#####
#### STEP 3. QUALITY CONTROL ####
#####

```

```

# Ensure that all observations (locations) have data values for all variables
sdata = sdata[complete.cases(sdata),] #'complete.cases' command returns only those rows in the
dataframe that have non-NA
head(sdata)
dim(sdata)
summary(sdata)
# Values for all columns
str(sdata)
nrow(sdata)
missingdata = data.frame(response, p.data) #creates a new dataframe called 'missingdata'
missingdata = missingdata[!complete.cases(missingdata),] #!interested in rows that are NOT complete
cases
dim(missingdata)

#### THIS IS POTENTIAL TO INCLUDE IF WANT ####

## Increase prevalence by sub-sampling absence data --

# calculate number of presences
npres <- sdata %>%
  filter(pa=='1') %>%
  nrow()
# Calculate prevalence
preval <- npres/nrow(sdata)
preval

# If prevalence threshold of 5% is not met subsample absences to match it
if (preval <0.05) {
sdata <- sdata %>%
  filter(pa=='1') %>%
  bind_rows(sdata %>%
    filter(pa=='0') %>%
    sample_n(20*npres))
}

#####

# Save the data frame and labels
save(sdata,envlab,
  file = paste0('Models/',rvar,'_Data.RData'))

#####
#### STEP 4. VARIABLE ELIMINATION/SELECTION #####
#####

# This is for backwards compatibility for code for now
numvars <- prnames

```

```

# Define the number of class levels
numclass <- nlevels(sdata[[1]])

### Data exploration ----

## Summary statistics --
# All
sumtable(sdata, simple.kable = TRUE)
# Environmental variables for presences only
sumtable(sdata[sdata$pa=="1"], simple.kable = TRUE)

## Covariance of environmental variables --
# Correlation
corr <- cor(sdata[-1])
colnames(corr) <- envlab[colnames(corr)]
rownames(corr) <- envlab[rownames(corr)]
# Correlation plot
corplot <- ggcorrplot::ggcorrplot(corr, method='circle',type = 'upper',hc.order = TRUE)
corplot

### Preliminary full model to compare variable importance ----

## Build model --
prelRF <- randomForest(pa~.,
                      data=sdata,
                      importance=TRUE)
prelRF

## Extract importance and place in order --
full.importance <- data.frame(Predictor=rownames(prelRF$importance),prelRF$importance)
full.importance <- full.importance[order(full.importance[,MeanDecreaseGini],decreasing=T),]
full.importance

## Plot partial dependence --

# Set up object to save plot data to
plotdata <- NULL
predselnf <- numvars # for backwards compatibility for now

# Define class to plot
cl = '1'

# Loop through environmental variables to create data for partial plots
for (j in 1:length(predselnf)) {

  pdata <- partial(prelRF,pred.var = predselnf[j],which.class = cl,
                  plot = FALSE,train=sdata,grid.resolution=100,prob = TRUE)
}

```

```

predname <- predselnf[j]
temp <- data.frame(predvar=predselnf[j],class=cl,x=pdata[[1]],y=pdata[[2]])
plotdata <- rbind(plotdata,temp)

}

# Round values
plotdata.r <- plotdata %>%
  mutate(y= round(y, 1))

# Create list of partial plots
fullRP.list <- list()

# Loop through each predictor variable to plot partial dependence and add to the list
for (i in predselnf) {

fullRP.list[[i]] <- ggplot(plotdata[plotdata$predvar==i,],aes(x=x,y=y,col=class)) +
  geom_smooth(linewidth=0.8,se=FALSE,span = 0.3,col='#172957') +
  facet_wrap(~ predvar,scales = "free_x", ncol=3) +
  ylim(c(min(c(0,plotdata$y)),max(plotdata$y))) +
  theme(axis. Title.y = element_text(vjust=0.5, size=12,colour="black"),
        axis. Text.y = element_text(vjust=0.5, size=12,colour="black"),
        axis. Text.x = element_text(vjust=0.5, size=12,colour="black"),
        axis. Title.x = element_blank(),
        plot. Title = element_text(size=12,colour="white", face = "bold",vjust=2),
        strip.background = element_rect(fill="grey90"),
        strip. Text.x = element_text(size=12, face="bold"),
        panel.grid.major = element_line(colour="grey80"),
        panel.grid.minor = element_line(colour="grey80"),
        panel.background = element_rect(fill="white"),
        legend. Title = element_blank(),
        legend.key = element_rect(fill = NA),
        legend. Text = element_text(size=12,colour="black"))
}

# Write a pdf with all partial plots to check
pdf(paste0("Models/combined_plots_",rvar,".pdf"), width = 8, height = 11)
# Loop through the plots and arrange them in a 2x4 grid, 8 plots per page
for (i in seq(1, length(fullRP.list), by = 8)) {
  combined_plot <- wrap_plots(fullRP.list[i:min(i+7, length(fullRP.list))], ncol = 2, nrow = 4)
  print(combined_plot)
}
# Close the PDF device
dev.off()

#### Select uncorrelated variables to keep ----

## Correlation matrix with variables in order of full model importance --

```

```

vl <- full.importance[[1]] # Variable list
vl <- vl[vl %in% numvars] # Numeric variables only - for back compatibility
cr <- cor(sdata[,vl]) # correlation matrix
# Remove variables correlated to a higher importance variable
for(j in 1:length(cr[1,])){
  if (j == 1){
    pl <- c(names(cr[j,][1]),names( cr[j,][sqrt((cr[j,])^2)<0.65]))
    pl1 <- pl
  } else if (names(cr[j,])[j] %in% pl1){
    rem <- names(cr[j,-c(1:j)][sqrt((cr[j,-c(1:j)])^2)>0.65])
    if (length(rem) != 0L){
      pl <- pl[!pl %in% rem]
    }
  }
}
next
}
# Show list of selected variables
pl

## Calculate Variance Inflation Factors (vif) for selected variables --
# Null model vif function
corvif = function(dataz) {
  dataz <- as.data.frame(dataz)

  #vif part
  form <- formula(paste("fooy ~ ",paste(strsplit(names(dataz)," "),collapse=" + "))
  dataz <- data.frame(fooy=1 + rnorm(nrow(dataz)) ,dataz)
  lm_mod <- lm(form,dataz)

  cat("\n\nVariance inflation factors\n\n")
  print(data.frame(vif=car::vif(lm_mod)))
}

# Table kept variables and their vif
crval <- as.data.frame(pl)
crval$vif <- corvif(sdata[,pl])
crval

## This process can be redone to decrease allowed correlation if vif values
## remain too high

### Choose the set of variables to use in model ----

# List selected variables
predsel <- crval[[1]] # keeping this line for backwards compatibility

# List of all variables including response
clms <- c(names(sdata)[1],predsel)

```



```

# Data to use in model
mdata <- sdata[,clms]
summary(mdata)

### Save preliminary model and model data ----
save(preIRF,full.importance,crval,clms,mdata,plotdata.r,
file=paste0("Models/RF_Prelim_",rvar,".RData"))
save(sdata,mdata,clms,predsel,file = paste0("Models/",rvar,'_Data.RData'))

#####
#### STEP 5. BUILD MODEL & VALIDATION #####
#####

# We are building the model inside the 10 loops and predict with it and get the validation at the
# same time.

### Set up data and constants ----

# Set name of response variable to be used in results tables
tax = 'Sea pens'
# Set the name of the positive class
pcl = '1'
# Predictor variable constants
preds <- predsel # for backwards compatibility
facvars <- NULL # gear code when running full model
predselnf <- predsel # this code doesn't do anything here, is backwards compatibility

# Rename the response column (just to fit with existing code)
setnames(mdata,1,'resp')

### Set number of cross-validation runs required ---
nruns <- 10

### Set up training data ---

# Set up empty lists for looping through
train.sets <- list()
test.sets <- list()

# Split for 10 random subsets (list of row numbers), selects 90% of rows, keeping balance of classes equal,
times = 10 runs
trainIndexP <- createDataPartition(mdata$resp, p = .90, # Repeated sampling
times = nruns)
trainIndexK <- createFolds(mdata$resp,k=nruns) # K-fold

# Create 10 x separate train and tests sets using K-fold
for (j in 1:nruns){

```

```

train.sets[[j]] <- mdata[unname(unlist(trainIndexK[-j])),]
test.sets[[j]] <- mdata[trainIndexK[[j]],]

next}

# Save the datasets
save(train.sets,test.sets,file=paste0("Models/Train_Test_",rvar,".RData"))

#### Drop unnecessary layers from predictors ---
dr <- names(predictors)
dr <- dr[!dr %in% predsel]
predictors <- dropLayer(predictors, dr)
predictors

#### Set up lists and tables for outputs ---

ffs <- list() # Create empty list for forests
imps <- list() # create empty list for importances
res <- list() # create empty list for results
tshs = NULL # create empty object for a list of optimal thresholds
cvpred <- NULL # create empty object for a stack of model class predictions
cvpred.cps <- list() # create empty list for model probability predictions
plotdata <- NULL # create empty object for partial plot data

# Create empty table for collecting all model performance statistics
class.res.all <- data.frame(Name=character(0),
                           Run=character(0),
                           N=character(0),
                           Acc=numeric(0),
                           NIR=numeric(0),
                           P=numeric(0),
                           Kappa=numeric(0),
                           Sensitivity=numeric(0),
                           Specificity=numeric(0),
                           BalancedAcc=numeric(0),
                           TSS=numeric(0),
                           stringsAsFactors =F)

# Below code is a loop that runs x 10
# Before running the whole loop, test the code by running just 1 model (run j=1)

for (j in 1:10){

  train <- train.sets[[j]]
  test <- test.sets[[j]]

  ffs[[j]] <- randomForest(resp ~.,data=train,

```

```

      ntree=500,
      strata=resp,
      replace=FALSE,
      importance=T,
      keep.forest= T)

results <- as.data.frame(rownames(test)) #check results
results$actual <- test[[1]] #adds column to results - P/A as factor
results$PA <- as.numeric(as.character(test[[1]])) # changes factor to numeric

# Predict class with model j
results$predicted <- as.data.frame(predict(ffs[j],test))[,1] # outputs factor

# Predicted probability is of PRESENCE
results$predprob <- as.data.frame(predict(ffs[j],test,type='prob'))[,2] # Check second column is
presence!
names(results)[1] <- "id"

# Choose own optimal probability threshold: ID,observed, predicted. Various threshold methods,
# but 'Sens=Spec' returns equal amounts true and false positive classifications

require(PresenceAbsence)

opttsh <- results %>%
  dplyr::select(id,PA,predprob) %>%
  optimal.Thresholds(opt.methods = 'Sens=Spec') %>%
  pull(predprob)
tshs <- c(tshs, opttsh)

# Presence by threshold, adds column for optimal thresholded class
results <- results %>%
  mutate(optimal=as.factor(case_when(predprob>=opttsh ~ '1',
                                     TRUE~'0')))

# Calculate confusion matrix for predictions by model i
results.matrix <- confusionMatrix(results$optimal, results$actual,positive = '1',)
results.matrix

# Get overall accuracy measures for model validation run i
class.res.all[j,2] <- j
class.res.all[j,3] <- nrow(test)
class.res.all[j,4] <- results.matrix[[3]][[1]]
class.res.all[j,5] <- results.matrix[[3]][[5]]
class.res.all[j,6] <- results.matrix[[3]][[6]]
class.res.all[j,7] <- results.matrix[[3]][[2]]
class.res.all[j,8] <- results.matrix[[4]][[1]]
class.res.all[j,9] <- results.matrix[[4]][[2]]

```

```

class.res.all[j,10] <- results.matrix[[4]][[11]]
class.res.all[j,11] <- results.matrix[[4]][[1]] + results.matrix[[4]][[2]] - 1

class.res.all$Name <- tax
class.res.all

imps[[j]] <- list(round(randomForest::importance(ffs[[j]]), 2))

require(pdp)

for (p in 1:length(predselnf)) {

  pdata <- partial(ffs[[j]],pred.var = predselnf[p],which.class = pcl,
    plot = FALSE,train=mdata,grid.resolution=100,prob = TRUE)
  predname <- predselnf[p]
  temp <- data.frame(Name=tax,run=j,predvar=predselnf[p],class=pcl,x=pdata[[1]],y=pdata[[2]])
  plotdata <- rbind(plotdata,temp)

}

## Predict rasters
rnn <- paste0('Run',j) # Set layer name
# Check if there is already a raster stack - if not create one
if (is.null(cvpred)){
  # Probabilities for each class
  cvpred.cps[[rnn]] <- predict(predictors,ffs[[j]],type='prob',index=1:numclass)
  # Presence/Absence raster from applying to the threshold to presence probability
  cvpred <- stack(cut(cvpred.cps[[rnn]]$layer.2,breaks=c(-1,tshs[j],1)))
  cvpred <- cvpred - 1
  names(cvpred) <- rnn
} else {
  # Probabilities for each class
  cvpred.cps[[rnn]] <- predict(predictors,ffs[[j]],type='prob',index=1:numclass)
  # Presence/Absence raster from applying to the threshold to presence probability
  tmp1 <- cut(cvpred.cps[[rnn]]$layer.2,breaks=c(-1,tshs[j],1))-1
  names(tmp1) <- rnn
  cvpred <- addLayer(cvpred,tmp1)
}

next
}

# Save models and validation results, plot data and importances
save(ffs,plotdata,class.res.all,imps,file = paste0('Models/RF_Results_',rvar,'.RData'))

##### Look at the Validation statistics ----
require(matrixStats)

```

```

# Calculate averages and standard deviations for validation statistics
callavevalsB <- colMeans(class.res.all[,4:11])
callsdvalsB <- colSds(as.matrix(class.res.all[,4:11]))

# Combine values in a table
BCallvalsT <- data.frame(Accmean=round(callavevalsB[1],2),
  Accsd=round(callsdvalsB[1],2),
  Pmean=round(callavevalsB[3],2),
  Psd=round(callsdvalsB[3],2),
  Kmean=round(callavevalsB[4],2),
  Ksd=round(callsdvalsB[4],2),
  Sensmean=round(callavevalsB[5],2),
  Senssd=round(callsdvalsB[5],2),
  Specmean=round(callavevalsB[6],2),
  Specsds=round(callsdvalsB[6],2),
  BAmean=round(callavevalsB[7],2),
  BASd=round(callsdvalsB[7],2),
  TSSmean=round(callavevalsB[8],2),
  TSSsd=round(callsdvalsB[8],2))

# Rename columns
names(BCallvalsT) <- c("Accmean", "Accsd", "Pmean", "Psd", "Kmean", "Ksd",
  "Sensmean", "Senssd", "Specmean", "Specsd",
  "BAmean", "BASd", "TSSmean", "TSSsd")

# Print table
BCallvalsT

asg.perf <- data.Table(N = nrow(train),
  'Sensitivity' = paste(BCallvalsT$Sensmean, '/u00B1', BCallvalsT$Senssd),
  'Specificity' = paste(BCallvalsT$Specmean, '/u00B1', BCallvalsT$Specsd),
  'Kappa' = paste(BCallvalsT$Kmean, '/u00B1', BCallvalsT$Ksd),
  'Balanced Accuracy' = paste(BCallvalsT$BAmean, '/u00B1', BCallvalsT$BASd),
  'TSS' = paste(BCallvalsT$TSSmean, '/u00B1', BCallvalsT$TSSsd))

asg.perf[, data.Table(t(.SD), keep.rownames=TRUE,)] %>%
  kbl('html', digits = 2, escape = FALSE, col.names = c('Statistic', 'Mean /u00B1 SD'),
  caption='Performance statistics') %>%
  kable_classic(full_width = F, position = "left", fixed_thead = T) %>%
  row_spec(0, bold = T) %>%
  column_spec(1:2, width = "3cm")

#### Plot variable importance ----

# Importance plot
imppl <- data.Table(Var=rownames(imps[[1]][[1]]))

```

```

for (i in 1:10){

  imppl <- cbind(imppl,as.data. Table(imps[[i]][[1]]),MeanDecreaseGini)

}

setnames(imppl,c('Var','Imp1','Imp2','Imp3','Imp4','Imp5','Imp6','Imp7','Imp8','Imp9','Imp10'))

imppl[,
  c("Mean",'Sd','Se') :=
  .(rowMeans(.SD, na.rm = TRUE),
  apply(.SD, 1, sd, na.rm = TRUE),
  apply(.SD, 1, plotrix::std.error, na.rm = TRUE)),
  .SDcols = 2:11]

imppl[,Var:=factor(Var,levels=Var[order(Mean)])]

impplot <- ggplot(imppl,(aes(x=Var,y=Mean))) +
  geom_bar(stat = 'identity',fill='#66afad', col='#172957',) +
  scale_x_discrete(labels=envlab[levels(imppl$Var)]) +
  geom_linerange(inherit.aes=FALSE,
    aes(x=Var, ymin=Mean-Se, ymax=Mean+Se),
    colour='#172957', alpha=0.9, linewidth=1.3) +
  ylab(label = 'Mean decrease in Gini coefficient') +
  coord_flip() +
  theme(axis. Title.y = element_blank(),
    axis. Text.y = element_text(vjust=0.5,hjust = 1, size=12,colour="black"),
    axis. Text.x = element_text(vjust=0.5, size=12,colour="black"),
    axis. Title.x = element_text(vjust=-4, size=12,colour="black"),
    plot. Title = element_text(size=12,colour="white", face = "bold",vjust=2),
    strip.background = element_rect(fill="grey90"),
    strip. Text.x = element_text(size=12, face="bold"),
    panel.grid.major = element_line(colour="grey80"),
    panel.grid.minor = element_line(colour="grey80"),
    panel.background = element_rect(fill="white"),
    legend. Title = element_blank(),
    legend.key = element_rect(fill = NA),
    legend. Text = element_text(size=12,colour="black"),
    plot.margin = ggplot2::margin(0.5, 0.5, 1, 0.5, "cm"),)
impplot

ggsave(impplot,
  filename=paste0('Models/',rvar,'_VariableImportance.png'),
  device = 'png', width = 15, height=16, units='cm', dpi=300, scale=1)

#### Partial response plots ----

pplotdata <- plotdata # this is here if need to make any changes to plotdata

```

```

# Create list for plots
cvRP.list <- list()

# Loop through predictors to create plots
for (i in predsel) {

  mxy <- max(pplotdata[pplotdata$class==pcl,'y'])

  cvRP.list[[i]] <- ggplot(pplotdata[pplotdata$predvar==i &
pplotdata$class==pcl,],aes(x=x,y=y,group=run)) +
  geom_smooth(method='loess',linewidth=0.01,se=FALSE,span = 0.2,col='#66afad') +
  geom_smooth(inherit.aes=FALSE,aes(x=x,y=y),
    method='loess',linewidth=0.9,se=FALSE,span = 0.2,col='#172957') +
  facet_wrap(~ predvar,scales = "free_x",ncol =3,labeller = labeller(predvar = envlab)) +
  ylim(c(min(c(0,pplotdata$y)),mxy)) +
  OneB_theme +
  theme(axis.Title.y = element_blank(),
    axis.Text.y = element_text(vjust=0.5, size=12,colour="black"),
    axis.Text.x = element_text(vjust=0.5, size=12,colour="black"),
    axis.Title.x = element_blank(),
    legend.Title = element_blank(),
    legend.position = 'none',
    legend.key = element_rect(fill = NA),
    legend.Text = element_text(size=12,colour="black"),
    plot.margin = unit(c(0.5, 0.5, 0,0), "cm"))
}

length(cvRP.list)

# Layout of all plots
cvRP <- wrap_plots(cvRP.list) + plot_layout(ncol=4)
cvRP

# Save the plot data
save(results,asg.perf,imppl, impplot,pplotdata,cvRP,cvRP.list,
file=paste0("Models/RF_",rvar,"_Results_Summary.RData"))

#### Raster outputs ----

### Create a raster stack for spatial confidence results
ROutput <- stack()

### Calculate most frequent class and its frequency
# Most frequent class - change to 0 and 1 for absence and presence
MaxClass <- modal(cvpred,freq=FALSE)
ROutput <- addLayer(ROutput,MaxClass)
# Frequency of most frequent class (fraction of runs)

```

```

MaxClassF <- modal(cvpred,freq=TRUE)/nruns
ROutput <- addLayer(ROutput,MaxClassF)

#### Calculate average probabilities for classes
classsums <- Reduce("+", cvpred.cps)
AvePclass <- classsums / nruns

#### Find average probability of maximum frequency class
MaxClassAveProb <- stackSelect(AvePclass, MaxClass+1)
ROutput <- addLayer(ROutput,MaxClassAveProb)

#### Calculate new layer for frequency x probability
CombConf <- MaxClassF * MaxClassAveProb
ROutput <- addLayer(ROutput,CombConf)

#### number of models predicting presence
cvPA <- stack(cvpred)
cvSum <- raster::calc(cvPA,sum)
ROutput <- addLayer(ROutput,cvSum)

#### Rename layers
names(ROutput) <- c("MaxClass","MaxClassF","MaxClassAveProb","CombConf","cvSum")

#### Plot layers
plot(ROutput)

#### Export Raster
raster::writeRaster(ROutput$MaxClass,      paste0("Models/",rvar,"_raster_output_maxclass.  Tif"),
format="GTiff",overwrite=T)
raster::writeRaster(ROutput$MaxClassF,    paste0("Models/",rvar,"_raster_output_maxclassf.  Tif"),
format="GTiff",overwrite=T)
raster::writeRaster(ROutput$MaxClassAveProb,
paste0("Models/",rvar,"_raster_output_maxclassaveprob. Tif"), format="GTiff",overwrite=T)
raster::writeRaster(ROutput$CombConf,    paste0("Models/",rvar,"_raster_output_combconf.  Tif"),
format="GTiff",overwrite=T)
raster::writeRaster(ROutput$cvSum,      paste0("Models/",rvar,"VME_raster_output_cvsum.  Tif"),
format="GTiff",overwrite=T)

#### Save raster stack to R workspace
save(AvePclass,ROutput,file= paste0("Models/",rvar,"_raster_output_all.RData"))

#### Extrapolation areas
library(dsmextra)

covariates.names <- predsel

allpred <-rasterToPoints(predictors)
p.data.all = data.frame(allpred)

```



```

aftt_crs <- sp::CRS("+proj=utm +zone=23 +datum=NAD83 +units=m +no_defs")

extrapolation.area <- compute_extrapolation(samples = mdata,
      covariate.names = predsel,
      prediction.grid = p.data.all,
      coordinate.system = aftt_crs)

plot(extrapolation.area$rasters$ExDet$analogue) # analogue areas
plot(extrapolation.area$rasters$ExDet$univariate) # univariate extrapolation
plot(extrapolation.area$rasters$ExDet$combinatorial) # combinatorial extrapolation
plot(extrapolation.area$rasters$mic$analogue) # most important variables causing analogue conditions
plot(extrapolation.area$rasters$mic$univariate) # most important variables causing univariate
extrapolation
plot(extrapolation.area$rasters$mic$combinatorial) # most important variables causing combinatorial
extrapolation
### Export Raster
writeRaster(extrapolation.area$rasters$ExDet$analogue, paste0("Models/",rvar,"_ext.analogue. Tif"),
format="GTiff",overwrite=T)
writeRaster(extrapolation.area$rasters$ExDet$univariate, paste0("Models/",rvar,"_ext.univariate. Tif"),
format="GTiff",overwrite=T)
writeRaster(extrapolation.area$rasters$ExDet$combinatorial,
paste0("Models/",rvar,"_ext.combinatorial. Tif"), format="GTiff",overwrite=T)
writeRaster(extrapolation.area$rasters$mic$analogue, paste0("Models/",rvar,"_mic.analogue. Tif"),
format="GTiff",overwrite=T)
writeRaster(extrapolation.area$rasters$mic$univariate, paste0("Models/",rvar,"_mic.univariate. Tif"),
format="GTiff",overwrite=T)
writeRaster(extrapolation.area$rasters$mic$combinatorial, paste0("Models/",rvar,"_mic.combinatorial.
Tif"), format="GTiff",overwrite=T)

```