A COMPARISON OF SPATIO-TEMPORAL SPECIES DISTRIBUTION MODELLING FRAMEWORKS USING *CUCUMARIA FRONDOSA* AS A CASE STUDY

by

Nathan Hebert

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
March 2023

To all my friends and family; they so graciously supported me throughout this journey.

Table of Contents

List of Tables

List of Figures

Abstract

Cucumaria frondosa—the most abundant sea cucumber species in the Maritimes—is the target of a relatively new offshore fishery on the Scotian Shelf. In an effort to ensure the long-term sustainability of the stock, spatial reserves for *C. frondosa* were set aside for the 2018/2019 fishing season. However, the expected habitat map used to design the reserve boundaries was acknowledged to be somewhat coarse at the time. Furthermore, high-resolution environmental data layers (including several from a digital elevation model) have recently become available for the area. We therefore incorporated these new data layers into three spatio-temporal species distribution models, with the aim of re-examining the distribution of *C. frondosa* on the Scotian Shelf. Each model accounted for spatio-temporal autocorrelation differently, allowing us to compare their underlying methodologies.

Specifically, we fitted a generalized additive model with spatio-temporal smooths and two generalized linear mixed models with spatio-temporal random effects, one of which implemented the random effects using Gaussian Markov random fields, and the other with nearest-neighbour Gaussian processes. All three were fitted to catch data (recorded 2000-2019) from Fisheries and Oceans Canada's annual Research Vessel and Snow Crab surveys. To accommodate the large number of zero catch values, each model consisted of two parts; the first part modelled the probability of sea cucumber presence using all tows, whereas the second used the Gaussian distribution to model positive catch for presence tows. Together, the three models suggested that several environmental covariates were possibly predictive of sea cucumber habitat and abundance in the region, including log-transformed depth and bottom stress. Furthermore, they indicated that considerable *C. frondosa* aggregations existed on Middle Bank, Banquereau Bank, and Sable Bank as of 2019. The models' predictions were generally very similar, although they provided somewhat different inferences regarding covariate effects; we discuss the implications of this result for the practitioner.

Acknowledgements

I would like to start by thanking Joanna Mills Flemming, my supervisor. She introduced me to the exciting world of spatio-temporal statistics, and graciously guided me throughout the duration of this project.

Next, I want to thank Jessica Sameoto and David Keith of Fisheries and Oceans Canada. I would not have made it far without them. Each provided a second set of (highly knowledgeable) eyes on my work, and patiently answered my many questions. Their colleague, Melanie Barrett, also lent this project her invaluable support and sea cucumber fishery expertise; for that, I am grateful.

I would like to acknowledge the efforts of Mike Dowd, who read this document and provided helpful feedback. Three others directly aided me in my research. Many thanks go out to Ethan Lawler and Sean Anderson for their diligence and enthusiasm in answering any queries related to *starve* and *sdmTMB* respectively. I also much appreciate the support Raphaël McDonald and Ethan Lawler provided leading up to my thesis defense. Ethan, Sean, and Raphaël are exceedingly proficient in spatio-temporal statistics and greatly enjoyed discussing their work.

I want to thank Bailey Drew for making me a radio star. Finally, I want to acknowledge my entire extended family. Not only were my parents, grandparents, and other relatives highly supportive, but they also put up with my sea cucumber talk for far too long. Sorry about that everyone!

Chapter 1

Introduction

As a result of burgeoning market demand and declining finfish stocks, the exploitation of benthic invertebrate species has exploded since the 1950s (Shackell et al., 2013). In addition, regulation and the scientific knowledge underlying it have been unable to keep up with the increase in their exploitation (Anderson et al., 2008; Anderson et al., 2011). This potent combination of fisheries growth and slow management response has alarming implications not only for benthic invertebrates, but also for the ecosystems in which they reside (Shackell et al., 2013). Of particular concern here are sea cucumber species, which have been over-exploited in many parts of the world (Purcell et al., 2010; Anderson et al., 2011).

Permanent or rotational fishery reserves are now considered integral to the protection of sea cucumber populations (Humble et al., 2007). Of course, the creation and maintenance of effective reserves requires extensive knowledge of how the species of interest is distributed. This thesis focuses on *Cucumaria frondosa*, the most abundant sea cucumber species on the Scotian Shelf (Hamel and Mercier, 2008). The analyses described herein were undertaken with two motivations. Firstly, we aimed to develop a species distribution model for *C. frondosa* on the Scotian Shelf using a framework that accounts for spatio-temporal autocorrelation.

This effort necessitated the selection of an approach, and the myriad of ways to account for spatio-temporal dependence in ecological data can be daunting. Consequently, ecologists well-versed in several such frameworks have to either select one, or move forward with an ensemble approach based on a few. Meanwhile, those familiar with only one framework have to decide whether to work with the method to which they are accustomed, or to learn how to apply others. In all cases, a non-trivial choice must be made and justified. Thus, our second motivation was to compare a selection of spatio-temporal species distribution modelling (SDM) frameworks using this investigation as a case study. Employing multiple frameworks would allow us to

examine the impact of choosing a particular one in this setting. Of course, if their predictions all concurred, we would have greater confidence in our overall inference.

This chapter will discuss spatio-temporal SDM, the three particular modelling frameworks we applied, and the motivating biological population (i.e., *C. frondosa* on the Scotian Shelf). The following chapter will describe the available data and the overarching modelling approach used. It will also provide specifics on how each framework was employed to address our particular problem. Chapter 3 will compare the resulting models in terms of their predictions and interpretations. Finally, Chapter 4 will discuss the results and what they mean for sea cucumbers on the Scotian Shelf, and spatio-temporal ecologists at large.

1.1 Species Distribution Modelling

In fields concerned with natural resource management, such as fisheries science, it is often necessary to develop a map of a species' actual or potential geographic distribution (Franklin and Miller, 2010). Reserve design is only one possible application of such maps; once obtained, they can also be used to support biodiversity assessment and ecological monitoring, among other applications (Zhang et al., 2018). Various techniques are employed to generate them. On occasion, the use of kernel density estimation is deemed sufficient; the technique enables the non-parametric estimation of a continuous probability density function, and has been extended to estimate the spatial distribution of species (Kenchington et al., 2009; Kenchington et al., 2014).

Kernel density estimation is notably limited in that it does not directly incorporate or require environmental variables, whereas SDM does (Franklin and Miller, 2010). Also known as habitat suitability modelling or ecological niche modelling, SDM as a methodology involves the use of statistical models or machine learning algorithms to predict a species' spatial or spatio-temporal distribution (Franklin and Miller, 2010). Generally, environmental variables believed to correlate with habitat suitability are chosen as predictors or used by the selected algorithm (Franklin and Miller, 2010). In the case of a benthic marine species for example, the bottom temperature or bathymetry (depth) at a given location may be incorporated.

A number of algorithms and modelling approaches are now used for SDM; in practice, the method a researcher selects generally depends on the type of data available

to them (Boitani and Powell, 2014). Information on where the species of interest has been present (i.e., presence data) is required by all SDM approaches (Boitani and Powell, 2014). However, presence-absence techniques necessitate absence data as well—that is, information on locations where the species is known to not have been present (Boitani and Powell, 2014). Presence-only data is highly prevalent, often emerging from patrol records and historical collections (Zhang et al., 2018). Absence data is more difficult to obtain, usually only arising via carefully designed surveys (Zhang et al., 2018). Absence data can also be somewhat unreliable; one can readily imagine situations where a species was present but undetected (Boitani and Powell, 2014).

To accommodate the ubiquity of presence-only data, methods such as ecological niche factor analysis (Hirzel et al., 2002), the MaxEnt algorithm (Phillips et al., 2006), and Mahalanobis distance modelling (Farber and Kadmon, 2003) have been developed. They all assume presence locations represent ideal habitat conditions, and then determine the suitability of other locations by examining how similar they are (Boitani and Powell, 2014). Meanwhile, when presence-absence data is available, SDM is frequently conducted via traditional statistical learning approaches such as generalized linear models (GLMs), generalized additive models (GAMs), and tree-based models (Zhang et al., 2018). These methods are often adequate to model the complex predictor-response relationships that can arise in SDM (Franklin and Miller, 2010). They can also move beyond the presence-absence paradigm and model species abundance if that information is available.

As a complication, the considered environmental covariates may not sufficiently describe the species' distribution; important environmental variables may be unobserved or neglected, and ecological processes such as dispersal, predation, and competition can have some influence (Bakka et al., 2016). Moreover, the measurement process used to study the species may exhibit a spatio-temporal dependence structure, especially when environmental information is retrieved via raster maps that were created independently of the biological survey (Gelfand et al., 2006). Somewhat surprisingly then, a brief perusal of the literature suggests that spatio-temporal autocorrelation is often neglected in SDM. A failure to account for correlated model residuals can lead to a multitude of issues, including biased parameter estimates,

overly-optimistic estimates of uncertainty, and even misidentification of important environmental variables (Record et al., 2013; Bakka et al., 2016). Alternatively, disregarding the spatio-temporal relationships between observations can simply be seen as casting aside potentially crucial information.

1.2 Accounting for Spatio-temporal Autocorrelation

In a regression setting, researchers often attempt to account for spatial autocorrelation through the inclusion of a trend surface involving geographic coordinates (Gelfand et al., 2006). This approach is perhaps most natural when the trend surface is incorporated into a GAM using splines. Introduced by Hastie and Tibshirani (1986) and popularized by implementations such as the *mgcv* R package (Wood, 2011), GAMs can be viewed as a generalization of GLMs where the link function is modeled as the sum of smooth non-parametric functions of the covariates. One can easily incorporate a function of geographic position (i.e., $\mathbf{s}$) into their GAM (with response Y and link function $g(\cdot)$) such that it can be specified as

$$g(E(Y)) = f(\mathbf{s}) + ... \tag{1.1}$$

The function of space can then be expanded such that it also involves time, resulting in a spatio-temporal model.

This method incorporates information on the spatial or spatio-temporal relationships between observations through the model's systematic component (Stock et al., 2020). In contrast, the inclusion of random effects through Gaussian random fields (GRFs) is another commonly-used approach which focuses on the random component (Stock et al., 2020). GRFs have been noted to improve prediction in fishery science settings (Kallasvuo et al., 2017), and they are now commonly used to model species distributions (Commander et al., 2022).

In classical geostatistics, a random field $Y(\cdot)$ is a continuous stochastic process defined by $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$, where $\mathbf{s}$ is a point within the region of interest D (Zimmerman and Stein, 2010). However, sampled observations constitute only a partial realization of this process; as such, it is considered to have a sample size of one (Moraga, 2020). To accommodate this, geostatistical models must be carefully structured

(Moraga, 2020). Traditionally, they are given the form

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + e(\mathbf{s}), \tag{1.2}$$

where $\mu(\cdot)$ is a deterministic mean function and $e(\cdot)$ is a stationary stochastic error process with a mean of zero (Zimmerman and Stein, 2010).

It is prudent to briefly define stationarity in a spatial context. Let $C(\cdot)$ be the field's covariance function, which describes the degree of spatial autocorrelation between any two locations in D (Bakka et al., 2016). Then second-order stationarity posits that

$$C(e(\mathbf{s_i}), e(\mathbf{s_j})) = C(\mathbf{s_i} - \mathbf{s_j}) \tag{1.3}$$

for any two points $\mathbf{s}_i$ and $\mathbf{s}_j$ (Zimmerman and Stein, 2010). In essence, second-order stationarity implies that the covariance function is translation invariant (Moraga, 2020). Alternatively, intrinsic stationarity can be assumed (Zimmerman and Stein, 2010). Less strict than second-order stationarity, it says that for a semivariogram $\gamma(\cdot)$ and any two points $\mathbf{s}_i$ and $\mathbf{s}_j$,

$$Var(e(\mathbf{s_i}) - e(\mathbf{s_j})) = 2\gamma(\mathbf{s_i} - \mathbf{s_j}) \tag{1.4}$$

(Zimmerman and Stein, 2010). One may in addition suppose second-order or intrinsic isotropy, where the covariance function or semivariogram depends only on $||\mathbf{s_i} - \mathbf{s_j}||$ (Moraga, 2020). However, the assumptions of stationarity and isotropy are easily violated in marine environments (especially near land masses); thus, they are sometimes relaxed through various means (Thorson et al., 2015; Bakka et al., 2016).

GRFs are continuously indexed collections of random variables where every finite collection is multivariate normal (Moraga, 2020). Because a GRF can be completely expressed by its mean and covariance function, they are frequently used in practice (Van Lieshout, 2019). However, the covariance function must be specified; Matérn functions are common selections (Moraga, 2020). For a random field $Y(\cdot)$, the Matérn covariance function takes the form

$$C(\mathbf{s_i}, \mathbf{s_j}) = Cov(Y(\mathbf{s_i}), Y(\mathbf{s_j})) = \frac{\sigma^2_{MV}}{2^{\nu-1}\Gamma(\nu)}(\kappa||\mathbf{s_i} - \mathbf{s_j}||)^{\nu} K_{\nu}(\kappa||\mathbf{s_i} - \mathbf{s_j}||), \tag{1.5}$$

where $\mathbf{s}_i$ and $\mathbf{s}_j$ are the two locations of interest, σ^2_{MV} is the spatial field's marginal variance, κ controls the spatial scale, and $K_\nu(\cdot)$ is the modified Bessel function of second kind and order $\nu > 0$ (Moraga, 2020). Generally, ν is held fixed (at, say, $\nu = 1$) and $\rho = \sqrt{8\nu}/\kappa$ is reported as the distance where the spatial autocorrelation diminishes to 0.1 (Carson et al., 2017). We further note that named special cases of the Matérn exist, such as the exponential covariance function (where $\nu = 0.5$) (Moraga, 2020).

1.3 Gaussian Markov Random Fields and Nearest-Neighbour Gaussian Processes

GRFs are relatively convenient for spatial and spatio-temporal modelling, but they can be computationally expensive (Lindgren et al., 2011; Carson et al., 2017). The difficulty arises because they require dense $n \times n$ covariance matrix factorization; this issue has been dubbed "the big n problem" (Lindgren et al., 2011). The Stochastic Partial Differential Equations (SPDE) approach rectifies this by representing a continuously indexed GRF with a discretely indexed Gaussian Markov random field (GMRF) (Lindgren et al., 2011). GMRFs are often used to model dependence within a grid or lattice structure, and have a neighbourhood-based construction where non-neighbouring components are conditionally independent (Lindgren et al., 2011; Carson et al., 2017). Their use provides a computational gain because only factorization of the sparse precision matrix $Q = \Sigma^{-1}$ is required (where $\Sigma = C(\mathbf{s_1}, \mathbf{s_2})$) (Lindgren et al., 2011).

In practice, the SPDE procedure involves assuming an underlying GRF with covariance matrix Σ (which originated from a Matérn covariance function), finding a GMRF such that Q is close to Σ^{-1}, and then conducting the computations with the GMRF (Lindgren et al., 2011). However, the substitution of a GMRF in place of a GRF requires a discretely indexed representation of a continuously indexed process to be found (Lindgren et al., 2011). The SPDE approach uses triangulation to do this, rather than a simple grid—this allows the cells to vary in size, and as a direct result, also enables them to appropriately portray the resolution provided by the data in each part of the spatial domain (Lindgren et al., 2011; Carson et al., 2017). The resulting triangulation is often called a mesh (Lindgren et al., 2011).

Various software implementations of the SPDE approach exist, including Rue et al.'s *R-INLA* package (2009). For our purposes, the R package *sdmTMB* (Anderson et al., 2022) is especially notable. Designed with ecologists and SDM in mind, *sdmTMB* can fit generalized linear mixed models (GLMMs) with spatial or spatio-temporal random effects. All of this is handled in a highly efficient manner, because of the package's reliance on *Template Model Builder* (*TMB*)(Kristensen et al., 2016). *TMB* enables the fast optimization of user-coded likelihood functions using the Laplace approximation and automatic differentiation. When applied to fit complex spatio-temporal models for fishery-based applications, including those involving GMRFs, *TMB* has seen great success (Thorson et al., 2015; McDonald et al., 2022).

The SPDE approach is not the only method that has been proposed to approximate a Gaussian process. Of particular interest here is the nearest-neighbour Gaussian process (NNGP) proposed by Datta et al. (2016), which is based on the Vecchia approximation (Vecchia, 1988). Briefly, they begin its development by assuming that $\mathbf{w}(\mathbf{s})$ follows a Gaussian process with a mean of zero and covariance function $C(\mathbf{s_i}, \mathbf{s_j})$. Next, they let S = $\{\mathbf{s_1}, \mathbf{s_2}, ...\mathbf{s_k}\}$ be a reference set of points located in the spatial domain of interest D such as—but not necessarily—the observed locations. Subsequently, the joint distribution of $\mathbf{w}_S$ can be expressed as a product of conditional densities, giving

$$p(\mathbf{w}_S)) = p(\mathbf{w}(\mathbf{s_1}))p(\mathbf{w}(\mathbf{s_2})|\mathbf{w}(\mathbf{s_1}))...p(\mathbf{w}(\mathbf{s_k})|\mathbf{w}(\mathbf{s_{k-1}}), \mathbf{w}(\mathbf{s_{k-2}}), ...\mathbf{w}(\mathbf{s_1})). \quad (1.6)$$

They note that the computational burden is eased when the conditioning sets on the right hand side are replaced with sets of size m such that m << k; in a NNGP these sets are chosen to be each $\mathbf{s_i}$'s m-nearest neighbours (in terms of Euclidean distance). Generally, modelling frameworks that implement NNGP-based random effects specify one or more underlying directed acyclic graphs (DAGs) on a set of random effects nodes; the graphs fill an analogous role to the basis functions used by spline-based models or the triangulation used in the SPDE approach (Lawler et al., 2023).

This is all exemplified by the *starve* R package (Lawler et al., 2023), which utilizes NNGPs to allow for efficient likelihood-based fitting of hierarchical GLMMs to spatio-temporal point-referenced data. *starve* is designed to be accessible to non-statisticians, and relies on the efficient and likelihood-based *TMB*. Thus, it is akin to

sdmTMB as a modelling framework.

1.4 Case Study: *Cucumaria frondosa* on the Scotian Shelf

Despite being the most abundant sea cucumber species on the Scotian Shelf, little is known about *C. frondosa*'s life-history characteristics in the region (Pantin et al., 2018). Some understanding has been developed by extrapolating what is known of *C. frondosa* living in the Gulf of St. Lawrence area; however, regional differences are expected (Pantin et al., 2018). Furthermore, the offshore fishery for *C. frondosa* on the Scotian Shelf has only existed since 2004, with fishing restricted to designated areas in Northwest Atlantic Fisheries Organization Divisions 4W, 4Vs, and 4X (Figure 1.1) (DFO, 2021a).

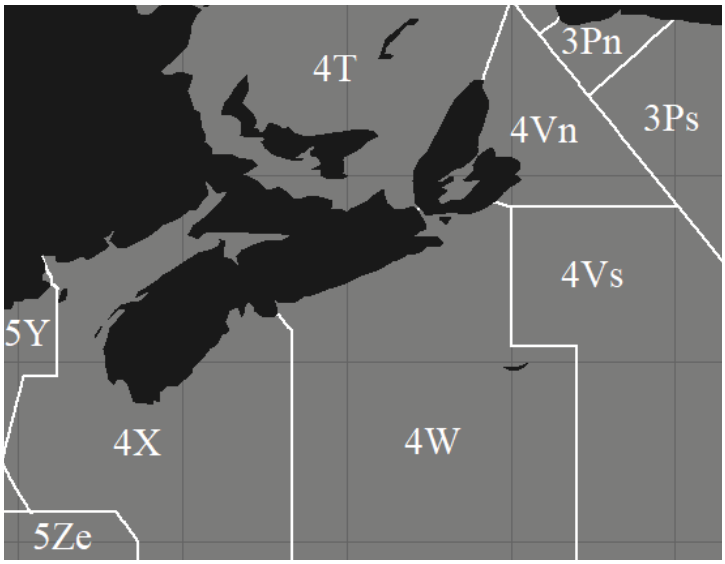


Figure 1.1: An illustration of relevant Northwest Atlantic Fisheries Organization Divisions (as well as others in proximity).

As mentioned previously, reserves are now believed to be critical to the protection of sea cucumber populations (Humble et al., 2007). *C. frondosa* tends to aggregate in high densities and is considered a low-mobility species; such traits make their populations potentially vulnerable to hyperstability—i.e., population decline masked by stable catch-based indicators (Orensanz et al., 1998). As a result, conventional management approaches (such as the setting of a Maximum Sustainable Yield) have been shown to be problematic, and the use of reserves is recommended instead (Purcell

et al., 2010; DFO, 2021a).

For the 2018/2019 fishing season, Fisheries and Oceans Canada (DFO) set aside reserves in the 4Vs Offshore fishing area (DFO, 2021a). Commercial data was limited for the region, and challenges arose in identifying predictive environmental variables. Thus, DFO's annual Research Vessel (RV) and Snow Crab surveys were used to construct an expected habitat map via a kernel density approach (DFO, 2021a). Each habitat was classified as being either primary, secondary, or tertiary, and then 30% of their combined area was set aside (Figure 1.2) (Purcell et al., 2010; DFO, 2021a). This endeavour has been accompanied by more targeted investigations into *C. frondosa*'s distribution, such as drop-camera surveys of 4W Offshore and 4W Midshore (Harper, 2020).

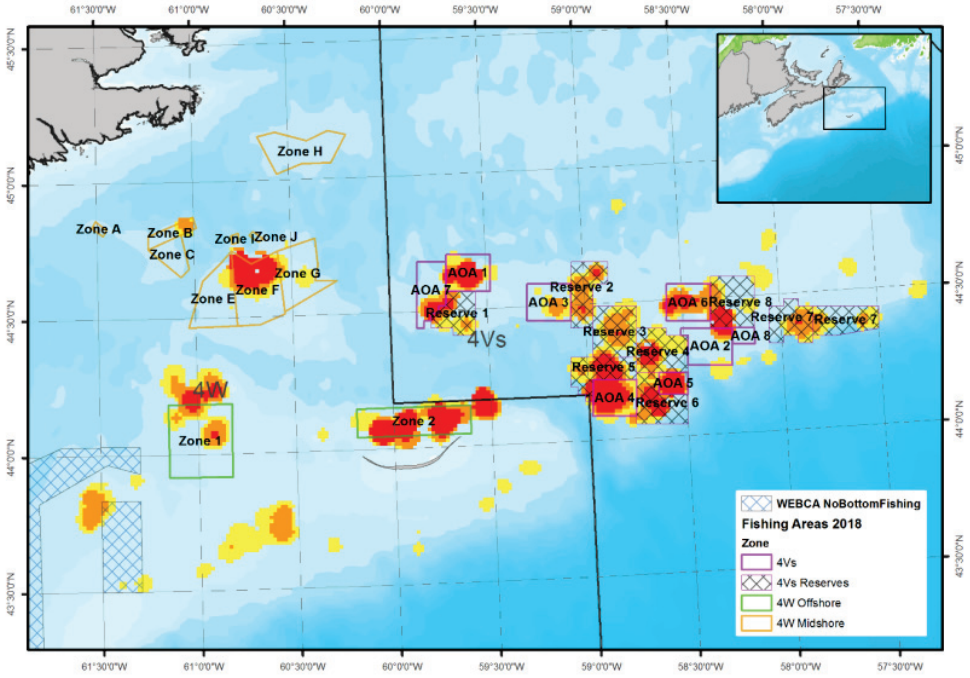


Figure 1.2: The map of expected *C. frondosa* habitat created by DFO, overlaid with where *C. frondosa* reserves were set aside for the 2018/2019 fishing season. Nearby fishing zones and Areas of Access are also shown. This figure was taken from DFO, 2021a.

It was recommended that the expected habitat map be updated as new information emerges (DFO, 2021a). Multibeam echosounders allow for detailed bathymetric measurements to be taken and used to create digital elevation models (or DEMs)

(Lecours et al., 2016). Terrain attributes, such as slope and slope aspect, are commonly derived from such models and exploited to produce benthic habitat maps (Lecours et al., 2016). Environmental data layers sourced from a novel DEM and the Bedford Institute of Oceanography North Atlantic Model (BNAM) (Brickman et al., 2016; Wang et al., 2018) have recently been made available. Thus, we aimed to broadly reassess the spatial distribution of *C. frondosa* using these new layers and a SDM framework that adequately accounts for spatio-temporal dependencies.

Using *C. frondosa* on the Scotian Shelf as a case study, this thesis will consider and informally compare the following three spatio-temporal SDM frameworks:

- GAMs with spatio-temporal smooth functions (as implemented by *mgcv* (Wood, 2011)),
- spatio-temporal GLMMs based on NNGPs (as implemented by *starve* (Lawler et al., 2023)),
- and spatio-temporal GLMMs based on the SPDE approach (as implemented by *sdmTMB* (Anderson et al., 2022)).

These three were chosen because they are relevant and accessible to ecologists. We initially implemented our GMRF-based model with *TMB* (Kristensen et al., 2016) directly. However, using a widely available package specifically tailored for such models (i.e., *sdmTMB*) seemed more prudent for the purposes of reproducibility. Using *TMB* directly gave similar results.

Chapter 2

Methods

We will now develop our spatio-temporal models for *C. frondosa* on the Scotian Shelf. As is usually the case with SDM, this task required that several data sources of various types (i.e., survey data and environmental rasters) be integrated in a cohesive and ecologically sensible way. Necessarily then, a considerable proportion of this chapter is devoted to describing the data and various concerns surrounding them. The rest of the chapter will specify how the three modelling frameworks were used to produce a trio of comparable species distribution models. All references to log transformations refer to uses of the natural log. Analyses were conducted in the R programming language (R Core Team, 2021), as was any data pre-processing.

2.1 Survey Data

DFO conducts an annual RV survey of the Scotian Shelf and Bay of Fundy to monitor the state of the ecosystem (DFO, 2021b). Stratified random sampling based on depth is used to select the fishing stations, and then standardized 30-minute tows are conducted at 3.5 knots using Western II-A bottom trawl gear (Tremblay et al., 2007). Generally, the *CCGS Alfred Needler* has been the vessel used, although others—most commonly the *CCGS Teleost*—have conducted the survey on occasion (Tremblay et al., 2007). Both the *Needler* and the *Teleost* were employed in 2005, and their catch rates were not found to be consistently different (Fowler and Showell, 2009). DFO supplied data on RV Survey tows conducted during the years 2000-2019; most occurred midsummer, while the rest generally took place in the winter. Only 4X saw survey coverage in 2018 because of mechanical issues with the survey vessel (DFO, 2021b). Pre-2000 tows were not provided, as invertebrates were only reliably identified to the species level starting in 2000.

DFO also supplied data on tows from the 2002-2019 Snow Crab Surveys. The annual Snow Crab Survey occurs in the late summer and early fall, and follows a fixed

station design where five-minute tows are conducted at a vessel speed of approximately two knots (Zisserson, 2015). While designed to aid stock assessment for snow crabs on the Scotian Shelf, it also catalogues information on other species encountered through bycatch (Zisserson, 2015). The Snow Crab Survey is more limited than the RV Survey in its spatial extent, however. It focused only on a small portion of 4Vs upon its inception in 1996, although it has gradually expanded since to include parts of 4Vn, 4W, and 4X (Zisserson, 2015). Throughout this period of early expansion (1996-2003), five vessels were used to conduct the survey (Zisserson, 2015). Their successor, the *F/V Gentle Lady*, served as the survey vessel until 2013, when it sank and was replaced by the *F/V Ms. Jessie* (Zisserson et al., 2018). A modified version of the Bigouden Nephrops trawl net has been employed across all years for consistency (Zisserson, 2015).

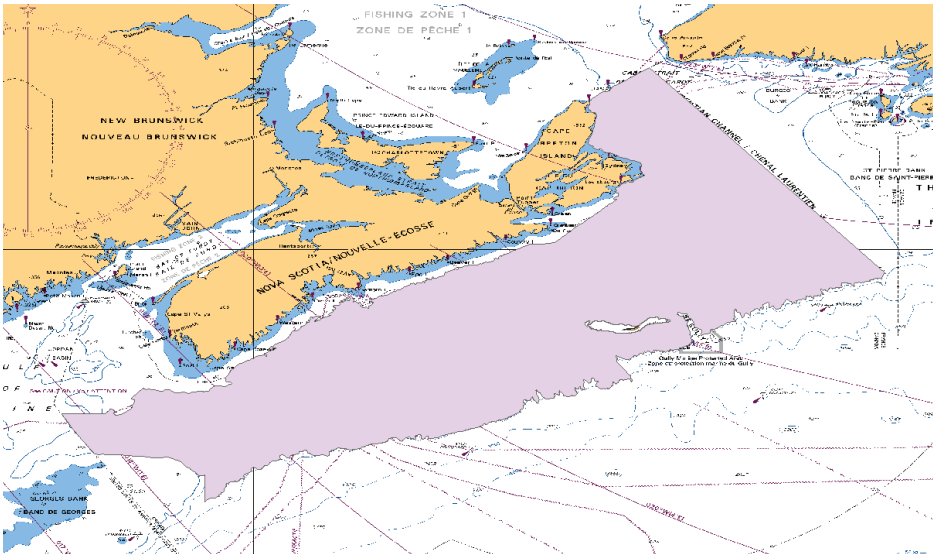


Figure 2.1: An illustration of the spatial domain considered for modelling and prediction.

We combined the available data from the two surveys and then subsetted the result to reflect only our spatial domain of interest. While most of the offshore sea cucumber fishery occurs in 4Vs and 4W (DFO, 2021a), Shackell et al. (2013) showed that *C. frondosa* aggregations exist as far south as Browns Bank. Thus, modelling was restricted to the Scotian Shelf, bounded by the Laurentian Channel in the northeast and the Northeast Channel in the southwest. It was further decided that predictions would not be made in areas with depths shallower than 23 metres (i.e., the minimum

depth value associated with our survey tows). This avoided potentially untenable extrapolation in the area immediately surrounding Sable Island, which was also not of interest for prediction (Figure 2.1).

There were a total of 10327 tows used in modelling. The record associated with each tow included the date, latitude, and longitude at which the observation occurred, as well as a standardized total weight of *C. frondosa* caught. Therefore, we had not only presence-absence data, but catch per unit effort (CPUE) quantities as well. We note that the Snow Crab Survey weights from before 2005 were considered unreliable due to the absence of a scale on the survey vessel, and thus they were not used in the analyses. 3168 (30.68%) of the available tows encountered *C. frondosa*, although only 2806 (27.17%) recorded a reliable non-zero weight.

Each survey dataset used different units for total weight. The RV Survey weights were provided in kilograms standardized to a 1.75 nautical mile tow length, whereas the Snow Crab Survey weights were given in kilograms per square kilometre. Using an assumed standard width of 41 feet for the RV Survey trawl (Shackell et al., 2013), the RV Survey weights were converted to kilograms per square kilometre. Further standardization was then required, because differences in gear and gear efficiency resulted in the Snow Crab Survey tows having higher CPUE. To control for this disparity, an indicator variable was included in each model; it was set to one for observations that occurred as part of the Snow Crab Survey (and zero otherwise). The indicator also implicitly accounted for the dissimilarity in the timing of the Snow Crab and RV surveys.

This temporal variability in tow occurrence had to be carefully considered because size-dependent seasonal migration has been identified in *C. frondosa* populations in other regions. In the St. Lawrence estuary for example, sexually mature sea cucumbers have been noted to move to deeper water in the fall (Hamel and Mercier, 1996). Preliminary analyses suggested that any confounding seasonality effects were likely negligible, however. Specifically, the available environmental data layers indicated that bottom temperature, stress, and salinity measurements were relatively stable in hypothesized *C. frondosa* habitat across any given year; this indicated that an extensive migration might not be of concern in such areas. Furthermore, predictions from our models were relatively insensitive to whether winter RV Survey tows were used

during model fitting. Seasonality differences were thus not explicitly accounted for during modelling, and no observations were dropped based on when they occurred.

The tows were projected onto the Universal Transverse Mercator (UTM) system, and then the resulting coordinates were converted from metres to kilometres. This transformation ensured that distance remained constant throughout the study region, and simplified our eventual interpretation of the spatial parameters (McDonald et al., 2021).

2.2 Environmental Covariates

BNAM is an eddy-resolving model of the North Atlantic Ocean (Brickman et al., 2016; Wang et al., 2018). Based on the NEMO 2.3 (Nucleus for European Modelling of the Ocean) framework, it is designed to support various DFO monitoring programs by providing both hindcast simulations and future climate projections. Monthly-averaged BNAM hindcast output was available at a resolution of 0.10° for 2000-2019. To incorporate relevant BNAM output into our modelling, we joined each tow with the bottom temperature, stress, and salinity values which corresponded spatio-temporally (i.e., to the month and year the tow was conducted, as well as its geographic middle point). Bottom stress was of particular interest here because it is believed to exert a mechanistic influence on benthic communities and has shown utility in mapping some benthic habitats (Jackson-Bué et al., 2022). Further, it can serve as a proxy for sediment grain size (Ward et al., 2015); this is a useful trait given that *C. frondosa* seems to prefer harder substrate types on the Scotian Shelf (Harper, 2020).

Preliminary analyses suggested that the BNAM values were generally close to in situ measurements, and thus validated our subsequent use of the BNAM data layers in modelling. Additional layers were also derived from BNAM's bottom temperature, stress, and salinity values such as the average and absolute minimum, maximum, and range. The "average" layers averaged across years, whereas the "absolute" layers took the absolute minimum, maximum, and range values as in Beazley et al. (2018). Of these, only (absolute) stress and temperature range were utilized. Once spatially intersected with the tows, additional layers deemed potentially relevant were correlated with others of interest. To avoid multicollinearity issues, we ensured that all pairwise $|r| < 0.70$ and all VIFs < 3 for the final set used in modelling (Zuur et al., 2010).

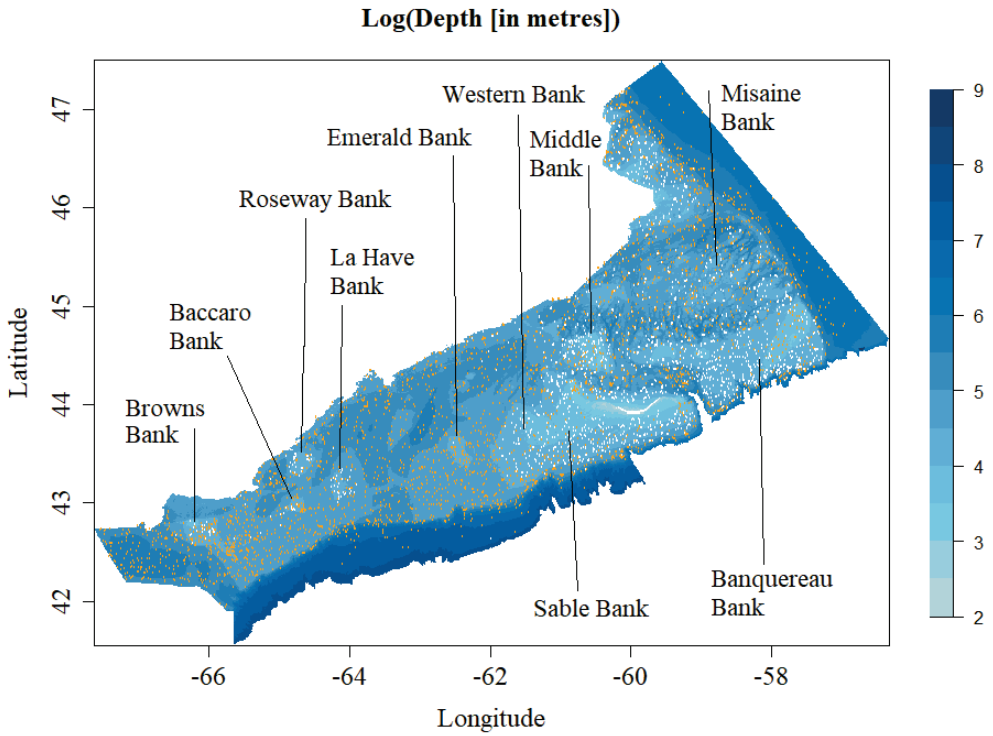


Figure 2.2: A bathymetric map of our spatial domain of interest (and the excluded area around Sable Island) created using values from the DEM. The scale is discrete because the depth values were log-transformed and then binned. Prominent banks in the area are labelled. Also overlaid on the plot are points denoting the survey tows used in the analysis, where white points illustrate the tows that encountered sea cucumbers.

The tows were augmented with information from a DEM as well. Lecours, Devillers, et al. (2016) recommended the consideration of six terrain attributes for ecological studies: relative deviation from mean value (RDMV; which identifies local peaks and valleys), rugosity (which measures roughness of terrain), local mean, slope, and two measures of slope aspect (i.e., easterness and northerness). They argued that this combination maximizes the amount of information extracted from the terrain while reducing the covariation and redundancy. However, they also noted that the local mean is often strongly correlated with depths from the original DEM, and thus its use is only recommended if the DEM itself it very noisy.

The available 100 m resolution DEM (Figure 2.2) was standardized to mean water level and compiled with gridded data from the General Bathymetric Chart of

the Oceans, Canadian Hydrographic Service NONNA-100 products, and DFO multi-beam echosounders. In consultation with its creators, DFO supplied the DEM and the recommended terrain attributes for use in this project. Rugosity and the local mean were provided in metres (i.e., the original units of the DEM), while the slope values were given in degrees and the other three quantities were dimensionless. We used values from the DEM itself for depth measurements, rather than the local mean. Furthermore, we dropped rugosity from consideration in the models because it was almost perfectly correlated with slope ($r \approx 1.00$). Although slope and rugosity explained different components of the topography (Lecours et al., 2016), they provided redundant information. It should also be noted that RDMV was initially undefined for some observations because the slope at their location was zero; any such RDMV values were set to zero too.

Table 2.1: Environmental covariates

Covariate	**Data source**
Bottom temperature (°C)	BNAM (2000-2019)
Bottom temperature range (°C)	BNAM (2000-2019)
Bottom salinity (psu)	BNAM (2000-2019)
Log(Bottom stress $[kg * m^{-1}s^{-2}]$)	BNAM (2000-2019)
Log(Bottom stress range $[kg * m^{-1}s^{-2}]$)	BNAM (2000-2019)
Log(depth [m])	DEM
$\sqrt{\text{Slope}(\circ)}$	DEM
Northerness (unitless)	DEM
Easterness (unitless)	DEM
RDMV (unitless)	DEM

To ensure the analysis was robust to the presence of several large slope values, a square root transformation was applied to slope. Bottom stress, stress range, and depth were log-transformed for similar reasons. The final set of environmental covariates is summarized by Table 2.1. All were centered and scaled prior to their use in modelling.

2.3 Overall Modelling Approach

As expected from a species with a patchy spatial distribution, many tows did not catch any *C. frondosa*. To accommodate this large proportion of zero catch values, we employed a delta modelling approach. In a delta model, positive abundances and the occurrence of a zero value are modelled separately; zero occurrence is treated as a Bernoulli response, whereas positive abundances are generally modelled with a gamma or lognormal distribution (Stefansson, 1996; Fletcher et al., 2005). The predictions from the two sub-models can then be combined to obtain final abundance estimates (Stefansson, 1996; Fletcher et al., 2005).

Define Y_i as the *C. frondosa* catch in kg/km^2 associated with tow i, which occurred at location $\mathbf{s}$ in year t (where t $\in [2000, 2019]$). Further, let $K_i \sim \text{Bernoulli}(p_i)$, where $K_i = 1$ if *C. frondosa* was present in the tow and $K_i = 0$ otherwise. Here, we allowed the positive abundance sub-model to model $log(Y_i|K_i = 1)$ as a Gaussian response, whereas the zero occurrence sub-model treated K_i as a Bernoulli response. We assumed $(Y_i|K_i = 1)$ and K_i to be independent quantities, and related the two sub-models' output through the following relationship:

$$E(Y_i) = E[(Y_i|K_i = 1)(K_i)] = p_i E(Y_i|K_i = 1) = p_i \exp\left(log[E(Y_i|K_i = 1)]\right). \quad (2.1)$$

The positive abundance sub-model provided $E[log(Y_i|K_i = 1)]$ rather than $log[E(Y_i|K_i = 1)]$. However, the two quantities were linked (see Crow and Shimizu (1988)) in that

$$log[E(Y_i|K_i = 1)] = E[log(Y_i|K_i = 1)] + 0.5\sigma^2, \quad (2.2)$$

where σ^2 is the variance of the errors from the positive abundance sub-model. Henceforth, the zero occurrence and positive abundance sub-models will be referred to as the presence and CPUE sub-models respectively.

As implied previously, we fit three spatio-temporal delta models of this form to the data. The first used a GAM framework, whereas the second and third used *starve* and GMRFs respectively. Despite their differences in implementation, all six sub-models incorporated the Snow Crab Survey indicator as well as our entire set of environmental covariates (which are listed in Table 2.1). Model-specific concerns will

now be discussed.

2.4 Model Descriptions

2.4.1 The GAM

Using a GAM framework for the first delta model enabled any non-linear relationships between the environmental covariates and the response to be accurately represented, and it also allowed for the inclusion of a spatio-temporal smooth in each sub-model. To fit GAMs in R, Wood's *mgcv* library (2011) was applied, and smoothness selection was conducted with REML (rather than GCV) as he recommended. Using the notation defined previously, the presence and CPUE sub-models took the form

$$p_i = logit^{-1}(\mathbf{X_i}^*\boldsymbol{\beta_1^*} + f_1(\mathbf{s}, t) + f_2(X_{1i}) + ... + f_{10}(X_{9i})), \tag{2.3}$$

and

$$E[log(Y_i|K_i = 1)] = \mathbf{X_i}^*\boldsymbol{\beta_2^*} + f_{11}(\mathbf{s}, t) + f_{12}(X_{1i}) + ... + f_{20}(X_{9i}) \tag{2.4}$$

respectively. Here, $\mathbf{X_i}^*$ is tow i's row in the model matrix for the parametric components (i.e., the intercept and the survey indicator), $\boldsymbol{\beta_1^*}$ and $\boldsymbol{\beta_2^*}$ are the corresponding parameter vectors, $f_1(\mathbf{s}, t)$ and $f_{11}(\mathbf{s}, t)$ are spatio-temporal smooths, and the remaining terms are smooth functions of the environmental covariates evaluated at tow i's values. As northerness and easterness jointly describe slope aspect, they were incorporated together in a 2D smooth function; the rest of the environmental covariates were given univariate smooth functions.

We further note that each $f_j(\mathbf{s}, t)$ was implemented as the tensor product of a one-dimensional marginal smooth for year and a two-dimensional marginal smooth for space (using eastings and northings). For space, Duchon splines based on first derivative penalisation were used, with their s parameter set to 1/2 (Wood, 2017). This selection was made in an effort to prevent inflated prediction away from the observed data (which was particularly a concern for the CPUE sub-model) (Miller and Wood, 2014; Wood, 2017). All other smooths were incorporated using *mgcv*'s implementation of thin plate splines.

When fitting a GAM in *mgcv*, the basis dimension k must be chosen for each

smooth function. This selection subsequently sets an upper limit on the smooth's potential complexity, with higher values generally leading to greater computational expense (Wood, 2011). Here, this determination was most critical for the spatio-temporal smooths. k was set to 400 and 300 for the presence and CPUE sub-models' respective space marginal smooths, and 5 for each year marginal smooth. Increasing these values seemed to result in little change to model predictions. Meanwhile, the smooth functions for the environmental covariates were each granted a basis dimension of 3 (except for slope aspect, for which k was set to $3 \times 3 = 9$ to accommodate the two variables used to specify it).

2.4.2 The *starve* Model

The second delta model was developed using Lawler et al.'s *starve* framework and R package (2023). In this version, the presence and CPUE sub-models were given the form

$$p_i = logit^{-1}(\mathbf{X_i}\boldsymbol{\beta_1} + w_t(\mathbf{s})), \tag{2.5}$$

and

$$E[log(Y_i|K_i = 1)] = \mathbf{X_i}\boldsymbol{\beta_2} + z_t(\mathbf{s}) \tag{2.6}$$

respectively. Here, $\mathbf{X_i}$ is tow i's row in the model matrix, $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are vectors of fixed effect coefficients, and $w_t(\mathbf{s})$ and $z_t(\mathbf{s})$ are spatio-temporal random effects. Note that intercepts are implicit in *starve* models. Furthermore, *mgcv*-style regression splines were not available in *starve*, and so any non-linear effects had to be approximated through polynomials. Ultimately, a quadratic term for depth was included in both sub-models. In addition, bottom temperature, salinity, and temperature range were granted quadratic terms in the presence sub-model; such terms appeared unnecessary for modelling non-zero CPUE.

The random effect vectors $\mathbf{w}_t(\mathbf{s})$ and $\mathbf{z}_t(\mathbf{s})$ each arose from their own separate time series of random fields. These two time series were described by analogous sets of parameters and temporal random effects, although they are not all written out here, as to avoid cumbersome notation. For example, $\mathbf{w}_t(\mathbf{s})$ could be fully described with

$$\mathbf{w}_t(\mathbf{s})|[\mathbf{w}_{t-1}(\mathbf{s}), \boldsymbol{\epsilon}] \sim NNGP(\phi(\mathbf{w}_{t-1}(\mathbf{s}) - \epsilon_{t-1}) + \epsilon_t, C(\mathbf{s_1}, \mathbf{s_2})), \tag{2.7}$$

and

$$\boldsymbol{\epsilon} \sim N(\mu, \frac{\sigma^2}{(1-\phi^2)}\boldsymbol{\Sigma}), \tag{2.8}$$

where

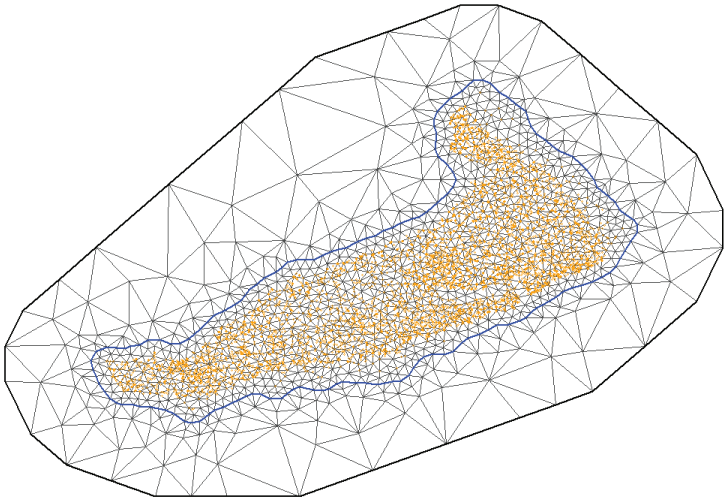
$$\boldsymbol{\Sigma}_{ij} = \phi^{||i-j||}. \tag{2.9}$$

Each field in the time series was assumed to follow a NNGP with a covariance function $C(\mathbf{s_1}, \mathbf{s_2})$, and a mean function defined such that each location temporally evolved through an AR(1) process. Convergence issues were encountered in preliminary models that used a Matérn covariance function with $\nu = 1$. Because of this, we let $C(\mathbf{s_1}, \mathbf{s_2})$ be an exponential covariance function with an unknown spatial range ρ and spatial standard deviation $\tau = \sigma_{MV}/\rho$. Meanwhile, the mean function was informed by a random vector $\boldsymbol{\epsilon}$, which consisted of a random effect ϵ_t for each year in the model. This temporal AR(1) process was directed by its mean μ, one-step-ahead variance σ^2, and AR(1) parameter ϕ (where $|\phi| \leq 1$), all of which were assumed unknown.

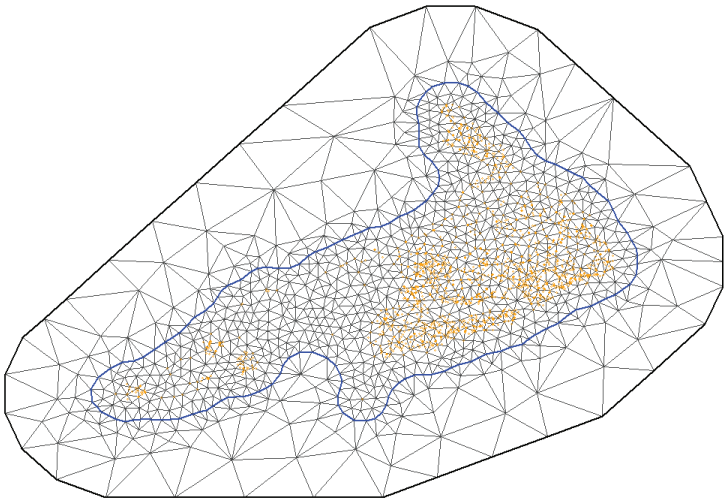
starve allows one to specify the maximum number of parents to be used for each random effect node in the DAGs. Some experimentation indicated that setting this to a value of five was reasonable. Leaving the hyperparameter at its default of 10 led to substantially greater computational expense and little change in the estimated parameters or random effects. This was not particularly surprising; Datta et al. (2016) showed that NNGP inference is fairly robust to such choices. It is also worth mentioning that *starve* models use a persistent DAG to encode the reference set S, and transient graphs (i.e., one for each time step) to encode any observed location not in S (Lawler et al., 2023). We included the locations of all observations in the reference set S, as is the default in *starve*.

2.4.3 The GMRF Model

The final delta model was fit using *sdmTMB* (Anderson et al., 2022). This model differed from its *starve* counterpart in that it included explicit intercepts and specified $\mathbf{w}_t(\mathbf{s})$ and $\mathbf{z}_t(\mathbf{s})$ differently. Of these two main distinctions, the one involving the random effects is more interesting. Once again, each sub-model treated their associated random effects vector as arising from a time series of random fields (which evolved through an AR(1) process). However, these random fields were parameterized in a



a.)



b.)

Figure 2.3: a: The Delaunay triangulation used for the GMRF presence sub-model. b: The Delaunay triangulation used for its CPUE sub-model counterpart. In both cases, the orange dots represent the locations of tows used in model fitting.

slightly different fashion, and the SPDE approach was used to model the random effects rather than NNGPs. We will begin by defining how the random effects were parameterized, before exploring computational concerns.

As with before, we will consider only $\mathbf{w}_t(\mathbf{s})$ for simplicity. We let $W_t(\cdot)$ represent the random field underlying $\mathbf{w}_t(\mathbf{s})$ at time t, and assumed that

$$W_t(\cdot) = \phi W_{t-1}(\cdot) + \sqrt{1-\phi^2} Y_t(\cdot) \tag{2.10}$$

for the temporally independent zero-mean Gaussian random field $Y_t(\cdot)$ (see Anderson et al., 2022). We then gave the process the initial condition

$$W_1(\cdot) = Y_1(\cdot), \tag{2.11}$$

and represented the $Y_t(\cdot)$'s as GMRFs. A Matérn covariance function was used to describe the spatial relationships implicit in each $Y_t(\cdot)$. For this model, we held $\nu = 1$, and estimated the remaining Matérn parameters—i.e., the spatial scale parameter κ and the spatial marginal variance σ^2_{MV}. However, we instead reported the more interpretable spatial range parameter ρ in place of κ. ϕ again served as the unknown AR(1) correlation, now restricted by *sdmTMB*'s implementation so that $|\phi| < 1$.

Fitting a model using the SPDE approach requires the creation of a triangulation (Lindgren et al., 2011). The R-INLA package (Rue et al., 2009) was used to create a separate one for each sub-model (Figure 2.3). Care had to be taken to choose an appropriate number of nodes for each, because a more complex mesh can potentially increase both the computational burden and the resolution of the approximation (Lindgren et al., 2011). In some sense then, these decisions were comparable to those required to setup the DAGs used by the *starve* model or the basis functions used by the GAM. Ultimately, the presence sub-model mesh was given 1365 nodes, whereas its CPUE counterpart was granted 1069 nodes. Model fitting was not prohibitively expensive with these meshes, and further increases in their complexity led to little change in the resulting approximations.

2.5 Model Predictions and Validation

Once residual plots were deemed satisfactory, we used each model to produce prediction maps of *C. frondosa* CPUE across our domain. The predictions were log-transformed to ease their interpretation. Maps of uncertainty were also produced for the log-transformed predictions using Shono's (2008) standard error approximation:

$$SD(log(\widehat{Y_i})) \approx \sqrt{SD[logit(\widehat{p_i})]^2(1-\widehat{p_i})^2 + SD(\widehat{log(Y_i|K_i=1)})^2}. \tag{2.12}$$

Such standard errors rely on Taylor expansions and the delta method.

We started by producing high resolution maps ($\approx 0.02°$) for 2019, the last year for which we had survey coverage. This allowed us to closely examine recent spatial patterns and evaluate whether the models gave plausible and similar fine scale predictions. When we were satisfied that the maps seemed reasonable, they were shown to experts on the sea cucumber fishery, including those from industry, as a qualitative assessment of the models' predictions in 4W and 4Vs; these experts deemed our predictions sensible. Next, with the aim of identifying any macroscale historical trends, a set of lower resolution ($\approx 0.06°$) maps were generated for each year for which we had survey coverage. All decisions regarding output resolution were made as a compromise between output quality and computational expense.

It was necessary to preprocess the environmental data layers before the maps could be created. The layers were first subsetted to our domain of interest and projected onto the UTM system. From there, they were bilinearly resampled such that they shared a common grid at the required resolution. The temporally varying data layers required slightly more preprocessing for this task than their static counterparts; for them, the model was given each location's year-specific average value (i.e., the average across all months in the year of interest). In addition, the Snow Crab Survey indicators were set to one during prediction. This decision to assume Snow Crab Survey gear only affected the scale of the predicted catch values, and not the relative differences between locations and across years.

Chapter 3

Results

This chapter will begin by individually examining our three model fits and their predictions. From there, it will thoroughly compare the models. Model summary tables are given in Appendix A, and supplementary figures (including residual plots) are provided in Appendix B. A 5% significance level was used where necessary.

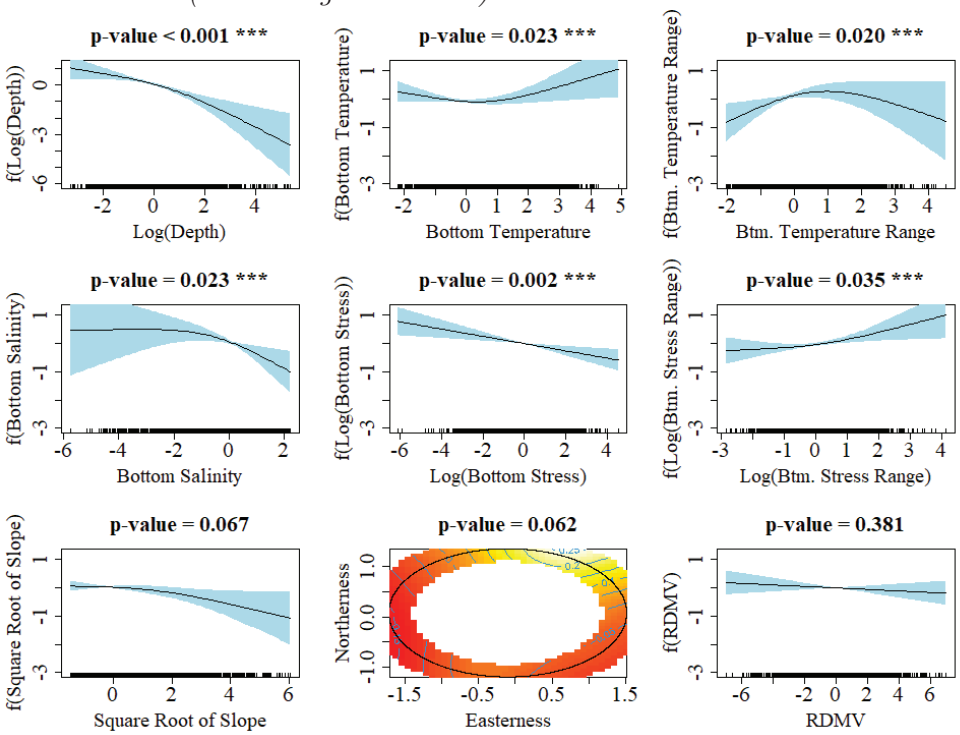
3.1 Model Fits and Predictions

3.1.1 The GAM

The smooth functions of the covariates are illustrated (with p-values and 95% confidence bounds) in Figure 3.1. Plots of the spatio-temporal terms (i.e., $f_1(\mathbf{s}, t)$ and $f_{11}(\mathbf{s}, t)$) are located in Appendix B (Figures B.10 and B.11). The majority of the model's explanatory power can be attributed to the spatio-temporal smooths. The presence sub-model explained 45.30% of its null deviance, a substantially higher quantity than the 17.20% explained if it was refit with the spatio-temporal smooth function removed. A similar phenomenon occurred with the CPUE sub-model, which saw 72.69% vs 34.90%. Thus, our covariates appeared insufficient alone to explain the variability in sea cucumber abundance.

RDMV, $\sqrt{\text{slope}}$, easterness, and northerness were not significantly informative to the GAM. Most of the other covariates, including log(bottom stress) and log(bottom stress range), were significant solely when predicting sea cucumber presence. Log(depth) and bottom temperature range were unique among our environmental variables in that their effects were significant in both sub-models. They were also the only significant environmental predictors of non-zero CPUE. Log(depth) was the more informative of the two; the wide ranges associated with its smooth functions implied that they contributed strongly to predictions. Both catch values and the probability of a presence tow were substantially lower at higher depths (all else being equal). Meanwhile,

Presence sub-model (on the log odds scale)



CPUE sub-model

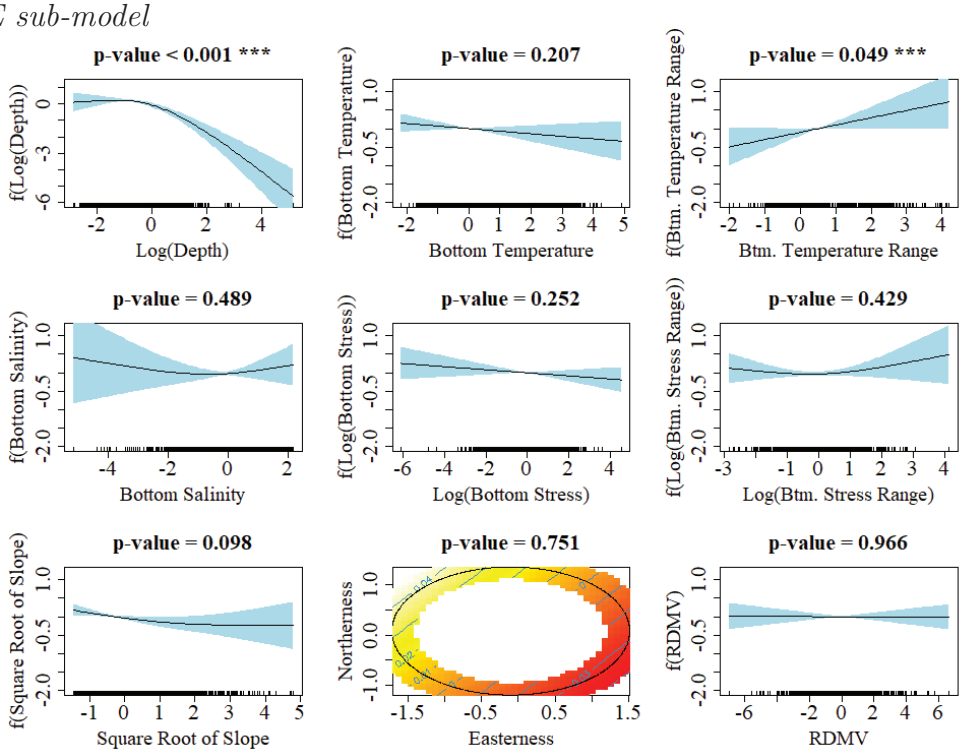


Figure 3.1: Partial effect plots (with 95% confidence bounds) for all the smooth functions in the GAM except for the spatio-temporal smooths. For the easterness-northerness smooth, lighter shades denote areas where the effect is more positive. The axis ticks illustrate the values of the predictors for which data were available. *** indicates the term is significant at the 5% significance level.

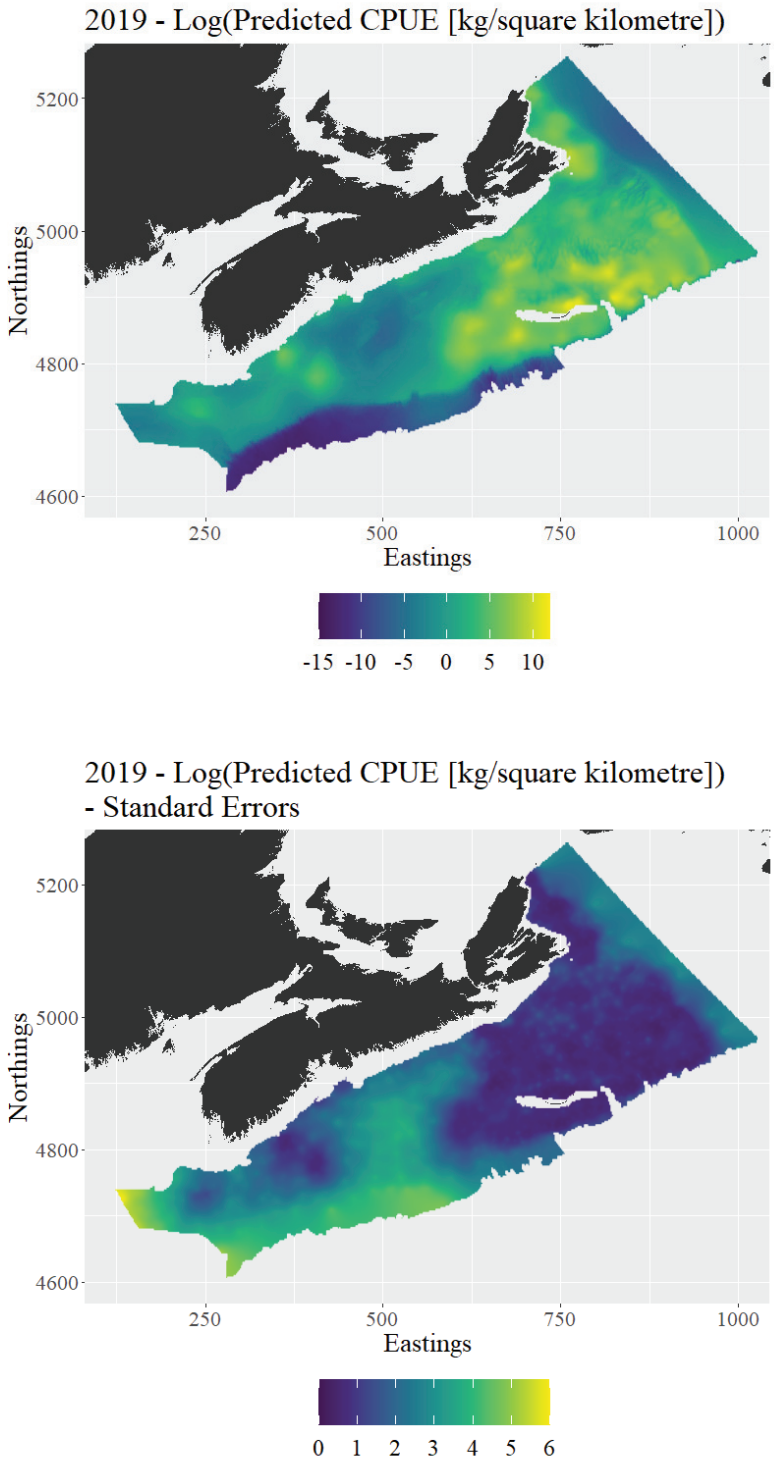


Figure 3.2: The GAM's prediction and associated uncertainty for 2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicators set to one).

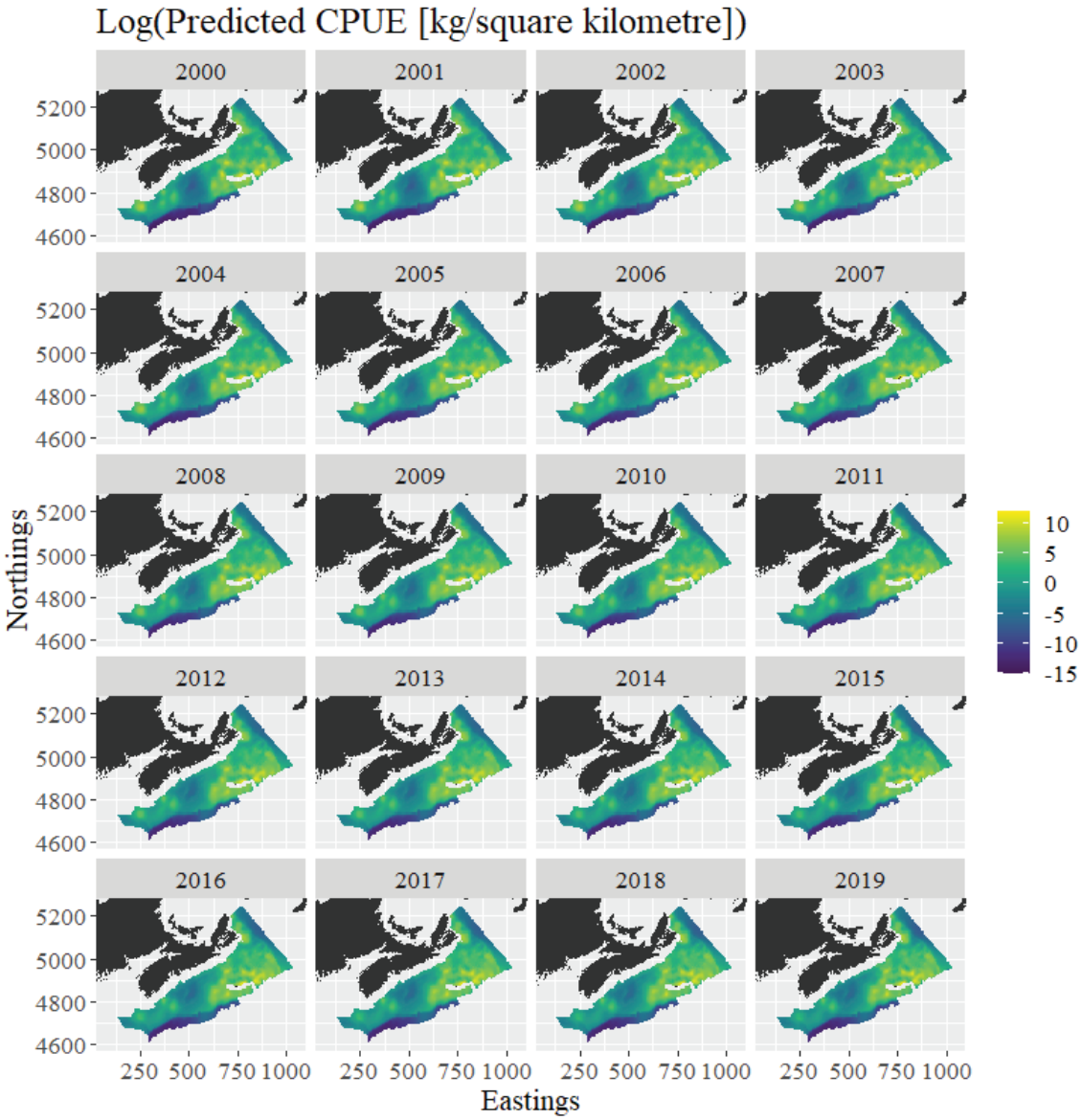


Figure 3.3: Predictions from the GAM for the years 2000-2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicators set to one).

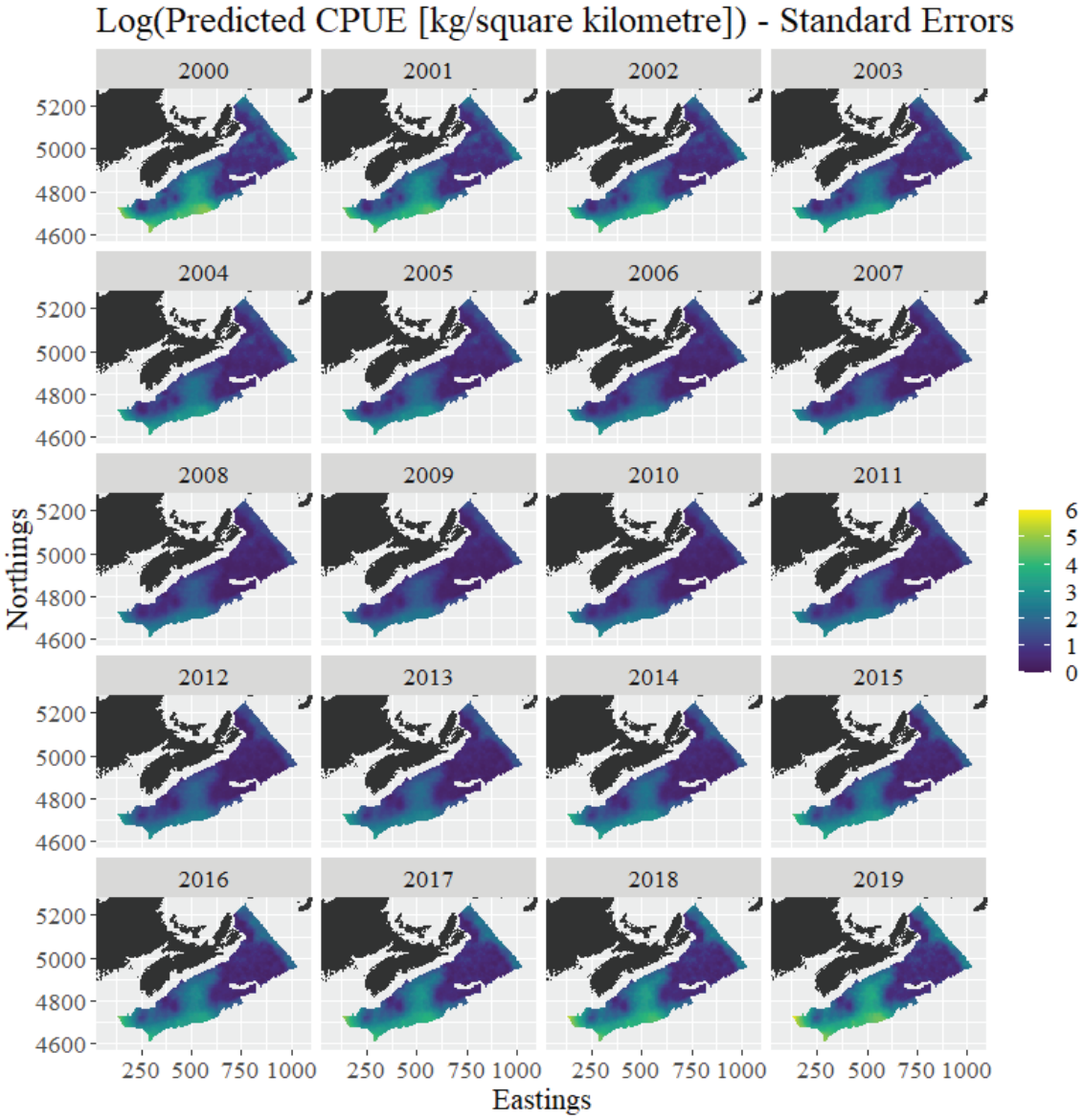


Figure 3.4: Uncertainty estimates for the GAM's 2000-2019 predictions.

higher temperature ranges were clearly associated with higher catch values where *C. frondosa* was encountered. However, while the presence sub-model's temperature range smooth was estimated to be approximately parabolic in shape, its confidence bounds made it challenging to interpret with any degree of certainty.

According to the GAM's predictions (Figures 3.2 and 3.3), most *C. frondosa* habitat existed in the eastern half of our spatial domain—i.e., off of Cape Breton, or on Sable Bank, Middle Bank, or Banquereau Bank. However, some pockets of habitat (albeit with a lower probability of sea cucumber presence) were also predicted in the west on Browns Bank, Baccaro Bank, Roseway Bank, and La Have Bank. The model further indicated that sea cucumber abundance was not uniformly distributed across habitat areas; particularly strong hotspots of CPUE were located on the three aforementioned banks in the east. Naturally, these predictions were interpreted with prediction uncertainty in mind (Figure 3.4). The GAM estimated low standard errors inside core sea cucumber habitat, and relatively higher ones elsewhere. This disparity was more pronounced in earlier and later years. The standard errors were clearly highest for the westernmost tip of our spatial domain, because of high uncertainty in the predicted log odds of presence there. They were second highest on the slopes, where only one tow encountered sea cucumbers.

The GAM suggested that sea cucumber abundance was relatively stable across our time series in the east; however, it did detect some shifts in distribution there (Figure 3.3). For example, its CPUE predictions for the northernmost point of Sable Bank declined early in the time series. This decrease was predicted due to a drop in *C. frondosa* catch values when the species was caught there. The GAM also predicted an increase in catch rates in the late 2010s south of that area and a gradual increase in abundance on the northwestern tip of Banquereau Bank. The former change resulted from a regional increase in the predicted probability of presence. Meanwhile, the latter was driven by a local increase in predicted non-zero CPUE across most of our time series.

More substantial changes occurred in the west (Figure 3.3). In the early 2000s, a part of Browns Bank was predicted to be the dominant sea cucumber aggregation in the area. However, *C. frondosa* abundance there and at nearby Baccaro Bank slowly declined throughout the remainder of the time series. These western shifts

in predicted abundance were mainly driven by changes in the presence sub-model's predictions.

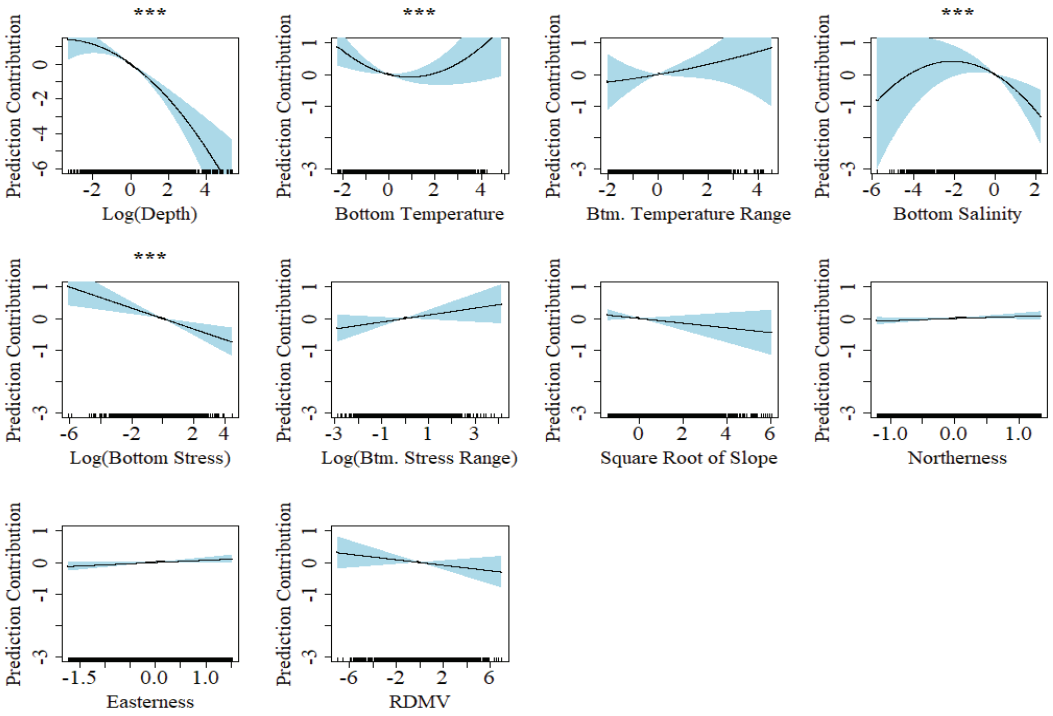
3.1.2 The *starve* Model

Partial effect plots for the environmental variables (with 95% confidence bounds) are given in Figure 3.5. According to the *starve* model, log(depth) was again predictive. All else held constant, *starve* suggested that this covariate had a concave-down parabolic relationship with the log odds of presence. However, the covariate's relationship with non-zero CPUE was more uncertain; a 95% confidence interval for log(depth)'s first-order term would be negative across its entire range, but an analogous interval for the second-order term would contain zero. Bottom temperature and salinity were the only other variables significantly informative to both sub-models. This model suggested they had, respectively, concave-up and concave-down parabolic relationships with the log odds of a presence tow (after holding the other predictors fixed). Furthermore, CPUE was predicted to decrease with increasing bottom temperature, whereas CPUE increased with increasing bottom salinity. Log(bottom stress) was a significant predictor of sea cucumber presence, while bottom temperature range and log(bottom stress range) had significant effects in the CPUE sub-model. All other environmental predictors were deemed not significant by *starve*.

We now shift our attention to two of the spatio-temporal parameters, starting with the AR(1) correlation ϕ. *starve* ϕ estimates were close to unity ($\widehat{\phi}_{presence} = 0.997$, standard error = 0.004; $\widehat{\phi}_{CPUE} = 0.941$, standard error = 0.009), suggesting that each sub-model's time series of random fields arose from an approximation of a random walk process. Random walks are often appropriate for modelling the distribution of sedentary and patchily-distributed species (Commander et al., 2022). The spatial range ρ was sensibly larger for the presence sub-model ($\widehat{\rho}_{presence} = 32.642$, standard error = 3.763; $\widehat{\rho}_{CPUE} = 9.705$, standard error = 0.361); this suggested that the random fields associated with CPUE exhibited more fine-scale spatial variation than their presence counterparts.

These properties of the random effects were better understood by examining model predictions (Figures 3.6 and 3.7). The 2000-2004 predictions differed from their GAM analogues, and thus some of the early shifts in abundance that model detected were

Presence sub-model (on the log odds scale)



CPUE sub-model

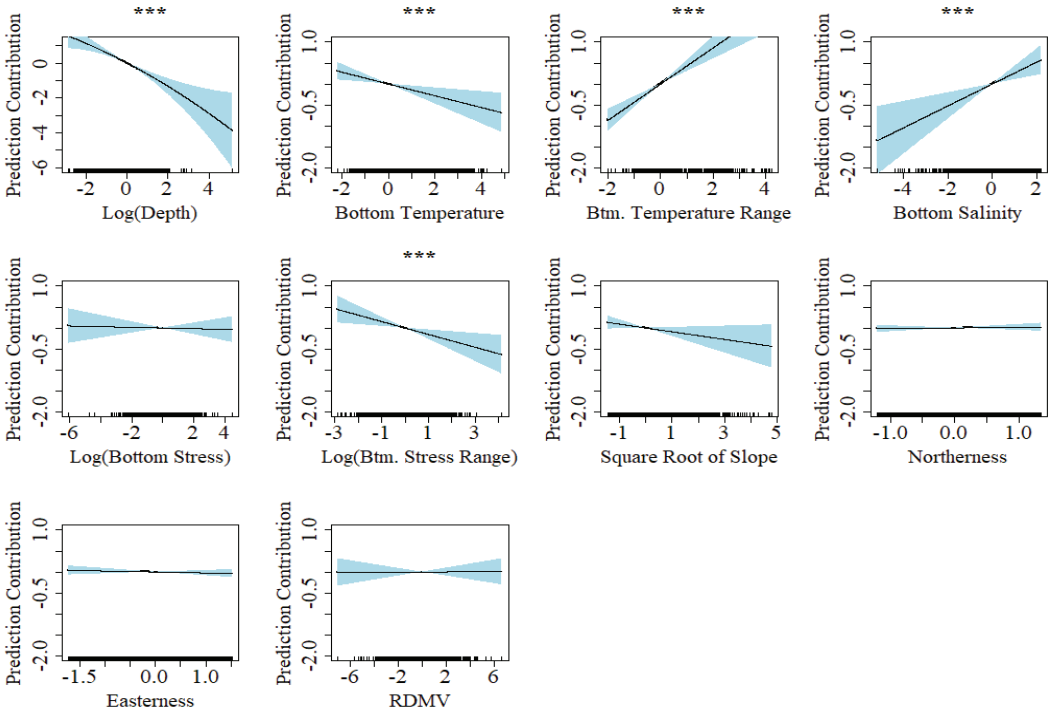


Figure 3.5: Partial effect plots (with 95% confidence bounds) for all environmental variables incorporated into the *starve* model. The axis ticks illustrate the values of the predictors for which data were available. *** indicates the variable has a term significant at the 5% significance level.

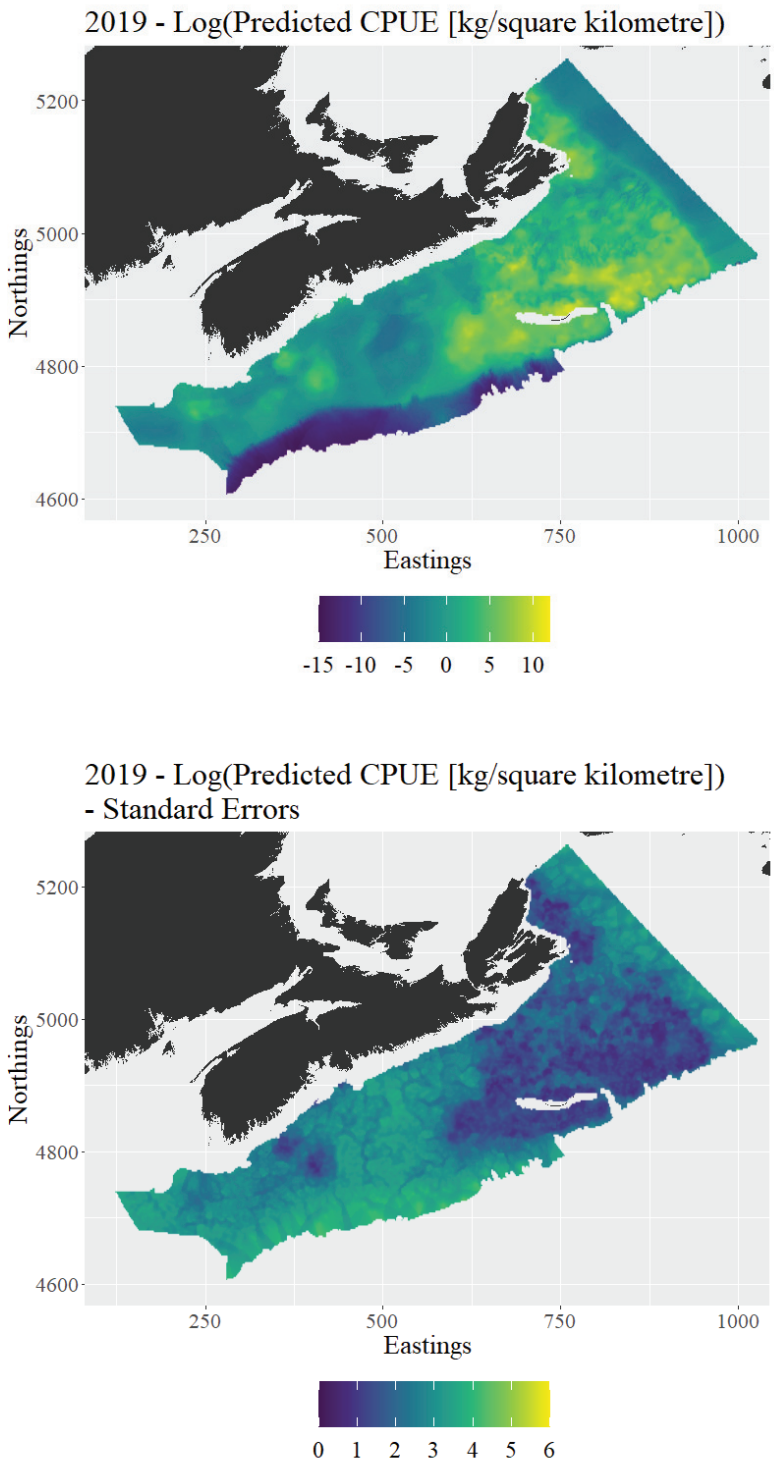


Figure 3.6: The *starve* model's prediction and associated uncertainty for 2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicators set to one).

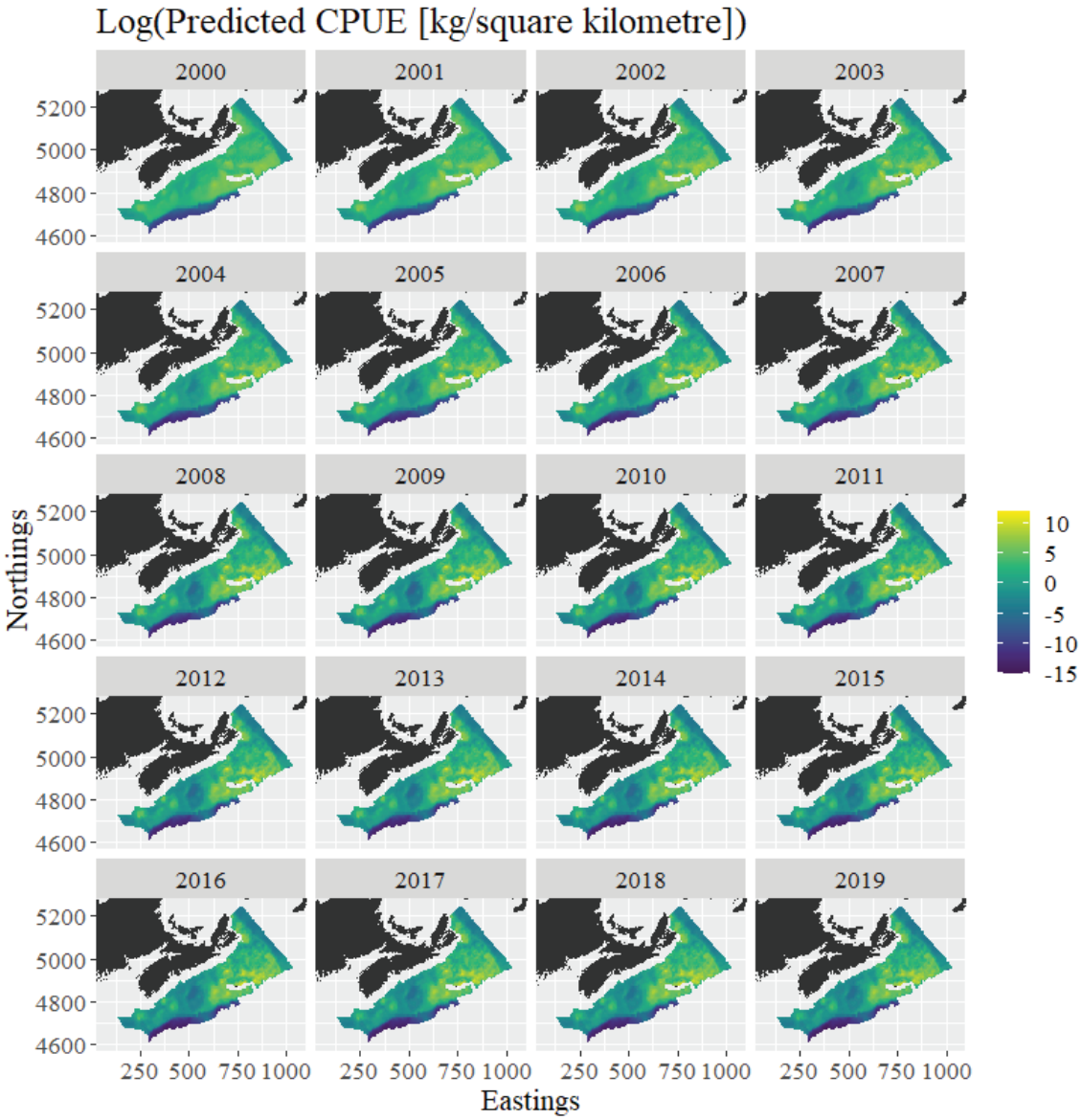


Figure 3.7: Predictions from the *starve* model for the years 2000-2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicators set to one).

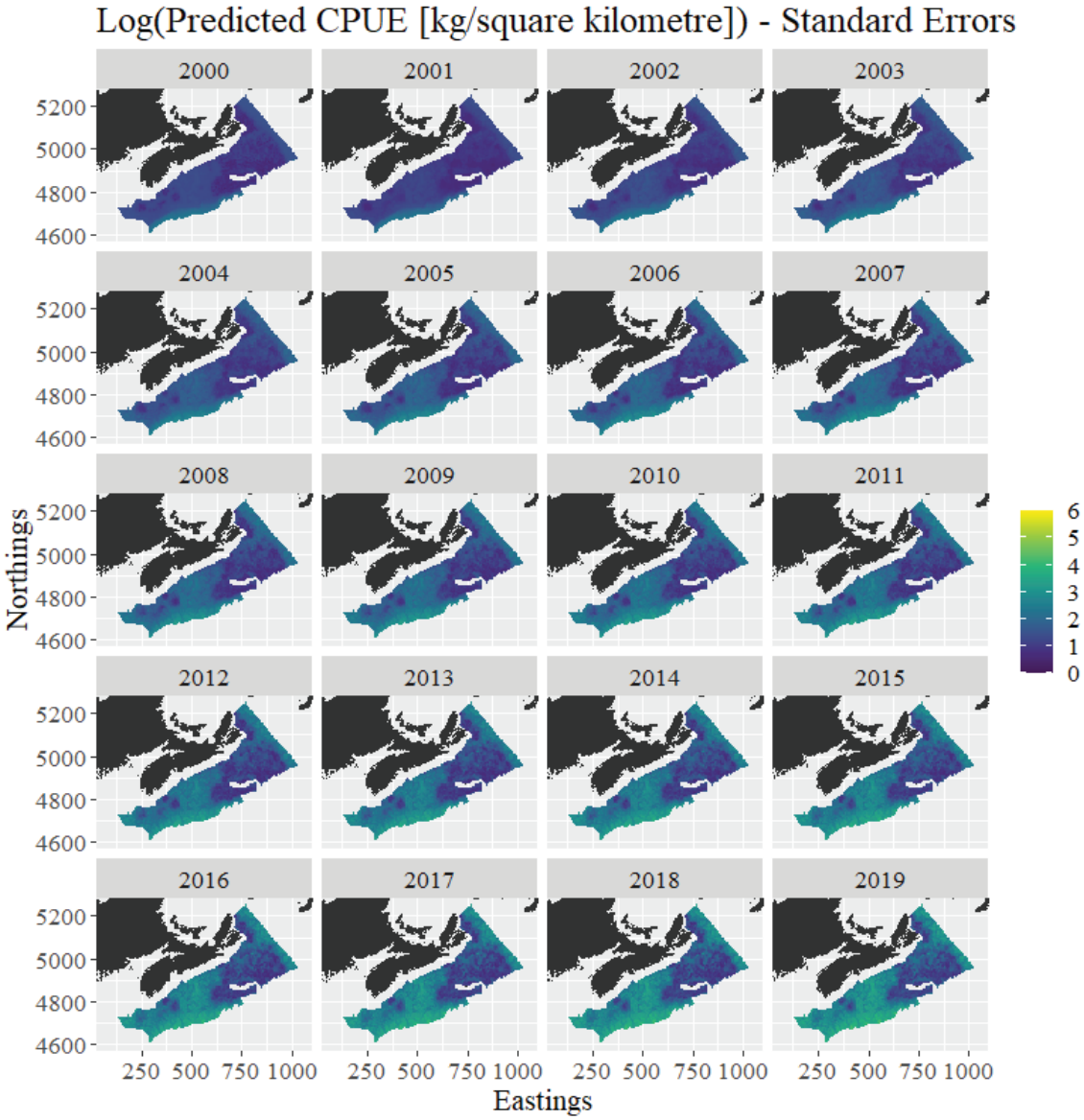


Figure 3.8: Uncertainty estimates for the *starve* model's 2000-2019 predictions.

less evident here. After 2004, the *starve* model's predictions were essentially consistent with those from the GAM with three exceptions. First, its predictions were noticeably less smooth than the GAM's. Second, *starve* predicted higher abundances on Middle Bank. Third, it detected a noticeable late 2010s decline in CPUE immediately northeast of Sable Island and on Middle Bank. Note that the *starve* model's prediction uncertainty steadily grew throughout the time series (Figure 3.8). For most years, *starve* was more uncertain of its predictions inside core sea cucumber habitat than the GAM was; this was particularly evident on Browns Bank. However, *starve*'s standard errors were often much lower in the far west and on the slopes.

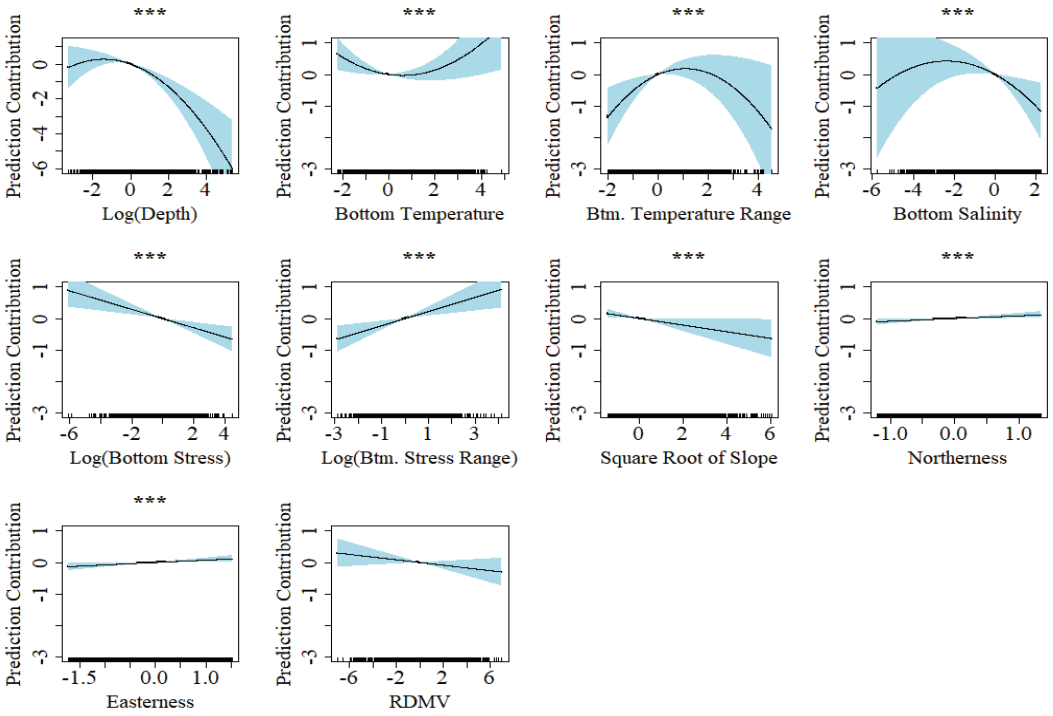
3.1.3 The GMRF Model

Partial effect plots for the environmental variables (with 95% confidence bounds) are provided in Figure 3.9. The model considered all environmental covariates, with the exception of RDMV, to be significant predictors of *C. frondosa* presence. In contrast, log(depth), bottom salinity, and bottom temperature range were the only significant environmental predictors of abundance when the species was caught. The model indicated that CPUE decreased with increasing bottom temperature, and presence log odds troughed at intermediate bottom temperatures (after holding our other covariates constant). Presence log odds peaked at intermediate bottom temperature ranges, whereas CPUE increased with increasing bottom temperature range. Bottom salinity's effect on presence log odds was uncertain as the covariate's quadratic term was not significant. However, bottom salinity clearly had a positive effect on CPUE.

The ϕ parameter values were near their upper boundary ($\widehat{\phi}_{presence} = 0.990$, standard error = 0.003; $\widehat{\phi}_{CPUE} = 0.985$, standard error = 0.004). This signified—as it did with our *starve*-based approach—that the random effects temporally evolved through an approximation of a random walk. Furthermore, the presence sub-model's ρ estimate was larger than its CPUE sub-model counterpart ($\widehat{\rho}_{presence} = 56.118$, standard error = 6.810; $\widehat{\rho}_{CPUE} = 28.756$, standard error = 3.667). This again hinted that the CPUE random fields exhibited more fine-scale spatial variation than their presence counterparts.

Compared to the GAM, the GMRF model predicted higher abundances on Middle Bank and near the center of our spatial domain (Figures 3.10 and 3.11). Outside of

Presence sub-model (on the log odds scale)



CPUE sub-model

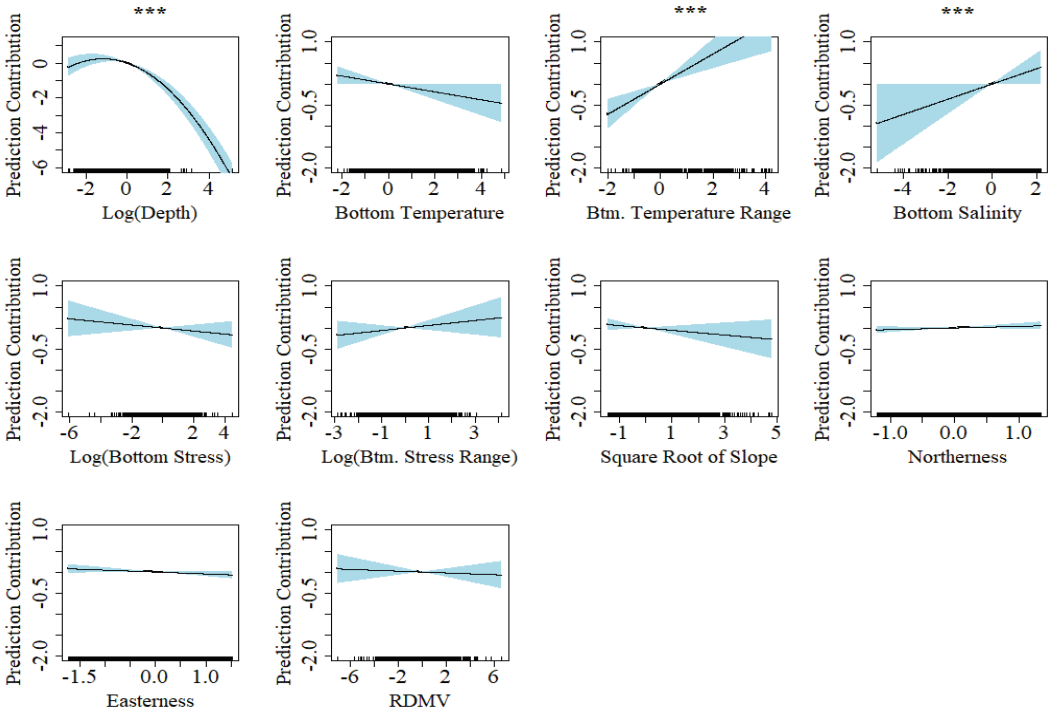


Figure 3.9: Partial effect plots (with 95% confidence bounds) for all environmental variables incorporated into the GMRF model. The axis ticks illustrate the values of the predictors for which data were available. *** indicates the variable has a term significant at the 5% significance level.

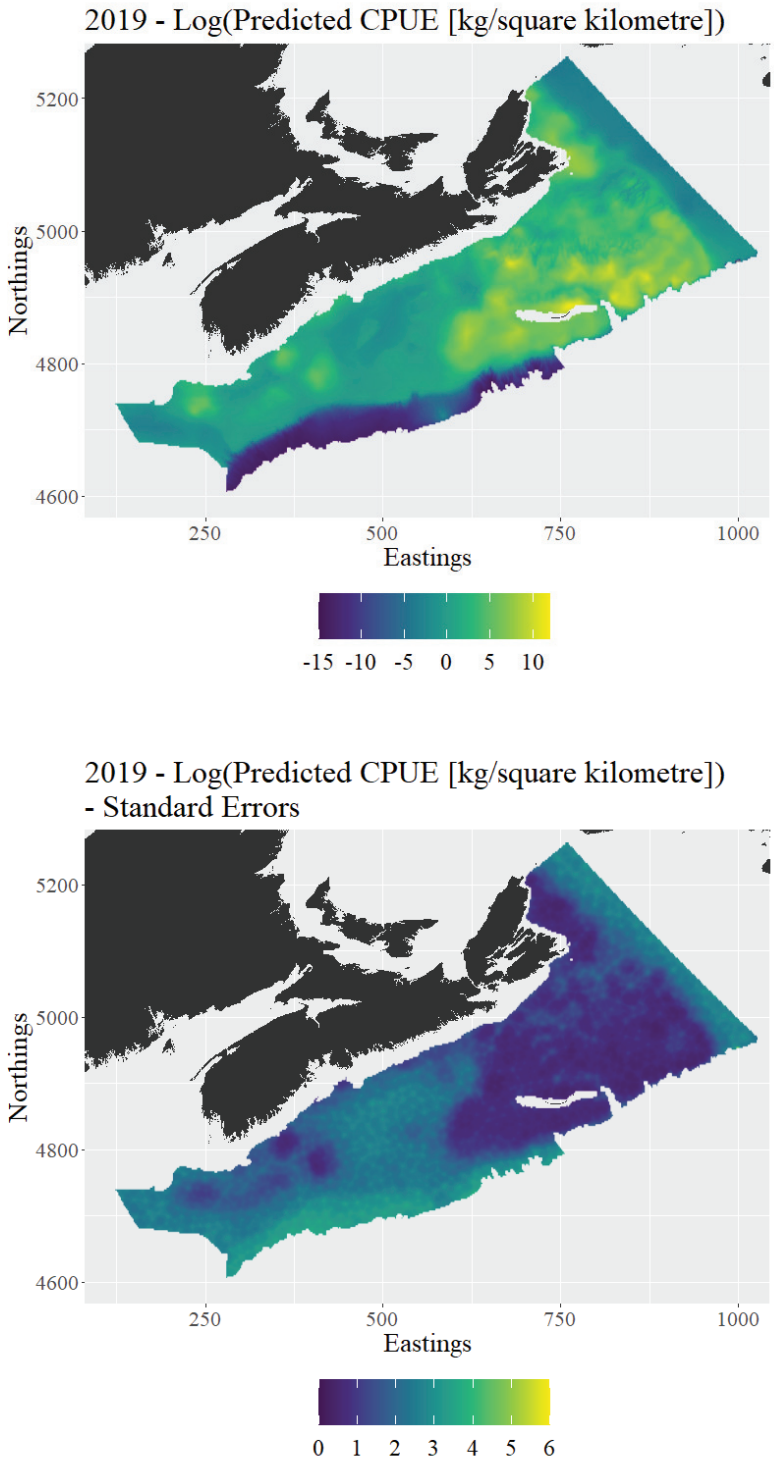


Figure 3.10: The GMRF model's prediction and associated uncertainty for 2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicators set to one). Note that the two triangulations underlying the model are somewhat visible in the bottom plot.

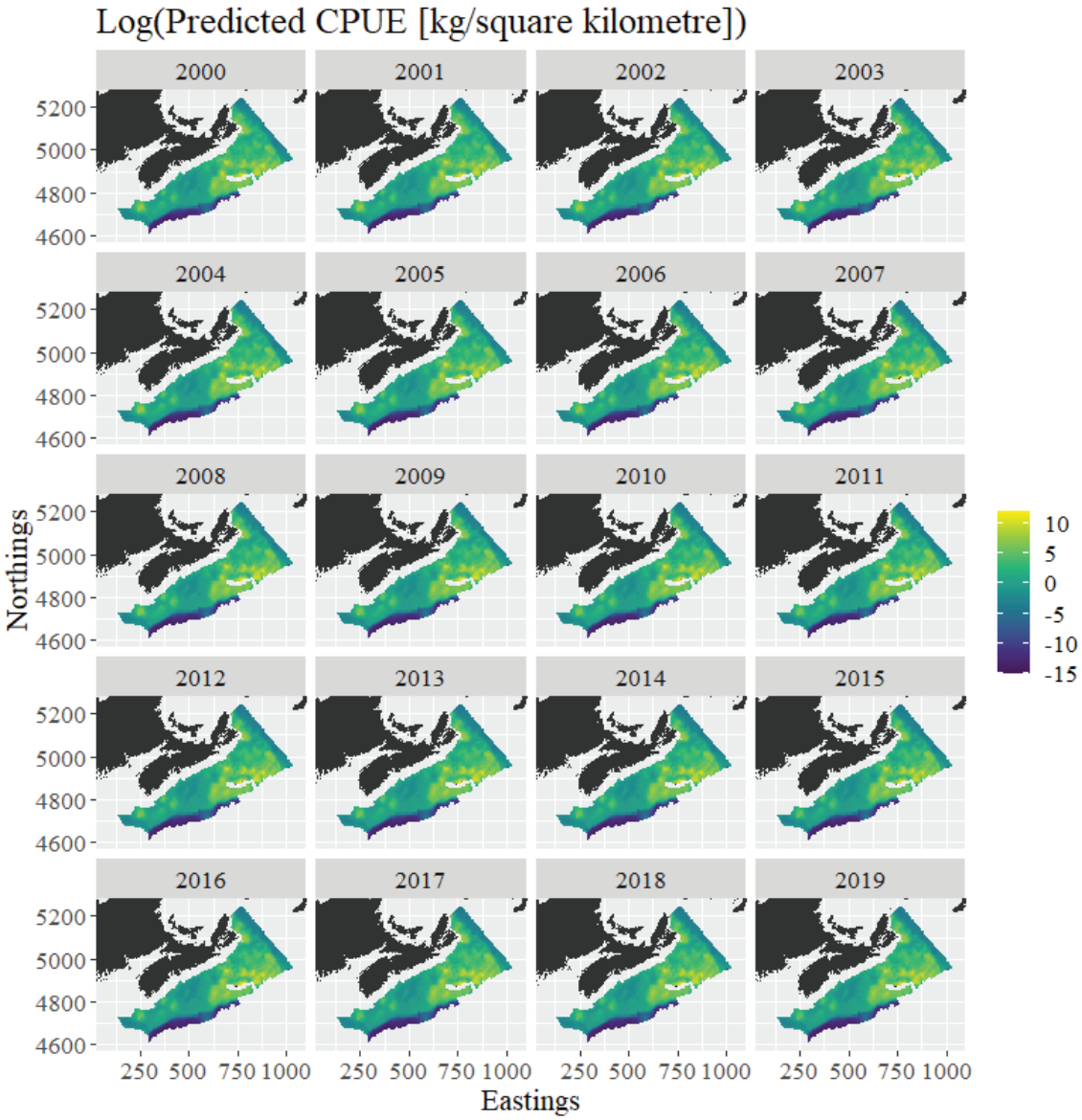


Figure 3.11: Predictions from the GMRF model for the years 2000-2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicators set to one).

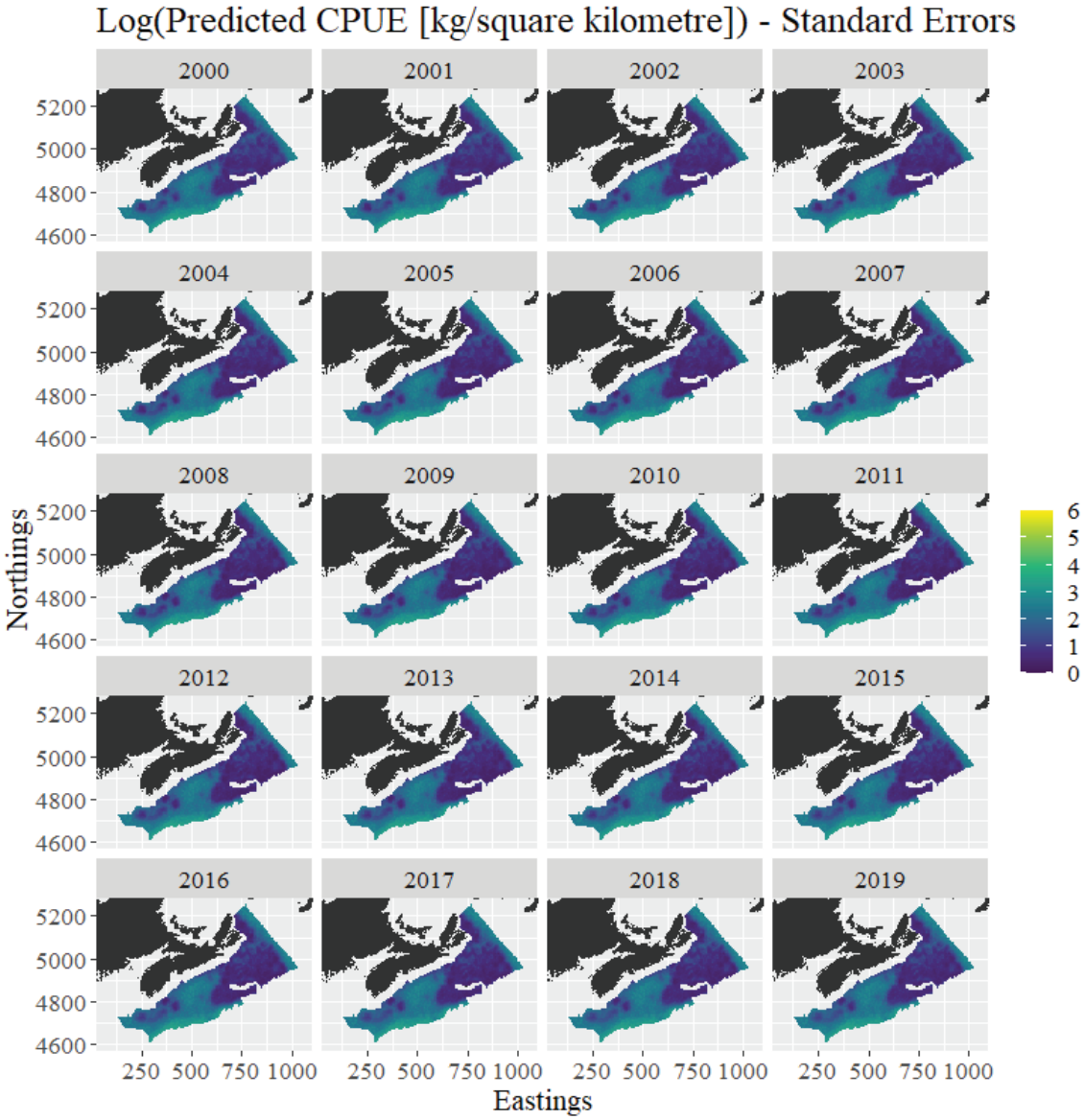


Figure 3.12: Uncertainty estimates for the GMRF model's 2000-2019 predictions.

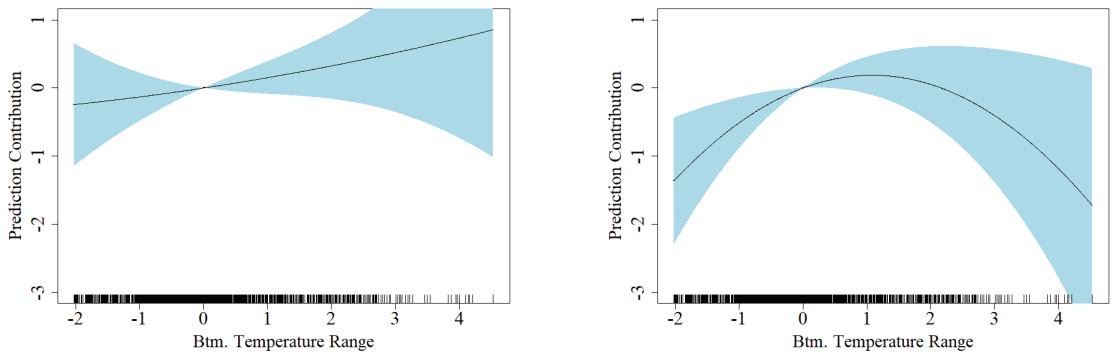
these regions, the two models' predictions were essentially consistent. Their estimates of prediction uncertainty were somewhat less congruous (Figure 3.12). The standard errors estimated by the GMRF model changed substantially less year-over-year, and they were generally lower in the far west. Furthermore, mesh vertices were visible in the GMRF-based standard errors (Figure 3.10).

3.2 Model Comparisons

As previously discussed, the GAM used smooth functions to account for any spatio-temporal autocorrelation, whereas our other two models relied on random effects which evolved temporally through an AR(1) process. The effects of the environmental covariates were also implemented differently between our three approaches; while the GAM utilized regression splines and thus flexibly represented any nonlinear relationships, its mixed model counterparts used a set of fixed effect coefficients. Because of these broad differences, the GAM considered bottom temperature range and log(depth) to be the only significant environmental predictors of non-zero CPUE, while our other approaches deemed additional covariates informative (such as bottom salinity).

The models based on random effects sometimes disagreed with each other as well, despite being structurally similar. For example, our GMRF-based approach indicated that bottom temperature range had a concave-down parabolic relationship with the log odds of a presence tow, and implied that log(depth) had a similar relationship with non-zero CPUE. By comparison, our *starve* model deemed log(depth)'s second-order term and both bottom temperature range terms not significant when predicting non-zero CPUE and presence, respectively. We additionally note that northerness, easterness, and $\sqrt{\text{slope}}$ were significant predictors of presence solely in the GMRF model. The standard errors associated with the GMRF model's fixed effects were similar to their *starve* model analogues. Nonetheless, because of distinctions in how the two models implemented random effects and intercepts, their fixed effect estimates sometimes diverged substantially (e.g., see Figure 3.13).

Clearly, each approach produced different overall conclusions regarding the significance of the covariate effects. Yet, in instances where their conclusions agreed, we had stronger evidence that the effects were meaningful. This was perhaps most



(a) *starve* model (b) GMRF model

Figure 3.13: Partial effect plots for bottom temperature range from the two presence sub-models based on random effects. 95% confidence bounds are included. The axis ticks illustrate the values of the predictors for which data were available.

notably the case with the effects of log(depth) and log(bottom stress). Log(depth) was significant in all six sub-models. Similarly, all three models deemed log(bottom stress) a significant predictor of sea cucumber presence, but not non-zero CPUE.

We note that the random effects used by the GMRF and *starve* models shared some properties. The ρ estimates from the GMRF-based approach were substantially larger than those found by *starve*. This was likely another consequence of subtle model specification differences. Both approaches indicated that ρ was larger for the presence random fields, however. Furthermore, their ϕ estimates were alike, and implied that the random effects evolved temporally through an approximation of a random walk process. Of course, the two models' random effects were most easily compared by examining predictions.

Despite any dissimilarities mentioned thus far, our models concurred in their predictions overall. Still, a few differences between the three sets of predictions (Figure 3.14) should be highlighted. For example, the GAM somewhat de-emphasized the sea cucumber aggregation on Middle Bank. Simultaneously, the GMRF model produced slightly higher predictions near the middle of our spatial domain; only one tow encountered sea cucumbers in this region, and the frameworks therefore had to rely on limited spatio-temporal information there. We further note that the *starve* model's random fields for the years 2000-2004 were affected by how *starve* parameterizes its

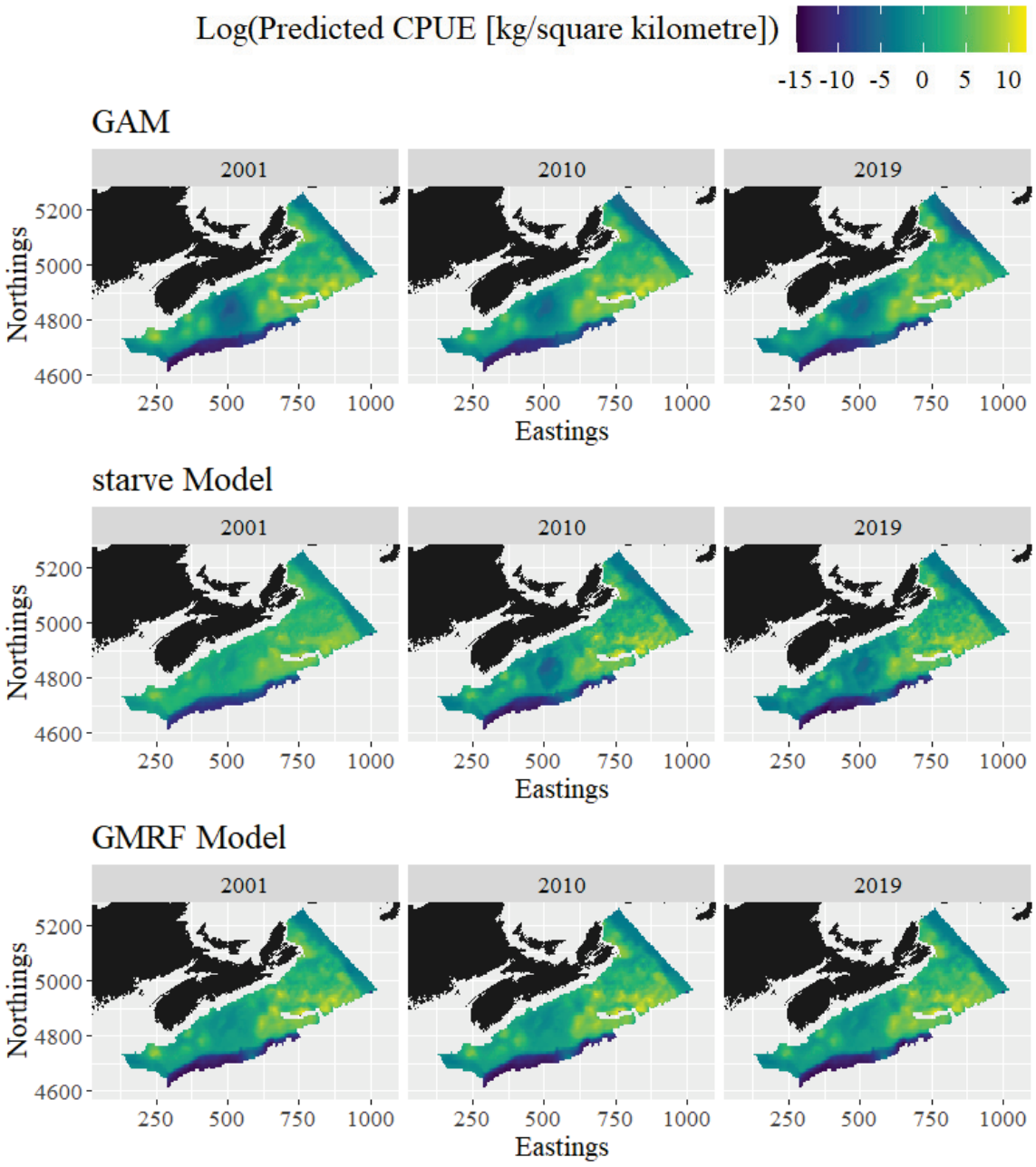


Figure 3.14: The models' predictions for three years in our time series, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicators set to one).

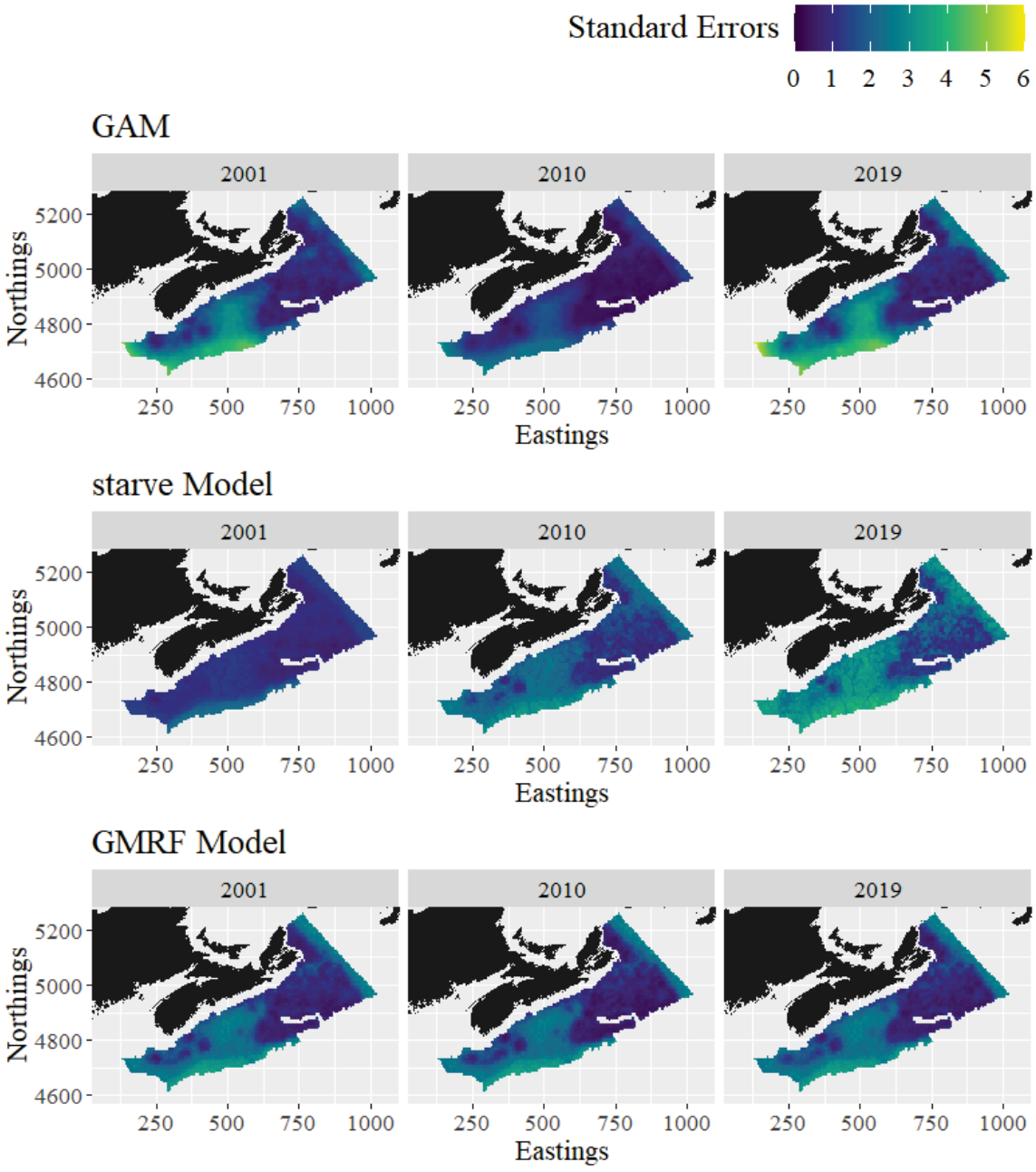


Figure 3.15: The models' estimates of prediction uncertainty for three years in our time series.

initial conditions (as of writing). This phenomenon led the model to give rather different predictions for these initial years. Once beyond this time period, its predictions became highly similar to those from the other models—albeit less smooth, due to its use of the exponential covariance function. However, *starve* detected an appreciable late 2010s decline in CPUE immediately northeast of Sable Island and on Middle Bank. The other approaches did not.

The models' estimates of prediction uncertainty (Figure 3.15) were arguably more sensitive to methodology differences than the predictions themselves. This was exemplified by the GMRF model's standard errors, which visibly depended on the triangulations used to incorporate the random effects (Figure 3.10). Those standard errors changed minimally across the years. In contrast, the *starve* model's standard errors generally grew over time, and the GAM's were smallest around 2010. The former model generally had the highest predictive uncertainty inside core *C. frondosa* habitat, whereas the latter often had the most uncertain predictions for the far west and on the slopes. The large fluctuations in the GAM's standard errors were caused by uncertainty in the time marginal smooths for earlier and later years. The approaches based on random effects did not use smoothing to model temporal variability; rather, they allowed their random fields to evolve over discrete timesteps.

The spatio-temporal components of the models outperformed the environmental covariates. Similarly, the covariates were unable to predict noteworthy spatio-temporal changes in abundance (such as on Browns Bank). This all indicated that substantial information would have been lost had we ignored the dependencies in the data.

Chapter 4

Conclusion

Because of the inherently spatial characteristics of *C. frondosa*, such as its sedentary nature and tendency to aggregate (Orensanz et al., 1998), we anticipated the need for a rigorous spatio-temporal approach. This choice was validated, as considerable variability in the species' distribution could not be explained by the examined covariates alone. Our predictions were greatly affected by our decision to account for underlying spatio-temporal dependence. However, they were less affected by our choice of framework. While the three frameworks differed somewhat in their predictive uncertainty, their post-2004 predictions only markedly diverged on small portions of the banks, and at one location far from core sea cucumber habitat. This outcome should provide reassurance to the practitioner whose primary interest is to simply map the distribution of a species. When troubled by the decision of which framework to use, we recommend using the one most familiar, if this is the only objective of an analysis.

Yet, we find that comparing the output from multiple frameworks can be useful as a validation technique. It is also helpful when the identification of predictive environmental covariates is of principal concern. While our models generally concurred in their predictions, each gave different conclusions regarding the significance of the covariates. $\sqrt{\text{Slope}}$, easterness, and northerness were significant only in the GMRF presence sub-model. In contrast, log(bottom stress) was significant in all three presence sub-models; all else equal, the log odds of presence appear to decrease linearly as this covariate increases. Because bottom stress can serve as a proxy for sediment grain size (Ward et al., 2015), its utility here was relatively unsurprising. Log(depth) was unique in that it was statistically significant in all six sub-models; they essentially suggested that *C. frondosa* prefers the relatively lower depths afforded by the banks. We are confident these covariates have meaningful effects, given that the models agreed in their significance. We cannot say the same for the effects of $\sqrt{\text{slope}}$, easterness, and northerness. Examining multiple models allows us to make these

distinctions between the predictors.

Our results involving the covariates seem reasonable. However, there is little to compare them to because no ecological studies of a similar nature have been conducted on this population. It is also worth noting that the smooth functions and fixed effect estimates we obtained might have been somewhat influenced by spatial confounding. The phenomenon generally biases covariate effect estimates, and arises when some subset of a model's predictors are correlated with its spatial/spatio-temporal effects (Clayton et al., 1993; Reich et al., 2006). It can be expected to some degree whenever the predictors used in a spatial or spatio-temporal model exhibit strong autocorrelation themselves (Hanks et al., 2015). Fortunately, the presence of spatial confounding does not hinder prediction to unobserved locations (and in truth, may actually help) (Page et al., 2017).

Our predictions support the findings of Shackell et al. (2013), who noted the existence of large sea cucumber aggregations on Middle Bank, Banquereau Bank, and Sable Bank, and smaller ones on La Have Bank, Roseway Bank, Baccaro Bank, Browns Bank, and off of Cape Breton. Yet, our models indicate that the aggregations on Browns Bank and Baccaro Bank have diminished substantially in abundance since the mid-2000s. We further state that our predictions for 2019 generally concur with DFO's expected habitat map (see Figures 4.1 and 4.2). It is more challenging to meaningfully compare our results with those described in Harper (2020). They showed that one can estimate *C. frondosa* biomass from drop-camera images. Nonetheless, they used their images to study counts instead, and found evidence that the areas with the highest count-based densities did not have the largest sea cucumbers.

To our knowledge, all other inquiries into this population's distribution have taken purely spatial approaches. Given the characteristics of sea cucumbers, this is understandable. Nonetheless, we posit that future examinations would benefit from taking a spatio-temporal approach (if it is feasible). The spatio-temporal methodologies we used here give further insight into how *C. frondosa* habitat and abundance have changed over time. They allowed us to detect the apparent decrease in abundance on Browns Bank, for example. This decline does not appear to be an artifact of the survey data, and could be related to an increase in scallop dredging on the bank (Hubley et al., 2014). If this is the case, than the models' spatio-temporal smooth

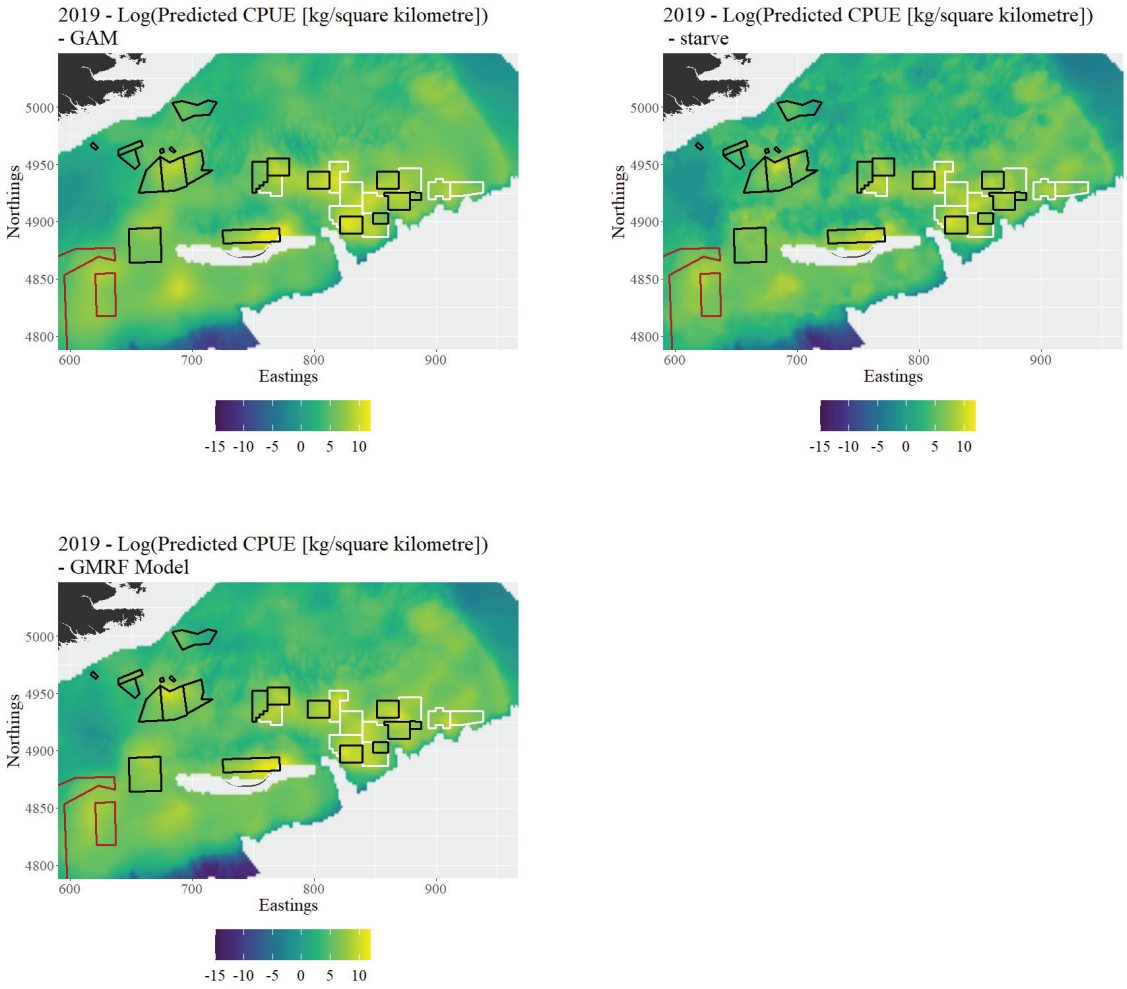


Figure 4.1: These plots provide a closer look at each model's abundance predictions for 2019 in the eastern half of our spatial domain. The brown lines denote the boundaries of the Western/Emerald Banks Conservation Area, while the white and black lines delineate sea cucumber reserves and fishing zones respectively.

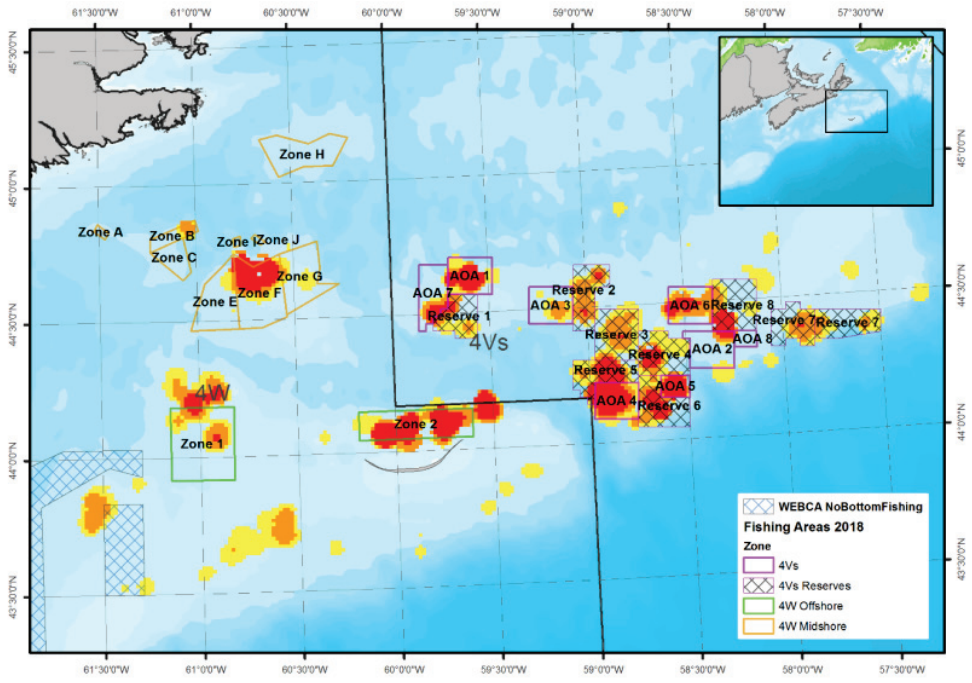


Figure 4.2: A duplicate of Figure 1.2, provided here for convenience. The figure was originally taken from DFO, 2021a.

functions and random effects captured anthropogenic-driven factors (such as shifts in fishing pressure for other species) without requiring them to be explicitly included as covariates. Our models based on random effects additionally indicated that *C. frondosa*'s distribution evolves temporally through an approximation of a random walk process. Accounting for spatio-temporal autocorrelation with smooth functions, while convenient, does not allow for such inferences on the underlying dependence structure.

The species' fishery on the Scotian Shelf is restricted to a number of small fishing zones (Figure 4.2), most of which were sparsely covered by our two surveys. Any fishery-driven trends would therefore be challenging to capture with RV and Snow Crab survey data alone. With that said, it is worth highlighting the situation in 4W Offshore Zone 2, as that zone's data coverage was significant. The *starve* model predicted a noticeable (and recent) drop in CPUE for the area immediately northeast of Sable Island, which 4W Offshore Zone 2 notably rests on. The other models did not detect any noticeable decrease inside this zone. The situation on the eastern Scotian Shelf should continue to be carefully monitored.

4.1 Future Directions

The modelling framework we proposed for Scotian Shelf *C. frondosa* can be readily expanded upon. For example, additional benthic environmental data layers can be considered for inclusion in our models as they become available for the region. In particular, we suggest that a layer denoting sediment grain size be evaluated when one is procurable. High resolution sediment information was unavailable at the time of writing, and we thus had to rely on bottom stress to serve as a proxy (Ward et al., 2015); ideally, this would be unnecessary. Data layers that provide information on relevant anthropogenic-driven factors, including fishing effort, would also be worth including if generated. A broadened suite of covariates could potentially lead to improved predictions and scientific advice.

In a similar vein, expanding our framework to consider drop-camera survey data could improve resolution in the fishing zones and reserves. A more ambitious proposition, but one worth contemplating, would be the incorporation of fishery-dependent data. This would be nontrivial because commercial fishing ventures and scientific surveys clearly differ in their overall goals and the data they collect. The preferential sampling nature of the new data would have to be considered, for example (Pennino et al., 2016). Furthermore, integrating fishery-dependent and fishery-independent data would require several sources of variability (such as the vessel, gear type, etc.) to be more rigorously accounted for. In recent work, some of these challenges have been addressed—with success—to produce species distribution models based on integrated survey and commercial data. In particular, Rufener et al. (2021) proposed a spatio-temporally explicit model for the western Baltic cod stock based on integrated data, and then compared it to a model that relied only on fishery-independent sources. They showed that the integrated model provided more information on spatio-temporal abundance dynamics and had reduced uncertainty in the fixed effects and predictions.

The potential effect of seasonality on the distribution of Scotian Shelf *C. frondosa* warrants further investigation. Our preliminary work suggested that (on average) bottom temperature, stress, and salinity were relatively stable across any given year in core *C. frondosa* habitat; this indicated extensive migrations may be of little concern. Furthermore, our results were insensitive to whether survey tows conducted in

the winter were used during model fitting. Nonetheless, given that size-dependent migration has been observed in other regions (Hamel and Mercier, 1996), the seasonality issue calls for a future study of its own. If evidence for an underlying seasonal trend is found, our current paradigm for studying and protecting this population may need to be reevaluated.

These models were not intended to provide a stock assessment. Regardless, we believe they have the potential to support such an endeavour, if expanded to consider data beyond our two surveys. Currently, assessments for Scotian Shelf *C. frondosa* rely heavily on commercial catch rates (DFO, 2021a). However, because of the species' sedentary nature, sea cucumber catch rates can be sensitive to the spatial relocation of fishing effort (DFO, 2021a). This was demonstrated within 4W Midshore Zone F on the Shelf; after six years of focused effort on a particular area, fishers relocated to a new location, leading to an increase in catch rates that may not have reflected conditions across the entire zone (DFO, 2021a). Yearly indices derived from the output of our models, such as the mean model prediction across a given zone, would not have such concerns. These ideas should be further developed. Indices derived from spatio-temporal models hold great promise in decreasing the uncertainty surrounding the stock assessment of sedentary benthic species such as *C. frondosa*.

Bibliography

Anderson, S. C., Lotze, H. K., & Shackell, N. L. (2008). Evaluating the knowledge base for expanding low-trophic-level fisheries in Atlantic Canada. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(12), 2553–2571. https://doi.org/10.1139/F08-156

Anderson, S. C., Mills Flemming, J., Watson, R., & Lotze, H. K. (2011). Rapid global expansion of invertebrate fisheries: Trends, drivers, and ecosystem effects (S. J. Bograd, Ed.). *PLoS ONE*, *6*(3). https://doi.org/10.1371/journal.pone.0014735

Anderson, S. C., Ward, E. J., English, P. A., & Barnett, L. A. K. (2022). sdmTMB: An R package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields. *bioRxiv*. https://doi.org/10.1101/2022.03.24.485545

Bahlai, C. (2016). Long-term dynamics in invaded lady beetle communities. *2016 International Congress of Entomology*. https://doi.org/10.1603/ICE.2016.93349

Bakka, H., Vanhatalo, J., Illian, J., Simpson, D., & Rue, H. (2016). Accounting for physical barriers in species distribution modeling with non-stationary spatial random effects.

Beazley, L., Guijarro, J., Lirette, C., Wang, Z., & Kenchington, E. (2018). *Characteristics of Environmental Data Layers for Use in Species Distribution Modelling in the Eastern Canadian Arctic and Sub-Arctic Regions* (Can. Tech. Rep. Fish. Aquat. Sci. 3248). Fisheries and Oceans Canada.

Boitani, L., & Powell, R. A. (2014). *Carnivore ecology and conservation*. OUP Oxford.

Brickman, D., Wang, Z., & DeTracey, B. (2016). *High resolution future climate ocean model simulations for the Northwest Atlantic Shelf region* (Can. Tech. Rep. Hydrogr. Ocean. Sci. 315). Fisheries and Oceans Canada.

Carson, S., Shackell, N., & Mills Flemming, J. (2017). Local overfishing may be avoided by examining parameters of a spatio-temporal model (D. E. Duplisea, Ed.). *PLOS ONE*, *12*(9), e0184427. https://doi.org/10.1371/journal.pone.0184427

Clayton, D. G., Bernardinelli, L., & Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, *22*(6), 1193–1202. https://doi.org/10.1093/ije/22.6.1193

Commander, C. J. C., Barnett, L. A. K., Ward, E. J., Anderson, S. C., & Essington, T. E. (2022). The shadow model: How and why small choices in spatially explicit species distribution models affect predictions. *PeerJ*, *10*, e12783. https://doi.org/10.7717/peerj.12783

Crow, E. L., & Shimizu, K. (1988). *Lognormal distributions: Theory and applications*. M. Dekker.

Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, *111*(514), 800–812. https://doi.org/10.1080/01621459.2015.1044091

DFO. (2021a). *Guidance for setting reference points for the sea cucumber (Cucumaria frondosa) fishery in the Maritimes region, and status of the SWNB sea cucumber fishery 2019* (Can. Sci. Advis. Sec. Sci. Advis. Rep. 2021/007).

DFO. (2021b). *Maritimes Research Vessel Survey trends on the Scotian Shelf and Bay of Fundy for 2020* (Can. Sci. Advis. Sec. Sci. Advis. Rep. 2021/022).

Farber, O., & Kadmon, R. (2003). Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, *160*(1-2), 115–130. https://doi.org/10.1016/s0304-3800(02)00327-7

Fletcher, D., MacKenzie, D., & Villouta, E. (2005). Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics*, *12*(1), 45–54. https://doi.org/10.1007/s10651-005-6817-1

Fowler, G. M., & Showell, M. A. (2009). *Calibration of bottom trawl survey vessels: comparative fishing between the Alfred Needler and Teleost on the Scotian Shelf during the summer of 2005* (Can. Tech. Rep. Fish. Aquat. Sci. 2824). Fisheries and Oceans Canada.

Franklin, J., & Miller, J. A. (2010). *Mapping species distributions: Spatial inference and prediction (ecology, biodiversity and conservation)*. Cambridge University Press.

Gelfand, A. E., Holder, M., Latimer, A., Lewis, P. O., Rebelo, A. G., Silander, J. A., & Wu, S. (2006). Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis*, *1*(1). https://doi.org/10.1214/06-BA102

Hamel, J.-F., & Mercier, A. (1996). Early development, settlement, growth and spatial distribution of the sea cucumber *Cucumaria frondosa* (echinodermata: Holothuroidea). *Canadian Journal of Fisheries and Aquatic Sciences*, *53*(2), 253–271. https://doi.org/10.1139/f95-186

Hamel, J.-F., & Mercier, A. (2008). Population status, fisheries and trade of sea cucumbers in temperate areas of the Northern Hemisphere. In V. Toral-Granda, A. Lovatelli and M. Vasconcellos (Eds)., *Sea cucumbers. A global review of fisheries and trade*. (FAO Fisheries and Aquaculture Technical Paper. No. 516, pp. 257-291). FAO.

Hanks, E. M., Schliep, E. M., Hooten, M. B., & Hoeting, J. A. (2015). Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, *26*(4), 243–254. https://doi.org/https://doi.org/10.1002/env.2331

Harper, D. L. (2020). *Densities, distributions, and individual size of orange-footed sea cucumber (Cucumaria frondosa) on the Scotian Shelf* (Master's thesis). Acadia University.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3). https://doi.org/10.1214/ss/1177013604

Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, *83*(7), 2027–2036. https://doi.org/10.1890/0012-9658(2002)083[2027:enfaht]2.0.co;2

Hubley, P. B., Reeves, A., Smith, S. J., & Nasmith, L. (2014). *Georges Bank 'a' and Browns Bank 'North' Scallop (Placopecten magellanicus) Stock Assessment* (Can. Sci. Advis. Sec. Res. Doc. 2013/079). Fisheries and Oceans Canada.

Humble, S. R., Hand, C. M., & de la Mare, W. K. (2007). *Review of data collected during the annual sea cucumber (Parastichopus californicus) fishery in British Columbia and recommendations for a rotational harvest strategy based on simulation modelling* (Can. Sci. Advis. Sec. Sci. Res. Doc. 2007/054). Fisheries and Oceans Canada.

Jackson-Bué, T., Williams, G. J., Whitton, T. A., Roberts, M. J., Goward Brown, A., Amir, H., King, J., Powell, B., Rowlands, S. J., Llewelyn Jones, G., & Davies, A. J. (2022). Seabed morphology and bed shear stress predict temperate reef habitats in a high energy marine region. *Estuarine, Coastal and Shelf Science*, *274*, 107934. https://doi.org/https://doi.org/10.1016/j.ecss.2022.107934

Kallasvuo, M., Vanhatalo, J., & Veneranta, L. (2017). Modeling the spatial distribution of larval fish abundance provides essential information for management. *Canadian Journal of Fisheries and Aquatic Sciences*, *74*(5), 636–649. https://doi.org/10.1139/cjfas-2016-0008

Kenchington, E. L., Cogswell, A., Lirette, C., & Murillo, F. (2009). The use of density analyses to delineate sponge grounds and other benthic vmes from trawl survey data.

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

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*(5). https://doi.org/10.18637/jss.v070.i05

Lawler, E., Field, C., & Mills Flemming, J. (2023). Starve: An r package for spatio-temporal analysis of research survey data using nearest-neighbour gaussian processes. *Methods in Ecology and Evolution*, 2041–210X.14053. https://doi.org/10.1111/2041-210X.14053

Lecours, V., Brown, C. J., Devillers, R., Lucieer, V. L., & Edinger, E. N. (2016). Comparing selections of environmental variables for ecological studies: A focus on terrain attributes. *PLOS ONE*, *11*(12). https://doi.org/10.1371/journal.pone.0167128

Lecours, V., Devillers, R., Simms, A. E., Lucieer, V. L., & Brown, C. J. (2017). Towards a framework for terrain attribute selection in environmental studies. *Environmental Modelling & Software*, *89*, 19–30. https://doi.org/10.1016/j.envsoft.2016.11.027

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

McDonald, R. R., Keith, D., Sameoto, J., Hutchings, J., & Flemming, J. (2021). Explicit incorporation of spatial variability in a biomass dynamics assessment model. *ICES Journal of Marine Science*, *78*. https://doi.org/10.1093/icesjms/fsab192

McDonald, R. R., Keith, D., Sameoto, J., Hutchings, J., & Flemming, J. (2022). Incorporating intra-annual variability in fisheries abundance data to better capture population dynamics. *Fisheries Research*, *246*. https://doi.org/10.1016/j.fishres.2021.106152

Miller, D. L., & Wood, S. N. (2014). Finite area smoothing with generalized distance splines. *Environmental and Ecological Statistics*, *21*(4), 715–731. https://doi.org/10.1007/s10651-014-0277-4

Moraga, P. (2020). *Geospatial health data: Modeling and visualization with R-INLA and Shiny*. CRC Press/Taylor Francis Group.

Orensanz, J. M., Parma, A. M., & Hall, M. A. (1998). The analysis of concentration and crowding in shellfish research. In G.S. Jamieson & A. Campbell (Eds.), *Proceedings of the North Pacific Symposium on Invertebrate Stock Assessment and Management* (Can. Spec. Publ. Fish. Aquat. Sci. No. 125, pp. 140–157). NRC Research Press.

Page, G. L., Liu, Y., He, Z., & Sun, D. (2017). Estimation and prediction in the presence of spatial confounding for spatial linear models. *Scandinavian Journal of Statistics*, *44*(3), 780–797. https://doi.org/https://doi.org/10.1111/sjos.12275

Pantin, J. R., Coughlan, E. J., Hynick, E. M., & Skanes, K. R. (2018). *An assessment of the sea cucumber (Cucumaria frondosa) resource on the St. Pierre Bank (NAFO Subdivision 3Ps) in 2016* (Can. Sci. Advis. Sec. Res. Doc. 2018/024). Fisheries and Oceans Canada.

Pennino, M. G., Conesa, D., López-Quílez, A., Muñoz, F., Fernández, A., & Bellido, J. M. (2016). Fishery-dependent and -independent data lead to consistent estimations of essential habitats. *ICES Journal of Marine Science*, *73*(9), 2302–2310. https://doi.org/10.1093/icesjms/fsw062

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3-4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Purcell, S. W., Lovatelli, A., Vasconcellos, M., & Yimin, Y. (2010). *Managing sea cucumber fisheries with an ecosystem approach* (FAO Fisheries and Aquaculture Technical Paper No. 520). FAO.

R Core Team. (2021). *R: A language and environment for statistical computing* [computer software].

Record, S., Fitzpatrick, M. C., Finley, A. O., Veloz, S., & Ellison, A. M. (2013). Should species distribution models account for spatial autocorrelation? A test of model projections across eight millennia of climate change: Projecting spatial species distribution models. *Global Ecology and Biogeography*, *22*(6), 760–771. https://doi.org/10.1111/geb.12017

Reich, B. J., Hodges, J. S., & Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, *62*(4), 1197–1206. https://doi.org/10.1111/j.1541-0420.2006.00617.x

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B*, *71*, 319–392.

Rufener, M.-C., Kristensen, K., Nielsen, J. R., & Bastardie, F. (2021). Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species. *Ecological Applications*, *31*(8), e02453. https://doi.org/https://doi.org/10.1002/eap.2453

Shackell, N. L., Brickman, D. W., & Frank, K. T. (2013). Reserve site selection for data-poor invertebrate fisheries using patch scale and dispersal dynamics: A case study of sea cucumber (Cucumaria frondosa): Fishery reserves for data-poor species. *Aquatic Conservation: Marine and Freshwater Ecosystems*. https://doi.org/10.1002/aqc.2369

Shono, H. (2008). Confidence interval estimation of cpue year trend in delta-type two-step model. *Fisheries Science*, *74*(4), 712–717. https://doi.org/10.1111/j.1444-2906.2008.01581.x

Stefansson, G. (1996). Analysis of groundfish survey abundance data: Combining the GLM and delta approaches. *ICES Journal of Marine Science*, *53*(3), 577–588. https://doi.org/10.1006/jmsc.1996.0079

Stock, B. C., Ward, E. J., Eguchi, T., Jannot, J. E., Thorson, J. T., Feist, B. E., & Semmens, B. X. (2020). Comparing predictions of fisheries bycatch using multiple spatiotemporal species distribution model frameworks. *Canadian Journal of Fisheries and Aquatic Sciences*, *77*(1), 146–163. https://doi.org/10.1139/cjfas-2018-0281

Thorson, J. T., Shelton, A. O., Ward, E. J., & Skaug, H. J. (2015). Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for west coast groundfishes. *ICES Journal of Marine Science*, *72*(5), 1297–1310. https://doi.org/10.1093/icesjms/fsu243

Tremblay, M. J., Black, G. A. P., & Branton, R. M. (2007). *The distribution of common decapod crustaceans and other invertebrates recorded in annual ecosystem surveys of the Scotian Shelf 1999-2006* (Can. Tech. Rep. Fish. Aquat. Sci. 2762). Fisheries and Oceans Canada.

Van Lieshout, M. N. M. (2019). *Theory of spatial statistics: A concise introduction*. CRC Press.

Vanhatalo, J., Veneranta, L., & Hudd, R. (2012). Species distribution modeling with Gaussian processes: A case study with the youngest stages of sea spawning whitefish (coregonus lavaretus l. s.l.) larvae. *Ecological Modelling*, *228*, 49–58. https://doi.org/https://doi.org/10.1016/j.ecolmodel.2011.12.025

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, *50*(2), 297–312. https://doi.org/10.1111/j.2517-6161.1988.tb01729.x

Wang, Z., Lu, Y., Greenan, B., Brickman, D., & DeTracey, B. (2018). *BNAM: An eddy-resolving North Atlantic Ocean model to support ocean monitoring* (Can. Tech. Rep. Fish. Aquat. Sci. 327). Fisheries and Oceans Canada.

Ward, S. L., Neill, S. P., Van Landeghem, K. J., & Scourse, J. D. (2015). Classifying seabed sediment type using simulated tidal-induced bed shear stress. *Marine Geology*, *367*, 94–104. https://doi.org/https://doi.org/10.1016/j.margeo.2015.05.010

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Chapman Hall/CRC.

Zhang, G., Zhu, A.-X., Windels, S. K., & Qin, C.-Z. (2018). Modelling species habitat suitability from presence-only data using kernel density estimation. *Ecological Indicators*, *93*, 387–396. https://doi.org/https://doi.org/10.1016/j.ecolind.2018.04.002

Zimmerman, D. L., & Stein, M. (2010). Classical geostatistical methods. In A. E. Gelfand, P. Diggle, M. Fuentes, & P. Guttorp (Eds.), *Handbook of spatial statistics*. Taylor Francis.

Zisserson, B. M., Cameron, B. J., Glass, A. C., & Choi, J. S. (2018). *Assessment of Scotian Shelf snow crab in 2018* (Can. Sci. Advis. Sec. Res. Doc. 2021/048). Fisheries and Oceans Canada.

Zisserson, B. (2015). *Maritimes Region Snow Crab Trawl Survey: Detailed technical description* (Can. Tech. Rep. Fish. Aquat. Sci. 3128). Fisheries and Oceans Canada.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems: Data exploration. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

Appendix A

Model Summary Tables

Table A.1: GAM presence sub-model

	Parametric Term	Estimate	Std. Error	Z	P-value
*	Intercept	-1.506	0.079	-18.950	< 0.001
*	Snow Crab Survey	-0.162	0.077	-2.094	0.036

	Smooth Function	Effective Df	Reference Df	χ^2	P-value
*	f(Eastings, Northings, Year)	442.409	567.567	1960.152	< 0.001
*	f(Log(Depth))	1.758	1.900	25.333	< 0.001
*	f(Bottom Temperature)	1.895	1.982	7.350	0.023
*	f(Btm. Temperature Range)	1.896	1.967	8.323	0.020
*	f(Bottom Salinity)	1.820	1.941	8.186	0.023
*	f(Log(Bottom Stress))	1.000	1.001	9.851	0.002
*	f(Log(Btm. Stress Range))	1.600	1.789	5.359	0.035
	f($\sqrt{\text{Slope}}$)	1.651	1.860	4.150	0.067
	f(Easterness, Northerness)	4.659	5.834	11.772	0.062
	f(RDMV)	1.001	1.002	0.769	0.381

*Term significant at the 5% significance level.

Table A.2: GAM CPUE sub-model

	Parametric Term	**Estimate**	**Std. Error**	**Z**	**P-value**
*	Intercept	3.779	0.049	76.686	< 0.001
*	Snow Crab Survey	2.022	0.074	27.405	< 0.001

	Smooth Function	**Effective Df**	**Reference Df**	***F***	**P-value**
*	f(Eastings, Northings, Year)	366.132	469.93	6.545	< 0.001
*	f(Log(Depth))	1.949	1.980	22.385	< 0.001
	f(Bottom Temperature)	1.000	1.001	1.596	0.207
*	f(Btm. Temperature Range)	1.009	1.014	3.881	0.049
	f(Bottom Salinity)	1.605	1.800	0.866	0.489
	f(Log(Bottom Stress))	1.001	1.002	1.312	0.252
	f(Log(Btm. Stress Range))	1.637	1.799	0.675	0.429
	f($\sqrt{\text{Slope}}$)	1.525	1.736	3.047	0.098
	f(Easterness, Northerness)	2.009	2.016	0.287	0.751
	f(RDMV)	1.001	1.002	0.003	0.966

*Term significant at the 5% significance level.

Table A.3: *starve* presence sub-model

	Fixed Effect	**Estimate**	**Std. Error**
*	Log(Depth)	-0.783	0.101
*	Log(Depth)^2	-0.108	0.048
	Bottom Temperature	-0.178	0.096
*	Bottom Temperature^2	0.099	0.034
	Btm. Temperature Range	0.141	0.144
	Btm. Temperature Range^2	0.011	0.056
*	Bottom Salinity	-0.399	0.143
*	Bottom Salinity^2	-0.094	0.038
*	Log(Bottom Stress)	-0.167	0.051
	Log(Btm. Stress Range)	0.111	0.077
	$\sqrt{\text{Slope}}$	-0.074	0.061
	Northerness	0.073	0.043
	Easterness	0.073	0.043
	RDMV	-0.044	0.037
	Snow Crab Survey	-0.089	0.105

Spatio-temporal Parameter	**Estimate**	**Std. Error**
Spatial Standard Deviation (τ)	0.152	0.010
Spatial Range (ρ)	32.642	3.763
Global Mean (μ)	-2.116	2.122
AR(1) Correlation (ϕ)	0.997	0.004
Temporal Standard Deviation (σ)	0.161	0.087

*Term significant at the 5% significance level.

Table A.4: *starve* CPUE sub-model

	Fixed Effect	**Estimate**	**Std. Error**
*	Log(Depth)	-0.613	0.082
	Log(Depth)^2	-0.027	0.037
*	Bottom Temperature	-0.140	0.049
*	Btm. Temperature Range	0.433	0.067
*	Bottom Salinity	0.262	0.081
	Log(Bottom Stress)	-0.008	0.035
*	Log(Btm. Stress Range)	-0.154	0.057
	$\sqrt{\text{Slope}}$	-0.091	0.054
	Northerness	0.016	0.036
	Easterness	-0.023	0.032
	RDMV	0.001	0.024
*	Snow Crab Survey	1.991	0.089

Spatio-temporal Parameter	**Estimate**	**Std. Error**
Spatial Standard Deviation (τ)	0.207	0.015
Spatial Range (ρ)	9.705	0.971
Global Mean (μ)	2.862	0.361
AR(1) Correlation (ϕ)	0.941	0.009
Temporal Standard Deviation (σ)	0.146	0.057

*Term significant at the 5% significance level.

Table A.5: GMRF presence sub-model

	Fixed Effect	**Estimate**	**Std. Error**
*	Log(Depth)	-0.389	0.094
*	Log(Depth)^2	-0.136	0.048
	Bottom Temperature	-0.100	0.081
*	Bottom Temperature^2	0.086	0.030
*	Btm. Temperature Range	0.348	0.156
*	Btm. Temperature Range^2	-0.161	0.057
*	Bottom Salinity	-0.361	0.147
	Bottom Salinity^2	-0.076	0.039
*	Log(Bottom Stress)	-0.147	0.044
*	Log(Btm. Stress Range)	0.226	0.074
*	$\sqrt{\text{Slope}}$	-0.105	0.051
*	Northerness	0.085	0.042
*	Easterness	0.076	0.037
	RDMV	-0.043	0.033
*	Snow Crab Survey	-0.177	0.085
*	Intercept	-2.578	0.451

Spatio-temporal Parameter	**Estimate**	**Std. Error**
Spatial Range (ρ)	56.118	6.810
Marginal Spatial Standard Deviation (σ_{MV})	2.787	0.221
AR(1) Correlation (ϕ)	0.990	0.003

*Term significant at the 5% significance level.

Table A.6: GMRF CPUE sub-model

	Fixed Effect	**Estimate**	**Std. Error**
*	Log(Depth)	-0.408	0.084
*	Log(Depth)^2	-0.171	0.040
	Bottom Temperature	-0.094	0.048
*	Btm. Temperature Range	0.359	0.090
*	Bottom Salinity	0.183	0.093
	Log(Bottom Stress)	-0.036	0.036
	Log(Btm. Stress Range)	0.060	0.060
	$\sqrt{\text{Slope}}$	-0.056	0.049
	Northerness	0.039	0.034
	Easterness	-0.045	0.031
	RDMV	-0.011	0.025
*	Snow Crab Survey	2.032	0.071
*	Intercept	2.798	0.173

Spatio-temporal Parameter	**Estimate**	**Std. Error**
Spatial Range (ρ)	28.756	3.667
Marginal Spatial Standard Deviation (σ_{MV})	1.555	0.077
AR(1) Correlation (ϕ)	0.985	0.004

*Term significant at the 5% significance level.

Appendix B

Supplementary Figures

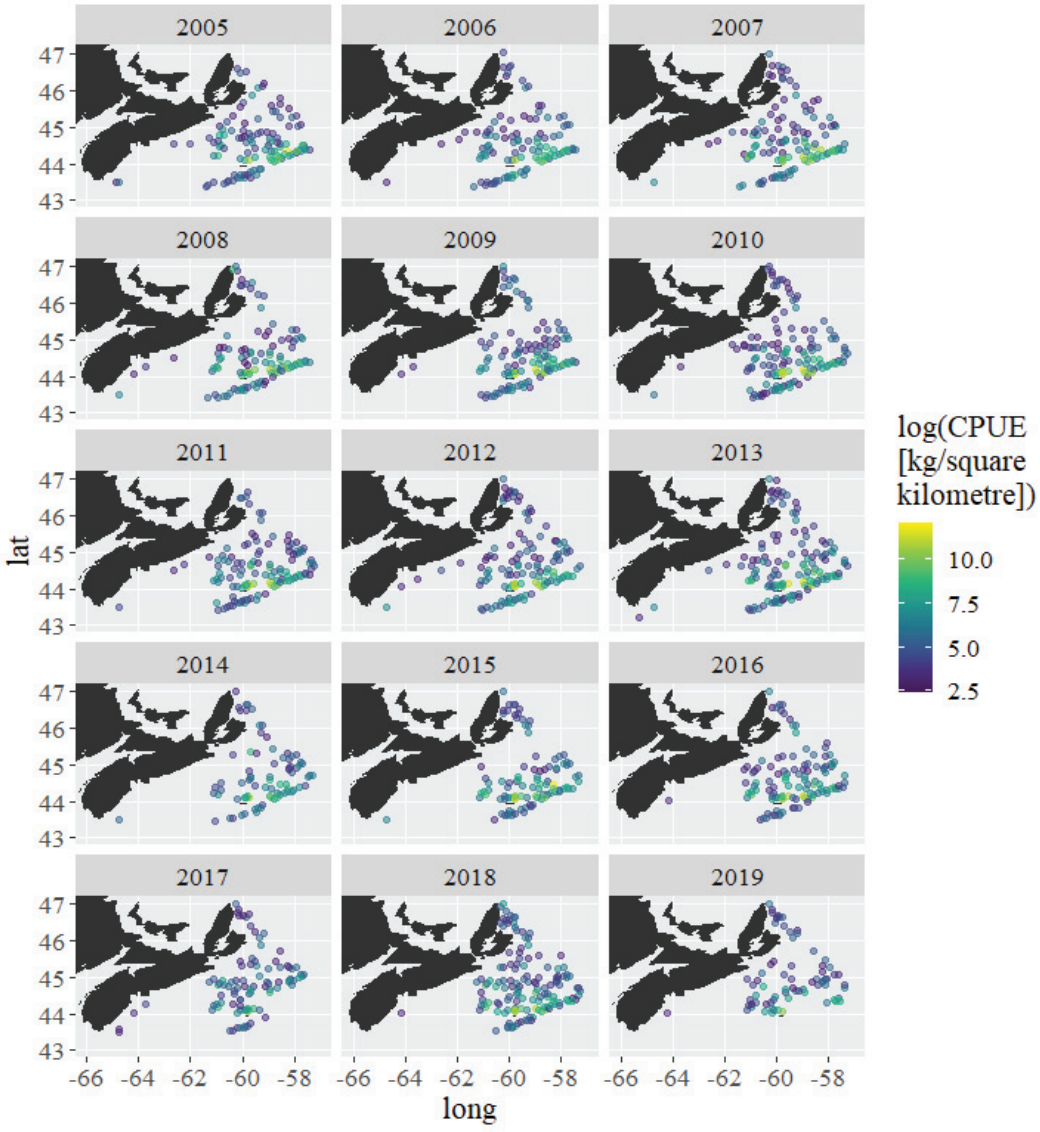


Figure B.1: An illustration of where non-zero weights of sea cucumber were caught during the available Snow Crab Survey tows. The first five years of CPUE data came exclusively from the RV Survey because only the Snow Crab Survey weights from 2005 onwards were considered reliable.

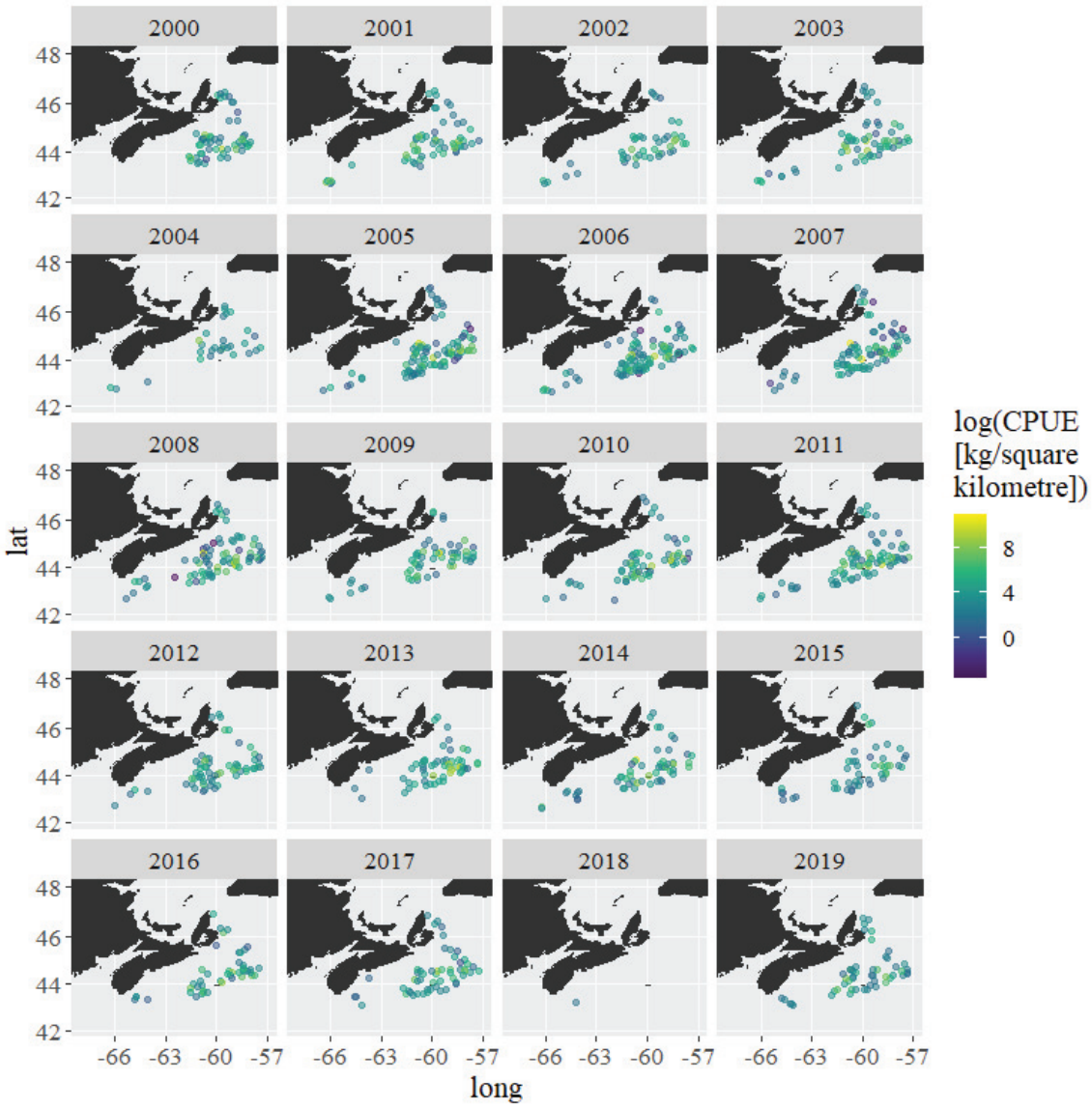
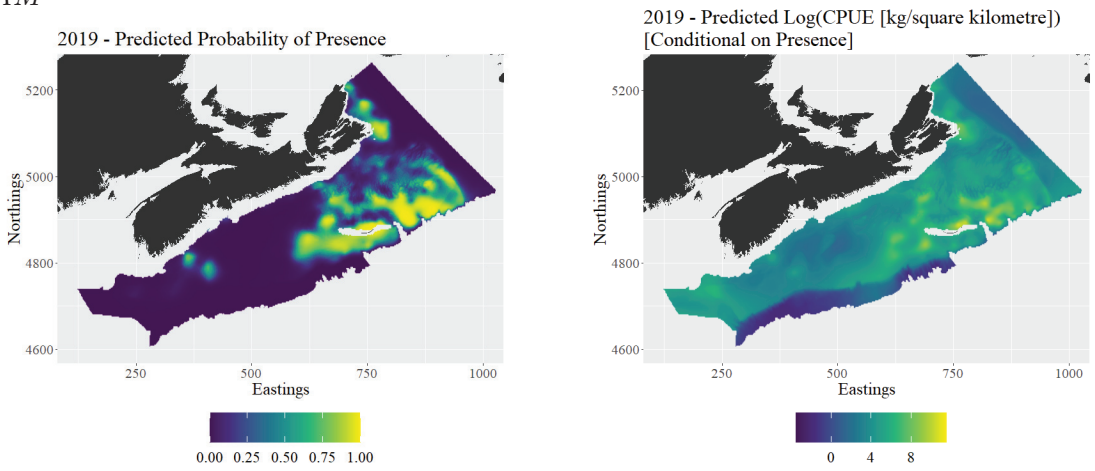
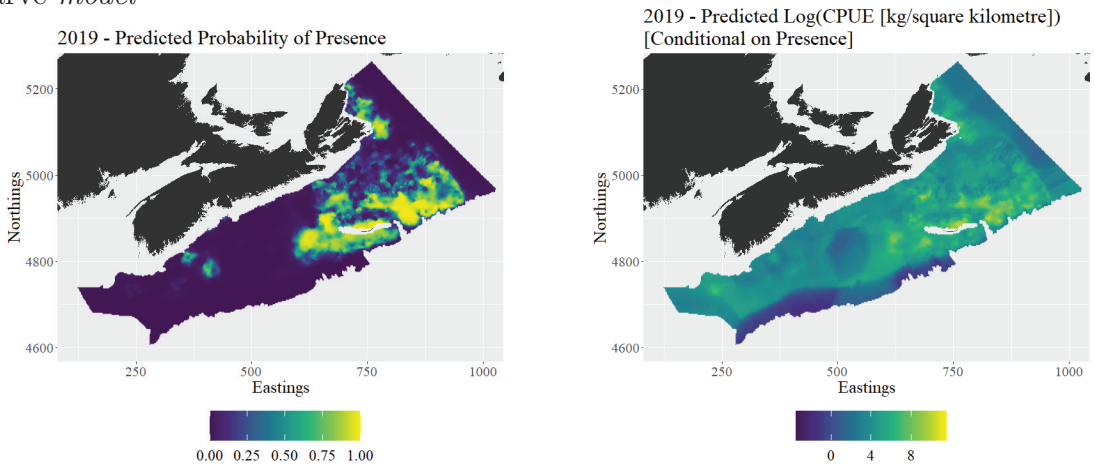


Figure B.2: An illustration of where non-zero weights of sea cucumber were caught during the available RV Survey tows. The 2018 survey was not completed due to mechanical issues with the survey vessel (DFO, 2021b).

GAM



starve *model*



GMRF model

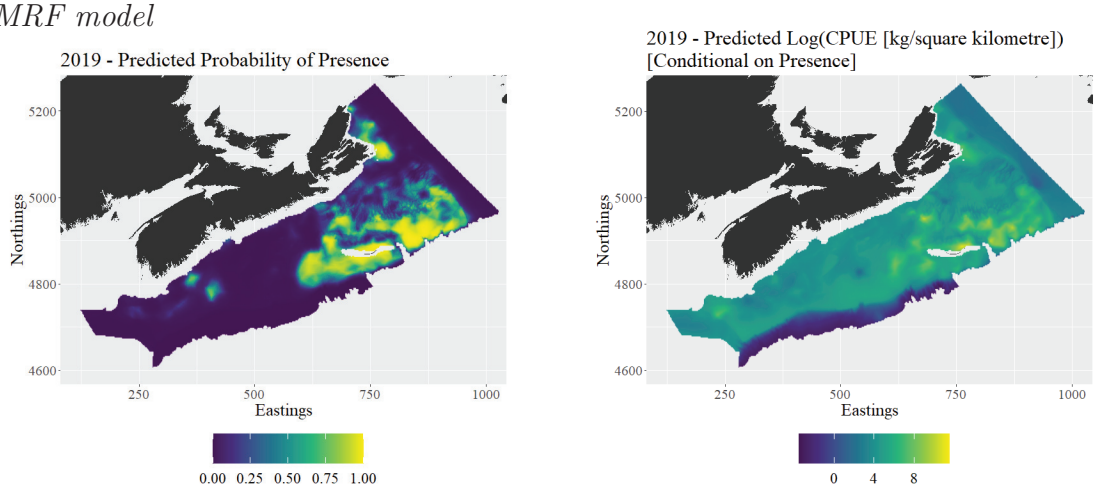


Figure B.3: The predictions from the six sub-models for 2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicators set to one).

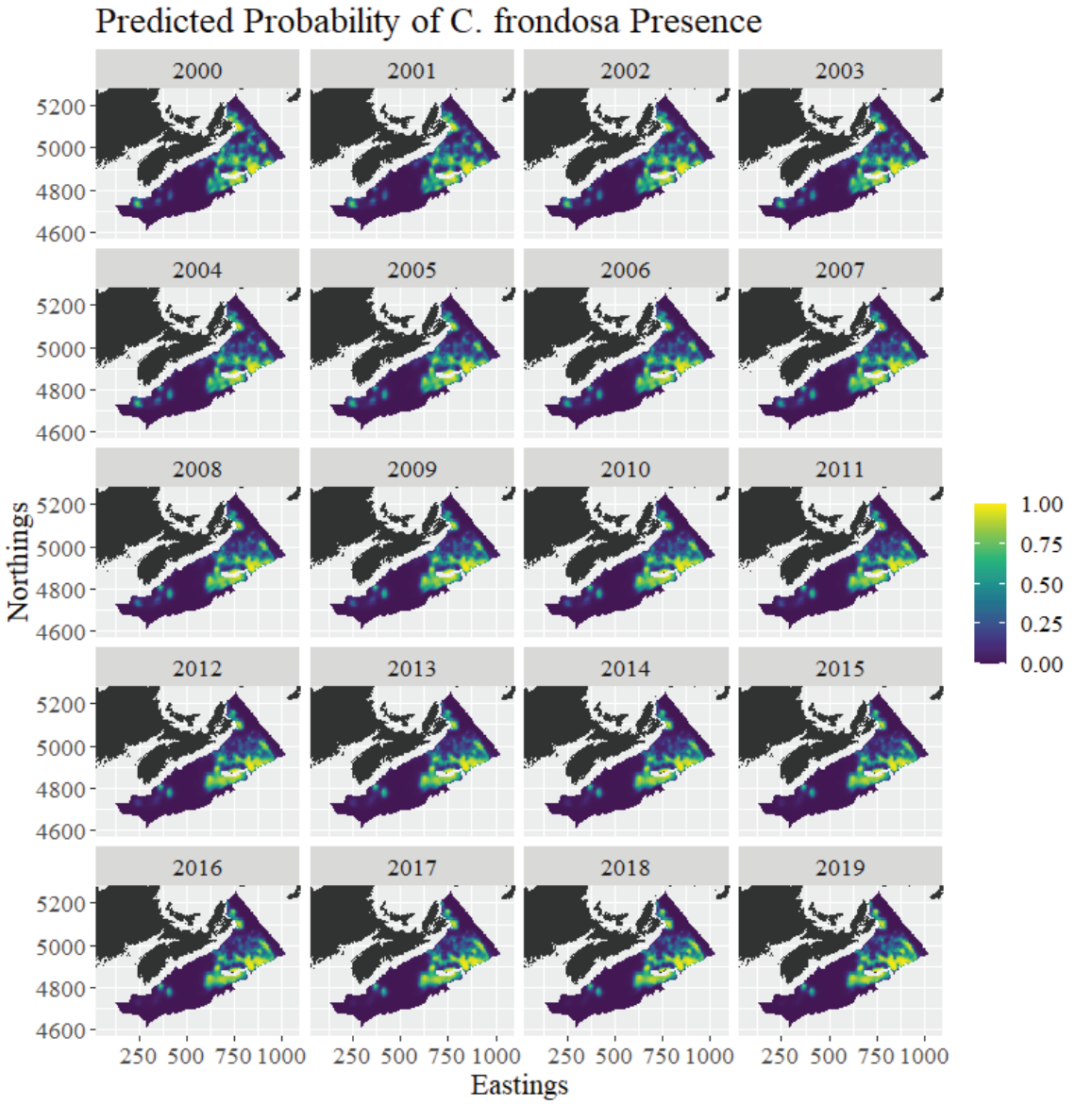


Figure B.4: Predictions from the GAM presence sub-model for the years 2000-2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicator set to one).

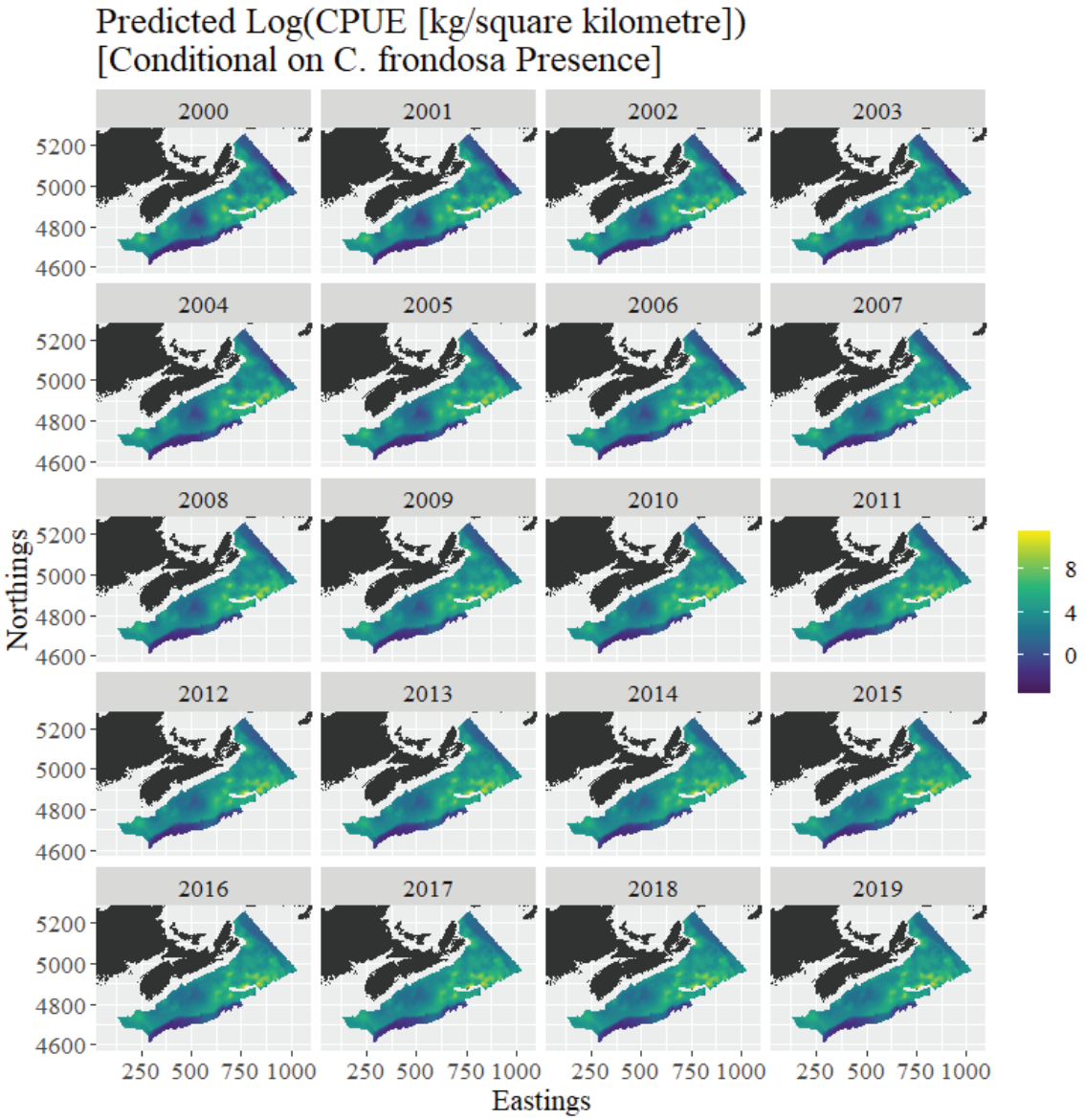


Figure B.5: Predictions from the GAM CPUE sub-model for the years 2000-2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicator set to one).

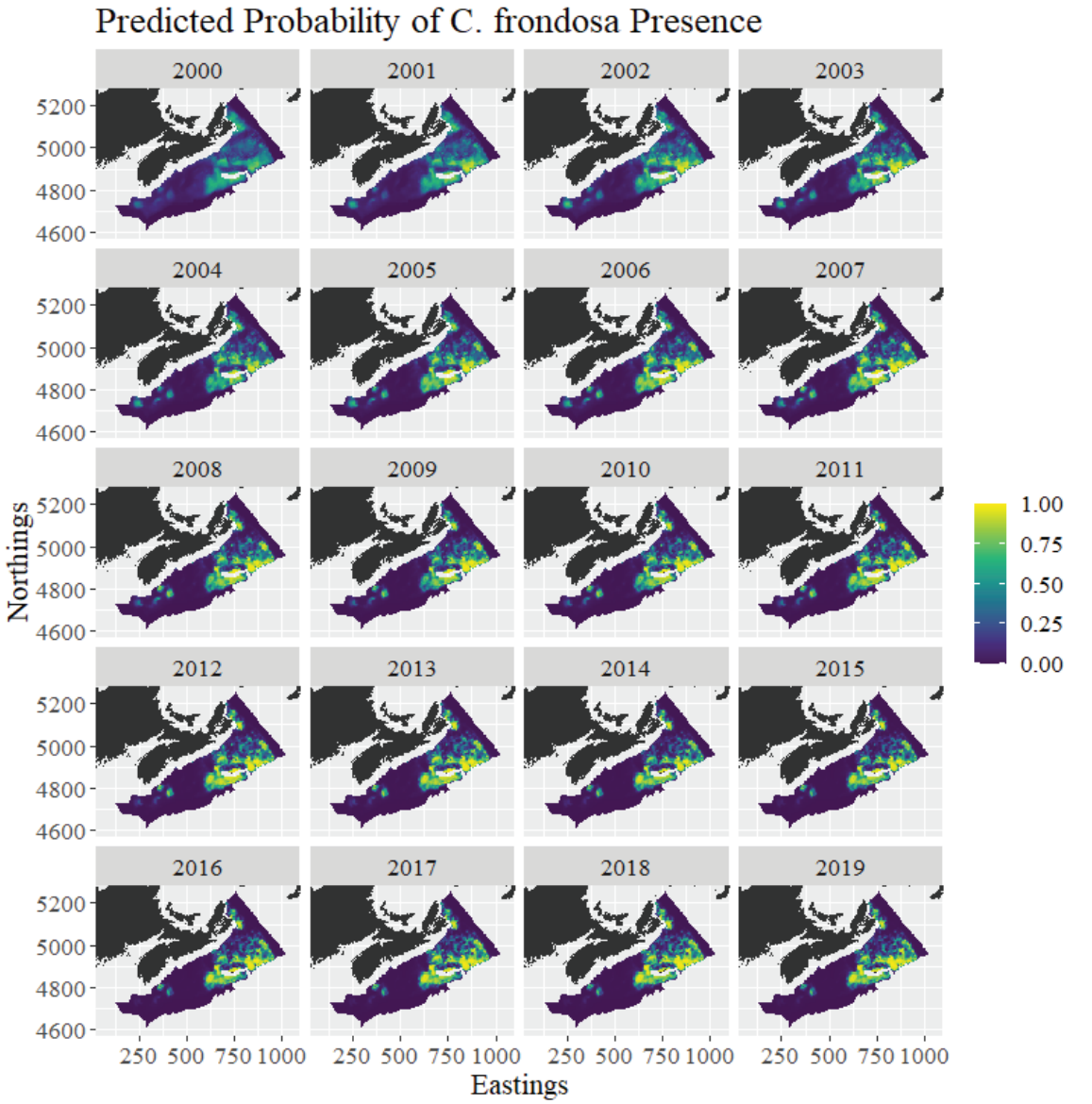


Figure B.6: Predictions from the *starve* presence sub-model for the years 2000-2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicator set to one).

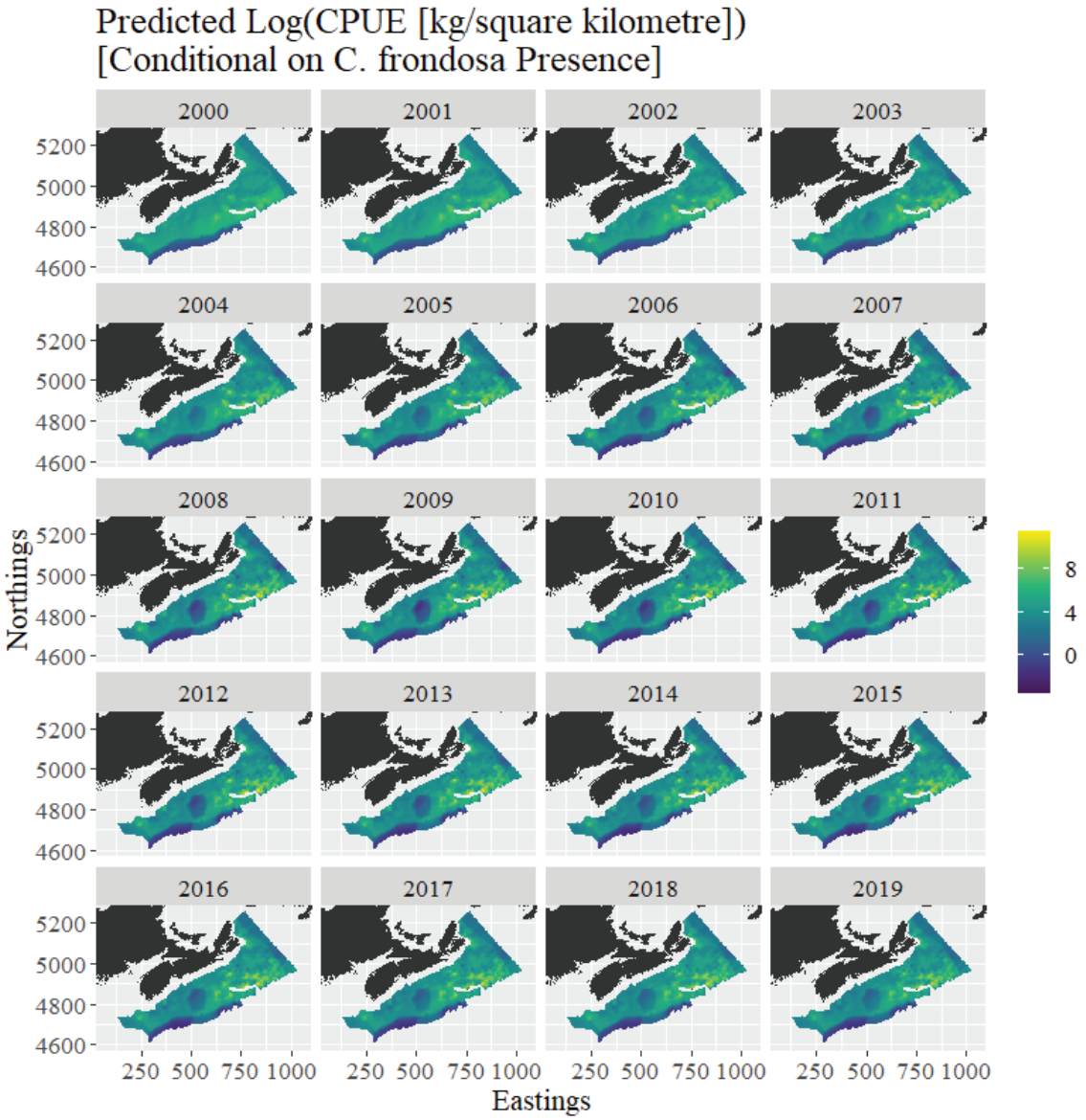


Figure B.7: Predictions from the *starve* CPUE sub-model for the years 2000-2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicator set to one).

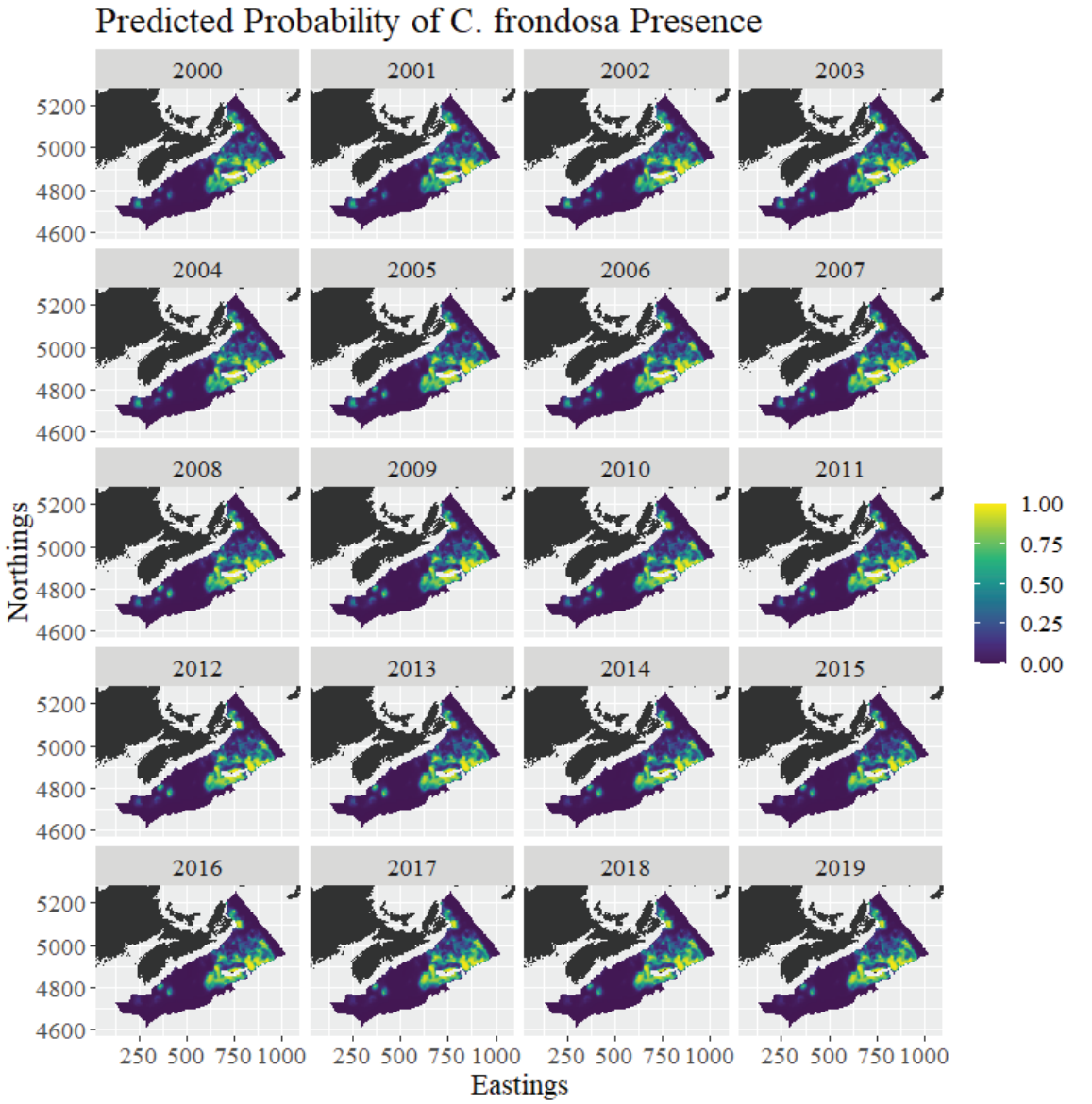


Figure B.8: Predictions from the GMRF presence sub-model for the years 2000-2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicator set to one).

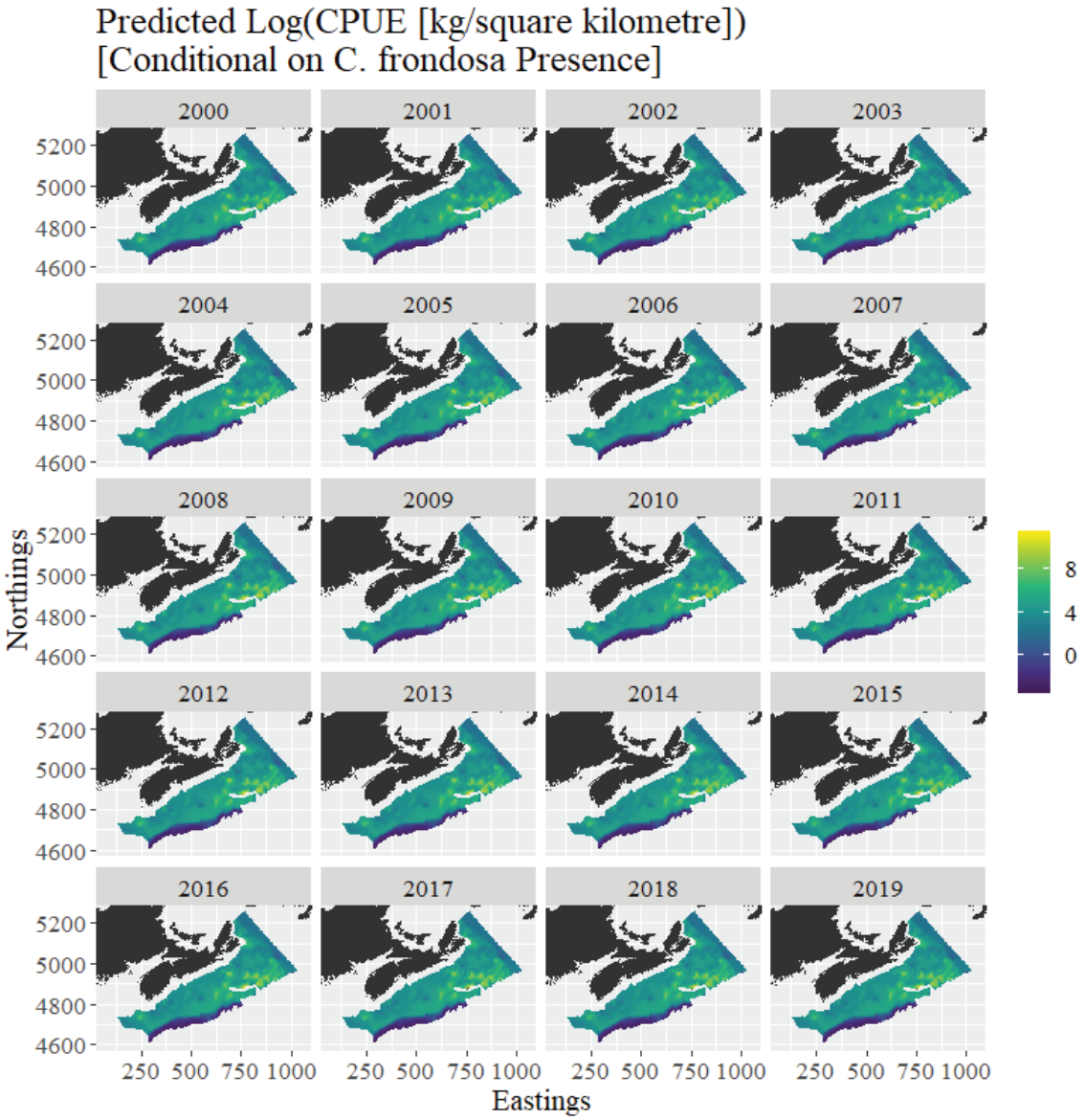


Figure B.9: Predictions from the GMRF CPUE sub-model for the years 2000-2019, obtained using the available BNAM and DEM raster data (and with the Snow Crab Survey indicator set to one).

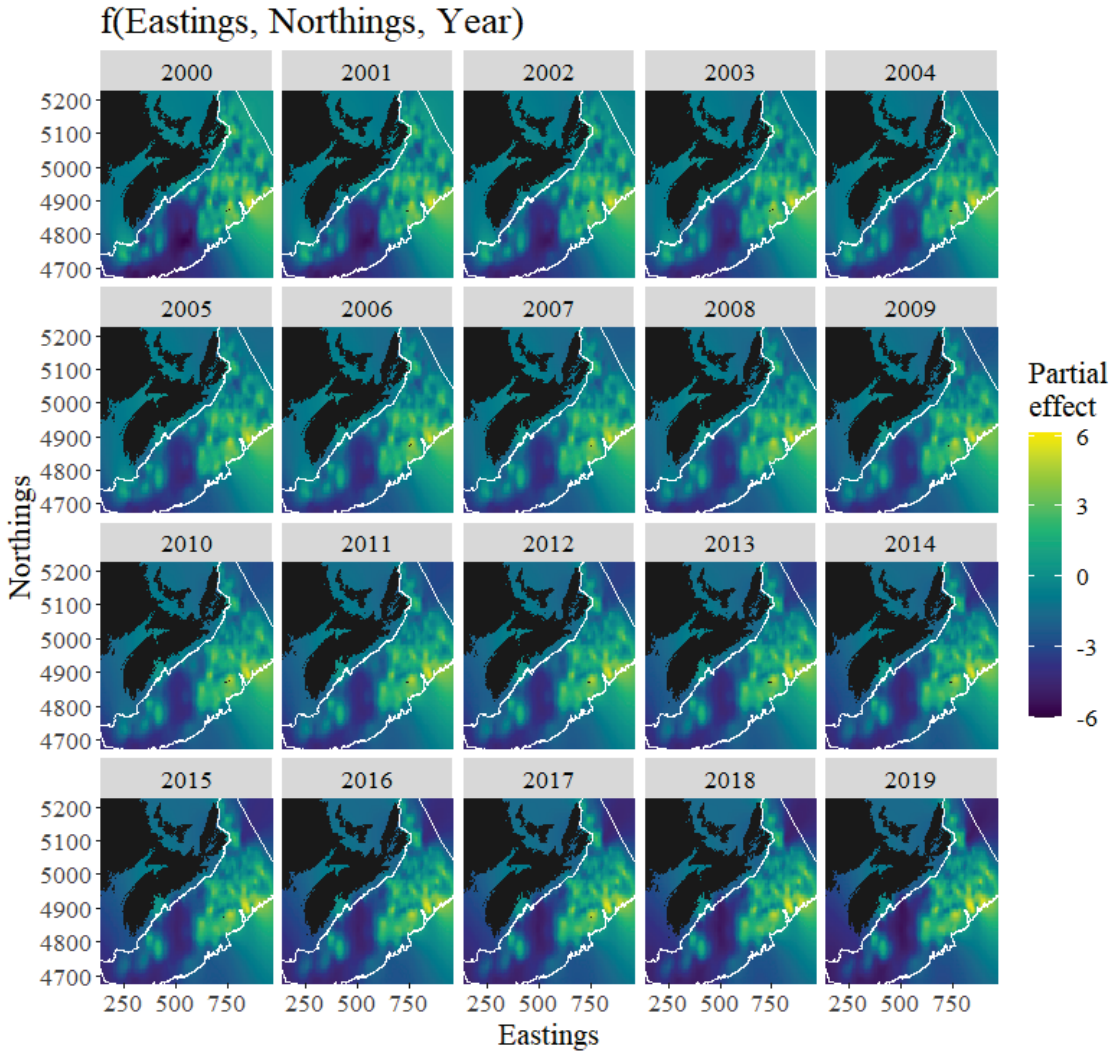


Figure B.10: A partial effects plot illustrating the GAM presence sub-model's spatio-temporal smooth function. Interpretations should not be made outside of our study area (outlined in white).

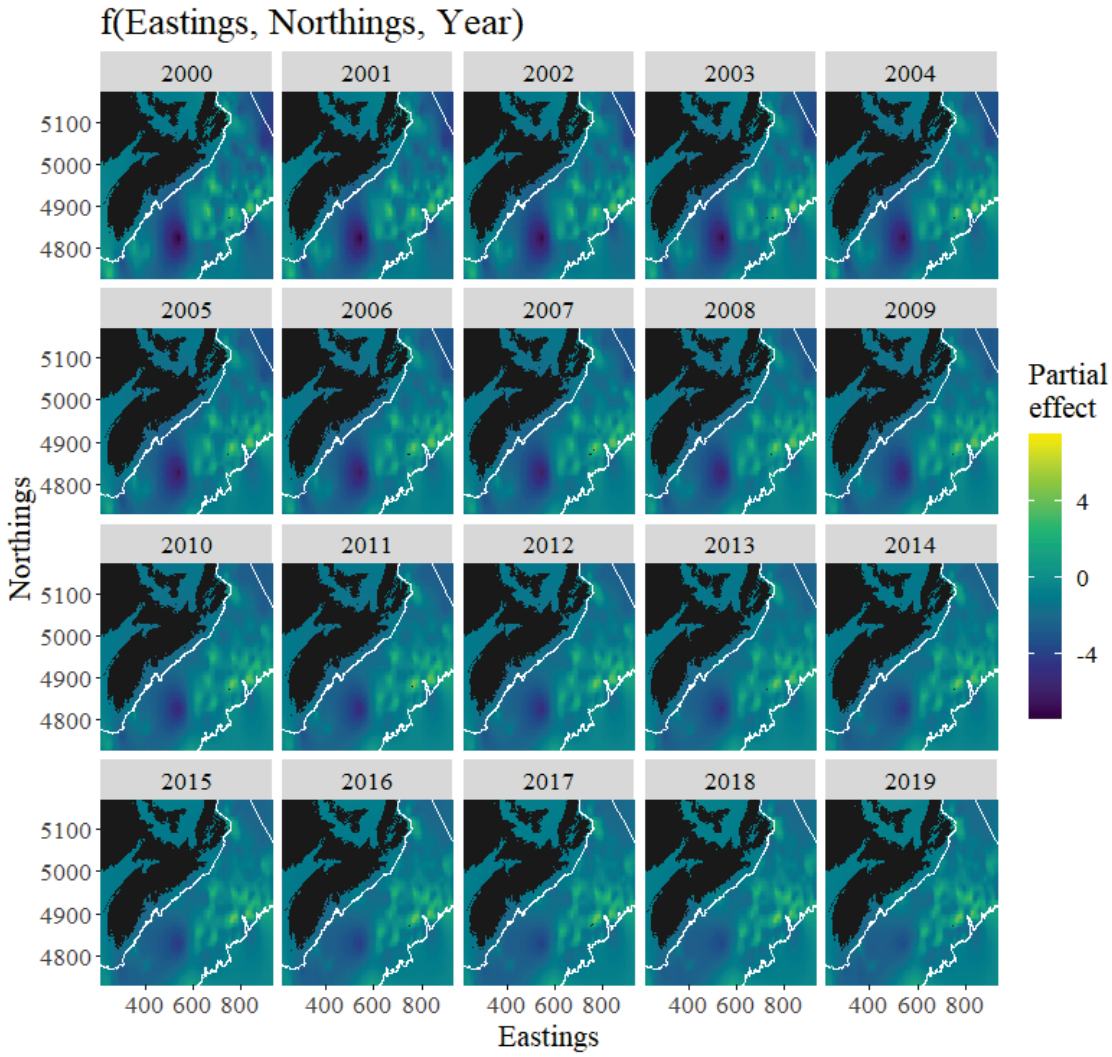


Figure B.11: A partial effects plot illustrating the GAM CPUE sub-model's spatio-temporal smooth function. Interpretations should not be made outside of our study area (outlined in white).

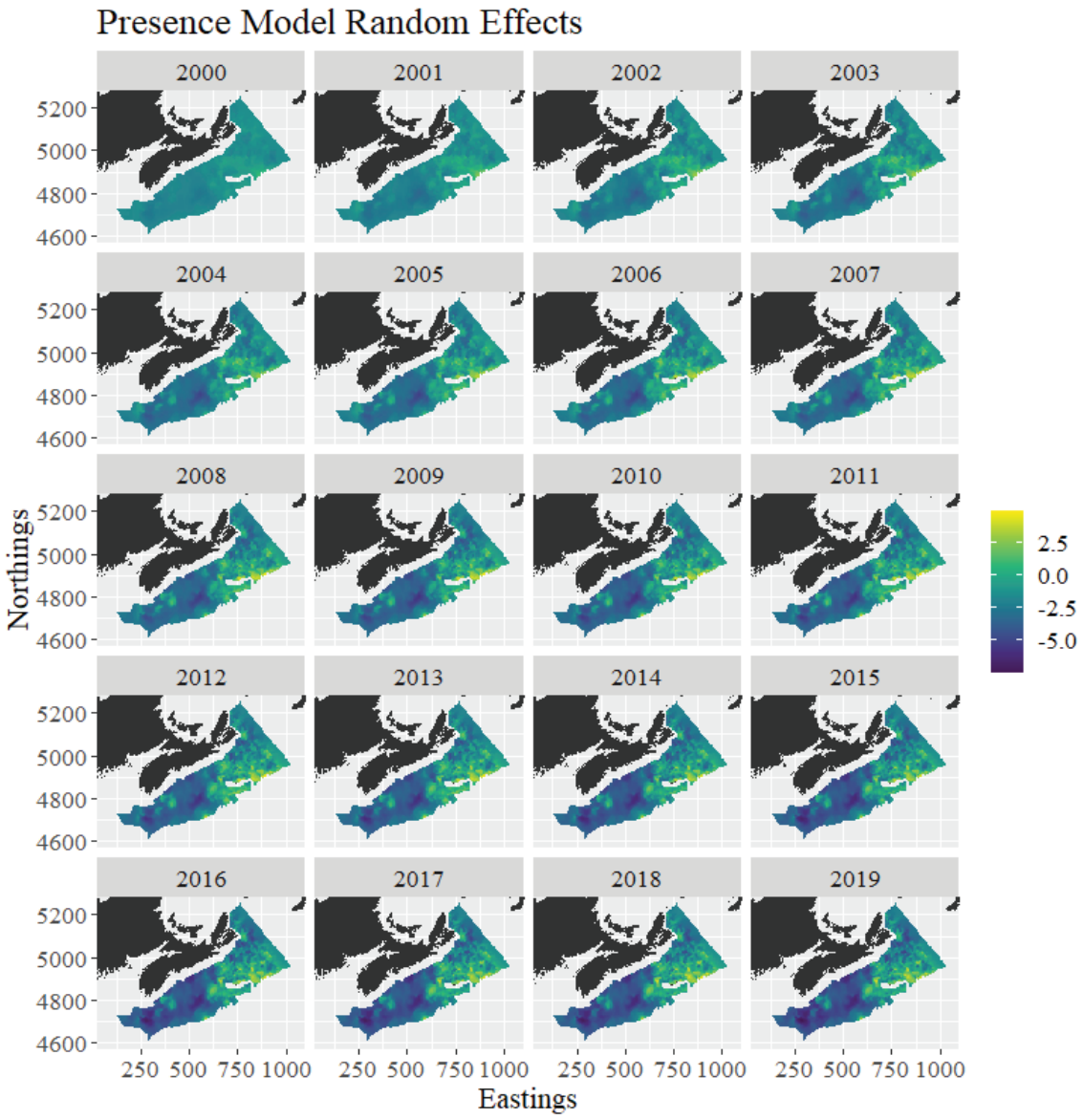


Figure B.12: The random effect contributions to the predictions shown in B.6. The random effects are in link space.

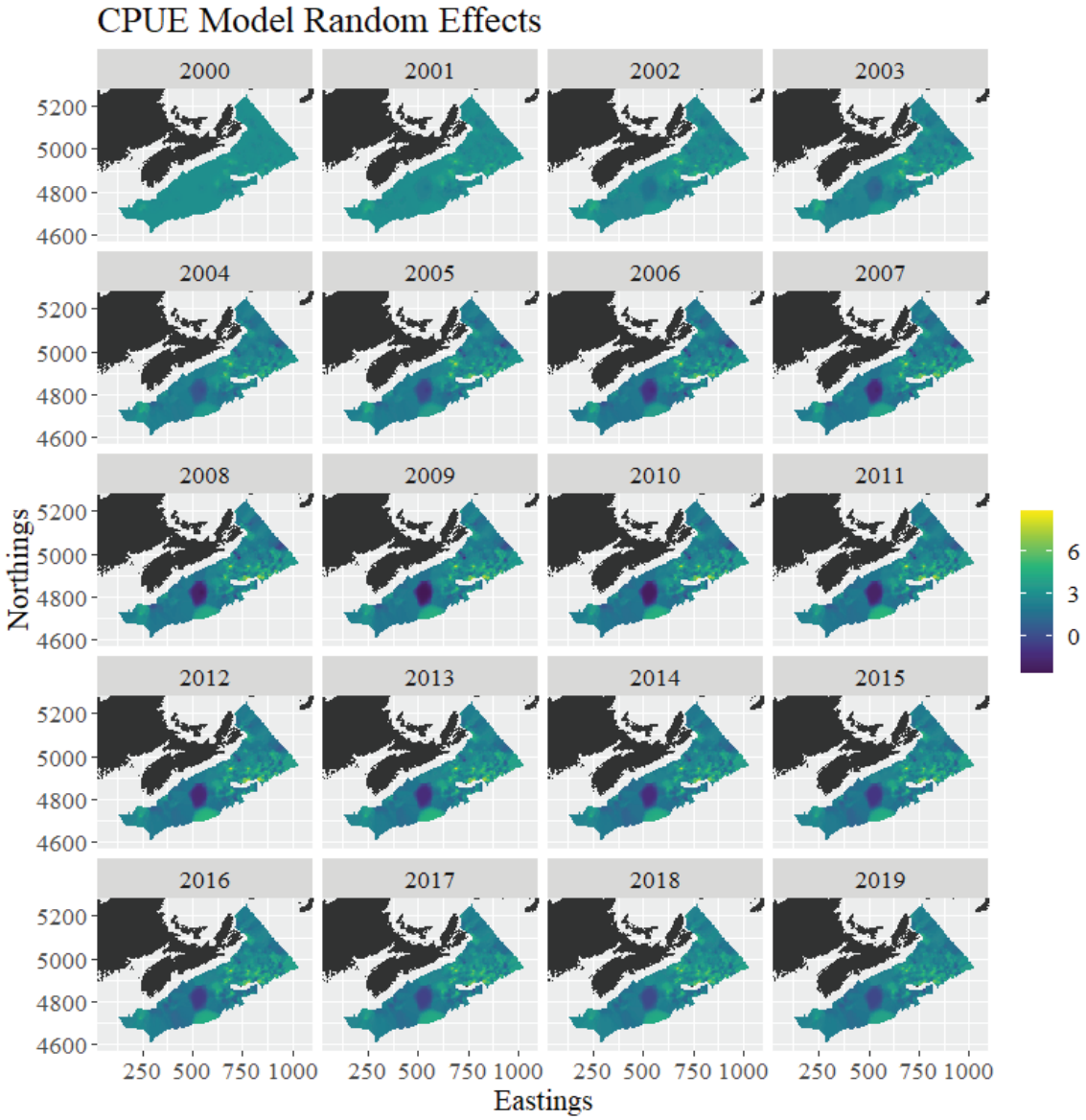


Figure B.13: The random effect contributions to the predictions shown in B.7.

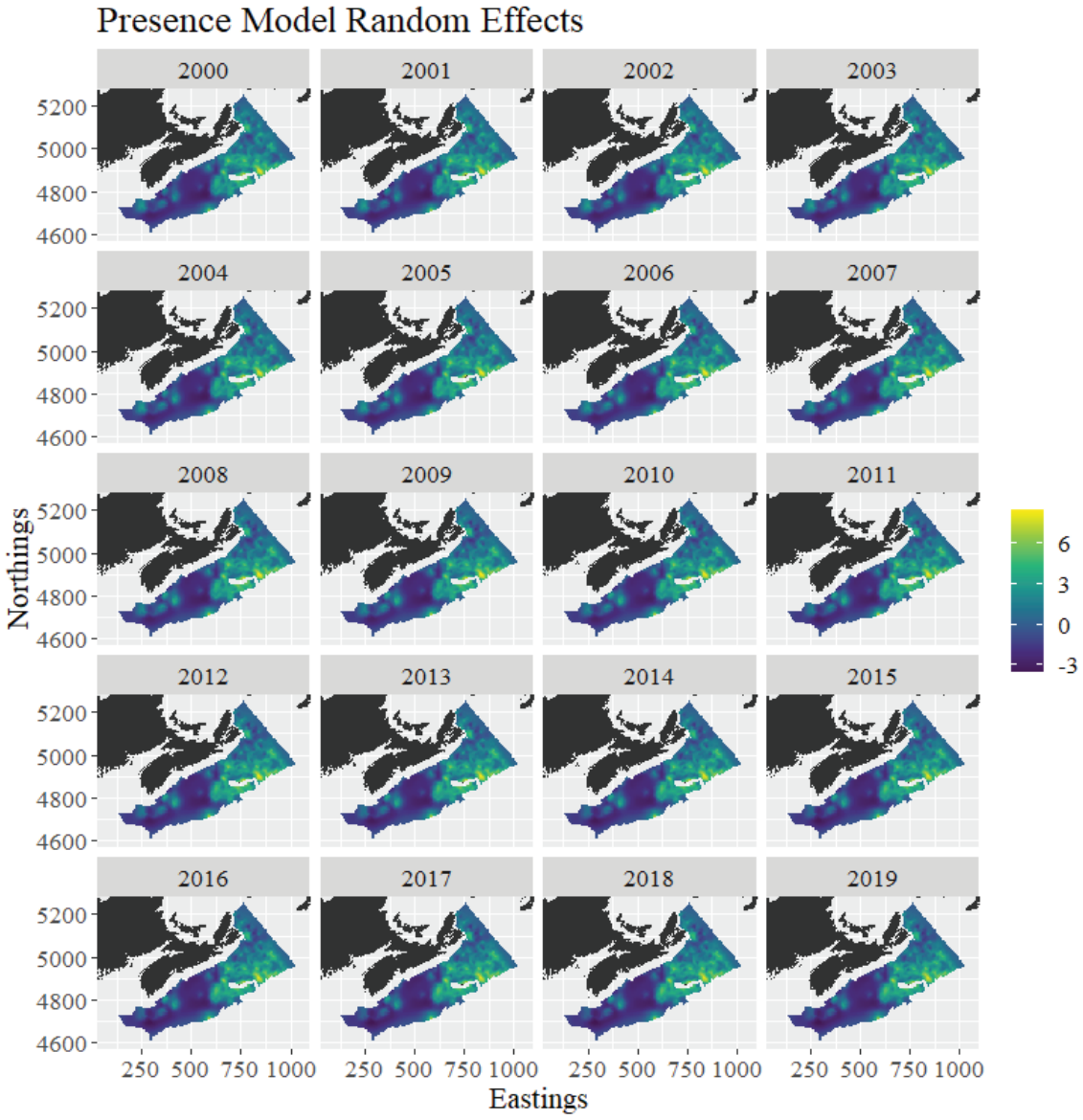


Figure B.14: The random effect contributions to the predictions shown in B.8. The random effects are in link space.

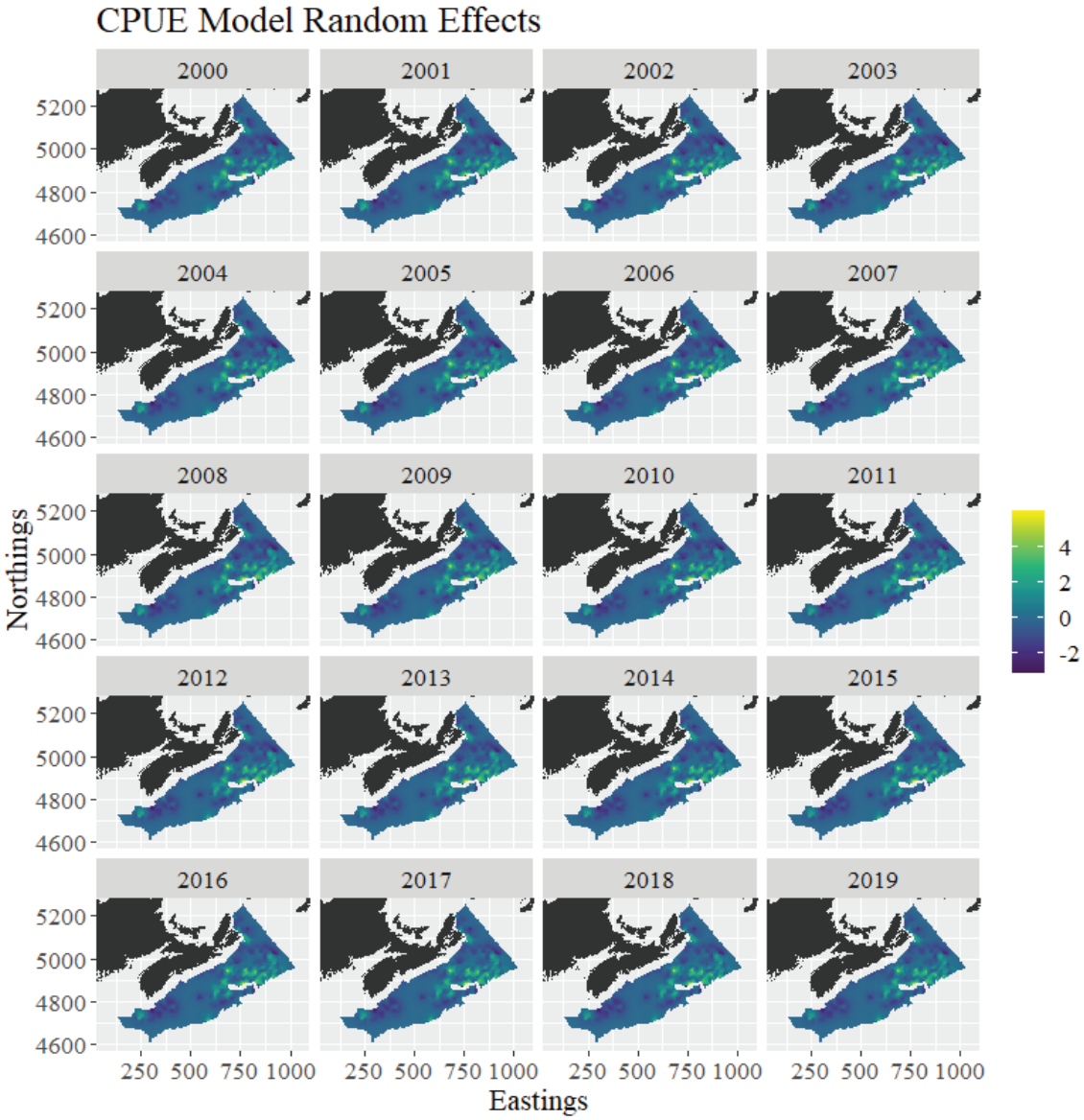


Figure B.15: The random effect contributions to the predictions shown in B.9.

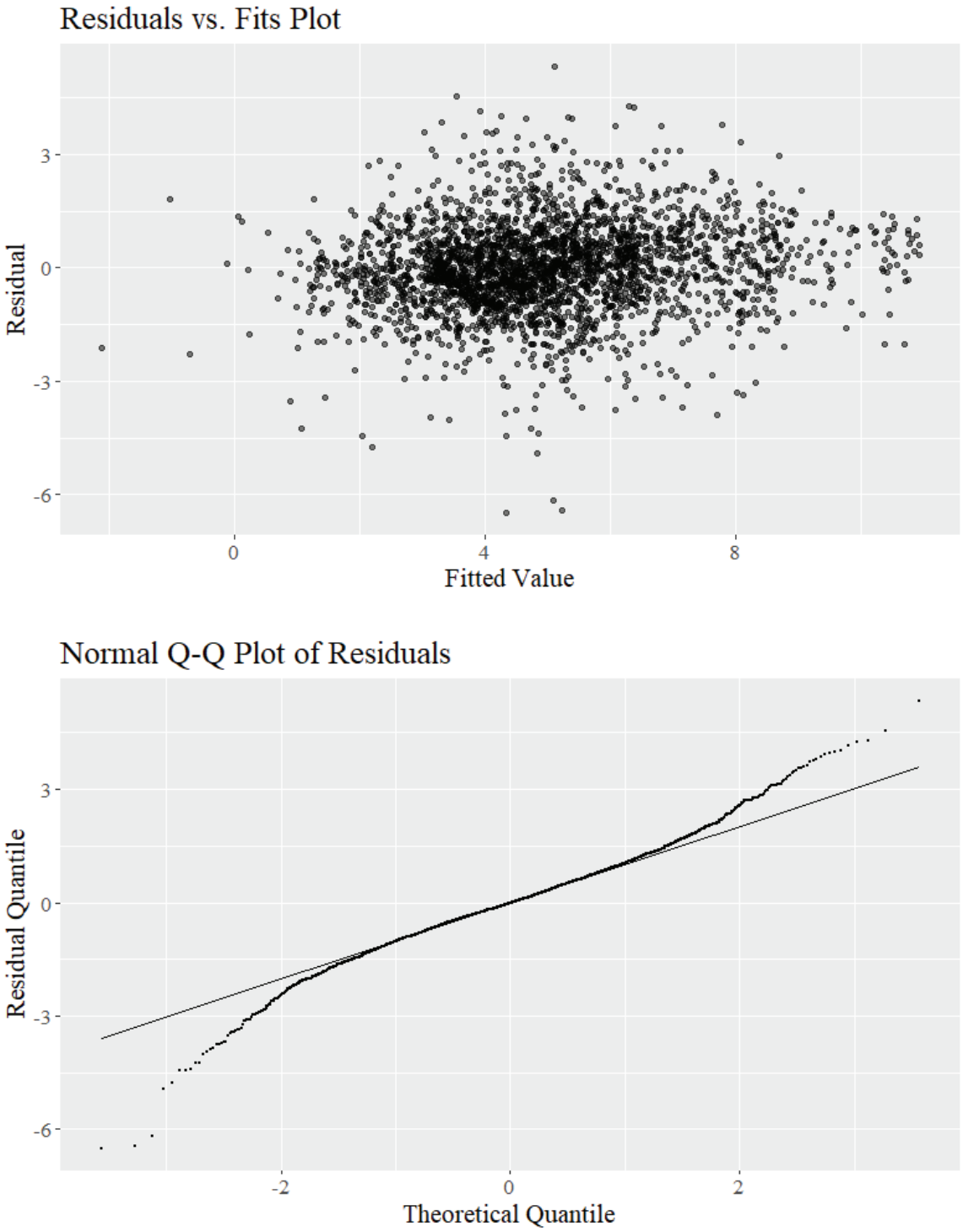


Figure B.16: Plots of the response residuals from the GAM CPUE sub-model.

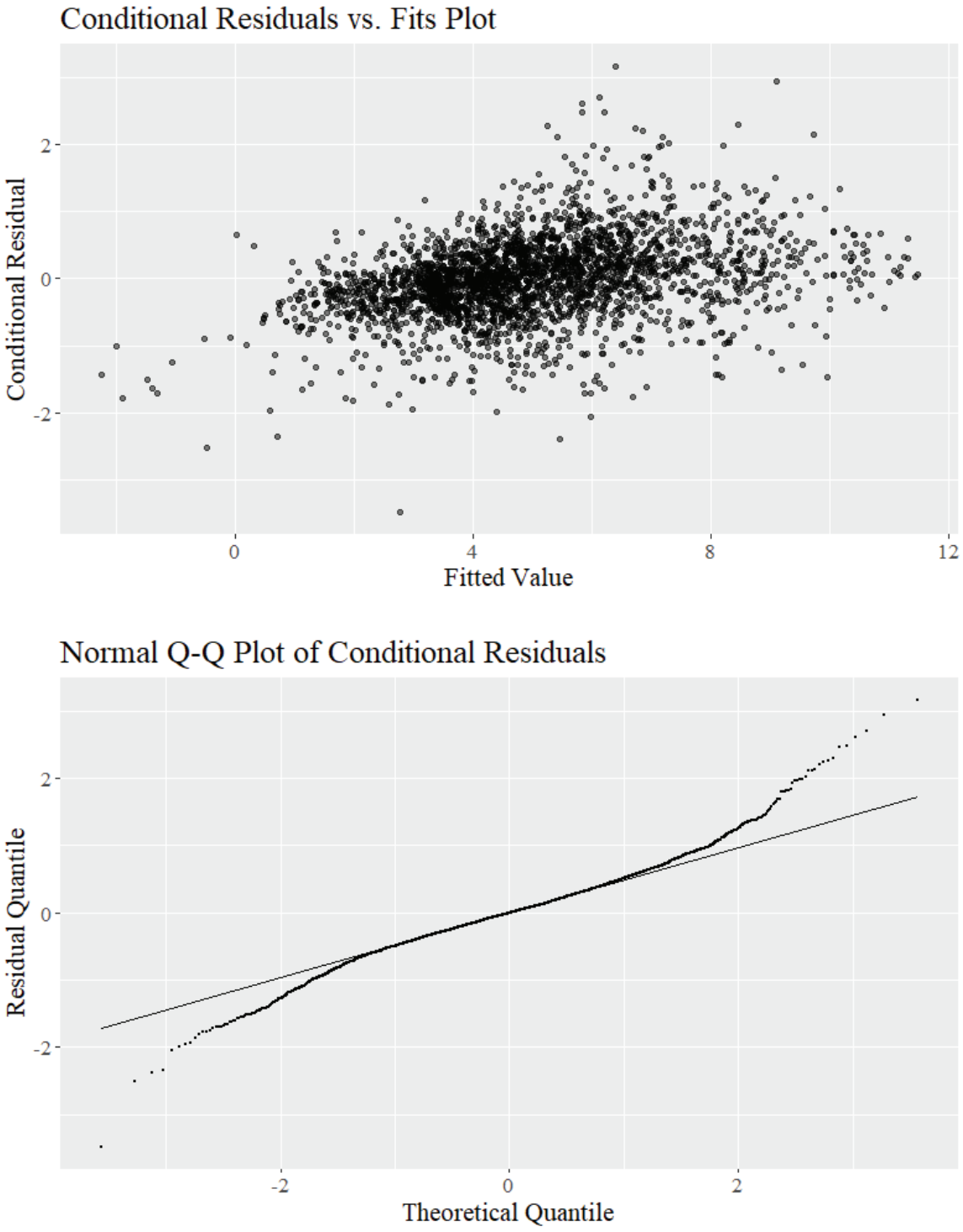


Figure B.17: Plots of the conditional response residuals from the *starve* CPUE submodel.

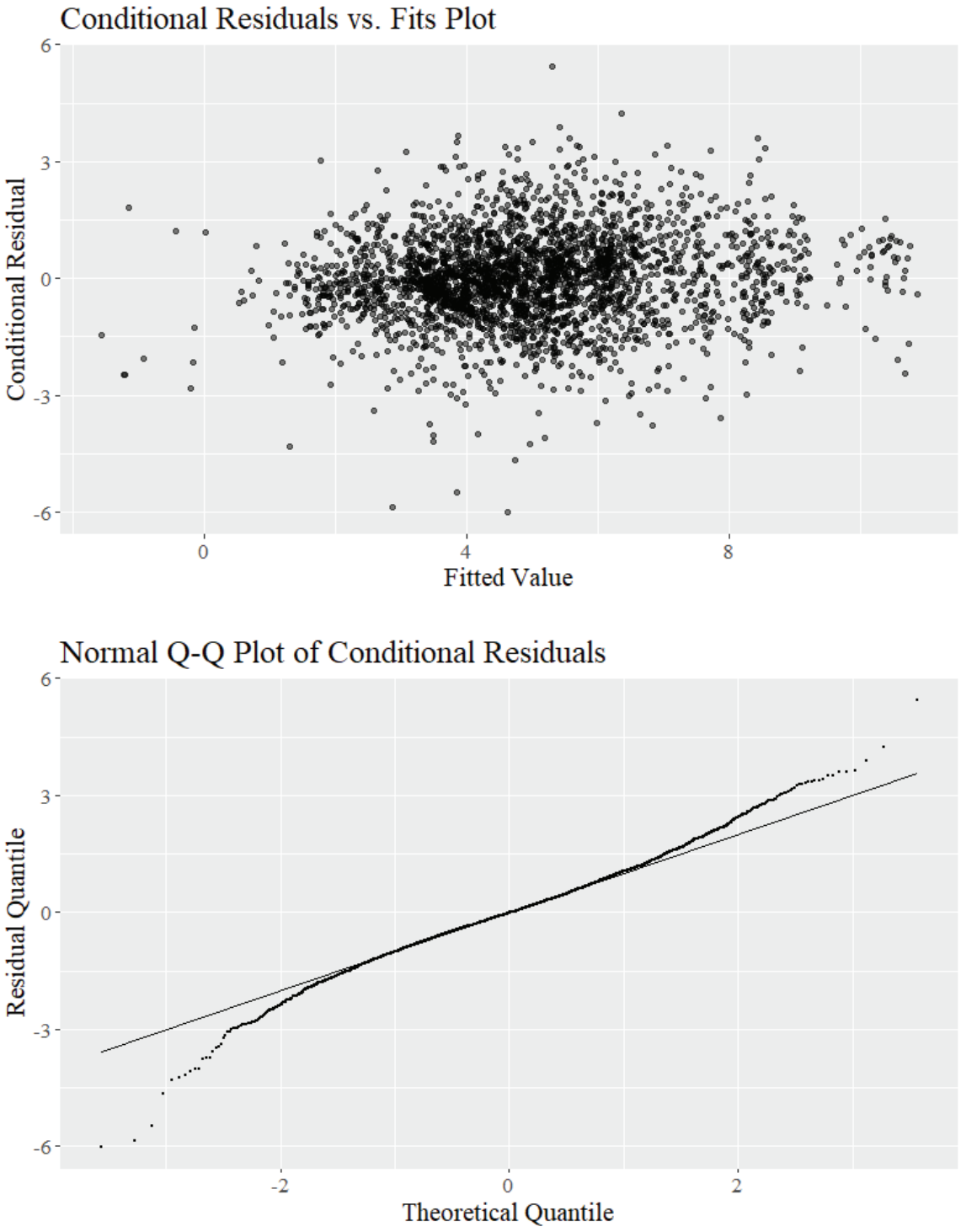


Figure B.18: Plots of the conditional response residuals from the GMRF CPUE submodel.