Northwest Atlantic **Fisheries Organization**

Serial No. N6276 **NAFO SCR Doc. 13/078**

SC ECOSYSTEM SCIENCE AND ASSESSMENT WORKING GROUP (WGESA) – NOVEMBER 2013

Species Distribution Models of Black Corals, Large Gorgonian Corals and Sea Pens in the NAFO Regulatory Area

A. Knudby[1], C. Lirette[2], E. Kenchington[2], F.J. Murillo[2]

[1] Department of Geography, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6

[2] Department of Fisheries and Oceans, Bedford Institute of Oceanography, 1 Challenger Drive, Dartmouth, Nova Scotia, Canada B2Y 4A2

Abstract

Random forest generated species distribution models have been produced for black corals, large gorgonian corals and sea pens in the NAFO regulatory area using a suite of 23 poorly correlated environmental variables. All models performed well, producing cross-validated AUC values of 0.937, 0.885 and 0.888 respectively. Prediction surfaces for the three species groups produced clearly defined areas of high occurrence probability. These can be used to identify areas for conservation where black corals and the vulnerable marine ecosystem indicators, the large gorgonian corals and sea pens, are likely to occur.

Introduction

Species distribution models (SDMs) predict the probability of occurrence of a species or habitat using a suite of environmental predictor variables. SDMs can be trained in data rich areas and extrapolated to predict occurrence in data poor areas. They have particular relevance to the protection of vulnerable marine ecosystems (VMEs) under the FAO International Guidelines for the Management of Deep-sea Fisheries in the High Seas (FAO, 2009) that calls for the identification of "*areas where VMEs are known **or likely to occur**".*

Recently Knudby *et al.* (2013a,b) applied such models to the large-size sponge species (*Geodia* spp.) and sponge grounds in the northwest Atlantic, including the NAFO regulatory area. Here we take advantage of the extensive environmental data set created for those analyses and apply the models to the black corals and VME indicator species of large gorgonian corals and sea pens. The black corals in the NAFO regulatory area generally occur as isolated individuals and do not form dense aggregations. They are

mostly a single species, *Stauropathes arctica*. The gorgonian corals do aggregate and comprise a number of species and genera. The sea pens also aggregate and are comprised of a number of species, most notably *Anthoptilum grandiflorum*, *Halipteris finmarchica* and *Pennatula aculeata* (Murillo *et al*., 2010). For the sea pens, the model was also run to determine whether sea pen field VME (locations with catches greater than 1.4 kg) can be predicted from environmental variables as was shown for sponge grounds (Knudby *et al*., 2013 a, b). Because the data coverage in the NRA is very extensive, the value of these models is greatest in areas where there is no or little survey coverage due to the nature of the bottom (e.g., rocky) or depth. Such locations where the models also predict high probability of VME occurrence should be the focus of future *in situ* camera data collection to validate the model outputs. Those within the same depth range of the training data (to approximately 1800 m) are likely to have the greatest likelihood of representing reality.

Methods

Environmental Data

Collinearity among variables (Graham, 2003) was reduced through selective removal of correlated variables (Knudby *et al*., 2013a), leaving a set of 23 environmental predictor variables (Table 1). Bathymetric data were derived from the 30 arc-second General Bathymetric Chart of the Oceans (GEBCO) data (BODC 2009) from which the slope was derived using ArcGIS's Spatial Analyst tool. Data on surface chlorophyll-*a* concentrations were derived from Level 3 SeaWiFS data for the period January 2001 – December 2010, while shear stress at the seafloor, as well as temperature, salinity and current for the sea surface and seafloor were derived from the GLORYS2V1 ocean reanalysis at ¼ ° resolution. All data layers were standardized to 0.017 degree cell resolution as detailed in Knudby *et al*. (2013a).

Presence/Absence Data on Black and Large Gorgonian Corals and Sea Pens

Data on the presence and absence of black corals, large gorgonian corals and sea pens were drawn from European Union (EU)-Spanish and Canadian DFO-Newfoundland and Labrador (NL) research vessel surveys as well as from NEREIDA rock and scallop dredge samples and benthic imagery (Table 2). For black coral the data consisted of 163 presences and 4185 absences, primarily drawn from the trawl surveys. For the large gorgonian corals there were 1643 presences and 3540 absences. There were 24 data cells that had both presence and absence records within them coming from the photo transects which provided very detailed records over small spatial scales causing the SDM to perform poorly. This was corrected by allowing each 0.017 degree data cell to have only one record. Any presence in the cell was used to indicate presence for the model, even if absences were also recorded. Absence in a cell meant that there were no presences. This created a large gorgonian coral response data set of 214 presences and 3192 absences. The sea pen data included 1327 presence records and 2183 absences, with 3409 of these records coming from the trawl survey data. For the sea pen VME (sea pen fields), only trawl survey data where catches greater than 1.4 kg were used. These constituted 31 presences and 3378 absences.

Table 1. Subset of environmental predictor variables used in the SDMs with measurement unit, data source and native resolution.

Predictor variable	**Quantification**	**Unit**	**Data source**	**Native resolution**
Depth		m	GEBCO	30”
Slope		degrees	GEBCO	30”
Annual chlorophyll	Min	mg m^{-3}	OceanColor	9 km
Winter chlorophyll	Range, Min, Mean	mg m^{-3}	OceanColor	9 km
Spring chlorophyll	Range	mg m^{-3}	OceanColor	9 km
Summer chlorophyll	Range, Min	mg m^{-3}	OceanColor	9 km
Fall chlorophyll	Range, Min	mg m^{-3}	OceanColor	9 km
Surface temperature	Range, Min	°C	GLORYS	¼°
Surface salinity	Range, Max	PSU	GLORYS	¼°
Bottom temperature	Min, Max	°C	GLORYS	¼°
Bottom salinity	Max	PSU	GLORYS	¼°
Surface current	Range, Min, Mean	m/s	GLORYS	¼°
Shear	Range, Min	Pa	GLORYS	¼°

Table 2. Sources and numbers for the presence (P) and absence (A) of response variables (Black corals, Large gorgonian corals and Sea pens) used in the SDMs.

Data source	Period	Black corals		Large gorgonian corals		Sea pens		Sea pens VME	
		P	A	P	A	P	A	P	A
Spanish/EU and DFO-NL Research Vessel Surveys	2000 – 2013	148	4097	174	3540	1310	2099	31	3378
NEREIDA Rock dredge and scallop gear	2009 – 2010	7	88	21	74	11	84		
NEREIDA Benthic imagery	2009 – 2010	8	-	1448	-	6	-		

Modeling

The methodology used in this study follows that of Knudby *et al*. (2013a) for their "FC" (i.e., Flemish Cap) area. A random forest model (Breiman *et al*., 1984) with variable elimination using jackknifing was applied with a spatial extent of the NAFO regulatory area to 2500 m depth. Random forests model results can vary from run to run due to the randomization procedure. Repeated runs stabilize the prediction surface, however, it is critical to examine the stability of the prediction surfaces created under different AUC thresholds. Variable importance as well should only be considered in a relative context and not be over-interpreted (Shih, 2011).

Jackknifing was used to eliminate predictor variables that did not contribute substantially to improving model fit, which was assessed using 10-fold cross-validation (Knudby *et al*., 2013b). Model fit was assessed using the commonly used Area Under the Curve (AUC) value. The AUC value is calculated as the area under the Receiver Operating Characteristic (ROC), which depicts the relationship between true positives and false positives for binary classifiers (Fawcett, 2006). AUC values quantify the likelihood that such a classifier will predict a higher presence probability for a randomly chosen presence location than for a randomly chosen absence location. As such, AUC values range from 0 to 1, with 0.5 indicating a classifier performance no better than random. Starting from a model using all predictor variables, each individual predictor was removed, and the change in model fit was assessed. The predictor that, when removed, resulted in the largest increase, or the smallest decrease, in AUC value was then removed from the set of predictor variables. This process was repeated until no variable could be removed without reducing the AUC by more than a set threshold value. The threshold selected was subjective and chosen to achieve a good balance between the AUC fit and the number of predictor variables in the model. For the black coral SDM variable elimination continued until the reduction in the AUC value was less than 0.005. For the large gorgonian corals a value of 0.0085 was chosen after rejecting values of

0.0075, 0.005 and 0.01. For the sea pens a value of 0.005 was chosen after rejecting values of 0.00625, 0.0075, and 0.01. In all cases the prediction surfaces created with the different thresholds were nearly identical when inspected visually.

Model development was done in R (R Core Development Team, 2012) using the "randomForest" package (Liaw and Wiener, 2002). AUC values were used to measure model performance (Fawcett, 2006) with values greater than 0.9 considered indicative of excellent model fits. All AUC values were based on 10-fold cross-validation repeated 10 times, as outlined in Knudby *et al*. (2013a, b).

Partial dependence plots were produced for all variables used in the SDM. Single-variable plots were produced by training the model using only a single environmental predictor. Multi-variable plots were produced by training the model using all the available predictors, and then observing the predictions made by this model when all predictors but one were kept at their mean value, while using the range of input values that exist for the predictor (Knudby *et al*. , 2013a). Variable importance for each predictor was assessed following Knudby *et al*. (2010) and Knudby *et al*. (2013a,b).

Results and Discussion

Black Coral

The AUC model fit was 0.937, which is considered excellent. Seven environmental variables were included in the model (Table 3) with maximum surface salinity and mean bottom temperature being the most important. The partial dependence plots suggest that the black corals prefer fully saline water with mean bottom water temperature of less than 4 °C (Figure 1).

Table 3. Predictor variable combinations remaining after the AUC-based variable elimination, for the NRA area under a random forest species distribution model for black coral (AUC = 0.937).

Response variable	**Predictor Variables in Order of Importance**	**Variable Importance**
Black Corals	Surface salinity, maximum	0.077
	Bottom temperate, mean	0.060
	Surface temperature, minimum	0.054
	Depth	0.053
	Winter Chl *a*, mean	0.031
	Bottom shear, minimum	0.013
	Bottom salinity, maximum	0.011

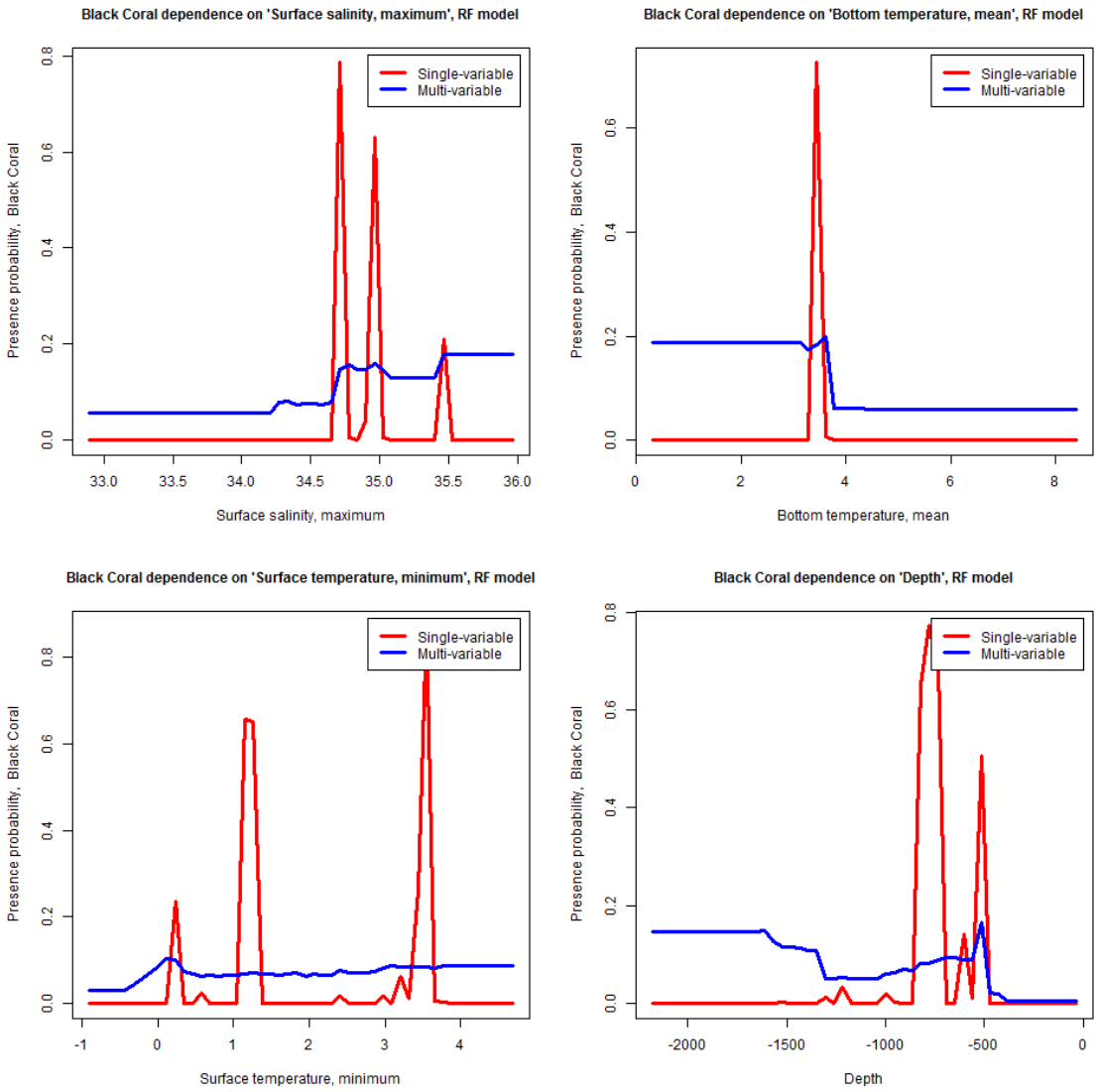

Figure 1. Partial dependence plots for the four predictor variables contributing most to the model fit (Table 3). Red lines are from a model trained on a single variable. Blue lines indicate performance when all other variables are held at their mean values (multi-variable).

The prediction surface shows an excellent fit with the presence/absence data used to train the model (Figure 2) with only a few records on the tail of Grand Bank not fitting with prediction. The highest probability of occurrence is around the Flemish Cap where a tight predicted distribution is seen between 500 and 1000 m water depth (Figure 3). Maximum probability of occurrence is found on the eastern edge of that distribution.

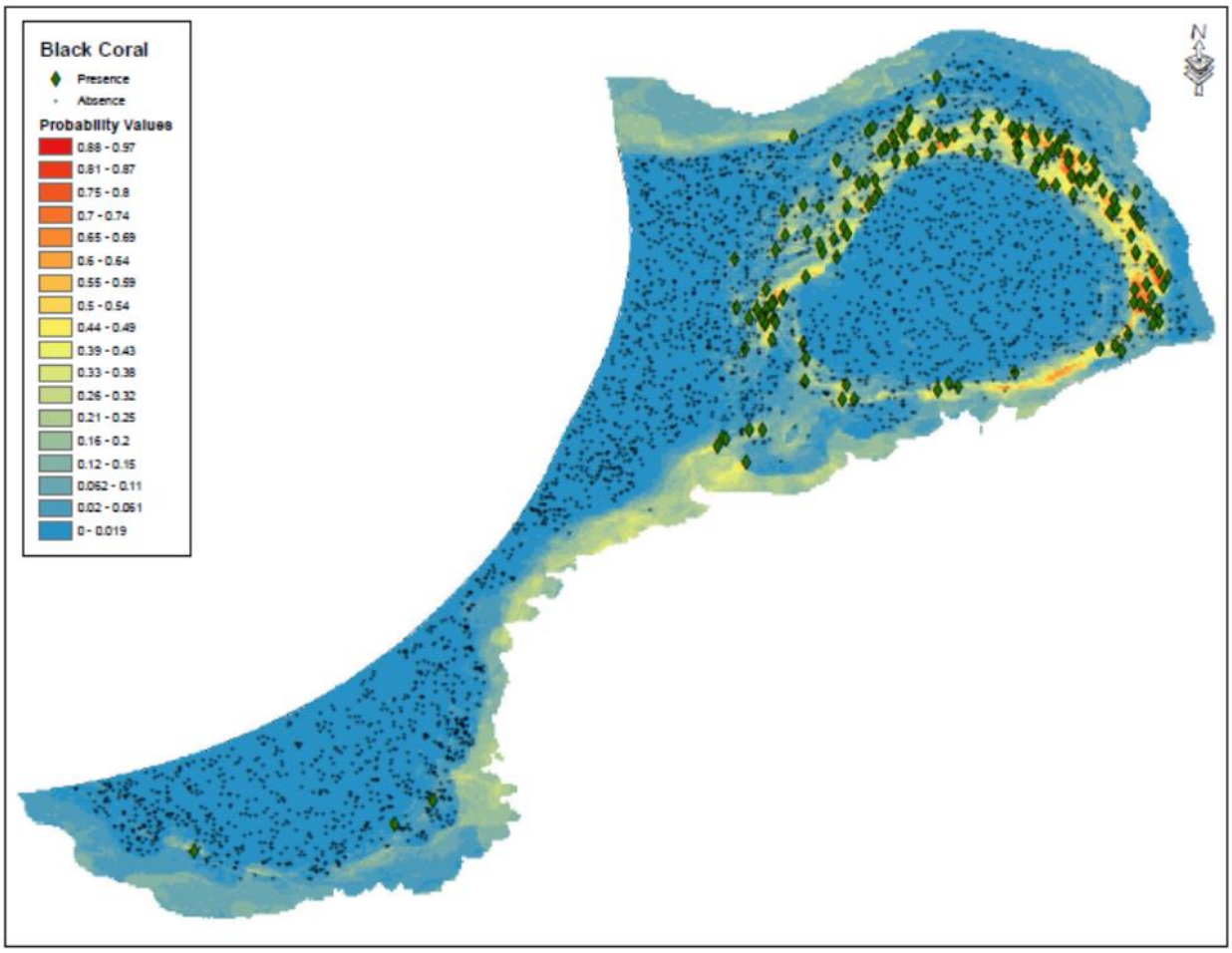

Figure 2. The location of known black coral presence and absence data overlain on the prediction surface produced from the random forest species distribution model.

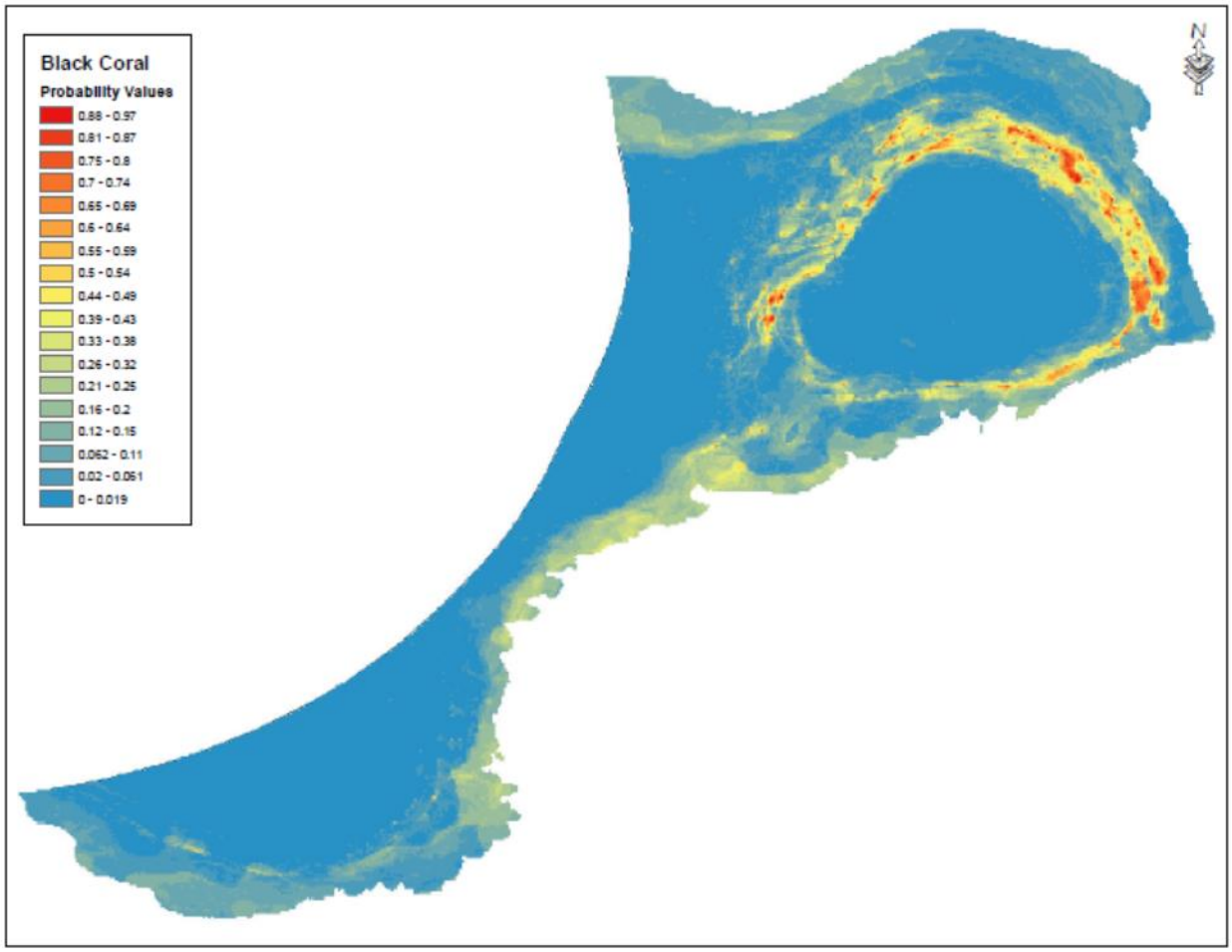

Figure 3. The probability of occurrence of black coral in the NAFO regulatory area produced with a random forest species distribution model.

Large Gorgonian Corals

The AUC model fit was 0.885, which is considered to be a very good fit. Eleven environmental variables were included in the model (Table 4) with minimum bottom shear and depth being the most important. The partial dependence plots for the two most important variables influencing the SDM for the large gorgonian corals are not highly informative, except to show relatively low probability of occurrence in shallow areas (Figure 4).

The prediction surface shows a very good fit to the known presences and absences although a number of the presences are found in areas of only moderate probability of occurrence (Figure 5). The southern and eastern slopes of Flemish Cap have very high probability of occurrence of large gorgonian corals (Figure 6).

Table 4. Predictor variable combinations remaining after the AUC-based variable elimination, for the NRA area under a random forest species distribution model for large gorgonian corals (AUC = 0.885).

Response variable	Predictor Variables in Order of Importance	Variable Importance
Large Gorgonian Corals	Bottom shear, minimum	0.049
	Depth	0.045
	Bottom temperature, mean	0.019
	Summer Chl *a*, minimum	0.019
	Sea surface salinity, maximum	0.014
	Bottom temperature, minimum	0.014
	Surface current, mean	0.013
	Slope	0.013
	Annual Chl *a*, minimum	0.005
	Spring Chl *a*, minimum	0.005
	Fall Chl *a*, range	0.004

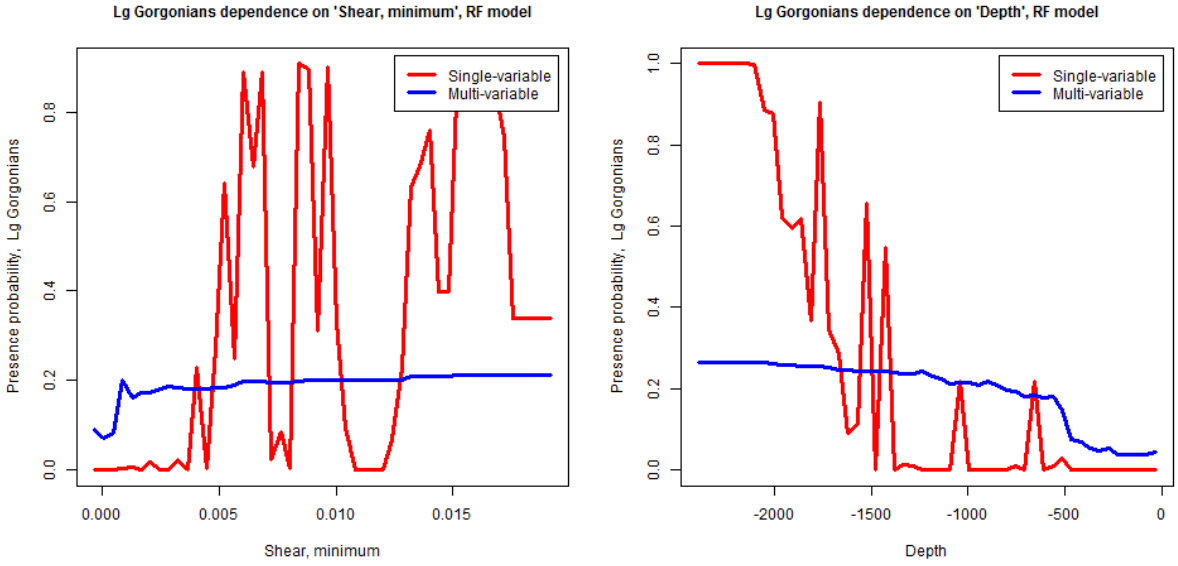

Figure 4. Partial dependence plots for the two predictor variables contributing most to the model fit (Table 4). Red lines are from a model trained on a single variable. Blue lines indicate performance when all other variables are held at their mean values (multi-variable).

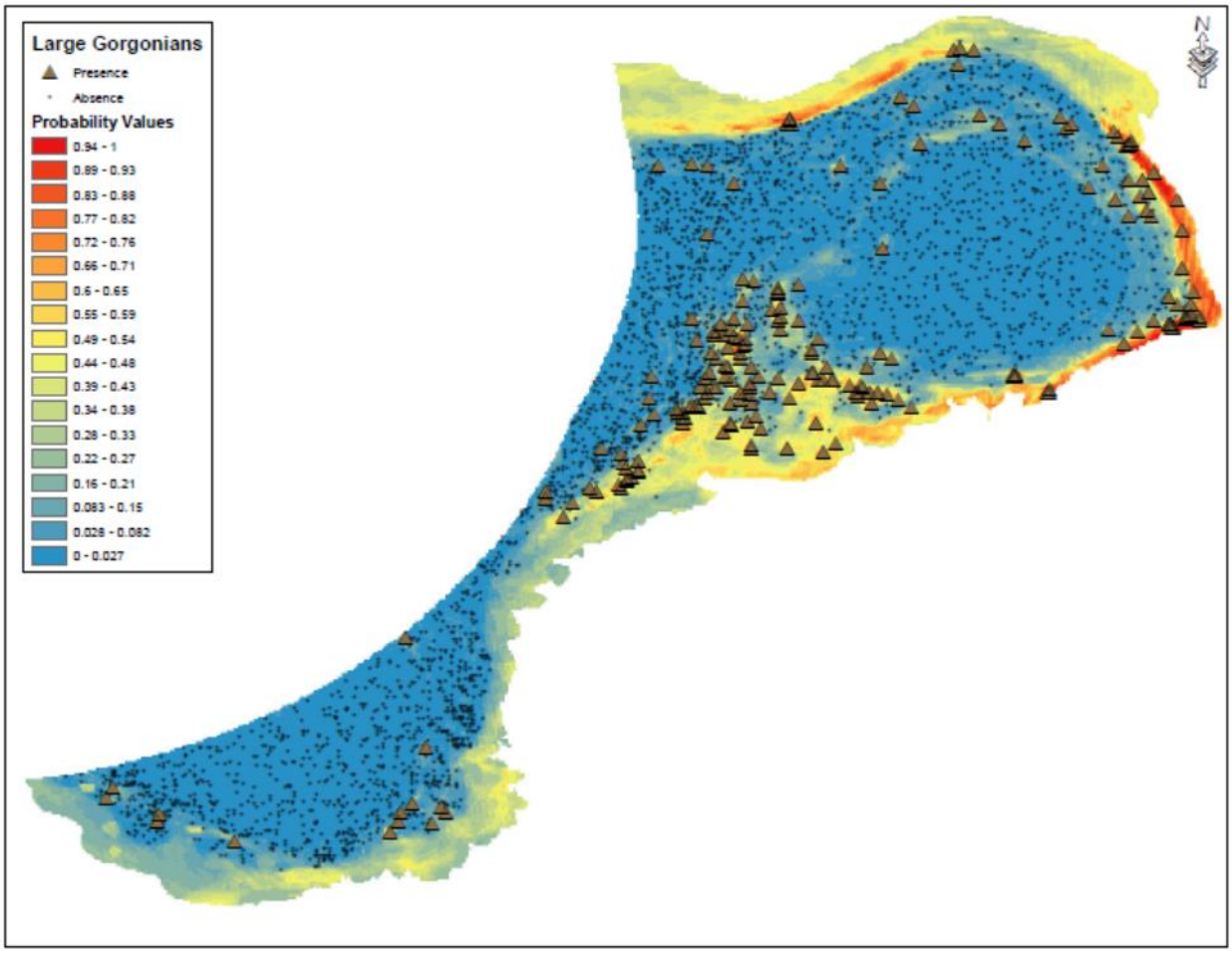

Figure 5. The location of known large gorgonian coral presence and absence data overlain on the prediction surface produced from the random forest species distribution model.

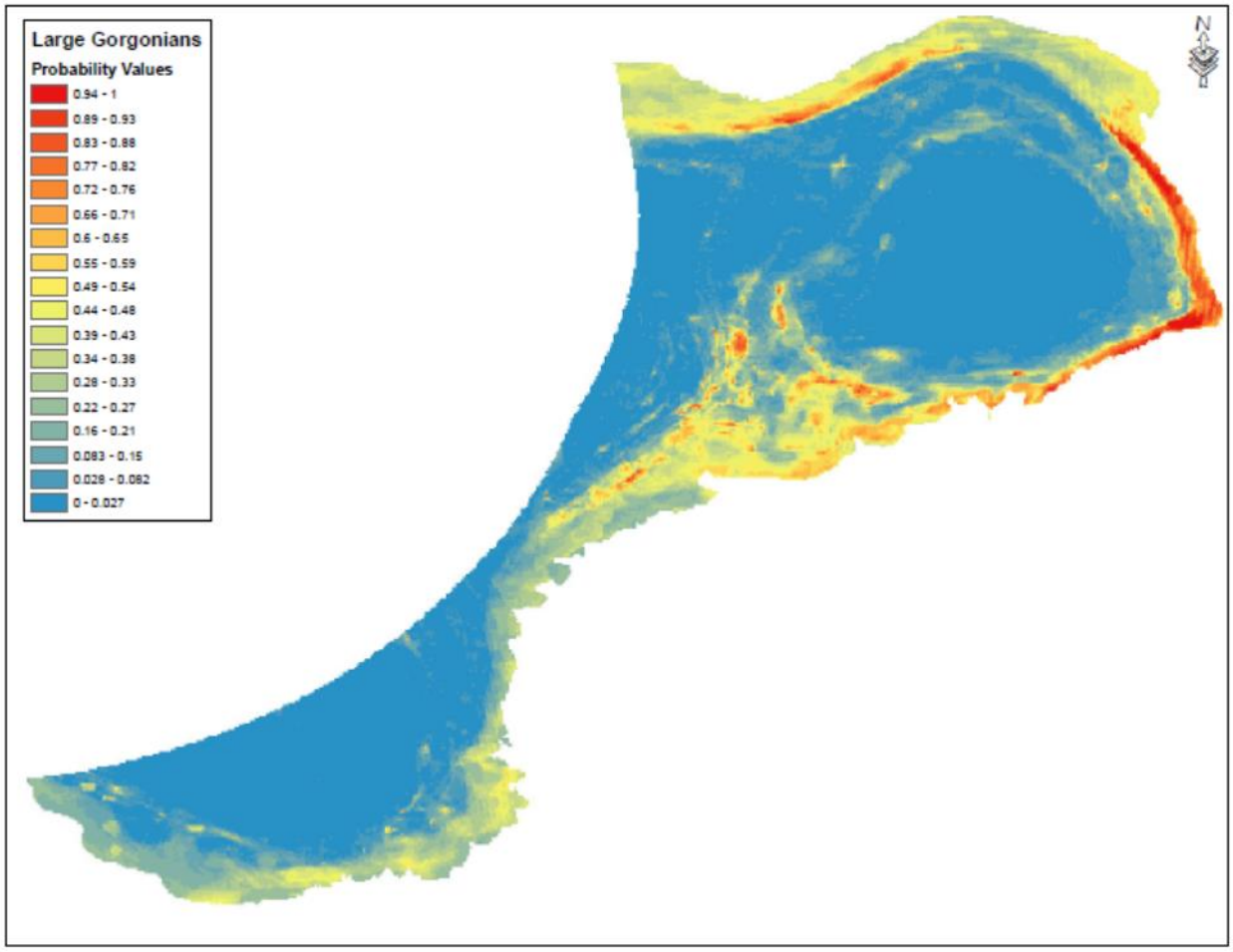

Figure 6. The probability of occurrence of large gorgonian corals in the NAFO regulatory area produced with a random forest species distribution model.

Sea Pens

The AUC model fit for the presence of sea pens was 0.888, which is considered very good. Eleven environmental variables were included in the model, with depth being the most important followed by maximum surface salinity and mean bottom temperature. Four of the variables showed negative importance (Table 5). Negative importance values can arise when there are paradoxes in the data, i.e., pairs of records with almost identical predictors have very different outcomes, when the model overfits to the training data. For predictors with negligible influence on the response, very small negative variable importance values like the ones seen in Table 5 can also be caused by the randomness inherent in the model. The partial dependence plots suggest that the sea pens prefer depths greater than 500 m and fully saline water (Figure 7).

The prediction surface shows a very good fit to the known presences and absences (Figure 8) and this surface was stable under different AUC cut off values and model runs. There is a high probability of occurrence on Flemish Cap and in the deeper waters on the tail of Grand Bank (Figure 9).

Table 5. Predictor variable combinations remaining after the AUC-based variable elimination, for the NRA area under a random forest species distribution model for sea pens (AUC = 0.888).

Response variable	Predictor Variables in Order of Importance	Variable Importance
Sea Pens	Depth	0.099
	Surface salinity, maximum	0.022
	Bottom temperate, minimum	0.020
	Summer Chl *a*, range	0.010
	Bottom salinity, maximum	0.003
	Slope	0.003
	Surface current, mean	0.002
	Annual Chl *a*, minimum	-0.001
	Surface current, minimum	-0.001
	Winter Chl *a*, range	-0.003
	Bottom shear, minimum	-0.004

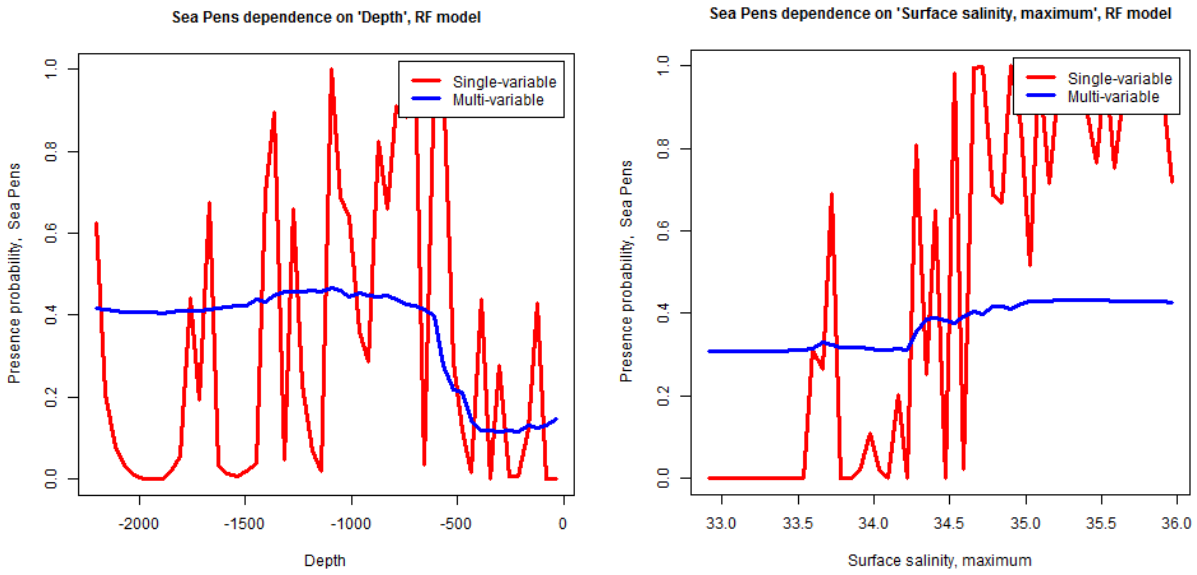

Figure 7. Partial dependence plots for the two predictor variables contributing most to the model fit (Table 5). Red lines are from a model trained on a single variable. Blue lines indicate performance when all other variables are held at their mean values (multi-variable).

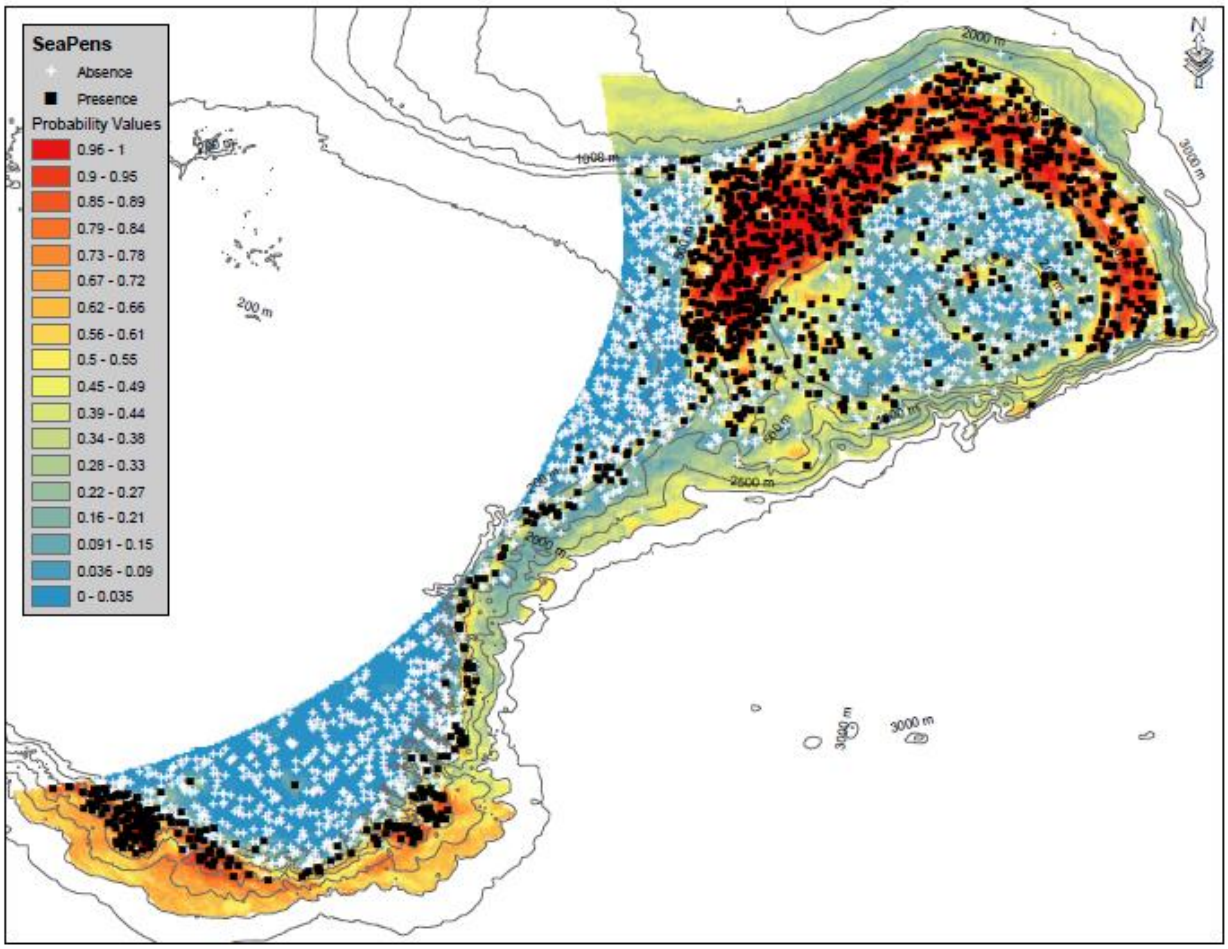

Figure 8. The location of known sea pen presence and absence data overlain on the prediction surface produced from the random forest species distribution model.

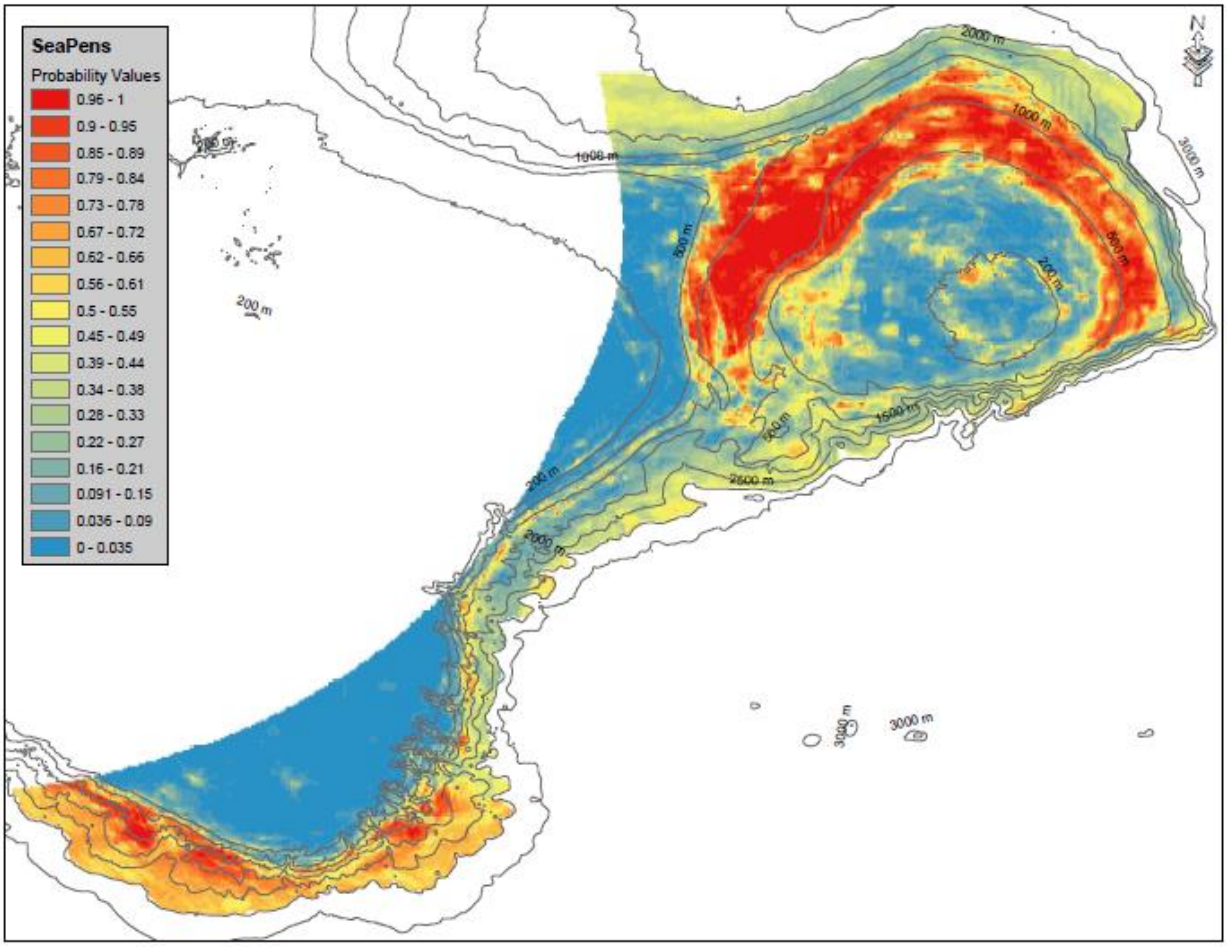

Figure 9. The probability of occurrence of sea pens in the NAFO regulatory area produced with a random forest species distribution model.

The AUC model fit for sea pen VME (catches greater than 1.4 kg per tow) was 0.942, which is considered excellent. Sixteen environmental variables were included in the model (Table 6) with maximum surface salinity being the most important followed by depth. As for the sea pen presence/absence model, six of the variables showed negative importance indicating mixed response signals for those variables (Table 6). Despite the excellent fit to the data, the model did not predict occurrences to exist outside of the known catch locations to any substantial extent (Figure 10). It is possible that the distribution of these significant concentrations within the broader distribution is not determined by the environment (to any great extent), at least not as described by the environmental variables we have access to, and at the spatial scale we have described them. They could also arise through stochastic biological processes related to recruitment. Another issue is that for this model we have a very big difference between the number of known presences vs. absences (see Table 2). Consequently the model is able to predict "absence" for the large majority of locations with a high degree of accuracy.

Table 6. Predictor variable combinations remaining after the AUC-based variable elimination, for the NRA area under a random forest species distribution model for sea pens (AUC = 0.942).

Response variable	**Predictor Variables in Order of Importance**	**Variable Importance**
Sea Pen > 1.4 kg/tow	Surface salinity, maximum	0.180
	Depth	0.118
	Bottom temperate, mean	0.052
	Bottom shear, minimum	0.020
	Surface current, range	0.019
	Bottom temperature, range	0.014
	Surface temperature, minimum	0.014
	Spring Chl *a*, mean	0.006
	Winter Chl *a*, range	0.005
	Bottom temperature, minimum	0.002
	Slope	-0.001
	Fall Chl *a*, range	-0.005
	Surface salinity, range	-0.008
	Winter Chl *a*, minimum	-0.009
	Surface current, minimum	-0.010
	Fall Chl *a*, minimum	-0.027

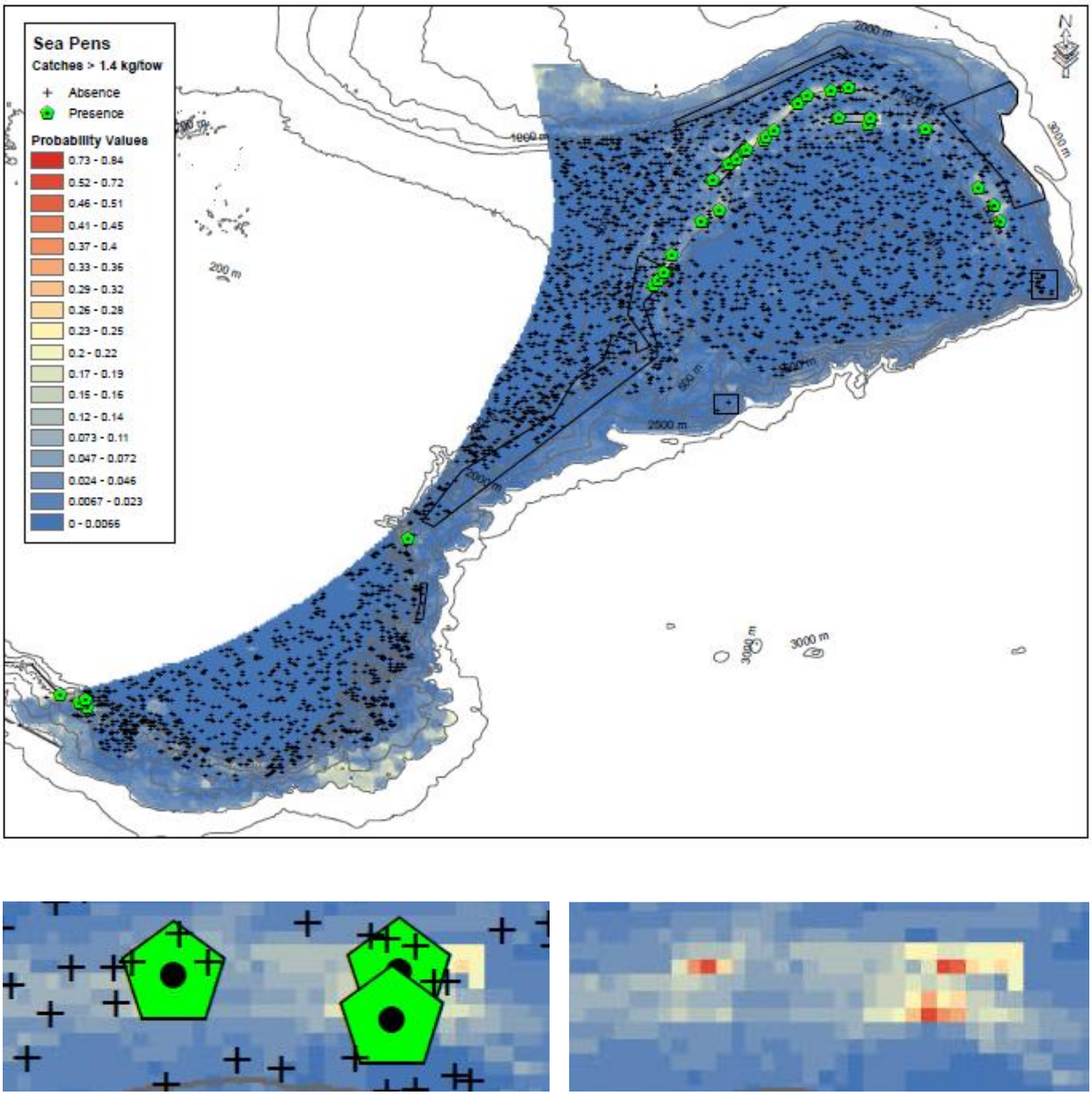

Figure 10. Upper. The probability of occurrence of sea pen fields in the NAFO regulatory area produced with a random forest species distribution model. **Lower Left.** Close up of the location of sea pen catches > 1.4 kg/tow and absences. **Lower Right.** Close up of SDM model probability of occurrence of catches > 1.4 kg/tow showing high probability in red centered on the single catch locations.

References

BODC. GEBCO Gridded Global Bathymetry Data. 2009. http://www.gebco.net/.

Breiman, L., J. Friedman, C.J. Stone and R. A. Olshen. 1984. Classification and Regression Trees. Chapman & Hall/CRC: Boca Raton, FL, USA. 358 p.

Fawcett, T. 2006. An Introduction to ROC Analysis. Pattern Recognition Letters 27: 861-874.

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

Graham, M. 2003. Confronting multicollinearity in ecological multiple regression. Ecology 84: 2809-2815.

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

Knudby, A., E. Kenchington, A.T. Cogswell, C.G. Lirette, F.J. Murillo. 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., E. Kenchington and F.J. Murillo. 2013b. Modeling the distribution of *Geodia* sponges and sponge grounds in the northwest Atlantic Ocean. PLoS ONE (accepted).

Liaw, A., and M. Wiener. 2002. Classification and Regression by Random Forest. R News 2: 18-22.

Murillo, F.J., E. Kenchington, C. Gonzalez, M. Sacau. 2010. The use of density analyses to delineate significant concentrations of Pennatulaceans from trawl survey data. NAFO Scientific Council Research Document 10/07, Serial No. N5753, 7 pp.

R Core Development Team. 2012. R: A Language and Environment for Statistical Computing. Version 2.15.

Shih, S. 2011. Random Forests for Classification Trees and Categorical Dependent Variables: an informal Quick Start R Guide. http://www.stanford.edu/~stephsus/R-randomforest-guide.pdf Accessed 13/12/2013.