NEW APPROACHES TO MODELLING DRIVERS OF SPECIES DISTRIBUTION AND ABUNDANCE IN THE SOUTHERN OCEAN

Lisa-Marie K. Harrison (BSc Hons)

Marine Predator Research Group

Department of Biological Sciences

Faculty of Science and Engineering

Macquarie University

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

2017

MACQUARIE University
SYDNEY·AUSTRALIA

Cover image courtesy of Professor Steve Nicol, University of Tasmania

Table of Contents

Summary

Inferring ecological patterns from marine survey data is difficult due to the large spatial and temporal scales at which processes operate and the challenges associated with collecting comprehensive and balanced survey data. In this thesis I use large scale survey data and cutting edge modelling techniques to examine the drivers of species distribution in the Southern Ocean at three trophic levels – primary producers, grazers and top predators. I develop a model to predict phytoplankton abundance in a 3D environment from temperature, salinity and depth. This framework is widely applicable to other marine settings regardless of their survey design and provides a robust method for dealing with complex data sets. An important grazer on phytoplankton, Antarctic krill (*Euphausia superba*), has previously been regarded as passively drifting with large scale current systems. I provide quantitative evidence that they actively swim, demonstrating that krill consistently aggregate around resources over an immense survey area spanning 1.3 million km^2. Krill distribution is patchy, and predators must locate these dynamic swarms across vast expanses of ocean. Islands may provide predictable and reliable feeding areas due to the Island Mass Effect. I find that krill swarms at the Balleny Islands, a Southern Ocean archipelago, are three times more numerous than in the adjacent open ocean, and are also denser and more compact. Around the islands, humpback whales (*Megaptera novaeangliae*) aggregate in areas of high productivity, medium krill density and waters greater than 350m deep. Two chapters of this thesis required manual processing of active acoustics data for detecting krill, which is time consuming and suffers from a lack of reproducibility. To automate this process, I developed an R package which drastically reduces processing time and is useful for any scientists using acoustic data. This thesis fills knowledge gaps about the mechanisms structuring the distribution of animals in the Southern Ocean and the statistical methods and software library developed are applicable to many other problems arising in complex environments.

Declaration

I, Lisa-Marie Katarina Harrison, certify that this thesis entitled "New approaches to modelling drivers of species distribution and abundance in the Southern Ocean" is an original piece of work and has not been submitted in whole or in part for a higher degree at any institution other than Macquarie University. This work was undertaken at the Marine Predator Research Group at Macquarie University under the supervision of Professor Robert Harcourt (Macquarie University), Dr Martin Cox (Australian Antarctic Division), Dr Steven Candy (SCandy Statistical Modelling) and Assistant Professor Leslie New (Washington State University).

This thesis was prepared and written by me. All assistance in the preparation of this thesis has been acknowledged and all references and sources of information used in this thesis are listed within. Ethics approval was not required for this thesis.

Lisa-Marie K. Harrison

17th of March 2017

Acknowledgments

I have so many people to thank for their assistance and support during my candidature. Firstly, I would like to thank my supervisors Robert Harcourt, Martin Cox, Steve Candy and Leslie New. They have not only guided and encouraged me over the past three and a half years, but have made it a fun and incredibly rewarding journey. Thank you to Leslie for hosting me for a lab visit at Washington State University and to Martin for hosting me many times at the Australian Antarctic Division.

Thank you to my lab group, the Marine Predator Research Group: Gemma, Vanessa, Ben, Nicolette, Kaja, Dustin, Ian, Justin, Adam, Monique, Marcus, Kim, Marine, Yuna, Kate, Alex, Paolo, Sam and Sally. I've enjoyed sharing an office and many lunch times with you, along with many dinners at our favourite restaurant! Your advice in our lab meetings has always made me feel more confident about my papers and presentations and I appreciate the feedback and encouragement that you have given me over the years.

As a part of this thesis, I undertook many aerial surveys by helicopter off coastal New South Wales, looking at the distribution and habitat use of marine megafauna. Unfortunately, this chapter did not end up making it in to my thesis, however a technical report detailing the many months of work that I spent on this project has been included in Appendix D. I enjoyed the fieldwork and would like to thank Vic Peddemors of the Department of Primary Industries for the opportunity. Thank you to Brett Kiteley, Joshua Wales and Rochelle Tonkin of Touchdown Helicopters for so many safe and enjoyable flights. Thank you to Jill of Park Meadows for accommodating us in Wollongong during our fieldwork, and especially for understanding when we cancelled at short notice due to bad weather.

Thank you to my family for being so incredibly supportive from over 3000km away. To my parents, you have always encouraged me to ask questions and think outside the box and I am grateful for the support that you have given me during this thesis. Thank you to my Mum, Vesna, for all your encouragement and support. Thank you to my Dad, Ian, for proof reading my thesis chapters and offering many pieces of advice that have improved the thesis. Thank you to my sister Liana, for always being supportive and cheerful. To my Baka and Grandy, who both sadly passed away during my candidature; your work ethic and strength has always been an inspiration to me.

Lastly, thank you to my husband Caspar, whose support has been unwavering. Thank you for listening to (and improving!) every presentation that I gave over the course of my degree. You encouraged me at times when I doubted myself and have always been there to celebrate every achievement along the way with me. Thank you for putting up with my absences for many field trips and for proofreading my thesis. Living in Sydney with you has been so much fun and I look forward to our next chapter in San Francisco.

Statement of Contribution

Contributors to this thesis

Principal Supervisor	Professor Robert Harcourt Department of Biological Sciences Macquarie University Sydney, Australia
Adjunct Supervisor	Dr Martin Cox Australian Antarctic Division Kingston, Australia
Adjunct Supervisor	Dr Steven Candy SCandy Statistical Modelling Pty Ltd Hobart, Australia
Adjunct Supervisor	Assistant Professor Leslie New Department of Mathematics and Arts Washington State University Vancouver, USA
Co-author – Chapter 3	Dr Guy Williams Institute of Marine and Antarctic Sciences University of Tasmania Hobart, Australia
Co-author – Chapter 4	Dr Georg Skaret Institute of Marine Research Bergen, Norway
Co-author – Chapter 6	Dr Kimberley Goetz National Institute of Water and Atmospheric Research Wellington, New Zealand
Co-author – Appendix Technical Report	Dr Vic Peddemors New South Wales Department of Primary Industries, Sydney, Australia

Chapter Declarations

Chapter 2: Theoretical foundation for models

This literature review was undertaken and written by myself. Leslie New, Martin Cox, Steve Candy and Rob Harcourt provided comments on the chapter.

Chapter 3: Modelling spatially autocorrelated phytoplankton fluorescence around East Antarctica using linear mixed models with cubic splines

This chapter uses previously collected data from the 2006 Baseline Research on Oceanography, Krill and the Environment (BROKE-West) survey. The analysis for this chapter was carried out by myself under the guidance of Steve Candy. I wrote the manuscript, and it was edited by Steve Candy, Martin Cox, Rob Harcourt and Guy Williams.

Chapter 4: The R package *EchoviewR* for automated processing of active acoustic data using Echoview

This chapter presents an R package developed to facilitate faster data processing for the other chapters in this thesis. The programming of the R package was carried out by myself and Martin Cox. Georg Skaret tested the package for errors and bugs. I wrote the manuscript and it was revised by Martin Cox, Rob Harcourt and Georg Skaret. Echoview (Myriax) kindly provided me with a full licence for their software during the development of this package. The example data used to demonstrate the package is from the 2003 Krill Acoustic and Oceanographic Survey (KAOS). This chapter was published in *Frontiers in Marine Science* in February 2015 and the published version of the manuscript is included in Appendix A.

Chapter 5: The world's most abundant predator is not a passive drifter: Antarctic krill aggregate around food and oxygen

This chapter uses the same BROKE-West survey data as Chapter 3. The analysis for this chapter was carried out by myself with guidance from Martin Cox. I wrote the manuscript and it was revised by Martin Cox and Robert Harcourt. Leslie New and Simon Wotherspoon (University of Tasmania) gave comments on the analysis. Steve Nicol (University of Tasmania) gave comments on the manuscript. I presented the results of this chapter as an oral presentation at the XXVIIIth International Biometric Conference in Canada, July 2016, where I received positive feedback on the analysis from biometricians.

Chapter 6: A Southern Ocean archipelago enhances feeding opportunities for a krill predator

The fieldwork to collect the data used in this chapter was carried out in 2015 by Kimberly Goetz (National Institute of Water and Atmospheric Research). I carried out the analysis with guidance from Martin Cox, and wrote the manuscript, which was revised by Martin Cox, Rob Harcourt and Kimberly Goetz.

Appendix D: Aerial survey final report

The aerial survey work detailed in the final report in Appendix D was originally intended to be analysed and included as a chapter in this thesis, however due to time constraints the work was not included. The fieldwork was carried out by myself and Vic Peddemors with the helicopter operated by Brett Kiteley, Joshua Wales and Rochelle Tonkin of Touchdown Helicopters. Sally Dupont volunteered as a photographer during three field surveys and Girish Vijayaraghavan and Tim Kett assisted with data transcription. The analysis in this report was carried out by myself and the report was written by myself.

Conference presentations during candidature

Conference: XXVIIIth International Biometric Conference, Canada, July 2016. Presented talk entitled: "*A hurdle mixed model for linking Antarctic Krill presence/absence and density to phytoplankton and environmental factors"*

Awards and grants during candidature

- Macquarie University Postgraduate Research Fund travel grant, $4800, 2016
- Best Modelling Based presentation, Macquarie Biological Sciences Higher Degree Research conference, 2014
- Runner-up Best Modelling Based presentation, Macquarie University Biological Sciences Higher Degree Research conference, 2015 and 2016

Accepted publications during candidature

Harrison L-MK, Cox MJ, Skaret G and Harcourt R (2015) The R package *EchoviewR* for automated processing of active acoustic data using Echoview. *Front. Mar. Sci*. **2**:15. doi: 10.3389/fmars.2015.00015

"We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions."

- *Ronald A. Fisher (1955)*

Chapter 1: General Introduction

Authors:

Lisa-Marie K. Harrison[1]

Affiliations:

[1]Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, New South Wales, Australia

Ecology and in particular ecosystem research requires the ability to detect signals using complex and often noisy data sets. It has long been recognised that this complexity is increased when studying marine ecosystems, which are more dynamic and challenging to survey than terrestrial ecosystems (Cassie, 1956). This thesis develops and applies sophisticated analytical techniques to answer key ecological questions in the Southern Ocean. These techniques range from advances in data processing, an area becoming increasingly important in ecology (Michener and Jones, 2012), to statistical approaches that enable us to ask new questions and detect ecological signals in large and highly correlated data sets. These methods allow us to unlock the potential of the many large and expensive marine data sets that already exist - in particular active acoustic data.

The questions addressed in this thesis involve determining how key functional groups in the Southern Ocean food web are distributed and identifying the drivers behind these patterns. Key functional groups are groups with multiple links in the food web which include multiple dependent species and have a widespread distribution influenced by many drivers operating at different spatial and temporal scales (Mills et al., 1993). The processing and analytical techniques used in this thesis aim to tease apart the drivers of distribution when direct inference from the raw data is not possible as a result of unbalanced survey design, spatial autocorrelation or the spatial and temporal ranges the data were collected over.

It is necessary to understand contemporary ecosystem conditions in order to accurately predict the future effects of climate change (Shaver et al., 2000). Key functional groups, such as primary producers, grazers and top predators, are important to study because their fate will affect many other species who depend on them as a food source or for providing

other ecosystem services (Mills et al., 1993, Grimm, 1995, Douglass et al., 2008). Developing models of species-environment relationships is a necessary first step in developing predictive models into which environmental and resource variables can be input to predict change in abundance and distribution (Guisan et al., 2006). The models developed in this thesis all have the capacity for prediction using new input, or predictor, data sets. However actually modelling predictions was beyond the scope of a 3-year thesis. Prediction remains an important future direction for work in this area.

It is difficult to accurately assess species distribution and build robust predictive models from sparse data sets (Ovaskainen and Soininen, 2011). Marine data sets are often unavoidably sparse due to logistic difficulties in even getting to a study site, sampling underwater and the difficulty in using a balanced and replicated survey design over the huge scales involved in open ocean research (Lawless, 2014; pg 19). Many marine processes operate over large spatial and temporal scales that are difficult to fully sample and model, requiring analysis methods that can detect survey-wide patterns from discrete sampling locations and at the same time can incorporate the spatial and temporal dynamics of the data (Kaiser, 2011; pg 208, Godø et al., 2014). These problems are exacerbated in remote survey areas because of the limited time available for sampling and the large cost of the survey. The Southern Ocean is a prime example of a remote and harsh environment, where survey data often spans enormous areas and where both sampling effort and survey design vary as a result (Atkinson et al., 2012).

1. Why study the Southern Ocean?

The Southern Ocean is a unique environment, dominated by large prevailing current systems including the Antarctic Circumpolar Current and the Antarctic Slope Front (Talley et al., 2011; pg 438). The Antarctic Circumpolar Current is one of the strongest currents in the world and, through its connection with three ocean basins, it is considered a vehicle of transport between the world's oceans (Talley et al., 2011; pg 439). Sea ice plays an extremely important role in shaping the Southern Ocean ecosystem and contributes to the uniqueness of this environment. The wide ice sheets around the Antarctic continent have resulted in a shelf break that occurs 2 – 4 times deeper than around other continents (Knox, 2007; pg 4). Changes in sea ice influence the biology of the ecosystem, including productivity and the timing of krill spawning, both of which have flow-on effects on the rest of the ecosystem (Murphy et al., 2007, Smith and Nelson, 1986). Extensive sea ice complicates in situ sampling because it is difficult for ships to travel through the ice and then observe the ecosystem in an undisturbed state. Autonomous Underwater Vehicles are a technological solution to this problem; the 'Autosub' is now a proven technology which has been successful in measuring oceanographic conditions under ice sheets (Nicholls et al., 2008, Nicholls et al., 2006) as well as krill (Brierley et al., 2002). However, the Autosub is expensive, requires a research ship to support operations and is extremely limited in its range of operations. Another, much older high-latitude sampling method, although not without its risks (e.g. the sinking of the Endurance (Shackleton, 1920)), is simply allowing a ship to be surrounded by ice and carried as the ice drifts. The upcoming 2019-2020 Polarstern Arctic voyage aims to drift with the ice for one year (MOSAiC, 2016).

It is not only the physical features of the Southern Ocean that make the area unique, but also the management regime. The Antarctic Treaty, established in 1959, is an agreement between all nations involved in research below latitude 60^{o}S and dictates the terms of research and cooperation in the Southern Ocean and Antarctic continent (Hanessian, 1960). Ecosystem management is governed by the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR), including the management of commercial fisheries and research (Miller, 2011). Antarctic krill is an important component of CCAMLR ecosystem management as it is an important prey item to many marine predators including fish, squid, seals and whales and is also the target species of the Southern Ocean's largest fishery (Siegel, 2016; pg 387, Nicol et al., 2012).

The importance of the Southern Ocean to global climatic processes, its current vulnerability to climate change and the many species that have adapted to the harsh environmental conditions make the Southern Ocean an important area to study (Caldeira and Duffy, 2000, Constable et al., 2014). It may be argued that relative to some other marine ecosystems the Southern Ocean has been reasonably well studied, but there remains much work to be done, especially in the development of models that can make sound inference from the large and complex survey data that exists (Boyd, 2002). One of the benefits to studying animals in the Southern Ocean is the relatively simple food web (Figure 1). In this thesis I use sophisticated modelling techniques to study drivers of distribution and density of three key components of this food web: phytoplankton, Antarctic krill (*Euphausia superba*) and humpback whales (*Megaptera novaeangliae*).

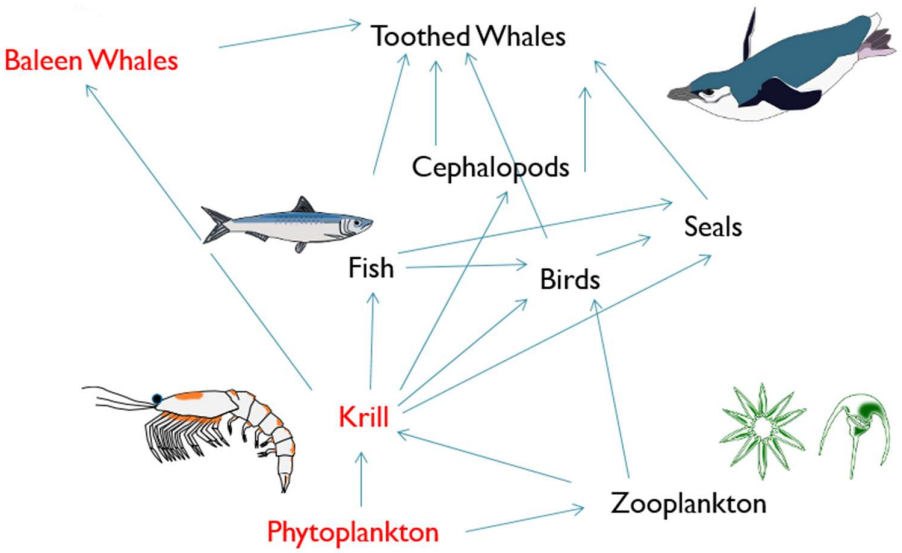


Figure 1 *Basic food web of the Southern Ocean. Animal images from Gemma Carroll, Macquarie University (with permission).*

2. Phytoplankton, krill and whales: Key components of the Southern Ocean food web

2.1. Phytoplankton

Phytoplankton form the base of the food web and worldwide are responsible for almost half of global primary productivity, fixing an estimated 30 – 50 billion tonnes of carbon annually (Field et al., 1998, Falkowski, 1994, Saba et al., 2011). Annual primary productivity in the Southern Hemisphere below 50°S is estimated at 2.9 billion tonnes (Moore and Abbott, 2000). This means that phytoplankton have enormous influence over the world's atmosphere and climate and so are an important component of earth system modelling. The Southern Ocean is characterised as a High-Nutrient/Low-Chlorophyll area, largely due to iron limitation (Boyd et al., 2000, Pollard et al., 2009). Phytoplankton levels in the East-Antarctic are higher near the ice-edge than in the open ocean during summer and are thought to be limited by iron levels and grazing by krill (Westwood et al., 2010, Wright et al., 2010).

2.2. Antarctic krill

Antarctic krill provide the largest link by biomass between primary producers and higher trophic levels, forming a major part of the diet for seabirds, seals and whales. Despite their small individual size, they have an estimated biomass of 379 million tonnes and occupy 19 million km^2 of the Southern Ocean during summer (Atkinson et al., 2009). It is important that this substantial resource is appropriately managed, especially since Antarctic krill are the target of the world's largest krill fishery and are an important food source for many predators. Krill distribution is highly patchy, with swarm density depending on environmental conditions and swarm shape varying with oxygen levels and predation risk (Godlewska et al., 1988, Brierley and Cox, 2010). This patchiness makes the analysis of krill distribution very difficult. Historically, krill were sampled using net sampling, although we know that they show net avoidance which biases biomass estimates (Atkinson et al., 2004, Wiebe et al., 2004). An alternative approach is to use active acoustics, and advances in computing power since the 1980s have made species identification and biomass estimation possible from active acoustics (Hewitt et al., 2004). Accordingly in situ sampling now typically occurs using echosounders (Horne, 2000).

2.3. Whales

Whales are an iconic group in the Southern Ocean and are key predators of krill. It has been shown that they contribute significantly to ecosystem function through the input of iron and nitrogen into the water during defecation (Nicol et al., 2010, Lavery et al., 2014). Whales are also responsible for the transport of nutrients from the polar feeding grounds to the temperate and tropical breeding grounds (Roman et al., 2014). As wide-ranging,

highly migratory foragers they are difficult to sample systematically (Kaiser, 2011; pg 209). The most common sampling method for whales are ship-based visual surveys (Kinzey et al., 2000), often undertaken concurrently with acoustic surveys for prey including fish and krill (Murase et al., 2006). Reconciling predator-prey distribution data is difficult because of a lack of available analytical techniques for data collected on different scales, which is often combined with opportunistic survey design (Fauchald et al., 2000). Spatial analysis of predator-prey relationships is important for understanding how habitat shapes predator distribution and vice versa (Willems and Hill, 2009).

This classic phytoplankton-krill-whale food chain serves as a motivation for the selection of study species in this thesis. Improving our understanding of these species' distribution will assist with describing ecosystem function, serve as a guide and framework for future work and enable other researchers to work with the mass of expensive but under-utilised data that already exists.

3. Thesis outline

This thesis aims to fill knowledge gaps associated with drivers of productivity and animal distribution in the Southern Ocean. Each chapter builds on the previous as we move up the food chain from primary producers (Chapter 3) to grazers (Chapter 5) and finally to top predators (Chapter 6). In Chapter 2 I provide a theoretical review of statistical methods in ecology and in Chapter 4 I develop a software package for data processing, which is then applied in chapters 5 and 6. There are several analytical and data processing challenges that complicate statistical analyses of marine species distribution and can make the extraction of a true ecological signal difficult.

3.1. Challenges when extracting ecological signals from marine data

Marine data are generally highly complex and successfully extracting an ecological signal is difficult. Spatial autocorrelation, which occurs when points closer in space are more similar than those further away, potentially occurs in three dimensions – latitude, longitude and depth (Sahlin et al., 2014). Spatial correlation violates a fundamental assumption of many statistical models, i.e. the assumption that observations are independent of one another (Haining, 2015). Autocorrelation can lead to incorrect model predictions and spurious inference, i.e. incorrectly attributing change (Diniz et al., 2003).

Given the difficulties in sampling the oceans on a regular spatial grid, or at regular time intervals, the data are generally irregularly positioned. This complicates analysis, especially where autocorrelation is concerned because many software packages cannot routinely account for 3-dimensional autocorrelation, particularly when sampling sites occur irregularly in space. This irregular sampling can also result in highly uneven between-group sample sizes. This problem is exacerbated in remote areas like the Southern Ocean, where inclement weather limits sampling opportunities.

In addition to the logistic and statistical difficulties of conducting research in the oceans, surveying and understanding marine ecosystems is more complex than on land. Processes in the ocean are more dynamic on both small and large scales, which makes teasing them apart more difficult (Kaiser, 2011; pg 208). For example, food resources in the oceans are highly variable depending on the environment (e.g. phytoplankton blooms) and in the pelagic realm, often do not have a fixed location, making the predators' search for prey more challenging than on land (Steele, 1989, Sims et al., 2006). Many of these processes that we aim to understand in the Southern Ocean are on scales that cannot be replicated in the laboratory, or affect organisms that cannot be held in captivity in order to conduct

manipulative experiments. A key example is understanding the effect of climate change on phytoplankton, where the processes are too subtle and much too complex to be represented accurately in the lab (Boyd et al., 2008). This makes field and modelling studies critical tools in answering these questions. As we have a limited ability to control factors that we are not interested in, statistical and sampling methodology must be sophisticated enough to avoid confounding ecological signal with noise or extraneous variables.

It is important that we do not discount these challenges when modelling marine data. Ignoring characteristics of data such as pseudo-replication, spatial autocorrelation and sampling design can lead to incorrect inference. Overlooking spatial autocorrelation can bias coefficient estimates and reduce model goodness-of-fit (Dormann, 2007). Not accounting for pseudo-replication can increase Type I errors (false positive), through the underestimation of the true variation and misrepresentation of its sources (Heffner et al., 1996). Overfitting reduces the reproducibility of the model and can lead to significant findings that are not actually true (Babyak, 2004). These problems are inherent in ecological data and analyses must consider them to avoid seriously impacting our conclusions.

3.2. Mixed models: a statistical solution to these problems?

Mixed models are becoming more common in ecology due to their ability to handle complex data (Bolker et al., 2009). They are especially applicable for marine data, where we are often limited by the environment and high data collection costs and are hence unable to perfectly follow a desired survey design. The key feature of mixed models is that they can facilitate inference at the population level, which is very important when

we are interested in the overall processes operating over a large area but are using data collected at many sites through this area (Bolker et al., 2009). Mixed models can incorporate complex correlation structures, allowing for different spatial autocorrelation patterns, and variance structures, allowing for different within-group variances for data collected across different sites or groups. Mixed models are particularly useful for data sets which are pseudo-replicated and can include complex structures of nesting (Chaves, 2010). Because there are many models available to ecologists, Chapter 2 of this thesis provides a theoretical overview of commonly used modelling methodology and introduces the analyses used in each subsequent chapter.

3.3. Phytoplankton in vertical profiles

Phytoplankton play a key role in global climate and are the base of the food chain in the Southern Ocean (Murphy and Hofmann, 2012). Phytoplankton blooms are strongly linked to environmental factors including light availability, nutrients, mixing of the water column and grazing (Barnes and Hughes, 2009; pg 32). In the Southern Ocean, sea ice levels are of great importance and melting ice can support blooms through the input of nutrients, seeding of algae and stabilisation of the water column (Sedwick and DiTullio, 1997, Smith and Nelson, 1986). Phytoplankton and environmental data are often collected through the water column using profiling instruments. The conductivity temperature depth (CTD) probe is a commonly used instrument and can collect data through the depths of the water column to over 1 km deep (Thomson and Emery, 2014; pg 19).

When collecting data at fixed stations, such as CTD deployment sites, it is important to use techniques that account for pseudo-replication of measurements within each station as well as vertical correlation. As mentioned previously, ignoring this can result in spurious findings. Chapter 3 of this thesis develops a mixed model for assessing the drivers of phytoplankton distribution in the East Antarctic from CTD profiles that can deal with these problems. The model I developed quantifies trends across large survey areas while recognising that the data are grouped into spatially autocorrelated vertical profiles.

3.4. Remote sensing: fisheries acoustics and data processing

Fisheries acoustics is a highly valuable tool for collecting high resolution ecological data through the water column as the ship travels (Benoit-Bird and Lawson, 2016). Acoustic data enable rapid sampling and provide a non-extractive method of estimating the density, distribution and biomass of many pelagic species over large survey areas (Kaiser, 2011; pg 211). Active acoustics have been used since the 1980s to study krill and processing (Horne, 2000) and identification methods are constantly being improved (Fallon et al., 2016, Korneliussen et al., 2016).

Active acoustics data sets are often collected incidentally while a ship is conducting other research or commercially fishing and hence there are many data sets already in existence. There are a number of initiatives for collecting and making available acoustics data in the Southern Hemisphere. The Integrated Marine Observing System Bio-Acoustic program collects 38kHz data from participating scientific vessels and commercial fishing vessels as they travel through the ocean basins of the Southern Hemisphere (IMOS Bio-Acoustic,

2016). The Southern Ocean Network of Acoustics is a related initiative aiming to collect and catalogue acoustics data, with the aim of mapping and identifying changes in the distribution of mid-level trophic level organisms in the Southern Ocean (SONA, 2016). I have contributed to this expanding area of research by developing an R package (R Development Core Team, 2014), EchoviewR (Harrison et al., 2015) – a flexible approach to the automation of acoustic data processing.

Acoustic data sets can be extremely large, often comprising of billions of raw data points that must be processed, integrated, cleaned to remove noise and then analysed appropriately. This is extremely time consuming since it must be done manually by human operators using available software programs. There is often a lack of automation and consistency when processing fisheries acoustic data. Reproducibility is difficult because often there is no record of the data processing techniques that have been applied to a given data set. If an error is subsequently found or an improvement in methods occurs, this requires reprocessing and must be done again manually. In Chapter 4 of this thesis I address these problems by developing an R package, EchoviewR, that automates data processing by acting as a scripting interface for one of the major acoustic processing software programs, Echoview (Echoview, 2015). EchoviewR contributes to reproducible research because the code script acts as a record of processing methods and can be modified and re-run on different data sets. This vastly cuts down the large number of hours required for manual active acoustics data processing. EchoviewR is of use to anyone using active acoustic data and applications may include biomass estimation of krill or fish, seafloor mapping or identification of features in the water column such as oil seeps.

3.5. Drift or swim: what drives the distribution of krill?

Antarctic krill are a key grazer on phytoplankton and provide the primary food source for many species in the Southern Ocean. Krill occur in large swarms which are distributed patchily throughout the Southern Ocean (Siegel, 2016; pg 279). The extent to which these swarms passively drift on large current systems, versus actively swimming, has been debated for decades. Larval krill are certainly passive drifters, relying on circulation to assist them in successfully completing their descent-ascent hatching cycle and transport them to suitable locations for maturation (Thorpe et al., 2004). Despite its ecological importance, passive drifting versus active swimming has been rarely studied in marine species (Putman et al., 2016).

There are important implications for animals being passive drifters or active swimmers (Richerson et al., 2015). Active swimmers may be able to take advantage of patchy or sparsely distributed resources that passive drifters might not be able to access. Habitat models will vary based on whether the target animal shows habitat preferences and can follow these preferences (swimming) or whether they are simply physically transported around the habitat by circulation (drifting). It also has implications for energetics, with lab-based studies estimating that active swimming in krill could account for 73% of metabolic expenditure during summer (Swadling et al., 2005). The movements of krill may cause mixing in the water column (Leshansky and Pismen, 2010) and if this is true, whether they drift or swim could influence mixing patterns. Drifting versus swimming is an important consideration for management approaches, most of which assume that krill are passive drifters (Richerson et al., 2015).

Studies assessing krill swarm drifting versus swimming have mostly been theoretical or observational. Overlaying historical krill distribution over large scale circulation patterns shows similarities between the two (Nicol, 2006, Amos, 1984). Lagrangian particle tracking also suggests that passive transport of larval krill causes intermixing of populations in different areas (Hofmann and Murphy, 2004). However, there is also evidence that krill could be active swimmers. Life history modelling suggests that there are strong selection advantages for active swimming in krill, including a 70% increase in reproductive success (Richerson et al., 2015). Current profiling has also shown that krill swarms can move in relation to local currents (Tarling and Thorpe, 2014).

The highly skewed and zero-inflated krill densities that arise from swarming make simple modelling of behaviour difficult because log-transformation is not possible with zeros present in the data. Some studies have side-stepped this issue by adding a constant before transformation of krill densities (Atkinson et al., 2004), however this approach can bias the fit of the model (O'Hara and Kotze, 2010). New modelling approaches, such as hurdle models, can be helpful here because they separate out the data into two separate models: i) presence/absence, where probability of presence is estimated and ii) conditional count, which models the remaining non-zero data (Zuur et al., 2009). In Chapter 5, I develop a hurdle model for assessing whether krill aggregate around resources. I have extended traditional hurdle models to incorporate continuous density data (hurdle models are currently only available for discrete count data) and to include a random effect at each level to allow for pseudo-replicated data within sites. There are other methods available for modelling skewed data that include zeros, such as generalised linear models, however these methods struggle to deal with the zero-inflation present in krill density data.

3.6. Combining predator-prey observations: the Balleny islands

Marine mammals are the predominant krill predators in the Southern Ocean and are estimated to be responsible for 30-60% of total krill predation by biomass (Siegel, 2016; pg 325). Observational data of marine mammals are often collected along with active acoustic observations of krill. However due to the nature of the observations and a likelihood of mismatch in sampling scales, quantitative analysis is difficult. Krill data are continuously collected at high resolution along the ships track line (Hewitt and Demer, 2000), but sparsely distributed marine mammal sightings occur only at the surface and visibility may depend on environmental variables and group size (Barlow et al., 2001). Complications include this spatial discrepancy between sightings and prey, opportunistic survey designs, non-linear predator-prey relationships and perception bias in marine mammal sightings. Distance sampling is a widely used method that accounts for perception bias in line and point transect data, and involves correcting for the fact that animals further away are less likely to be seen (Thomas et al., 2002).

Coincident predator-prey data are collected for many reasons. They are not only useful for quantifying predator-prey interactions, but also to characterise valuable regions in a survey area for conservation planning (Schmitt et al., 2016). This type of data can also be used to answer ecological questions and estimate energetics, which can then be used in population and species distribution modelling (Hatton et al., 2015, Trainor et al., 2014). Mapping prey distribution is especially important when prey is a patchily distributed and mobile resource, that itself relies on environmental features and habitat of the survey area, as this will affect a predator's foraging choices (Vijayan et al., 2017). Productivity is also an important consideration because it can indicate areas with a general enhancement of

the water column (Perissinotto et al., 1992). Productivity in the oceans is non-uniform and bathymetric features, frontal zones and islands are known to be highly productive. The increase in productivity around islands is termed the Island Mass Effect (Elliott et al., 2012). This can occur through the input of nutrients such as iron from upwelling and the stabilisation of the water column from melting ice and freshwater runoff (Planquette et al., 2007, Perissinotto et al., 1992). The fixed location of islands may be attractive for migratory predators seeking high food availability in a large expanse of ocean.

The Balleny Islands (67°S, 164°E) are a Southern Ocean archipelago that have received little research time due to their remote location, but are a known humpback whale feeding ground (Constantine et al., 2014). In Chapter 6 I investigate whether the waters around the Balleny Islands contain more krill swarms than the surrounding open ocean to assess why whales are attracted to the islands. I then use coincident whale sighting, krill acoustic and environmental data to describe habitat use of whales around the islands using a density surface model. Density surface models are a recent statistical advance that incorporate i) distance sampling to correct for perception bias, ii) generalised additive models to account for non-linear relationships, iii) survey design to account for opportunistic surveys with unequal effort and iv) a spatial surface to map unexplained spatial variability (Miller et al., 2013). This makes them a very useful technique for extracting predator-prey-environment information over large survey areas.

In summary, I have developed a suite of analytical techniques to assess interactions between key functional groups in the Southern Ocean food web and their environment and energy sources. The different nature and complexities of the three key components studied – phytoplankton, Antarctic krill and humpback whales – meant that they required different approaches. The models developed in this thesis can be used to predict future

distributions and abundance under different environmental scenarios and are readily adaptable for other marine ecosystems. Active acoustic data for sampling krill requires a large amount of manual processing so I developed automated processing software, EchoviewR, which I then applied in the final two chapters in this thesis. The statistical techniques used in this thesis are complex and a review of current methods has been conducted in Chapter 2 as a preface to the applied chapters. Except for this theoretical review in Chapter 2, the chapters of this thesis were written for publication and each contains the relevant background information, methodology and discussion.

4. References

AMOS, A. F. 1984. Distribution of krill (Euphausia superba) and the hydrography of the Southern Ocean: Large-scale processes. *Journal of Crustacean Biology,* 4**,** 306-329.

ATKINSON, A., SIEGEL, V., PAKHOMOV, E. A., JESSOPP, M. J. & LOEB, V. 2009. A re-appraisal of the total biomass and annual production of Antarctic krill. *Deep Sea Research Part I: Oceanographic Research Papers,* 56**,** 727-740.

ATKINSON, A., SIEGEL, V., PAKHOMOV, E. A. & ROTHERY, P. 2004. Long-term decline in krill stock and increase in salps within the Southern Ocean. *Nature,* 434**,** 100-103.

ATKINSON, A., WARD, P., HUNT, B., PAKHOMOV, E. & HOSIE, G. 2012. An overview of Southern Ocean zooplankton data: abundance, biomass, feeding and functional relationships. *CCAMLR Science,* 19**,** 171-218.

BABYAK, M. A. 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine,* 66**,** 411-421.

BARLOW, J., GERODETTE, T. & FORCADA, J. 2001. Factors affecting perpendicular sighting distances on shipboard line-transect surveys for cetaceans. *Journal of Cetacean Research and Management,* 3**,** 201-212.

BARNES, R. S. K. & HUGHES, R. N. 2009. The Planktonic System of Surface Waters. *An Introduction to Marine Ecology.* 3rd Edition ed. Great Britian: Blackwell Publishing Ltd.

BENOIT-BIRD, K. J. & LAWSON, G. L. 2016. Ecological insights from pelagic habitats acquired using active acoustic techniques. *Annual review of marine science,* 8**,** 463-490.

BOLKER, B. M., BROOKS, M. E., CLARK, C. J., GEANGE, S. W., POULSEN, J. R., STEVENS, M. H. H. & WHITE, J.-S. S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution,* 24**,** 127-135.

BOYD, P. W. 2002. Environmental factors controlling phytoplankton processes in the Southern Ocean. *Journal of Phycology,* 38**,** 844-861.

BOYD, P. W., DONEY, S. C., STRZEPEK, R., DUSENBERRY, J., LINDSAY, K. & FUNG, I. 2008. Climate-mediated changes to mixed-layer properties in the Southern Ocean: assessing the phytoplankton response. *Biogeosciences,* 5**,** 847-864.

BOYD, P. W., WATSON, A. J., LAW, C. S., ABRAHAM, E. R., TRULL, T., MURDOCH, R., BAKKER, D. C., BOWIE, A. R., BUESSELER, K. & CHANG,

H. 2000. A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization. *Nature,* 407**,** 695-702.

BRIERLEY, A. & COX, M. J. 2010. Shapes of Krill Swarms and Fish Schools Emerge as Aggregation Members Avoid Predators and Access Oxygen. *Current Biology,* 20**,** 1758-1762.

BRIERLEY, A. S., FERNANDES, P. G., BRANDON, M. A., ARMSTRONG, F., MILLARD, N. W., MCPHAIL, S. D., STEVENSON, P., PEBODY, M., PERRETT, J., SQUIRES, M., BONE, D. G. & GRIFFITHS, G. 2002. Antarctic Krill Under Sea Ice: Elevated Abundance in a Narrow Band Just South of Ice Edge. *Science,* 295**,** 1890-1892.

CALDEIRA, K. & DUFFY, P. B. 2000. The role of the Southern Ocean in uptake and storage of anthropogenic carbon dioxide. *Science,* 287**,** 620-622.

CASSIE, R. M. 1956. The Sampling Problem, with particular reference to Marine Organisms. *Proceedings (New Zealand Ecological Society)***,** 37-39.

CHAVES, L. F. 2010. An entomologist guide to demystify pseudoreplication: data analysis of field studies with design constraints. *Journal of medical entomology,* 47**,** 291-298.

CONSTABLE, A. J., MELBOURNE-THOMAS, J., CORNEY, S. P., ARRIGO, K. R., BARBRAUD, C., BARNES, D. K. A., BINDOFF, N. L., BOYD, P. W., BRANDT, A., COSTA, D. P., DAVIDSON, A. T., DUCKLOW, H. W., EMMERSON, L., FUKUCHI, M., GUTT, J., HINDELL, M. A., HOFMANN, E. E., HOSIE, G. W., IIDA, T., JACOB, S., JOHNSTON, N. M., KAWAGUCHI, S., KOKUBUN, N., KOUBBI, P., LEA, M.-A., MAKHADO, A., MASSOM, R. A., MEINERS, K., MEREDITH, M. P., MURPHY, E. J., NICOL, S., REID, K., RICHERSON, K., RIDDLE, M. J., RINTOUL, S. R., SMITH, W. O., SOUTHWELL, C., STARK, J. S., SUMNER, M., SWADLING, K. M., TAKAHASHI, K. T., TRATHAN, P. N., WELSFORD, D. C., WEIMERSKIRCH, H., WESTWOOD, K. J., WIENECKE, B. C., WOLF-GLADROW, D., WRIGHT, S. W., XAVIER, J. C. & ZIEGLER, P. 2014. Climate change and Southern Ocean ecosystems I: how changes in physical habitats directly affect marine biota. *Global Change Biology,* 20**,** 3004-3025.

CONSTANTINE, R., STEEL, D., ALLEN, J., ANDERSON, M., ANDREWS, O., BAKER, C. S., BEEMAN, P., BURNS, D., CHARRASSIN, J.-B., CHILDERHOUSE, S., DOUBLE, M., ENSOR, P., FRANKLIN, T., FRANKLIN, W., GALES, N., GARRIGUE, C., GIBBS, N., HARRISON, P., HAUSER, N., HUTSEL, A., JENNER, C., JENNER, M.-N., KAUFMAN, G., MACIE, A., MATTILA, D., OLAVARRÍA, C., OOSTERMAN, A., PATON, D., POOLE, M., ROBBINS, J., SCHMITT, N., STEVICK, P., TAGARINO, A., THOMPSON, K. & WARD, J. 2014. Remote Antarctic feeding ground important for east Australian humpback whales. *Marine Biology,* 161**,** 1087-1093.

DINIZ, J. A. F., BINI, L. M. & HAWKINS, B. A. 2003. Spatial autocorrelation and red herrings in geographical ecology. BLACKWELL PUBLISHING LTD.

DORMANN, C. F. 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global ecology and biogeography,* 16**,** 129-138.

DOUGLASS, J. G., DUFFY, J. E. & BRUNO, J. F. 2008. Herbivore and predator diversity interactively affect ecosystem properties in an experimental marine community. *Ecology Letters,* 11**,** 598-608.

ECHOVIEW, H., AUSTRALIA 2015. Echoview Software, version 6.1.35.26153.

ELLIOTT, J., PATTERSON, M. & GLEIBER, M. Detecting 'Island Mass Effect'through remote sensing. Proceedings of the 12th International Coral Reef Symposium, 2012 Cairns, Australia.

FALKOWSKI, P. G. 1994. The role of phytoplankton photosynthesis in global biogeochemical cycles. *Photosynthesis Research,* 39**,** 235-258.

FALLON, N. G., FIELDING, S. & FERNANDES, P. G. 2016. Classification of Southern Ocean krill and icefish echoes using random forests. *ICES Journal of Marine Science,* 73**,** 1998-2008.

FAUCHALD, P., ERIKSTAD, K. E. & SKARSFJORD, H. 2000. Scale-Dependent Predator-Prey Interactions: The Hierarchical Spatial Distribution of Seabirds and Prey. *Ecology,* 81**,** 773-783.

FIELD, C. B., BEHRENFELD, M. J., RANDERSON, J. T. & FALKOWSKI, P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science,* 281.

GODLEWSKA, M., KLUSEK, Z. & WARSZAWY, P. 1988. The density structure of krill aggregations and their diurnal and seasonal changes (BIOMASS III, October-November 1986 and January 1987). *Pol. Polar Res,* 9**,** 357-366.

GODØ, O. R., HANDEGARD, N. O., BROWMAN, H. I., MACAULAY, G. J., KAARTVEDT, S., GISKE, J., ONA, E., HUSE, G. & JOHNSEN, E. 2014. Marine ecosystem acoustics (MEA): quantifying processes in the sea at the spatio-temporal scales on which they occur. *ICES Journal of Marine Science,* 71**,** 2357-2369.

GRIMM, N. B. 1995. Why link species and ecosystems? A perspective from ecosystem ecology. *Linking species & ecosystems.* Springer.

GUISAN, A., LEHMANN, A., FERRIER, S., AUSTIN, M., OVERTON, J. M. C., ASPINALL, R. & HASTIE, T. 2006. Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology,* 43**,** 386-392.

HAINING, R. 2015. Spatial Autocorrelation. *International Encyclopedia of the Social & Behavioral Sciences (Second Edition).* Oxford: Elsevier.

HANESSIAN, J. 1960. The Antarctic Treaty 1959. *The International and Comparative Law Quarterly,* 9**,** 436-480.

HARRISON, L.-M. K., COX, M. J., SKARET, G. & HARCOURT, R. 2015. The R package EchoviewR for automated processing of active acoustic data using Echoview. *Frontiers in Marine Science,* 2.

HATTON, I. A., MCCANN, K. S., FRYXELL, J. M., DAVIES, T. J., SMERLAK, M., SINCLAIR, A. R. & LOREAU, M. 2015. The predator-prey power law: Biomass scaling across terrestrial and aquatic biomes. *Science,* 349**,** aac6284.

HEFFNER, R. A., BUTLER, M. J. & REILLY, C. K. 1996. Pseudoreplication revisited. *Ecology,* 77**,** 2558-2562.

HEWITT, R. P. & DEMER, D. A. 2000. The use of acoustic sampling to estimate the dispersion and abundance of euphausiids, with an emphasis on Antarctic krill, Euphausia superba. *Fisheries Research,* 47**,** 215-229.

HEWITT, R. P., WATKINS, J., NAGANOBU, M., SUSHIN, V., BRIERLEY, A. S., DEMER, D., KASATKINA, S., TAKAO, Y., GOSS, C., MALYSHKO, A., BRANDON, M., KAWAGUCHI, S., SIEGEL, V., TRATHAN, P., EMERY, J., EVERSON, I. & MILLER, D. 2004. Biomass of Antarctic krill in the Scotia Sea in January/February 2000 and its use in revising an estimate of precautionary yield. *Deep Sea Research Part II: Topical Studies in Oceanography,* 51**,** 1215-1236.

HOFMANN, E. E. & MURPHY, E. J. 2004. Advection, krill, and Antarctic marine ecosystems. *Antarctic Science,* 16**,** 487-499.

HORNE, J. K. 2000. Acoustic approaches to remote species identification: a review. *Fisheries oceanography,* 9**,** 356-371.

IMOS BIO-ACOUSTIC. 2016. Available: http://imos.org.au/bioacoustic.html [Accessed 3/3/2017].

KAISER, M. J. 2011. *Marine ecology: processes, systems, and impacts*, Oxford University Press.

KINZEY, D., OLSON, P. & GERRODETTE, T. 2000. Marine mammal data collection procedures on research ship line-transect surveys by the Southwest Fisheries Science Center. *NOAA, SWFSC Administrative Report LJ-00-08*.

KNOX, G. A. 2007. *Biology of the Southern Ocean,* USA, CRC Press.

KORNELIUSSEN, R. J., HEGGELUND, Y., MACAULAY, G. J., PATEL, D., JOHNSEN, E. & ELIASSEN, I. K. 2016. Acoustic identification of marine species using a feature library. *Methods in Oceanography,* 17**,** 187-205.

LAVERY, T. J., ROUDNEW, B., SEYMOUR, J., MITCHELL, J. G., SMETACEK, V. & NICOL, S. 2014. Whales sustain fisheries: Blue whales stimulate primary production in the Southern Ocean. *Marine Mammal Science,* 30**,** 888-904.

LAWLESS, J. F. 2014. *Statistics in Action: A Canadian Outlook*, Chapman and Hall/CRC.

LESHANSKY, A. & PISMEN, L. 2010. Do small swimmers mix the ocean? *Physical Review E,* 82**,** 025301.

MICHENER, W. H. & JONES, M. B. 2012. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in ecology & evolutions,* 27**,** 85-93.

MILLER, D. 2011. Sustainable management in the Southern Ocean: CCAMLR science. *Science diplomacy: Antarctica, science, and the governance of international spaces***,** 103-121.

MILLER, D. L., BURT, M. L., REXSTAD, E. A. & THOMAS, L. 2013. Spatial models for distance sampling data: recent developments and future directions. *Methods in Ecology and Evolution,* 4**,** 1001-1010.

MILLS, L. S., SOUL, XE, E., M. & DOAK, D. F. 1993. The Keystone-Species Concept in Ecology and Conservation. *BioScience,* 43**,** 219-224.

MOORE, J. K. & ABBOTT, M. R. 2000. Phytoplankton chlorophyll distributions and primary production in the Southern Ocean. *Journal of Geophysical Research,* 105**,** 709-722.

MOSAIC 2016. Multidisciplinary drifting Observatory for the Study of Arctic Climate (http://www.mosaicobservatory.org/documents.html). International Arctic Science Committee

MURASE, H., KIWADA, H., MATSUOKA, K. & NISHIWAKI, S. 2006. Results of the cetacean prey survey using a quantitative echo sounder in JARPA from 1998/99 to 2004/2005.

MURPHY, E. J. & HOFMANN, E. E. 2012. End-to-end in Southern Ocean ecosystems. *Current Opinion in Environmental Sustainability,* 4**,** 264-271.

MURPHY, E. J., TRATHAN, P. N., WATKINS, J. L., REID, K., MEREDITH, M. P., FORCADA, J., THORPE, S. E., JOHNSTON, N. M. & ROTHERY, P. 2007. Climatically driven fluctuations in Southern Ocean ecosystems. *Proceedings of the Royal Society B: Biological Sciences,* 274**,** 3057-3067.

NICHOLLS, K. W., ABRAHAMSEN, E. P., BUCK, J. J. H., DODD, P. A., GOLDBLATT, C., GRIFFITHS, G., HEYWOOD, K. J., HUGHES, N. E., KALETZKY, A., LANE-SERFF, G. F., MCPHAIL, S. D., MILLARD, N. W., OLIVER, K. I. C., PERRETT, J., PRICE, M. R., PUDSEY, C. J., SAW, K., STANSFIELD, K., STOTT, M. J., WADHAMS, P., WEBB, A. T. &

WILKINSON, J. P. 2006. Measurements beneath an Antarctic ice shelf using an autonomous underwater vehicle. *Geophysical Research Letters,* 33.

NICHOLLS, K. W., ABRAHAMSEN, E. P., HEYWOOD, K. J. & STANSFIELD, K. 2008. High-latitude oceanography using the Autosub autonomous underwater vehicle. *Limnology and Oceanography,* 53**,** 2309-2320.

NICOL, S. 2006. Krill, Currents, and Sea Ice: Euphausia superba and Its Changing Environment. *BioScience,* 56**,** 111-120.

NICOL, S., BOWIE, A., JARMAN, S., LANNUZEL, D., MEINERS, K. M. & VAN DER MERWE, P. 2010. Southern Ocean iron fertilization by baleen whales and Antarctic krill. *Fish and Fisheries,* 11**,** 203-209.

NICOL, S., FOSTER, J. & KAWAGUCHI, S. 2012. The fishery for Antarctic krill – recent developments. *Fish and Fisheries,* 13**,** 30-40.

O'HARA, R. B. & KOTZE, D. J. 2010. Do not log-transform count data. *Methods in Ecology and Evolution,* 1**,** 118-122.

OVASKAINEN, O. & SOININEN, J. 2011. Making more out of sparse data: hierarchical modeling of species communities. *Ecology,* 92**,** 289-295.

PERISSINOTTO, R., LAUBSCHER, R. & MCQUAID, C. 1992. Marine productivity enhancement around Bouvet and the South Sandwich Islands (Southern Ocean). *Marine Ecology Progress Series,* 88**,** 41-41.

PLANQUETTE, H., STATHAM, P. J., FONES, G. R., CHARETTE, M. A., MOORE, C. M., SALTER, I., NÉDÉLEC, F. H., TAYLOR, S. L., FRENCH, M., BAKER, A. R., MAHOWALD, N. & JICKELLS, T. D. 2007. Dissolved iron in the vicinity of the Crozet Islands, Southern Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography,* 54**,** 1999-2019.

POLLARD, R. T., SALTER, I., SANDERS, R. J., LUCAS, M. I., MOORE, C. M., MILLS, R. A., STATHAM, P. J., ALLEN, J. T., BAKER, A. R., BAKKER, D. C. E., CHARETTE, M. A., FIELDING, S., FONES, G. R., FRENCH, M., HICKMAN, A. E., HOLLAND, R. J., HUGHES, J. A., JICKELLS, T. D., LAMPITT, R. S., MORRIS, P. J., NEDELEC, F. H., NIELSDOTTIR, M., PLANQUETTE, H., POPOVA, E. E., POULTON, A. J., READ, J. F., SEEYAVE, S., SMITH, T., STINCHCOMBE, M., TAYLOR, S., THOMALLA, S., VENABLES, H. J., WILLIAMSON, R. & ZUBKOV, M. V. 2009. Southern Ocean deep-water carbon export enhanced by natural iron fertilization. *Nature,* 457**,** 577-580.

PUTMAN, N. F., LUMPKIN, R., SACCO, A. E. & MANSFIELD, K. L. Passive drift or active swimming in marine organisms? Proc. R. Soc. B, 2016. The Royal Society.

R DEVELOPMENT CORE TEAM 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

RICHERSON, K., WATTERS, G. M., SANTORA, J. A., SCHROEDER, I. D. & MANGEL, M. 2015. More than passive drifters: a stochastic dynamic model for the movement of Antarctic krill. *Marine Ecology Progress Series,* 529**,** 35-48.

ROMAN, J., ESTES, J. A., MORISSETTE, L., SMITH, C., COSTA, D., MCCARTHY, J., NATION, J. B., NICOL, S., PERSHING, A. & SMETACEK, V. 2014. Whales as marine ecosystem engineers. *Frontiers in Ecology and the Environment,* 12**,** 377-385.

SABA, V., FRIEDRICHS, M., ANTOINE, D., ARMSTRONG, R., ASANUMA, I., BEHRENFELD, M., CIOTTI, A., DOWELL, M., HOEPFFNER, N. & HYDE, K. 2011. An evaluation of ocean color model estimates of marine primary productivity in coastal and pelagic regions across the globe.

SAHLIN, J., MOSTAFAVI, M. A., FOREST, A. & BABIN, M. 2014. Assessment of 3D Spatial Interpolation Methods for Study of the Marine Pelagic Environment. *Marine Geodesy,* 37**,** 238-266.

SCHMITT, E. L., LUCKENBACH, M. W., LEFCHECK, J. S. & ORTH, R. J. 2016. Predator–prey interactions in a restored eelgrass ecosystem: strategies for maximizing success of reintroduced bay scallops (Argopecten irradians). *Restoration Ecology,* 24**,** 558-565.

SEDWICK, P. N. & DITULLIO, G. R. 1997. Regulation of algal blooms in Antarctic shelf waters by the release of iron from melting sea ice. *Geophysical Research Letters,* 24**,** 2515-2518.

SHACKLETON, E. H. 1920. *South: the story of Shackleton's last expedition, 1914-1917*, Macmillan.

SHAVER, G. R., CANADELL, J., CHAPIN, I. F. S., GUREVITCH, J., HARTE, J., HENRY, G., INESON, P., JONASSON, S., MELILLO, J., PITELKA, L. & RUSTAD, L. 2000. Global Warming and Terrestrial Ecosystems: A Conceptual Framework for Analysis. *BioScience,* 50**,** 871-882.

SIEGEL, V. 2016. Biology and Ecology of Antarctic Krill. *Advances in polar ecology*.

SIMS, D. W., WITT, M. J., RICHARDSON, A. J., SOUTHALL, E. J. & METCALFE, J. D. 2006. Encounter success of free-ranging marine predator movements across a dynamic prey landscape. *Proceedings of the Royal Society of London B: Biological Sciences,* 273**,** 1195-1201.

SMITH, W. O. & NELSON, D. M. 1986. Importance of ice edge phytoplankton production in the Southern Ocean. *BioScience,* 36**,** 251-257.

SONA. 2016. Available: https://sona.aq/ [Accessed 3/3/2017].

STEELE, J. H. 1989. The ocean 'landscape'. *Landscape ecology,* 3**,** 185-192.

SWADLING, K. M., RITZ, D. A., NICOL, S., OSBORN, J. E. & GURNEY, L. J. 2005. Respiration rate and cost of swimming for Antarctic krill, Euphausia superba, in large groups in the laboratory. *Marine Biology,* 146**,** 1169-1175.

TALLEY, L. D., PICKARD, G. L., EMERY, W. J. & SWIFT, J. H. 2011. The Southern Ocean. *Descriptive Physical Oceanography: An Introduction.* 6th Edition ed. USA: Elsevier Science.

TARLING, G. A. & THORPE, S. E. 2014. Instantaneous movement of krill swarms in the Antarctic Circumpolar Current. *Limnology and Oceanography,* 59**,** 872-886.

THOMAS, L., BUCKLAND, S. T., BURNHAM, K. P., ANDERSON, D. R., LAAKE, J. L., BORCHERS, D. L. & STRINDBERG, S. 2002. Distance Sampling. *In:* EL-SHAARAWI, A. H. & PIEGORSCH, W. W. (eds.) *Encyclopedia of Environmetrics.* Chichester: John Wiley & Sons Ltd.

THOMSON, R. E. & EMERY, W. J. 2014. *Data analysis methods in physical oceanography*, Newnes.

THORPE, S. E., HEYWOOD, K. J., STEVENS, D. P. & BRANDON, M. A. 2004. Tracking passive drifters in a high resolution ocean model: implications for interannual variability of larval krill transport to South Georgia. *Deep Sea Research Part I: Oceanographic Research Papers,* 51**,** 909-920.

TRAINOR, A. M., SCHMITZ, O. J., IVAN, J. S. & SHENK, T. M. 2014. Enhancing species distribution modeling by characterizing predator–prey interactions. *Ecological Applications,* 24**,** 204-216.

VIJAYAN, S., KOTLER, B. P. & ABRAMSKY, Z. 2017. A predator equalizes rate of capture of a schooling prey in a patchy environment. *Behavioural Processes,* 138**,** 91-95.

WESTWOOD, K. J., BRIAN GRIFFITHS, F., MEINERS, K. M. & WILLIAMS, G. D. 2010. Primary productivity off the Antarctic coast from 30°–80°E; BROKE-West survey, 2006. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57**,** 794-814.

WIEBE, P. H., ASHJIAN, C. J., GALLAGER, S. M., DAVIS, C. S., LAWSON, G. L. & COPLEY, N. J. 2004. Using a high-powered strobe light to increase the catch of Antarctic krill. *Marine Biology,* 144**,** 493-502.

WILLEMS, E. P. & HILL, R. A. 2009. Predator-Specific Landscapes of Fear and Resource Distribution: Effects on Spatial Range Use. *Ecology,* 90**,** 546-555.

WRIGHT, S. W., VAN DEN ENDEN, R. L., PEARCE, I., DAVIDSON, A. T., SCOTT, F. J. & WESTWOOD, K. J. 2010. Phytoplankton community structure and stocks in the Southern Ocean (30–80°E) determined by CHEMTAX analysis of HPLC pigment signatures. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57**,** 758-778.

ZUUR, A. F., IENO, E. N., WALKER, N. J., SAVELIEV, A. A. & SMITH, G. M. 2009. *Mixed Effects Models and Extensions in Ecology with R,* New York, Springer.

Chapter 2: Theoretical foundation for models

Authors:

Lisa-Marie K. Harrison[1]

Affiliations:

[1]Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, New South Wales, Australia

This chapter focuses on the modelling techniques that are available for the analysis of ecological data and introduces the methods that I have used in my thesis. This review was undertaken because it is essential to choose modelling methods appropriate to the systems being studied and the data available. This review has been compiled to fulfil the mandate that modelling biological systems is undertaken to enable a better understanding of their function as well as providing a mechanism by which to predict future outcomes with the best chance that the predictions are correct.

There are two main approaches to statistical modelling: frequentist and Bayesian. There is a prolonged debate on the pros and cons of each method, and disagreement over which is more correct in different circumstances (Vallverdu, 2016; pg 61). The debate ranges from profound and technical, to relatively indecisive; for instance Bland & Altman (1998) suggested that a person's choice of approach will depend on the university they attended. In this chapter I present a concise comparison of frequentist and Bayesian methodology and an overview of different models used in the analysis of ecological data. In this chapter I review both frequentist and Bayesian methods, but for the data chapters that follow I use only frequentist inference.

1. Frequentist Inference

Frequentist approaches are by far the most common in biological and ecological modelling. A frequentist model is so called, because it is based on the long-run frequencies of events (Vallverdu, 2016; pg 49). Essentially, a frequentist asks "what is the probability of the data given this model/parameter is correct" (McCarthy, 2007; pg 8). Frequentist methods are often the only methods taught to biologists during higher education. Hence most applied statistics in the biological sciences is from a frequentist perspective.

1.1. Model Selection

1.1.1. Null hypothesis testing

Frequentist testing methods revolve around the formulation of a refutable null hypothesis, and the presentation of an alternate hypothesis. For example, the null hypothesis might be that a parameter is not different to 0, and the alternate hypothesis will be that the parameter does differ from zero. Hypothesis tests have a wide number of uses, including testing whether two populations are different at a statistically significant level.

Null hypothesis testing involves the formulation of a p-value to decide whether to reject the null hypothesis based on a chosen α cut-off, which is normally chosen as 0.05. The p-value is defined as "the probability, under the null hypothesis, of a test statistic as or more extreme than actually observed" (Rohde, 2014; pg 52). However, many people incorrectly assume that the p-value is the probability that the hypothesis is correct given the data (Congdon, 2006).

1.1.2. Hypothesis testing errors

There are two types of error that can occur: we can falsely reject the null hypothesis when it is actually true (Type I error) or fail to reject it when we should (Type II) (Rohde, 2014; pg 42). The Type I and Type II errors are not conditional on the strength of the support for the hypothesis by the data. Rather, they depend only on whether the data lies within the constructed acceptance or rejection region (Dass and Berger, 2003). Unfortunately, many scientists use the p-value as if it represents truth exactly and doggedly adhere to the $\alpha = 0.05$ threshold for significance, rather than using it as a continuum by which to judge strength for the hypothesis, as Fisher first proposed (Halsey et al., 2015). Johnson (2013) suggested revising the standard α threshold to 0.005 or even 0.001 for highly significant findings, however 0.05 is still most commonly used. The p-value threshold value will

also depend on the consequences of making a Type I or Type II error. For example, in climate science you might need to be highly certain that your result is not a false negative, because there may be limited time to react and the consequences of not doing so might be severe. In contrast, a medical study might require that there is a very small chance of a false positive, so that a person will not receive treatment for a condition they do not have.

1.1.3. Accounting for multiple testing

Null hypothesis testing using p-values can only test one hypothesis at a time. Running many hypothesis tests at once increases the probability of type II error (failure to reject the null hypothesis when you should) (Westfall and Young, 1993; pg 2). For example, if we run 100 hypothesis tests, 5 of them may be falsely significant based on $\alpha = 0.05$. The number of p-values reported in the journal *Ecology* has been above 3000 every year since 1984 (Anderson et al., 2000), potentially leading to 150 falsely significant findings each year. There are numerous corrections that can be applied to allow for testing of multiple hypotheses, the most common of which is the Bonferroni correction. Rather than using the predetermined α for all hypotheses, the Bonferroni correction uses α/k, where k is the number of hypotheses to be tested (Westfall and Young, 1993; pg 44). Hence if 100 hypotheses were to be tested at $\alpha = 0.05$, the p-value would need to be < 0.0005 for a null hypothesis to be rejected.

There are arguments against corrections to the p-value such as the Bonferroni correction, like how to choose the scale it applies to (only on one study, to all papers in a journal or even a lifetime of work) (Moran, 2003) and how logical it is to choose the significance level to reject a hypothesis based on how many questions you plan to ask (Perneger, 1998). Armstrong (2014) suggests only three situations in which the Bonferroni

correction should be considered: 1) one universal hypothesis (H_0) that all tests are non-significant is tested, 2) avoiding Type I errors is critical and 3) when conducting many tests without a pre-defined hypothesis.

The growing criticisms of the p-value and null hypothesis testing are causing a shift in model selection methods, as seen in a 15% decrease in conservation biology papers using null hypothesis testing in two leading journals from 2000 (93%) to 2005 (78%) (Fidler et al., 2006). Instead, authors are turning towards likelihood-based approaches, Bayesian statistics and confidence intervals.

1.1.4. Likelihood and Information Theoretics

Likelihood based approaches are another tool for model selection, and can be found in both frequentist and Bayesian methods. The frequentist Likelihood Ratio Tests involve calculating a test statistic from the ratio of the likelihood of two models. The test statistic is calculated as (West et al., 2007; pg 35):

$$\boldsymbol{l = -2\left(log_e(L_i) - log_e(L_j)\right)} \qquad \textit{Equation 1}$$

where L_i and L_j are the likelihoods of two nested models. Models are defined as nested if all terms in the simpler model also occur in the more complex model. This test statistic is then tested for significance by computing the p-value from a Chi-square distribution. If there are more than two models that require comparing, multiple Likelihood Ratio Tests can be run although this method only works for nested models (Johnson and Omland, 2004). The models must also be fit to the same subset of data (West et al., 2007; pg 35). Wald tests are similar to Likelihood Ratio Tests but can be used to assess the significance of a single model compared to a null model (Everitt, 2006; pg 416).

Information Theoretic methods can be used compare multiple plausible models and are an alternative to traditional hypothesis testing using p-values (Anderson et al., 2000). The Akaike Information Criterion (AIC) (Akaike, 1998) is often used in frequentist Information Theoretics methods. It is a measure of the amount of information retained by a model while penalising for the number of parameters and is based on the Kullback-Leibler information criterion (Demidenko, 2004). The AIC is given by:

$$AIC = -2l_{max} + 2k \qquad \textit{Equation 2}$$

where l_{max} is the maximised log likelihood and k is the number of parameters. The AIC is calculated for each candidate model, and the model with the lowest AIC is preferred. The lowest AIC value does not indicate truth, only that the model is the most plausible of the candidate models. Hence it is important to assess that all models are plausible before undertaking model selection. The difference in AIC between a candidate model and the model with the smallest AIC may be small (<2), in which case there is substantial support for the second model also. In this case, model averaging can be used. It is important to note that model averaging can cause problems with inference in cases that require the simultaneous interpretation of multiple coefficients, such as interaction terms or polynomial predictors (Cade, 2015). In some cases standardisation, where the data are scaled to a mean of 0 and standard deviation of 1, can help with this issue, but this must be assessed on a case-by-case basis.

Model selection using criterion such as AIC is well regarded because it provides a rank of all models in a set and allows for model averaging if many models are similar (Johnson and Omland, 2004). Despite its intention to be used with a small set of predetermined and plausible models, Information Theoretic methods are still used inappropriately for 'data-dredging', where many models are compared regardless of their biological credibility

(Anderson, 2008; pg 47, 64). However, it should be noted that there are situations where data dredging is appropriate. An extreme example of this are a number of papers comparing over 1 million models when their sample size is small, at less than 100 observations (Anderson, 2008; pg 6). AIC can under or over-estimate model complexity for singular models or small sample size, and the Frequentist Information Criterion (FIC), which aims to reduce this problem, was proposed in 2015; however this work is currently only in pre-print (LaMont and Wiggins, 2015).

A corrected version of AIC, the AIC_c, is available for small sample sizes and prevents the criterion from favouring larger models by adding an extra bias term (Sugiura, 1978). The corrected AIC is:

$$\boldsymbol{AIC_c = AIC + \frac{2k(k+1)}{n-k-1}} \qquad \textit{Equation 3}$$

where *k* is the number of parameters and *n* is the sample size (Anderson, 2008; pg 60). Information Criteria are also used in the Bayesian Framework.

2. Bayesian Inference

Bayesian models are based on Bayes theorem. Bayes theorem is named after Thomas Bayes, who proposed it in the 1740s, however it wasn't until after his death in 1761 that the work was found and published (McCarthy, 2007, Bayes, 1763). In direct contrast to frequentist approaches, Bayesian methods ask the question "what is the probability of my model/parameter given the data". In some situations, this may be a more natural way to approach questions in ecology (Wade, 2000).

Bayesian models use prior information in addition to the available data to inform a potential model. The posterior probability distribution, defined as the probability of the hypothesis given the data, is calculated based on the product of the likelihood function and the prior distribution (Wade, 2000). For a finite number of hypotheses, the posterior of hypothesis *i* using Bayes Theorem is (McCarthy, 2007; pg 12):

$$P(H_i|D) = \frac{P(H_i).\ P(D|H_i)}{\sum_j P(H_j).P(D|H_j)} \qquad \textit{Equation 4}$$

where $P(H_i)$ = prior probability of hypothesis i

$P(D|H_i)$ = probability of obtaining the data given hypothesis i

D = data

j = other hypotheses

The three steps to Bayesian model fitting are (Denison et al., 2002):

1. Assign priors to all parameters and states to be estimated
2. Define the likelihood of the data given the parameters and states
3. Calculate the posterior distribution of the parameters and states given the data using Bayes theorem as shown above

2.1. Prior Information

The prior distribution for a parameter θ, $p(\theta)$, describes the probability of different values of θ without considering the data (Wasserman, 2000). A prior can be any piece of previously known information, which can have differing levels of uncertainty around it.

McCarthy (2007) presents the following diagram to represent the relationship between the data, priors and posterior:

prior + data $\longrightarrow$ posterior

Each parameter in the model can be assigned a different prior, which may simply be an informed guess about the distribution, mean and standard deviation of the parameter. For example, if the heights of a study species are already known to have a Gaussian distribution rather than a uniform distribution, this information could be used as a prior. A prior with a large variance is likely to be uninformative and the posterior distribution will be dominated by the data, giving results similar to a frequentist analysis (McCarthy, 2007).

Uniform priors (also called flat priors) are common when little information is known about the questions being asked, because they will be overwhelmed by the data. A uniform prior is considered improper because its integral is infinity, regardless of the constant chosen (Christensen et al., 2011). However it does lead to a proper posterior, making it an appropriate choice for inference when little is known about the parameter or state of interest. Since the prior does not carry much weight and the posterior is dominated by the data, the results of the analysis will be the same as the frequentist approach to the problem. The Gibbs sampler will give a reasonable looking output when improper priors are used, meaning that it is not a method which can be used to determine if the priors are improper. Uninformative priors have been more common in ecology, possibly because of a concern that informative priors may reduce accuracy; however, a recent study found that appropriate informative priors increased precision, although the effects on accuracy were variable (Morris et al., 2015).

2.2. Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) can be used to estimate the posterior distribution of a model in a Bayesian framework (McCarthy, 2007). While Bayes theorem can be simple enough that it can be computed by hand, in more complex and higher dimensional formulations it can be difficult to solve the integral in the denominator, which may have a dimension in the thousands (Cressie and Wikle, 2011). In this case, MCMC is used to avoid the necessity of calculating this denominator. The samples from the MCMC algorithms are equivalent to a sample from the posterior distribution, hence eliminating the need to calculate the often complex integral of the denominator of equation 4.

There are a several algorithms that are used to sample from the posterior via MCMC, including the Metropolis-Hastings algorithm and Gibbs sampler (Zuur et al., 2009). The general method behind these algorithms is that for every parameter we wish to estimate, a new value is drawn from that parameter's candidate-generating density, and the models with the new and old parameter values are compared using a pre-defined acceptance criteria (Chib and Greenberg, 1995). If the new parameter estimate is accepted, the current model is updated; if not the current model remains the same until the next iteration. The Gibbs sampler is a component-wise Metropolis-Hastings algorithm that is less general because it requires knowledge of the conditional posterior distributions of the parameters (Congdon, 2006, Denison et al., 2002).

A defining property of Markov Chains is that they are memory-less, so future states depend only on the current state, or in the case of higher-order chains states further back than a single time-step (Ching et al., 2013; pg 1). When running an MCMC algorithm, the first values (known as the burn in) must be removed because they show strong dependence on the arbitrarily chosen first value (Christensen et al., 2011; pg 145). The

point at which the chain no longer depends on the first value is called reaching stationarity. After numerous iterations, the MCMC algorithm may converge. This convergence should not be confused with that of an algorithm converging to a numeric solution; rather, convergence occurs when each realisation of the Markov Chain has the same distribution as the stationary distribution of the Markov Chain (Cressie and Wikle, 2011).

2.3. Model Selection

There is no single accepted method for Bayesian model selection (Hooten and Hobbs, 2015). The posterior probability of a model is often used for model selection. If comparing multiple models, the model with the highest posterior probability is chosen (Wasserman, 2000). If we have a total of R candidate models, the posterior probability for model *i* is calculated as (Posada and Buckley, 2004):

$$P(M_j \mid D) = \frac{P(D \mid M_i)\, P(M_i)}{\sum_{r=1}^{R} P(D \mid M_r)\, P(M_r)} \qquad \textit{Equation 5}$$

where D is the data, M_i is model *i* and $P(M_i)$ is the prior probability of M_i. To compare two models, *i* and *j*, the Bayes Factor can be used. It is a ratio of the evidence for each model and is calculated as (Congdon, 2006; pg 26):

$$B_{ij} = \frac{\mathrm{P}(D \mid M_i)}{\mathrm{P}(D \mid M_j)} \qquad \textit{Equation 6}$$

where M_i is model *i*, and D is the data. B_{ij} can be interpreted using Jefferey's scale (Wasserman, 2000; pg 99), where $B_{ij} = 5$ would mean there is 5x stronger evidence for model *i*. The Bayes Factor does not require models to be nested and automatically penalises model complexity because complex models are able to make a larger variety of predictions and hence *P(D | M)* for our observed data will be lower than for a more simple model (Berger et al., 1994).

2.3.1. Information Criteria

Like the frequentist AIC, the Bayesian Information Criterion (BIC) and Deviance Information Criterion (DIC) can be used for selection of the model parameters. The BIC is a measure of the evidence favouring a model compared to other models and does not require the specification of priors (Weakliem, 1999) and was primarily developed for model averaging within a well justified candidate set of models (Hooten and Hobbs, 2015). The DIC is very similar to the AIC, in that it measures the information content as a model, taking into account the model's complexity. Retaining the most information with the simplest model is preferable (McCarthy, 2007). The DIC can return a negative number of effective model parameters in missing data models (Celeux et al., 2006). This is due to the posterior mean settling on a value that is between two different modes of the posterior density. The BIC (Stoica and Selen, 2004) and DIC are calculated as follows:

$$\boldsymbol{BIC = -2\,log_e\,P\left(y, \hat{\theta}^n\right) + n\,log_e\,N} \qquad \textit{Equation 7}$$

$$\mathrm{DIC} = \widehat{\mathrm{D}} + 2\mathrm{p}_D$$

where y = the observed data of length N, $P()$ = the likelihood of the model, p_D is the effective number of parameters, $\hat{\theta}$ = the maximum likelihood estimate of the parameter vector containing n parameters and $\widehat{D}$ = deviance using mean of the parameter's posterior distributions. A model with a DIC close to the model with the lowest DIC (difference of <10) may also be the best model, and should not simply be removed in favour of the best model. Rather, model averaging could be used in this situation. When averaging across the parameters of multiple candidate models, care must be taken as the interpretation of each parameter will vary depending on the structure of the models (Posada and Buckley, 2004). A new information criteria was proposed in 2013, the widely applicable AIC

(WAIC) (Watanabe, 2013). It is a generalised fully Bayesian form of AIC which, unlike DIC, can be used for singular models whose Fisher-information matrix is not invertible.

2.4. Model Uncertainty

Bayesian models automatically include uncertainty around the parameters (Wade, 2000), and assume the data are fixed while the model can vary. In contrast, most frequentist approaches are built around the assumption that the model is true and that the data can vary (Chatfield, 2006). There are three main ways in which uncertainty enters a model:

1. Uncertainty surrounding the model's type and structure (i.e. hierarchical)
2. Uncertainty around choice in the model's parameters
3. Random unexplained variation in the dependent variable

Ignoring the uncertainty surrounding a model can be a dangerous practice and can lead to unsound conclusions. One method to reduce model uncertainty is model averaging, where instead of selecting the single best model, the posterior distribution is an average of a number of 'best' models, using the posterior model probabilities (Raftery et al., 1996).

3. Comparison between Bayesian and frequentist methods

While some of the components in frequentist and Bayesian methods may appear to be similar or even identical, there are important differences. For example, frequentist confidence intervals are different to Bayesian credible intervals. For a Bayesian credible interval, the interpretation is that 'there is a 95% chance that the true value of the parameter is within the interval'. This contrasts with a confidence interval where, if the experiment were repeated many times, 95% of the time the confidence interval would encompass the true value of the parameter. Credible intervals will have the same

numerical value as confidence intervals if the prior is uninformative (McCarthy, 2007). It is not uncommon for ecologists to misinterpret 95% confidence intervals by thinking that the definition is the same as for the credible interval.

Bayesian and frequentist methods can give different results depending on how the experiment was conducted. Frequentist models depend on the method by which the data was collected. For example, a frequentist hypothesis test can give different results if the sample was randomly collected to a predetermined number (for example, it was predetermined that 12 koalas with pouch young will be sampled), or was collected using a stopping method (it is predetermined that koalas will be sampled until 3 koalas with pouch young are obtained) (McCarthy, 2007). A Bayesian analysis would not give different answers in these two situations, and neither would Information Theoretic methods, which are based on maximum likelihood. Since frequentist and Bayesian methods will give different answers based on data collection methods (even without priors taken into account), caution must be exercised when comparing two studies that used frequentist and Bayesian methods.

Bayesian and frequentist methods can give the same results if the same model is being used. They are most alike when little or no prior information is available, and hence the priors have a large variance or are simply objective distributions (Bayarri and Berger, 2004). Theoretical and empirical studies show that the preferred method relies heavily on the quality of prior information available (Samaniego and Reneau, 1994). Prior information can be incorporated into frequentist statistics to some extent by constraining parameters. The knowledge that the two methods may produce the same results in some cases is certainly not new, with this being shown decades ago for one sided hypotheses (Casella and Berger, 1987). Bayesian methods allow for better understanding of uncertainty around the parameters (Congdon, 2006) and are also not restricted by sample

size, which can allow for quantitative studies on rare species (Dorazio, 2016). However, they are often far more computationally expensive and therefore may not be as efficient as frequentist methods. With that in mind, both methods have much to offer in the field of statistics (Bayarri and Berger, 2004).

3.1. Hybrid methods

Some methods blur the lines between frequentist and Bayesian analysis, presenting a 'unified' result. For example, if frequentist methods are heavily conditional, they can also give the same results as a Bayesian analysis (Berger et al., 1994, Berger et al., 1997). This involves the calculation of a test statistic that represents the strength of support for the null hypothesis by the data (Dass and Berger, 2003). Others have suggested a compromise, called the "calibrated Bayes", where both approaches should be used in analyses; frequentist methods would be used for model development and assessment, but the inference under the model would be from a Bayesian perspective (Little, 2006). This approach makes use of the strengths of both frequentist and Bayesian statistics. In the late 1990s a third paradigm, "evidential statistics", was proposed (Royall, 1997). Evidential statistics also draw on many concepts in both frequentist and Bayesian statistics, and brings the idea that the result space in model selection is a continuum divided into three areas: i) strong support for model A, ii) strong support for model B and iii) weak support for both models (Taper and Ponciano, 2016).

4. Models in ecology

There has been a rise in complex statistics in ecology, with researchers moving away from the traditional ANOVAs and t-tests and employing more complicated methodology such as Bayesian statistics and mixed models (Touchon and McCoy, 2016). Despite this, only 64.6% of ecological papers from 1990 - 2013 used any statistics (n = 30,190), and only 6.5% of ecological doctoral programs surveyed (n = 154) in the United States of America taught methods more complex than traditional statistical methods (Touchon and McCoy, 2016). This section reviews some of the models applicable to ecology. Due to the wide variety of models used, only the most relevant for this thesis are reviewed. Figure 1 shows a theoretical representation of how many of these models are related.

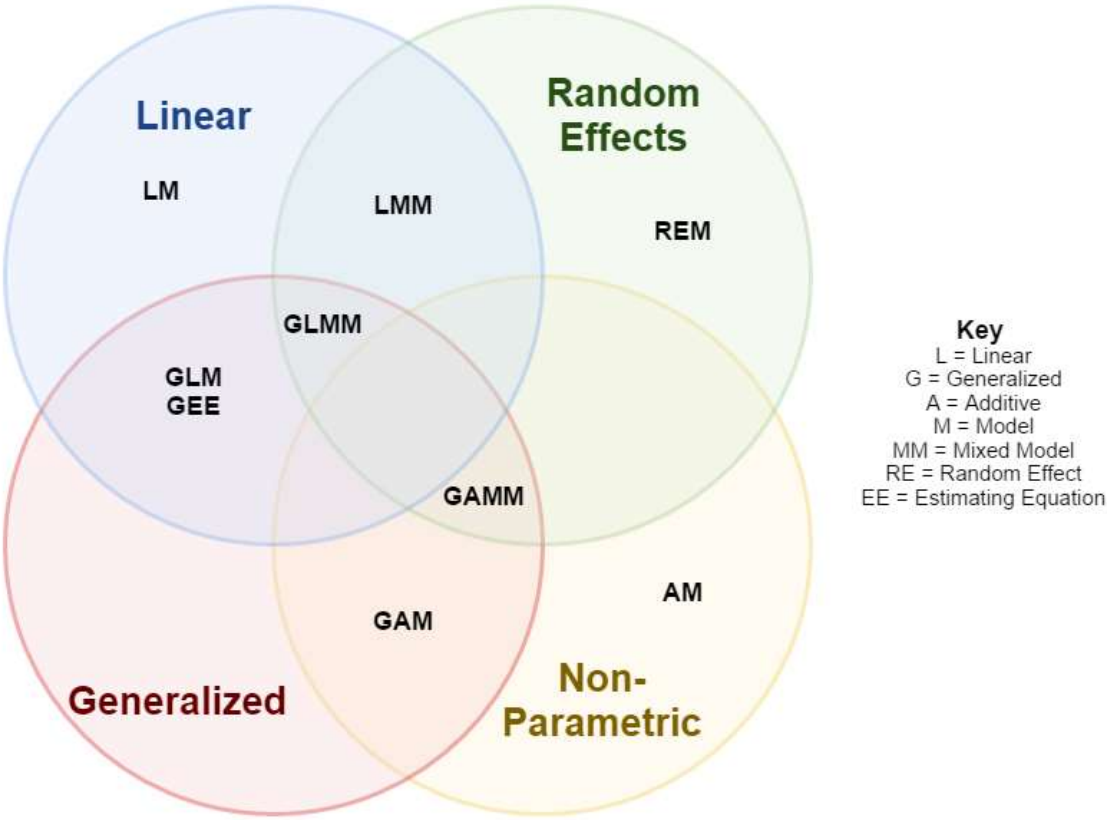


Figure 1 *Theoretical representation of how the commonly used models are related.*

4.1. Linear Models

Linear models are the simplest form of statistical model and are often the first that students are taught in school. They include any model that is linear in the parameters (Khuri, 2010; pg 1). The general form is:

$$Y = X\boldsymbol{\beta} + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2) \qquad \textit{Equation 8}$$

where $\boldsymbol{\beta}$ is a vector of coefficients, X are the explanatory variables, Y is the dependent variable and ε are the residuals (Clarke, 2008). There are two subsets of linear models: those with continuous explanatory variables are *regression* models and those with categorical explanatory variables are *Analysis of Variance* (*ANOVA)* models (Khuri, 2010; pg 2-3). Linear models can also accommodate a combination of categorical and continuous variables at the same time. Unfortunately the complexity of many ecological systems means that linear models are often not suitable for these analyses.

4.2. Generalised Linear Models (GLMs)

Generalised Linear Models (GLMs) were proposed by Nelder and Wedderburn in 1972 and are an extension of linear models to allow for error distributions other than the Normal distribution (Nelder and Wedderburn, 1972). They are appropriate for many of the commonly encountered types of ecological data including presence/absence and proportions (Binomial, Bernoulli), counts and densities (Poisson, Negative Binomial, Geometric) and non-negative continuous data (Exponential, Gamma, Inverse Gaussian). The Negative Binomial, Quasi-Poisson and Geometric distributions are commonly used when the data are overdispersed in comparison to the distribution that is otherwise appropriate.

The general form uses a link function, g(), to relate the mean, μ, to the covariates (Myers et al., 2010; pg 5):

$$\boldsymbol{g(\mu) = X\beta + \varepsilon} \qquad \textit{Equation 9}$$

where **β** is a vector of coefficients, X are the explanatory variables and ε are the residuals. Often transformations (log, square-root, inverse) are performed to normalise residuals and coerce the data into being suitable for a linear model (Osborne, 2005, Bartlett, 1947). However, GLMs have been shown through simulation to produce better results than transformation (O'Hara and Kotze, 2010). Transformation is particularly problematic when the data contain zeros because log transformation will make the zeros infinite. This issue is often side-stepped by adding a constant to all values to raise them above zero, after which a log-transformation is performed, however this is less than ideal because the choice of constant is arbitrary but can change the outcome of the model (Fletcher et al., 2005). Transformation can also result in the new transformed mean not being equivalent to the raw mean on the untransformed scale, which complicates interpretation. Hence, GLMs should be investigated before transforming data, especially when zeros are present.

4.3. Hierarchical Models

Hierarchical models are useful when there is nesting in the data, such as observations within sites or samples taken from individuals (Raudenbush and Bryk, 2002; pg 5 - 7). There can be any number of levels to the hierarchical model (McCarthy, 2007; pg 75) and they are present in both the Bayesian and frequentist framework.

The hierarchical model can have a 'data model' at the top level, which expresses the distribution of the observed data (Cressie and Wikle, 2011; pg 361 - 362). Underneath

this is a 'process model' which represents the uncertainty in the process (the true process is unobservable). This specific case is known as a state-space model because it describes the relationship between the underlying state and the observed data.

The hierarchical model is Bayesian if there is a 'parameter model' that specifies the joint probability of all unknown parameters. This parameter model relies on the theory that a joint distribution can be broken down into a conditional model: [A, B, C] = [A| B, C][B, C][C] (Wikle, 2003). For example, for studies with multiple locations (e.g.: different sites or quadrats), a Bayesian Hierarchical Model can allow the parameters to vary by site, while still having the same distribution (Borsuk et al., 2001, McCarthy, 2007).

4.4. Mixed Models

Unlike fixed effects models such as linear models and GLMs, Mixed Models include both fixed and random effects. Random effects account for variability between groups, such as individuals or sites, and allow for both subject specific and population level inference (Wu, 2009; pg 39). Generally the random effects measure factors whose levels are chosen randomly from a population and would not be the same if the experiment was repeated again (i.e. individuals within a population or quadrats within a site). Mixed Models can hence incorporate survey design (i.e. nesting of sites) and ensure that the data are not pseudoreplicated within the model.

Mixed models are a form of hierarchical model, in that they model data where the observations are nested within levels (Wu, 2009; pg 39). They can take many forms, including but not limited to random intercepts, random slope, generalised additive mixed models (GAMM) and generalised linear mixed models (GLMM) (Zuur et al., 2009; pg 101, 323). The general form of a mixed model is:

$$Y = X\boldsymbol{\beta} + Z\mathbf{v} + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2) \qquad \textit{Equation 10}$$

Where **β** is a vector of fixed effects, ***v*** is a vector of random effects, X and Z are design matrices and ε is a vector of unobserved errors. For the above model, $v \sim N(0, G)$ and $\varepsilon \sim N(0, R)$, where $R = \sigma^2 I$ and **I** is an identity matrix (Wolfinger, 1993). A covariance matrix can be used to indicate the structure of correlation between variables. There are several choices for covariance matrices, including but not limited to diagonal, compound symmetry, unstructured, Toeplitz and autoregressive (Wolfinger, 1993; pg 1081 - 1082). The choice of which to use is driven by the structure of correlation between the variables as seen in the data.

Mixed models are generally estimated using Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML). REML is preferred for the estimation of the variance components in mixed models because it does not depend on the correct estimation of the fixed effects, so the random effect estimates are not biased downward (Verbeke and Molenberghs, 2009). ML and REML require integration of the likelihood over all possible values of the random effects, making model fitting slow and often unfeasible for complex mixed models. To get around this, there are several ways to approximate the likelihood, including quasi-likelihood (pseudo and penalised), Laplace approximations, Gauss-Hermite quadrature and MCMC methods (Bolker et al., 2009).

4.5. Models for zero-inflation

Zero-inflation is especially common in animal count data because the number of cells, quadrats or sites where no animals are observed is often higher than the expected number of zeros under most theoretical distributions. Failing to account for zero-inflation can bias the parameter estimates and standard errors, cause overdispersion and mask the true ecological patterns (Martin et al., 2005, Zuur et al., 2009). There are two broad classes of models for dealing with zero-inflated data: hurdle models and zero-inflated models.

Models for zero-inflated data can be modified for spatial count data and repeated measures by incorporating random effects (Agarwal et al., 2002, Hall, 2000, Min and Agresti, 2005).

4.5.1. Hurdle models

Hurdle models use separate processes to model the zero and non-zero values. It can be interpreted as modelling the probability of presence/absence separate from conditional (given presence) counts. Hurdle models are recommended over zero-inflated models when the zeros are known to be 'true zeros', i.e. those arising from true absences rather than systematic errors such as the observer missing a sighting, or environmental conditions concealing a sighting (Martin et al., 2005). Generally, the zero model is a binomial model and the count model is a zero-truncated Poisson, Negative-Binomial or Geometric model. The general form of a hurdle model (Zuur et al., 2009; pg 287) is:

$$\boldsymbol{f} = \begin{cases} \boldsymbol{f_{binomial}}(\boldsymbol{y} = \boldsymbol{0};\ \boldsymbol{\gamma}) & \boldsymbol{y} = \boldsymbol{0} \\ \left(\boldsymbol{1} - \boldsymbol{f_{binomial}}(\boldsymbol{y} = \boldsymbol{0};\ \boldsymbol{\gamma})\right) * \frac{\boldsymbol{f_{count}}(\boldsymbol{y};\boldsymbol{\beta})}{\boldsymbol{1} - \boldsymbol{f_{count}}(\boldsymbol{y}=\boldsymbol{0};\boldsymbol{\beta})} & \boldsymbol{y} > \boldsymbol{0} \end{cases} \qquad \textit{Equation 11}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are vectors of covariates in the zero and count models respectively and $\boldsymbol{f}$ are the Probability Mass Functions. Hurdle models can also model zero-deflation (Min and Agresti, 2005), although this is much less commonly seen in ecology. In **Chapter 5** of this thesis, I use a hurdle model to assess whether Antarctic krill aggregate around resources or are passive drifters. I extended the traditional hurdle model by adding random effects in both stages of the model, because the data were collected across multiple sampling stations and are hence pseudoreplicated if sampling station is not incorporated into the model. Standard hurdle models use a count model for non-zero

values, however I modified this to be a continuous model because krill densities are not discrete counts.

4.5.2. Zero-inflated models

Zero-inflated models supplement a regular count model by modelling only *excess* zeros separately and the true zeros are retained in the count model. They are useful when we are cannot distinguish between the true and false zeros in the data (Martin et al., 2005). The zero-inflated Poisson and zero-inflated Binomial are the two common forms of these models, both of which use a binomial model for the excess zeros relative to the original count distribution. The general forms of zero-inflated models can be found in Zuur et al (2009; pg 276). The Poisson and Negative Binomial options can be compared using a likelihood ratio test (Zuur et al., 2009; pg 288).

4.6. Non-Parametric Models

4.6.1. Generalised Additive Models (GAMs)

GAMs use non-parametric smoothers to model the relationship between the dependent variable and each independent variable, which gives an advantage over a Linear Model because they can fit data-driven relationships between the predictors and the dependent variable (Guisan et al., 2002). This is particularly useful when we don't know much about the nature of this relationship (Denison et al., 2002). GAMs can also accommodate non-normal error structures. The general form of a GAM is similar to a linear model:

$$\boldsymbol{Y = s_o + \sum_{j=1}^{P} s_j(X_j) + \varepsilon} \qquad \textit{Equation 12}$$

where s_0 is an intercept and the s() terms are smoothers for each explanatory variable, $\mathbf{X}_j$ (Hastie and Tibshirani, 1986). Interactions are also possible and take the form s(x, y),

however this becomes complicated when x and y do not vary on the same scale. The smooth term can take a number of forms, most of which are based on smoothing splines. First proposed in 1922 (Whittaker, 1922), splines are piecewise polynomials that are smooth at the joining points, which are called knots (Wold, 1974). Splines are named by the degree of the polynomials that comprise them. For example, a spline made up of cubic functions is called a cubic smoothing spline. An important part of smoothing models is the question of how much to smooth (Lee et al., 2006). Too much smoothing can mask an underlying pattern but not enough will cause overfitting. The number of knots and the smoothing parameter (if present) can influence smoothness. There are several ways to estimate the smoothing parameter. If it is estimated from the data, it is called 'data-driven smoothing' (Lee et al., 2006). Generalised cross-validation and AIC can be used to estimate the smoothing parameter from the data.

4.6.2. *Generalised Additive Mixed Models*

While GAMs can deal with non-parametric relationships, they don't allow for random effects that can accommodate differences between locations or subjects. Rather than fitting separate models per study, which could reduce the power of the study as well as reduce the usefulness of the models, the GAM can be extended to a GAMM to allow for the addition of random effects. The GAMM extends the GAM by adding **Z**, a vector of random effects.

4.6.3. Spline Mixed Models (ASReml)

Non-linearity can be modelled in a linear mixed model framework using penalised splines, an approach available in the statistical package ASReml (VSNi, 2009). For each coefficient, the fixed component will model the linear trend and the cubic splines, fit as random effects, will model the departures from this linearity (Butler et al., 2009). It

should be noted that while the splines are fit as random effects, they are not true random effects, rather this is a mechanism for fitting the model. As with GAMs, we need to estimate the amount of smoothing required in our model. In **Chapter 3**, I use spline mixed models in ASReml to assess drivers of phytoplankton distribution off East-Antarctica. This modelling method was chosen because the splines allow for non-linear relationships, the random effect is required because the data were collected across multiple sampling stations (observations within a station are replicates) and ASReml allows for easy fitting of a 3D correlation structure to account for 3D spatial autocorrelation.

4.7. Generalised Estimating Equations

GLMs for count and binomial data can be extended to allow for correlation between observations, resulting in Generalised Estimating Equations (GEE) (Liang and Zeger, 1986, Ziegler, 2011). They differ from GLMs because you don't need to specify the full distribution, only the mean structure. GEE works on the idea that observations within a cluster of data will be correlated, while observations from different clusters won't be (Ziegler et al., 1998). Some examples of these clusters are longitudinal analysis, family studies and spatial analysis.

AIC cannot be directly used with GEEs because they are not likelihood based, although there is a modified AIC available that uses the quasi-likelihood (Pan, 2001a). Other model selection methods include minimisation of the expected predictive bias (Pan, 2001b) and Wald tests (Zuur et al., 2009; pg 318).

4.8. Autoregressive Models

For spatial data, conditional autoregressive models (CAR) and simultaneously autoregressive models (SAR) are often used to incorporate neighbourhood values into the

model. The covariance structure includes spatial dependence as a function of the neighbourhood matrix (Wall, 2004). This can be modelled either explicitly or implicitly.

SAR models can have different forms depending on whether the spatial autocorrelation is present only in the dependent variable or also in the explanatory variables (Dormann et al., 2007). If the autocorrelation is only present in the dependent variable the model takes the form:

$$Y = p\boldsymbol{W}Y + X\boldsymbol{\beta} + \varepsilon \qquad \textit{Equation 13}$$

Y is the dependent variable, p is the autoregression parameter, **W** is the spatial weights matrix, X is the vector of explanatory variables and **β** is a vector of slopes. If the autoregression is present in both the dependent and independent variables the model will take the form:

$$Y = p\boldsymbol{W_1}Y + X\boldsymbol{\beta} + \boldsymbol{W_2}X\gamma + \varepsilon \qquad \textit{Equation 14}$$

The X term is a matrix of spatially lagged predictors, $\mathbf{W_i}$ are spatial weight matrices which are usually assumed to be the same and ɣ is the regression coefficient for this matrix. The spatial autoregression can also occur only in the error term, in which case the model becomes:

$$Y = X\boldsymbol{\beta} + \varepsilon + \lambda\boldsymbol{W}\mu \qquad \textit{Equation 15}$$

In the above model, W is the weighted spatial structure of μ, the spatially dependent error term, ε is the error term and $\boldsymbol{\lambda}$ is a spatial coefficient to be estimated. Two different SAR models were used by Santora et al (2010) to assess the spatial dependence of baleen whales and Antarctic krill, one with a lagged dependent variable and one with an autoregressive error term (although the models were called spatial regression models).

Two dimensional correlograms were then used to assess correlation between krill and whale distribution.

CAR models are similar to SAR models, however the distance matrix, W, must be symmetric. The general form of a CAR model is:

$$Y = X\boldsymbol{\beta} + p\mathbf{W}(Y - X\boldsymbol{\beta}) + \varepsilon \qquad \textit{Equation 16}$$

Where *p* is an autoregression parameter, W is a spatial weighting matrix, X are the independent variables and β are a vector of slopes.

4.9. Density Surface Models

Density Surface Models (DSMs) combine numerous techniques to account for common problems seen in ecological data and are reviewed in Miller et al (2013). DSMs use GAMs to model data with non-linear trends, use an x-y surface to account for spatial variability and correct for the decrease in sightings as they get further from the transect using a detection function. This decrease in sightings with distance is a form of observation bias and occurs primarily because objects become harder to detect further away. If an observer measures the distance from transect to each sighting a function can be fit to correct the abundance and density estimates, in a method known as 'Distance Sampling' (Thomas et al., 2002). Commonly used detection functions include half-normal, hazard-rate and gamma (Figure 2). DSMs combine these strengths of GAMs and Distance Sampling and in addition incorporate survey design, which enables their use for surveys with unequal effort as is commonly seen from platforms of opportunity.

I use a DSM in **Chapter 6** of this thesis to link whale distribution at an island feeding area to food levels (krill) and environmental conditions. This model allowed for the unequal sample effort, the missed whale sightings as distance from transect increased and the non-linear relationships between whale count and environmental conditions.

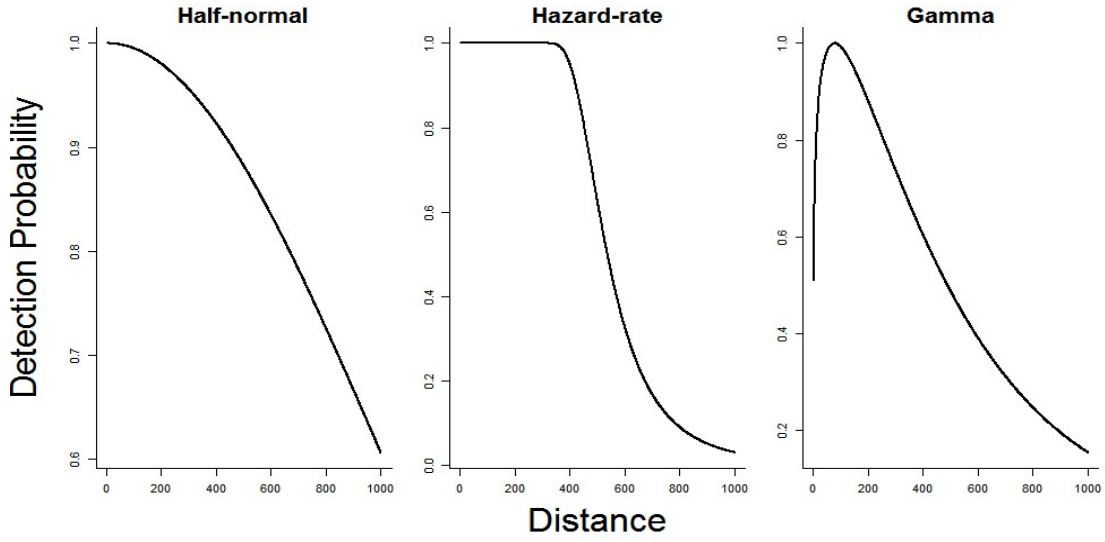


Figure 2 *Examples of detection functions for decreasing probability of detection from the transect. Half-normal detection functions are used when there is immediate decreasing probability of detection from transect while hazard-rate models assume perfect detectability until a certain distance, after which sightings decrease. Gamma functions are used when the peak probability of detection is not on the transect.*

4.10. Spatial Autocorrelation

Spatial autocorrelation is common in ecological data, where observations closer together are likely to be more similar than those further away (Dormann et al., 2007). This can occur when variables are measured on a finer scale than they vary, when variables are observed in 'blocks' with different observers which may introduce observer bias, if variables depend on an underlying spatial process or if the model omits an important

spatially varying trend (Haining, 2015). Spatial autocorrelation is problematic when correlation is visible in the model's errors. Many studies use conventional methods that ignore this problem. However making inference on models where the data doesn't satisfy the assumption of independence required by the model can lead to flawed conclusions (Haining, 2003, Kühn, 2007) and it is important that the final fitted model does not have spatially autocorrelated residuals which would violate the assumption of independence (Haining, 2003). Spatial autocorrelation can take many forms including autoregressive, moving average and autoregressive moving average (ARMA) among others.

There are many methods available for accounting for spatial autocorrelation of a model's errors and some of the most common are summarised in Table 1. In addition, the experimental design can be chosen to minimise the chance that sites will be correlated, such as spacing sites to maximise their distance apart, although there is no guarantee that closer sites will not still be autocorrelated (Haining, 2015).

Table 1 *Statistical methods to account for spatial autocorrelation in ecological data. Summarised from Dormann et al (2007).*

Method	How it works	Comments
Autocovariate models	Adds distance weighted function of nearby response variables to GLM	Applicable to binomial, normal and Poisson models
Spatial Eigenvector Mapping	Spatial arrangement of data is translated into explanatory variables	Computationally intensive for >200 data points
Correlation structures	Models the spatial covariance in the variance-covariance matrix	Available for many GLMs and mixed models
Conditional Autoregressive Models (CAR)	Weighted distance matrices to specify the strength of interaction between points	Unsuitable for directional processes
Simultaneous Autoregressive Models (SAR)	Weighted distance matrices to specify the strength of interaction between points	Like CAR but distance matrices don't need to be symmetric
Mixed models	Nesting of spatial autocorrelation structures within locations	Spatial autocorrelation in mixed models can be specified in the covariance G-matrix or in the error structure
Generalised Estimating Equations (GEE)	Correlation matrix to specify within cluster correlations	Better for parameter estimation than prediction. Correlations are reflected in the ordering of the data.

There are several tests to identify spatial autocorrelation. Mantel tests assess distance and similarity between sites (Mantel, 1967). If autocorrelation is not consistent across the entire study area, localised correlation can be tested for using Moran's I test, where a test statistic is computed for each region on the map (Haining, 2015). Plotting the residuals

of a model across each spatial dimension can identify if there is residual autocorrelation remaining in a model. Semivariograms are a good tool for visualising spatial autocorrelation and making a guess about the nature, whether it be Gaussian, exponential, spherical or something different (Stroup, 2013). A semivariogram is a plot of the semivariance vs the lag distance between observations (Figure 3). The 'nugget' is the variation that cannot be explained by the distance between observations, the 'sill' is the estimated variance, and the difference between these values is the observed variation that can be explained by distance. Semivariograms can be calculated for single directional distances or for 2-dimensional data such as latitude/longitude or row/column survey designs.

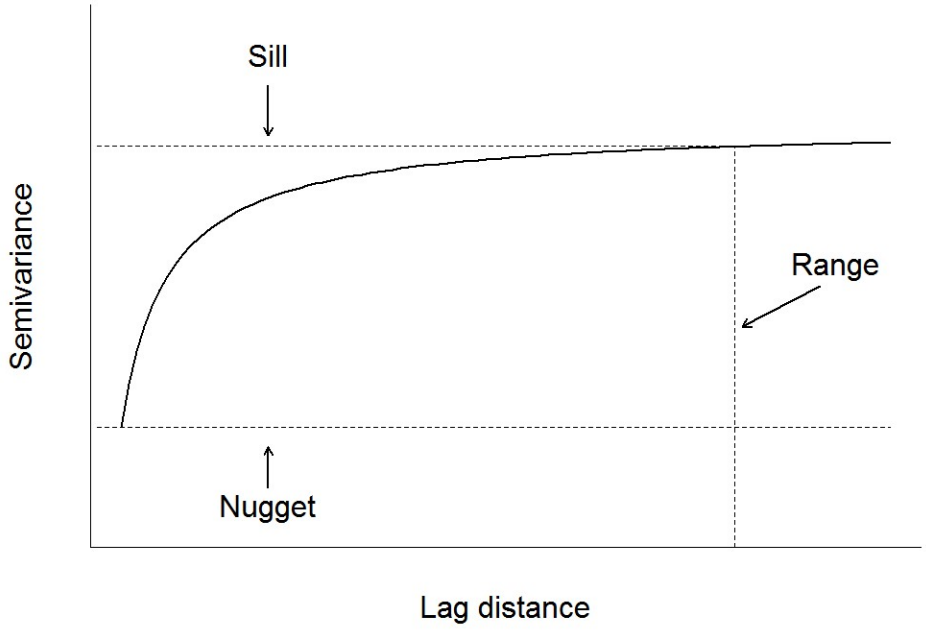


Figure 3 *Conceptual diagram of semivariogram showing the sill (estimated variance), nugget (variance unexplained by distance between observations) and range (distance to reach sill).*

5. Conclusion and methods used in this thesis

Ecological data is inherently complex due to the often non-linear relationships between variables, the spatial and temporal variability, inter-subject and inter-site differences and the high number of zeros due to patchily distributed animals or plants. Sophisticated modelling methods are required to answer the questions that we want to extract from the data. This chapter has provided a review of some useful models for ecologists and a comparison of the frequentist and Bayesian approaches to using them. The remaining chapters of this thesis focus on applying and extending some of these methods to answer important questions about the distribution of phytoplankton, Antarctic krill and humpback whales in the Southern Ocean. I have used a frequentist approach and Information Theoretics to reduce computational time and because in most cases there was little prior information available. The methods used in each chapter of this thesis are listed in Table 2. They are described in greater detail in the Methods and Discussion segments of each chapter. The software package developed in Chapter 4 does not use statistical methods for ecology and is hence not included in this table.

Table 2 *Methods used in each analysis chapter of this thesis.*

Thesis chapter	Complexities in data	Modelling method
Chapter 3 Modelling drivers of phytoplankton distribution off North-East Antarctica	▪ Data collected over multiple sites ▪ Non-linear relationships ▪ 3D spatial autocorrelation	Spline mixed models (in ASReml) with 3D autocorrelation structure
Chapter 5 Are Antarctic krill passive drifters or do they aggregate around resources?	▪ Data collected over multiple sites ▪ Continuous krill density data (regular hurdle models are only for counts) ▪ Zero-inflation (over half cells surveyed contained no krill)	Hurdle mixed model for semi-continuous data
Chapter 6 Modelling the effects of food distribution and environmental parameters at the Balleny Islands on feeding humpback whales	▪ Non-linear relationships ▪ Imperfect detectability of whales ▪ Uneven sampling effort in survey area	Density Surface Model

6. References

AGARWAL, D. K., GELFAND, A. E. & CITRON-POUSTY, S. 2002. Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics,* 9**,** 341-355.

AKAIKE, H. 1998. Information theory and an extension of the maximum likelihood principle. *Selected Papers of Hirotugu Akaike.* Springer.

ANDERSON, D. R. 2008. *Model Based Inference in the Life Sciences: A Primer on Evidence,* USA, Springer.

ANDERSON, D. R., BURNHAM, K. P. & THOMPSON, W. L. 2000. Null Hypothesis Testing: Problems, Prevalence, and an Alternative. *The Journal of Wildlife Management,* 64**,** 912-923.

ARMSTRONG, R. A. 2014. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics,* 34**,** 502-508.

BARTLETT, M. S. 1947. The Use of Transformations. *Biometrics,* 3**,** 39-52.

BAYARRI, M. J. & BERGER, J. O. 2004. The Interplay of Bayesian and Frequentist Analysis. *Statistical Science,* 19**,** 58-80.

BAYES, T. R. 1763. An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions,* 53**,** 370-418.

BERGER, J. O., BOUKAI, B. & WANG, Y. 1997. Unified Frequentist and Bayesian Testing of a Precise Hypothesis. *Statistical Science,* 12**,** 133-160.

BERGER, J. O., BROWN, L. D. & WOLPERT, R. L. 1994. A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing. *The Annals of Statistics,* 22**,** 1787-1807.

BLAND, J. M. & ALTMAN, D. G. 1998. Bayesians and frequentists. *BMJ,* 317**,** 1151-1160.

BOLKER, B. M., BROOKS, M. E., CLARK, C. J., GEANGE, S. W., POULSEN, J. R., STEVENS, M. H. H. & WHITE, J.-S. S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution,* 24**,** 127-135.

BORSUK, M. E., HIGDON, D., STOW, C. A. & RECKHOW, K. H. 2001. A Bayesian hierarchical model to predict benthic oxygen demand from organic matter loading in estuaries and coastal zones. *Ecological Modelling,* 143**,** 165-181.

BRUNSDON, C., FOTHERINGHAM, A. S. & CHARLTON, M. E. 1996. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis,* 28**,** 281-298.

BRUNSDON, C., FOTHERINGHAM, S. & CHARLTON, M. 1998. Geographically Weighted Regression-Modelling Spatial Non-Stationarity. *Journal of the Royal Statistical Society. Series D (The Statistician),* 47**,** 431-443.

BUTLER, D. G., CULLIS, B. R., GILMOUR, A. R. & GOGEL, B. J. 2009. Mixed models for S language environments: ASReml-R reference manual. Queensland Department of Primary Industries and Fisheries and NSW Department of Primary Industries, VSNi, UK.

CADE, B. S. 2015. Model averaging and muddled multimodel inferences. *Ecology,* 96**,** 2370-2382.

CASELLA, G. & BERGER, R. L. 1987. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association,* 82**,** 106-111.

CELEUX, G., FORBES, F., ROBERT, C. P. & TITTERINGTON, D. M. 2006. Deviance information criteria for missing data models. *Bayesian Analysis,* 1**,** 651-673.

CHATFIELD, C. 2006. Model Uncertainty. *Encyclopedia of Environmetrics.* John Wiley & Sons, Ltd.

CHIB, S. & GREENBERG, E. 1995. Understanding the Metropolis-Hastings Algorithm. *The American Statistician,* 49**,** 327-335.

CHING, W.-K., HUANG, X., NG, M. K. & SIU, T.-K. 2013. Markov Chains. Springer US.

CHRISTENSEN, R., JOHNSON, W., BRANSCUM, A. & HANSON, T. E. 2011. *Bayesian Ideas and Data Analysis: An introduction for scientists and statisticians,* Boca Raton, FL, Chapman & Hall/CRC Press.

CLARKE, B. R. 2008. *Linear Models: The theory and application of analysis of variance,* New Jersey, USA, John Wiley & Sons Inc.

CONGDON, P. 2006. *Bayesian Statistical Modelling,* England, Wiley.

CRESSIE, N. & WIKLE, C. K. 2011. *Statistics for Spatio-Temporal Data,* Hoboken, New Jersey, Wiley.

DASS, S. C. & BERGER, J. O. 2003. Unified Conditional Frequentist and Bayesian Testing of Composite Hypotheses. *Scandinavian Journal of Statistics,* 30**,** 193-210.

DEMIDENKO, E. 2004. *Mixed Models: Theory and Applications,* Hoboken, New Jersey, Wiley.

DENISON, D. G. T., HOLMES, C. C., MALLICK, B. K. & SMITH, A. F. M. 2002. *Bayesian methods for nonlinear classification and regression,* England, Wiley.

DORAZIO, R. M. 2016. Bayesian data analysis in population ecology: motivations, methods, and benefits. *Population Ecology,* 58**,** 31-44.

DORMANN, C. F., MCPHERSON, J. M., ARAÚJO, M. B., BIVAND, R., BOLLIGER, J., CARL, G., DAVIES, R. G., HIRZEL, A., JETZ, W. & DANIEL KISSLING, W. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography,* 30**,** 609-628.

EVERITT, B. 2006. *The Cambridge Dictionary of Statistics,* UK, Cambridge University Press.

FIDLER, F., BURGMAN, M. A., CUMMING, G., BUTTROSE, R. & THOMASON, N. 2006. Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology. *Conservation Biology,* 20**,** 1539-1544.

FLETCHER, D., MACKENZIE, D. & VILLOUTA, E. 2005. Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics,* 12**,** 45-54.

GUISAN, A., EDWARDS JR, T. C. & HASTIE, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling,* 157**,** 89-100.

GUISAN, A., WEISS, S. & WEISS, A. 1999. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology,* 143**,** 107-122.

HAINING, R. 2003. *Spatial Data Analysis - Theory and Practice,* Cambridge, Cambridge University Press.

HAINING, R. 2015. Spatial Autocorrelation. *International Encyclopedia of the Social & Behavioral Sciences (Second Edition).* Oxford: Elsevier.

HALL, D. B. 2000. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics,* 56**,** 1030-1039.

HALSEY, L. G., CURRAN-EVERETT, D., VOWLER, S. L. & DRUMMOND, G. B. 2015. The fickle P value generates irreproducible results. *Nat Meth,* 12**,** 179-185.

HASTIE, T. & TIBSHIRANI, R. 1986. Generalized Additive Models. *Statistical Science,* 1**,** 297-310.

HOOTEN, M. B. & HOBBS, N. T. 2015. A guide to Bayesian model selection for ecologists. *Ecological Monographs,* 85**,** 3-28.

JOHNSON, J. B. & OMLAND, K. S. 2004. Model selection in ecology and evolution. *Trends in ecology & evolution,* 19**,** 101-108.

JOHNSON, V. E. 2013. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences,* 110**,** 19313-19317.

KHURI, A. I. 2010. *Linear Model Methodology,* USA, Chapman & Hall/CRC Press.

KÜHN, I. 2007. Incorporating spatial autocorrelation may invert observed patterns. *Diversity and Distributions,* 13**,** 66-69.

LAMONT, C. H. & WIGGINS, P. A. 2015. The Frequentist Information Criterion (FIC): The unification of information-based and frequentist inference. *arXiv preprint arXiv:1506.05855.*

LEE, Y., NELDER, J. A. & PAWITAN, Y. 2006. *Generalized linear models with random effects: Unified analysis via H-likelihood,* Boca Raton, FL, Chapman & Hall/CRC.

LIANG, K.-Y. & ZEGER, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika,* 73**,** 13-22.

LITTLE, R. J. 2006. Calibrated Bayes: a Bayes/frequentist roadmap. *The American Statistician,* 60**,** 213-223.

MANTEL, N. 1967. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research,* 27**,** 209-220.

MARTIN, T. G., WINTLE, B. A., RHODES, J. R., KUHNERT, P. M., FIELD, S. A., LOW-CHOY, S. J., TYRE, A. J. & POSSINGHAM, H. P. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters,* 8**,** 1235-1246.

MCCARTHY, M. A. 2007. *Bayesian Methods for Ecology,* UK, Cambridge University Press.

MILLER, D. L., BURT, M. L., REXSTAD, E. A. & THOMAS, L. 2013. Spatial models for distance sampling data: recent developments and future directions. *Methods in Ecology and Evolution,* 4**,** 1001-1010.

MIN, Y. & AGRESTI, A. 2005. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling,* 5**,** 1-19.

MORAN, M. D. 2003. Arguments for Rejecting the Sequential Bonferroni in Ecological Studies. *Oikos,* 100**,** 403-405.

MORRIS, W. K., VESK, P. A., MCCARTHY, M. A., BUNYAVEJCHEWIN, S. & BAKER, P. J. 2015. The neglected tool in the Bayesian ecologist's shed: a case study testing informative priors' effect on model accuracy. *Ecology and Evolution,* 5**,** 102-108.

MYERS, R. H., MONTGOMERY, D. C., VINING, G. G. & ROBINSON, T. J. 2010. *Generalized Linear Models: With applications in Engineering and Science,* New Jersey, USA, John Wiley & Sons Inc.

NELDER, J. A. & WEDDERBURN, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General),* 135**,** 370-384.

O'HARA, R. B. & KOTZE, D. J. 2010. Do not log-transform count data. *Methods in Ecology and Evolution,* 1**,** 118-122.

OSBORNE, J. 2005. Notes on the use of data transformations. *Practical Assessment, Research and Evaluation,* 9**,** 42-50.

PAN, W. 2001a. Akaike's information criterion in generalized estimating equations. *Biometrics,* 57**,** 120-125.

PAN, W. 2001b. Model Selection in Estimating Equations. *Biometrics,* 57**,** 529-534.

PERNEGER, T. V. 1998. What's wrong with Bonferroni adjustments. *BMJ,* 316**,** 1236-1238.

POSADA, D. & BUCKLEY, T. R. 2004. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Systematic Biology,* 53**,** 793-808.

RAFTERY, A. E., MADIGAN, D. & VOLINSKY, C. T. 1996. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian statistics,* 5**,** 323-349.

RAUDENBUSH, S. W. & BRYK, A. S. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods,* USA, Sage Publications.

ROHDE, C. A. 2014. *Introductory Statistical Inference with the Likelihood Function,* Switzerland, Springer.

ROYALL, R. 1997. *Statistical evidence: a likelihood paradigm*, CRC press.

SAMANIEGO, F. J. & RENEAU, D. M. 1994. Toward a Reconciliation of the Bayesian and Frequentist Approaches to Point Estimation. *Journal of the American Statistical Association,* 89**,** 947-957.

SANTORA, J. A., REISS, C. S., LOEB, V. J. & VEIT, R. R. 2010. Spatial association between hotspots of baleen whales and demographic patterns of Antarctic krill Euphausia superba suggests size-dependent predation. *Marine Ecology Progress Series,* 405**,** 255-269.

STOICA, P. & SELEN, Y. 2004. Model-order selection: a review of information criterion rules. *Signal Processing Magazine, IEEE,* 21**,** 36-47.

STROUP, W. W. 2013. *Generalized Linear Mixed Models: Modern concepts, methods and applications,* Florida, CRC Press.

SUGIURA, N. 1978. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-Theory and Methods,* 7**,** 13-26.

TAPER, M. L. & PONCIANO, J. M. 2016. Evidential statistics as a statistical modern synthesis to support 21st century science. *Population Ecology,* 58**,** 9-29.

TER BRAAK, C. J. & VERDONSCHOT, P. F. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic sciences,* 57**,** 255-289.

THOMAS, L., BUCKLAND, S. T., BURNHAM, K. P., ANDERSON, D. R., LAAKE, J. L., BORCHERS, D. L. & STRINDBERG, S. 2002. Distance Sampling. *In:* EL-

SHAARAWI, A. H. & PIEGORSCH, W. W. (eds.) *Encyclopedia of Environmetrics.* Chichester: John Wiley & Sons Ltd.

TOUCHON, J. C. & MCCOY, M. W. 2016. The mismatch between current statistical practice and doctoral training in ecology. *Ecosphere,* 7**,** e01394.

VALLVERDU, J. 2016. *Bayesians vs Frequentists: A philosophical debate on statistical reasoning*, Springer.

VERBEKE, G. & MOLENBERGHS, G. 2009. *Linear Mixed Models for Longitudinal Data,* New York, Springer.

VSNI, U. K. 2009. ASReml.

WADE, P. R. 2000. Bayesian Methods in Conservation Biology. *Conservation Biology,* 14**,** 1308-1316.

WALL, M. M. 2004. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference,* 121**,** 311-324.

WASSERMAN, L. 2000. Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology,* 44**,** 92-107.

WATANABE, S. 2013. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research,* 14**,** 867-897.

WEAKLIEM, D. L. 1999. A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research,* 27**,** 359-397.

WEST, B., WELCH, K. B., GALEKI, A. T. & GILLESPIE, B. W. 2007. *Linear mixed models : a practical guide using statistical software,* Boca Raton, USA, Chapman & Hall/CRC.

WESTFALL, P. H. & YOUNG, S. S. 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment,* USA, John Wiley & Sons.

WHEELER, D. C. 2014. Geographically Weighted Regression. *In:* FISCHER, M. M. & NIJKAMP, P. (eds.) *Handbook of Regional Science.* Berlin, Heidelberg: Springer Berlin Heidelberg.

WHITTAKER, E. T. 1922. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society,* 41**,** 63-75.

WIKLE, C. K. 2003. Hierarchical Bayesian Models for Predicting the Spread of Ecological Processes. *Ecology,* 84**,** 1382-1394.

WOLD, S. 1974. Spline functions in data analysis. *Technometrics,* 16**,** 1-11.

WOLFINGER, R. 1993. Covariance structure selection in general mixed models. *Communications in Statistics - Simulation and Computation,* 22**,** 1079-1106.

WU, L. 2009. *Mixed Effects Models for Complex Data,* Boca Raton, USA, CRC Press, Taylor & Francis Group.

ZIEGLER, A. 2011. *Generalized Estimating Equations,* USA, Springer.

ZIEGLER, A., KASTNER, C. & BLETTNER, M. 1998. The generalised estimating equations: an annotated bibliography. *Biometrical Journal,* 40**,** 115-139.

ZUUR, A. F., IENO, E. N., WALKER, N. J., SAVELIEV, A. A. & SMITH, G. M. 2009. *Mixed Effects Models and Extensions in Ecology with R,* New York, Springer.

Chapter 3

MODELLING SPATIALLY AUTOCORRELATED PHYTOPLANKTON FLUORESCENCE AROUND EAST ANTARCTICA USING LINEAR MIXED MODELS WITH CUBIC SPLINES

Authors:

Lisa-Marie K. Harrison[1], Steven Candy[2], Martin J. Cox[3], Guy Williams[4,5], Robert Harcourt[1]

Affiliations:

[1] Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, New South Wales, Australia

[2] SCandy Statistical Modelling Pty Ltd, Blackmans Bay, Tasmania, Australia

[3]Australian Antarctic Division, 203 Channel Highway, Kingston, Tasmania, Australia

[4] Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Australia

[5] Antarctic Climate and Ecosystems Cooperative Research Centre, University of Tasmania, Hobart, Australia

Abstract

Productivity in the Southern Ocean is important for global oxygen levels and climate. Determining the relationship between phytoplankton density and environmental variables enables us to understand and predict the effects of environmental change on phytoplankton. Marine environmental data, such as phytoplankton fluorescence, are commonly collected using vertical profiling instruments which are time consuming to use in costly remote environments. Standard modelling techniques typically do not account for the 3D autocorrelated and non-linear nature of marine profiling data, resulting in incorrect inference and biased predictions. Here we use a spline mixed model with a 3D correlation structure to model environmental correlates with phytoplankton fluorescence collected using a Conductivity-Temperature-Depth (CTD) system during the BROKE-West research cruise along the East Antarctic margin (30-80ºE) in January-February 2006. Our modelling procedure was tested via simulation and found to be unbiased. We found that the variables depth, in situ temperature, salinity and dissolved oxygen were significant predictors of phytoplankton fluorescence. Strong spatial autocorrelation was found in the latitude and depth dimensions (φ_{depth}= 0.92, $\varphi_{latitude}$ = 0.85). Ignoring correlation led to over fitting, negatively biased variance estimates and spurious inference, highlighting the importance of considering correlation when modelling CTD data. This study identified important drivers of phytoplankton distribution in the East-Antarctic and provided a method for predicting future scenarios using the vast array of survey data that already exists.

Keywords: conductivity temperature depth, random effect, Chlorophyll-a, spatial autocorrelation, non-linear relationship

1. Introduction

The Southern Ocean is of global importance due to i) its key role in global climate (Mayewski et al., 2009), ii) its unique and endemic-rich ecosystems (Murphy and Hofmann, 2012) and iii) its hosting of multi-national fisheries, including commercial finfish and the world's largest krill fishery (Nicol et al., 2012). Key to these fisheries are the high level of primary production that occurs in the Southern Ocean (Smetacek and Nicol, 2005). This is largely due to the high phytoplankton densities which are responsible for more than half of the world's annual photosynthesis (Chisholm, 2000). In the Southern Ocean, phytoplankton are the primary food source for a number of marine species, including euphausiids, larval fish, tunicates and cephalopods (Gurney et al., 2001, Dubischar and Bathmann, 1997, Murphy et al., 2012). Accurately determining the drivers of phytoplankton abundance is therefore crucial to our understanding of productivity in the oceans as a whole.

There is increasing concern over the environmental status of the Southern Ocean because the region is experiencing unprecedented changes (Constable et al., 2014). A focal point for research has been the Western Antarctic Peninsula, which is warming significantly, and at one of the fastest rates on the planet (Montes-Hugo et al., 2009, Clarke et al., 2007). This change is characterised by decreasing winter sea-ice levels, salinification and warming surface waters (Meredith and King, 2005), with major observed changes in Chlorophyll-a distribution and phytoplankton community composition (Massom and Stammerjohn, 2010). Chlorophyll-a is a proxy for primary productivity by phytoplankton, however, as pigmentation is sensitive to photoacclimation and nutrient-driven physiological responses, it is not directly proportional to productivity (Behrenfeld et al., 2015). In contrast to the warming seen in the Western Antarctic Peninsula, the Ross Sea has experienced increased extent and duration of sea ice (Smith Jr et al., 2012). The

phytoplankton biomass there is strongly linked to sea ice, mixed layer depth and light availability (Smith Jr et al., 2014). On a global scale, it is expected that there will be large range expansions of warm-water species towards the poles, changes in abundance, growth levels, timing of peak production and trickle-down effects through the marine food web (Hallegraeff, 2010). In light of the divergent environmental shifts in the Southern Ocean, it is becoming increasingly important to understand how phytoplankton may be affected by complex regional climate change. The predicted environmental changes are too subtle to examine using perturbation experiments (Boyd et al., 2008), necessitating modelling based methods to answer these questions.

The oceans are heavily vertically stratified and experience mixing and convection, transporting heat from the tropics to the poles (Wunsch and Ferrari, 2004, Ganachaud and Wunsch, 2002). To capture the stratification in the surface layers, oceanographic data collection during ship-based marine research voyages primarily involves lowering instrumented platforms through the water column. The base unit sensor/platform is the CTD, measuring conductivity, temperature and depth. Ancillary sensors such as fluorometers and dissolved oxygen sensors can be included on the CTD package. However, establishing relationships between the fluorescence of Chlorophyll-a and coincident oceanographic properties is a complex modelling task, caused by non-linear relationships, strong autocorrelation of data within each vertical profile, and the potential for an observation station level effect on the profiles.

Generalized Additive Models (GAMs) have been used to model similar complex data, as they can account for non-linear relationships by using smoothers (Hastie & Tibshirani, 2000, Wood, 2006). In the marine setting, GAMs have previously been used to model the relationship between profiles of bioluminescent zooplankton sources and environmental

variables (Craig et al., 2010, Heger et al., 2008), drivers of phytoplankton productivity (Lamont et al., 2014), the distribution and biomass of euphausiid aggregations (Lawson et al., 2008) and nano-microplankton and meso-zooplankton biomass (Zarauz et al., 2007). Modelling with data collected at multiple sites or stations may mean a random effect is appropriate, in which case a GAM must be extended to a Generalized Additive Mixed Model (GAMM). Mixed models include fixed effects (parameters of specific interest where the levels are not randomly chosen) and random effects (parameters whose levels are not of particular interest but are required to avoid pseudoreplication). Ignoring random effects for a data set collected at many sites can result in a large loss in degrees of freedom and the inability to make population level inference (Crawley, 2002). This in turn makes it difficult to make valid predictions at a new location, where no data is used to develop the initial model.

Often present in ecological data is spatial autocorrelation, where closer points are more similar than those further away (Dormann, 2007). Spatial autocorrelation is difficult to address in marine environments since there is the potential for 3-dimensional correlation, i.e. correlation between stations and the observations collected during vertical profiling. This is especially likely if there are missing covariates (i.e. lurking variables) for which data are not available (Joiner, 1981). In a complex and difficult to quantify environment such as the Southern Ocean, it is unlikely that representative data on all variables of interest can or will be collected. Therefore, an error structure needs to be included in the model to specify the type of correlation that is still present in the residuals. Ignoring spatial autocorrelation can reduce model fit, bias parameter estimates, cause inverted relationships and result in false conclusions (Dormann, 2007, Lichstein et al., 2002, Kühn, 2007). Parameters may also become falsely significant, causing spatially varying

parameters, defined as parameters that co-vary with latitude or longitude i.e. Sea Surface Temperature, to appear more significant than they really are (Lennon, 2000).

Here, we use linear mixed models including cubic smoothing splines (Verbyla et al., 1999) combined with a 3-dimensional error structure to account for i) non-linear relationships between phytoplankton fluorescence and environmental predictors ; ii) a site level effect and iii) 3D spatial autocorrelation. The spline component allows for non-linear response vs explanatory variable relationships; the mixed effects model allows for a station random effect, where stations are considered as samples drawn from a population of stations; and the error structure also includes additional dependencies due to spatial autocorrelation arising from two-dimensional station locations as well as the third dimension, depth within the water column. We use our model to explore the generalised survey level relationships between phytoplankton fluorescence and environmental variables. Our independent variables include factors known to influence phytoplankton growth and distribution (temperature, salinity, sea ice, vertical mixing, current velocity) as well as one product of photosynthesis (dissolved oxygen). We then use simulation, based on the characteristics of our observations, to assess the validity of our modelling approach and therefore the accuracy of the conclusions we have drawn.

The data come from a multi-disciplinary marine science survey called the Baseline Research into Oceanography, Krill and the Environment (BROKE-West – see Nicol et al. (2010) and references therein). The survey took place from January – March 2006 along the East Antarctic margin south of 60^{o}S and between 30 – 80^{o}E and covered 1.3 million km^{2}. The major zonal current systems in the BROKE-West survey area are the eastward-flowing southern Antarctic Circumpolar Current front zone (sACCf) to the north and the westward-flowing Antarctic Slope Current (ASC) and coastal current to the south (Meijers et al., 2010, Williams et al., 2010). The Antarctic Circumpolar Current

(ACC) waters are characterised as warm, high nutrient, weakly stratified waters strongly influenced by wind stress (Mitchell et al., 1991). Two large-scale clockwise circulations influence the western (outer Weddell Gyre) and eastern (greater Prydz Bay Gyre/Australian-Antarctic Basin Gyre extension) boundaries of the survey. There was more sea ice in the west due to the Weddell Gyre and as a result the summer mixed layers (SMLs) were most developed from the north-east (Williams et al., 2010). The depth and thickness of the seasonal pycnocline increased from the western to eastern boundaries of the survey area, as a result of the deepening of the SML.

Phytoplankton distribution in the Southern Ocean depends on an interplay between bottom-up physical processes (light, nutrient and iron availability and mixing in the water column) and top-down grazing (Smith and Lancelot, 2004, Boyd, 2002). During BROKE-West, nitrate concentrations were more strongly regulated by uptake during photosynthesis than mixing and water masses and were correlated with dissolved oxygen levels (Pearson coefficient = 0.44). Nitrate concentration was negatively correlated with chlorophyll biomass, however silicate, phosphate and nitrate levels were all above limiting levels. While iron levels were not measured during BROKE-West due to sampling difficulty, it is thought that iron exhaustion due to grazing and sedimentation limited the growth and sustainability of blooms (Wright et al., 2010). The release of iron by melting sea ice plays an important role in the formation of phytoplankton blooms (Sedwick and DiTullio, 1997). Salinity greatly influences seawater density and the melting sea ice also creates a stable surface layer that is amenable to phytoplankton blooms (Smith and Nelson, 1986). Primary productivity was higher in the summertime sea ice zone than the open ocean, although blooms possibly associated with high iron levels were observed in the open ocean (Westwood et al., 2010). High silicate drawdown

and low assimilation numbers in the north-eastern region indicate high diatom growth earlier in the season, in conjunction with the earlier onset of sea ice melt. The number of days since full ice cover has a significant effect on phytoplankton community composition (Wright et al., 2010). Overall, the observations during the BROKE-West survey support the high-nutrient/low-chlorophyll status of the Southern Ocean (Westwood et al., 2010).

Primary productivity during BROKE-West depended heavily on mixing and water mass characteristics (Westwood et al., 2010, Wright et al., 2010), with monthly variability in surface chlorophyll explained by sea surface temperature and wind stress (Schwarz et al., 2010). In the water column, high productivity in the Marginal Ice Zone (MIZ) was associated with a shallow Mixed Layer Depth (MLD) and the MLD was shallower near the ice than in the open ocean (Westwood et al., 2010). MLD can be used to predict the upper limit of a phytoplankton bloom size, with shallower MLDs favouring the formation of large blooms (Mitchell and Holm-Hansen, 1991). In contrast to the relationships seen near the ice edge, the MLD varied considerably in the South Antarctic Circumpolar Current Zone (SACCZ) without any visible effect on phytoplankton stocks (Wright et al., 2010). On the shelf region of the Western Antarctic Peninsula, MLD also showed no correlation with Chlorophyll-a concentration or phytoplankton community composition (Prézelin et al., 2000). There were significant differences in phytoplankton taxa between the Southern Boundary and the sACCf which could be attributed to turbulent mixing and advection of Antarctic Circumpolar Current (ACC) water (Wright et al., 2010). Turbulent mixing affects the competition for light between taxa and causes a shift in community composition with mixing favouring sinking phytoplankton (Huisman et al., 2004), while Chlorophyll-a in general correlates strongly with vertical stability of the water column

(Garibotti et al., 2003). The depth of the euphotic zone, below which there is insufficient light for photosynthesis, was 55-100m at most stations and throughout the survey area light availability was above the levels required for maintenance of phytoplankton populations and promoted growth (Westwood et al., 2010). The only observed case of light limitation was self-shading at two stations near the sea-ice edge (Wright et al., 2010) and no significant difference in photosynthesis was seen between day and night time CTD stations (Westwood et al., 2010).

The factors regulating phytoplankton growth and distribution are complex and therefore simple modelling methods are unlikely to capture the spatial complexity and non-linear nature of these processes. Here our aim is to develop a predictive model of phytoplankton fluorescence based on selected environmental data collected throughout the BROKE-West survey area. The model we develop here can deal with the aforementioned problems with spatial autocorrelation and non-linearity and is widely applicable regardless of survey design.

2. Materials and methods

2.1. Oceanographic Data

We used conductivity temperature depth (CTD) data (Rosenberg, 2006) collected using the 2006 Baseline Research on Oceanography, Krill and the Environment survey (BROKE-West) in East Antarctica (see Nicol et al (2010) and the map of the survey area with CTD stations marked as circles). The CTD was a SeaBird SBE9plus with attached dissolved oxygen sensor (SBE43), fluorometer (Wet Labs ECO) and twenty two 10L Niskin bottles (General Oceanics). During the survey there were 118 CTD locations,

where data on depth, temperature, conductivity, phytoplankton fluorescence and dissolved oxygen were resolved at 2m depth intervals. The first station was a test, so only the subsequent 117 stations were included in our analysis. Only depths between the surface and 250m were used because no fluorescence was observed below 250m. The Acoustic Doppler Current Profiler (ADCP) data from the BROKE-West survey (Meijers and Klocker, 2006) were collected using an RDI 150kHz broadband ADCP (Rosenberg et al., 1999).

The fluorescence data were collected using a fluorometer attached to the CTD and calibrated using High Performance Liquid Chromatography (HPLC) pigments from water from the top 10m of the water column collected in Niskin bottles attached to the CTD rosette (Westwood et al., 2010). The calibration was performed at the time of data collection and the calibrated fluorescence values had units of μg Chl a L^{-1}. While the calibration was performed using total Chl a, it should be noted that divinyl chlorophyll (indicating the presence of Prochlorococcus) was not present.

2.2. Statistical Analysis

The analysis was undertaken using R 3.1 (R Development Core Team, 2014) running in R-Studio 0.98.932 (RStudio, 2014). The mixed models were fit using generalized least squares for fixed effect parameters combined with Residual Maximum Likelihood (REML) for variance parameters, and Best Linear Unbiased Prediction (BLUP) to estimate random effects (Diggle et al., 2002, Gilmour et al., 1995, Patterson and Thompson, 1971) using the R package ASReml-R, version 3.0 (Butler, 2009). ASReml-R was chosen over other mixed modelling packages in R because it can fit the 3D error structure required here (depth, latitude, longitude) as well as accommodating irregularly spaced CTD stations that were closely spaced near the ice edge and sparser offshore.

2.3. Linear Mixed Model Specification

Our linear mixed model uses cubic smoothing splines which are piecewise third order polynomials joined smoothly at locations within the data known as knots points, or simply knots (Wold, 1974). Knot location and number controls spline smoothness, with a higher number of knots causing the spline to more closely follow the data but at the risk of poor model development due to overfitting. A range of different knot points (5 to 50) were tested and since the model did not change depending on the number of knots for each spline, 10 knots were chosen for computational efficiency. A station random effect was fitted to allow prediction across East Antarctica. The model was fitted by minimising the Residual Maximum Likelihood (REML) criterion using the Average Information algorithm. REML is robust to misspecification of the correlation structure when using penalized splines with correlated data (Krivobokova and Kauermann, 2007). Furthermore, REML is readily accessible via ASreml (VSNi) and to our knowledge provides the only off-the-shelf solution for fitting 3D correlation structures.

All explanatory variables (Table 1) were collected *in situ* using the CTD with the exception of ice-free days (the number of days since full ice cover at the CTD station) and distance from the sea ice edge (at time of sampling), which were calculated using satellite data. These variables were extracted using remotely sensed environmental data accessed using the R package *raadtools* (Sumner and Raymond, 2015). The ice data are from the environmental data sets included with *raadtools* from the National Snow and Ice Data Centre and are available on a daily 25km resolution. The ice edge was defined using a contour that followed the convex hull of locations around Antarctica where ice cover had declined to zero. Distance to this polygon was calculated using an elliptical distance calculation to account for curvature of the earth. The vertical stratification region

(VSR) variable is a 4 level categorical variable that describes the water mass relative to the summer stratification and mixed layers/boundaries. Continuous explanatory variables were centred and scaled, as required in ASReml if missing values are present (Butler et al., 2009), and fluorescence was log transformed to normalize the residuals. These explanatory variables were chosen because they were collected on a scale that reflects the sampling stations and biologically could plausibly effect phytoplankton.

Table 1 *List of explanatory variables considered with summary statistics. For the parameter 'ice free days', negative values indicate that the region was not yet ice free at time of sampling i.e.: -35 = Region became ice free 35 days after sampling*

Explanatory Variable	Symbol	Mean	Standard deviation	Min	Max
Temperature (°C)	*t*	-0.47	1.287	-1.99	1.97
Depth in water column (m)	*z*	126	72	2	250
Dissolved oxygen (μ mol L^{-1})	*o*	293	59.25	174	407
Salinity (psu)	*s*	34.25	0.309	32.72	34.69
Ice free days (days since full ice cover)	*i*	34.69	30.77	-35	103
Distance from ice edge (km)	*d*	277	200	2.31	635
Current speed (m s^{-1})	*c*	0.11	0.09	0	0.97
CTD station (categorical)	*stn*	Factor levels 1 - 117			
Vertical Stratification Region (categorical)	*vsr*	Factor levels 1- in the summer mixed layer[1] 2- in the seasonal pycnocline 3- in the Tmin layer 4- below the Tmin layer			

[1] Mixed layer depth was visually assessed using vertical profiles of salinity, potential temperature and potential density rather than using a fixed algorithm to allow for high accuracy across the survey (Williams et al 2010). Mixed layer depths were consistent with the surface gradient offset methods used in other studies.

The full model is specified in Equation 1 using the symbols from Table 1 with the shorthand spl() to denote a spline and re() to denote a random effect. Collinearity between variables was assessed using the diagonal of the Cholesky decomposition of the correlation matrix. All values along the diagonal were > 0.37 indicating that no variables were strongly collinear.

$$\log(Fl) = p + z + o{:}vsr + t{:}\ vsr + s{:}\ vsr + i + d + c + \text{spl}(p) + \text{spl}(z) + \text{spl}(o) : vsr + \text{spl}(t){:}\ vsr + \text{spl}(s) : vsr + \text{spl}(i) + \text{spl}(d) + \text{spl}(c) + \text{re}(stn) + \varepsilon \qquad \textit{Equation 1}$$

Unlike other GAM packages, ASReml fits the cubic spline in two separate parts as a linear fixed effect term combined with a random effect spline that captures departures from linearity (Verbyla et al., 1999). Despite this specification, the splines are not true random components. Hence the station (*stn*) variable in the model above is the only true random component. The *:vsr* and spl(*):vsr* terms denote an interaction between a variable and VSR, which fits a separate intercept and spline for that variable at each level of VSR. As VSR is a proxy for mixing, interactions with dissolved oxygen, salinity and temperature were considered. Pair plots revealed a very weak correlation between temperature and ice free days, however there was so much variance that an interaction term was not considered.

Starting with the full model including all explanatory variables (Equation 1), backwards selection with Akaike Information Criteria was used to select the best model (Cheng et al., 2010). Conditional R^2 and Root Mean Square Error (RMSE) were used to assess model goodness of fit. The conditional R^2 value takes into account variation explained by both the fixed and random effects (Nakagawa and Schielzeth, 2013).

The vertical profiles from the BROKE-West CTD data were collected at stations with a fixed position; the potential exists for data to be correlated in three-dimensions (latitude, longitude and depth). We used an autoregressive first order AR(1) process to account for correlation with depth at each station (Butler et al., 2009) and an anisotropic Gaussian surface across latitudes and longitudes at each depth. To include this correlation structure in the mixed model specification we replace the conventional homogenous error variance structure, $\mathbf{R} = \sigma^2 I$, with a 3D correlation structure (Equation 2). In Equation 2 the typical element of the modelled correlation matrix is specified for observations defined for two generalised stations indexed by i and j (where $i=j$, or $i \neq j$), with corresponding latitudes and longitudes and a pair of depths indexed by k and k' (where $k=k'$ or $k \neq k'$) using the direct product of matrices formulation (Butler et al., 2009).

$$[\mathbf{R}_{\mathbf{spatial}} \otimes \mathbf{R}_{\mathbf{depth}}] = \phi_{lat}^{(lat_i - lat_j)^2} \phi_{long}^{(long_i - long_j)^2} \phi_{depth}^{(depth_k - depth_{k\prime})^2}$$

Equation 2

The three ϕ parameters are the correlation coefficients in each direction and have absolute values less than one. For two stations, i and j, the term $(lat_i - lat_j)^2$ represents the squared latitudinal distance (km) between the stations and $(long_i - long_j)^2$ is the squared longitudinal distance (km). Since observations are resolved at equal depth intervals of 2 m, these unit intervals allow a pure AR(1) process to be fitted. The AR(1) process is assumed identical at each station so that the term $(Depth_k - Depth_{k'})$ where $k'=k-1$ (i.e. proceeding down the water column) in Equation 2, can be set to 1 for all k. This AR(1) error model gives corresponding model error term:

$\varepsilon_{ijk} = \varphi_{depth}\ \varepsilon_{ij,k-1} + Z_{ijk}$ where Z is a random Gaussian (white-noise) error with variance $\sigma^2(1-\varphi_{depth}^2)$ and σ^2 is the residual variance (Diggle et al., 2002). Note also that the random station effect introduces a further additive, constant, and positive covariance between observations at different depths for the same station where this covariance is equal to the station random effect variance, σ^2_s. The correlation between model residuals for a pair of adjacent depths, adjusted for the spatial autocorrelation terms, is $(\sigma^2_s + \varphi_{depth}\ \sigma^2)/(\sigma^2_s + \sigma^2)$ where REML parameter estimates are obtained by implicitly averaging across stations.

2.4. Simulation Study

A simulation study was developed to verify that ASReml was able to accurately return the variance components of a complex model. This simulation was designed to mirror the BROKE-West CTD data as closely as possible. We used the coordinates of each BROKE-West CTD station and the same depths to ensure that the data set maintained the irregular spatial nature of the real data. Temperature and Photosynthetically Active Radiation (PAR) trends were built in as explanatory variables using exponential and Weibull probability density functions and these were combined to create a fluorescence explanatory variable.

To make the simulation more realistic, random error was generated and applied to the simulated trends. This also allowed us to check that the model would not fit to noise and would correctly identify random error, inter-station variation, spatial autocorrelation and actual relationships with independent variables. To mirror the real data, the generated random error was correlated in the latitude, longitude and depth planes using an AR1

process for the depth at each station and an anisotropic Gaussian surface across latitudes and longitudes. To do this, a random value was generated for each point from the normal distribution, with a fixed noise standard deviation. The anisotropic Gaussian surface was calculated using the distance between each station and using the formula (φ_x^dist_x) * (φ_y^dist_y). The inverse of the Cholesky factorisation of this matrix is then multiplied by the random error values previously generated. The AR1 depth process is then included by adding $\varphi_z * k$' to each value, where k' = k – 1, the previous depth value at a station. A station random effect was added using randomly generated values from a normal distribution, with mean = 0 and a fixed standard deviation across all stations.

The model fit, using this simulated data set, was of the same form as the model used for the BROKE-West CTD data. Model selection was not done because we were mainly interested in assessing whether the variance components could be estimated correctly. This simulation was run 200 times and the estimated variance component for each variable was calculated for each simulated data set. For each simulation, a new simulated data set was generated using the same input variance values. This allowed each data set to be different due to randomness but have the same global variance components for the model to estimate.

2.5. Cross-validation

Due to the logistic costs of Antarctic oceanographic surveys of this nature, a second data set was not available for model validation. Therefore 6-fold cross-validation was run on the BROKE-West data set to assess how well the model could predict the observed values at stations excluded in each run. As there were 6 vertical transects aligned in the north-

south direction in the BROKE-West survey design, K was chosen to be 6. The 6 vertical transects were dropped one-by-one and the model was fitted to the horizontal transect and the 5 remaining vertical transects. The log(fluorescence) at each station in the dropped transect was then predicted using the explanatory variables: observed depth, temperature, dissolved oxygen, salinity and VSR.

3. Results

3.1. Simulation Study

The model accurately estimated all variance components over the 200 simulations (Table 2). The average conditional R^2 value across the 200 fitted models was 0.997 indicating that the model had a high goodness of fit to each simulated data set. A very high R^2 value such as this is expected from a model with good fit because the noise component in the simulation is simplistic compared to real life scenarios and there are not any variables used to form the data that aren't input into the model.

Table 2 *Relative bias and 95% Confidence Interval (CI) Coverage Probabilities of variance component estimators (n = 200 simulations).*[1]

	True variance component ($\boldsymbol{\theta}$)	**Relative Bias ($E[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta}$)/ $\boldsymbol{\theta}$**	**95% Confidence Interval**
Station random effect (*stn*)	0.22	0.0021	(0.217, 0.221)
Error variance (ε)	0.45	-0.0019	(0.448, 0.450)
Latitude correlation ($\varphi_{latitude}$)	0.50	0.0001	(0.4996, 0.5003)
Longitude correlation ($\varphi_{longitude}$)	0.40	0.0035	(0.399, 0.402)
Depth correlation (φ_{depth})	0.35	-0.0074	(0.346, 0.349)

3.2. BROKE-West model fit

The variables – distance from ice edge, ice-free days and current speed – and interactions with VSR, salinity and dissolved oxygen were dropped during backward AIC-based model selection. Table 3 gives the parameter estimates and variance components for all variables in the best model. Note: the temperature values have been averaged across the 4 VSRs to give an average for the full survey area. The conditional R^2 value is 0.81, indicating that a high proportion of the total observed variance is explained by the model. The temperature spline variance component was ten times smaller than the other variance components. As the variance components are related to the splines' smoothing parameters, this indicates that the temperature spline relationship is closer to a straight line than the other variables, rather than a smaller contribution to the model (Verbyla et al., 1999).

Table 3 *Parameter estimates (on log scale) and variance components for fixed and random effects. The spline variance components are equivalent to smoothing parameters*

	Fixed Effects	
	Parameter estimate	Spline variance component
Temperature (°C)	-0.453	0.024
Depth (m)	-0.016	0.338
Dissolved oxygen (μ mol L^{-1})	-0.434	0.130
Salinity (psu)	-0.283	0.400

	Random effects and residual variance	
	Variance	Standard Error
Station	0.292	0.061
Residual	0.970	0.052
Depth correlation	0.934	0.004
Latitude correlation	0.841	0.046
Longitude correlation	0.000	NA

3.3. Spatial autocorrelation

Spatial autocorrelation in the depth and latitude directions was estimated to be very high with $\emptyset_{depth}$= 0.92 (SE = 0.004) and $\emptyset_{lat}$ = 0.85 (SE = 0.04) respectively while there was no correlation in the longitude direction with $\emptyset_{long}$ = 9.35 x 10^{-8} (SE = NA). Overall the model tends to underestimate the true fluorescence, especially at the higher fluorescence values near the ice edge (figure 1).

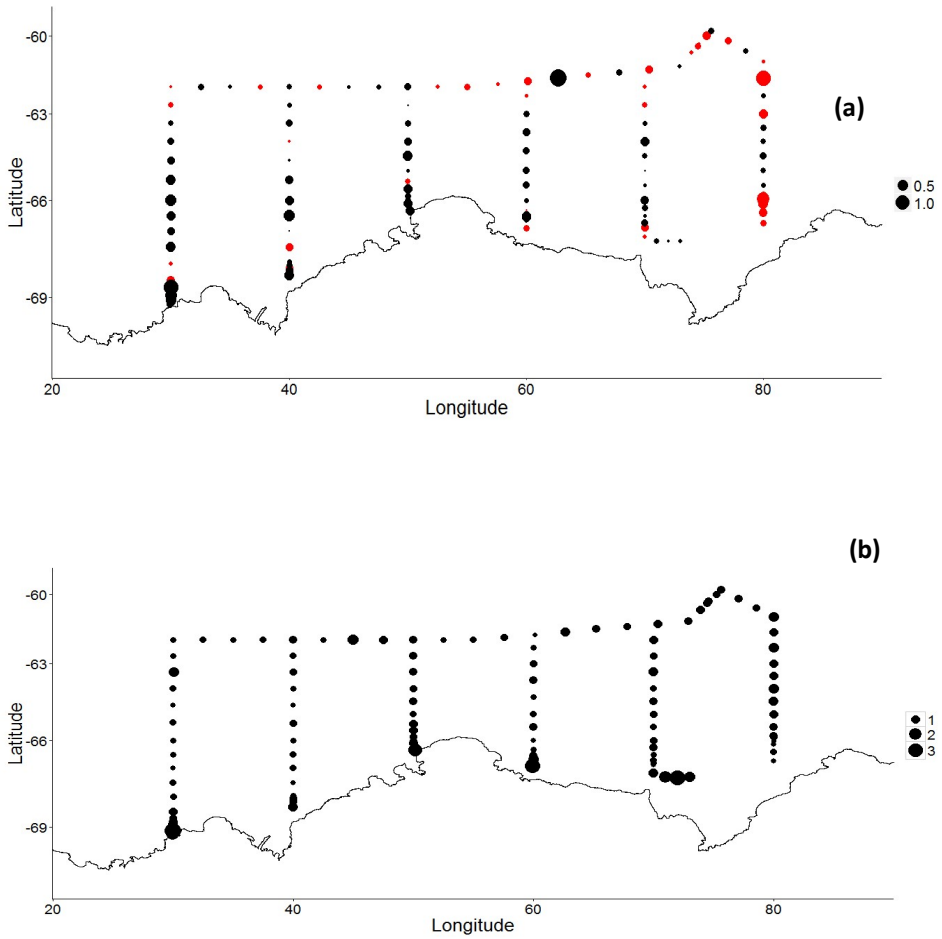


Figure 1 *(a) Mean conditional residual at each station. Negative = red, Positive = black. Random terms are included in the residuals and circle size is proportional to the absolute value of the residual; (b) Mean fluorescence (ug* L^{-1}*) measured at each station. Higher mean fluorescence is evident at stations near the ice edge. The Antarctic coastline was sourced from the National Snow and Ice Data Centre (Haran et al., 2005, Scambos et al., 2007).*

3.4. Average trends

The average spline trends between each explanatory variable and fluorescence in the presence of other fixed effects were extracted from the model to quantify how fluorescence varies with each variable (Figure 2).

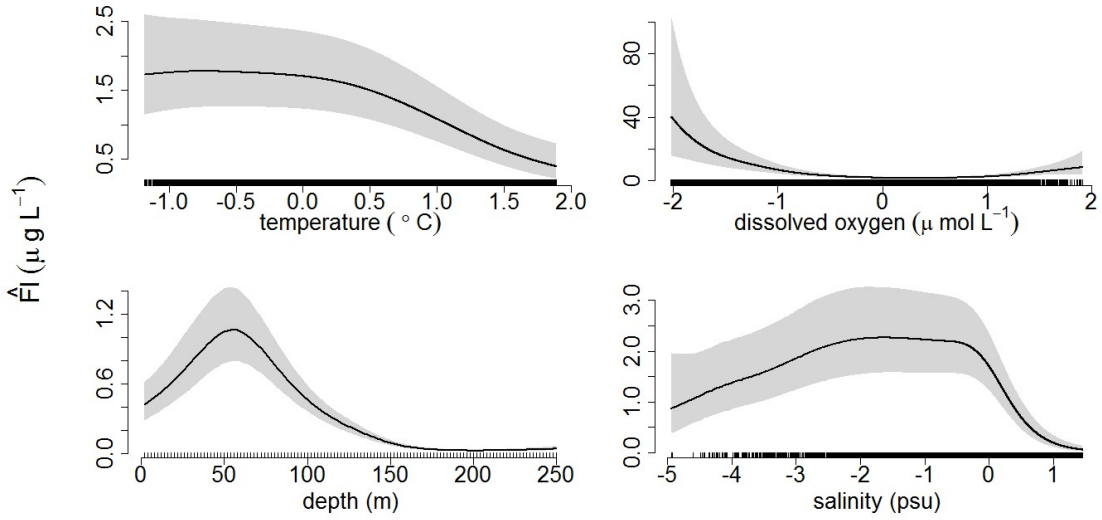


Figure 2 *Average partial smooth relationships for fluorescence (µg L^{-1}) vs environmental variables with 95% confidence envelopes. Smooth terms are calculated as the addition of the linear fixed effect and random spline term for each parameter. All variables except depth have been centred and scaled to mean = 0 and standard deviation = 1.*

The average fluorescence-depth trend shows a Chlorophyll maximum at a depth of 50m. Average fluorescence values at the Chlorophyll maximum were 1.0 µg L^{-1}. There is decreasing uncertainty (as indicated by the narrowing grey confidence envelopes) as depth increases and fluorescence tends to zero because the observed fluorescence at all stations was near-zero at depths greater than 150m, compared with higher variation seen at shallower depths.

3.5. Correlation structure

A likelihood ratio test between the model with (full model) and without (null model) the correlation structure indicated the correlation structure significantly improves the model fit ($P < 0.0001$; Table 4).

Table 4 *Model selection results for intercept model (no correlation structure), null model (no correlation structure) and full model (with correlation structure).*

	K	Log likelihood	AIC	ΔAIC	p
Intercept Model	1	-11482	22966		
Null Model	6	-2216	4445	-18521	<0.0001
Full Model	9	5553	-11089	-15531	<0.0001

The standardised conditional residual autocorrelation function (ACF) for the null models showed a particularly strong correlated trend in the depth dimension compared with the full model (Figure 3). While there appears to still be a small amount of correlation present, it is clear that the error structure has reduced the depth autocorrelation (Figure 3). The negative spike for the full model at lag 1 is caused by most stations still showing a small amount of residual correlation at lag 1. Furthermore, the average trends for temperature and salinity were overfit, picking up too much curvature from the data (Figure 4). In some circumstances sharp boundary layers, such as the low salinity pocket that surrounds the ice edge, could cause a spline to appear overfit while reflecting a real difference, however it is unlikely that that this is occurring in Figure 4 because the splines are consistently undulating and this is not reflected in the real data.

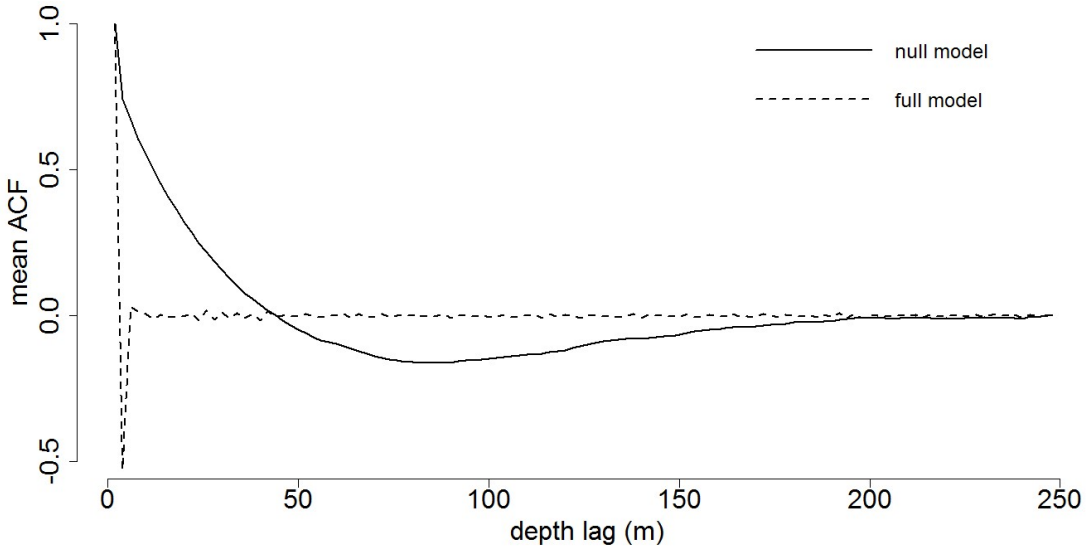


Figure 3 *Mean autocorrelation of standardised conditional residuals in depth dimension for model with (full) and without (null) a correlation structure.*

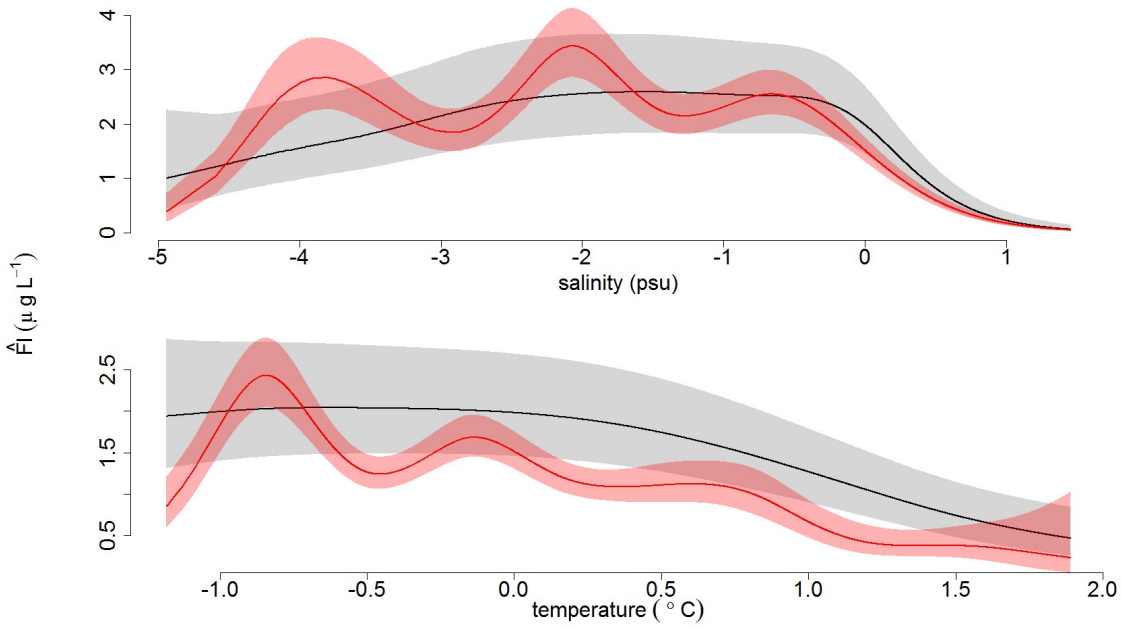


Figure 4 *Salinity and temperature average trends for the null model without a correlation structure (red) compared to the full model (black). Shaded areas are 95% confidence envelopes. Salinity and temperature have been centred and scaled as a requirement for variables with missing values in ASReml.*

3.6. Cross-validation

The overall RMSE for the cross validation is 0.89. This value is calculated using only predicted values at stations that were not used in the model fit and hence will be higher than the RMSE for the fitted model. The RMSE is approximately 8% of the range of the log(fluorescence) values, which indicates a low standard deviation of unexplained variance and good predictive strength. In general, the goodness of fit varied between stations (Figure 5). The station random effect accounts for some of the unknown variable influencing the station variances when calculating fitted values, however it cannot be used directly for predictive purposes because there is no random effect estimate for a novel station. This may result in some stations being poorly predicted because they are influenced by a factor that was not measured. One method to account for this is to generate a new random effect estimate for the novel station using the estimated random effect variance from the other stations. The station random effect may account for some of the spatial correlation present in the data set and hence be unnecessary in some situations because the model contains a correlation structure. In the simulation study it was found that the ASReml model assigned the spatial autocorrelation to the correct component of the correlation structure, with the station random effect only picking up the extra random station variation in the model.

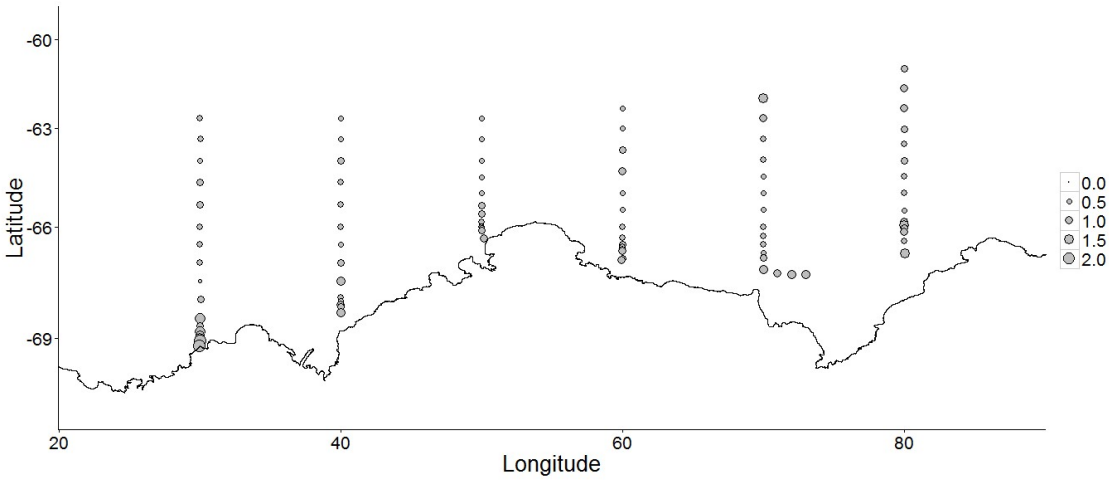


Figure 5 *RMSE for each station (n = 92 predicted stations) shows a range in goodness of fit across stations, with stations near the sea ice edge having the poorest predictions. Antarctic coastline as in previous figure.*

4. Discussion

This study successfully quantified phytoplankton-environment relationships using data collected offshore from the East Antarctic margin during the Broke-west 2006 survey, revealing that phytoplankton levels correlate most strongly with temperature and dissolved oxygen levels. Unusually high Chlorophyll-a levels in comparison to previous cruises were seen during the North-Eastern region of this survey and along the coastline, while unusually low levels were seen to the west (Schwarz et al., 2010), which may be attributed to a negative Southern Annular Model index and coastal upwelling. Our study aimed to identify environmental drivers of phytoplankton distribution.

Central to the success of this study was the application of a linear mixed modelling approach that included a 3D correlation structure based on observation location (latitude/longitude) and depth, as well as random effects of the observation unit ("CTD station" in this study) that accounted for inter-station differences. We did not input station as a fixed effect because this would result in a loss of 116 degrees of freedom as there were 117 stations and to do so may have complicated inference and lead to model convergence problems. We were also more interested in generalisation of the model across the survey area, as exactly the same CTD stations would unlikely be chosen if the survey were repeated.

Our model performed well at estimating unbiased random effects parameters under a range of simulation scenarios, thereby ensuring population level processes, such as temperature and salinity processes, are accurately represented to ensure unbiased inference. The inclusion of random effects was important to avoid pseudoreplication of data within stations, and accurate estimation of the variance components was important to avoid bias in the fixed effects, enabling population level processes within the survey area to be inferred. Biased fixed effects can also occur due to miss-specification of the correlation structure, a problem which is not remedied by increasing sample size (Gurka et al., 2011). We chose a biologically realistic correlation structure where data in the depth plane were subject to autoregressive level 1 correlation while in the latitude-longitude plane they were subject to anisotropic Gaussian correlation. The purpose of the simulation was to check that the model could correctly estimate the variance components and attribute them to the correct source (random error, inter-station variation, spatial autocorrelation or actual relationships with independent variables) rather than assess goodness of fit, so the high R^2 value is not a concern.

Whilst it is attractive to work with simple models such as simple linear regression, these models are unsuitable for complex data as they cannot accommodate correlated data or non-linear relationships and hence are unable to facilitate population level inference in this complex environment. We believe that our model provides a powerful, comprehensive tool to predict how anticipated environmental changes could affect the abundance of phytoplankton.

This model is particularly suitable for marine surveys due to its ability to accommodate irregular sampling with a 3d correlation structure. Vertical profile data sets are ubiquitous in marine science and our modelling method could easily be applied retrospectively, regardless of the spatial arrangement of the vertical profile stations. Drivers of phytoplankton variability including temperature, salinity, current velocity, sea ice presence and nutrients are collected on all oceanographic cruises and hence a large amount of data already exists. The model allows for general survey-wide trends to be established as well as accounting for 3-dimensional autocorrelation, an important advance on many previous studies. The relaxing of the regular grid constraint is important as much marine sampling is conducted in harsh environments, such as the Southern Ocean, where ensuring complete sampling of all stations in pre-designed surveys is difficult. The regular spacing of stations on a grid may also be impossible due to environmental features or weather conditions.

4.1. Spatial autocorrelation in a 3D survey area

Using the BROKE-West data, we have shown that ignoring 3D correlation adversely affects inference through severe over-fitting of the temperature and salinity splines, as seen by the extra curvature in Figure 4. We found correlation in only two of the three dimensions (latitude and depth), which may in part be explained by environmental conditions varying more with latitude than longitude with the former relating to proximity

to the Antarctic sea ice edge and the latter not. Stations along more extreme environmental gradients will display more similarity to nearby stations than more distant stations. This could also reflect the survey design since transects were run in the north-south plane and stations within transects were closer both spatially and temporally to each other than they were to other transects. Temporal autocorrelation will not be discussed separately here because the survey design makes it difficult to discern temporal effects from spatial effects.

4.2. Environmental parameters influencing phytoplankton distribution

In the Southern Ocean the regulation of primary productivity is a complex interplay of many factors (Geider and La Roche, 1994). The correlation detected in the latitudinal direction likely indicates missing environmental variables. Silicic acid and iron concentration have been flagged as important limiting factors (Martin et al., 1990, Boyd, 2002, Boyd et al., 2000) but were not measured during BROKE-West. Large seasonal phytoplankton blooms are often seen in the marginal ice zone during the period of ice melt after winter, the size and location of which is influenced by wind strength, vertical mixing and grazing pressure (Lancelot et al., 1993). The primary ice edge bloom during BROKE-West is reported to have occurred 35 days before ice melt (Wright et al., 2010). Ice edge blooms proved problematic for our model's predictions, with under-prediction occurring at most ice edge stations. While the station random effect will partly account for differences in ice-edge stations, it would also be possible to partition the model into different systems using indicator variables or even fitting separate models if there were obviously distinct systems present in a survey area. It is interesting that the 'distance to ice edge' variable was not retained in the model when there is an edge effect remaining in the residuals. One reason may be that the edge effect was less pronounced, or even

negative on the later transects during the survey which may be due to the temporal difference in when these transects were sampled. The transects sampled towards the end of the survey were closer to winter than summer and ice melt may have started to decline. A possible temporal effect could be considered in future work.

4.3. Salinity

Our model identified that fluorescence headed towards zero in areas of high salinity, which may be more influenced by depth gradients rather than indicating a latitudinal or longitudinal pattern in the surface waters since the highest salinity waters were below the euphotic zone at over 150m deep, where there was hence no fluorescence. This could also reflect the ice-edge melt which causes a pocket of low-salinity waters whose stability assists in the formation of phytoplankton blooms (Smith and Nelson, 1986). To further explore the relationship between salinity and fluorescence, it would be useful to compare the relationship between these variables in the surface mixed layer and t_{min} layers. Information on species composition could help us understand why we found low fluorescence outside the salinity range of 33.4 – 34.3 psu, with diatoms believed to correlate more strongly with salinity than flagellates (Kang and Lee, 1995).

4.4. Temperature

Satellite data during the BROKE-West survey identified Sea Surface Temperature and wind stress as the factors most correlated with monthly surface Chlorophyll-a (Schwarz et al., 2010). It was noted during the survey that temperature was relatively invariant across the survey area and its functional relationship with fluorescence would be difficult to discern (Wright et al., 2010). Our model has allowed us to quantify this relationship while accounting for vertical stratification and other environmental parameters and we

found that across the survey area there was low fluorescence at temperatures above 1^{O}C. This is likely because these higher temperatures were present either at the surface or due to Circumpolar Deep Water intrusions at depths deeper than 100m, where there is generally low availability of light for photosynthesis depending on the currents and mixing at the station. While we found a decrease in fluorescence at high temperatures, an incubation experiment showed that low temperatures may be a limiting factor controlling phytoplankton growth and nutrient uptake (Reay et al., 2001).

4.5. Dissolved Oxygen

Dissolved oxygen was one of the strongest predictors in the model and was the only variable that was a product of photosynthesis rather than a direct influence on fluorescence. While it might seem illogical to include a variable that does not influence fluorescence, as a product of photosynthesis dissolved oxygen is a strong indicator of phytoplankton presence, and is measured on all oceanographic surveys, so we have included it in our model. We found that fluorescence was highest at both low and high levels of dissolved oxygen, and lower at moderate levels. Phytoplankton produce oxygen during photosynthesis and while the observed positive relationship between phytoplankton and dissolved oxygen is expected, we do not know what proportion of our dissolved oxygen measurements are phytoplankton derived. Other factors may explain the more puzzling increase in fluorescence at lower dissolved oxygen levels, for example, mixing and currents may have transported phytoplankton to a new location, where they have been observed before photosynthesis has occurred. Alternatively, nutrient limitation could inhibit photosynthesis despite there being a large amount of phytoplankton, or high respiration by grazers and bacteria could balance production of oxygen by photosynthesis. A longitudinal study with measurements on nutrients, currents, Chlorophyll-a and dissolved oxygen may be better able to separate these relationships.

As dissolved oxygen levels are a product of photosynthesis rather than a direct driver of phytoplankton growth or distribution, this parameter is primarily of interest when modelling fluorescence in oceanographic surveys, rather than in models forecasting future scenarios.

4.6. Sea ice and currents

Ice concentration, distance from the ice edge and current strength did not contribute any explanatory power and were not included in the final model. Ocean currents have been shown to influence circumpolar phytoplankton distribution (Sullivan et al., 1993). In the BROKE-West study, ocean current strength varied little throughout the survey area and in addition, malfunctioning equipment resulted in large areas with no data recorded. Therefore, it is unsurprising that it made little contribution. More surprising was the lack of statistical significance of ice levels as a predictor, given the importance of nutrient release by melting ice to phytoplankton bloom formation (Sedwick and DiTullio, 1997, Gerringa et al., 2012). Time since full ice cover affected phytoplankton community composition during BROKE-West (Wright et al., 2010), however our results indicate that it had a minimal effect on overall fluorescence. This may be a scale issue and requires further exploration. Distance from ice edge was also expected to be an important variable because the physical and biological ocean dynamics change drastically depending on the proximity to the mainland and most transects were traversed longitudinally. For example, microbial grazing on phytoplankton was higher at the western ice-edge (>100% primary production d^{-1}) than the survey wide average (65% primary production d^{-1}) (Pearce et al., 2010). Distance from ice edge could also have acted as a proxy for missing nutrient data, if a latitudinal nutrient gradient influenced phytoplankton distribution or density.

4.7. Comparison to other studies

Our model is widely applicable in analysing marine data to produce a predictive model based on environmental parameters having dealt with 3D autocorrelation, non-linear relationships and multiple sampling stations that were neither on a regular grid nor randomly placed. There have been several comprehensive studies which have analysed a range of marine data and environmental conditions around the world with characteristics similar to our study, including the assessment of drivers of bioluminescent zooplankton in the Mediterranean Sea (Craig et al., 2010), phytoplankton production in the Benguela upwelling system (Lamont et al., 2014), krill biomass estimation along the West Antarctic Peninsula (Lawson et al., 2008) and oceanographic effects on meso-zooplankton and nano-microplankton in the Bay of Biscay (Zarauz et al., 2007). The data sets used in these studies were diverse and differed in their aims, but we believe our model could be used to improve upon their foundations by including random effects for multiple sampling stations, splines for the non-linear relationships seen in many of the studies and a 3-D irregularly-gridded correlation structure to account for any spatial autocorrelation that may be present, regardless of sampling design.

4.8. Future developments

This study advances modelling techniques commonly used in the marine environment by incorporating 3D spatial autocorrelation, non-linear relationships and random effects. Based on our simulations and the prediction of missing data with a high degree of confidence, we believe that our model, developed using the BROKE-West 2006 survey, provides the ability to accurately predict phytoplankton fluorescence from commonly collected oceanographic data. However to fully validate this model, especially in other areas of the Antarctic, additional surveys will be required. An interesting extension to the simulation study would be to assess the Signal to Noise Ratio (SNR) to ensure that the

model is robust enough to correctly estimate fixed effects with high, medium and low noise and correlation levels. SNR was not investigated in our simulation study because we were primarily interested in whether the model could accurately recover the different components of spatially autocorrelated error.

There are many avenues to explore to further this work. Iron limitation in the open ocean is a possible reason for the higher productivity in the MIZ compared to the open ocean, and nitrate levels are strongly correlated with depth integrated productivity (Westwood et al., 2010) so the inclusion of bottled nutrient data could add valuable extra information. The inclusion of mixed layer depth/depth of the euphotic zone could also be investigated to account for the effect of mixing on phytoplankton photoadaptation. As our model allows for predictions in a 3D environment, the incorporation of animal tag and active acoustics data is possible. For example, tagged elephant seals can provide information on foraging behaviour (Jouma'a et al., 2015), environmental conditions surrounding prey fields (Vacquié-Garcia et al., 2015) and oceanographic features such as bottom water production (Williams et al., 2016). Utilising active acoustics to include krill distribution as a low level predator would also be a valuable addition to the model. Our model could also be used to make 3D predictions under different future environmental scenarios, and a longitudinal analysis would be possible if data were available over multiple sampling periods.

5. Conclusion

Marine profile data are very common due to the need to quantify and understand layering and mixing in the ocean. Sophisticated modelling techniques are often necessary to understand complex ecosystems such as this, especially where the method of data collection introduces additional problems such as spatial autocorrelation. Marine data are inherently spatially complex, due to the 3-dimensional survey area and difficulty in sampling at regularly spaced locations. Simple modelling methods are often inappropriate for complex data such as this because they cannot model non-linear relationships and do not deal with 3D spatial autocorrelation or site random effects, both of which could lead to inaccurate inference. Our model offers a robust method to make unbiased population level inference about non-linear organism-environment relationships in a 3-dimensional study area and make predictions of potential change under different environmental scenarios.

We used spline mixed models with a 3-dimensional correlation structure to model phytoplankton-environment relationships from CTD data across the BROKE-West survey area in East Antarctica. We quantified the partial responses of phytoplankton fluorescence to temperature, salinity and dissolved oxygen levels. Temperature and dissolved oxygen levels were most strongly correlated with fluorescence, while distance from ice edge, current strength and number of ice free days had no effective predictive power and were not included in the final model. The inclusion of iron levels, predation by krill and phytoplankton community structure may further improve the model. Despite these missing covariates, the model performed well under simulation. The correlated residuals and over-fitted spline trends seen when omitting the error structure highlight the need to include spatial correlation. Our modelling method could be extended to describe the 3D habitat surrounding animal tag data or make predictions based on future

expected climactic scenarios and is widely applicable to both marine and terrestrial data, regardless of sampling design.

7. Acknowledgements

We would like to thank Steve Nicol for his advice and Mike Sumner (AAD) for extracting the 'distance from ice edge' variable. M.J.C. was funded by Australian Research Council grant FS11020005. L-M.H. is funded by a Macquarie University Research Excellence Scholarship. G.D.W. is funded by the Australian Research Council, the Antarctic Climate and Ecosystem Cooperative Research Centre and the Centre of Excellence for Climate System Science.

8. References

BEHRENFELD, M. J., O'MALLEY, R. T., BOSS, E. S., WESTBERRY, T. K., GRAFF, J. R., HALSEY, K. H., MILLIGAN, A. J., SIEGEL, D. A. & BROWN, M. B. 2015. Revaluating ocean warming impacts on global phytoplankton. *Nature Climate Change,* 6**,** 323-330.

BOYD, P. W. 2002. Environmental factors controlling phytoplankton processes in the Southern Ocean. *Journal of Phycology,* 38(5)**,** 844-861.

BOYD, P. W., DONEY, S. C., STRZEPEK, R., DUSENBERRY, J., LINDSAY, K. & FUNG, I. 2008. Climate-mediated changes to mixed-layer properties in the Southern Ocean: assessing the phytoplankton response. *Biogeosciences,* 5(3)**,** 847-864.

BOYD, P. W., WATSON, A. J., LAW, C. S., ABRAHAM, E. R., TRULL, T., MURDOCH, R., BAKKER, D. C., BOWIE, A. R., BUESSELER, K. & CHANG, H. 2000. A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization. *Nature,* 407(6805)**,** 695-702.

BUTLER, D. 2009. asreml: asreml() fits the linear mixed model. R package version 3.0 ed.: VSN International, UK.

BUTLER, D. G., CULLIS, B. R., GILMOUR, A. R. & GOGEL, B. J. 2009. Mixed models for S language environments: ASReml-R reference manual. Queensland Department of Primary Industries and Fisheries and NSW Department of Primary Industries, VSNi, UK

CHENG, J., EDWARDS, L. J., MALDONADO-MOLINA, M. M., KOMRO, K. A. & MULLER, K. E. 2010. Real longitudinal data analysis for real people: building a good enough mixed model. *Statistics in medicine,* 29(4)**,** 504-520.

CHISHOLM, S. W. 2000. Oceanography: Stirring times in the Southern Ocean. *Nature,* 407(6805)**,** 685-687.

CLARKE, A., MURPHY, E. J., MEREDITH, M. P., KING, J. C., PECK, L. S., BARNES, D. K. A. & SMITH, R. C. 2007. Climate change and the marine ecosystem of the western Antarctic Peninsula. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 362(1477)**,** 149-166.

CONSTABLE, A. J., MELBOURNE-THOMAS, J., CORNEY, S. P., ARRIGO, K. R., BARBRAUD, C., BARNES, D. K. A., BINDOFF, N. L., BOYD, P. W., BRANDT, A., COSTA, D. P., DAVIDSON, A. T., DUCKLOW, H. W., EMMERSON, L., FUKUCHI, M., GUTT, J., HINDELL, M. A., HOFMANN, E. E., HOSIE, G. W., IIDA, T., JACOB, S., JOHNSTON, N. M., KAWAGUCHI, S., KOKUBUN, N., KOUBBI, P., LEA, M.-A., MAKHADO, A., MASSOM, R. A., MEINERS, K., MEREDITH, M. P., MURPHY, E. J., NICOL, S., REID, K., RICHERSON, K., RIDDLE, M. J., RINTOUL, S. R., SMITH, W. O., SOUTHWELL, C., STARK, J. S., SUMNER, M., SWADLING, K. M.,

TAKAHASHI, K. T., TRATHAN, P. N., WELSFORD, D. C., WEIMERSKIRCH, H., WESTWOOD, K. J., WIENECKE, B. C., WOLF-GLADROW, D., WRIGHT, S. W., XAVIER, J. C. & ZIEGLER, P. 2014. Climate change and Southern Ocean ecosystems I: how changes in physical habitats directly affect marine biota. *Global Change Biology,* 20(10)**,** 3004-3025.

CRAIG, J., JAMIESON, A. J., HUTSON, R., ZUUR, A. F. & PRIEDE, I. G. 2010. Factors influencing the abundance of deep pelagic bioluminescent zooplankton in the Mediterranean Sea. *Deep Sea Research I,* 57**,** 1474-1484.

CRAWLEY, M. J. 2002. *Statistical Computing: An Introduction to Data Analysis using S-Plus,* West Sussex, England, John Wiley & Sons Ltd.

DIGGLE, P., HEAGERTY, P., LIANG, K.-Y. & ZEGER, S. 2002. *Analysis of longitudinal data*, Oxford University Press.

DORMANN, C. F. 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global ecology and biogeography,* 16(2)**,** 129-138.

DUBISCHAR, C. D. & BATHMANN, U. V. 1997. Grazing impact of copepods and salps on phytoplankton in the Atlantic sector of the Southern ocean. *Deep Sea Research Part II: Topical Studies in Oceanography,* 44(1–2)**,** 415-433.

GANACHAUD, A. & WUNSCH, C. 2002. Oceanic nutrient and oxygen transports and bounds on export production during the World Ocean Circulation Experiment. *Global Biogeochemical Cycles,* 16(4)**,** 1057.

GARIBOTTI, I., A. , VERNET, M., FERRARIO, M., E. , SMITH, R. C., ROSS, R. M. & QUETIN, L. B. 2003. Phytoplankton spatial distribution patterns along the western Antarctic Peninsula (Southern Ocean). *Marine Ecology Progress Series,* 261**,** 21-39.

GEIDER, R. & LA ROCHE, J. 1994. The role of iron in phytoplankton photosynthesis, and the potential for iron-limitation of primary productivity in the sea. *Photosynthesis Research,* 39(3)**,** 275-301.

GERRINGA, L. J., ALDERKAMP, A.-C., LAAN, P., THUROCZY, C.-E., DE BAAR, H. J., MILLS, M. M., VAN DIJKEN, G. L., VAN HAREN, H. & ARRIGO, K. R. 2012. Iron from melting glaciers fuels the phytoplankton blooms in Amundsen Sea (Southern Ocean): Iron biogeochemistry. *Deep Sea Research Part II: Topical Studies in Oceanography,* 71**,** 16-31.

GILMOUR, A. R., THOMPSON, R. & CULLIS, B. R. 1995. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics,* 51(4)**,** 1440-1450.

GURKA, M. J., EDWARDS, L. J. & MULLER, K. E. 2011. Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine,* 30(22)**,** 2696-2707.

GURNEY, L., FRONEMAN, P., PAKHOMOV, E. & MCQUAID, C. 2001. Trophic positions of three euphausiid species from the Prince Edward Islands (Southern

Ocean): implications for the pelagic food web structure. *Marine Ecology Progress Series,* 217**,** 167-174.

HALLEGRAEFF, G. M. 2010. Ocean climate change, phytoplankton community responses, and harmful algal blooms: a formidable predictive challenge. *Journal of Phycology,* 46(2)**,** 220-235.

HARAN, T. J., BOHLANDER, J., SCAMBOS, T., PAINTER, T. & FAHNESTOCK, M. 2005. MODIS Mosaic of Antarctica 2003-2004 (MOA2004) Image Map, Version 1. [Coastline] (updated 2013). Boulder, Colorado USA: NSIDC: National Snow and Ice Data Center.

HASTIE, T.J., & TIBSHIRANI, R.J., 1990, Generalized Additive Models. *Chapman & Hall*

HEGER, A., IENO, E. N., KING, N. J., MORRIS, K. J., BAGLEY, P. M. & PRIEDE, I. G. 2008. Deep-sea pelagic bioluminescence over the Mid-Atlantic Ridge. *Deep Sea Research Part II: Topical Studies in Oceanography,* 55(1–2)**,** 126-136.

HUISMAN, J., SHARPLES, J., STROOM, J. M., VISSER, P. M., KARDINAAL, W. E. A., VERSPAGEN, J. M. H. & SOMMEIJER, B. 2004. Changes in turbulent mixing shift competition for light between phytoplankton species. *Ecology,* 85(11)**,** 2960-2970.

JOINER, B. L. 1981. Lurking variables: Some examples. *The American Statistician,* 35(4)**,** 227-233.

JOUMA'A, J., LE BRAS, Y., RICHARD, G., VACQUIÉ-GARCIA, J., PICARD, B., EL KSABI, N. & GUINET, C. 2015. Adjustment of diving behaviour with prey encounters and body condition in a deep diving predator: the Southern Elephant Seal. *Functional Ecology,* 30(4)**,** 636-648.

KANG, S. H. & LEE, S. H. 1995. Antarctic phytoplankton assemblage in the western Bransfield Strait region, February 1993: composition, biomass, and mesoscale distributions. *Marine Ecology Progress Series,* 129**,** 253-267.

KRIVOBOKOVA, T. & KAUERMANN, G. 2007. A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association,* 102(480)**,** 1328-1337.

KÜHN, I. 2007. Incorporating spatial autocorrelation may invert observed patterns. *Diversity and Distributions,* 13(1)**,** 66-69.

LAMONT, T., BARLOW, R. & KYEWALYANGA, M. 2014. Physical drivers of phytoplankton production in the southern Benguela upwelling system. *Deep Sea Research Part I: Oceanographic Research Papers,* 90**,** 1-16.

LANCELOT, C., MATHOT, S., VETH, C. & DE BAAR, H. 1993. Factors controlling phytoplankton ice-edge blooms in the marginal ice-zone of the northwestern

Weddell Sea during sea ice retreat 1988: Field observations and mathematical modelling. *Polar Biology,* 13(6)**,** 377-387.

LAWSON, G. L., WIEBE, P. H., ASHJIAN, C. J. & STANTON, T. K. 2008. Euphausiid distribution along the Western Antarctic Peninsula—Part B: Distribution of euphausiid aggregations and biomass, and associations with environmental features. *Deep Sea Research Part II: Topical Studies in Oceanography,* 55(3–4)**,** 432-454.

LENNON, J. J. 2000. Red-shifts and red herrings in geographical ecology. *Ecography,* 23(1)**,** 101-113.

LICHSTEIN, J. W., SIMONS, T. R., SHRINER, S. A. & FRANZREB, K. E. 2002. Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs,* 72(3)**,** 445-463.

MARTIN, J. H., FITZWATER, S. E. & GORDON, R. M. 1990. Iron deficiency limits phytoplankton growth in Antarctic waters. *Global Biogeochemical Cycles,* 4(1)**,** 5-12.

MASSOM, R. A. & STAMMERJOHN, S. E. 2010. Antarctic sea ice change and variability–physical and ecological implications. *Polar Science,* 4(2)**,** 149-186.

MAYEWSKI, P. A., MEREDITH, M., SUMMERHAYES, C., TURNER, J., WORBY, A., BARRETT, P., CASASSA, G., BERTLER, N. A., BRACEGIRDLE, T. & NAVEIRA GARABATO, A. 2009. State of the Antarctic and Southern Ocean climate system. *Reviews of Geophysics,* 47(1).

MEIJERS, A. & KLOCKER, A. 2006. ADCP current velocity data for CTD stations of the BROKE-West Survey. Australian Antarctic Data Centre - CAASM Metadata. http://data.aad.gov.au/aadc/metadata/metadata_redirect.cfm?md=/AMD/AU/BROKE-West_ADCP.

MEIJERS, A. J. S., KLOCKER, A., BINDOFF, N. L., WILLIAMS, G. D. & MARSLAND, S. J. 2010. The circulation and water masses of the Antarctic shelf and continental slope between 30 and 80 degrees East. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57(9–10)**,** 723-737.

MEREDITH, M. P. & KING, J. C. 2005. Rapid climate change in the ocean west of the Antarctic Peninsula during the second half of the 20th century. *Geophysical Research Letters,* 32(19)**,** L19604.

MITCHELL, B. G., BRODY, E. A., HOLM-HANSEN, O., MCCLAIN, C. & BISHOP, J. 1991. Light limitation of phytoplankton biomass and macronutrient utilization in the Southern Ocean. *Limnology and Oceanography,* 36(8)**,** 1662-1677.

MITCHELL, B. G. & HOLM-HANSEN, O. 1991. Observations of modeling of the Antartic phytoplankton crop in relation to mixing depth. *Deep-Sea Research,* 38(8)**,** 981-1007.

MONTES-HUGO, M., DONEY, S. C., DUCKLOW, H. W., FRASER, W., MARTINSON, D., STAMMERJOHN, S. E. & SCHOFIELD, O. 2009. Recent Changes in Phytoplankton Communities Associated with Rapid Regional Climate Change Along the Western Antarctic Peninsula. *Science,* 323(5920)**,** 1470-1473.

MURPHY, E., CAVANAGH, R., HOFMANN, E., HILL, S., CONSTABLE, A., COSTA, D., PINKERTON, M., JOHNSTON, N., TRATHAN, P. & KLINCK, J. 2012. Developing integrated models of Southern Ocean food webs: including ecological complexity, accounting for uncertainty and the importance of scale. *Progress in Oceanography,* 102**,** 74-92.

MURPHY, E. J. & HOFMANN, E. E. 2012. End-to-end in Southern Ocean ecosystems. *Current Opinion in Environmental Sustainability,* 4(3)**,** 264-271.

NAKAGAWA, S. & SCHIELZETH, H. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution,* 4(2)**,** 133-142.

NICOL, S., FOSTER, J. & KAWAGUCHI, S. 2012. The fishery for Antarctic krill – recent developments. *Fish and Fisheries,* 13(1)**,** 30-40.

NICOL, S., MEINERS, K. & RAYMOND, B. 2010. BROKE-West, a large ecosystem survey of the South West Indian Ocean sector of the Southern Ocean, 30°E–80°E (CCAMLR Division 58.4.2). *Deep Sea Research Part II: Topical Studies in Oceanography,* 57(9–10)**,** 693-700.

PATTERSON, H. D. & THOMPSON, R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika,* 58(3)**,** 545-554.

PEARCE, I., DAVIDSON, A. T., THOMSON, P. G., WRIGHT, S. & VAN DEN ENDEN, R. 2010. Marine microbial ecology off East Antarctica (30 - 80°E): Rates of bacterial and phytoplankton growth and grazing by heterotrophic protists. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57(9–10)**,** 849-862.

PRÉZELIN, B. B., HOFMANN, E. E., MENGELT, C. & KLINCK, J. M. 2000. The linkage between Upper Circumpolar Deep Water (UCDW) and phytoplankton assemblages on the west Antarctic Peninsula continental shelf. *Journal of Marine Research,* 58(2)**,** 165-202.

R DEVELOPMENT CORE TEAM 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

REAY, D. S., PRIDDLE, J., NEDWELL, D. B., WHITEHOUSE, M. J., ELLIS-EVANS, J. C., DEUBERT, C. & CONNELLY, D., P. 2001. Regulation by low temperature of phytoplankton growth and nutrient uptake in the Southern Ocean. *Marine Ecology Progress Series,* 219**,** 51-64.

ROSENBERG, M. 2006. BROKE West Survey, Marine Science Cruise AU0603 - Oceanographic Field Measurements and Analysis. NASA Global Change Master

Directory. http://gcmd.nasa.gov/KeywordSearch/Keywords.do?KeywordPath=%5BParent_DIF%3D%27BROKE-West%27%5D&Portal=GCMD&MetadataType=0.

ROSENBERG, M., BINDOFF, N., CURRAN, C., HELMOND, I., MILLER, K., LACHLAN, D., CHURCH, J., RICHMAN, J. & LEFFANUE, H. 1999. Amery Ice Shelf Experiment (AMISOR). *Marine Science Cruises AU0106 and AU0207—Oceanographic Field Measurements and Analysis. Antarctic CRC Research Report,* 30**,** 1-119.

RSTUDIO 2014. RStudio: Integrated development environment for R (version 0.98.932). Boston, MA.

SCAMBOS, T., HARAN, T. J., FAHNESTOCK, M., PAINTER, T. & BOHLANDER, J. 2007. MODIS-based Mosaic of Antarctica (MOA) data sets: continent-wide surface morphology and snow grain size. *Remote Sensing of Environment,* 111**,** 242-257.

SCHWARZ, J. N., RAYMOND, B., WILLIAMS, G. D., PASQUER, B., MARSLAND, S. J. & GORTON, R. J. 2010. Biophysical coupling in remotely-sensed wind stress, sea surface temperature, sea ice and chlorophyll concentrations in the South Indian Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57(9–10)**,** 701-722.

SEDWICK, P. N. & DITULLIO, G. R. 1997. Regulation of algal blooms in Antarctic shelf waters by the release of iron from melting sea ice. *Geophysical Research Letters,* 24(20)**,** 2515-2518.

SMETACEK, V. & NICOL, S. 2005. Polar ocean ecosystems in a changing world. *Nature,* 437**,** 362-368.

SMITH JR, W. O., AINLEY, D. G., ARRIGO, K. R. & DINNIMAN, M. S. 2014. The oceanography and ecology of the Ross Sea. *Annual review of marine science,* 6**,** 469-487.

SMITH JR, W. O., SEDWICK, P. N., ARRIGO, K. R., AINLEY, D. G. & ORSI, A. H. 2012. The Ross Sea in a sea of change. *Oceanography,* 25(3)**,** 90-103.

SMITH, W. O. & LANCELOT, C. 2004. Bottom-up versus top-down control in phytoplankton of the Southern Ocean. *Antarctic Science,* 16(4)**,** 531-539.

SMITH, W. O. & NELSON, D. M. 1986. Importance of ice edge phytoplankton production in the Southern Ocean. *BioScience,* 36(4)**,** 251-257.

SULLIVAN, C. W., ARRIGO, K. R., MCCLAIN, C. R., COMISO, J. C. & FIRESTONE, J. 1993. Distributions of Phytoplankton Blooms in the Southern Ocean. *Science,* 262**,** 1832-1837.

SUMNER, M. & RAYMOND, B. 2015. R tools for spatial data at the Australian Antarctic Division (AAD). Hobart,

https://github.com/AustralianAntarcticDivision/raadtools: Australian Antarctic Division.

VACQUIÉ-GARCIA, J., GUINET, C., LAURENT, C. & BAILLEUL, F. 2015. Delineation of the southern elephant seal′s main foraging environments defined by temperature and light conditions. *Deep Sea Research Part II: Topical Studies in Oceanography,* 113**,** 145-153.

VERBYLA, A. P., CULLIS, B. R., KENWARD, M. G. & WELHAM, S. J. 1999. The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 48(3)**,** 269-311.

VSNI, U. K. ASReml.

WESTWOOD, K. J., BRIAN GRIFFITHS, F., MEINERS, K. M. & WILLIAMS, G. D. 2010. Primary productivity off the Antarctic coast from 30°–80°E; BROKE-West survey, 2006. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57(9–10)**,** 794-814.

WILLIAMS, G., HERRAIZ-BORREGUERO, L., ROQUET, F., TAMURA, T., OHSHIMA, K., FUKAMACHI, Y., FRASER, A., GAO, L., CHEN, H., MCMAHON, C., HARCOURT, R. & HINDELL, M. 2016. The suppression of Antarctic bottom water formation by melting ice shelves in Prydz Bay. *Nature Communications,* 7**,** 12577.

WILLIAMS, G. D., NICOL, S., AOKI, S., MEIJERS, A. J. S., BINDOFF, N. L., IIJIMA, Y., MARSLAND, S. J. & KLOCKER, A. 2010. Surface oceanography of BROKE-West, along the Antarctic margin of the south-west Indian Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57(9–10)**,** 738-757.

WOLD, S. 1974. Spline functions in data analysis. *Technometrics,* 16(1)**,** 1-11.

WOOD, S. 2006. *Generalized additive models: an introduction with R*, CRC press.

WRIGHT, S. W., VAN DEN ENDEN, R. L., PEARCE, I., DAVIDSON, A. T., SCOTT, F. J. & WESTWOOD, K. J. 2010. Phytoplankton community structure and stocks in the Southern Ocean (30–80°E) determined by CHEMTAX analysis of HPLC pigment signatures. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57(9–10)**,** 758-778.

WUNSCH, C. & FERRARI, R. 2004. Vertical mixing, energy, and the general circulation of the oceans. *Annual Review of Fluid Mechanics,* 36(1)**,** 281-314.

ZARAUZ, L., IRIGOIEN, X., URTIZBEREA, A. & GONZALEZ, M. 2007. Mapping plankton distribution in the Bay of Biscay during three consecutive spring surveys. *Marine Ecology Progress Series,* 345**,** 27-39.

Chapter 4

THE R PACKAGE *ECHOVIEWR* FOR AUTOMATED PROCESSING OF ACTIVE ACOUSTIC DATA USING ECHOVIEW

Published Journal Article:

Journal: *Frontiers in Marine Science*

Published Online: 25th of February 2015

Citation: Harrison L-MK, Cox MJ, Skaret G and Harcourt R (2015) The R package *EchoviewR* for automated processing of active acoustic data using Echoview. *Front. Mar. Sci*. **2**:15. doi: 10.3389/fmars.2015.00015

Authors:

Lisa-Marie K. Harrison[1], Martin J. Cox[1, 2], Georg Skaret[3], Robert Harcourt[1]

Affiliations:

[1]Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, New South Wales, Australia

[2]Australian Antarctic Division, Department of the Environment, Australian Government, Kingston, Tasmania, Australia

[3]Institute of Marine Research, Bergen, Norway

Abstract

Acoustic data is time consuming to process due to the large data size and the requirement to often undertake some data processing steps manually. Manual processing may introduce subjective, irreproducible decisions into the data processing work flow, reducing consistency in processing between surveys. We introduce the R package *EchoviewR* as an interface between R and Echoview, a commercially available acoustic processing software package. *EchoviewR* allows for automation of Echoview using scripting which can drastically reduce the manual work required when processing acoustic surveys. This package plays an important role in reducing subjectivity in acoustic data processing by allowing exactly the same process to be applied automatically to multiple surveys and documenting where subjective decisions have been made. Using data from a survey of Antarctic krill, we provide two examples of using *EchoviewR*: krill biomass estimation and swarm detection.

Keywords: active acoustic, Antarctic krill, data processing, echosounder, Echoview, R, package

1. Introduction

Active acoustics is a tool widely used for seabed mapping, seabed type classification, underwater tracking and resource monitoring. A suite of active acoustic instruments are available to carry out imaging (e.g. scanning sonars) and more quantitative tasks (e.g. multibeam and scientific echosounders). Echosounders have evolved from being instruments used primarily for mapping and navigation, to precision instruments capable

of resolving organisms a few millimeters in length and providing quantitative estimates of, for example, biomass.

This advance has seen widespread use of echosounders to detect organisms in the upper water column of both freshwater and marine environments for commercial fisheries and scientific purposes. In the marine environment, echosounders are routinely used to provide data informing commercial fishery stock assessments (Gerlotto et al., 1999) and to investigate ecological relationships such as predator-prey interactions (Benoit-Bird et al., 2013). Oceanographic applications include seabed habitat mapping (Brown et al., 2004) and environmental monitoring, e.g. oil seep and methane bubble monitoring after the Deepwater Horizon oil spill (Weber et al., 2014). Echosounders are commonly used in conjunction with image/video (McGonigle et al., 2009) and sediment sampling (van Walree et al., 2005) to verify seabed type, or trawls to verify a biological species' presence, size and target strength (McGonigle et al., 2009).

Echosounder transducers are most commonly embedded in a ship's hull or drop keel, although other platforms such as landers (Johansen et al., 2009), gliders (Guihen et al., 2014) and other autonomous underwater vehicles (Brierley et al., 2002) have been used. Regardless of platform, datasets from active acoustics are invariably extremely large and time consuming to process.

In active acoustic surveys, a conventional split-beam echosounder collecting data to a range of 500 m and pinging once per second typically collects around 8 GB of data per day (note: this depends on settings such as range resolution and pulse length). This may be compounded by the need to use multiple echosounder frequencies, sometimes up to six, operating simultaneously, further inflating the size of the raw data sets. Moreover, the routine use of broadband systems like the Simrad EK80 on board scientific and

commercial vessels is not far away. The amount of data from such systems vastly exceeds those from conventional sounders, and will again push storage and processing capacity. With advances in data storage capacity, data storage is no longer a significant constraint and enhanced computational power has enabled the development of powerful acoustic data processing software.

There are several software packages suitable for the processing of echosounder data, e.g. Echoview (Myriax, Hobart; www.echoview.com), LSSS (MAREC, Christian Michelsen Research, Norway, http://www.cmr.no/index.cfm?id=421565) and Sonar5-Pro (University of Oslo, Norway, http://folk.uio.no/hbalk/sonar4_5/). However, processing acoustic data remains time consuming and frequently requires subjective, often undocumented, decisions to be made by the user, such as removal of noise or bad data and allocation of backscatter to targets. Subjective decisions can potentially bias outputs from processed active acoustic data, for example biomass estimates.

Here we present the R package *EchoviewR* as a tool to: 1) reduce the processing time requiring a human operator, 2) document processing steps, thereby generating reproducible methodology, and 3) provide a framework within which additional functionality can be built by members of the acoustics community, so reducing the number of subjective decisions. The *EchoviewR* package is an interface between the widely used and freely available R program (http://www.R-project.org/) and Echoview (Myriax, Hobart; www.echoview.com). The methods used are generic and can be transferred to other acoustic processing software with scripting options, but the package as such is incompatible with other acoustic software.

EchoviewR uses Component Object Model (COM) scripting to run Echoview using R. This removes a large portion of the manual processing time and enables entire acoustic

surveys to be mostly processed automatically. It also increases consistency in processing because the same methods and thresholds can be applied in exactly the same way to multiple data sets. Hence *EchoviewR* provides a reproducible and transparent automated method for processing acoustic data using Echoview. Some examples of its use include filtering of data, automated biomass estimation and detection of krill swarms.

Using two examples, we illustrate *EchoviewR* functionality. Both examples are based on data collected during surveys of Antarctic krill (*Euphausia superba*; herein krill) using a Simrad EK60 echosounder (Horten, Norway) with downward facing hull-mounted transducers. The first example estimates regional krill biomass, and the second example detects krill swarms.

EchoviewR is intended to speed up processing of already clean acoustic data and is not currently capable of removing false bottom effects, time varied gain or noise spikes, although the package can access Echoview virtual variables to do some of these tasks, e.g. the 'Background noise removal algorithm' virtual variable (De Robertis and Higginbottom, 2007). The package is intended only as a method of automating processing using Echoview and is not a standalone method for processing acoustic data.

2. Methods

2.1. Implementation and Dependencies

EchoviewR was created using R 3.1 (R Development Core Team, 2014; available from http://cran.r-project.org/) with R-Studio 0.98.932 (RStudio, 2014; available from http://www.rstudio.com/), and Echoview 6.1 (Myriax, 2015; available from http://www.echoview.com/). Both R and Echoview are required to use the package. COM

objective handling is achieved using the *RDCOMClient* library. Additional *EchoviewR* functionality uses the *sp, lubridate, geosphere, maptools* and *rgeos* R libraries. To run Echoview via COM the following modules are required: base, bathymetric, analysis export, and scripting. Worked example one also requires the virtual echogram module and worked example two requires the virtual echogram and schools detection modules.

The *EchoviewR* package is available as open source on the GitHub repository (https://github.com/lisamarieharrison/EchoviewR) and can be downloaded and installed as an R library using the 'install from .zip file' option in R, or via devtools:install_github().

2.2. Expected data input for the package and worked examples

EchoviewR can work with any data type accommodated in Echoview that is accessible via COM. The worked examples provided here have been built using data collected using a Simrad EK60 echosounder (www.simrad.com/ek60). In itself, *EchoviewR* does not create Echoview templates or calibration files, but it can use both of these via COM.

2.3. Functions of the package

There are 46 functions available in *EchoviewR,* which are described in Table 1. A working example for each of these functions is given in the package documentation in the supplementary material. Not all Echoview functions are currently available in the package, however any functionality in Echoview that has COM accessibility could be added by the user.

Table 1 *Functions available in EchoviewR.*

Function	Description
EVOpenFile	Opens an existing .EV file
EVSaveFile	Saves an existing .EV file
EVSaveFileAs	Saves an existing .EV file to a new file name
EVCloseFile	Closes an open .EV file
EVNewFile	Creates a new .EV file
EVCreateFileset	Creates a new fileset
EVFindFilesetByName	Finds a fileset by name
EVAddRawData	Adds .RAW files to a fileset
EVCreateNew	Creates a new .EV file from a template
EVminThresholdSet	Sets the minimum dB threshold for an acoustic variable
EVSchoolsDetSet	Sets schools detection parameters
EVAcoVarNameFinder	Finds an acoustic variable by name
EVRegionClassFinder	Finds a region class by name
EVSchoolsDetect	Runs schools detection on an acoustic variable
EVIntegrationByRegionExport	Exports integration by region for an acoustic object
msDateConversion	Converts an Echoview date to readable format
EVAddCalibrationFile	Adds a calibration file to an .EV file
EVFilesInFileset	Finds the names of all .RAW files in the fileset
EVClearRawData	Clears all .RAW files from a fileset
EVFindFilesetTime	Finds the start and end date and time of a fileset
EVNewRegionClass	Creates a new region class
EVImportRegionDef	Imports a regions definition file
EVExportRegionSv	Exports Sv data for a region
EVAdjustRegionBitmap	Adjusts the settings of a region bitmap object
EVFindLineByName	Finds an Echoview line by name
EVChangeVariableGrid	Changes the horizontal and vertical grid for an acoustic variable
EVExportIntegrationByCells	Exports integration by cells for an acoustic variable
EVAddNewAcousticVar	Adds a new acoustic variable
EVShiftRegionDepth	Changes the depth of a region

Function	Description
EVShiftRegionTime	Changes the time of a region
EVGetCalibrationFileName	Finds the calibration file name
EVNewLineRelativeRegion	Creates a new line relative region
EVNewFixedLineDepth	Creates a new fixed depth line
EVDeleteLine	Deletes a line object
EVRenameLine	Renames a line object
EVExportRegionDef	Exports region definitions for a single region
EVFindRegionByName	Finds a region object by name
EVFindRegionClass	Finds a region class by name
EVExportRegionDefByClass	Exports region definitions for an entire region class
EVIntegrationByRegionByCellsExport	Exports integration by region by cells for an acoustic variable
lawnSurvey	Generate coordinates for a rectangular lawn survey design
zigzagSurvey	Generate coordinates for a zig-zag survey design
centreZigZagOnPosition	Centers a zig-zag survey on a given position
centreLawnOnPosition	Centers a lawn survey on a given position
exportMIF	Write a map information file for import into Echoview
EVImportLine	Imports an Echoview Line object

3. Examples

Here we present two examples using *EchoviewR:* 1) krill biomass estimation, and 2) krill swarm detection and classification. The purpose of these examples is to demonstrate that these analyses can be run automatically using *EchoviewR* and to show how Echoview output can be seamlessly linked to analyses carried out using R. Both examples assume that the reader is familiar with Echoview and are not intended to be a tutorial on Echoview. It is also assumed that the reader is familiar with R and programming concepts such as for loops.

The data are a subset of the EK60 split-beam data collected during the Krill Acoustics and Oceanography Survey (KAOS) carried out from the Aurora Australis. The KAOS survey was undertaken in January - March 2003 off North Eastern Antarctica. Data from 38, 120 and 200 kHz were written to RAW files. For clarity in the worked examples, we have used the 38 and 120 kHz data because these frequencies are the most useful for detecting and identifying the example species, Antarctic krill.

To demonstrate that biomass estimation and swarm detection can be automatically run on multiple transects where the data are too large to practically read in to Echoview at once, as is the case for most acoustic surveys, segments of six KAOS transects are provided and each 10-20 km transect segment is processed separately (Figure 1).

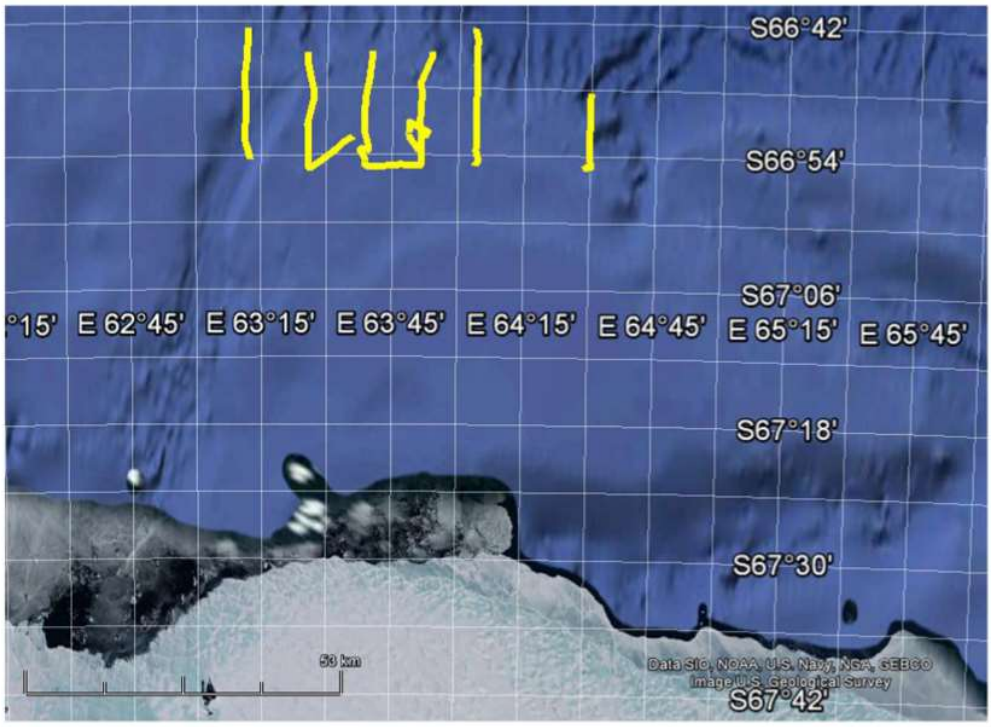


Figure 1 *Map showing location of 6 the example transects in yellow. Map created using Google Earth 7.1.2.2041.*

Both these examples have been tested using R 0.98.932 and Echoview 6.1.32.26088. The data to run these examples are available at the Australian Antarctic Division Data Centre [doi: 10.4225/15/54CF081FB955F]. An example of the data flow for the template used in this example is available in the supplementary material.

Before running each example some pre-processing is demonstrated to get the data in to a convenient format for analyzing each transect in a separate .EV file. In this pre-processing phase, the six transects are imported separately into Echoview and the following tasks are performed:

1. Create a new .EV file for the transect using the Echoview template file;
2. Import the EK60.RAW data files for that transect;
3. Add an Echoview .ecs calibration file;
4. Import .evr region definitions files to remove off effort data;
5. Import a seabed exclusion line (lineKAOS.evl)
6. Close and save the file and repeat for remaining transects.

These steps and the code to run them are demonstrated in the "Read data using the R package EchoviewR to control Echoview via COM" pdf vignette that is available with the supplementary material. Pre-processing must take place before examples 1 and 2 are run.

3.1. Example 1 – Krill biomass estimation

Automated biomass estimation of krill is demonstrated by processing the six transects separately in Echoview and exporting the data into R for density and biomass calculation. For each transect, the following steps are taken in Echoview:

1. Open the transect's .EV file
2. Set the grid for 38kHz and 120kHz noise removed values to 50 ping * 5m depth
3. Export integration by cells for 38kHz and 120kHz noise removed values

This produces two .csv files for each transect, one containing 38kHz and one containing 120kHz integrated data (i.e. a mean volume backscattering strength value for each cell). Then, the following steps are taken in R:

1. Import the 38kHz and 120kHz files for the transect
2. Remove no data values (set -999 and 999dB as NA) and depths < 0
3. Calculate the krill difference window of 120kHz – 38kHz for each integration cell using the following formula:

$$\Delta Sv_{ij} = Sv_{120_{ij}} - Sv_{38_{ij}}$$

where $Sv_{120_{ij}}$ = mean 120kHz backscattering strength for cell at horizontal integration interval j at depth i and $Sv_{38_{ij}}$ = mean 38kHz backscattering strength for cell at interval j at depth i.

4. Apply the dB difference technique (e.g. Watkins and Brierley (2002)) by setting $Sv_{120_{ij}}$ values outside the survey-specific dB difference range of 1.04 >= ΔSv_{ij} <= 14.75 dB to NA as these windows are unlikely to contain krill.

5. Convert the backscattering strength, $Sv_{120_{ij}}$ for each cell to linear scale, $sv_{120_{ij}}$ (Echoview uses a log scale by default):

$$sv_{ij} = 10^{\frac{Sv_{ij}}{10}}$$

6. Calculate mean volume backscattering strength (MVBS) across all depths for each 50 ping integration interval using the following formula:

$$MVBS_j = 10 \log_{10} \frac{1}{n_j} \sum_{i=0}^{n_j} sv_{120_{ij}}$$

where j = integration interval, n = maximum depth within integration interval j and $sv_{120_{ij}}$= backscattering strength at 120kHz for interval j at depth i.

7. Calculate the density, $\hat{p}_j$, for each integration interval:

$$\hat{p}_j = n_j * 10^{\left\{\frac{MVBS_j - TS}{10}\right\}}$$

where n_j = maximum depth of integration interval j, $MVBS_j$ = mean volume backscattering strength for interval j as calculated above and TS = target strength for 1kg of krill at 120kHz.

8. Calculate the overall transect density, $\hat{p}_k$ for transect k:

$$\hat{p}_k = \frac{1}{s_k}\sum_{j=1}^{s_k}\hat{p}_j$$

where j = integration interval, k = transect and s_k = number of integration intervals within transect k

9. The full survey density is then estimated using the Jolly & Hampton (1990) method, which uses the weighted density of each transect by length to calculate total survey density. Note that the formula has been modified to remove stratum as no strata were used in the KAOS example survey design:

$$\hat{p} = w_k\hat{p}_k$$

where k = transect, $w_k = \frac{L_k}{L}$, L_k = length of transect k in km, L = length of all survey transects in km and $\hat{p}_k$ = estimated density for transect k.

10. The full survey biomass estimate, $\hat{b}$, is then calculated by multiplying the weighted survey density by survey area:

$$\hat{b} = \hat{p}A$$

where $\hat{p}$ = estimated survey biomass and A = survey area in km^2.

Both the Echoview and R components above are run within loops to allow each transect to be run separately. This is done to demonstrate how looping over transects or days of a large survey is possible, rather than manually loading and processing each set of files. Users could format their code to process transects in parallel if processing time becomes too long. The *EchoviewR* and R code for the above analysis is shown in the "Biomass estimation using the R package EchoviewR to control Echoview via COM" pdf vignette that is available with the supplementary material. Table 2 shows the estimated density, length and biomass for the sample transects and survey area.

Table 2 *Estimated transect krill areal density and survey biomass for the six example transects.*

Transect Number	**Mean estimated density (gm^{-2})**	**Transect length (km)**	**Biomass (tonnes)**
1	3.26	13	42
2	20.66	22	454
3	43.74	15	656
4	22.57	22	467
5	6.66	18.5	123
6	4.99	21.5	107
Full Survey Area	16.79	112	43, 497

Example 1 has demonstrated the use of *EchoviewR* to automatically process and extract data by transect from Echoview. Krill density and biomass are then calculated in R using the extracted .csv files.

3.2. Example 2 – Swarm detection and classification

Automated swarm detection and classification of krill aggregations is demonstrated here using *EchoviewR*. The code for this example is available in the "Schools detection using the R package EchoviewR to control Echoview via COM" pdf vignette file available with the supplementary material. Each transect is processed separately to demonstrate how a full survey can be processed automatically using loops. Schools detection is run in Echoview and then detected aggregations are classified and clustered in R. The following steps are undertaken in Echoview using *EchoviewR*:

1. Open the transect's .EV file
2. Run schools detection on the variable *120 7x7 convolution*, assigning all detected schools to the region class "aggregations".
3. Export 120 and 38 kHz data for regions of class "aggregations" to a .csv file using the *EVIntegrationByRegionExport* function. This exports a single mean Sv for each aggregation

In this example, all detected aggregations are exported. However, it is also possible to export only aggregations classified as krill using the *120-38 aggregation dB difference filter* variable included in the template. The filter sets the *Krill aggregations* data to NULL if the *120-38 aggregation dB difference* value for that cell is outside the [1.04, 14.75] dB difference window for the KAOS survey.

The exported aggregations can now be classified and clustered in R. Each transect is run separately using a loop:

1. Import the 120kHz and 38kHz export by regions files
2. Remove null values (-999)
3. Calculate the 120 – 38 kHz difference window and subset data to only include difference values between [1.04, 14.75]
4. If no aggregations were classified as krill, exit and move to next transect
5. If krill aggregations are found, run Partition Around Medoids cluster analysis using the *ClusterSim* library using selected metrics (tutorial and example metrics available in Appendix B)
6. Print a summary table of the number of aggregations assigned to each identified cluster. Table 3 shows the number of krill swarms identified and the number of clusters detected for each transect

Table 3 *Number of unique krill aggregation clusters identified for each transect.*

Transect Number	**Number of krill swarms**	**Number of Clusters**
1	0	0
2	37	9
3	105	6
4	64	3
5	8	3
6	14	6

This example has demonstrated how school detection, data export and cluster analysis can be run automatically for an entire acoustic survey.

4. Discussion and future directions

EchoviewR is a free interface between R and Echoview that provides automated acoustic data processing. It drastically decreases manual processing time and reduces subjectivity by providing an easy way to implement exactly the same method across surveys. This package enables reproducible methodology, which is a vital part of the scientific method. We have given examples of automated krill biomass estimation and school detection using *EchoviewR* that demonstrate the use of the package on a subset of the KAOS survey. This method can easily be extended to run a full survey by transect, day or any other subset required.

There are a number of limitations to the package. Currently it is only available for use for single and split beam sonar data. *EchoviewR* is also unable to handle removal of noise and false bottom effects, which must be completed prior to using the package. Not all functions in Echoview are currently available using *EchoviewR*, however any COM functionality in Echoview can be implemented in R. The COM hierarchy help page is a useful starting point for those wishing to add extra functions.

EchoviewR is accessible as free software from the *EchoviewR* GitHub repository (https://github.com/lisamarieharrison/EchoviewR) and is readily available for community development. An important next step is the implementation of false bottom and noise removal using *EchoviewR*, and it is our hope that the acoustic community will take the tools that we are providing and extend the package to include the functionality that they require. We also underline that the methods described here are generic, and hope the work can inspire the implementation of scripting interface in other acoustic processing software.

5. Acknowledgement

We would like to thank Echoview for their support of this project. This research is a contribution to Australian Antarctic Division science programme Project 4104 and project 4102. MJC is funded by Australian Research Council grant FS11020005. LMH is funded by a Macquarie University Research Excellence Scholarship.

6. Supplementary Material

The published pdf of this chapter is included in Appendix A.

The supplementary materials available in Appendix B are:

1. Pdf vignettes
    a. Read data
    b. Biomass estimation
    c. Schools detection
2. Plot of data flow

7. References

BENOIT-BIRD, K. J., BATTAILE, B. C., HEPPELL, S. A., HOOVER, B., IRONS, D., JONES, N., KULETZ, K. J., NORDSTROM, C. A., PAREDES, R., SURYAN, R. M., WALUK, C. M. & TRITES, A. W. 2013. Prey Patch Patterns Predict Habitat Use by Top Marine Predators with Diverse Foraging Strategies. *PLoS ONE,* 8**,** e53348.

BRIERLEY, A. S., FERNANDES, P. G., BRANDON, M. A., ARMSTRONG, F., MILLARD, N. W., MCPHAIL, S. D., STEVENSON, P., PEBODY, M., PERRETT, J., SQUIRES, M., BONE, D. G. & GRIFFITHS, G. 2002. Antarctic Krill Under Sea Ice: Elevated Abundance in a Narrow Band Just South of Ice Edge. *Science,* 295**,** 1890-1892.

BROWN, C. J., HEWER, A. J., MEADOWS, W. J., LIMPENNY, D. S., COOPER, K. M. & REES, H. L. 2004. Mapping seabed biotopes at hastings shingle bank, eastern English Channel. Part 1. Assessment using sidescan sonar. *Journal of the Marine Biological Association of the UK,* 84**,** 481-488.

DE ROBERTIS, A. & HIGGINBOTTOM, I. 2007. A post-processing technique to estimate the signal-to-noise ratio and remove echosounder background noise. *ICES journal of marine science,* 64**,** 1282-1291.

GERLOTTO, F., SORIA, M. & FRÉON, P. 1999. From two dimensions to three: the use of multibeam sonar for a new approach in fisheries acoustics. *Canadian Journal of Fisheries and Aquatic Sciences,* 56**,** 6-12.

GUIHEN, D., FIELDING, S., MURPHY, E. J., HEYWOOD, K. J. & GRIFFITHS, G. 2014. An assessment of the use of ocean gliders to undertake acoustic measurements of zooplankton: the distribution and density of Antarctic krill (Euphausia superba) in the Weddell Sea. *Limnology and Oceanography: Methods,* 12**,** 373-389.

JOHANSEN, G. O., GODØ, O. R., SKOGEN, M. D. & TORKELSEN, T. 2009. Using acoustic technology to improve the modelling of the transportation and distribution of juvenile gadoids in the Barents Sea. *ICES Journal of Marine Science: Journal du Conseil,* 66**,** 1048-1054.

JOLLY, G. & HAMPTON, I. 1990. A stratified random transect design for acoustic surveys of fish stocks. *Canadian Journal of Fisheries and Aquatic Sciences,* 47**,** 1282-1291.

MCGONIGLE, C., BROWN, C., QUINN, R. & GRABOWSKI, J. 2009. Evaluation of image-based multibeam sonar backscatter classification for benthic habitat discrimination and mapping at Stanton Banks, UK. *Estuarine, Coastal and Shelf Science,* 81**,** 423-437.

MYRIAX 2015. Echoview. Hobart, Tasmania.

R DEVELOPMENT CORE TEAM 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

RSTUDIO 2014. RStudio: Integrated development environment for R (version 0.98.932). Boston, MA.

VAN WALREE, P. A., TĘGOWSKI, J., LABAN, C. & SIMONS, D. G. 2005. Acoustic seafloor discrimination with echo shape parameters: A comparison with the ground truth. *Continental Shelf Research,* 25**,** 2273-2293.

WATKINS, J. L. & BRIERLEY, A. S. 2002. Verification of the acoustic techniques used to identify Antarctic krill. *ICES Journal of Marine Science: Journal du Conseil,* 59**,** 1326-1336.

WEBER, T. C., JERRAM, K. & MAYER, L. 2014. Acoustic Sensing of Gas Seeps in the Deep Ocean with Split-beam Echosounders. *Proceedings of Meetings on Acoustics,* 17**,** -.

Chapter 5

THE WORLD'S MOST ABUNDANT PREDATOR IS NOT A PASSIVE DRIFTER: ANTARCTIC KRILL AGGREGATE AROUND FOOD AND OXYGEN

Authors:

Lisa-Marie K. Harrison[1], Martin J. Cox[2], Robert Harcourt[1]

Affiliations:

[1]Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, 2109, New South Wales, Australia

[2]Australian Antarctic Division, 203 Channel Highway, Kingston, 7050, Tasmania, Australia

Abstract

Antarctic krill (*Euphausia superba*) are often thought to be solely passive drifters and opportunistic feeders, however here we show that krill are more mobile than previously thought. Using a hurdle model, we simultaneously examined environmental drivers of 1) the probability of krill being present and 2) density (gm^{-2}) given presence. We found that krill consistently position themselves in response to environmental variables across a large area (1.3 million km^2) of East Antarctica. Krill were more likely to be present in shallow, more saline, warmer water. This is unlikely to be solely due to current systems since the survey design crossed two major oceanographic boundaries and there was no reflection of these circulation patterns in the model's residuals or random effects. High krill densities occurred in areas with both high oxygen concentration and increasing food availability. This indicates that krill actively aggregate around these essential resources and do not simply coalesce into high-density swarms by chance. Our conclusions suggest that models based primarily on krill transport by currents need to be supplemented to include active behaviour based on environmental conditions and food sources if we are to accurately predict krill distribution in the Southern Ocean.

Keywords: Antarctic krill, Southern Ocean, phytoplankton, hurdle model, acoustic survey, distribution

1. Introduction

The extent to which organisms are able to influence their population distribution by active movement as opposed to passively drifting within dynamic ocean systems is of major importance to conservation and management in dynamic pelagic ecosystems, yet has rarely been tested (Putnam et al 2016). This is exemplified by Antarctic krill (*Euphausia*

superba) as the degree to which Antarctic krill are passive drifters or active swimmers has been deliberated for decades, with management approaches assuming they are passive drifters, despite evidence of strong selection for active behaviour (Richerson et al., 2015). Antarctic krill are both the focus of the world's largest krill fishery (Siegel, 2016; pg. 387) as well as an essential part in the Southern Ocean food web (Siegel, 2016; pg 321).

While large amounts of krill survey data exist, much remains unknown about the drivers of their highly patchy distribution. Physical forcing by current systems, local retention and Circumpolar Deep Water intrusions have been invoked to explain large scale krill distribution patterns (Piñones et al., 2013). While there is no single study that has mapped circumpolar krill distribution and currents, overlaying historical krill distribution on maps of surface circulation indicates that overall distribution may be related to gyres (Nicol, 2006, Amos, 1984) and distribution is limited to the extent of the Southern Boundary (Tynan, 1998). Krill may be distributed around these major oceanographic features by advection or through actively choosing to be at these locations because they are attractive due to abundant food or beneficial environmental conditions. These two potential causes are difficult to disentangle and doing so is the aim of our study. At a small scale relative to these large oceanographic features, krill form swarms and must actively swim to do so, however the extent of active swimming on a larger scale remains a mystery. We assess the mechanisms behind large scale krill distribution using data collected over a large area (1.3 million km^2) and spanning two important frontal systems in the East Antarctic.

Large scale circulation patterns are certainly important during the larval stage, aggregating larvae in areas where Circumpolar Deep Water encroaches on to the continental shelf and conditions are favourable for egg development before ascent and hatching (Piñones et al., 2016). After metamorphosis, juvenile and adult krill form

aggregations, or swarms, and patchiness varies with overall biomass (Siegel, 2016; pg. 279, Brierley and Cox, 2015). Swarm type changes throughout the life-cycle, with larger, denser swarms occurring when krill are young and immature, and swarms transitioning to small more diffuse swarms as the krill mature (Tarling et al., 2009). Coalescing swarms may form layers of krill, with the edges of swarms touching but the discrete swarms not fully merged (Watkins and Murray, 1998). These layers are different to those of individual swarms because the intra-layer population differences are just as large as between-layer differences, while individual swarm characteristics are more distinct. While there is biological evidence that environmental factors are important for the life cycle and distribution of krill, there are few quantitative studies of the drivers of krill density at regional scales (Siegel, 2016; pg. 26-28, Nicol, 2006, Nicol, 2003a).

The key requirements for krill survival include adequate food, suitable habitat and predation avoidance (Alonzo and Mangel, 2001). We might expect that if krill are not passive drifters, swarm characteristics (e.g. location, size and density) will have drivers based on these key requirements to maximise the probability of survival. Vertical distribution of swarms depends on mixing depth, with most krill swarms occurring above the thermocline (Godlewska et al., 1988), and dense swarms more likely to occur in higher water temperature (Krafft et al., 2012). The need to find food and avoid predators is believed to be more responsible for the formation of localised krill swarms in the Marginal Ice Zone than the direct impact of physical forcing (Daly and Macaulay, 1991). Instantaneous swarm shape is influenced by the competing needs of accessing oxygen for respiration and avoiding predation, with these two requirements being absolutely necessary for an individual to survive to be able to forage (Brierley and Cox, 2010). Swarms closer to the Antarctic coastline are larger and denser than their counterparts further offshore, which may be because krill are clustering for protection from land based

air-breathing predators (Klevjer et al., 2010). Subsequent to the immediate demands of finding oxygen and avoiding predation, foraging is a longer term goal (Brierley and Cox, 2010).

Observational studies demonstrate that krill are capable of active foraging rather than only feeding opportunistically on food that they come upon by chance. Krill have been observed to forage actively in captivity (e.g. (Hamner and Hamner, 2000, Kawaguchi et al., 2010)) and similar behaviours have been inferred from survey data (e.g. (Quetin and Ross, 1991)). In small tanks in captivity, krill have been observed to use chemoreception to locate phytoplankton, then use localized area foraging upon reaching the bloom (Hamner and Hamner, 2000). Calculations suggest that krill aggregations can rapidly exhaust food supplies in a phytoplankton bloom and must change locations constantly to maintain their energy intake (Nicol, 2003b). The formation of larger swarms may allow for more efficient foraging if individuals within a swarm can communicate and larger swarms have been observed in areas of higher surface productivity than low productivity (Tarling et al., 2009). In addition to these instantaneous effects on krill swarm size and location, food resources are known to have long term effects on population size. Food availability constrains local population growth rates, with simulation showing that the optimal growth strategy is for krill to switch between low and high metabolic states based on food abundance (Groeneveld et al., 2015). Large phytoplankton blooms can trigger an early start to the mating season (Schmidt et al., 2012) and sustained high food levels can increase reproductive success, resulting in higher juvenile krill recruitment in the following season (Saba et al., 2014). This has been observed in the West Antarctic Peninsula, where years of high productivity lead to an increase in krill stocks in the following year (Saba et al., 2014). Food stocks clearly have short and long term effects

on krill populations and individual swarms and are an important consideration when studying drivers of krill distribution.

We hypothesise that krill respond to their environment and seek to distribute themselves within environmentally favourable areas, particularly areas with high food concentration. We define this to be 'active swimming', which contrasts with 'passive drifting' which we define as large scale krill transport by currents. We test our hypothesis using data collected over 1.3 million km^2 of ocean in the East Antarctic and a hurdle model, which is a flexible statistical model that can accommodate different sets of variables for krill presence/absence and density given presence. If our hypothesis is incorrect and krill are not aggregating through active swimming, then there should be strong evidence of this in the model's residuals because the survey design traversed three large oceanographic current systems. Our methodology overcomes the limitations of standard modelling methods which do not adequately deal with both zero-inflation and random effects in continuous data and have enabled us, for the first time, to develop predictive models of krill to quantitatively examine the passive drifting hypothesis.

To visualise how the observed distribution of krill can help us understand their habitat preferences and let us test our hypothesis, a conceptual figure demonstrating observations under passive drifting and active swimming is shown in Figure 1. Under a simulated uniform temperature environment (Figure 1a), krill with no environmental preference, or alternatively no capability to aggregate around a preferred temperature, will show uniform probability of presence across all temperatures (Figure 1b). In contrast, under a hypothetical water temperature preference of 0.5^{o}C (Atkinson et al., 2006), krill will display a non-uniform probability of presence and cluster about the their preferred 0.5^{o}C water temperature (Figure 1c) . The combination of water temperature (a) and

environmental preference (b and c) will determine the observed krill distribution under passive drifting (Figure 1d) and active swimming (Figure 1e). Our model uses observations of krill (i.e. Figures 1d or e) to try and understand which underlying environmental preference is occurring (i.e. Figures 1b or c).

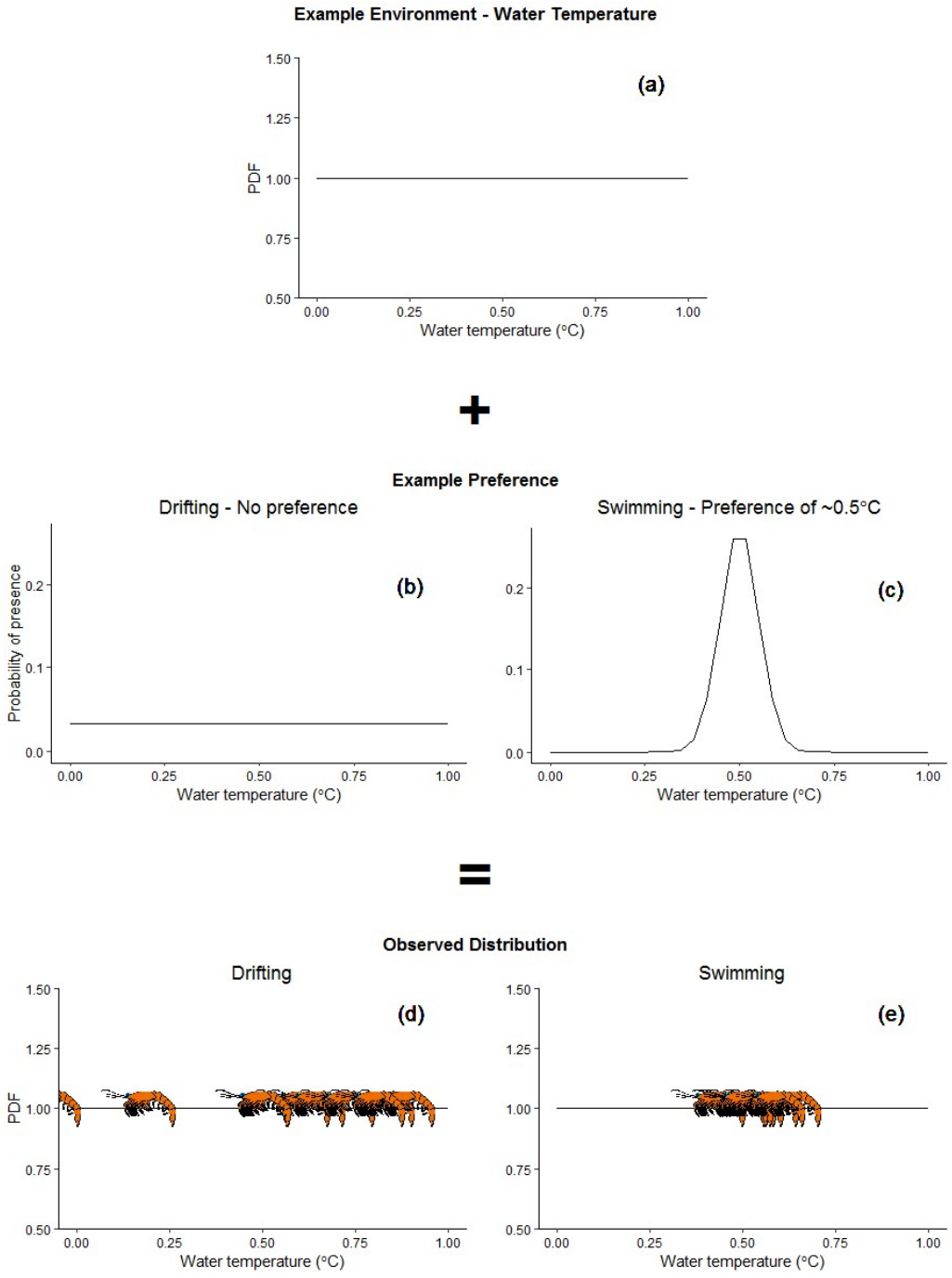


Figure 1 *Conceptual figure of krill distribution under passive drifting (left column) and active swimming (right column) with temperature preference of 0.5°C. Data for figure are simulated. In a hypothetical environment (a), the preference of krill (b = drifting and c = swimming) will determine their final location (d & e). The same number of krill appear in (d) and (e) however in (e) they are strongly clustered around their temperature preference of 0.5°C. Krill image from Gemma Carroll, Macquarie University (with permission).*

2. Methods

We used data from the 2006 Baseline Research on Oceanography Krill and the Environment (BROKE-West) survey in the South-East Indian Ocean (2010). The data are from 37 Conductivity Temperature Depth (CTD) stations with coincident EK60 scientific echosounder data in the top 250m of the water column. CTD data were collected using a SeaBird SBE9plus with an attached dissolved oxygen sensor (SBE43), fluorometer (Wet Labs ECO) and Photosynthetically Active Radiation (PAR) sensor (LI-COR).

The echosounder data were processed and integrated across a regular 50 m horizontal by 10 m vertical grid. Krill were identified and krill density calculated from 120 kHz and 38 kHz acoustic data using standard methods (Jarvis et al., 2010) and the software programs Echoview (Myriax, 2015), R (R Development Core Team, 2014) and the R package EchoviewR (Harrison et al., 2015). The full acoustic data set contains over a billion data points, and the EchoviewR package allowed us to process the data around the CTD stations automatically by providing a scripting interface between R and Echoview. Krill

densities were averaged across the closest 100 bins at 10 m depth increments and paired to the CTD measurements.

To examine our hypothesis that environmental conditions influence krill presence we extended traditional hurdle model methods to accommodate the acoustic grid data being zero inflated (half of the grid cells contained no krill) and site random effects. We developed our model using R with the probability of krill presence modelled using a binomial mixed model (Zuur et al., 2009; pg. 324) in *lme4* (Bates et al., 2015) and conditional krill density using a linear mixed model (*nlme*; (Pinheiro et al., 2016)). In both models, we used sampling station as a random effect to account for extraneous inter-station differences. The density model contained an additional identity variance structure with station as a categorical variable because heterogeneity of variances between stations occurred. We were unable to use the standard hurdle or zero-inflated models because our density data are continuous and these functions can't accommodate random effects. Potentially, krill presence-absence and krill conditional density could be driven by the same environmental conditions so for both models we used the same candidate explanatory variables of: cell depth, temperature, salinity, dissolved oxygen, time of day and phytoplankton fluorescence. Krill conditional density data were $\log_e$ transformed, and explanatory variables were centred and scaled. Backwards Akaike Information Criterion (AIC) based model selection was used and model fit was assessed using 40-fold cross-validation (Arlot and Celisse, 2010), where one station was dropped at a time. During each iteration of the cross-validation, a random effect was generated for the dropped station from a normal distribution with mean = 0 and the estimated standard deviation from the fitted model. This is required to avoid skewing the predictions during back-transformation resulting in biased inference. Area Under Curve (AUC) of the Receiver Operator Characteristic (ROC) was used to assess cross-validation goodness-

of-fit for the presence/absence model and marginal and conditional R^2 were used for the density model (Hanley and McNeil, 1982, Nakagawa and Schielzeth, 2013).

3. Results

The best presence/absence model, selected by AIC, had depth, temperature and salinity as explanatory variables along with a station random effect (Figure 2 and Table 1). See supplementary materials (Appendix C) for details on model selection and model diagnostics. The drop one station cross-validation AUC was 0.72 indicating good predictive power even at novel stations (see supplementary materials for ROC curve). The Variance Inflation Factors of all variables were less than 5 indicating low co-linearity.

Table 1 *Model summary for best model of y ~ depth + temperature + salinity + re(stn) with family binomial (link = logit). Note: Parameter estimates are on the link scale.*

Coefficient	**Estimate**	**Standard Error**	**Variance Inflation Factor**
Depth	-1.203	0.144	2.51
Temperature	0.368	0.110	1.22
Salinity	0.310	0.142	2.68
Random Effects			
	Variance Estimate	**Standard Error**	
Station	0.709	0.842	

Krill were found in shallower, warmer and saltier water (Figure 2). While the presence/absence component of the model indicates that krill presence is linked to environmental conditions, krill density given presence was strongly influenced by both phytoplankton concentration and oxygen ($\log_e$(density) = $\log_e$(phytoplankton) * oxygen

+ re(stn)) (Figure 3). A plot of raw dissolved oxygen and phytoplankton is available in Appendix C. A variance structure to account for heterogeneity of variance between stations was required. The marginal and conditional R^2 were 0.26 and 0.81 respectively indicating that the CTD station random effect accounts for a large amount of inter-station variation.

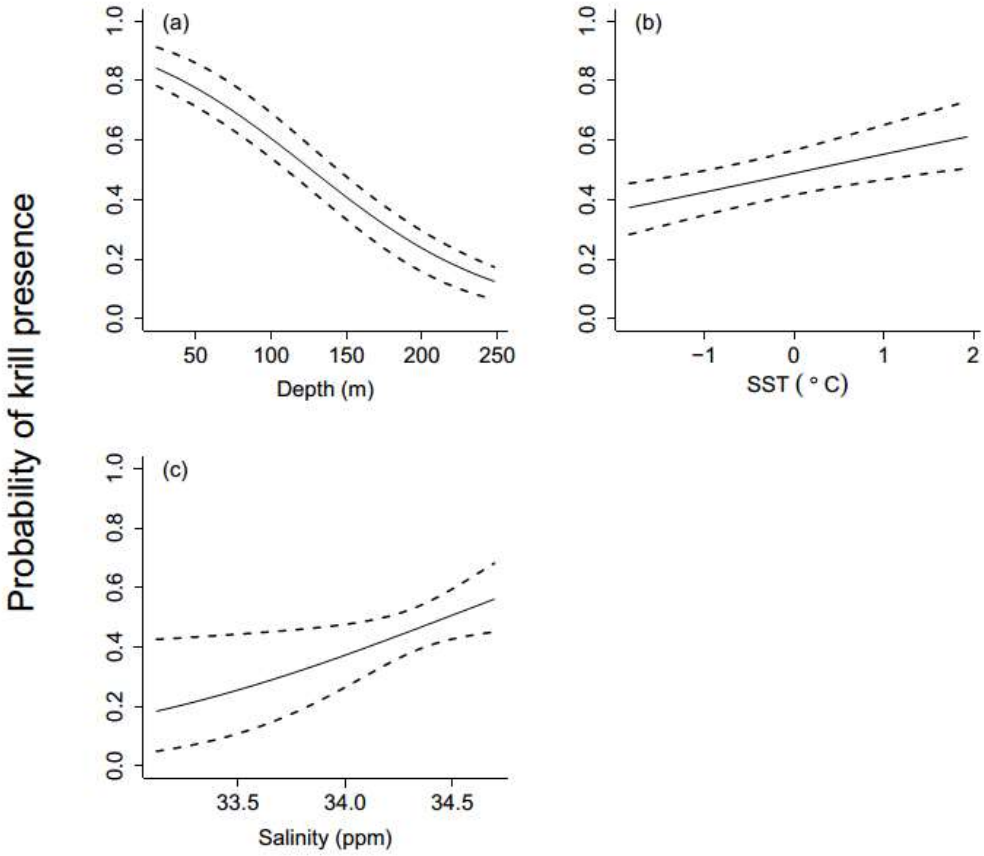


Figure 2 *Predicted probability of krill presence for our selected model across the BROKE-West survey area. Significant variables were a) depth, b) temperature and c) salinity. Dotted lines are 95% confidence intervals. Plots only cover the range of the observed data so the $y \rightarrow 0$ and $y \rightarrow 1$ asymptotes are not always visible.*

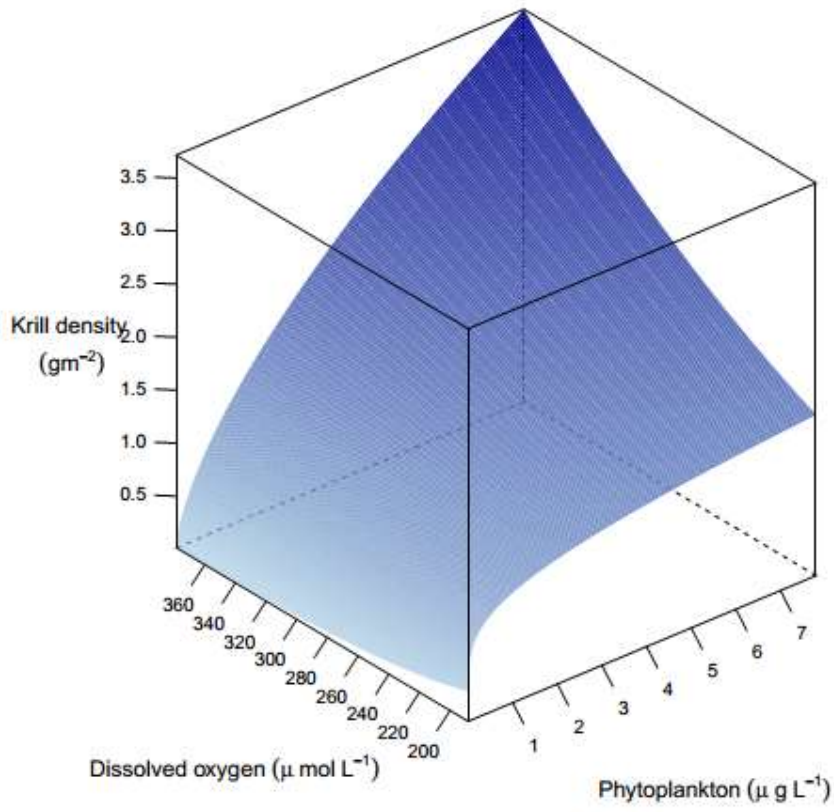


Figure 3 *Krill density given presence averaged over the random effects to give survey wide inference showing the interaction between Dissolved Oxygen and Phytoplankton Fluorescence. High conditional densities of krill only occur when both oxygen concentration and fluorescence are high.*

To assess whether there are residual patterns that align with the prevailing current systems in the survey area, we plotted the station random effects (subplots a and c) and model residuals (subplots b and d) for the presence/absence and density models with the locations of the front systems overlayed (Figure 4). An obvious pattern of residuals or random effects on either side of the front boundaries would provide evidence of aggregation by passive drifting. There is no obvious pattern in these plots which supports our hypothesis that the aggregation we observed was achieved through active swimming rather than passive drifting on these large scale current systems.

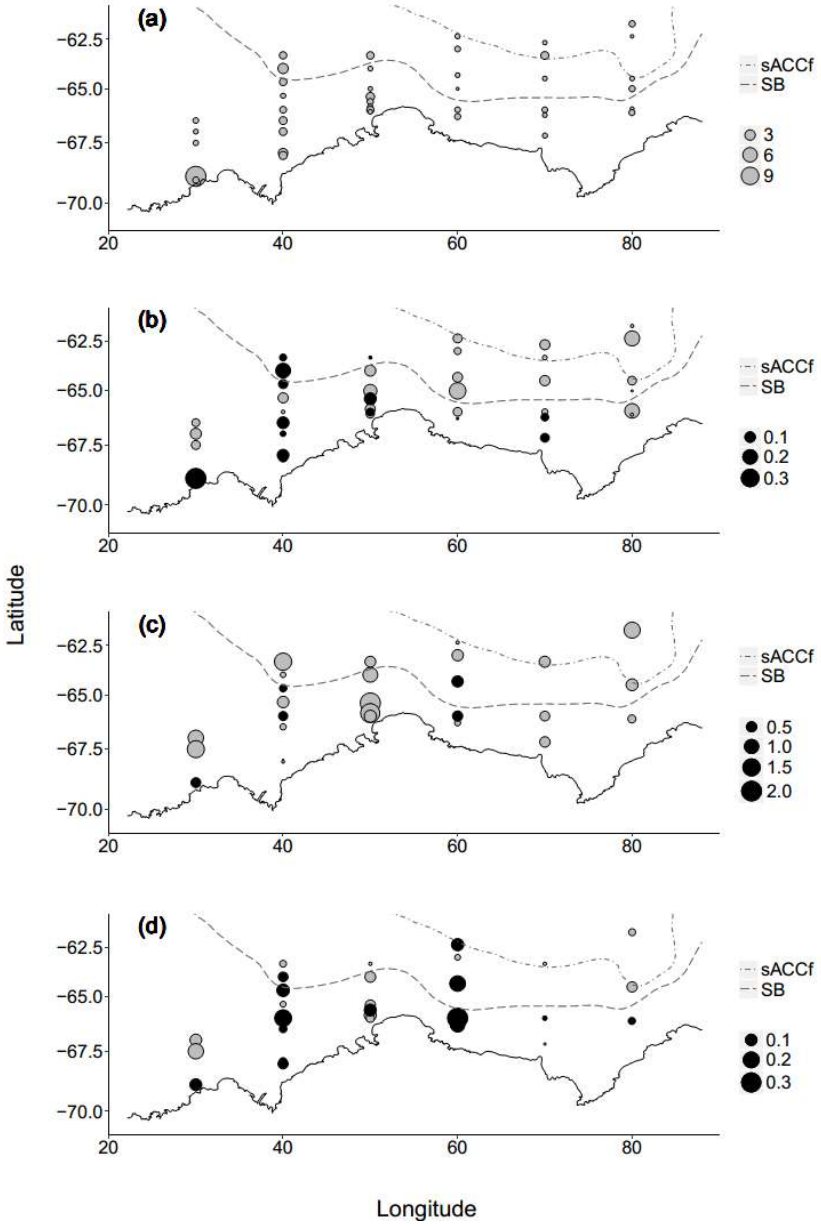


Figure 4 *Bubble plots of presence/absence model station random effect (a) and average station residuals (b), and density model station random effect (c) and average station residuals (d). Grey values are positive and black values are negative. Plots are shown with the Antarctic continent (solid line), southern Antarctic Circumpolar Current front (sACCf; dot-dash line) and Southern Boundary of the Antarctic Circumpolar Current (SB; dashed line). Shapefiles for Antarctic coastline were sourced from the National Snow and Ice Data Center (Scambos et al., 2007) and shapefiles for the fronts were sourced from the Australian Antarctic Data Centre (Orsi and Harris, 2001 (Updated 2015)).*

4. Discussion

Here we have clearly demonstrated that krill distribute actively in relation to both their environment and food availability, rather than simply drifting passively. These relationships are valid across the BROKE-West survey region (30-80^{o}E) in the East Antarctic and we would not have found strong relationships between krill presence and environmental variables if krill were merely distributed at random. Furthermore, these relationships are unlikely to be solely due to krill transport by current systems since the BROKE-West survey pattern cut two oceanographic boundaries (the Southern Antarctic Circumpolar Current Front and the Southern Boundary Front) (Meijers et al., 2010, Nicol et al., 2010), and these patterns were not seen in our model's residuals or random effects.

Advection has been suggested as the dominant force behind krill distribution (Hofmann and Murphy, 2004, Amos, 1984, Murphy et al., 2004) and while it certainly plays an important role in krill circumpolar distribution, our modelling suggests that on the scale of the BROKE West survey, krill are able to seek out environmentally preferred areas and source localised food and oxygen. There are important implications arising from active distribution rather over passive drifting. Life history modelling reveals that active behaviour in krill would result in higher survival probability and increase reproductive success by 70% (Richerson et al., 2015). Our results corroborate a life history modelling study which showed reproductive and survival benefits for active swimming in krill (Richerson et al., 2015) and crucially, our results show that forecasting of krill distribution requires much more sophisticated data collection and modelling than relying on the assumption of passive drifting (Kock et al., 2007, Marin and Delgado, 2001).

4.1. Krill relationship to environment

We found a strong link between krill density given presence, high oxygen concentration and high food availability and provide quantitative evidence supporting observational accounts of active feeding in krill (Hamner and Hamner, 2000, Quetin and Ross, 1991, Kawaguchi et al., 2010). Links occur between krill and Chlorophyll-a at large spatial and temporal scales in the Scotia Sea and Antarctic Peninsula (Santora et al., 2012, Silk et al., 2016), however our modelling clearly showed that across the East Antarctic (30 to 80^{o}E) food availability is insufficient alone to explain high-density krill areas; high dissolved oxygen concentration is also required. These high oxygen concentrations may be necessary to sustain dense krill swarms during grazing. At a local scale, accessing oxygen and avoiding predation influences swarm shape (Brierley and Cox, 2010), and our results show that this trade off extends to foraging, with food as well as oxygen and anti-predation requirements underpinning localised krill distribution. High oxygen concentration could also result from persistent phytoplankton patches, indicating that krill may specifically target areas with stable blooms. High oxygen levels may also allow krill to graze for longer before depleting localised oxygen levels. Krill can maintain constant respiration down to oxygen levels of 55% air saturation, after which krill oxygen consumption declines (Tremblay and Abele, 2016). Oxygen levels were always above 60% air saturation in our data set which could explain why oxygen was not a significant factor in the presence/absence model: oxygen saturation never fell below 55% and so did not adversely affect krill. The relationships revealed during our modelling of krill actively sourcing food and oxygen further demonstrates that krill are not solely opportunistic feeders, and aggregate around patchy resources.

Our model found that the probability of krill presence increased with higher temperature, salinity and deeper depths in the water column. This likely reflects areas where the warm

and nutrient rich Modified Circumpolar Deep Water (water south of the Antarctic Circumpolar Current with similar properties to Circumpolar Deep Water) flows over the continental slope, around which the most krill was seen during BROKE-West (Williams et al., 2010). Circumpolar Deep Water intrusions benefit krill by providing optimal conditions for egg hatching, with the warmer water hastening development and causing hatching to occur at a shallower depth leaving the larvae with a shorter distance to ascend (Hofmann and Hüsrevoğlu, 2003). Circumpolar Deep Water intrusions also increase primary productivity through the input of nutrients into the system (Prézelin et al., 2004, Nicol, 2006). This theory is supported by observational reports of high krill densities in association with Circumpolar Deep Water and Lower Circumpolar Deep Water in the Ross Sea (Sala et al., 2002, Taki et al., 2008). However, while a positive correlation between krill presence and Circumpolar Deep Water was expected in the West Antarctic Peninsula, no relationship was found using Generalised Additive Models (Lawson et al., 2008).

While our data set covered a large spatial extent (1.3 million km), like many surveys in remote and environmentally extreme locations, we viewed a small snapshot in time at each site. Whether krill had just arrived, had already depleted the resources or were passing through to more favourable areas remains unknown. Even if behaviour was known, krill feeding rates alone are highly variable with individual feeding rates varying from 0.37 – 86 µg Chlorophyll-a d^{-1} (Perissinotto et al., 1997). The site random effects partly account for this unknown behaviour and reduce the confounding that could occur in a study with numerous sites separated spatially and temporally (Davies and Gray, 2015). This gives us confidence that the relationships we have demonstrated are real, rather than an artefact caused by the survey design or the small temporal window through which we view the ecosystem.

4.2. Future directions

Krill are highly seasonal animals, displaying vastly different behaviour in winter when food and light become limited (Meyer, 2012). In this work we have extended current modelling techniques to clearly support the hypothesis that krill actively position themselves in relation to their environment. For this information to be actively used to model and predict krill movement it is necessary to determine whether the relationships we observed extend across different seasons as we believe is likely. This will require additional data across seasons. The two-part nature of our model is ideal for answering this question because it allows us to partition the importance of variables to presence/absence and density. The decline in krill respiration (30-50%) and feeding (80-86%) during autumn and winter (Meyer et al., 2010) would likely alter the shape of the surface in Figure 3, as the priorities for an individual's survival change to reflect the increased difficulty in finding sufficient food. To examine whether krill avoid areas without any phytoplankton when food sources are scarce in winter, the significance of phytoplankton fluorescence in the presence/absence model during winter could be tested. A seasonal analysis could identify variables that krill are particularly sensitive to throughout different times of the year and help us understand the effects of extreme environmental events during winter.

Our hurdle model does not include a current model, however the BROKE-West survey cutting two large current boundaries allowed us to infer that krill were not passively transported by these systems because there was no evidence of this in the model diagnostics. It is an important future direction to match krill densities to high resolution current models to further assess the extent of active positioning, especially at smaller scales than the frontal systems encountered during BROKE-West. There is weak

evidence for krill swarm movement in relation to local current systems (Tarling and Thorpe, 2014) but this needs further quantitative study.

Krill spatial distribution and feeding preferences differ among age classes (Schaafsma et al., 2016, Atkinson et al., 2002, Siegel et al., 2013), and a similar approach to our hurdle model could be used to compare the relative importance of each variable to the presence/absence and density of krill at different stages of the life cycle. High-spatial resolution trawl data would be required to differentiate between age classes, as this is not currently possible from acoustic data alone. Modelling of important variables at each life stage could help understand the potential consequences of todays' environmental conditions on next (and future) years' stock.

4.3. Summary

In summary, we have shown that krill aggregate around key resources, and presence-absence and density are driven by different sets of conditions. The probability of krill presence is highest around CDW intrusions and where krill are present they aggregate around food and oxygen resources. Together, this contradicts the long-held belief that krill are solely passive drifters. Adding missing variables such as nutrients (in particular, ammonium and iron) could further improve our results. Krill data have been harvested globally for decades (Siegel, 2016; pg. 21, Everson, 2008) and our two-part mixed modelling methodology could be retrospectively applied to assess whether similar behavioural patterns occur in other euphausiids, and whether these relationships display seasonality or change over time. The answer to these questions is a missing piece of the puzzle in our understanding of a widespread and ecologically important, but costly to study, group of organisms.

5. Acknowledgements

We thank Dr Leslie New and Dr Simon Wotherspoon for their advice on the analysis and Dr Stephen Nicol for his comments on the manuscript. M.J.C. was funded by Australian Research Council grant FS11020005. L-M.H. is funded by a Macquarie University Research Excellence Scholarship.

6. Data availability

CTD: https://data.aad.gov.au/metadata/records/BROKE-West_CTD_au0603

Acoustic: https://data.aad.gov.au/metadata/records/BROKE-West_hydroacoustic_dataset

7. References

ALONZO, S. H. & MANGEL, M. 2001. Survival strategies and growth of krill: avoiding predators in space and time. *Marine Ecology Progress Series,* 209**,** 203-217.

AMOS, A. F. 1984. Distribution of krill (Euphausia superba) and the hydrography of the Southern Ocean: Large-scale processes. *Journal of Crustacean Biology,* 4**,** 306-329.

ARLOT, S. & CELISSE, A. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys,* 4**,** 40-79.

ATKINSON, A., MEYER, B., BATHMANN, U., STÜBING, D., HAGEN, W. & SCHMIDT, K. 2002. Feeding and energy budget of Antarctic krill Euphausia superba at the onset of winter-II. Juveniles and adults. *Limnology and Oceanography,* 47**,** 953-966.

ATKINSON, A., SHREEVE, R., HIRST, A. G., ROTHERY, P., TARLING, G. A., POND, D. W., KORB, R. E., MURPHY, E. J. & WATKINS, J. L. 2006. Natural growth rates in Antarctic krill (Euphausia superba): II. Predictive models based on food, temperature, body length, sex, and maturity stage. *Limnology and Oceanography,* 51**,** 973-987.

BATES, D., MAECHLER, M., BOLKER, B. & WALKER, S. 2015. Fitting Linear Mixed-Effects Models Using lme4 *Journal of Statistical Software,* 67**,** 1-48.

BRIERLEY, A. & COX, M. J. 2010. Shapes of Krill Swarms and Fish Schools Emerge as Aggregation Members Avoid Predators and Access Oxygen. *Current Biology,* 20**,** 1758-1762.

BRIERLEY, A. S. & COX, M. J. 2015. Fewer but not smaller schools in declining fish and krill populations. *Current Biology,* 25**,** 75-79.

DALY, K. & MACAULAY, M. 1991. Influence of physical and biological mesoscale dynamics on the seasonal distribution and behavior of Euphausia superba in the Antarctic marginal ice zone. *Marine ecology progress series. Oldendorf,* 79**,** 37-66.

DAVIES, G. M. & GRAY, A. 2015. Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). *Ecology and Evolution,* 5**,** 5295-5304.

EVERSON, I. 2008. *Krill: biology, ecology and fisheries*, John Wiley & Sons.

GODLEWSKA, M., KLUSEK, Z. & WARSZAWY, P. 1988. The density structure of krill aggregations and their diurnal and seasonal changes (BIOMASS III, October-November 1986 and January 1987). *Pol. Polar Res,* 9**,** 357-366.

GROENEVELD, J., JOHST, K., KAWAGUCHI, S., MEYER, B., TESCHKE, M. & GRIMM, V. 2015. How biological clocks and changing environmental conditions determine local population growth and species distribution in Antarctic krill (Euphausia superba): a conceptual model. *Ecological Modelling,* 303**,** 78-86.

HAMNER, W. M. & HAMNER, P. P. 2000. Behavior of Antarctic krill (Euphausia superba): schooling, foraging, and antipredatory behavior. *Canadian Journal of Fisheries and Aquatic Sciences,* 57**,** 192-202.

HANLEY, J. A. & MCNEIL, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143**,** 29-36.

HARRISON, L.-M. K., COX, M. J., SKARET, G. & HARCOURT, R. 2015. The R package EchoviewR for automated processing of active acoustic data using Echoview. *Frontiers in Marine Science,* 2.

HOFMANN, E. E. & HÜSREVOĞLU, Y. S. 2003. A circumpolar modeling study of habitat control of Antarctic krill (Euphausia superba) reproductive success. *Deep Sea Research Part II: Topical Studies in Oceanography,* 50**,** 3121-3142.

HOFMANN, E. E. & MURPHY, E. J. 2004. Advection, krill, and Antarctic marine ecosystems. *Antarctic Science,* 16**,** 487-499.

JARVIS, T., KELLY, N., KAWAGUCHI, S., VAN WIJK, E. & NICOL, S. 2010. Acoustic characterisation of the broad-scale distribution and abundance of Antarctic krill (*Euphausia superba*) off East Antarctica (30 - 80E) in January-March 2006. *Deep Sea Research II,* 57**,** 916-933.

KAWAGUCHI, S., KING, R., MEIJERS, R., OSBORN, J. E., SWADLING, K. M., RITZ, D. A. & NICOL, S. 2010. An experimental aquarium for observing the schooling behaviour of Antarctic krill (Euphausia superba). *Deep Sea Research Part II: Topical Studies in Oceanography,* 57**,** 683-692.

KLEVJER, T. A., TARLING, G. A. & FIELDING, S. 2010. Swarm characteristics of Antarctic krill Euphausia superba relative to the proximity of land during summer in the Scotia Sea. *Marine Ecology Progress Series,* 409**,** 157-170.

KOCK, K.-H., REID, K., CROXALL, J. & NICOL, S. 2007. Fisheries in the Southern Ocean: an ecosystem approach. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 362**,** 2333-2349.

KRAFFT, B. A., SKARET, G., KNUTSEN, T., MELLE, W., KLEVJER, T. A. & SØILAND, H. 2012. Antarctic krill swarm characteristics in the Southeast Atlantic sector of the Southern Ocean. 465**,** 69-83.

LAWSON, G. L., WIEBE, P. H., ASHJIAN, C. J. & STANTON, T. K. 2008. Euphausiid distribution along the Western Antarctic Peninsula—Part B: Distribution of euphausiid aggregations and biomass, and associations with environmental features. *Deep Sea Research Part II: Topical Studies in Oceanography,* 55**,** 432-454.

MARIN, VICTOR H. & DELGADO, L. E. 2001. A Spatially Explicit Model of the Antarctic Krill Fishery off the South Shetland Islands. *Ecological Applications,* 11**,** 1235-1248.

MEIJERS, A. J. S., KLOCKER, A., BINDOFF, N. L., WILLIAMS, G. D. & MARSLAND, S. J. 2010. The circulation and water masses of the Antarctic shelf and continental slope between 30 and 80 degrees East. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57**,** 723-737.

MEYER, B. 2012. The overwintering of Antarctic krill, Euphausia superba, from an ecophysiological perspective. *Polar Biology,* 35**,** 15-37.

MEYER, B., AUERSWALD, L., SIEGEL, V., SPAHIC, S., PAPE, C., FACH, B., TESCHKE, M., LOPATA, A. L. & FUENTES, V. 2010. Seasonal variation in body composition, metabolic activity, feeding, and growth of adult krill Euphausia superba in the Lazarev Sea. *Marine Ecology Progress Series,* 398**,** 1-18.

MURPHY, E. J., THORPE, S. E., WATKINS, J. L. & HEWITT, R. 2004. Modeling the krill transport pathways in the Scotia Sea: spatial and environmental connections generating the seasonal distribution of krill. *Deep Sea Research Part II: Topical Studies in Oceanography,* 51**,** 1435-1456.

MYRIAX 2015. Echoview Software, version 6.1.35.26153. Hobart, Australia.

NAKAGAWA, S. & SCHIELZETH, H. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution,* 4**,** 133-142.

NICOL, S. 2003a. Krill and currents—physical and biological interactions influencing the distribution of Euphausia superba. *Ocean and Polar Research,* 25**,** 633-644.

NICOL, S. 2003b. Living krill, zooplankton and experimental investigations: a discourse on the role of krill and their experimental study in marine ecology. *Marine and Freshwater Behaviour and Physiology,* 36**,** 191-205.

NICOL, S. 2006. Krill, Currents, and Sea Ice: Euphausia superba and Its Changing Environment. *BioScience,* 56**,** 111-120.

NICOL, S., MEINERS, K. & RAYMOND, B. 2010. BROKE-West, a large ecosystem survey of the South West Indian Ocean sector of the Southern Ocean, 30°E–80°E (CCAMLR Division 58.4.2). *Deep Sea Research Part II: Topical Studies in Oceanography,* 57**,** 693-700.

ORSI, A. H. & HARRIS, U. 2001 (Updated 2015). Locations of the various fronts in the Southern Ocean Australian Antarctic Data Centre - CAASM Metadata (https://data.aad.gov.au/metadata/records/southern_ocean_fronts).

PERISSINOTTO, R., PAKHOMOV, E. A., MCQUAID, C. D. & FRONEMAN, P. W. 1997. In situ grazing rates and daily ration of Antarctic krill Euphausia superba feeding on phytoplankton at the Antarctic Polar Front and the Marginal Ice Zone. *Marine Ecology Progress Series,* 160**,** 77-91.

PINHEIRO, J., BATES, D., DEBROY, S., SARKAR, D. & TEAM, R. C. 2016. nlme: Linear and Nonlinear Mixed Effects Models R package version 3.1-128.

PIÑONES, A., HOFMANN, E. E., DALY, K. L., DINNIMAN, M. S. & KLINCK, J. M. 2013. Modeling the remote and local connectivity of Antarctic krill populations along the western Antarctic Peninsula. *Marine Ecology Progress Series,* 481**,** 69-92.

PIÑONES, A., HOFMANN, E. E., DINNIMAN, M. S. & DAVIS, L. B. 2016. Modeling the transport and fate of euphausiids in the Ross Sea. *Polar Biology,* 39**,** 177-187.

PRÉZELIN, B. B., HOFMANN, E. E., MOLINE, M. & KLINCK, J. M. 2004. Physical forcing of phytoplankton community structure and primary production in continental shelf waters of the Western Antarctic Peninsula. *Journal of Marine Research,* 62**,** 419-460.

QUETIN, L. B. & ROSS, R. M. 1991. Behavioral and Physiological Characteristics of the Antarctic Krill, Euphausia superba. *American Zoologist,* 31**,** 49-63.

R DEVELOPMENT CORE TEAM 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

RICHERSON, K., WATTERS, G. M., SANTORA, J. A., SCHROEDER, I. D. & MANGEL, M. 2015. More than passive drifters: a stochastic dynamic model for the movement of Antarctic krill. *Marine Ecology Progress Series,* 529**,** 35-48.

SABA, G. K., FRASER, W. R., SABA, V. S., IANNUZZI, R. A., COLEMAN, K. E., DONEY, S. C., DUCKLOW, H. W., MARTINSON, D. G., MILES, T. N., PATTERSON-FRASER, D. L., STAMMERJOHN, S. E., STEINBERG, D. K. & SCHOFIELD, O. M. 2014. Winter and spring controls on the summer food web of the coastal West Antarctic Peninsula. *Nature Communications,* 5.

SALA, A., AZZALI, M. & RUSSO, A. 2002. Krill of the Ross Sea: distribution, abundance and demography of Euphausia superba and Euphausia crystallorophias during the Italian Antarctic Expedition (January-February 2000). *Scientia Marina,* 66.

SANTORA, J. A., SYDEMAN, W. J., SCHROEDER, I. D., REISS, C. S., WELLS, B. K., FIELD, J. C., COSSIO, A. M. & LOEB, V. J. 2012. Krill space: a comparative assessment of mesoscale structuring in polar and temperate marine ecosystems. *ICES Journal of Marine Science: Journal du Conseil,* 69**,** 1317-1327.

SCAMBOS, T., HARAN, T. J., FAHNESTOCK, M., PAINTER, T. & BOHLANDER, J. 2007. MODIS-based Mosaic of Antarctica (MOA) data sets: continent-wide surface morphology and snow grain size. *Remote Sensing of Environment,* 111**,** 242-257.

SCHAAFSMA, F. L., DAVID, C., PAKHOMOV, E. A., HUNT, B. P. V., LANGE, B. A., FLORES, H. & VAN FRANEKER, J. A. 2016. Size and stage composition of age class 0 Antarctic krill (Euphausia superba) in the ice–water interface layer during winter/early spring. *Polar Biology,* 39**,** 1515-1526.

SCHMIDT, K., ATKINSON, A., VENABLES, H. J. & POND, D. W. 2012. Early spawning of Antarctic krill in the Scotia Sea is fuelled by "superfluous" feeding on non-ice associated phytoplankton blooms. *Deep Sea Research Part II: Topical Studies in Oceanography,* 59–60**,** 159-172.

SIEGEL, V. 2016. Biology and Ecology of Antarctic Krill. *Advances in polar ecology*.

SIEGEL, V., REISS, C. S., DIETRICH, K. S., HARALDSSON, M. & ROHARDT, G. 2013. Distribution and abundance of Antarctic krill (Euphausia superba) along the Antarctic Peninsula. *Deep Sea Research Part I,* 77**,** 63-74.

SILK, J. R. D., THORPE, S. E., FIELDING, S., MURPHY, E. J., TRATHAN, P. N., WATKINS, J. L. & HILL, S. L. 2016. Environmental correlates of Antarctic krill distribution in the Scotia Sea and southern Drake Passage. *ICES Journal of Marine Science*.

TAKI, K., YABUKI, T., NOIRI, Y., HAYASHI, T. & NAGANOBU, M. 2008. Horizontal and vertical distribution and demography of euphausiids in the Ross Sea and its adjacent waters in 2004/2005. *Polar biology,* 31**,** 1343-1356.

TARLING, G. A., KLEVJER, T., FIELDING, S., WATKINS, J., ATKINSON, A., MURPHY, E., KORB, R., WHITEHOUSE, M. & LEAPER, R. 2009. Variability and predictability of Antarctic krill swarm structure. *Deep Sea Research Part I: Oceanographic Research Papers,* 56**,** 1994-2012.

TARLING, G. A. & THORPE, S. E. 2014. Instantaneous movement of krill swarms in the Antarctic Circumpolar Current. *Limnology and Oceanography,* 59**,** 872-886.

TREMBLAY, N. & ABELE, D. 2016. Response of three krill species to hypoxia and warming: an experimental approach to oxygen minimum zones expansion in coastal ecosystems. *Marine Ecology,* 37**,** 179-199.

TYNAN, C. T. 1998. Ecological importance of the Southern Boundary of the Antarctic Circumpolar Current. *Nature,* 392**,** 708-710.

WATKINS, J. & MURRAY, A. 1998. Layers of Antarctic krill, Euphausia superba: are they just long krill swarms? *Marine Biology,* 131**,** 237-247.

WILLIAMS, G. D., NICOL, S., AOKI, S., MEIJERS, A. J. S., BINDOFF, N. L., IIJIMA, Y., MARSLAND, S. J. & KLOCKER, A. 2010. Surface oceanography of BROKE-West, along the Antarctic margin of the south-west Indian Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57**,** 738-757.

ZUUR, A. F., IENO, E. N., WALKER, N. J., SAVELIEV, A. A. & SMITH, G. M. 2009. *Mixed Effects Models and Extensions in Ecology with R,* New York, Springer.

Chapter 6

A SOUTHERN OCEAN ARCHIPELAGO ENHANCES FEEDING OPPORTUNITIES FOR A KRILL PREDATOR

Authors:

Lisa-Marie K. Harrison[1], Kimberly Goetz[2], Martin J. Cox[3], Robert Harcourt[1]

Affiliations:

[1]Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, 2109, New South Wales, Australia

[2]National Institute of Water and Atmospheric Research, 301 Evans Bay Parade, Wellington, 6021, New Zealand

[3]Australian Antarctic Division, 203 Channel Highway, Kingston, 7050, Tasmania, Australia

Abstract

Productivity in the oceans is heightened around oceanographic and bathymetric features such as fronts and islands. This can have a flow-on effect, providing increased food availability for higher trophic level species. Using data from a combined visual and acoustic survey, we examine the hypothesis that higher Antarctic krill (*Euphausia superba*) density provides a lucrative resource for humpback whales (*Megaptera novaeangliae*) at a remote Antarctic feeding area, the Balleny Islands (67°S, 164°E). We assess whale presence at the foraging area in relation to prey, productivity and environmental variables using density surface modelling. We found stark differences in krill swarms at the islands compared with the adjacent open water. Swarms were twice as dense and three times more numerous at the Balleny Islands compared with open water, suggesting that the islands offer a profitable feeding opportunity. At the feeding area, humpback whales were found in deeper and more productive waters with medium krill densities. These relationships, along with the high krill availability around the islands, may relate to the Island Mass Effect. Our krill swarm and spatial analysis suggests that island feeding areas are important resources. We have provided the first quantitative study of habitat use by whales in an area that has rarely been visited, but has recently become a part of the world's largest marine protected area.

Keywords: Antarctic krill, humpback whale, Southern Ocean, density surface model, foraging, Island Mass Effect, prey field

1. Introduction

Productivity has a bottom up effect on ecosystems and is critical for sustaining large Antarctic krill (*Euphausia superba*) populations and the Southern Ocean predators that forage on them (Groeneveld et al., 2015, Ware and Thomson, 2005). Four key features limit productivity in the oceans: light, nutrients, mixing and grazing (Barnes and Hughes, 2009; pg 32). Light and macronutrient/iron availability are the most important limiting factors for phytoplankton growth in the Southern Ocean (Lancelot et al., 2000, Smith and Lancelot, 2004). Melting ice creates amenable conditions for productivity by: 1) releasing nutrients and trace metals, 2) stabilizing the upper water column through stratification of low salinity water and, 3) seeding blooms through the release of algae from the ice (Smith and Nelson, 1986, Sedwick and DiTullio, 1997). Oceanic features such as bathymetry, islands and frontal zones are also highly productive areas due to the nutrients brought up by upwelling (Bost et al., 2009, Laubscher et al., 1993, Gove et al., 2016). Large swarms of krill that measure tens to hundreds of kilometres can be concentrated around these locations for months at a time (Siegel, 2016; pg 300).

Frontal zones might provide a lucrative feeding area, but lack the predictability brought about by the fixed location of high productivity around islands, which is likely highly important to migratory predators searching for foraging areas. The increased productivity around islands, due to modification of the physical oceanography, is known as the Island Mass Effect (Elliott et al., 2012, Gove et al., 2016). There are several mechanisms through which the Island Mass Effect operates. At South Georgia and the Kerguelen and Crozet Islands, the Island Mass Effect causes increased productivity through the release of iron (Blain et al., 2001, Planquette et al., 2007, Atkinson et al., 2001), a limiting nutrient that has been flagged as a possible cause of the high-nutrient low-chlorophyll status observed

throughout most of the Southern Ocean (Boyd et al., 2000). Increases in productivity and zooplankton around the Prince Edward Archipelago have also been attributed to the Island Mass Effect through nutrient inputs (Boden, 1988). Increased productivity and a general boosting of higher trophic levels in the food-chain also occurs through the creation of a stable surface layer by meltwater and rainwater run-off, as seen around Bouvet and the South Sandwich Islands (Perissinotto et al., 1992).

Apex predators aggregate around biological hotspots, such as fronts and seamounts (Bost et al., 2009, Scales et al., 2014, Morato et al., 2010), although there are likely temporal lags between productivity and predator presence. In Southern Ocean ecosystems, krill populations are known to increase after multiple years of high productivity due to increased spawning success, which could create an annual lag (Saba et al., 2014). Shorter time lags also occur, due to the time taken for predators to find prey (Sims et al., 2008).

The world's largest predators are the baleen whales, who collectively consume an estimated 3 – 120 million tonnes of Antarctic krill annually (Siegel, 2016; pg 325). They migrate annually from temperate and tropical breeding grounds in winter to cold but energetically-rich polar waters in summer, where they must consume enough prey to sustain themselves until the next summer. Hence the availability of krill at the summer foraging grounds is critical for survival and population growth (Mori and Butterworth, 2004, Nicol et al., 2008). Baleen whales are known to aggregate around frontal zones for feeding where there is increased productivity (Doniol-Valcroze et al., 2007). Resource partitioning is evident between humpback, fin and minke whales, which preferentially, although not exclusively, target different age classes and species of krill (Santora et al., 2010, Friedlaender et al., 2009).

Globally, there have been numerous studies linking whale sightings to surface productivity and prey density. In the Northern Hemisphere, positive correlations between feeding humpback whales and Chlorophyll-a, an indicator of productivity, have been recorded in the California Current system (Tynan et al., 2005, Thompson et al., 2012). Areas of upwelling are important, with feeding humpback whales in the Gulf of St. Lawrence clustering around frontal zones (Doniol-Valcroze et al., 2007). Regression analyses have revealed different but highly non-linear trends recorded between whale sightings and Chlorophyll-a in the Bering Sea (Zerbini et al., 2015) and the Western Antarctic Peninsula (Friedlaender et al., 2006). The reason for the different non-linear trends could be due to temporal lags between Chlorophyll-a and whales, or a spatial mismatch between Chlorophyll-a and whale sightings. While the reported relationships of humpback whale sightings and Chlorophyll-a have been varied, consistent relationships have been observed with krill. Positive relationships between humpback whale sightings and krill density have been documented in the Antarctic (Herr et al., 2016, Murase et al., 2002, Reid et al., 2000), the Barents Sea (Ressler et al., 2015), North Greenland (Laidre et al., 2010) and the California Current system (Benson et al., 2002).

The foraging grounds of humpback whales in the Southern Hemisphere, are divided into six areas. There is an interchange between breeding populations in all areas with the exception of the West Antarctic Peninsula (Area 1), where whales from the south-eastern Pacific Ocean (Breeding Stock G) reliably return (Amaral et al., 2016). In addition to coastal Antarctica, numerous Southern Ocean islands are feeding grounds for humpback whales including South Georgia, the South Sandwich Islands (Horton et al., 2011, Zerbini et al., 2006) and the Balleny Islands (Constantine et al., 2014).

The Balleny Islands are an uninhabited cluster of Antarctic islands located between New Zealand and the Antarctic continent (67ºS, 164ºE; see Figure 1 for map) and have recently been identified as a feeding ground for the east-Australian (E1) population of humpback whales (Constantine et al., 2014). The E1 population migrates annually from the warm calving grounds of the Great Barrier Reef to the krill-rich Antarctic waters to feed (Gales et al., 2009, Smith et al., 2012), and photo-ID/genotyping has matched 38 whales seen at the Balleny Islands to previous sightings off Eastern Australia, New Zealand and New Caledonia (Constantine et al., 2014). The Balleny Islands cover only a small area, extending approximately 150 km latitudinally and longitudinally, and it is not known whether they provide preferred foraging habitat for whales or how krill is distributed around the islands. This is important information for baseline monitoring and future management of the Ross Sea Marine Protected Area, the world's largest Marine Protected Area, which indirectly benefits whales through fishing bans and currently allows whaling.

We assess whether humpback whales are attracted to the Balleny Islands due to high prey availability by comparing krill swarm metrics list them around the islands to an adjacent area of open ocean. Around the islands we evaluate how prey availability, bathymetry, productivity and indicators of upwelling such as salinity and temperature relate to whale distribution at the feeding ground. We hypothesise that there will be more krill at the Balleny Islands than in open water and that around the islands, whales will aggregate in areas of high krill density and upwelling, as indicated by high productivity, high salinity and low temperature. We use Kolmogorov-Smirnov tests to compare swarms at and away from the islands, and a Density Surface Model to test our hypotheses about whale

distribution around the islands using data from a 2015 wildlife survey of the Balleny Islands.

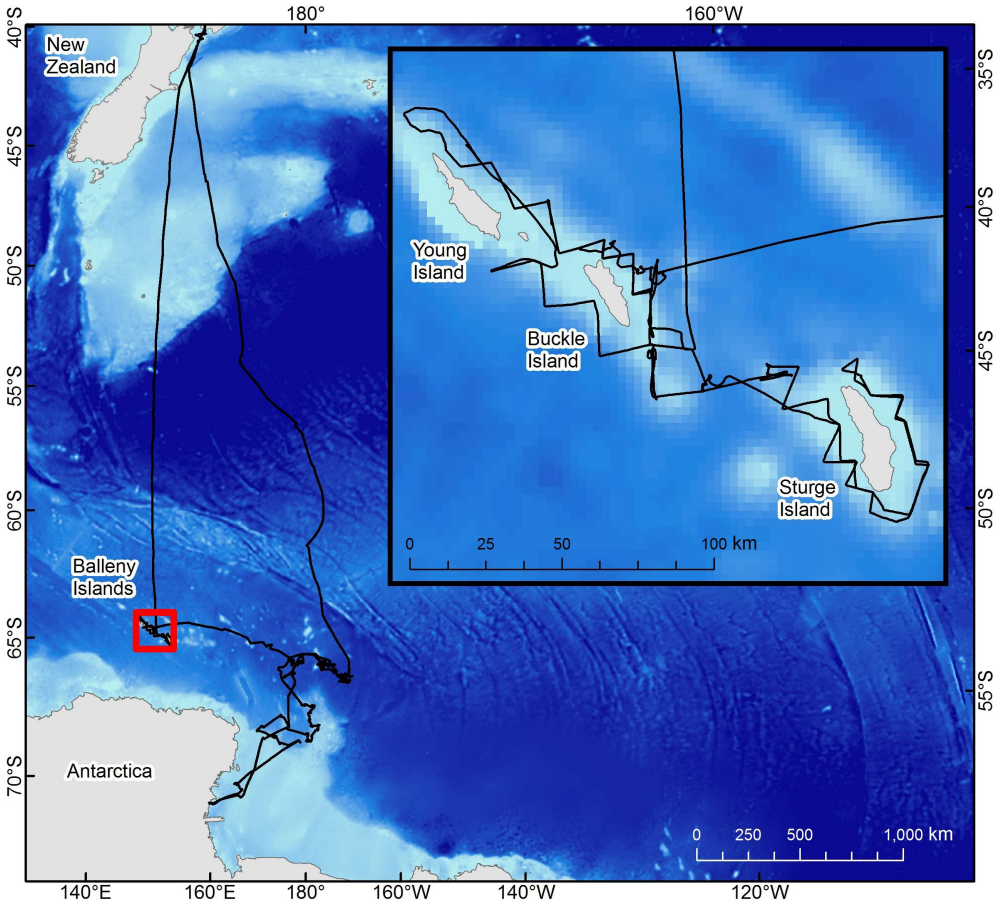


Figure 1 *Map showing the Balleny Islands with the cruise track overlayed in black.*

2. Methods

2.1. Survey data

The data are from a 2015 marine mammal survey with contemporaneous prey field mapping at the Balleny Islands undertaken by the National Institute of Water and Atmospheric Research on board the RV Tangaroa. The survey around the islands took place from the 2nd – 6th of February 2015 (Figure 1). Marine mammal sighting data were collected as per Kinzey et al (2000) by two observers located in the ship's "Monkey Island", located one level above the bridge but below the crow's nest. Both observers scanned the waters 180 degrees in front of the vessel with 7X50 handheld Fujinon binoculars and recorded the reticle and angle between the trackline and the ship for each sighting. Perpendicular distance to each sighting was calculated using the reticle, the angle between the trackline and the sighting and mean observer eye height (15 m) using the equation on page 12 of Kinzey et al (2000). On-effort times were defined as times when two observers were actively searching for sightings from Monkey Island. Off-effort times are when there were one or no observers actively searching for whales.

Underway oceanographic data were collected using a Wetlabs ECO-TRIPLET (Chlorophyll-a) and a Seabird 21 thermosalinograph (salinity and temperature). Acoustic data were collected using a calibrated Simrad EK60 Echosounder operating at 38 and 120 kHz frequencies and krill identified using standard 'dB-difference' techniques (Cox et al., 2011). The acoustic backscatter was processed to obtain volumetric density of krill (gm^{-3}) using the software package Echoview (Echoview, Hobart, Australia, 2015) and the R package EchoviewR (Harrison et al., 2015). Integration intervals were vertically integrated to the shallowest of either the seabed echo or the top 250 m of the water column, and had a mean length of 1600 m.

2.2. Statistical Analysis

2.2.1. Comparison with outside the Balleny Islands

To assess whether krill were more abundant around the Balleny Islands compared to adjacent open waters, swarms around the islands were compared to swarms encountered one day after leaving the Balleny Islands. Swarms were identified and extracted using Echoview (Cox et al., 2011). A threshold of -70dB was used, and krill swarms were identified using a decibel difference window ($Sv_{120kHz} - Sv_{38kHz}$) between 0.37 and 12 dB. Encounter rate was calculated as "swarms encountered per kilometre" to account for differences in ship speed. Swarm internal volumetric density (gm^{-3}) and swarm length (m) distributions at and away from the Balleny Islands were compared using a one-sided Kolmogorov-Smirnov test.

2.2.2. Whale distribution around the islands

Density Surface Models (DSMs) can provide a flexible method of modelling line transect environmental and sighting data by merging Generalised Additive Modelling (GAM), to incorporate non-linear environmental coefficients, with i) distance sampling, to account for imperfect detection, and ii) survey design, to allow for opportunistic surveys with repeat sampling and unequal effort (Miller et al., 2013). A DSM with a half-normal detection function was used to evaluate whether whale sightings are correlated with environmental or biological features whilst accounting for imperfect detectability. The detection function accounts for imperfect detectability, which may vary if larger whale groups are easier to see, so a coefficient for sighting group size was considered. The best detection function (hazard-rate, half-normal) and the support for the group size covariate was assessed using Akaike Information Criterion (AIC). The DSM was fit using the best detection function, and a GAM linking whale sightings to environmental variables which

includes a s(x, y) surface. A soap-film smoother (Wood et al., 2008) was used so that the s(x, y) smoothed around islands, rather than through, with island boundaries input into the smoother as polygons. As the Poisson family with an over-dispersion parameter was used for the GAM component of the DSM, AIC was not available because the model is not based on a full likelihood so a p-value backwards step-wise model selection was used instead ($\alpha = 0.05$).

Variables considered in the DSM are summarised in Table 1. AMSR2 satellite-sensed daily sea ice coverage data were sourced from the University of Bremen (Spreen et al., 2008; http://www.iup.uni-bremen.de:8084/amsr2data/) however sea ice coverage was low throughout the survey area as the RV Tangaroa did not travel in areas of heavy ice coverage, so this variable was not included in the model. For the DSM, krill data were vertically and horizontally integrated and mean density per integration interval was calculated, rather than extracting individual swarms as per the island/open water comparision. The DSM was fit using the *dsm* package (Miller et al., 2016) in R (R Development Core Team, 2014; version 3.2.3) and R-studio (RStudio, 2014; version 0.99.892).

Table 1 *Summary of explanatory environmental and biological variables measured around the Balleny Islands that are investigated as correlates with humpback whale sightings using a Density Surface Model.*

Variable	**Short name**	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
Bottom depth (m)	*depth*	85	2283	741	542
Chlorophyll-a (μgL^{-1})	*chl*	0.22	4.41	1.49	1.13
Salinity (psu)	*salinity*	33.41	34.15	33.78	0.14
Sea Surface Temperature (^oC)	*SST*	-1.6	-0.6	-1.11	0.24
Krill density (gm^{-3})	*krill*	0	2307	46.54	221.6
Easting (m)	*x*	372842	504886	438471	37622
Northing (m)	*y*	2495177	2666548	2579171	47198

DSMs can incorporate both unequal effort over a survey area and repeated transects. To ensure that repeated visits to the same area were considered independently in the DSM, the survey was coded in four separate transects based on parts of the survey with continuous on-effort times, where observers were actively searching for sightings. 'Segments' within each transect were the krill integration intervals and the underway environmental data were interpolated onto these intervals. As sighting distances reached up to 13.8 km from the transect, sightings were matched to the closest segment to ensure that the most relevant set of environmental variables were associated with that sighting.

DSMs allow the user to specify the segment area to account for unequal effort. At times, the vessel was much closer to the islands than the segment width (taken to be the

maximum distance a sighting was observed at = 13.8 km), causing a reduction in segment area because the land restricted the field of view. To account for this, a polygon of each segment was overlayed over a polygon of the Balleny Islands land masses and the percentage overlap was calculated. For segments where an overlap occurred, the segment area was reduced by the percentage of overlap with land.

3. Results

There were 63 sightings of humpback whales over 39.2 hours of time on effort. Group sizes ranged from 1 – 7 individuals (mean = 2.16; SD = 1.35).

3.1. Comparison to areas outside of the Balleny Islands

The sighting rate of 1.7 sightings/hrs around the Balleny Islands dropped to only 0.17 sightings/hr the next day. The encounter rate of krill swarms around the Balleny Islands was also much higher than the day after the ship left the island area. Krill swarms around the islands were encountered at a rate of 0.15 swarms/hr while on the next day the encounter rate was 0.05 swarms/hr. While there were more swarms encountered around the islands, they were significantly shorter in length than those encountered the next day ($p < 0.001$, Table 2). Despite being shorter in length, the island based krill swarms were denser ($p = 0.061$, Table 2), however this difference was not statistically significant.

Table 2 *Comparison of krill swarm metrics around the Balleny Islands and in the adjacent open water.*

		Open Water	**Balleny Islands**
	Effort (km)	512.2	1083.2
	Number of swarms	25	160
	Swarms/hour	0.05	0.15
Swarm length (m)	Minimum	159	32
	Maximum	1012	2789
	Mean	350	288
Swarm density (gm^{-3})	Minimum	3	2
	Maximum	198	1463
	Mean	33	57

3.2. Whale distribution around the islands

The best DSM included easting, northing, Chlorophyll-a, krill density, salinity and bottom depth (Table 3) and had a Deviance Explained of 46.8%. Model selection results for backwards p-value based selection are available in the supplementary materials (Table S2). Except for Chlorophyll-a, which displayed a linear relationship; all other variables had non-linear relationships with whale count and hence were modelled as smooth terms (Figure 3).

Table 3 *Output of GAM results from Density Surface Model. Chlorophyll-a and the intercept are parametric coefficients and s() represents a smooth term.*

Family = Poisson with a log-link function. The over dispersion parameter was estimated as 3.33, resulting in 46.8% deviance explained.			
Parametric Coefficients			
	Estimate	**Standard Error**	**P-value**
Intercept	-18.229	0.433	<0.001
chl	0.514	0.192	0.008
Smooth terms			
	Estimated Degrees of Freedom	**P-value**	
s(*x, y*)	5.1	0.002	
s(*krill*)	3.6	0.017	
s(*salinity*)	5.7	0.029	
s(*depth*)	2.2	0.002	

The s(*x, y*) surface models extra spatial variation that is not accounted for by the other variables in the model. There was a 'hotspot' of high counts of humpback whales to the East of Young Island, and an area of low counts between the two southern islands that was not accounted for by the other variables (Figure 2).

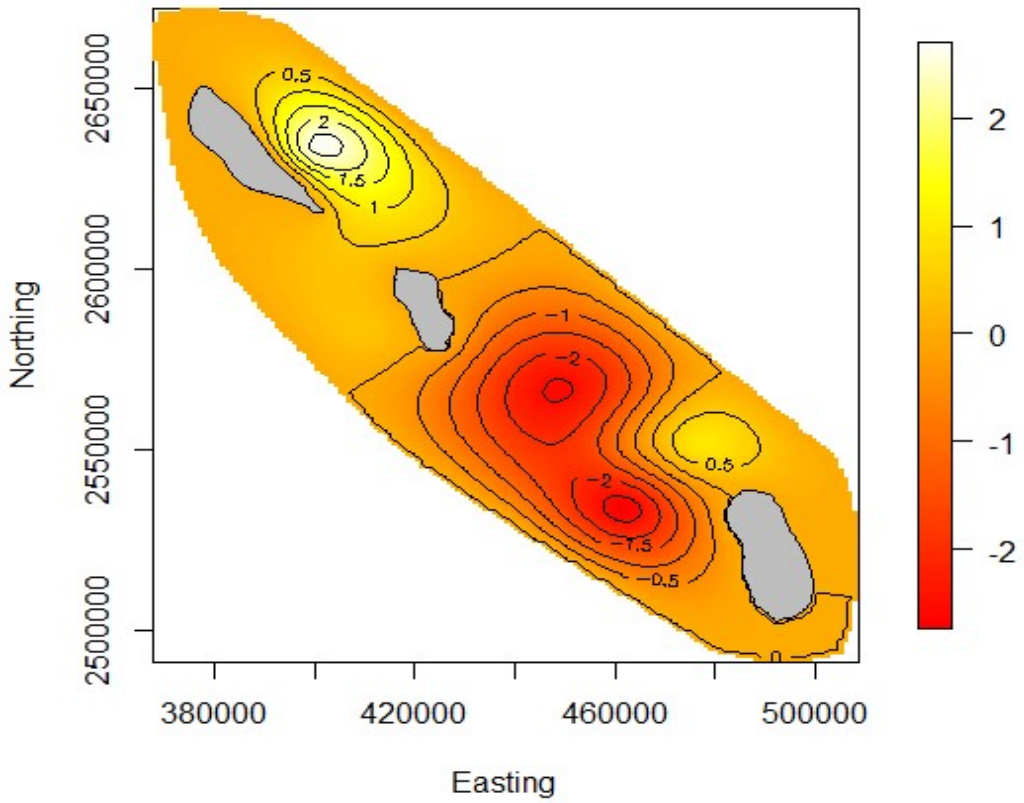


Figure 2 *Contour plot of s(x, y) Easting-Northing surface of relative whale count on the natural log scale from the density surface model. The grey polygons are the Balleny Islands land masses. This surface models the extra spatial variability not accounted for by the other predictors. The outer boundary was calculated from a convex hull of the outer transect points and was increased in width by the sighting distance.*

While the krill-whale relationship (Figure 3a & 3b) indicates that there is a quadratic like relationship when krill is >500 gm^{-2}, there were only 4 observations driving this relationship, so it should be interpreted with caution. This is reflected in the large confidence intervals in this plot. Relative frequencies of krill decline exponentially (Figure 4).

The relationship between whale count and salinity (Figure 3c) appeared to overfit even when the basis dimension was restricted. The 'wavy' line hovers around zero, indicating

that while this relationship was statistically significant, its effective influence is likely low particularly because the range of salinity values was low (33.41 – 34.15 psu). The relationship between whale count and bottom depth (Figure 3d) appeared to be the most robust of these smooth terms, showing higher whale numbers in deeper water. The estimated number of whales decreases sharply at 1900 m depth and the confidence intervals become extremely large, which is driven by the lack of data (only three observations) at depths greater than 1900 m.

Chlorophyll-a was initially included as a smooth term in the DSM, however its effective degrees of freedom was 1 and the smooth term plot was linear so this term was changed to a parametric coefficient (Figure 3e). For every 1 μgL^{-1} increase in Chlorophyll-a, there was an estimated increase of 1.7 whale sightings.

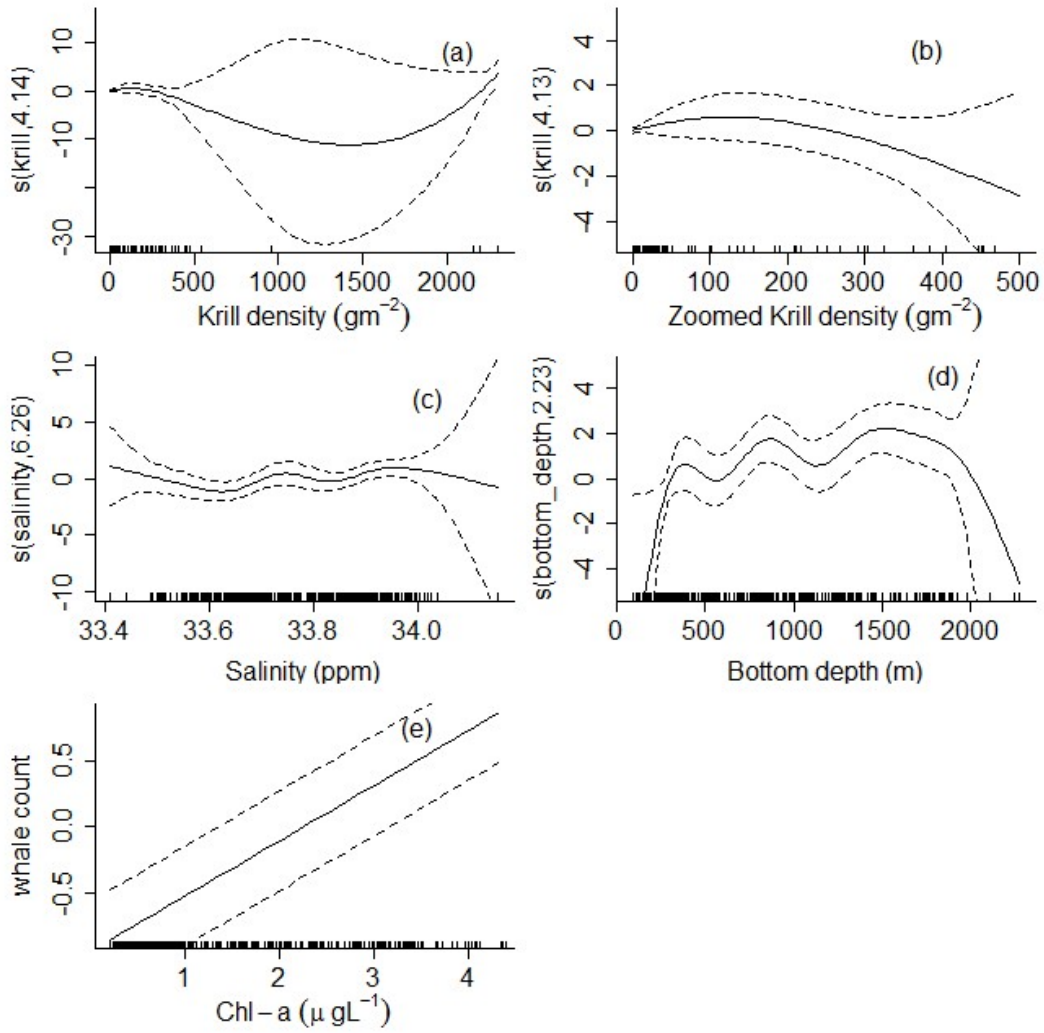


Figure 3 *Smooth effects of a) Krill density, b) Krill density zoomed to 0-500 gm^{-2} , c) Salinity and d) Bottom depth on humpback whale sightings from the density surface model. Linear Chlorophyll-a term is shown in e) and is centred to mean=0 for consistency with other plots. Dashed lines are 2*Standard Error; the distribution of observations is given as a rug plot along the x-axes.*

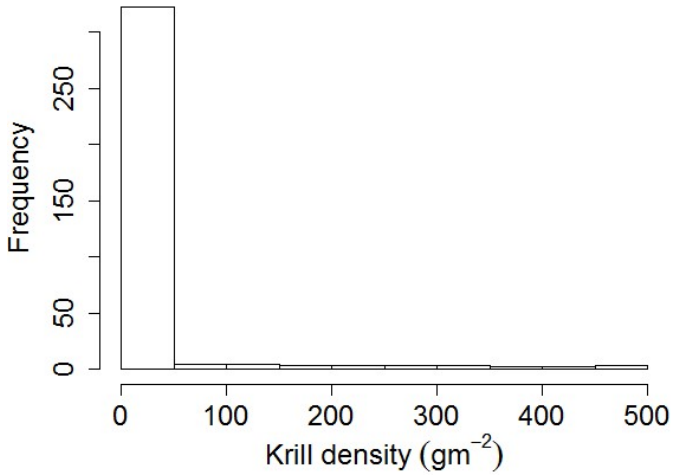


Figure 4 *Histogram of krill density to demonstrate exponential decline in frequency of higher krill densities*

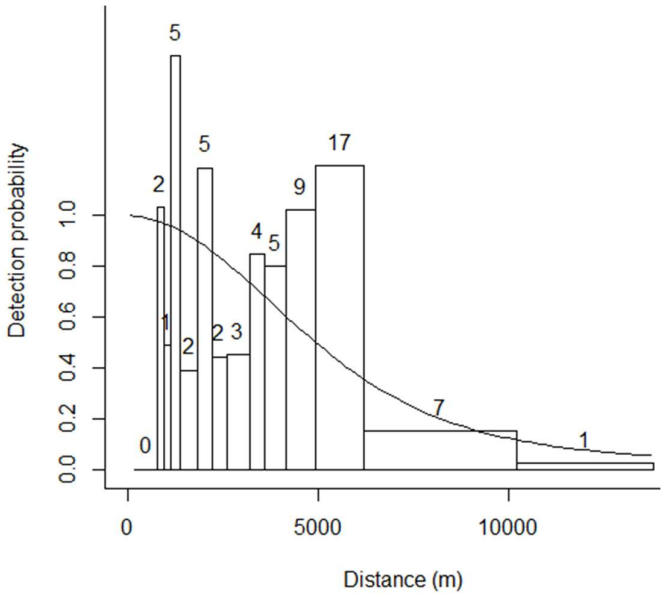


Figure 5 *Fitted half-normal detection function (solid line) with whale group size covariate. Custom sighting frequency bins of unequal size were used because the distance between reticle marks on the binoculars increases with distance from transect. The number of observations in each bin is displayed above each bar in the histogram. A truncation distance was not used because the detection function would not converge when various truncation distances were tested.*

The best detection function included a coefficient for group size (AIC_{Null} = -40.3; AIC_{Size} = -46.2), as larger groups were seen further from transect than smaller groups, most likely because they were easier to see (Figure 5). See the supplementary materials for AIC-based model selection results (Table S1).

Humpback whale abundance was predicted over a grid of 10x10 km cells, in locations where all variables were measured (Figure 6). The highest number of individuals were predicted to be in the North-East of the survey area, and this area also had the lowest Coefficient of Variation. In general, the Coefficient of Variation for each cell was high, with 36% of cells having a Coefficient of Variation of 2 or higher, indicating that the Standard Error was at least twice the value of the estimate for those cells (Figure 6b).

The total estimated humpback whale abundance from the DSM is 182 individuals (SE = 27). This prediction only includes animals in the shaded 10x10 km grid shown in Figure 6, as a count can only be calculated in areas where all environmental variables were measured. This estimate includes a correction for decreasing detectability of whales further away from the transect, i.e. the detection function.

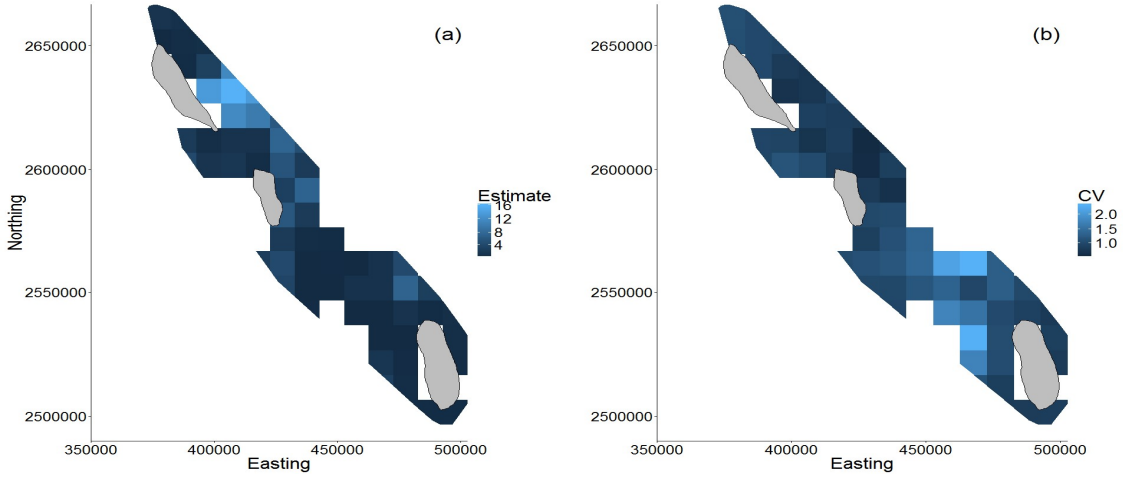


Figure 6 *Predictions from density surface model of (a) Predicted whale count and (b) Coefficient of Variation (CV) over a 10 km grid. Note: Predictions can only be made in cells where all variables were sampled.*

4. Discussion

Islands can offer a predictable and profitable feeding ground for migratory predators through enhanced productivity caused by the Island Mass Effect. We found dense, small and numerous krill swarms at the Balleny Islands compared to large diffuse layers of krill in the adjacent open water, which is likely responsible for the high number of humpback whales seen at the islands. We found that around the islands, whales aggregate in more productive areas with medium krill densities and bottom depths greater than 350m.

The number of whales seen around the Balleny Islands (1.6 sightings/hr) was not only higher than in nearby open water, but also higher than previously observed in other areas of the Southern Ocean. In the West Antarctic Peninsula, surveys in the early 2000s found encounter rates of 0.54, 0.32 and 0.55 sightings/hr (Friedlaender et al., 2006, Thiele et al., 2004) and off East Antarctica the encounter rate for a 1995 survey was 0.23 sightings/hr (Thiele et al., 2000). The sighting rate at the Balleny Islands is also higher

than previously seen around other island groups. At the South Shetland Islands, the sighting rate of humpback whales over a 5-year survey was 0.39 sightings/hr (Santora et al., 2010) and at South Georgia, only one humpback whale group was seen over 110 hours of effort (Rossi-Santos et al., 2007). The high number of whale sightings and high krill abundance at the Balleny Islands suggests this area provides important foraging habitat for humpback whales.

The threefold increase in the number of krill swarms around the Balleny Islands compared to open water could be due to increased productivity around the islands caused by the Island Mass Effect, which has been documented around other islands in the Southern Ocean (Blain et al., 2001, Planquette et al., 2007, Boden, 1988, Perissinotto et al., 1992). The smaller ($p < 0.001$), denser ($p = 0.061$) and more populous krill swarms around the Balleny Islands are likely less energetically expensive to find and consume than the few long but sparsely populated swarms in open water. Lunging during feeding incurs a large energetic cost for humpback whales, with each extra lunge resulting in decreased dive times and longer surface intervals (Goldbogen et al., 2008). Humpback whales can take advantage of dense krill swarms by repeatedly targeting a dense swarm within the same dive using a reverse-looping behaviour (Ware et al., 2011), filtering and then swallowing the prey while positioning for the next lunge (Simon et al., 2012). Feeding in an area of higher food occurrence and density could hence conserve energy if it allows for fewer lunges and a shorter time to locate swarms.

Whale distribution at the feeding ground is influenced by biological (Chlorophyll-a and krill) and environmental factors (salinity and bottom depth). The non-linear relationships with these factors (except Chlorophyll-a) highlight the need for non-parametric approaches when modelling complex data. Here we discuss the importance of the

biological and environmental factors that we found were correlated with humpback whale distribution around the Balleny Islands.

4.1. Chlorophyll-a and Krill

Productivity in the water column can have an effect on higher up trophic levels through enhancement of the food chain (Ware and Thomson, 2005). Our observed positive linear relationship between whale sightings and Chlorophyll-a indicates that the whales around the Balleny Islands are seeking areas of high productivity. These patches of high productivity could be facilitated by nutrient inputs from the Island Mass Effect, either from upwelling or nutrient runoff. The Southern Ocean is generally considered to be iron limited, and iron inputs increasing productivity due to the Island Mass Effect have been observed at three other Southern Ocean island groups (Atkinson et al., 2001, Blain et al., 2001, Planquette et al., 2007).

Whales may aggregate in high productivity areas due to high krill availability. However, we found no relationship between krill density and Chlorophyll-a. Rather than indicating a true lack of relationship between krill and phytoplankton, this may be an issue of scale. Another possibility is that Chlorophyll-a could be more persistent than krill swarms, offering whales a stable indicator of areas with generally high krill densities.

A secondary explanation for a positive correlation between whale sightings and Chlorophyll-a could be fertilization, where the nutrients (particularly iron) released by defecating whales cause phytoplankton blooms (Nicol et al., 2010). We believe this is unlikely to be the explanation for our observed relationship because the time scale for bloom formation after whale presence is in the order of 13-16 weeks (Visser et al., 2011) and we observed the relationship on a much smaller temporal and spatial scale.

For krill densities below 500gm^{-3}, we found an increasing relationship with whale count to 220gm^{-3} of krill, after which whale count decreased. This occurred despite krill frequencies decreasing exponentially after 50gm^{-2}. This indicates that the relationship we observed is likely to be due to a preference rather than in incidental occurrence because krill up to 220gm^{-2} are far more common and hence more likely to be found. At krill densities higher than 500gm^{-3}, the relationship was largely influenced by four points of unusually high krill density and should therefore be regarded with caution. A recent study in the West Antarctic Peninsula found a similar result to our model, with humpback whales seen in areas of 'medium' krill density rather than low or high density (Herr et al., 2016). Although this is a comparable *pattern* to the one we observed, in that study there were lower overall krill densities, so 'medium' was only 20 – 40 gm^{-2}. Our krill densities were significantly higher and were skewed.

When prey resources such as krill are patchily distributed, predators must decide how long to remain at the current patch and whether it is worth seeking a more lucrative patch at the risk of expending energy searching and travelling with no guarantee that such a patch exists. Optimal foraging theory dictates that the predator should leave the patch when the patch's marginal capture rate drops below the average capture rate in the habitat (Charnov, 1976). When faced with high uncertainty, optimisation is not a reliable strategy and robust satisficing, which involves maximising the robustness to uncertainty of a satisfactory outcome, is a preferred strategy (Schwarz et al., 2010). Information-gap robust satisficing is thought to occur in ecological systems more often than optimal foraging (Carmel and Ben-Haim, 2005). We found that whale sightings at the Balleny Islands were highest at medium krill densities, which could be because average patches still provide the best chance of satisfying current energy needs, given the high uncertainty

about i) whether there are better patches, ii) how much energy is required to locate these lucrative patches and iii) how far away these patches are likely to be. For whales at low krill densities, this risk becomes worth taking, or even necessary, hence the low whale numbers observed at low krill densities.

4.2. Bottom depth

We found a lower number of whale sightings where bottom depths were greater than 350m, after which the smooth relationship hovers around zero, indicating that there are fewer whales close to the islands' shores. This may be due to the dynamics of the Island Mass Effect, if upwelling caused by the disruption of currents by the islands occurs slightly offshore. Future research to characterise the island mass effect around the Balleny Islands could test this hypothesis. Reported correlations between humpback whales and bottom depth in the literature show that whales (especially females and calves) prefer shallow water (Smultea, 1994, Guidino et al., 2014, Felix and Haase, 2005), however these studies are from winter breeding grounds where the priorities of the whales are different. In our study, there may be some discrepancy between the bottom depth at the location of the actual sighting and the recorded bottom depth because we only have depth data directly below the ship, and sightings are paired to the closest point on transect. High resolution bathymetric data could be matched to the true location of each sighting to assess the discrepancy between depth at the ship and depth at the sighting. However, this variable couldn't easily be included in the DSM because bottom depth varies with distance from transect rather than being constant within a segment.

4.3. Salinity

Whale sightings showed a highly non-linear relationship with salinity, with peaks at low, medium and high densities. This occurs despite the basis dimension for the smooth term being reduced to force smoothing and minimise the chance of overfitting. The undulating relationship could be because salinity is much more consistent (33.41-34.15 psu) around the Balleny Islands than many other locations, because, apart from melting sea ice, there are no large freshwater inputs. A contrasting example is the Dominican Republic, where a salinity gradient of 1 – 32 ‰ occurs and humpback whales avoid the low salinity areas around the freshwater inputs (Mattila et al., 1994). If salinity is indicative of upwelling, its importance may depend on the whale's activity. This has been seen in British Colombia, where salinity is only associated with whale foraging in shallow waters (Gregr and Trites, 2001). We found no evidence of a salinity – bottom depth interaction, which could be because our DSM was only fit to the portion of the survey around the islands, so whales migrating over deep water are not included. If salinity depends on upwelling generated by the Island Mass Effect around the Balleny Islands, the non-linear relationship we observed could also be influenced by current patterns around the islands, which are not included in our model but would be partly accounted for by the spatial smooth, s(x, y). A dedicated investigation of any Island Mass Effect occurring around the Balleny Islands would help us better understand the relationship we have found with salinity.

4.4. Limitations and future directions

One variable missing from our model is ocean current strength and direction. The ship's Acoustic Doppler Current Profiler was not operating at the same time as the echosounder so we do not have underway current data. Satellite based current data were considered

but are collected at a larger temporal and spatial resolution than our DSM segments (mean length = 1596m) and often don't cover the Southern Ocean. The most important feature that current data could add would be an indication of upwelling locations and a characterisation of the Island Mass Effect, however because our Chlorophyll-a, SST and salinity variables would reflect areas where upwelling causes high productivity, we believe this did not significantly impact on the outcomes of our study.

The short survey time of three days makes it impossible to assess how long the whales remain around the Balleny Islands. We cannot tell from our data if the Balleny Islands are a 'stop-over' foraging ground for whales on the way to the Antarctic continent or whether they provide a foraging ground in which whales are resident there the entire summer. If the same individuals return to this area each year, these animals would be particularly vulnerable to changes in krill stocks around the Balleny Islands. While Southern Hemisphere humpback whales are believed to follow the classical feeding model, where feeding occurs only in the Southern Ocean during summer, stable isotope analysis and direct observations indicate that E1 humpback whales may diverge from this strategy and supplement their diet in temperate waters (Eisenmann et al., 2016, Owen et al., 2016, Owen et al., 2015). Due to the remote location of the Balleny Islands it is financially and logistically difficult to conduct long term monitoring to assess the role that this feeding ground plays in E1 humpback whale feeding strategies. However as technology advances it might not be long before it becomes possible to count whales in geographically isolated locations using unmanned aerial surveys (Linchant et al., 2015), gliders (Baumgartner et al., 2013), fixed passive acoustic sensors (Marques et al., 2009) or high resolution satellite imagery (Fretwell et al., 2014).

The large distances to sightings (up to 13 km) meant that often individuals could not be identified and that animals may have been recounted on a previous day. The DSM partly

accounts for this by including survey design through the specification of unique transect/sampling blocks, however our confidence in the abundance estimate would be higher if individuals could be identified. Satellite tracking of humpback whales in the West Antarctic Peninsula feeding grounds has revealed that they travel large distances of 17 – 75 km/day (Dalla Rosa et al., 2008). Given that the Balleny Islands survey area is only approximately 150km in both latitude and longitude, the whales could have potentially moved anywhere in the survey area over the three days that we observed them. Hence, identifying individuals to avoid recounting would help give a more robust abundance estimate and would allow for fluke matching to known individuals from the E1 and Oceania populations. Another issue that could potentially have down-biased our population estimate is availability bias, where whales were in the survey area but were not available for sampling because they were diving or inclement weather conditions made it impossible to see them.

5. Conclusion

The unusually high numbers of humpback whales seen at the Balleny Islands are likely attracted by the threefold increase in krill swarms compared to the adjacent open water. These abundant and concentrated krill swarms may be easier to find and forage on than the spread-out swarms seen in open water. Whale distribution around the islands was non-uniform, with a hot-spot on the North-Eastern side of the islands. There were higher counts in water > 350m deep with medium krill density and high productivity. We believe that the abundance of krill and hence higher whale numbers at the Balleny Islands may be due to an Island Mass Effect increase in productivity in an otherwise relatively featureless expanse of ocean. Further long-term studies are needed to quantify annual trends and identify whether the same individuals return to the Balleny Islands each summer. This is essential information if we are to adequately manage the marine protected area that encompasses the Balleny Islands and ensure that they remain a pristine feeding area for whales.

6. Acknowledgements

We thank scientific staff for processing the trawl samples. We thank the officers and crew of the RV Tangaroa during voyage TAN1502 for their invaluable help. This research is a contribution to Australian Antarctic Division science programme Project 4104 and project 4102 and National Institute of Water and Atmospheric Research (NIWA) project number VES15303. TAN1502 was funded under by Antarctica New Zealand, New Zealand Ministry for Business, Innovation and Employment, Australian Antarctic Division, and NIWA. MJC is funded by Australian Research Council grant FS11020005. L-M.H. is funded by a Macquarie University Research Excellence Scholarship.

7. References

AMARAL, A. R., LOO, J., JARIS, H., OLAVARRIA, C., THIELE, D., ENSOR, P., AGUAYO, A. & ROSENBAUM, H. C. 2016. Population genetic structure among feeding aggregations of humpback whales in the Southern Ocean. *Marine Biology,* 163**,** 132.

ATKINSON, A., WHITEHOUSE, M., PRIDDLE, J., CRIPPS, G., WARD, P. & BRANDON, M. 2001. South Georgia, Antarctica: a productive, cold water, pelagic ecosystem. *Marine Ecology Progress Series,* 216**,** 279-308.

BARNES, R. S. K. & HUGHES, R. N. 2009. The Planktonic System of Surface Waters. *An Introduction to Marine Ecology.* 3rd Edition ed. Great Britian: Blackwell Publishing Ltd.

BAUMGARTNER, M. F., FRATANTONI, D. M., HURST, T. P., BROWN, M. W., COLE, T. V. N., VAN PARIJS, S. M. & JOHNSON, M. 2013. Real-time reporting of baleen whale passive acoustic detections from ocean gliders. *The Journal of the Acoustical Society of America,* 134**,** 1814-1823.

BENSON, S. R., CROLL, D. A., MARINOVIC, B. B., CHAVEZ, F. P. & HARVEY, J. T. 2002. Changes in the cetacean assemblage of a coastal upwelling ecosystem during El Niño 1997–98 and La Niña 1999. *Progress in Oceanography,* 54**,** 279-291.

BLAIN, S., TRÉGUER, P., BELVISO, S., BUCCIARELLI, E., DENIS, M., DESABRE, S., FIALA, M., MARTIN JÉZÉQUEL, V., LE FÈVRE, J., MAYZAUD, P., MARTY, J.-C. & RAZOULS, S. 2001. A biogeochemical study of the island mass effect in the context of the iron hypothesis: Kerguelen Islands, Southern Ocean. *Deep Sea Research Part I: Oceanographic Research Papers,* 48**,** 163-187.

BODEN, B. P. 1988. Observations of the island mass effect in the Prince Edward archipelago. *Polar Biology,* 9**,** 61-68.

BOST, C. A., COTTÉ, C., BAILLEUL, F., CHEREL, Y., CHARRASSIN, J. B., GUINET, C., AINLEY, D. G. & WEIMERSKIRCH, H. 2009. The importance of oceanographic fronts to marine birds and mammals of the southern oceans. *Journal of Marine Systems,* 78**,** 363-376.

BOYD, P. W., WATSON, A. J., LAW, C. S., ABRAHAM, E. R., TRULL, T., MURDOCH, R., BAKKER, D. C., BOWIE, A. R., BUESSELER, K. & CHANG, H. 2000. A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization. *Nature,* 407**,** 695-702.

CARMEL, Y. & BEN-HAIM, Y. 2005. Info-Gap Robust-Satisficing Model of Foraging Behavior: Do Foragers Optimize or Satisfice? *The American Naturalist,* 166**,** 633-641.

CHARNOV, E. L. 1976. Optimal foraging, the marginal value theorem. *Theoretical population biology,* 9**,** 129-136.

CONSTANTINE, R., STEEL, D., ALLEN, J., ANDERSON, M., ANDREWS, O., BAKER, C. S., BEEMAN, P., BURNS, D., CHARRASSIN, J.-B., CHILDERHOUSE, S., DOUBLE, M., ENSOR, P., FRANKLIN, T., FRANKLIN, W., GALES, N., GARRIGUE, C., GIBBS, N., HARRISON, P., HAUSER, N., HUTSEL, A., JENNER, C., JENNER, M.-N., KAUFMAN, G., MACIE, A., MATTILA, D., OLAVARRÍA, C., OOSTERMAN, A., PATON, D., POOLE, M., ROBBINS, J., SCHMITT, N., STEVICK, P., TAGARINO, A., THOMPSON, K. & WARD, J. 2014. Remote Antarctic feeding ground important for east Australian humpback whales. *Marine Biology,* 161**,** 1087-1093.

COX, M. J., WATKINS, J. L., REID, K. & BRIERLEY, A. S. 2011. Spatial and temporal variability in the structure of aggregations of Antarctic krill (Euphausia superba) around South Georgia, 1997–1999. *ICES Journal of Marine Science: Journal du Conseil,* 68**,** 489-498.

DALLA ROSA, L., SECCHI, E. R., MAIA, Y. G., ZERBINI, A. N. & HEIDE-JØRGENSEN, M. P. 2008. Movements of satellite-monitored humpback whales on their feeding ground along the Antarctic Peninsula. *Polar Biology,* 31**,** 771-781.

DONIOL-VALCROZE, T., BERTEAUX, D., LAROUCHE, P. & SEARS, R. 2007. Influence of thermal fronts on habitat selection by four rorqual whale species in the Gulf of St. Lawrence. *Marine Ecology Progress Series,* 335**,** 207-216.

ECHOVIEW, H., AUSTRALIA 2015. Echoview Software, version 6.1.35.26153.

EISENMANN, P., FRY, B., HOLYOAKE, C., COUGHRAN, D., NICOL, S. & BENGTSON NASH, S. 2016. Isotopic Evidence of a Wide Spectrum of Feeding Strategies in Southern Hemisphere Humpback Whale Baleen Records. *PLOS ONE,* 11**,** e0156698.

ELLIOTT, J., PATTERSON, M. & GLEIBER, M. Detecting 'Island Mass Effect'through remote sensing. Proceedings of the 12th International Coral Reef Symposium, 2012 Cairns, Australia.

FELIX, F. & HAASE, B. 2005. Distribution of humpback whales along the coast of Ecuador and management implications. *Journal of Cetacean Research and Management,* 7**,** 21-31.

FRETWELL, P. T., STANILAND, I. J. & FORCADA, J. 2014. Whales from Space: Counting Southern Right Whales by Satellite. *PLOS ONE,* 9**,** e88655.

FRIEDLAENDER, A. S., HALPIN, P. N., QIAN, S. S., LAWSON, G. L., WIEBE, P. H., THIELE, D. & READ, A. J. 2006. Whale distribution in relation to prey abundance and oceanographic processes in shelf waters of the Western Antarctic Peninsula. *Marine Ecology Progress Series,* 317**,** 297-310.

FRIEDLAENDER, A. S., LAWSON, G. L. & HALPIN, P. N. 2009. Evidence of resource partitioning between humpback and minke whales around the western Antarctic Peninsula. *Marine Mammal Science,* 25**,** 402-415.

GALES, N., DOUBLE, M. C., ROBINSON, S., JENNER, C., JENNER, M., KING, E., GEDAMKE, J., PATON, D. & RAYMOND, B. 2009. Satellite tracking of southbound East Australian humpback whales (Megaptera novaeangliae): challenging the feast or famine model for migrating whales. *Int Whal Comm: SC61/SH17*.

GOLDBOGEN, J. A., CALAMBOKIDIS, J., CROLL, D. A., HARVEY, J. T., NEWTON, K. M., OLESON, E. M., SCHORR, G. & SHADWICK, R. E. 2008. Foraging behavior of humpback whales: kinematic and respiratory patterns suggest a high cost for a lunge. *Journal of Experimental Biology,* 211**,** 3712-3719.

GOVE, J. M., MCMANUS, M. A., NEUHEIMER, A. B., POLOVINA, J. J., DRAZEN, J. C., SMITH, C. R., MERRIFIELD, M. A., FRIEDLANDER, A. M., EHSES, J. S., YOUNG, C. W., DILLON, A. K. & WILLIAMS, G. J. 2016. Near-island biological hotspots in barren ocean basins. *Nature Communications,* 7**,** 10581.

GREGR, E. J. & TRITES, A. W. 2001. Predictions of critical habitat for five whale species in the waters of coastal British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences,* 58**,** 1265-1285.

GROENEVELD, J., JOHST, K., KAWAGUCHI, S., MEYER, B., TESCHKE, M. & GRIMM, V. 2015. How biological clocks and changing environmental conditions determine local population growth and species distribution in Antarctic krill (Euphausia superba): a conceptual model. *Ecological Modelling,* 303**,** 78-86.

GUIDINO, C., LLAPAPASCA, M. A., SILVA, S., ALCORTA, B. & PACHECO, A. S. 2014. Patterns of Spatial and Temporal Distribution of Humpback Whales at the Southern Limit of the Southeast Pacific Breeding Area. *PLOS ONE,* 9**,** e112627.

HARRISON, L.-M. K., COX, M. J., SKARET, G. & HARCOURT, R. 2015. The R package EchoviewR for automated processing of active acoustic data using Echoview. *Frontiers in Marine Science,* 2.

HERR, H., VIQUERAT, S., SIEGEL, V., KOCK, K.-H., DORSCHEL, B., HUNEKE, W. G. C., BRACHER, A., SCHRÖDER, M. & GUTT, J. 2016. Horizontal niche partitioning of humpback and fin whales around the West Antarctic Peninsula: evidence from a concurrent whale and krill survey. *Polar Biology,* 39**,** 799-818.

HORTON, T. W., HOLDAWAY, R. N., ZERBINI, A. N., HAUSER, N., GARRIGUE, C., ANDRIOLO, A. & CLAPHAM, P. J. 2011. Straight as an arrow: humpback whales swim constant course tracks during long-distance migration. *Biology Letters*.

KINZEY, D., OLSON, P. & GERRODETTE, T. 2000. Marine mammal data collection procedures on research ship line-transect surveys by the Southwest Fisheries Science Center. *NOAA, SWFSC Administrative Report LJ-00-08*.

LAIDRE, K. L., HEIDE-JØRGENSEN, M. P., HEAGERTY, P., COSSIO, A., BERGSTRÖM, B. & SIMON, M. 2010. Spatial associations between large baleen whales and their prey in West Greenland. *Marine Ecology Progress Series,* 402**,** 269-284.

LANCELOT, C., HANNON, E., BECQUEVORT, S., VETH, C. & DE BAAR, H. J. W. 2000. Modeling phytoplankton blooms and carbon export production in the Southern Ocean: dominant controls by light and iron in the Atlantic sector in Austral spring 1992. *Deep Sea Research Part I: Oceanographic Research Papers,* 47**,** 1621-1662.

LAUBSCHER, R. K., PERISSINOTTO, R. & MCQUAID, C. D. 1993. Phytoplankton production and biomass at frontal zones in the Atlantic sector of the Southern Ocean. *Polar Biology,* 13**,** 471-481.

LINCHANT, J., LISEIN, J., SEMEKI, J., LEJEUNE, P. & VERMEULEN, C. 2015. Are unmanned aircraft systems (UASs) the future of wildlife monitoring? A review of accomplishments and challenges. *Mammal Review,* 45**,** 239-252.

MARQUES, T. A., THOMAS, L., WARD, J., DIMARZIO, N. & TYACK, P. L. 2009. Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales. *The Journal of the Acoustical Society of America,* 125**,** 1982-1994.

MATTILA, D. K., CLAPHAM, P. J., VÁSQUEZ, O. & BOWMAN, R. S. 1994. Occurrence, population composition, and habitat use of humpback whales in Samana Bay, Dominican Republic. *Canadian Journal of Zoology,* 72**,** 1898-1907.

MILLER, D. L., BURT, M. L., REXSTAD, E. A. & THOMAS, L. 2013. Spatial models for distance sampling data: recent developments and future directions. *Methods in Ecology and Evolution,* 4**,** 1001-1010.

MILLER, D. L., REXSTAD, E. A., BURT, M. L., BRAVINGTON, M. V. & HEDLEY, S. L. 2016. dsm: Density Surface Modelling of Distance Sampling Data. R package version 2.2.12.

MORATO, T., HOYLE, S. D., ALLAIN, V. & NICOL, S. J. 2010. Seamounts are hotspots of pelagic biodiversity in the open ocean. *Proceedings of the National Academy of Sciences,* 107**,** 9707-9711.

MORI, M. & BUTTERWORTH, D. S. 2004. Consideration of multispecies interactions in the Antarctic: a preliminary model of the minke whale – blue whale – krill interaction. *African Journal of Marine Science,* 26**,** 245-259.

MURASE, H., MATSUOKA, K., ICHII, T. & NISHIWAKI, S. 2002. Relationship between the distribution of euphausiids and baleen whales in the Antarctic (35°E – 145°W). *Polar Biology,* 25**,** 135-145.

NICOL, S., BOWIE, A., JARMAN, S., LANNUZEL, D., MEINERS, K. M. & VAN DER MERWE, P. 2010. Southern Ocean iron fertilization by baleen whales and Antarctic krill. *Fish and Fisheries,* 11**,** 203-209.

NICOL, S., WORBY, A. & LEAPER, R. 2008. Changes in the Antarctic sea ice ecosystem: potential effects on krill and baleen whales. *Marine and Freshwater Research,* 59**,** 361-382.

OWEN, K., KAVANAGH, A. S., WARREN, J. D., NOAD, M. J., DONNELLY, D., GOLDIZEN, A. W. & DUNLOP, R. A. 2016. Potential energy gain by whales outside of the Antarctic: prey preferences and consumption rates of migrating humpback whales (Megaptera novaeangliae). *Polar Biology***,** 1-13.

OWEN, K., WARREN, J. D., NOAD, M. J., DONNELLY, D., GOLDIZEN, A. W. & DUNLOP, R. A. 2015. Effect of prey type on the fine-scale feeding behaviour of migrating east Australian humpback whales. *Marine Ecology Progress Series,* 541**,** 231-244.

PERISSINOTTO, R., LAUBSCHER, R. & MCQUAID, C. 1992. Marine productivity enhancement around Bouvet and the South Sandwich Islands (Southern Ocean). *Marine Ecology Progress Series,* 88**,** 41-41.

PLANQUETTE, H., STATHAM, P. J., FONES, G. R., CHARETTE, M. A., MOORE, C. M., SALTER, I., NÉDÉLEC, F. H., TAYLOR, S. L., FRENCH, M., BAKER, A. R., MAHOWALD, N. & JICKELLS, T. D. 2007. Dissolved iron in the vicinity of the Crozet Islands, Southern Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography,* 54**,** 1999-2019.

R DEVELOPMENT CORE TEAM 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

REID, K., BRIERLEY, A. S. & NEVITT, G. A. 2000. An initial examination of relationships between the distribution of whales and Antarctic krill Euphausia superba at South Georgia. *Journal of Cetacean Research and Management,* 2**,** 143-149.

RESSLER, P. H., DALPADADO, P., MACAULAY, G. J., HANDEGARD, N. & SKERN-MAURITZEN, M. 2015. Acoustic surveys of euphausiids and models of baleen whale distribution in the Barents Sea. *Marine Ecology Progress Series,* 527**,** 13-29.

ROSSI-SANTOS, M. R., BARACHO, C., CIPOLOTTI, S. & MARCOVALDI, E. 2007. Cetacean sightings near South Georgia islands, South Atlantic Ocean. *Polar Biology,* 31**,** 63.

RSTUDIO 2014. RStudio: Integrated development environment for R (version 0.98.932). Boston, MA.

SABA, G. K., FRASER, W. R., SABA, V. S., IANNUZZI, R. A., COLEMAN, K. E., DONEY, S. C., DUCKLOW, H. W., MARTINSON, D. G., MILES, T. N., PATTERSON-FRASER, D. L., STAMMERJOHN, S. E., STEINBERG, D. K. & SCHOFIELD, O. M. 2014. Winter and spring controls on the summer food web of the coastal West Antarctic Peninsula. *Nature Communications,* 5.

SANTORA, J. A., REISS, C. S., LOEB, V. J. & VEIT, R. R. 2010. Spatial association between hotspots of baleen whales and demographic patterns of Antarctic krill Euphausia superba suggests size-dependent predation. *Marine Ecology Progress Series,* 405**,** 255-269.

SCALES, K. L., MILLER, P. I., HAWKES, L. A., INGRAM, S. N., SIMS, D. W. & VOTIER, S. C. 2014. On the Front Line: frontal zones as priority at-sea conservation areas for mobile marine vertebrates. *Journal of Applied Ecology,* 51**,** 1575-1583.

SCHWARZ, J. N., RAYMOND, B., WILLIAMS, G. D., PASQUER, B., MARSLAND, S. J. & GORTON, R. J. 2010. Biophysical coupling in remotely-sensed wind stress, sea surface temperature, sea ice and chlorophyll concentrations in the South Indian Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57**,** 701-722.

SEDWICK, P. N. & DITULLIO, G. R. 1997. Regulation of algal blooms in Antarctic shelf waters by the release of iron from melting sea ice. *Geophysical Research Letters,* 24**,** 2515-2518.

SIEGEL, V. 2016. Biology and Ecology of Antarctic Krill. *Advances in polar ecology*.

SIMON, M., JOHNSON, M. & MADSEN, P. T. 2012. Keeping momentum with a mouthful of water: behavior and kinematics of humpback whale lunge feeding. *The Journal of Experimental Biology,* 215**,** 3786-3798.

SIMS, D. W., SOUTHALL, E. J., HUMPHRIES, N. E., HAYS, G. C., BRADSHAW, C. J. A., PITCHFORD, J. W., JAMES, A., AHMED, M. Z., BRIERLEY, A. S., HINDELL, M. A., MORRITT, D., MUSYL, M. K., RIGHTON, D., SHEPARD, E. L. C., WEARMOUTH, V. J., WILSON, R. P., WITT, M. J. & METCALFE, J. D. 2008. Scaling laws of marine predator search behaviour. *Nature,* 451**,** 1098-1102.

SMITH, J. N., GRANTHAM, H. S., GALES, N., DOUBLE, M. C., NOAD, M. J. & PATON, D. 2012. Identification of humpback whale breeding and calving habitat in the Great Barrier Reef. *Marine Ecology Progress Series,* 447**,** 259-272.

SMITH, W. O. & LANCELOT, C. 2004. Bottom-up versus top-down control in phytoplankton of the Southern Ocean. *Antarctic Science,* 16**,** 531-539.

SMITH, W. O. & NELSON, D. M. 1986. Importance of ice edge phytoplankton production in the Southern Ocean. *BioScience,* 36**,** 251-257.

SMULTEA, M. A. 1994. Segregation by humpback whale (Megaptera novaeangliae) cows with a calf in coastal habitat near the island of Hawaii. *Canadian Journal of Zoology,* 72**,** 805-811.

SPREEN, G., KALESCHKE, L. & HEYGSTER, G. 2008. Sea ice remote sensing using AMSR-E 89 GHz channels. *Journal of Geophysical Research,* 113.

THIELE, D., CHESTER, E. T. & GILL, P. C. 2000. Cetacean distribution off Eastern Antarctica (80–150°E) during the Austral summer of 1995/1996. *Deep Sea Research Part II: Topical Studies in Oceanography,* 47**,** 2543-2572.

THIELE, D., CHESTER, E. T., MOORE, S. E., ŠIROVIC, A., HILDEBRAND, J. A. & FRIEDLAENDER, A. S. 2004. Seasonal variability in whale encounters in the Western Antarctic Peninsula. *Deep Sea Research Part II: Topical Studies in Oceanography,* 51**,** 2311-2325.

THOMPSON, S. A., SYDEMAN, W. J., SANTORA, J. A., BLACK, B. A., SURYAN, R. M., CALAMBOKIDIS, J., PETERSON, W. T. & BOGRAD, S. J. 2012. Linking predators to seasonality of upwelling: Using food web indicators and path analysis to infer trophic connections. *Progress in Oceanography,* 101**,** 106-120.

TYNAN, C. T., AINLEY, D. G., BARTH, J. A., COWLES, T. J., PIERCE, S. D. & SPEAR, L. B. 2005. Cetacean distributions relative to ocean processes in the northern California Current System. *Deep Sea Research Part II: Topical Studies in Oceanography,* 52**,** 145-167.

VISSER, F., HARTMAN, K. L., PIERCE, G. J., VALAVANIS, V. D. & HUISMAN, J. 2011. Timing of migratory baleen whales at the Azores in relation to the North Atlantic spring bloom. *Marine Ecology Progress Series,* 440**,** 267-279.

WARE, C., FRIEDLAENDER, A. S. & NOWACEK, D. P. 2011. Shallow and deep lunge feeding of humpback whales in fjords of the West Antarctic Peninsula. *Marine Mammal Science,* 27**,** 587-605.

WARE, D. M. & THOMSON, R. E. 2005. Bottom-Up Ecosystem Trophic Dynamics Determine Fish Production in the Northeast Pacific. *Science,* 308**,** 1280-1284.

WOOD, S. N., BRAVINGTON, M. V. & HEDLEY, S. L. 2008. Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 70**,** 931-955.

ZERBINI, A. N., ANDRIOLO, A., HEIDE-JØRGENSEN, M. P., PIZZORNO, J. L., MAIA, Y. G., VANBLARICOM, G. R., DEMASTER, D. P., SIMÕES-LOPES, P. C., MOREIRA, S. & BETHLEM, C. 2006. Satellite-monitored movements of humpback whales Megaptera novaeangliae in the Southwest Atlantic Ocean. *Marine Ecology Progress Series,* 313**,** 295-304.

ZERBINI, A. N., FRIDAY, N. A., PALACIOS, D. M., WAITE, J. M., RESSLER, P. H., RONE, B. K., MOORE, S. E. & CLAPHAM, P. J. 2015. Baleen whale abundance and distribution in relation to environmental variables and prey density in the Eastern Bering Sea. *Deep Sea Research Part II: Topical Studies in Oceanography,* 134**,** 312-330.

Chapter 7: General Conclusion

Authors:

Lisa-Marie K. Harrison[1]

Affiliations:

[1]Marine Predator Research Group, Department of Biological Sciences,

Faculty of Science and Engineering, Macquarie University,

North Ryde, New South Wales, Australia

In this thesis I set out to provide new insights into the relationship between key ecosystem components through the development and application of advanced quantitative techniques. I successfully applied these techniques to data collected in the Southern Ocean to examine drivers of the distribution and abundance of key organisms in the Southern Oceans that span three trophic levels, starting with phytoplankton, then plankton grazers (krill) and finally krill predators (whales). The processes that drive spatial distribution at each trophic level were found, as predicted, to centre on food availability or in the case of phytoplankton, energy sources and environmental conditions that promote photosynthesis. Here I synthesise the findings of the thesis, discuss their implications for our understanding of the Southern Ocean ecosystem, and close with ideas for future research.

There is a wide array of statistical methods available for modelling ecological data, many of which are necessarily intricate to deal with the complex characteristics of the data. In Chapter 2 I conducted a literature review of the more common statistical modelling methods, detailing a comparison between the two main approaches to ecological models, the frequentist and Bayesian paradigms. This chapter provides background information about the modelling methodology used in this thesis and a rational for why these methods are necessary to account for complexities such as correlations and population level inference.

1. Processing large acoustic data sets

Counting animals at each level of the food web requires different in situ data collection methods. For example, phytoplankton data are collected using fluorometry, Niskin bottle sampling or tows; mid-level predators (primarily krill and fish) are counted from active acoustics or net surveys and high level predators (air-breathing marine mammals and seabirds) using visual surveys, passive acoustics or tracking. While these methods all involve a large time cost during collection, the size and complexity involved in processing active acoustics data sets requires a second large investment of time in the data processing stage – something often overlooked when planning ecosystem surveys. Raw active acoustics data sets can easily contain billions of data points, which must be correctly integrated, identified and then interpreted in the appropriate manner for each different purpose.

To more easily manage the data processing required in this thesis I developed an R package (R Development Core Team, 2014), EchoviewR, to automate the processing of active acoustic data sets using the commercially available software, Echoview (Echoview, 2015). EchoviewR increases reproducibility and cuts down on user error through the use of code scripts. As processing techniques advance, or new thresholds for the identification of noise or species are adopted, it allows us to automatically re-run the data processing by simply modifying a line in the original script. EchoviewR is a valuable tool for any scientist using active acoustics, whether it be for fish or krill detection, sea floor mapping or oil/gas plume detection. It has been made widely available on the software repository GitHub and in the publication Harrison et al (2015) in the journal *Frontiers in Marine Science*. I employed this package to efficiently process the acoustic data used to calculate Antarctic krill densities in a further two chapters in this thesis.

2. Phytoplankton distribution in a 3D environment

Phytoplankton form the base of the Southern Ocean food web and their abundance and distribution shapes ecosystem dynamics and supports the entire food web. In such large, dynamic and complex ecosystems it is difficult to quantify the drivers of distribution over a survey area while reducing confounding by extraneous effects. The spline mixed model I developed to assess drivers of phytoplankton distribution in Chapter 3 is widely applicable in both terrestrial and marine settings, regardless of survey design and could hence retrospectively be applied to the large amount of vertical profile data that already exists.

In Chapter 3 I found that phytoplankton density off North-East Antarctica correlated with salinity, temperature, depth in the water column and dissolved oxygen levels. Sea ice, distance from coastline and current strength and direction were not significantly associated with phytoplankton levels and were not included in the final model. This model is the first to assess drivers of phytoplankton distribution while accounting for 3D spatial autocorrelation, non-linear relationships using data from multiple irregularly spaced sampling stations. It was found to be unbiased using simulation. I demonstrated the importance of accounting for spatial complexities such as 3D autocorrelation, which can bias results if not included. Ignoring spatial autocorrelation led to significant over-fitting problems in the model along with high residual correlation.

An important practical application of this research is to predict phytoplankton abundance based on future expected climate scenarios. The model offers a fast and flexible method to predict both localised and survey wide trends. It can make predictions in a 3D environment and is applicable to other surveys regardless of survey design. Understanding how the base of the food chain will be affected by environmental changes

is the first step in forecasting the outcome on other Southern Ocean species and in identifying where we can act quickly and effectively to effect mitigation or adaptive management.

3. Antarctic krill: drifting or swimming?

Actively sourcing resources as opposed to passively drifting and feeding solely opportunistically allows animals to avoid unfavourable conditions and take advantage of patchy resources. Despite the importance for conservation, drifting versus swimming has rarely been tested (Putman et al., 2016). Antarctic krill are a key link in the food chain, transporting the energy in phytoplankton to the higher trophic levels (Siegel, 2016; pg 322). They are often treated as though they are passive drifters, completely at the mercy of the current systems surrounding Antarctica. While there have been observations in captivity of krill actively seeking out food (Hamner and Hamner, 2000, Kawaguchi et al., 2010), the extent to which they are able to combat currents and circulation to actively position themselves in the wild is difficult to assess and has hence remained a large but important knowledge gap over many decades. This thesis provided the first quantitative evidence of Antarctic krill aggregating around important resources. In Chapter 4 I showed quantitatively that over a large survey area (1.3 million km^2) krill aggregate in areas favourable to them, i.e. those with high food availability and high dissolved oxygen levels. These findings were only made possible through the application of a hurdle model to partition the presence/absence and conditional densities into separate models. Without this approach, I would not have been able to separately assess the drivers of presence and density and would not have discovered the otherwise 'masked' signal of krill aggregating around oxygen and phytoplankton.

The ability for krill to actively seek out areas of high dissolved oxygen and phytoplankton may help krill adapt to future predicted changes in phytoplankton community composition and loss of sea ice. However, active swimming will offer little protection against large scale ocean acidification, predicted to have catastrophic consequences for krill (Kawaguchi et al., 2011), and larval and juvenile krill are still reliant on current systems to move to areas that offer the right conditions to progress to the next stage in the life cycle. The finding of their ability to aggregate in proximity to resources is important for the management of the krill fishery and forecasting future changes because management approaches based only on passive krill flux will fail to adequately capture the localised active swimming behaviour of krill swarms.

4. Prey hotspots around islands: the Balleny Islands

The Southern Ocean is an important feeding ground for migratory predators, such as whales and seabirds, who every season must efficiently locate areas of high prey availability. Islands, frontal systems and bathymetric features are known to cause high productivity and an enhancement of the food chain through upwelling and stabilisation of the water column (Bost et al., 2009, Laubscher et al., 1993, Gove et al., 2016). This creates a highly dynamic, patchily distributed resource for krill predators. In the Southern Ocean, higher trophic predators must search widely for the highly dynamic krill swarms, and high krill abundance around islands could benefit whales by providing a relatively predictable food source in the ever-changing seascape.

In Chapter 5, I found that Antarctic krill aggregate in locations of high food availability, indicating that they might be abundant around these zones of high productivity. In Chapter 6 I tested this theory at a remote Southern Ocean archipelago, the Balleny

Islands. I found that krill swarms around the islands were more numerous (with a three times higher encounter rate than in the open ocean), denser and more compact than those in nearby open water, and this is likely to attract the unusually high number of humpback whales seen around the islands. In close proximity to the islands, whales aggregate in areas of high productivity, medium krill density and with a bottom depth >350m. These findings demonstrate that islands in a large expanse of open water can offer a profitable feeding opportunity. The high levels of krill and whales found at the Balleny Islands, along with the results of my analysis of whale habitat use at the feeding area, have implications for conservation and management of this resource. In 2016, the islands became part of the world's largest marine protected area and we have an obligation to protect and conserve significant biological hotspots, such as this important krill hotspot at the Balleny Islands.

In Chapters 3 and 5 I used different methods to incorporate the spatial components (3D autocorrelation structure versus smoothing surface). The reason I chose different methods was because the BROKE-West data used in Chapter 3 was on a much larger and sparser scale than the Balleny Islands data in Chapter 5. The BROKE-West data are also 3-dimensional because they include a depth through the water column component, and current software packages are unable to incorporate 3-dimensional volumetric spline surfaces. Hence using a smoothing surface for BROKE-West would not have worked well. The Balleny Islands dependent variable data (whale sightings) was also collected through visual surveys so a distance sampling component was required in the analysis which was not needed for the phytoplankton analysis. DSMs can incorporate both the spatial smooth surface and distance sampling which is why that method was selected rather than trying to modify a spline mixed model to be suitable.

5. Future Directions

5.1. Predicting the future

A key reason for modelling ecosystems is to use the available information and the developed model to forecast what might happen if environmental conditions or food availability changes. The phytoplankton, krill and whale models developed in this thesis can all be used to predict both survey wide and localised distributions under different environmental input parameters. While it was a key motivation for this work, the predictive capacity of the models in this thesis was not explored beyond cross-validation given the time constraints of a three-year thesis and remain an important future direction for studies building on this research. There are many ways in which these models can be extended and built upon, and I will now outline briefly some areas that I would have liked to explore further.

5.2. Survey locations

Most ecological and oceanographic surveys in the Southern Ocean take place only in a single sector due to a combination of time and logistical constraints. The chapters in this thesis used data from single sectors because that was the available data. However, the Antarctic continent is experiencing quite contrasting environmental changes in different sectors, with some warming and some cooling (Constable et al., 2014). Accordingly the inference drawn from one sector is unlikely to be directly transferrable to other areas. An important future direction would therefore be to incorporate data from these other regions, such as the West Antarctic Peninsula, to assess which processes drive species distributions at regional or continental levels. Regional effects that incorporate data from other areas around the continent could be assessed using the mixed effects models I developed in this thesis. For example, the Antarctic Circumpolar Expedition (ACE, 2016)

will provide valuable circumpolar information about the key systems studied in this thesis and applying these models to that dataset would be a highly rewarding extension of this work.

5.3. Currents and nutrients

The models in this thesis primarily focus on data collected from underway ship data or Conductivity Temperature Depth (CTD) vertical profiles, both of which measure environmental variables such as salinity, temperature, dissolved oxygen concentration and chlorophyll-a levels. Given the importance of nutrients for productivity, it would be informative to add nutrient data to these models. Nutrient data was omitted from our study because the available data were collected on a different scale to the other oceanographic variables, or were in some cases defective or missing (e.g. phosphate data during the BROKE-West cruise on which the phytoplankton and krill studies were based) and were in some cases not collected at all (e.g. iron data are commonly not collected because cross-contamination of the samples is difficult to avoid). As iron is believed to be a limiting factor in phytoplankton productivity and growth (Boyd et al., 2000) its inclusion in the phytoplankton drivers model might prove highly illuminating. Iron might well account for some of the variation seen between stations, especially those near the ice edge that we predict could be experiencing phytoplankton blooms simulated by iron release from melting ice.

The Southern Ocean current systems and large scale circulation are important in shaping both Southern Ocean ecosystems, and have an important link to world climate through thermohaline circulation. Underway current data, collected with an Acoustic Doppler Current Profiler (ADCP), was not included in the models in this thesis because during the BROKE-West survey much of the data were missing due to equipment malfunction

and during the Balleny Islands survey ADCP data were not collected. The use of satellite collected current strength and direction was considered but the data were not high resolution enough to be useful for our models. The inclusion of current data would be useful in all studies in this thesis because it likely underpins areas of high productivity due to upwelling. This is especially likely around the Balleny Islands, where an Island Mass Effect may occur.

5.4. Trophic levels

Each chapter of this thesis investigated drivers of animal distribution, looking at the food or energy sources and environmental conditions for each study species. However, I did not consider all trophic levels in a single model because in the first instance I needed to establish what were the drivers within each level. A whole ecosystem model could be accomplished with a hierarchical model to explain complex species interactions and make predictions across all trophic levels under different scenarios. This may require a Bayesian approach to take into account prior information, ensure convergence of the highly complex model and handle uncertainty at each level. This model would be complicated to set up and computationally expensive but is the logical next step and could be based on the knowledge gained from each chapter of this thesis.

5.5. Seasonal and annual differences

The Southern Ocean is highly seasonal, with summer conditions vastly different to those in winter. Many of the species have adapted to this either through migration or through overwintering strategies (Meyer, 2012, Dawbin, 1966). Most Antarctic shipboard surveys only occur in summer because of the harsh conditions and almost complete lack of daylight in winter. This means that much of what we know, along with the research in

this thesis, is about the summertime cycles and ecology of the animals we study. Contrasting these to winter or spring, during the ice melt, would provide a more rounded picture and allow us to understand how changes will affect species survival through winter.

The Southern Ocean experiences climactic forcing over large time scales, such as the El Niño Southern Oscillation (ENSO) and Southern Annular Mode (SAM) (Kwok and Comiso, 2002, Lovenduski and Gruber, 2005). This thesis used only surveys from a single summer in each analysis, which means that the fluctuations seen with ENSO and SAM cannot be captured in these models. Time series data collected in the same location could extend our models by including these systems. It would also allow us to understand how short term annual differences in sea-ice and productivity affect the ecosystem. Importantly, time series data could help forecast the future.

5.6. Life stages and community composition

Chlorophyll-a and fluorescence capture only the magnitude of phytoplankton sampled but offer no information about community composition. Climate change is expected to affect not only the magnitude of phytoplankton abundance but also community composition (Tortell et al., 2008, Moline et al., 2004). This will affect krill, who are known to preferentially feed on diatoms and avoid phytoplankton that are too large or small to consume efficiently. The chapters in this thesis do not take this into account because data on community composition was only available for Niskin bottle water samples, which were collected much more irregularly than fluorescence data. A survey designed to include phytoplankton community composition at a high resolution

concurrently with krill and environmental data could extend the work in this thesis to understand and predict the flow-on effects of changes in community composition.

Due to their life cycle, krill are vulnerable to different processes at different ages. Larval krill are entirely reliant on currents and circulation to carry them to a location where they can hatch, ascend and survive as a juvenile. In Chapter 5, I have shown that adult krill are not passive drifters but are able to seek out beneficial areas, such as those with high prey availability and dissolved oxygen. We are not yet able to separate backscattering values from adult and juvenile krill from active acoustic data alone, relying on trawls to know the length-frequency histograms in the survey area. A model such as the hurdle model presented in Chapter 5 could be used to assess processes affecting juvenile krill, which are heavily reliant on sea ice to forage and may demonstrate different behavioural characteristics to adult krill.

5.7. Advances in technology

Technological advances have already allowed us insights into the Southern Ocean that were not possible mere decades ago. For example, advances in active acoustics have made it routine to count animals *in situ* that would previously have been trawled such as fish and krill. Single- or split- beam acoustics, giving a 2-dimensional picture below the ship, have been the most common form of this technology, but 3-dimensional multibeam acoustics are becoming more viable as data storage becomes larger. This technology will require automated processing methods to reduce operator time and ensure that data processing is manageable and efficient, which could build on the scripting interface I provided in EchoviewR.

Marine data are already inherently large and are becoming more complex as technology allows us to collect more information. Mixed modelling techniques, as used in this thesis,

will become more common in ecology as we become able to repeatedly sample multiple individuals and collect data at more sites/locations. Without the use of mixed modelling or other sophisticated techniques, pseudo-replication will make it difficult to draw correct and meaningful conclusions.

Data collection methods are advancing quickly and through improvements in technology such as drones, high resolution satellite imagery and tracking devices on wildlife, we can collect data in places and at times we have previously been unable to survey (Palumbi et al., 2009, Chelton et al., 2004, Hussey et al., 2015). For example, it is now possible to detect whales from satellite imagery (Fretwell et al., 2014), and although currently expensive, this could provide time-series data for assessing whale presence around island feeding areas such as the Balleny Islands. Further advances in technology will allow us to detect patterns in the marine environment which we cannot detect now and help guide our research and management of areas that have recently become far more accessible.

6. Final remarks

The Southern Ocean is already far from pristine, and continuing to change rapidly. It is important that we understand how environmental conditions affect animal distribution and abundance to predict how they will be affected as the oceans change. In this thesis I have identified environmental and biological influences on key taxa in the Southern Ocean food web: phytoplankton, krill and marine mammals. Using sophisticated modelling techniques I have developed a flexible and widely applicable model for predicting phytoplankton distribution in a 3D environment, showed that Antarctic krill aggregate around phytoplankton and dissolved oxygen, and demonstrated the benefits that Antarctic islands offer predators through heightened krill availability. This knowledge, along with the predictive capacity of the models used, fills a large knowledge

gap in drivers of animal distribution in the Southern Ocean using real survey data. There is still much left to discover to give us the power to fully understand this remote but highly relevant environment, and allow us to recognise and potentially mitigate the dramatic changes that are now occurring.

7. References

ACE 2016. Antarctic Circumnavigation Expedition (ACE). https://www.unige.ch/forel/files/5214/6104/9200/ACE_Project_Description.pdf. Swiss Polar Institute.

BOST, C. A., COTTÉ, C., BAILLEUL, F., CHEREL, Y., CHARRASSIN, J. B., GUINET, C., AINLEY, D. G. & WEIMERSKIRCH, H. 2009. The importance of oceanographic fronts to marine birds and mammals of the southern oceans. *Journal of Marine Systems,* 78**,** 363-376.

BOYD, P. W., WATSON, A. J., LAW, C. S., ABRAHAM, E. R., TRULL, T., MURDOCH, R., BAKKER, D. C., BOWIE, A. R., BUESSELER, K. & CHANG, H. 2000. A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization. *Nature,* 407**,** 695-702.

CHELTON, D. B., SCHLAX, M. G., FREILICH, M. H. & MILLIFF, R. F. 2004. Satellite measurements reveal persistent small-scale features in ocean winds. *science,* 303**,** 978-983.

CONSTABLE, A. J., MELBOURNE-THOMAS, J., CORNEY, S. P., ARRIGO, K. R., BARBRAUD, C., BARNES, D. K. A., BINDOFF, N. L., BOYD, P. W., BRANDT, A., COSTA, D. P., DAVIDSON, A. T., DUCKLOW, H. W., EMMERSON, L., FUKUCHI, M., GUTT, J., HINDELL, M. A., HOFMANN, E. E., HOSIE, G. W., IIDA, T., JACOB, S., JOHNSTON, N. M., KAWAGUCHI, S., KOKUBUN, N., KOUBBI, P., LEA, M.-A., MAKHADO, A., MASSOM, R. A., MEINERS, K., MEREDITH, M. P., MURPHY, E. J., NICOL, S., REID, K., RICHERSON, K., RIDDLE, M. J., RINTOUL, S. R., SMITH, W. O., SOUTHWELL, C., STARK, J. S., SUMNER, M., SWADLING, K. M., TAKAHASHI, K. T., TRATHAN, P. N., WELSFORD, D. C., WEIMERSKIRCH, H., WESTWOOD, K. J., WIENECKE, B. C., WOLF-GLADROW, D., WRIGHT, S. W., XAVIER, J. C. & ZIEGLER, P. 2014. Climate change and Southern Ocean ecosystems I: how changes in physical habitats directly affect marine biota. *Global Change Biology,* 20**,** 3004-3025.

DAWBIN, W. H. 1966. The seasonal migratory cycle of humpback whales. *In:* NORRIS, K. S. (ed.) *Whales, dolphins and porpoises. University of California Press, Berkeley.* USA: University of California Press.

ECHOVIEW, H., AUSTRALIA 2015. Echoview Software, version 6.1.35.26153.

FRETWELL, P. T., STANILAND, I. J. & FORCADA, J. 2014. Whales from Space: Counting Southern Right Whales by Satellite. *PLOS ONE,* 9**,** e88655.

GOVE, J. M., MCMANUS, M. A., NEUHEIMER, A. B., POLOVINA, J. J., DRAZEN, J. C., SMITH, C. R., MERRIFIELD, M. A., FRIEDLANDER, A. M., EHSES, J. S., YOUNG, C. W., DILLON, A. K. & WILLIAMS, G. J. 2016. Near-island biological hotspots in barren ocean basins. *Nature Communications,* 7**,** 10581.

HAMNER, W. M. & HAMNER, P. P. 2000. Behavior of Antarctic krill (Euphausia superba): schooling, foraging, and antipredatory behavior. *Canadian Journal of Fisheries and Aquatic Sciences,* 57**,** 192-202.

HARRISON, L.-M. K., COX, M. J., SKARET, G. & HARCOURT, R. 2015. The R package EchoviewR for automated processing of active acoustic data using Echoview. *Frontiers in Marine Science,* 2.

HUSSEY, N. E., KESSEL, S. T., AARESTRUP, K., COOKE, S. J., COWLEY, P. D., FISK, A. T., HARCOURT, R. G., HOLLAND, K. N., IVERSON, S. J. & KOCIK, J. F. 2015. Aquatic animal telemetry: a panoramic window into the underwater world. *Science,* 348**,** 1255642.

KAWAGUCHI, S., KING, R., MEIJERS, R., OSBORN, J. E., SWADLING, K. M., RITZ, D. A. & NICOL, S. 2010. An experimental aquarium for observing the schooling behaviour of Antarctic krill (Euphausia superba). *Deep Sea Research Part II: Topical Studies in Oceanography,* 57**,** 683-692.

KAWAGUCHI, S., KURIHARA, H., KING, R., HALE, L., BERLI, T., ROBINSON, J. P., ISHIDA, A., WAKITA, M., VIRTUE, P., NICOL, S. & ISHIMATSU, A. 2011. Will krill fare well under Southern Ocean acidification? *Biology Letters,* 7**,** 288-291.

KWOK, R. & COMISO, J. C. 2002. Southern Ocean Climate and Sea Ice Anomalies Associated with the Southern Oscillation. *Journal of Climate,* 15**,** 487-501.

LAUBSCHER, R. K., PERISSINOTTO, R. & MCQUAID, C. D. 1993. Phytoplankton production and biomass at frontal zones in the Atlantic sector of the Southern Ocean. *Polar Biology,* 13**,** 471-481.

LOVENDUSKI, N. S. & GRUBER, N. 2005. Impact of the Southern Annular Mode on Southern Ocean circulation and biology. *Geophysical Research Letters,* 32.

MEYER, B. 2012. The overwintering of Antarctic krill, Euphausia superba, from an ecophysiological perspective. *Polar Biology,* 35**,** 15-37.

MOLINE, M. A., CLAUSTRE, H., FRAZER, T. K., SCHOFIELD, O. & VERNET, M. 2004. Alteration of the food web along the Antarctic Peninsula in response to a regional warming trend. *Global Change Biology,* 10**,** 1973-1980.

PALUMBI, S. R., SANDIFER, P. A., ALLAN, J. D., BECK, M. W., FAUTIN, D. G., FOGARTY, M. J., HALPERN, B. S., INCZE, L. S., LEONG, J.-A. & NORSE, E. 2009. Managing for ocean biodiversity to sustain marine ecosystem services. *Frontiers in Ecology and the Environment,* 7**,** 204-211.

PUTMAN, N. F., LUMPKIN, R., SACCO, A. E. & MANSFIELD, K. L. Passive drift or active swimming in marine organisms? Proc. R. Soc. B, 2016. The Royal Society.

R DEVELOPMENT CORE TEAM 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

SIEGEL, V. 2016. Biology and Ecology of Antarctic Krill. *Advances in polar ecology*.

TORTELL, P. D., PAYNE, C. D., LI, Y., TRIMBORN, S., ROST, B., SMITH, W. O., RIESSELMAN, C., DUNBAR, R. B., SEDWICK, P. & DITULLIO, G. R. 2008. CO2 sensitivity of Southern Ocean phytoplankton. *Geophysical Research Letters,* 35.

Appendix A

THE R PACKAGE *ECHOVIEWR* FOR AUTOMATED PROCESSING OF ACTIVE ACOUSTIC DATA USING ECHOVIEW

Original PDF of the Published Journal Article:

Journal: *Frontiers in Marine Science*

Published Online: 25th of February 2015

Citation: Harrison L-MK, Cox MJ, Skaret G and Harcourt R (2015) The R package *EchoviewR* for automated processing of active acoustic data using Echoview. *Front. Mar. Sci*. **2**:15. doi: 10.3389/fmars.2015.00015

Authors:

Lisa-Marie K. Harrison[1], Martin J. Cox[1, 2], Georg Skaret[3], Robert Harcourt[1]

Affiliations:

[1]Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, New South Wales, Australia

[2]Australian Antarctic Division, Department of the Environment, Australian Government, Kingston, Tasmania, Australia

[3]Institute of Marine Research, Bergen, Norway

frontiers in
MARINE SCIENCE

METHODS ARTICLE
published: 25 February 2015
doi: 10.3389/fmars.2015.00015

The R package *EchoviewR* for automated processing of active acoustic data using Echoview

Lisa-Marie K. Harrison[1]*, Martin J. Cox[1,2], Georg Skaret[3] and Robert Harcourt[1]

[1] Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, NSW, Australia
[2] Australian Antarctic Division, Department of the Environment, Australian Government, Kingston, TAS, Australia
[3] Institute of Marine Research, Bergen, Norway

Edited by:
Xabier Irigoien, King Abdullah University of Science and Technology, Saudi Arabia

Reviewed by:
Guillermo Boyra, AZTI, Spain
Anders Rastad, King Abdullah University of Science and Technology, Saudi Arabia

***Correspondence:**
Lisa-Marie K. Harrison, Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, Balaclava Road, North Ryde, NSW 2109, Australia
e-mail: lisamarie.k.harrison@gmail.com

Acoustic data is time consuming to process due to the large data size and the requirement to often undertake some data processing steps manually. Manual processing may introduce subjective, irreproducible decisions into the data processing work flow, reducing consistency in processing between surveys. We introduce the R package *EchoviewR* as an interface between R and Echoview, a commercially available acoustic processing software package. *EchoviewR* allows for automation of Echoview using scripting which can drastically reduce the manual work required when processing acoustic surveys. This package plays an important role in reducing subjectivity in acoustic data processing by allowing exactly the same process to be applied automatically to multiple surveys and documenting where subjective decisions have been made. Using data from a survey of Antarctic krill, we provide two examples of using *EchoviewR*: krill biomass estimation and swarm detection.

Keywords: active acoustic, Antarctic krill, data processing, echosounder, Echoview, R package

INTRODUCTION

Active acoustics is a tool widely used for seabed mapping, seabed type classification, underwater tracking and resource monitoring. A suite of active acoustic instruments are available to carry out imaging (e.g., scanning sonars) and more quantitative tasks (e.g., multibeam and scientific echosounders). Echosounders have evolved from being instruments used primarily for mapping and navigation, to precision instruments capable of resolving organisms a few millimeters in length and providing quantitative estimates of, for example, biomass.

This advance has seen widespread use of echosounders to detect organisms in the upper water column of both freshwater and marine environments for commercial fisheries and scientific purposes. In the marine environment, echosounders are routinely used to provide data informing commercial fishery stock assessments (Gerlotto et al., 1999) and to investigate ecological relationships such as predator-prey interactions (Benoit-Bird et al., 2013). Oceanographic applications include seabed habitat mapping (Brown et al., 2004) and environmental monitoring, e.g., oil seep and methane bubble monitoring after the Deepwater Horizon oil spill (Weber et al., 2014). Echosounders are commonly used in conjunction with image/video (McGonigle et al., 2009) and sediment sampling (Van Walree et al., 2005) to verify seabed type, or trawls to verify a biological species' presence, size and target strength (McGonigle et al., 2009).

Echosounder transducers are most commonly embedded in a ship's hull or drop keel, although other platforms such as landers (Johansen et al., 2009), gliders (Guihen et al., 2014), and autonomous underwater vehicles (Brierley et al., 2002) have been used. Regardless of platform, datasets from active acoustics are invariably extremely large and time consuming to process.

In active acoustic surveys, a conventional split-beam echosounder collecting data to a range of 500 m and pinging once per second typically collects around 8 GB of data per day (Note: this depends on settings such as range resolution and pulse repetition rate). This may be compounded by the need to use multiple echosounder frequencies, sometimes more than six, operating simultaneously, further inflating the size of the raw data sets. Moreover, the routine use of broadband systems like the Simrad EK80 on board scientific and commercial vessels is not far away. The amount of data from such systems vastly exceeds those from conventional sounders, and will again push storage and processing capacity. With advances in data storage capacity, data storage is no longer a significant constraint and enhanced computational power has enabled the development of powerful acoustic data processing software.

There are several software packages suitable for the processing of echosounder data e.g., Echoview (Myriax, Hobart; www.echoview.com), LSSS (MAREC, Christian Michelsen Research, Norway, http://www.cmr.no/index.cfm?id=421565) and Sonar5-Pro (University of Oslo, Norway, http://folk.uio.no/hbalk/sonar4_5/). However, processing acoustic data remains time consuming and frequently requires subjective, often undocumented, decisions to be made by the user, such as removal of noise or bad data and allocation of backscatter to targets. Subjective decisions can potentially bias outputs from processed active acoustic data, for example biomass estimates.

Here we present the R package *EchoviewR* as a tool to: (1) reduce the processing time requiring a human operator, (2) document processing steps thereby generating reproducible methodology, and (3) provide a framework within which additional functionality can be built by members of the acoustics community, so reducing the number of subjective decisions. The *EchoviewR* package is an interface between the widely used and freely available R program (http://www.R-project.org/) and Echoview (Myriax, Hobart; www.echoview.com). The methods used are generic and can be transferred to other acoustic processing software with scripting options, but the package as such is incompatible with other acoustic software.

EchoviewR uses Component Object Model (COM) scripting to run Echoview using R. This removes a large portion of the manual processing time and enables entire acoustic surveys to be mostly processed automatically. It also increases consistency in processing because the same methods and thresholds can be applied in exactly the same way to multiple data sets. Hence *EchoviewR* provides a reproducible and transparent automated method for processing acoustic data using Echoview. Some examples of its use include filtering of data, automated biomass estimation and detection of krill swarms.

Using two examples, we illustrate *EchoviewR* functionality. Both examples are based on data collected during surveys of Antarctic krill (*Euphausia superba*; herein krill) using a Simrad EK60 echosounder (Horten, Norway) with downward facing hull-mounted transducers. The first example estimates regional krill biomass, and the second example detects krill swarms.

EchoviewR is intended to speed up processing of already clean acoustic data and is not currently capable of removing false bottom effects, time varied gain or noise spikes although the package can access Echoview virtual variables to do some of these tasks, e.g., "Background noise removal algorithm" virtual variable (De Robertis and Higginbottom, 2007). The package is intended only as a method of automating processing using Echoview and is not a standalone method for processing acoustic data.

METHODS

IMPLEMENTATION AND DEPENDENCIES

EchoviewR was created using R 3.1 (R Development Core Team, 2014; available from http://cran.r-project.org/) with R-Studio 0.98.932 (Rstudio, 2014; available from http://www.rstudio.com/), and Echoview 6.1 (Myriax, 2015; available from http://www.echoview.com/). Both R and Echoview are required to use the package. COM objective handling is achieved using the *RDCOMClient* package. Additional *EchoviewR* functionality uses the *sp, lubridate, geosphere, maptools,* and *rgeos* R libraries (Pebesma and Bivand, 2005; Grolemund and Wickham, 2011; Hijmans, 2014; Bivand and Lewin-Koh, 2014; Bivand and Rundel, 2014). To run Echoview via COM the following modules are required: base, bathymetric, analysis export, and scripting. Worked example one also requires the virtual echogram module and worked example two requires the virtual echogram and schools detection modules.

The *EchoviewR* package is available open source on the GitHub repository (https://github.com/lisamarieharrison/EchoviewR) and can be downloaded and installed as an R package using the "install from.zip file" option in R, or via devtools:install_github().

EXPECTED DATA INPUT FOR THE PACKAGE AND WORKED EXAMPLES

EchoviewR can work with any data type accommodated in Echoview that is accessible via COM. The worked examples provided here have been built using data collected using a Simrad EK60 echosounder (www.simrad.com/ek60). In itself, *EchoviewR* does not create Echoview templates or calibration files, but can use both of these via COM.

FUNCTIONS OF THE PACKAGE

There are 46 functions available in *EchoviewR*, which are described in **Table 1**. A working example for each of these functions is given in the package documentation in the Supplementary Material. Not all Echoview functions are currently available in the package; however any functionality in Echoview that has COM accessibility could be added by the user.

EXAMPLES

Here we present two examples using *EchoviewR:* (1) krill biomass estimation, and (2) krill swarm detection and classification. The purpose of these examples is to demonstrate that these analyses can be run automatically using *EchoviewR* and to show how Echoview output can be seamlessly linked to analyses carried out using R. Both examples assume that the reader is familiar with Echoview and are not intended to be a tutorial on Echoview. It is also assumed that the reader is familiar with R and programming concepts such as for loops.

The data are a subset of the EK60 split-beam data collected during the Krill Acoustics and Oceanography Survey (KAOS) carried out from RV *Aurora Australis*. The KAOS survey was undertaken in January–March 2003 off North Eastern Antarctica. Data from 38, 120, and 200 kHz were written to RAW files. For clarity in the worked examples, we have used the 38 and 120 kHz data because these frequencies are the most useful for detecting and identifying the example species, Antarctic krill.

To demonstrate that biomass estimation and swarm detection can be automatically run on multiple transects where the data are too large to practically read in to Echoview at once, as is the case for most acoustic surveys, segments of six KAOS transects are provided and each 10–20 km transect segment is processed separately (**Figure 1**).

Both these examples have been tested using R Studio and 0.98.932 and Echoview 6.1.32.26088. The data to run these examples are available at the Australian Antarctic Division Data Centre [doi: 10.4225/15/54CF081FB955F]. An example of the data flow for the template used in this example is available as **Figure S1** in the Supplementary Material.

Before running each example some pre-processing is demonstrated to get the data in to a convenient format for analyzing each transect in a separate .EV file. In this pre-processing phase, the six transects are imported separately into Echoview and the following tasks are performed:

1. Create a new .EV file for the transect using the Echoview template file;

Appendix A

Table 1 | Functions available in *EchoviewR*.

Function	Description
EVOpenFile	Opens an existing .EV file
EVSaveFile	Saves an existing .EV file
EVSaveFileAs	Saves an existing .EV file to a new file name
EVCloseFile	Closes an open .EV file
EVNewFile	Creates a new .EV file
EVCreateFileset	Creates a new fileset
EVFindFilesetByName	Finds a fileset by name
EVAddRawData	Adds .RAW files to a fileset
EVCreateNew	Creates a new .EV file from a template
EVminThresholdSet	Sets the minimum dB threshold for an acoustic variable
EVSchoolsDetSet	Sets schools detection parameters
EVAcoVarNameFinder	Finds an acoustic variable by name
EVRegionClassFinder	Finds a region class by name
EVSchoolsDetect	Runs schools detection on an acoustic variable
EVIntegrationByRegionExport	Exports integration by region for an acoustic object
msDateConversion	Converts an Echoview date to readable format
EVAddCalibrationFile	Adds a calibration file to an .EV file
EVFilesInFileset	Finds the names of all .RAW files in the fileset
EVClearRawData	Clears all .RAW files from a fileset
EVFindFilesetTime	Finds the start and end date and time of a fileset
EVNewRegionClass	Creates a new region class
EVImportRegionDef	Imports a regions definition file
EVExportRegionSv	Exports Sv data for a region
EVAdjustRegionBitmap	Adjusts the settings of a region bitmap object
EVFindLineByName	Finds an Echoview line by name
EVChangeVariableGrid	Changes the horizontal and vertical grid for an acoustic variable
EVExportIntegrationByCells	Exports integration by cells for an acoustic variable
EVAddNewAcousticVar	Adds a new acoustic variable
EVShiftRegionDepth	Changes the depth of a region
EVShiftRegionTime	Changes the time of a region
EVGetCalibrationFileName	Finds the calibration file name
EVNewLineRelativeRegion	Creates a new line relative region
EVNewFixedLineDepth	Creates a new fixed depth line
EVDeleteLine	Deletes a line object
EVRenameLine	Renames a line object
EVExportRegionDef	Exports region definitions for a single region
EVFindRegionByName	Finds a region object by name
EVFindRegionClass	Finds a region class by name
EVExportRegionDefByClass	Exports region definitions for an entire region class
EVIntegrationByRegionByCellsExport	Exports integration by region by cells for an acoustic variable
lawnSurvey	Generate coordinates for a rectangular lawn survey design
zigzagSurvey	Generate coordinates for a zig-zag survey design
centreZigZagOnPosition	Centers a zig-zag survey on a given position
centreLawnOnPosition	Centers a lawn survey on a given position
exportMIF	Write a map information file for import into Echoview
EVImportLine	Imports an Echoview Line object

(Continued)

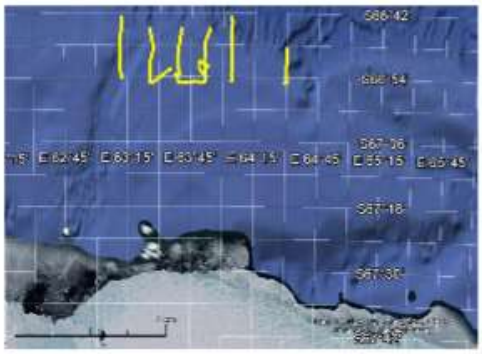


FIGURE 1 | Map showing location of 6 the example transects in yellow. Map created using Google Earth 7.1.2.2041.

2. Import the EK60 .RAW data files for that transect;
3. Add an Echoview.ecs calibration file;
4. Import .evr region definitions files to remove off effort data;
5. Import a seabed exclusion line (lineKAOS .evl);
6. Close and save the file and repeat for remaining transects.

These steps and the code to run them are demonstrated in the "Read data using the R package EchoviewR to control Echoview via COM" pdf vignette that is available with the Supplementary Material. Pre-processing must take place before examples 1 and 2 are run.

EXAMPLE 1—KRILL BIOMASS ESTIMATION

Automated biomass estimation of krill is demonstrated by processing the six transects separately in Echoview and exporting the data into R for density and biomass calculation. For each transect, the following steps are taken in Echoview:

1. Open the transect's .EV file.
2. Set the grid for 38 and 120 kHz noise removed values to 50 ping * 5 m depth.
3. Export integration by cells for 38 and 120 kHz noise removed values.

This produces two .csv files for each transect, one containing 38 kHz and one containing 120 kHz integrated data (i.e., a mean volume backscattering strength value for each cell. Then, the following steps are taken in R.

1. Import the 38 and 120 kHz files for the transect.
2. Remove no data values (set -999 and 999 dB as NA) and depths <0.
3. Calculate the krill difference window of 120–38 kHz for each integration cell using the following formula:

$$\Delta Sv_{ij} = Sv_{120_{ij}} - Sv_{38_{ij}}$$

where $Sv_{120_{ij}}$ = mean 120 kHz backscattering strength for cell at interval j at depth i and $Sv_{38_{ij}}$ = mean 38 kHz backscattering strength for cell at interval j at depth i.

4. Apply the dB difference technique (e.g., Watkins and Brierley, 2002) by setting $Sv_{120_{ij}}$ values outside the survey-specific dB difference range of 1.04 >=ΔSv_{ij} < = 14.75 dB to NA as these windows are unlikely to contain krill.
5. Convert the backscattering strength, $Sv_{120_{ij}}$ for each cell to linear scale, $sv_{120_{ij}}$ (Echoview uses a log scale by default):

$$sv_{ij} = 10^{\frac{Sv_{ij}}{10}}$$

6. Calculate mean volume backscattering strength (MVBS) across all depths for each 50 ping integration interval using the following formula:

$$MVBS_j = 10\log_{10}\frac{1}{n_j}\sum_{i=0}^{n_j} sv_{120_{ij}}$$

where j = integration interval, n = maximum depth within integration interval j and $sv_{120_{ij}}$ = backscattering strength at 120 kHz for interval j at depth i.

7. Calculate estimates of krill density, $\hat{p}_j$, for each integration interval:

$$\hat{p}_j = n_j^* 10^{\left\{\frac{MVBS_j - TS}{10}\right\}}$$

where n_j = maximum depth of integration interval j, $MVBS_j$ = mean volume backscattering strength for interval j as calculated above and TS = target strength for 1 kg of krill at 120 kHz.

8. Calculate the overall transect density, $\hat{p}_k$ for transect k:

$$\hat{p}_k = \frac{1}{s_k}\sum_{j=1}^{s_k}\hat{p}_j$$

where j = integration interval, k = transect and s_k = number of integration intervals within transect k.

9. The full survey density is then estimated using the Jolly and Hampton (1990) method, which uses the weighted density of each transect by length to calculate total survey density. Note that the formula has been modified to remove stratum as no strata were used in the KAOS example survey design:

$$\hat{p} = w_k\hat{p}_k$$

where k = transect, $w_k = \frac{L_k}{L}$, L_k = length of transect k in km, L = length of all survey transects in km and $\hat{p}_k$ = estimated density for transect k.

10. The full survey biomass estimate, $\hat{b}$, is then calculated by multiplying the weighted survey density by survey area:

$$\hat{b} = \hat{p}A$$

where $\hat{p}$ = estimated survey biomass and A = survey area in km^2.

Both the Echoview and R components above are run within loops to allow each transect to be run separately. This is done to demonstrate how looping over transects or days of a large survey is possible, rather than manually loading and processing each set of files. The *EchoviewR* and R code for the above analysis is shown in the "Biomass estimation using the R package EchoviewR to control Echoview via COM" pdf vignette that is available with the Supplementary Material. **Table 2** shows the estimated density, length and biomass for the sample transects and survey area.

Example 1 has demonstrated the use of *EchoviewR* to automatically process and extract data by transect from Echoview. Krill density and biomass are then calculated in R using the extracted .csv files.

EXAMPLE 2—SWARM DETECTION AND CLASSIFICATION

Automated swarm detection and classification of krill aggregations is demonstrated here using *EchoviewR*. The code for this example is available in the "Schools detection using the R package EchoviewR to control Echoview via COM" pdf vignette file available with the Supplementary Material. Each transect is processed separately to demonstrate how a full survey can be processed automatically using loops. Schools detection is run in Echoview and then detected aggregations are classified and clustered in R. The following steps are undertaken in Echoview using *EchoviewR*:

Table 2 | Estimated transect krill areal density and survey biomass for the six example transects.

Transect number	Mean estimated density, gm^{-2}	Transect length, km	Biomass, tones
1	3.26	13	42
2	20.66	22	454
3	43.74	15	656
4	22.57	22	497
5	6.66	18.5	123
6	4.99	21.5	107
Full survey area	16.79	112	43, 497

1. Open the transect's .EV file.
2. Run schools detection on the variable *120 7x7 convolution*, assigning all detected schools to the region class "aggregations."
3. Export 120 and 38 kHz data for regions of class "aggregations" to a .csv file using the *EVIntegrationByRegionExport* function. This exports a single mean Sv for each aggregation.

In this example, all detected aggregations are exported. However, it is also possible to export only aggregations classified as krill using the *120-38 aggregation dB difference filter* variable included in the template. The filter sets the *Krill aggregations* data to NULL if the *120-38 aggregation dB difference* value for that cell is outside the [1.04, 14.75] dB difference window for the KAOS survey.

The exported aggregations can now be classified and clustered in R. Each transect is run separately using a loop:

1. Import the 120 and 38 kHz export by regions files.
2. Remove null values (-999).
3. Calculate the 120–38 kHz difference window and subset data to only include difference values between [1.04, 14.75].
4. If no aggregations were classified as krill, exit here and move to next transect.
5. If krill aggregations are found, run cluster analysis using the *ClusterSim* library using selected metrics.
6. Print a summary table of the number of aggregations assigned to each identified cluster. **Table 3** shows the number of krill swarms identified and the number of clusters detected for each transect.

This example has demonstrated how school detection, data export and cluster analysis can be run automatically for an entire acoustic survey.

DISCUSSION AND FUTURE DIRECTIONS

EchoviewR is a free interface between R and Echoview that provides automated acoustic data processing. It drastically decreases manual processing time and reduces subjectivity by providing an easy way to implement exactly the same method across surveys. This package enables reproducible methodology, which is a vital part of the scientific method. We have given examples of automated krill biomass estimation and school detection using *EchoviewR* that demonstrate the use of the package on a subset of the KAOS survey. This method can easily be extended to run a full survey by transect, day or any other subset required.

Table 3 | Number of unique krill aggregation clusters identified for each transect.

Transect number	Number of krill swarms	Number of clusters
1	0	0
2	37	9
3	105	6
4	64	3
5	8	3
6	14	6

There are a number of limitations to the package. Currently it is only available for use for single and split beam echo sounder data. *EchoviewR* is also unable to handle removal of noise and false bottom effects, which must be completed prior to using the package. Not all functions in Echoview are currently available using *EchoviewR*, however any COM functionality in Echoview can be implemented in R. The COM hierarchy help page is a useful starting point for those wishing to add extra functions.

EchoviewR is accessible as free software from the *EchoviewR* GitHub repository (https://github.com/lisamarieharrison/EchoviewR) and is readily available for community development. An important next step is the implementation of false bottom and noise removal using *EchoviewR*, and it is our hope that the acoustic community will take the tools that we are providing and extend the package to include the functionality that they require. We also underline that the methods described here are generic, and hope the work can inspire the implementation of scripting interface in other acoustic processing software.

ACKNOWLEDGMENTS

We would like to thank Echoview for their support of this project. This research is a contribution to Australian Antarctic Division science programme Project 4104 and project 4102. MC is funded by Australian Research Council grant FS11020005. LH is funded by a Macquarie University Research Excellence Scholarship.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmars.2015.00015/abstract

Figure S1 | Screenshot of the Echoview data flow used for the examples. Template made using Echoview 6.0.

REFERENCES

Benoit-Bird, K. J., Battaile, B. C., Heppell, S. A., Hoover, B., Irons, D., Jones, N., et al. (2013). Prey patch patterns predict habitat use by top marine predators with diverse foraging strategies. *PLoS ONE* 8:e53348. doi: 10.1371/journal.pone.0053348

Bivand, R., and Lewin-Koh, N. (2014). *maptools: Tools for Reading and Handling Spatial Objects. R Package Version 0.8-29.* Available online at: http://CRAN.R-project.org/package=maptools

Bivand, R., and Rundel, C. (2014). *rgeos: Interface to Geometry Engine - Open Source (GEOS). R package version 0.3-4.* Available online at: http://CRAN.R-project.org/package=rgeos

Brierley, A. S., Fernandes, P. G., Brandon, M. A., Armstrong, F., Millard, N. W., McPhail, S. D., et al. (2002). Antarctic krill under sea ice: elevated abundance in a narrow band just south of ice edge. *Science* 295, 1890–1892. doi: 10.1126/science.1068574

Brown, C. J., Hewer, A. J., Meadows, W. J., Limpenny, D. S., Cooper, K. M., and Rees, H. L. (2004). Mapping seabed biotopes at hastings shingle bank, eastern English Channel. Part 1. Assessment using sidescan sonar. *J. Mar. Biol. Ass. U.K.* 84, 481–488. doi: 10.1017/S0025315404009492Xh

De Robertis, A., and Higginbottom, I. (2007). A post-processing technique to estimate the signal-to-noise ratio and remove echosounder background noise. *ICES. J. Mar. Sci.* 64, 1282–1291. doi: 10.1093/icesjms/fsm112

Gerlotto, F., Soria, M., and Fréon, P. (1999). From two dimensions to three: the use of multibeam sonar for a new approach in fisheries acoustics. *Can. J. Fish. Aquat. Sci.* 56, 6–12. doi: 10.1139/cjfas-56-1-6

Grolemund, G., and Wickham, H. (2011). Dates and times made easy with lubridate. *J. Stat. Softw.* 40, 1–25.

Guihen, D., Fielding, S., Murphy, E. J., Heywood, K. J., and Griffiths, G. (2014). An assessment of the use of ocean gliders to undertake acoustic

measurements of zooplankton: the distribution and density of Antarctic krill (Euphausia superba) in the Weddell Sea. *Limnol. Oceanogr.* 12, 373–389. doi: 10.4319/lom.2014.12.373

Hijmans, R. J. (2014). *geosphere: Spherical Trigonometry. R Package Version 1.3-8.* Available online at: http://CRAN.R-project.org/package=geosphere

Johansen, G. O., Godø, O. R., Skogen, M. D., and Tørkelsen, T. (2009). Using acoustic technology to improve the modelling of the transportation and distribution of juvenile gadoids in the Barents Sea. *ICES. J. Mar. Sci.* 66, 1048–1054. doi: 10.1093/icesjms/fsp081

Jolly, G., and Hampton, I. (1990). A stratified random transect design for acoustic surveys of fish stocks. *Can. J. Fish. Aquat. Sci.* 47, 1282–1291. doi: 10.1139/f90-147

McGonigle, C., Brown, C., Quinn, R., and Grabowski, J. (2009). Evaluation of image-based multibeam sonar backscatter classification for benthic habitat discrimination and mapping at Stanton Banks, UK. *Estuar. Coast. Shelf Sci.* 81, 423–437. doi: 10.1016/j.ecss.2008.11.017

Myriax. (2015). *Echoview*. Hobart, TAS. Available online at: http://www.echoview.com/

Pebesma, E. J., and Bivand, R. S. (2005). *Classes and methods for spatial data in R. R News 5 (2)*, Available online at: http://cran.r-project.org/doc/Rnews/

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rstudio. (2014). *RStudio: Integrated Development Environment for R (version 0.98.932)*. Boston, MA. Available online at: http://www.rstudio.com/

Van Walree, P. A., Tęgowski, J., Laban, C., and Simons, D. G. (2005). Acoustic seafloor discrimination with echo shape parameters: a comparison with the ground truth. *Cont. Shelf Res.* 25, 2273–2293. doi: 10.1016/j.csr.2005.09.002

Watkins, J. L., and Brierley, A. S. (2002). Verification of the acoustic techniques used to identify Antarctic krill. *ICES. J. Mar. Sci.* 59, 1326–1336. doi: 10.1006/jmsc.2002.1309

Weber, T. C., Jerram, K., and Mayer, L. (2014). "Acoustic sensing of gas seeps in the deep ocean with split-beam echosounders," in *Proceedings of Meetings on Acoustics* (Edinburgh, EB), 17.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 19 January 2015; accepted: 11 February 2015; published online: 25 February 2015.

Citation: Harrison L-MK, Cox MJ, Skaret G and Harcourt R (2015) The R package EchoviewR for automated processing of active acoustic data using Echoview. Front. Mar. Sci. 2:15. doi: 10.3389/fmars.2015.00015

This article was submitted to Marine Ecosystem Ecology, a section of the journal Frontiers in Marine Science.

Appendix B

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

This appendix provides the supplementary materials that were published in *Frontiers in Marine Science* along with Chapter 4 of this thesis. The supplementary materials contain three pdf tutorials detailing the use of the *EchoviewR* software along with an image of the Echoview workflow used to process the data. The 3 vignettes included are:

1. Biomass estimation using the R package *EchoviewR* to control Echoview via COM
2. Read data using the R package *EchoviewR* to control Echoview via COM
3. Schools detection using the R package *EchoviewR* to control Echoview via COM

Authors:

Lisa-Marie K. Harrison[1], Martin J. Cox[1, 2], Georg Skaret[3], Robert Harcourt[1]

Affiliations:

[1]Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, New South Wales, Australia

[2]Australian Antarctic Division, Department of the Environment, Australian Government, Kingston, Tasmania, Australia

[3]Institute of Marine Research, Bergen, Norway

Biomass estimation using the R package EchoviewR to control Echoview via COM

Lisa-Marie Harrison

February 3, 2015

1 Introduction

This vignette provides an example of biomass calculation using the R package R-acoustic as an interface to Echoview (Myriax, Hobart). The data used are from the 2003 Aurora Australis survey Krill Acoustics and Oceanography Survey (KAOS). For this example, segments of 6 transects are used, each approximately 10km in length. Each transect is processed separately to show how this package can be used to process multiple transects or surveys automatically. It is assumed that the user is familiar with Echoview and concepts such as regions, variables, filesets and integration intervals.

To run this example, you will need the following files:

1. **Echosounder data:** Simrad EK60 RAW data collected during the KAOS voyage.
2. **An Echoview template file:** An Echoview (.EV) file containing the virtual variables used during data processing and must be copied to the Echoview templates folder.
3. **Echoview calibration file:** The Echoview .ECS file containing the EK60 calibration parameters.
4. **Start and end file regions:** Start and end times of each transect stored as an Echoview region file.

All these files are available from ... and should be downloaded to the data directory specified in the dd object. `If you want to reproduce this vignette, you should assign the location of the KAOS example data to dd.`.

This example assumes that the Read Data vignette has been run first, to prepare the data for processing. This example contains two components:

1. **Processing and exporting in Echoview** For each transect, the data are processed in Echoview and are exported as integration intervals.
2. **Biomass estimation in R** The exported data are used in R to calculate the density for each integration interval and estimate biomass.

2 Processing and exporting each transect in Echoview

Firstly, a connection to Echoview is established and the working directory is set.

```
> dd='C:/Users/Lisa/Desktop/KAOS'
> library(acoustic)
> EVAppObj <- COMCreate('EchoviewCom.EvApplication')
```

A list of all raw files and the transect that they correspond to is imported:

```
> tF <- read.csv(paste(dd,'vignette_file_list_subset.csv',sep='/'))
> head(tF)

  transectNumber                          filename
1              1 L0055-D20030115-T171028-EK60.raw
2              1 L0055-D20030115-T182914-EK60.raw
3              2 L0056-D20030118-T000714-EK60.raw
4              3 L0056-D20030118-T131501-EK60.raw
5              3 L0056-D20030118-T143357-EK60.raw
6              3 L0056-D20030118-T155252-EK60.raw

> tF$filename <- paste(dd,'raw',tF$filename,sep='/')
> uniqueTransect=unique(tF$transectNumber)
>
```

In the Read Data vignette, the calibration file, raw data and off effort region definitions were to each transect's .EV file so it is not necessary to add them again here.

Next, the grid distance for the two acoustic variables of interest (38kHz and 120kHz) needs to be set. The grid size corresponds to the integration interval size when exporting the data. For this example a grid of 5m depth * 50 pings width is used. The 38kHz and 120kHz integration by cells are exported as .csv files. Each line in the exported .csv files represents a single integration interval.

Each transect is processed separately using a loop:

```
> for (i in 1:length(uniqueTransect)) {
+
+   #open the correct .ev file for the tranect
+   #note: the correct raw data files were pre-loaded when the .ev file was created
+   EVFile <- EVOpenFile(EVAppObj, paste(dd,'/kaos-transect-',i,'.ev',sep=''))$EVFile
+
+   #set the integration interval size using grid settings for 38kHz and 120kHz acousti
+   #for 38kHz
+   varObj <- EVAcoVarNameFinder(EVFile, acoVarName = "38 seabed and surface excluded")
+   EVChangeVariableGrid(EVFile = EVFile, acousticVar = varObj, verticalType = 4, horiz
+
+   #for 120kHz
+   varObj <- EVAcoVarNameFinder(EVFile, acoVarName = "120 seabed and surface excluded'
```

```
+   EVChangeVariableGrid(EVFile = EVFile, acousticVar = varObj, verticalType = 4, horiz
+
+   #export the acoustic variables using integration by cells for the specified grid si
+   message(paste(Sys.time(), "Exporting data for transect ", i, "..."))
+   EVExportIntegrationByCells(EVFile = EVFile, variableName = '38 seabed and surface e
+   EVExportIntegrationByCells(EVFile = EVFile, variableName = '120 seabed and surface
+
+   message(paste(Sys.time(), "Finished exporting data for transect ", i))
+
+   #close the .ev file
+   EVCloseFile(EVFile = EVFile)
+
+ }
>
```

3 Processing the exported data in R

Now that the 38kHz and 120kHz integration intervals have been exported from Echoview, the calculation of survey density and biomass can be calculated in R. In this example, mean density is calculated for each vertical bin through the water column, rather than for each integration interval.

A dB difference window (120kHz - 38kHz) is used to determine which intervals contain krill. Non krill intervals are removed and krill density is calculated. The densities for each interval are then appended to the same .csv file for all transects.

```
> #create an empty .csv file to export the density data to
> file.create("C:/Users/Lisa/Desktop/KAOS/combined_density_intervals.csv")

[1] TRUE

> #for each transect, calculate krill density
> for (i in 1:length(uniqueTransect)) {
+
+   acoustic_38  <- read.csv(file = paste(dd, "/exported integrations/kaos_38_integrati
+   acoustic_120 <- read.csv(file = paste(dd, "/exported integrations/kaos_120_integrat
+
+   #sort by interval
+   acoustic_38 <- acoustic_38[order(acoustic_38$Interval), ]
+   acoustic_120 <- acoustic_120[order(acoustic_120$Interval), ]
+   acoustic_120 <- acoustic_120[c(1:nrow(acoustic_38)), ]
+
+   #remove NULL layers (layer < 0)
+   acoustic_38 <- acoustic_38[acoustic_38$Layer > 0, ]
+   acoustic_120 <- acoustic_120[acoustic_120$Layer > 0, ]
+
```

```
+   #remove NuLL values (-900 or 900) from Sv values
+   sv_38 <- acoustic_38$Sv_mean
+   sv_120 <- acoustic_120$Sv_mean
+   sv_38[sv_38 > 500 | sv_38 < -500] <- NA
+   sv_120[sv_120 > 500 | sv_120 < -500] <- NA
+
+   #calculate 120kHz - 38kHz for each 5m*50ping window
+   sv_diff <- sv_120 - sv_38
+
+   #remove 120 - 38 kHz values outside of [1.02, 14.75]
+   #dB difference window is from Potts AAD report for KAOS data
+   sv_diff[sv_diff < 1.02 | sv_diff > 14.75] <- NA
+
+   #for windows that don't contain krill, remove 120kHz values
+   sv_120[is.na(sv_diff)] <- NA
+
+   #convert Sv values (for integration intervals that contain krill) back from log sca
+   sv <- 10^(sv_120/10)
+
+   #work out the mean volume backscattering strength (MVBS) for each vertical slice th
+   n.layers <- acoustic_120$Layer
+   mvbs  <- 10*log10(aggregate(matrix(sv, ncol = 1), by = list(rep(c(1:(length(sv)/max
+   mvbs[mvbs == -Inf] <- NA
+
+   #convert to density using target strength (units = kg/m2 per interval)
+   #formula is multiplied by 250m because this is the depth of slice through the water
+   p <- 250*10 ^((mvbs - -42.22)/10)*1000
+   p <- p[p < 700] #remove noise values
+
+   #calculate transect density
+   p_transect <- mean(na.omit(p))
+
+   #write the density per vertical slice to a single .csv file
+   write.table(p_transect, file = "C:/Users/Lisa/Desktop/KAOS/combined_density_interva
+
+   message(paste(Sys.time(), "Finished calculating krill density for transect ", i))
+ }
```

Now that the mean densities for all 6 example transects are in the same .csv file, the biomass for the survey area can be calculated. The survey coordinates and the `geosphere` package in R were used to calculate the survey area. Survey density is calculated using the Jolly and Hampton (1990) method, where transect densities are weighted by transect length to give the final survey density. This is then multiplied by survey area to give the overall biomass estimate.

```
> #Calculate average survey density
> transect.density <- read.csv("C:/Users/Lisa/Desktop/KAOS/combined_density_intervals.c
```

```
> #calculate survey area in km^2
> library(geosphere)
> coords <- read.csv(paste(dd, "/survey_coordinates.csv", sep =""), header = T)
> survey.area <- areaPolygon(coords)*10^-6
> #specify transect length in km
> transect.length <- c(13, 22, 15, 22, 18.5, 21.5)
> total.length <- sum(transect.length)
> length.weight <- transect.length/total.length
> #weight transect density by transect length to get survey density
> p.survey <- sum(transect.density*length.weight)
> #multiply by survey area to get survey biomass (units = tonnes)
> survey.biomass <- p.survey * survey.area * 10^6 / 1000000
```

The total estimated biomass for the 2591.21 square km sample area of the survey is 43497.07 tonnes. Note that this is only a simple example of how the package `acoustic` can be used to automatically estimate krill density. It is not intended as a reference for estimating biomass.

Read data using the R package EchoviewR to control Echoview via COM

Lisa-Marie Harrison and Martin Cox

February 3, 2015

This vignette provides an example of preparing and importing multiple transects of acoustic data into Echoview using the R package acoustic. It is necessary to run this file before running the Schools Detection or Biomass Estimation examples. It is assumed that the user is already familiar with Echoview.

To run this example, you will need the following files:

1. **Echosounder data:** Simrad EK60 RAW data collected during the KAOS voyage.
2. **An Echoview template file:** An Echoview (.EV) file containing the virtual variables used during data processing and must be copied to the Echoview templates folder.
3. **Echoview calibration file:** The Echoview .ECS file containing the EK60 calibration parameters.
4. **Start and end file regions:** Start and end times of each transect stored as an Echoview region file.

All these files are available from ... and should be downloaded to the data directory specified in the dd object. `If you want to reproduce this vignette, you should assign the location of the KAOS example data to dd.`

First, set the working directory:

```
> dd='C:/Users/Lisa/Desktop/KAOS'
```

1 Loading RAW data into Echoview

In this section we load the necessary packages into the R workspace, open a COM connection between R and Echoview, then populate the Echoview template file with RAW Simrad EK60 data files.

```
> library(acoustic)
```

The EV file `KAOStemplate.EV` **must be copied into the c:/Program Files (x86)/Myriax/Echoview/Echoview6/Templates/directory.** Open a COM connection between `R` and `Echoview`:

```
> EVAppObj <- COMCreate('EchoviewCom.EvApplication')
```

Now we get a vector of the RAW data files to add to the Echoview template:

```
> pathAndFn=list.files(paste(dd,'raw',sep='/'),full.names=TRUE)
```

We remove any '.evi' files from the `pathAndFn`:

```
> eviLoc=grep('.evi',pathAndFn)
> if(length(eviLoc)>0) pathAndFn=pathAndFn[-eviLoc]
```

Important: Any filenames that are passed via COM must contain the full directory path.

2 Populate an Echoview template

Now that we have a vector of RAW data file locations, we can populate the Echoview template file specified in `EVCreateNew(templateFn)`.

```
> EVFile <- EVCreateNew(EVAppObj=EVAppObj,
+   templateFn=paste(dd, "/KAOStemplate.EV", sep = ""),
+   EVFileName=paste(dd,'kaosAll.ev',sep='/'),
+   filesetName="038-120-200",
+   dataFiles=pathAndFn)$EVFile
```

We can also populate a template in a loop, so if there are multiple transects, each containing a large amount of RAW data, we can create an EV file for each transect. To help illustrate this, there is a .csv file, `vignette_file_list_subset.csv`, included in the example data that specifies which transect the RAW data files are assigned to. First of all, we load this file into the `R` workspace:

```
> tF <- read.csv(paste(dd,'vignette_file_list_subset.csv',sep='/'))
> head(tF)
```

```
  transectNumber                          filename
1              1 L0055-D20030115-T171028-EK60.raw
2              1 L0055-D20030115-T182914-EK60.raw
3              2 L0056-D20030118-T000714-EK60.raw
4              3 L0056-D20030118-T131501-EK60.raw
5              3 L0056-D20030118-T143357-EK60.raw
6              3 L0056-D20030118-T155252-EK60.raw
```

Next, we append the raw data directory path to the raw data filenames:

```
> tF$filename <- paste(dd,'raw',tF$filename,sep='/')
```

Now we will loop over the transect file and create one EV file per transect. A calibration file is added to each .EV file. Regions definitions files for off transect times are also imported, which will remove off effort time from the analysis.

```
> uniqueTransect=unique(tF$transectNumber)
> for(i in 1:length(uniqueTransect)) {
+
+   EVFile <- EVCreateNew(EVAppObj=EVAppObj,
+   templateFn=paste(dd, "KAOStemplate.EV", sep = "/"),
+   EVFileName=paste(dd,'/kaos-transect-',i,'.ev',sep=''),
+   filesetName="038-120-200",
+   dataFiles=as.character(tF$filename[tF$transectNumber==uniqueTransect[i]]),
+   CloseOnSave = FALSE)$EVFile
+
+     #add a calibration file
+   EVAddCalibrationFile(EVFile = EVFile, filesetName = "038-120-200", calibrationFile
+
+     #get a list of the regions definitions files to import and import individually
+   off_transect_files <- list.files(paste(dd, "/off transect regions", sep = ''),
+                                    full.names = T)
+   for (j in 1:length(off_transect_files)) {
+     EVImportRegionDef(EVFile, off_transect_files[j], paste("region_", j, sep = ""))
+   }
+
+   #add an EV line object and rename to 'seabed line'
+   evLine <- EVImportLine(EVFile, pathAndFn = 'C:/Users/Lisa/Desktop/KAOS/lineKAOS.evl
+   EVRenameLine(EVFile = EVFile, evLine = evLine, newName = "seabed line")
+
+   #save the open .EV file
+   EVSaveFile(EVFile = EVFile)
+
+   #close the current transect
+   EVCloseFile(EVFile = EVFile)
+
+   }
>
```

This method allows subsets of a full survey to be processed automatically which is useful for large data sets. Rather than running transects separately as shown above, the same method could also be used to run each survey day separately. The Biomass Estimation and Schools Detection vignettes follow on from this point and require this example to have been run first.

Schools detection using the R package EchoviewR to control Echoview via COM

Lisa-Marie Harrison and Martin J. Cox

February 3, 2015

1 Introduction

In this vignette we will use the example acoustic data collected during the KOAS voyage to demonstrate how to carry out schools detection using `acoustic` to control Echoview via COM. We assume the reader has knowledge of the schools detection algorithm implemented in Echoview. If not please visit `http://support.echoview.com/WebHelp/Echoview.htm/` to see the Echoview help file and also ... We also assume the reader is familar with Echoview concepts such as filesets, regions, files and virtual variables.

To run this example, you will need the following files:

1. **Echosounder data:** Simrad EK60 RAW data collected during the KAOS voyage.
2. **An Echoview template file:** An Echoview (.EV) file containing the virtual variables used during data processing and must be copied to the Echoview templates folder.
3. **Echoview calibration file:** The Echoview .ECS file containing the EK60 calibration parameters.
4. **Start and end file regions:** Start and end times of each transect stored as an Echoview region file.

All these files are available from ... and should be downloaded to the data directory specified in the `dd` object. `If you want to reproduce this vignette, you should assign the location of the KAOS example data to dd.`. This vignette requires you to have run the Read Data example first.

First, set the working directory. Echoview requires the full file path to be specified for every file name passed using COM.

```
> dd <- 'c:/Users/Lisa/Desktop/KAOS'
```

2 Loading RAW data into Echoview

In this section we load the necessary packages into the R workspace, open a COM connection between R and Echoview, then populate the Echoview template file with RAW Simrad EK60 data files.

```
> library(acoustic)
```

Open a COM connection between R and Echoview:

```
> EVAppObj <- COMCreate('EchoviewCom.EvApplication')
```

Get a list of the raw files required for each transect:

```
> tF=read.csv(paste(dd,'vignette_file_list_subset.csv', sep = '/'))
> head(tF)
```

```
  transectNumber                         filename
1              1 L0055-D20030115-T171028-EK60.raw
2              1 L0055-D20030115-T182914-EK60.raw
3              2 L0056-D20030118-T000714-EK60.raw
4              3 L0056-D20030118-T131501-EK60.raw
5              3 L0056-D20030118-T143357-EK60.raw
6              3 L0056-D20030118-T155252-EK60.raw
```

```
> tF$filename <- paste(dd,'raw',tF$filename,sep='/')
> uniqueTransect <- unique(tF$transectNumber)
```

3 Add seabed line

Automatic seabed detection performed poorly so the seabed line was manually edited and here we overwrite an existing editable line and replace it with lines saved in the lines directory of the example data.

4 Schools detection

We are now ready to start schools detection. First, we get the ful path and name of all the EV files:

```
> fnEVVec <- list.files(dd,full.name=TRUE,pattern=paste("(", "transect", ").*\\.ev$", s
```

Schools detection is run on the 120 kHz 7x7 convolution variable separately for each transect. Using a loop, a single transect is opened, processed and closed before moving on to the next transect. For each transect, the regions definitions and Sv values for each aggregation detected are exported.

```
>   for(i in 1:length(fnEVVec)){
+     EVLog <- NULL
+     message("Processing ",fnEVVec[i])
+     opens <- EVOpenFile(EVAppObj, fileName = fnEVVec[i])
+     EVFile <- opens$EVFile
+     EVLog <- c(EVLog,opens$msg)
+
+     schDet<-EVSchoolsDetect(EVFile = EVFile,
+                             acoVarName = '120 7x7 convolution',
+                             outputRegionClassName = 'aggregations',
+                             deleteExistingRegions = TRUE,
+                             distanceMode = "GPS distance",
+                             maximumHorizontalLink = 15,#m
+                             maximumVerticalLink = 5,#m
+                             minimumCandidateHeight = 1,#m
+                             minimumCandidateLength = 10,#m
+                             minimumSchoolHeight = 2,#m
+                             minimumSchoolLength = 15, #m
+                             dataThreshold = -70)
+     EVLog = c(EVLog, schDet$msg)
+     EVLog = c(EVLog, EVSaveFile(EVFile)$msg)
+
+     #if aggregations were detected, export the data
+     if (schDet$nbrOfDetectedschools > 0) {
+
+     #export region definitions all aggregations
+     regionClass <- EVRegionClassFinder(EVFile, "aggregations")$regionClass
+     EVExportRegionDefByClass(regionClass, paste(dd, "/exported aggregations/aggregati
+
+     #export Sv data for 38kHz and 120kHz all aggregations by region
+     EVIntegrationByRegionsExport(EVFile = EVFile, acoVarName = "120 seabed and surfac
+
+         EVIntegrationByRegionsExport(EVFile = EVFile, acoVarName = "38 seabed and sur
+
+     }
+
+     #close the file
+     EVCloseFile(EVFile = EVFile)
+   }
```

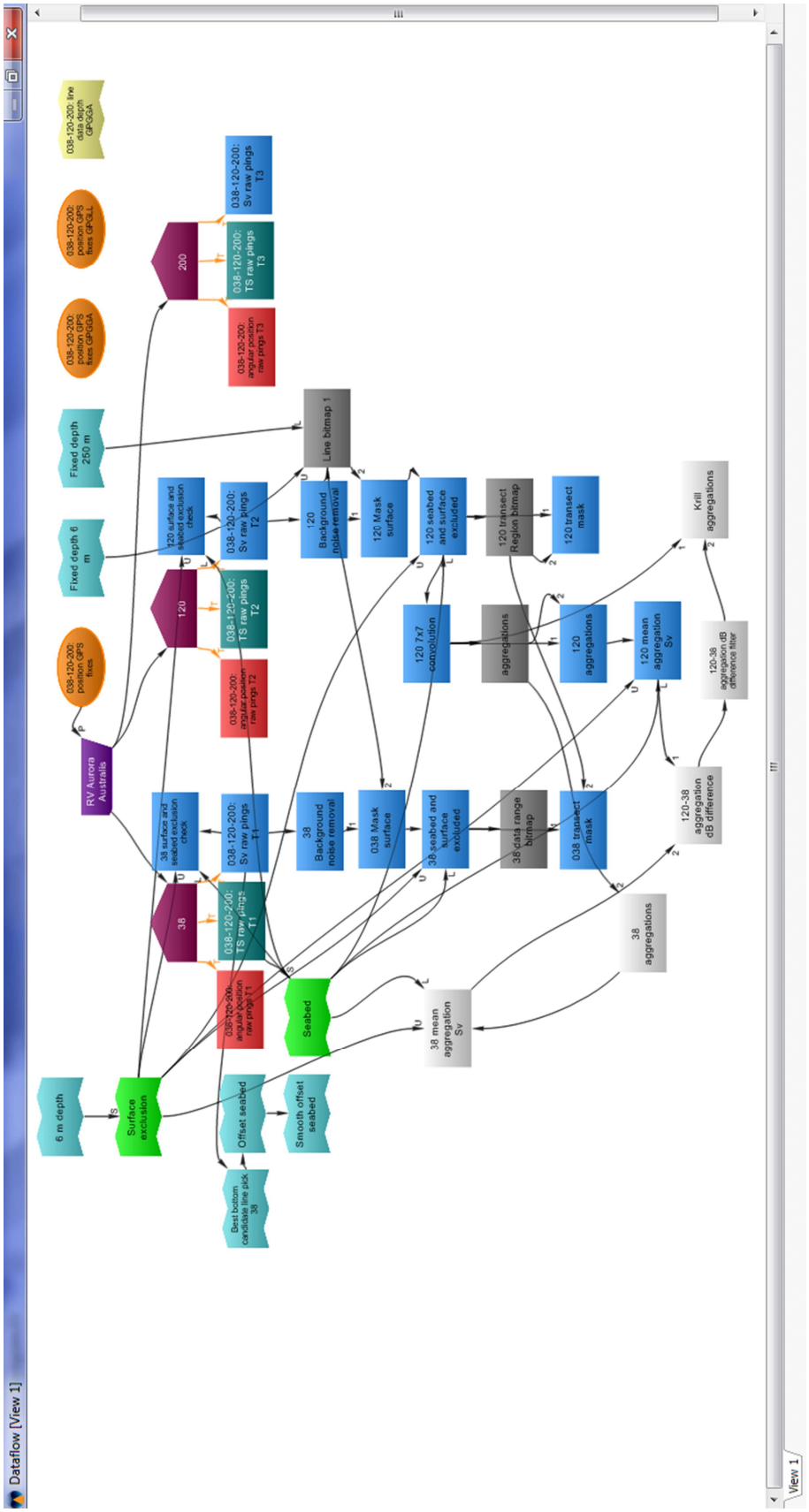
The exported aggregations are now analysed in R. Firstly, aggregations are subsetted to only include krill using the [1.04, 14.75]dB difference window. Krill aggregations are then clustered using the texxtClusterSim library. These steps are run separately for each transect. It is possible to aggregate all krill aggregations data into one .csv file and run the cluster analysis on the entire survey at once, however this is not demonstrated here.

```
> library(clusterSim)
>   for(i in 1:length(fnEVVec)){
+
+ #read in data files - the results of integration by regions for each aggregation
+
+ f038 <- read.csv(paste("C:/Users/Lisa/Desktop/KAOS/exported aggregations/38_aggregati
+ f120 <- read.csv(paste("C:/Users/Lisa/Desktop/KAOS/exported aggregations/120_aggregat
+ Sv038 <- cbind.data.frame(Region_name = f038$Region_name, Sv038 = f038$Sv_mean)
+
+ #merge Sv_mean from 38 to 120 kHz data
+ ag <- merge(f120, Sv038, by = 'Region_name')
+
+ #remove NA values
+ ag$Sv_mean[ag$Sv_mean == -999] <- NA
+ ag$ag$Sv038[ag$Sv038 == -999] <- NA
+
+ #calculate dB difference and isolate krill swarms
+ ag$dBdiff <- ag$Sv_mean - ag$Sv038
+ swarms <- subset(as.matrix(ag), ag$dBdiff > 1.02 & ag$dBdiff < 14.75)
+
+ #if no swarms are found, move to the next transect, otherwise run a cluster analysis
+ if (nrow(swarms) == 0) {
+   message("No swarms within difference window of 1.02 - 14.75dB found")
+ } else {
+
+ #select swarm metrics for PCA - there are loads more we could chose, these are just e
+ swarms <- swarms[, c("Sv_mean","Sv_max","Sv_min","Corrected_length","Height_mean",
+   "Depth_mean","Corrected_thickness","Corrected_perimeter",
+   "Corrected_area","Image_compactness",
+   "Corrected_mean_amplitude","Coefficient_of_variation",
+   "Horizontal_roughness_coefficient",
+   "Vertical_roughness_coefficient")]
+
+ swarms <- apply(swarms, 2, as.numeric)
+
+ #scale the data
+ scaleSwarm <- scale(swarms)
+
+ #determine number of clusters in scaled krill swarm data using the gap-stat. see:
+ #Tibshirani, R., Walther, G., Hastie, T. (2001), Estimating the number of clusters ir
+
+ #the code that follows is lifted from the
+ #index.Gap function in clusterSim:
+ # nc - number_of_clusters
+
+ min_nc <- 1
```

```
+ max_nc <- 10
+ if (nrow(swarms) < 10) {
+   max_nc <- nrow(swarms) - 2
+ }
+ min    <- 0
+ clopt  <- NULL
+
+ res <- array(0, c(max_nc - min_nc + 1, 2))
+ res[, 1] <- min_nc:max_nc
+ found <- FALSE
+
+ for (nc in min_nc:max_nc){
+   cl1 <- pam(scaleSwarm, nc, diss = FALSE)
+   cl2 <- pam(scaleSwarm, nc + 1, diss = FALSE)
+   clall <- cbind(cl1$clustering, cl2$clustering)
+   gap <- index.Gap(scaleSwarm, clall, B = 20, method = "pam")
+   res[nc - min_nc + 1, 2] <- diffu <- gap$diffu
+   if ((res[nc-min_nc + 1, 2] >= 0) && (!found)){
+     nc1 <- nc
+     min <- diffu
+     clopt <- cl1$cluster
+     found <- TRUE
+   }
+ }
+ if (found){
+   print(paste("Minimal number of clusters where diffu >= 0 is", nc1, "for diffu = ",
+ }else{
+   print(paste("Transect", i, "I have not found clustering with diffu>=0", quote = FAI
+ }
+ plot(res, type = "p", pch = 0, xlab = "Number of clusters", ylab = "diffu", xaxt = "n
+ abline(h = 0, untf = FALSE)
+ axis(1, c(min_nc:max_nc))
+ title(paste("Clustering for transect", i, ""))
+
+ swarms <- as.data.frame(swarms)
+ swarms$type <-  c(pam(scaleSwarm, nc1, diss = FALSE)$clustering)
+
+ #print a summary table of the number of swarms assigned to each type
+ t <- table(swarms$type) #krill swarm types determined by PAM)
+ print(t)
+
+ #the following code can print example results summary for depth, height, length by ty
+ #lapply(split(swarms[, c("Depth_mean" ,'Height_mean','Corrected_length')], swarms$typ
+ }
+
+ message(paste("Finished classifying aggregations for transect", i))
```

```
+ }

[1] Minimal number of clusters where diffu >= 0 is 9 for diffu =  0.0295

 1  2  3  4  5  6  7  8  9
 8  4  7  1 11  1  3  1  1
[1] Minimal number of clusters where diffu >= 0 is 6 for diffu =  0.0184

 1  2  3  4  5  6
17 46 23  8  8  3
[1] Minimal number of clusters where diffu >= 0 is 3 for diffu =  0.003

 1  2  3
28  9 27
[1] "Transect 5 I have not found clustering with diffu>=0 FALSE"

1 2 3
5 2 1
[1] Minimal number of clusters where diffu >= 0 is 6 for diffu =  0.0214

1 2 3 4 5 6
1 2 4 2 1 4
```

This vignette has demonstrated an automated method for performing schools detection in texxtEchoview and then using texxtR to run classification and cluster analysis on the detected aggregations.


Dataflow [View 1]
038-120-200: line data depth GPGGA
038-120-200: position GPS fixes GPGLL
038-120-200: position GPS fixes GPGGA
Fixed depth 250 m
Fixed depth 6 m
038-120-200: position GPS fixes
RV Aurora Australis
200
120
38
038-120-200: Sv raw pings T3
038-120-200: TS raw pings T3
038-120-200: angular position raw pings T3
038-120-200: Sv raw pings T2
038-120-200: TS raw pings T2
038-120-200: angular position raw pings T2
038-120-200: Sv raw pings T1
038-120-200: TS raw pings T1
038-120-200: angular position raw pings T1
120 surface and seabed exclusion check
38 surface and seabed exclusion check
Line bitmap 1
120 Background noise removal
38 Background noise removal
120 Mask surface
038 Mask surface
120 seabed and surface excluded
38 seabed and surface excluded
120 transect Region bitmap
038 transect mask
120 transect mask
120 7x7 convolution
aggregations
38 data range bitmap
120 aggregations
120 mean aggregation Sv
Krill aggregations
120-38 aggregation dB difference filter
120-38 aggregation dB difference
38 aggregations
38 mean aggregation Sv
Seabed
6 m depth
Surface exclusion
Offset seabed
Smooth offset seabed
Best bottom candidate line pick 38
View 1

Appendix C

SUPPLEMENTARY MATERIAL FOR CHAPTER 5

This appendix provides the supplementary materials for Chapter 5, including model selection results and model diagnostics for the presence/absence and density components of the hurdle model

Authors:

Lisa-Marie K. Harrison[1], Steven Candy[2], Martin J. Cox[3], Guy Williams[4, 5], Robert Harcourt[1]

Affiliations:

[1] Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, New South Wales, Australia

[2] SCandy Statistical Modelling Pty Ltd, Blackmans Bay, Tasmania, Australia

[3] Australian Antarctic Division, 203 Channel Highway, Kingston, Tasmania, Australia

[4] Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Australia

[5] Antarctic Climate and Ecosystems Cooperative Research Centre, University of Tasmania, Hobart, Australia

Supplementary Tables for Model Selection and Validation

Traditional hurdle models were designed for zero-inflated count data (Poisson, Negative Binomial and Geometric distributions) however our data are continuous (krill densities) so these models are not appropriate. Additionally, our data includes observations over multiple sites (CTD stations) which necessitates the use of a random effect to account for extraneous differences between observations at different stations. There is currently no function in R that can incorporate both problems because hurdle mixed models for continuous data are still under development. To extend the traditional hurdle models to accommodate our data we modelled the two stages separately using a logistic mixed model (presence/absence) and a linear mixed model (conditional density). As we are interested in conditional inference rather than marginal inference there was no need to calculate a marginal likelihood.

Presence/Absence

The best model was logit(presence) = depth + temperature + salinity + re(stn), with the Akaike Information Criterion (AIC) and Area Under Curve (AUC) of the Receiver Operator Characteristic for candidate models shown in Table S1.

Table S1 *Model selection results – Presence/absence logistic mixed model showing Akaike Information Criterion (AIC) and Area Under Curve (AUC) for Receiver Operator Characteristic. The re(stn) term denotes a station random effect*

Model	AIC	AUC
y ~ depth + temperature + salinity + oxygen + phytoplankton + re(stn)	821.5	0.62
y ~ depth + temperature + salinity + phytoplankton + re(stn)	819.6	0.62
y ~ depth + temperature + salinity + oxygen + re(stn)	821.1	0.70
y ~ depth + temperature + salinity + re(stn)	**819.5**	**0.71**
y ~ depth + re(stn)	832.0	0.69
y ~ salinity + temperature + re(stn)	849.7	0.60

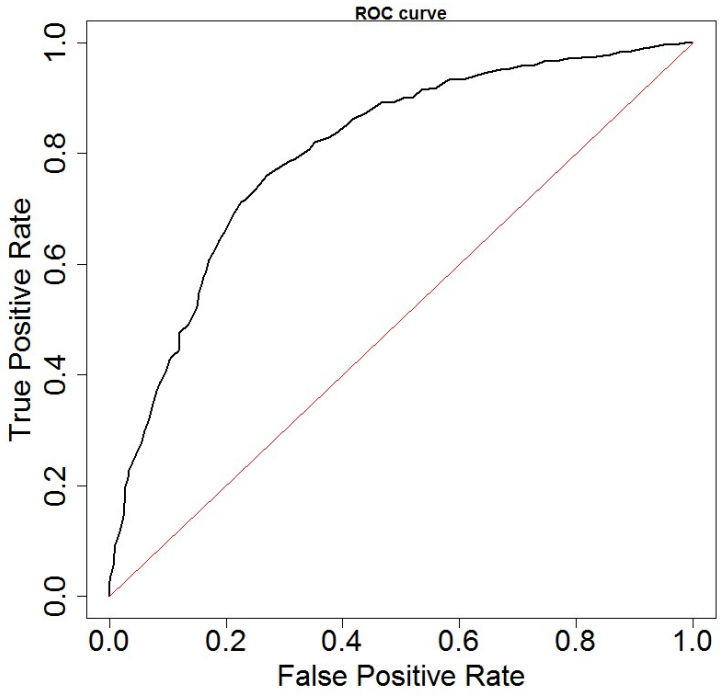


Figure S1 *Receiver Operator Characteristic (ROC) curve for the best presence/absence logistic model*

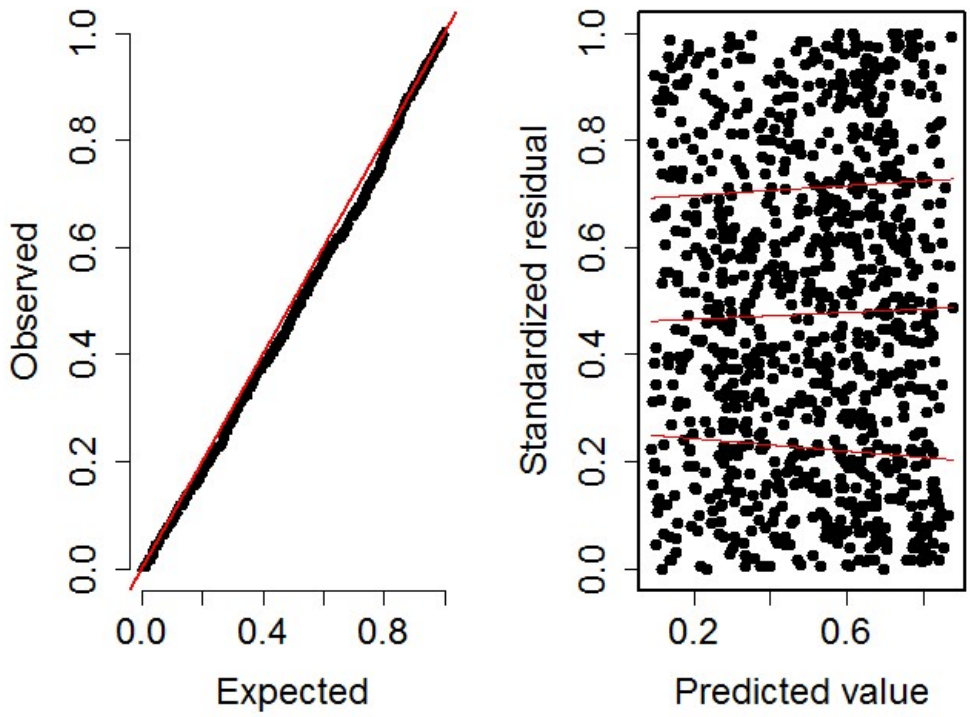


Figure S2 *Scaled quantile residuals for logistic mixed model for presence/absence. Normal QQplot (left) shows good model fit and Standardised residual plot (right) shows linear quantile lines nearly horizontal at 0.25, 0.5 and 0.75. Plots produced using R package DHARMa (Hartig, F (2017), DHARMa: Residual Diagnostics for Hierarchical (Multilevel/Mixed) Regression Models, https://CRAN.R-project.org/package=DHARMa)*

Density given presence

The best model was $\log_e$(density) ~ oxygen * $\log_e$(phytoplankton) + re(stn), where * indicates an interaction between oxygen and phytoplankton and re() denotes a random effect. AIC results for candidate models are shown in Table S2.

Table S2 *Model selection results – Linear mixed model for krill density showing Akaike Information Criterion (AIC) for candidate models. The re(stn) term denotes a station random effect. Cross-validation Root Mean Square Error (RMSE) is shown as a measure of goodness of fit*

Model	AIC	RMSE
$\log_e$(density) ~ oxygen * $\log_e$(phytoplankton) + re(stn)	**879.0**	13.00
$\log_e$(density) ~ oxygen + $\log_e$(phytoplankton) + re(stn)	882.1	13.07
$\log_e$(density) ~ depth + temperature + salinity + oxygen * $\log_e$(phytoplankton) + re(stn)	881.3	13.13
$\log_e$(density) ~ depth + temperature + oxygen * $\log_e$(phytoplankton) + re(stn)	881.2	13.04
$\log_e$(density) ~ depth + oxygen * $\log_e$(phytoplankton) + re(stn)	882.5	13.06

Table S3 *Model summary for best model of $\log_e$(krill density) ~ $\log_e$(phytoplankton) * oxygen + re(stn)*

Coefficient	Estimate	Standard Error	P-value	Variance Inflation Factor
Intercept	-0.51	0.16	0.002	

Log(phytoplankton)	0.66	0.08	<0.001	5.39
Oxygen	-0.11	0.09	0.19	4.51
Interaction	0.21	0.05	<0.001	1.78

Random Effects	
Coefficient	**Variance**
Station	0.78
Residual	0.45

The residuals (Figure S3) do not have any obvious problems and Figure S4 shows a plot of phytoplankton vs oxygen to show the coverage of the data.

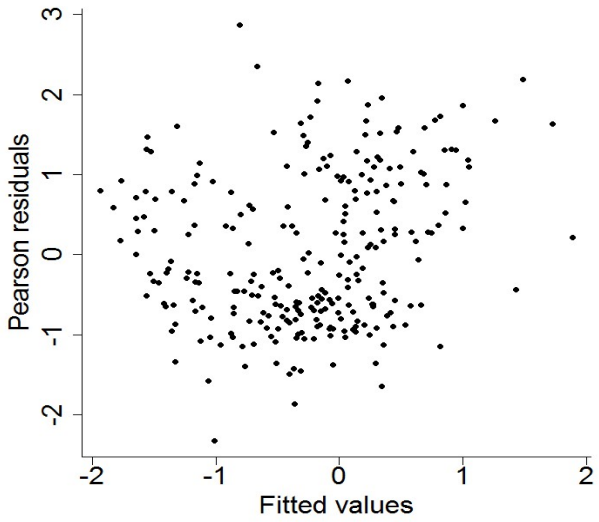


Figure S3 *Pearson residual plot for density model* log_e*(krill density) ~* log_e*(phytoplankton) * oxygen + re(stn)*

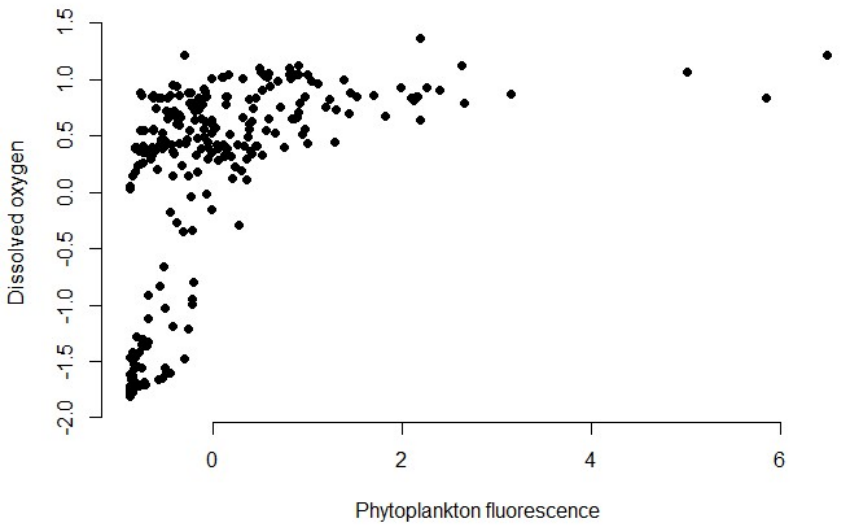


Figure S4 *Plot of phytoplankton and oxygen to show coverage of data*

Note on Random Effects in Cross Validation

There is no estimated random effect for the dropped station, however ignoring random effects when predicting the dropped station causes skewing during back-transformation, hence we simulated the random effect for the new station using the estimated standard deviation.

Appendix D

AERIAL SURVEY FINAL REPORT: A TECHNICAL REPORT FOR THE DEPARTMENT OF PRIMARY INDUSTRIES

This pdf report was prepared at the conclusion of the aerial survey fieldwork conducted with the Department of Primary Industries during the candidature of this thesis. It was intended that this data would be used to assess drivers of marine megafauna distribution off coastal New South Wales, Australia, however due to time constrains this data was not used in this thesis.

Authors:

Lisa-Marie K. Harrison[1]

Affiliations:

[1]Marine Predator Research Group, Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, North Ryde, New South Wales, Australia

AERIAL SURVEY FINAL REPORT

December 2013 – October 2015

Lisa-Marie Harrison

Executive Summary

Between December 2013 and October 2015 aerial surveys were conducted between Wollongong and Newcastle for bather protection. There were a total of 53 aerial survey days, including 19 inter-observer discrepancy trials. Total marine wildlife sightings were dominated by baitfish schools (n = 1708) and bottlenose dolphin pods (n = 471), while shark sightings rates were low. There were 34 White, 1 Bull, 2 Whaler, 144 Hammerhead and 7 unidentified shark sightings. The highest number of sightings per survey occurred in spring for baitfish and bottlenose dolphins, and summer for all sharks. Sea state and Turbidity had little effect on sighting rates except for bottlenose dolphins, for which sighting rates declined with increasing sea state. Most shark sightings occurred in the bay between Swansea Heads and Redhead beach, while baitfish and bottlenose dolphins were seen throughout the entire transect.

Using distance sampling methods, an estimate of 150 baitfish schools (CV = 0.109) and 202 bottlenose dolphins (CV = 0.195) are present within 1km of the coast and between Wollongong and Newcastle. There were too few shark sightings to calculate abundance. For southbound transects where the observer is looking offshore, the average probability of detection was 12.7% for baitfish, 25.4% for bottlenose dolphins and 11.8% for non-hammerhead sharks.

Contents

List of Tables and Figures

Methods

Flight path

Between December 2013 and October 2015 there were a total of 53 aerial survey days, including 19 inter-observer discrepancy trials. The same flight path was flown during each survey, with a northbound and southbound transect each day. The northbound transect ran between South Wollongong and the Sygna Wreck at Newcastle, and the southbound transect ran in the opposite direction, from the Sygna Wreck to South Wollongong. The northbound transect commenced at 8am (AEST) on each survey date. Between the transects, the helicopter lands at Newcastle Regional Heliport to refuel and an hour break is taken. This allows the observer to rest and the marine wildlife to move in and out of the survey area to decrease the chance of recounting. Each transect takes approximately two hours from take-off to landing, however this varies slightly with wind conditions and the number of sightings requiring the helicopter to circle. A map of the survey track is shown in **Error! Reference source not found.**.

Figure 1 Flight path (yellow) over google earth map

Survey methods and equipment

Each survey used the same method to standardise the results as much as possible. The aircraft was a Robinson 44 with the pilot seated on the front right, the observer on the front left, and the photographer on the back left. Both left hand doors are removed (note: the front left hand door was left on during two surveys as a trial, however it was determined that the door should be taken off). The northbound transect was flown approximately 300m from the back surf line with the observers looking onshore. The southbound transect was flown over the surf line, with the observers looking offshore.

The observer used a voice recorder to record all information to minimise the time spent looking away from the transect. The observer used a handheld Garmin 76 GPS to record a waypoint at each event code and an inclinometer to measure the angle to each sighting. The following events were recorded:

- Start and end of transect
- Leaving and returning to transect (i.e.: due to circle backs for photographs)
- Marine wildlife sightings (Sharks, cetaceans, pinnipeds, turtles, rays and baitfish)
- Changes in environmental conditions

As these surveys were collecting distance sampling data, it was crucially important that the observer be the only person calling sightings. The pilot and photographer were only allowed to notify the observer of a sighting once it was past 90 degrees to the observer, and hence deemed a "missed" sighting.

In addition to notifying the observer of missed sightings, the photographer was responsible for photographing marine wildlife during both transects. They also captures images of each netted beach during the northbound transect for beach usage counts. From the Spring 2014 season onwards, the photographer also photographed all rock fishermen seen during the northbound transect. The netted beaches to be photographed on the northbound trip are:

South Wollongong	Manly	Umina	Merewether
North Wollongong	North Steyne	Kilcare	Dixon Park
Thirroul	Queenscliff	McMasters	Bar
Austinmer	Freshwater	Copacabana	Newcastle
Coledale	Curl Curl	Avoca	Nobby's
Garie	Dee Why	North Avoca	Stockton
Wattamolla	Narabeen	Terrigal	
Cronulla	North Narabeen	Shelly	
North Cronulla	Warriewood	The Entrance	
Elouera	Mona Vale	Soldiers	
Wanda	Newport	Lakes	
Maroubra	Bilgola	Catherine Hill Bay	
Coogee	Avalon	Caves	
Bronte	Whale	Swansea-Blacksmiths	
Bondi	Palm	Redhead	

Inter-observer surveys

During the 2014-2015 survey season, there were 18 inter-observer discrepancy surveys. These were conducted with Lisa-Marie Harrison as the first observer, seated on the front left of the helicopter, and Vic Peddemors as the second observer, seated on the back left. As only three people (including the pilot) could take part in the surveys, the second observer also acted as the photographer. Only the southbound surveys were flown as inter-observer surveys, because the second observer was required to take beach usage images during the northbound transect which would have changed their sighting effort, hence making it an unfair inter-observer trial. If an image was required during the southbound inter-observer survey, the transect was stopped, the helicopter circled to allow the second observer to capture photographs, and only restarted once the photographs have been taken.

To get independent results, the two observers must be completely isolated and unable to alert each other to a sighting by either sight or voice. A chipboard screen was made and fastened to the helicopter with cable ties to prevent the observers from seeing each other. The helicopter COM system was set up so that the observers could record sightings into their voice recorders without the information being heard by the other members of the survey team.

Survey Dates

A total of 53 survey dates were flown between summer 2013 and October 2015. Lisa-Marie Harrison was the first observer for all surveys except those marked by * in Table 1, where Vic Peddemors was primary observer.

*Table 1 Aerial survey dates by Season. *Vic Peddemors was observer*

	Summer 2013	**Autumn 2014**	**Spring 2014**	**Summer 2014**	**Autumn 2015**	**Spring 2015**	
	21/12/2013	9/4/2014	13/09/2014*	20/12/2014	1/04/2015	9/09/2015	
	24/12/2013	21/4/2014	21/09/2014	25/12/2014*	6/04/2015	12/09/2015	
	25/12/2013	23/4/2014	24/09/2014	27/12/2014*	11/04/2015	16/09/2015	
	28/12/2013	1/5/2014	2/10/2014*	31/12/2014	14/04/2015	19/09/2015	
	31/12/2013	14/5/2014	5/10/2014	1/01/2015	29/04/2015	30/09/2015	
	1/1/2014	17/5/2014	9/10/2014	3/01/2015	18/05/2015	4/10/2015	
	4/1/2014		16/10/2014	7/01/2015	19/05/2015	5/10/2015	
	8/1/2014			14/01/2015	20/05/2015	14/10/2015	
	11/1/2014			17/01/2015			
	15/1/2014*			21/01/2015			
	18/1/2014*			25/01/2015			
	27/1/2014*			26/01/2015			
Number of Surveys	**12**	**6**	**7**	**12**	**8**	**8**	**TOTAL = 53**

5

Results

Marine Wildlife observations

Sighting numbers

Sighting numbers per season are shown in Table 2. Full sighting numbers for each survey are provided in the appendix. Baitfish and bottlenose dolphins were the most commonly seen species, with very few sharks seen. Except in the inter-observer section of the results, all tables and figures do not include secondary sightings by observer 2 during inter-observer surveys for consistency because most surveys were not inter-observer.

Table 2 Marine wildlife sightings by season. Note: Group numbers are shown (e.g.: Dolphin pods), not individuals. n = number of surveys in each season. Secondary sightings from inter-observer surveys are not included.

		Summer 2013 (n = 12)	Autumn 2014 (n = 6)	Spring 2014 (n = 7)	Summer 2014 (n = 12)	Autumn 2015 (n = 8)	Spring 2015 (n = 8)	
Fish	Baitfish	192	175	335	303	226	477	**1708**
	Sunfish				43			**43**
	Large Fish	3	2	1	4	1	2	**13**
Mammals	Bottlenose Dolphin	83	39	95	119	34	101	**471**
	Common Dolphin				1			**1**
	Humpback Whale		1	7			3	**11**
	Seal	2	5	4	9	8	4	**32**
Sharks	White Shark	27		2	5			**34**
	Bull Shark					1		**1**
	Whaler Shark				2			**2**
	Hammerhead Shark	44	10	8	104	20	2	**188**
	Unidentified Shark	2		1	1	2	1	**7**
Other	Turtle	1	14	5	4	16	12	**52**
	Ray	5	17	28	26	22	19	**117**
		359	**263**	**486**	**621**	**330**	**621**	**2680**

Group size

Bottlenose dolphin group size (Figure 2) varied considerably, from 1 to 120 individuals (mean = 16). No relationship was seen with bottlenose dolphin group size and distance from transect.

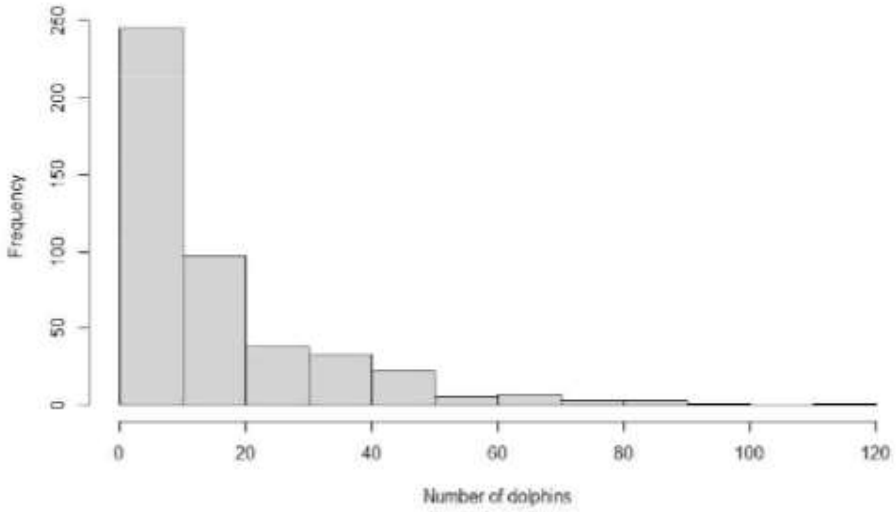


Figure 2 Estimated bottlenose dolphin group size

Baitfish schools were classified as "small", "medium" or "large". An very small opposite relationship is seen in the north and south survey directions, with larger schools been seen offshore and smaller schools onshore in both directions (Figure 3). This may be because small baitfish schools are most visible in very shallow water and are likely to be missed further offshore.

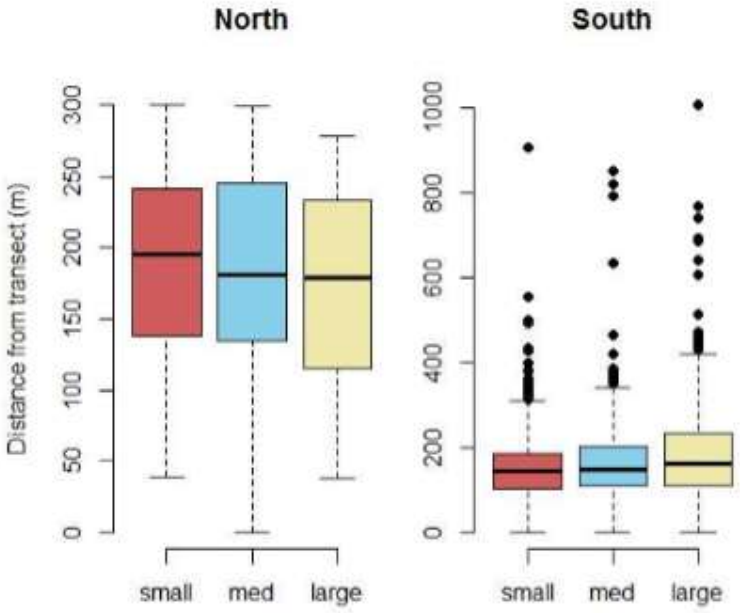


Figure 3 Distance from transect by baitfish school size. Note: North direction distances are truncated at 300m because observers are looking onshore

Distance Offshore

During southbound transects the helicopter is positioned at the edge of the water with the observer looking out to sea. As the transect roughly follows the shoreline, the distance of a sighting from transect is approximately equal to distance offshore. Cetaceans and baitfish schools were seen the furthest from shore, while seals, rays sharks and turtles were never seen more than 500m from shore (Figure 4). It was not possible to look at distance offshore by time of day because the helicopter always travelled south at the same time of day.

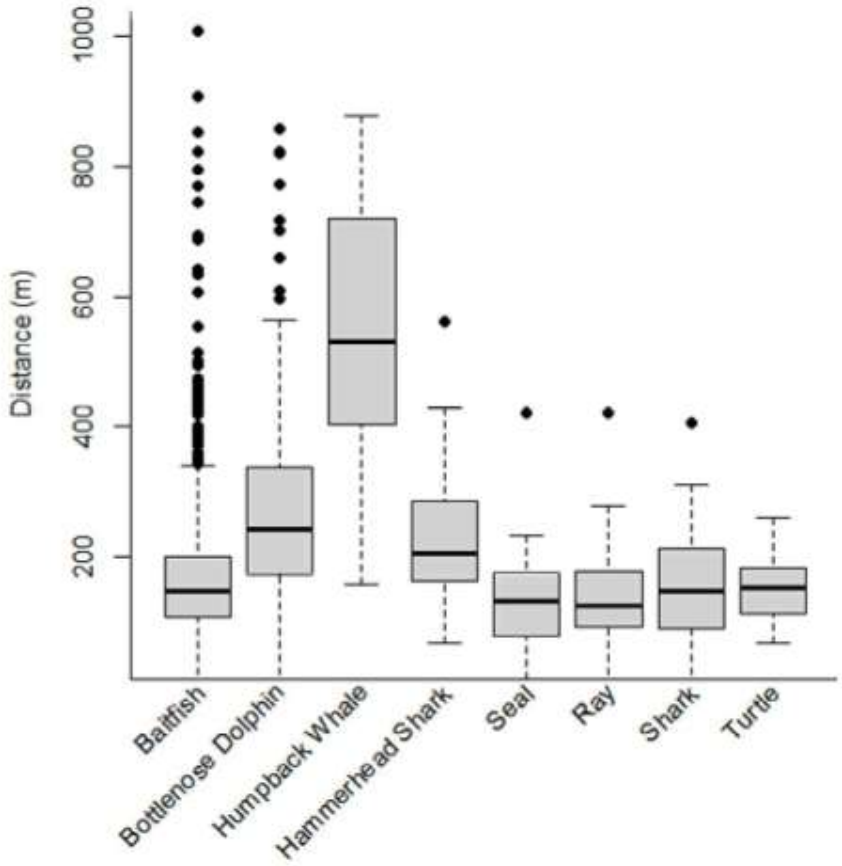


Figure 4 Distance offshore (m) by species. All non-hammerhead shark sightings have been combined into 'Shark'

Sighting locations

Sighting locations are only discussed for baitfish, bottlenose dolphins and sharks. Baitfish and dolphins were seen all along the length of the transect, while a non-hammerhead shark was only seen once south of The Entrance (Figure 5). In particular, the bay between Swansea Heads and Redhead beach was where most shark sightings occurred.

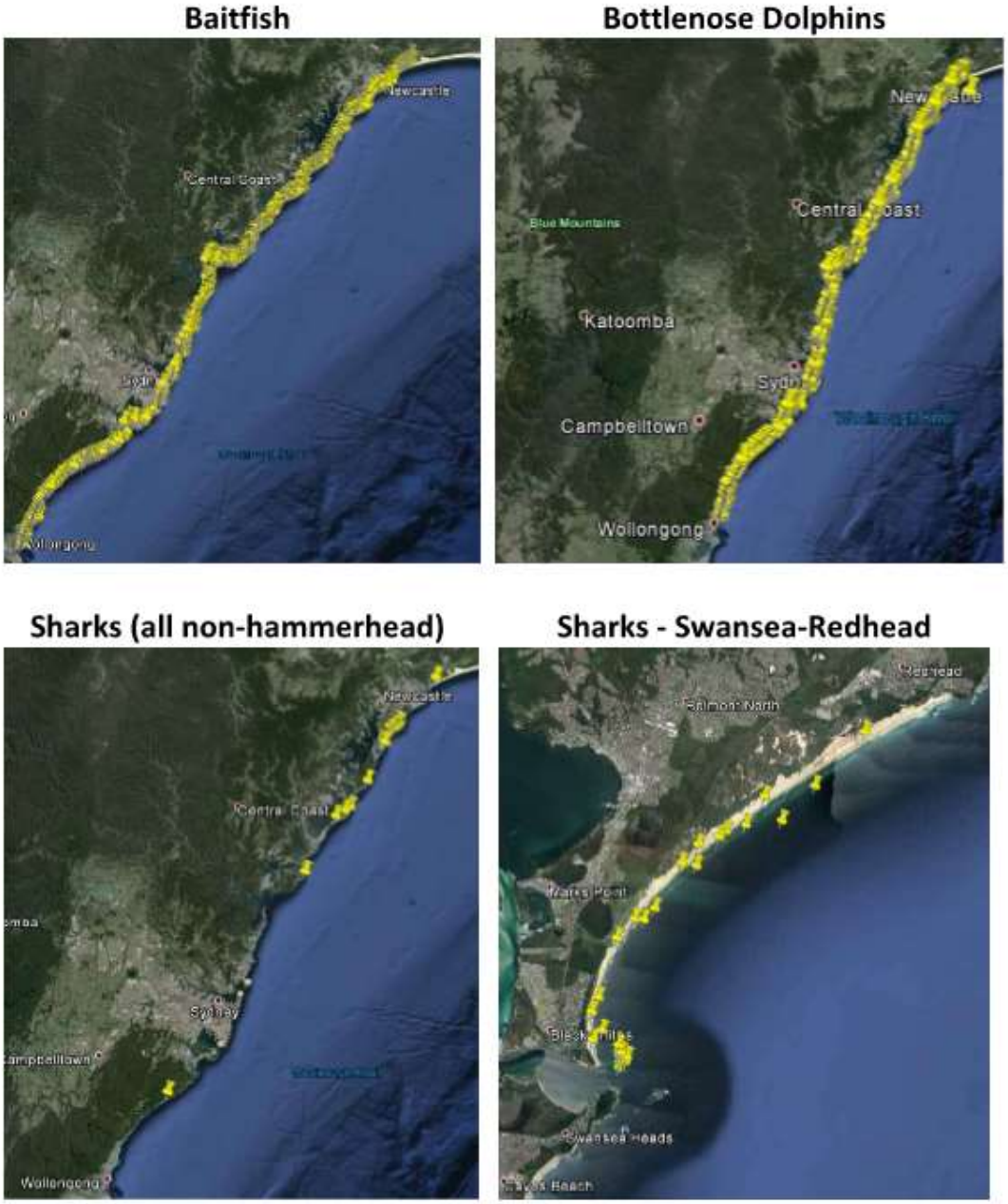


Figure 5 GPS sighting locations of baitfish, bottlenose dolphins and sharks during the 53 surveys

Seasonal and environmental differences

The number of sightings per survey differed between seasons and within each season there was a large amount of variation in number of sightings (Figure 6). Shark species were seen more often in summer while baitfish and dolphins were seen most in spring.

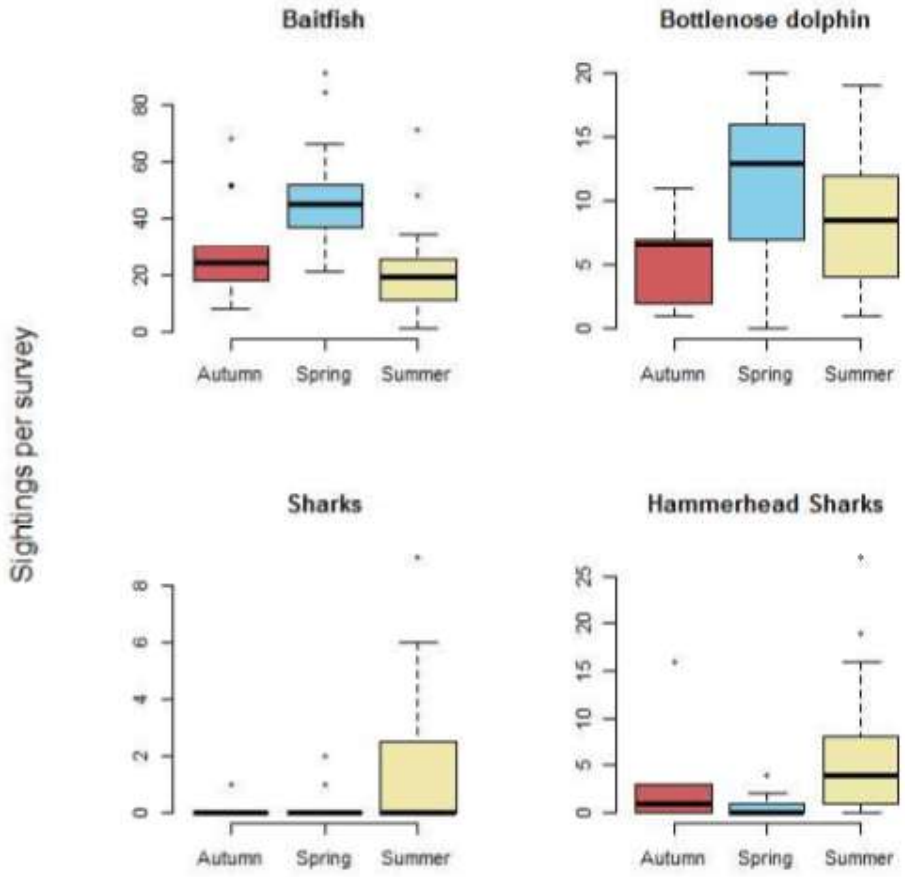


Figure 6 Sightings per survey by season. Note: 'Sharks' includes all non-hammerhead species

To assess whether these differences may be due to environmental factors limiting sightability in some seasons, proportion of survey time at each level of Sea State, Turbidity and Cloud Cover is plotted (Figure 7). Sea state and turbidity were fairly consistent among seasons, while Cloud Cover was lower in spring than in the other seasons.

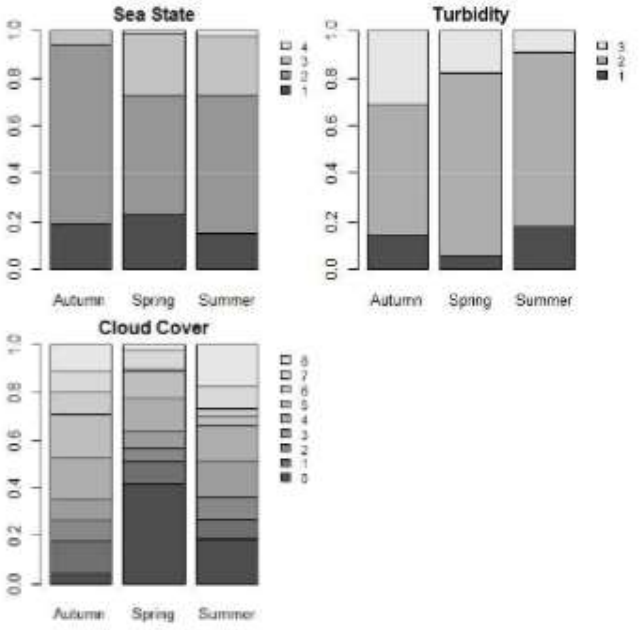


Figure 7 Proportion of flight time at environmental factor levels during each season. Turbidity 1 = excellent, 2 = good, 3 = turbid; Cloud Cover 0 = no clouds to 8 = full cover

The number of sightings per hour was calculated using the effort at each environmental variable level (Figure 8). For baitfish, there was no difference in sighting rate at different Sea State or Turbidity levels. Sighting rates for dolphins were the same with Turbidity, but decreased with every increase in Sea State. This could be because splashing and foam from dolphins' fins are a major visual cue for observers, but are more difficult to see when there are breaking wave crests with Beaufort Sea States 3 and higher. With the low sighting rates for sharks, it is more difficult to see any patterns that may be present with sighting rate and environmental conditions.

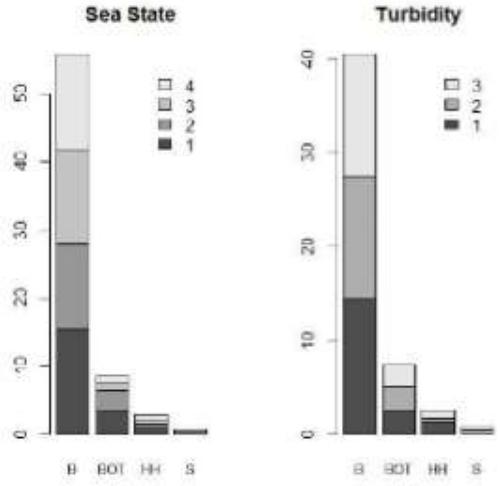


Figure 8 Number of sightings/hr at Beaufort Sea States 1 - 4 and Turbidity (1 = excellent clarity, 2 = good, 3 = turbid), for Baitfish (B), Bottlenose dolphins (BOT), Hammerhead Sharks (HH) and all other sharks (S)

Inter-observer surveys

The total number of missed sightings by each observer during the 19 surveys is shown in Figure 9. These results show the importance of conducting inter-observer surveys if surveys conducted with different observers are to be compared. See Table 6 in the Appendix for the full inter-observer sighting data by survey.

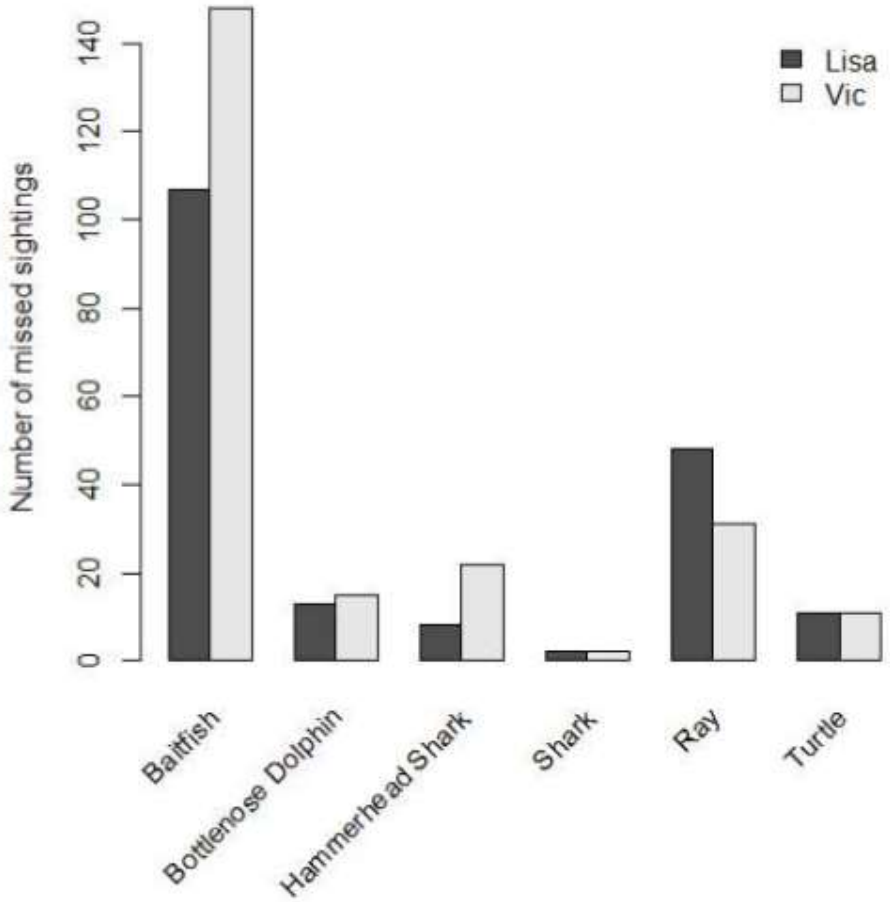


Figure 9 Missed sightings during 18 inter-observer surveys

Distance Sampling

A distance sampling analysis was used to calculate the probability of detection of each species and estimate abundance within the survey region. For this analysis, only southbound flights were used because the strip width of the northbound flights is limited to 300m because the observers are looking onshore. Distance sampling revolves around the notion that sighting rates will decrease with distance from transect because animals further away are smaller and harder to see. The probability of an available animal being detected will hence decrease with increasing distance from transect. For Bottlenose dolphins, baitfish schools and non-hammerhead sharks, the probability of detection curves are shown in Figure 10. A gamma detection function fit best for baitfish and bottlenose dolphins while a hazard-rate function was best for shark species.

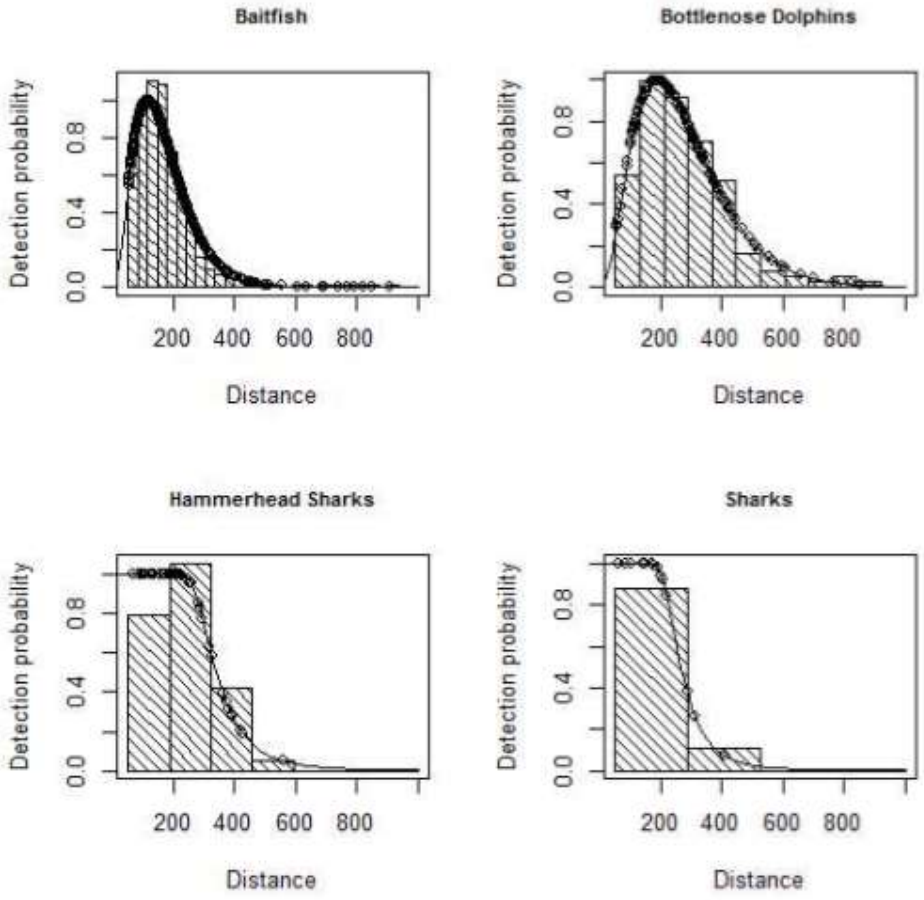


Figure 10 Detection functions with distance from transect (m)

Inter-observer data between Lisa-Marie Harrison and Vic Peddemors was used to calculate the probability of detection on the transect line, p(0). This value does not take into account missed sightings due to distance. Average p is the average probability of detection across all distances. To get the corrected probability of detection (for distance and inter-observer) these two values are multiplied together. From Table 3, the probability of detection for baitfish is 12.7%, for bottlenose dolphins is 25.4% and for sharks is 11.8%.

Table 3 Probability of detection (p) for Lisa-Marie Harrison

	Probability of detection (p)		
	Average p (corrects for distance)	**p(0) (corrects using interobserver)**	**Corrected Average p (Average p*p(0))**
Baitfish	0.187 (CV: 0.020)	0.678 (CV: 0.038)	0.127 (CV: 0.043)
Bottlenose Dolphin	0.322 (CV:0.054)	0.788 (CV: 0.067)	0.254 (CV: 0.086)
Sharks	0.235 (CV: 0.181)	0.500 (CV: 0.500)	0.118 (CV: 0.532)

Density and abundance estimates, shown in Table 4, are calculated using the formula:

$$\hat{d} = \frac{N}{wLp_a}$$

Where $\hat{d}$ = density, N = mean animals seen per survey, w = strip width, L = transect length and p_a = corrected average probability of detection from Table 3. The abundance estimates are for the 265km * 1km survey area. The estimates for sharks are very low because there were too few sightings to get a good approximation of the detection function.

Table 4 Density and Abundance estimates corrected for perception bias. 'Sharks' includes all non-hammerhead shark sightings, including white, whaler, bull and unidentified sharks.

	Density (N/km^2)	**Abundance (N)**	**Coefficient of Variation (CV)**
Baitfish	0.57	150	0.109
Bottlenose Dolphin Pods	0.05	13	0.158
Individuals	0.76	202	0.195
Sharks	0.01	3	0.627

Conclusion

Baitfish and bottlenose dolphins made up most marine wildlife sightings between Wollongong and Newcastle from 2013 – 2015. Density estimates from distance sampling indicate that there are 0.57 baitfish/km^2 and 0.76 bottlenose dolphin s/km^2. There were not enough shark sightings to get an accurate abundance or density estimate. Of sharks, the only species with >35 sightings in total over the six seasons was hammerhead sharks.

There were large seasonal differences seen, with spring having the highest number of baitfish and bottlenose dolphin sightings and summer having the most shark sightings. Environmental condition levels did not greatly effect sighting rate in general, except for a decrease in bottlenose dolphin sightings at higher sea states. The practice of recording environmental variables whenever they change, rather than at each sighting, is vital because otherwise the time in minutes spent at each environmental level cannot be accurately calculated. Recording via the headset and a voice recorder is also important during distance sampling surveys from aircraft because it allows the observer to always have their eyes on transect, reducing missed sightings. The inter-observer surveys were highly valuable for estimating the probability of detection on the transect line (distance = 0) for the primary observer.

Table 5 Total sightings by survey date (secondary sightings for inter-observer surveys are not included)

	Fish		Mammals				Sharks					Others		
	Baitfish	Sun Fish	Bottlenose Dolphin	Common Dolphin	Humpback Whale	Seal	White	Bull	Whaler	Hammer-head	Unid	Ray	Turtle	Total
21/12/2013	11		4							1		1		17
24/12/2013	9		6							1				16
25/12/2013	16		3							4		3		27
28/12/2013	19		14				4			14				51
31/12/2013	26		10							9			1	46
1/01/2014	11		11				3			7		1		33
4/01/2014	23		3				2							28
8/01/2014	25		3				6			2				37
11/01/2014	24		7			1	9			2				44
15/01/2014	11		2											13
18/01/2014	11		10				2			1	2			26
27/01/2014	6		10			1	1			3				21
9/04/2014	8		7							1		1		17
21/04/2014	19		10							2		1	4	36
23/04/2014	29		7			1						3	2	43
1/05/2014	21		1			1				1		4		28
14/05/2014	68		9			3				3		6	5	94
17/05/2014	30		5		1					3		2	3	45
13/09/2014	57		1									5		63
21/09/2014	22		7			1						3		33
24/09/2014	52		15		1	1						4		74

2/10/2014	37		16		2		1			2	1	9	3	71
5/10/2014	35		12				1					1	1	50
9/10/2014	45		16							1		5	1	68
11/10/2014	66		18		2					4				90
16/10/2014	21		10		2	2				1		1		37
20/12/2014	34		13			3	1			6		2		59
25/12/2014	19		13			1				4		2	1	40
27/12/2014	1		6									1		8
31/12/2014	1		4				1			7	1	1	1	16
1/01/2015	15	4	19							7		3		48
3/01/2015	48	7	15			1				16				88
7/01/2015	33	6	6			1				27		1		74
14/01/2015	19	6	4							19		1		50
17/01/2015	71	6	10	1		2				10		6		106
21/01/2015	20	5	11							5		6		47
25/01/2015	31	9	17				3			2		2	2	68
26/01/2015	11		1			1			2	1		1		17
1/04/2015	51		7							16	1	7	3	85
6/04/2015	52		7										1	60
11/04/2015	26		1			2						3	2	34
14/04/2015	22		2			2		1		1		1	1	30
29/04/2015	13		1									1	1	16
18/05/2015	18		3			2					1	4	3	31
19/05/2015	16		6			1						5	2	30
20/05/2015	28		7			1				3		1	3	44
9/09/2015	91		17			3								112
12/09/2015	83		20							1		2	4	110
16/09/2015	52		9									1	2	65
19/09/2015	46		5			1								52
30/09/2015	51		5									5		61

16

4/10/2015	38		13								1			52	
5/10/2015	37		13			1					2	4		57	
9/10/2015	38							1			1	1		41	
14/10/2015	41		19		3						7	1		71	
Grand Total	**1708**	**43**	**471**	**1**	**11**	**32**	**34**	**1**	**2**	**188**	**7**	**117**	**52**	**2680**	

Table 6 Number of missed sightings during each inter-observer survey. For each species, 'Total Seen' is the total unique objects seen and Lisa and Vic's columns contain the number of those objects that seen of them missed. E.g.: There were 11 unique baitfish schools seen on 21/04/2014, of which Lisa missed 1 and Vic missed 8. 'Unique' means that objects seen by both observers are not counted twice.

	Baitfish			Bottlenose Dolphins			Hammerhead Sharks			Rays			Sharks			Turtles		
	Missed		Total	Missed		Total	Missed		Total	Missed		Total	Missed		Total	Missed		Total
	Lisa	Vic	Seen	Lisa	Vic	Seen	Lisa	Vic	Seen	Lisa	Vic	Seen	Lisa	Vic	Seen	Lisa	Vic	Seen
21/04/2014	1	8	11	1	1	4		2	2		1	1					4	4
23/04/2014	12	10	30			2				1	1	4				1	1	3
14/05/2014	3	20	46	1		4		2	2		2	4				1	2	3
24/09/2014	3	10	34		2	5				2	1	6						
5/10/2014	6	10	30	1	2	5				4	1	5						
9/10/2014	6	8	32	2	2	14		1	1	1	2	4					1	1
16/10/2014	2	1	11	1	2	6		1	1	2		3						
20/12/2014	5	9	27			6	2	3	6	3		3		1	1			
7/01/2015	9	6	24	1		2	2	9	17	3	1	4						
14/01/2015	2	1	15	1	1	3	1		1	9	1	10						
17/01/2015	10	12	44					1	1	4	4	8						
21/01/2015		4	10		1	6	1		1	2	5	8						
25/01/2015	7	5	18		1	7				2	2	4			2	1		1
26/01/2015		4	10					1	1		1	1		1	1			
1/04/2015	20	10	54	3	1	5	1	2	3	5	5	12	2		2	1		2
11/04/2015	6	11	26	1	1	2	1		1		2	3				1	1	2
14/04/2015	3	11	22		1	1				4	1	5				3		3
20/05/2015	12	8	33	1		4				6	1	7				3	2	5
Total	**107**	**148**	**477**	**13**	**15**	**76**	**8**	**22**	**37**	**48**	**31**	**92**	**2**	**2**	**6**	**11**	**11**	**24**