Modelling spatio-temporal preferences for albacore tuna

by

Craig Marsh

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Master of Science
in Biological Science.

Victoria University of Wellington
2015

Abstract

This research looks at using a spatially explicit population modeling technique to predict the relative spatial distribution of albacore tuna *Thunnus alalunga* around New Zealand. Spatial Population Model (SPM) software implements preference processes to move a population around a spatial structure within the modeled area. The preference attributes that best reproduced the relative spatial and temporal distribution of longline fishery commercial catch rates were estimated as sea surface temperature as a logistic function and sea surface gradient (magnitude of front) as a logistic function. Albacore tuna were found to prefer areas with warmer temperatures, and areas with high seas surface gradient, which represents oceanic fronts. This study also investigated incorporating dependence structures between preference attributes using copulas. Future research in the area of spatially explicit population modeling is then suggested.

Acknowledgments

Firstly a massive thanks goes to Matt Dunn, Nokuthaba Sibanda and Alistair Dunn. Without their knowledge, support, enthusiasm and guidance I wouldn't have gained the knowledge and skill set used to complete this thesis. A special shout out to Petros Hadjacostos for helping me understand the Copula framework. Andy Rae your advice and help with working out shortest paths and geodetic distances was greatly appreciated. MPI for funding, and institutes that have made environmental attributes public and easily accessible, for scientific use (Aviso, NOAA, Koordinates and Oregan University).

Contents

List of Tables

Table 1: Glossary of fisheries terms

Term	Definition
By-catch	Part of a catch of a fishing unit taken incidentally in addition to the target species towards which fishing effort is directed
Catch (C)	The total number of fish caught by fishing operations, estimated on board
Catch-per-unit-effort	CPUE: The amount of catch that is taken per unit of fishing effort
Catchability (q)	The extent to which a stock is susceptible to fishing
Ecosystem	A spatio-temporal system of the biosphere, including its living components and the non-living components of their environment
Fishing Effort (E)	The amount of fishing gear of a specific type used on the fishing grounds over a given unit of time
Pelagic	organisms that spend most of their life in the water column with little contact with or dependency on the bottom.
Population	A group of interbreeding organisms that represents the level of organization at which speciation begins
Population Size (N)	The number of individuals in a population
Stock	The part of a fish population which is under consideration from the point of view of actual or potential utilization
Stock Assessment	Collecting and analysing biological and statistical information to determine the changes in the abundance of fishery stocks in response to fishing
Stakeholder	An individual or group of individuals with an interest or claim who could potentially be impacted by or have an impact on a given project and its objectives

Table 2: Fisheries Abbreviations

Term	Definition
AIC	Akaikes Information Criteria
CPUE	Catch per unit effort
CR	Chatham Rise
EEZ	Exclusive Economic Zone
GLM	General Linear Model
GAM	General Additive Model
MPI	Ministry for Primary Industries
NIWA	National Institute for Water and Atmospheric Research
NPP	Net Primary Productivity
SPM	Spatial Population Model
SSH	Sea Surface Height
SST	Sea Surface Temperature
QMS	Quota Management System

Chapter 1

Introduction

In nature, organisms are neither distributed uniformly nor at random (Legendre and Fortin, 1989). Spatial heterogeneity, as it is known, can arise in populations because of many factors such as; biological variation (Begg et al., 1999), environmental variability (Lande et al., 1999, Lovett et al., 2005, Williams et al., 2002), habitat availability (Botsford et al., 2009, Spencer, 2008), or spatially variable exploitation (Booth, 2000, Prince and Hilborn, 1998, Stelzenmuller et al., 2008).

Traditionally spatial variation has usually been ignored in tactical fish population models. Rather, fish populations have been modeled with an implicit assumption of being spatially homogeneous within a defined area. Fish 'stocks' is the name given to a management unit of fish, assumed to be self-reproducing, with members of each group having common life history characteristics (Hilborn and Walters, 1992). Modeling stocks has been the focal point of fisheries modeling in the past (Haddon, 2010). Stock assessments aim to estimate the relationship between the rate of fishing and sustainable yield, and to determine the current status of an

exploited fish population within the context of that relationship (Botsford et al., 2009). Fisheries are often divided up into multiple stocks (e.g. New Zealand Snapper (*Pagrus Auratus*) fishery is thought to consist of four discrete stocks (for Primary Industries, 2014), where each stock is modeled as a closed population. This can create an element of discrete spatial management for a fishery, however stock boundaries are often difficult to distinguish, and in some instances, thought not to be discrete (Begg et al., 1999). Other ways in which space has been included in stock assessments is allowing for spatial structure implicitly, such as allowing spatially distinct fleets to have different catches and selectivities for the same stock (Anderson and Dunn, 2011). This creates an alias for space, provided fleets do not move from area to area. Explicitly a stock assessment traditionally assumes that there is spatially homogeneity within the stock geographical boundary.

There are a number of reasons why traditional tactical fish population models have not incorporated spatial structure, such as a poor understanding of the spatial dynamics of fish and fisheries (Guan et al., 2013), additional computational complexity (Levin et al., 1997), and management goals which do not allow for spatial structure. Historically, marine population models were generally conducted to answer and investigate management questions such as, how many fish can be sustainably utilised in the future for given management area? Haddon (2010). To answer these questions managers need to know information such as, the productivity and mortality of the stock. Fisheries stock assessment models are important tools that are central to fisheries management advice, and to ensuring the sustainable utilisation of fish stocks.

Fisheries management and ecologists has been recognising the importance of space and spatial structures in population models (Cope and Punt, 2011, Gimenez et al., 2014), for reasons such as; avoidance of local stock depletion (Ying et al., 2011), investigating spatial protected areas (Botsford et al., 2009), investigating bias in model outcomes (Su et al., 2012) and understanding spatial ecological processes (Chandler et al., 2014). This means spatially explicit theoretical frameworks for answering these queries need to be investigated and developed. There exists software available that can implement spatially explicit population models, such as SEAPODYM (Lehodey et al., 2008), Spatial Population Model (SPM) (Dunn et al., 2015). There is also software such as AD Model builder `http://admb-project.org/` where users can create their own models.

Along with spatial structure, climate variability has also been an important dynamic to capture that influences productivity, distribution, and exploitation of fish populations (Hofmann and Powell, 1998, Lehodey et al., 2006, McCarty, 2001). As mentioned earlier, populations spatial variability can be attributed to habitat heterogeneity. Making the link between environment (habitat) and population dynamics can provide insight, and potentially be used to forecast the effects of climate variability on populations. This information can aid in management and policy decisions for the future (Allison et al., 2009).

Population models simulate a population through time over a geographical area. Historically, these models have only kept track of numbers or biomass for particular groups or age classes of interest within a defined area. Spatially explicit population models (such as SPM) contain spatial constructs that allow the population to move within an area in a manner

determined by environmental or biological factors.

The spatial population model used in this investigation is implemented in the Spatial Population Model (SPM) software (Dunn et al., 2015), which applies a generalised spatially demographic model that can model movement of fish within areas according to demographic characteristics and movement preferences. The spatial movement idea in SPM is that populations are distributed based on preferred attributes for their environment. These attributes include any biotic and abiotic factor in the environment that affects the population's spatial distribution. In SPM, only closed populations can be explicitly handled. Spatial population models have been applied with a similar preference or habitat based movement mechanism on blue marlin (*Makaira nigricans*) Su et al. (2012) and shiners (*Tanakia limbata*) Sekine et al. (1997). SPM itself has also been applied on the Antarctic toothfish (*Dissostichus mawsoni*) exploratory fishery in the Ross Sea region (Mormede et al., 2013a). In another application of SPM on the Antarctic toothfish fishery Mormede et al. (2013b), found potential bias may enter stock assessments when spatial structure is ignored, giving further support to the conclusion that spatial variation is important to capture in population models.

SPM can use an array of preference attributes (abiotic or biotic properties) to describe the spatial distribution for a population. For each attribute, SPM uses a clear specified preference function to determine population relative preferences for a range of attribute values. SPM has the ability to incorporate many preference attributes to explain the movement of a population. When SPM applies multiple preference attributes, it describes the joint preference (that is the relative preference probability when mul-

tiple attributes are considered), by assuming independence among preference attributes. This is an assumption that is investigated in this thesis through the use of a multivariate technique called the Copula. The copula is a flexible technique for constructing joint preference probabilities based on a specified dependence structure and marginal preference functions.

Highly migratory species (HMS) such as tuna are ideal populations to apply a spatial model framework on. Typically HMS populations vary through space and time and migrations are linked to environmental conditions and habitat (Clemens, 1961, Nakamura, 1969). The species modeled in this thesis is the South Pacific albacore tuna (*Thunnus alalunga*), caught around New Zealand by method of longline fishing. It is thought that there is single stock of albacore tuna in the South Pacific (Murray, 1994). The stock is thought to be confined to an area with borders going from the equator to $50^{\circ}S$ and from $140^{\circ}E$ to $80^{\circ}W$. However, the extent of the fishery modeled in this study is only around New Zealand's Exclusive Economic Zone (EEZ) (Figure 1.1), this was due to data availability.

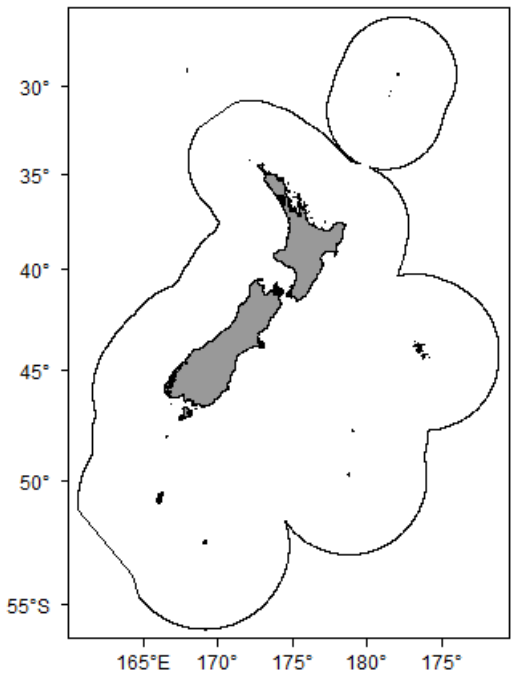

Figure 1.1: New Zealand's EEZ (line), which is the spatial scope of this study

Spawning of the south pacific albacore stock is thought to occur in the tropics and sub tropics between $10°S$ and $25°S$ during the New Zealand summer (Ramon and Bailey, 1996). Juveniles migrate south from this area to the sub tropical convergence zone (about $40°S$) a year later (Langley, 2004). This is the part of the stock that is observed in New Zealand's EEZ (Hoyle et al., 2013). This seasonal variation in the population distribution is evidenced by the high seasonal trends observed in New Zealand's long-line fishery (Slack, 1969), where the largest catches occurring in Autumn and Winter (see Figures 3.10-3.12).

Albacore tuna are a temperate pelagic species usually found in water

temperatures from about $10°C$-$20°C$ and attaining a maximum weight of 45 kg (Graham and Laurs, 1982). They are pelagic and to a limited extent warm blooded species that use physiological mechanisms to control heat loss and gain, to defend against the ambient environmental surroundings (Graham and Dickson, 1981). Albacore have supported commercially important fisheries and have been exploited within New Zealand's EEZ since the 1960s (for Primary Industries, 2014).

Previous studies have shown that albacore tuna presence and absence data are associated with environmental conditions such as Sea surface temperature (McGregor and Horn, 2013). Another study has found the relationship between CPUE data and Sea surface temperature (Lee et al., 2005)). The preference attributes used in this investigation were; sea surface temperature, Sea surface height, net primary productivity, sea surface gradient, depth and distance. These attributes are considered our explanatory variables for the relative spatial distribution of albacore tuna around New Zealand.

The objective of this study was to evaluate the performance of different preference attributes in predicting the observed fishery catch rates using a spatially explicit population model framework (SPM). Chapter 2 describes how SPM is used to model populations with emphasis on the spatially explicit movement process and the assumed processes. Chapter 3 describes the environmental (Preference) and observed fisheries data sets and how these datasets were processed for model runs, and explores the relationships between the various data. Chapter 4 sets out; model assumptions that were investigated, modification to how SPM calculates distances within the model, and how estimation and model selection was

undertaken. Chapter 5 contains results from model runs with Chapter 6 discussing findings, limitations and future work.

Chapter 2

The SPM model

Spatial Population Model (SPM) is a software program for modeling population dynamics and spatial distribution of a fish stock (Dunn et al., 2015). SPM has the ability to run as a deterministic model (fixed parameters), calculate parameter point estimates with observed data, run likelihood profiles on parameters, run MCMC chains using Bayesian estimation and simulate observations (based on an input file). The population model implemented in SPM is a generalised spatially explicit demographic population model. SPM is flexible in the sense that it can be applied to many marine species stocks. At its core SPM is an age based demographic model (Haddon, 2010) that keeps track of numbers of fish in specific age classes. Like all demographic models, SPM can apply population processes such as; recruitment, growth, and mortality. SPM differs from most demographic models in it's explicit spatial component. SPM has an array of spatial population movement processes that the user can implement. The movement processes used in this study are based on the concept of preference, where the population is distributed over an area based on known (observed) spa-

tially and temporally varying characteristics. That is, highly preferable areas contain a higher proportion of the population compared to less preferable areas at a point in time. SPM assumes a closed population within the modeled area throughout time. This means that external migration outside of the modeled area is not modeled in SPM explicitly. There was a way of artificially migrating the population in and out of the core modelling area, this is explained in section 2.3 of this thesis. SPM can integrate a range of observed datasets that are common in fisheries data i.e abundance or biomass indices, age composition data, and tagging data. This adds information available to the model when estimating parameters.

SPM simulates a population abundance over a spatial 'state' through discrete time steps based on deterministic processes (controlled by parameters $\boldsymbol{\theta}$). A populations state represent an estimated version of the spatial and age structure of the population at a point in time. The processes act on the state of the population at time $t-1$ to create an estimated state of the population at time t. This is analogous to traditional population models (Equation 2.1). These deterministic processes can be either population or movement processes in SPM, where population processes alter the size of the population, and movement processes that reorganise the population within the state. SPM can fit models to data by estimation of $\boldsymbol{\theta}$. This can be done via optimistion of a likelihood objective function or Bayesian estimation. SPM models these states through discrete time steps to capture temporal change in the population, these discrete times steps are the temporal resolution of the model. A state can be measured at any time interval and spatial structure the user wishes to investigate for a given stock. Figure 2.1 shows a range of spatial structures that could be assumed to represent the

state of a population. SPM can model populations in discrete time steps that can range from; days, weeks or even years.

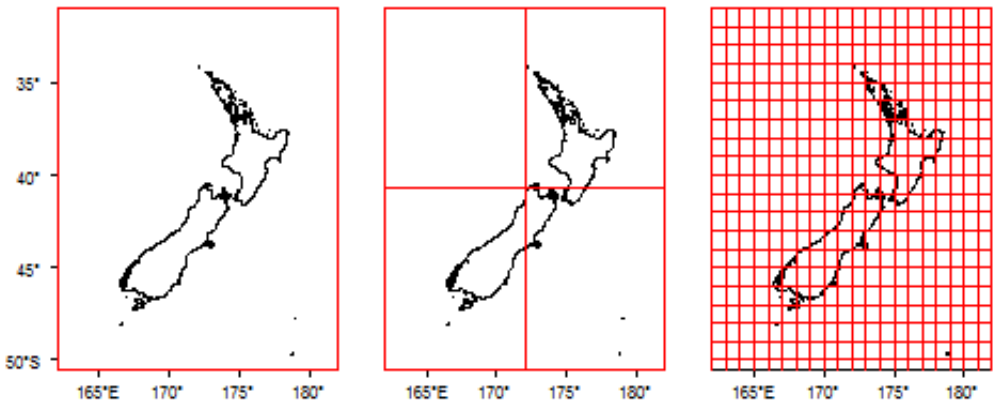

Figure 2.1: Examples of different spatial structures for a 'state' at a given time step. Moving from the simplest spatial structure (a single cell) on the left to more complex spatial structures to the right

2.1 Population Models

A traditional deterministic population model, with an implicit assumption of spatial homogeneity, would be modeled as Equation 2.1. This corresponds to the state in Figure 2.1 (left panel) and is specified as,

$$N_{a,t} = N_{a-1,t-1} + f(N_{a-1,t-1}; \boldsymbol{\theta}) \tag{2.1}$$

where $N_{a,t}$ is the expected abundance of the stock at time step t for age class a, $f(N_{a-1,t-1}; \boldsymbol{\theta})$ is a population process that reflects how abundance at time $t-1$ changes to that at t e.g growth processes and mortality. $\boldsymbol{\theta}$ is a vector of parameters that control these population processes.

SPM can implement Equation 2.1 when the spatial structure has one

cell,

$$N_{a,t,k} = N_{a-1,t-1,k} + f(N_{a-1,t-1,k}; \boldsymbol{\theta}) \tag{2.2}$$

where $k = 1, 2, 3, ..., RC$

for $R =$ number of rows and $C =$ number of columns of the spatial state. Equation 2.2 is special case of Equation 2.1 for $RC = 1$. SPM has a large number of population processes $f(N_{a-1,t-1,k}; \boldsymbol{\theta})$. These can include recruitment relationships (Ricker model (Ricker, 1954)), mortality processes, and more, which are described in the SPM manual (Dunn et al., 2015).

Before applying Equation 2.2 an intialisation phase is required. The aim of this phase is to set and spatially distribute a virgin/initial abundance N_0, from which modeling can commence. This phase can run for as many time steps as the user wishes or until the initial population stabilises. In fisheries stock assessments this stability assumes that the population is at carrying capacity. This phase can include its own population parameters.

2.2 Movement Processes

This study is focused on the spatially explicit component of SPM. This is Equation 2.2, when $RC > 1$ (e.g. Figure 2.1). When that is the case, $f(N_{a-1,t-1,k}; \boldsymbol{\theta})$ is expanded to capture movement processes $f(N_{a-1,t-1,k}; \boldsymbol{\theta}, \boldsymbol{\phi})$, where $\boldsymbol{\theta}$ controls population processes and $\boldsymbol{\phi}$ controls movement processes. There are a range of movement processes that SPM can apply. These are; adjacent cell movement, preference movement, and directed migration. Adjacent cell movement moves a proportion of individuals in each cell to

its four neighbouring cells, and hence mimics a simple diffusion process. Migration movement moves individuals from one sets of locations (the source) to another set of locations (the sink) within the state. More detail is available in the SPM manual (Dunn et al., 2015).

The movement process used in this study is preference movement. This was chosen as, highly migratory species are thought to take advantage of the oceanographic heterogeneity to seek optimal conditions, forage prey or migrate to favored habitat Xu et al. (2013), this implies environmental preference is a useful mechanism in predicting populations spatial distribution. This is done by assuming a relationship known as a preference function, between a known oceanographic attribute and the species preference for that attribute. The list of preference functions are described later in section 3.3. Suppose there exists an attribute X, and that its values x_k are known for all cells in of the structure for all time steps. With an assumed functional form $f(x_k, \boldsymbol{\phi})$, SPM can calculate preference value p_k for each cell k. SPM can also incorporate multiple attributes $X_1, X_2, X_3 ... X_n$ calculating the joint preference value p_k of preference for cell k (Dunn et al., 2015),

$$p_k = f_1(x_{1k}; \boldsymbol{\phi}_1)^{\alpha_1} \times f_2(x_{2k}; \boldsymbol{\phi}_2)^{\alpha_2} \times ... \times f_n(x_{nk}; \boldsymbol{\phi}_n)^{\alpha_n} \tag{2.3}$$

where $k = 1, 2, 3, ..., RC$ and α_n is a weighting factor for preference attribute n. α_n is a parameter that controls the importance of its corresponding preference attribute and function. One α_n is always fixed to the value of one, and if other attributes are present their α_n is estimable. This represents how important attributes with a certain functional form are com-

pared to the attribute and form that is fixed to one. For example if there were two preference attributes with their respective functional forms being modeled, the first attributes α is =1 and say the other is estimated to be 0.08, then the second attribute contributes 8% of the information. Equation 2.3 creates a preference distribution over the state's spatial structure. SPM applies the preference as a relative preference r_k

$$r_k = \frac{p_k}{\sum_{l=1}^{RC} p_l} \tag{2.4a}$$

$$N_{t,k} = N_t r_{t,k}, \tag{2.4b}$$

where $N_{t,k}$ is the expected abundance for cell k in time step t and $r_{t,k}$ is the relative preference probability for cell k in time step t. This ensures that,

$$\sum_{k=1}^{RC} r_{t,k} = 1 \tag{2.5a}$$

$$\sum_{k=1}^{RC} N_{t,k} = N_t \tag{2.5b}$$

This means that a model equation that incorporates preference movement and other population processes can be written as,

$$N_{a,t,k} = f_{a-1}(N_{a-1,t-1}, \boldsymbol{X}_{t,k}; \boldsymbol{\theta}, \boldsymbol{\phi}) \tag{2.6}$$

where $N_{a,t,k}$ is abundance in cell k at time step t for age class a, $\boldsymbol{X_{t,k}}$ is

the vector of oceanic attributes for cell k in time step t, $f_{a-1}(N_{a-1,t-1}, \boldsymbol{X}_{t,k}; \boldsymbol{\theta})$ is the preference function controlled by $\boldsymbol{\phi}$ and or population processes controlled by $\boldsymbol{\theta}$ for age class $a-1$.

From Equation 2.3 we see that SPM describes the joint preference function by multiplying marginal preference functions. This method assumes that preference functions are independent of one another. This assumption is investigated in Chapter 4.

Another assumption applied in this movement process is one of stationarity, explained by Tveito et al., 2006, pg. 40 *" the condition that the probability distribution of the variable is constant in time and space, meaning that the same distribution function should be expected anywhere/any time..."*. This assumption is a result of simplifying the model in order to maintain parsimony. Although SPM has flexibility in adjusting for temporal stationarity e.g., within a year a preference function can take different functional forms, it will always have the assumption of spatial stationarity. SPM also assumes that each movement process occurs simultaneously and instantaneously over all cells (synchronous updating).

Equation 2.6 describes the the deterministic process model controlled by parameters $\boldsymbol{\theta}$ and or $\boldsymbol{\phi}$. SPM has the ability to estimate these parameters using observed data. This is explained in section 4.3.

2.3 Model Structure Setup

This study aims at recreating the relative spatial distribution of albacore tuna in New Zealand's EEZ through time. We were not interested in where specific age classes are distributed (we also do not have the data). For this reason, the age structure of the population was ignored in this study. All individuals were assumed to be of the same age. Along with no age structure, there is no growth assumed for the population. Model runs in this investigation started with an initialisation phase. This phase ran for one year, and within this year two processes occurred. The first process was recruitment. This seeded an initial population of 1000000, denoted P. The second process distributed this population (P) evenly to the northern latitudes (26-28$^{\circ}S$). After initialisation no more population processes were applied within the model, no catch was removed so the population remained constant during model runs. This implies $\boldsymbol{\theta} = 0$ in Equation 2.6, see appendix A.6 for input script. This assumes a fixed and constant population in this investigation. After this phase was the model of interest was conducted, which started in Summer 2003.

The model ran for the years 2003-2012. Within each year there were four seasonal time steps; summer (december, january, february), autumn (march, april and may), Winter (june, july, and august, and spring (september, october, and november)). Discrete time steps of season was a trade off between observed data availability and variability of oceanic characteristics (see Section 3.1). At each time step the spatial structure of the state consisted of a rectangular grid of 23 rows by 21 columns, which were representative of 1° latitude and longitude bins. This was assumed due to

availability of oceanic characteristic data resolution (See section 3.1). The extent of this spatial state extended from $26°S$ to $49°S$ and $164°E$ to $175°W$, to encompass the range of observed data.

Previous studies have suggested that Albacore migrate south into the subtropics from the tropics in summer, with a return migration in winter Hoyle et al. (2013). SPM assumes a closed population. Since SPM assumes a closed population, migration outside of the area of interest was artificially imposed by adding an extra time step in between Spring and Summer. This was a migration time step that redistributed all the biomass to the northen latitudes (26 - 28°S). This assumed every summer the population had to re-enter New Zealand's EEZ from the northern latitudes, and that all fish had to leave New Zealand's EEZ after Winter. These assumptions seemed reasonable given the tropical to temperate migration of the species.

Although SPM is an integrated population model, this study did not have other observed datasets that described population characteristics other than CPUE. This meant the integrated aspect of SPM was not used in the scope of this study, and left the final model as,

$$N_{tk} = f(P, \boldsymbol{X}_{tk}; \boldsymbol{\phi}) \tag{2.7}$$

where N_{tk} is the abundance in cell k at time step t, P is our fixed population (1000000), $\boldsymbol{X}_{tk}$ is the vector of preference attributes that represent cell the conditions of cell k at time step t, $\boldsymbol{\phi}$ is the vector of preference parameters and catchability parameter q. Equation 2.7 describes how SPM models the relative abundance for each cell through time. The model pre-

diction is compared to observed data in cells where abundance was observed to evaluate the suitability of $\boldsymbol{X}$ (preference attributes) and their parameters ϕ. The following chapter describes how the preference attributes and observed abundance indices were obtained, and the data manipulations done for SPM to allow them to be used in model runs.

Chapter 3

Exploratory Data Analysis

Two sets of data were used. The first consists of spatial layers of environmental and biological characteristics, used in modeling movement processes as explanatory variables (preference attributes). The second dataset is of historical albacore tuna catch and effort in New Zealand's EEZ (the response variable).

New Zealand's EEZ was divided up into spatial (Figure 3.1) and temporal cells, where each cell represents a particular location specified by longitude and latitude coordinates, at a given time. Each cell had a value for each of the preference attributes and the response variable where available. An important aspect of the analysis was identifying a suitable spatial and temporal structure to help answer the question of interest. In this thesis the temporal and spatial resolution was restricted by preference attribute resolution and observed catch and effort data availability. The finest temporal resolution within a year that could be modeled was monthly due, to preference data resolution. This was considered to fine Monthly resolution meant only 32% of the monthly time steps contained

at least one catch and effort observation. We therefore aggregated the temporal resolution into seasons. This was still adequate for capturing environmental variation over time (e.g. SST change, Figure 3.2). Seasonal time steps also contain a higher number of observations where 70% contained at least one cell with a observed catch and effort observation. This meant for each year, our model has five time steps, with respected months of each season in brackets:

- Summer (December, January, and February)
- Autumn (March, April, and May)
- Winter (June, July, and August)
- Spring (September, October, and November)
- Migration (instantaneous time step, no preference attribute related, or observed data is associated with this time step)

this meant we modeled 5 time steps over 10 years (2003-2012) so $t = 1, 2, 3, ...50$. The spatial structure is shown in Figure 3.1,

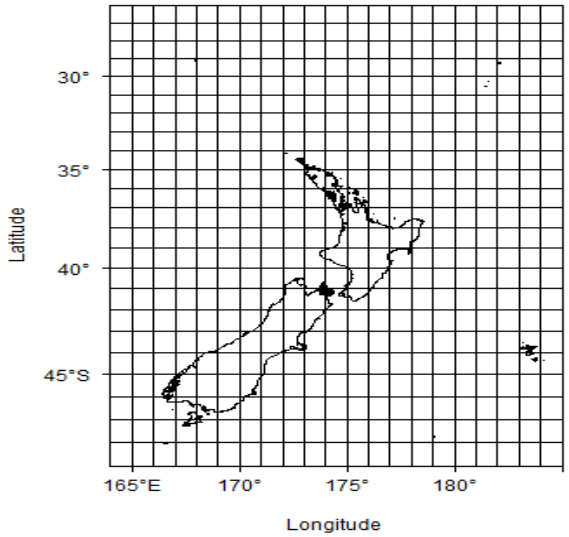

Figure 3.1: Spatial resolution for this model, where each cell is a 1° latitude and longitude bin

All data was handled in the statistical software R (R Core Team, 2013), with the exception being net primary productivity, which was handled using MATLAB (Guide, 1998).

3.1 Preference attributes

All preference attributes are produced using a combination of remote sensing via using satellite imagery and in situ data with interpolated models. There is a range of methods used for collecting *in situ* data such as; research vessel instruments, controlled and uncontrolled buoys. *In situ* data and satellite imagery are used to inform interpolating models that fill in the gaps where observations are missing. These create point estimates and are assumed to reflect that attribute accurately, for an area at a point

in time. Preference attributes were chosen on the basis of availability, spatial/temporal coverage (1° latitude and longitude and seasonal time steps) and biological significance (McGregor and Horn, 2013). This criteria left us with the following variables:

- Sea Surface Temperature (SST)
- Net Primary Productivity (NPP)
- Sea Surface Height (SSH)
- Sea Surface Gradient (MAG)
- Bathymetry
- Distance

The following section explains the processes used to obtain and manipulate each attribute

3.1.1 Sea Surface Temperature (SST)

SST represents the mean global skin (top 20μm of the surface water) temperature of the ocean measured in $^{\circ}C$. In this study it is assumed that it represents the temperature of at least the fished water column. Albacore tuna have been associated with SST ranges (McGregor and Horn, 2013, Ramos et al., 1996). SST is thought to influence tuna thermo-regulation and food availability (Laurs et al., 1984). The SST values were sourced from `http://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html`, where the product used was based on the interpolated

model NOAA OI.v2 SST (Reynolds and Smith, 1994, Smith and Reynolds, 1998). The temporal and spatial resolution is in monthly time intervals, on a 1° latitude and longitude spatial resolution. SST has the coarsest spatial resolution relative to the other attributes, and as a consequence the 1° spatial resolution was maintained for all other data and model runs, and no spatial manipulation was needed for SST. The seasonal SST values for each cell were created by taking the mean of the respective monthly values for each season and location For each cell i (representing a unique season and location combination $\bar{x}_i^t$), the average SST value was given by

$$x_i^t = \frac{\sum_{m=1}^{M} x_{i,m}^t}{M}, \tag{3.1}$$

where m is the month, and M is the number of months in the season. In this case $M = 3$.

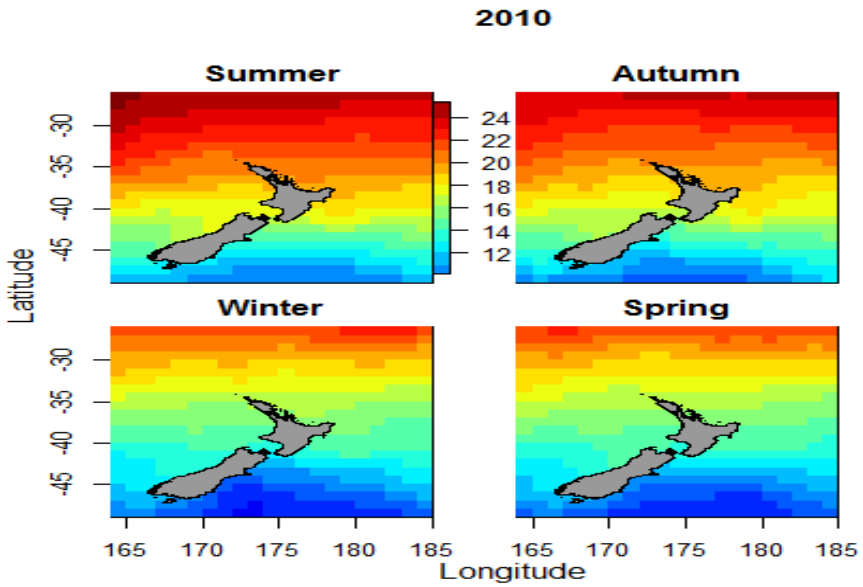

Figure 3.2: Sea surface temperature $°C$ for four seasons in 2010. In this study area around New Zealand

Figure 3.2 shows the typical seasonal variability of SST in a year. Warmer waters occur during Summer and Autumn, with cooler water occurring in Winter and Spring.

3.1.2 Net Primary Productivity (NPP)

NPP is the average rate of net primary productivity, amount of carbon per square meter per day ($mg\ C/m^2/d$). This attribute represents the oceans productivity, and thus can be assumed to be related to potential carbon available that an ecosystem can utilize for all other metabolic processes (Falkowski et al., 2003). Tuna species such as yellow fin tuna have been associated with stable NPP between 220-380 $mg\ C/m^2/d$ (Lan et al., 2013). NPP is sourced from the Oregon states ocean productivity website `http:`

//www.science.oregonstate.edu/ocean.productivity/. NPP is derived via the Vertically Generalized Production Model (VGPM). VGPM is a light-dependent, vertically integrated model that partitions environmental factors affecting primary production into two categories; those that influence the relative vertical distribution of primary production, and those that control the optimal efficiency of the productivity profile (Behrenfeld and Falkowski, 1997). The inputs used in the VGPM are; SST, day length, depth of euphotic zone and chlorophyll concentration. There are two time series (1997-2009 and 2002-2014) that are available for NPP. Each time series relates to a different sensor used (seaWiFS and MODIS respectively). This meant that they cannot be used as a single time series from 1997-2012. Therefore, in this study the latter time series was used (MODIS, 2002-2014). The MODIS time series started halfway through 2002, so we only included 2003-2012. As a result from the inclusion of NPP, all model runs and hence data, were restricted for that period. These data come in the format Hierarchical Data Format (HDF), which was handled in the program Matlab (MATLAB 7.14, (Guide, 1998)). The temporal resolution of this product was monthly, on a $0.25°$ latitude and longitude spatial scale. To obtain a value for each cell at a seasonal and $1°$resolution, the NPP values were averaged over all values in the cell as shown in Equation 3.2,

$$\bar{x}_i^p = \frac{\sum_{j=1}^{J} \sum_{m=1}^{M} x_{j,m}}{JM} \tag{3.2}$$

where $\bar{x}_i^p$ is the mean value for cell i for the season, that is covered from month m to month M (M=3). j in this case are the $0.25°$ resolution cells within i ($j = 16$).

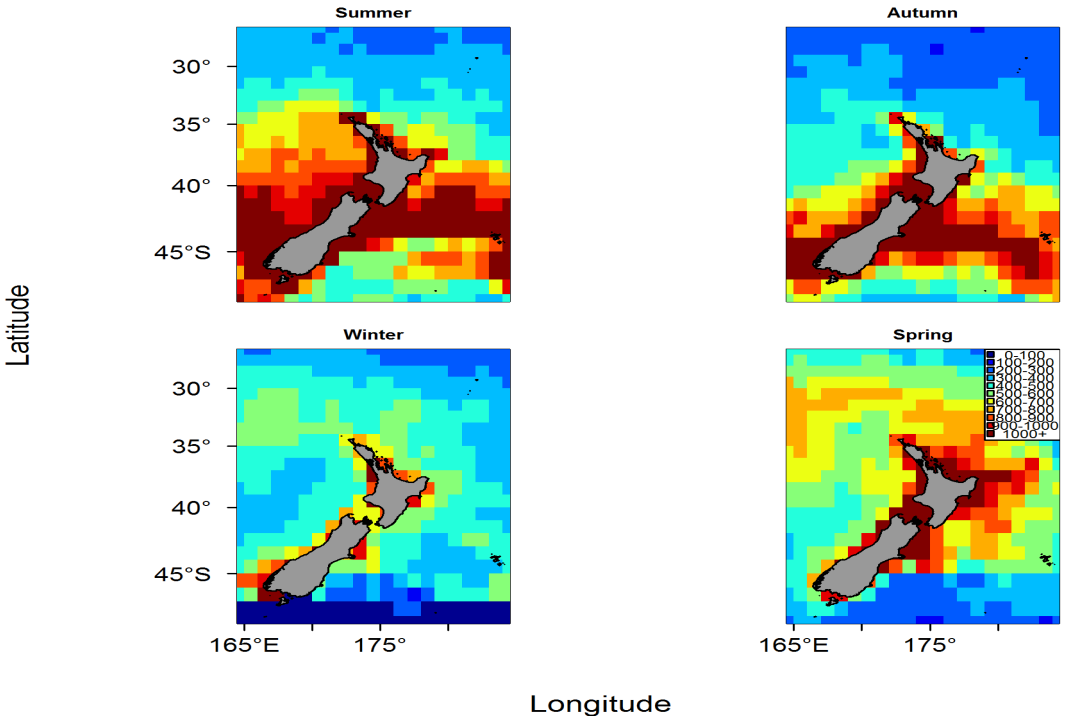

Figure 3.3: Net Primary Productivity From the VGBM model for 2010 measured in $mg\ C/m^2/day$ using the MODIS data

There is seasonal and temporal variation in NPP with higher NPP associated with coastal areas and other features such as Chatham Rise. NPP is generally higher in summer and autumn than in spring and winter.

3.1.3 Sea Surface Height (SSH)

SSH represents the height of the ocean (in cm's) from a reference ellipsoid `http://www.aviso.altimetry.fr/en/techniques/altimetry/principle.html`. SSH is determined by the mass of water at a given location and by the water density (a function of temperature, salinity, and pressure). This gives an indication of fluctuation in both current strength and direction

(Bradshaw et al., 2004). SSH is considered a habitat preference attribute because of it's relationship with areas of upwelling which have been associated with tuna (Xu et al., 2013). SSH consists of sea surface height anomaly (irregularities) added to a Mean Dynamic Topography (MDT). The MDT is part of SSH that was due to permanent currents and processes. More information about the MDT used can be found on the AVISO web site `http://www.aviso.altimetry.fr/en/home.html`. The reference period for MDT that was used in this study is from 1993 to the 1999.

SSH comes in weekly time intervals and in 0.25 degree latitude and longitude cells and seasonal resolution by averaging over all values within a cell.

$$\bar{x}_k^h = \frac{\sum_{j=1}^{J} \sum_{w=1}^{W} x_{k,j,w}}{JW}, \tag{3.3}$$

where $\bar{x}_k^h$ is the mean value for cell k for the season, that is covered from week w to week W ($W = 13$). j in this case are the $0.25°$ resolution cells within k ($j = 16$).

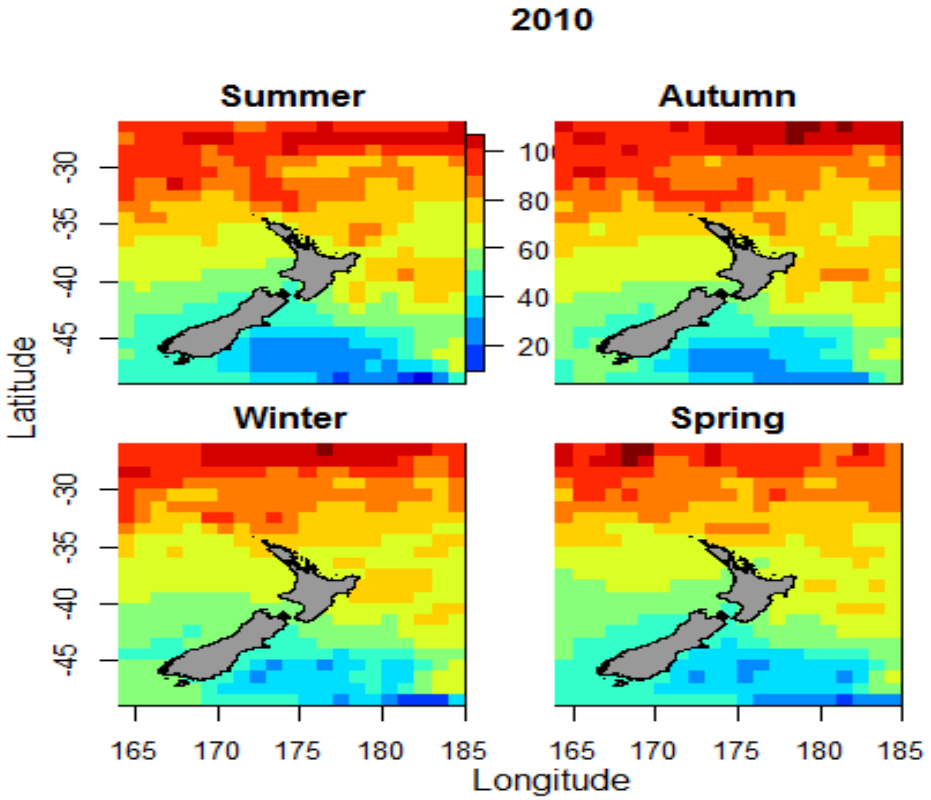

Figure 3.4: Sea Surface Height(SSH) for the four seasons of 2010, summer top left, Autumn top right, winter bottom left and spring bottom right. Measured in cm

There is spatial variation in SSH, but relatively limited seasonal/temporal variation. SSH is higher in the north than the south.

3.1.4 Sea Surface Gradient (MAG)

Oceanic fronts are regions where water mass properties and SSH show strong surface gradients (Fernandez, 2012). Sea surface gradient reinterprets SSH to create another attribute that reflects oceanic fronts. This was used for the frontal attribute which has been thought to be associ-

ated with tuna (Xu et al., 2013). How oceanic fronts are identified is by calculating the gradient of SSH. This is done by calculating the magnitude of the meridional and zonal axis of SSH (Equation 3.4). Meridional and Zonal gradients are the difference in SSH over a distance of 0.5° in there respected direction (Figure 3.4). The total gradient was measured via calculating the hypotenuse between the meridional and zonal gradients as done in the following Figure/equation. Sea surface gradient was calculated on the original SSH temporal and resolution. This was then converted to a 1° latitude and longitude seasonal resolution by averaging the attribute over all values within a cell (Equation 3.3). The measurement units of this variable are in $cm\ km^{-1}$ which represent change in Sea Surface height (cm) over an area in km.

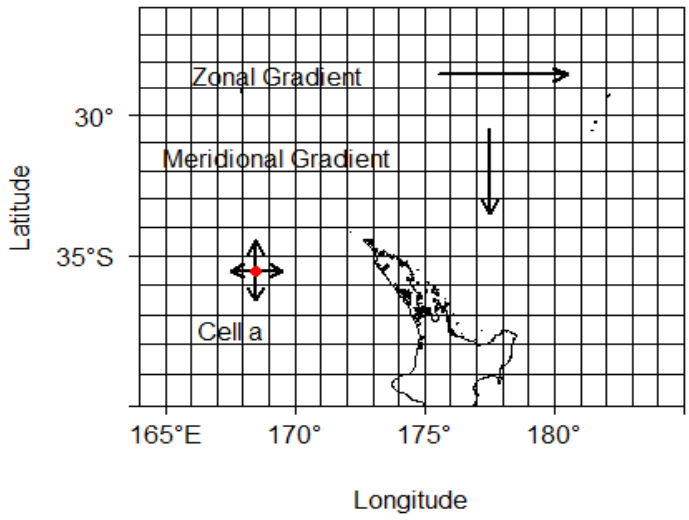

Figure 3.5: Calculating the zonal z_a and meridional gradient m_a for cell a

$$x_a^g = \sqrt{z_a^2 + M_a^2} \tag{3.4}$$

Where x_a^g is the sea surface gradient for cell a and z_a and m_a are the zonal and meridional gradients. This generated a value that represents the frontal magnitude for cell a. An example of what this preference attribute looks like for the four seasons in the year 2010, is given in Figure 3.6, which shows large frontal areas to the far west coast and patchy fronts to the south and east coast of the south island.

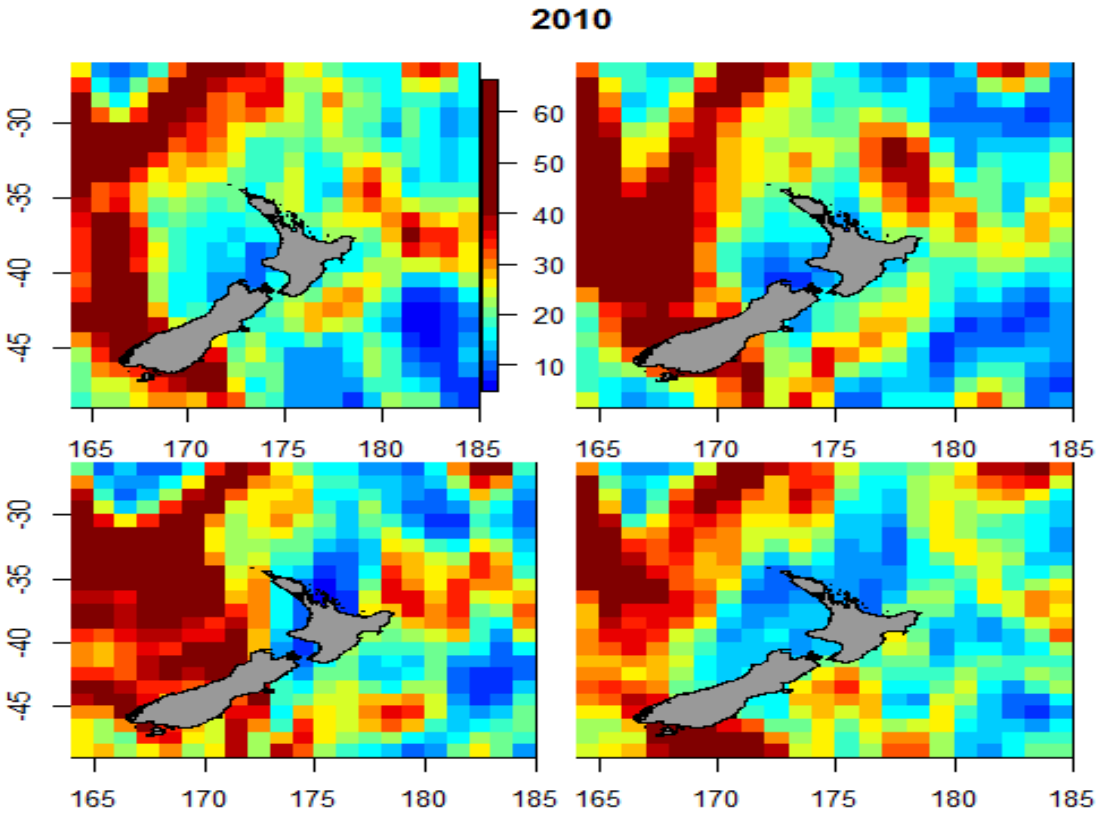

Figure 3.6: Sea Surface Gradient (MAG) for the four seasons of 2010, summer top left, Autumn top right, winter bottom left and spring bottom right. Measured in cm

3.1.5 Bathymetry

Bathymetry or bottom depth is the average depth from the ocean floor to a reference ellipsoid. This product was considered as a preference attribute as others have considered it for other tuna species (Schick et al., 2004). Depth was sourced from `https://koordinates.com/layer/1541-new-zealand-region-bathymetry/`. The resolution of this product was in $250m^2$ cells. Bathymetry was assumed to be constant through time.

The only manipulation was taking the mean bathymetry over 1° latitude and longitude bins (Figure 3.7).

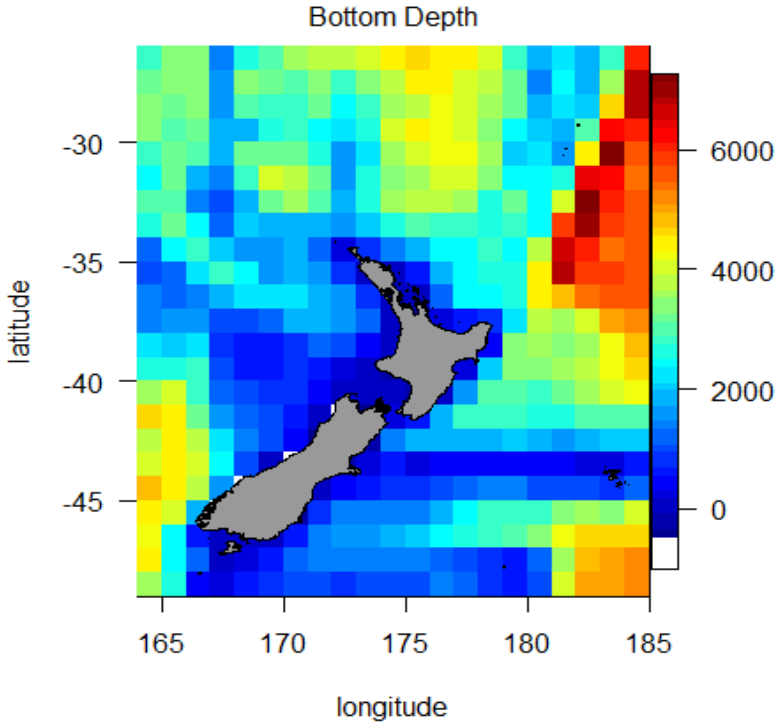

Figure 3.7: Bottom Depth of New Zealand,s EEZ, measured in meters below reference sea level.

3.1.6 Distance

Distance was the last preference attribute to be considered a preference attribute. Distance was included for simple biology and realism. Suppose there are cells with equal preference but one is $100km$ away and another is $1000km$ away. Then there should be a higher preference for cells that are closer, as less energy would be used for the same benefit. Distance was

calculated within SPM between cells based on square cells. The published version of SPM also uses Euclidean straight line distances (Dunn et al., 2015). These two assumptions create problems when a large range of latitudes are considered, and when islands are present in the area. These problems are addressed later in Chapter 4.4.

3.2 Commercial Fisheries Data

There are two potential sources of commercial fisheries data in New Zealand. First, data collected by observers on surface longline vessels are stored in the centralised observer database (*Cod*) administered by NIWA for the Ministry for Primary Industries (MPI). Sampling of individual surface longline fishing sets involve the observers identifying and counting all the catch during the time they are observing. Second, information on estimated catch and effort recorded by commercial fishers at the end of each trip, on the Tuna Longline Catch Effort Return (TLCER) forms is stored on the Quota Management System (QMS) database (*Warehou*) administered by MPI. For the current analysis, however, the cod database was used in preference to *Warehou*, because the TLCER forms underestimate bycatch and catches of non-quota species since much of it is discarded at sea and not recorded (Francis et al. 2000). In addition albacore tuna were only introduced into the QMS in 2007, so QMS *Warehou* data on the distribution and abundance of these species would likely not be comprehensive or reliable. The data available from *Cod* for this analysis comprised information from 6515 observed surface longline sets collected between 25 March 1994 and 28 August 2012. This was subsequently reduced to 3642 observations

when the dataset was truncated due to including NPP as a preference attribute (2003-2012).

The first data manipulation was to create a relative index of abundance by converting Catch, to Catch Per Unit Effort (CPUE). The catch recorded in the *Cod* dataset was number of individuals caught (abundance) per fishing event. Each fishing event was allocated to a cell defined by start latitude, longitude, and season for each year. The measurement of effort used was the number of hooks present on the longline. Each catch was divided by number of hooks, to get a CPUE of number of fish per hook. This helped standardise the catch data for effort, which is otherwise a source of bias in the raw catch data; That is the more hooks a fisher puts out, the higher their catch is likely to be. Often, other standardisations are carried out on CPUE (Maunder and Punt, 2004), which aim at removing other false signals from CPUE. No further standardisations were considered in this investigation to ensure spatial variability was not removed through standardisation of variables such as, duration or vessel that could alias for space. The relationship between estimated abundance $N_{t,k}$ and CPUE $I_{t,k}$ is expected to be proportional, given by,

$$I_{t,k} = qN_{t,k}e^{\epsilon_{tk}}, \tag{3.5}$$

where $N_{t,k}$ is abundance in cell k where $k = 1, 2, 3, ...RC$, at time step t, $I_{t,k}$ is the Catch Per Unit Effort, q is the catchability coefficient, and ϵ_k is the observational error. In the scope of this study q is a nuisance parameter; just a scaler used to compare expected ($N_{t,k}$) with observed ($I_{t,k}$) within

the model. Not all cells in the spatial state will have. This is a result of SPM constraining the spatial structure to be rectangle, and fishermen not applying effort over the entire spatial state in each time step, for economic reasons. This comes with a simplified assumption that the relative abundance index is proportional to the expected population abundance over space and time with associated observation error, this is a common assumption in fisheries science (Haddon, 2010, Maunder and Punt, 2004). Since q is not a time varying parameter and we have assumed a fixed population P, over the entire model run, this creates a potential problem of varying abundance by year. When there is higher abundance in New Zealand's EEZ, due to for example a high recruitment or external immigration from the greater South Pacific Ocean, this may increase the average catch rate for the year, which may create poor fitting in the model. This issue, referred to as year class effect, is addressed in the following section (Section 3.2.1. The following figure shows the spatial coverage of effort collapsed over time from the observed dataset used.

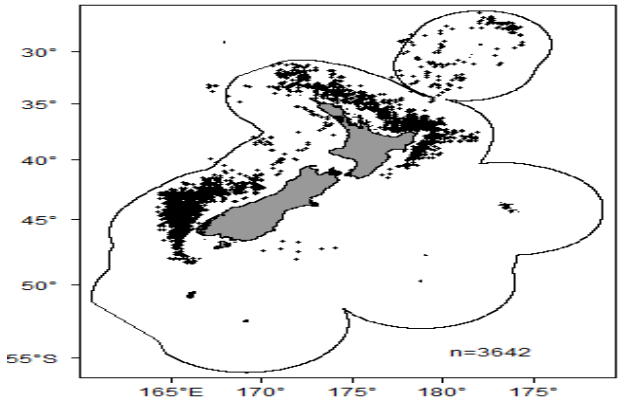

Figure 3.8: The spatial locations of observed effort in this dataset, between the years 2003-2012.

3.2.1 Handling Variability

Fisheries catch rates may provide snapshots of the relative abundance of particular areas at particular times of the year. Factors that affect the fishing operation (e.g. different gear, vessels, and area) and also the vulnerability of the fish to capture, such as schooling behaviour, and vertical movement, complicate interpretation of catch rates (Hilborn and Walters, 1992). Uncertainty in CPUE is incorporated into observations by taking the best point estimate of CPUE, and capturing the variation around these point estimates. Uncertainty is an important property, due to the large temporal and spatial range for a given cell (a single model cell spans three months (one Season) over an average area of 7920 km^2). In each cell potentially multiple observations need to be aggregated into a single observation for that cell k for a given time step t for use as a single observation within SPM, e.g. in summer of 2012 in the cell 40-41°S and 170-171°E there are 6 fishing events). For model fitting observations are weighted in SPM using the Coefficient of Variation (C.V).

$$c_{t,k} = \frac{\hat{\sigma_{t,k}}}{\hat{\mu}_{t,k}}, \tag{3.6a}$$

$$\hat{\sigma_{t,k}} = \sqrt{\frac{\sum_{j=1}^{n_{t,k}} (x^c_{t,k,j} - \hat{\mu_{t,k}})}{n_{t,k} - 1}}, \tag{3.6b}$$

$$\hat{\mu_{t,k}} = \frac{\sum_{j=1}^{n_{t,k}} x^c_{t,k,j}}{n_{t,k}} \tag{3.6c}$$

for $k = 1, 2, 3, ..., RC$, t is the time step, j denotes the j^{th} observation within cell k at time t, $x^c_{t,k,j}$ represents the observed CPUE for cell k at time t. $\hat{\mu_{t,k}}$

is our resulting single observation for that cell in time that SPM uses, and $c_{t,k}$ is the CV representing the uncertainty around that observation. The CV is used within SPM because it does not depend on the measurement unit, and it gives an indication of the relative size of the standard deviation with respect to the mean. For each cell (that contain at least one observation) the mean (Equation 3.6c) and CV (Equation 3.6a) was calculated. This meant that an *a priori* error was related to every observation prior to model run. These are used in the parameter estimation to weight the importance of observations. As the mean (Equation 3.6b) was used our observations will be sensitive to extreme values, an alternative could have been the geometric mean (Limpert et al., 2001), which is considered in discussion. Once all cells in all seasons where summarized, the problem of year class effect mentioned in section 3.2 was addressed. To account for this year class effect, all cells with observations within in each year were divided by the yearly CPUE mean $\bar{X}^c_y$. To demonstrate the standardisation of time to our observed data time steps t were described by seasons s and years y. This results in now in describing $x^c_{t,k}$ as $x^c_{s,y,k}$, where

$$x^c_{1,k}, x^c_{2,k}, x^c_{3,k}, x^c_{4,k}, x^c_{5,k}, x^c_{6,k}, ..., x^c_{50,k} = x^c_{1,1,k}, x^c_{2,1,k}, x^c_{3,1,k}, x^c_{4,1,k}, x^c_{5,1,k}, x^c_{1,2,k}, ..., x^c_{5,10,k}$$

$$\bar{X}_y^c = \frac{\sum_{s=1}^{4}\sum_{k=1}^{RC} x_{s,y,k}^c \, 1\{Year = y\}}{n_y}, \tag{3.7a}$$

$$\tilde{x}_{s,y,k}^c = \frac{x_{s,y,k}^c}{\bar{X}_y^c} \tag{3.7b}$$

$$= \tilde{x}_{t,k}^c, \tag{3.7c}$$

where $\tilde{x}_{s,y,k}^c$ is the year-standardised CPUE observation for cell k in season s in year y, n_y is the number of cells in year y that have observations, and $y \in \{2003, 2004, ..., 2012\}$. From now on time is indexed back as t for simplicity of working. The final problem addressed with our observed data was how zero catch observations were handled within SPM. This is discussed in the following section.

3.2.2 Dealing with Zero Observations

Given our assumed likelihood distribution (log normal see section 4.3) SPM has problems when observed data contains zero observations, this is where effort was applied in that cell but no albacore tuna were caught. Zero observations are not allowed due to the fact that within the likelihood we log transform the following $(I_{t,k}/qZ\,(N_{t,k})$ (see Equation 4.11). If the observed CPUE in cell k at time t ($I_{t,k}$) is a zero then the whole term becomes a zero and so the resulting likelihood transformation is $\log(0) = \infty$. There have been many arguments over whether absence data should be ignored or not; for reasons such as being unrepresentative when species are hard to catch, or whether they represent suitable habitat that is unoccupied (Elith

and Leathwick, 2009). Two options were investigated for dealing with this issue. First we can impute a constant for zeros that effectively is seen as zero within the model along, with a corresponding $c_{t,k}$. This will be investigated by plotting our year-standardised CPUE ($\tilde{x}^c_{t,k}$) versus its respective CV ($c_{t,k}$) to find a reasonable CPUE constant and $c_{t,k}$ to impute. Secondly, we ignore the zero catch observations. SPM is restricted in how small a value can be imputed for a zero observation ($\approx 1e^{-6}$). If an observation is less than this by an order of magnitude then it has a large effect on the likelihood; this was observed in the preliminary model runs. These two methods were described as potential methods to deal with catch observations in a review of CPUE analysis (Maunder and Punt, 2004). Other methods these authors suggested were using the delta method, or zero inflated models. SPM is not set up to apply these latter methods, hence why we are using the first two methods. Figure 3.9 shows the relationship between mean CPUE and $c_{t,k}$, both log transformed because SPM estimates in log space (so this is how SPM views the relationship).

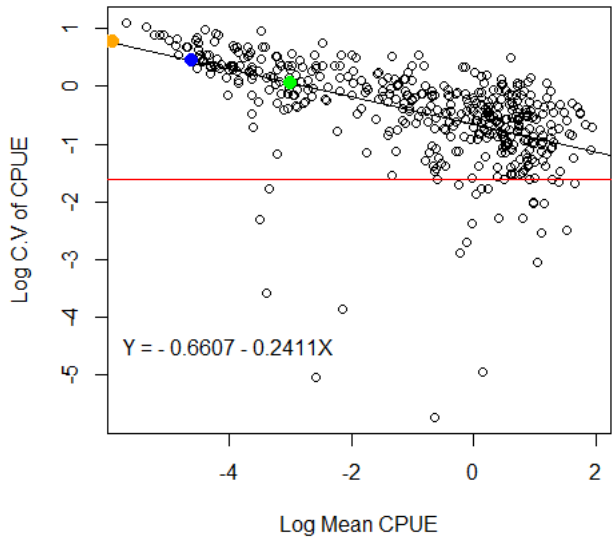

Figure 3.9: Log of $c_{t,k}$ versus log of CPUE ($I_{t,k}$), red line corresponds to a $c_{t,k}$ of 0.2 and the blue dot represents initial imputation value. Orange and green dots represent imputations for sensitivity analysis

The red line in Figure 3.9 represents an assumed cut off point that an observations $c_{t,k}$ was allowed to take. This was done because a previous study has shown that $c_{t,k}$ for fishing is typically around 0.2 (Francis, 2000), and it also stops these observations having a large influence on the likelihood. Observations with small $c_{t,k}$'s have more weight on the model because they are considered closer to the truth. In doing this correction, we are assuming that commercial fishing events must have at least this $c_{t,k}$. The blue dot was chosen as the initial imputation value for zero. This is because it is at the tail end of the observations, so signifies very small abundance whilst sill within the linear trend. This value, and $c_{t,k}$ were used as the imputed value for zero in model runs. It was thought that

including small values for zeros would be more important than ignoring them due to the information they represent (edge effects). Sensitivity tests were conducted to see if the model is sensitive to this constant and corresponding $c_{t,k}$. Three sensitivity tests were run: two were other imputation values either side of the blue dot (orange and green Figure 3.9), and the last was with the zero observations removed. The sensitivity were conducted on the final selected model. The observed data as input to the model with imputed zeros, are shown in the following Figures 3.10, 3.11 and 3.12.

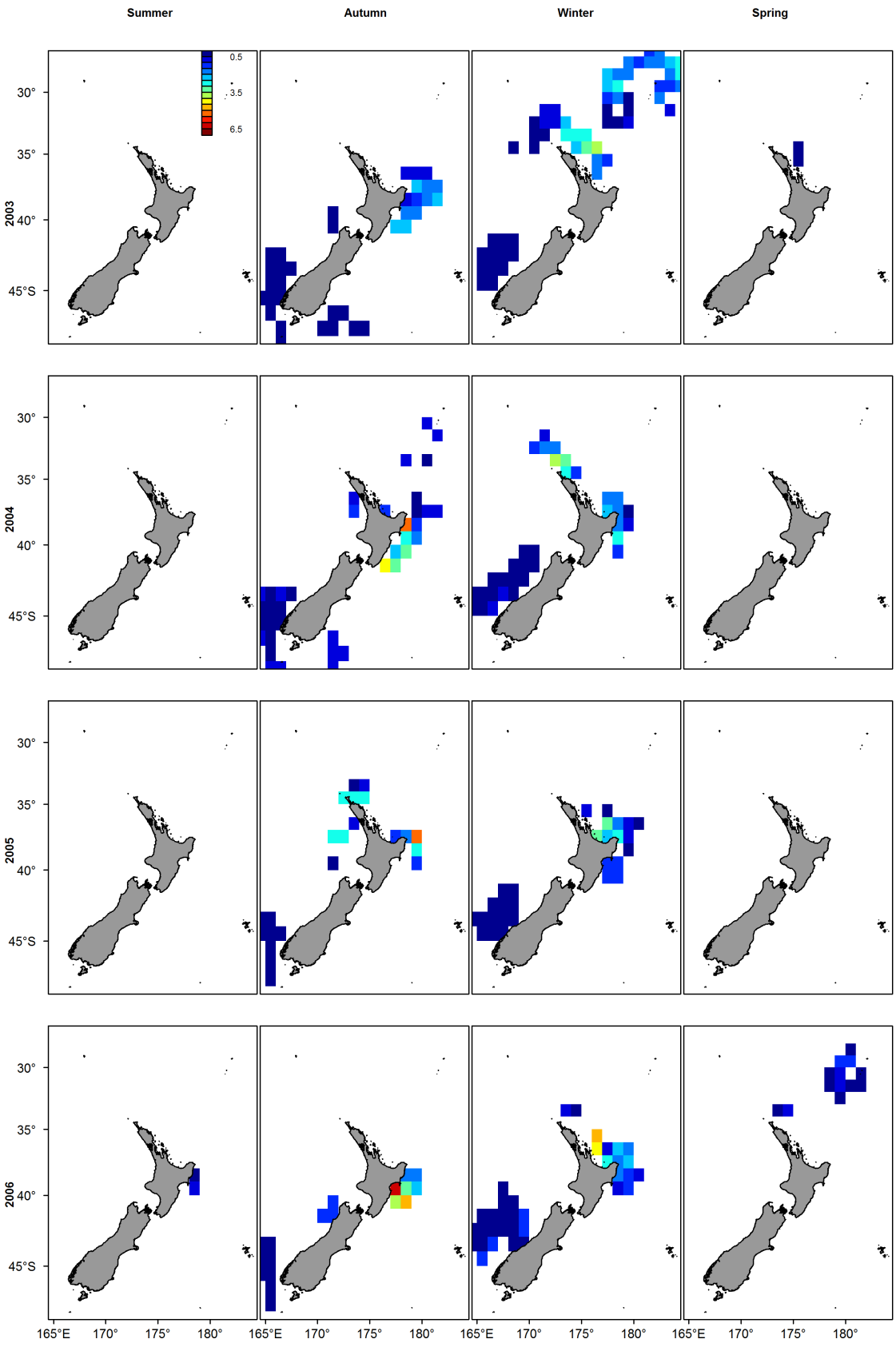

Figure 3.10: Observed albacore standardised CPUE for the seasons between the years 2003-2006

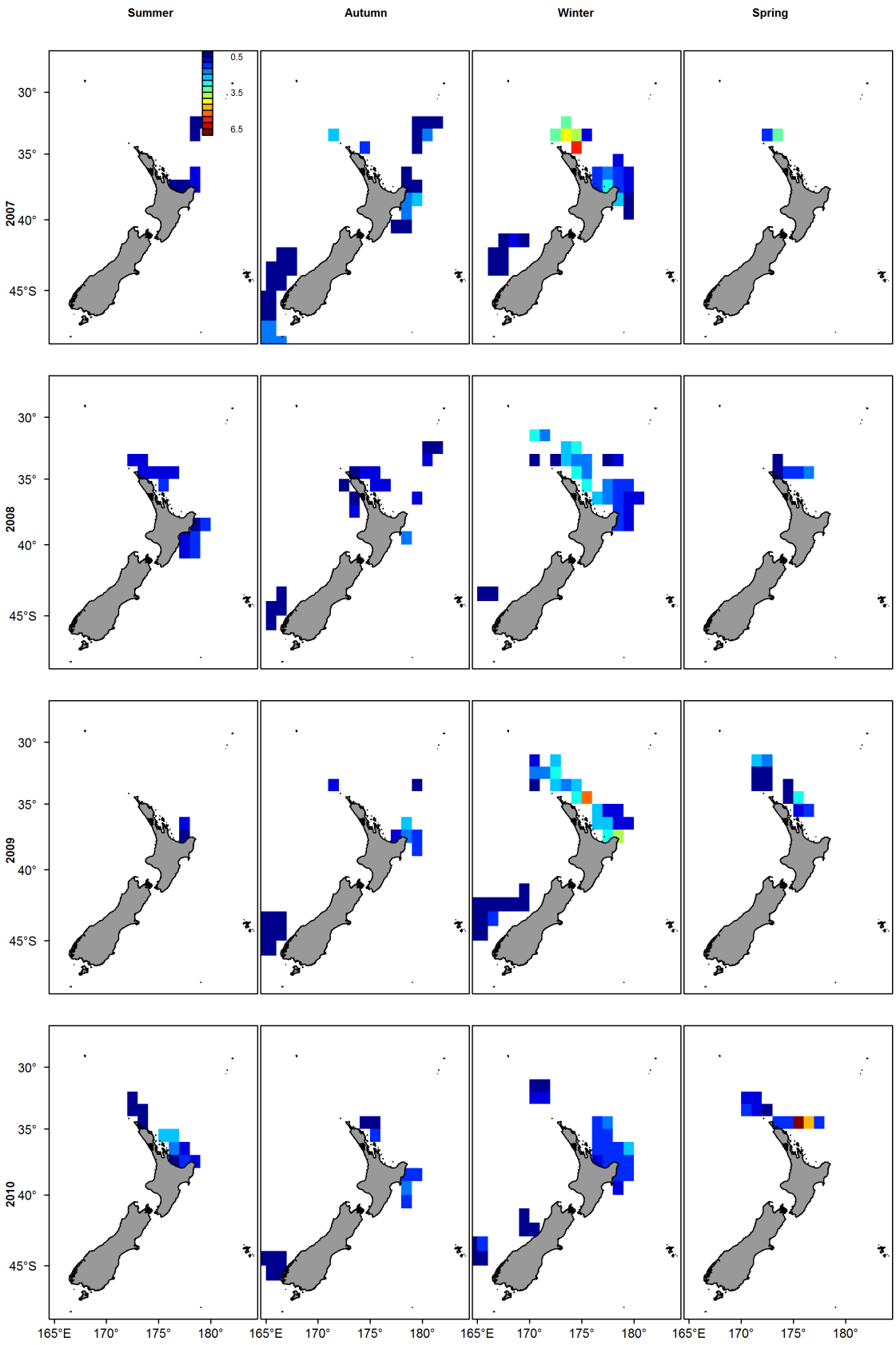

Figure 3.11: Observed albacore standardised CPUE for the seasons between the years 2007-2010

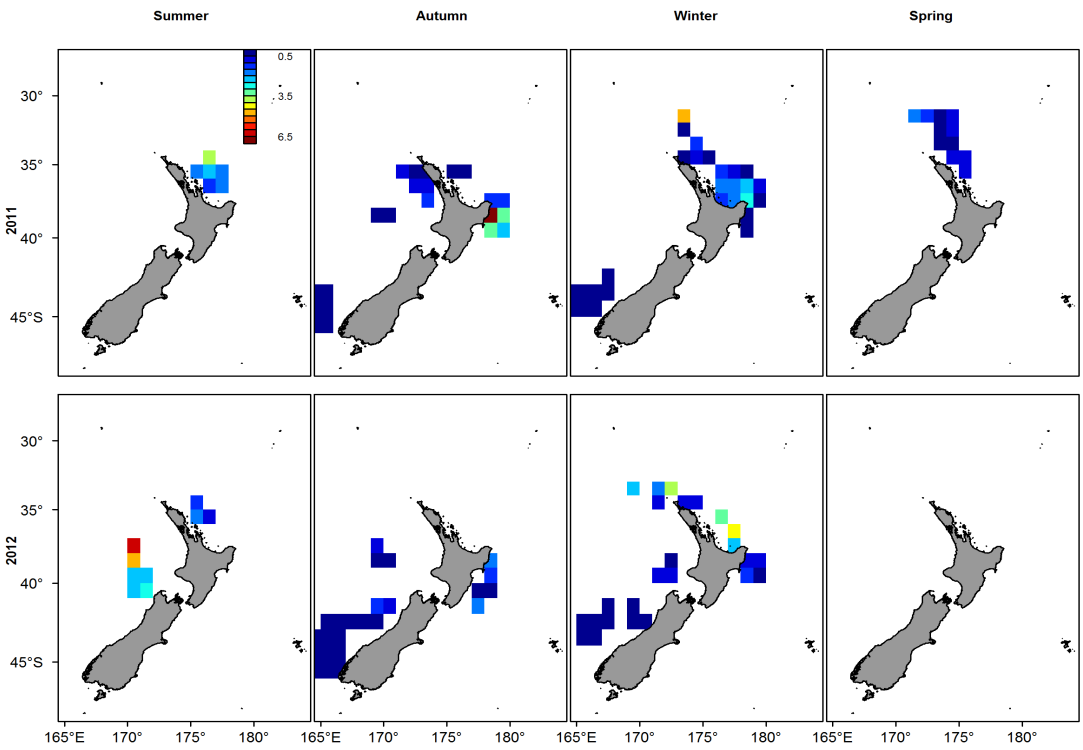

Figure 3.12: Observed albacore standardised CPUE for the seasons between the years 2011-2012

The equivalent figure, but with $c_{t,k}$ instead of CPUE, is shown in the Appendix A.2. From Figures 3.10, 3.11 and 3.12 we see that spatial distribution of catch varies over time, with observations predominately occurring along the east coast of New Zealand. In years 2003-2005 little or no effort was applied in summer and spring. The area south west of New Zealand is dominated by little CPUE values. There is hints of a latitude effect, for example in 2011 summer there is high observed CPUE at 35°S, in autumn it drops down to 40°S. Then in winter the high CPUE is north at 31°S and in spring no high catch rates.

3.3 Candidate preference functions

The concept of spatial movement within SPM is based on populations being distributed based on preferred physical or environmental conditions. This section aims at investigating if there are potential relationships between the preference attribute dataset and the observed year-standardised CPUE dataset. The preference functions ($f(x;\boldsymbol{\phi})$) available within SPM that connect population preference to environmental attributes are listed below,

- The Normal preference function for preference attribute x has parameters $\boldsymbol{\phi} = (\mu, \sigma)$, defined as,

$$f(x;\mu,\sigma) = 2^{-[(x-\mu)/\sigma]^2} \tag{3.8}$$

- The double-Normal preference function for preference attribute x has parameters $\boldsymbol{\phi} = (\mu, \sigma_L, \sigma_R)$, defined as,

$$f(x;\mu,\sigma_L,\sigma_R) = \begin{cases} 2^{-[(x-\mu)/\sigma_L]^2}, & \text{if } x \leq \mu \\ 2^{-[(x-\mu)/\sigma_R]^2}, & \text{if } x \geq \mu \end{cases} \tag{3.9}$$

- The Logistic preference function for preference attribute x has parameters $\boldsymbol{\phi} = (a_{50}, a_{to95})$, defined as,

$$f(x;a_{50},a_{to95}) = 1/[1 + 19^{(a_{50}-x)/a_{to95}}] \tag{3.10}$$

  where $x \in (-\infty, \infty)$

- The inverse-Logistic preference function for preference attribute x has parameters $\boldsymbol{\phi} = (a_{50}, a_{to95})$, defined as,

$$f(x; a_{50}, a_{to95}) = 1 - 1/[1 + 19^{(a_{50}-x)/a_{to95}}], \tag{3.11}$$

  where $x \in (-\infty, \infty)$

- The Exponential preference function for preference attribute x has parameter $\boldsymbol{\phi} = (\lambda)$, defined as,

$$f(x; \lambda) = \exp(-\lambda x), \text{where } x \geq 0 \text{ and } 0 \text{ otherwise} \tag{3.12}$$

These preference functions are graphically demonstrated in Figures 3.13 and 3.14 with arbitrary parameter values.

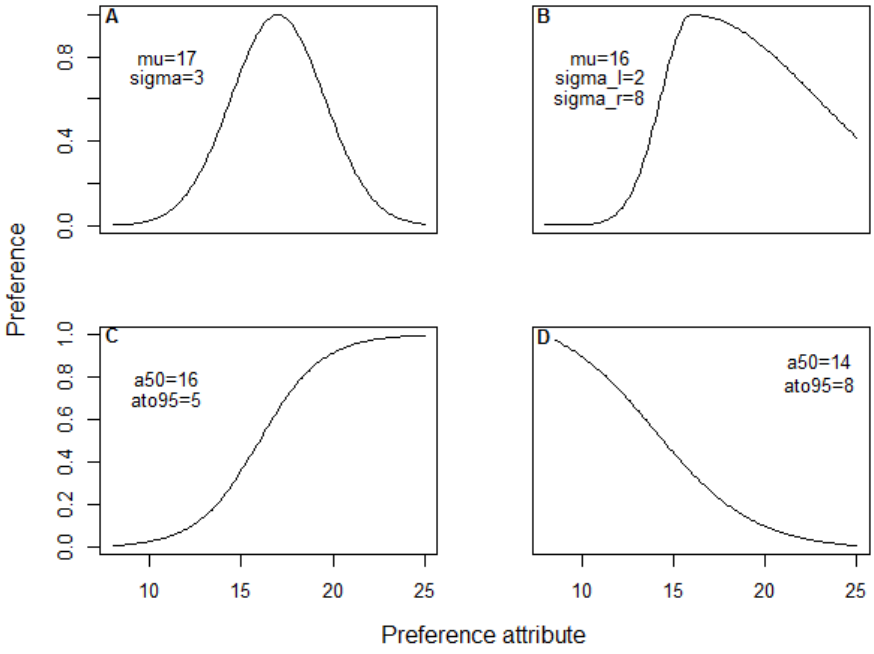

Figure 3.13: Demonstrating functional shape of preference functions in SPM, A= Normal, B= Double Normal, C= Logistic, D= Inverse Logistic

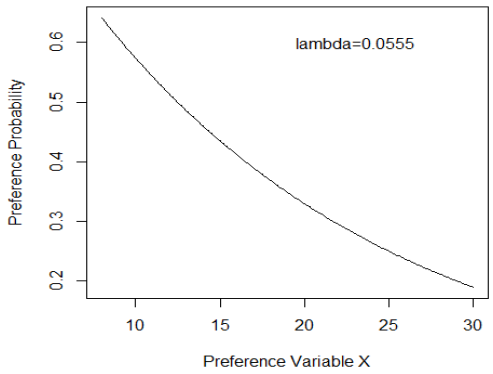

Figure 3.14: Exponential preference function, with parameters Lambda = 0.0555

The potential functional forms for each preference attribute were restricted based upon visualisation of the data (Figures 3.15 3.14). This was done to reduce the permutations of functional forms and preference attributes considered when fitting the model.

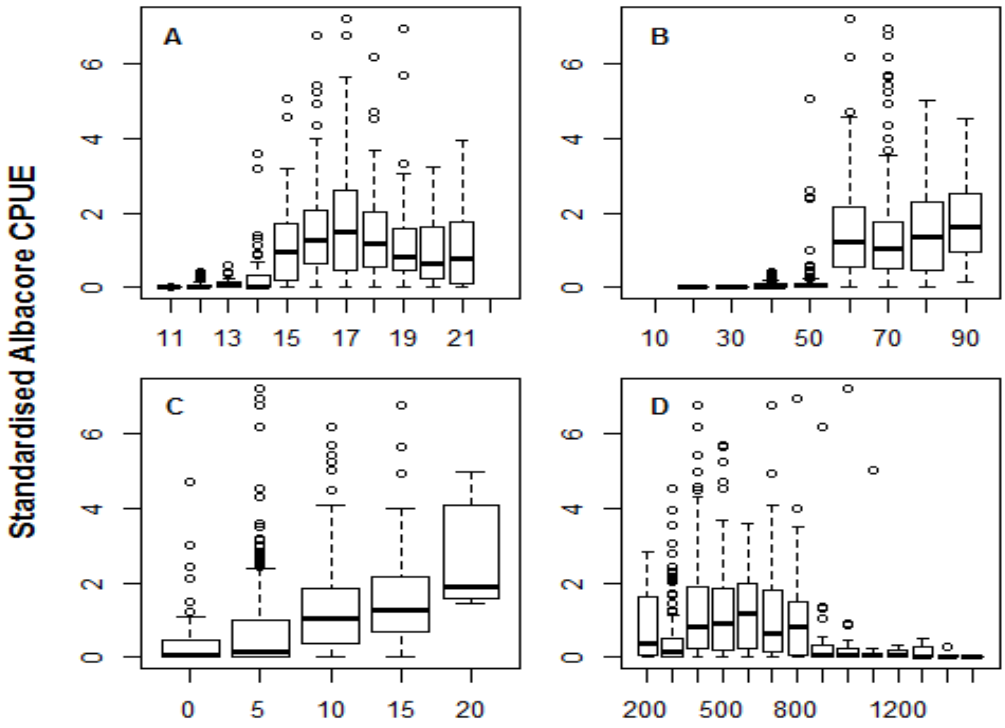

Figure 3.15: Boxplots of observed standardised albacore CPUE with preference attributes, A= CPUE vs SST$^{\circ}C$, B= CPUE vs SSH cm, C= CPUE vs MAG $\frac{cm}{km}$, D= CPUE vs NPP $mg\ C/m^2/d$

For SST the functional forms that were considered for model selection were the Normal, Double normal, and Logistic functions. This was due unimodal or S-shaped curves being functional shapes (Figure 3.15 panel **A**). SSH and MAG showed a positive and increasing trend with CPUE,

and the considered forms were therefore logistic, normal, or double normal. NPP showed a negative relationship with CPUE, on average. There was also a potential mode at value of 700 $mg\ C/m^2/day$. The candidate forms were therefore the normal, double normal, inverse logistic, and exponential.

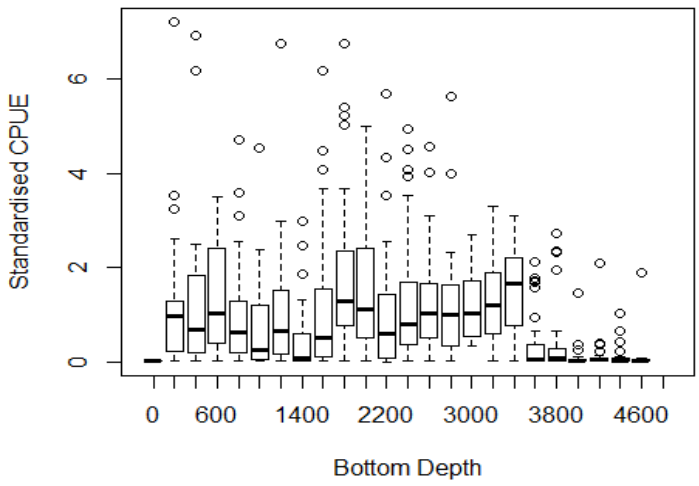

Figure 3.16: Boxplots of observed standardised albacore CPUE with depth preference attributes

Depth (Figure 3.16) showed a fairly uniform relationship with CPUE from $100m$ to $3500m$, and then a decrease after $3500m$. As a result, inverse logistic and exponential were the only forms considered for the variable depth. There was a preference attribute that cannot be visualised, which is distance. Biological reasoning was used to constrain the functional form. This was based on energy, where cells further away from the origin were considered less preferable, so only decaying functions (Exponential and

Inverse Logistic) were considered.

If attributes are highly correlated, the parameters describing them can also be correlated. This can lead to a problem of non-identifiability in the parameters estimates (Li and Vu, 2013). To avoid this, highly correlated attributes were not considered in model selection. Correlations were identified by scatter-plot matrix of preference attributes (Figure 3.17).

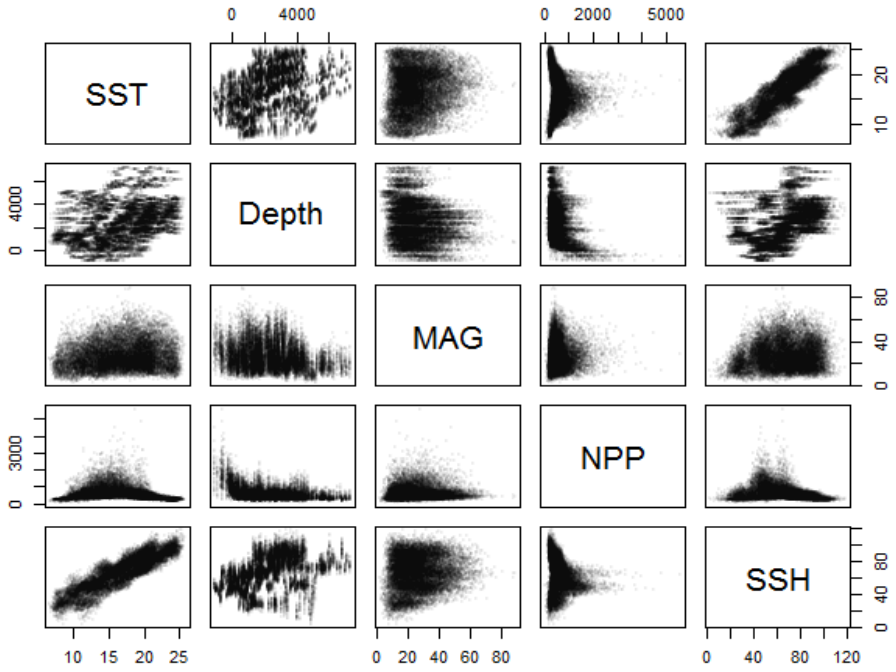

Figure 3.17: Scatter plot matrix of all preference attributes described in Section 3.1.

From Figure 3.17 we see that the only two variables that are highly correlated ($r > 0.8$) were SST and SSH. Thus these are the only two preference attributes that were not considered in the same model.

Chapter 4

Model assumptions, estimation and evaluation

Mathematical models represent estimated versions of the real world, with simplifications. Simplifications require assumptions. The following Chapter describes some assumptions within SPM, and how we evaluated them, as well as describing the estimation and model selection criteria used in this thesis.

4.1 Independence of preference attributes

As noted earlier (section 2.2) SPM calculates the joint preference by multiplying the marginal preference functions of interest (Equation 2.3). This assumes that those preference functions are acting independently. Although SPM has an assumed independence structure between preference variables, it has the ability of capturing relative importance of preference variables via the weighting factor (α in Equation (2.3)), which can be esti-

mated. The different effects that alpha has on a bivariate model are shown in Figure 4.4.

The aim of this section was to find, and describe, alternative joint preference distributions for spatial movement. Multidimensional (joint) distributions can be complex, and most of the time unknown. SPM's current set up adds difficulty, due to the different functional forms that marginal preference functions can take. There exists a technique in the literature called the Copula that allowed us to explore the joint distribution. Copulas have been applied in survival analysis (Clayton, 1978, Oakes, 1989), finance (Cherubini et al., 2004, Patton, 2006) and actuarial science (Frees and Valdez, 1998).

The idea behind copulas is that most of the time we have knowledge of marginal distributions, and have an idea about the dependence structure. The copula is the method of "copuling" or linking the marginal distributions with an assumed dependence structure, to describe the joint distribution. The benefits of this method are that, the user is not restricted to having marginal distributions from the same probability family (e.g. multivariate normal), and the copula method has the ability to describe the joint distribution in any number of dimensions. For this thesis, we will be restricting all joint distributions to the bivariate case, for ease of demonstration and application.

In this study we relax the assumption of independent preference functions. We evaluate alternative dependence structures on a predefined range of probability marginal distributions using copulas. There is one other change needed to apply this method compared to the current approach in

SPM. That is, in order to use the copula theory we need to use true probability distributions, where $f(\boldsymbol{x};\boldsymbol{\phi})$ the Probability Density Function (PDF) integrates to one, that is

$$\int_{-\infty}^{\infty} f(\boldsymbol{x};\boldsymbol{\phi})dx = 1 \tag{4.1}$$

The above property does not hold for current continuous preference functions (section 3.5), due to preference functions being reinterpreted to have biological meaning, that is

$$\int_{-\infty}^{\infty} f(\boldsymbol{x};\boldsymbol{\phi})dx \neq 1$$

When applying copulas within SPM, true probability distributions were used that captured the same functional shapes to those in current use, new code was required and written during this thesis. A particular problem arises when trying to find probability density functions that are similar in shape to the logistic and inverse logistic preference functions. These functional forms do not decrease at both tails, and this results in,

$$\int_{-\infty}^{\infty} f(\boldsymbol{x};\boldsymbol{\phi})dx \to \infty$$

Due to interest of time and illustration four continuous probability distributions functions where coded into SPM and these did not include logistic and inverse logisitic (see Table 4.1). For a more technical description of copulas theory and an example of code (implemented in C++), see Appendix A.4 and A.5. The table of continuous distributions that are used in subsequent analysis are in Table 4.1,

Table 4.1: Table of densities used on aopulas demonstration

Distribution	Probability density function
Normal	$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp - \left(\frac{(x-\mu)^2}{2\sigma^2}\right)$
Log Normal	$f(x;\mu,\sigma) = \frac{1}{x_i\sigma\sqrt{2\pi}} exp - \left(\frac{(\log x-\mu)^2}{2\sigma^2}\right)$
Exponential	$f(x;\lambda) = \lambda exp(-\lambda x)$
Uniform	$f(x;a,b) = \frac{1}{b-a}$

The normal and exponential distributions have similar preference functions in SPM. This makes these two probability distributions directly comparable to SPM preference functions. The uniform PDF was used to be mimic the constant preference function, the difference being that the uniform distribution has parameters to estimate, whereas the constant preference function does not. Code for the log-normal distribution was added for future use, but was not eventually used in this study.

In the bivariate case, the joint distribution for two random variables X_1 and X_2 with known marginal densities $f_1(x_1;\phi_1)$ and $f_2(x_2;\phi_2)$ is given by

$$f_{x_1,x_2}(x_1,x_2;u_1,u_2,\phi_1,\phi_2,\phi) = c(u,v;\phi)f_1(x_1;\phi_1)f_1(x_1;\phi_1) \tag{4.2}$$

where, $c(u_1,u_2;\phi)$ is a copula density representing the structure of dependence between X_1 and X_2 with $0 \leq u_1,u_2 \leq 1$, and $f_{x_1,x_2}(x_1,x_2;u_1,u_2,\phi_1,\phi_2,\phi)$ is the bivariate joint distribution of X_1 and X_2, more detail is given in Appendix A.4. There were four assumed dependence structures used in

this investigation. The copula probability density distribution with different dependence structures and their respective formulas are shown in Table 4.2.

Derivation of the gaussian copula is shown in appendix A.4.2. This study applied the gaussian, gumbel, and frank copula. The clayton copula was not implemented in SPM in time for analysis, and the independence copula is equivalent to SPM current set up if α in Equation 2.3 were equal to one, for the bivariate case. The independence copula was implemented in SPM for comparison to the existing preference, this involved checking estimated parameters converged to the same estimates with equivalent functional forms with the original preference method. This was satisfied and so the independence copula was ignored for the rest of analysis. Independence refers to the original SPM preference setting from now on. The copulas chosen in this thesis are commonly applied in the literature (Chen and Fan, 2006, Li, 1999), they also cover a range of functional shapes. Definition 4.1.1 defines dependence relationships that are used to describe how each copula distribution changes compared to the independence structure. Positive quadrant dependence (PQD) describes the dependence where large values of one random variable tend to be associated with large values of the other random variable, and vice versa with negative dependence.

Table 4.2: Table of bivariate copula dependence structures and their formulas

Name	Bivariate copula density $c(u_1, u_2; \phi) f_1(x_1; \phi_1) f_1(x_1; \phi_1)$
Gumbel	$e^{\left\{-\left[(-ln(u_1))^\phi + (-ln(u_2))^\phi\right]^{1/\phi}\right\}} \frac{(-ln(u_1))^\theta - 1}{u_1} \frac{(-ln(u_2))^\phi - 1}{u_2} \left\{\left[(-ln(u_1))^\phi + (-ln(u_2))^\phi\right]^{\frac{2}{\phi}-2} + (\phi - 1)\left[(-ln(u_1))^\phi + (-ln(u_2))^\phi\right]^{\frac{1}{\theta}-2}\right\} f_1(x_1; \phi_1) f_2(x_2; \phi_2)$
Frank	$\phi e^{-\phi u_1} e^{-\phi u_2} \left[(e^{-\phi} - 1) + (e^{-\phi u_1} - 1)(e^{-\phi u_2} - 1)\right]^{-1} \left\{(e^{-\phi u_1} - 1)(e^{-\phi u_2} - 1)((e^{-\phi} - 1) + (e^{-\phi u_1} - 1)(e^{-\phi u_2} - 1))^{-1} - 1\right\} f_1(x_1; \phi_1) f_2(x_2; \phi_2)$
Clayton	$\frac{(1+\phi)}{(u_1 u_2)^{\phi+1}} \left[\frac{(u_1 u_2)^\phi}{u_1^{-\phi} + u_2^{-\phi} - (u_1 u_2)^\phi}\right]^{\frac{1+2\phi}{\phi}} f_1(x_1; \phi_1) f_2(x_2; \phi_2)$
Gaussian	$\frac{1}{2\pi\sqrt{1-\rho}} exp\left(-\frac{\xi_1^2 - 2\rho\xi_1\xi_2 + \xi_2^2}{2(1-\rho^2)}\right) f_1(x_1; \phi_1) f_2(x_2; \phi_2)$
Independence	$f_1(x_1; \phi_1) f_2(x_2; \phi_2)$

Definition 4.1.1 *Quadrant dependence:*
(Lehmann 1966) Let X_1 and X_2 be random variables. X_1 and X_2 are positively quadrant dependent if for all (x_1, x_2) in $\boldsymbol{R^2}$

$$P[X_1 \leq x_1, X_2 \leq x_2] \geq P[X_1 \leq x_1], P[X_2 \leq X_2] \tag{4.3}$$

or,

$$P[X_1 > x_1, X_2 > x_2] \geq P[X_1 > x_1], P[X_2 > x_2] \tag{4.4}$$

where $\boldsymbol{R^2}$ is the two dimensional space of real numbers. Negative quadrant dependence (NQD) is defined analogously by reversing the sense of the inequalities. PQD says the probability that random variables are jointly large is greater than when they are looked at independently.

The Frank copula in Table 4.2 and Figure 4.1 can describe both PQD and NQD. When $\theta > 0$ it displays PQD, and when $\theta < 0$ it displays NQD, as shown below, and independence copula represents lack of dependence. For illustration purposes of copulas, two marginal distributions were arbitrarily chosen, the distribution on the vertical axis of subsequent plots (X2) is an exponentially distributed variable with parameter $\lambda = 0.0033$, and the distribution on the horizontal axis (X1) is a normally distributed variable with parameters $\mu = 21$ and $\sigma = 4$,

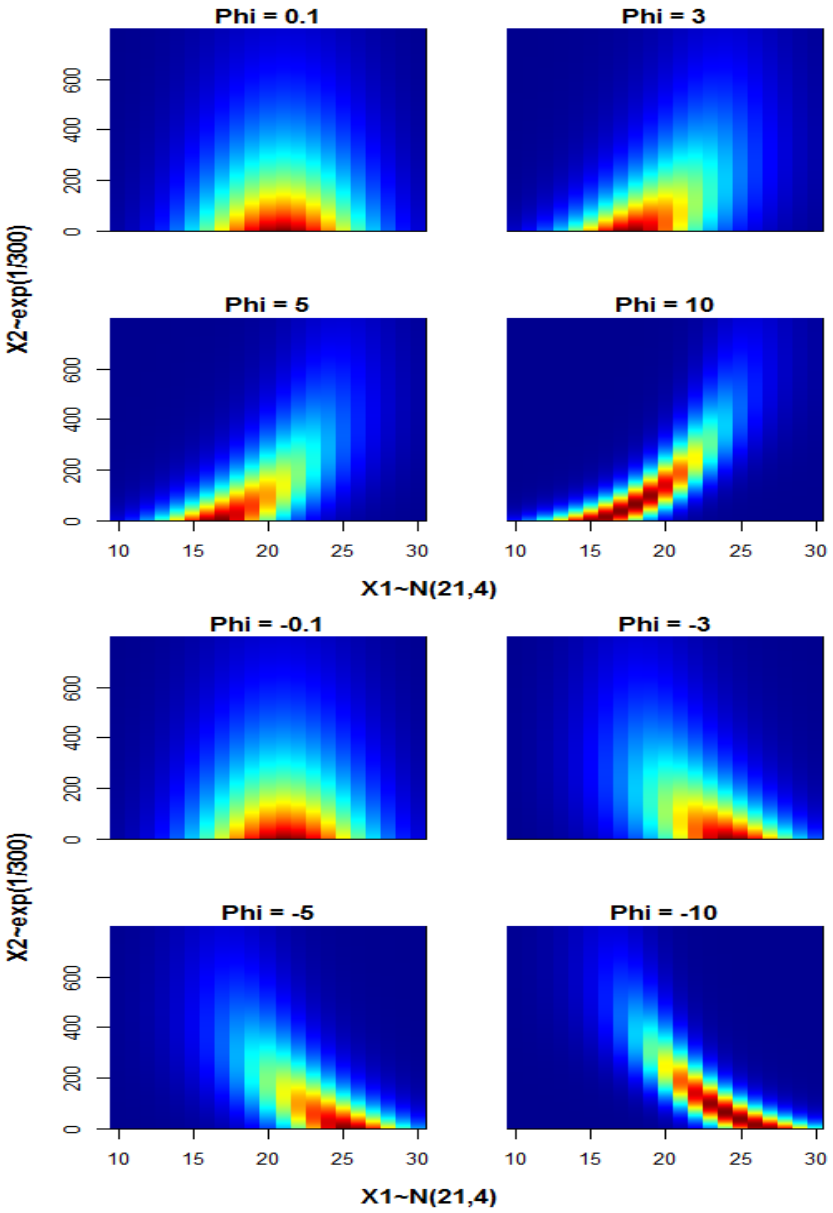

Figure 4.1: The Frank copula showing a range of forms controlled by ϕ indicated by top of panel. Key: red color represents high probability at that (x_1, x_2) coordinate, blue represents low probability for that coordinate

The Gumbel copula (Table 4.2,Figure 4.2) cannot account for negative dependence but is well suited for the case when there is strong right tail

dependence (Strong association at high values) but weak left tail dependence (weak association at low values) Figure 4.2.

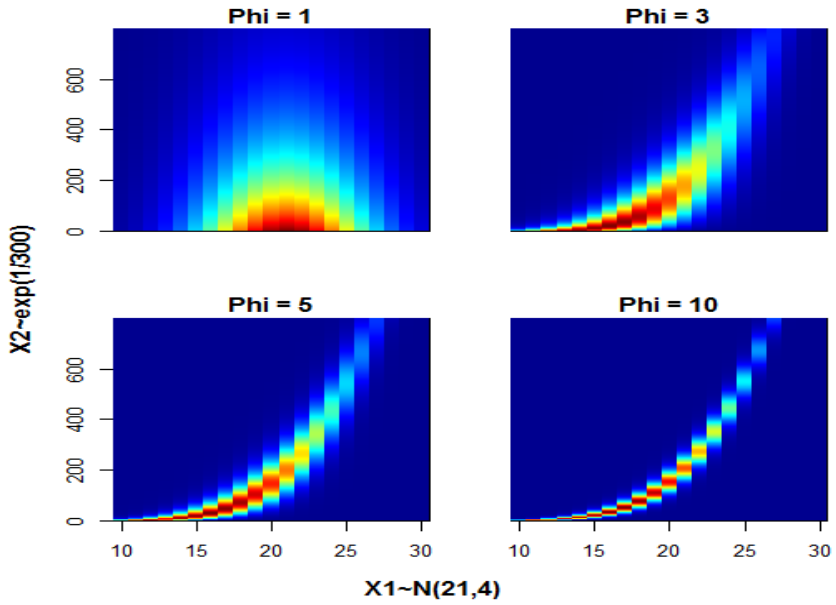

Figure 4.2: The Gumbel copula with varying ϕ. Key: red color represents high probability at that (x_1, x_2) coordinate, blue represents low probability for that coordinate

There is also the Gaussian copula, an elliptical copula which captures the full range of positive and negative dependence. The Gaussian Copula is described as having radial symmetry and is elliptical shaped naturally (Figure 4.3).

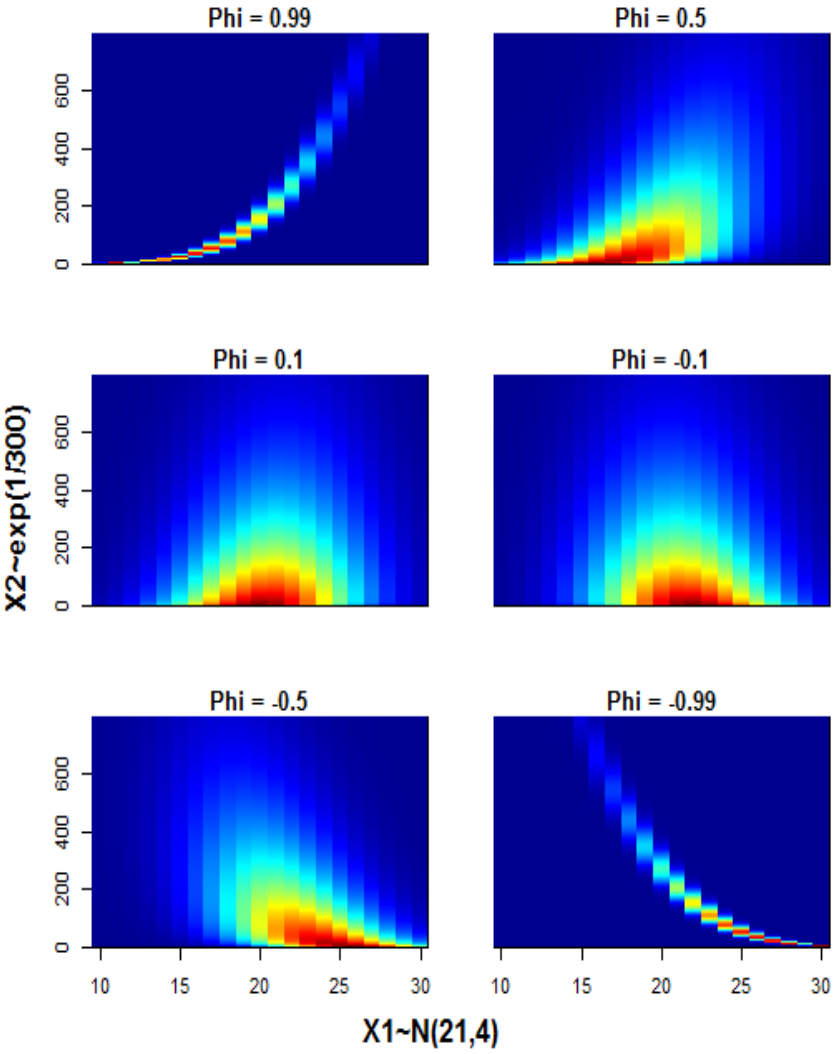

Figure 4.3: The Gaussian copula with varying ϕ. Key: red color represents high probability at that (x_1, x_2) coordinate, blue represents low probability for that coordinate

The following figure shows the limitations of SPM existing independent preference function, where the alpha is the parameter that affects the relative weight of each function (Equation 2.3),

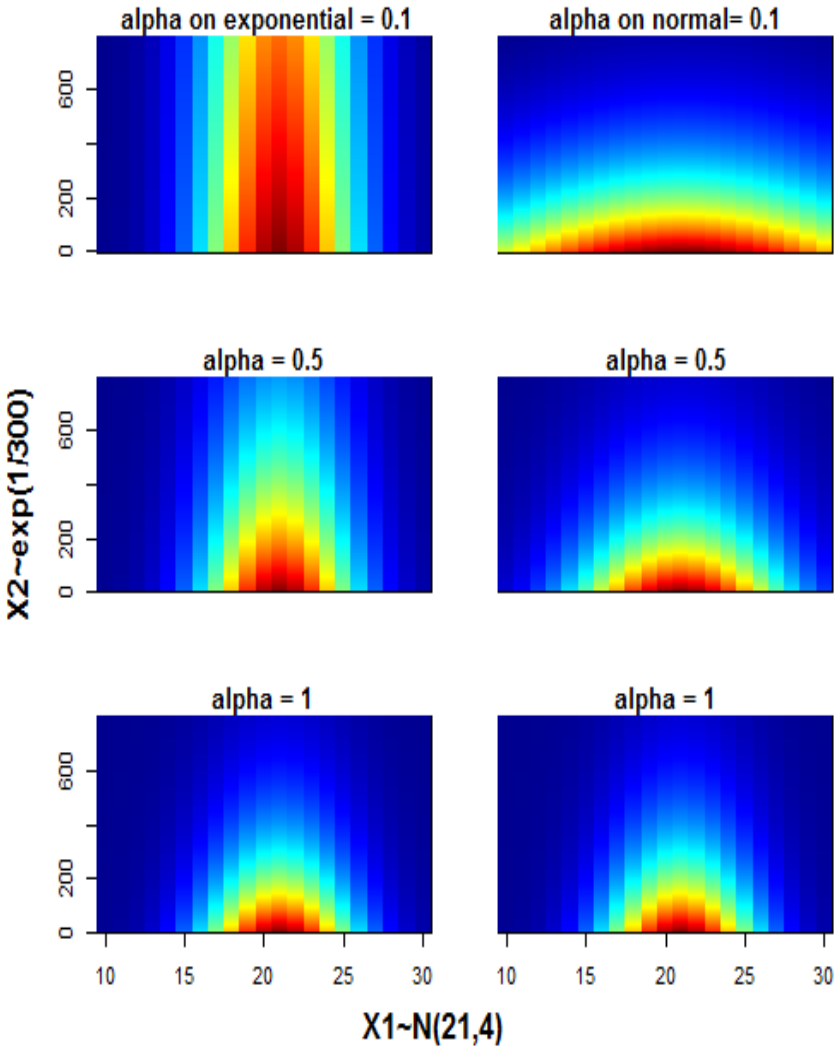

Figure 4.4: The range of forms a bivariate preference function in SPM can take with the same preference functions as demonstrated in the copula figures. Left panels are when alpha is on the exponential function (alpha =1 for normal function), and vice verse for right hand panels.

Figure 4.4 shows how the independence structure is not as flexible as other copula forms. The bottom panels in Figure 4.4 represent true independence (alpha =1 for both functions), where as the top left panel is dominated by $X1$ variable and vice versa for the top right panel. Figure 4.4 is

a reference to compare the structural shapes of copula.

Due to the restrictions on available marginal distributions within this framework, the model runs with the original SPM preference function were rerun with functions restricted to Normal and Exponential forms when finding the best bivariate model. This provides a comparison between the independent (existing SPM) and the copula technique. All dependence structures were evaluated relative to using the best restricted independent model. How the best model was identified, and model fit were compared, is described in section 4.3.

4.2 Dealing with Distance in SPM

An important preference layer in SPM, which is not considered an environmental variable but is considered important, is distance. Distance is a measure of how far away a potential target cell is from the cell a fish is in, measured in kilometers. Distance is important in the sense that it adds realistic properties to the model. SPM calculates a distance layer before the model is run, but this is based on square cells, and calculates straight line Euclidean distance.

This method inherits two problems. The first is that the world is a spherical/Geoid shape, and in SPM it assumes space is a perfect rectangle, made up of square cells. This is problematic given latitude and longitude cells get narrower towards the poles, due to the spherical chape of the world. This can create a bias when calculating distances between cells, especially over large ranges of latitudes, and as we get closer to the poles. Figure 4.5 shows how SPM currently assumes cell structure (left panel)

compared to reality (right panel).

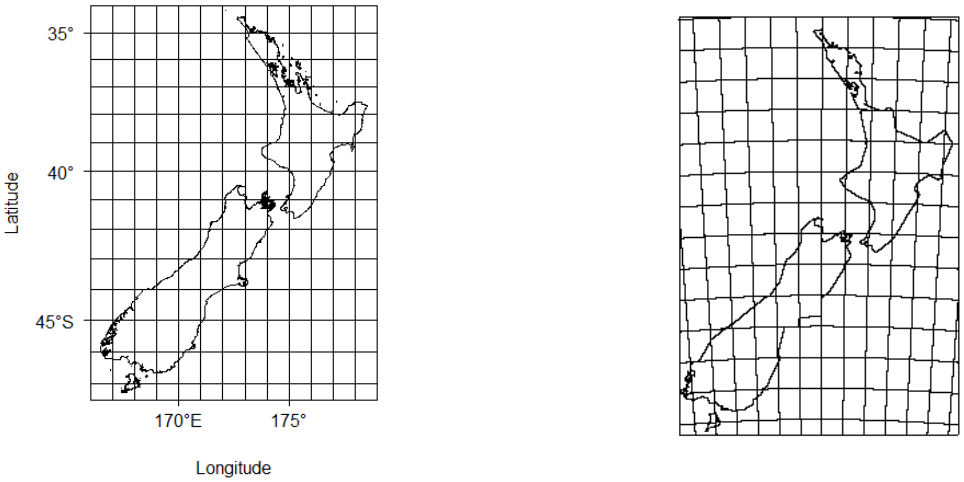

Figure 4.5: Left panel shows how SPM assumes cell structure to be. Right panel shows the real world cell structure (Note the narrowing of the cells in southern latitudes)

The more southerly a cell is, the narrower the cell becomes. A cell at $29°S$ has the approximate length and width of 111.12km by 97.19km, resulting in an area of 10799.5km^2, where as a cell at $55°S$ has the dimensions 111.12km by 63.73km, resulting in an area of 7081.678km^2. SPM currently ignores this attribute of the world and assumes each cell is a perfect symmetric square. This has the potential of penalising lateral movement of fish in the southern cells if the existing distance function is used.

The second problem inherited by SPM's existing method pronounced for this particular study area, is there are islands in the middle of our spatial state. Suppose we wanted to calculate the distance from the top right cell to a cell in the bottom left. SPM currently calculates it "as the crow

flies" or straight line Euclidean distance as shown in the Figure 4.6,

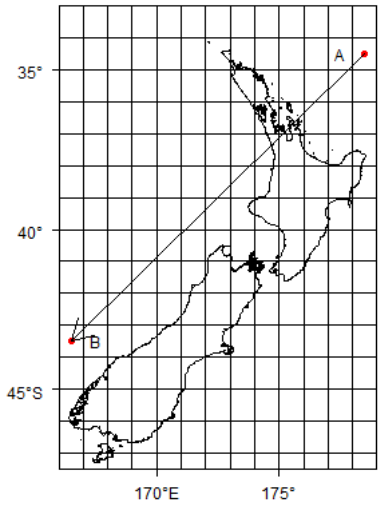

Figure 4.6: Demonstrating how SPM calculates the distance from one cell (A)to another (B) direction given by arrow

This is obviously a problem, because due to land, fish cannot physically swim this route. They either have to go around the North Island, through the Cook strait, or around the bottom of the South Island.

These two problems cause misrepresentations of the true distance layer, so here I proposed to fix both these potential problems at the same time. This is done by running a shortest path algorithm from a source cell to all other cells, whilst taking into account the curved surface of the world. The shortest path problem is a frequently tackled problem in graph theory, where it is commonly known as the "Travelling sales man problem",

that is "A salesman is required to visit each of the n given cities once and only once, starting from any city and returning to the original place of departure. What route, or tour, should he choose in order to minimise the total distance traveled?" (Lin, 1965). The same methodology was applied to solve the shortest path problem in this thesis. This problem is conceptualised as a graph. Graphs consist of nodes (cell mid points) connected by branches (distance connecting adjacent nodes). A shortest path algorithm will find the shortest path from a source node to a target node, whilst considering all connecting pathways (nodes and branches). Treating our spatial state as a graph, where cells are nodes and the distance between them as branches, adds flexibility of allowing land cells to be turned off, thus distance can be calculated by traveling around. This method also allows us to calculate the distances between cells using geodetic functions. Geodetic functions calculate the distance between two coordinates on the earth's curved surface in terms of straight line distances. The structure of our graph has nodes for all cell midpoints that a fish can visit (i.e. only ocean), and branches calculated from each node to all adjacent nodes calculated by a geodetic function, these branches connect up the nodes to create the graph. The geodetic function that was implemented to calculate distances between adjacent cells was Haversine's Formula,

$$a = (\sin(\nabla lat/2))^2 + \cos(lat_1) \times \cos(lat_2) \times (\sin(\nabla lon/2))^2 \tag{4.5a}$$

$$c = 2 \times a \tan 2(\sqrt{a}, \sqrt{1-a}) \tag{4.5b}$$

$$D = R \times c, \tag{4.5c}$$

where D is the distance between two coordinates (lon_i,lat_i, where $i = 1, 2$), R is the radius of the earth. lon_i is the longitude and lat_i latitude for coordinate i, a and c are functions that adjust the distance for a curved surface.

This geodetic function assumes that the earth surface is a perfect sphere (which it isn't it is slightly ellipsoid with a bulge around the equator). There is more than one function that can be used for calculating distances over curved surfaces. The Haversine function was chosen due to it being the computationally simplest, whilst not loosing too much accuracy($< 0.3\%$) compared to an alternative, such as Vincenty inverse formula for ellipsoids. Vincenty inverse function would yield more precise distances but comes at a computational cost (see `http://jsperf.com/vincenty-vs-haversine-distance-calculations`). For this reason we chose the computationally simplest method (Haversine).

The issue of curved surface is addressed in the construction of the graph, when distances between cells were calculated. The last component of the new distance function applies a shortest path algorithm across the graph to find the shortest path from source cell to target cell. This solved the issue of islands, by not allowing land cells to be nodes in the graph essentially a closed off area is formed where land is. This forces the algorithm to find the shortest path by going around these closed off areas. The algorithm I chose to use, was Dijkstra's algorithm (Dijkstra, 1959). There are other algorithms for solving the shortest path problem, such as Bellman and Fords algorithm (Bellman, 1958, Ford and Lester, 1956) and Gabow's algorithm (Gabow, 1983). I chose to implement Dijkstra's algorithm for its trade off of simplicity and speed. Dijkstra's algorithm starts

from the source cell and assesses all paths from source cell. With each node it visits calculating the cumulative distance to get there. If two separate paths reach the same node it will choose the one with the smallest distance traveled. It does this, till all cells have been reached. A hypothetical shortest path is demonstrated to compare how this new distance function would calculate the distance given the same problem in Figure 4.6,

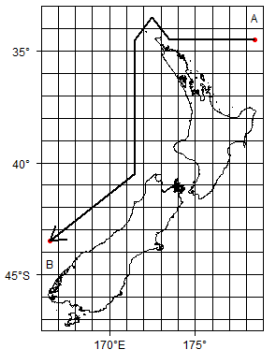

Figure 4.7: Demonstrating how Dijkstra's algorithm would calculate the distance from one cell (A) to another (B) direction given by arrow

In model selection, distance cannot be considered a preference variable for univariate selection. This is due to the conditional nature of distance i.e. distance is calculated from where you are. This means that after the initialisation phase the population would not move because of the decaying preference function forced on the variable (i.e higher preference to stay). For this reason it was only considered for model selection after the first preference attribute had been selected.

Adding distance as a preference attribute will slightly modify the original preference equation Equation 2.3. This is because distance is condi-

tional on where you are, so changes from calculating a preference value for all cells to calculating a preference cell based on where you are,

$$p_k \rightarrow p_l = f_1(x_{dl}; \boldsymbol{\phi}_1)^{\alpha_1} \times f_2(x_{2l}; \boldsymbol{\phi}_2)^{\alpha_2} \times ... \times f_n(x_{nl}; \boldsymbol{\phi}_n)^{\alpha_n} \tag{4.6}$$

where $p_k \rightarrow p_l$ is the preference value of moving from cell k to cell l and $f_1(x_{dl}; \boldsymbol{\phi}_1)^{\alpha_1}$ is the distance function, where,

$$x_{dl} = g(x_k, x_l) \tag{4.7}$$

where $g()$ is the Dijkstra's shortest path using Haversine geodetic function between x_k and x_l. This describes the preference probability of movement, but SPM uses relative preference probability for the set of all cells from cell k so,

$$r_{lt} = \frac{p_{lt}}{\sum_{j=1}^{RC} p_{jt}} \tag{4.8}$$

where $\{p_{jt} : j = 1, ..., RC\}$ is described as the set of preference probabilities possible from source cell k including cell k and l at time step t. Population in cell l at time t that move from cell k is.

$$N_{lt} = N_{kt} r_{lt} \tag{4.9}$$

Adding this new method within SPM improved the distance function making it a more realistic assumption of the world. This gives users more confidence in applying distance as a preference attribute.

4.3 Model Estimation and Selection

4.3.1 Parameter estimation

From section 2.3 the only processes applied in this study, were movement processes controlled by parameters ϕ (Equation 2.7). These parameters can be fixed or estimated within SPM, by optimising a likelihood objective function or Bayesian estimation. This study focuses on the former method. This requires assuming a likelihood function which links the observations to the process model,

$$L(\boldsymbol{\phi}; \boldsymbol{N}, \boldsymbol{X}) = \prod_{t}^{T} \prod_{k}^{RC} f(N_{t,k}; X_{t,k}, \boldsymbol{\phi}) \tag{4.10a}$$

$$logL(\boldsymbol{\phi}; \boldsymbol{N}, \boldsymbol{X}) = \sum_{t}^{T} \sum_{k}^{RC} logf(N_{t,k}; X_{t,k}, \boldsymbol{\phi}) \tag{4.10b}$$

where T is the number of time steps, $N_{t,k}$ is the abundance in time step t in cell k, L is the likelihood, $logL$ is the log likelihood, and RC is the total number of cells. Alternatively a negative can be taken of the likelihood to get a negative likelihood $-logL$ in which minimisation can be used to find parameters values. Parameters estimates are values of $\boldsymbol{\phi}$ that minimise $-logL$ or maximise L.

The likelihood assumed in this study for observation $I_{t,k}$ and expectation $N_{t,k}$, is based on a lognormal distribution. Derivation of the likelihood and log-likelihood function is given in Appendix A.3. After dropping out constant terms from the negative log likelihood, we get the following ob-

jective function that is passed to the optimiser

$$-\log L = \sum_{t}^{T} \sum_{k}^{RC} \left(\log\left(\sigma_{tk}\right) + 0.5 \left(\frac{\log\left(I_{t,k}/qZ\left(N_{t,k},\delta\right)\right)}{\sigma_{t,k}} + 0.5\sigma_{t,k} \right)^2 \right) \tag{4.11}$$

where $Z\left(\gamma,\delta\right)$ is a robustifying function to prevent division by zero errors, with parameter $\delta > 0$. $Z\left(\gamma,\delta\right)$ is defined as,

$$Z\left(\gamma,\delta\right) = \begin{cases} \gamma, & \text{where } \gamma \geq \delta \\ \delta/\left(2-\gamma/\delta\right), & \text{otherwise} \end{cases} \tag{4.12}$$

The default value of δ was 1×10^{-11}.

where $-\log L$ is the objective function. The likelihood takes into account the *a priori* uncertainty of observations through $c_{t,k}$, which can be calculated from data, as in this investigation, or assumed constant. In addition to observation error, SPM can estimate process error for observations. Additional process error has the effect of increasing the observation error in the data, and hence decreasing the relative weight given to those data in the fitting process. In this case process error is added to the $c_{t,k}$ and the resulting combined error ($c'_{t,k}$) is specified as,

$$c'_{t,k} = \sqrt{c^2_{t,k} + c^2_{process\ error}}, \tag{4.13}$$

where $c^2_{process\ error}$ can be estimated within SPM. This was ignored in this study and the observations variability was assumed to be accurately represented by their corresponding observation error.

The lognormal likelihood was chosen over the alternative (Normal)

due to the nature of CPUE data. CPUE typically have many small catch rates, and few very large catch rates, creating skewed distributions. This is a common error distribution applied with fishery abundance error (Haddon, 2010, Robson, 1966), others have log transformed CPUE and assumed gaussian error distribution (Fujii, 2015, Gavaris, 1980).

Equation 4.11 describes how the likelihood objective function is derived for fixed parameters ϕ. How SPM estimates ϕ is via optimisation algorithms that search over the multidimensional objective surface to find a global minimum that satisfies the best fit to the data, for a given model structure and parameters. SPM has two optimizers that can be used to change parameter values to search over the likelihood surface to find global minima. The first uses a quasi-Newton minimiser, which is a slightly modified implementation of the main algorithm of Dennis Jr. & Schnabel (Dennis Jr and Schnabel, 1996), while the second uses a genetic algorithm developed by Storn & Price (Storn and Price, 1995), the differential evolution minimiser. At the beginning of model selection both optimisers were used for each model run. As a result of the differential evolution poor performance it was dropped and model selection was done solely based on quasi-Newton minimiser. Poor performance was concluded from model run time and number of non converged model runs.

During estimation, preference parameters are restricted to physically meaningful ranges (i.e the range observed in the preference attributes for this investigation, see section 3.3), this restricted optimisers from searching outside of these limits.

4.3.2 Model assessment

Model selection involved finding the 'best' combination of preference attributes with preference functions as well as estimating the optimal values for the parameters. The 'best' model in the context of this thesis is defining the most informative model, whilst considering parsimony, given the information available.

To find the 'best' preference model a forward selection procedure was conducted. First, to find the best univariate model and then the best bivariate model, and so on. With each additional preference attribute added, we compare to the previous model to observe the differences. How competing models were compared is by a combination of; Akaike Information Criterion (AIC), summarised residual fit, and parameter estimates. AIC (Akaike, 1974) is a useful model comparison criteria as it takes into account goodness of fit whilst penalising for over parameterising. AIC is defined as,

$$AIC = 2(-\log L + K) \tag{4.14}$$

where $-\log L$ is the objective function (Equation 4.11), and K is the number of estimated parameters in that given model.

Although AIC is a good overall indication of competing model performance, it does not give information on goodness of fit. Once a best model has been chosen via AIC, goodness of fit is assessed through the use of residual patterns. Normalised residuals ($r_{t,k}$, see Equation 4.15) for observation $I_{t,k}$ were used to evaluate model fit. There are many residuals (a residual for every observation in Figures 3.10, 3.11, and 3.12) and as a result of the layout and visual presentation of 40 panels of residual plots, it

is very difficult to identify patterns in residuals, especially when comparing competing model fits. For this reason, the same set of residuals were visualised on three axes; latitude, longitude, and year. This gave an idea of fit through time and space without trying to analyse patterns from 40 maps. Normalised residuals were defined as,

$$\sigma_{t,k} = \sqrt{\log(1 + c_{t,k}^2)} \tag{4.15a}$$

$$r_{t,k} = (log(I_{t,k}/N_{t,k}) + 0.5\sigma_{t,k}^2)/\sigma_{t,k}, \tag{4.15b}$$

where $I_{t,k}$ is CPUE, $N_{t,k}$ is expected abundance, $c_{t,k}$ is the coefficient of variation for that CPUE observation, and $r_{t,k}$ is the normalised residual for cell k in time step t.

If parameter estimates ran off to parameter bounds (beyond the plausible attribute space) during optimisation, then that preference function was deemed an unacceptable option for that attribute, and another function was considered. Precision of parameter estimates will be investigating by obtaining the margin of error MOE_j for parameter j at a 95% confidence level, using the square root of the inverse Hessian matrix $S.E_j$ for parameter j and assuming central limit theorem applies.

$$MOE_j = 1.96 \times S.E_j \tag{4.16}$$

Convergence was also evaluated for each model run. Confirming model convergence is always a potential issue when dealing with multi-parameter models. Convergence was deemed satisfied when a global minimum was found. In reality we will never really know if the global minimum was

found. How convergence was evaluated was by changing starting values of parameters three times for each model run randomly over the attribute space. If all three converge then a global minimum was assumed. If, however, there were different ending parameter values, then model parameters with the lowest likelihood were used as the starting values for an additional model run. If this run converged to the starting estimates, then convergence was assumed; otherwise, the run with the lowest likelihood was deemed satisfactory. This method was used as a trade off between accuracy and computational time (where some models take a day to converge). Parameter profiles were run on the final models to confirm the location of the minimum likelihood for each parameter.

The first model run was the Null model, that is a model with uniform spatial distribution. This model was the initial point for forward model selection.

The process of forward model selection was,

1. Model each attribute sequentially
2. For each attribute try functional forms described in exploratory analysis (section 3.4)
3. Pick the best attribute and its best functional form, based on previous criteria
4. Fix the functional form of the best univariate attribute along with setting the weighting factor (α) equal to one. Test the remaining preference attributes, and their respective functional forms. If an additional preference attribute improves model performance, fix it and its functional form.

5. Continue until no more attributes remain, or no improvement can be found (AIC), or parameters become unidentifiable.

After the final model was selected, an analysis of the residuals was conducted using a Generalised Additive Model (GAM) (Hastie and Tibshirani, 1990). This models residuals with the explanatory variables; latitude, longitude, and time (defined as year). This was used as an analytical method for checking the zero mean assumption of residuals through time and space. This method also has the benefit of identifying any trends in residuals. These trends could give insight into additional preference attributes.

Chapter 5

Results

In this chapter we present model performance and comparison from the final models in the forward model selection process.

5.1 Null Model

The first model run was the spatial null model. This model has an implicit assumption of uniform spatial distribution over the state (equal abundance in each cell). The following Figures 5.1-5.3 show summarised normalised residuals for time, longitude and latitude.

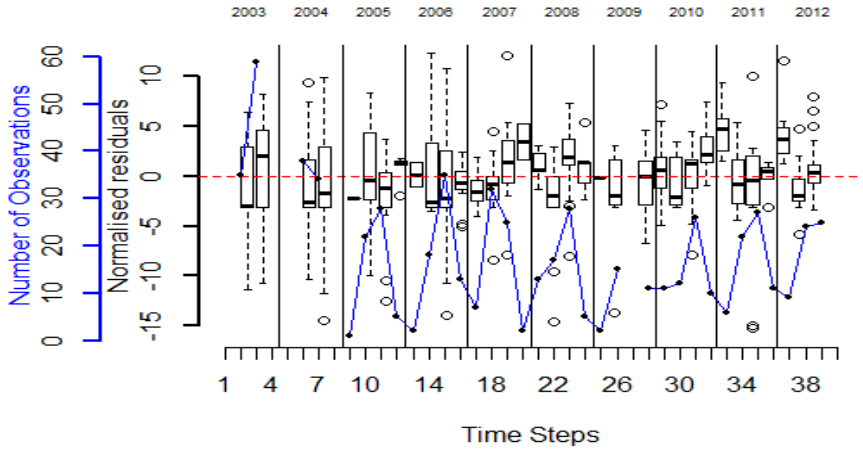

Figure 5.1: Boxplots of normalised residuals summarised into each time step, Blue legend is number of observations in each time step, the dashed red line indicates 0, and the black legend represents the value of the residuals

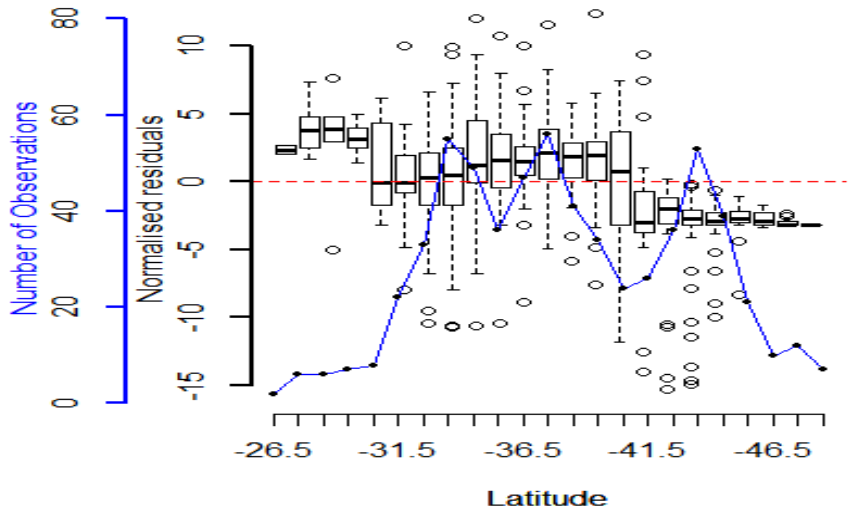

Figure 5.2: Boxplots of normalised residuals summarised into latitude bins, Blue legend is number of observations in each latitude bin, the dashed red line indicates 0, and the black legend represents the value of the residuals

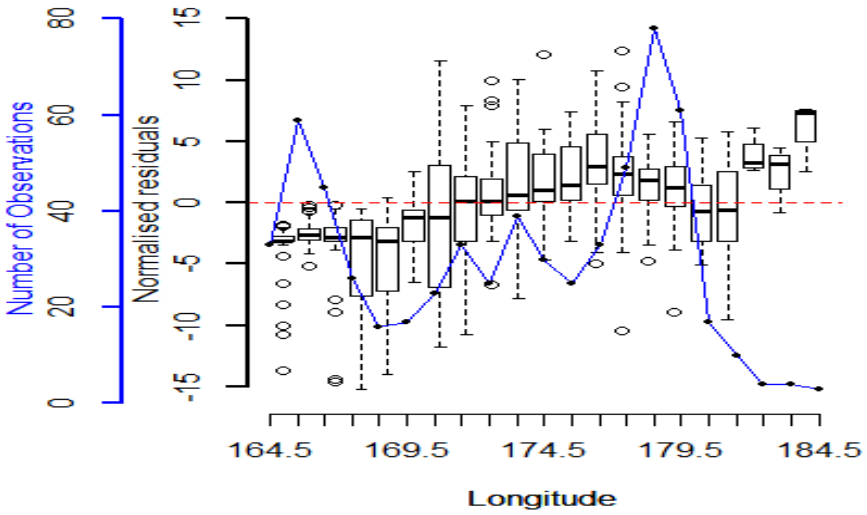

Figure 5.3: Boxplots of normalised residuals summarised into longitude bins, Blue legend is number of observations in each longitude bin, the dashed red line indicates 0, and the black legend represents the value of the residuals

The Null model shows a variable fit over time, 2004 shows the poorest fit relative to other years, where medians are well below zero. The Null model fails to fit well in the tails of both latitude and longitude (Figures 5.2– 5.3). The null model over estimates at high latitudes and under estimates at lower latitudes. There are fewer observations on both tails. This is expected since latitudes tails are at fringes of fisheries area. Longitude is underestimated in the western longitudes and over estimated in the eastern longitudes. The objective function for the null model was 4948.21, with an AIC of 9898.42. The next model was fitted with a univariate preference attribute, with the aim of better capturing of the spatial variability in CPUE

5.2 Univariate Model

Table 5.1: Univariate model summary Table

Attribute	Function	$-LogL$	Params	AIC
Null Model	uniformt	4948.21	1	9898.42
SST	Normal	3344.85	3	6695.7
	Double Normal	4043.66	4	8095.32
	Logistic	3132.47	3	6270.94
SSH	Normal	4971.29	3	9948.58
	Double Normal	5032.25	4	10072.5
	Logistic	4952.19	3	9910.38
MAG	Normal	4363.21	3	8732.42
	Double Normal	4349.91	4	8707.82
	Logistic	4441.16	3	8888.32
NPP	Normal	4832.45	3	9670.9
	Double Normal	4648.46	4	9304.92
	Inverse Logistic	4929.72	3	9865.44
	Exponential	4921.6	2	9847.2
Depth	Inverse Logistic	4858.55	3	9723.1
	Exponential	119153	2	238310

Based on AIC, the best model was with the preference attribute SST (Table 5.1). This is with a logistic preference function (Figure 5.4. With the final estimates from Table 5.2 we get the estimated functional form shown in Figure 5.4. Table 5.2 also shows small margin of error for the parameters suggesting high confidence in the parameter estimates. Normalised residuals are shown in Figures 5.5–5.7 for assessing goodness of fit.

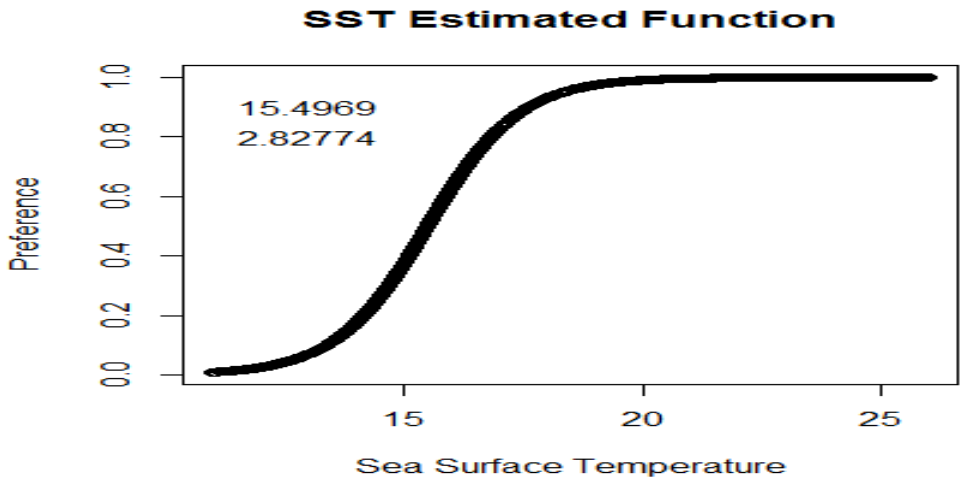

Figure 5.4: Estimated univariate preference for Albacore towards SST

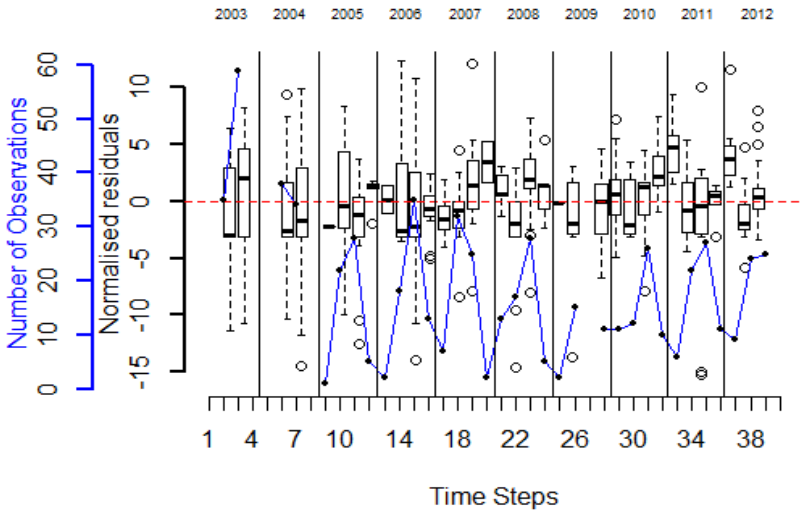

Figure 5.5: Boxplots of normalised residuals summarised into each time step, Blue legend is number of observations in each time step, the red line indicates 0, and the black legend represents the value of the residuals

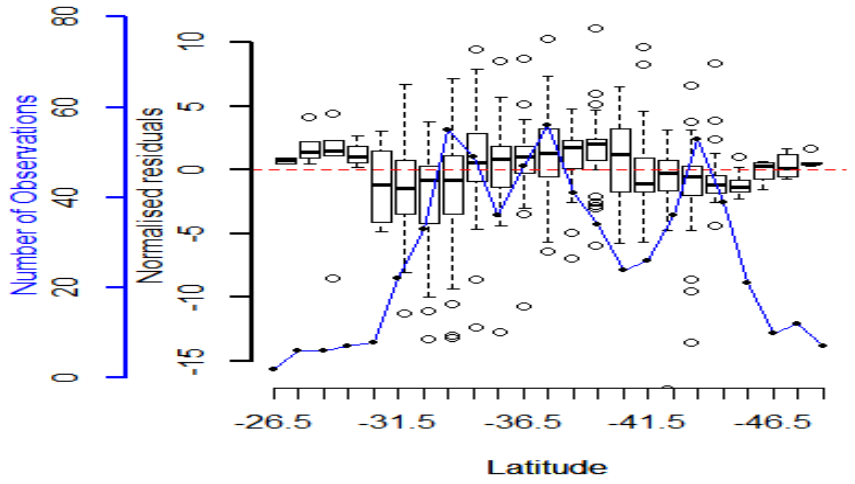

Figure 5.6: Boxplots of normalised residuals summarised into latitude bins, Blue legend is number of observations in each latitude bin, the red line indicates 0, and the black legend represents the value of the residuals

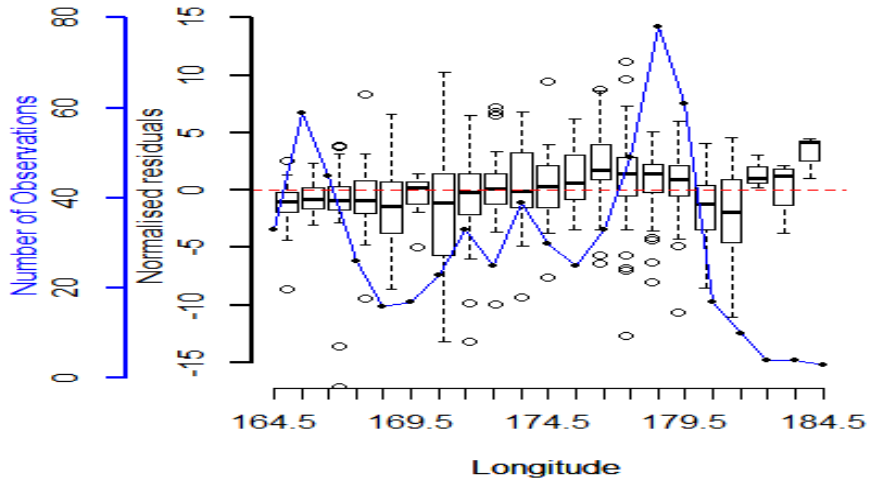

Figure 5.7: Boxplots of normalised residuals summarised into longitude bins, Blue legend is number of observations in each longitude bin, the red line indicates 0, and the black legend represents the value of the residuals

The univariate model with SST had, in general more closely scattered residuals around zero, compared to the null model. Residuals over time steps still show variability, most of the poorest fit time steps have small number of observations. Spatial fits were improved, most substantially in the extremes of both longitude and latitude bins (Figures 5.7 5.6). Although in general the residual fit was better there were still evident non-linear patterns in the residuals e.g. still a wave in latitude. The parameter estimates and corresponding margin of error shown in Table 5.2,

Table 5.2: Final univariate parameter point estimates with corresponding margin of error

Parameter	Estimate	Margin of Error
q	0.00029	$1.67182e^{-11}$
SST_{a50}	15.5	$5.79177e^{-03}$
SST_{ato95}	2.83	$6.68105e^{-03}$

The margin of error from Table 5.2 for all parameters is quite small, suggesting precise estimates.Profiles were run on the final parameter in the univariate SST logistic model (Figure 5.8). All profiles show clear minimuma over the SST range.

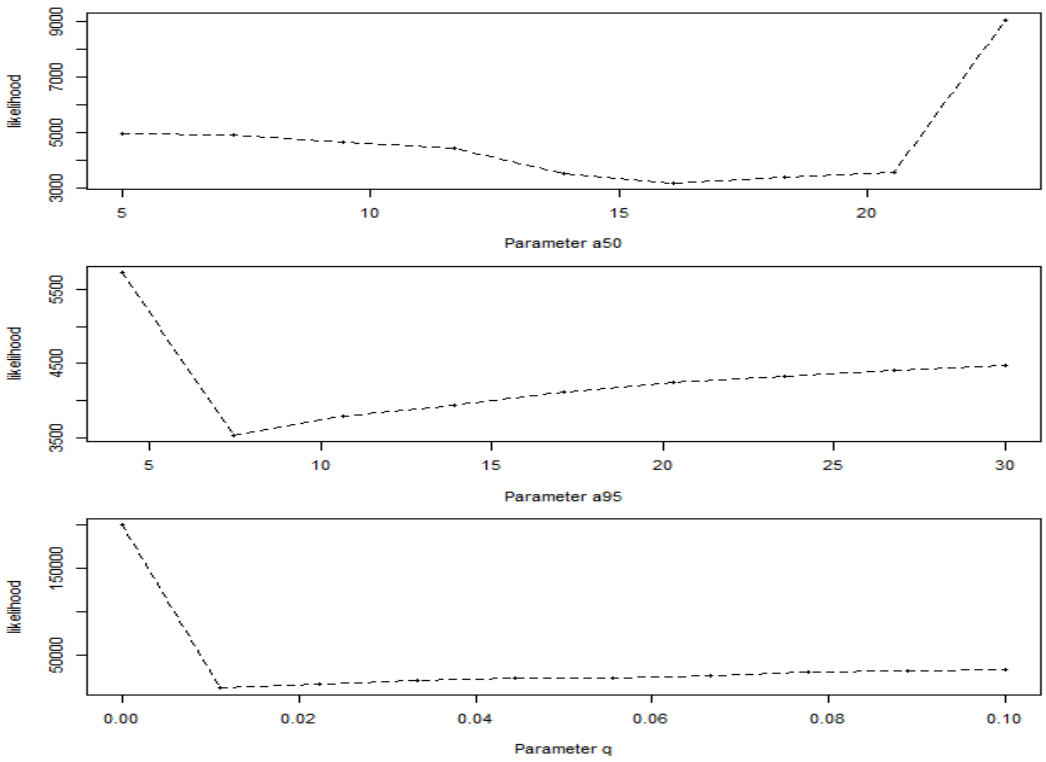

Figure 5.8: Profiles for the parameters in the logistic preference model

The expected model fit for two years (2010 and 2005), and in two different seasons (summer and winter) is shown in Figure 5.9,

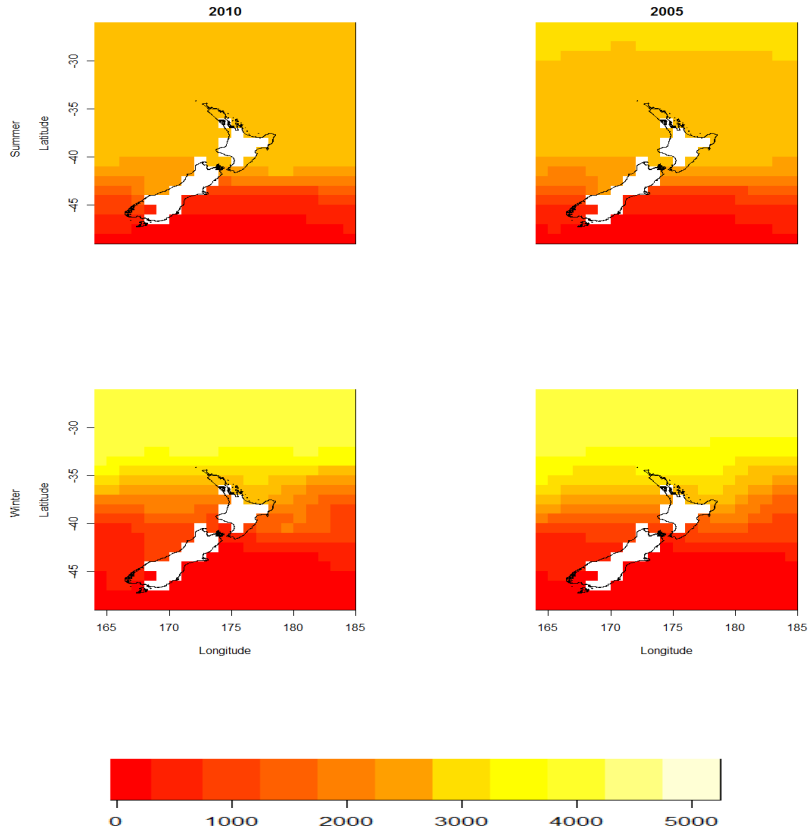

Figure 5.9: Expected spatial distribution with univariate model. Key is in numbers of individuals per cell

Figure 5.9 shows each season has two main spatial areas with near homogeneous population distribution in each. With high proportion of the population in the northern cells where the temperature is warmer (see Figure 3.2).

5.3 Bivariate Model

Model selection was continued to see if an extra preference attribute could explain more variation than the univariate model. Table 5.3 shows the

summary results from all the allowed additional preference attributes and functions.

Table 5.3: Bivariate preference model summary, where each model is summarised by objective function, number of parameters, and AIC

Variable	Function	$-LogL$	Params	AIC
Univariate	SST Logistic	3132.47	3	6270.94
MAG	Normal	3050.98	6	6113.96
	Double Normal	3050.46	7	6114.92
	Logistic	3050.4	6	6112.8
NPP	Normal	3178.34	6	6368.68
	Double Normal	3110.71	7	6235.42
	Inverse Logistic	3132.45	6	6276.9
	Exponential	3132.47	5	6274.94
Depth	Inverse Logistic	3132.31	6	6276.62
	Exponential	3083.69	5	6177.38
Distance	Inverse Logistic	3131.96	6	6275.92
	Exponential	3132.48	5	6274.96

Based on AIC, the preference attribute MAG was the best secondary preference attribute (Table 5.3). All of preference functions applied, gave similar global fits, all of which were better than the univariate model. Based on AIC the best bivariate model, is with the logistic function. Because the AIC was close between the normal and logistic functional form for MAG attribute, they were investigated (Figure 5.10),

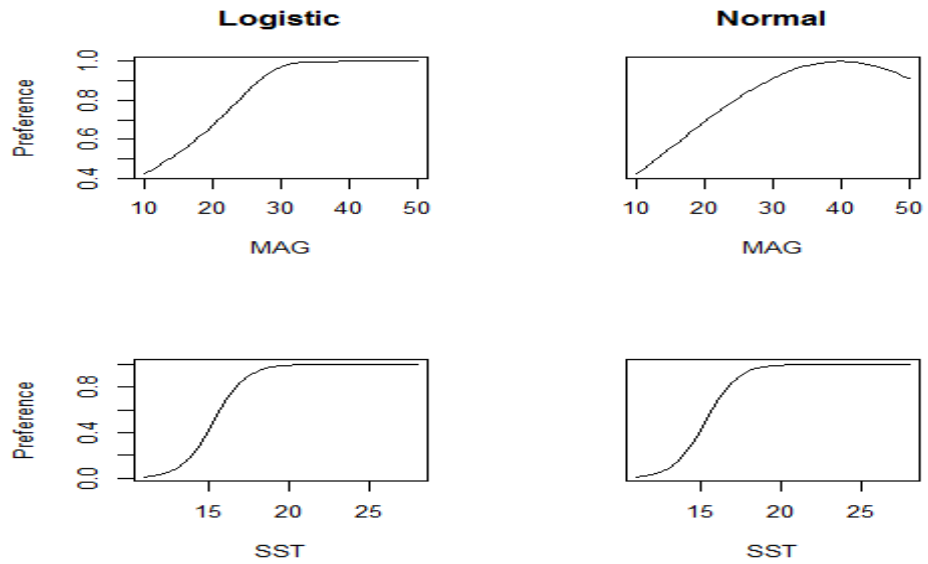

Figure 5.10: Estimated preference functions left two panels, MAG∼logistic form, right two panels MAG∼normal form

Figure 5.10 shows there was little functional difference between the two competing SST + MAG models, with the exception being the high end of the MAG variable. Logistic was chosen for MAG because only the left hand side of the normal function was being utilised which is what the logistic function describes. For this reason MAG described by a logistic function was considered the best second attribute in the bivariate model. Parameter estimates and margin of error are shown in Table 5.4,

Table 5.4: Final bivariate parameter estimates with margin of error

Parameter	estimate	Margin of error
q	0.000286	0.000297
SST_{a50}	15.31	1.285
SST_{ato95}	2.89	0.638
MAG_{a50}	28.50	0.99
MAG_{ato95}	5.08	0.56
MAG_{alpha}	0.080	0.014

Point estimates from Table 5.4 give the corresponding preference function.

$$p_k = f(x_{1k}; 15.31, 2.89)^1 \times f(x_{2k}; 28.50, 5.08)^{0.08}, \tag{5.1}$$

where $f()$ is the logistic function with parameters defined in section 3.3. Table 5.4 also shows that, margin of error for parameters are orders of magnitude larger than the univariate parameters. From Table 5.4 the effect of the preference attribute MAG is only 8% of SST, denoted by MAG_{alpha}. The bivariate model with parameter estimates in Table 5.4 was investigated for local fit through normalised residual fit (Figures 5.11-5.13).

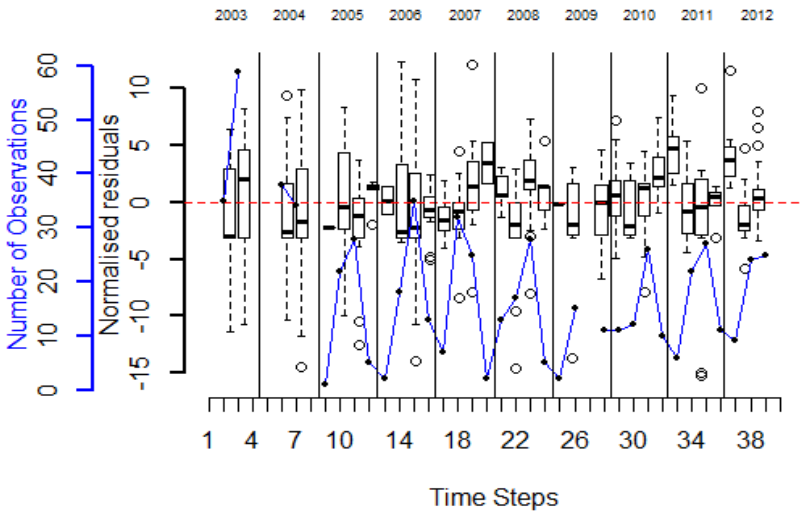

Figure 5.11: Boxplots of normalised residuals summarised into each time step, Blue legend is number of observations in each time step, the red line indicates 0, and the black legend represents the value of the residuals

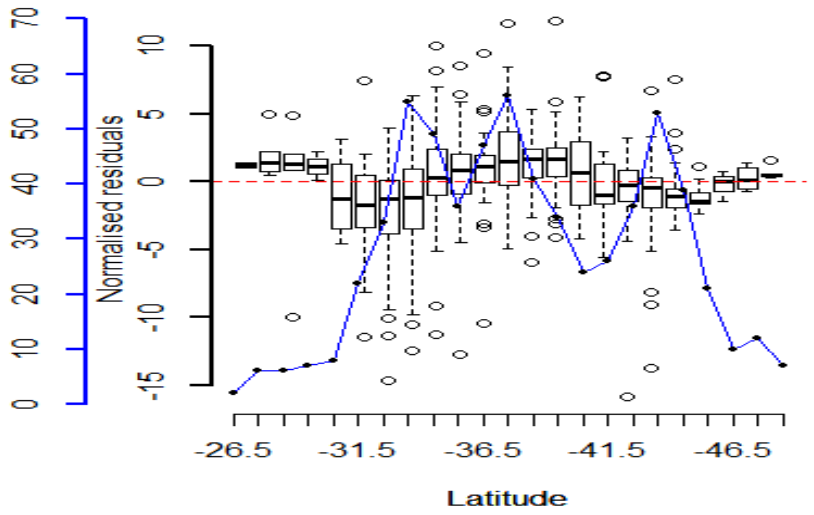

Figure 5.12: Boxplots of normalised residuals summarised into latitude bins, Blue legend is number of observations in each latitude bin, the red line indicates 0, and the black legend represents the value of the residuals

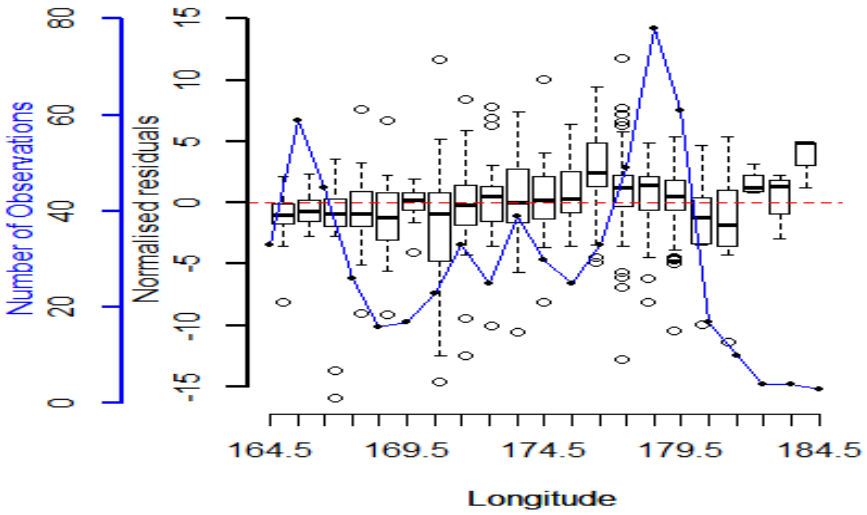

Figure 5.13: Boxplots of normalised residuals summarised into longitude bins, Blue legend is number of observations in each longitude bin, the red line indicates 0, and the black legend represents the value of the residuals

Residual fit (Figure 5.11-5.13) for the bivariate model was quite similar to the univariate model, where the bivariate model does fit better is in the western longitudes (180E-185E). In this region the residuals are tighter around zero with fewer outliers. Although it fits better, the number of observations are small and this could be why, only a small change in objective function is observed. There is still a clear nonlinear trend in the spatial residuals. This was nonlinear trend was summarized using the GAM method to find the general trend over space.

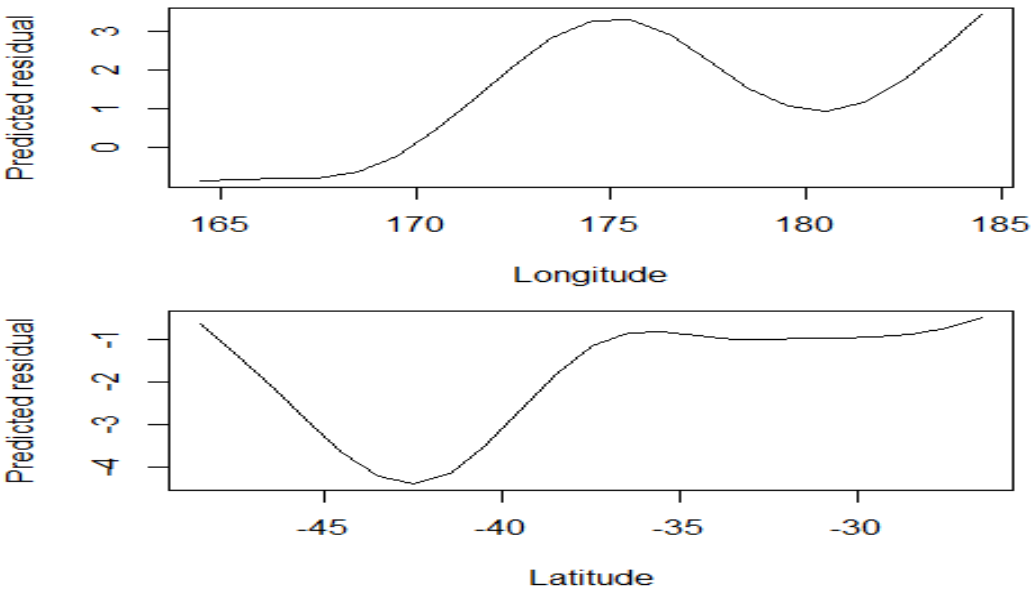

Figure 5.14: Predicted residual effect along latitude and longitude, from the GAM

Figure 5.14 shows the general nonlinear trend observed in the residuals. We can see over fitting occurs along the western longitudes, and major under-fitting between -45 and -40 $^{\circ}S$. The GAM with latitude and longitude as main effects explained 18% of the deviance in residual pattern.

Profiles were run on all parameters of the final bivariate model to ensure convergence and global minimum (Figure 5.13).

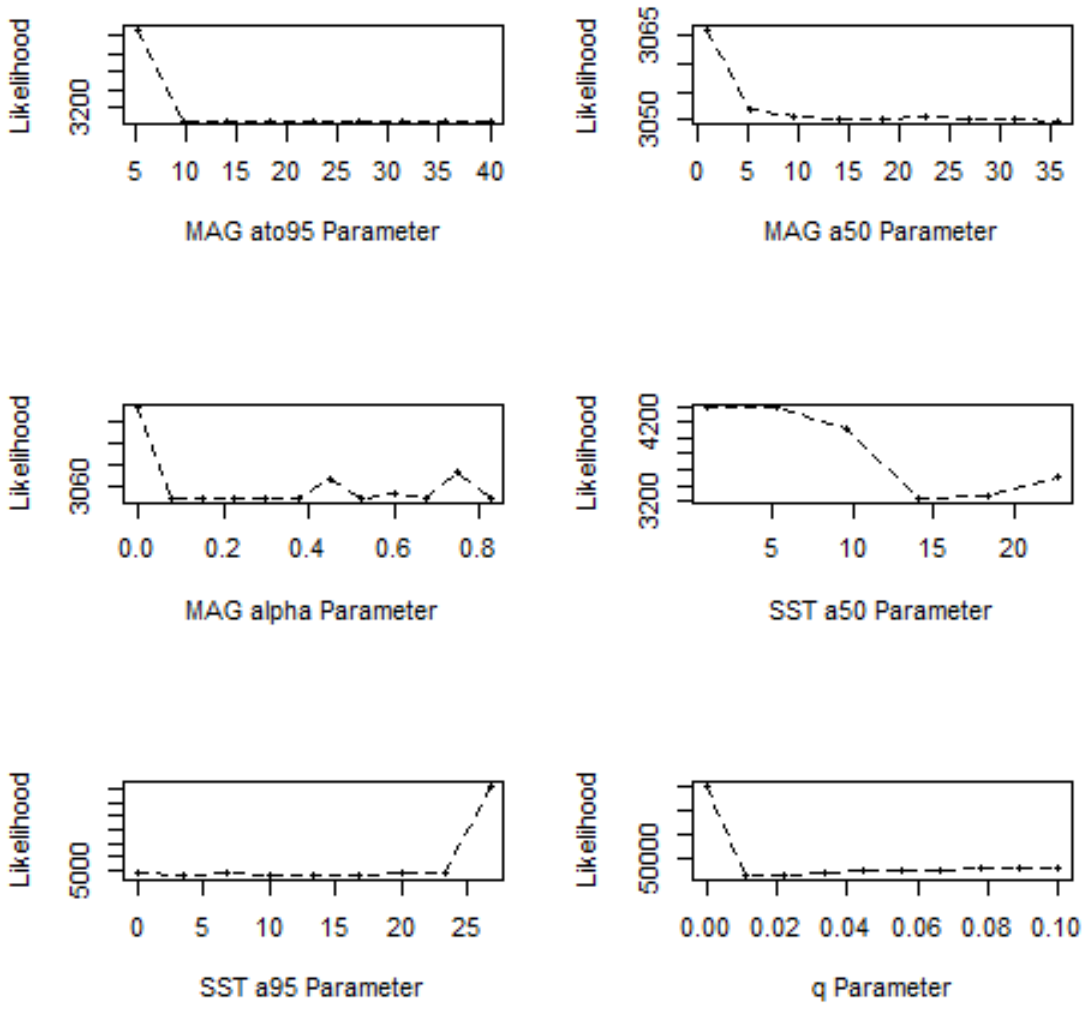

Figure 5.15: Profiles

SST $a50$, and q both had identifiable minima over the variable space. Parameters that were not well estimated were MAG $a50$, $ato95$, alpha and SST $ato95$. It seems that although MAG adds more information (i.e., a lower objective function and AIC) its parameter estimates are poorly de-

termined. For this reason model selection was stopped at the bivariate model as it was assumed any further parameters would be equally poorly determined.

Expected abundance from the model with parameter estimates from Table 5.4 are shown for two years (2010 and 2005) and two seasons (summer and winter) in Figure 5.16.

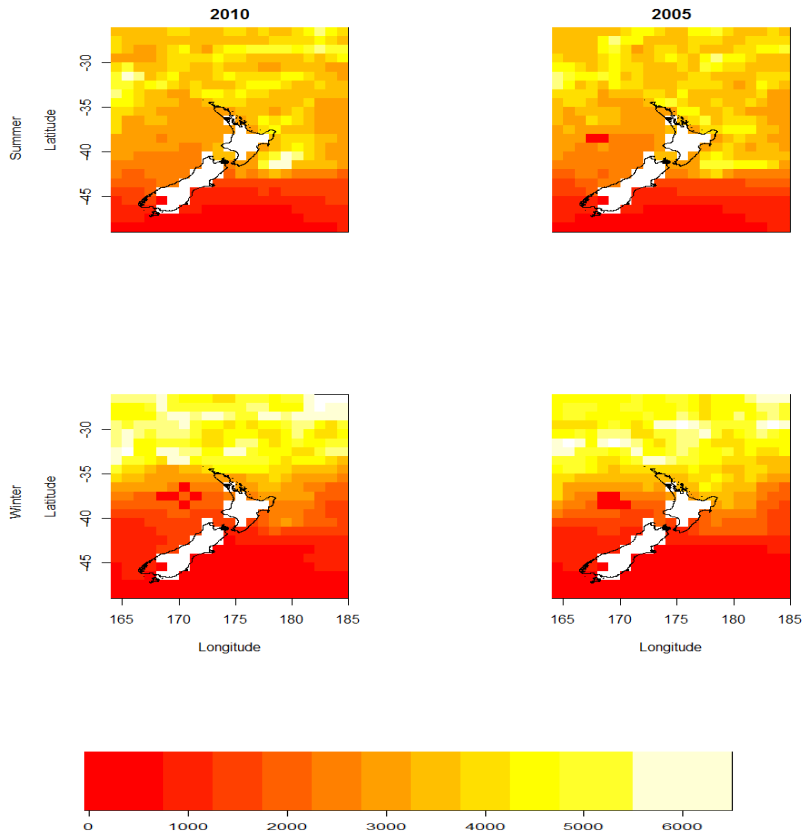

Figure 5.16: Expected spatial distribution from the bivariate model for a population fixed at one million individuals

Figure 5.16 shows the distributions of fish in areas with warm SST values, had relatively high densities around areas with high MAG. This was considered the best model for describing the relative distribution of albacore given the available data and model framework

5.4 Copulas

To compare with copula approach, the model selection process was conducted with restricted preference functions (only Normal or Exponential).

The best bivariate (non-copula) model with restricted preference functions contained SST and MAG as attributes, with parameter estimates as in Equation (5.2).

$$p_{i,j} = \mathcal{N}_{SST}(18.97, 3.33)\mathcal{N}_{MAG}(49.84, 10.40)^{0.117} \tag{5.2}$$

This yielded the bivariate preference function over both variables space shown in Figure 5.17,

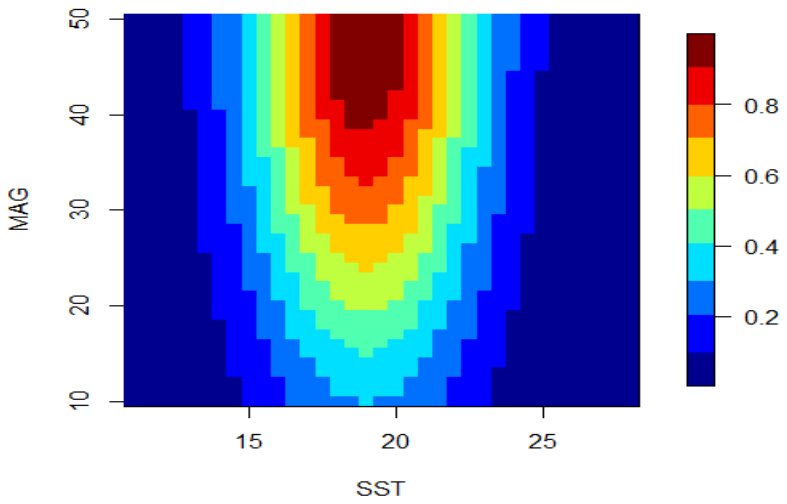

Figure 5.17: Best bivariate preference function for non-copula model, over the preference attributes space The key represents preference probabilities, so high preference is attributed to red-dark red

The residuals are drawn in Figures 5.18-5.20

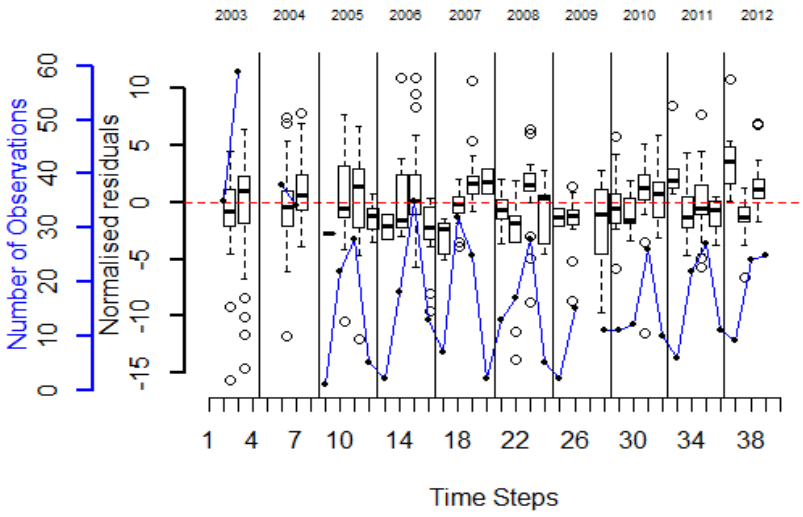

Figure 5.18: Boxplots of normalised residuals summarised into each time step, Blue legend is number of observations in each time step, the red line indicates 0, and the black legend represents the value of the residuals

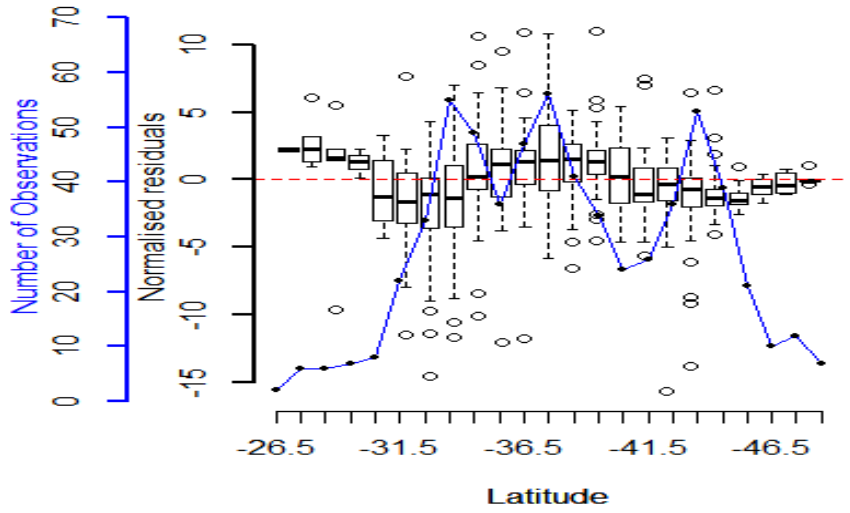

Figure 5.19: Boxplots of normalised residuals summarised into latitude bins, Blue legend is number of observations in each latitude bin, the red line indicates 0, and the black legend represents the value of the residuals

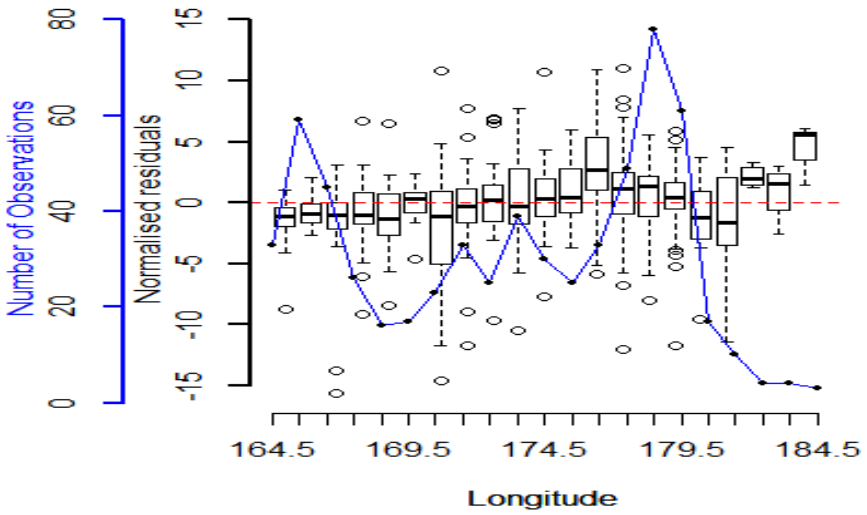

Figure 5.20: Boxplots of normalised residuals summarised into longitude bins, Blue legend is number of observations in each longitude bin, the red line indicates 0, and the black legend represents the value of the residuals

From the residual plots, we observe a similar fit to the bivariate model summarised in the earlier section.

All the dependence structures described in Section 4.1 were then fitted as potential descriptions of the joint distribution. The results are summmarised in the following Table 5.5,

Table 5.5: Copulas vs existing Preference SST + MAG bivariate model, with summary statistics

Copulas or Preference	objective function	Parameters
Preference	3088.25	6
Frank	3066.7	6
Gaussian	3015.36	6
Gumbel	3088.07	6

From Table 5.5, we see that all global fits were close to one another (difference ranged between 51-73 objective function points bewteen the models). The best model, based on global fit, was with the Gaussian dependence structure. This had the estimated preference function displayed in Figure 5.21.

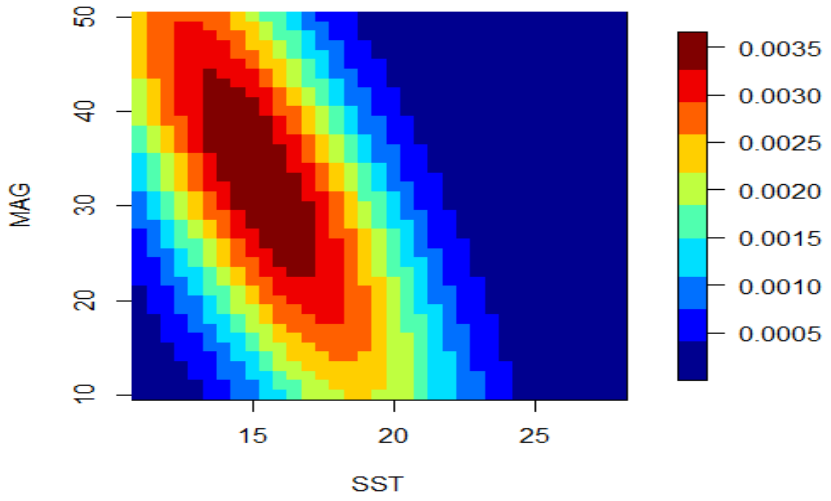

Figure 5.21: Gaussian bivariate copula function with estimated θ. Over the preference attribute space.

Assuming albacore are associated to SST and MAG with a Gaussian dependence yielded a decrease in objective function. From this model we, see that SST and MAG, are negatively correlated ($\rho = -0.78$) when explaining albacore relative spatial preference. Residuals were plotted to see if there were substantial improvements in local fit.

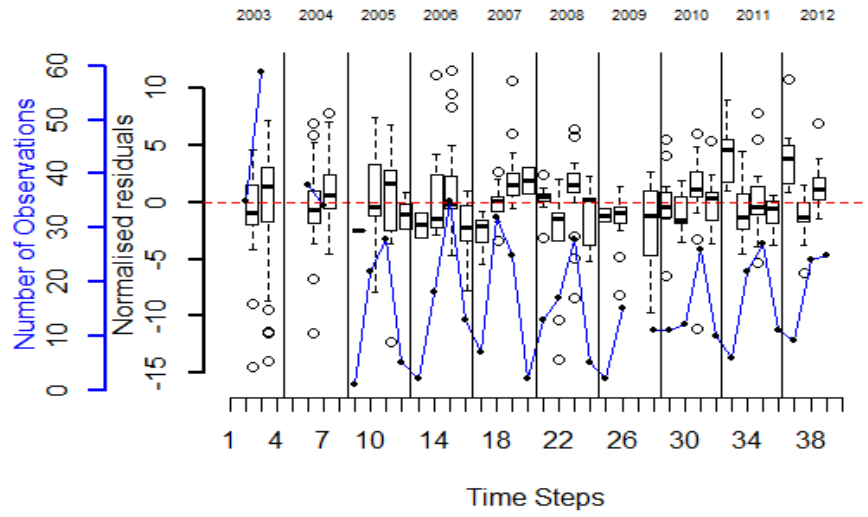

Figure 5.22: Normalised residuals collapsed into time step bins from the Gaussian Copula model.

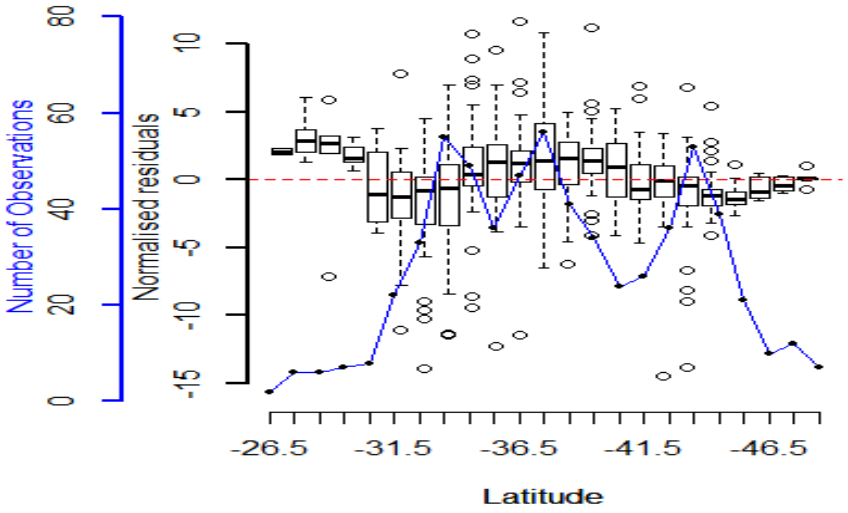

Figure 5.23: Normalisedresiduals collapsed into Latitude bins from the Gaussian Copula model.

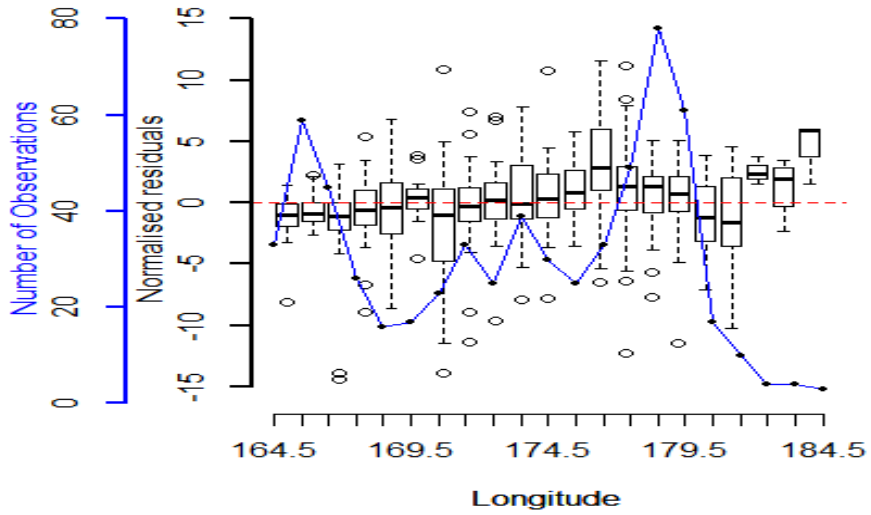

Figure 5.24: Normalised Residuals collapsed into Longitude bins From the Gaussian Copula model

There was no clear difference in residual plots between SPM's current preference model and the Gaussian copula model, suggesting that although copulas has made an improved fit it did not change the fit in a substantial way.

5.5 Zero observation sensitivities

Table 5.6: Summarising models from sensitivity analysis

CPUE value	CV value	objective function	q, SST_{a50}, SST_{ato95}, MAG_{a50}, MAG_{ato95}, MAG_{alpha}
0.00248	2.19	3101.99	0.00029, 15.26, 2.76, 38.6, 7.1, 0.11
0.01	1.57	3050.4	0.00029, 15.31, 2.89, 28.5, 5.09, 0.08
0.0498	0.51	3274.3	0.00028, 15.64, 3.65, 30.7, 5.51, 0.099
removed	removed	2895.44	0.00029, 15.25, 2.8, 30.28, 9.52, 0.15

Table 5.6 shows that the objective function was sensitive to whether zeros were imputed or ignored. If imputation is done, then the model objective function was sensitive to different imputed values. Although the global fit varied the estimated parameter values remained relatively constant over all model runs(e.g. SST_{a50} 15.26-15.64). This suggested that, although fit changed somewhat, the preference did not, and therefore preference was not sensitive to imputation or removal of zeros.

Chapter 6

Discussion

6.0.1 Biological Findings

This study suggests albacore tuna in New Zealand's EEZ prefer areas associated with high sea surface temperature and high sea surface magnitude. Specifically, a preference for areas with sea surface temperature greater than 15°C, and preference for areas with sea surface gradient greater than 10 $cm\ km^{-1}$. These were identified as the best oceanic attributes for predicting the relative spatial distribution of albacore tuna catch rates. The bivariate model explained more spatial and temporal variation than an assumed uniform spatial model (null model), and univariate a model (SST alone).

Previous studies on albacore tuna have suggested SST as a significant environmental variable towards albacore presence/absence, and CPUE (McGregor and Horn, 2013, Xu et al., 2013), with preferred temperature ranges from 15-21°C and 10-25°C respectively. The lower SST limits similar to those identified in McGregor and Horn (2013) study of (15°C). When

applying the logistic preference function, it implies there is no reduction in preference as temperature continues to increase. We can only confirm that range extends to $26°C$, as this was the observed maximum temperature limit in this investigation (Figure 3.15). The second most important preference variable was Sea Surface Gradient. This had only 8% of the weight of SST as denoted by α (Equation 5.1). Previous studies have attributed tuna aggregations to frontal areas (Fiedler and Bernard, 1987), this was supported here as Sea Surface Gradient is a latent attribute for intensity of oceanic fronts. Although Sea Surface Gradient was considered the second most important preference variable, the parameters describing the functional form proved hard to estimate, shown by the flat likelihood profiles (Figure 5.15). Other studies have suggested that NPP is a significant environmental attribute for tuna catch rates, where the the preference is around $220-380\, mg\ C/m^2/d$ (Lan et al., 2013). NPP presumably does not attract tuna directly, but rather it is the secondary production that attract tunas. It is thought that oceanic fronts play an important role in aggregating plankton and micronekton (Flament et al., 1996, Power, 1996), which promote NPP. We did not observe a direct correlation between NPP and MAG in the area of interest (Figure 3.17). This lack of relationship could be due to constraints/assumptions applied to the data discussed in the following section.

Results found in this thesis have potential implication for differing stakeholders. There is the potential for fishermen to use this to be more efficient at catching their quota. It could be used by fisheries managers to obtain a better understanding of why fish stocks and catches fluctuate through time and space. Also the model might be run with preference at-

tributes from climate models to evaluate potential albacore responses to climate change and ocean warming.

6.0.2 How model assumptions have affected results

One of the major assumptions and constraints in this model was spatial and temporal resolution. The problem results from amalgamating detailed catch rates and environmental characteristic data over large areas and time periods, this caused a loss of information. When environmental attributes are averaged, we loose small scale spatial and temporal trends. For example, small scale eddies or local fronts systems. An example is the Sub Tropical Front which runs along Chatham Rise. This front is reported to be relatively narrow ($150km$) (Sutton, 2001). Cells here, were $111km$ in height, therefore averaging over a cell that is of similar size may dilute the signal from the front. This is where environmental data resolution and small sample sizes can be restrictive for this type of investigation. Higher resolution products will add power to preference models by reducing the averaging of data. Another idea that could yield higher utilisation of information that is currently lost due to amalgamation, is add hierarchical structure for observations within SPM.

From the spatial residuals from the final bivariate model (Figure 5.12–5.13) there was a poor fit of the model to CPUE in some areas. Although we investigated this pattern against the spatial resolution attributes of the model; latitude, and longitude, only 18% of the deviance could be explained using these as covariates in a non linear relationship in a generalised additive model (GAM). This trend in the residuals suggested there

was potentially another wide-scale preference attribute or process that was missed in this investigation. Biotic factors that affect a population's distribution could be due to competition (i.e. density dependence Lindberg et al. 2006), predator/prey presence (Cosner et al., 1999), or an abiotic variable that we did not have access to (e.g. depth of mixing layer). These would be interesting to incorporate for future model runs.

The observed data were commercial fishing catch and effort, thus the spatial coverage of our observed dataset was based on the prior expectation of commercial fishermen. This means the coverage of observed data may not to be ideal for fully exploring the spatial the spatial distribution and preferences of the population of interest. Research data are usually very expensive and time consuming, although have better (controlled) spatial and temporal coverage, so are not always available and so we use the data available. This raises question as to whether we are modeling the albacore tuna movement, or the fishers expectation on albacore tuna. This could affect results through missing or biased coverage. Suppose there exists a scenario where there was a movement corridor that had not been discovered by the fishery. This could be a source of valuable information and insight into preferences of New Zealand's albacore tuna. Another potential bias in the commercial fishing data is different vessels may have different CPUE because of different fishing practices and gear. The study data here did not have vessel information available so we could not standardise it in our observed data.

Knowing whether a multivariate model has identified a global minimum in the objective function is difficult to assess. In this study we had access to two optimizers for finding a unique global minimum. Although

in theory one was better suited for global minimum problems (Storn and Price, 1995) this optimiser did not perform well compared to the local optimiser (Dennis Jr and Schnabel, 1996). For reasons such as getting stuck in local minimums and number of un converged model runs. This meant we scrapped the global optimizer for a local minimiser in this study. This came with a computational cost, for each model was run with multiple starting parameter values to check convergence. Potentially a better global optimiser would be suited for this model.

Changing assumptions on zero observations in CPUE turned out to have a moderate influence on the objective function, but little effect on the preference parameters. Zero catch observations are a common occurrence in abundance or biomass records from multi-species fisheries (Maunder and Punt, 2004). The methods for handling zero observations in this investigation were quite crude. Ignoring zeros is equivalent to ignoring information in a spatial investigation. Other techniques are needed to address zero catch within SPM such as those applied in other spatially explicit models. SEAPODYM manages zero observations in the likelihood, where the likelihood follows a negative binomial distribution with inflated zeros (Senina et al., 2008). Other analyses on CPUE have used the delta approaches or hurdle models (Cragg, 1971). These two latter approaches could be considered future modification within SPM, because this information needs to be accurately handled for these investigations.

6.0.3 The challenge of model assessment in spatially explicit population models

Model comparison is easily achievable through summary statistics such as AIC and BIC. These statistics give an idea of relative model performance, but not absolute model performance. In this investigation model performance was assessed by local residual fit, summarised over three axes (latitude, longitude and year). However given the nature of the data (spatial and temporal) this approach was still not ideal, as it could not identify residual patterns over both space and time (interactions). An alternative is to look at residual maps for every time step, this method seemed difficult and arduous when comparing competing models, given there were forty plots per model run. There is little evidence of investigation into local residual diagnostics for spatially explicit population models in the literature. This is a future area of research, with spatially explicit population models likely to grow in use in the future. Assessing the fit, taking into account the nature of the problem (spatial and temporal interacting patterns), would be of great benefit to future modelers. One suggestion is the use of analytical tools, such as cluster analysis or tree regression techniques to analyse (rather than just visualise) residual patterns.

6.0.4 Copula based modelling

In nature, environmental variables are related. Oceanic fronts promote mixing of nutrients and aggregation of plankton, with greater NPP as a result (Moore and Abbott, 2000). Coastal areas generally attain higher levels of NPP relative to the open ocean (Carpenter, 1998), suggesting

correlations between NPP and bathymetry. SST and SSH are commonly highly correlated (Figure 3.17). The natural correlation of environmental attributes gave motivation to study dependence when using multiple variables, and to describe the joint preference distribution. It seems likely multiple preference attributes affect spatial distributions so treating them as independent processes seems an unrealistic constraint.

From the copula results we found that using the Gaussian Copula gave a slightly better fit in the objective function compared to the restricted independent preference model (Table 5.5), but no substantial difference in the summarised residuals (Figure 5.19- 5.24). Given the relatedness between environmental variables that make up habitats for marine organisms, the idea of copula in this setting makes sense. More work is needed to describe more distributions that more realistically explain preference relationships (i.e logistic and inverse logistic PDF's). This method will be better suited when both preference attributes have a significant effect. Parameter estimates were poorly determined for the second preference attribute (MAG). I believe this diluted the effect of applying a copula. For another species applying copula theory may have a more noticeable effect.

This thesis looked into applying a preference based movement process describing the spatial distribution of a fishery, in a population model framework. Some spatial and temporal heterogeneity of the albacore longline fishery was successfully explained by applying this method. There were still unexplained patterns in residuals, which suggested more insight on the movement of albacore tuna around New Zealand is needed. This study also encouraged incorporating dependence structures for multiple naturally related preference attributes, when explaining the spatial dis-

tribution of a population. Copulas were used in this instance but more refinement for application is needed. The next step after this thesis, is to apply SPM, with population processes to a distinct stock, this will be when SPM is most powerful.

Appendix A

Appendix

A.1 Summary of Observed CPUE

Table A.1: Mean sandardised CPUE, summarised for each year and season

Year	Summer	Autumn	Winter	Spring
2003	NA	0.670	1.223	0.246
2004	NA	0.894	1.123	NA
2005	0.0596	1.365	0.764	0.913
2006	0.658	1.501	0.906	0.614
2007	0.450	0.497	1.678	2.524
2008	0.880	0.339	1.451	1.035
2009	0.522	0.515	1.255	1.007
2010	0.946	0.461	0.887	2.001
2011	2.182	1.039	0.844	0.666
2012	2.746	0.337	1.009	NA

A.2 Plots of *a priori* coefficient of variation

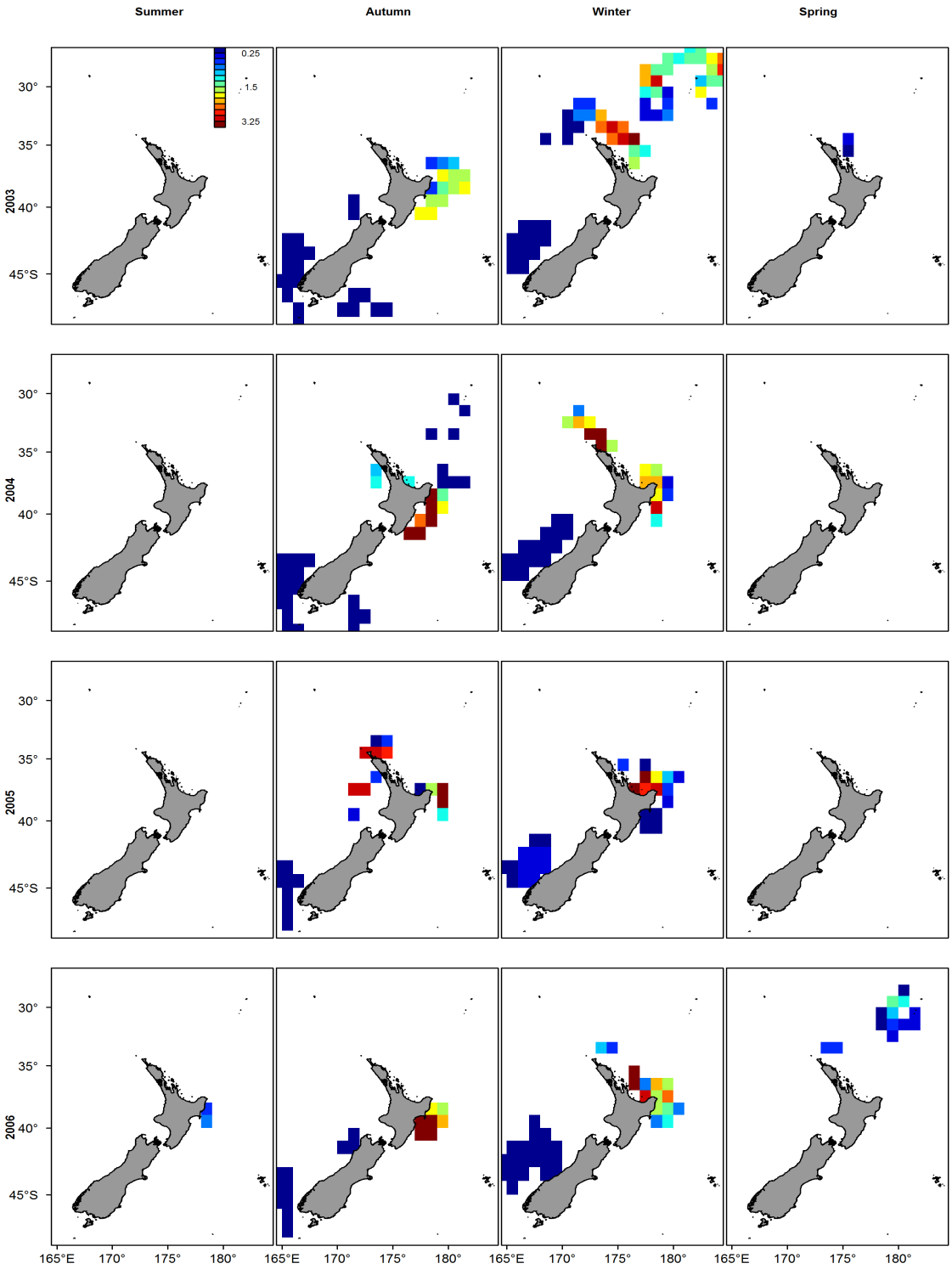

Figure A.1: Observed albacore coefficient of variation for the seasons between the years 2003-2006

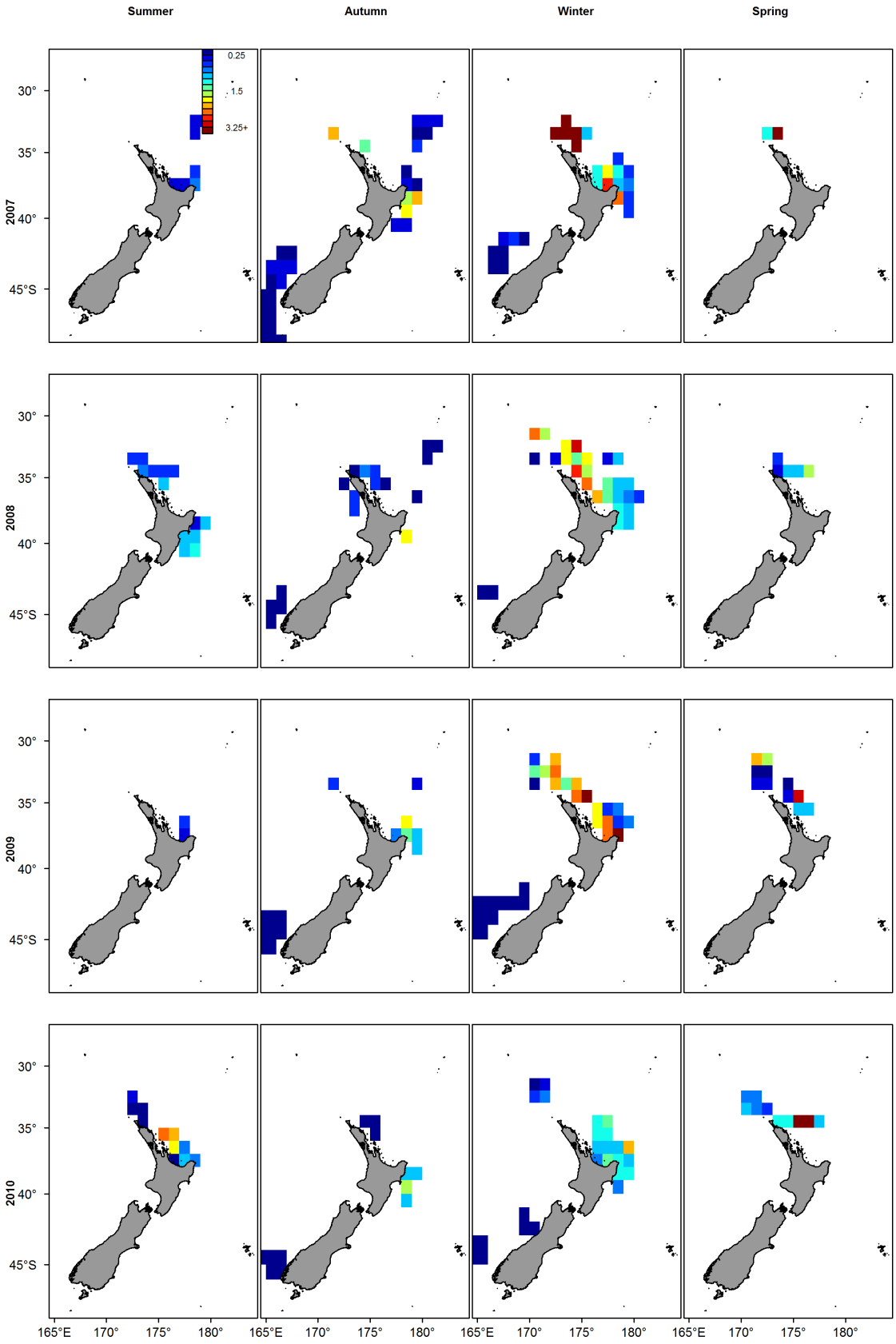

Figure A.2: Observed albacore coefficient of variation the seasons between the years 2007-2010

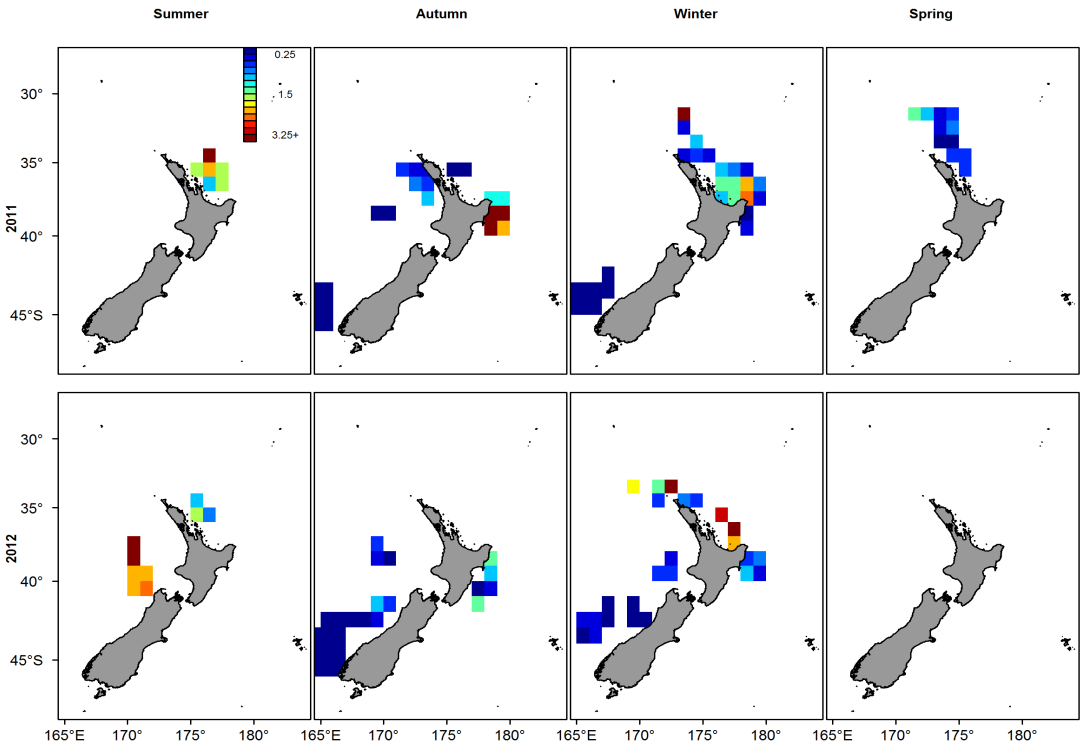

Figure A.3: Observed albacore coefficient of variation for the seasons between the years 2011-2012

A.3 Likelihood correction

the negative log likelihood derivation, assume $I_{t,k} \sim LN(\mu_{t,k}, \sigma_{t,k})$ with mean $N_{t,k}$. N.B for log Normal $E[I_{t,k}] \neq \mu_{t,k}$, but $E[I_{t,k}] = e^{\mu_{t,k}+\sigma^2_{t,k}/2}$,

$$f(I_{t,k}) = \frac{1}{I_{t,k}\sqrt{2\pi\sigma^2_{t,k}}} exp\left(-\frac{1}{2}\left(\frac{log(I_{t,k}) - \mu_{t,k}}{\sigma_{t,k}}\right)^2\right) \tag{A.1}$$

Thus the likelihood for RCT observations.

$$\prod_{t=1}^{T}\prod_{k=1}^{RC} f(I_{t,k}) = \prod_{t=1}^{T}\prod_{k=1}^{RC} \frac{1}{I_{t,k}\sqrt{2\pi\sigma^2_{t,k}}} exp\left(-\frac{1}{2}\left(\frac{log(I_{t,k}) - \mu_{t,k}}{\sigma_{t,k}}\right)^2\right) \tag{A.2}$$

Negative log likelihood $-LogL$.

$$-LogL = -log \prod_{t=1}^{T} \prod_{k=1}^{RC} f(I_{t,k}) \tag{A.3a}$$

$$= \sum_{t=1}^{T} \sum_{k=1}^{RC} \left(log(I_{t,k}) + log(\sigma_{t,k}) + \frac{1}{2} log(2\pi) + \frac{1}{2} \left(\frac{log(I_{t,k}) - \mu_{t,k}}{\sigma_{t,k}} \right)^2 \right) \tag{A.3b}$$

When we ignore constants from the likelihood,

$$-LogL = \sum_{t=1}^{T} \sum_{k=1}^{RC} \left(log(\sigma_{t,k}) + \frac{1}{2} \left(\frac{log(I_{t,k}) - \mu_{t,k}}{\sigma_{t,k}} \right)^2 \right) \tag{A.4}$$

Since $E[I_{t,k}] = qN_{t,k}$ we have,

$$E[I_{t,k}] = qN_{t,k} \tag{A.5a}$$

$$= e^{\mu_{t,k} + \sigma_{t,k}^2/2} \tag{A.5b}$$

so

$$log(qN_{t,k}) = \mu_{t,k} + \sigma_{t,k}^2/2 \tag{A.6}$$

$$\mu_{t,k} = log(qN_{t,k}) - \sigma_{t,k}^2/2 \tag{A.7}$$

when Equation A.7 is substituted back into the likelihood (Equation A.4), we get

$$-LogL = \sum_{t=1}^{T} \sum_{k=1}^{RC} \left(log(\sigma_{t,k}) + \frac{1}{2} \left(\frac{log(I_{t,k}) - log(qN_{t,k}) + \frac{1}{2}\sigma_{t,k}^2}{\sigma_{t,k}} \right)^2 \right) \tag{A.8}$$

Which can be written as in section 4.3 (Equation 4.11),

$$-LogL = \sum_{t=1}^{T}\sum_{k=1}^{RC}\left(log(\sigma_{t,k}) + \frac{1}{2}\left(\frac{log(I_{t,k}) - log(qN_{t,k})}{\sigma_{t,k}} + \frac{1}{2}\sigma_{t,k}\right)^2\right) \quad \text{(A.9)}$$

$$-LogL = \sum_{t=1}^{T}\sum_{k=1}^{RC}\left(log(\sigma_{t,k}) + \frac{1}{2}\left(\frac{log(I_{t,k}/qN_{t,k})}{\sigma_{t,k}} + \frac{1}{2}\sigma_{t,k}\right)^2\right) \quad \text{(A.10)}$$

The only difference being an extra robustifying function is added to prevent the likelihood to error out if the model expectation equals zero ($N_{t,k} = 0$)

A.4 Copula Theory

Most of the working follows here comes from Nelsen (2007), a very useful source for learning about the copula technique. Copulas is the multidimensional technique used on probability distributions. Copula theory describes the joint cumulative distribution of random variables. This results in the technique using Cumulative distributions of the margins ($F_X(x)$)

Definition A.4.1 *A distribution function is a function F with domain $\boldsymbol{R}$*

$$F_X(x) = \int_{-\infty}^{x} f(t)dt \quad \text{(A.11)}$$

$$U = F_X(x)$$

where,

$$U \sim Uniform(0,1)$$

Knowing this then we can always use u to get back to x, since $F_X(x)$ is a nondecreasing function with inverse (or generalised inverse) function, defined as any value of u between 0 and 1 as:

$$F_X^{(-1)} = inf\{x : F_X(x) \geq u\} \qquad 0 \leq u \leq 1,$$

$$x = F_X^{(-1)}(U)$$

or more explicitly,

$$x = F_X^{(-1)}(F_X(x))$$

Now we can express a joint Cumulative distribution in terms of uniform random variables,

A.4.1 Definition and Basic Properties of Copula Function

A bivariate copula function is a function that links or marries univariate marginals to their full multivariate distribution. For two uniform random variables, U_1, U_2 the joint distribution function C, defined as

$$C(u_1, u_2) = P(U_1 \leq u_1, U_2 \leq u_2) \tag{A.12}$$

can also be called a copula function. Copula functions can be used to link marginal distributions with a joint distribution. For a given univariate

marginal distribution functions $F_1(x_1), F_2(x_2)$, the function

$$C(F_1(x_1), F_2(x_2)) = F_{12}(x_1, x_2) \tag{A.13}$$

which is defined using a copula function C, results in a multivariate distribution function with univariate marginal distributions as specified $F_1(x_1), F_2(x_2)$. This property can be easily shown:

$$\begin{aligned} C(F_1(x_1), F_2(x_2)) &= P(U_1 \leq F_1(x_1), U_2 \leq F_2(x_2)) && \text{(A.14a)} \\ &= P(F_1^{(-1)}(U_1) \leq x_1, F_2^{(-1)}(U_2) \leq x_2) && \text{(A.14b)} \\ &= P(X_1 \leq x_1, X_2 \leq x_2) && \text{(A.14c)} \\ &= F_{12}(x_1, x_2) && \text{(A.14d)} \end{aligned}$$

This demonstrates that any joint Cumulative Distribution Function (CDF) of random variables can be expressed by a copula function with uniform random variables. These uniform random variables can be considered reinterpretations through CDF of random variables with an assumed probabilistic distribution function. A bivariate copula is a function $C : [0, 1]^2 \rightarrow [0, 1]$. That adheres to the following criteria defined,

Definition A.4.2 *A copula is a function C from $\boldsymbol{I}^n$ to $\boldsymbol{I}$ with the following properties, where $\boldsymbol{I}$ is the unit matrix:*

1. *For every u_1, u_2 in $\boldsymbol{I}$,*

$$C(u_1, 0) = 0 = C(0, u_2)$$

and

$$C(u_1, 1) = u \text{ and } C(1, u_2) = u_2;$$

2. *For every* u_{11}, u_{12}, u_{21}, u_{22} *in* $\boldsymbol{I}$ *such that* $u_{11} \leq u_{12}$ *and* $u_{21} \leq u_{22}$, *where for* u_{ij} $i =$ *marginal family and* $j =$ *observation*

$$C(u_{12}, u_{22}) - C(u_{12}, u_{21}) - C(u_{11}, u_{22}) + C(u_{11}, u_{21}) \geq 0 \qquad \text{(A.16)}$$

Sklar (1959) established the above. He showed that any multivariate (in our case bivariate) distribution function F_{12} can be written in the form of a copula function. He proved the following: If $F(x_1, x_2)$ is a joint bivariate distribution function with univariate marginal distribution functions $F_1(x_1), F_2(x_2)$, then there exists a copula function $C(u_1, u_2)$ such that

$$F_{12}(x_1, x_2) = C(F_1(x_1), F_2(x_2)) \qquad \text{(A.17)}$$

For all bivariate copulas there are limits within their respective dimensions. These are known as Frèchet-Hoeffiding bounds of inequality, described in the following theorem.

Theorem A.4.1 *Let* C *be a copula. Then for every* u_1, u_2 *in Dom C,*

$$max(u_1 + u_2 - 1, 0) \leq C(u_1, u_2) \leq min(u_1, u_2) \qquad \text{(A.18)}$$

$proof$. Let (u_1, u_2) be an arbitrary point in Dom C. Now $C(u_1, u_2) \leq C(u_1, 1) = u_1$ and $C(u_1, u_2) \leq C(1, u_2) = u_2$ yield $C(u_1, u_2). \leq min(u_1, u_2)$. Furthermore, $V_c([u_1, 1] \times [u_2, 1]) \leq 0)$ implies $C(u_1, u_2) \leq max(u_1 + u_2 + 1, 0)$. Thus

for every copula C and every (u_1, u_2) in $\boldsymbol{I^2}$

$$max(u_q + u_2 - 1, 0) \leq C(u_1, u_2) \leq min(u_1, u_2)$$

These are graphically represented in figure for the bivariate case(??)

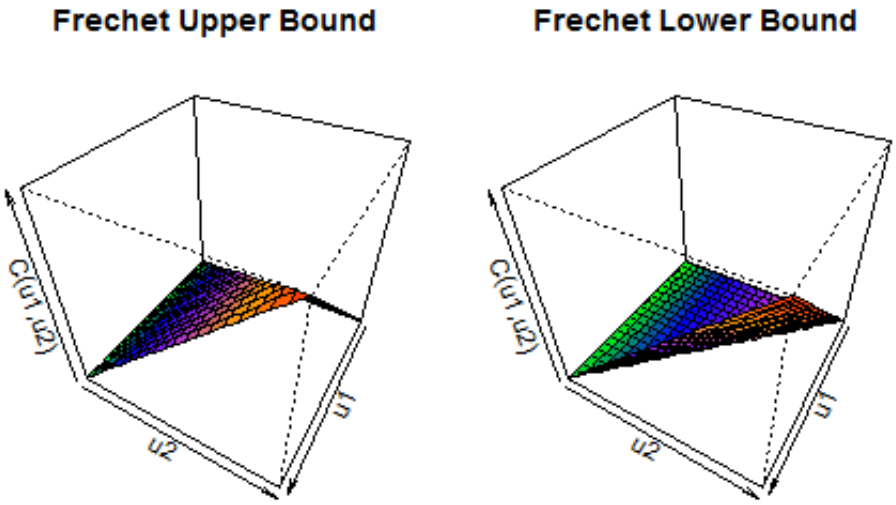

Figure A.4: Frèchet-Hoeffiding bounds of inequality. All bivariate copulas are restricted between the upper and lower bound

In this context we are interested in the joint Probability Density Function (PDF) so in order to obtain this we use partial differentiation along

with the chain rule on our joint cumulative copula as follows,

$$c(u_1, u_2, ..., u_n) = \frac{\partial^n}{\partial u_1, \partial u_2, ..., \partial u_n} C(u_1, u_2, ..., u_n) \tag{A.19a}$$

$$= \frac{\partial^n}{\partial u_1, \partial u_2, ..., \partial u_n} C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \tag{A.19b}$$

$$= c(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \prod_{i=1}^{n} f_i(x_i) \tag{A.19c}$$

The copulas used in this study come under two groups. There is the elliptical copulas and the Archimedean copulas. The Gaussian Copula is the only copula from the elliptical copula family,

A.4.2 The Gaussian Copula

Demonstrating the generalisation of this method on a specific example with two Randomly distributed variables with an assumed standard normal dependence structure (Gaussian copula).

Lets say we have two environmental attributes such as SST (X_1) and NPP (X_2). We have a range of proposed distributions that the margins may follow in which we would use to generate u_i,

$$u_i = F_i(x_i; \phi)$$

along with,

$$f_{x_i} = f_i(x_i; \phi)$$

So if we assume that our random variables dependencies can be described by a bivariate standard Gaussian/Normal copula, we can follow

the definition of the Gaussian copula Li (1999),

$$C(u_1, u_2) = \mathcal{N}_2(\Phi^{(-1)}(u_1), \Phi^{(-1)}(u_2))$$

Where $\mathcal{N}_2 \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{I})$ and $\Phi \sim \mathcal{N}(0, 1)$, Firstly we check weather choosing the bivariate normal distribution satisfies the definition of a copulas/joint distribution. Obviously because the bivariate normal distribution is a well defined distribution, it is known that is satisfies the definition of a copula, that is having the attributes of a distribution. The bivariate normal CDF distribution can be expressed with parameter $\rho \in [-1, 1]$. Since our proposed copulas is the standard normal with $\mu_i = 0$ and $\sigma_i = 1$ and margins $\Phi_i^{(-1)}(u_i)$ then we can write our Copula as,

$$C(u_1, u_2) = \frac{1}{2\pi\sqrt{1-\rho}} \int_{-\infty}^{\zeta_1} \int_{-\infty}^{\zeta_2} exp\left(- \frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{2(1-\rho^2)} \right) dz_1 dz_2 \quad \text{(A.20)}$$

where $\zeta_i = \Phi^{-1}(u_i)$

Finding the joint pdf of our normal copula involves differentiation of our CDF with respect to $\Phi_i^{(-1)}(u_i)$. Obtaining the following result

$$c(u_1, u_2) = \frac{1}{2\pi\sqrt{1-\rho}} exp\left(- \frac{\zeta_1^2 - 2\rho\zeta_1\zeta_2 + \zeta_2^2}{2(1-\rho^2)} \right) \frac{\partial}{\partial u_1} \Phi_1^{(-1)}(u_1) \frac{\partial}{\partial u_2} \Phi_2^{(-1)}(u_2) f_{x_2} f_{x_2} \quad \text{(A.21)}$$

Denoting as $\varphi = \phi'$ the pdf of the $\mathcal{N}(0, 1)$ law,

$$\frac{\partial}{\partial u_i}\Phi_i^{(-1)}(u_i) = \frac{1}{\varphi(\Phi_i^{(-1)}(u_i)} = \sqrt{2}\pi exp(\zeta_i^2/2) \tag{A.22}$$

So that,

$$c(u_1, u_2) = \frac{1}{2\pi\sqrt{1-\rho}} exp\left(- \frac{\zeta_1^2 - 2\rho\zeta_1\zeta_2 + \zeta_2^2}{2(1-\rho^2)} \right) exp\left(\frac{\zeta_1^2 + \zeta_2^2}{2} \right) f_{x_2} f_{x_2} \tag{A.23}$$

This copula gives us an elliptical and radial dependence structure (Figure 4.3). This is the general framework but note now we have the flexibility of choosing how u_i is calculated (i.e the marginal distribution). The other class of copula functions applied in this thesis are Archimedean copulas explained in the following section.

A.4.3 Archimedean copulas

The other class of copulas that will be used to represent alternative dependence structures are the Archimedean copulas. Table ? gives the formula for some of the most commonly applied Archimedean copulas. Archimedean copulas can be defined as,

$$C(u_1, u_2) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2)) \tag{A.24}$$

where φ is a specific function known as a generator of the copula. All the Archimedean copulas that we will be using for the bivariate case contain only one parameter so we can re write the definition as,

$$C(u_1, u_2; \theta) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2)) \tag{A.25}$$

where θ is the dependence parameter which is embedded with in the functional form of the generator. There are a range of copulas in this family which can be found in Nelsen (2007). The archimedean structures chosen are commonly used in biostatistics and actuarial science. Our copulas also had to fit the criteria of being differentiable. This is because we are actually interested in the Joint PDF in the end result. Since Copulas are generally expressed in joint CDF being differentiable is an important property.

Table A.2: Archimedean copulas formula (CDF), with generator and parameter space

Name	Bivariate copula $C(u_1, u_2)$	generator$\varphi_\theta(t)$	generator inverse $\varphi_\theta^{-1}(t)$	Parameter θ
Gumbal	$exp\Big(-\big((-log(u_1))^\theta + (-log(u_2))^\theta\big)^{\frac{1}{\theta}}\Big)$	$(-log(t))^\theta$	$exp\big(-t^{\frac{1}{\theta}}\big)$	$\theta \in [1, \infty)$
Frank	$-\frac{1}{\theta} log\left(1 + \frac{(exp(-\theta_1)-1)(exp(-\theta u_2)-1)}{exp(-\theta)-1)}\right)$	$-log\left(\frac{exp(-\theta t)-1}{exp(-\theta)-1}\right)$	$-\frac{1}{\theta log(1+exp(-t)(exp(-\theta)-1)}$	$\theta \in \mathbb{R} \setminus \{0\}$
Clayton	$\big(max\{u_1^{-\theta} + u_2^{-\theta}; 0\}\big)^{\frac{1}{\theta}}$	$\frac{1}{\theta}(t^{-\theta-1})$	$(1+\theta t)^{-\frac{1}{\theta}}$	$\theta \in$ $[-1, \infty) \setminus \{0\}$
Independence	uv			

A.5 Gaussian Copula C++ Code

The following code is in C++, this code was implemented in SPM. This example is of the Gaussian Copula (Table 4.2). The normal inverse functions within the code, is from Wichura (1988)

```
//*******************************************
// Function Calculates standard normal Quantile
   function
// used in Copula formula
//*******************************************
double CGaussianPreferenceFunction::NormalInverse(
   double p)
{
        double q, r, val;
        q = p - 0.5;
        /*-- use AS 241 --- */
        /* double ppnd16_(double *p, long *ifault)*/
        /*  37, NO. 3
        Produces the normal deviate Z corresponding to
            a given lower
        tail area of P; Z is accurate to about 1 part
           in 10**16.
        */
        if (p <= 0.00001) {
                val = -5;   // return a large negative
                    number (in std normal space)
        } else  if(p >= 0.99999) {
        // return a large number (in std normal space)
        val = 5;
} else if (fabs(q) <= 0.425) {/* 0.075 <= p <= 0.925
   */
r = 0.180625 - q * q;
val = q * ((((((r * 2509.0809287301226727 +
33430.575583588128105) * r + 67265.770927008700853) *
   r +
45921.953931549871457) * r + 13731.693765509461125) *
   r +
```

```
1971.5909503065514427) * r + 133.14166789178437745) *
   r +
3.3871328727963666608)
/ (((((((r * 5226.495278852854561 +
28729.085735721942674) * r + 39307.89580009271061) * r
    +
21213.794301586595867) * r + 5394.1960214247511077) *
   r +
687.18700749205790083) * r + 42.313330701600911252) * r
    + 1.0);
} else { /* closer than 0.075 from {0,1} boundary */
/* r = min(p, 1-p) < 0.075 */
if (q > 0.0)
r = 1.0 - p;
else
r = p;
r = sqrt(-log(r));
/* r = sqrt(-log(r))  <==>  min(p, 1-p) = exp( - r^2 )
    */
if (r <= 5.0) { /* <==> min(p,1-p) >= exp(-25) ~=
   1.3888e-11 */
        r += -1.6;
        val = (((((((r * 7.7454501427834140764e-4 +
        0.022723844989269184583) * r +
           0.24178072517745061177) *
        r + 1.27045825245236838258) * r +
        3.64784832476320460504) * r +
           5.7694972214606914055) *
        r + 4.6303378461565452959) * r +
        1.42343711074968357734)
        / (((((((r *
        1.05075007164441684324e-9 +
           5.475938084995344946e-4) *
        r + 0.015198666563616457196) * r +
        0.14810397642748007459) * r +
           0.68976733498510000455) *
        r + 1.6763848301838038494) * r +
        2.05319162663775882187) * r + 1.0);
} else { /* very close to  0 or 1 */
```

```
r += -5.0;
val = (((((((r * 2.0103343992922881326 5e-7 +
2.711555568743487578 15e-5) * r +
0.0012426609473880784386) * r +
   0.026532189526576123093) *
r + 0.29656057182850489123) * r +
1.7848265399172913358) * r + 5.4637849111641143699) *
r + 6.6579046435011037772)
/ (((((((r *
2.0442631033893997856 4e-15 + 1.4215117583164458887e-7)
    *
r + 1.8463183175100546818e-5) * r +
7.8686913114561325 91e-4) * r +
   0.014875361290850614852 5)
* r + 0.13692988092273580531) * r +
0.59983220655588793769) * r + 1);
}
if (q < 0.0) {
        val = -val;
}
}
return val;
}
**************************
//Function For evaluating bivariate Copula, given
   attribute x1 //and x2
**************************
// Obtain Attributes
double x1 = vLayers[0]->getValue(TRIndex, TCIndex,
   RIndex, CIndex);
double x2 = vLayers[1]->getValue(TRIndex, TCIndex,
   RIndex, CIndex);
// Calculate their assumed marginal PDF's
double dPDF1 = vPDFs[0]->getPDFResult(x1);
double dPDF2 = vPDFs[1]->getPDFResult(x2);
// Calcualte the uniform variable base on marginal CDF
   's
double dCDF1 = vPDFs[0]->getCDFResult(x1);
double dCDF2 = vPDFs[1]->getCDFResult(x2);
```

```
// Calculate Inverse normal for the CDF
double dI1 = NormalInverse(dCDF1);
double dI2 = NormalInverse(dCDF2);
// Evaluate Copula given parameter dRho called Theta
   in thesis
// formula
dRet = 1.0/(1.0 - dRho*dRho) * exp(-(dI1*dI1 - 2.0 *
   dRho* dI1*dI2 + dI2*dI2)/(2.0 - 2.0 * dRho* dRho))
* exp((dI1*dI1 + dI2*dI2)/2.0) * dPDF1 * dPDF2;
```

A.6 Input script for SPM

The following code is a typical input script for this thesis, that SPM reads to run a model

```
## Config file for SPM estimation for model selection
   for years 2003-2012
## The model spatial structure
@model
nrows 23
ncols 21
layer Base
## The population
categories stock
## Age structure
min_age 1
max_age 1
age_plus_group True

## Initialisation phase
initialisation_phases Phase1
## Model over years
initial_year 2003
current_year 2012
## Average cell size based on square cells. this is
   ignored when applying the new distance function
cell_length 111.2   ## Approx how many kilometers are
   in a Degree.
```

```
## Time steps in each year
time_steps one two three four five

## Growth of individuals at a certain age
age_size none
@age_size none     ### Not specifiing usually is Von
   Bertalanffy
type none

## Calculate biomass given length and age of fish
size_weight none
@size_weight none ## Not interested in biomass so is
   ignored - This is useful for managment
type none

## Describe initialisation phase
@initialisation_phase Phase1
years 1
time_steps initial_step_one
lambda 1e-10  ## This is the threshold that SPM uses
   as satisfying the initial popualtion has stabilised
    (see section 2.3)
lambda_years 1

## Processes that happen in initialisation
@time_step initial_step_one
processes Move_north recruitment

## Initialisation movement functions This is to seed
   the population in the north
@process Move_north
type preference
categories stock
preference_functions Move_North


@preference_function Move_North
type normal
```

```
alpha 1
layer Int_move
mu 100
sigma 1
## The layer Int_move only had values of 100 in the
   northern latitudes, there were zeros everywhere
   else

## Population processe : Note this is only applied in
   Initialisation phase
@process recruitment
type constant_recruitment
categories stock
proportions 1
r0 1000000 ## Seeds our fixed population
age 1
layer seed_recruits

## Annual Cycle in each model year
@time_step one # Summer
processes Movestock_sum

@time_step two # Autum
processes  Movestock_aut

@time_step three # Winter
processes Movestock_win

@time_step four # Spring
processes Movestock_spr

@time_step five ## reset Back up North (implementing
   migration outside of spatial state)
processes Move_north

## Define the moving processs using preference
   movement

@process Movestock_sum
```

```
type preference
categories stock
preference_functions stock_SST_sum

@process Movestock_aut
type preference
categories stock
preference_functions stock_SST_aut

@process Movestock_win
type preference
categories stock
preference_functions stock_SST_win

@process Movestock_spr
type preference
categories stock
preference_functions stock_SST_spr

## Preference functions
## type is the function layer is the preference
   attribute data followed by parameters
@preference_function stock_SST_sum
type logistic
alpha 1
layer sst_sum
a50 18
ato95 3

@preference_function stock_SST_aut
type logistic
alpha 1
layer sst_aut
a50 18
ato95 3

@preference_function stock_SST_win
type logistic
alpha 1
```

```
layer sst_win
a50 18
ato95 3

@preference_function stock_SST_spr
type logistic
alpha 1
layer sst_spr
a50 18
ato95 3

## Tell SPM where the neccassary layers are
@include "layers/Base.spm"
@include "layers/Depth_layer.spm"
@include "layers/Int_move.spm"
@include "layers/seed_recruits.spm"
@include "layers/obs_cat.spm"

### SST
@include "SST/sst_sum.spm"
@include "SST/sst_aut.spm"
@include "SST/sst_win.spm"
@include "SST/sst_spr.spm"

@include "SST/sst_03_sum.spm"
@include "SST/sst_04_sum.spm"
@include "SST/sst_05_sum.spm"

@include "SST/sst_12_spr.spm"


## Catchability coeffecient. This is the only other
   paramteter estimated
@catchability CPUEq
q 1e-5

############  Estimation Section ########
@estimation
minimiser simp
```

```
@minimiser simp
type numerical_differences       ## optimisation
   algorithm
iterations 500
tolerance 0.002
covariance True

## Parameter estimates with bounds to search over
@estimate
parameter catchability[CPUEq].q
type uniform
lower_bound 1e-7
upper_bound 1e-1

@estimate
parameter preference_function[stock_SST_sum].a50
same      preference_function[stock_SST_aut].a50
            preference_function[stock_SST_win].a50
   preference_function[stock_SST_spr].a50
type uniform
lower_bound 1
upper_bound 26

@estimate
parameter preference_function[stock_SST_sum].ato95
same      preference_function[stock_SST_aut].ato95
          preference_function[stock_SST_win].ato95
            preference_function[stock_SST_spr].ato95
type uniform
lower_bound 1
upper_bound 20


## Observation Section
## Attached is the following input for some observed
   year-standardised CPUE  with CV for August 2003
## Spatial cell is denoted by row = r and column =c
@observation aut_03_obs
```

```
type abundance
year 2003
time_step two ## Summer
## proportion time step 0.5 ## THink about that later.
catchability CPUEq
categories stock
selectivities One
layer obs_cat
likelihood lognormal
obs r20-c1 0.01
obs r17-c2 0.0127196915446577
obs r18-c2 0.0300135612558533
obs r19-c2 0.0552378544028614
obs r20-c2 0.0192991985426614
obs r21-c2 0.0109627878585166
obs r17-c3 0.0148178042906203
error_value r20-c1 1.567695
error_value r17-c2 1.73205080756888
error_value r18-c2 0.771761871280192
error_value r19-c2 1.21831255617915
error_value r20-c2 1.43309292475985
error_value r21-c2 1.37505166681712
error_value r17-c3 0.2
error_value r18-c3 1.14808239012968

### End of input config
```

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Allison, E., Perry, A., Badjeck, M., Adger, W., et al. (2009). Vulnerability of national economies to the impacts of climate change on fisheries. *Fish and Fisheries*, 10:173–196.

Anderson, O. and Dunn, M. (2011). Assessment of the mid-east coast orange roughy stock (orh 2a south, orh 2b and orh 3a) to the end of the 2009-2010 fishing year. New zealand fisheries assessment report 2011/62, National Institute of Water and Atmosphere.

Begg, G., Friedland, K., and Pearce, J. (1999). Stock identification and its role in stock assessment and fisheries management: an overview. *Fisheries Research*, 43:1–8.

Behrenfeld, M. and Falkowski, P. (1997). A consumer's guide to phytoplankton primary productivity models. *Limnology and Oceanography*, 42:1479–1491.

Bellman, R. (1958). "on a routing problem. *Quarterly of Applied Mathematics*, 16:87–90.

Booth, A. (2000). Incorporating the spatial component of fisheries data into stock assessment models. *Journal of Marine Science*, 57:858–865.

Botsford, L., Brumbaugh, D., Grimes, C., et al. (2009). Connectivity, sustainability, and yield: bridging the gap between conventional fisheries management and marine protected areas. *Rev. Fish Biology*, 19:69–95.

Bradshaw, C., Higgins, J., Michael, K., Wotherspoon, S., and Hindell, M. (2004). At-sea distribution of female southern elephant seals relative to variation in ocean surface properties. *Journal of Marine Science*, 61:1014–1027.

Carpenter, K. (1998). An introduction to the oceanography, geology, biogeography, and fisheries of the tropical and subtropical western and central pacific. Technical report, The Living Marine Resources of the Western Central Pacific.

Chandler, R., , and Clark, J. (2014). Spatially explicit integrated population models. *Methods in Ecology and Evolution*, 5:1351–1360.

Chen, X. and Fan, Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of econometrics*, 135(1):125–154.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. Wiley, Chichester, West Sussex, England.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151.

Clemens, H. (1961). The migration, age and growth of pacific albacore(*Thunnus germo*) 1951-1958. *California Department of Fish and Game, Bulletin*, 115:128.

Cope, J. and Punt, A. (2011). Reconciling stock assessment and management scales under conditions of spatially varying catch histories. *Fisheries Research*, 107:22–38.

Cosner, C., DeAngelis, D., Ault, J., and Olson, D. (1999). Effects of spatial grouping on the functional response of predators. *Theoretical Population Biology*, 56:65–75.

Cragg, J. (1971). Some statistical models for limited dependent variables with applications to the demand of durable goods. *Econometrica*, 39:829–844.

Dennis Jr, J. and Schnabel, R. (1996). *Numerical methods for unconstrained optimisation and nonlinear equations*. Classics in Applied Mathematics. Prentice Hall.

Dijkstra, E. (1959). a note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

Dunn, A., Rasmussen, S., and Mormede, S. (2015). *Spatial population model user manual, SPM*. Hobart, Australia, v1.1-2015-03-05 (rev. 1248) edition.

Elith, J. and Leathwick, J. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics*, 40:677–697.

Falkowski, P., Laws, E., Barber, R., and Murray, J. (2003). *Phytoplankton and Their Role in Primary, New, and Export Production*, chapter 4, pages 99–121. Springer.

Fernandez, D. (2012). Do winds control the confluence of subtropical and subantarctic surface waters east of new zealand? Master's thesis, School of Geography, Environment and Earth Sciences Victoria University of Wellington.

Fiedler, P. and Bernard, H. (1987). Tuna aggregation and feeding near fronts observed in satellite imagery. *Continental Shelf Research*, 7:871–881.

Flament, P., Kennan, S., Knox, R., Niiler, P., and Bernstein, R. (1996). The three-dimensional structure of an upper ocean vortex in the tropical pacific ocean. *Nature*, 383:610–613.

for Primary Industries, M. (2014). Fisheries assessment plenary, november 2014: stock assessments and stock status. compiled by the fisheries science group. Technical report, Ministry for Primary Industries, Wellington, New Zealand.

Ford, J. and Lester, R. (1956). Network flow theory. *California: RAND Corporation*.

Francis, C. (2000). Data weighting in statistical fisheries stock assessment models. *Canadian Journal Fisheries Aquaitc Science*, 68:1124–1138.

Frees, E. and Valdez, E. (1998). Understanding relationships using copulas. *North America Actuarial Journal*, 2(1):1–24.

Fujii, T. (2015). Temporal variation in environmental conditions and the structure of fish assemblages around an offshore oil platform in the north sea. *Marine Environmental Research*.

Gabow, H. (1983). Scaling algorithms for network problems. In *SFCS '83' Proceedings of the 24th Annual Symposium on Foundations of Computer Science*.

Gavaris, S. (1980). Use of a multiplicative model to estimate catch rate and effort from commercial data. *Canadian Journal of Fisheries and Aquatic Science*, 37:2272–2275.

Gimenez, O., Buckland, S. T., Morgan, B. J., Bez, N., Bertrand, S., Choquet, R., Dray, S., Etienne, M.-P., Fewster, R., Gosselin, F., et al. (2014). Statistical ecology comes of age. *Biology letters*, 10(12):20140698.

Graham, J. and Dickson, K. (1981). Physiological thermoregulation in the albacore *Thunnus Alalunga*. *Physiological Zoology*, 54(4):470–486.

Graham, J. and Laurs, R. (1982). Metabolic trate of the albacore tuna *Thunnus alalunga*. *Marine Biology*, 72:1–6.

Guan, W., Cao, J., Chen, Y., and Cieri, M. (2013). Impacts of population and fidhery spatial structures on fishery stock assesment. *Canadian Journal Fisheries and Aquatics Sciences*, 70.

Guide, M. U. (1998). The mathworks. *Inc., Natick, MA*, 5:333.

Haddon, M. (2010). *Modelling and quantitative methods in fisheries*. CRC press.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, Florida, USA.

Hilborn, R. and Walters, C. J. (1992). *Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty/Book and Disk*. Springer Science & Business Media.

Hofmann, E. and Powell, T. (1998). Environmeneal variability effects on marine fisheries:four case histories. *Ecological Applications*, 8:23–32.

Hoyle, S., Hampton, J., and Davies, N. (2013). Stock assessment of albacore tuna in the south pacific ocean. Wcpfc-sc8-2012/sa-wp-04-rev1, Western and Cntral Pacific Fisheries Commission.

Lan, K., Evans, K., and Lee, M. (2013). Effects of climate variability on the distribution and fishing conditions of yellowfin tuna (*Thunnus albacares*) in the wetern indian ocean. *Climate Change*, 119:63–77.

Lande, R., Engen, S., and Saether, B. (1999). Spatial scale of populations synchrony environmental correlation versus dispersal and density regulation. *American Naturalist*.

Langley, A. (2004). An examination of the influence of recent oceanographic conditions on the catch rate of albacore in the main domestic longline fisheries. Technical report, Seventeenth Meeting of the Standing Committee on Tuna and Billfish.

Laurs, R., Fiedler, P., and Montgomery, D. (1984). Albacore tuna *Thunnus alalunga* catch distribution relative to environment features observed from satellites. *Deep Sea Research Part I*, 31:1085–1099.

Lee, P., Chen, I., and Tzeng, W. (2005). Distribution of albacore (*Thunnus alalunga*) in the indian ocean and its relation to environmental factors. *Fisheries Oceanography*, 14:71–80.

Lehodey, P., Alheit, J., Barange, M., Baumgartner, T., Beaugrand, G., et al. (2006). Climate variability, fish, and fisheries. *Journal of Climate*, 19:5009–5030.

Lehodey, P., Senina, I., and Murtugudde, R. (2008). A spatial ecosystem and populations dynamics model (seapodym)–modeling of tuna and tuna-like populations. *Progress in Oceanography*, 78(4):304–318.

Lengendre, P. and Fortin, M. (1989). Spatial pattern ecological analysis. *Vegetation*, 80:107–138.

Levin, B., Lipsitch, M., and Bonhoeffer, S. (1997). Mathematical and computational challenges in population biology and ecosystem science. *Science*, 274:334–343.

Li, D. X. (1999). On default correlation: A copula function approach. *Available at SSRN 187289*.

Li, P. and Vu, Q. D. (2013). Identification of parameter correlations for parameter estimation in dynamic biological models. *BMC systems biology*, 7(1):91.

Limpert, E., Stahel, W., and Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51(5):341–352.

Lin, S. (1965). Computer solutions of the traveling salesman problem. *Bell System Tech Journal*, 44:2245–2269.

Lindberg, W., Frazer, T., Portier, K., et al. (2006). Density-dependent habitat selection and performance by a large mobile reef species. *Ecological Applications*, 16:731–746.

Lovett, G., Jones, C., Turner, M., and Weathers, K. (2005). *Ecosystem Function in Heterogeneous Landscapes*. Springer.

Maunder, M. and Punt, A. (2004). Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*, 70:141–159.

McCarty, J. (2001). Ecologicla consequences of recent climate change. *Conservative Biology*, 15(2):320–331.

McGregor, V. and Horn, P. (2013). Factors affecting the distribution of highly migratory species in new zealand waters. Technical report, New Zealand Aquatic Environment and Biodiversity Report No. 146.

Moore, K. and Abbott, M. (2000). Phytoplankton chlorophyll distributions and pimary production in the southern ocean. *Journal of Geophysical Research*, 105:709–722.

Mormede, S., Dunn, D., Parker, S., and Hanchet, D. (2013a). Further development of a spatially explicit population dyamics operating model for antarctic toothfish in the ross sea region. Technical report, CCAMLR, Hobart, Australia.

Mormede, S., Dunn, D., Parker, S., and Hanchet, D. (2013b). Investigation of potential biases in the assessment of antarctic toothfish in the ross sea fishery using outputs from a spatially explicit operating model. Technical report, CCAMLR, Hobart, Australia.

Murray, T. (1994). A review of the biology and fisheries for albacore, *Thunnus alalunga*, in the south pacific. *Fish.Tech.Pap*, pages 188–206.

Nakamura, H. (1969). *Tuna distribution and migration*. Fishing News, London, England.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84:487–493.

Patton, A. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review*, pages 527–556.

Power, J. (1996). Simulations of the effect of advective-diffusive processes on observations of plankton abundance and population rates. *Journal of Plankton Research*, 18:1881–1896.

Prince, J. and Hilborn, R. (1998). Concentration profiles and invertebrate fisheries management. In *Proceedings of the North Pacific Symposium on invertebrate Stock Assessment and Management*.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramon, D. and Bailey, K. (1996). Spawning seasonality of albacore, *Thunnus alalunga*, in the south pacicfic ocean. *Fishery Bulletin*, 94:725 – 733.

Ramos, A., Santiago, J., Sangra, P., and Canton, M. (1996). An application of satellite-derived sea surface temperature data to the skipjack (*Katsuwonus pelamis Linnaeus, 1758*) and albacore tuna (*Thunnus alalunga* bonaterre,1788) fisheries in the north-east atlantic. *International Journal of Remote Sensing*, 17:749–759.

Reynolds, R. and Smith, T. (1994). A high-resolution global sea surface temperature climatology. *Journal of Climate*, 8:1571–1583.

Ricker, W. E. (1954). Stock and recruitment. *Journal of the Fisheries Board of Canada*, 11(5):559–623.

Robson, D. (1966). Estimation of the relative fishing power of individual ships. *ICNAF Research Bulletin*, 3:5–14.

Schick, R., Goldtein, J., and Lutcavage, M. (2004). Bluefin tuna (*Thunnus thynnus*) distribution in relation to sea surface temperature fronts in the gulf of maine (1994-96). *Fisheries Oceanography*, 13:225–238.

Sekine, M., Imai, T., and Ukita, M. (1997). A model of fish distribution in rivers according to their preference for environmental factors. *Ecological Modelling*, 104:215–230.

Senina, I., Sibert, J., and Lehodey, P. (2008). Parameter estimation for basin-scale ecosystem-linked population models of large pelagic predators: application to skipjack tuna. *Progress in Oceanography*, 78(4):319–335.

Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Slack, E. (1969). A commercial catch of albacore (*Thunnus alalunga Bonnaterre*) in new zealand. Technical Report 46, New Zealand Marine Department Fisheries.

Smith, T. and Reynolds, R. (1998). A high-resolution global sea surface temperature climatology for the 196190 base period. *Journal of Climate*, 11:3320–3323.

Spencer, P. (2008). Density-independent and density-dependent factors affectingtemporal changes in spatial distributions of eastern bering sea flatfish. *Fisheries Oceanography*, 17:396–410.

Stelzenmuller, V., Rogers, S., and Mills, C. (2008). Spatio-temporal patterns of fishing pressure on uk marine landscapes, and their implications for spatial planning and management. *ICES Journal of Marine Sience*, 64.

Storn, R. and Price, K. (1995). Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces. Technical Report TR-95-012, International Computer Science Institute, Berkeley, CA.

Su, N., Sun, C., Punt, A., Yeh, S., and Dinardo, G. (2012). Incorporating habitat preference into the stock assessment and management of blue marlin (*Makaira nigricans*) in the pacific ocean. *Marine Freshwater Research*, 63:565–575.

Sutton, P. (2001). Detailed structure of the subtropical front over the chatham rise, east of new zealand. *Journal of Geophysical Research*, 106:31045–31056.

Tveito, O., Wegehenkel, M., Van der Wel, F., and Dobesch, H. (2006). The use of geographic information systems in climatology and meteorology. Final report cost action 719, Earth System Science and Environmental Management.

Wichura, M. J. (1988). Algorithm as 241: The percentage points of the normal distribution. *Applied Statistics*, pages 477–484.

Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002). *Analysis and management of animal populations*. Academic Press.

Xu, Y., Teo, S., and Holmes, J. (2013). Environmental influences on albacore tuna (thunnus alalunga) distribution in the coastal and open oceans of the northeast pacific: Preliminary results from boosted regression trees models. Technical Report SC/13/ALBWG/01, NOAA Fisheries.

Ying, Y., Chen, Y., Lin, L., and Gao, T. (2011). Risks of ignoring fish population spatial structure in fisheries management. *Canadian Journal of Fisheries and Aquatic Science*, 68:2101–2120.