Modelling the probability of capture for New Zealand's longfin eels (*Anguilla dieffenbachii*) and shortfin eels (*Anguilla australis*)

by

Anthony R. Charsley

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

Victoria University of Wellington
2019

Abstract

Longfin eel and shortfin eel probability of capture models can be used to build probability of capture maps. These maps can help identify eel encounter hotspots in New Zealand and are useful for managing and conserving the species. This research models longfin eel and shortfin eel presence/absence data using regularized random forest (RRF) models, vector-autoregressive spatial-temporal (VAST) models and Bayesian Gaussian random field (GRaF) models. Probability of capture maps built under VAST and GRaF remain approximately consistent with the maps built under RRF models. That is, longfin eels have high probabilities of capture around the coast of New Zealand's North Island and have low probabilities of capture throughout the centre of New Zealand's South Island. Shortfin eels have high probabilities of capture in small isolated regions of New Zealand's North Island and have very low probabilities of capture throughout most of New Zealand's South Island. Cross validation and spatial cross validation was used to compare the models. Cross validation results show that, compared to RRF models, VAST models improve predictive accuracy for the longfin eel and shortfin eel. Whereas, GRaF only improves predictive performance for the longfin eel. However, spatial cross validation shows no significant difference between VAST and RRF models. Hence, VAST models have higher predictive accuracy than RRF models for the longfin eel and shortfin eel when the training set is spatially correlated to the test set.

Acknowledgments

I would like to give thanks to the following people for the help they've given me.

I would like to thank my primary supervisor Dr. Nokuthaba Sibanda. Thank you for your support throughout my postgraduate studies and for your mentorship. I am very grateful for the help and guidance that you've given me.

Thank you to my co-supervisor Eric Graynoth. I appreciate the guidance you've given me throughout my thesis and your help in understanding the eel literature. I would also like to thank Dr. Simon Hoyle and Dr. Shannan Crow for your help in developing my project and for the ongoing support.

I would also like to thank the Ministry for Primary Industries (MPI) and the National Institute of Water and Atmospheric Research (NIWA) for your resources and your financial support. Additionally, I would like to thank Victoria University of Wellington for their co-funding.

Thank you to Ash, Ploi and Sam for their support throughout my postgraduate studies. "It's us against them".

I would like to thank my mum and dad who have always been there for me. Thank you for your love, support and guidance.

Lastly, I would like to thank all those that I have not mentioned explicitly by name. Your contribution is greatly appreciated, thank you.

Contents

Glossary

angaus The label used in the New Zealand Freshwater Fish Database (NZFFD) for the shortfin eel.

angdie The label used in the New Zealand Freshwater Fish Database (NZFFD) for the longfin eel.

areaswept The area over which the sample has taken place. Is a measure of the amount of effort put into sampling.

card A unique identifier for each record of the New Zealand Freshwater Fish Database (NZFFD).

catchability covariate A covariate which describes differences in catch rates between sampling occasions.

density covariate A covariate which describes variability in the density of a species in question.

nzsegment Identifies a segment of the River Environment Classification (REC).

organisation A categorical variable within the New Zealand Freshwater Fish Database (NZFFD) that identifies the organisation that has sampled a particular card.

River Environment Classification A database containing GIS variables relating to environmental classification at each of the nzsegments. The database does not vary temporally.

Acronyms

AUC area under the receiver operator characteristic curve.

BRT boosted regression tree.

DOC Department of Conservation.

ESA Eel Statistical Area.

GIS geographic information systems.

GRaF Gaussian random field.

MPI Ministry of Primary Industries.

NIWA National Institute of Water and Atmospheric Research.

NZFFD New Zealand Freshwater Fish Database.

QMS quota management system.

REC River Environment Classification.

ROC receiver operator characteristic.

RRF regularized random forest.

SPDE stochastic partial differential equation.

SSB spawning stock biomass.

TAC total allowable catch.

VAST vector-autoregressive spatio-temporal.

Chapter 1

Introduction

Species distribution models are used as a tool for estimating the distribution of a particular species of interest. These models can take spatial data on a species distribution, such as the occurrence of a species (known as presence/absence data) or the abundance of a species (i.e. species count or catch weight) at given locations, and relate this to geographical information of the locations (otherwise known as geographic information systems (GIS)) (Elith & Leathwick, 2009). These models are known as probability of capture models when dealing with freshwater fish such as *Anguilla dieffenbachii* (known as the longfin eel hereafter) and *Anguilla australis* (known as the shortfin eel hereafter).

Species distribution models can be implemented on a wide variety of species; including terrestrial, freshwater and marine species. These models can, among other things, identify species habitat and range, highlight the risk of invasive species to native species, help environmental managers design conservation areas (e.g. marine protected areas), and identify hot spots for species richness and decline (Martínez-Minaya et al., 2018).

Anthropogenic activities have modified New Zealand's landscape from pristine conditions (Gluckman, 2017). This has had a negative effect on freshwater quality and habitats that longfin and shortfin eels rely on (McDowall, 1990; Jellyman, 2012; Gluckman, 2017). This is reflected through

freshwater quality measures described by Gluckman (2017). Hence, knowledge on longfin and shortfin eel distributions are important for monitoring the species relative to anthropogenic changes. Additionally, estimates of the distributions of the species can aid conservation decisions by identifying areas of high and low occurrence.

1.1 Research objectives

The overall objective of this thesis is to develop models which estimate the probability of capture for New Zealand's longfin and shortfin eels. In order to achieve this, these general steps will be carried out:

1. A data set will be found and appropriate modelling techniques will be decided on,

2. The data set will be processed according to the needs of the modelling techniques,

3. Longfin and shortfin eel models will be constructed to estimate probability of capture,

4. Model comparisons and conclusions will be drawn.

The following section describes the longfin and shortfin eel modelling literature. The section gives detail on probability of capture models used for New Zealand freshwater fish and describes the modelling techniques to be used in this thesis. Following this, a thesis outline section is given. This section outlines the exact methods that will be used in this thesis for model building and comparison. It also gives an outline of the thesis structure.

1.2 Literature review

Probability of capture models describe the spatial distribution of a particular animal. Studies such as Leathwick et al. (2008b) and Crow et al. (2014) use machine learning approaches to make probability of capture estimates for many freshwater fish in New Zealand. Ecological studies tend to use machine learning approaches less frequently than traditional statistical methods (i.e. regression methods) (Elith et al., 2008). However, machine learning techniques such as the boosted regression tree (BRT) approach and the regularized random forest (RRF) approach offer advantages over traditional analysis. Machine learning techniques do not assume what the data-generating process is; instead they consider the process to be complex and unknown (Elith et al., 2008). By observing the inputs and the associated response, the machine learning algorithm tries to learn the response by finding dominating patterns (Elith et al., 2008). The following section discusses BRT and RRF probability of capture models.

1.2.1 Machine learning models

Boosted regression trees (BRT's)

Leathwick et al. (2008b) used boosted regression trees to predict the probability of capture for 30 New Zealand freshwater fish. The New Zealand Freshwater Fish Database (NZFFD) was used as the source of presence/absence data (see Chapter 2 for more detail) for each of the 30 fish species. Leathwick et al. (2008b) predicted the probability of capture using environmental predictors from the River Environment Classification (REC) GIS database (see Chapter 2). Each of these GIS predictors are associated with a segment of river (Leathwick et al., 2008b) which is known as a 'nzsegment'.

BRT models draw upon two algorithms, namely regression trees and boosting. Regression trees work by partitioning the possible values that

the predictors can take into rectangles (Elith et al., 2008). This is achieved by establishing a set of rules which determine rectangles of the predictors which produce similar responses (Elith et al., 2008). Then, a simple model is fit to each rectangle; this is usually a constant (Hastie et al., 2009) such as the mean response in the rectangle (Elith et al., 2008). A node of a regression tree represents a point at which a predictor may split a predictor space. The initial node contains all observations then this node may be split into two nodes by a predictor; these two nodes then contain a subset of the observations (Hastie et al., 2009). See Hastie et al. (2009) for more details on tree based methods and regression trees. Also see De'ath & Fabricius (2000) for more details on classification and tree based methods in the ecological context.

The theory behind boosting is that it is difficult to find one model that has highly accurate predictions, instead one can more easily find many models which, on average, produce accurate results (Elith et al., 2008). The study by Leathwick et al. (2008b) iteratively fits regression trees where a form of gradient descent is used to minimise a loss function (Elith et al., 2008). This builds up a network of, possibly, hundreds or thousands of trees. Elith et al. (2008) highlights how the procedure is stagewise because existing trees are unchanged but the fitted values for each observation are re-estimated at each step as a result of adding new regression trees. They describe how a BRT can be thought of as a regression model consisting of a linear combination of trees whereby each tree is a term in this regression model. See Hastie et al. (2009) for more details on boosting.

The results of the Leathwick et al. (2008b) study show that shortfin eels are more likely to be found in:

- Warm coastal, maritime environments with infrequent high intensity rain;
- Small sandy streams with unstable flows;
- Streams containing low downstream gradients;

- Streams containing low cover of native vegetation when in upstream catchments;
- Streams containing low riparian shading; and
- Streams containing high levels of nitrogen concentrations.

Leathwick et al. (2008b) found that longfin eels are more likely to be found in:

- Coastal and inland locations (especially of low gradient);
- Mild maritime climates having high intensity rain events of moderate frequency;
- Small, gravelly streams with unstable flows;
- Streams and catchments containing moderate downstream gradients;
- Streams and catchments containing very steep gradients; and
- Upstream catchments containing a moderate level of vegetation cover with a wide variety of riparian shading.

Leathwick et al. (2008b) used residual deviance and area under the receiver operator characteristic curve (AUC) to measure goodness of fit and prediction accuracy. The models for both the longfin and shortfin eels resulted in an AUC greater than 0.8 which indicate good prediction accuracy. Leathwick et al. (2008b) concluded that the estimates of probability of capture are subject to bias inherent in the NZFFD (see Chapter 2). They give particular reference to larger rivers and lakes which are difficult to accurately measure compared to small rivers and streams.

Regularized Random forests (RRF's)

Following the probability of capture estimates through BRT modelling, updates were made to the NZFFD and the REC (Crow et al., 2014). Updates to the geographical information formed an updated River Environment Classification known as REC2 (Crow et al., 2014). Additionally, the

updates improved the assignment of data from the NZFFD to river segments of the REC (Crow et al., 2014). See Crow et al. (2014) for further detail.

Crow et al. (2014) estimates the probability of capture under REC2 using regularized random forests. Crow et al. (2014) uses environmental, spatial and hydrological predictor variables to predict the probability of capture for 33 New Zealand freshwater fish. See Section 2.2.1 of this thesis for details on the model predictors.

A random forest consists of multiple decision trees such as the regression trees described in Deng & Runger (2013). Each tree is built around a bootstrap of the training data (Deng & Runger, 2013). Random forests use a technique known as bagging which reduces the variance of the overall model by taking an average of many complex but unbiased models (Hastie et al., 2009). A set of regression trees are established and the random forest takes a model average of these trees.

Deng & Runger (2012, 2013) describe RRF's as follows. Regularized Random Forests apply the tree regularization framework to the random forest framework described by Deng & Runger (2013). The tree regularization framework is a feature selection framework for decision trees (Deng & Runger, 2012). As always, the goal is to select a parsimonious model. In the case of decision trees, this is a tree consisting of the most compact set of features F (where $F \subset \{x_1, x_2, ..., x_M\}$ of M features) while retaining a strong predictive performance (Deng & Runger, 2012). Features of machine learning models are known as predictors in regression modelling. A feature x_ε not currently belonging to F is selected to belong in F if $\Lambda \times gain(x_\varepsilon)$ is greater than $\max_m gain(x_m)$ (where there are $m = 1, ...M$ features in the model) (Deng & Runger, 2012). The function $gain(\cdot)$ is a measure of information gain, Λ is a penalty taking a value between zero and one, and $x_m \in F$ (Deng & Runger, 2012). A larger penalty is implemented by setting Λ smaller (i.e. the feature is less likely to be selected).

At each node ϕ features are considered out of N possible features (Deng

& Runger, 2012), where the ϕ features are randomly selected and ϕ is given by $\phi = \sqrt{N}$ (Deng & Runger, 2012). Overall, B trees are built using the tree regularization framework. The features selected across all trees are given in the set F.

Crow et al. (2014) produced separate models under each sampling method (electric fishing, netting, trapping and visual) in order to minimise the influence of sampling methodology on catch rate. Similar to the approach of Leathwick et al. (2008b), Crow et al. (2014) used the area under the receiver operator characteristic curve (AUC) as a measure of model performance. In some cases, models for the same species but under different fishing methods produced very different results (Crow et al., 2014). Modelling by fishing method should be considered in future analysis in order to reduce sampling bias.

Crow et al. (2014) found that the predictions made were subject to the sampling bias inherent to the NZFFD (see Chapter 2). However, unlike Leathwick et al. (2008b), they attempted to account for biases associated with fishing methods by producing separate models under each sampling method. Additionally, they attempted to account for differences in sampling patterns between regions by including spatial predictors.

Lastly, Crow et al. (2014) concludes that their study produced very similar predictive performance compared to that of Leathwick et al. (2008b). Therefore, the conclusions made for the longfin and shortfin eels (outlined in the previous subsection) hold. Any differences between the two models are likely to be due to differences in the statistical model used and differences in the data sets (Crow et al., 2014). The next section addresses how we can improve on the probability of capture predictions for longfin and shortfin eels.

Gaps in probability of capture models

The study by Joy & Death (2004) was an early machine learning approach in predicting the probability of capture for New Zealand freshwater fish.

The study used artificial neural networks (ANN) to make these predictions. They confirmed that environmental variables and spatial variables have a strong influence on freshwater fish community structure. The study acknowledged the importance of GIS data in predicting fish assemblage for future studies.

The RRF model constructed by Crow et al. (2014) offered similar probability of capture predictions for New Zealand freshwater fish compared to that of Leathwick et al. (2008b). An issue addressed by Crow et al. (2014) was that RRF models do poorly in extrapolating outside of the space that was sampled. Hence, segments containing longfin and shortfin eels which aren't represented by the NZFFD will be poorly modelled. Additionally, anthropogenic activities such as habitat loss and degradation, land development (e.g. dam construction), and the drainage of wetlands (McDowall, 1990; Graynoth et al., 2008a) need to be accounted for. This can be done by accounting for temporal effects in the modelling process. However, Crow et al. (2014) did not consider these effects.

Crow et al. (2016) developed probability of capture models for New Zealand freshwater fish based on temporal changes. They found that accounting for temporal effects had a negligible impact on predicting shortfin eel presence but had a larger (but still a small) impact on predicting longfin eel presence. They used AUC values to measure this. When spatial, environmental, hydrological and methodological (i.e. method of sampling and organisation that sampled) variables were included in the model, they found AUC measures to be significant.

Based on the results of the RRF and BRT machine learning models and on Crow et al. (2016)'s temporal analysis, future modelling work should attempt to account for spatial and temporal effects, and spatial and temporal autocorrelation. The inclusion of environmental and hydrological predictor variables in the machine learning approaches allowed for longfin and shortfin eel habitat type and quality preferences to be accounted for. Future modelling work should also account for these eel preferences.

1.2.2 Stock assessment models

Longfin and shortfin eel stock assessment methods consist of two broad classes: conventional stock assessment models and GIS based models (Hoyle, 2016). Conventional stock assessment models have been developed for the longfin eel population in Southland, New Zealand by Dunn et al. (2009) and Fu et al. (2012).

Dunn et al. (2009) developed a model based on the age structure of longfin eels. They studied single-area and two-area spatial model structures and used these models to estimate virgin and current spawning stock biomass (SSB). Whereas, Fu et al. (2012) estimated virgin and current SSB through the development of a two-area spatial model based on the length structure of the longfin eel. However, both the length and age structured models made many assumptions on model input variables (Hoyle, 2016). These assumptions involved longfin eel growth; recruitment to Eel Statistical Area (ESA); ageing; density dependence; protected area range; and habitat (Hoyle, 2016). As a result, the Ministry of Primary Industries (MPI) rejected these models as a method of determining the current status of longfin eels (Hoyle, 2016).

GIS modelling methodologies have been developed for New Zealand longfin eels by Graynoth et al. (2008b) (GJB) and Graynoth & Booker (2009). Graynoth et al. (2008b) estimated eel biomass per km using the relationship between eel biomass per km, river flow and gradient. Generalised Additive Models were used but the estimates were inadequate for medium to large rivers (Graynoth et al., 2008b). Graynoth & Booker (2009) builds upon the GJB model by using 'weighted useable area' as a predictor variable in a Generalised Additive Model. Biomass estimates were then made for all rivers in New Zealand (including large rivers) (Graynoth & Booker, 2009).

The longfin eel stock assessment review by Hoyle (2016) concludes that GIS based methods are reasonable and offer an advantage over conventional models. This advantage stems from the fact that longfin and short-

fin eels are highly dependent on their habitat. GIS approaches account for this dependence (Hoyle, 2016).

Gaps in stock assessment models

The 2016 review of longfin eel stock assessment research by Hoyle (2016) addresses the gaps in knowledge with regards to longfin eel stock assessment work. The review concluded that the GIS approaches by Graynoth et al. (2008b) and Graynoth & Booker (2009) made invalid model assumptions due to gaps in knowledge that need additional work. Hoyle (2016) noted that sex ratios, variation between catchments and temporal variability were not considered in the modelling. Major issues associated with longfin and shortfin eel modelling are that eel populations tend to be highly fragmented, unmixed, and vary spatially in population parameters (in particular between fished and unfished locations) (Hoyle, 2016).

Hoyle (2016) recommends the development of additional research such as sex ratio models, and long term monitoring of spawning biomass of fished and unfished populations (fishery independent data) is continued. See Hoyle (2016) for additional details on stock assessment modelling.

1.2.3 Gaussian random fields

Gaussian random fields enable flexibility in the approach taken to fit the statistical model (Rasmussen & Williams, 2006) and have been used to model species' distributions within a frequentist (e.g. Thorson & Barnett (2017)) and Bayesian framework (e.g. Vanhatalo et al. (2012) and Golding & Purse (2016)). Gaussian random fields within the Bayesian framework often require Markov chain Monte Carlo (MCMC) methods for numerical approximations. However, this tends to require a great deal of computing power (Rue et al., 2009).

If we consider the spatial locations s within a spatial domain $\mathcal{D}$ ($\mathrm{s} \in \mathcal{D}$) of real numbers $\mathbb{R}^d$ ($\mathcal{D} \in \mathbb{R}^d$). Lindgren et al. (2011) defines a Gaussian

random field $\iota(\mathrm{s})$ as a joint distribution of all finite collections of $\{\iota(\mathrm{s}_i)\}$, where there are $i = 1, ..., n$ spatial locations. This joint distribution follows a multivariate Normal distribution with a specified expectation function $\mu(\cdot)$ and covariance function $C(\cdot, \cdot)$. The Gaussian random field has mean $\mu = \mu(\mathrm{s}_i)$ and a covariance matrix between spatial locations s_i and s_j of $\mathbf{\Sigma} = C(\mathrm{s}_i, \mathrm{s}_j)$ (Lindgren et al., 2011).

The system is said to be stationary if the covariance function only depends on the positions of two spatial locations (Lindgren et al., 2011). Additionally, the system is said to be isotropic if the covariance function is only dependent on the euclidean distance between locations (Lindgren et al., 2011). The covariance matrix of a Gaussian random field is often defined using a Matérn function.

1.2.4 Laplace approximation

The Laplace approximation is a deterministic method which has been developed to overcome computational restrictions (Rasmussen & Williams, 2006). Computationally restrictive methods such as Markov chain Monte Carlo (MCMC) are often used in Bayesian modelling because of its ability to handle complex models (Rue et al., 2009). Additionally, MCMC can reduce approximation error by increasing the number of iterations (Golding & Purse, 2016). Whereas, the Laplace approximation has a fixed error because of its deterministic approach (Golding & Purse, 2016).

A simulation approach such as MCMC is computationally expensive when dealing with Gaussian random fields (Rue et al., 2009). Hence, the deterministic Laplace approximation is often used for Gaussian random fields to lessen computation time.

1.2.5 Vector-Autoregressive Spatio-Temporal (VAST)

Thorson & Barnett (2017) proposed a vector-autoregressive spatio-temporal (VAST) model for modelling the population distribution of fisheries. The

approach uses a spatial delta generalised linear mixed model to model fisheries catch data b_i, where $i = 1, 2, ..., n$ observations. The approach can work for single categories or multiple categories of fish species sampled at different locations and time (Thorson & Barnett, 2017). The model is implemented in R (R Core Team, 2017) through the VAST R package (see `https://github.com/James-Thorson/VAST`) (Thorson & Barnett, 2017).

Thorson & Barnett (2017) and Thorson (2018) describe the VAST model as follows. Fisheries modellers aim to make measures of a fish species of interest. These measures can help fisheries modellers make inferences about the population. VAST can be a helpful tool to fisheries modellers because of the variety of functions it can perform. One potential outcome of modelling catch data with VAST is estimating fish density $\xi(s, c, t)$, where c is the category (i.e. fish species, taxon group etc.) which is found at the spatial location s at year t.

VAST decomposes the probability distribution of the catch data into two components. These components are 1. the probability of capture $\eta_1(i)$ and 2. the positive catch rates $\eta_2(i)$. These linear predictors are able to incorporate spatial, temporal and vessel effects; and density and catchability covariates. A density covariate is a covariate which accounts for variability in the density of the species in question and a catchability covariate is a covariate which describes differences in catch rates between sampling occasions. See Section 3.2 for details of the linear predictors with respect to the NZFFD data. Also see Thorson (2018) for full details on these linear predictors. The spatial and spatial-temporal components of these linear predictors are specified as Gaussian random fields by VAST (Thorson, 2018).

The user can control the link function of $\eta_1(i)$ and the observation model used for $\eta_2(i)$. An example of a link function used may be the logit-link:

$$\psi_1(i) = logit^{-1}(\eta_1(i)), \tag{1.1}$$

where the logistic function (i.e. inverse logit) is applied to $\eta_1(i)$ (Thorson,

2018) and $\psi_1(i) = Pr(b_i > 0)$. The term $\psi_1(i)$ gives the probability of capture for the ith observation and is what's of interest when modelling presence/absence data. In this case the model ignores $\eta_2(i)$ and $\psi_2(i)$ (described below).

When the user of VAST is interested in modelling abundance data then a potential model could be a delta lognormal model. If $\eta_2(i)$ was deemed to have a continuous support, e.g. biomass modelled with the Gamma distribution, then the probability distribution of the catch data is given by:

$$f(b_i) = \begin{cases} 1 - \psi_1(i) & b_i = 0 \\ \psi_1(i) \times g(b_i|\psi_2(i), \sigma^2_\chi(c)) & b_i > 0 \end{cases} \tag{1.2}$$

where $g(\cdot)$ is a probability density function for b_i and is a Gamma distribution for this example. But $g(\cdot)$ can be any probability density function (supported by VAST) with a continuous support above zero (Thorson, 2018). We define $\psi_2(i)$ as the mean of $g(\cdot)$ and $\sigma^2_\chi(c)$ as the variability of $g(\cdot)$ for category c. The term $\psi_2(i)$ is given by:

$$\psi_2(i) = a_i \times \exp(\eta_2(i)), \tag{1.3}$$

where the area swept for observation i is defined as a_i (Thorson, 2018). The terms $\psi_1(i)$ and $\psi_2(i)$ are parameterised so that $\mathrm{E}(b_i) = \psi_1(i) \times \psi_2(i)$. If the observation model is deemed to have a discrete support (e.g. count data) then the probability distribution of the catch data is given by:

$$Pr(b_i = B) = \begin{cases} (1 - \psi_1(i)) + g(0|\psi_2(i), ...) & B = 0 \\ \psi_1(i) \times g(B|\psi_2(i), ...) & B > 0 \end{cases} \tag{1.4}$$

where the terms are similarly defined and ... is used to indicate that there may be more terms depending on the probability mass function used (Thorson, 2018). Finally, quantities can be derived from these results. For example, a species (c) density can be found by:

$$\xi(s, c, t) = \psi_1(s, c, t) \times \psi_2(s, c, t) \tag{1.5}$$

where $\psi_1(s, c, t)$ and $\psi_1(s, c, t)$ are equivalent to $\psi_1(i)$ and $\psi_2(i)$ respectively but they do not incorporate catchability covariates (Thorson, 2018).

When using presence/absence data (binary data), VAST reduces to a logistic regression model which predicts probability of capture across a specified domain (Thorson, 2018). This model reduces the delta model to a single component $\eta_1(i)$, where the probability of capture is given by $\psi_1(i)$.

Thorson & Barnett (2017) used VAST to model multiple US Pacific Coast rockfishes. They then compared this multi-species model against relative single species models. They wanted to investigate whether or not accounting for correlation amongst species would result in improved biomass and fish distribution predictions. This is measured by the standard errors of the predictions and by measures of the overall predictive performance such as AIC.

Thorson & Barnett (2017) found the overall predictive performance improved when using the multi-species VAST model but the confidence intervals were estimated slightly wider (i.e. more uncertainty in the predictions) in the multi-species VAST model. The advantage of using VAST is its flexibility. The user can specify whether or not they would like a single species or a multi-species model. In some cases it may not make sense to use a multi-species model or it may over complicate the interpretation of the model (Thorson & Barnett, 2017). Additionally, the user can specify whether or not they would like to use certain modelling parameters. For example, the user can 'switch off' temporal variability. This means that the model will only account for spatial effects. Likewise, the user can incorporate important catchability covariates and/or density covariates. With respect to marine fisheries, the user is also able to account for 'vessel effects', i.e. the effect that different survey vessels have on the results. See Thorson (2018) for more details on the flexibility of VAST.

1.2.6 VAST parameter estimation

VAST begins by defining a spatial domain which is represented by a mesh (a set of knots which are connected by vertices). The user specifies the number of knots and then VAST uses a K-means algorithm (see Hartigan (1975)) to determine the location of these knots with respect to the sampling locations (Thorson, 2018). The algorithm positions knots by minimising the total distance between the sampling locations and the knots (Thorson, 2018).

The stochastic partial differential equation (SPDE) approach is used to approximate a Gaussian random field as the solution to the SPDE:

$$(\zeta^2 - \Delta)^{\varphi'/2}\iota(\mathbf{u}) = W(\mathbf{u}), \quad \mathbf{s} \in \mathbb{R}, \tag{1.6}$$

where $\iota(\mathbf{u})$ is the Gaussian field of interest, ζ is the spatial scale parameter, Δ is the Laplacian, φ' is a smoothness parameter and $W(\mathbf{u})$ is spatial Gaussian white noise (Lindgren et al., 2011). The solution is found through an approximation of the SPDE which involves generating a triangulated mesh (spatial domain) with vertices (the corners of the meeting triangles) at each knot (Thorson, 2018; Lindgren et al., 2011). VAST implements the R software package R-INLA (`www.r-inla.org`) (Rue et al., 2009) to do this. Next, the solution to the SPDE is found through the construction of a basis representation:

$$\iota(\mathbf{u}) = \sum_{a=1}^{A} \upsilon_a(\mathbf{u})\Xi_a, \tag{1.7}$$

where $\upsilon_a(\mathbf{u})$ are basis functions and are equal to 1 at vertex a and 0 otherwise (Lindgren et al., 2011). The term Ξ_a is a Gaussian distributed weight which determine the values of the field at the vertices and the values not on a vertex are determined by linear interpolation (Lindgren et al., 2011).

Model fixed effects are estimated using maximum likelihood estimation, where maximum likelihood estimates are found by integrating the joint likelihood of the fixed effects with respect to the random effects (Thor-

son, 2018). The integral is defined as:

$$L(\boldsymbol{\theta}) = \int_{\epsilon} P(\mathbf{D}|\boldsymbol{\theta}, \boldsymbol{\epsilon}) P(\boldsymbol{\epsilon}|\boldsymbol{\tau}) d\boldsymbol{\epsilon}, \tag{1.8}$$

where $L(\boldsymbol{\theta})$ is the marginal likelihood which we seek to maximise to estimate the fixed effects $\boldsymbol{\theta}$ (Skaug & Fournier, 2006). The term $\boldsymbol{\tau}$ are the parameters governing the distribution of random effects $\boldsymbol{\epsilon}$, $P(\mathbf{D}|\boldsymbol{\theta}, \boldsymbol{\epsilon})$ is the probability of the data D conditional on the fixed and random effects and $P(\boldsymbol{\epsilon}|\boldsymbol{\tau})$ is the probability of the random effects conditional on the parameters governing their distribution. The integral is approximated using the Laplace approximation. Full details on this approximation is given in Section 3.2.4.

The Laplace approximation is implemented through VAST which uses Template Model Builder (Kristensen et al., 2015) to make the approximations (Thorson, 2018). Maximum likelihood estimates are made through an R optimisation method and a generalisation of the delta method is used to make standard error estimates (Kass & Steffey, 1989).

1.2.7 The Gaussian random field (GRaF) model

The Gaussian random field (GRaF) model was proposed by Golding & Purse (2016) as a species distribution model. The GRaF model is built under the assumption that similar covariate values will result in similar response values (Golding et al., 2013). Hence, the model is built based on the similarity or dissimilarity between the sampled locations (Golding et al., 2013).

GRaF models have a hierarchical structure which involves inference over latent variables and hyperparameters (Golding & Purse, 2016). These models can be defined using a Bayesian framework or a classical statistical framework and posterior computation can be implemented using either Laplace approximation or an expectation-propagation (EP) algorithm (Golding & Purse, 2016).

GRaF models are highly flexible in how they can be specified. When using a Bayesian framework, prior knowledge can be incorporated into the model which enables the user to incorporate any information they may have about how the probability of capture $\mathbf{q}$ changes with model covariates. As an example we may want to model a freshwater fish species which is highly sensitive to pollution. If we have pollution measures (covariates) in the waterways of interest then we could incorporate our knowledge of pollution sensitivity into a GRaF model. Hence, the user could give a prior which describes probability of capture for the freshwater fish species of interest to be low in waterways with high levels of pollution.

GRaF models were first implemented for describing the spatial distribution of vector mosquitoes in the United Kingdom (Golding et al., 2013). The models made use of a Bayesian framework to incorporate expert knowledge on mosquito assemblage. This enabled the first high resolution spatial maps of vector mosquitoes in the United Kingdom to be constructed (Golding et al., 2013). The results of this study and that of Golding & Purse (2016) showed that GRaF models outperform many of the previously used species distribution models, including BRT machine learning models. The success of these GRaF models has been attributed with their ability to allow for a range of complex functions through specification of a covariance function (Golding & Purse, 2016).

Following the notation of Golding & Purse (2016), presence/absence data that has been collected from sampling sites are represented by the vector $\mathbf{y}$, where there are $i = 1, ..., n$ observations. Hence,

$$\mathbf{y} \sim Bern(\mathbf{q}), \tag{1.9}$$

where $\mathbf{q}$ is a vector for the probability of capture of each observation and

$$\mathbf{q} = \text{Probit}(\mathbf{z}). \tag{1.10}$$

The vector $\mathbf{q}$ is given by the probit transformation of the latent variable $\mathbf{z}$. A latent variable is found for each observation and $\mathbf{z}$ is a Gaussian random

field. The latent variables are defined by a user specified prior (when using Bayesian estimation).

A squared exponential term is used to define the covariance of the Gaussian random field because it produces smooth curves which are considered ecologically plausible (Golding & Purse, 2016). The covariance is further defined by a hyperparameter l_g known as a lengthscale where $g = 1, ..., n_g$. A lengthscale defines the correlation between probability of capture and covariate values (Golding & Purse, 2016). One must be given (or estimated) for each covariate. A small lengthscale indicates strong correlation and a large lengthscale indicates weak correlation. The natural log of the lengthscale $\ln(l_g) = \Phi_g$ is given by:

$$\Phi_g \sim N(\mu_\Phi, \sigma_\Phi^2), \tag{1.11}$$

where μ_Φ and σ_Φ^2 are user specified hyperparameters for the mean and variance of the Normal distribution. The prior over the mean function defines how the probability of capture changes with a covariate and the lengthscale defines how rapidly this change occurs (Golding & Purse, 2016). Hence, the priors work to describe the ecology of the species to the model (Golding & Purse, 2016).

1.2.8 Methods for model validation

The purpose of model validation is to assess how well a model fits the data and how well a model makes predictions. The former is usually tested by examining the residuals of a model and assessing whether or not there are underlying patterns in the residuals which the model has not accounted for. However, when assessing model predictions, there are a number of validation techniques which could be used. Additional considerations must also be made when dealing with spatially and temporally dependent data.

K-fold cross validation is a commonly used technique for assessing a models predictive ability. The method works by dividing the data set into

'K' groups (known as folds hereafter) randomly. 'K' is typically 10 but the user can define this as any number. A model is built using a training set of K-1 folds and then predictions are made to the fold which wasn't used, i.e. the test set. This is repeated until every fold is used as a test set and K models are built. The predictions made for each test set can be assessed against the observed values in each test set. The method for assessing these predictions is dependent on the data. This thesis uses presence/absence data which is typically assessed using the area under the receiver operator characteristic curve (AUC).

K-fold cross validation doesn't account for spatial and/or temporal dependence in a data set. For these data sets K-fold cross validation results in overly optimistic evaluation estimates (Mosteller & Tukey, 1977; Picard & Cook, 1984). When the data is spatially and/or temporally dependent, K-fold cross validation assesses how well the model performs when the training set contains spatial information about the test set. Spatial K-fold cross validation can be used to account for this lack of independence.

Rather than dividing data into folds randomly, spatial K-fold cross validation uses a K-means algorithm to spatially partition data into K clusters that maximise spatial correlation within clusters, while minimising spatial correlation between clusters. Each cluster is then used as a fold in the cross validation. This is designed to ensure that the training data set and the test data set have as little autocorrelation as possible (Ruß & Brenning, 2010). This means that the test sets are approximately independent of the training set which is an underlying assumption of any cross validation method (Pohjankukka et al., 2017). Hence, the results of a spatial K-fold cross validation indicate how well a model performs in a spatially distinct location to the training data.

A potential shortcoming of K-fold cross validation and spatial K-fold cross validation is that multiple models need to be run. This means that models which are computationally intensive and take a long time to compute would take a long time to cross validate. This is often a restriction

when dealing with spatial or spatio-temporal models. Hence, K-fold cross validation may need to be applied with a low number for K (less models).

1.2.9 Longfin and shortfin eel biology and importance

This thesis concentrates on the New Zealand native shortfin eel (*Anguilla australis*) and the endemic longfin eel (*Anguilla dieffenbachii*). The two species are catadromous, meaning that they predominately live in freshwater and migrate to the ocean at the end of their lives to spawn and then die (semelparous) (Ministry of Primary Industries, 2014). They coexist but shortfin eels tend to prefer lowland waterways whereas longfin eels may occupy waterways that are at longer distances inland and high country (Jellyman, 2012). Both adult shortfin and longfin eels prefer slow flowing water (Jellyman, 2012). Shortfin eels prefer finer sediment environments whereas longfin eels prefer coarser environments such as gravels and boulders (Jellyman, 2012; Jellyman et al., 2003). However, the species are known to coexist (Jellyman, 2012) hence there is potential for correlation in the spatial distribution of the two species. Therefore a multi-species model (a model which accounts for the correlation between multiple species) may have better predictive performance then separate single species models.

The size and weight distribution of the two eel species are very different; longfin eels can reach a maximum length of 2m and a weight of 25kg or greater, whereas shortfin eels can only reach a maximum length of 1.1m and a weight of 3kg (Jellyman, 2003; Graynoth & Taylor, 2005). However, eel growth rate is highly dependent on environmental influence such as eel density, water temperature and food availability (Graynoth & Taylor, 2005).

Both species are long lived where South Island shortfin eels take c. 12.8 years to reach the minimum legal fishing size and South Island longfin eels take c. 17.5 years (Ministry of Primary Industries, 2014; McDowall, 1990). In comparison, the equivalent times for the North Island shortfin

and longfin eels are c. 5.8 years and c. 8.7 years (Ministry of Primary Industries, 2014; McDowall, 1990). However, the Ministry of Primary Industries (2014) expresses that these measures are highly variable.

Augilla sex differentiation occurs during their 'yellow eel' life cycle phase where wild populations tend to have highly skewed sex ratios (Davey & Jellyman, 2005). Sex differentiation is said to be influenced by environmental conditions and eel density (Colombo & Grandidr, 1996; Beullens et al., 1997). As a result, sex ratios vary significantly at all scales and therefore vary between islands and between habitat types within a waterway (Hoyle & Jellyman, 2002).

Longfin and shortfin eels migrate to the sea at the end of their lives. The age at which an eel migrates (known as age at migration) is dependent on the growth rate of the eel (McDowall, 1990). This is because migration is thought to be dependent on reaching a certain combination of length and weight (McDowall, 1990). Male longfin and shortfin eels are known to migrate at a smaller size than female longfin and shortfin eels (McDowall, 1990).

Longfin and shortfin eels serve important intrinsic, ecological, customary and commercial purposes (Jellyman, 2012). These eels are intrinsic to New Zealand and are important to New Zealand's heritage (Jellyman, 2012). *Augilla* serve an important ecological role in New Zealand waterways. Longfin eels prey on introduced and native freshwater fish (including eels) and play an important role in controlling fish populations (Jellyman, 2012). Chisnall et al. (2003) found that when larger longfins were removed from a waterway, smaller longfin and shortfin eels moved into the waterway in large numbers. Hence, the larger longfins were controlling the presence of smaller longfin and shortfin eels.

Longfin eels are described as ecological generalists (Glova et al., 1998; Jellyman et al., 2003). One of the main reasons for this description is because of their highly variable diet which consists of whatever food is available (Jellyman, 2012). Smaller longfins feed on aquatic invertebrates and

once they are large enough they will begin feeding on fish (Jellyman, 2012).

Longfin and shortfin eels are culturally significant to Māori. Māori eel fisheries have been established in recognition of this (Jellyman, 2012). New Zealand longfin and shortfin eels are managed commercially through a quota management system (QMS). The QMS was introduced to South Island eels in 2000 and North Island eels in 2004 (Jellyman, 2012). A total allowable catch (TAC) has been set for longfin and shortfin eels in the North Island and for longfin and shortfin eels combined in the South Island (Jellyman, 2012). Figure 1.1 shows the quota management areas in New Zealand. These areas are further broken down into ESA's used for reporting eel commercial catch (Jellyman, 2012). See Jellyman (2012) for further details on New Zealand's eel fishing industry.

1.3 Thesis outline

The NZFFD will be used as the source of data for this thesis. The database gives spatially and temporally extensive presence/absence data for longfin and shortfin eels as well as model covariates. Probability of capture estimates will be made on this data using RRF models, VAST models and GRaF models. Covariates will be selected by the RRF model and then VAST and GRaF models will be built on these covariates. This ensures that all models are built on the same covariates and enables model comparisons. Models will be compared using K-fold cross validation techniques which estimate AUC.

The following chapter details the data used for modelling and the procedure that was followed in the data processing stage. Chapter 3 describes how each of the models cross validation techniques work. Chapter 4 gives the results of the models, their probability of capture predictions and the model comparisons results. Lastly, Chapter 5 gives the discussion and conclusion. This chapter discusses the findings of the thesis, possible problems with the research, future research considerations and the conclusions

that can be made from the research.

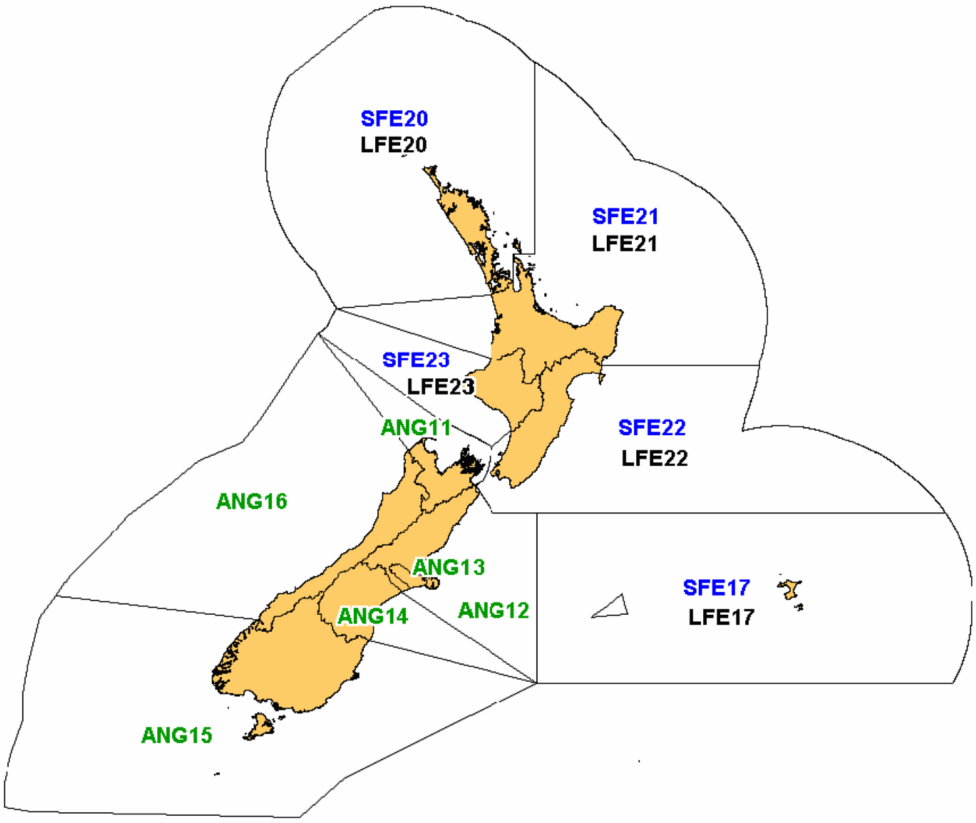


Figure 1.1: A map of the quota management areas in New Zealand. The South Island areas ANG denotes the combined longfin and shortfin eel stocks. The North Island areas and the Chatham Islands area LFE denotes the longfin eel stocks and SFE donotes the shortfin eel stocks.

Source: Jellyman (2012)

Chapter 2

Data

This section examines the longfin and shortfin eel data used. We begin by discussing the New Zealand Freshwater Fish Database (NZFFD) and its strengths and weaknesses.

2.1 The New Zealand Freshwater Fish Database (NZFFD)

The NZFFD is a voluntary database containing the records of the occurrence of New Zealand freshwater fish (Richardson, 2005). Organisations such as the National Institute of Water and Atmospheric Research (NIWA), the Department of Conservation (DOC), New Zealand regional councils, tertiary institutes, and other private and public organisations have contributed data towards the NZFFD voluntarily. The database contains mainly freshwater fish presence/absence data (Hoyle, 2016) along with other variables such as site location (Northing and Easting) (Richardson, 2005) and a REC identifier (known as nzsegment). The fish sampling method is also given, where the most common methods used are electric fishing, fish trapping and visual inspections. See Richardson (2005) for a full account on how to use the NZFFD and see Joy et al. (2013) for a guide on freshwa-

ter fish sampling protocols.

Data is recorded on the NZFFD voluntarily; this introduces selection bias. Sites which are 'harder' to sample are likely to be sampled at a lower rate than a site that is 'easier' to sample. For example, sites which are easily accessible from a road may be sampled at a higher rate than sites further away from roads. As a result, model estimates will have a stronger representation of the sites which were more easily accessed.

Lakes and large rivers are impossible to sample through electric fishing methods. Therefore, estimates of probability of capture are not representative of longfin or shortfin eel occurrence in large rivers or lakes. Leathwick et al. (2008b) notes that the NZFFD not only under represents large rivers but also partially saline waters. This is due to the difficultly in electric fishing large rivers and saline waterways (Leathwick et al., 2008b).

A disadvantage of the NZFFD is that the data has been collected by different organisations, often for different purposes (Jowett & Richardson, 2003). As a result, an organisation may put more effort or less effort into sampling a longfin or shortfin eel from a waterway, depending on the organisation's objective.

Administrators for the NZFFD perform quality control checks for each entry submitted to the NZFFD (Crow et al., 2016). This is designed to reduce human error and to ensure that entries are complete (Crow et al., 2016).

2.2 The longfin and shortfin eel data

The data being used for this research is provided by NIWA and was used in the 2014 study by Crow et al. (2014). The data originates from the NZFFD and contains presence/absence data for all known native freshwater fish. There are different data sets for each sampling method used. This thesis focuses on the electric fishing data set.

The exact date of fishing was recorded in the data set along with spa-

tial information (i.e. the REC identifier known as 'nzsegment' and New Zealand Transverse Mercator (NZTM) easting and northing coordinates). This research incorporates temporal effects and therefore any records missing their exact date of sampling are excluded. Additionally, the records taken before 1972 are excluded because data wasn't recorded on a yearly basis before 1972.

The years 1972 and 1973 are also excluded because 1973 contained no encounters for shortfin eels, therefore 1972 is excluded to have data on a yearly basis. This is done because each year of data must contain at least one encounter and non-encounter in order for the VAST modelling software to calculate probabilities of capture. Hence, the final data set contains data spanning from 1974 to 2014. Figures 2.1 and 2.2 show the sampling location for each year of the data set. Shortfin eels appear to have a large number of absences whereas longfin eels appear to have a large number of presences.

Figures 2.3 and 2.4 display the observed proportion of longfin and shortfin eels respectively. The observed proportions were calculated within each NZMS 260 map series grid square (30km by 40km northing by easting) using NZFFD presence/absence data measured between 1974 and 2014. Each observation of the NZFFD corresponds to a grid square of the NZMS 260 map. Therefore, the proportion of eels encountered in each grid square was calculated and each point was then plotted with a colour corresponding to the proportion calculated.

Figure 2.3 shows high observed proportions of presence of longfin eels throughout the North Island of New Zealand. This is particularly true on the North Island's coast. The Auckland region of New Zealand shows a low proportion of presence for New Zealand longfin eels. The central areas of New Zealand's South Island shows low proportions of longfin eel presence whereas the central west coast of the South Island shows higher proportions of longfin presence.

Figure 2.4 shows that shortfin eels are unlikely to be observed in the

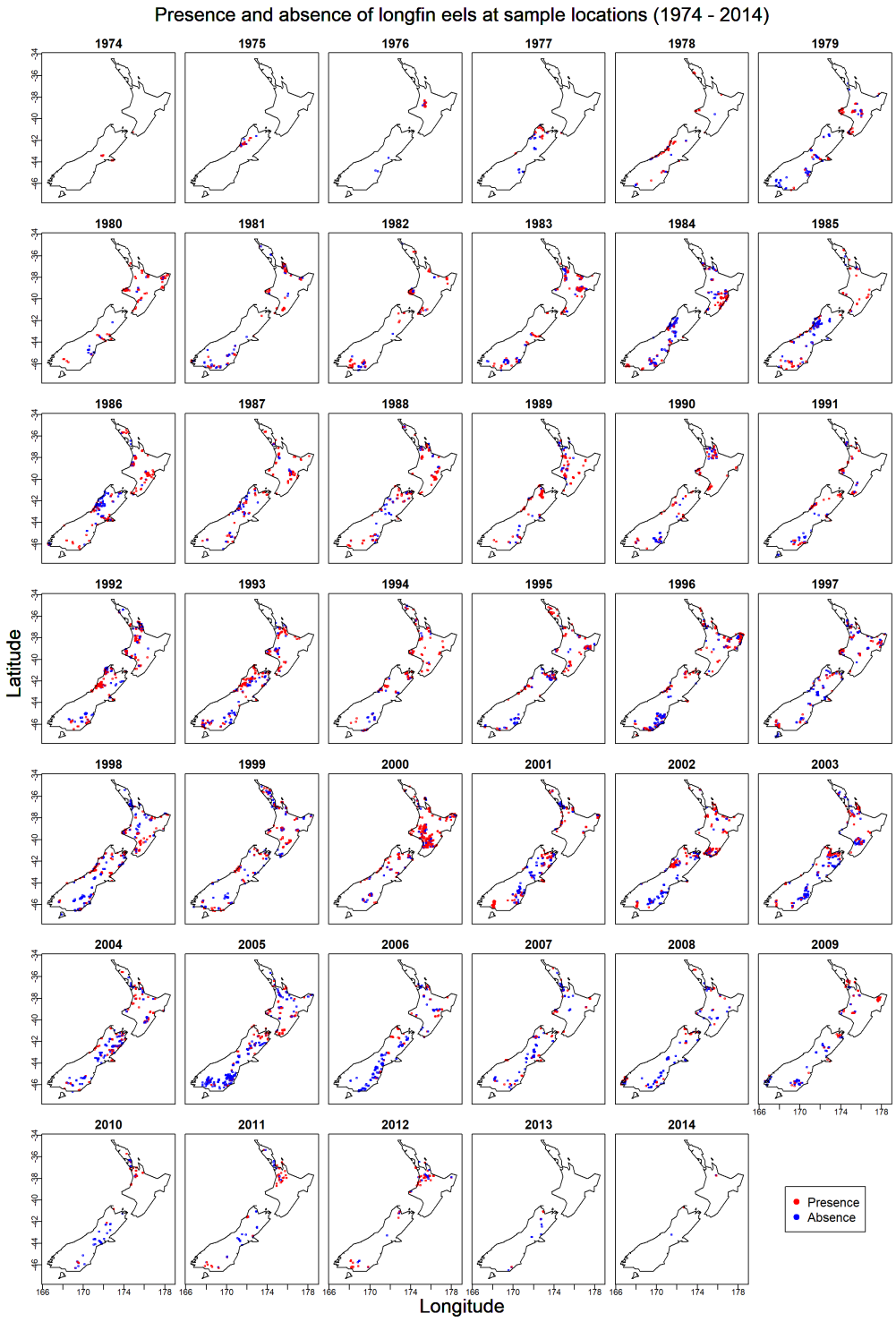


Figure 2.1: Presence/absence of longfin eels by sample locations from 1974 to 2014. The data is displayed on a latitude-longitude grid.

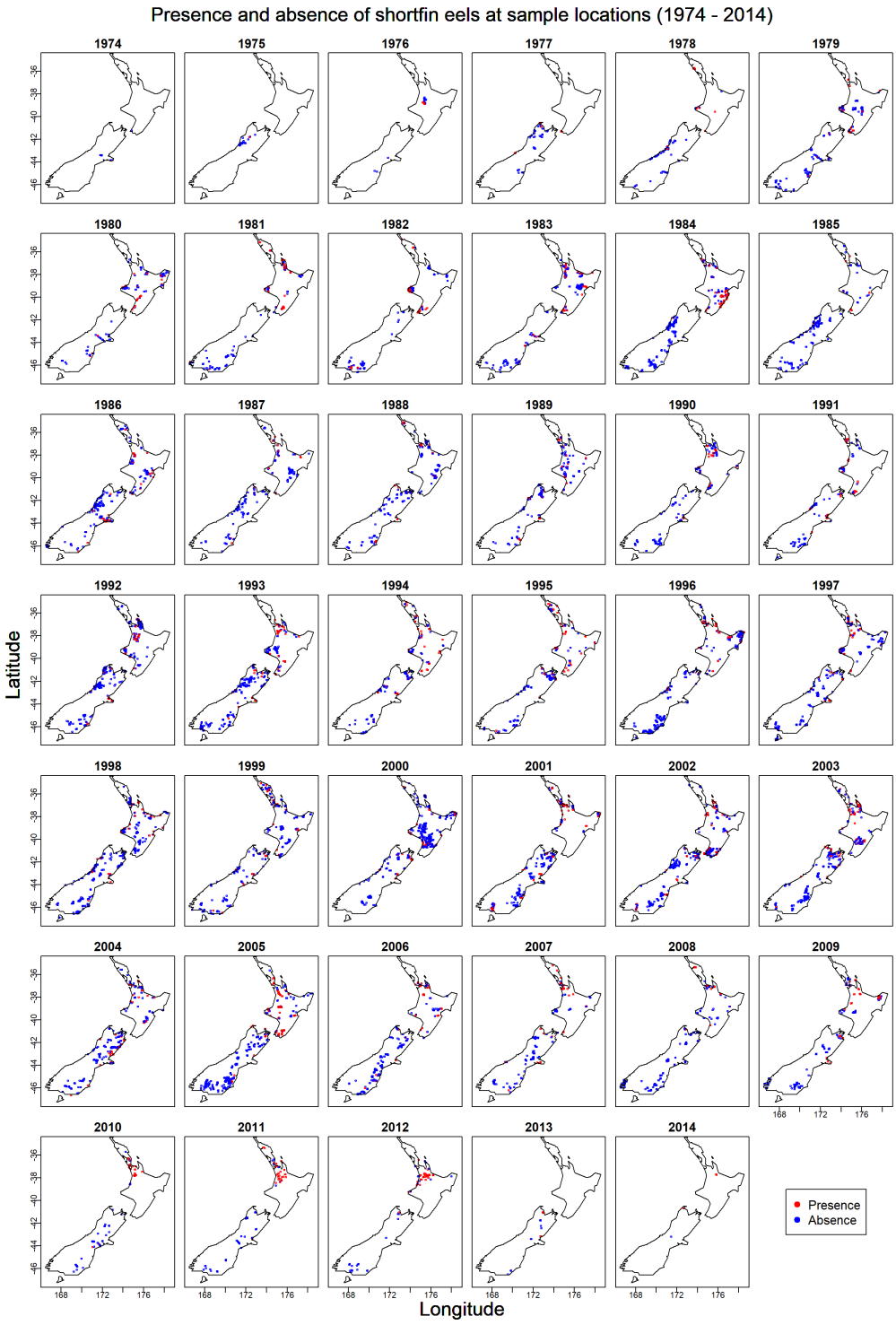


Figure 2.2: Presence/absence of shortfin eels by sample locations from 1974 to 2014. The data is displayed on a latitude-longitude grid.

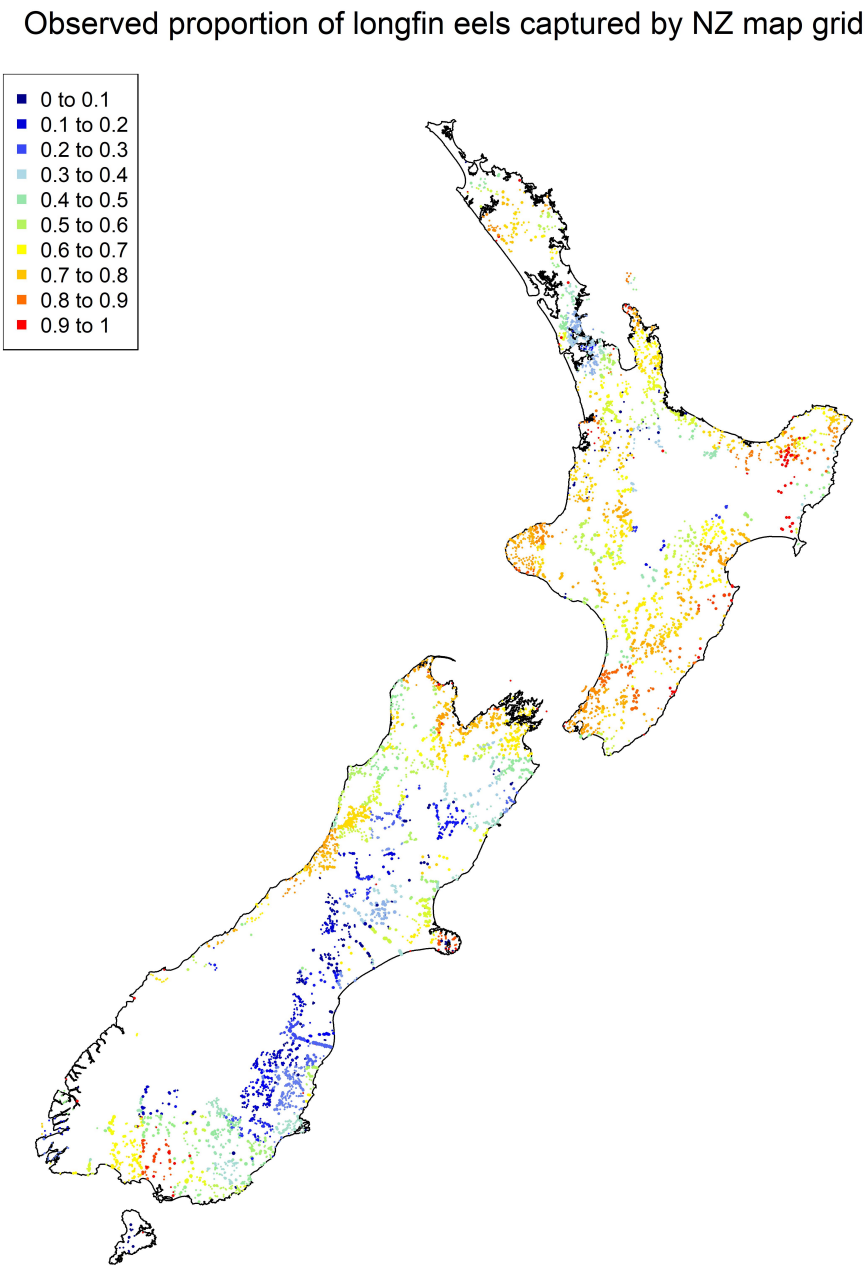


Figure 2.3: Map of the observed proportion of longfin eels captured within each NZMS 260 map series grid square. The data comes from the NZFFD and was measured between 1974 to 2014. Blank areas are unsampled.

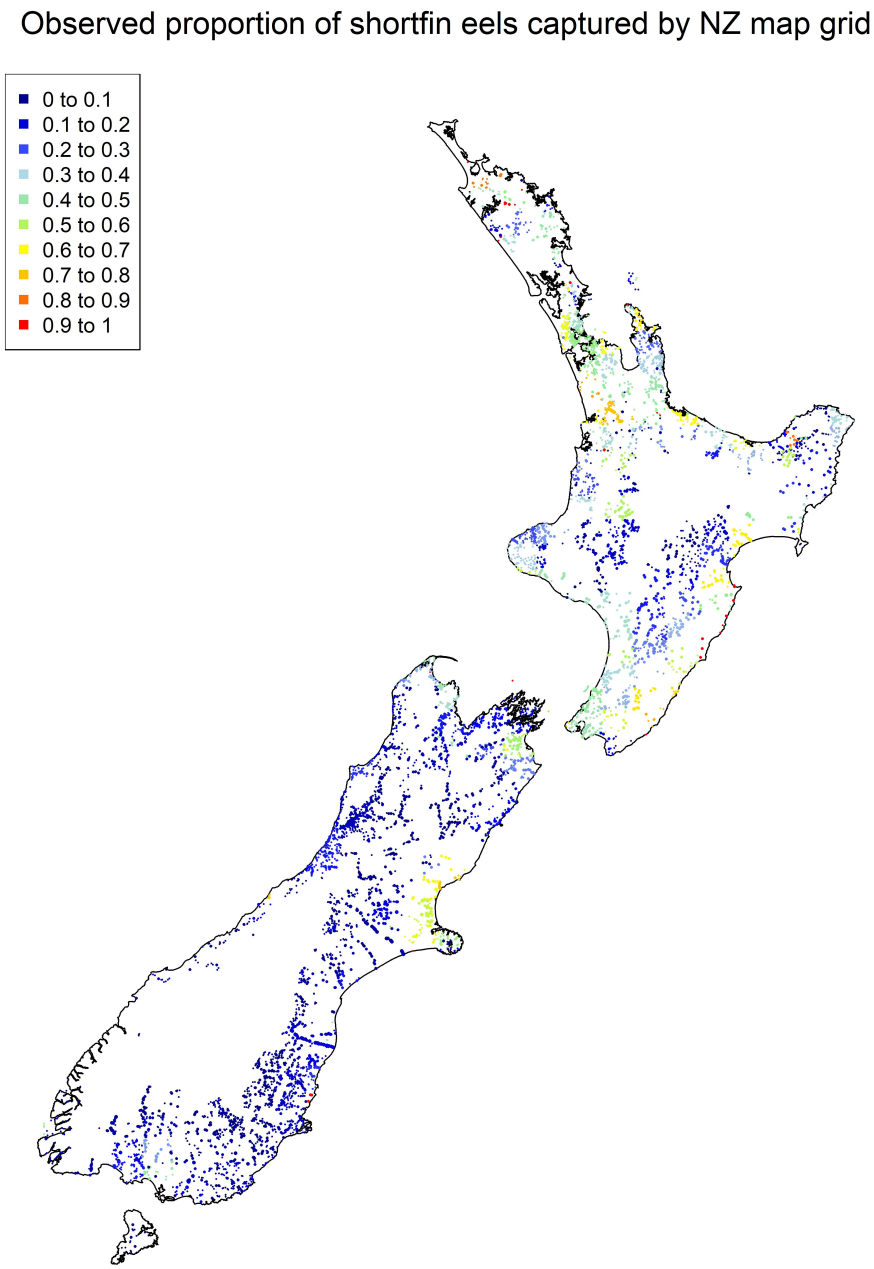


Figure 2.4: Map of the observed proportion of shortfin eels captured within each NZMS 260 map series grid square. The data comes from the NZFFD and was measured between 1974 to 2014. Blank areas are unsampled.

central North Island of New Zealand. But the proportion of shortfin eels observed increases slightly in the North Island's coast, in particular, the south-east coast of New Zealand's North Island. Shortfin eels have a low proportion of presence throughout New Zealand's South Island. However, Christchurch and the surrounding area have higher proportions of observed presence of approximately 0.5 to 0.7.

There are various cards (a unique identifier for each record of the NZFFD) in the data set that were sampled on the same exact date and location (nzsegment). This means that for any given location and date there may be multiple records. This may be due to an organisation taking various samples at the same location and date or multiple organisations taking samples at the same location and date (or both). Only one card is kept if the samples (occurring at the same location and date) were taken by the same organisation. But if the samples (occurring at the same location and date) were taken by different organisations then one card is selected from each organisation. This is a result of the assumption that the same organisations sample with the same catch rate whereas different organisations sample with different catch rates. Hence, it is important to keep replicates if the organisation were different and to include the organisation variable in the model.

Only one card of the samples occurring at the same location and date, and sampled by the same organisation was taken. This is because the covariates of each card are the same for each location. This occurred because the established covariates of Leathwick et al. (2008b) and Crow et al. (2014) were taken from a GIS database and were therefore not sampled independently at each visit. Hence, using only one card from each replicate location and date (given that they are sampled from the same organisation) ensured that an organisation is not overrepresented.

The NZFFD contains many cards taken by an organisation on the same day and location. This is because an organisation has decided to take multiple-pass electric fishing samples. As an example, Jowett & Richard-

son (1996) takes single-pass (otherwise known as first-pass) and multiple-pass electric fishing samples to compare single-pass catches and multiple-pass population estimates. The use of multiple-pass sampling is often used as a measure of freshwater fish depletion. In the context of this thesis, we are interested in obtaining the results of the first-pass because the first-pass is a sample of the location that is unfished for that day. Hence, we would like to select the first sample of the day. The 'time' variable (in 24 hours) is used to distinguish the earliest record of the day. Where possible, the earliest record of the day is selected and the others (for that day, location and organisation) are removed from the data set.

It is not always possible to distinguish records by time as the 'time' variable contains missing values or, in some cases, 'time' was recorded as a categorical variable such as 'day' or 'night'. In these cases the card is determined by selecting the card with the largest number of fish presence's across all freshwater fish. This assumes that the first sample of the day had the greatest diversity of fish species found for that day. This, theoretically, should be true because we expect that consecutive catches will be less than the first pass (fish have been sampled in the first-pass without replacement). Hence, the final data set contains records only for the same location and date if the organisations taking the sample were different.

The final data set contains a unique identifier (known as 'card' in the NZFFD), spatial information ('nzsegment', latitude and longitude, catchment name, etc.), date of sample, sampling organisation, longfin and shortfin eel presence/absence (known in the NZFFD as 'angdie' for the longfin eel and 'angaus' for the shortfin eel), and the 87 covariates used in the Crow et al. (2014) study (see Table A.1).

There are a wide variety of organisations who have contributed data towards the NZFFD. Additionally, organisations such as DOC and Fish and Game New Zealand have a number of departments around New Zealand that have contributed. Departments within an organisation appear as different levels in the organisation variable of the NZFFD. In order to ensure

that the model is not over saturated by organisation dummy variables, departments within an organisation are collapsed into one level. Furthermore, organisations that contributed less than 100 data points are excluded from the data set. Contributors to the NZFFD recorded as 'unknown' or as 'private individuals' are removed because catch rates are likely to be inconsistent and we cannot verify whether or not they would comply with sampling protocols such as those of Joy et al. (2013).

These changes reduced the number of organisation levels from 109 to 14. As a result, all departments within the same organisation and individual organisations as a whole are assumed to sample with equal catch rates. These are strong but necessary assumptions because we do not have information on how electric fishing catch rates differ by organisation.

A full list of all the organisations that were used and how many presences and absences of longfin and shortfin eels they found is shown in Tables 2.1 and 2.2. Longfin eels have 45% absences and 55% presences, whereas shortfin eels have 78% absences and 22% presences. This pattern is shown in Figures 2.1 and 2.2. The variable for organisation was used in the VAST model as a 'gear' effect otherwise known as a catchability covariate. Note that collapsing the organisation departments into single organisations is done before removing time 'replicates'. This ensures the final data set has no replicates in time/space by the same organisation. Regardless, replicates are unlikely because departments tend to be separated in space.

The VAST modelling software is able to account for the area swept. In marine fisheries research, the area swept is the area over which a sample is taken. Equivalently, in freshwater electric fishing, the area swept is the area over which electric fishing occurred. However, this is often termed as a measure of effort because when a sampler fishes in a larger area, the sampler is more likely to encounter a fish. Additionally, when more effort is put into sampling we are likely to observe a greater catch. The NZFFD has an effort variable but this variable contains extensive missing values.

Organisation	Absent	Present	Total	% Present
Auckland Regional Council	51	83	134	61.9%
Fish and Game	687	1141	1828	62.4%
Bioresearchers	455	273	728	37.5%
Carter Holt Harvey Forests	96	182	278	65.5%
Cawthron Institute	50	66	116	55.9%
The Department of Conservation	1746	1062	2808	37.8%
Marlborough District Council	38	54	92	58.7%
NIWA	1001	2101	3102	67.7%
Taranaki Regional Council	54	174	228	76.3%
University of Canterbury	71	87	158	55.1%
Massey University	77	313	390	80.3%
University of Otago	266	82	348	23.6%
Victoria University of Wellington	48	45	93	48.4%
Wellington Regional Council	102	167	269	62.1%
Total	4711	5828	10539	55.3%

Table 2.1: A table of the presence and absence of longfin eels throughout New Zealand by each organisation. The table also gives the percent present by each organisation to 1dp.

Organisation	Absent	Present	Total	% Present
Auckland Regional Council	81	53	134	39.6%
Fish and Game	1610	218	1828	11.9%
Bioresearchers	428	300	728	41.2%
Carter Holt Harvey Forests	268	10	278	3.6%
Cawthron Institute	96	20	116	17.2%
The Department of Conservation	2559	249	2808	8.9%
Marlborough District Council	62	30	92	32.6%
NIWA	1958	1144	3102	36.9%
Taranaki Regional Council	189	39	228	17.1%
University of Canterbury	149	9	158	5.7%
Massey University	328	62	390	15.9%
University of Otago	344	4	348	1.1%
Victoria University of Wellington	28	65	93	69.9%
Wellington Regional Council	139	130	269	48.3%
Total	8207	2332	10539	22.1%

Table 2.2: A table of the presence and absence of shortfin eels throughout New Zealand by each organisation. The table also gives the percent present by each organisation to 1dp.

As a result, the effort variable is not suitable to use for modelling and there are no other suitable variables. Therefore it was necessary to assume that the area swept (i.e. effort put into sampling) was the same for each sample taken.

The data set was modified by Crow et al. (2014) before it was received for this analysis. Crow et al. (2014) removed cards from the data set that poorly represented the natural environmental conditions of the fish species. Cards that represented artificial waterways or canals, or were upstream of a artificial physical barrier were removed by Crow et al. (2014). Additionally, cards that are within a lake, lake outlet, estuary, wetland or pond were removed as these waterways cannot be sampled efficiently through electric fishing.

2.2.1 Covariates

Crow et al. (2014) identified 87 environmental, hydrological and spatial covariates which related New Zealand freshwater fish to being encountered. RRF models are used to determine a compact subset of covariates without impeding predictive performance (Deng & Runger, 2013). The covariates selected under the longfin eel RRF model are used in subsequent longfin eel models. Likewise, the covariates selected under the shortfin eel RRF model are used in subsequent shortfin eel models.

Covariates were measured at each nzsegment and were taken from the REC2 GIS database. The covariates were also used in the freshwater fish modelling studies by Leathwick et al. (2008b) and Crow et al. (2014).

Table A.2 indicates which covariates have been selected by the RRF models and Table A.1 describes each of the covariates considered in the RRF models. 69 covariates were selected for the longfin eel models and 55 covariates were selected for the shortfin eel models. Correlation plots of the selected covariates (Table A.2) are given in Figures 2.5 and 2.6. Figure 2.5 shows the correlation amongst the longfin eel covariates and Figure 2.6

shows the correlation amongst the shortfin eel covariates. Stronger correlation is indicated by a darker and larger red square (negative correlation) or blue square (positive correlation). Both plots indicate large positive or negative correlation amongst many of the covariates.

Figure 2.5 shows very strong negative correlation between hydrological variables. For example, the 'duration between flow' variables (e.g. FRE1.MaxDurBetween) and the 'pulse length' variable (MeanPulseLength-High) show strong positive correlations. Additionally, some environmental variables and hydrological variables show strong correlations between one another. For example, the 'rainfall' and 'runoff' variables have strong positive correlations against one another and the 'FRE5.Count', 'l1', 'l2', 'nPulsesHigh' and 'nPulsesLow' variables. The spatial variables selected for the longfin eel models ('y.1', 'xy', 'xy2' and 'yx2') have strong positive positive correlations against one another, and 'seg_tmin', 'us_tmin', 'seg_june' and 'us_june'. See Figure 2.5 for all the longfin eel covariate correlations.

Figure 2.6 shows that there are less strong positive and negative correlations for the shortfin eel covariates compared to the longfin eel covariates. The only spatial covariates selected for the shortfin eel models was 'x3', hence we do not see strong correlations amongst spatial covariates and against other covariates. However, the 'rainfall' variable 'seg_rain' and the 'runoff' variable 'seg_ro_mm' show strong positive correlation against one another and against the 'FRE1.Count', 'FRE10.Count', 'FRE5.Count', 'l2', and 'nPulsesLow' variables. Additionally, 'seg_rain' and 'seg_ro_mm' show strong negative correlation against 'FRE1.MaxDurBetween', 'MeanPulseLengthHigh', 'MeanPulseLengthLow' and 'Predictability'. See Figure 2.6 for all the shortfin eel covariate correlations.

The correlation plots of Figures 2.5 and 2.6 show that multicollinearity may be a problem. One of the main issues with multicollinearity is that we can expect to see increases in the variability of estimated model parameters (O'brien, 2007). As a result, any changes to the data set may

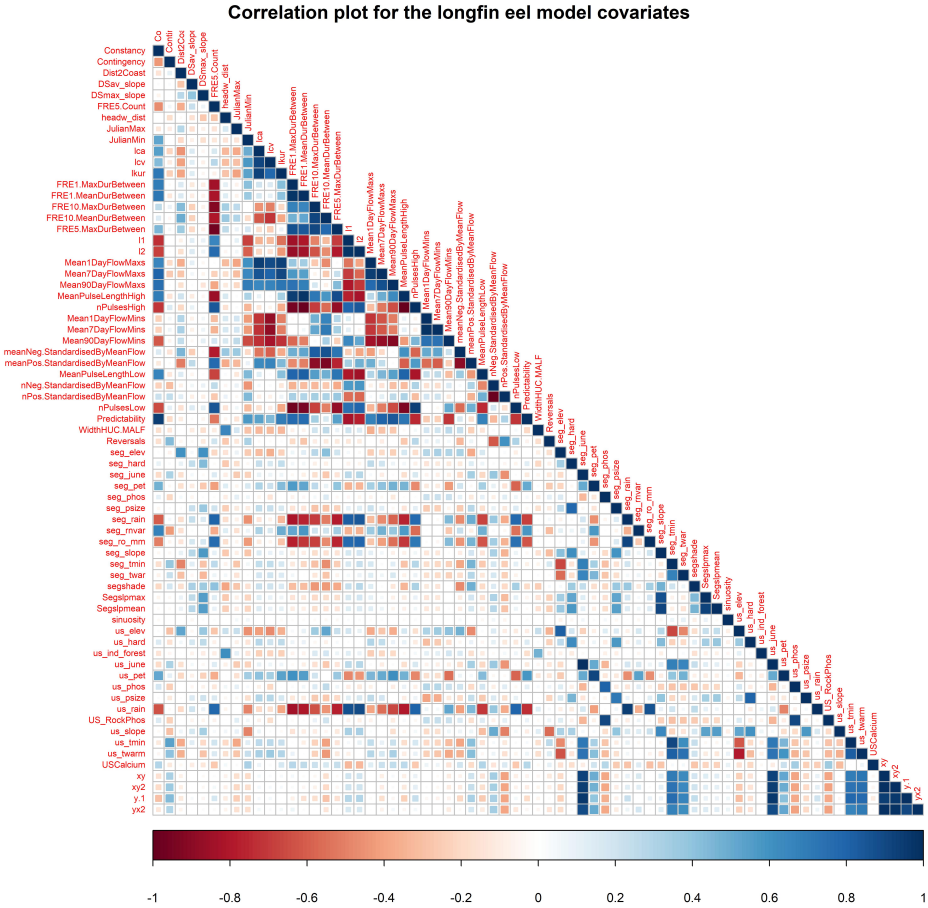


Figure 2.5: Correlation plot of the covariates used in the longfin eel models.

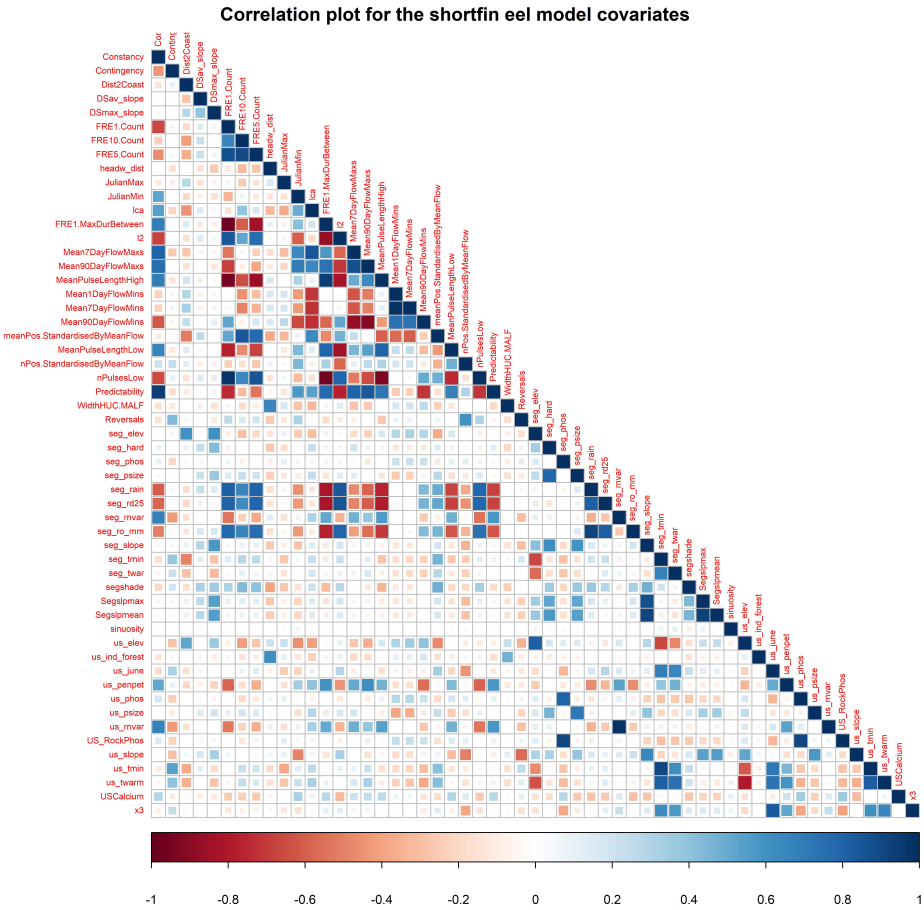


Figure 2.6: Correlation plot of the covariates used in the shortfin eel models.

result in much different parameter estimates and may even result in model non-convergence (O'brien, 2007). Variance inflation factors (VIFs) can be used as a measure of multicollinearity in a given covariate. The variance inflation factor for a covariate g is given by:

$$VIF = \frac{1}{1 - R_g^2}, \tag{2.1}$$

where there are $g = 1, ..., n_g$ covariates and R_g^2 is the R^2 value for covariate g which is found regressing the covariate g against the remaining $n_g - 1$ covariates. If the variability of covariate g cannot be explained well by the remaining covariates (i.e. small R^2 and little multicollinearity) then this will result in a small VIF score. However, if the variability of covariate g can be explained well by the remaining covariate (i.e. large R^2 and large multicollinearity) then this will result in a large VIF score. A VIF score of 10 or greater is considered to show large multicollinearity.

Given the large correlations between the selected covariates, variance inflation factors (VIF) were examined for the covariates selected by the RRF for longfin eels and for the covariates selected by the RRF for shortfin eels. These are shown in Tables A.3 and A.4 of the appendix. The VIF scores are ordered from smallest to largest in each of the tables. 23 covariates have VIF scores less than 10 for the longfin eel covariates and 25 covariates have VIF scores less than 10 for the shortfin eel covariates.

VIF scores were used to assess the selected covariates but no action was taken based on these scores. These scores were included to inform the reader on the shortcomings of the covariates being used and to highlight covariate selection as an area for possible future research.

The following subsections discuss the environmental, hydrological and spatial covariates.

Environmental covariates

The environmental covariates were determined by Leathwick et al. (2008b) based on how relevant they were to freshwater organisms. Fish access related covariates (e.g. distance to coast) were identified as important for fish species with highly mobile behaviour (e.g. diadromous fish such as the longfin and shortfin eels) (Leathwick et al., 2008a). Many of the environmental covariates were measured at multiple scales (upstream scale and catchment scale) in an attempt to account for the hierarchical structure that exists in the distribution of freshwater fish species and their environment (Elith & Leathwick, 2009).

From the Leathwick et al. (2008b) study, Crow et al. (2014) selected covariates to use in the RRF study. The Crow et al. (2014) RRF model selected covariates at the upstream scale and catchment scale as well as all the non-hydrological covariates (Crow et al., 2014). Various segment scale covariates and a number of other miscellaneous covariates were also selected. Inappropriate covariates such as fishing method and presence of dams were excluded because the analysis was only concerned with electric fishing methods and NZFFD cards above dams were excluded from the data (Crow et al., 2014). The environmental covariates make up a total of 44 covariates out of 87 and these covariates at least reflect the same information as the Leathwick et al. (2008b) non-hydrological covariates (Crow et al., 2014).

Hydrological covariates

Crow et al. (2014) selected hydrological covariates which reflect the hydrology of the nzsegments. They predicted these variables using the hydrology modelling methods of Booker & Woods (2014) for the REC2 database. Crow et al. (2013) has shown that the selected hydrological covariates explain unique and significant amounts of variability in the NZFFD. Hence, it was important that Crow et al. (2014) included these variables in the

RRF study. Multicollinearity would be a major issue if all the covariates of Crow et al. (2013) were used. So Crow et al. (2014) only used a subset of the covariates identified in Crow et al. (2013). This is broken down into a further subset through the RRF algorithm in the modelling stage. See Table A.2 for details on what hydrological covariates were used for this research.

Spatial covariates

The spatial covariates were constructed from New Zealand Transverse Mercator (NZTM) coordinates of the downstream end of each segment of the REC2 (Leathwick et al., 2008b). The coordinates (easting and northing) of the downstream end of each nzsegment were determined through the REC2 by Crow et al. (2014). The cubic trend surface regression formula (proposed by Legendre (1990)) was then used to determine x (easting) and y (northing) geographical coordinates of a two dimensional matrix (Crow et al., 2014). This matrix and subsequent x and y coordinates allow us to account for the complex geographical patterns of freshwater fish (Crow et al., 2014). These complex patterns include patches or gaps in the spatial distribution of freshwater fish (Crow et al., 2014).

Chapter 3

Methodology

This chapter outlines the methodology used to make estimates of the probability of capture for longfin and shortfin eels. RRF, VAST and GRaF were used to model the NZFFD longfin and shortfin eel presence/absence data. The models were then used to make probability of capture estimates. The theoretical construct of the regularized random forest (RRF), vector-autoregressive spatio-temporal (VAST) and Gaussian random field (GRaF) models are described.

3.1 The RRF model

Firstly, boosted regression tree (BRT) models are described. These models are then linked to RRF models. Hastie et al. (2009) describes BRT models as follows. Regression trees partition the predictor space into a set of $j = 1, ..., J$ rectangles R_j which represent tree nodes (Hastie et al., 2009). A constant Γ_j is fit to the rectangle and the predictive rule for each rectangle is described as:

$$x \in R_j \implies f(x) = \Gamma_j,$$

where x is a predictor variable. Therefore, we can express a single regression tree as:

$$T(x; \boldsymbol{\Theta}) = \sum_{j=1}^{J} \Gamma_j \mathbb{1}_{x \in R_j}, \tag{3.1}$$

where $\mathbb{1}_{x \in R_j}$ is an indicator variable (1 when $x \in R_j$ and 0 otherwise) and $\boldsymbol{\Theta} = \{(R_1, \Gamma_1), (R_2, \Gamma_2)..., (R_J, \Gamma_J)\}$. The parameters of $\boldsymbol{\Theta}$ are found by empirical risk minimisation:

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \sum_{j=1}^{J} \sum_{x_i \in R_j} \tilde{L}(y_i, \Gamma_j), \tag{3.2}$$

where $\tilde{L}(y_i, \Gamma_j)$ is the loss function of y_i (the response variable) and Γ_j. Additionally, i indexes the data set, where $i = 1, ..., n$. The BRT takes the form of:

$$f_B(x) = \sum_{d=1}^{B} T(x; \boldsymbol{\Theta}_d), \tag{3.3}$$

for $d = 1, ..., B$ tree models. The BRT takes on a stagewise procedure which must solve $\boldsymbol{\Theta}$ at each stage, based on the previous model i.e.

$$\hat{\boldsymbol{\Theta}}_d = \arg\min_{\boldsymbol{\Theta_d}} \sum_{i=1}^{n} \tilde{L}(y_i, f_{d-1}(x_i) + T(x_i; \boldsymbol{\Theta}_d)), \tag{3.4}$$

where $f_{d-1}(x_i)$ is the previous model (Hastie et al., 2009). See Hastie et al. (2009) for further details on parameter estimation.

A RRF model is a machine learning model which applies a regularization framework to a random forest model (Deng & Runger, 2012). See Section 1.2.1 for details on how RRF models work in general. A RRF model was used by Crow et al. (2014) to estimate the probability of capture for New Zealand freshwater fish. The NZFFD longfin and shortfin eel data was modelled using a RRF to compare it directly against the VAST and GRaF methods. The theoretical construct of the longfin and shortfin eel RRF models are discussed below. The model was implemented in R using the 'RRF' package (Deng, 2013; Deng & Runger, 2013, 2012).

3.1.1 The RRF model structure

A total of $N = 87$ features were considered for feature selection in the RRF model. These features are denoted by $x_1, ..., x_{87}$ and are given in Table A.1 of the appendix. Table A.2 indicates which variables were selected by the longfin eel RRF model and by the shortfin eel RRF model. The target variable contained two classes: "True" or "False", where "True" indicates that an eel was observed and "False" indicates that an eel wasn't observed. Each data point came from a certain spatial location (i.e. a nzsegment) which was sampled at a certain time.

Unlike the Crow et al. (2014) RRF models, each nzsegment was not equally weighted in the model. An equal weighting would mean replicating nzsegments that were rarely sampled and removing data from nzsegments that were frequently sampled. This may cause a bias in the results. As the desired outcome of this research is to compare modelling approaches, weighting was not examined in detail. Each model did not incorporate weights in order to remain consistent with one another.

The model takes the form of:

$$f_{RRF}(x) = \frac{1}{B}\sum_{d=1}^{B} T(x; \mathbf{\Theta}_d), \tag{3.5}$$

where there are $d = 1, ..., B$ trees and B is set to 1000. Additionally, the number of trees considered at each node is set to $\phi = \sqrt{87} = 9$ (1 sf) and the minimum node size is set as 1. These settings reflect the findings of Díaz-Uriarte & De Andres (2006) who showed that the performance of a random forest does not significantly improve when the number of trees is between 1000 and 40000. Díaz-Uriarte & De Andres (2006) also showed that $\sqrt{N}$ is a good measure of ϕ.

The individual tree $T(x; \mathbf{\Theta}_d)$ is given by Equation 3.1 and the parameter $\mathbf{\Theta}$ is given by empirical risk minimisation of Equation 3.2.

The features $\mathbf{x}$ in the dth random forest model are determined by measuring the information gain. A feature is selected for the random forest

model if its information gain ($gain(x_\varepsilon)$) is substantially greater than the maximum information gain of the features contained in the model. A penalty of $\Lambda = 0.8$ is applied. Hence, for a feature to be selected for the model the information gain of $0.8 \times gain(x_\varepsilon)$ would need to be larger than $max_m gain(x_m)$.

Each particular tree consists of features $1, ..., m$. The features of all the models are stored in a set known as F. Hence, the overall RRF model is said to contain the features F.

3.1.2 Feature importance

A RRF model is able to evaluate how important a feature is to predicting the outcome variable (Deng & Runger, 2013). This is achieved through an importance score. Deng & Runger (2013) describe the process as follows. A Gini index score is calculated at each node ν,

$$Gini(\nu) = \sum_{I=1}^{2} \hat{\varrho}_I^\nu (1 - \hat{\varrho}_I^\nu), \tag{3.6}$$

where $\hat{\varrho}_I^\nu$ is the proportion of observations in class I at node ν. There are $I = \{1, 2\}$ classes (either presence or absence) in the NZFFD longfin and shortfin eel data set. The information gain made by a particular feature x_m is then measured by

$$Gain(x_m, \nu) = Gini(x_m, \nu) - (\alpha^{(1)} Gini(x_m, \nu^{(1)}) + \alpha^{(2)} Gini(x_m, \nu^{(2)})). \tag{3.7}$$

Weights, $\alpha^{(1)}$ and $\alpha^{(2)}$, are applied to the proposed left and right child nodes $\nu^{(1)}$ and $\nu^{(2)}$, respectively. The difference between the Gini index score of the parent node and the sum of the weighted Gini index score of the child nodes gives the Gini information gain. At each node, 9 (ϕ) features are randomly selected to be evaluated. A 'seed' is set before running each of the RRF models so that differences between the models are not due to differences in the ϕ randomly selected features at each node. Additionally, this ensures replicability of the models. The feature which increases

gain with a penalty (Λ) applied is selected for splitting node ν. That is,

$$Gain_{node}(x_m, \nu) = \begin{cases} 0.8 \times Gain(x_m, \nu) & m \notin F \\ Gain(x_m, \nu) & m \in F \end{cases} \tag{3.8}$$

hence, $Gain_{node}(x_m, \nu)$ is measured at each node. The importance score for a particular variable can be found by,

$$Imp_m = \frac{1}{1000} \sum_{\nu \in V_{x_m}} Gain_{node}(x_m, \nu) \tag{3.9}$$

where there are 1000 regression trees and V_{x_m} are all the nodes (across all trees) which have split by x_m.

Additionally, each of the covariates selected by the longfin eel and the shortfin eel RRF models were assessed using variance inflation scores. The scores for each covariate are given in Tables A.3 and A.4 of the appendix. However, no action was taken based on these scores. These scores are included to highlight the shortcomings on the features selected by the RRF models.

3.2 The VAST model

The VAST model was proposed by Thorson & Barnett (2017) as a way of modelling fisheries data using a spatial-temporal approach. Fisheries data embodies data types such as presence/absence data, count data (i.e. abundance data), and biomass data. The tool can be used for single species or for multi-species modelling, where a multi-species approach can account for correlations among species (Thorson & Barnett, 2017). Models can be built in R statistical software (R Core Team, 2017) using the VAST R modelling package (Thorson & Barnett, 2017; Thorson, 2019).

VAST incorporates a delta model in order to separately model encounter probability and positive catch rates. This research is only interested in the encounter probability component. The VAST modelling methodology

makes use of Gaussian random fields and SPDE approximations to account for spatial and spatio-temporal effects. Additionally, VAST uses a maximum likelihood estimation approach where maximum likelihood estimates are made through the Laplace approximation.

Three VAST models were built: a longfin eel single species model, a shortfin eel single species model, and a longfin and shortfin eel multi-species model.

3.2.1 The VAST model structure

As described in Section 1.2.5, the VAST model consists of the linear predictors $\eta_1(i)$ and $\eta_2(i)$, where $\eta_1(i)$ is associated with the encounter probability and $\eta_2(i)$ is associated with the positive catch rate within a delta model structure (Thorson, 2018). VAST probability of capture models were constructed using the NZFFD presence/absence data for longfin and shortfin eels. In this case we are only interested in the encounter probability (otherwise known as probability of capture) component of the delta model. Therefore, we 'switch off' the components of the positive catch rate in the delta model. See Appendix D for details on how this was done in R.

Estimates for the probability of capture incorporated spatial effects, spatio-temporal effects, density covariates and catchability covariates. Therefore, $\eta_1(i)$ is given by:

$$\begin{aligned}\eta_1(i) =& \beta_1(c_i, t_i) + \sum_{f=1}^{n_{\Omega_1}} L_{\Omega_1}(c_i, f)\Omega_1(s_i, f) + \sum_{f=1}^{n_{\epsilon_1}} L_{\epsilon_1}(c_i, f)\epsilon_1(s_i, f, t_i) \\ &+ \sum_{g=1}^{n_g} \gamma_1(c_i, t_i, g)\varpi(x_i, t_i, g) + \sum_{k=1}^{n_k} \lambda_1(k)Q(i, k),\end{aligned} \tag{3.10}$$

where $\eta_1(i)$ is the linear predictor for encounter probability in a delta model for observation i (Thorson, 2018). We define t_i as the year for observation i, s_i as the spatial location for observation i and c_i as the category (i.e species) for observation i. When constructing the single species models, Equation

3.10 ignores the term for category. Model variables have been indexed with 1 to distinguish them from the model variables of $\eta_2(i)$. Although this model purposely excludes $\eta_2(i)$, the index of 1 has been included in Equation 3.10 to remain consistent with Thorson (2018).

The term $\beta_1(c_i, t_i)$ is the intercept for category c_i and year t_i, $\Omega_1(s_i, f)$ is the spatial variation at location s_i for factor f and $\epsilon_1(s_i, f, t_i)$ is the spatio-temporal variation at location s_i and year t_i for factor f (Thorson, 2018). Factors f are built into the model to enable new terms to be constructed for each category. Hence, f is made equal to the number of categories ($n_{\Omega_1} = 2$ and $n_{\epsilon_1} = 2$ for the multi-species VAST model, and $n_{\Omega_1} = 1$ and $n_{\epsilon_1} = 1$ for both the longfin eel VAST model and the shortfin eel VAST model). The terms $L_{\Omega_1}(c_i, f)$ and $L_{\epsilon_1}(c_i, f)$ are loadings matrices that generate spatial and spatial-temporal covariation respectively for each factor f (Thorson, 2018). These matrices collapse to a single scalar value in a VAST single species model. A summation is taken over the factors f.

The summation $\sum_{g=1}^{n_g} \gamma_1(c_i, t_i, g)\varpi(x_i, t_i, g)$ occurs across $g = 1, 2, ...n_g$ covariates, where $\varpi(x_i, t_i, g)$ are the density covariates for covariate value x_i at observation i, year t_i and covariate g. Density covariates are defined as covariates which account for the density of a species. This research uses GIS information (data with a defined spatial location) of New Zealand river segments as density covariates. The term $\gamma_1(c_i, t_i, g)$ is the estimated impact of the density covariates for category c_i, year t_i and covariate g (Thorson, 2018). The term $Q(i, k)$ is an element in the design matrix $\mathbf{Q}$ which is one for catchability covariate k observed on observation i. Catchability covariates are covariates which describe differences in catch rates between sampling occasions. The term $\lambda_1(k)$ is the estimated impact of the kth catchability covariate. The sampling organisation is included as a catchability covariate. The matrix $\mathbf{Q}$ will therefore be an $n \times k - 1$ matrix of dummy variables with:

$$Q(i, k) = \begin{cases} 1 & \text{if sampling organisation } k \text{ sampled observation } i \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

The term $\lambda_1(k)$ is equal to zero for the most common sampling organisation to sample (NIWA). This is set by the user and therefore does not need to be the most common sampling organisation. Lastly, the summation $\sum_{k=1}^{n_k} \lambda_1(k)Q(i,k)$ occurs over $k = 1$ to n_k, where $n_k = 14$ (the number of sampling organisations).

The predicted probability of capture $\psi_1(i)$ is given by:

$$\psi_1(i) = logit^{-1}(\eta_1(i)) = \frac{\exp(\eta_1(i))}{1 + \exp(\eta_1(i))}, \tag{3.12}$$

where the link function $logit^{-1}$ is the logistic transformation function (inverse logit function) applied to $\eta_1(i)$ (Thorson, 2018).

Model fit is assessed using Pearson's residuals. This is given by:

$$r_{Pearson} = \frac{y_i - \hat{\psi}_1(i)}{\sqrt{\hat{\psi}_1(i)(1 - \hat{\psi}_1(i))}}, \tag{3.13}$$

where y_i is a binary variable (1 if the species is captured, 0 otherwise), $y_i - \hat{\psi}_1(i)$ is the raw residual, $\hat{\psi}_1(i)$ is the fitted value for y_i and $\hat{\psi}_1(i)(1 - \hat{\psi}_1(i)) = var(y_i)$. Hence, the Pearson residual accounts for the variance function of the GLM.

3.2.2 Establishing the spatial domain

Section 1.2.6 begins by describing the spatial construct of the VAST models. The section outlines how the user constructs a spatial 'mesh' by specifying the number of knots used. A total of 400 knots were used to construct the mesh for each VAST model. This offered a fine mesh triangulation while simultaneously not burdening computation speed. It was also the finest resolution capable of running with bias corrections (see Thorson & Kristensen (2016)) under a NIWA server with 245GB of RAM. The mesh is constructed as a part of the SPDE approximation. A solution to the SPDE is found through a basis representation which estimates a Gaussian random field.

Gaussian random fields are estimated for the spatial and spatio-temporal components of Equation 3.10.

3.2.3 Model parameters

The spatial term $\Omega_1(s_i, f)$ and spatio-temporal term $\epsilon_1(s_i, f, t_i)$ of Equation 3.10 are specified in VAST as random effects (Thorson, 2018). These random effects are defined through Gaussian random fields (Thorson & Barnett, 2017). Hence,

$$\Omega_1(\mathbf{s}, f) \sim MVN(\mathbf{0}, \sigma^2_{\Omega_1} \mathbf{\Psi}_1(s, s+h')),$$
$$\epsilon_1(\mathbf{s}, f, t) \sim MVN(\mathbf{0}, \sigma^2_{\epsilon_1} \mathbf{\Psi}_1(s, s+h')),$$

where the spatial term $\Omega_1(\mathbf{s}, f)$ is a vector indexed by s for a given factor f and has a Multivariate Normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2_{\Omega_1} \times \mathbf{\Psi}_1(s, s+h')$. This spatial autocorrelation ($\sigma^2_{\Omega_1} \times \mathbf{\Psi}_1(s, s+h')$) is constant over time. The term $\sigma^2_{\Omega_1}$ is the variance of $\Omega_1(s, f)$ and $\mathbf{\Psi}_1(s, s+h')$ is the correlation matrix between locations s and $s+h'$. The spatio-temporal term $\epsilon_1(\mathbf{s}, f, t)$ is a vector indexed by s for a given t and f. It has a Multivariate Normal distribution with mean zero and covariance matrix $\sigma^2_{\epsilon_1} \times \mathbf{\Psi}_1(s, s+h')$. This means that the spatial-temporal effect is independent from year-to-year. The term $\sigma^2_{\epsilon_1}$ is the variance of $\epsilon_1(s, f, t)$ and $\mathbf{\Psi}_1(s, s+h')$ is defined the same as before. The variance terms $\sigma^2_{\Omega_1}$ and $\sigma^2_{\epsilon_1}$ are both set to one by default.

The correlation matrix $\mathbf{\Psi}_1(s, s+h')$ is defined as following a Matérn function:

$$\mathbf{\Psi}_1(s, s+h') \;=\; \frac{1}{2^{\varphi-1}\Gamma(\varphi)} \times (\kappa_1|h'\mathbf{H}|)^{\varphi} \times K_{\varphi}(\kappa_1|h'\mathbf{H}|), \tag{3.14}$$

where $\mathbf{H}$ is a 2×2 matrix for geometric anisotropy and is defined as:

$$\mathbf{H} = \begin{bmatrix} \exp(h_1) & h_2 \\ h_2 & \exp(-h_1)(1+h_2^2) \end{bmatrix}, \tag{3.15}$$

where h_1 is Northing anisotropy and h_2 is anisotropic correlation in $\mathbf{H}$ (Thorson et al., 2015).

Additionally, φ is the Matérn smoothness parameter which is fixed at 1, h' is the distance between locations s and $s + h'$, κ_1 governs the distance of decorrelation (i.e. the distance at which two locations are considered to be uncorrelated) and K_φ is the Bessel function (Thorson, 2018; Thorson & Barnett, 2017). The two components of geometric anisotropy $\mathbf{H}$ and κ_1 are estimated as fixed effects. The spatial correlation $\mathbf{\Psi}_1(s, s + h')$ between the locations s and $s + h'$ is expected to decline as $|h'|$ increases (Thorson & Barnett, 2017), as defined by Tobler's first law of geography (Tobler, 1970).

The model components $\beta_1(t+1)$, $\gamma_1(t, g)$ and $\lambda_1(k)$ are specified as random effects. Hence, they are governed by a distribution. These are:

$$\begin{aligned}
\beta_1(t+1) &\sim N(\rho_{\beta_1}\beta_1(t), \sigma^2_{\beta_1}), \\
\gamma_1(t, g) &\sim unif(-20, 20), \\
\lambda_1(k) &\sim unif(-20, 20),
\end{aligned}$$

where $\beta_1(t+1)$ is the intercept term for year $t+1$ and is defined as a random effect. The term takes on a Normal distribution with mean $\rho_{\beta_1} \times \beta_1(t)$ and variance $\sigma^2_{\beta_1}$. This approach allows dependence between years to be incorporated in the model structure. The term ρ_{β_1} represents weight of the previous year and $\beta_1(t)$ is the value of the intercept for year t. The term ρ_{β_1} is set equal to one. Hence, $\beta_1(t+1)$ is specified as a random walk. The variance term $\sigma^2_{\beta_1}$ is a fixed effect and is therefore estimated. The initial intercept term $\beta_1(1)$ is defined as $\beta_1(1) \sim N(0, \sigma^2_{\beta_1})$.

The density covariate effects $\gamma_1(t, g)$ follow a uniform distribution for covariate g and for all years t. The bounds are set between -20 and 20. The catchability covariate effects $\lambda_1(k)$ follow a uniform distribution for each catchability term k. Likewise, the bounds of its distribution are set between -20 and 20.

3.2.4 Parameter estimation

The VAST model has the form of Equation 3.10, and incorporates a number of fixed effects and random effects. VAST takes a maximum likelihood estimation approach to parameter estimation. The goal is to find a solution of the joint likelihood $L(\boldsymbol{\theta})$ with respect to the random effects $\boldsymbol{\epsilon}$:

$$L(\boldsymbol{\theta}) = \int_{\epsilon} P(\mathbf{D}|\boldsymbol{\theta}, \boldsymbol{\epsilon}) P(\boldsymbol{\epsilon}|\boldsymbol{\tau}) d\boldsymbol{\epsilon} = \int_{\boldsymbol{\epsilon}} \exp(\ell(\boldsymbol{\epsilon}, \boldsymbol{\theta})) d\boldsymbol{\epsilon}. \tag{3.16}$$

The terms $\boldsymbol{\theta}$, $\mathbf{D}$ and $\boldsymbol{\tau}$ give the fixed effects, data and parameters governing the distribution of the random effects respectively. Additionally, $\ell(\boldsymbol{\epsilon}, \boldsymbol{\theta})$ is the log-likelihood of the mixed effects model i.e.

$$\ell(\boldsymbol{\epsilon}, \boldsymbol{\theta}) = \log(P(\mathbf{D}|\boldsymbol{\theta}, \boldsymbol{\epsilon}) P(\boldsymbol{\epsilon}|\boldsymbol{\tau})).$$

To denote the maximiser of $\ell(\boldsymbol{\epsilon}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\epsilon}$ we use $\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})$. Hence,

$$\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\epsilon}} \ell(\boldsymbol{\epsilon}, \boldsymbol{\theta}).$$

The Hessian of $\ell(\boldsymbol{\epsilon}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\epsilon}$ is denoted by $\boldsymbol{\omega}(\boldsymbol{\theta})$. The Hessian is given by:

$$\boldsymbol{\omega}(\boldsymbol{\theta}) = -\left[\frac{\partial^2}{\partial \boldsymbol{\epsilon}^2} \ell(\boldsymbol{\epsilon}, \boldsymbol{\theta}) \bigg|_{\boldsymbol{\epsilon} = \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})} \right]. \tag{3.17}$$

Therefore,

$$\frac{\partial^2}{\partial \boldsymbol{\epsilon}^2} \ell(\boldsymbol{\epsilon}, \boldsymbol{\theta}) \bigg|_{\boldsymbol{\epsilon} = \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})} = -\boldsymbol{\omega}(\boldsymbol{\theta}).$$

We are interested in finding a solution to the integral $\int_{\boldsymbol{\epsilon}} \exp(\ell(\boldsymbol{\epsilon}, \boldsymbol{\theta})) d\boldsymbol{\epsilon}$. Hence, we apply Laplace's approximation. Firstly, a Taylor series expansion is applied to $\ell(\boldsymbol{\epsilon}, \boldsymbol{\theta})$ evaluated at $\boldsymbol{\epsilon} = \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})$:

$$\begin{aligned} \ell(\boldsymbol{\epsilon}, \boldsymbol{\theta}) \;\approx\; & \ell(\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}), \boldsymbol{\theta}) \;+\; (\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})) \frac{\partial \ell(\boldsymbol{\epsilon}, \boldsymbol{\theta})}{\partial \boldsymbol{\epsilon}} \bigg|_{\boldsymbol{\epsilon} = \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})} \\ & + \frac{1}{2} (\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}))^{\top} \frac{\partial^2 \ell(\boldsymbol{\epsilon}, \boldsymbol{\theta})}{\partial \boldsymbol{\epsilon}^2} \bigg|_{\boldsymbol{\epsilon} = \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})} (\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})), \end{aligned}$$

where we ignore higher order terms and we know that $\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\epsilon}} \ell(\boldsymbol{\epsilon}, \boldsymbol{\theta})$. Therefore,

$$\ell(\boldsymbol{\epsilon}, \boldsymbol{\theta}) \;\approx\; \ell(\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}), \boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\epsilon} \;-\; \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}))^{\top}\mathbf{H}(\boldsymbol{\theta})(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})),$$

where $\boldsymbol{\omega}(\boldsymbol{\theta})$ is given in Equation 3.17. Using this Taylor series expansion, we can apply the Laplace approximation to the integral of interest. Hence,

$$\begin{aligned}\int_{\boldsymbol{\epsilon}} \exp(\ell(\boldsymbol{\epsilon}, \boldsymbol{\theta}))d\boldsymbol{\epsilon} \;&\approx\; \int_{\boldsymbol{\epsilon}} \exp\left(\ell(\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}), \boldsymbol{\theta})) \;-\; \frac{1}{2}(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}))^{\top}\boldsymbol{\omega}(\boldsymbol{\theta})(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}))\right) d\boldsymbol{\epsilon} \\ &=\; \exp(\ell(\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}), \boldsymbol{\theta})) \int_{\boldsymbol{\epsilon}} \exp\left(-\frac{1}{2}(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}))^{\top}\boldsymbol{\omega}(\boldsymbol{\theta})(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}))\right) d\boldsymbol{\epsilon},\end{aligned}$$

where $\exp\left(-\frac{1}{2}(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}))^{\top}\boldsymbol{\omega}(\boldsymbol{\theta})(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}))\right)$ is the kernel of the multivariate Normal distribution so that,

$$\int_{\boldsymbol{\epsilon}} \exp\left(-\frac{1}{2}(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}))^{\top}\boldsymbol{\omega}(\boldsymbol{\theta})(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}))\right) d\boldsymbol{\epsilon} \;=\; 2\pi^{\vartheta/2}\det(\boldsymbol{\omega}(\boldsymbol{\theta}))^{-1/2},$$

where ϑ is the number of dimensions, i.e. the number of random effects. Therefore, the Laplace approximation of the integral of interest is:

$$\begin{aligned}L^{*}(\boldsymbol{\theta}) \;&=\; \int_{\boldsymbol{\epsilon}} \exp(\ell(\boldsymbol{\epsilon}, \boldsymbol{\theta}))d\boldsymbol{\epsilon} \\ &\approx\; \exp(\ell(\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}), \boldsymbol{\theta}))2\pi^{\vartheta/2}\det(\boldsymbol{\omega}(\boldsymbol{\theta}))^{-1/2},\end{aligned}$$

where estimates of $\boldsymbol{\theta}$ minimise the negative log of the Laplace approximation (Kristensen et al., 2015):

$$-\log L^{*}(\boldsymbol{\theta}) \;=-\frac{\vartheta}{2}\log 2\pi \;+\; \frac{1}{2}\log\det(\boldsymbol{\omega}(\boldsymbol{\theta})) \;+\; \ell(\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}), \boldsymbol{\theta}). \tag{3.18}$$

VAST uses Template Model Builder (Kristensen et al., 2015) to implement the Laplace approximation given in Equation 3.18 (Thorson & Barnett, 2017). Maximum likelihood estimates for the fixed effects are found by using a gradient based non-linear minimiser (Thorson & Barnett, 2017). At each iteration $\hat{\epsilon}(\boldsymbol{\theta})$ is re-evaluated through this optimiser (Skaug &

Fournier, 2006). Automatic differentiation is used to evaluate $\boldsymbol{\omega}(\boldsymbol{\theta})$. See Skaug & Fournier (2006) for details on automatic differentiation. Standard errors of the fixed effects are then calculated using a generalisation of the delta method (Kass & Steffey, 1989). Estimates for the probability of capture $\psi_1(i)$ are then derived by:

$$\hat{\psi}_1(i) = logit^{-1}(\hat{\eta}_1(i)). \tag{3.19}$$

Whenever a random effect is transformed then the mean and variance which define the random effect will be transformed (Thorson & Kristensen, 2016). If this transformation is non-linear, such as the logistic transformation function in Equation 3.19, then the estimator will be biased (Thorson & Kristensen, 2016). In order to account for this, a bias correction algorithm can be implemented. VAST is able to implement a bias correction algorithm through Template Model Builder (TMB) (Kristensen et al., 2015) using the epsilon method (Thorson & Kristensen, 2016).

The bias correction algorithm proposed by Thorson & Kristensen (2016) is given as follows. If we have a model containing fixed effects $\boldsymbol{\theta}$, random effects $\boldsymbol{\epsilon}$ and data $\mathbf{D}$ then we seek to derive an unbiased estimate of $\Upsilon = f(\hat{\boldsymbol{\theta}}, \boldsymbol{\epsilon}|\mathbf{D})$. Here $\hat{\boldsymbol{\theta}}$ is the estimated fixed effect and is given by:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left(\log \left(\int \exp(\ell(\boldsymbol{\theta}, \boldsymbol{\epsilon}; \mathbf{D})) d\boldsymbol{\epsilon} \right) \right),$$

where $\ell(\boldsymbol{\theta}, \boldsymbol{\epsilon}; \mathbf{D})$ is the joint log-likelihood of the fixed and random effects. The minimum variance unbiased estimator of Υ is given by:

$$E[\Upsilon|\mathbf{D}] = \frac{\int \exp(\ell(\boldsymbol{\theta}, \boldsymbol{\epsilon}; \mathbf{D})) f(\hat{\boldsymbol{\theta}}, \boldsymbol{\epsilon}) d\boldsymbol{\epsilon}}{\int \exp(\ell(\boldsymbol{\theta}, \boldsymbol{\epsilon}; \mathbf{D})) d\boldsymbol{\epsilon}}. \tag{3.20}$$

A nuisance parameter ς is introduced into the calculation of the expected value of Equation 3.20. Then the gradient of the marginal likelihood is calculated with respect to ς. Firstly, the function g is defined as:

$$g(\boldsymbol{\theta}, \boldsymbol{\epsilon}, \varsigma; \mathbf{D}) = \log \left(\int \exp(\ell(\boldsymbol{\theta}, \boldsymbol{\epsilon}; \mathbf{D}) - \varsigma f(\boldsymbol{\theta}, \boldsymbol{\epsilon})) d\boldsymbol{\epsilon} \right), \tag{3.21}$$

and the first derivative with respect to ϵ of this function is given by:

$$\frac{\partial}{\partial \varsigma}(g(\boldsymbol{\theta}, \boldsymbol{\epsilon}, \varsigma; \mathbf{D})) = \frac{\int \exp(\ell(\boldsymbol{\theta}, \boldsymbol{\epsilon}; \mathbf{D}) - \varsigma f(\boldsymbol{\theta}, \boldsymbol{\epsilon})) f(\boldsymbol{\theta}, \boldsymbol{\epsilon}) d\boldsymbol{\epsilon}}{\int \exp(\ell(\boldsymbol{\theta}, \boldsymbol{\epsilon}; \mathbf{D}) - \varsigma f(\boldsymbol{\theta}, \boldsymbol{\epsilon})) d\boldsymbol{\epsilon}}. \tag{3.22}$$

If we evaluate this derivative at $\varsigma = 0$ given the estimated fixed effects then Equation 3.22 becomes:

$$\frac{\partial}{\partial \varsigma}(g(\hat{\boldsymbol{\theta}}, \boldsymbol{\epsilon}, \varsigma; \mathbf{D})) = \frac{\int \exp(\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\epsilon}; \mathbf{D})) f(\hat{\boldsymbol{\theta}}, \boldsymbol{\epsilon}) d\boldsymbol{\epsilon}}{\int \exp(\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\epsilon}; \mathbf{D})) d\boldsymbol{\epsilon}} = E[\Upsilon | \mathbf{D}]. \tag{3.23}$$

Equation 3.23 is the minimum variance unbiased estimator for $\Upsilon = f(\hat{\boldsymbol{\theta}}, \boldsymbol{\epsilon} | \mathbf{D})$, where the numerator and denominator of Equation 3.23 are estimated by the Laplace approximation (see above). See Thorson & Kristensen (2016) for more details on the bias correction method used.

3.3 The GRaF model

The GRaF model was proposed by Golding & Purse (2016) as a method for modelling species distributions within a relatively fast and flexible framework. The method enables parameter estimation through a Bayesian framework or a maximum likelihood estimation framework. Additionally, the method allows the user to make approximations through either the Laplace approximation or the expectation-propagation algorithm. This research makes use of the model's Bayesian framework and makes relatively fast approximations through the Laplace approximation. Golding & Purse (2016) uses Gaussian random fields (see Section 1.2.3) to model species distributions.

GRaF is used to model the NZFFD longfin and shortfin eel presence/absence data, and to predict the probability of capture for the longfin and shortfin eels. The approach can be implemented in the statistical software R (R Core Team, 2017) using the R package 'GRaF' (Golding, 2017; Golding et al., 2013; Golding & Purse, 2016).

3.3.1 The GRaF model structure

The Bayesian network directed acyclic graph (DAG) for the GRaF model is shown in Figure 3.1. The DAG shows how the different variables of a GRaF model are related to each other. There are two GRaF models: the longfin eel GRaF model and the shortfin eel GRaF model. The results of each are shown in Chapter 4. From Figure 3.1, y_i is the ith NZFFD longfin or shortfin eel presence/absence, where $i = 1, ..., n$. Each y_i follows a Bernoulli distribution with a probability of success of q_i, where q_i is the probability of capture for the ith data point. The probability of capture q_i is given by a probit transformation of the latent variable z_i, where $\mathbf{z}$ is given by:

$$\mathbf{z} \sim MVN(\boldsymbol{\delta}, \boldsymbol{\Sigma}), \tag{3.24}$$

and $\mathbf{z}$ is defined as a Gaussian random field with mean $\boldsymbol{\delta}$ and covariance $\boldsymbol{\Sigma}$. The mean of the Gaussian random field is given by a user defined function

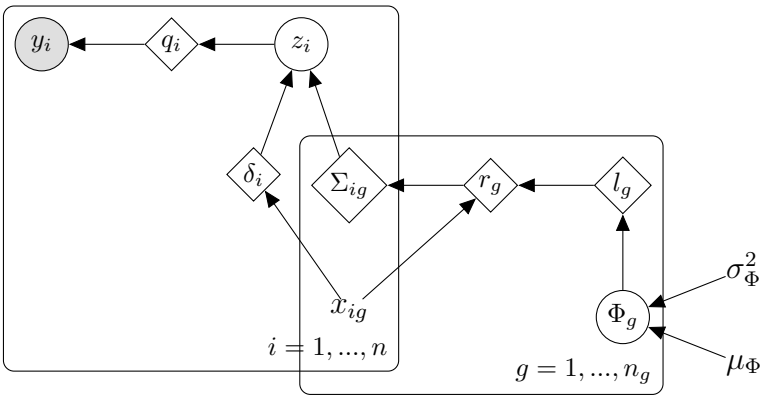


Figure 3.1: A Bayesian network directed acyclic graph of the GRaF model constructed for the NZFFD longfin and shortfin eel presence/absence data. The shaded circle represents observations, circles represent latent variables, diamonds represent deterministic variables, variables by themselves represent constants and the square plates show the construct of the variables within the plates.

of the covariates $\mathbf{x}$. This function describes how the probability of capture changes with each covariate (Golding & Purse, 2016). All covariates are defined in Table A.1 and the covariates that were used for each of the eel species are given in Table A.2. In addition, each of the models used year as a covariate to account for temporal variability in the data.

The function over $\mathbf{x}$ is considered a prior. An uninformative prior was used on $\boldsymbol{\delta}$ for the longfin eel GRaF model and shortfin eel GRaF model. The prior has a flat distribution across each of the covariates in the models. The flat prior gives the probability of capture as the probability of being observed at any given site, p_0. This is defined as:

$$p_0 = \frac{\sum_{i=1}^{n} y_i}{n}.$$

For the longfin eel GRaF model p_0 is 0.55 (2dp) and for the shortfin eel GRaF model p_0 is 0.22 (2dp).

The covariance of the Gaussian random field is given by a squared exponential term:

$$\mathbf{\Sigma} = \exp(\frac{-\mathbf{r}^2}{2}), \tag{3.25}$$

where $\mathbf{r}$ is given by:

$$\mathbf{r} = \sqrt{\sum_{g=1}^{n_g} \left(\frac{\mathbf{x_g} - \mathbf{x_g}'}{l_g^2} \right)^2}. \tag{3.26}$$

The covariance is dependent on each covariate $\mathbf{x_g}$, where $g = 1, ...n_g$, and the hyperparameter l_g. The hyperparameter l_g is known as a lengthscale parameter which defines how rapidly the probability of capture changes with a covariate (Golding & Purse, 2016). A smaller lengthscale indicates that the probability of capture changes much more rapidly with a covariate and a larger lengthscale indicates that the probability of capture changes slower with a covariate. The natural log of the lengthscale $\ln(l_g)$ is given by Φ_g and is defined by:

$$\Phi_g \sim N(\mu_\Phi, \sigma_\Phi^2), \tag{3.27}$$

which is a prior distribution with mean $\mu_\Phi = \log(10)$ and variance $\sigma_\Phi^2 = 1$. This places an informative prior on the lengthscales which indicates that the covariates will have a smooth effect on species' niches (Golding et al., 2013). We allow the models to estimate the lengthscales as oppose to providing each lengthscale. A large lengthscale gives a function of low complexity (flatter distribution) whereas a small lengthscale gives a function of high complexity (Golding & Purse, 2016).

3.3.2 Parameter estimation

There are two steps to inference with the GRaF model. There is inference over the Gaussian random field $\mathbf{z}$ and inference over the hyperparameters defining $\mathbf{z}$ (Golding & Purse, 2016).

The variable of interest is the probability of capture variable q_i which is given by a probit transformation of $\mathbf{z}$. The full posterior distribution for the GRaF model is:

$$\pi(\mathbf{z}, \boldsymbol{\delta}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}|\mathbf{y}) = \frac{f(\mathbf{y}|\mathbf{z})\pi(\mathbf{z}|\boldsymbol{\delta}, \boldsymbol{\Sigma})\pi(\boldsymbol{\delta})\pi(\boldsymbol{\Sigma}|\boldsymbol{\Phi})\pi(\boldsymbol{\Phi})}{f(\mathbf{y})}, \tag{3.28}$$

where

$$f(\mathbf{y}) = \int f(\mathbf{y}|\mathbf{z})\pi(\mathbf{z})d\mathbf{z}. \tag{3.29}$$

The term $\pi(\mathbf{z}, \boldsymbol{\delta}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}|\mathbf{y})$ is the full posterior distribution. Where $f(\mathbf{y}|\mathbf{z})$ is given by a Bernoulli distribution, $\pi(\mathbf{z}|\boldsymbol{\delta}, \boldsymbol{\Sigma})$ is given by a multivariate Normal distribution, $\pi(\boldsymbol{\delta})$ is the prior function across $\boldsymbol{\delta}$, $\pi(\boldsymbol{\Sigma}|\boldsymbol{\Phi})$ is defined by a squared exponential term (given in Equation 3.25), and $\pi(\boldsymbol{\Phi})$ is defined by a Normal distribution. The evidence $f(\mathbf{y})$ is found through a Laplace approximation.

To demonstrate the Laplace approximation we use the derivation which comes from Blangiardo & Cameletti (2015, p. 105). The Laplace approximation can be used to approximate an integral of interest. As an example, we may be interested in approximating the following integral:

$$\int f(y)dy = \int \exp(\log(f(y)))dy, \tag{3.30}$$

where $f(y)$ is a function of the random variable Y. The term $\log(f(y))$ can be expanded through the Taylor series expansion evaluated at $y = y_0$:

$$\log f(y) \approx \log f(y_0) + (y - y_0) \left.\frac{\partial \log f(y)}{\partial y}\right|_{y=y_0} + \frac{(y - y_0)^2}{2} \left.\frac{\partial^2 \log f(y)}{\partial y^2}\right|_{y=y_0}.$$

The mode is given by $y^* = \arg\max_y \log f(y)$. Therefore, if we set $y_0 = y^*$ then $\left.\frac{\partial \log f(y)}{\partial y}\right|_{y=y^*} = 0$. Hence,

$$\log f(y) \approx \log f(y^*) + \frac{(y - y^*)^2}{2} \left.\frac{\partial^2 \log f(y)}{\partial y^2}\right|_{y=y^*}.$$

We can then find the integral of interest (Equation 3.30) by:

$$\int f(y)dy \approx \int \exp\left(\log f(y^*) + \frac{(y - y^*)^2}{2} \left.\frac{\partial^2 \log f(y)}{\partial y^2}\right|_{y=y^*}\right) dy$$

$$= \exp(\log f(y^*)) \int \exp\left(\frac{(y - y^*)^2}{2} \left.\frac{\partial^2 \log f(y)}{\partial y^2}\right|_{y=y^*}\right) dy,$$

where we set $\sigma^2* = -1/\frac{\partial^2 \log f(y)}{\partial y^2} |_{y=y^*}$ so that the integral forms the kernel of the Normal distribution. Hence, the result of the approximated integral is given by:

$$\int f(y)dy \approx f(y^*) \int \exp\left(\frac{-(y - y^*)^2}{2\sigma^{2*}}\right) dy, \tag{3.31}$$

where we obtain the Normal distribution kernel with mean y^* and variance σ^{2*}.

3.3.3 Bayesian inference

Bayesian statistics differs from classical statistics (otherwise known as frequentist statistics) in that it takes the unknown parameter(s) as a random quantity as opposed to a fixed quantity (Carlin & Louis, 2008). A likelihood function $f(\mathbf{y}|\theta)$ is specified for the data $\mathbf{y} = (y_1, ..., y_n)$ given an unknown parameter θ. In Bayesian statistics a prior distribution $\pi(\theta)$ is

specified for the unknown parameter θ. The prior distribution expresses prior knowledge that we may (or may not) have on θ. Hence, the extent to which we have knowledge on θ can be defined through an informative or non-informative prior (or somewhere between). A disadvantage to the Bayesian inference approach is that the prior is often subjective and therefore leaves the model open to criticism by objecting parties.

The target distribution in Bayesian inference is the posterior distribution. The posterior distribution $\pi(\theta|\mathbf{y})$ describes the probability of θ given the observed data $\mathbf{y}$. This is given by Bayes theorem:

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})}, \tag{3.32}$$

where $f(\mathbf{y})$ is the marginal density of $\mathbf{y}$ (otherwise known as the evidence) and is given by:

$$f(\mathbf{y}) = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta. \tag{3.33}$$

The integral of Equation 3.33 is often difficult to evaluate. Markov Chain Monte Carlo (MCMC) have been developed as a simulation method of approximating the integral. However, MCMC can often be computationally expensive (i.e. long computing time and high computing power is needed) (Golding & Purse, 2016). In these cases deterministic methods such as the Laplace approximation are less computationally restrictive and are therefore preferred.

3.4 Model validation

This section describes the model validation methods implemented. See Section 1.2.8 for details on cross validation techniques in general. Two methods of model validation were considered: spatial K-fold cross validation and K-fold cross validation. Spatial K-fold cross validation accounts for spatial correlation in the data. Even though the training sets and test sets are not independent, K-fold cross validation was performed.

The RRF models and VAST models could be assessed using spatial K-fold cross validation and K-fold cross validation. The results under each validation tell us different things about the model. The GRaF models were assessed using K-fold cross validation but with $K = 5$. This was implemented so that validations could run within a practical time. Therefore, 5-fold cross validation was used to assess the GRaF models, and the RRF models and VAST models. This meant that only the RRF models and VAST models could be compared with spatial K-fold cross validation and K-fold cross validation, and the RRF models, VAST models, and GRaF models could be compared with 5-fold cross validation. Spatial 5-fold cross validation was not attempted due to convergence issues with the VAST model.

The RRF models and VAST models were validated using a 50-fold spatial cross validation. A K of 50 was selected because the VAST models struggled to converge when less folds were used. The likely reason for this is that when K was smaller, test sets were larger and the training sets were missing data points at spatial locations which enabled model convergence.

100 'balancing steps' were implemented in the K-means clustering. This means that the K-means clustering was performed 100 times and the clustering which kept each of the 50 folds as even as possible was selected. The spatial clustering was implemented with the 'sperrorest' R package proposed by Brenning (2005). The same 50-fold spatial partitioning was used to validate each model. One main difference between the model cross validations is that RRF models are a feature selection model. Hence, each of the 50 models may have selected a different subset of the variables given in Table A.1 to the variables that were used in the full models.

Receiver operator characteristic (ROC) curves are constructed at each fold of the spatial K-fold cross validation and the K-fold cross validation. The area under the receiver operator characteristic (ROC) curve (AUC) is then estimated and used as a measure of model performance. Estimates of AUC are made using the 'ROCR' R package (Sing et al., 2005).

A wrapper to the 'ROCR' package, known as the 'cvAUC' R package (LeDell et al., 2015), is used to make standard error and 95% confidence intervals of AUC. The variability estimates follow the methodology proposed by LeDell et al. (2015) who used an influence curve based approach to make variance estimates. An influence based approach was demonstrated to be much more computationally efficient than a bootstrap approach which is computationally expensive when using complex models (LeDell et al., 2015). These measures of variability are essential for comparing classifier models (Fawcett, 2006).

Chapter 4

Results

This chapter presents the longfin and shortfin eel probability of capture modelling results. A section is given for each modelling method. Under each section, the model results are given for the longfin eel and shortfin eel separately. Cross validation results are given for each modelling method and a final section is given for the model comparison results.

The observed proportion of longfin eel and shortfin eel capture (Figures 4.1 and 4.2) are repeated here from Chapter 2. This is to aid comparisons between the observed proportion of longfin eel and shortfin eel capture, and the predicted longfin eel and shortfin eel probability of capture.

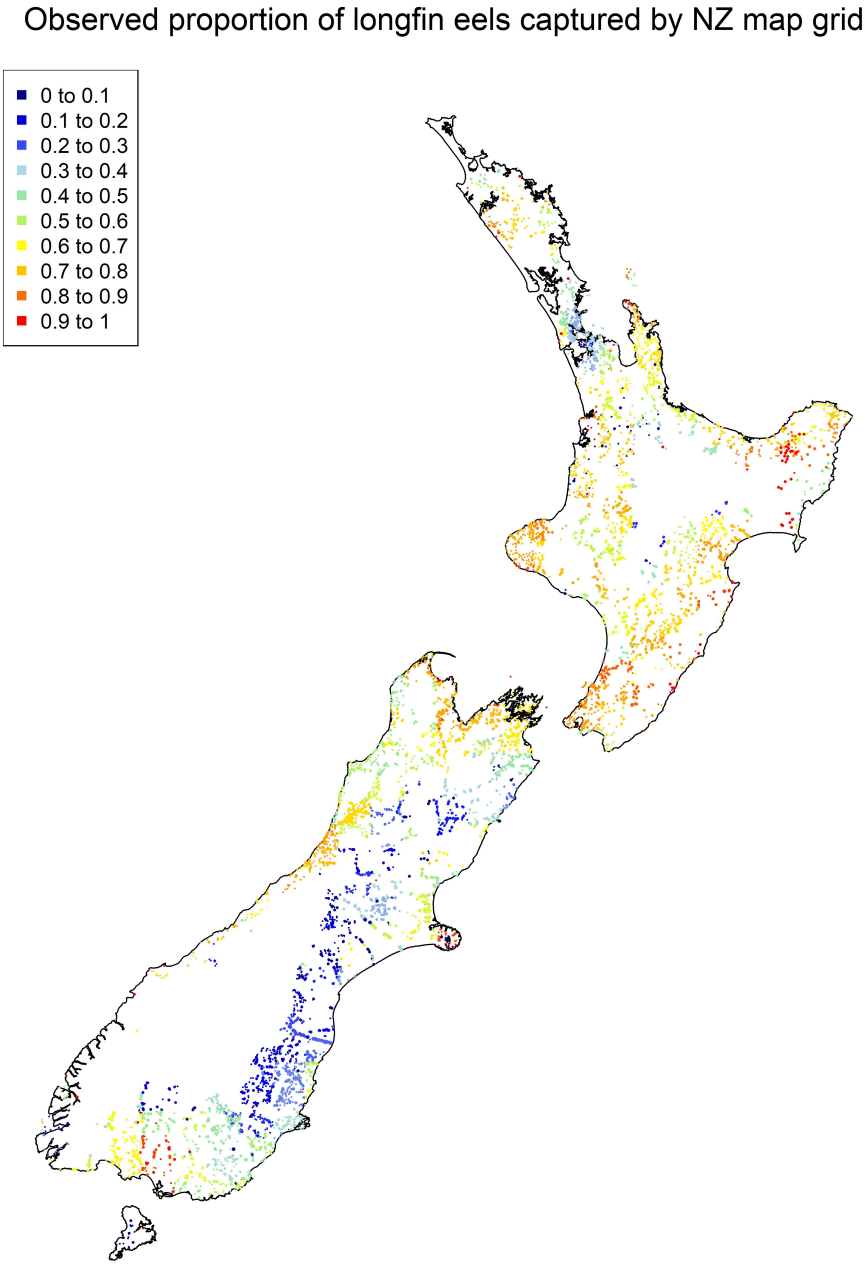


Figure 4.1: Map of the observed proportion of longfin eels captured within each NZMS 260 map series grid square. This map is repeated from Figure 2.3.

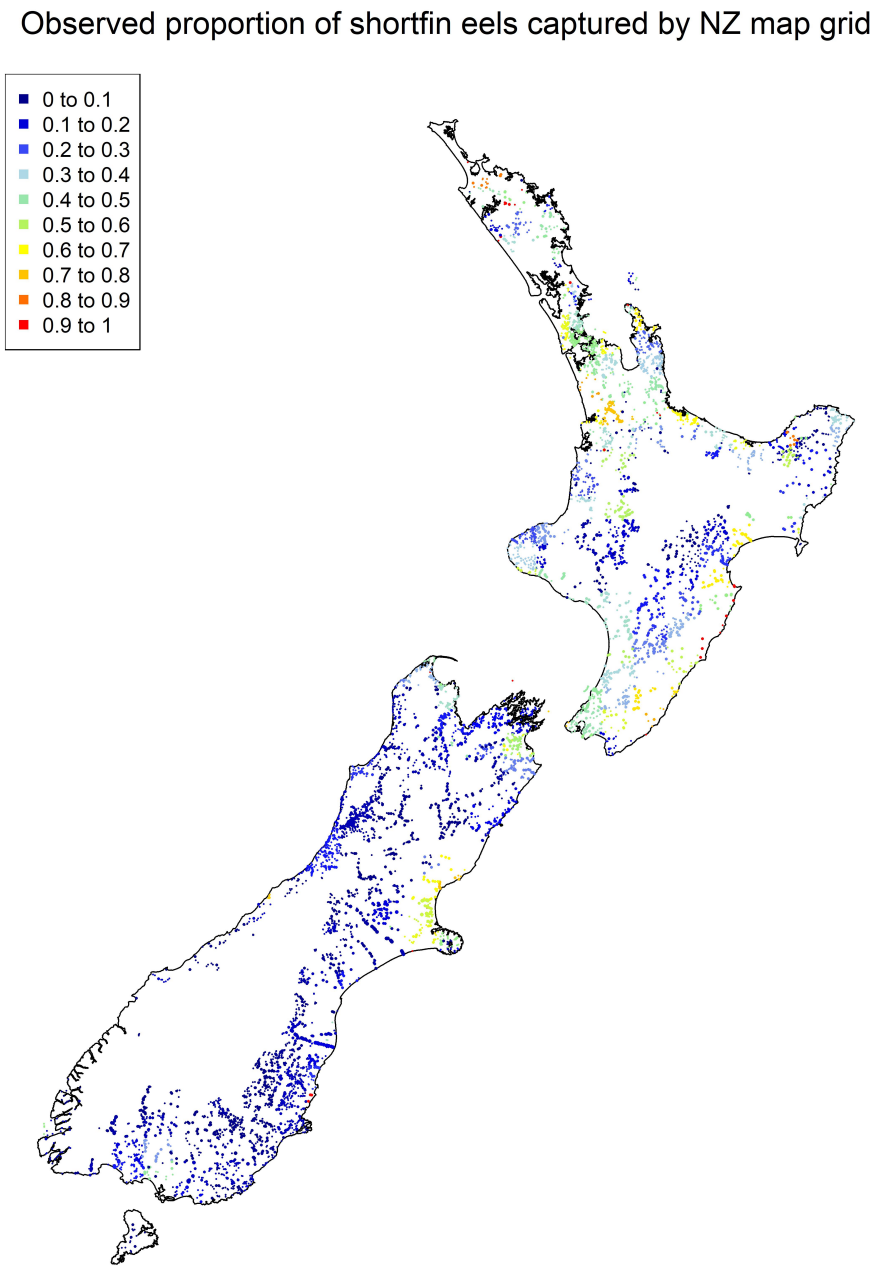


Figure 4.2: Map of the observed proportion of shortfin eels captured within each NZMS 260 map series grid square. This map is repeated from Figure 2.4.

4.1 RRF modelling results

This section presents the results of the longfin eel RRF model and shortfin eel RRF model. The data used for each model is detailed in Chapter 2 and the RRF methodology is detailed in Section 3.1.

The RRF models perform feature selection, where the selected features were used as covariates in subsequent models. The features that were selected by the longfin eel RRF model and shortfin eel RRF model are shown in Table A.2 (determined by obtaining a score greater than 0 from Equation 3.9). The full set of features supplied to the RRF models are described in Table A.1.

Longfin and shortfin eel probability of capture estimates are made under each of their respective models. Estimates are made at particular spatial points of New Zealand. These spatial points come from the centre of each nzsegment (river segment) in the REC2 database. Each nzsegment has a value for each of the model covariates. Hence, predictions are made to each nzsegment of the REC2 database.

50-fold cross validation, spatial 50-fold cross validation and 5-fold cross validation were used to validate the RRF models. Mean AUC estimates and 95% confidence intervals were obtained under each cross validation method.

4.1.1 Longfin eel results

Model results

The longfin eel RRF model selected 69 features in its feature selection process. The importance score for each of the features is shown in Figure 4.3. The features of Figure 4.3 which have an importance score of zero were not selected by the longfin eel RRF model. Any feature which had an importance score greater than zero was selected by the model. Many of the spatial parameters were excluded by the model. The features scor-

ing the highest importance scores tended to be environmental variables. This indicates that environmental variables are important in predicting longfin eel presence or absence. The variable 'seg_tmin' (the mean minimum wintertime air temperature for a segment) achieved an importance score much larger than any others of c.560. This is inconsistent with the findings of Crow et al. (2014) where 'seg_tmin' was not selected by the RRF algorithm.

The probability of capture estimates made across the entire REC2 database using the longfin eel RRF model are shown in Figure 4.4. This can be compared directly to the observed proportions of longfin eel capture as shown in Figure 4.1. The North Island of New Zealand has high probability of capture estimates (≥ 0.6) in the central east and west coast, the Wellington region, and northern areas on the island. This same pattern is seen in Figure 4.1. The central North Island has probability of capture estimates ranging from 0 to 0.5. Very low probabilities of capture are seen in the nzsegments surrounding Mount Ruapehu (located in New Zealand's central North Island). However, Figure 4.1 shows that the observed proportion of longfin eel capture ranged from 0.5 to 0.6 (although there are a few nzsegments with very low probabilities of capture). Various small areas of the central North Island do not have any observed data or were above dams and waterfalls (as shown in Figure 2.1), hence direct comparisons cannot be made.

The South Island of New Zealand mainly consists of very low probability of capture estimates (0 to 0.3) throughout the centre of the island. This same pattern was observed in Figure 4.1. However, some parts of the central South Island do not have any observations (as shown in Figure 2.1). Therefore, it is difficult to assess the probability of capture estimates for this area. Probabilities of capture exceeding 0.6 are seen in the south coast, north coast, central west coast, and in the region surrounding Christchurch. Stewart Island has longfin eel probability of capture estimated around 0 to 0.4. We see this same approximate pattern in Figure

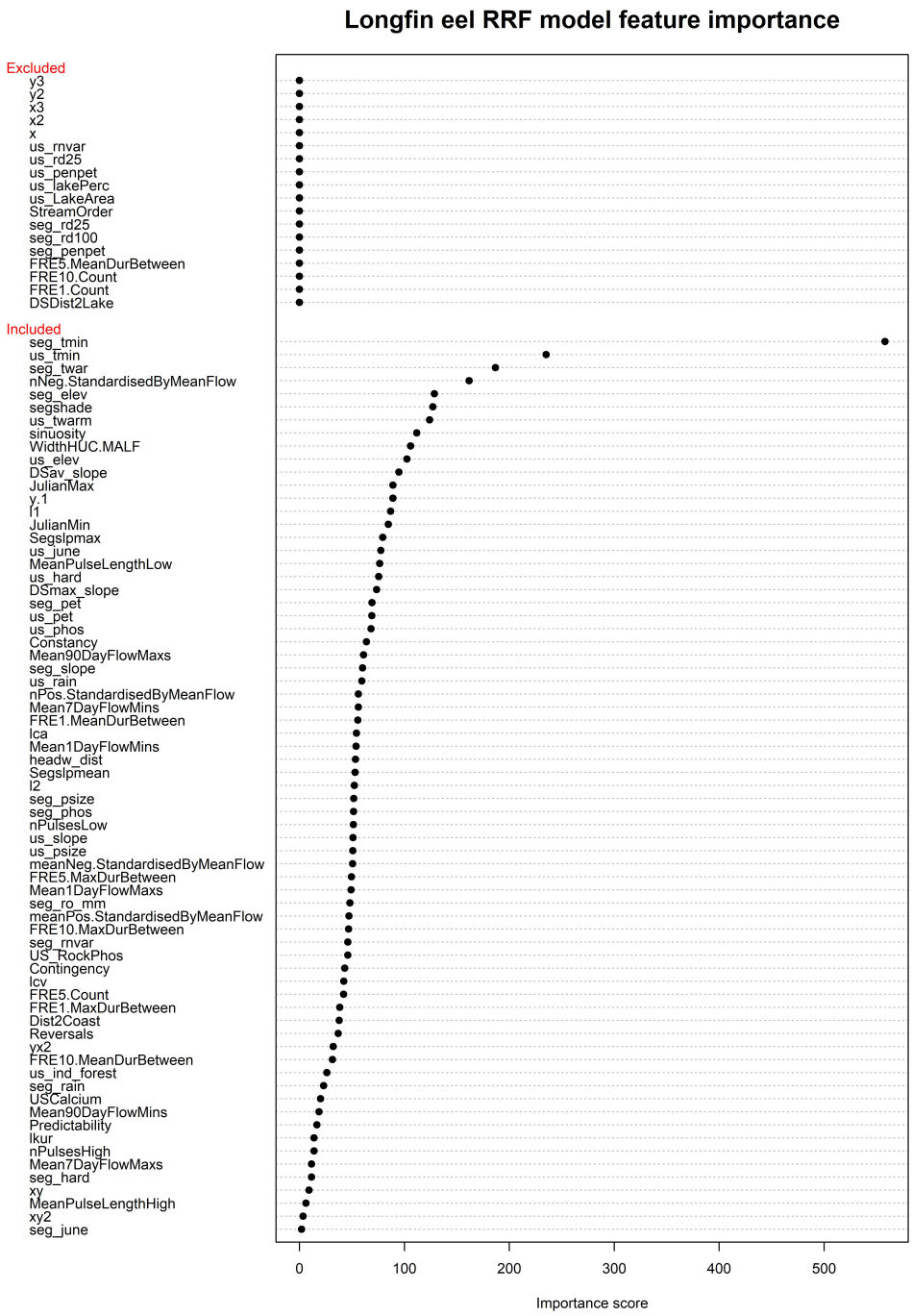


Figure 4.3: Importance scores for the features of the longfin eel RRF model. Features with importance scores greater than 0 were included in the model. For the features included in the model, the features are ordered from largest importance score to smallest importance score.

4.1. However, the north-west corner of the South Island makes low estimates (0 to 0.4) of probability of capture (as shown in Figure 4.4). But the observed proportions of longfin eel capture in this area appear to be 0.5 or greater (as shown in Figure 4.1).

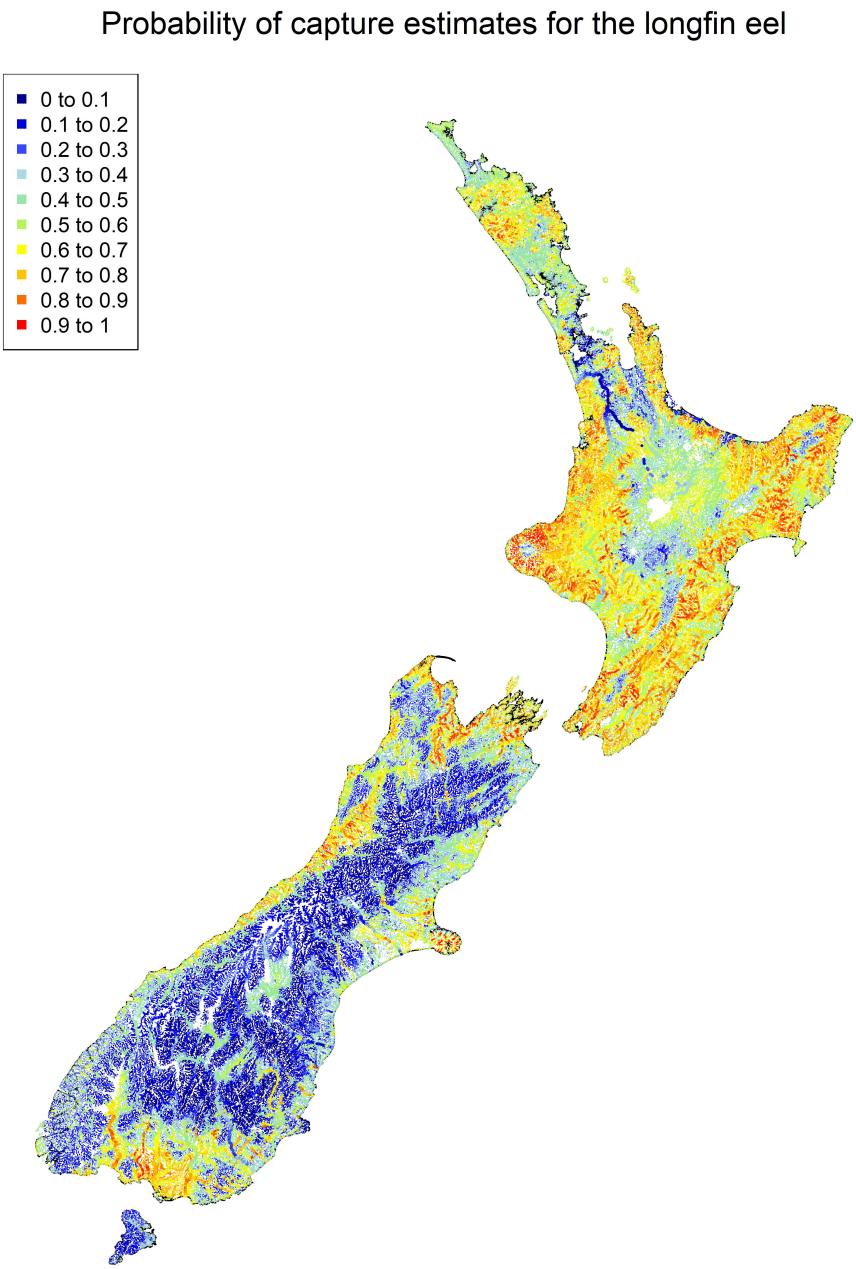


Figure 4.4: Probability of capture estimates for the longfin eel using the REC2 database. These estimates were made using the longfin eel RRF model. Larger points have a larger stream order.

Cross validation results

The spatial 50-fold cross validation for the longfin eel RRF model resulted in a mean AUC of 0.6550 (4dp) with a 95% confidence interval based on an influence curve of 0.6444 and 0.6656 (standard error of 0.0054 (4dp)). The ROC curves for each fold of the 50 spatial folds and the mean ROC curve are shown in Figure 4.5. The boxplots of Figure 4.5 show the spread of the ROC curves.

There appears to be a significant amount of variation in the ROC curves. This indicates that some of the spatial areas are estimated much better than others. A number of curves are producing AUC estimates less than

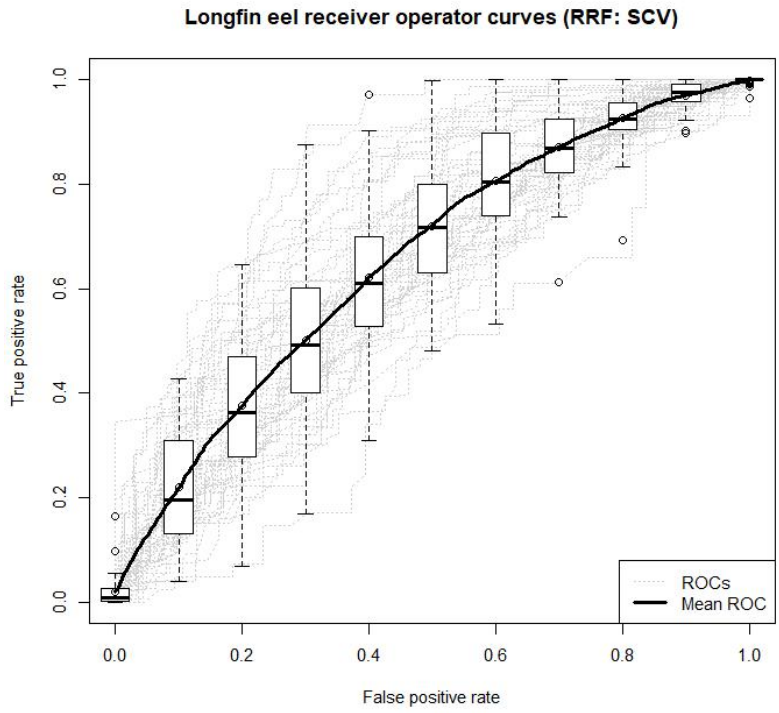


Figure 4.5: ROC curves under each of the 50-folds in the longfin eel RRF model spatial cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

or close to 0.5. This is because predicting longfin eels is highly dependent on their spatial location. Hence, the model is lacking crucial longfin eel data and therefore the model makes poor predictions on these locations. For these spatial folds, estimates for longfin eel presence or absence under the longfin eel RRF model are no better than simply guessing or are estimating effects in the opposite direction as to what's occurring in the area. A spatial fold in the Auckland region results in an AUC of 0.4336. Here, probability of capture is being predicted in the opposite direction to what's occurring.

Overall, when using a RRF for the longfin eel NZFFD data, areas which are spatially distinct to the model data are, on average, estimated poorly. However, there is significant variation, hence, some areas are predicted much better than others.

The 50-fold cross validation for the longfin eel RRF model resulted in a mean AUC of 0.7798 (4dp) with a 95% confidence interval based on an influence curve of 0.7709 and 0.7887 (standard error of 0.0045 (4dp)). The ROC curves for the 50 randomly selected folds and the mean ROC curve are shown in Figure 4.6.

The curves show little variation from one another, especially if compared to the ROC curves of the spatial 50-fold cross validation (Figure 4.5). Each fold results in a AUC greater than 0.5, with the lowest AUC being 0.6964 (4dp) and the highest AUC being 0.8405 (4dp). When the models have some knowledge of the spatial location being estimated (through data which is close spatially) the longfin eel RRF model returns relatively accurate estimates of the probability of capture for longfin eels. This is shown by the relatively high AUC values and the little variation that appears between the ROC curves for each fold.

5-fold cross validation was performed for the longfin eel RRF model. The ROC curve for the cross validation is given in Figure 4.7. The 5-fold cross validation returned almost identical results to the 50-fold cross validation. The mean AUC is 0.7799 (4dp) with 95% confidence interval 0.7710

and 0.7890 (with a standard error of 0.0045).

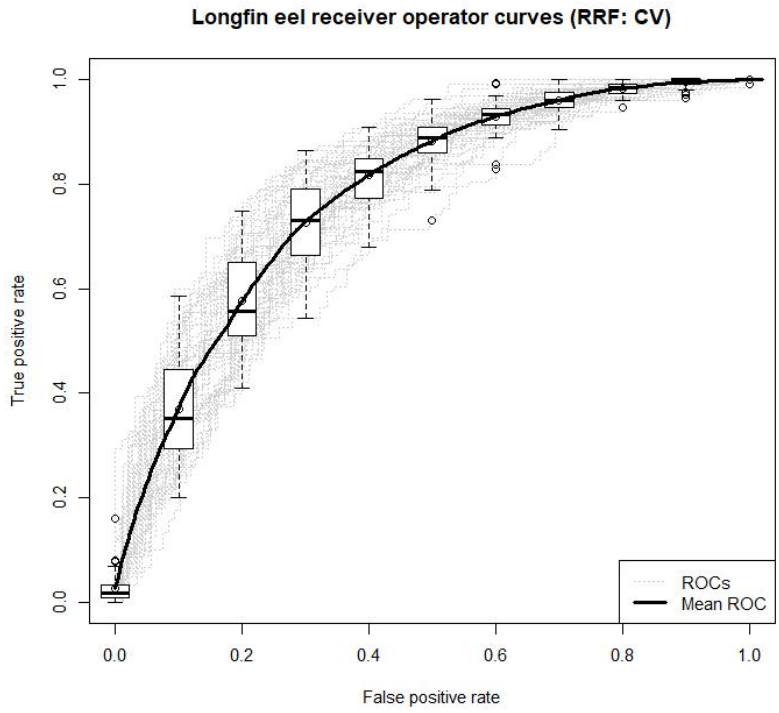


Figure 4.6: ROC curves under each of the 50-folds in the longfin eel RRF model cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

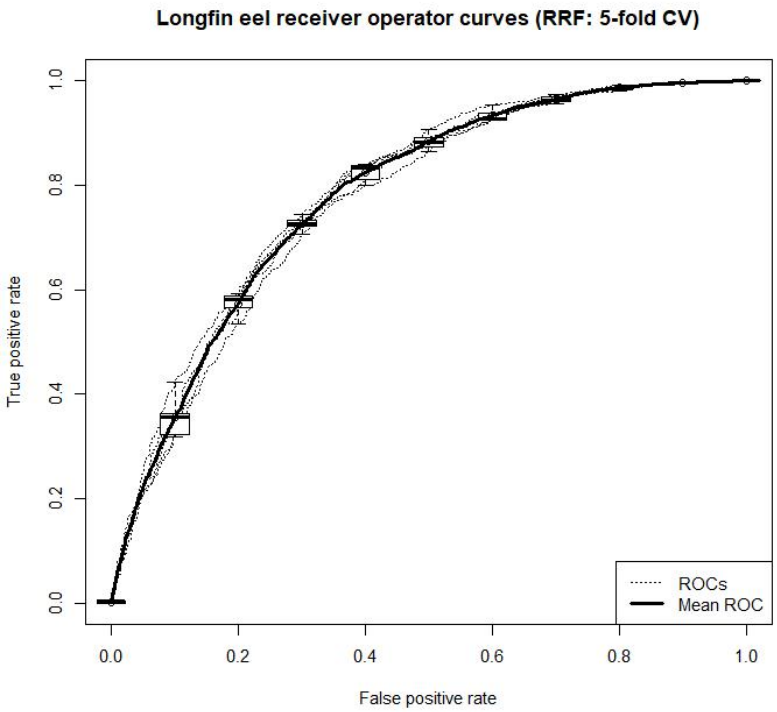


Figure 4.7: ROC curves under each of the 5-folds in the longfin eel RRF model cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

4.1.2 Shortfin eel results

Model results

The shortfin eel RRF model selected 55 features in the model feature selection process. There are some differences between the features selected by the shortfin eel RRF model and by the longfin eel RRF model. Figure 4.8 shows the importance score of each of the features. Each of the features are grouped into 'excluded' or 'included'. The features in the 'excluded' group have an importance score of zero and were excluded from the model, and the features in the 'included' group had an importance score greater than zero and were included in the model.

The feature 'seg_twar' (the average January temperature within a segment of river in deg. C $\times$ 10) achieved the highest importance score of c.660 in the shortfin eel RRF model. This was the third highest importance score in the longfin eel RRF model. Similar to the longfin eel RRF model, many of the spatial features were excluded from the model (only 'x3' was included). However, the shortfin eel RRF model also excluded many of the hydrological and environmental features. Many of the same features such as 'StreamOrder' (the stream order) were excluded from both the longfin and shortfin eel RRF models.

The probability of capture estimates made across the entire REC2 database using the shortfin eel RRF model are shown in Figure 4.9. The North Island of New Zealand contains probability of capture estimates around 0 to 0.3 throughout the central North Island which spreads to the east, west and southern coast. Observed proportions of shortfin eel capture values (shown in Figure 4.2) were found to be of similar size, in similar locations.

Large rivers in the northern Waikato region estimated probabilities of capture close to 1. Observed proportions of shortfin eel capture did not show such high values for the same Waikato rivers. The northern North Island contains probabilities of capture mainly around 0.2 to 0.7. The

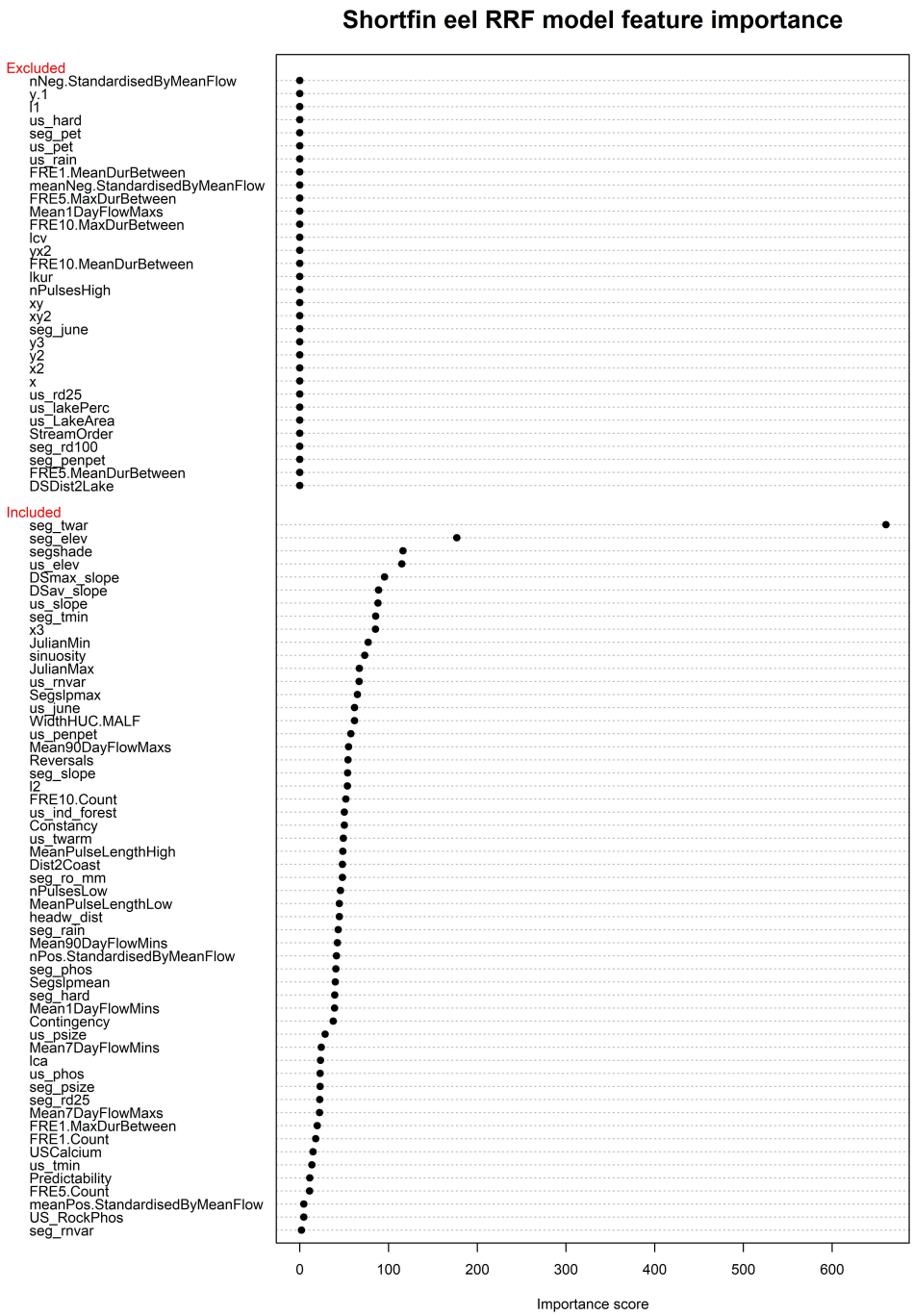


Figure 4.8: Importance scores for the features of the shortfin eel RRF model. Features with importance scores greater than 0 were included in the model. For the features included in the model, the features are ordered from largest importance score to smallest importance score.

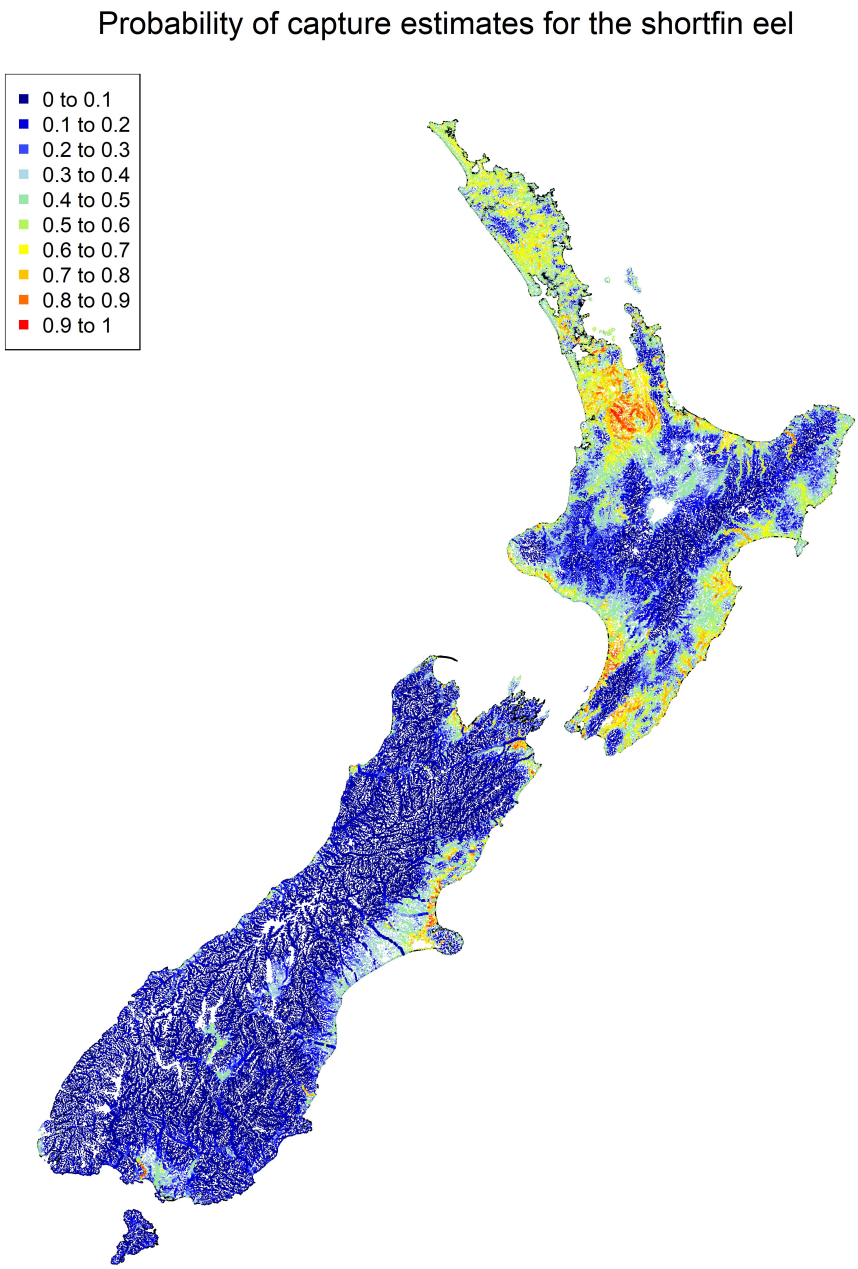


Figure 4.9: Probability of capture estimates for the shortfin eel using the REC2 database. These estimates were made using the shortfin eel RRF model. Larger points have a larger stream order.

observed proportions for this area were very similar but included some lower estimates of 0 to 0.1 and 0.1 to 0.2.

The west coast of New Zealand's Manawatu region has high probabilities of capture (around 0.7 to 0.9). The observed proportions for this region was lower than this (0.3 to 0.6). As noted with the longfin eels, various small areas of the central North Island do not have any data (as shown in Figure 2.1) and therefore direct comparisons between what was observed and what was predicted cannot be made.

The majority of the South Island of New Zealand estimated low probabilities of capture (0 to 0.3) for the shortfin eel. Similarly, the observed proportions of shortfin eel capture mostly consisted of estimates ranging between 0 to 0.3 (as shown in Figure 4.2). Very small areas of Invercargill, Nelson, Blenheim, and a large area surrounding Christchurch have probabilities of capture exceeding 0.6. This is unsurprising given that these same areas have high observed proportions (as shown in Figure 4.2). Stewart Island has very low probabilities of capture (approximately 0 to 0.1) which is consistent with Figure 4.2.

Cross validation results

The spatial 50-fold cross validation for the shortfin eel RRF model resulted in a mean AUC of 0.7443 (4dp) with a 95% confidence interval of 0.7329 and 0.7557 (standard error of 0.0058). The ROC curves for each fold of the 50 spatial folds and the mean ROC curve are shown in Figure 4.10. The boxplots in Figure 4.10 show the spread of the ROC curves.

Similar to the spatial 50-fold cross validation for the longfin eel RRF model (Figure 4.5), the ROC curves of Figure 4.10 have very larger variability. The variability appears to be greater than that of the spatial cross validation for the longfin eels. The smallest AUC was 0.5193 (4dp) and

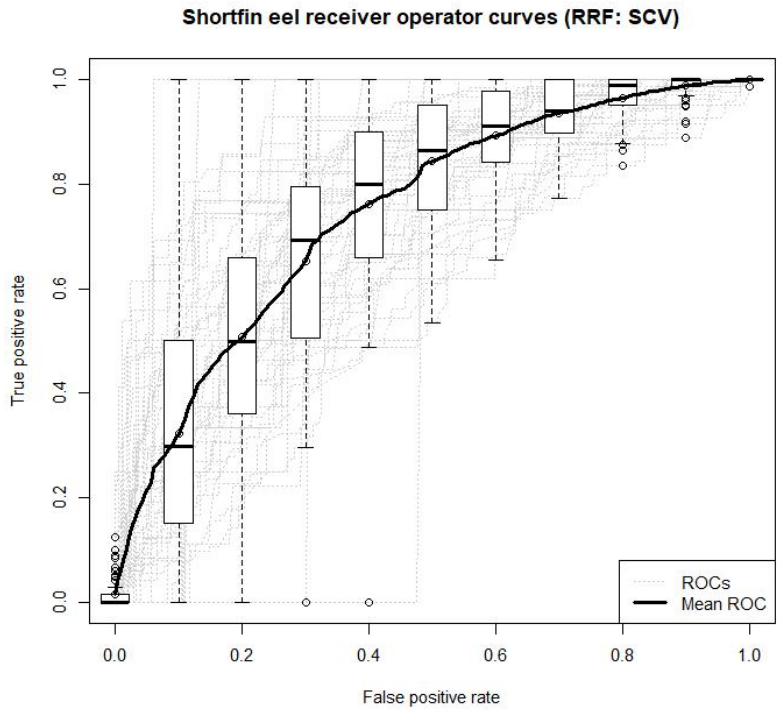


Figure 4.10: ROC curves under each of the 50-folds in the shortfin eel RRF model spatial cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

the largest was 0.8372 (4dp). When the AUC was 0.5193, the model was only just performing better than guessing shortfin eel presence or absence. This occurred in a spatial fold close to Dunedin and is due to the model being data poor in this area. The large variability in ROC curves shows that some spatial areas were predicted much better than others. But, on average, a RRF model for the shortfin eel do reasonably well in predicting to areas which are spatially distinct to the training data.

The 50-fold cross validation for the shortfin eel RRF model resulted in a mean AUC of 0.8692 (4dp) with a 95% confidence interval of 0.8613 and 0.8771 (standard error of 0.0040). The ROC curves for the 50 randomly selected folds are shown in Figure 4.11. The boxplots at fixed intervals of the ROC curves show little variability. The minimum AUC value was 0.8046 (4dp) and the maximum was 0.9152 (4dp).

In comparison to the shortfin eel spatial 50-fold cross validation results (Figure 4.10), the AUC values are larger and have less variability. Hence, when the shortfin eel RRF model contains training data which is in close proximity to where the model is predicting then the predictions are more accurate to when the model doesn't have this knowledge. Overall the shortfin eel RRF model does very well in predicting to locations which are spatially dependent.

5-fold cross validation was performed for the shortfin eel RRF model. The ROC curves for the cross validation is given in Figure 4.12. The 5-fold cross validation returned almost identical results to the 50-fold cross validation. The mean AUC is 0.8674 (4dp) with 95% confidence interval 0.8595 and 0.8754 (with a standard error of 0.0041).

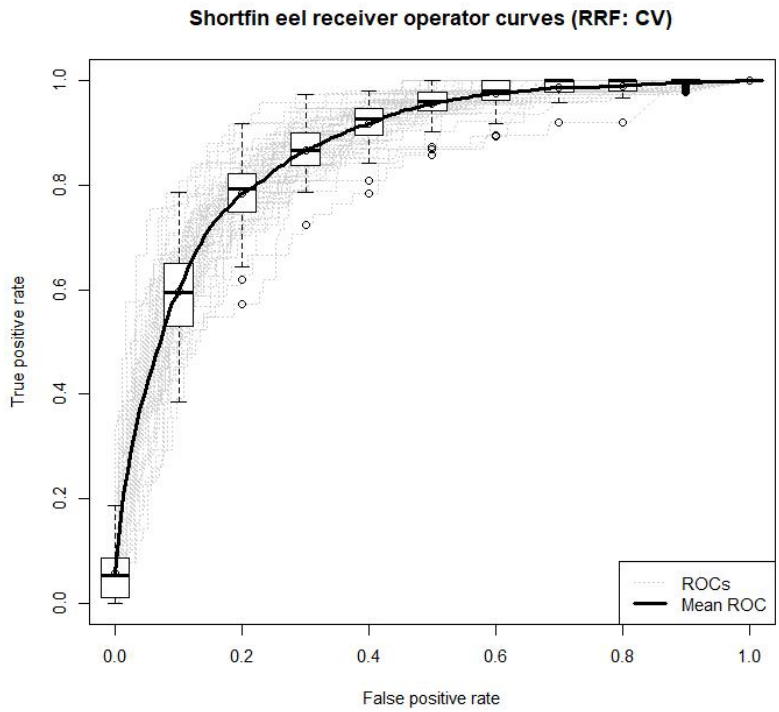


Figure 4.11: ROC curves under each of the 50-folds in the shortfin eel RRF model cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

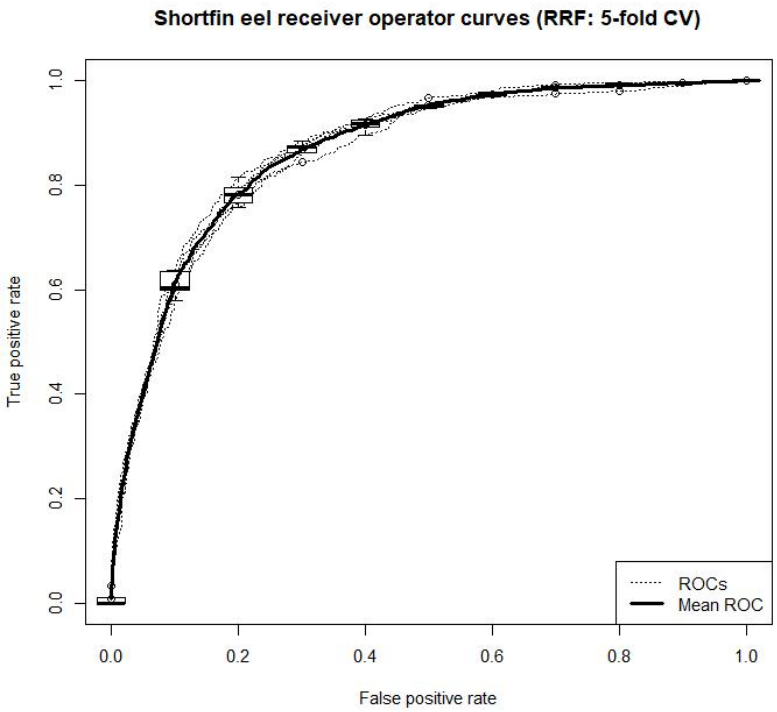


Figure 4.12: ROC curves under each of the 5-folds in the shortfin eel RRF model cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

4.2 VAST modelling results

This section describes the results of the VAST models. The data used for the models are detailed in Chapter 2 and the VAST method is detailed in Section 3.2.

A probability of capture model for longfin eels, shortfin eels and for both species (known as a multi-species model) were constructed. The models were run using a NIWA server with 245GB of RAM. This enabled bias correction (Thorson & Kristensen, 2016) to be used at a finer resolution (i.e. more knots) than a standard computer would allow for. VAST estimates spatial and spatio-temporal variation at each of 400 knots which are distributed across the domain. Probability of capture maps are then built by interpolating within the domain specified. In this case the domain is the whole of New Zealand.

Figure 4.13 shows a map of the knots (on a northing-easting coordinate grid) and of the interpolation areas (on a latitudinal-longitudinal and northing-easting coordinate grid) for all the VAST models. The interpolation areas were within a maximum distance of 15km from any knot. This was deemed an acceptable distance for interpolation given that the further away one interpolates, the less correlated the sampling point is to the interpolated point.

50-fold spatial cross validation, 50-fold cross validation and 5-fold cross validation was performed on each of the models. Validation was performed on models with 1000 knots and without bias correction. Bias correction would significantly increase computation time and 1000 knots was used to increase the resolution of the results.

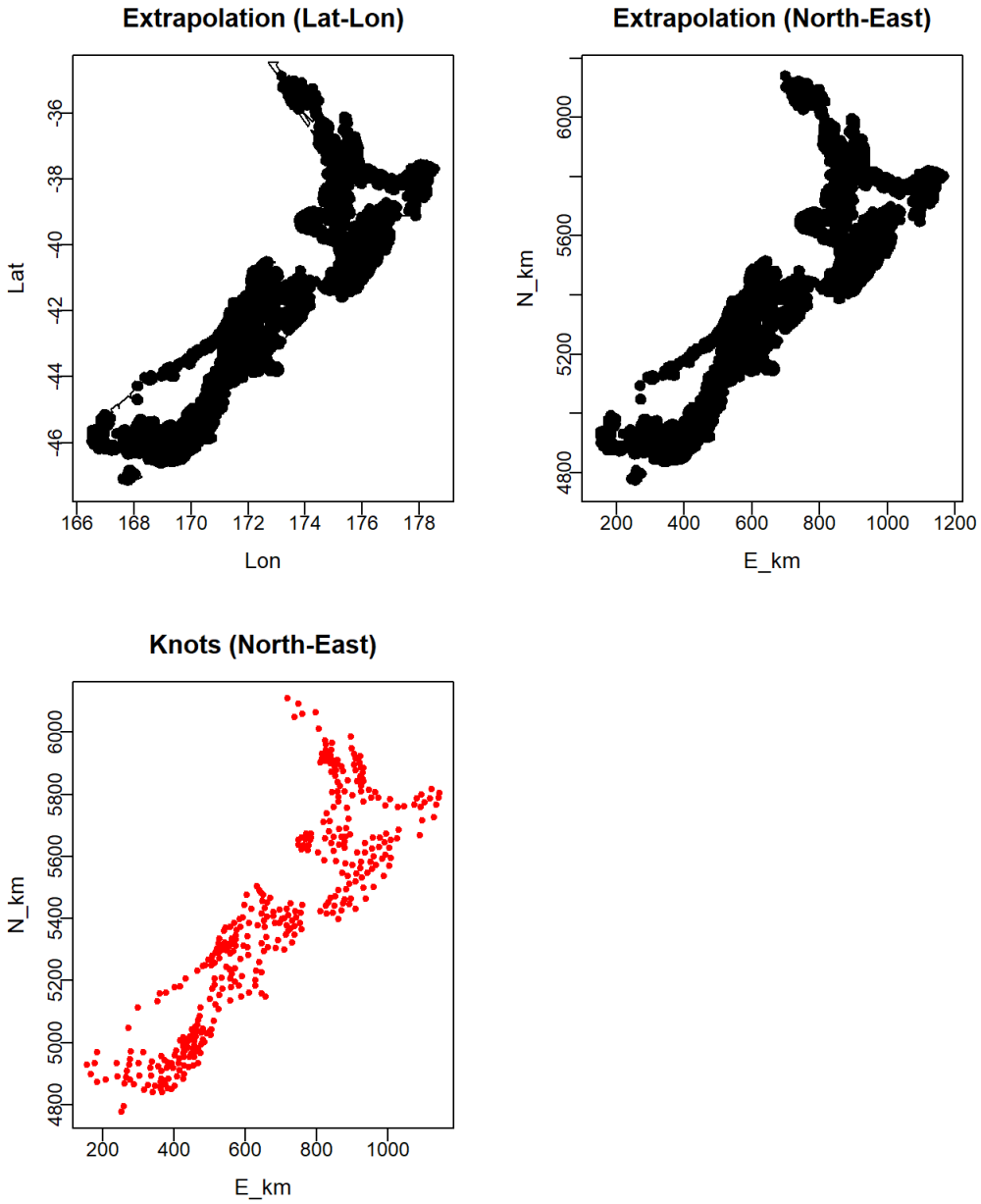


Figure 4.13: The upper maps show interpolation areas in a latitude-longitude coordinate grid and a northing-easting coordinate grid. The bottom map shows the locations of the knots used by VAST on a northing-easting coordinate grid.

4.2.1 Longfin eel results

Model results

Estimates of the fixed effects for the longfin eel VAST model are given in Table 4.1. The terms of the matrix $\mathbf{H}$ and the parameter κ_1 which define the correlation function $\mathbf{\Psi}_1(s, s+h')$ are given. The natural log of the terms defining the $\mathbf{H}$ matrix are -0.5627 ($\ln h_1$) and 0.2550 ($\ln h_2$) with standard errors 0.2017 and 0.2238 respectively. These are large standard errors relative to their estimated value and indicates uncertainty in these estimates. The term $\log \kappa_1$ is estimated as -3.5958 with standard error 0.1338. This is used to find the distance at which spatial correlation was 10% of the original correlation.

Parameter	Estimate	Standard error	C.V. (%)
$\ln h_1$	-0.5627	0.2017	35.8%
$\ln h_2$	0.2550	0.2238	87.8%
$\log \kappa_1$	-3.5958	0.1338	3.7%
L_{Ω_1}	0.8979	0.1222	13.6%
L_{ϵ_1}	1.2434	0.0721	5.8%
$\log \sigma_{\beta_1}$	-2.9477	0.7043	23.9%

Table 4.1: The estimated fixed effects (4dp), associated standard errors (4dp) and coeffient of variation (C.V.) (1dp) of the longfin eel VAST model.

Table 4.1 gives the coefficient of variation (C.V.) for each of the estimated model parameters. The C.V. for $\ln h_1$ and $\ln h_2$ are very large and indicate large uncertainty in these parameters. Hence, we lack certainty in how longfin eel data is correlated in space and direction. Caution should be taken when interpreting model predictions. The C.V. for the other model parameters shown in Table 4.1 are all less than 25%. These parameters show reasonable level of precision.

Figure 4.14 displays the direction in which geometric anisotropy oc-

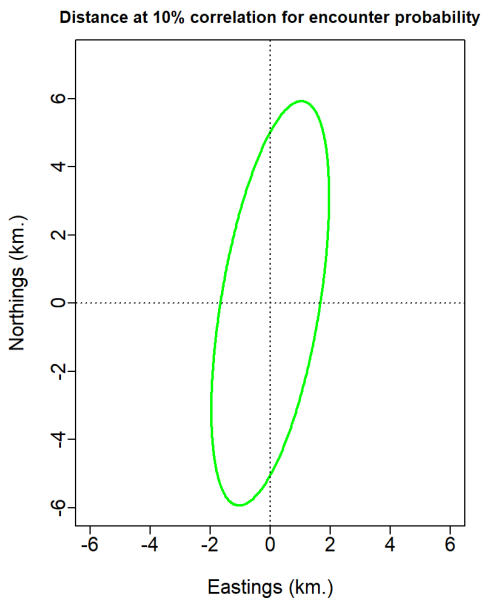


Figure 4.14: Plot representing geometric anisotropy for encounter probability from the longfin eel VAST model on a easting-northing coordinate grid.

curs for encounter probability. The plot shows that spatial decorrelation occurred slower in a slightly north-east and south-west direction. 10% correlation occurred at a distance of approximately 6km in this direction.

The terms L_{Ω_1} and L_{ϵ_1} which define spatial covariation and the spatio-temporal covariation respectively are also given in Table 4.1. The term L_{Ω_1} was estimated as 0.8979 with standard error 0.1222, and L_{ϵ_1} was estimated as 1.2434 with standard error 0.0721. Lastly, the terms defining the intercept parameter of Equation 3.10 is defined by the fixed effect term $\sigma^2_{\beta_1}$. From Table 4.1 we can see that $\log \sigma_{\beta_1}$ was estimated as -2.9477 with standard error 0.7043. Hence, σ_{β_1} was estimated as 0.0011 (4dp) and indicates small variability in the distribution of the autoregressive term $\beta_1(t+1)$. Hence, the 'baseline' effect of probability of capture for the longfin eel VAST model changes very little with time. The final gradients for each of the fixed effects of Table 4.1 were all approximately zero which indicates

that the model converged successfully.

Longfin eel probability of capture estimates from 1974 to 2014 are shown in Figure 4.15. The estimates are mapped on a northing-easting coordinate grid. Areas in dark red indicate that longfin eels had a high probability of being observed whereas areas in dark blue indicate that longfin eels had a low probability of being observed. Areas in a green or yellow colour indicate that longfin eels had a probability of being observed which ranges from approximately 0.4 to 0.6.

The 2014 probability of capture estimates for the longfin eel are shown in Figure 4.16. From Figure 4.16 we can see that the east and west coast of the North Island of New Zealand had high probabilities of capturing a longfin eel. These regions often exceeded probabilities of capture of 0.7. This is consistent with the observed proportions of longfin eel capture in Figure 4.1. The tip of the North Island had probabilities of capture around 0.7 to 0.8 and the central North Island showed very low probabilities of capture for longfin eels. A small area within the central North Island of New Zealand was outside of coverage due to the construction of knots. The observed proportions of longfin eel capture showed values ranging from 0.4 to 0.8 in the tip of New Zealand's North Island. However, the observed proportions of Figure 4.1 were mostly between 0.5 to 0.6 in the central North Island.

The South Island of New Zealand showed high probabilities of capture in the north of the South Island, the centre of the west coast, and around the cities of Invercargill and Christchurch. The central South Island extending towards the east coast were estimated as very low for probability of capture of longfin eel (0 to 0.3). Likewise, a small area in the south-west of the South Island and all of Stewart Island showed low probabilities of capture. The remaining areas tended to show probabilities of capture ranging from 0.3 to 0.7. These patterns are consistent with the observed proportions of longfin eel capture in Figure 4.1. A large area within the South Island does not have any estimates for the probability of capture

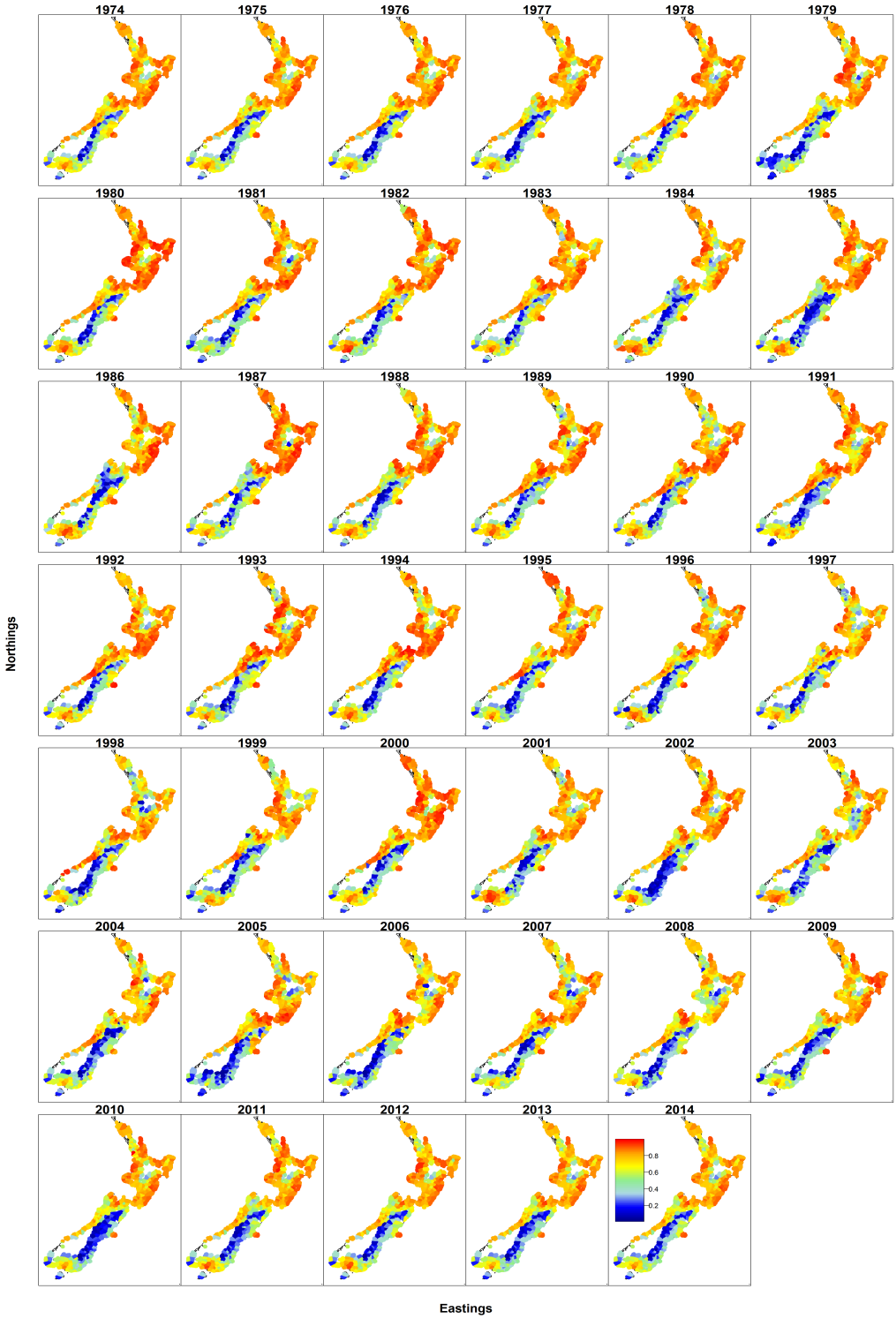


Figure 4.15: Longfin eel probability of capture estimates for 1974 to 2014 from the longfin eel VAST model. The estimates are shown on a northing-easting coordinate grid.

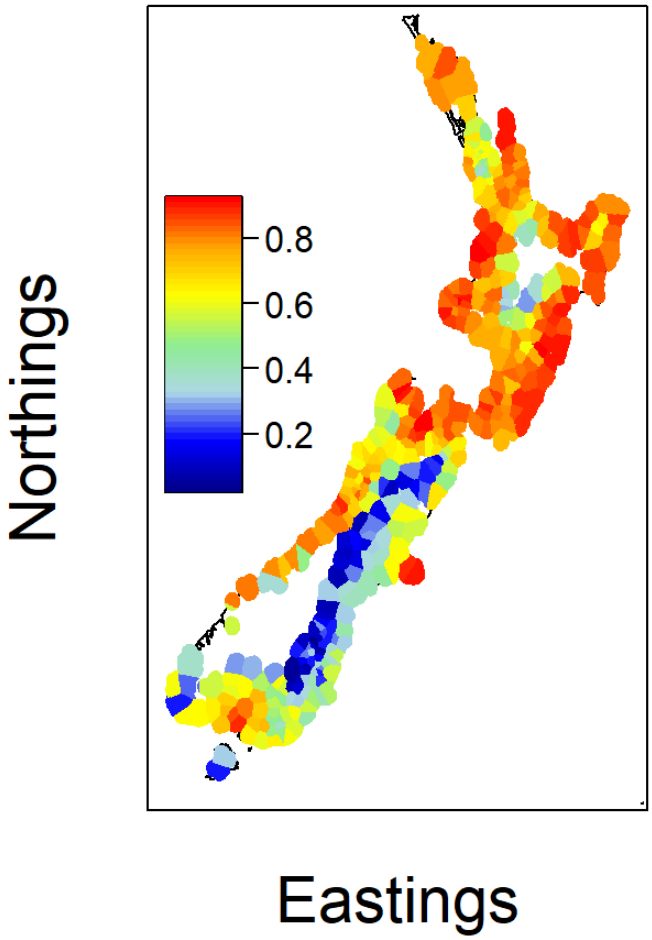


Figure 4.16: Map of the 2014 longfin eel probability of capture estimates made from the longfin eel VAST model. The estimates are shown on a northing-easting coordinate grid.

because of the way knots were placed within the domain.

These approximate patterns for longfin eel probability of capture appeared from 1974 to 2014 (as shown by Figure 4.15). However, there were some strong differences such as in 1980 and 1987 where the North Island of New Zealand had very high probabilities of capture (apart from the central North Island).

The Pearson residuals vs. fitted values for the longfin eel VAST model are shown in Figures 4.17 and 4.18. Pearson's residuals are given by Equation 3.13.

Figure 4.17 shows the Pearson residuals against the fitted values for each year and Figure 4.18 shows all the Pearson residuals against the fitted values on one plot. Some of the years from Figure 4.17 show unequal variability in the residuals, while others show constant variance but non-random scatter. Each plot of Figure 4.17 (except plots with very little data) tend to have residuals clustering at higher fitted values and/or at lower fitted values. This is most obvious in Figure 4.18 which gives an overall picture of the residuals. Clustering patterns indicate that there may still be an underlying correlation in the data which hasn't been fully accounted for in the model.

It is clear from Figure 4.18 that residuals do not vary around zero evenly. Residuals appear to be larger at small fitted values and smaller at higher fitted values with larger groupings at both ends. Additionally, Figure 4.18 shows a decreasing trend in the residuals. This indicates that there may be an underlying pattern in the data which hasn't been fully accounted for.

Figure 4.19 gives the Pearson's residuals on maps. When accounting for year, the residuals appear to be evenly scattered above and below zero, and are not very large. This is because each map has light blue and red residuals and none of the maps show one colour more prominently.

Figure 4.20 is a QQ-plot for the longfin eel VAST model. The plot indicates that the residuals follow a Normal distribution as the residuals follow the line very closely.

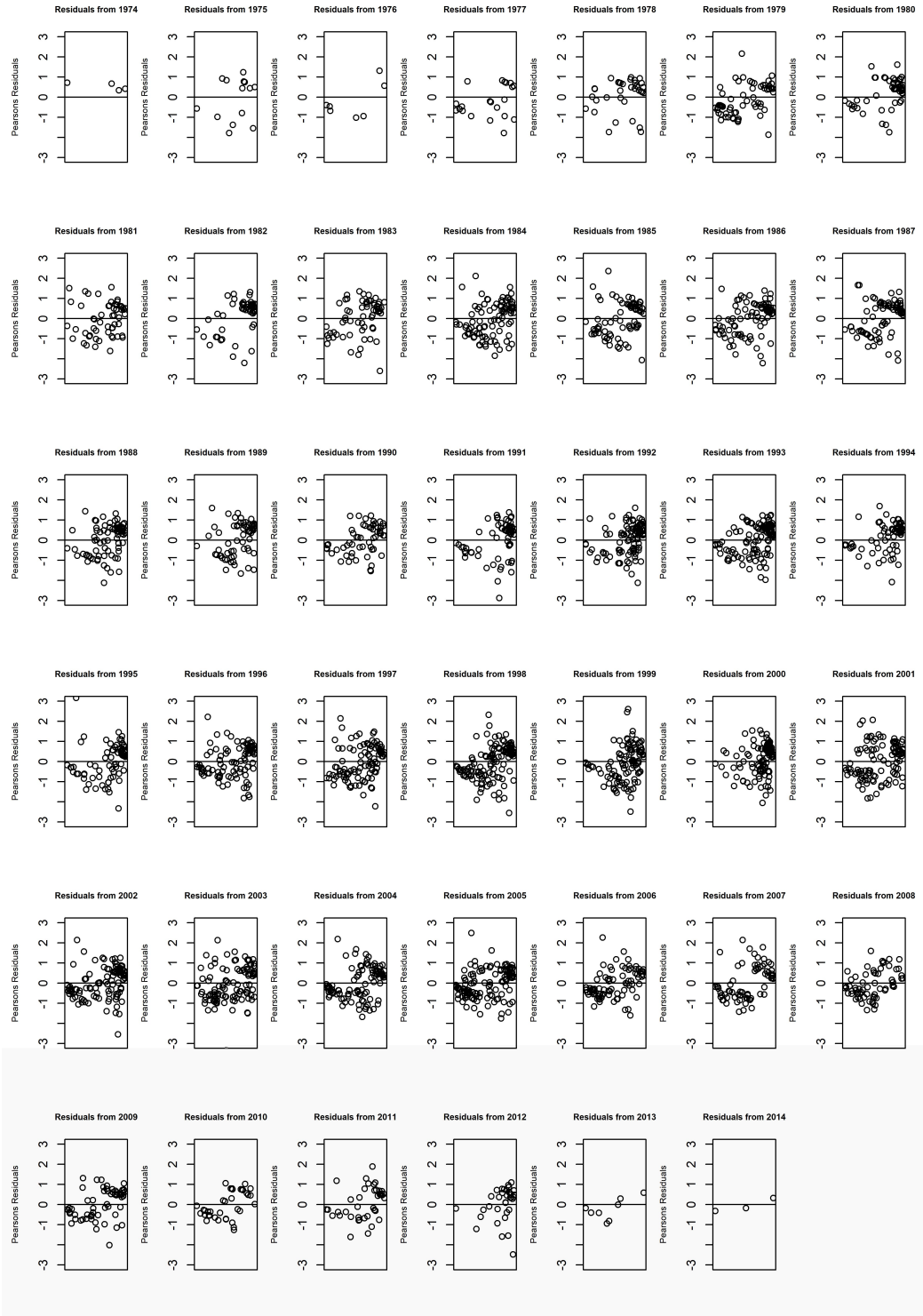


Figure 4.17: Plots of the Pearson residuals vs. fitted values for the longfin eel VAST model. These are given across every year of sampling from 1974 to 2014.

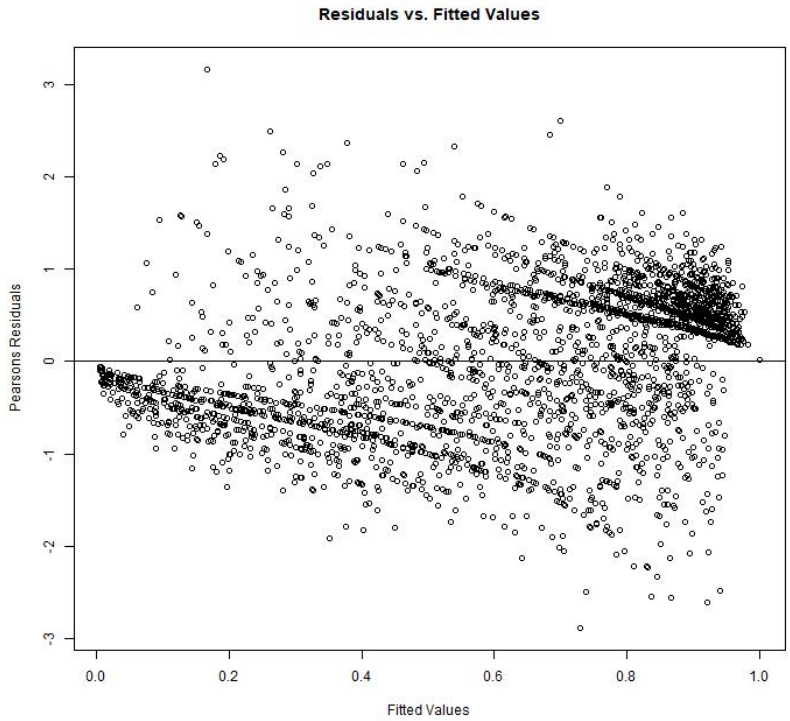


Figure 4.18: Plot of all the Pearson residuals vs. fitted values for the longfin eel VAST model.

Figure 4.21 shows observed encounter frequency against the predicted encounter probability (i.e. predicted probability of capture). The red area shows the 95% confidence interval band for the predicted probability of capture. The points at the upper and lower ends of the plot do not fall within the 95% predictive interval.

Observed points fall below the 95% predictive interval when the predicted encounter is low and observed points are above this band at high predicted encounter probability. But points tend to fall within this band when the observed encounter frequency is between 0.2 to 0.7. This indicates that the model is over-estimating small observed encounter frequency and under-estimating large observed encounter frequencies. This is not surprising given what was observed in the residuals. However,

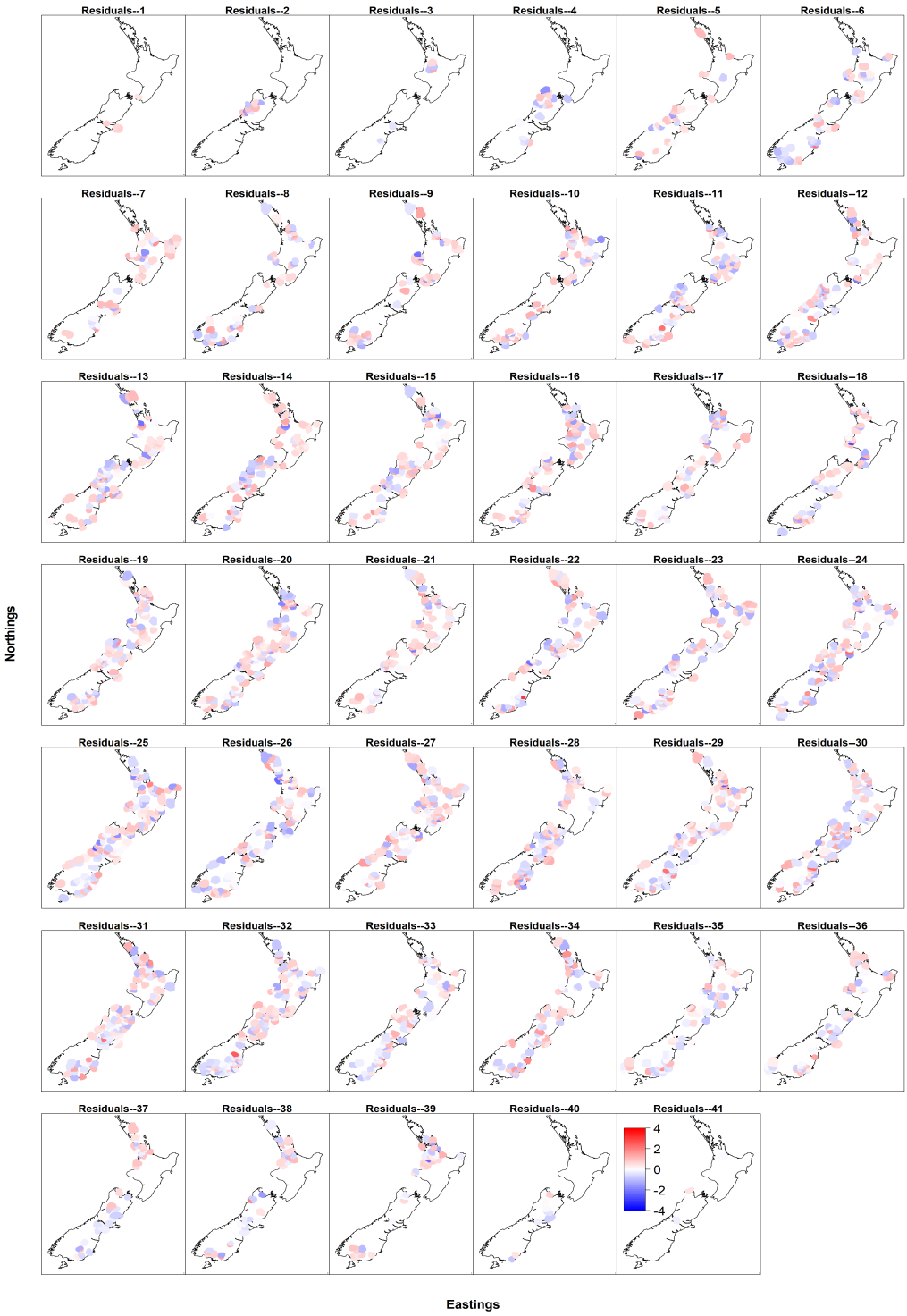


Figure 4.19: Heat maps of the longfin eel VAST model's Pearson residuals.

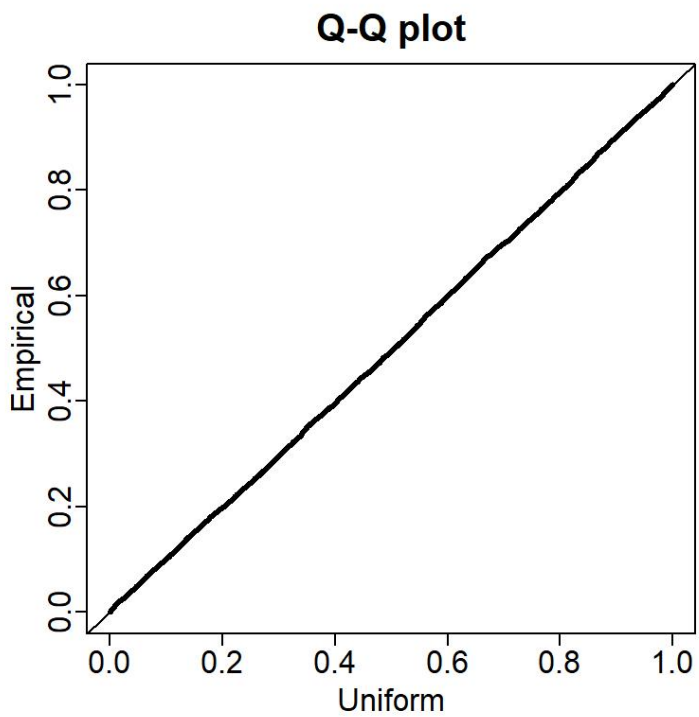


Figure 4.20: QQ-plot for the longfin eel VAST model.

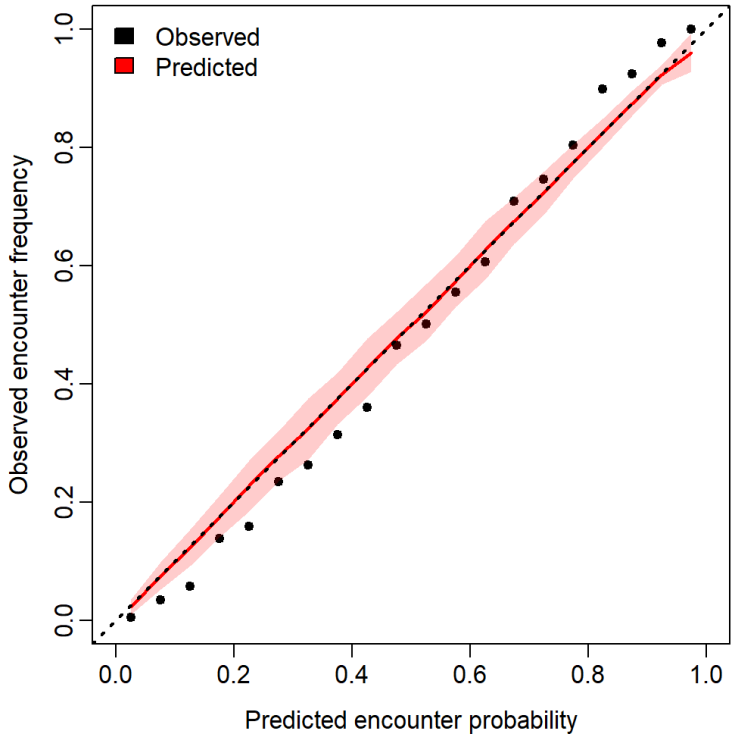


Figure 4.21: A diagnostic plot for observed encounter frequency against the predicted encounter probability for the longfin eel VAST model.

the observed encounter frequencies only just fall outside of this band and therefore is not a major concern. Nevertheless, this must still be taken into consideration when examining the probability of capture estimates.

Cross validation results

A spatial 50-fold cross validation was performed on the longfin eel VAST model. The area under the ROC curve (AUC) was calculated for each fold. The ROC curves are shown in Figure 4.22 where the mean ROC curve is shown by the dark line and the variability of the curves are shown by the boxplots. The mean AUC was 0.6646 (4dp) with 95% confidence interval 0.6542 and 0.6751 (standard error of 0.0053).

The ROC curves are highly variable with the lowest AUC being 0.4339 (4dp) and the largest being 0.8471 (4dp). The lowest estimate was found to be in a spatial fold in west Waikato. Hence, predictions made to this

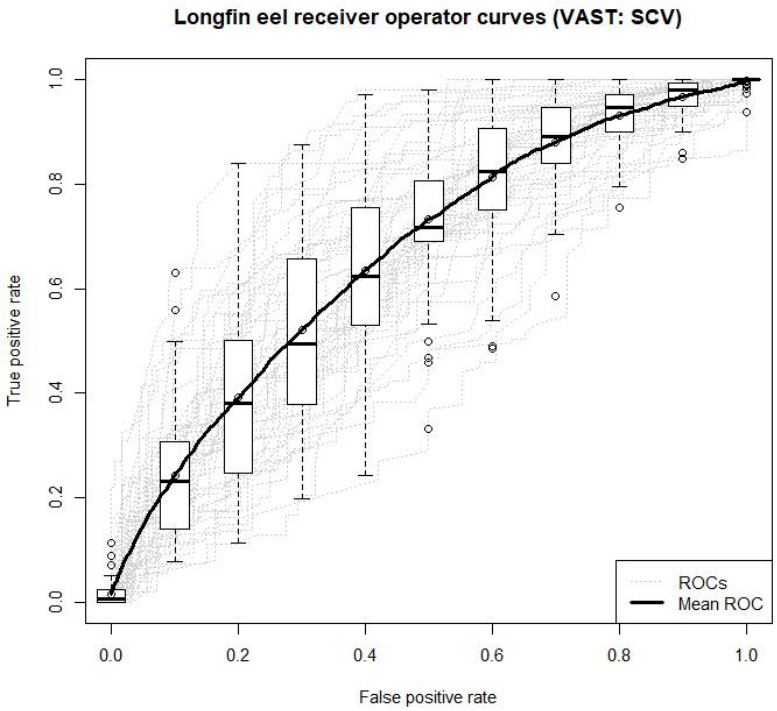


Figure 4.22: ROC curves under each of the 50 folds in the longfin eel VAST model spatial cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

area are very poor when the model lacks these spatial data points. When the AUC is less than 0.5, the model is estimating effects in the opposite direction to the true effect. On average, the model is performing fairly poorly when making predictions in areas outside of the spatial domain of the training data.

The 50-fold cross validation for the longfin eel VAST model resulted in a mean AUC of 0.8321 (4dp) with 95% confidence interval 0.8243 and 0.8399 (standard error 0.0040). The ROC curves for each of the 50 folds are shown in Figure 4.23. There is little variability in the receiver operating curves. The smallest AUC was 0.7727 (4dp) and the largest was 0.8798 (4dp). This shows that all 50 folds were predicted very well and that, on average, the longfin eel VAST model performs well when the test data set is spatially correlated to the training data set.

5-fold cross validation for the longfin eel VAST model found very similar results to the 50-fold cross validation. The ROC curves and mean ROC curve is shown in Figure 4.24. The mean AUC is 0.8269 (4dp) with a 95% confidence of 0.8190 and 0.8348 (with a standard error of 0.0040).

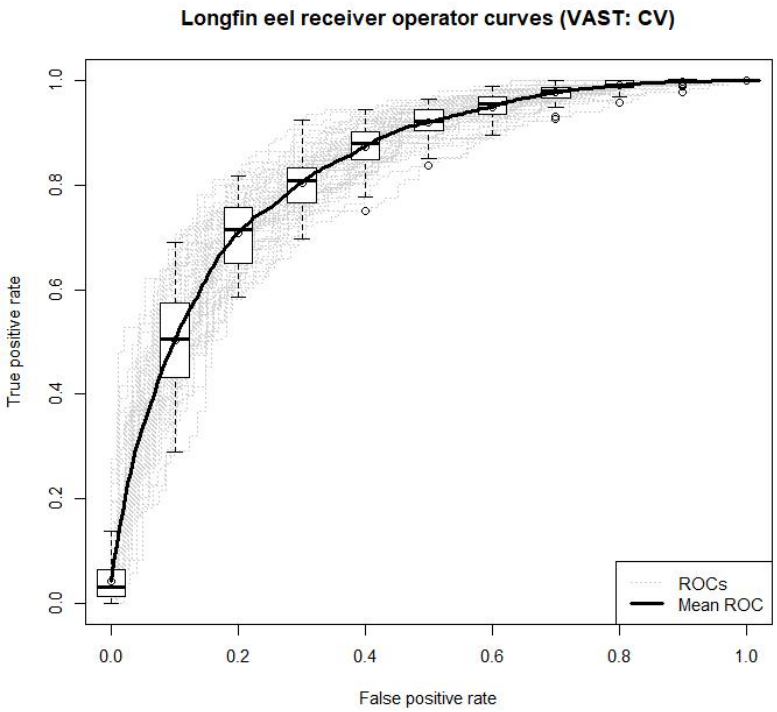


Figure 4.23: ROC curves under each of the 50 folds in the longfin eel VAST model cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

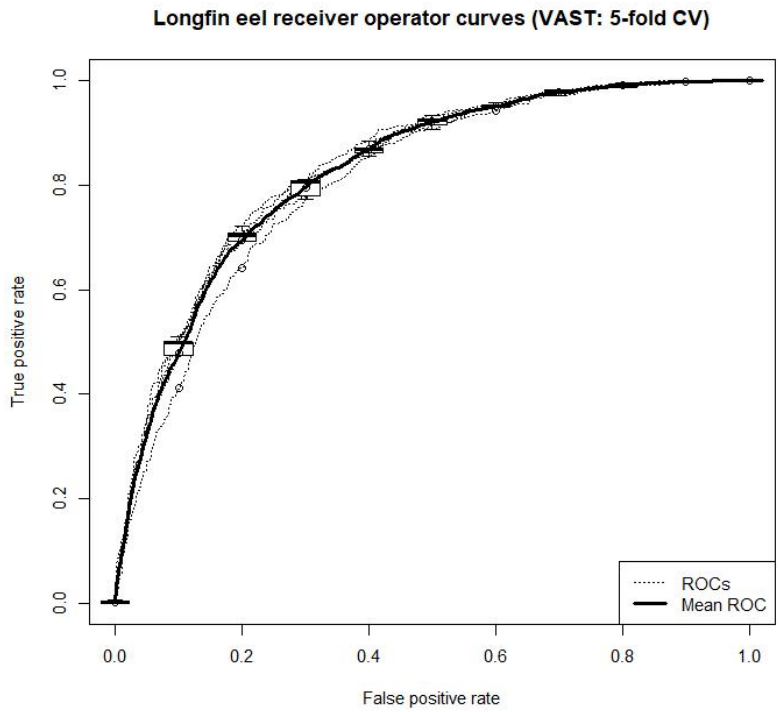


Figure 4.24: ROC curves under each of the 5 folds in the longfin eel VAST model cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

4.2.2 Shortfin eel results

Model results

Table 4.2 gives the estimates of the fixed effects for the shortfin eel VAST model. The natural log of the parameters defining **H** were given as -0.7503 ($\ln h_1$) and -0.2756 ($\ln h_2$) with respective standard errors of 0.3355 and 0.3746. The standard errors were large relative to their estimated values. The model set $\log \kappa_1$ as -3.2739 with standard error 0.2635. Figure 4.25 shows the distance at which spatial data points had a 10% correlation for encounter probability. Spatial decorrelation occurs slower in a slightly north-west and south-east direction. 10% correlation occurs at approximately 7km in those directions. This was the opposite direction of the longfin eel VAST model. The matrix **H** and κ_1 define the correlation function $\mathbf{\Psi}_1(s, s + h')$ defined in Section 3.2.3.

Parameter	Estimate	Standard error	C.V.
$\ln h_1$	-0.7503	0.3355	44.7%
$\ln h_2$	-0.2756	0.3746	135.9%
$\log \kappa_1$	-3.2739	0.2635	8.0%
L_{Ω_1}	1.0878	0.1504	13.8%
L_{ϵ_1}	1.2363	0.1247	10.1%
$\log \sigma_{\beta_1}$	-1.3666	0.3097	22.7%

Table 4.2: The estimated fixed effects (4dp), associated standard errors (4dp) and coefficient of variation (C.V.) (1dp) of the shortfin eel VAST model.

Table 4.2 give coefficient of variation (C.V) estimates for each of the model parameters. The anisotropy parameters $\ln h_1$ and $\ln h_2$ show very large C.V values where $\ln h_2$ is extremely large at 135.9%. This indicates very large imprecision in these parameters. Hence, caution should be taken in interpreting these parameters and model predictions. The C.V.

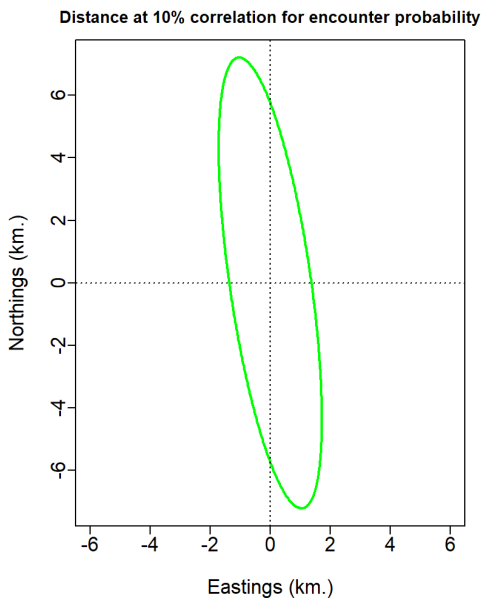


Figure 4.25: Plot representing geometric anisotropy for encounter probability from the shortfin eel VAST model on a easting-northing coordinate grid.

for the other model parameters shown in Table 4.2 are all less than 25%. These parameters show reasonable levels of precision.

The term which defines spatial covariation L_{Ω_1} was estimated as 1.0878 with standard error 0.1504 and the term which defines spatio-temporal covariation L_{ϵ_1} was estimated as 1.2363 with standard error 0.1247. The term defining the variability of the intercept parameter (see Equation 3.10) is termed $\sigma^2_{\beta_1}$. Where $\log \sigma_{\beta_1}$ was estimated as -1.3666 with standard error 0.3097. Hence, σ_{β_1} was estimated as 0.0430 (4dp). This gives the variability in the model autoregressive intercept term $\beta_1(t+1)$. There is very little variability in the distribution of this term which means that the 'baseline' probability of capture for shortfin eels stayed approximately the same with time. The final gradients for each of the fixed effects of Table 4.1 were all approximately zero.

Probability of capture estimates were made for shortfin eels from 1974

to 2014. These estimates are shown on heatmaps in Figure 4.26. The same colour scheme of the longfin eel heatmaps was used. Figure 4.27 is a heatmap of the probability of capture estimates for 2014 and were made using the shortfin eel VAST model.

In reference to Figure 4.27: in the North Island of New Zealand shortfin eels are unlikely to be encountered in the centre of the island (probability of capture is c.0.2). Similar estimates are made in a small area to the east of the central North Island and a larger area to the west of the central North Island. This is seen in the observed proportions of shortfin eel capture of Figure 4.2. There are gaps in shortfin eel probability of capture estimates in the central North Island of New Zealand and we therefore cannot make direct comparisons.

The northern areas of the North Island tend to have high probability of capture estimates of 0.7 or greater. The observed proportions of shortfin eel capture in this area are highly variable (between 0 and 1). The southern North Island, and east and west coast of the North Island tend to have probability of capture estimates between 0.6 to 1. This is inconsistent with the observed proportions of shortfin eel capture in Figure 4.2 which were less than 0.6.

In reference to Figure 4.27: in the South Island of New Zealand, shortfin eels are unlikely to be encountered throughout the central South Island, west coast, south coast and the southern east coast of the South Island. These areas tended to have probabilities of capture of 0.1 or less. This is approximately consistent with the observed proportions of shortfin eel capture in Figure 4.2. However, the majority of the east coast of the South Island (with the exception of the southern east coast) showed probability of capture estimates between 0.4 to 0.6. This is inconsistent with the observed proportions of shortfin eel capture in the east coast of the South Island. With the exception of the Christchurch region (where observed proportions of shortfin eel capture was between 0.4 to 0.7), the observed proportions of shortfin eel capture in the east coast of the South Island

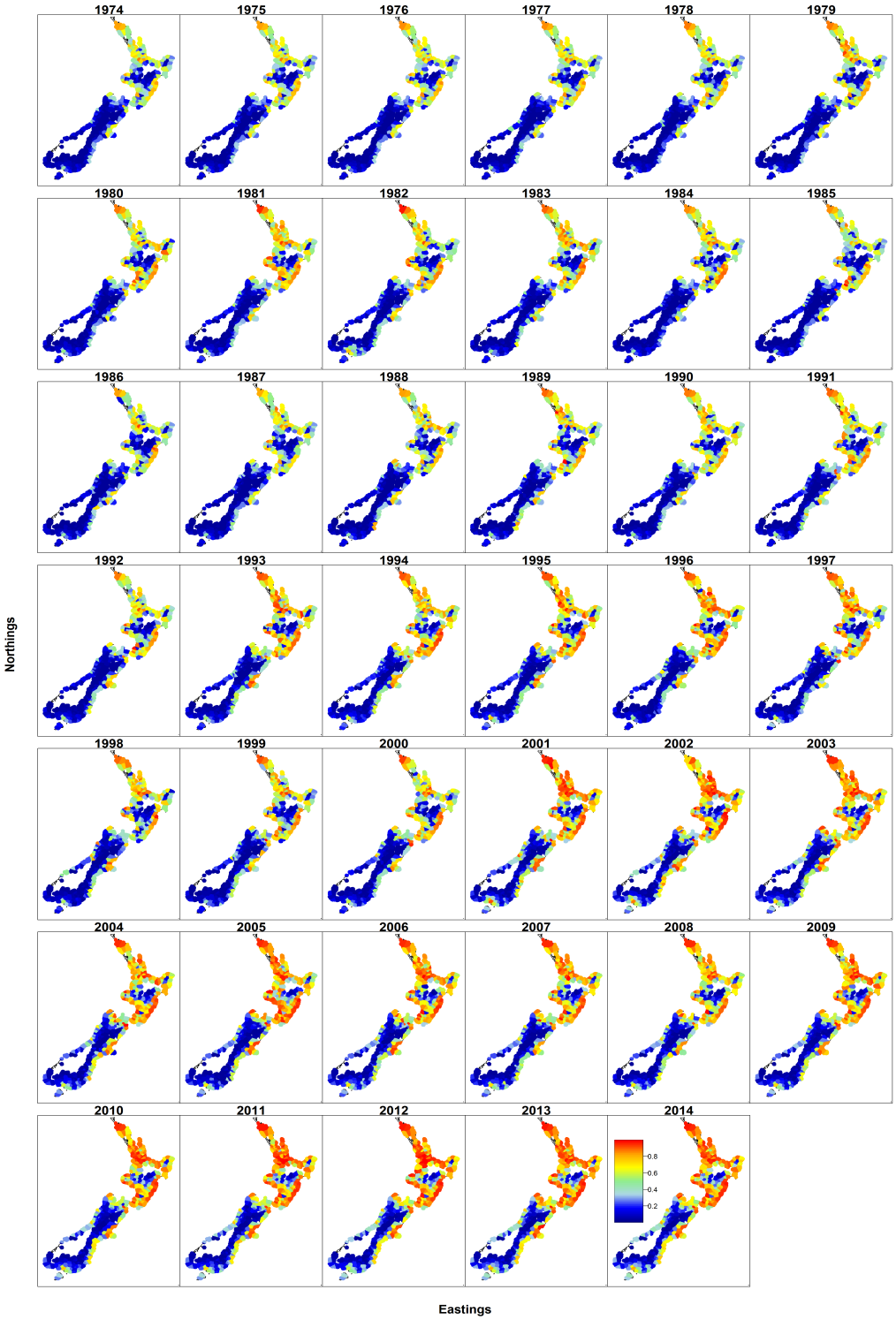


Figure 4.26: Shortfin eel probability of capture estimates for 1974 to 2014 from the shortfin eel VAST model. The estimates are shown on a northing-easting coordinate grid.

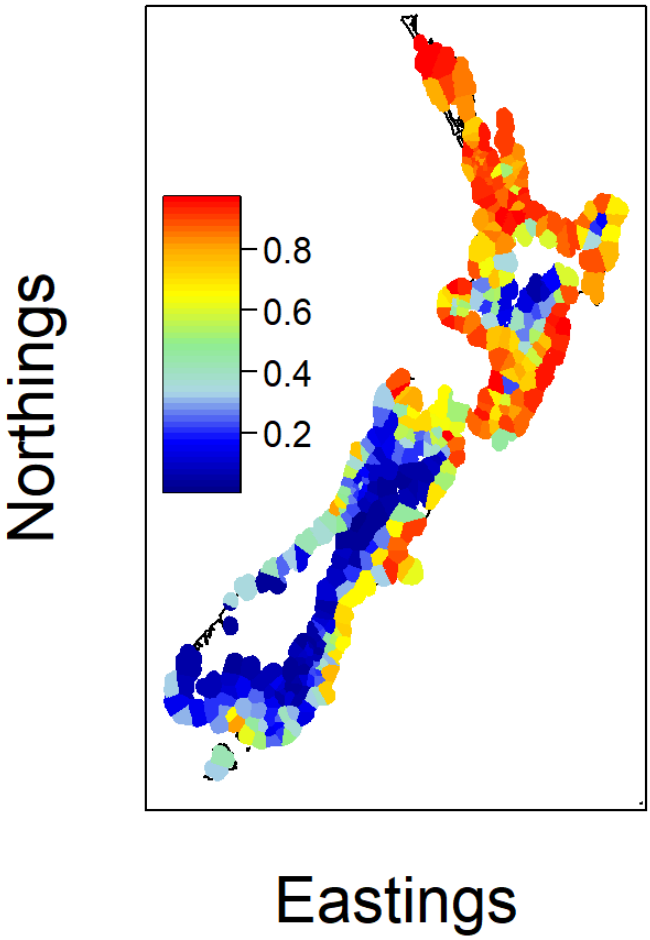


Figure 4.27: Map of the 2014 shortfin eel probability of capture estimates made from the shortfin eel VAST model. The estimates are shown on a northing-easting coordinate grid.

tended to range between 0 and 0.3.

The South Island of New Zealand has very few areas where the probability of capturing a shortfin eel was greater than 0.7. There is a large gap in probability of capture estimates in the South Island of Figure 4.27 because of the way knots were constructed

Similar to the longfin eel probability of capture patterns, the shortfin eel showed these approximate patterns throughout the years of 1974 to 2014. However, Figure 4.26 did show some changes. The North Island of New Zealand tends to show higher probabilities of capture as time increases. This is most obvious along the North Island's coasts and from 2001 to 2014. The coasts of the South Island appear to also increase in probability of capture as time increases. But this occurs only in smaller areas and continues to remain at low probabilities. However, the Christchurch region increases to probabilities exceeding 0.8.

Pearson's residuals of the shortfin eel VAST model were examined. Figure 4.28 shows the Pearson residuals vs. fitted values for each year and Figure 4.29 shows Pearson's residuals vs. fitted values for all the residuals. We can see from Figure 4.28 that the residuals, for some years, tend to be further from zero at larger fitted values. Hence, we observe a 'funnelling out' effect. This is most obvious in 1981, 2000, 2001 and 2009. But this doesn't appear to be a significant feature in the majority of the plots.

Equivalent to the residuals of the longfin eel VAST model, Pearson's residuals of the shortfin eel VAST model appear to show an underlying structure. This is most obvious in Figure 4.29, where the residuals with a fitted value close to zero are not randomly scattered. Instead they are clustered together.

Figure 4.29 also shows some outliers at lower fitted values. Most notable, there is one large outlier with a Pearson residual value close to 9 and a fitted value around 0.1. There were four outliers with a residual value greater than 4. These came from the years 1993, 1994, 1997 and 1999.

A residual of 8.96 (2dp) came from the year 1997. We can see this most

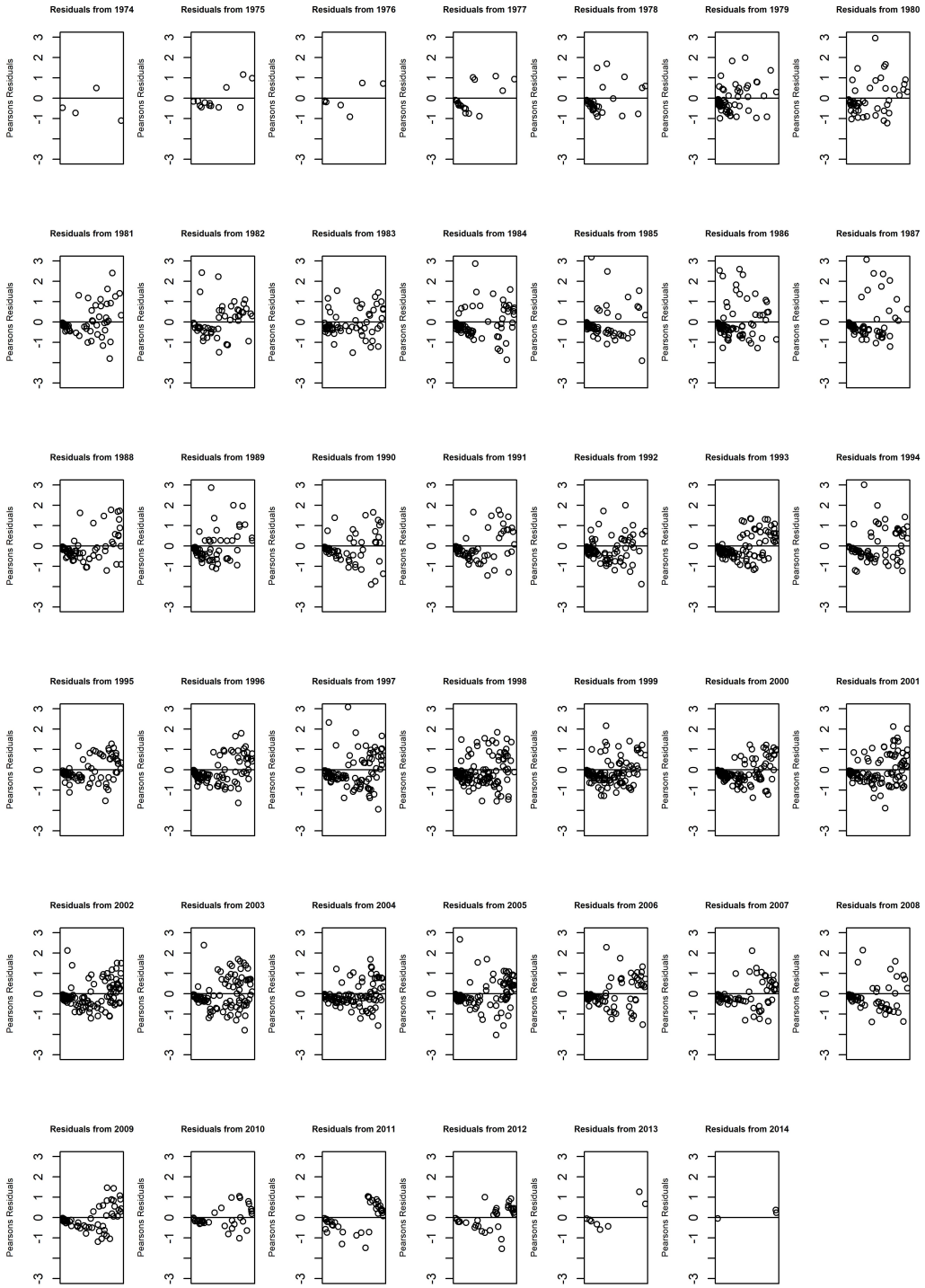


Figure 4.28: Plots of the Pearson residuals vs. fitted values for the shortfin eel VAST model. These are given across every year of sampling from 1974 to 2014.

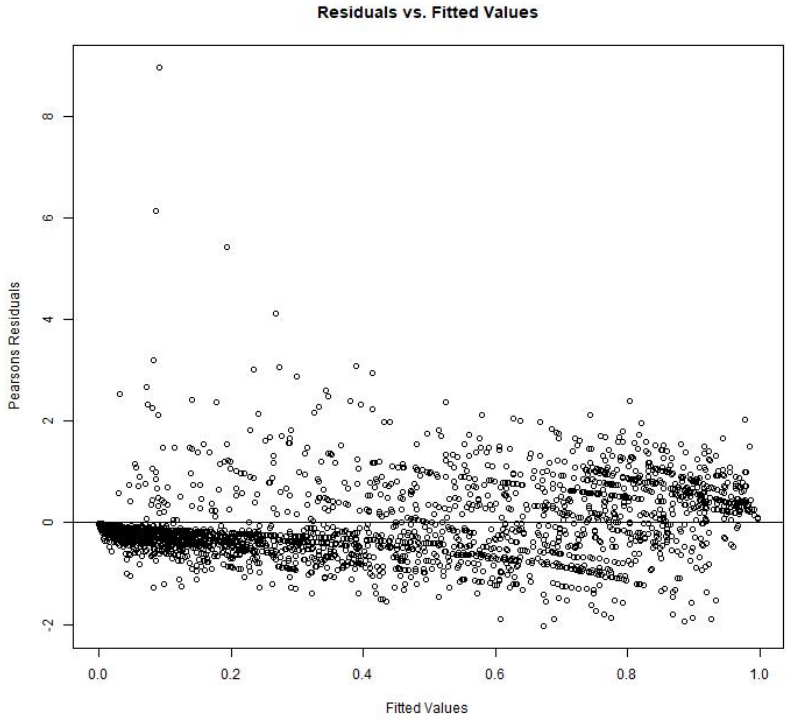


Figure 4.29: Plot of all the Pearson residuals vs. fitted values for the shortfin eel VAST model.

obviously in residual 24 (1997) of Figure 4.30. There is a bright red residual on the west coast of the South Island. This indicates that for 1997, that location in the South Island of New Zealand was over estimated. However, Figure 4.30 indicates that the residuals are small and scattered approximately around zero (equal number of red and blue points).

Figure 4.29 shows that the residuals have approximately homogenous variability around 0 when not accounting for the outliers. In general, these residuals do not appear to show any strong cause for concern.

Figure 4.31 is a QQ plot for the shortfin eel VAST model. The plot indicates that Pearson's residuals follow a Normal distribution. This is because the residuals follow the QQ line very well.

An observed encounter frequency vs. predicted encounter probabil-

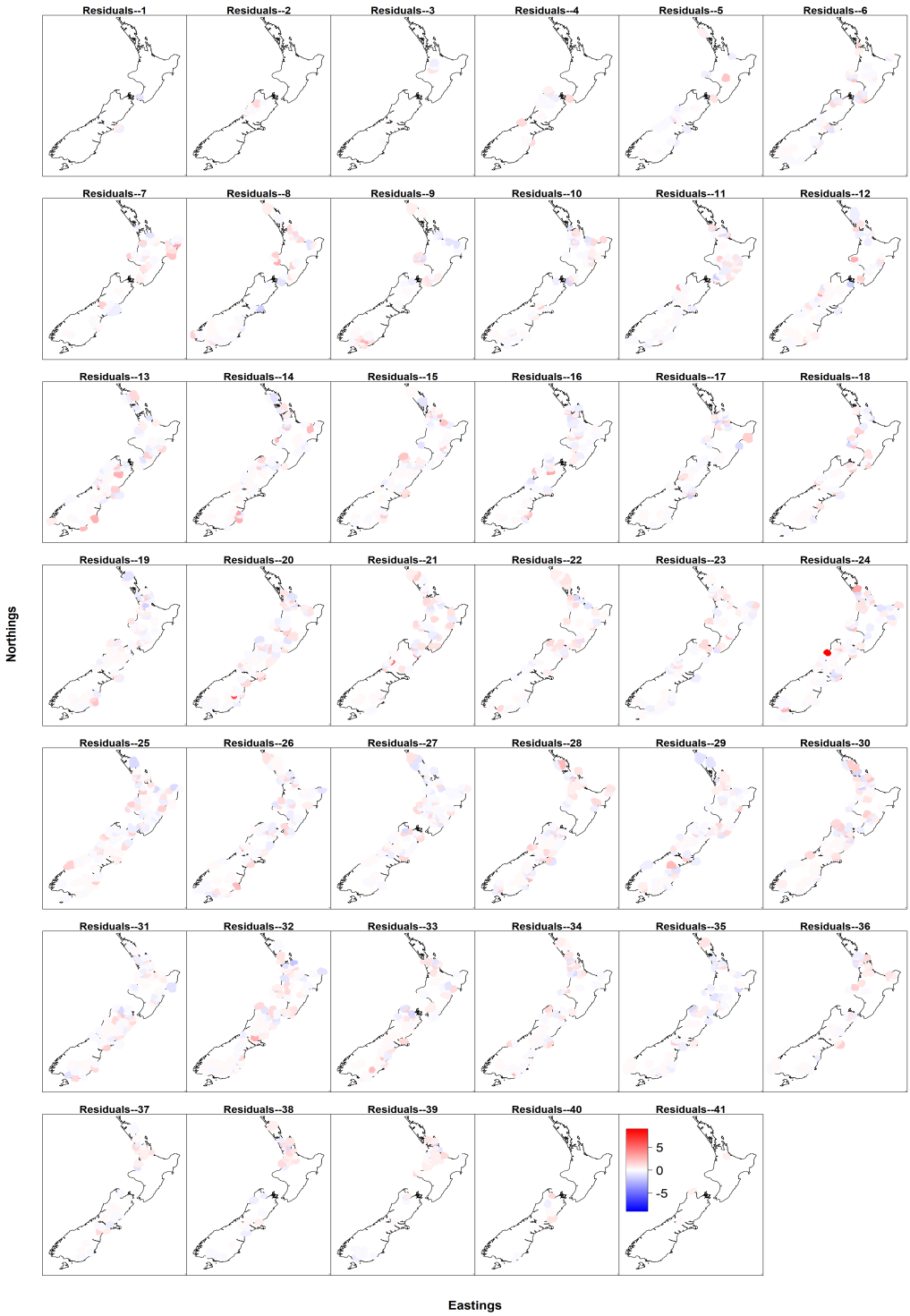


Figure 4.30: Heat maps of the shortfin eel VAST model's Pearson residuals.

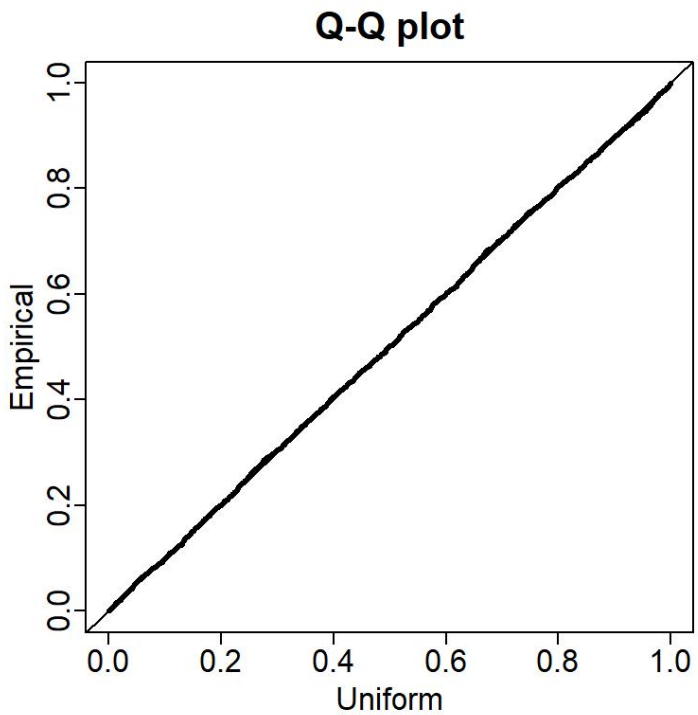


Figure 4.31: QQ-plot for the shortfin eel VAST model.

ity plot is shown in Figure 4.32. The observed encounter probability approximately falls within the 95% confidence interval for the predicted encounter probability when the predicted encounter probability is less than 0.7. However, observed encounter probabilities with a predicted encounter probability greater than 0.7 do not fall within this interval. In these cases the model is under predicting the true probability of encounter. However, these points fall just outside the 95% confidence interval range and therefore isn't a major cause for concern.

The majority of the observed encounter frequencies fall within the predicted encounter probability range so we can be relatively satisfied with the performance of the model.

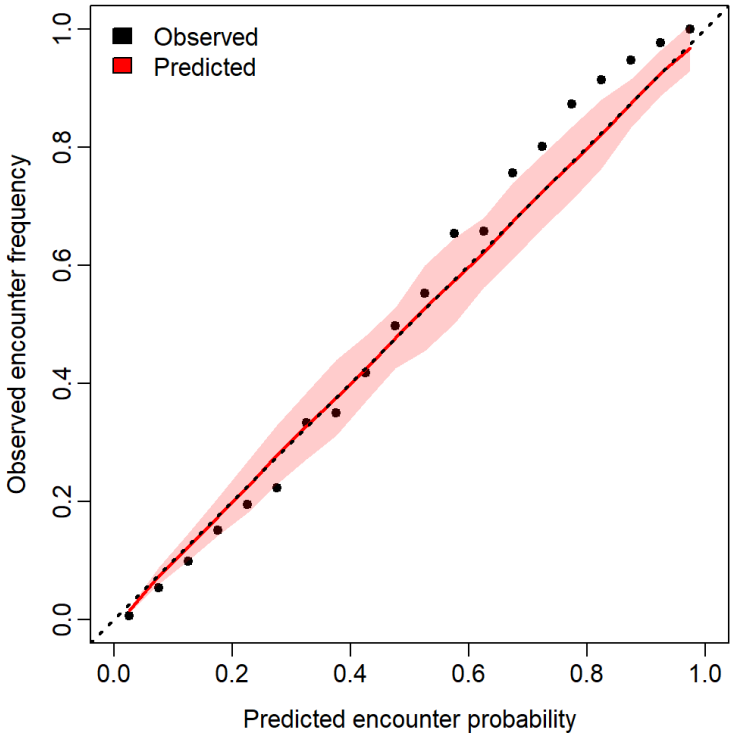


Figure 4.32: A diagnostic plot for observed encounter frequency against the predicted encounter probability for the shortfin eel VAST model.

Cross validation results

A spatial 50-fold cross validation was performed on the shortfin eel VAST model. The ROC curves for each of the 50 folds are shown in Figure 4.33. The mean AUC was 0.7864 (4dp) with 95% confidence interval 0.7754 and 0.7974 (standard error 0.0056). Figure 4.33 shows significant variability in the ROC curves with the smallest AUC being 0.5979 (4dp) and the largest being 0.9982 (4dp).

All curves return AUC values greater than 0.5 and therefore all do reasonably well. However, the lowest AUC nevertheless shows poor performance in model estimation. On average the shortfin eel VAST model

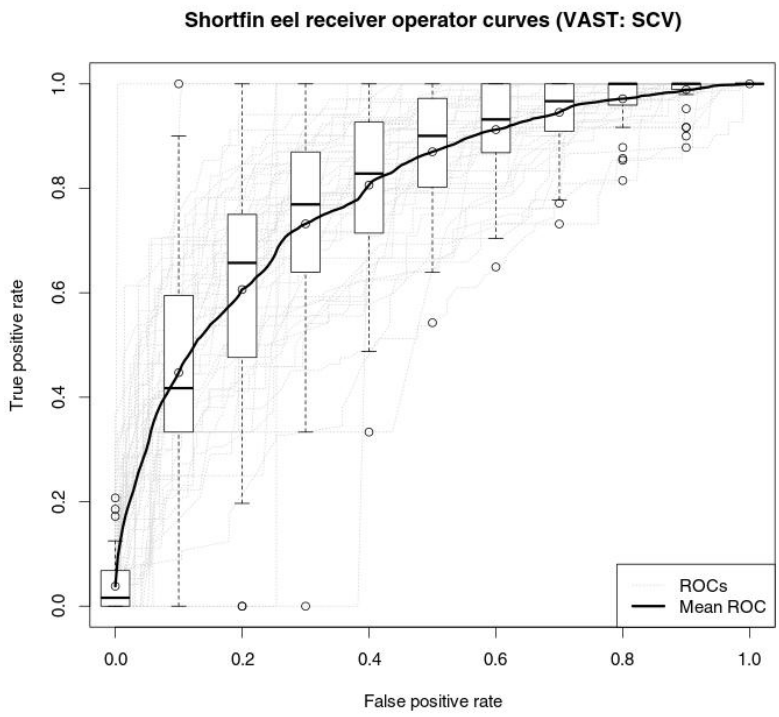


Figure 4.33: ROC curves under each of the 50 folds in the shortfin eel VAST model spatial cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

does well in predicting the probability of capture for areas which are spatially distinct to the spatial locations of the training data. In some cases the model predicts the probability of capture very well.

50-fold cross validation was performed on the shortfin eel VAST model which resulted in the ROC curves shown in Figure 4.34. The mean AUC (shown by the black line in Figure 4.34) was 0.9046 (4dp) with 95% confidence interval 0.8981 and 0.9111 (standard error 0.0033).

Figure 4.34 shows very little variability in AUC; the smallest AUC was 0.8415 (4dp) and the largest AUC was 0.9454 (4dp). Hence, the shortfin eel VAST model performs very well when predicting to areas that are not

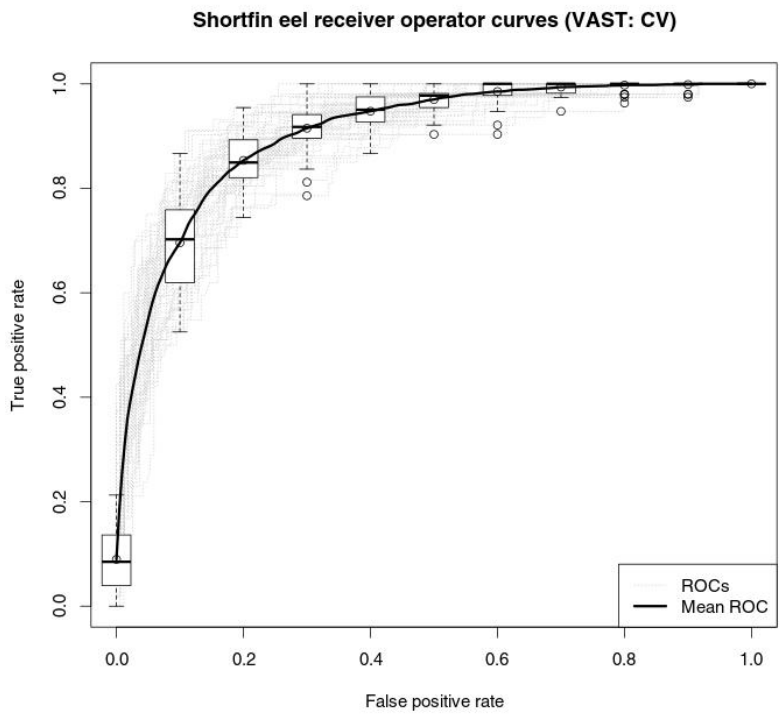


Figure 4.34: ROC curves under each of the 50 folds in the shortfin eel VAST model cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

spatially distinct to the training data set. This is shown by the large mean AUC value and the fairly small variability in AUC from this mean.

5-fold cross validation was implemented on the shortfin eel VAST model. The ROC curves for the cross validation and mean ROC curve are shown in Figure 4.35. The mean AUC is 0.9006 (4dp) with a 95% confidence interval of 0.8940 and 0.9073 (standard error of 0.0034). These results are very similar to the 50-fold cross validation results.

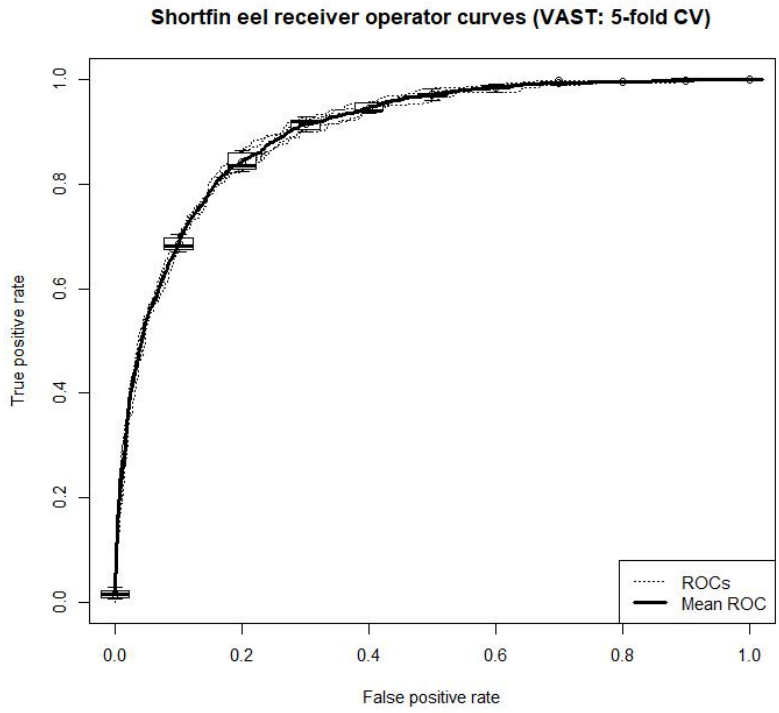


Figure 4.35: ROC curves under each of the 5 folds in the shortfin eel VAST model cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

4.2.3 Multi-species results

Model results

A multi-species probability of capture VAST model was run using the NZFFD longfin and shortfin eel presence/absence data. The model's estimated fixed effects are given in Table 4.3. The natural log of the values of the matrix **H** are -0.4162 ($\ln h_1$) and 0.0106 ($\ln h_2$) with standard errors 0.1555 and 0.1471 respectively. Both have high standard errors relative to the size of the effects. The fixed effect $\log \kappa_1$ was estimated as -3.4002 with standard error 0.1056. The term **H** and κ_1 define the Matérn function (given in Equation 3.14) and κ_1 defines decorrelation in the multi-species VAST model.

Parameter	Estimate	Standard error	C.V. (%)
$\ln h_1$	-0.4162	0.1555	37.4%
$\ln h_2$	0.0106	0.1471	1387.7%
$\log \kappa_1$	-3.4002	0.1056	3.1%
$L^{(1)}_{\Omega_1}$	0.9185	0.1067	11.6%
$L^{(2)}_{\Omega_1}$	0.1168	0.1550	132.7%
$L^{(3)}_{\Omega_1}$	0.8601	0.1250	14.5%
$L^{(1)}_{\epsilon_1}$	1.2420	0.0676	5.4%
$L^{(2)}_{\epsilon_1}$	0.1918	0.0948	49.4%
$L^{(3)}_{\epsilon_1}$	1.1156	0.0894	8.0%
$\log \sigma_{\beta_1}$	-1.6424	0.2990	18.2%

Table 4.3: The estimated fixed effects (4dp), associated standard errors (4dp) and coefficient of variation (C.V.) (1dp) of the multi-species VAST model (to 4dp).

Table 4.3 gives the coefficient of variation (C.V.) for each of the model parameters. The C.V. estimates for $\ln h_1$, $\ln h_2$, $L^{(2)}_{\Omega_1}$ and $L^{(2)}_{\epsilon_1}$ are all very large and therefore indicate large imprecision in these parameters. In par-

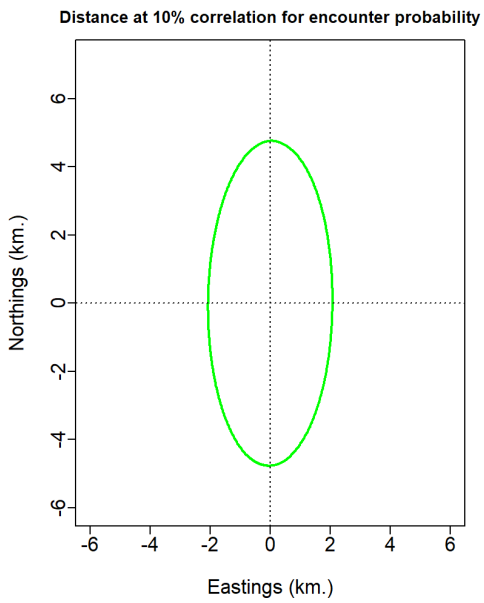


Figure 4.36: Encounter probability geometric anisotropy for the multi-species VAST model on a easting-northing coordinate grid.

ticular $\ln h_2$ results in a C.V. of 1387.7% which indicates that we have very little certainty in anisotropy correlation. Extreme caution should be taken in reviewing predictions made by the multi-species VAST model as it has much greater uncertainty in many of its parameters in comparison to the single species VAST models. However, the rest of the parameters C.V. estimates shown in Table 4.3 are less than 20%. This indicates that these estimates have reasonable precision.

Figure 4.36 shows the distance at which correlation is at 10% of the original correlation. This plot shows that decorrelation occurs slower in a northwards and southwards direction and that 10% decorrelation occurs at just over 4km in these directions.

There are three fixed effects of the loadings matrix $\mathbf{L}_{\mathbf{\Omega_1}}$, these are estimated as 0.9185 ($L^{(1)}_{\Omega_1}$), 0.1168 ($L^{(2)}_{\Omega_1}$) and 0.8601 ($L^{(3)}_{\Omega_1}$) with standard errors 0.1067, 0.1550 and 0.1250 respectively. This generates spatial variation within each eel species and covariation amongst the longfin and shortfin

eel.

There are three fixed effects of the loadings matrix $\mathbf{L}_{\epsilon_1}$, these are estimated as 1.2420 ($L_{\epsilon_1}^{(1)}$), 0.1918 ($L_{\epsilon_1}^{(2)}$) and 1.1156 ($L_{\epsilon_1}^{(3)}$) with standard errors 0.0676, 0.0948 and 0.0894 respectively. This generates spatial-temporal variation within each eel species and covariation amongst the longfin and shortfin eel.

The intercept term ($\beta_1(c_i, t_i)$) of Equation 3.10 was defined as a random effect which follows a Normal distribution with a random walk variance structure. The variance term $\log \sigma_{\beta_1}$ is estimated as -1.6424 with standard error 0.2990. Hence, σ_{β_1} is estimated as 0.0228 (4dp). This indicates that the $\beta_1(c_i, t_i)$ will change very little with time due to its very small estimated variability. We can expect the 'baseline' probability of capture effect to change very little from one year to the next.

Probability of capture estimates were made across New Zealand from 1974 to 2014 using the multi-species VAST model. Figure B.1 of the appendix shows the longfin eel estimates from 1974 to 2014 and Figure B.2 of the appendix shows the longfin eel estimates in 2014. Figure B.3 of the appendix shows the shortfin eel estimates from 1974 to 2014 and Figure B.4 of the appendix shows the shortfin eel estimates in 2014. The estimates under the multi-species model for both species shows subtle differences between the probability of capture estimates made by the longfin eel VAST model and the shortfin eel VAST model.

The following describes the diagnostic plots used to assess the multi-species VAST model. Figure 4.37 shows the multi-species VAST model Pearson residuals by knot. The maps of Figure 4.37 appear to show an even amount of blue and red colouring and do not have any brightly coloured points (large residuals). Hence, we can be satisfied that the multi-species model does not have any particularly unusual residuals or unusual spread of residuals.

Figures 4.38a and 4.38b are QQ plots for the longfin eel results and shortfin eel results based on the multi-species VAST model. Residuals for

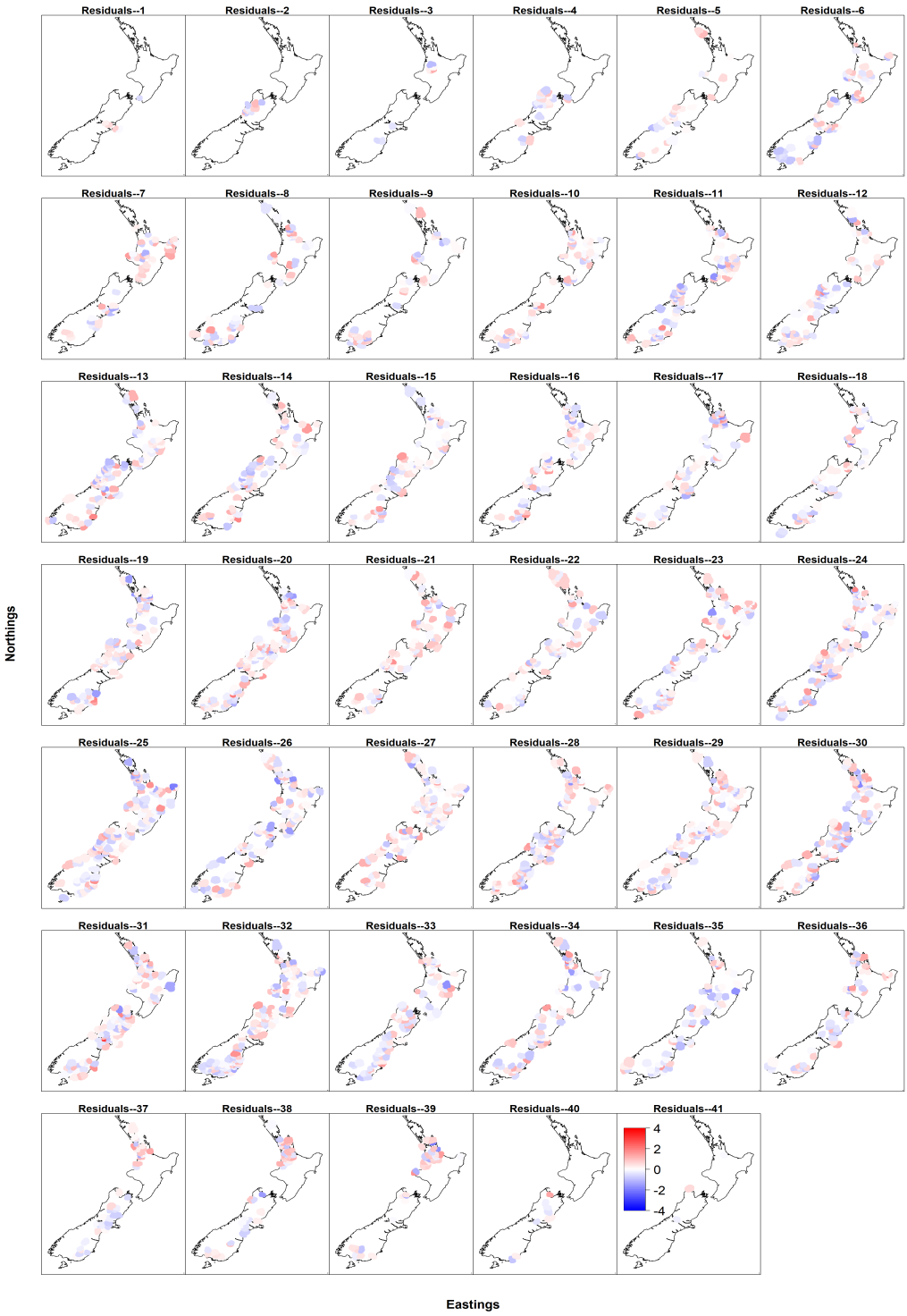
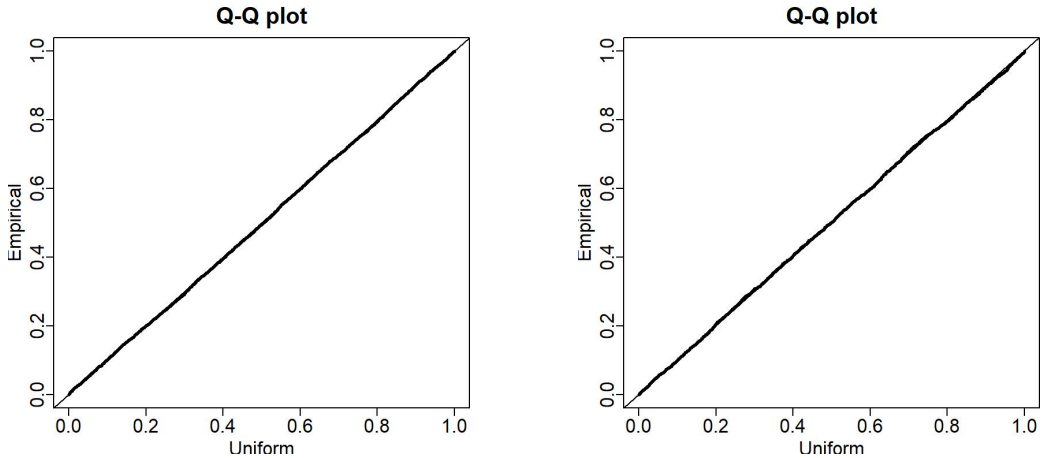


Figure 4.37: Heat maps of the multi-species VAST model's Pearson residuals.

both species follow the QQ line very well. Therefore we can conclude that the longfin eel and shortfin eel Pearson's residuals follow a Normal distribution.



(a) QQ-plot of the longfin eel results.

(b) QQ-plot of the shortfin eel results.

Figure 4.38: QQ-plots for the multi-species VAST model.

An observed encounter frequency vs. predicted encounter probability plot is given in Figure 4.39. The plot shows that the model tends to underestimate encounter probability when the observed encounter frequency is greater than c.0.7, and the model tends to overestimate encounter probability when the observed encounter frequency is less than c.0.3. Similar biases are seen in the single species models. Once again, one must take this into consideration when using the multi-species VAST model.

The estimated spatial and spatio-temporal correlation is displayed in Figure 4.40 for the multi-species VAST model. The estimated spatial correlation between the longfin eels and the shortfin eels is 0.1. The estimated spatio-temporal correlation between the longfin eels and shortfin eels is 0.2. The spatial and spatio-temporal correlation between the longfin eels and shortfin eels are both very small positive correlations. This indicates that the probability of catching a longfin eel slightly increases when a

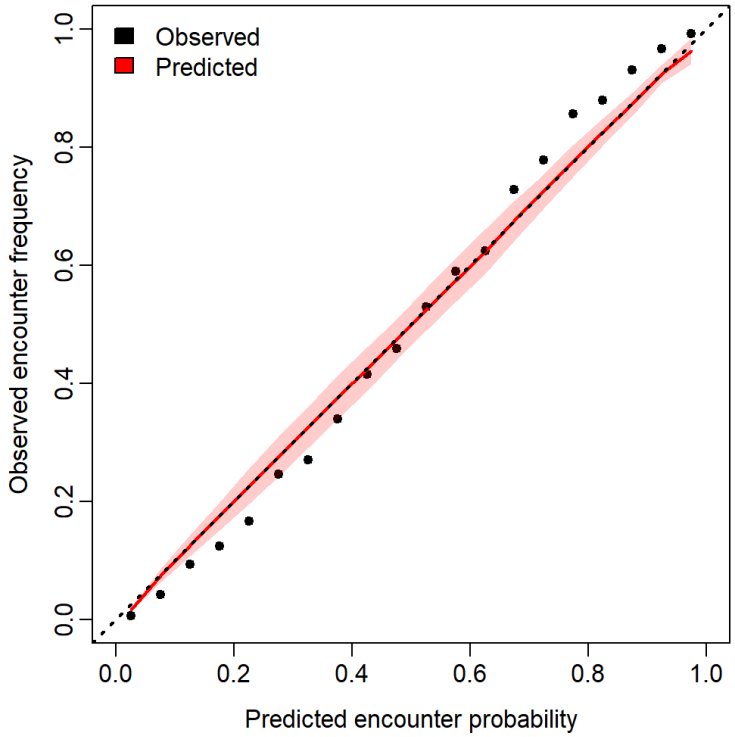


Figure 4.39: A diagnostic plot for observed encounter frequency against the predicted encounter probability for the multi-species VAST model.

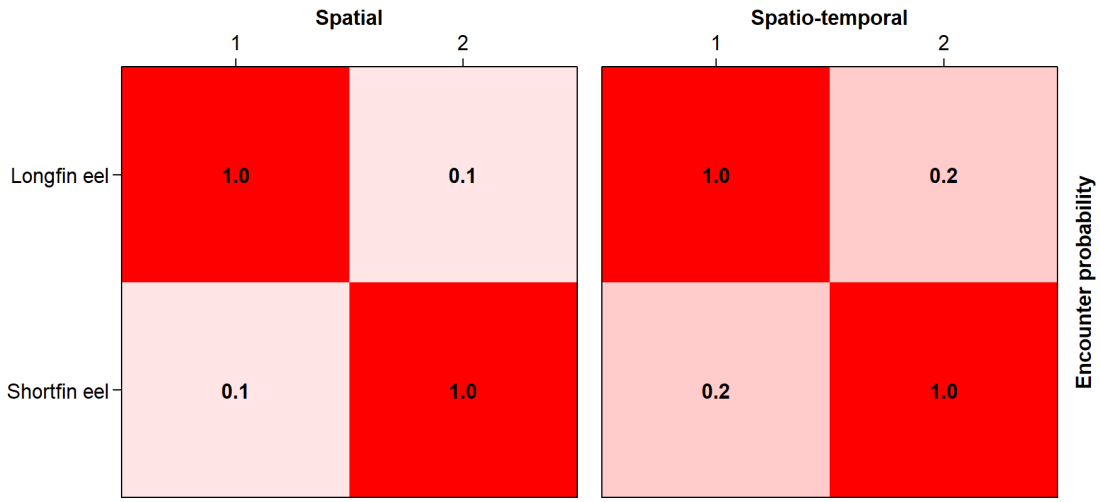


Figure 4.40: The estimated spatial and spatio-temporal correlation for longfin eels and shortfin eels from the multi-species VAST model. The columns numbered 1 corresponds to the longfin eel and the columns numbered 2 correspond to the shortfin eel.

shortfin eel is present (and vice-versa). But this increase is likely to be minimal.

Cross validation results

Spatial cross validation, 50-fold cross validation and 5-fold cross validation was performed on the VAST multi-species model. AUC was calculated using 1) all the probability of capture estimates, 2) only the longfin eel estimates, and 3) only the shortfin eel estimates. Examining the results separately allows us to see how the model performs for each of the species individually and allows us to compare these results to the single species results.

The ROC curves for the 50 folds of the spatial cross validation are shown in Figures 4.41, 4.42 and 4.43. The figures give the spatial cross validation ROC curves for the multi-species VAST model. Figure 4.41 gives the curves using both the longfin eel and shortfin eel probability of capture results. Figure 4.42 gives the curves using only the longfin eel probability of capture results. Figure 4.43 gives the curves using only the shortfin eel probability of capture results.

Figures 4.41, 4.42 and 4.43 show ROC curves which are highly variable. In particular, Figure 4.41 contains an outlier ROC curve with a corresponding AUC of 0.4738. This means that the probability of capture for one of the folds (which is spatially distinct to the training data) is being estimated very poorly. As the AUC is less than 0.5, probability of capture is being estimated in the wrong direction, i.e. high probabilities of capture assigned to sites where eels have not been found.

The highly variable nature of the ROC curves from the spatial cross validation is consistent with other spatial cross validation results. Hence, the multi-species model performs very well in some folds and very poorly in other folds.

The ROC curves for the 50 folds of the cross validation are shown in Figures 4.44, 4.45 and 4.46. Figure 4.44 shows the ROC curves for the multi-species model results with both the species combined whereas Figure 4.45 and 4.46 show the ROC curves with the results only of the longfin eel and shortfin eel respectively.

The mean AUC for the multi-species model of both species is 0.8890 (4dp) with a 95% confidence interval of 0.8846 and 0.8933 (standard error 0.0022). As shown in the boxplots, the variability in Figure 4.44 is very small. These results indicate that areas which are not spatially distinct to the training data can predict either longfin or shortfin eel presence/absence very well when using a multi-species model.

The mean AUC for the multi-species model of longfin eels is 0.8310 (4dp) with a 95% confidence interval of 0.8232 and 0.8387 (standard error

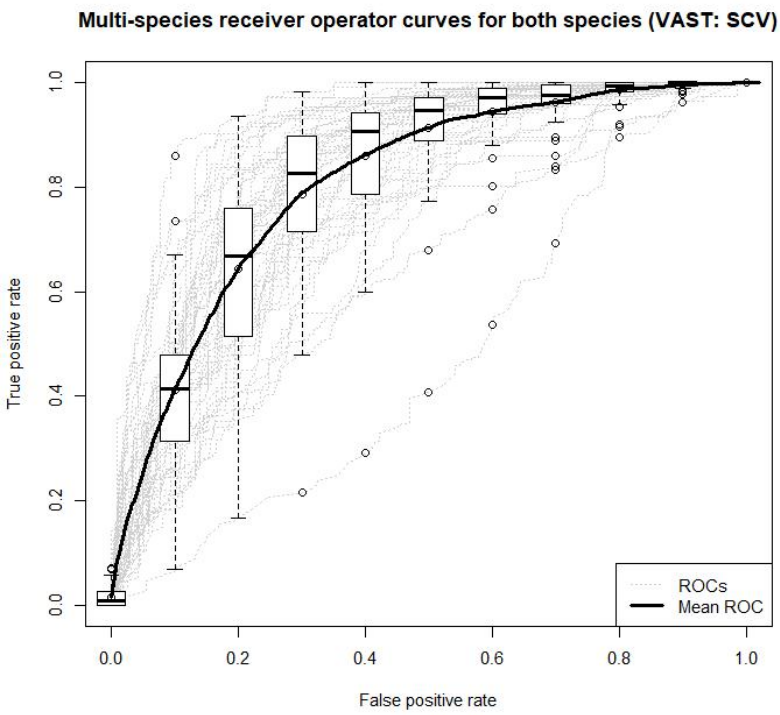


Figure 4.41: ROC curves under each of the 50 folds in the multi-species VAST model spatial cross validation. The ROC curves are built under the results for both the longfin eel and shortfin eel. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

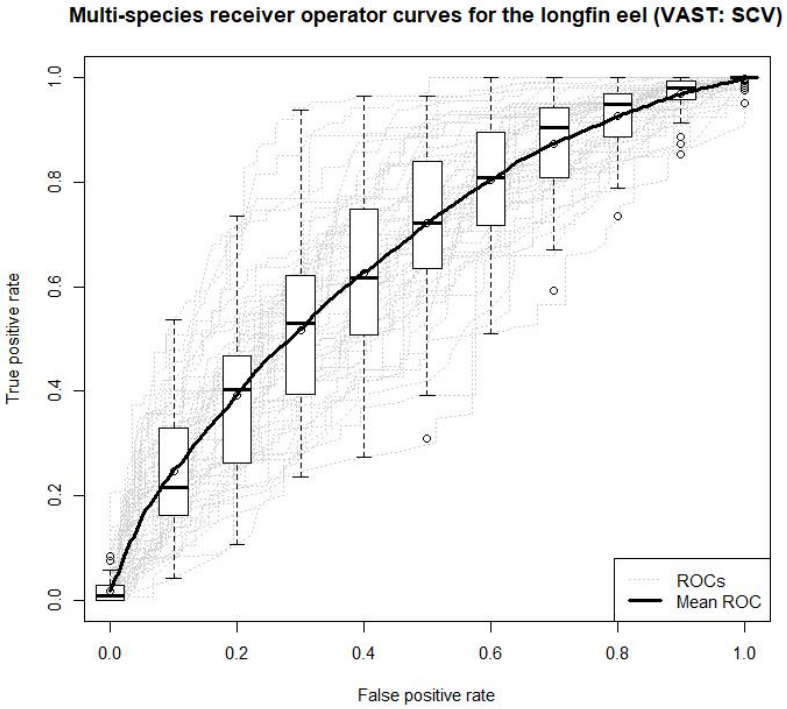


Figure 4.42: ROC curves under each of the 50 folds in the multi-species VAST model spatial cross validation. The ROC curves are built under the results for the longfin eel only. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

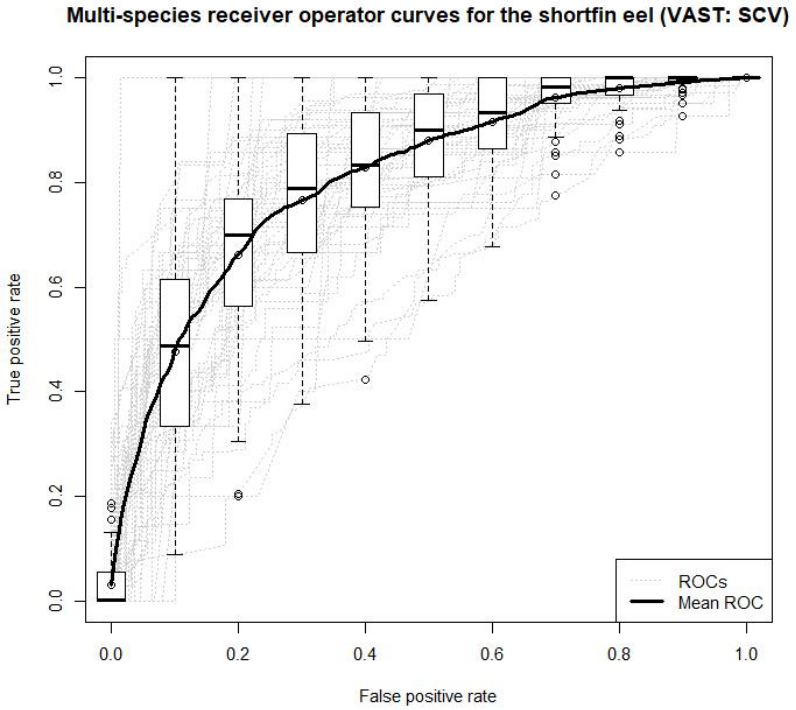


Figure 4.43: ROC curves under each of the 50 folds in the multi-species VAST model spatial cross validation. The ROC curves are built under the results for the shortfin eel only. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

of 0.0040). The boxplots of Figure 4.45 show the variability in the ROC curves to be small. However, these curves have more variability than the curves of the multi-species model with both species results and approximately the same as the curves of the multi-species model with shortfin eel results only. The multi-species model performs very well in predicting presence/absence for longfin eels.

The mean AUC for the multi-species model of the shortfin eels is 0.9025 (4dp) with a 95% confidence interval of 0.8960 and 0.9091 (standard error of 0.0034). The boxplots of Figure 4.46 show similar variability in ROC curves as the ROC curves of the multi-species model results for the longfin

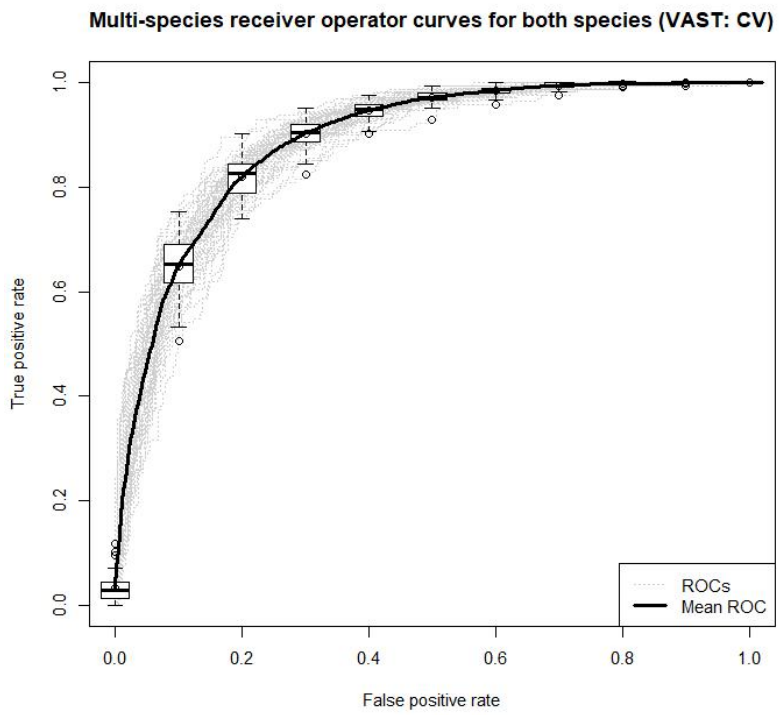


Figure 4.44: ROC curves under each of the 50 folds in the multi-species VAST model cross validation. The ROC curves are built under the results for both the longfin eel and shortfin eel. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

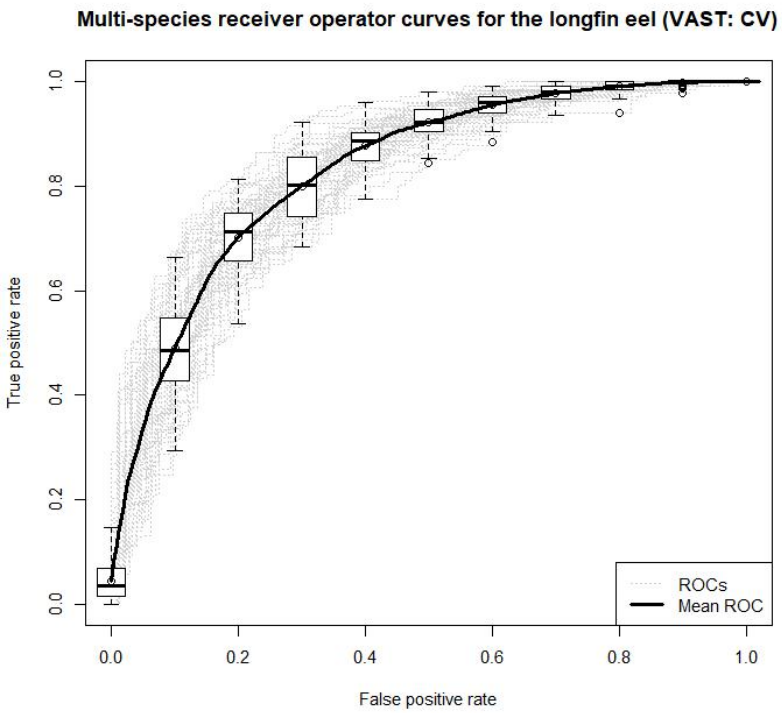


Figure 4.45: ROC curves under each of the 50 folds in the multi-species VAST model cross validation. The ROC curves are built under the results for the longfin eel only. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

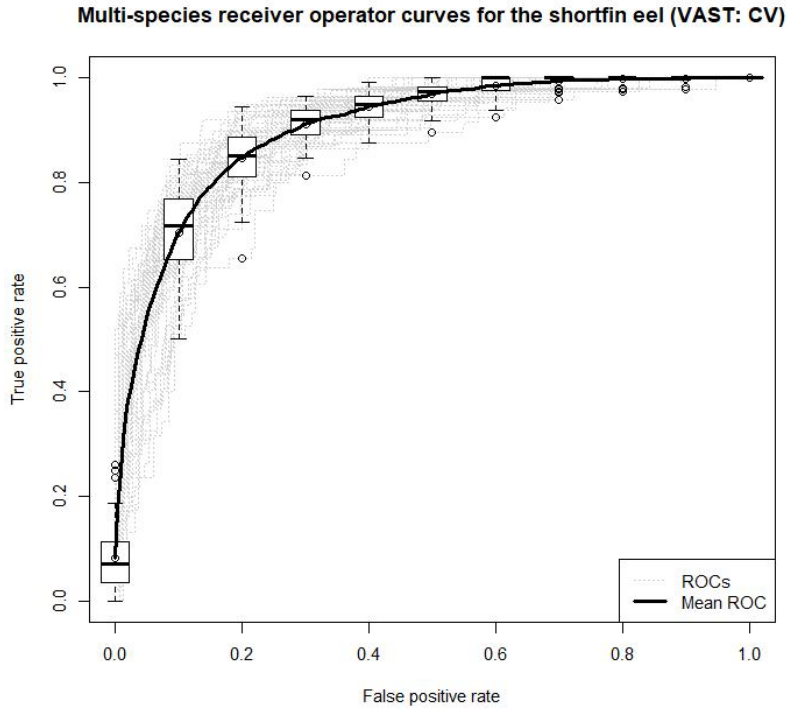


Figure 4.46: ROC curves under each of the 50 folds in the multi-species VAST model cross validation. The ROC curves are built under the results for the shortfin eel only. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

eels (Figure 4.45). This variability is fairly small but is reduced when using the results of the species combined (Figure 4.44). This is shown by the standard errors which is much smaller when using both species results. The mean AUC and 95% confidence interval indicates that the multi-species model performs well in predicting shortfin eel presence/absence.

5-fold cross validation was performed on the multi-species VAST model. This was implemented to compare directly against the GRaF longfin and shortfin eel models. Therefore we are interested in the AUC results based on only the longfin eel probability of capture estimates and based on only the shortfin eel probability of capture estimates. The ROC curves based on

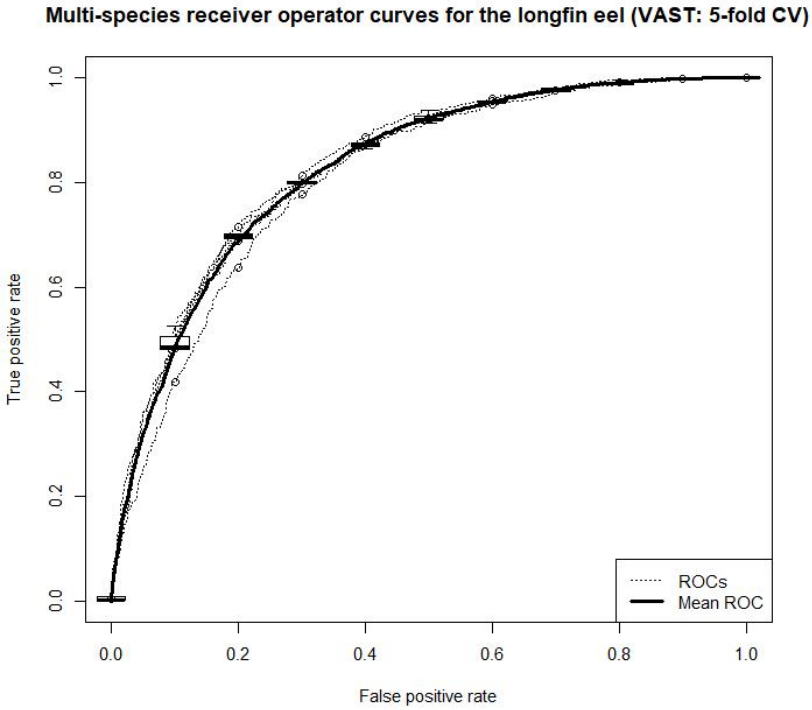


Figure 4.47: ROC curves under each of the 5 folds in the multi-species VAST model cross validation. The ROC curves are built under the results for the longfin eel only. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

the longfin eel estimates and the shortfin eel estimates are given in Figures 4.47 and 4.48 respectively. The mean AUCs and 95% confidence intervals are almost identical to that of the 50-fold cross validation using the longfin eel results only and the shortfin eel results only.

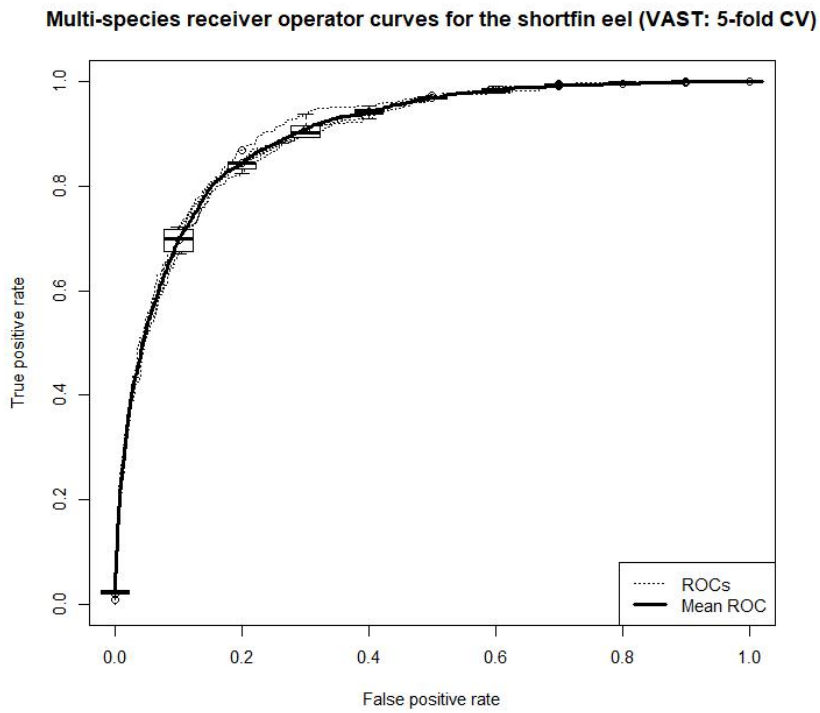


Figure 4.48: ROC curves under each of the 5 folds in the multi-species VAST model cross validation. The ROC curves are built under the results for the shortfin eel only. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

4.3 GRaF modelling results

This section outlines the findings of the longfin eel GRaF model and the shortfin eel GRaF model. See Chapter 2 for details on the data used and Section 3.3.1 for details on the GRaF methodology.

The probability of capture models were built using an uninformative prior on the probability of capture for each model covariate. However, an informative prior was placed on the model length scales to allow for a smooth function in how probability of capture changes with a covariate.

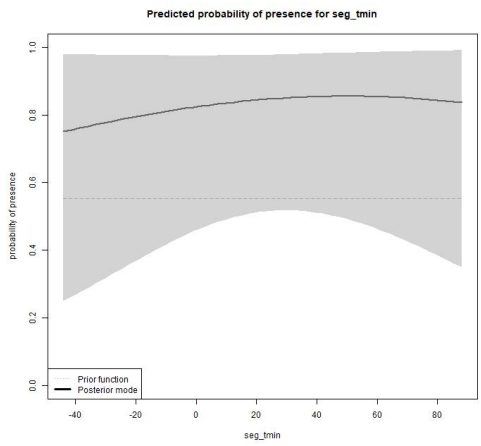
4.3.1 Longfin eel results

Model results

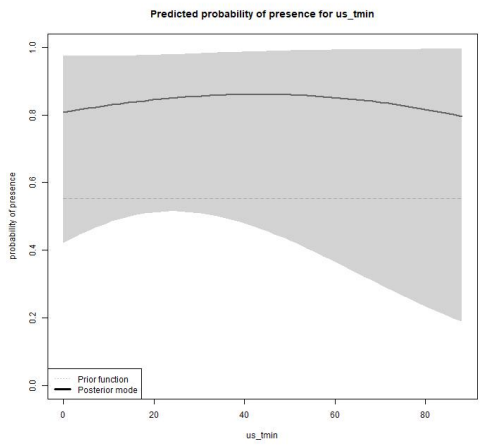
A total of 70 covariates were used for the longfin eel GRaF model. Of these, 69 were determined by the longfin eel RRF model (see Table A.2) and an additional covariate was used to account for temporal variability in the data set (known as 'year'). Table C.1 of the appendix gives the estimated lengthscales ordered from smallest to largest. Year was estimated to have the most complex function.

The plots of the posterior distribution of seg_tmin, us_tmin, seg_twar and year are given in Figures 4.49a, 4.49b, 4.49c and 4.49d. The variables seg_tmin, us_tmin and seg_twar achieved the highest importance scores in the longfin eel RRF model (see Figure 4.3). The function for seg_twar has the largest lengthscale (50.64) of the figures and as a result achieves a posterior distribution which is very flat (Figure 4.49a).

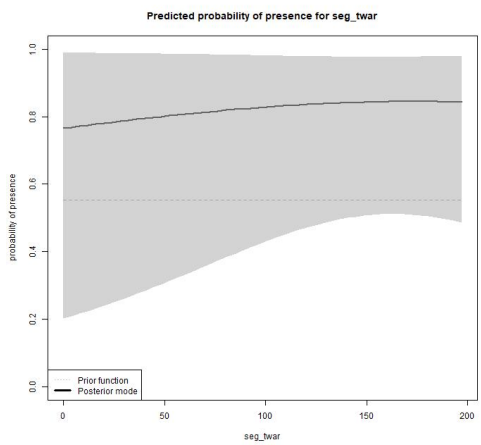
The posterior modes for seg_tmin and us_tmin are also quite flat due to their fairly large lengthscales (13.84 and 6.39 respectively). The 95% credible intervals for seg_tmin, us_tmin and seg_twar are very wide. The modes and credible intervals for each of these variables peak at a certain value of the covariate and tend to be indicative of high probability of presence across all values of the covariate. However, the 95% credible intervals



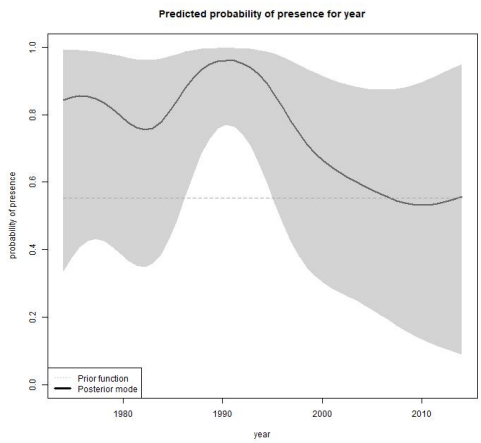
(a) Probability of presence vs. seg_tmin (mean minimum wintertime air temperature for a river segment (deg. C $\times$ 10)) for the longfin eel GRaF model. Lengthscale was estimated as 13.83.



(b) Probability of presence vs. us_tmin (mean minimum wintertime air temperature upstream of a river segment (deg. C $\times$ 10)) for the longfin eel GRaF model. Lengthscale was estimated as 6.39.



(c) Probability of presence vs. seg_twar (mean January air temperature for a river segment ((deg. C $\times$ 10))) for the longfin eel GRaF model. Lengthscale was estimated as 50.64.



(d) Probability of presence vs. year for the longfin eel GRaF model. Lengthscale was estimated as 0.35.

Figure 4.49: Plots of probability of presence (i.e. capture) vs. covariate for seg_tmin, us_tmin, seg_twar and year. Dotted lines show the mean (δ), solid lines show the predicted probability of capture (posterior mode) and shaded areas show the 95% credible interval.

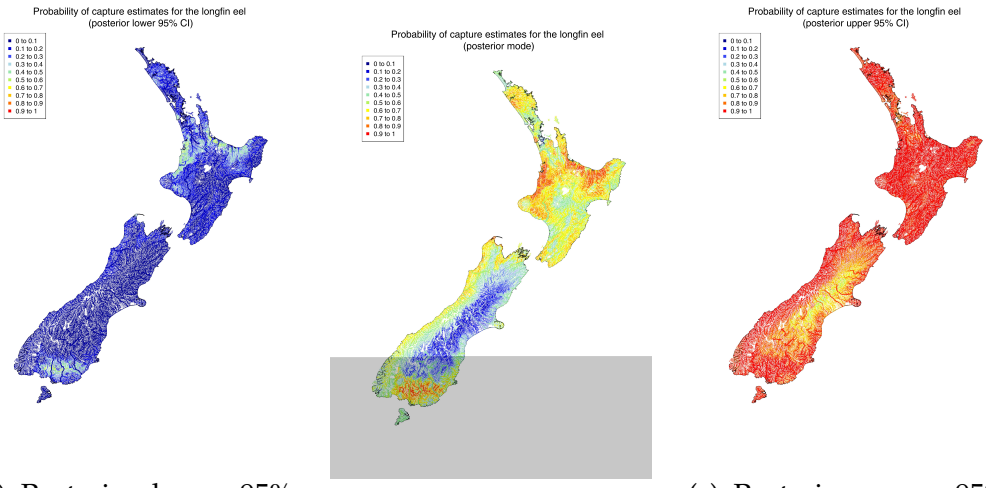
for Figures 4.49a, 4.49b and 4.49c tends to include probability of presence values less than 0.5 across all values of the covariate. This shows that the model wouldn't predict probability of capture with strong certainty.

Figure 4.49d shows the posterior mode and 95% credible interval for the probability of presence against year. The function has a very small length scale of 0.35 and therefore changes probability of presence rapidly in comparison to functions with a large lengthscale. The posterior mode has high probability of presence in the first 10-15 years and then rapidly increases in the late 1980s and peaks in the early 1990s. Probability of presence then drops rapidly. The 95% credible intervals show that there is little certainty in the estimates over the first 10-15 years and the last c.15 years. However, from c.1987 to c.1995 the high probability of presence is estimated with small 95% credible intervals.

The probability of capture estimates (posterior mode) made at each segment of river in the REC database is shown in Figure 4.50b. Figures 4.50a and 4.50c show the posterior 95% lower confidence interval and the posterior 95% upper confidence interval respectively. Figure 4.51 repeats Figure 4.50b on a full page to ease probability of capture comparisons.

From Figure 4.50b one can see that the North Island of New Zealand has very high probability of capture estimates (in red) of 0.8 to 1 in larger coastal rivers of the Waikato region. The Waikato shows lower observed proportions of longfin eel capture, as shown in Figure 4.1. The central east and west coast, and parts of the northern North Island have high probabilities of capture estimates (yellow/red) of 0.6 to 0.9. This is approximately consistent with the observed proportions of longfin eel capture in Figure 4.1. The rest of the North Island tends to show probabilities of capture around 0.4 to 0.7 (green/yellow). This is also consistent with the observed proportions of capture in Figure 4.1.

From Figure 4.50b one can see that longfin eels are unlikely to be found throughout the centre of the South Island (probability of capture of 0.1 to 0.3). This is seen in the observed proportions of capture for longfin eels



(a) Posterior lower 95% confidence interval probability of capture estimates for the longfin eel.

(b) Posterior mode probability of capture estimates for the longfin eel.

(c) Posterior upper 95% confidence interval probability of capture estimates for the longfin eel.

Figure 4.50: Probability of capture estimates using the REC2 database. These estimates were made using the longfin eel GRaF model. Larger points have a larger stream order.

(Figure 4.1). High probabilities of capture can be seen in larger rivers at the south of the South Island (0.7 to 0.9) and the west coast of the South Island (0.6 to 0.8). Parts of the west coast of the observed proportions of longfin eel capture are consistent with this result but the south of the South Island tends to have low observed proportions of longfin eel capture (c.0.3 to c.0.6). The remaining parts of the Island and Stewart Island tend to have probabilities of capture for the longfin eel around 0.4 to 0.6. Figure 4.1 shows observed proportions to be lower than this in Stewart Island and more variable throughout the South Island.

Figures 4.50a and 4.50c confirm the large uncertainty in probability of capture estimates from the longfin eel GRaF model. The lower 95% confidence interval (Figure 4.50a) shows very low probabilities of capture (0 to 0.3) throughout New Zealand. The bottom of the South Island and

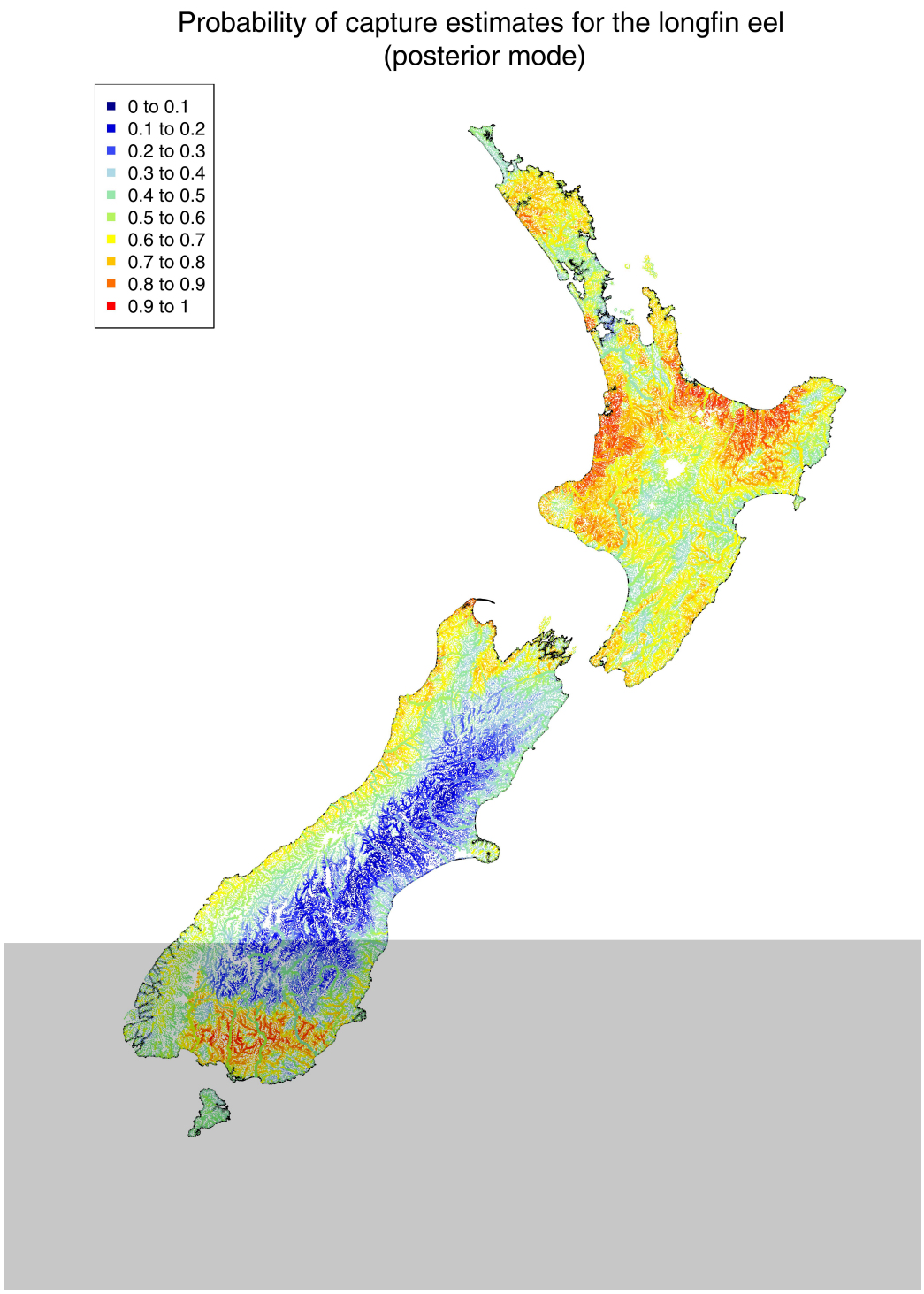


Figure 4.51: Posterior mode probability of capture estimates for the longfin eel using the REC2 database. These estimates were made using the longfin eel GRaF model. Larger points have a larger stream order. This is repeated from Figure 4.50b.

Waikato's east and west coast show slightly higher probabilities of 0.3 to 0.6.

The upper 95% confidence interval (Figure 4.50c) shows very high probabilities of capture (0.8 to 1) throughout New Zealand. The central South Island and some large streams in the the North Island's Waikato region have slightly lower probabilities of capture. The central South Island has rivers with upper 95% credible interval probabilities of 0.5 to 0.7. Whereas the Waikato region has a small number of rivers with upper 95% credible interval probabilities of 0.6 to 0.7.

Cross validation results

5-fold cross validation was performed on the longfin eel GRaF model. The ROC curves for this validation are shown in Figure 4.52. The cross validation found a mean AUC of 0.8272 (4dp) with a 95% confidence interval of 0.8194 and 0.8350 (standard error of 0.0040). Since the mean AUC is quite large and the 95% confidence interval is small, then we can conclude that the model is performing very well in predicting longfin eel probability of capture with strong certainty (small standard error and confidence interval).

There is very little variability in the ROC curves of Figure 4.52. How-

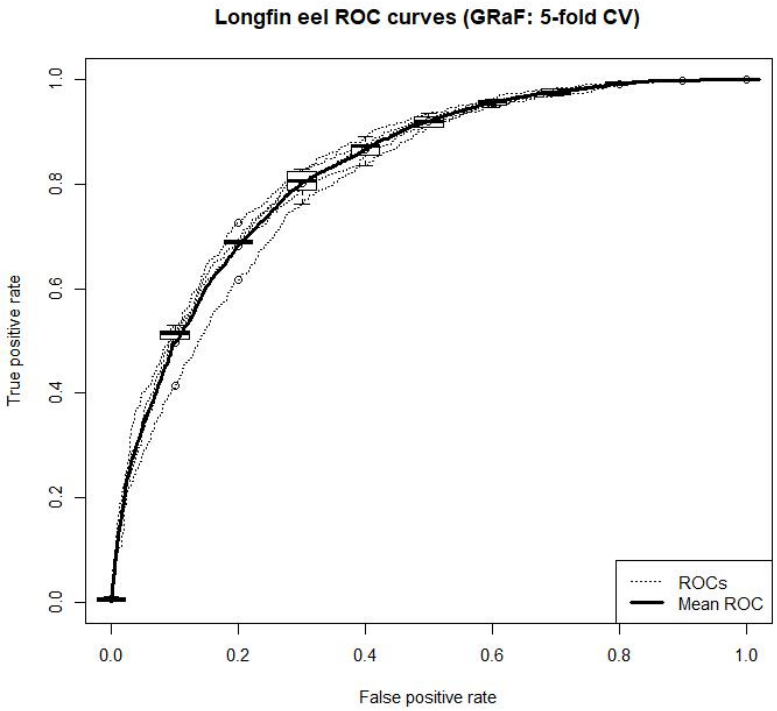


Figure 4.52: ROC curves under each of the 5 folds in the longfin eel GRaF model cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

ever, one of the folds does show smaller true positive rates at false positive rates of 0.1 and 0.2. This fold resulted in an AUC of 0.8025 (4dp) which is still quite large. Hence, probability of capture is being predicted fairly well in this fold, regardless of the outliers. Overall, the ROC curves in Figure 4.52 show very little variability and high certainty in the results. Therefore, we can conclude that the longfin eel GRaF model predicts probability of capture very well in spatial locations which are spatially dependent to the training data.

4.3.2 Shortfin eel results

Model results

The shortfin eel GRaF model used 56 covariates. 55 of these were the covariates selected by the shortfin eel RRF model (see Table A.2). The final covariate is year which enables temporal variability to be accounted for. Table C.2 of the appendix gives the estimated lengthscales for each of the covariates in the model.

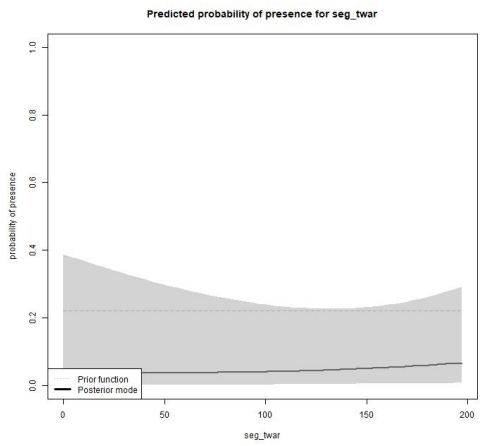
Plots of the posterior distribution of the covariates scoring the three highest importance scores in the shortfin eel RRF model (see Figure 4.8) are given in Figures 4.53a, 4.53b and 4.53c. These covariates are seg_twar, seg_elev and segshade. Additionally, Figure 4.53d gives a plot of the posterior distribution of the year covariate.

The posterior distributions for seg_twar and seg_elev (Figures 4.53a and 4.53b) show very low probabilities of capture regardless of covariate value (flat distributions). The probability of capture for seg_twar slightly increases as seg_twar increases and the probability of capture for seg_elev is slightly curved (largest at low segment elevation). Figure 4.53c shows that the posterior distribution for segshade has a decreasing trend and that shortfins are most likely to be found at low segment shade.

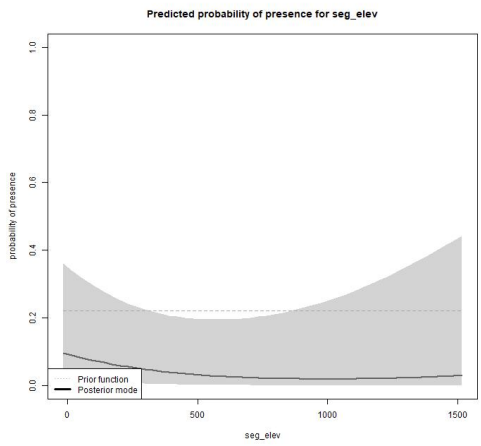
The 95% credible intervals for seg_twar, seg_elev and segshade all contain probabilities that do not exceed c.0.5. This indicates that shortfin eel capture has a 95% chance of being less than c.0.5 across all values of these covariates.

The posterior distribution for year (Figure 4.53d) is quite variable. This is determined by the very low lengthscale of 0.27. Probability of capture remains very low across time but peaks in 2014. The 95% credible intervals are fairly wide but we obtain fairly accurate (small credible intervals) in the late 1980s and late 1990s.

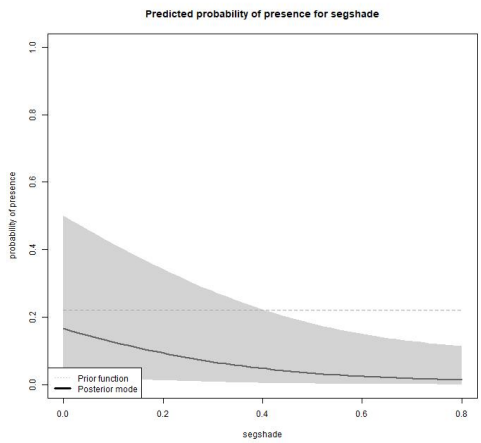
Figure 4.54b shows the posterior mode probability of capture estimates for the shortfin eel. Figures 4.54a and 4.54c show the lower and upper 95%



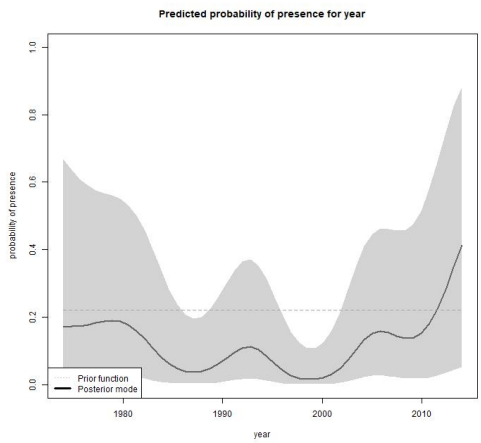
(a) Probability of presence vs. seg_twar (mean January air temperature for a river segment ((deg. C × 10))) for the shortfin eel GRaF model. Lengthscale was estimated as 62.62.



(b) Probability of presence vs. seg_elev (elevation of a river segment above sea level (m)) for the shortfin eel GRaF model. Lengthscale was estimated as 17.47.

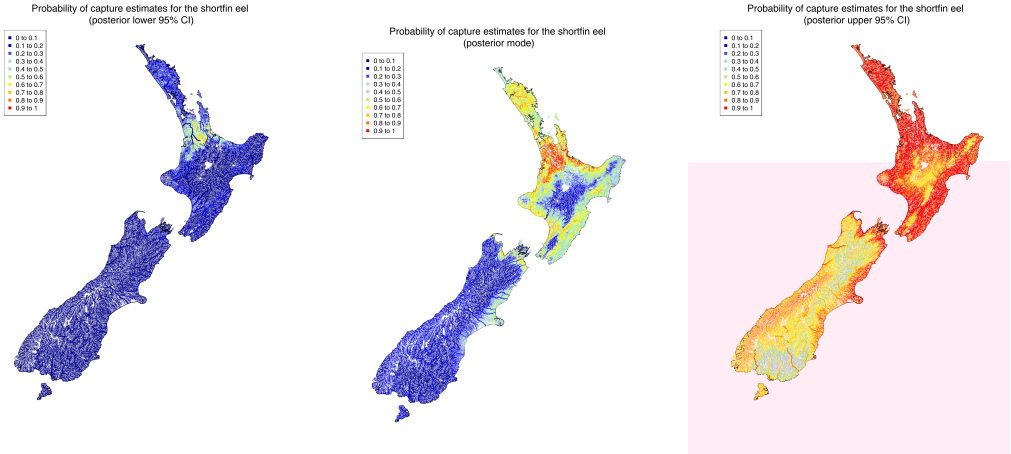


(c) Probability of presence vs. segshade (proportion of riparian shade area in a segment of river (%)) for the shortfin eel GRaF model. Lengthscale was estimated as 13.59.



(d) Probability of presence vs. year for the shortfin eel GRaF model. Lengthscale was estimated as 0.27.

Figure 4.53: Plots of probability of presence (i.e. capture) vs. covariate for seg_twar, seg_elev, segshade and year. Dotted lines show the mean (δ), solid lines show the predicted probability of capture (posterior mode) and shaded areas show the 95% credible interval.



(a) Posterior lower 95% confidence interval probability of capture estimates for the shortfin eel.

(b) Posterior mode probability of capture estimates for the shortfin eel.

(c) Posterior upper 95% confidence interval probability of capture estimates for the shortfin eel.

Figure 4.54: Probability of capture estimates using the REC2 database. These estimates were made using the shortfin eel GRaF model. Larger points have a larger stream order.

credible intervals for the probability of capture of shortfin eels. Figure 4.55 repeats Figure 4.54b on a full page to ease probability of capture comparisons.

We can see from Figure 4.54b that the North Island of New Zealand has high probabilities of capture (0.7 to 1) around the Waikato and Auckland region. This is particularly true for large rivers in these regions. The observed proportion of shortfin eel capture (Figure 4.2) is less than 0.8 for these same regions. The central North Island and small parts of the central east North Island, central west North Island and south of the North Island show very low probabilities of capture (0 to 0.3). This is approximately consistent with the observed proportions of shortfin eel capture in Figure 4.2. The rest of the North Island tend to show probabilities around 0.4 to 0.7. The observed proportions for the rest of the North Island tend to be

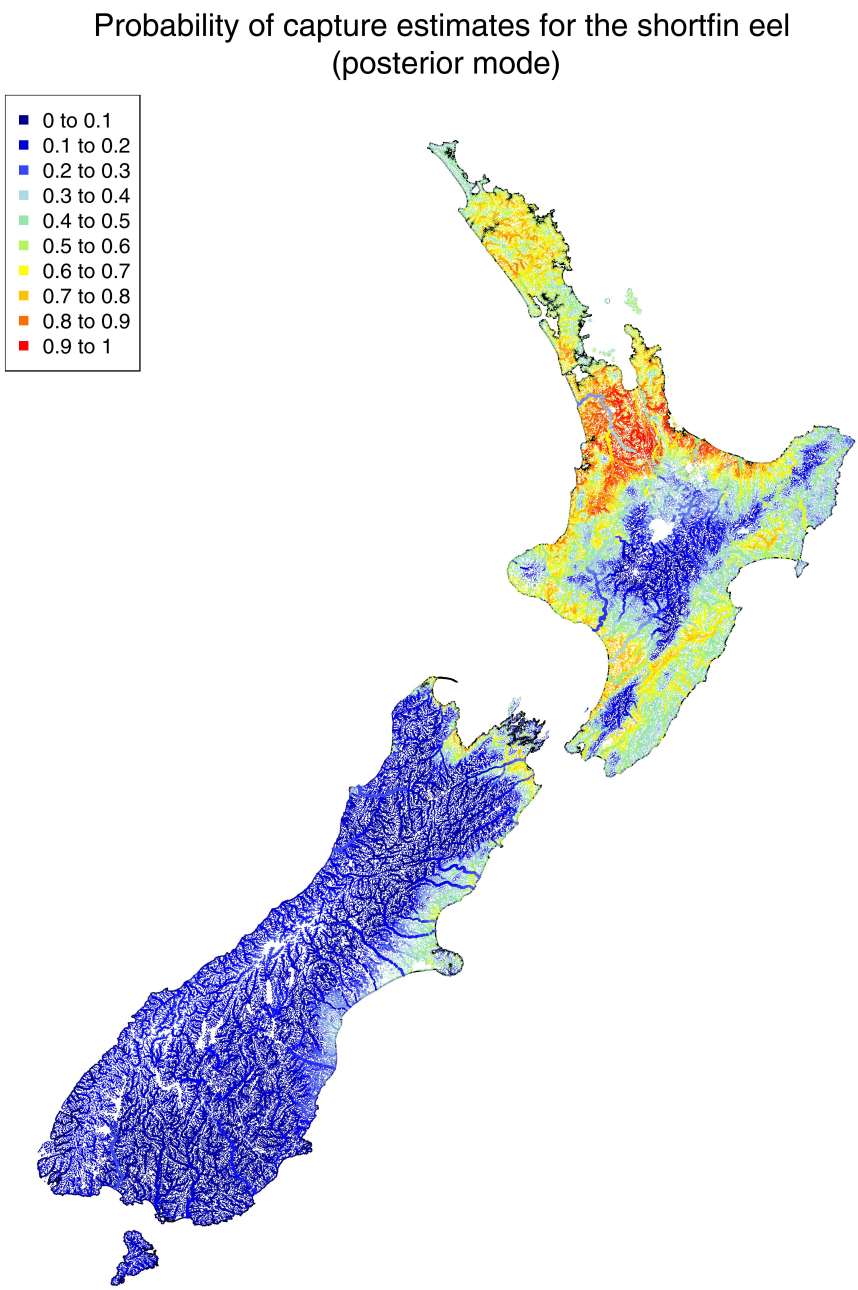


Figure 4.55: Posterior mode probability of capture estimates for the shortfin eel using the REC2 database. These estimates were made using the shortfin eel GRaF model. Larger points have a larger stream order. This is repeated from Figure 4.54b.

quite low (0.3 to 0.5).

The probabilities of capture in Figure 4.54b for shortfin eels are very low throughout the majority of the South Island and Stewart Island (0 to 0.3). However, parts of the central-east coast and north coast of the South Island show slightly higher probabilities ranging from 0.3 to 0.7. This same approximate pattern is shown in the observed proportions of shortfin eel capture (Figure 4.2).

The lower 95% credible interval for shortfin eel probability of capture (Figure 4.54a) shows very low probabilities throughout the majority of the country (0 to 0.2). The Auckland and Waikato region show probabilities ranging from 0.4 to 0.7.

The upper 95% credible interval for shortfin eel probability of capture (Figure 4.54c) shows very high probabilities throughout the majority of the North Island (0.8 to 1). Whereas the central North Island and small parts of the central east North Island, central west North Island and south of the North Island show upper credible interval values of 0.6 to 0.7. The South Island shows upper credible interval values of 0.4 to 0.7 throughout the centre of the island and probabilities around 0.7 to 0.8 across areas of the east and west coast. Some parts of the coast show high probabilities (0.8 to 1) for upper credible interval values.

The large 95% credible intervals show that the model is very uncertain about shortfin eel probabilities of capture. However, we are more certain about areas in the central South Island having low probabilities of capture as these areas range from 0 to 0.4.

Cross validation results

5-fold cross validation was performed on the shortfin eel GRaF model. The ROC curves are shown in Figure 4.56. The cross validation found a mean AUC of 0.8861 (4dp) with a 95% confidence interval of 0.8790 and 0.8933 (with a standard error of 0.0036). The mean AUC is large and the 95% confidence interval shows certainty in the estimate. Hence, the shortfin eel GRaF model performs well in predicting the probability of capture in areas spatially dependent to the training set.

There is very little variability in the 5 ROC curves of Figure 4.56. This indicates that the model is consistently performing well.

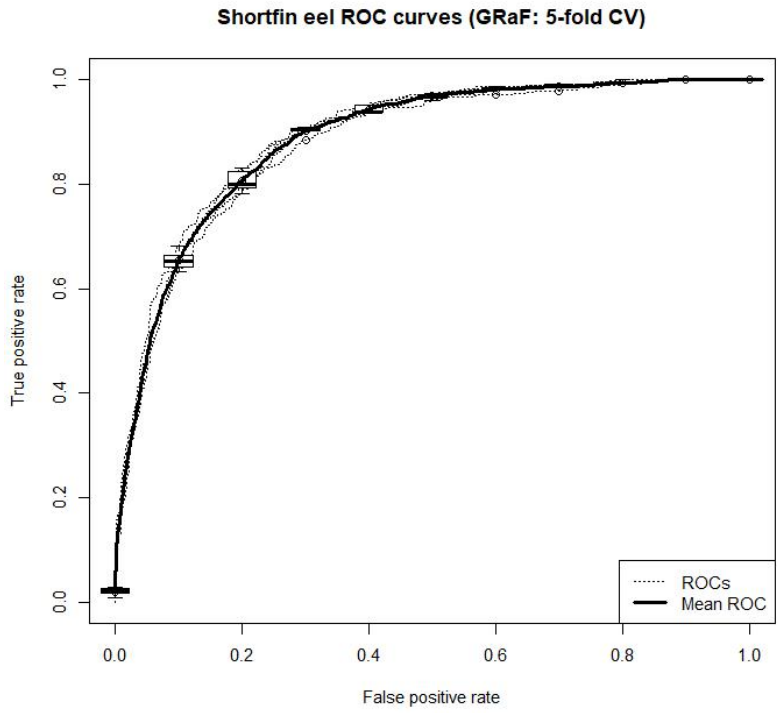


Figure 4.56: ROC curves under each of the 5 folds in the shortfin eel GRaF model cross validation. These are shown in grey and the mean ROC curve is shown in black. Boxplots show the spread of the curves.

4.4 Model comparison results

Tables 4.4 and 4.5 give AUC estimates, 95% confidence intervals and standard errors under various modelling techniques and validation techniques for the longfin and shortfin eel respectively. The tables are split into three groups: 50-SCV (50-fold spatial cross validation), 50-CV (50-fold cross validation) and 5-CV (5-fold cross validation). Models can be compared amongst the same cross validation technique.

By looking at Tables 4.4 and 4.5 it's clear that 50-fold cross validation and 5-fold cross validation achieve very similar results for the models evaluated in both cases, under the longfin eel and shortfin eel. Using an influence curve based approach to calculating AUC confidence intervals and standard errors (Hampel, 1974; LeDell et al., 2015), the estimates under 50-fold cross validation and 5-fold cross validation are very similar. Hence, we can conclude that the models are fairly stable. This means that using a larger number of folds does not change the results significantly.

Model	Validation	AUC	95% CI	SE
RRF	50-SCV	0.6550	0.6444, 0.6656	0.0054
VAST	50-SCV	0.6646	0.6542, 0.6751	0.0053
VAST-MS	50-SCV	0.6619	0.6515, 0.6722	0.0053
RRF	50-CV	0.7798	0.7709, 0.7887	0.0045
VAST	50-CV	0.8321	0.8243, 0.8399	0.0040
VAST-MS	50-CV	0.8310	0.8232, 0.8387	0.0040
RRF	5-CV	0.7799	0.7710, 0.7890	0.0045
VAST	5-CV	0.8269	0.8190, 0.8348	0.0040
VAST-MS	5-CV	0.8262	0.8183, 0.8341	0.0040
GRaF	5-CV	0.8272	0.8194, 0.8350	0.0040

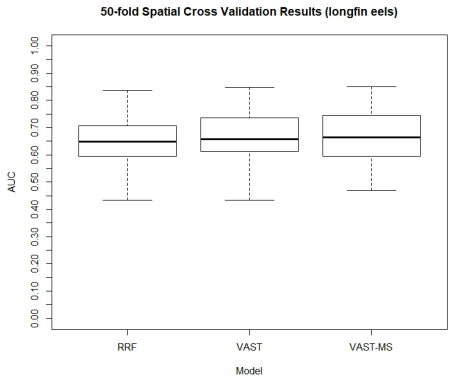
Table 4.4: AUC estimates (4dp) for longfin eel models under 50-fold spatial cross validation (50-SCV), 50-fold cross validation (50-CV) and 5-fold cross validation (5-CV).

Model	Validation	AUC	95% CI	SE
RRF	50-SCV	0.7443	0.7329, 0.7557	0.0058
VAST	50-SCV	0.7864	0.7754, 0.7974	0.0056
VAST-MS	50-SCV	0.8053	0.7950, 0.8156	0.0053
RRF	50-CV	0.8692	0.8613, 0.8771	0.0040
VAST	50-CV	0.9046	0.8981, 0.9111	0.0033
VAST-MS	50-CV	0.9025	0.8960, 0.9091	0.0034
RRF	5-CV	0.8674	0.8595, 0.8754	0.0041
VAST	5-CV	0.9006	0.8940, 0.9073	0.0034
VAST-MS	5-CV	0.8998	0.8931, 0.9065	0.0034
GRaF	5-CV	0.8861	0.8790, 0.8933	0.0036

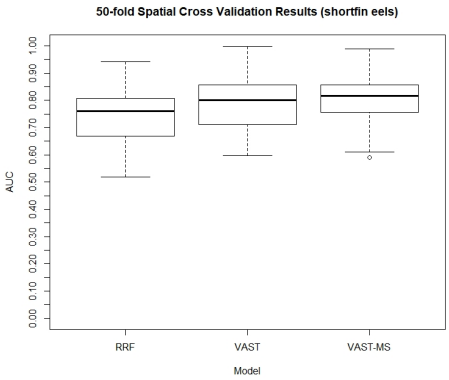
Table 4.5: AUC estimates (4dp) for shortfin eel models under 50-fold spatial cross validation (50-SCV), 50-fold cross validation (50-CV) and 5-fold cross validation (5-CV).

The findings from Tables 4.4 and 4.5 are displayed by boxplots in Figures 4.57a and 4.57b, 4.58a and 4.58b, and 4.59a and 4.59b. Figures 4.57a and 4.57b display the AUC estimates of each of the 50-folds of the spatial cross validation for the longfin eel and shortfin eel respectively. Figures 4.57a and 4.57b display the AUC estimates of each of the 50-folds of the cross validation for the longfin eel and shortfin eel respectively. Figures 4.59a and 4.59b display the AUC estimates for 5-fold cross validation. 5-fold cross validation was used to compare the GRaF models.

From Table 4.4 and Figure 4.57a we can see that the AUC estimates under the longfin eel RRF model, longfin eel VAST model and multi-species VAST model are not significantly different from one another. The boxplots have significant overlap between one another and the 95% confidence intervals have large overlap. Therefore, we cannot distinguish between the models through 50-fold spatial cross validation. These results indicate that these models perform approximately the same at estimating the probability of capture for longfin eels in spatial areas outside of the training data.



(a) Longfin eel AUC estimates for RRF, VAST single species and VAST multi-species modelling techniques.



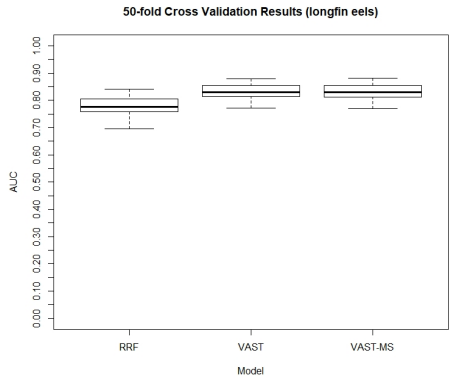
(b) Shortfin eel AUC estimates for RRF, VAST single species and VAST multi-species modelling techniques.

Figure 4.57: 50-fold spatial cross validation AUC estimates.

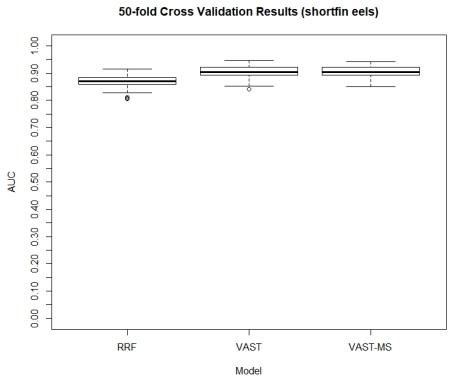
Additionally, these estimates will not be very accurate (low AUC).

The boxplots in Figure 4.57b shows fairly large variability in the shortfin eel 50-fold spatial cross validation AUC estimates. However, the 95% confidence interval (see Table 4.5) based on influence curves did not overlap for the RRF model and VAST single species model, and the RRF model and the VAST multi-species model. This indicates that we can say with 95% confidence that the VAST model (either single species or multi-species) will perform slightly better than the RRF in predicting probability of capture for shortfin eels in spatial areas outside of the training data. The multi-species VAST model performs, on average, much better than the RRF model in making these estimates. This is shown by a great difference in confidence intervals.

Figure 4.59a and Table 4.4 show that there is very little variability in the 5 folds of the cross validation, regardless of the model used. The VAST single species, VAST multi-species and GRaF longfin eel models all perform better than the RRF longfin eel model. This is because the 95% confidence intervals do not overlap. However, the 5-fold cross validation shows no



(a) Longfin eel AUC estimates for RRF, VAST single species and VAST multi-species modelling techniques.



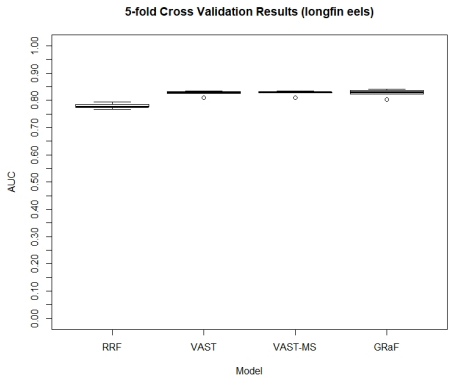
(b) Shortfin eel AUC estimates for RRF, VAST single species and VAST multi-species modelling techniques.

Figure 4.58: 50-fold cross validation AUC estimates.

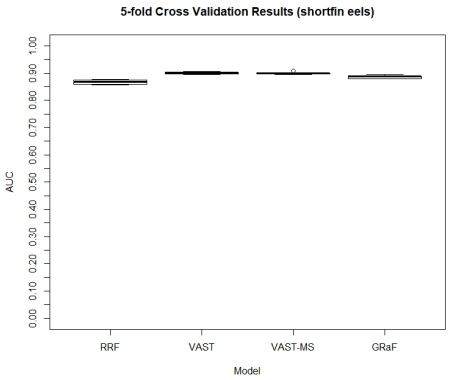
difference in 95% confidence intervals between the VAST single species, VAST multi-species and GRaF models.

Figures 4.59b and 4.5 show that the VAST single species and VAST multi-species shortfin eel models perform better than the RRF model. This is because the 95% confidence intervals do not overlap. The 95% confidence interval for the GRaF shortfin eel model (0.8790, 0.8933) only just falls outside of the 95% confidence interval for the RRF shortfin eel model (0.8595, 0.8754). This occurs at the third decimal place. Since this difference is so small, we cannot be certain that the GRaF shortfin eel model truely improves on probability of capture predictions, compared to the RRF shortfin eel model, or if this difference is because of the small number of folds used. Hence, we cannot justify concluding a difference between these models.

The 95% confidence intervals for the GRaF shortfin eel model (0.8790, 0.8933), VAST single species model (0.8940, 0.9073), and VAST multi-species model (0.8931, 0.9065) all overlap or only just fall outside of each others confidence band (to the third decimal place). Hence, we cannot justify



(a) Longfin eel AUC estimates for RRF, VAST single species, VAST multi-species and GRaF modelling techniques.



(b) Shortfin eel AUC estimates for RRF, VAST single species, VAST multi-species and GRaF modelling techniques.

Figure 4.59: 5-fold cross validation AUC estimates.

concluding that these models are statistically different from one another.

In all the evaluated cross validations, the VAST multi-species model does not perform significantly better than the VAST single species model when predicting probability of capture for longfin eels or shortfin eels (see Tables 4.4 and 4.5). This is not surprising given the lack of correlation between the two species (see Figure 4.40). However, when comparing the VAST models against the RRF model for the shortfin eel and under spatial cross validation, the multi-species VAST model estimates probability of capture significantly better than the shortfin eel RRF model. The overlap between the VAST single species model and multi-species model is very small. This indicates that it may be worth using the multi-species VAST modelling approach when attempting to estimate probability of capture for shortfin eels in spatial areas outside of the models domain.

Patterns of longfin eel and shortfin eel probability of capture appear approximately the same in the maps made by the RRF models, VAST models and GRaF.

Chapter 5

Discussion and conclusion

This research aimed to model longfin eel and shortfin eel probability of capture and to compare the modelling techniques. The RRF modelling technique has been used by Crow et al. (2014) as a method for predicting the probability of capture for New Zealand freshwater fish (including the longfin eel and shortfin eel). Hence, the research aimed to assess whether or not the VAST and GRaF modelling approaches improved upon the RRF approach.

This research found that the VAST single species approach, the VAST multi-species approach and the GRaF approach all significantly improved probability of capture estimation for the longfin eel, compared to the RRF approach. This improvement was measured using AUC and the improvements were only shown when making probability of capture predictions in spatial areas which are spatially dependent to the training data.

Only the VAST multi-species approach and the VAST single species approach showed an improvement (in AUC) over the RRF approach for the shortfin eel. These improvements were shown in spatial areas which were spatially dependent and spatially independent of the training data.

In the case where predictions would like to be made to areas of New Zealand where researchers contain no information of longfin eel presence or absence. The RRF performs just as well as the VAST approaches (com-

parisons of GRaF were not possible). Whereas the VAST approaches perform better (as measured by AUC) on average, compared to the RRF approach, when researchers have no information of shortfin eel presence or absence for an area of New Zealand. Hence, the VAST modelling approaches do no worse than a RRF modelling approach in these situations.

These results show that the VAST approach offers significant improvements over the RRF approach in modelling the probability of capture for both longfin eels and shortfin eels. This means that further longfin eel and shortfin eel conservation management should rely upon the results of the VAST models presented in this research. In particular, the VAST probability of capture maps shown in Figures 4.15, 4.16, 4.26, and 4.27 or in Figures B.1, B.2, B.3 and B.4 can be used as a tool for the management of these species.

The VAST probability of capture estimates can be used in an eel abundance model. This can be done within the VAST modelling software as a component of the delta model (Thorson & Barnett, 2017; Thorson, 2019) or as part of a separate stock assessment model (e.g. used as a model predictor).

The GRaF modelling approach significantly improves upon the regularized random forest (RRF) approach for predicting longfin eel probability of capture but performs just as well as the RRF in predicting shortfin eel probability of capture. Hence, the GRaF approach could be used over the RRF approach for conservation management of longfin eels. In particular, the map in Figure 4.50b could be used.

The VAST modelling approaches and the GRaF modelling approach perform approximately the same (as measured by AUC). This is shown in Tables 4.4 and 4.5. However, the GRaF models were computationally expensive to run in comparison to the VAST models. Using a powerful NIWA server (245GB of RAM), the GRaF models took c.4.3 days to converge for the longfin eel and c.4.4 days for the shortfin eel. In comparison, on this same server, the longfin eel VAST model took c.1.3 hours, the short-

fin eel VAST model took c.0.5 hours and the multi-species VAST model took c.6.0 hours. Hence, the VAST models are faster to run and perform just as well. Additionally, the VAST models can be run on an everyday computer (8GB of RAM) on lower model settings. Whereas, the GRaF models could not be (without removing data or covariates).

One should note that the cross validation results tell us how well probability of capture is estimated, given the approach the model takes to making these estimates. This is because the RRF models, VAST models and GRaF models can incorporate different information which the other models can't. The RRF models performs feature selection and therefore select different covariates at each fold of the cross validation. Whereas, the other models use the same covariates at each fold. The VAST models incorporate spatial and temporal effects in a way that the other models can't. VAST also incorporates catchability covariates which aren't included in the other models. Finally, the GRaF models incorporate prior knowledge and use year as a covariate in the model. Again, the other models do not do this.

This research used spatial K-fold cross validation and (non-spatial) K-fold cross validation to compare models. The advantage of using both validation approaches is that one can measure how well a model performs when training data is spatially correlated to the test data (K-fold cross validation) and when it is not (spatial K-fold cross validation). Hence, spatial K-fold cross validation will give the user a measure of how well a model can make predictions in locations which are un-sampled (distinct to the data set). However, what is often more useful is getting a measure of how well the model makes predictions to areas which are spatially correlated to the training data (K-fold cross validation). This is because we cannot be certain about whether or not our model applies to locations distinct to our sampled data. Hence, it's more useful to see how the model estimates spatially correlated areas. Therefore, it's recommended that K-fold cross validation is always used to validate models and that spatial K-fold cross

validation is used when the user wants to measure how well the model performs outside of the models spatial domain.

The study by Crow et al. (2014) built probability of capture maps for the longfin eel and shortfin eel. These maps show the same approximate pattern as the longfin eel and shortfin eel RRF, VAST and GRaF maps constructed in this study. One would expect the RRF to be identical to the maps by Crow et al. (2014). For the most part this is true, any differences between the maps are due to differences in the data set which was altered to be used in the VAST models (see Chapter 3) and in the number of trees used in the RRF. This research used 1000 trees compared to the study by Crow et al. (2014) which used 500. Comparisons between the maps of the VAST models and the maps of Crow et al. (2014) are difficult to make given the differences in spatial scale between them. However, the patterns can still be distinguished and are seen to be approximately the same.

The AUC estimates made from the RRF, VAST and GRaF models of this research cannot be compared against the estimates made in Crow et al. (2014) and Leathwick et al. (2008b). This is because of differences in the data sets used and in the model evaluation methods used. Hence, future comparisons in models should use the same data set and model evaluation techniques.

Studies such as that of Grüss et al. (2017) use the VAST modelling approach to produce probability of capture maps for a variety of species in the Gulf of Mexico. Grüss et al. (2017) produced maps across different life stages of fish whereas this research focused on how the distribution of eels change with time. This research and the study by Grüss et al. (2017) highlight the flexibility in the VAST approach. Given its flexibility, one can account for various aspects of eel biology and sampling procedure. This allows us to best model longfin and shortfin eel probability of capture.

It is surprising that the shortfin eel GRaF model did not outperform the shortfin eel RRF model. The study by Golding & Purse (2016) compared the proposed GRaF approach against a variety of other approaches, in-

cluding a boosted regression tree (BRT) machine learning approach. Golding & Purse (2016) found that the GRaF modelling approach outperformed BRT models. Given that BRT have been shown to perform very well in modelling species distributions (Elith et al., 2008), one would suspect machine learning approaches such as BRT and RRF to perform similarly. Hence, according to the results of Golding & Purse (2016), we would expect GRaF models to outperform RRF models. This is only the case for longfin eels and not shortfin eels.

A strong advantage that the GRaF approach has over the VAST approach is its ability to incorporate prior knowledge. However, in order to do so, suitable prior information must be found. Hence, the GRaF models may outperform the VAST model if independent information on how each of the covariates change with probability of capture is found. Studies of longfin eel and shortfin eel probability of capture modelling have made use of the NZFFD. Hence, this information would not be independent of the data used in this study and would therefore not be a suitable prior. We are in danger of over-fitting the model when prior knowledge is dependent on the model data.

In this research, the models constructed with VAST used a resolution of 400 knots. The higher the number of knots, the finer the spatial resolution will be. This means that more knots will produce maps with greater detail. This research was restricted to 400 knots as this was the highest number of knots that could be used (due to memory restrictions) while using the bias correction feature of the VAST model. Hence, the longfin eel and shortfin eel maps produced by VAST did not estimate probability of capture for some areas of New Zealand. This is because knots are placed at positions which reduce the distance between the sampling locations and knots. Since estimates are only interpolated at a maximum distance of 15km away from a knot, some areas of New Zealand are not estimated. This is a limitation to using the VAST modelling approach.

Many of the parameter estimates made through the VAST models showed

very large uncertainty in their estimates. In particular the natural log of the parameters h_1 and h_2 (Northing anisotropy and anisotropic correlation) showed very large coefficient of variation (C.V.) percentages for all the VAST models. This indicates that there is a lack of precision in estimating how longfin eel and shortfin eels are correlated amongst themselves and each other across New Zealand. Large C.V. values indicate that the predictions made through the VAST models should be treated with caution.

A limitation to using a spatial-temporal approach with data existing within a stream network is that the approach should take into account the dendritic stream network. As of writing this, neither the VAST approach nor the GRaF approach take into account stream networks. This means that two waterways which are close to one another will be correlated but they may not actually meet at any point. This gives an unrealistic representation of the true correlation occurring in the waterways. Additionally, probability of capture predictions have been made in locations outside of waterways or above impassable natural structures and man-made structures.

The GRaF models for this research used uninformative priors. These uninformative priors gave a flat distribution at p_0 (the probability of being observed at any given site) for each of the model covariates. In real freshwater systems, we know that longfin eels and shortfin eels have habitat preferences (see Booker & Graynoth (2013)). Hence, we are limiting the potential of the GRaF model by not introducing this information. For example, Leathwick et al. (2008b) found that shortfin eels have a preference to waterways with low riparian shading. Hence, the 'segshade' covariate could express this prior knowledge by expressing lower probabilities of capture at higher 'segshade' values. However, it should be noted that the findings of Leathwick et al. (2008b) come from the NZFFD and therefore would not be an independent source of prior knowledge.

Given the large computation time of the GRaF models, the models

could only be cross validated with 5 folds as opposed to 50 folds. This also meant we couldn't use spatial cross validation to compare all the models as VAST could not converge on such a low number of folds (spatial information missing in the models). However, the results of the 50-fold cross validation and the 5-fold cross validation were very similar. Therefore, we are not concerned about using a lower number of folds to validate the models.

This research made use of the extensive voluntary data of the NZFFD. However, since contributions to the database are purely voluntary, the data hasn't been collected through a random sampling scheme. Hence, we introduce bias (which cannot be measured) by using this data. This is a major limitation to the results and should be considered when evaluating the predictions made through any of the methods assessed in this research. However, this data was deemed the most suitable for this study because of it's extensive spatial and temporal range.

The covariates used for this study (see Table A.2) were selected by the longfin eel RRF model and the shortfin eel RRF model. This meant that every longfin eel model used the same covariates and every shortfin eel model used the same covariates. However, this does not necessarily mean that the most parsimonious set of covariates were selected for the models. Hence, this has the potential to over-fit the models and therefore poorly represent the probability of capture throughout New Zealand.

The full set of covariates (see Table A.1) contained covariates at various different scales. For example, spatial covariates were given at multiple scales and environmental covariates were given at the segment scale and the upstream scale. This has the potential to introduce multicollinearity. As an example, the longfin eel RRF model and the shortfin eel RRF model selected stream elevation at the segment scale and at the upstream scale. Elevation at the segment scale and at the upstream scale have a large positive correlation (see Figures 2.5 and 2.6). Hence, it is unnecessary to include both covariates in the models. This introduces multicollinearity

which has the potential to increase the variability in estimated model parameters (O'brien, 2007) and therefore, poorly predict probability of capture throughout New Zealand.

The NZFFD consists of covariates which have been drawn from the REC2. This means that the probability of capture predictions made at each segment of river by the RRF models and the GRaF models were made on a data set which contained entries in the training data set. It should be noted that there isn't a fully independent data set of covariates across each segment of river in New Zealand. Hence, the predictions made in Figure 4.4, 4.9, 4.50b and 4.54b are biased. This was a strong limitation to the predictions being made across the REC. An advantage with using VAST is that predictions are interpolated by the training data rather than being made to a separate data set.

As with the VAST models, probability of capture predictions made by the RRF and GRaF models have been made in locations above impassable natural structures and man-made structures. Additionally, the predictions reflect the electric fishing data (not fishing through other methods) and therefore poorly predict large waterways.

This research has identified three key areas where further research could be made to improve probability of capture predictions for the longfin eel and shortfin eel. These are:

1. VAST models could be improved by using a stream network. The approach accounts for spatial-temporal correlation within a stream as opposed to across all space (Hocking et al., 2018). This will give a more realistic correlation structure. As of writing this, the stream network proximity approach (outlined by Hocking et al. (2018)) is in development for use with VAST software.

2. Further research should be done in identifying sources of possible prior knowledge for each of the covariates in the NZFFD. This knowledge should arise from information outside of the NZFFD and could

be incorporated in GRaF models.

3. Models could be improved by using a more parsimonious set of covariates. Further research should be implemented into identifying a set of covariates which have low multicollinearity. A good starting point would be to look at variance inflation scores (VIF) (see Tables A.3 and A.4) for the longfin eel covariates and the shortfin eel covariates.

This research has found VAST modelling software (Thorson & Barnett, 2017; Thorson, 2019) to be an improvement over RRF models for modelling the probability of capture for longfin eels and shortfin eels. The VAST probability of capture models have the potential to be developed further through a stream network proximity approach and careful selection of model covariates.

Appendices

Appendix A

Modelling covariates

A.1 Model covariates

Covariate type	Label	Description
Environmental	Dist2Coast	Donwstream distance to the ocean
	StreamOrder	A number describing the Strahler order a reach in a network of reaches.
	sinuosity	Actual distance divided by the straight line distance giving the degree of curvature of the stream
	headw_dist	Distance of the furthermost source or headwater reach from any reach (m).
	Segslpmax	Maximum segment slope along length of reach.
	Segslpmean	Mean segment slope along length of reach.

seg_rain	Mean annual segment rain (mm)
us_rain	Mean annual upstream rain (mm)
seg_ro_mm	Annual segment runoff (mm)
seg_hard	Segment induration or hardness value. Ordinal scale
us_hard	Upstream induration or hardness value. Ordinal scale
seg_elev	Segment mean elevation above sea level of the watershed or basin (m)
us_elev	Upstream mean elevaltion above sea level of the watershed or basin (m)
seg_slope	Segment mean slope of the watershed or basin in degrees.
us_slope	Upstream mean slope of the watershed or basin in degrees.
seg_tmin	Segment mean minimum wintertime air temperature (deg C x 10)
us_tmin	Upstream mean minimum wintertime air temperature (deg C x 10)
seg_june	Segment June solar radiation. W/m2
us_june	Upstream June solar radiation. W/m2
seg_penpet	Segment penman potential evaporation measurement. mm

us_penpet	Upstream penman potential evaporation measurement. mm
seg_rnvar	Segment coefficient of variation of annual catchment rainfall. mm
us_rnvar	Upstream coefficient of variation of annual catchment rainfall. mm
seg_rd25	Segment catchment rain days (greater than 25mm/month). mean # days/mo
us_rd25	Upstream Catchment rain days (greater than 25mm/month). mean # days/mo
seg_rd100	Upstream Catchment rain days (greater than 100mm/month). mean # days/mo
seg_phos	Segment catchment average of phosphorous. ordinal scale.
us_phos	Upstream catchment average of phosphorous. ordinal scale.
seg_psize	Segment catchment average of particle size. ordinal scale.
us_psize	Upstream catchment average of particle size. ordinal scale.
seg_pet	Segment annual potential evapotranspiration of catchment. mm
us_pet	Upstream annual potential evapotranspiration of catchment. mm

seg_twar	Segment average within section mean January air temperature. deg C x10
us_twarm	Upstream average within section mean January air temperature. deg C x10
DSDist2Lake	Downstream Distance to lake (m). Set to 500 km if no lake present downstream
DSmax_slope	Maximum downstream slope (degrees)
DSav_slope	Average slope (degrees)
us_ind_forest	Upstream area with indigenous vegetation (m2)
US_RockPhos	Average phosphorous concentration of underlying rocks 1= very low to 5 = very high
USCalcium	Average calcium concentration of underlying rocks 1= very low to 5 = very high
us_LakeArea	Upstream area of the catchment covered by lakes (m2)
us_lakePerc	Upstream area of the catchment covered by lakes (%)
segshade	NZSegment area with riparian shade (proportion)

Spatial	x	Easting co-ordinates of the NZSegment center
	y	Northing co-ordinates of the NZSegment center
	xy	Multiple of XY from the cubic trend surface regression formula
	y2	Square of Y from the cubic trend surface regression formula
	x2	Square of X from the cubic trend surface regression formula
	x3	Cube of X from the cubic trend surface regression formula
	yx2	Multiple of X2Y from the cubic trend surface regression formula
	xy2	Multiple of XY2 from the cubic trend surface regression formula
	y3	Cube of Y from the cubic trend surface regression formula
Hydrological	Constancy	Constancy of mean-monthly flows (see Colwell (1974))
	Contingency	Consistancy mean-monthly flows among years (see Colwell (1974))
	FRE1.Count	Number of flows greater than the median. Expressed as ratio of mean flow.

FRE1.MaxDurBetween	Maximum duration between flows greater than the median/mean flow
FRE1.MeanDurBetween	Mean duration between flows greater than the median/mean flow.
FRE10.Count	Number of flows greater than ten times the median/mean flow
FRE10.MaxDurBetween	Maximum duration between flows greater than ten times the median/mean flow.
FRE10.MeanDurBetween	Mean duration between flows greater than ten times the median/mean flow.
FRE5.Count	Number of flows greater than five times the median/mean flow.
FRE5.MaxDurBetween	Maximum duration between flows greater than five times the median/mean flow.
FRE5.MeanDurBetween	Mean duration between flows greater than five times the median/mean flow.
JulianMax	Annual maximum flow/mean flow
JulianMin	Annual minimum flow/mean flow
l1	First linear moment of daily flows/catchment area

l2	Second linear moment of daily flows/catchment area
lca	Ratio of the first and second linear moment of daily flows/catchment area
lcv	Linear moments coefficient of variation/catchment area
lkur	Third linear moment of daily flows/catchment area
Mean1DayFlowMaxs	Mean annual maximum 1 day flow / mean flow
Mean1DayFlowMins	Mean annual minimum 1 day flow / mean flow
Mean7DayFlowMaxs	Mean annual maximum 7 day flow / mean flow
Mean7DayFlowMins	Mean annual minimum 7 day flow / mean flow
Mean90DayFlowMaxs	Mean annual maximum 90 day flow / mean flow
Mean90DayFlowMins	Mean annual minimum 90 day flow / mean flow
meanNeg.StandardisedBy MeanFlow	Mean number of all negative differences between days/mean flow
meanPos.StandardisedBy MeanFlow	Mean number of all positive differences between days/mean flow
MeanPulseLengthHigh	Mean duration of high pulses/mean flow

MeanPulseLengthLow	Mean duration of low pulses/mean flow
nNeg.StandardisedByMean Flow	Number of all negative differences between days/mean flow
nPos.StandardisedByMean Flow	Number of all positive differences between days/mean flow
nPulsesHigh	Number of high pulses within each water year/mean flow
nPulsesLow	Number of low pulses within each water year/mean flow
Predictability	Predictability of mean-monthly flows (Colwell 1974)
Reversals	Number of hydrologic reversals/mean flow
WidthHUC.MALF	Mean annual low flow in cumecs

Table A.1: The table is replicated from Crow et al. (2014) and describes the covariates considered for the RRF model and it's associated label. The table also details the type of covariate that it is.

A.2 Covariates selected by RRF

Covariate label	Species	
	Longfin eel	Shortfin eel
Constancy	1	1
Contingency	1	1
Dist2Coast	1	1
DSav_slope	1	1
DSDist2Lake	0	0
DSmax_slope	1	1
FRE1.Count	0	1
FRE10.Count	0	1
FRE5.Count	1	1
headw_dist	1	1
JulianMax	1	1
JulianMin	1	1
lca	1	1
lcv	1	0
lkur	1	0
FRE1.MaxDurBetween	1	1
FRE1.MeanDurBetween	1	0
FRE10.MaxDurBetween	1	0
FRE10.MeanDurBetween	1	0
FRE5.MaxDurBetween	1	0
FRE5.MeanDurBetween	0	0
l1	1	0
l2	1	1
Mean1DayFlowMaxs	1	0
Mean7DayFlowMaxs	1	1
Mean90DayFlowMaxs	1	1
MeanPulseLengthHigh	1	1

nPulsesHigh	1	0
Mean1DayFlowMins	1	1
Mean7DayFlowMins	1	1
Mean90DayFlowMins	1	1
meanNeg.StandardisedByMeanFlow	1	0
meanPos.StandardisedByMeanFlow	1	1
MeanPulseLengthLow	1	1
nNeg.StandardisedByMeanFlow	1	0
nPos.StandardisedByMeanFlow	1	1
nPulsesLow	1	1
Predictability	1	1
WidthHUC.MALF	1	1
Reversals	1	1
seg_elev	1	1
seg_hard	1	1
seg_june	1	0
seg_penpet	0	0
seg_pet	1	0
seg_phos	1	1
seg_psize	1	1
seg_rain	1	1
seg_rd100	0	0
seg_rd25	0	1
seg_rnvar	1	1
seg_ro_mm	1	1
seg_slope	1	1
seg_tmin	1	1
seg_twar	1	1
segshade	1	1
Segslpmax	1	1
Segslpmean	1	1

sinuosity	1	1
StreamOrder	0	0
us_elev	1	1
us_hard	1	0
us_ind_forest	1	1
us_june	1	1
us_LakeArea	0	0
us_lakePerc	0	0
us_penpet	0	1
us_pet	1	0
us_phos	1	1
us_psize	1	1
us_rain	1	0
us_rd25	0	0
us_rnvar	0	1
US_RockPhos	1	1
us_slope	1	1
us_tmin	1	1
us_twarm	1	1
USCalcium	1	1
x	0	0
x2	0	0
x3	0	1
xy	1	0
xy2	1	0
y.1	1	0
y2	0	0
y3	0	0
yx2	1	0

Table A.2: The covariates selected by the longfin eel RRF model (longfin eel column above) and the shortfin eel RRF model (shortfin eel column above). A 1 indicates that the covariate was used and a 0 indicates it wasn't used. The selected longfin eel covariates were used in the longfin eel VAST model, multi-species VAST model and the longfin eel GRaF model. The selected shortfin eel covariates were used in the shortfin eel VAST model and the shortfin eel GRaF model.

A.3 Variance Inflation factors

Covariates	VIF score
sinuosity	1.18
DSav_slope	1.80
USCalcium	1.81
REC1_WidthHUC.MALF_cumecs	2.10
us_ind_forest	2.31
segshade	2.57
JulianMax.StandardisedByMeanFlow	2.77
seg_twar	3.08
DSmax_slope	3.09
Dist2Coast	4.51
headw_dist	5.71
seg_rnvar	6.39
seg_phos	6.74
us_phos	6.90
JulianMin.StandardisedByMeanFlow	7.01
seg_hard	7.06
Reversals.StandardisedByMeanFlow	7.06
Contingency	7.57

seg_psize	7.90
us_slope	8.26
us_hard	8.54
Segslpmax	8.65
seg_ro_mm	9.65
us_psize	10.32
MeanPulseLengthLow.StandardisedByMeanFlow	10.43
seg_slope	11.07
Segslpmean	11.44
US_RockPhos	11.82
meanNeg.StandardisedByMeanFlow	18.58
seg_rain	22.36
us_tmin	23.71
us_rain	24.27
Log10_Mean90DayFlowMaxs.StandardisedByMeanFlow	29.24
seg_elev	31.87
seg_tmin	32.96
nPulsesLow.StandardisedByMeanFlow	37.62
meanPos.StandardisedByMeanFlow	40.56
Mean90DayFlowMins.StandardisedByMeanFlow	44.05
Log10_FRE10.MaxDurBetween.StandardisedByMeanFlow	45.14
lca.StandardisedByCatchArea	45.83
Log10_FRE10.MeanDurBetween.StandardisedByMeanFlow	48.96
lkur.StandardisedByCatchArea	50.08
lcv.StandardisedByCatchArea	51.67
Predictability	52.90
Constancy	55.27
seg_pet	59.53
Log10_Mean1DayFlowMaxs.StandardisedByMeanFlow	62.03
nNeg.StandardisedByMeanFlow	63.31
Log10_FRE5.MaxDurBetween.StandardisedByMeanFlow	64.71

nPos.StandardisedByMeanFlow	65.73
us_pet	66.30
Log10_Mean7DayFlowMaxs.StandardisedByMeanFlow	69.62
Mean1DayFlowMins.StandardisedByMeanFlow	71.66
Mean7DayFlowMins.StandardisedByMeanFlow	74.23
FRE5.Count.StandardisedByMeanFlow	74.88
us_elev	79.19
Log10_FRE1.MaxDurBetween.StandardisedByMeanFlow	82.56
us_twarm	84.76
Log10_FRE1.MeanDurBetween.StandardisedByMeanFlow	91.93
Log10_l1.StandardisedByCatchArea	155.18
Log10_l2.StandardisedByCatchArea	158.54
Log10_MeanPulseLengthHigh.StandardisedByMeanFlow	206.91
Log10_nPulsesHigh.StandardisedByMeanFlow	212.98
seg_june	463.72
us_june	506.24
xy	1230.39
y.1	1789.26
yx2	2922.77
xy2	9227.69

Table A.3: Variance inflation factors (VIF) for each of the longfin eel covariates, ordered by size.

Covariates	VIF score
sinuosity	1.17
USCalcium	1.69
DSav_slope	1.76
REC1_WidthHUC.MALF_cumecs	2.09
us_ind_forest	2.11
JulianMax.StandardisedByMeanFlow	2.15
segshade	2.41
DSmax_slope	3.01
seg_twar	3.07
Dist2Coast	4.00
seg_hard	4.22
headw_dist	4.57
us_psize	4.71
Reversals.StandardisedByMeanFlow	5.72
JulianMin.StandardisedByMeanFlow	6.03
seg_psize	6.21
us_penpet	6.38
seg_phos	6.61
us_phos	6.74
Contingency	6.95
nPos.StandardisedByMeanFlow	7.00
us_slope	7.46
x3	7.48
Segslpmax	8.55
MeanPulseLengthLow.StandardisedByMeanFlow	9.33
seg_rd25	10.17
seg_ro_mm	10.26
seg_slope	10.89

Segslpmean	11.31
US_RockPhos	11.64
seg_rain	12.74
FRE10.Count.StandardisedByMeanFlow	15.32
seg_elev	15.64
lca.StandardisedByCatchArea	16.98
us_tmin	18.92
meanPos.StandardisedByMeanFlow	23.08
seg_tmin	24.23
Log10_Mean90DayFlowMaxs.StandardisedByMeanFlow	26.89
Log10_l2.StandardisedByCatchArea	28.31
us_june	28.58
Mean90DayFlowMins.StandardisedByMeanFlow	36.00
FRE5.Count.StandardisedByMeanFlow	36.69
Log10_Mean7DayFlowMaxs.StandardisedByMeanFlow	38.31
nPulsesLow.StandardisedByMeanFlow	41.66
Log10_MeanPulseLengthHigh.StandardisedByMeanFlow	45.07
us_elev	45.56
Predictability	50.17
Constancy	52.36
seg_rnvar	54.03
us_rnvar	54.72
us_twarm	60.99
Log10_FRE1.MaxDurBetween.StandardisedByMeanFlow	65.37
Mean1DayFlowMins.StandardisedByMeanFlow	65.71
Mean7DayFlowMins.StandardisedByMeanFlow	71.08
FRE1.Count.StandardisedByMeanFlow	98.49

Table A.4: Variance inflation factors (VIF) for each of the shortfin eel covariates, ordered by size.

Appendix B

VAST probability of capture Figures

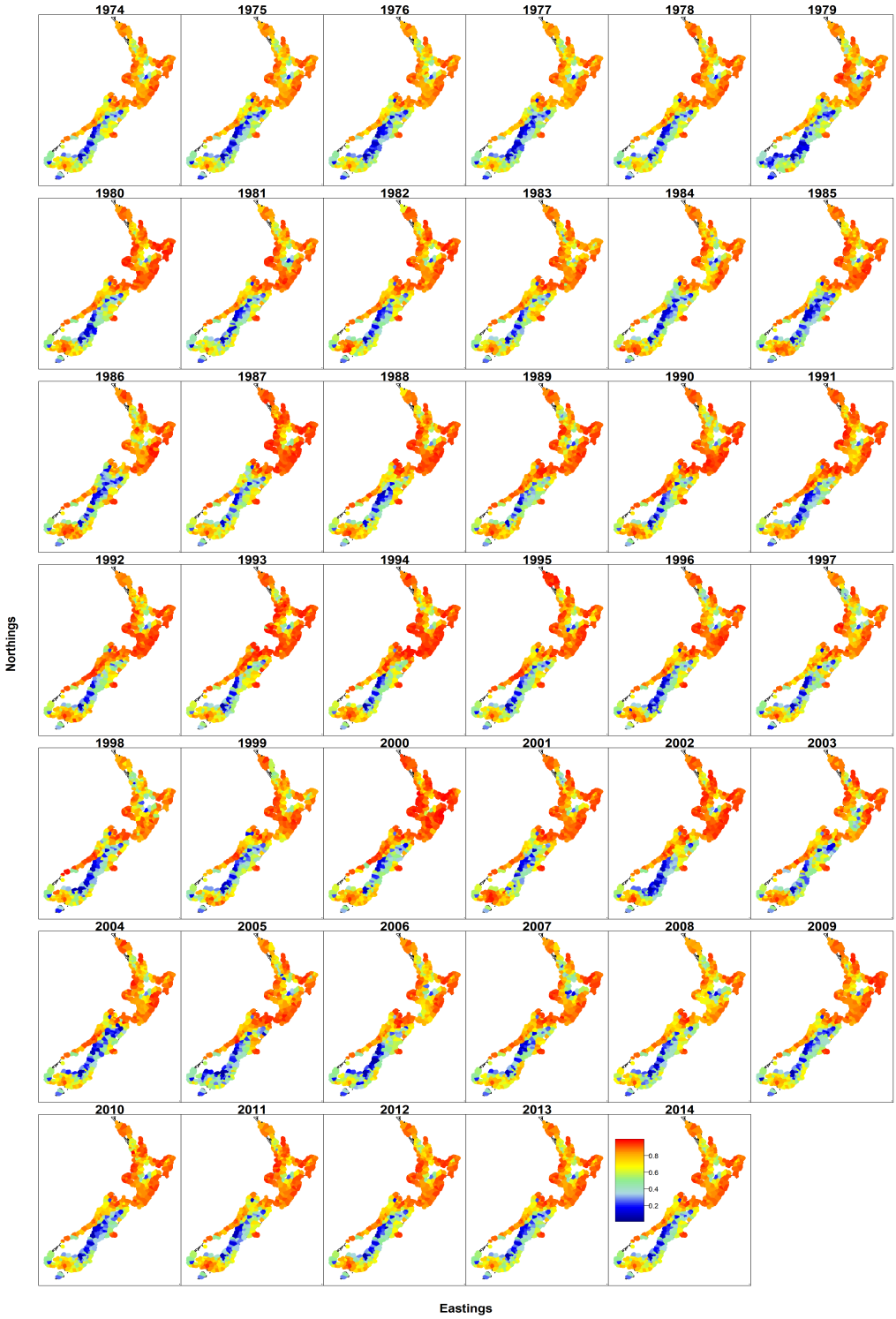


Figure B.1: Longfin eel probability of capture estimates for 1974 to 2014 from the multi-species VAST model. The estimates are shown on a northing-easting coordinate grid.

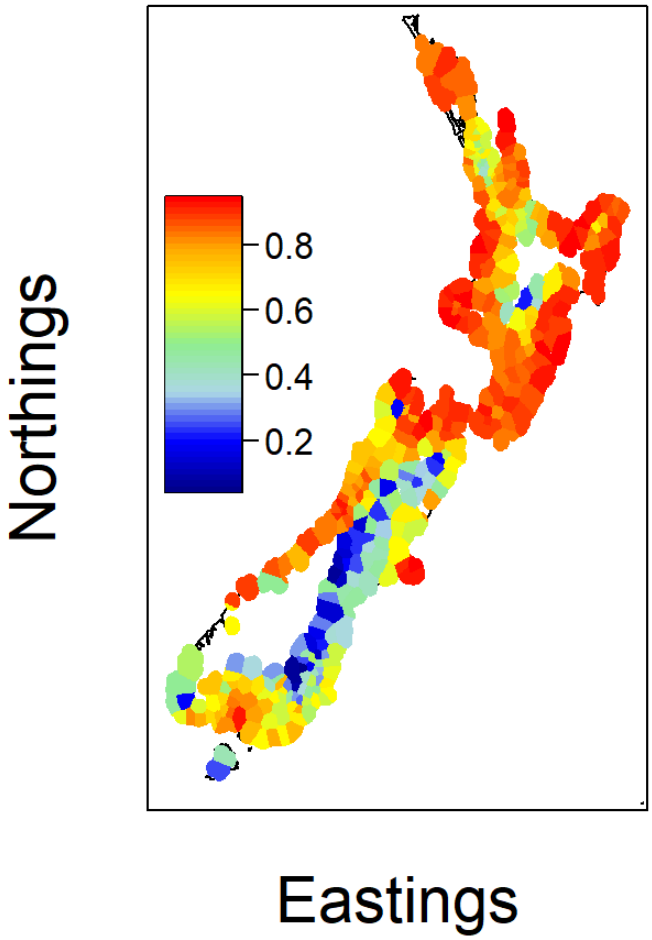


Figure B.2: Map of the 2014 longfin eel probability of capture estimates made from the multi-species VAST model. The estimates are shown on a northing-easting coordinate grid.

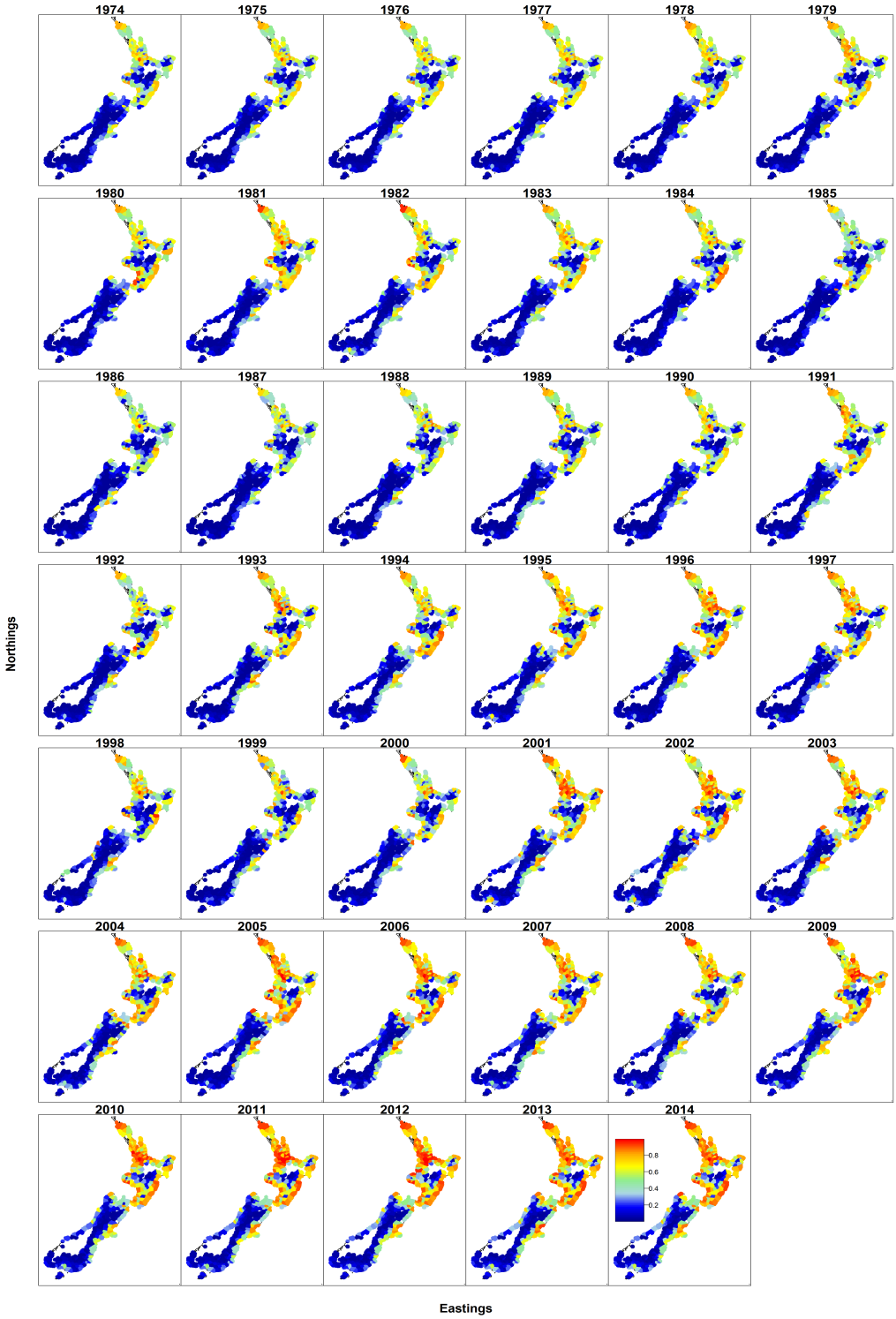


Figure B.3: Shortfin eel probability of capture estimates for 1974 to 2014 from the multi-species VAST model. The estimates are shown on a northing-easting coordinate grid.

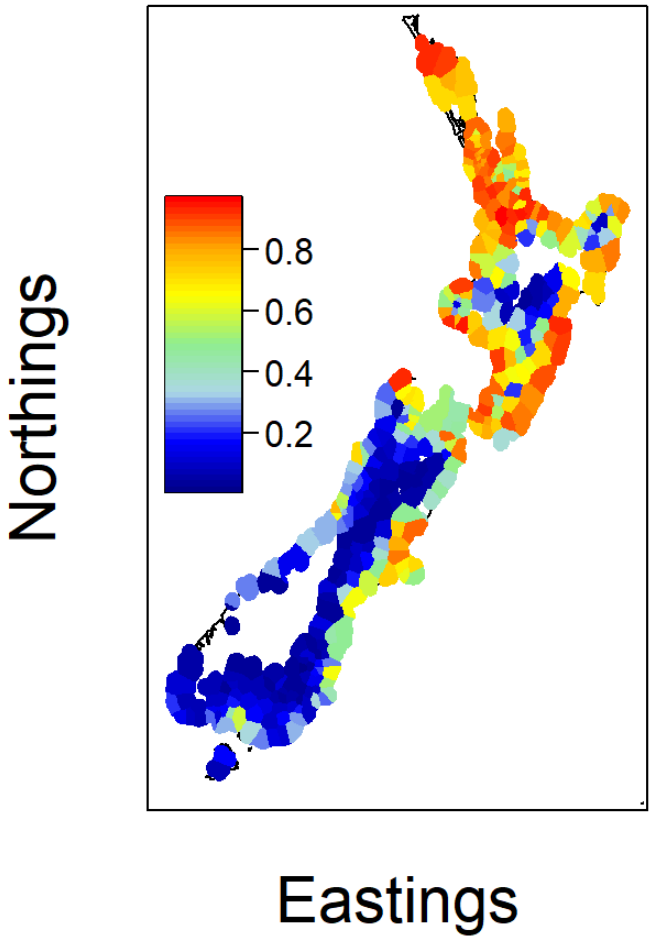


Figure B.4: Map of the 2014 shortfin eel probability of capture estimates made from the multi-species VAST model. The estimates are shown on a northing-easting coordinate grid.

Appendix C

GRaF lengthscale tables

Model Covariates	Length-scales
year	0.35
us_june	2.29
seg_rnvar	2.98
us_twarm	4.61
seg_june	5.95
xy	5.99
seg_elev	6.14
us_tmin	6.39
yx2	7.18
xy2	7.65
us_pet	8.14
y.1	8.35
Dist2Coast	9.49
nNeg.StandardisedByMeanFlow	11.88
Log10_Mean90DayFlowMaxs.StandardisedByMeanFlow	11.93
seg_tmin	13.83
us_elev	16.02

nPos.StandardisedByMeanFlow	16.87
segshade	17.46
MeanPulseLengthLow.StandardisedByMeanFlow	18.60
headw_dist	19.92
seg_pet	19.98
Contingency	20.41
us_phos	20.73
nPulsesLow.StandardisedByMeanFlow	20.81
JulianMax.StandardisedByMeanFlow	20.83
DSav_slope	21.25
us_slope	21.76
Log10_FRE5.MaxDurBetween.StandardisedByMeanFlow	22.01
Predictability	22.39
Log10_l2.StandardisedByCatchArea	22.86
Log10_l1.StandardisedByCatchArea	23.81
Constancy	24.19
meanNeg.StandardisedByMeanFlow	24.35
FRE5.Count.StandardisedByMeanFlow	24.69
US_RockPhos	24.92
lkur.StandardisedByCatchArea	25.03
Log10_nPulsesHigh.StandardisedByMeanFlow	26.36
Log10_FRE1.MaxDurBetween.StandardisedByMeanFlow	27.32
Log10_MeanPulseLengthHigh.StandardisedByMeanFlow	27.69
Log10_Mean7DayFlowMaxs.StandardisedByMeanFlow	27.76
seg_slope	28.53
JulianMin.StandardisedByMeanFlow	28.67
Log10_FRE1.MeanDurBetween.StandardisedByMeanFlow	28.68
Mean90DayFlowMins.StandardisedByMeanFlow	29.03
us_ind_forest	30.46
USCalcium	30.76
Mean7DayFlowMins.StandardisedByMeanFlow	31.48

Log10_Mean1DayFlowMaxs.StandardisedByMeanFlow	31.49
Log10_FRE10.MaxDurBetween.StandardisedByMeanFlow	33.22
meanPos.StandardisedByMeanFlow	33.80
Segslpmean	34.41
us_hard	34.46
Log10_FRE10.MeanDurBetween.StandardisedByMeanFlow	34.56
us_psize	35.06
Mean1DayFlowMins.StandardisedByMeanFlow	35.84
us_rain	36.09
seg_rain	36.44
seg_phos	37.93
lca.StandardisedByCatchArea	40.45
Reversals.StandardisedByMeanFlow	40.78
lcv.StandardisedByCatchArea	42.68
seg_ro_mm	44.58
DSmax_slope	48.16
seg_twar	50.64
Segslpmax	57.10
seg_hard	58.26
REC1_WidthHUC.MALF_cumecs	70.06
seg_psize	80.70
sinuosity	152.66

Table C.1: The lengthscales (2dp) for each of the covariates of the longfin eel GRaF model ordered from smallest (most complex function) to largest (least complex function). See Table A.1 for a description of the covariates.

Model.Covariates	Length-scales
year	0.27
x3	2.24
us_june	2.36
us_tmin	3.17
us_twarm	5.23
us_elev	8.94
seg_tmin	8.99
Dist2Coast	12.13
seg_rnvar	12.37
segshade	13.59
us_penpet	14.52
DSav_slope	16.10
Log10_l2.StandardisedByCatchArea	16.93
us_slope	16.94
seg_elev	17.47
seg_rd25	18.20
us_rnvar	18.28
JulianMin.StandardisedByMeanFlow	18.40
DSmax_slope	18.43
seg_rain	21.38
meanPos.StandardisedByMeanFlow	22.93
FRE5.Count.StandardisedByMeanFlow	24.02
nPulsesLow.StandardisedByMeanFlow	24.47
JulianMax.StandardisedByMeanFlow	24.91
FRE10.Count.StandardisedByMeanFlow	25.00
seg_ro_mm	25.45
FRE1.Count.StandardisedByMeanFlow	26.25
lca.StandardisedByCatchArea	26.47

Log10_MeanPulseLengthHigh.StandardisedByMeanFlow	26.56
headw_dist	27.31
us_phos	27.66
Log10_Mean90DayFlowMaxs.StandardisedByMeanFlow	28.55
Predictability	28.70
Log10_FRE1.MaxDurBetween.StandardisedByMeanFlow	29.02
Constancy	29.57
MeanPulseLengthLow.StandardisedByMeanFlow	31.52
nPos.StandardisedByMeanFlow	31.59
us_ind_forest	31.95
Log10_Mean7DayFlowMaxs.StandardisedByMeanFlow	34.47
Contingency	36.42
Mean90DayFlowMins.StandardisedByMeanFlow	37.46
Reversals.StandardisedByMeanFlow	40.92
us_psize	40.92
US_RockPhos	41.49
Segslpmean	46.90
Mean7DayFlowMins.StandardisedByMeanFlow	47.78
Mean1DayFlowMins.StandardisedByMeanFlow	48.68
USCalcium	49.59
seg_phos	51.35
seg_slope	53.34
seg_twar	62.62
seg_psize	64.75
Segslpmax	68.10
seg_hard	69.42
REC1_WidthHUC.MALF_cumecs	71.53
sinuosity	162.47

Table C.2: The lengthscales (2dp) for each of the covariates of the shortfin eel GRaF model ordered from smallest (most complex function) to largest (least complex function). See Table A.1 for a description of the covariates.

Appendix D

R modelling code

The following gives the R code for building the RRF longfin eel and shortfin eel models.

```
setwd("D:/Masters/RRF Model") #set working directory

####  PACKAGES
library(RRF) #RRF pacakage
library(ROCR) #auc function
library(cvAUC) #auc CI's
library(sperrorest) #K-means spatial partitioning function

############################ Load and edit data ##############################################
load(file = "D:/Masters/Data/My_NZFFD.REC2.Diad.EF.Rdata") #load data

diad.preds <- read.csv("D:/Masters/RF R stuff/Fish predictor list to use for RandForest models.csv") #all
      the predictor variables
Xvars <- diad.preds$predictors[which(diad.preds$diadromous == "T")] #use these predictors
rm(diad.preds) #remove as we no longer need this
Xvars <- as.character(Xvars) #set as a character

Xvars<-Xvars[-1] #fishmeth isn't needed as a covariate as we are only using EF data

############################ Run model ###############################################

#RRF functions
GetRRFModel <- function(y, myExplanatoryFrame, x, Classification = F, ...) {
if(Classification) { #when using RRF as a classifier
myExplanatoryFrame[[y]] <- factor(myExplanatoryFrame[[y]]) #set y as factor
levels(myExplanatoryFrame[[y]]) <- c("F", "T") #set levels
}
myOut <- RRF(y = myExplanatoryFrame[, y], x = myExplanatoryFrame[ , x], ...) #run RRF model
return(myOut) #return RRF model
}

#########################################################################################
#Longfin eel Model
set.seed(22)
ModelListMMQ.RRF.angdie <- lapply("angdie", GetRRFModel, myExplanatoryFrame = NZFFD.REC2.Diad.EF, x = Xvars,
Classification = T, ntree=1000) #run rrf model with 1000 Regression trees

```

```
angdie_vars_RRF = Xvars[ModelListMMQ.RRF.angdie[[1]]$feaSet] #Variables selected by the RRF feature selection

##########################################################################################
#Shortfin eel Model
set.seed(22)
ModelListMMQ.RRF.angaus <- lapply("angaus", GetRRFModel, myExplanatoryFrame = NZFFD.REC2.Diad.EF, x = Xvars,
Classification = T, ntree=1000) #run rrf model with 1000 Regression trees

angaus_vars_RRF = Xvars[ModelListMMQ.RRF.angaus[[1]]$feaSet] #Variables selected by the RRF feature selection


Gini_df <- data.frame(importance(ModelListMMQ.RRF.angdie[[1]]), importance(ModelListMMQ.RRF.angaus[[1]])) # obtain importance scores for lf/sf
colnames(Gini_df) <- c("angdie", "angaus") #rename columns
write.csv(Gini_df, file = "Gini_scores.csv") #save importance scores to csv

##########################################################################################
##########################################################################################
```

The following gives the R code for building the VAST longfin eel and shortfin eel single species models.

```
start1=Sys.time() #measure how long it takes
setwd("D:/Masters/VAST R stuff") #Set working directory

library("devtools")
library("Matrix")
library(TMB,lib.loc = .libPaths()[1]) #needed for VAST
library(INLA) #needed for VAST
library(SpatialDeltaGLMM) #needed for VAST
library(VAST,lib.loc = .libPaths()[1]) #VAST package
library(maps) #to visualise
library(ROCR) #auc function
library(cvAUC) #auc CI's
library(sperrorest) #K-means spatial partitioning function


## PART 1 - Load data and edit settings
## ------------------------------
##
## Load the data and select the species to model - angdie or angaus.

############################ Load and edit data ##############################################
load("D:/Masters/Data/My_NZFFD.REC2.Diad.EF.Rdata") #load data
species = c("angdie","angaus")[1] #Species to model
covariates = c("RRF_sel_angdie", "RRF_sel_angaus")[1] #covariates to use

#CSV of covariates to use
diad.preds <- read.csv("D:/Masters/RF R stuff/Fish predictor list to use for RandForest models.csv")
Xvars <- diad.preds$predictors[which(diad.preds$diadromous == "T")] #predictors considered
rm(diad.preds) #remove as we no longer need this
Xvars <- as.character(Xvars) #set as a character

Xvars<-Xvars[-1] #fishmeth isn't needed as a covariate as we are only using EF data

table(NZFFD.REC2.Diad.EF$year, NZFFD.REC2.Diad.EF[,species])

############################ Settings ##############################################
Data_set=paste("NZFFD.REC2.Diad.EF - ", species) #set the data set
Version=get_latest_version( package="VAST" ) #version of VAST to use

#Spatial Settings - Need to find optimum settings
Method = "Mesh"
grid_size_km=25 #the distance between grid cells for the 2D AR1 grid
n_x=400 #number of knots with bias correction
#Kmeans object for determining the location for a set of knots for approximating spatial variation
Kmeans_Config=list("randomseed"=1, "nstart"=100, "iter.max"=1000)

#controls number of spatial and spatio-temporal factors used for each component
FieldConfig=c(Omega1=1, Epsilon1=1, Omega2=0, Epsilon2=0)
# Turn off annual variation in the intercept for positive-catch rates (which we'll ignore anyway)
RhoConfig=c(Beta1=2, Beta2=3, Epsilon1=0, Epsilon2=0) #Beta2 is a constant intercept and Beta1 is a random
      walk
OverdispersionConfig=c(Delta1=0, Delta2=0) #Controls the number of spatial and spatio-temporal factors for
      the vessel effects
# Logit-link for encounter probability (positive catch rate distribution doesn't matter)
ObsModel=c(2,0)
#Control Output
Options =  c("SD_site_density"=0, "SD_site_logdensity"=0, "Calculate_Range"=1, "Calculate_evenness"=0,
"Calculate_effective_area"=1, "Calculate_Cov_SE"=0, 'Calculate_Synchrony'=0, 'Calculate_Coherence'=0)

Use_REML = TRUE #use restricted maximum likelihood
```

```

strata.limits<-data.frame(STRATA = "All_areas") #Perhaps this can be changed to ESA's later?? - stratification settings

#Region
Region="Other"

bias.cor <- "TRUE"

##Save settings
#Location for saving files
if(species == "angdie"){
DateFile=paste0(getwd(), '/VAST_EF_Eels_output_angdie/')
}
if(species == "angaus"){
DateFile=paste0(getwd(), '/VAST_EF_Eels_output_angaus/')
}

dir.create(DateFile)
#save settings for later reference
Record = ThorsonUtilities::bundlelist(c("Data_set", "Version", "species", "covariates", "Method", "grid_size_km", "n_x",
"FieldConfig", "RhoConfig","OverdispersionConfig", "ObsModel", "Kmeans_Config",
"bias.cor"))
save(Record, file = file.path(DateFile, "Record.RData"))
capture.output(Record, file = paste0(DateFile, "Record.txt"))


############################ Editing data ##############################################
#Data for longfin eel catch
Data_Geostat=data.frame(Lon=NZFFD.REC2.Diad.EF[,"long"],Lat=NZFFD.REC2.Diad.EF[,"lat"], Year=NZFFD.REC2.Diad.EF[,'year'],
Vessel="missing",Catch_KG=as.numeric(NZFFD.REC2.Diad.EF[,species]), Gear=NZFFD.REC2.Diad.EF$org)
set.seed(22)
Data_Geostat[,'Catch_KG'] = Data_Geostat[,'Catch_KG'] * exp(1e-3*rnorm(nrow(Data_Geostat)))
# Add 'empty' area_swept measure
Data_Geostat = cbind( Data_Geostat, "AreaSwept_km2"=1)
Data_Geostat = cbind(Data_Geostat, "PredTF_i"=0) #use this data in the likelihood

Cov_ep = as.matrix(NZFFD.REC2.Diad.EF[,Xvars]) #matrix of the covariates

if(covariates=="RRF_sel_angdie" & species=="angdie"){ #covariates for angdie
diad.gini = read.csv("D:/Masters/RRF Model/Gini_scores.csv") #covariates selected by the RRF to use
dontinclude_covs = diad.gini[diad.gini[,species] == 0 , 1] #the variables not to include
}
if(covariates=="RRF_sel_angaus" & species=="angaus"){
diad.gini = read.csv("D:/Masters/RRF Model/Gini_scores.csv") #covariates selected by the RRF to use
dontinclude_covs = diad.gini[diad.gini[,species] == 0 , 1] #the variables not to include
}

dontinclude_covs = match(dontinclude_covs, colnames(Cov_ep)) #match with Cov_ep
Cov_ep = Cov_ep[,-dontinclude_covs] #Disregard

##Final data set
pander::pandoc.table( Data_Geostat[1:6,], digits=6 ) #table of the first 6 observations


## PART 2 - Establish VAST objects
## ------------------------------
##
## Build Extrapolation information, spatial information, density covariates, catchability covariates
## and bundle all together.

#We generate a grid for extrapolation for a given region
```

```
Extrapolation_List = SpatialDeltaGLMM::Prepare_Extrapolation_Data_Fn(Region=Region,
strata.limits=strata.limits,
observations_LL =
Data_Geostat[,c("Lat","Lon")],
grid_in_UTM=TRUE,
maximum_distance_from_sample=15)

#bundle together the spatial information into a list
Spatial_List = SpatialDeltaGLMM::Spatial_Information_Fn(grid_size_km=grid_size_km, n_x=n_x,
Method=Method, Lon=Data_Geostat[,'Lon'],
Lat=Data_Geostat[,'Lat'],
Extrapolation_List=Extrapolation_List,
randomseed=Kmeans_Config[["randomseed"]],
nstart=Kmeans_Config[["nstart"]],
iter.max=Kmeans_Config[["iter.max"]],
DirPath=DateFile, Save_Results=FALSE )

# Add knots to Data_Geostat - used for spatial prediction
Data_Geostat=cbind(Data_Geostat, knot_i=Spatial_List$knot_i)

#Build covariate matrix and Gear design matrix
X_xtp = format_covariates(Lat_e = Data_Geostat$Lat , t_e = Data_Geostat$Year ,
Lon_e = Data_Geostat$Lon ,Cov_ep = Cov_ep, Extrapolation_List = Extrapolation_List,
Spatial_List = Spatial_List, na.omit = "time-average")

#Design matrix for the gear effects (organisation sampling) offset from NIWA (the most common sampler)
Q_ik = ThorsonUtilities::vector_to_design_matrix( Data_Geostat[,'Gear'] )
Q_ik = Q_ik[, -which(colnames(Q_ik) %in% "niwa")]


#Firstly, we build a list of data inputs used for parameter estimation, Data_Fn does this
#in built "dummy observations", excluding Q_ik (org) at this point
TmbData = VAST::Data_Fn("Version"=Version, "FieldConfig"=FieldConfig,
"OverdispersionConfig"=OverdispersionConfig, "RhoConfig"=RhoConfig,
"ObsModel"=ObsModel, "c_i"=rep(0,nrow(Data_Geostat)),
"b_i"=Data_Geostat[,'Catch_KG'], "a_i"=Data_Geostat[,'AreaSwept_km2'],
"v_i"=as.numeric(Data_Geostat[,'Vessel'])-1,
"s_i"=Data_Geostat[,'knot_i']-1, "t_i"=Data_Geostat[,'Year'],
"a_xl"=Spatial_List$a_xl, "MeshList"=Spatial_List$MeshList,
"GridList"=Spatial_List$GridList, "Method"=Spatial_List$Method,
"Options"=Options, "X_xtp"=X_xtp$Cov_xtp, "Q_ik"=Q_ik,
Aniso = 1, PredTF_i=Data_Geostat$PredTF_i)


#Builds the TMB object
TmbList = VAST::Build_TMB_Fn("TmbData"=TmbData, "RunDir"=DateFile, "Version"=Version,
"RhoConfig"=RhoConfig, "loc_x"=Spatial_List$loc_x,
"Method"=Method, "Use_REML"=Use_REML )

Obj = TmbList[["Obj"]] #Extract TMB object


## PART 3 - Run model
## -------------------------------
##

##Estimate fixed effects and predict random effects
Opt = TMBhelper::Optimize( obj=Obj, lower=TmbList[["Lower"]], upper=TmbList[["Upper"]], getsd=TRUE,
savedir=DateFile, bias.correct=bias.cor, newtonsteps=3,
control = list(eval.max = 100000, iter.max = 100000 ,trace = TRUE))

#Save the results
Report = Obj$report()
Save = list("Opt"=Opt, "Report"=Report, "ParHat"=Obj$env$parList(Opt$par), "TmbData"=TmbData)
```

```
save(Save, file=paste0(DateFile,"Save.RData"))

#Parameter results
Save$Opt$diagnostics[,c(1,4,6)]

end1=Sys.time()
time<-end1-start1 ; time
```

The following gives the R code for building the VAST multi-species model.

```
start1=Sys.time() #measure how long it takes
setwd("/am/courtenay/home1/charslanth/Masters/VAST R stuff")
library("devtools")
library("Matrix")
library(TMB,lib.loc = .libPaths()[2])
library(INLA)
library(SpatialDeltaGLMM)
library(VAST,lib.loc = .libPaths()[2])
library(maps) #to visualise
library(ROCR) #auc function
library(cvAUC) #auc CI's
library(sperrorest) #K-means spatial partitioning function


## PART 1 - Load data and edit settings

############################ Load and edit data ##############################################
load("/am/courtenay/home1/charslanth/Masters/Data/My_NZFFD.REC2.Diad.EF.Rdata")
covariates = c("RRF_sel_angdie", "RRF_sel_angaus")[1] #covariates to use
diad.preds <- read.csv("/am/courtenay/home1/charslanth/Masters/RF R stuff/Fish predictor list to use for RandForest models.csv")
Xvars <- diad.preds$predictors[which(diad.preds$diadromous == "T")] #use these predictors
rm(diad.preds) #remove as we no longer need this
Xvars <- as.character(Xvars) #set as a character
Xvars<-Xvars[-1] #fishmeth isn't needed as a covariate as we are only using EF data

############################ Settings ##############################################
Data_set="NZFFD.REC2.Diad.EF" #set the data set
Version="VAST_v4_4_0" #version of VAST to use

#Spatial Settings - Need to find optimum settings
Method = "Mesh"
grid_size_km=25 #the distance between grid cells for the 2D AR1 grid
n_x=400 #number of knots to use

#controls number of spatial and spatio-temporal factors used for each component
FieldConfig=c(Omega1=2, Epsilon1=2, Omega2=0, Epsilon2=0)
# Turn off annual variation in the intercept for positive-catch rates (which we'll ignore anyway)
RhoConfig=c(Beta1=2, Beta2=3, Epsilon1=0, Epsilon2=0) #Beta2 is a constant intercept and Beta1 is a random walk
OverdispersionConfig=c(Delta1=0, Delta2=0) #Controls the number of spatial and spatio-temporal factors for the vessel effects
# Logit-link for encounter probability (positive catch rate distribution doesn't matter)
ObsModel=c(2,0)
#Control Output
Options =  c("SD_site_density"=0, "SD_site_logdensity"=0, "Calculate_Range"=1, "Calculate_evenness"=0,
"Calculate_effective_area"=1, "Calculate_Cov_SE"=0, 'Calculate_Synchrony'=0, 'Calculate_Coherence'=0)

Use_REML = TRUE #use restricted maximum likelihood

strata.limits<-data.frame(STRATA = "All_areas") #Perhaps this can be changed to ESA's later?? - stratification settings

#Region
Region="Other"

bias.cor <- "TRUE"

##Save settings
#Location for saving files
```

```
DateFile=paste0(getwd(), '/VAST_EF_Eels_output_MS/')
dir.create(DateFile)
#save settings for later reference
Record = ThorsonUtilities::bundlelist(c("Data_set", "Version", "covariates" ,"Method", "grid_size_km", "n_x"
      , "FieldConfig", "RhoConfig",
"OverdispersionConfig", "ObsModel", "bias.cor"))
save(Record, file = file.path(DateFile, "Record.RData"))
capture.output(Record, file = paste0(DateFile, "Record.txt"))


############################ Editing data ##############################################
#Data for longfin eel catch
Data_angdie=data.frame(spp = "angdie", Lon=NZFFD.REC2.Diad.EF[,"long"],Lat=NZFFD.REC2.Diad.EF[,"lat"],
Year=NZFFD.REC2.Diad.EF[,'year'], Vessel="missing",Catch_KG=as.numeric(NZFFD.REC2.Diad.EF[,"angdie"]),
Gear=NZFFD.REC2.Diad.EF$org)
set.seed(22)
Data_angdie[,'Catch_KG'] = Data_angdie[,'Catch_KG'] * exp(1e-3*rnorm(nrow(Data_angdie)))

#Data for shortfin eel catch
Data_angaus=data.frame(spp = "angaus", Lon=NZFFD.REC2.Diad.EF[,"long"],Lat=NZFFD.REC2.Diad.EF[,"lat"],
Year=NZFFD.REC2.Diad.EF[,'year'], Vessel="missing",Catch_KG=as.numeric(NZFFD.REC2.Diad.EF[,"angaus"]),
Gear=NZFFD.REC2.Diad.EF$org)
set.seed(22)
Data_angaus[,'Catch_KG'] = Data_angaus[,'Catch_KG'] * exp(1e-3*rnorm(nrow(Data_angaus)))

Data_Geostat <- rbind(Data_angdie, Data_angaus) #bind longfin and shortfin data
rm(Data_angdie) ; rm(Data_angaus) #remove

# Add 'empty' area_swept measure
Data_Geostat = cbind( Data_Geostat, "AreaSwept_km2"=1)
Data_Geostat = cbind(Data_Geostat, "PredTF_i"=0) #use this data in the likelihood

Cov_ep = as.matrix(NZFFD.REC2.Diad.EF[,Xvars]) #matrix of the covariates
diad.gini = read.csv("/am/courtenay/home1/charslanth/Masters/RRF Model/Gini_scores.csv")

#I will use the variables of the LONGFIN eel RRF for the MS model
if(covariates=="RRF_sel_angdie"){ #covariates for angdie
dontinclude_covs = diad.gini[diad.gini[,"angdie"] == 0 , 1] #the variables not to include
}
if(covariates=="RRF_sel_angaus"){
dontinclude_covs = diad.gini[diad.gini[,"angaus"] == 0 , 1] #the variables not to include
}

dontinclude_covs = match(dontinclude_covs, colnames(Cov_ep)) #match with Cov_ep
Cov_ep = Cov_ep[,-dontinclude_covs] #Disregard

##Final data set
pander::pandoc.table( Data_Geostat[1:6,], digits=6 ) #table of the first 6 observations


## PART 2 - Establish VAST objects

#We generate a grid for extrapolation for a given region
Extrapolation_List = make_extrapolation_info(Region=Region, strata.limits=strata.limits,
observations_LL = Data_Geostat[,c("Lat","Lon")],
grid_in_UTM=TRUE, maximum_distance_from_sample=15)

#Kmeans object for determining the location for a set of knots for approximating spatial variation
Kmeans_Config=list("randomseed"=1, "nstart"=100, "iter.max"=1000)
#bundle together the spatial information into a list
Spatial_List = make_spatial_info(n_x = n_x, Lon=Data_Geostat[,'Lon'], Lat=Data_Geostat[,'Lat'],
Extrapolation_List=Extrapolation_List, Method=Method, grid_size_km=grid_size_km,
randomseed=Kmeans_Config[["randomseed"]], nstart=Kmeans_Config[["nstart"]],
iter.max=Kmeans_Config[["iter.max"]], DirPath=DateFile, Save_Results=FALSE )
```

```
# Add knots to Data_Geostat - used for spatial prediction
Data_Geostat=cbind(Data_Geostat, knot_i=Spatial_List$knot_i)

Whichcovs=c(1:nrow(NZFFD.REC2.Diad.EF)) #Covariates are identical for each species so no need to repeat for each
#Build covariate matrix and Gear design matrix
X_xtp = format_covariates(Lat_e = Data_Geostat$Lat[Whichcovs] , t_e = Data_Geostat$Year[Whichcovs] ,
Lon_e = Data_Geostat$Lon[Whichcovs] ,Cov_ep = Cov_ep, Extrapolation_List = Extrapolation_List,
Spatial_List = Spatial_List, na.omit = "time-average")

#Design matrix for the gear effects (organisation sampling) offset from NIWA (the most common sampler)
Q_ik = ThorsonUtilities::vector_to_design_matrix( Data_Geostat[,'Gear'] )
Q_ik = Q_ik[, -which(colnames(Q_ik) %in% "niwa")]

#Firstly, we build a list of data inputs used for parameter estimation, Data_Fn does this
#in built "dummy observations", excluding Q_ik (org) at this point
TmbData = VAST::Data_Fn("Version"=Version, "FieldConfig"=FieldConfig,
"OverdispersionConfig"=OverdispersionConfig, "RhoConfig"=RhoConfig,
"ObsModel"=ObsModel, "c_i"=as.numeric(Data_Geostat[,'spp'])-1,
"b_i"=Data_Geostat[,'Catch_KG'], "a_i"=Data_Geostat[,'AreaSwept_km2'],
"v_i"=as.numeric(Data_Geostat[,'Vessel'])-1,
"s_i"=Data_Geostat[,'knot_i']-1, "t_i"=Data_Geostat[,'Year'],
"a_xl"=Spatial_List$a_xl, "MeshList"=Spatial_List$MeshList,
"GridList"=Spatial_List$GridList, "Method"=Spatial_List$Method,
"Options"=Options, "X_xtp"=X_xtp$Cov_xtp, "Q_ik"=Q_ik,
Aniso = 1, PredTF_i=Data_Geostat$PredTF_i)


#Builds the TMB object
TmbList = VAST::Build_TMB_Fn("TmbData"=TmbData, "RunDir"=DateFile, "Version"=Version,
"RhoConfig"=RhoConfig, "loc_x"=Spatial_List$loc_x,
"Method"=Method, "Use_REML"=Use_REML )

Obj = TmbList[["Obj"]] #; beep(5)#Extract TMB object


## PART 3 - Run model
## ------------------------------
##
## Run and optimise the model

##Estimate fixed effects and predict random effects
Opt = TMBhelper::Optimize(obj=Obj, lower=TmbList[["Lower"]], upper=TmbList[["Upper"]], getsd=TRUE,
savedir=DateFile, bias.correct=bias.cor, newtonsteps=3,
control = list(eval.max = 100000, iter.max = 100000 ,trace = TRUE),
bias.correct.control=list(sd=FALSE, split=NULL, nsplit=1, vars_to_correct="Index_cyl"))

#Save the results
Report = Obj$report()
Save = list("Opt"=Opt, "Report"=Report, "ParHat"=Obj$env$parList(Opt$par), "TmbData"=TmbData)
save(Save, file=paste0(DateFile,"Save.RData"))

#Parameter results
Save$Opt$diagnostics[,c(1,4,6)]

end=Sys.time()
time<-end-start1 ; time
```

The following gives the R code for building the GRaF longfin eel and shortfin eel models.

```
setwd("D:/Masters/GRaF Model")
####  PACKAGES
library(devtools)
library(GRaF) #install GRaF from github (the version from goldingn's repo at least)
library(sperrorest) #K-means spatial partitioning function
library(ROCR) #auc function
library(cvAUC) #auc CI's

##############################################################################################
############################## Load and edit data ############################################
##############################################################################################

load(file = "D:/Masters/Data/My_NZFFD.REC2.Diad.EF.Rdata") #load data
species <- "angdie"
covariates = c("RRF_sel_angdie", "RRF_sel_angaus")[1] #covariates to use

diad.preds <- read.csv("D:/Masters/RF R stuff/Fish predictor list to use for RandForest models.csv") #all
      the predictor variables
Xvars <- diad.preds$predictors[which(diad.preds$diadromous == "T")] #use these predictors
rm(diad.preds) #remove as we no longer need this
Xvars <- as.character(Xvars) #set as a character

Xvars<-Xvars[-1] #fishmeth isn't needed as a covariate as we are only using EF data

#covariates to use
diad.preds <- read.csv("D:/Masters/RF R stuff/Fish predictor list to use for RandForest models.csv")
diad.gini = read.csv("D:/Masters/RRF Model/Gini_scores.csv") #covariates selected by the RRF to use
Model_covs = as.vector(diad.gini[diad.gini[,species] > 0 , 1]) #the variables to include
Model_covs <- c(Model_covs, "year") #use year as a covariate

NZFFD.REC2.Diad.EF[[species]] <- as.factor(NZFFD.REC2.Diad.EF[[species]]) #specify the species as a factor
levels(NZFFD.REC2.Diad.EF[[species]]) <- c("FALSE", "TRUE") #ensure the levels arestr F/T
pa <- as.numeric(as.logical(NZFFD.REC2.Diad.EF[,species])) #pa data


##############################################################################################
#################################### Prior ###################################################
myprior <- c("Informative_RRF", "Uninformative")[2]

##############################################################################################
################################### Record setting ###########################################
Record_list <- list() #empty list to record settings
Record_list$species <- species
Record_list$covariates <- covariates
Record_list$prior <- myprior
capture.output(Record_list, file = paste0("Record_GPSDM_",species,".txt"))

##############################################################################################
############################# Build longfin GRaF model #######################################
##############################################################################################
graf_model <- graf(y=pa, x=NZFFD.REC2.Diad.EF[,Model_covs], opt.l = TRUE, prior = RRF_pred,
verbose = TRUE) #with uninformative prior
```

Bibliography

Beullens, K., Eding, E. H., Gilson, P., Ollevier, F., Komen, J., & Richter, C. J. J. (1997). Gonadal differentiation, intersexuality and sex ratios of European eel (Anguilla anguill L.) maintained in captivity. *Aquaculture*, **153**(1), 135–150. https://doi.org/10.1016/S0044-8486(97)00018-5.

Blangiardo, M. & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. Chichester, United Kingdom: John Wiley & Sons.

Booker, D. & Graynoth, E. (2013). Relative influence of local and landscape-scale features on the density and habitat preferences of longfin and shortfin eels. *New Zealand Journal of Marine and Freshwater Research*, **47**(1), 1–20. https://doi.org/10.1080/00288330.2012.714389.

Booker, D. J. & Woods, R. A. (2014). Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. *Journal of Hydrology*, **508**, 227–239. https://doi.org/10.1016/j.jhydrol.2013.11.007.

Brenning, A. (2005). Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Science*, **5**(6), 853–862.

Carlin, B. P. & Louis, T. A. (2008). *Bayesian methods for data analysis*. Boca Raton, FL: CRC Press, 3 edition.

Chisnall, B. L., Martin, M. L., & Hicks, B. J. (2003). Effect of harvest on size, abundance, and production of freshwater eels anguilla australis and a. dieffenbachii in a new zealand stream. In *American Fisheries Society Symposium* (pp. 177–190).: American Fisheries Society.

Colombo, G. & Grandidr, G. (1996). Histological study of the development and sex differentiation of the gonad in the European eel. *Journal of Fish Biology*, **48**(3), 493–512. https://doi.org/10.1111/j.1095-8649.1996.tb01443.x.

Crow, S., Booker, D., Sykes, J., Unwin, M., & Shankar, U. (2014). *Predicting distributions of New Zealand freshwater fishes*. Technical Report CHC2014-145, National Institute of Water and Atmospheric Research.

Crow, S., Snelder, T., Jellyman, P., Greenwood, M., Booker, D., & Dunn, A. (2016). *Temporal trends in the relative abundance of New Zealand freshwater fishes: Analysis of New Zealand freshwater fish database records*. Technical Report CHC2016-049, National Institute of Water and Atmospheric Research.

Crow, S. K., Booker, D. J., & Snelder, T. H. (2013). Contrasting influence of flow regime on freshwater fishes displaying diadromous and nondiadromous life histories. *Ecology of Freshwater Fish*, **22**(1), 82–94. https://doi.org/10.1111/eff.12004.

Davey, A. J. & Jellyman, D. J. (2005). Sex determination in freshwater eels and management options for manipulation of sex. *Reviews in fish biology and fisheries*, **15**(1), 37–52.

De'ath, G. & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, **81**(11), 3178–3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

Deng, H. (2013). Guided random forest in the rrf package. *arXiv:1306.0237*, .

Deng, H. & Runger, G. (2012). Feature selection via regularized trees. In *Neural Networks (IJCNN), The 2012 International Joint Conference on* (pp. 1–8).: IEEE.

Deng, H. & Runger, G. (2013). Gene selection with guided regularized random forest. *Pattern Recognition*, **46**(12), 3483–3489. https://doi.org/10.1016/j.patcog.2013.05.018.

Díaz-Uriarte, R. & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, **7**(1), 3. https://doi.org/10.1186/1471-2105-7-3.

Dunn, A., Beentjes, M. P., & Graynoth, E. (2009). Preliminary investigations into the feasibility of assessment models for New Zealand longfin eels (anguilla dieffenbachii). *New Zealand Fisheries Assessment Report*, **30**, 42.

Elith, J. & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, **40**, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, **27**(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Fu, D., Beentjes, M. P., & Dunn, A. (2012). Further investigations into the feasibility of assessment models for New Zealand longfin eels (anguilla dieffenbachii). Final Research Report for Ministry of Fisheries

Project EEL200702. 77 p. (Unpublished report held by Ministry of Fisheries, Wellington.).

Glova, G. J., Jellyman, D. J., & Bonnett, M. L. (1998). Factors associated with the distribution and habitat of eels (anguilla spp.) in three New Zealand lowland streams. *New Zealand Journal of Marine and Freshwater Research*, **32**(2), 255–269. https://doi.org/10.1080/00288330.1998.9516824.

Gluckman, P. (2017). *New Zealands fresh waters: Values, state, trends and human impacts*. Technical report, Office of the Prime Ministers Chief Science Advisor. Retrieved from http://www.pmcsa.org.nz/wp-content/uploads/PMCSA-Freshwater-Report.pdf.

Golding, N. (2017). *GRaF: Species distribution modelling using gaussian processes*. R package version 0.1-15.

Golding, N. & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, **7**(5), 598–608. https://doi.org/10.1111/2041-210X.12523.

Golding, N., Rogers, D. J., Purse, B. V., & Nunn, M. A. (2013). *Mapping and understanding the distributions of potential vector mosquitoes in the UK: New methods and applications*. PhD thesis, Oxford University, UK. Retrieved from https://ndownloader.figshare.com/files/1146673.

Graynoth, E. & Booker, D. (2009). *Biomass of longfin eels in medium to large rivers*. Technical Report 2009/44, New Zealand Fisheries Assessment Report. Retrieved from http://docs.niwa.co.nz/library/public/FAR2009-44.pdf.

Graynoth, E., Francis, R. I. C. C., & Jellyman, D. J. (2008a). Factors influencing juvenile eel (Anguilla spp.) survival in lowland New Zealand streams. *New Zealand Journal of Marine and Freshwater Research*, **42**(2), 153–172. https://doi.org/10.1080/00288330809509945.

Graynoth, E., Jellyman, D. J., & Bonnett, M. (2008b). *Spawning escapement of female longfin eels*. Technical Report 2008/7, New Zealand Fisheries Assessment Report. Retrieved from http://docs.niwa.co.nz/library/public/FAR2008-07.pdf.

Graynoth, E. & Taylor, M. (2005). Influence of different rations and water temperatures on the growth rates of shortfinned eels and longfinned eels. *Journal of Fish Biology*, **57**(3), 681–699. https://doi.org/10.1111/j.1095-8649.2000.tb00268.x.

Grüss, A., Thorson, J. T., Babcock, E. A., Tarnecki, J. H., & editor: James Watson, H. (2017). Producing distribution maps for informing ecosystem-based fisheries management using a comprehensive survey database and spatio-temporal models. *ICES Journal of Marine Science*, **75**(1), 158–177. https://doi.org/10.1093/icesjms/fsx120.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, **69**(346), 383–393.

Hartigan, J. A. (1975). *Clustering algoritms*. New York, NY: John Wiley & Sons.

Hastie, T., Friedman, J. H., & Tibshirani, R. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer Science+Business Media, 2nd edition.

Hocking, D. J., Thorson, J. T., O'Neil, K., & Letcher, B. H. (2018). A geostatistical state-space model of animal densities for stream networks. *Ecological Applications*, . https://doi.org/10.1002/eap.1767.

Hoyle, S. D. (2016). *Feasibility of longfin eel stock assessment*. Technical Report 2016/29, New Zealand Fisheries Assessment Report. Retrieved from https://www.researchgate.net/profile/Simon_Hoyle/publication/307981164_Feasibility_of_longfin_eel_stock_assessment/links/

57d5e2f708ae601b39aa70cf/Feasibility-of-longfin-eel-stock-assessment.pdf.

Hoyle, S. D. & Jellyman, D. J. (2002). Longfin eels need reserves: Modelling the effects of commercial harvest on stocks of New Zealand eels. *Marine and Freshwater Research*, **53**(5), 887–895. https://doi.org/10.1071/MF00020.

Jellyman, D. J. (2003). The distribution and biology of the south pacific species of anguilla. In *Eel Biology* (pp. 275–292). Tokyo, Japan: Springer.

Jellyman, D. J. (2012). *The status of longfin eels in New Zealand-an overview of stocks and harvest*. Technical Report CHC2012-006, National Institute of Water and Atmospheric Research. Retrieved from https://www.pce.parliament.nz/media/1237/jellyman-report-final2.pdf.

Jellyman, D. J., Bonnett, M. L., Sykes, J. R. E., & Johnstone, P. (2003). Contrasting use of daytime habitat by two species of freshwater eel anguilla spp. in New Zealand rivers. In *American Fisheries Society Symposium* (pp. 63–78).: American Fisheries Society.

Jowett, I. G. & Richardson, J. (1996). Distribution and abundance of freshwater fish in New Zealand rivers. *New Zealand journal of marine and freshwater research*, **30**(2), 239–255. https://doi.org/10.1080/00288330.1996.9516712.

Jowett, I. G. & Richardson, J. (2003). Fish communities in New Zealand rivers and their relationship to environmental variables. *New Zealand Journal of Marine and Freshwater Research*, **37**(2), 347–366. https://doi.org/10.1080/00288330.2003.9517172.

Joy, M., David, B., & Lake, M. (2013). *New Zealand freshwater fish sampling protocols*. Palmerston North, New Zealand: Massey University.

Joy, M. K. & Death, R. G. (2004). Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. *Freshwater Biology*, **49**(8), 1036–1052. https://doi.org/10.1111/j.1365-2427.2004.01248.x.

Kass, R. E. & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, **84**(407), 717–726. https://doi.org/10.1080/01621459.1989.10478825.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. (2015). TMB: automatic differentiation and Laplace approximation. *Journal of Statistical Software*, **70**(5), 1–21. https://arxiv.org/ct?url=https

Leathwick, J. R., Elith, J., Chadderton, W. L., Rowe, D., & Hastie, T. (2008a). Dispersal, disturbance and the contrasting biogeographies of New Zealands diadromous and non-diadromous fish species. *Journal of Biogeography*, **35**(8), 1481–1497. https://doi.org/10.1111/j.1365-2699.2008.01887.x.

Leathwick, J. R., Julian, K., Elith, J., & Rowe, D. (2008b). *Predicting the distributions of freshwater fish species for all New Zealands rivers and streams*. Technical Report HAM2008-005, National Institute of Water and Atmospheric Research. Retrieved from https://www.niwa.co.nz/sites/niwa.co.nz/files/29_nativefishpredictionmaps.pdf.

LeDell, E., Petersen, M., & van der Laan, M. (2015). Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electronic journal of statistics*, **9**(1), 1583. https://dx.doi.org/10.1214

Legendre, P. (1990). Quantitative methods and biogeographic analysis. In *Evolutionary biogeography of the marine algae of the North Atlantic* (pp. 9–34). Springer.

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

Martínez-Minaya, J., Cameletti, M., Conesa, D., & Pennino, M. G. (2018). Species distribution modeling: a statistical review with focus in spatio-temporal issues. *Stochastic Environmental Research and Risk Assessment*, , 1–18.

McDowall, R. M. (1990). *New Zealand freshwater fishes: A natural history and guide*. Auckland, New Zealand: Heinemann Reed.

Ministry of Primary Industries (2014). Fisheries assessment plenary, May 2014: Stock assessments and stock status. Wellington, New Zealand: Compiled by the Fisheries Science Group, Ministry for Primary Industries.

Mosteller, F. & Tukey, J. W. (1977). Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, .

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, **41**(5), 673–690.

Picard, R. R. & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, **79**(387), 575–583.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., & Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, **31**(10), 2001–2019. https://doi.org/10.1080/13658816.2017.1346255.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Retrieved from http://www.gaussianprocess.org/gpml/chapters/RW.pdf.

Richardson, J. (2005). New Zealand freshwater fish database user guide. *NIWA Client Rep HAM2005-033*, , 1–28.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, **40**(8), 913–929. https://doi.org/10.1111/ecog.02881.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, **71**(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x.

Ruß, G. & Brenning, A. (2010). Data mining in precision agriculture: management of spatial information. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 350–359). Berlin, Heidelberg: Springer.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**(20), 3940–3941. https://doi.org/10.1093/bioinformatics/bti623.

Skaug, H. J. & Fournier, D. A. (2006). Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics & Data Analysis*, **51**(2), 699–709. https://doi.org/10.1016/j.csda.2006.03.005.

Thorson, J. (2018). *VAST user manual*. Retrieved from https://github.com/James-Thorson/VAST/blob/master/examples/VAST_user_manual.docx.

Thorson, J. T. (2019). Guidance for decisions using the vector autoregressive spatio-temporal (vast) package in stock, ecosystem, habitat and climate assessments. *Fisheries Research*, **210**, 143–161. https://doi.org/10.1016/j.fishres.2018.10.013.

Thorson, J. T. & Barnett, L. A. K. (2017). Comparing estimates of abundance trends and distribution shifts using single-and multispecies models of fishes and biogenic habitat. *ICES Journal of Marine Science*, **74**(5), 1311–1321. https://doi.org/10.1093/icesjms/fsw193.

Thorson, J. T. & Kristensen, K. (2016). Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples. *Fisheries Research*, **175**, 66–74. https://doi.org/10.1016/j.fishres.2015.11.016.

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

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, **46**, 234–240.

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