Manh Cuong Ngo

Modelling Marine Mammal Reactions

PHD THESIS

THIS THESIS HAS BEEN SUBMITTED TO THE PHD SCHOOL OF
THE FACULTY OF SCIENCE, UNIVERSITY OF COPENHAGEN

Department of Mathematical Sciences
University of Copenhagen

January 2022

Manh Cuong Ngo
cuong.ngo@math.ku.dk
Department of Birds and Mammals
Greenland Institute of Natural Resources
3900 Nuuk
Greenland

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark

Thesis title:	Modelling Marine Mammal Reactions
Supervisor:	Professor Susanne Ditlevsen University of Copenhagen
Co-supervisor:	Professor Mads Peter Heide-Jørgensen Greenland Institute of Natural Resource
Assessment committee:	Professor Helle Sørensen (chair) University of Copenhagen
	Professor Roland Langrock University of Bielefeld
	Senior scientist Jacob Nabe-Nielsen Aarhus University
Date of Submission:	January 31, 2022
Date of Defense:	April 1, 2022
ISBN:	978-87-7125-053-4

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen. It was supported by the Greenland Research Council.

Abstract

This Ph.D thesis consists of some works contributing to the field of ecology modelling for Arctic whales. Arctic cetaceans are facing challenges due to human activities, like changes of habitats, scarcity of prey due to fisheries and warmer water, pollution, anthropogenic noise, etc. Therefore, understanding their behaviour is very important for conservation management plans. Using several long-term datasets collected by tagging of individual animals, we apply different statistical and machine learning methods to understand the behaviour of two endemic Arctic whales, bowhead whales and narwhals. We present new hidden Markov models, taking into account the correlation between maximum depth and dive duration, to understand the diving behaviour of a narwhal using dive data of 83 days. Our models relax the contemporaneous conditional independence assumption, which is often used in ecological modelling, leading to improvement of the model fit. We also establish machine learning models using deep learning, predicting the prey capture attempts from accelerometer data, without the need to use resource-heavy acoustic data. Our models outperform the classical machine learning method random forest, and the statistical method logistic regression. Our results show that narwhals do not make instant change in acceleration which is often used as a proxy for prey captures in several other cetacean species. Finally, we propose Tweedie generalized linear models to understand the distribution of bowhead whales under warming water in the Arctic area. We exploit GPU computing to boost up the model performance, thus, allowing to fit to a higher resolution of environmental data and daily whale positions.

Resumé

Denne Ph.d.-afhandling består af nogle arbejder, der bidrager til området økologisk modellering af arktiske hvaler. Arktiske hvaler står over for nogle udfordringer udsprunget af menneskelige aktiviteter, såsom ændringer i deres habitat, mangel på føde på grund af fiskeri og varmere vand, forurening, menneskeskabt støj osv. Derfor er forståelsen af deres adfærd meget vigtig for forvaltning og planer for arternes bevarelse. Ved at bruge flere datasæt indsamlet over lang tid ved tagging af enkeltindivider, anvender vi forskellige statistiske og maskinlæringsmetoder til at forstå adfærden hos to endemiske arktiske hvaler, grønlandshvaler og narhvaler. Vi præsenterer nye Hidden Markov-modeller, der tager højde for sammenhængen mellem maksimal dybde af et dyk og dykkets varighed, for at forstå narhvalers dykkeadfærd ved hjælp af dykkedata indsamlet over 83 dage. Vores modeller kræver ikke at der er samtidig betinget uafhængighed, som ellers ofte bruges i økologisk modellering. Dette fører til en forbedring af modelfittet. Vi etablerer også maskinlæringsmodeller ved hjælp af deep learning, der forudsiger hvalernes forsøg på at fange byttedyr fra accelerometerdata uden at skulle bruge ressourcetunge akustiske data. Vores modeller udkonkurrerer den klassiske maskinlæringsmetode random forest og den statistiske metode logistisk regression. Vores resultater viser, at narhvaler ikke foretager en øjeblikkelig ændring i accelerationen, som ellers ofte bruges som proxy for byttefangst hos flere andre hvalarter. Til sidst foreslår vi Tweedie generaliserede lineære modeller for at forstå den spatio-temporale fordeling af grønlandshvaler som funktion af overfladetemperaturen i det arktiske område. Vi udnytter GPU-databehandling til at øge modellens ydeevne, så de kan passe til den højeste opløsning af miljødata og daglige hvalpositioner.

Preface

This thesis has been submitted in partial fulfillment of the requirements for the Ph.D. degree at the Department of Mathematical Sciences, Faculty of Science, University of Copenhagen. This work was written between October 2018 and January 2022 at the Greenland Institute of Natural Resource and Section for Statistics and Probability Theory. The research was funded by the Greenland Research Council.

Acknowledgments

First and foremost, I would like to express my gratitude to both of my supervisors Susanne Ditlevsen and Mads Peter Heide-Jørgensen to let me have a chance to work on two marvellous species. To me, it is a dream come true, something I never imagined I could have a chance, only six year ago. Thank you so much for your patience to teach me too many things that I never heard of when I came to Denmark the first time five years ago.

I wish to thank Prof. Helle Sørensen, Prof. Roland Langrock and Senior scientist Jacob Nabe-Nielsen for being members of my Ph.D thesis assessment committee. To all my co-authors and colleagues Eva Garde, Nynne Nielsen, Outi Tervo, Rikke Hansen, Raghav Selvan, Jonas Peters (in Denmark), and Aili Labansen, Adriana Nogueira, Fernando Ugarte, Klaus Nygaard, Julius Nielsen and other people at the Institute (in Greenland), thank you very much for the collaborations, the friendly welcome when I just arrived to the new lands, the fruitful discussion and enjoyable moments. I would like to thank my colleagues at the Math Department at KU for joyful events and great scientific discussions, even I did not see you usually. I thank Martin Emil Jakobsen to let his thesis open source, so I can borrow the beautiful LaTex template.

I would like to thank Gert Søndergaard and Tom Poes Jensen for the IT support; Anthon Møller, Kenneth Nielsen, Katrine Olsen, Lene Holm Kleist, Nina Weisse, Carina Belle Jensen, and other for the administrative helps.

To all my buddies in Østerbro, especially our SB group and my best friend there, thank you for the wonderful time we have spent together. I really felt younger when I were with you. And to my family, no matter how far the physical distance between us, we are always together.

Contents

Chapter 1

Introduction

1.1 Summary of Contributions

This thesis consists of five introductory chapters and three main papers. The main papers aim to contribute to the domain of ecological modelling using new theoretical and practical tools from statistics and machine learning. Chapter 3 introduces the background of general linear models (GLMs) and the uncommon Tweedie GLMs, chapter 4 presents hidden Markov models with correlation, and chapter 5 introduces deep learning and U-Net. The main papers consist of two published articles, and one work-in-progress paper. The three main papers are:

Paper I Manh Cuong Ngo, Mads Peter Heide-Jørgensen and Susanne Ditlevsen: Understanding narwhal diving behaviour using Hidden Markov Models with dependent state distributions and long range dependence. PLoS Computational Biology, 15(3): e1006425, 2019.

Paper II Manh Cuong Ngo, Raghavendra Selvan, Outi Tervo, Mads Peter Heide-Jørgensen, Susanne Ditlevsen: Detection of foraging behavior from accelerometer data using U-Net type convolutional networks. Ecological Informatics., 62, 101275, 2021.

Paper III Manh Cuong Ngo, Susanne Ditlevsen and Mads Peter Heide-Jørgensen: Sea surface temperatures drive the movements of bowhead whales (*work in progress*).

In Paper I, we propose the new correlated Hidden Markov Models (HMMs) for a narwhal dive data set consisting of time series of 83 days, which takes into account the correlation between two variables. HMMs have been used recently for animal movement modelling, due to their ability of taking into account the autocorrelation of time series data. Usually, *contemporaneous conditional independence* is assumed since it is simple to implement. However, in many cases, there exists from-medium-to-strong dependence between state dependent processes. Thus, it is unrealistic to ignore such dependence. In our case, the maximum depth and dive duration are correlated with medium dependence. We exploit the continuity of the response variables and propose correlated HMMs with correlation between maximum depth and dive duration, based on the log-normal distribution and the gamma distribution. We compare these models with independent HMMs with both a log-normal distribution and a gamma distribution. We clearly see a better fit of correlated models in both distributions. It indicates that one should consider the correlation between variables wherever possible.

In Paper II, we investigate the possibility of prey capture detection using accelerometer data for narwhals, without using the heavy resource demanding acoustic data. It is based on an assumption that prey capture events are closely related to buzzes, a high-rate series of echolocation clicks. We find that the sudden changes in acceleration, called jerks, which are often used as proxies for several different whale species, is not useful to detect capture

events in narwhals. It may be due to the narwhals using suction feeding when they are close enough to potential preys, or due to the tags being placed on the back too far away from the head of the narwhals. We then propose different models based on logistic regression, and two machine learning methods, random forest and U-Net deep learning, to capture more complex signals in the data. The results show, surprisingly, that random forest, which is often very good for tabular data, is worse than logistic regression. Our deep learning models show good potential by outperforming the two other methods and they give decent buzz detection rates. It encourages more applications of deep learning in the ecological field.

In Paper III, we study the relationship between sea surface temperature and the spatio-temporal distribution of Arctic bowhead whales. It has recently been shown that warmer water in the Arctic has a negative impact on bowhead whales, forcing them to move further north, and hence limit their habitat. We propose regression models based on the uncommon Tweedie generalized linear models, with the duration at different sites as the response variable, allowing for true zeros in the data. The GPU computing allow us to fit our models with daily SST with the finest possible resolution of $0.083 \times 0.083^{\circ}$, approximately 10×10 km. There are around 9.6 million data points, hence we avoid losing information by downsampling data. Our results confirm that the bowhead whales prefer staying in the colder water, hence the warming Arctic is a tough challenge for their future.

Four additional contributions to the applied sciences, which are not included in this thesis, are cited below.

- Olsen, M.T., Nielsen, N.H., Biard, V., Teilmann, J., Ngô, M.C., Víkingsson, G., Gunnlaugsson, T., Stenson, G., Lawson, J., Lah, L. and Tiedemann, R., 2022. *Genetic and behavioural data confirm the existence of a distinct harbour porpoise ecotype in West Greenland*. Ecological Genetics and Genomics, 22, p.100108.

- Heide-Jørgensen, M.P., Blackwell, S.B., Tervo, O.M., Samson, A.L., Garde, E., Hansen, R.G., Ngô, M.C., Conrad, A.S., Trinhammer, P., Schmidt, H.C., Sinding, M.H.S. and Williams, T.M., 2021. *Behavioral response study on seismic airgun and vessel exposures in narwhals*. Frontiers in Marine Science, p.665.

- Tervo, O.M., Ditlevsen, S., Ngô, M.C., Nielsen, N.H., Blackwell, S.B., Williams, T.M. and Heide-Jørgensen, M.P., 2021. *Hunting by the stroke: How foraging drives diving behavior and locomotion of East-Greenland narwhals (Monodon monoceros)*. Frontiers in Marine Science, 7, p.1244.

- Heide-Jørgensen, M.P., Blackwell, S.B., Williams, T.M., Sinding, M.H.S., Skovrind, M., Tervo, O.M., Garde, E., Hansen, R.G., Nielsen, N.H., Ngô, M.C. and Ditlevsen, S., 2020. *Some like it cold: Temperature-dependent habitat selection by narwhals*. Ecology and evolution, 10(15), pp.8073-8090.

Chapter 2

Arctic whales

In this chapter, we include basic knowledge of narwhals and bowhead whales.

2.1 Narwhals

Narwhal (*Monodon monoceros*, meaning "one tooth and one horn") is a mysterious species, in the past and even in the present. It is the original version of the legendary unicorn: a white horse with a spiralled tusk from its head, but sometime depicted as a fish-liked monster with a horn [Heide-Jørgensen and Laidre, 2006](Figure 2.1). The tusk is a grown version of the upper left canine that is protruding through the upper lip. [Garde and Heide-Jørgensen, 2022] estimates there are 97% males with tusk, while it is only 1,5% in females. They also found that 0,9% of narwhal have two tusks. Male tusk can reach the length of 267 cm, while usually its length is approximately 190 cm [Heide-Jørgensen, 2018]. Female tusk is shorter, around 150 cm [Garde and Heide-Jørgensen, 2022]. The tusk is not for feeding purpose but rather for male sexual selection, even though aggressive behaviours have not been observed [Graham et al., 2020] (Figure 2.2).

While the scientific name comes from their tusk, the common name narwhal, i.e. corps-whale in Old Norse language, comes from their dark brown skin with mottled patterns and white patches [Heide-Jørgensen and Laidre, 2006]. The male is often significantly bigger than the female when fully grown: 400 cm and 900 kg for females, 450 cm and 1600 kg for males on average [Heide-Jørgensen, 2018]. Using eye lens ageing methods, their age has been estimated to a maximum of 115 years [Garde et al., 2007]. Mating happens in April and May with estimated gestation between 13 and 16 months, and the female is believed to give birth every three years [Heide-Jørgensen, 2009, 2018].

This paragraph summarises the content in [Heide-Jørgensen, 2009, 2018]. As a winter cetacean like bowhead whale and its relative the beluga whale (*Delphinapterus leucas*), using ARGOS and GPS-based tagging systems, we only can find narwhals in Arctic Ocean and North Atlantic Ocean. In Greenland, there are two big narwhal populations with significant genetic difference between the East and the West, due to the geographical separation [Louis et al., 2020]. They often migrate in small groups (5-10 whales) of adult males, or of females and their calves and sometimes joined with immature males. Narwhal is a deep diver, with a record of 1864m, only less than Cuvier's beaked whales (*Ziphius cavirostris*) and sperm whales (*Physeter macrocephalus*). The exact reason is unknown, but an assumption is that they prefer some specific preys that only found at the bottom layer. Stomach samples show that their food include fish and squid, including Greenland halibut (*Reinhardtius hippoglossoides*), Arctic cods (*Arctogadus glacialis and Boreogadus saida*), squids (*Gonatus sp.*), etc. The dive depth varies between seasons, < 500 m in summer and increases slowly during autumn and winter to reach depth > 800 m. At such depth > 100 m, there is almost no light, hence narwhals use biosonar for locating and capturing prey and various acoustics for communication with contubernals. The low

Figure 2.1: A group of narwhals in Hjørnedal, Scoreby Sound, East Greenland. Credit: Greenland Institute of Natural Resource.

Figure 2.2: Narwhal tagging and measurement in Hjørnedal, Scoreby Sound, East Greenland. Credit: Greenland Institute of Natural Resource.

rates between 300 Hz and 18 kHz are believed to be used for communication [Ford and Fisher, 1978, Miller et al., 1995], whereas faster click rates of 110-115 clicks per second are for feeding [Heide-Jørgensen, 2018]. The highest rate is around 48 kHz of 3-10 clicks per second [Stafford et al., 2012].

Narwhals have two natural predators [Heide-Jørgensen, 2018]. Killer whales attack narwhals in open-water seasons, while polar bears pull them out of water from sea ices in fjords [Lefort et al., 2020]. Inuits have been hunting narwhals for thousand years for meat, tusks, and skin (*mattak*) [Heide-Jørgensen and Laidre, 2006]. From arial surveys, the narwhal's population has been estimated to around 75,000 [Innes et al., 2002, Heide-Jørgensen, 2004] but more recent survey surveys suggest an even larger world abundance [Doniol-Valcroze et al., 2019]. Most of the world population is in uninhabited areas of the Canadian Arctic Archipelago [Heide-Jørgensen, 2018]. Between 2000-2004, the quotas for hunting were 535, 100, and 433 whales in West Greenland, East Greenland, and Canada, respectively [Heide-Jørgensen, 2009]. Hunting has in some areas caused a substantial decrease in the abundance of narwhals in several sub-populations. This is especially pronounced for the populations in Southeast Greenland. Hence, the quotas have been reduced (50 whales in Southeast Greenland in 2022) [Naalakkersuisut, 2021]. However, since 2019 the biologists at North Atlantic Marine Mammal Commission (NAMMCO) have recommended a zero quota for Southeast Greenland [NAMMCO, 2019]. Many biologists are concerned that with the current hunting, the narwhal population in Southeast Greenland will go extinct in 2028 [NAMMCO, 2021].

The loss of sea ice is another risk for narwhals. It may help them to be less likely to be captured by polar bear, but the longer open-water season increases the chance of being attacked by killer whales [Siegstad and Heide-Jørgensen, 1994, Williams et al., 2011]. The activities of humans are another threat to narwhals. Unlike the relatively human-friendly beluga whales, the narwhal is a shy and skittish species, hence they try to avoid humans as much as possible. They are very sensitive to noise, and much recent research have shown that anthropogenic factors are threating their existence [Williams et al., 2017, Heide-Jørgensen et al., 2021, Tervo et al., 2021].

2.2 Bowhead whales

The bowhead whale (*Balaena mysticetus*) is an Arctic baleen whale. It is also called the Greenland right whale, because it is the only existing species of the genus *Balaena*, closely related to the right whales (*Eubalaena*)(Figure 2.3). It is described by Darwin as "one of the most wonderful animal in the world" [Darwin, 2004]. It is the heaviest baleen whale species, only smaller than blue whale among the world of cetaceans [George et al., 2021]. Inuit hunters and Yankee whaling documented individuals of lengths exceeding 24.5 m and mass exceeding 172 tons, however modern scientific studies show that bowhead whales have maximum length of 17-19m, with an estimated body mass of up to 100 tons [George et al., 2021]. They are believed to be able to live longer than any other mammal species ($>$ 200 years using eye lens aging method) [George et al., 1999, Wetzel et al., 2017].

The bowhead whale is one of only three cetacean species endemic to Arctic and sub-Arctic waters. Three main known stocks of bowhead whales are Bering-Chukchi-Beaufort Seas, East Canada-West Greenland (ECWG), and the East Greenland-Svalbard-Barents Sea (EGSB) stock, as well as a smaller stock in Okhotsk Sea [Givens and Heide-Jørgensen, 2021]. A hypothesis from Corkeron & Connor is that the warm water is a more ideal

Figure 2.3: Tagging bowhead whale in West Greenland. Credit: Greenland Institute of Natural Resource.

habitat for the killer whale than cold water, hence staying in the cold Arctic helps the bowhead whale to stay away from such predator [Corkeron and Connor, 1999].

To adapt to the cold environment throughout the year, the dorsal blubber thickness of bowhead whales can reach 38.5 cm, and the skin thickness range is from 1 mm (eyelid) to 25 mm (lower jaw) [Haldiman et al., 1985]. Another characteristic is the low body temperature: the mean core temperature is 33.8°C (range 32.4°C - 35.3°C), several degrees lower than the other whales, which allows them to have exceptionally low metabolism rates [George, 2009, Lefebvre et al., 2016]. The large amount of blubber in bowhead whales exceeds what is necessary to maintain thermal homeostasis and could even lead to overheating [Hokkanen, 1990]. It is therefore likely that it, together with the low metabolic rate, blubber also functions as an energy depository that allow the whales to survive for long periods without feeding [Burns et al., 1993, George et al., 2020]. This may be particularly useful in the Arctic where production and zooplankton concentrations show large annual spatial variations [Pomerleau et al., 2017].

Bowhead whales mostly feed on small zooplankton species, including copepods, euphausiids, mysids, and amphipods [George, 2009]. While most feeding happens in the water column, sometimes bowhead whales feed on the seafloor as well as the surface [Lowry, 1993, Lowry et al., 2004]. It leads to an abundant list of species [Lowry, 1993, Sheffield and George, 2021]. The stomach contents and fecal samples show that their main prey are copepods (12 species, especially three species of calanoid copepods *Calanus hyperboreus, C. finmarchicus and C. glacialis*) and euphausiids (2 species) [George, 2009, Lefebvre et al., 2016]. Besides that, they may also occasionally feed on fish (e.g. Arctic cod and sculpins), benthic and epibenthic invertebrate species [George, 2009, Lefebvre et al., 2016, Sheffield and George, 2021].

It is believed that due to their main distribution in Arctic and sub-Arctic regions, bowhead whales are sensitive to the ongoing warming that is amplified in the Arctic with rapid reduction in sea ice and increasing sea temperatures [Alexander et al., 2018]. Beside sea temperature, sea ice is another important factor for filter feeding like bowhead whales because it affects the prey distribution seriously [Heide-Jørgensen et al., 2013]. It has been observed that currently the bowhead whales from the ECWG stock depart 1.5 to 3 weeks earlier to the period 1780-1837 because the warm water has been coming earlier [Laidre and Heide-Jørgensen, 2012, Eschricht and Reinhardt, 2018]. Therefore, a warming Arctic will likely reduce the suitable habitat for bowhead whales and force them to move into more northern and currently ice-covered areas [Chambault et al., 2018]. The first signs that bowhead whales today are found further north than in previous centuries comes from the Svalbard where they apparently have abandoned their historical range where they amount during the whaling period in the 16th and 17th centuries: today the bowhead whales from the EGSB stock are primarily found north of Svalbard, at the northeast corner of Greenland and around Frantz Josef Land, in areas that were not possible to navigate in the whaling days [Kovacs et al., 2020].

Chapter 3

Generalized linear models

In statistics, the linear model is one of the simplest models for the relationship between the outcome and the predictors: the outcome $y \in \mathbb{R}^p$ is simply the weighted sum of the predictors $x \in \mathbb{R}^{p \times n}$ for some positive integer p and number of datapoints n. Hence, it is the standard model in many fields of science, e.g., medicine, sociology, psychology and biology [Molnar, 2020]. However, in many cases the linear model is "too simple to be useful", due to the violation of assumptions. For instance, the outcome might not be a linear combination of predictors (the assumption of linearity), the outcome does not follow the normal distribution (the assumption of normality), or the variance of error terms is not constant (the assumption of homoscedasticity). Count data, data of proportions, or positive continuous data are among such examples. Necessarily, the model needs some generalization to overcome these limitations. One such generalization was introduced by Nelder and Wedderburn, the generalized linear model (GLM) [Nelder and Wedderburn, 1972]. In the next sections, we introduce GLMs and a specific example, the Tweedie GLMs. The following presentation in this chapter summarise the contents of [Dunn and Smyth, 2018a,b].

3.1 Introduction

As a regression model, two components of a GLM need to be determined:

- the random component is determined by choosing an appropriate probability distribution for the outcome.
- the systematic component is determined by choosing the link function g between the linear predictor $\eta = \beta_0 + \sum_{k=1}^{p} \beta_k x_k$ and the mean $\mu = \mathbb{E}[y]$: $\eta = g(\mu)$. A commonly used link function is the log-link, i.e. $\eta = \log(\mu)$, which leads to the class of log-linear models.

GLMs assume that the response variable follows a distribution from the family of distributions called the *exponential dispersion model* family (EDMs). The normal distribution (in linear regression model) and the gamma distribution are examples of continuous EDMs, while the Poisson distribution and (negative) binomial distribution (in logistic regression) belong to discrete EDMs.

Definition 3.1.1. *[Dunn and Smyth, 2018a] The probability function of the EDMs has the form*

$$\mathcal{P}(y; \theta, \phi) = a(y, \phi) \exp\left(\frac{y\theta - \kappa(\theta)}{\phi}\right) \tag{3.1.1}$$

- *θ is the canonical parameter,*
- *ϕ is the dispersion parameter,*

- $\kappa(\cdot)$ *is the cumulant function,*
- $a(\cdot,\cdot)$ *is the normalizing function ensuring that* $\int \mathcal{P}(y;\theta,\phi)dy = 1$ *if* y *is continuous, and* $\sum_y \mathcal{P}(y;\theta,\phi) = 1$ *if* y *is discrete.*

3.2 Mean and variance

The *moment generating function* (MGF) $M(y)$ is defined for all $t \in \mathbb{R}$ such that $M(t)$ exists:

$$M(t) = \mathbb{E}[e^{ty}] = \begin{cases} \displaystyle\int_S \mathcal{P}(y)e^{ty}\,dy & \text{for } y \text{ continuous} \\ \\ \displaystyle\sum_{y\in S} \mathcal{P}(y)e^{ty} & \text{for } y \text{ discrete,} \end{cases} \tag{3.2.1}$$

where $\mathcal{P}(y)$ is the probability density/mass function of y, and S is the support of y. We then define the *cumulant generating function* (CFG) $K(t) = \log(M(t))$. The $i-$th cumulant is defined as

$$\kappa_i = \left.\frac{d^i K(t)}{(dt)^i}\right|_{t=0} \tag{3.2.2}$$

The mean and variance are the first and second cumulants, respectively.

Now we derive the mean and variance of EDMs. Define $\theta_t = \theta + t\phi$. From (3.1.1) and (3.2.1), the MGF of an EDM is

$$\begin{aligned} M(t) &= \mathbb{E}[\exp(ty)] \\ &= \int_S \exp(ty)a(y,\phi)\exp\left(\tfrac{y\theta-\kappa(\theta)}{\phi}\right)dy \\ &= \exp\left(\tfrac{\kappa(\theta_t)-\kappa(\theta)}{\phi}\right)\int_S a(y,\phi)\exp\left(\tfrac{y\theta_t-\kappa(\theta_t)}{\phi}\right)dy \\ &= \exp\left(\tfrac{\kappa(\theta_t)-\kappa(\theta)}{\phi}\right)\textstyle\int_S \mathcal{P}(y;\theta_t,\phi)dy = \exp\left(\tfrac{\kappa(\theta_t)-\kappa(\theta)}{\phi}\right). \end{aligned}$$

Thus, $K(t) = \log(M(t)) = \frac{\kappa(\theta_t)-\kappa(\theta)}{\phi}$. Because the Taylor series of $\kappa(t)$ is

$$\kappa(\theta_t) = \kappa(\theta) + \frac{\kappa'(\theta)}{1!}(\phi t) + \frac{\kappa''(\theta)}{2!}(\phi t)^2 + \frac{\kappa'''(\theta)}{3!}(\phi t)^3 + \cdots,$$

then

$$K(t) = \frac{\kappa'(\theta)}{1!}t + \phi\frac{\kappa''(\theta)}{2!}t^2 + \phi^2\frac{\kappa'''(\theta)}{3!}t^3 + \cdots \tag{3.2.3}$$

The mean and variance of an EDM are thus

$$\begin{aligned} \mathbb{E}[y] &= \left.\frac{dK(t)}{dt}\right|_{t=0} = \kappa'(\theta) = \mu \\ \mathrm{Var}[y] &= \left.\frac{d^2K(t)}{(dt)^2}\right|_{t=0} = \phi\kappa''(\theta) = \phi V(\mu) \end{aligned} \tag{3.2.4}$$

where $V(\mu) = d\mu/d\theta$ is the *variance function*. Since variance $\mathrm{Var}[y]$ and ϕ are both positive, $V(\mu) > 0$. For example:

- Normal distribution: $\mathbb{E}[y] = \theta$ and $\mathrm{Var}[y] = \sigma^2$.
- Poisson distribution: $\mathbb{E}[y] = \mathrm{Var}[y] = \mu$ since $\phi = 1$.

3.3 Deviance and dispersion model form

As mentioned above, $V(\mu) = d\mu/d\theta > 0$, therefore μ is a monotonely increasing function of θ. Thus, there exists an injective map between μ and θ. So we can replace the canonical parameter θ by the mean of the probability function μ in the representation of $\mathcal{P}(y; \phi, \mu)$. It allows for an easier interpresentation of the EDMs.

To do that, define $T(y, \mu) = y\theta - \kappa(\theta)$. Hence

$$\begin{aligned} \frac{\partial T(y,\mu)}{\partial \theta} &= y - \frac{\mathrm{d}\kappa(\theta)}{\mathrm{d}\theta} = y - \mu \\ \frac{\partial^2 T(y,\mu)}{\partial \theta^2} &= -\frac{\mathrm{d}\mu}{\mathrm{d}\theta} = -V(\mu) < 0. \end{aligned} \tag{3.3.1}$$

The second derivative of T is always negative, thus T is concave. The maximum of T is for $\mu = y$. The *unit deviance*, or the distance between μ and y, is then defined as

$$d(y, \mu) = 2\left(T(y, y) - T(y, \mu)\right) \geq 0. \tag{3.3.2}$$

Finally, we can use the unit deviance to represent the probability function in the *dispersion model form*:

$$\mathcal{P}(y; \mu, \phi) = b(y, \phi) \exp\left(-\frac{1}{2\phi} d(y, \mu)\right) \tag{3.3.3}$$

where $b(y, \phi) = a(y, \phi) \exp\left(T(y, y)/\phi\right)$.

The unit deviance is also used in fitting GLMs in R, under the function `glm.fit` [R Core Team, 2021]. The *residual deviance* (or the deviance function), defined as the (weighted) sum of the unit deviances,

$$D(y, \mu) = \sum_{i=1}^{n} w_i d(y_i, \mu_i), \tag{3.3.4}$$

where w_i is the $i-$th weight. It is used to find the maximum log-likelihood of GLMs in R [R Core Team, 2021].

3.4 Tweedie distribution

In this chaater, we introduce the Tweedie distributions, used in the paper III. The Tweedie distribution was first introduced by Maurice Tweedie [Tweedie, 1946]. It is identical to Taylor's law in empirical ecology, based on a power-law relationship [Taylor, 1961], which has been used in many practical applications in ecology, e.g. in [Taylor, 1961, Anderson et al., 1982, Kendal, 2002, 2004b].

The Tweedie EDMs' variance function has the form $V(\mu) = \mu^p$, where p is the power parameter. It is a generalisation of many well-known distributions:

- $p = 0$: the normal distribution has $V(\mu) = 1$,
- $p = 1$: the Poisson distribution has $V(\mu) = \mu$,
- $p = 2$: the Gamma distribution has $V(\mu) = \mu^2$.

Now we compute the cumulant function κ from both the canonical parameter θ and μ. Given $V(\mu) = d\mu/d\theta = \mu^p$, hence $d\theta/d\mu = \mu^{-p}$, so

$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p} & \text{for } p \neq 1 \\ \log(\mu) & \text{for } p = 1. \end{cases} \tag{3.4.1}$$

Equation (3.4.1) is indeed the *canonical link* for Tweedie EDMs. Therefore,

$$\mu = \begin{cases} ((1-p)\theta)^{\frac{1}{1-p}} & \text{for } p \neq 1 \\ \exp(\theta) & \text{for } p = 1. \end{cases} \tag{3.4.2}$$

From (3.2.4) where $\kappa'(\theta) = \mu$, given $\alpha = \frac{p-2}{p-1}$, we then obtain

$$\kappa(\theta) = \begin{cases} \frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^{\alpha} & \text{for } p \neq 1, 2 \\ -\log(-\theta) & \text{for } p = 2 \\ \exp(\theta) & \text{for } p = 1. \end{cases} \tag{3.4.3}$$

Now we calculate κ based on μ. We have

$$\frac{d\kappa}{d\mu} = \frac{d\kappa}{d\theta} \cdot \frac{d\theta}{d\mu} = \mu \cdot \mu^{-p} = \mu^{1-p},$$

then

$$\kappa(\mu) = \begin{cases} \frac{\mu^{2-p}}{2-p} & \text{for } p \neq 2 \\ \log(\mu) & \text{for } p = 2. \end{cases} \tag{3.4.4}$$

We are interested in computing the unit deviance from y and μ, as it is used to fit GLMs in `glm.fit` in R. Given $T(y, \mu) = y\theta - \kappa(\mu)$, (3.4.1), and (3.4.4), we have

$$T(y, \mu) = \begin{cases} y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p} & \text{for } p \neq 1, 2 \\ y\log(\mu) - \mu & \text{for } p = 1 \\ -y/\mu - \log(\mu) & \text{for } p = 2 \end{cases} \tag{3.4.5}$$

Then the unit deviance is

$$d(y, \mu) = \begin{cases} 2\left(\frac{y^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\right) & \text{for } p \neq 1, 2 \\ 2\left(y\log(y/\mu) - y + \mu\right) & \text{for } p = 1 \\ 2\left(y/\mu - \log(y/\mu) - 1\right) & \text{for } p = 2 \end{cases} \tag{3.4.6}$$

There does not exist EDMs where $0 < p < 1$, see Theorem 2 [Jørgensen, 1987]. The case of $1 < p < 2$ corresponds to the compound Poisson Gamma distribution [Delong et al., 2021]. The random variable Y following this distribution if $Y = \sum_{i=1}^{N} X_i$ where

the random variables X_i following Gamma distributions, and $N \sim \text{Poisson}(\lambda)$ for some positive λ [Ross, 2014]. When $N = 0$ is $Y = 0$, therefore the compound Poisson Gamma distribution allows for exact zero observations, as well as for positive continuous data when $N > 0$. It is suitable for our model of bowhead whale distributions in Paper III. It has been used in several applications, for example in biology [Kendal, 2004a, 2007], fisheries research [Foster and Bravington, 2013, Shono, 2008], insurance modelling [Jørgensen and Paes De Souza, 1994, Smyth and Jørgensen, 2002], and meteorology [Revfeim, 1984, Thompson, 1984].

Chapter 4

Hidden Markov Models

In a dynamical system, a state is a set of variables describing the system at some specific time [Morrissey, 2021]. An ecological system is an example of a dynamical system. The set of all possible states is denoted the state space [Terman and Izhikevich, 2008]. Hidden Markov Models (HMMs), also called state space models, are a class of models of dynamical systems that are often natural for describing ecological systems at all levels, from the smallest scale such as individuals, to the largest scale such as the entire ecosystem [McClintock et al., 2020]. It is a Markovian model that takes into account the temporal dependence in the system: the current state depends on the state at the previous time point. It is not as simple as the (generalized) linear model, where observations are assumed independent. By including temporal dependencies characterising the dynamics of the system, HMMs are often more suitable in many applications. Especially in ecology, [McClintock et al., 2020] list many studies involving HMMs at different levels: individual, population, community, and ecosystem. In the next sections, we introduce HMMs and its application to a narwhal dive dataset [Ngô et al., 2019].

4.1 Introduction

First, we introduce the definition of a simple discrete-time HMM (Figure 4.1).

Definition 4.1.1. *[Zucchini et al., 2016] Let $(C_t)_{t\geq 0}, (X_t)_{t\geq 0}$ be stochastic processes, $t \in \mathbb{N}$, and $m \in \mathbb{N}$. $(C_t, X_t)_{t\geq 0}$ is a hidden Markov model if:*

- $\Pr\left(X_t | C^{(t)}, X^{(t-1)}\right) = \Pr(X_t|C_t)$
- $\Pr\left(C_t | C^{(t-1)}\right) = \Pr(C_t|C_{t-1})$
- $C_t \in \{1, \ldots, m\}$ *for all* $t \geq 0$

where $X^{(t)} = \{X_1, X_2, \ldots, X_t\}$ and $C^{(t)} = \{C_1, C_2, \ldots, C_t\}$ for $t \geq 1$. $(C_t)_{t\geq 0}$ is called the (unobserved) parameter process*, while $(X_t)_{t\geq 0}$ is called the* state-dependent process*.*

Let $\omega_{ji}(t) = \Pr(C_{t+1} = j | C_t = i)$, then we define the transition probability matrix $\Omega(t)$:

$$\Omega(t) = \begin{bmatrix} \omega_{11}(t) & \ldots & \omega_{1m}(t) \\ \vdots & \ddots & \vdots \\ \omega_{m1}(t) & \ldots & \omega_{mm}(t) \end{bmatrix} \tag{4.1.1}$$

where $\omega_{ij} \geq 0$ and $\sum_{j=1}^{m} \omega_{ij} = 1$ for $i \in \{1, \ldots, m\}$.

There are many ways to extend the simple HMMs in Figure 4.1. One way is implemented in our work [Ngô et al., 2019]. In order to include a longer memory, we let the past state-dependent processes influence the hidden process (Figure 4.2).

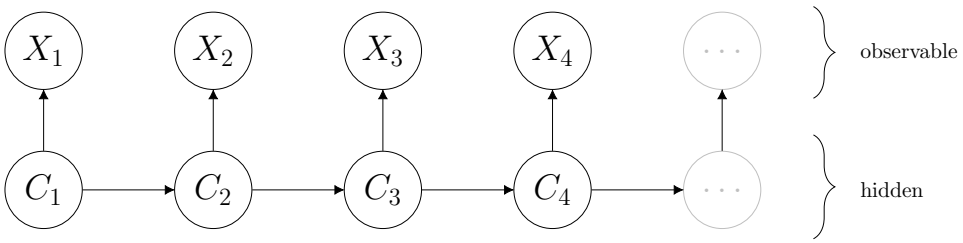

Figure 4.1: Hidden Markov Model.

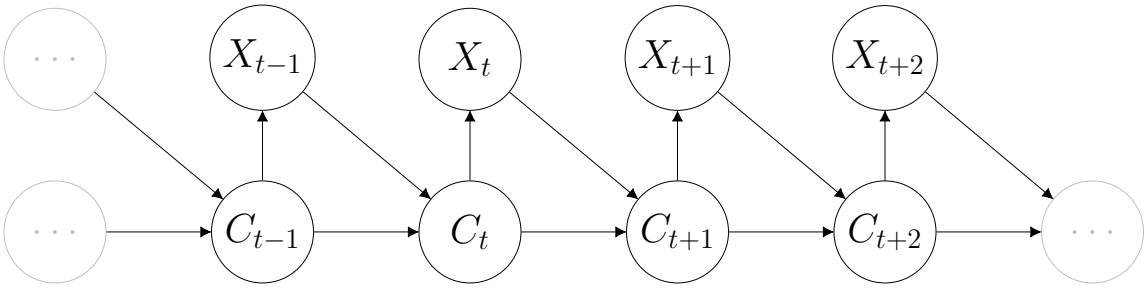

Figure 4.2: Hidden Markov Model with feedback processes, adapted from [Ngô et al., 2019].

4.2 Contemporaneous conditional independence relaxation

Another assumption that is often used is the *contemporaneous conditional independence* [Zucchini et al., 2016, Ngô et al., 2019]. If $(X_t)_{t\geq 0}$ is a multivariate process of dimension p: $X_t = (X_{1,t}, \ldots, X_{p,t})$, it states that

$$\Pr(X_t|C_t = i) = \prod_{k=1}^{p} \Pr(X_{k,t}|C_t = i). \tag{4.2.1}$$

for some $i \in \{1, \ldots, m\}$. This assumption is often assumed in multidimensional time series, because it is often not easy to model the correlation between state variables at the same time point. In [Ngô et al., 2019], we relax this assumption, to take into account the correlations between the three variables maximum depth (MD), dive duration (DT), and post-dive surface time (PD) in the narwhal dive data. It leads to significant improvements of the model fit. We will explain the model in details in the following sections, assuming the variables follow the Log-normal distribution and the Gamma distribution.

4.2.1 Correlated log-normal distribution

The random variable X follows a log-normal distribution if its logarithm $\log(X)$ follows a normal distribution. For a bivariate log-normal distribution [Aitchison and Brown, 1957], denote $Y = (Y_{1,t}, Y_{t,2}) = (\log(X_{1,t}), \log(X_{2,t}))$. Its probability density function f of the observation (x_1, x_2) is

$$f(x_1, x_2) = \left(2\pi x_1 x_2 \sigma_{Y_{1,t}} \sigma_{Y_{2,t}} \sqrt{1-\rho^2} \exp\left(\frac{a^2 + b^2 - 2\rho ab}{2\left(1-\rho^2\right)}\right)\right)^{-1},$$

where $x_1, x_2 > 0$, the product-moment correlation of $Y_{1,t}$ and $Y_{2,t}$ satisfies $-1 < \rho < 1$, while $\mu_{Y_{i,t}}$ and $\sigma_{Y_{i,t}} (i = 1, 2)$ are the mean and standard deviation of $Y_{i,t}$, and $a = \frac{\log x_1 - \mu_{Y_{1,t}}}{\sigma_{Y_{1,t}}}, b = \frac{\log x_2 - \mu_{Y_{2,t}}}{\sigma_{Y_{2,t}}}$. Then

$$\sigma_{Y_{i,t}} = \log\left(1 + \frac{\sigma^2_{X_{i,t}}}{\mu^2_{X_{i,t}}}\right)^{1/2}, \qquad \mu_{Y_{i,t}} = \log\left(\mu_{X_{i,t}}\right) - \frac{\sigma_{Y_{i,t}}}{2}.$$

We will now explain the trivariate log-normal distribution used in [Ngô et al., 2019]. Given $X_t = (X_{1,t}, X_{2,t}, X_{3,t})$ and $C_t = i$ for some $i \in \{1, \ldots, m\}$, let the mean and variance of $\log(X_{k,t})$ be μ_i^k and $(\sigma_i^k)^2$, respectively. Thus, the mean and variance of $X_{k,t}$ are $\exp(\mu_i^k + (\sigma_i^k)^2/2)$ and $(\exp((\sigma_i^k)^2) - 1)\exp(2\mu_i^k + (\sigma_i^k)^2)$, respectively. Denote $\rho_i^{k_1,k_2}$ the correlation between $\log(X_{k_1,t})$ and $\log(X_{k_2,t})$, where $k_1, k_2 \in \{1, 2, 3\}$, then the correlation between $X_{k_1,t}$ and $X_{k_2,t}$ is

$$\frac{\exp(\rho_i^{k_1,k_2}\sigma_i^{k_1}\sigma_i^{k_2}) - 1}{\sqrt{(\exp((\sigma_i^{k_1})^2) - 1)(\exp((\sigma_i^{k_2})^2) - 1)}}.$$

For some small $x \in \mathbb{R}$, $\exp(x) - 1 \approx x$. Hence, if $\sigma_i^{k_1}, \sigma_i^{k_2}$ are small, the correlation between $X_{k_1,t}$ and $X_{k_2,t}$ is aproximately

$$\frac{\exp(\rho_i^{k_1,k_2}\sigma_i^{k_1}\sigma_i^{k_2}) - 1}{\sqrt{(\exp((\sigma_i^{k_1})^2) - 1)(\exp((\sigma_i^{k_2})^2) - 1)}} \approx \frac{\rho_i^{k_1,k_2}\sigma_i^{k_1}\sigma_i^{k_2}}{\sqrt{\left(\sigma_i^{k_1}\right)^2\left(\sigma_i^{k_2}\right)^2}} = \rho_i^{k_1,k_2}.$$

Then, the state dependent probability density functions are:

$$f_i(X_t) \approx \frac{1}{(2\pi)^{3/2}\sqrt{|\Sigma_i|}\prod_{k=1}^3 \log(X_{k,t})} \exp\left(-\frac{1}{2}(\log X_t - \mu_i)^\top \Sigma_i^{-1}(\log X_t - \mu_i)\right), \tag{4.2.2}$$

where the covariance matrix Σ_i is

$$\Sigma_i = \begin{bmatrix} (\sigma_i^1)^2 & \rho_i^{12}\sigma_i^1\sigma_i^2 & \rho_i^{13}\sigma_i^1\sigma_i^3 \\ \rho_i^{12}\sigma_i^1\sigma_i^2 & (\sigma_i^2)^2 & \rho_i^{23}\sigma_i^2\sigma_i^3 \\ \rho_i^{13}\sigma_i^1\sigma_i^3 & \rho_i^{23}\sigma_i^2\sigma_i^3 & (\sigma_i^3)^2 \end{bmatrix}, \tag{4.2.3}$$

$|\cdot|$ denotes the determinant of a matrix, $\mu_i = (\mu_i^1, \mu_i^2, \mu_i^3)^\top$, and $\rho_i^{12}, \rho_i^{13}, \rho_i^{23}$ are the correlation coefficients between the three components for $C_t = i$, respectively. Hence

$$\begin{aligned} |\Sigma_i| &= \left(\sigma_i^1\sigma_i^2\sigma_i^3\right)^2\left(1 + 2\rho_i^{12}\rho_{13}\rho_i^{23} - \rho_{12}^2 - (^2 - (\rho_i^{23})^2\right) \\ &= \left(\sigma_i^1\sigma_i^2\sigma_i^3\right)^2\left((1 - (\rho_i^{12})^2)(1 - (\rho_i^{13})^2) - (\rho_i^{12}\rho_i^{13} - \rho_i^{23})^2\right). \end{aligned}$$

To ensure that the covariance matrix Σ_i is positive definite (so the density exists), all of its principal components need to be positive following Sylvester's condition [Horn and Johnson, 1985], i.e. $(\sigma_i^1)^2 > 0$, $(1 - (\rho_i^{12})^2)(\sigma_i^1)^2(\sigma_i^2)^2 > 0$, and

$$\begin{aligned} & 1 + 2\rho_i^{12}\rho_i^{13}\rho_i^{23} - (\rho_i^{12})^2 - (\rho_i^{13})^2 - (\rho_i^{23})^2 > 0 \\ \iff \quad & \rho_i^{12}\rho_i^{13} - \sqrt{\Delta} < \rho_i^{23} < \rho_i^{12}\rho_i^{13} + \sqrt{\Delta}, \end{aligned}$$

where $\Delta = (1-(\rho_i^{12})^2)(1-(\rho_i^{13})^2)$. It is equivalent to $\sigma_i^1, \sigma_i^2 > 0$, $-1 < \rho_i^{12} < 1$, and $\rho_i^{23} = \rho_i^{12}\rho_i^{13} + \alpha\sqrt{\Delta}$ for some $\alpha \in \mathbb{R}$ such that $-1 < \alpha < 1$.

One can determine Σ_i^{-1} by using cofactors following Cramer's rule. Denote $A = 1 + 2\rho_i^{12}\rho_i^{13}\rho_i^{23} - (\rho_i^{12})^2 - (\rho_i^{13})^2 - (\rho_i^{23})^2$, the cofactors of Σ_i are

$$
\begin{aligned}
C_{11} &= \left(\sigma_i^2\sigma_i^3\right)^2 \left(1-(\rho_i^{23})^2\right)/|\Sigma_i| = \frac{1-(\rho_i^{23})^2}{\left(\sigma_i^1\right)^2}\cdot\frac{1}{A} \\
C_{22} &= \left(\sigma_i^1\sigma_i^3\right)^2 \left(1-(\rho_i^{13})^2\right)/|\Sigma_i| = \frac{1-(\rho_i^{13})^2}{\left(\sigma_i^2\right)^2}\cdot\frac{1}{A} \\
C_{33} &= \left(\sigma_i^1\sigma_i^2\right)^2 \left(1-(\rho_i^{12})^2\right)/|\Sigma_i| = \frac{1-(\rho_i^{12})^2}{\left(\sigma_i^3\right)^2}\cdot\frac{1}{A} \\
C_{12} = C_{21} &= \sigma_i^1\sigma_i^2\left(\sigma_i^3\right)^2 \left(\rho_i^{13}\rho_i^{23}-\rho_i^{12}\right)/|\Sigma_i| = \frac{\rho_i^{13}\rho_i^{23}-\rho_i^{12}}{\sigma_i^1\sigma_i^2}\cdot\frac{1}{A} \\
C_{13} = C_{31} &= \sigma_i^1\left(\sigma_i^2\right)^2\sigma_i^3 \left(\rho_i^{12}\rho_i^{23}-\rho_i^{13}\right)/|\Sigma_i| = \frac{\rho_i^{12}\rho_i^{23}-\rho_i^{13}}{\sigma_i^1\sigma_i^3}\cdot\frac{1}{A} \\
C_{23} = C_{32} &= \left(\sigma_i^1\right)^2\sigma_i^2\sigma_i^3 \left(\rho_i^{12}\rho_i^{13}-\rho_i^{23}\right)/|\Sigma_i| = \frac{\rho_i^{12}\rho_i^{13}-\rho_i^{23}}{\sigma_i^2\sigma_i^3}\cdot\frac{1}{A}
\end{aligned}
$$

hence

$$
\Sigma_i^{-1} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}.
$$

4.2.2 Correlated gamma distribution

For the correlated gamma models in [Ngô et al., 2019], we only take into account the correlation between Maximum Depth and Dive Duration, and assume that there is no correlation between these two variables and the Post-dive duration. So we only discuss here the case of a bivariate gamma distribution. [Moran, 1969] introduces a numerical method to calculate the bivariate gamma distribution based on the normal distribution, that is used in our work [Ngô et al., 2019]. We follow the notions of [Yue et al., 2001]. Given $X_1 \sim \text{Gamma}(\alpha_1, \lambda_1), X_2 \sim \text{Gamma}(\alpha_2, \lambda_2)$ where α_i, λ_i are scale and shape parameters of the Gamma distributions for $i \in \{1, 2\}$, denote g_i, G_i the probability density and cumulative distribution functions of X_i, respectively. Given $\Phi(\cdot)$ the cumulative distribution function of the standard normal distribution, we define the normalized random variables X_1' and X_2' of X_1 and X_2, using the normal quantile transform (NQT), as

$$
\begin{aligned}
X_1' &= \Phi(G_1(\cdot; \alpha_1, \lambda_1)) \\
X_2' &= \Phi(G_2(\cdot; \alpha_2, \lambda_2))
\end{aligned}
$$

Denote ρ_{12} the correlation coefficient between X_1' and X_2'. The bivariate gamma joint density distribution between X_1 and X_2 is then

$$
g_{12}(x_1, x_2) = \frac{g_1(x_1)g_2(x_2)}{\sqrt{1-\rho_{12}^2}} \exp\left(-\frac{(\rho_{12}x_1')^2 + (\rho_{12}x_2')^2 - 2\rho_{12}x_1'x_2'}{1-\rho_{12}^2}\right).
$$

The Moran model above is one of several different methods to compute the bivariate gamma density distribution. Compared to other methods, it has several advantages: it allows the correlation coefficient to vary completely between -1 and 1, rather than to be

limited between $-1/3$ and $1/3$ like in the Farlie-Gumbel-Morgensen (FGM) model [Farlie, 1960, Gumbel, 1958, Morgenstern, 1956]. The density function can be calculated by numerical methods based on the standard normal distribution which is available in several languages (e.g. function `pnorm` in R), while the Izawa bigamma model [Izawa, 1953] and the Smith-Adelfang-Tubbs (SAT) models [Smith et al., 1982] require the computation of infinite series. For the other models, see [Yue et al., 2001] for details. Its generalized form, the bivariate meta-Gaussian model, allows arbitrary continuous marginal distributions that make its applications much broader [Kelly and Krzysztofowicz, 1997].

Chapter 5

Deep learning

In the last decade, the terminologies "Artificial intelligence" (AI) and "Deep learning" (DL) are mentioned everywhere, from media to academia. Designing a machine that can act like humans, or even more intelligent than humans, has been a dream for thousands of years, and many people feel that it is coming closer. Besides hypes which always exist in "hot technologies", we cannot deny that there are many breakthroughs in research and automation applications due to deep learning. There are several applications in which AI performs better than all or most humans, mainly in fields having clear rules, such as chess (Deep Blue) [Krauthammer, Charles, 1997], Go (AlphaGo/AlphaZero) [Silver et al., 2017], computer games [Vinyals et al., 2019], poker [Brown and Sandholm, 2018, Blair and Saffidine, 2019], and other applications [Wani et al., 2020]. It is also very strong in image recognition, on-par with humans in image classification, character recognition, etc. In many other fields, it approaches human performance, e.g. the GPT-3 language model can write code, poetry, and text in a similar way to humans [Brown et al., 2020]. In biology, AI models can predict the protein fold structures of humans and many animals which is helpful for drug designs [Jumper et al., 2021]. Even in the most abstract field like mathematics, recently some mathematicians discovered new patterns with the help of AI techniques [Davies et al., 2021].

Most of these recent breakthroughs of AI are based on DL. With the huge number of parameters, DL models can fit to many kinds of data. They outperform many past state-of-the-art methods like support vector machines, random forests, gradient boosting, and hidden Markov models in many practical applications. However, we have poor understanding of how DL works. DL models can easily overfit to data, but their performance are very impressive, contradicting classical statistical theories. In the next sections, we introduce the background of DL and how it is applied in our work in [Ngô et al., 2021].

5.1 Some concepts of machine learning

We review briefly some concepts of machine learning (ML), of which DL is a member. Similar to the concept of learning in humans, we are interested in letting ML algorithms learn to do some tasks by themselves. Given a task, in order to learn, we humans need some inputs, for ML algorithms it is training data of that task. If ML algorithms learn from labelled data, often labelled by humans, then it is denoted supervised learning, while letting the algorithms explore the dataset to find patterns or structures is called unsupervised learning [Goodfellow et al., 2016]. If the algorithm "collects" data by itself by interacting with environment, it is called reinforcement learning [Sutton and Barto, 2018].

When learning, we need to practice with exercises and examinations, for ML algorithms this corresponds to validation data and testing data, respectively. Specifically, ML algorithms, with some given structure with basic components, often have some fixed

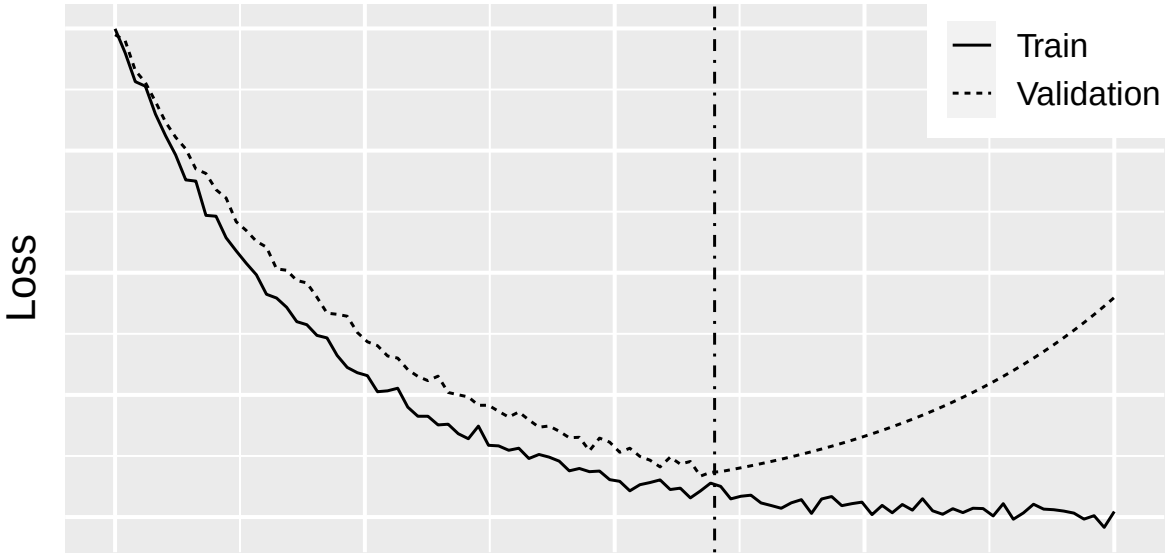

Figure 5.1: The training and validation losses. The vertical dashed line indicates when overfitting happens, i.e. the validation loss starts increasing.

hyper-parameters, which need to be specified, normally by humans. The number of hyper-parameters are often much smaller than the number of parameters of the model that needs to be trained. Validation set helps us to choose such hyper-parameters, called *hyper-parameter tuning*. It is especially important for DL to obtain good performance, because since the model has a huge number of parameters, it is easy to overfit. Test set evaluates the performance of the ML algorithm after tuning, by measuring the difference between the output of the model and the data. For different tasks, we need different measures, called *loss functions*. For example, in regression we use mean square error, or L^1 loss, etc.; while in classification, we can use for example cross entropy or Kullback–Leibler divergence [Kullback and Leibler, 1951].

When training DL, with huge amounts of data needed for good performance, we need to let the algorithm go though the data many times. We want to reduce the loss as much as possible. DL models often fit very well to the training data, but unfortunately, also to the noise if let it run for too long. We therefor need the validation data set to measure the loss on a different data set. Once the loss on the validation data stops decreasing and start increasing, we stop the algorithm. It is called *early stopping* to avoid model overfitting (Figure 5.1) [Prechelt, 1998].

5.2 Introduction to deep learning

The following exposition is summarised from [Calin, 2020]. The DL unit element is an artificial neuron, which is based on biological neurons. A biological neuron has three basic components: dendrites, an axon, and a body cell. A neuron cell collects signals x_i's from the other neurons through its dendrites. These signals are scaled by weights w_i's at the synapses of the dendrites, hence the total signal is $\sum_i w_i x_i$ inside the body cell. The cell transmits this total signal if it is greater than a threshold b, called *bias*.

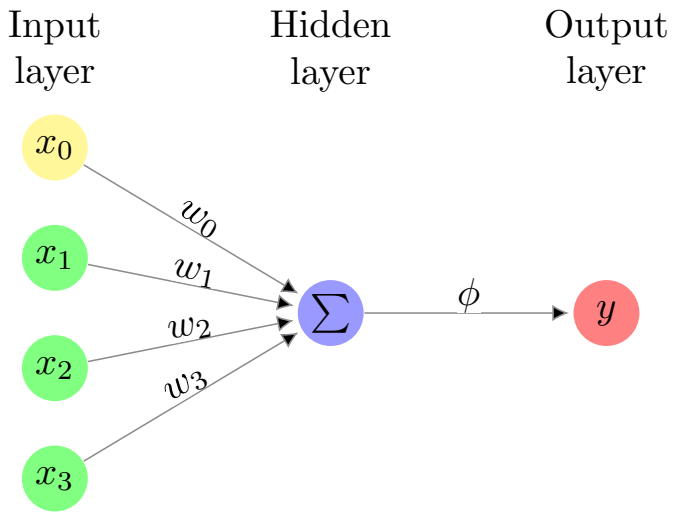

Figure 5.2: An abstract neuron with the activation function ϕ, input x, output y and weights w, such that $y = \phi \sum_{i=0}^{3} w_i x_i$. x_0 denotes the bias.

Mathematically, the output signal is

$$y(x, w) = \begin{cases} 0, & \text{if } \sum_{i=1} w_i x_i \leq b \\ 1, & \text{if } \sum_{i=1} w_i x_i > b, \end{cases}$$

given $x = (x_1, \ldots, x_n)$ and $w = (w_1, \ldots, w_n)$. It can be written in a more formal way, using activation function ϕ. In the case above, ϕ is the Heaviside function

$$y(x, w) = \phi\left(\sum_{i=0} w_i x_i\right) = \begin{cases} 0, & \text{if } \sum_{i=0} w_i x_i \leq 0 \\ 1, & \text{if } \sum_{i=0} w_i x_i > 0, \end{cases}$$

if we define $w_0 = -b$ and $x_0 = -1$.

Definition 5.2.1. *[Calin, 2020] An abstract neuron is a quadruple (x, w, ϕ, y), where $x^\top = (x_0, x_1, \ldots, x_n)$ is the input vector, $w^\top = (w_0, w_1, \ldots, w_n)$ is the weight vector, with $x_0 = -1$ and $w_0 = b$, the bias, and ϕ is an activation function that defines the outcome function $y = \phi(x^\top w) = \phi(\sum_i^n w_i x_i)$. The goal of the step function ϕ is to introduce the capacity of modelling non-linearities.*

One of the most well-known examples of this model is the *perceptron* [Rosenblatt, 1957], where the inputs are binary: $x_i \in \{0, 1\}$ for all i, and ϕ is the Heaviside function. It is the origin of multilayer neural network, or deep learning. As a simple model, it can only model simple data, e.g. a linear separable classification problem, i.e. there exists a hyperplane to separate the patterns in the data. To learn these patterns, the perceptron tunes the weights w_i according to the inputs x_i to determine the hyperplane that separates the patterns. However, if such hyperplane does not exist, it takes at least a two-layer perceptron neural network, published in 1969 in the book "Perceptrons: an introduction to computational geometry" [Minsky and Papert, 2017]. For example, one perceptron can learn (or model) the `AND` and `OR` logical function, but not the `XOR` function, i.e. produce the outputs from its inputs. This limitation is one of the reasons leading to the first AI crisis in 1970s, due to the misunderstanding that similar to a single perceptron,

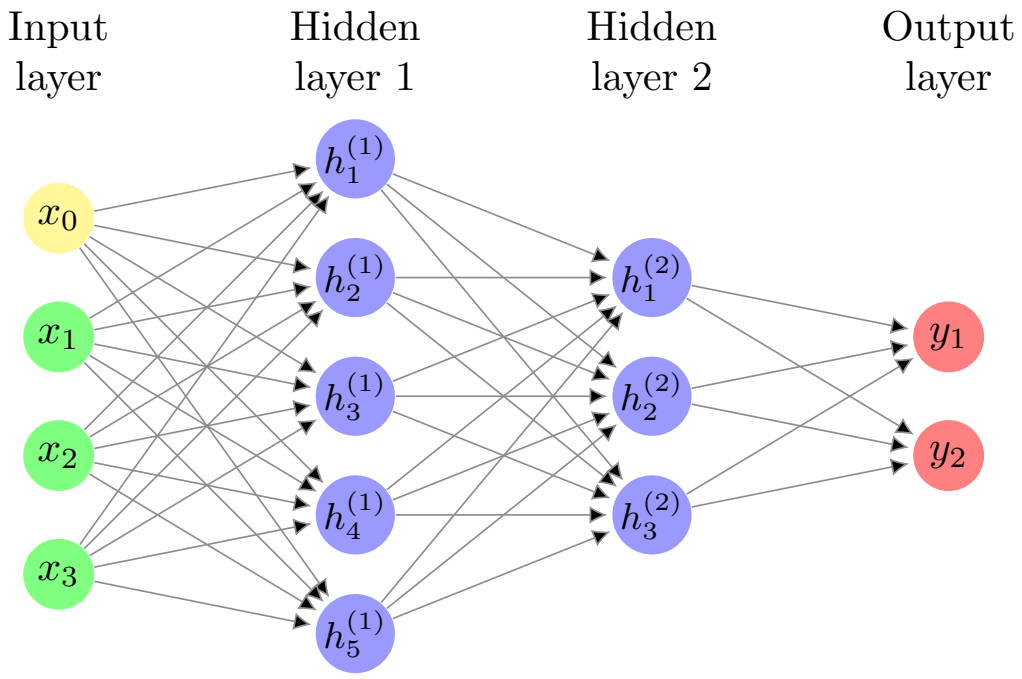

Figure 5.3: A 2-layer neural network, with input x, output y and hidden layers $h^{(1)}$ and $h^{(2)}$. x_0 denotes the bias.

the multilayer network is not able to learn non-linear separable patterns [Crevier, 1993] either. In a few cases, the neural network can learn the exact outputs, why in most of the cases when the data are too complex, using neural networks as an approximator is often good enough.

Now we explain why a single perceptron cannot learn the `XOR` function by an algebraic method [Calin, 2020]. Recall the definition of the `XOR` function

$$y := x_1 \oplus x_2 = \begin{cases} 0, & \text{if } x_1 = x_2 \\ 1, & \text{if } x_1 \neq x_2, \end{cases} \tag{5.2.1}$$

where $x_1, x_2 \in \{0, 1\}$. Assume that there exists w_1, w_2, b such that

$$y(x, w) = \phi\left(w_1 x_1 + w_2 x_2 - b\right) = \begin{cases} 0, & \text{if } w_1 x_1 + w_2 x_2 \leq b \\ 1, & \text{if } w_1 x_1 + w_2 x_2 > b, \end{cases}$$

where x_1, x_2, y satisfy 5.2.1. Since $y = 1$ if $x_1 \neq x_2$, hence $w_1, w_2 > b$. And $y = 0$ if $x_1 = x_2$, then $0 \leq b$ and $w_1 + w_2 \leq b$. Therefore, $2b < w_1 + w_2 \leq b \Rightarrow b < 0$, which contradicts that $0 \leq b$.

To boost the power of perceptrons or more general abstract neurons, one way is to make a network by stacking many layers, where each layer have several neurons, see Figure 5.3 for a 2-layer neural network. The layers $h^{(1)}$ and $h^{(2)}$ between input and output are called hidden layers. No feedbacks and information traverses back tothe input layer from the output layer, it is also called *feed-forward neural network*.

Now we show that the network of a 2-layer perceptron can learn the `XOR` function [Calin, 2020]. The first layer has two perceptrons $p_1^1 : (w_{11}^1, w_{12}^1, b^{11}) = (1, 1, 0.5)$ and $p_2^1 : (w_{21}^1, w_{22}^1, b^{12}) = (1, 1, 1.5)$. The second layer has one perceptron: $p^2 : (w_{21}^2, w_{22}^2, b^2) = (1, -1, 0.5)$. It can be verified easily that 5.2.1 is satisfied. It shows the potential of multilayer neural networks to overcome the limitations of a single perceptron.

Beside the perceptron using the Heaviside step function, choosing other activation functions create different types of artificial neurons. For example, an approximator of a perceptron is a sigmoid neutron where the activation function is a *sigmoidal function* [Calin, 2020].

Definition 5.2.2. *[Calin, 2020] A function $\phi : \mathbb{R} \to [0, 1]$ is sigmoidal if*

- $\lim_{x\to-\infty} \phi(x) = 0$,
- $\lim_{x\to\infty} \phi(x) = 1$.

One well-known example is the logistic function $\phi : \ \phi(x) = (1 + \exp(-x))^{-1}$, which leads to logistic regression. Other functions in this class is arctangent function $\frac{2}{\pi}\arctan(x)$, softside function $\frac{x}{1+|x|}$, or hyperbolic tangent $\frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$ [Calin, 2020]. These activation functions allow a continuous output between $(0, 1)$, rather than just discrete output as step functions. It is used in classification problems, allowing the probability of predicted classes to vary between 0 and 1.

Another important class of activation functions, when the output is continuous and not double-bounded, is the *hockey-stick function*, having the L-shape [Calin, 2020]. The simplest and most well-known is the Rectified Linear Unit (ReLU) function

$$y := \max(x, 0) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } x \geq 0. \end{cases} \tag{5.2.2}$$

Recently, it has become the *de-facto* activation function due to its simplicity [Ramachandran et al., 2017], especially in image recognition [Krizhevsky et al., 2017]. Another advantage is that it does not saturate, i.e. the output could come too close to the bounds like in sigmoidal functions. Other variants include Parametric ReLU and Exponential linear unit (ELU), but ReLU is used the most because it is fastest, several times faster than neural network using sigmoidal activation functions, while still having good performance [Krizhevsky et al., 2017]. When ϕ is the identity function $\phi(x) = x$, we have linear regression. For other classes of activation functions, see [Calin, 2020].

The above example of a 2-layer neural network which can learn the `XOR` function show the potential ability of multilayer neural network to learn complex functions. In fact, it has been shown that a multilayer neural network, also called feed-forward network, can approximate any continuous function [Cybenko, 1989]. A "multilayer" neural network can even have only one layer. However, if a network is shallow, i.e. do not have many layers, it needs to have a very large number of neurons, i.e. be arbitrary wide, to approximate well a complex function. The feed-forward neural network is thus a *universal approximator*.

In fact, the proof of [Cybenko, 1989] is for the sigmoidal activation function. [Hornik, 1991] shows that the multilayer structure is the main reason behind the approximation capacity, not the activation function, which is later expanded to a much larger class of any activation function which are not polynomial [Leshno et al., 1993, Pinkus, 1999]. The feed-forward network is not the only one having the universal approximation property. It has been shown that it is also the case for *convolutional neural networks* [Zhou, 2020] (which is used in our work [Ngô et al., 2021] and introduced later), recurrent neural networks [Schäfer and Zimmermann, 2006], and graph neural networks [Gabrielsson, 2020]. However, note that these works say nothing about the form of the approximated solution, or how to train the data to get that function. Hence, the classical ways are to use numerical methods. *Gradient descent algorithm* is one of the most canonical ways, which is discussed in the next section [Cauchy et al., 1847, Hadamard, 1908].

5.3 Deep learning optimization

Given a differentiable loss function f and parameters θ, the gradient descent algorithm is a first-order iterative method. It is preferable to (quasi) Newton methods, because one of its advantages is that it does not require to calculate the Hessian, which is infeasible due the huge number of models in DL. The gradient descent algorithm tries to reduce the value of f at each step, hopefully making f reaching some (local) optimum. Its big disadvantage is that it is much slower than Newton methods. For DL, however, it is not necessary to reach some optimum, as it is only expected to decrease the loss function as much as possible, so improve the learning ability [Goodfellow et al., 2016]. The *de-facto* gradient descent algorithm for DL, Adam Kingma and Ba [2014], does not converge to an optimum in many cases [Bae et al., 2019].

Formally, denote θ_0 the initial values of a parameter θ, and the time step $t \in \mathbb{N}$. The value of θ at time $t+1$ is

$$\theta_{t+1} = \theta_t - \eta_t \frac{df}{d\theta},$$

where η_t is the learning rate (or step size) at time t. For DL, the dataset is often very big, hence it does not fit into the memory. A common way is to divide the training data into many small sets, denoted *batches*. The gradient descent algorithm using such batches is called *stochastic gradient descent* (SGD). The point of view of SGD is to address the gradient as an expectation that can be estimated from some small samples. Given a batch $B_t = \{x_1^t, \ldots, x_m^t\}$ of size m, then the estimated g is the average of the gradient at each data point

$$\tilde{g}_t = \frac{1}{m} \times \frac{d}{d\theta} \sum_{i=1}^{m} f(x_i^t, \theta),$$

hence

$$\theta_{t+1} = \theta_t - \tilde{g}_t \frac{df}{d\theta}.$$

m is often not more than several hundreds, and does not grow with training set size.

How η_t varies over time step t is an ongoing study with numerous different approaches. The simplest way is to choose some small constant learning rate. However, it could get stuck at saddle points, and also oscillate in the area close to local optimum because the gradients of the parameters are different orders of magnitudes [Jin et al., 2017, Sutton, 1986]. To resolve these issues, adaptive optimization methods have been developed. A well-known way is to take into account the past gradients, which is called momentum methods [Qian, 1999, Nesterov, 1983]. Another approach is to adapt the different learning to different parameters due to their sparseness in the data. It includes Adagrad [Duchi et al., 2011], AdaDelta [Zeiler, 2012], and RMSProp [Hinton et al., 2012]. The next generation optimizer are often the combination of different approaches and/or their enhanced versions: Adam Kingma and Ba [2014] is the combination of Momentum and RMSProp; Nadam is combined of Adam and Nesterov accelerated gradient [Nesterov, 1983], and so on (for more details, see [Ruder, 2016]). Even Adam is a default choice for DL, it is not one-size-fits-all for every problem, as stated by the *no free-lunch* theorem [Wolpert and Macready, 1997]. As DL is still an experimental field, one may need to test different optimisers to compare and select the most suitable one for each problem and dataset.

Besides optimizers, several other tricks are helpful for improving DL performance [Ruder, 2016]. It includes training data shuffling to reduce bias, batch normalisation to re-

establish normalisation after each batch learning for SGD acceleration [Ioffe and Szegedy, 2015], and early stopping to reduce overfitting.

Last but not least, fast calculation of gradients in DL network is not less important than the above techniques, due to the huge number of parameters of the model. The common way is to use the *backpropagation* algorithm, which is a form of dynamic programming based on the chain rule. It has a rich history [Schmidhuber, 2014], and it has been rediscovered many times since 1960s (e.g. [Kelley, 1960, Bryson, 1961, Dreyfus, 1962]). The modern version used in well-known DL frameworks is arguably based on the Master thesis of [Linnainmaa, 1970]. [Rumelhart et al., 1986] apply the method for neural networks, hence names it backpropagation. It have became the standard method to calculate gradients since 2010s with the popular use of GPU in DL [Schmidhuber, 2015].

5.4 Convolutional neural networks & U-Net

Convolutional neural networks (CNNs) are one type of feed-forward neural networks having less parameters than the same size networks, thanks to the *convolution* instead of full matrix multiplication in fully connected neural networks [Goodfellow et al., 2016]. Hence it saves time when training significantly while retain the on-par performance [Goodfellow et al., 2016]. CNN thus attracts many different applications, especially become the *de facto* DL model for image-related tasks due to their huge amount of data [Valueva et al., 2020].

It is inspired by signal processing, where data are signal and kernel is filter. Formally, for a one-dimensional signal $s = (\dots, s_{-1}, s_0, s_1, \dots)$ and the filter w of length $N = n_1 + n_2 + 1$: $w = (w_{-n_1}, \dots, w_{-1}, w_0, w_1, \dots, w_{n_2}) \in \mathbb{R}^N$, the convolution product between s and w is

$$z_j = \sum_{k=-n_1}^{n_2} s_{j+k} w_k$$

where $z = (\dots, z_{-1}, z_0, z_1, \dots)$ is the convoluted signal [Calin, 2020]. For example, when $w_{-k} = \cdots = w_0 = w_1 = \cdots = w_k$, we have a $(2k+1)$-moving average, as

$$z_k = \frac{1}{2k+1} \sum_{i=-k}^{k} s_i.$$

Hence we can say that the convolution is a form of weighted sum.

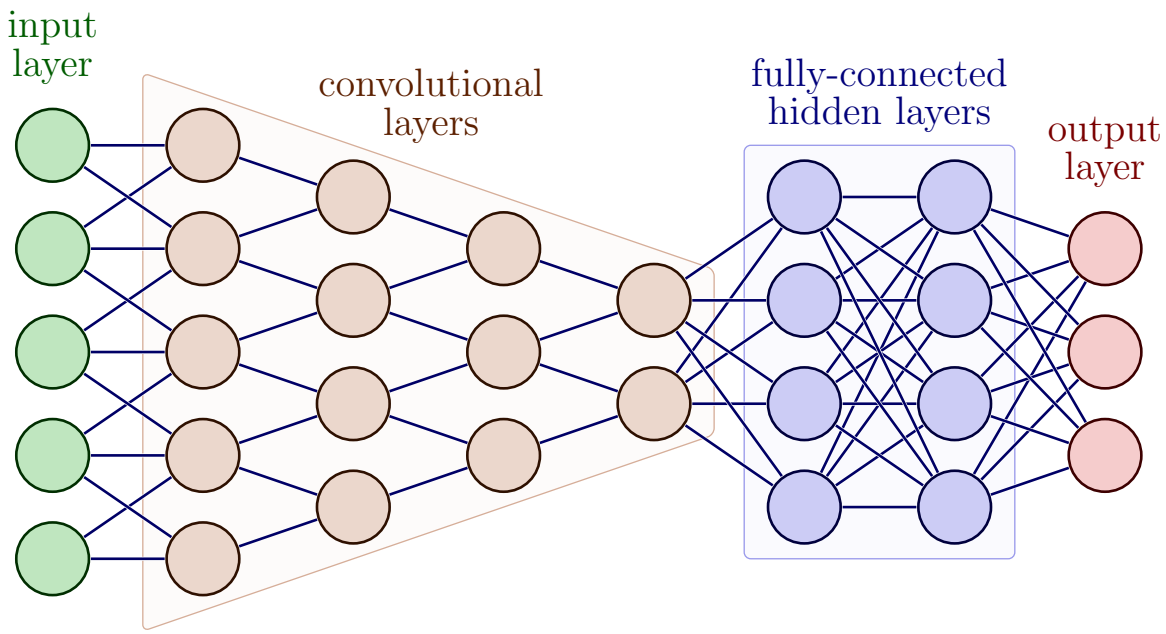

Figure 5.4: A neural network with CNN layers connected with fully connected hidden layers.[1]

In Figure 5.4, there are less connections between convolutional layers in the CNN than in fully connected neural network, it hence help boosting the training speed significantly. Similar to weight matrices in fully connected neural network, the kernels of CNN are automatically learned by the training. Another important component of CNN which is often used is Pooling layer, including Max Pooling and Average Pooling [Yamaguchi et al., 1990, Goodfellow et al., 2016]. Pooling layer decreases the dimension of the input data, allowing the CNN to "look" at scarcer versions of signals at different layers. Using Pooling, CNN can learn hierarchical structures in the data, from highest resolution with every detail to lower resolutions with more abstraction and less noise. For example, given $s = (1, 2, 3, 4)$ and the 2-size pooling layer P goes through s:

- if P is a Max pooling layer, then $P(s) = (\max(1, 2), \max(3, 4)) = (2, 4)$,
- if P is an Average pooling layer, then $P(s) = (\frac{1+2}{2}, \frac{3+4}{2}) = (1.5, 3.5)$.

[1]Image source from https://tikz.net/neural_networks/

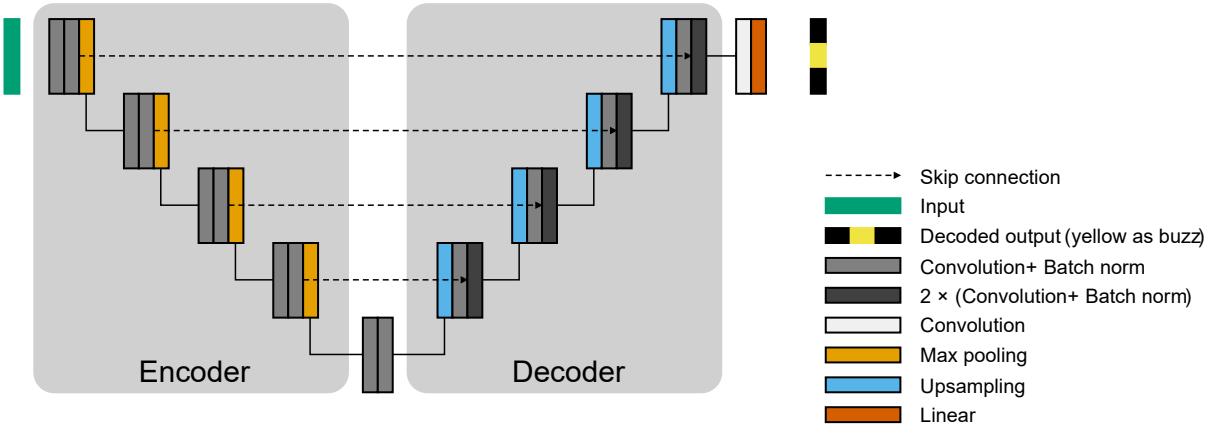

Figure 5.5: An example of a U-Net for buzz detection in our work [Ngô et al., 2021]. The encoder encodes data at different resolutions to feature maps, making the contracting path. The decoder decodes the corresponding messages from the encoder. The skip connections allow the feature maps skip lower layers to feed on the corresponding CNN decoder on the other side.[2]

We now focus on U-Net [Ronneberger et al., 2015], a specific kind of CNN used in our work [Ngô et al., 2021] for the time series data, inspired by the work for time series data by [Perslev et al., 2019]. It is based on an Encoder-Decoder structure (Figure 5.5). At the encoder side, CNN layers encode different resolutions of feature maps, outputted by CNN layers and downsampled by (Max or Average) pooling layers. These sequences of steps are called *contracting paths*. On the *decoder* side, CNN layers here act as translators to decode the encoded data, after upsampling to the same resolutions as in the encoder side. However, note that the input of the decoder loses information due to pooling layers at the encoder. Hence, U-Net introduces skip connections to allow the raw feature maps to go directly to the corresponding CNN layers in the decoder side, so they can learn the original version and the upsampling one together [Drozdzal et al., 2016]. It is one of the key ideas that make U-Net very successful in biomedical data, which require very high precision in pattern detection and low error rates [Siddique et al., 2021]. In our dataset, we let U-Net models explore the original data since we do not know what frequencies in the accelerometer data are useful for prey attempt detection.

[2]Image source from the Arxiv version of our paper: https://arxiv.org/abs/2101.01992

Bibliography

J. Aitchison and J. Brown. The lognormal distribution cambridge university press. *Cambridge UK*, 1957.

M. A. Alexander, J. D. Scott, K. D. Friedland, K. E. Mills, J. A. Nye, A. J. Pershing, A. C. Thomas, and E. C. Carmack. Projected sea surface temperatures over the 21st century: Changes in the mean, variability and extremes for large marine ecosystem regions of Northern Oceans. *Elementa: Science of the Anthropocene*, 6, 2018.

R. Anderson, D. Gordon, M. CraWley, and M. Hassell. Variability in the abundance of animal and plant species. *Nature*, 296(5854):245–248, 1982.

K. Bae, H. Ryu, and H. Shin. Does Adam optimizer keep close to the optimal point? *arXiv preprint arXiv:1911.00289*, 2019.

A. Blair and A. Saffidine. Ai surpasses humans at six-player poker. *Science*, 365(6456): 864–865, 2019.

N. Brown and T. Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

A. E. Bryson. A gradient method for optimizing multi-stage allocation processes. In *Proc. Harvard Univ. Symposium on digital computers and their applications*, volume 72, page 22, 1961.

J. Burns, J. Montague, and C. Cowles. The bowhead whale, Special Publication Number, 2. *The Society for Mammalogy*, 1993.

O. Calin. *Deep Learning Architectures: A Mathematical Approach*. Springer International Publishing, Cham, 2020. ISBN 978-3-030-36721-3. doi: 10.1007/978-3-030-36721-3_1. URL `https://doi.org/10.1007/978-3-030-36721-3_1`.

A. Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

P. Chambault, C. M. Albertsen, T. A. Patterson, R. G. Hansen, O. Tervo, K. L. Laidre, and M. P. Heide-Jørgensen. Sea surface temperature predicts the movements of an Arctic cetacean: the bowhead whale. *Scientific Reports*, 8(1):1–12, 2018.

P. J. Corkeron and R. C. Connor. Why do baleen whales migrate? *Marine Mammal Science*, 15(4):1228–1245, 1999.

D. Crevier. *AI: the tumultuous history of the search for artificial intelligence*. Basic Books, Inc., 1993.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

C. Darwin. *On the origin of species, 1859*. Routledge, 2004.

A. Davies, P. Veličković, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász, et al. Advancing mathematics by guiding human intuition with AI. *Nature*, 600(7887):70–74, 2021.

Ł. Delong, M. Lindholm, and M. V. Wüthrich. Making Tweedie's compound poisson model more accessible. *European Actuarial Journal*, pages 1–42, 2021.

T. Doniol-Valcroze, J.-F. Gosselin, D. G. Pike, J. W. Lawson, N. C. Asselin, K. Hedges, and S. H. Ferguson. Narwhal abundance in the Eastern Canadian High Arctic in 2013. *NAMMCO Scientific Publications*, 11, 2019.

S. Dreyfus. The numerical solution of variational problems. *Journal of Mathematical Analysis and Applications*, 5(1):30–45, 1962.

M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *Deep learning and data labeling for medical applications*, pages 179–187. Springer, 2016.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

P. K. Dunn and G. K. Smyth. *Chapter 5: Generalized Linear Models: Structure*, pages 211–241. Springer New York, New York, NY, 2018a. ISBN 978-1-4419-0118-7. doi: 10.1007/978-1-4419-0118-7_5. URL `https://doi.org/10.1007/978-1-4419-0118-7_5`.

P. K. Dunn and G. K. Smyth. *Chapter 12: Tweedie GLMs*, pages 457–490. Springer New York, New York, NY, 2018b. ISBN 978-1-4419-0118-7. doi: 10.1007/978-1-4419-0118-7_12. URL `https://doi.org/10.1007/978-1-4419-0118-7_12`.

D. Eschricht and J. Reinhardt. Om nordhvalen (Balaena mysticetus L.) navnlig med hensyn til dens udbredning i fortiden og nutiden og til dens ydre og indre saerkjender. *Scientific Reports*, 8(1):1–12, 2018.

D. J. Farlie. The performance of some correlation coefficients for a general bivariate distribution. *Biometrika*, 47(3/4):307–323, 1960.

J. K. Ford and H. D. Fisher. Underwater acoustic signals of the narwhal (Monodon monoceros). *Canadian Journal of Zoology*, 56(4):552–560, 1978.

S. D. Foster and M. V. Bravington. A poisson–gamma model for analysis of ecological non-negative continuous data. *Environmental and ecological statistics*, 20(4):533–552, 2013.

R. Gabrielsson. Universal function approximation on graphs. *Advances in Neural Information Processing Systems*, 33, 2020.

E. Garde and M. P. Heide-Jørgensen. Tusk anomalies in narwhals (Monodon monoceros) from Greenland. *Polar research*, XX(XX):to appear, 2022.

E. Garde, M. P. Heide-Jørgensen, S. H. Hansen, G. Nachman, and M. C. Forchhammer. Age-specific growth and remarkable longevity in narwhals (Monodon monoceros) from West Greenland as estimated by aspartic acid racemization. *Journal of Mammalogy*, 88(1):49–58, 2007.

J. George, J. Thewissen, A. Von Duyke, G. A. Breed, R. Suydam, T. L. Sformo, B. T. Person, and H. Brower. Life history, growth, and form. In *The bowhead whale: Balaena mysticetus: Biology and human interactions*, pages 87–115. Elsevier, 2021.

J. C. George, J. Bada, J. Zeh, L. Scott, S. E. Brown, T. O'Hara, and R. Suydam. Age and growth estimates of bowhead whales (balaena mysticetus) via aspartic acid racemization. *Canadian Journal of Zoology*, 77(4):571–580, 1999.

J. C. George et al. *The bowhead whale: Balaena mysticetus: Biology and human interactions*. Academic Press, 2020.

J. C. George. *Growth, morphology and energetics of bowhead whales (Balaena mysticetus)*. University of Alaska Fairbanks, 2009.

G. H. Givens and M. P. Heide-Jørgensen. Abundance. In *The bowhead whale: Balaena mysticetus: Biology and human interactions*, pages 77–86. Elsevier, 2021.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

Z. A. Graham, E. Garde, M. P. Heide-Jørgensen, and A. V. Palaoro. The longer the better: evidence that narwhal tusks are sexually selected. *Biology letters*, 16(3):20190950, 2020.

E. J. Gumbel. Distributions à plusieurs variables dont les marges sont données. *Comptes rendus hebdomadaires des seances de l'academie des sciences*, 246(19):2717–2719, 1958.

J. Hadamard. *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*, volume 33. Imprimerie nationale, 1908.

J. T. Haldiman, W. G. Henk, R. W. Henry, T. F. Albert, Y. Z. Abdelbaki, and D. W. Duffield. Epidermal and papillary dermal characteristics of the bowhead whale (Balaena mysticetus). *The Anatomical Record*, 211(4):391–402, 1985.

M. P. Heide-Jørgensen. Aerial digital photographic surveys of narwhals, Monodon monoceros, in northwest Greenland. *Marine Mammal Science*, 20(2):246–261, 2004.

M. P. Heide-Jørgensen. Narwhal: Monodon monoceros. In *Encyclopedia of marine mammals, 2th edition*, pages 754–758. Elsevier, 2009.

M. P. Heide-Jørgensen. Narwhal: Monodon monoceros. In *Encyclopedia of marine mammals, 3rd edition*, pages 627–631. Elsevier, 2018.

M.-P. Heide-Jørgensen and K. Laidre. *Greenland's winter whales: The beluga, the narwhal and the bowhead whale*. Ilinniusiorfik Undervisningsmiddelforlag, 2006.

M. P. Heide-Jørgensen, K. L. Laidre, N. H. Nielsen, R. G. Hansen, and A. Røstad. Winter and spring diving behavior of bowhead whales relative to prey. *Animal Biotelemetry*, 1(1):1–14, 2013.

M. P. Heide-Jørgensen, S. B. Blackwell, O. M. Tervo, A. L. Samson, E. Garde, R. G. Hansen, A. S. Conrad, P. Trinhammer, H. C. Schmidt, M.-H. S. Sinding, et al. Behavioral response study on seismic airgun and vessel exposures in narwhals. *Frontiers in Marine Science*, page 665, 2021.

G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

J. Hokkanen. Temperature regulation of marine mammals. *Journal of Theoretical Biology*, 145(4):465–485, 1990.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985. doi: 10.1017/CBO9780511810817.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

S. Innes, M. Heide-Jørgensen, J. L. Laake, K. L. Laidre, H. J. Cleator, P. Richard, and R. E. Stewart. Surveys of belugas and narwhals in the Canadian High Arctic in 1996. *NAMMCO Scientific Publications*, 4:169–190, 2002.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

T. Izawa. The bivariate gamma distribution. *Climate and statistics*, 4, 1953.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

B. Jørgensen. Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(2):127–145, 1987.

B. Jørgensen and M. C. Paes De Souza. Fitting Tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93, 1994.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

H. J. Kelley. Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954, 1960.

K. Kelly and R. Krzysztofowicz. A bivariate meta-gaussian density for use in hydrology. *Stochastic Hydrology and hydraulics*, 11(1):17–31, 1997.

W. S. Kendal. Spatial aggregation of the colorado potato beetle described by an exponential dispersion model. *Ecological modelling*, 151(2-3):261–269, 2002.

W. S. Kendal. A scale invariant clustering of genes on human chromosome 7. *BMC evolutionary biology*, 4(1):1–10, 2004a.

W. S. Kendal. Taylor's ecological power law as a consequence of scale invariant exponential dispersion models. *Ecological Complexity*, 1(3):193–209, 2004b.

W. S. Kendal. Scale invariant correlations between genes and snps on human chromosome 1 reveal potential evolutionary mechanisms. *Journal of theoretical biology*, 245(2):329–340, 2007.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

K. M. Kovacs, C. Lydersen, J. Vacquiè-Garcia, O. Shpak, D. Glazov, and M. P. Heide-Jørgensen. The endangered Spitsbergen bowhead whales' secrets revealed after hundreds of years in hiding. *Biology letters*, 16(6):20200148, 2020.

Krauthammer, Charles. Be Afraid. `https://www.washingtonexaminer.com/weekly-standard/be-afraid-9802`, 1997. Accessed: 2022-01-30.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

K. L. Laidre and M. P. Heide-Jørgensen. Spring partitioning of Disko Bay, West Greenland, by Arctic and subarctic baleen whales. *ICES Journal of Marine Science*, 69(7): 1226–1233, 2012.

K. A. Lefebvre, L. Quakenbush, E. Frame, K. B. Huntington, G. Sheffield, R. Stimmelmayr, A. Bryan, P. Kendrick, H. Ziel, T. Goldstein, et al. Prevalence of algal toxins in Alaskan marine mammals foraging in a changing arctic and subarctic environment. *Harmful Algae*, 55:13–24, 2016.

K. J. Lefort, C. J. Garroway, and S. H. Ferguson. Killer whale abundance and predicted narwhal consumption in the Canadian Arctic. *Global change biology*, 26(8):4276–4283, 2020.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

S. Linnainmaa. The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), Univ. Helsinki*, pages 6–7, 1970.

M. Louis, M. Skovrind, J. A. Samaniego Castruita, C. Garilao, K. Kaschner, S. Gopalakrishnan, J. S. Haile, C. Lydersen, K. M. Kovacs, E. Garde, et al. Influence of past climate change on phylogeography and demographic history of narwhals, Monodon monoceros. *Proceedings of the Royal Society B*, 287(1925):20192964, 2020.

L. F. Lowry. Foods and feeding ecology. *The bowhead whale. Society for marine mammalogy, Special publication*, 2:201–238, 1993.

L. F. Lowry, G. Sheffield, and J. C. George. Bowhead whale feeding in the Alaskan Beaufort Sea, based on stomach contents analyses. *Journal of Cetacean research and Management*, 6(3):215–223, 2004.

B. T. McClintock, R. Langrock, O. Gimenez, E. Cam, D. L. Borchers, R. Glennie, and T. A. Patterson. Uncovering ecological state dynamics with hidden markov models. *Ecology letters*, 23(12):1878–1903, 2020.

L. A. Miller, J. Pristed, B. Møshl, and A. Surlykke. The click-sounds of narwhals (Monodon monoceros) in Inglefield bay, northwest Greenland. *Marine Mammal Science*, 11 (4):491–502, 1995.

M. Minsky and S. A. Papert. *Perceptrons: An introduction to computational geometry*. MIT press, 2017.

C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

P. Moran. Statistical inference with bivariate gamma distributions. *Biometrika*, 56(3): 627–634, 1969.

D. Morgenstern. Einfache beispiele zweidimensionaler verteilungen. *Mitteilingsblatt fur Mathematische Statistik*, 8:234–235, 1956.

D. Morrissey. Introduction to dynamical systems. *Math Insight*, 2021.

Naalakkersuisut. 2022 kvoter for hvid- og narhvaler. `https://naalakkersuisut.gl/da/Naalakkersuisut/Nyheder/2021/12/2112_Qilalukkat`, 2021. Accessed: 2022-01-30.

NAMMCO. Report of the Scientific Committee 26th Meeting. Technical report, NAMMCO, 10 2019.

NAMMCO. Report of Meeting of the Ad hoc Working Group on Narwhal in East Greenland. Technical report, NAMMCO, 10 2021.

J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady an ussr*, volume 269, pages 543–547, 1983.

M. C. Ngô, M. P. Heide-Jørgensen, and S. Ditlevsen. Understanding narwhal diving behaviour using hidden markov models with dependent state distributions and long range dependence. *PLoS computational biology*, 15(3):e1006425, 2019.

M. C. Ngô, R. Selvan, O. Tervo, M. P. Heide-Jørgensen, and S. Ditlevsen. Detection of foraging behavior from accelerometer data using U-Net type convolutional networks. *Ecological Informatics*, 62:101275, 2021.

M. Perslev, M. H. Jensen, S. Darkner, P. J. Jennum, and C. Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *arXiv preprint arXiv:1910.11162*, 2019.

A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta numerica*, 8:143–195, 1999.

C. Pomerleau, M. P. Heide-Jørgensen, S. H. Ferguson, H. L. Stern, J. L. Høyer, and G. A. Stern. Reconstructing variability in West Greenland ocean biogeochemistry and bowhead whale (Balaena mysticetus) food web structure using amino acid isotope ratios. *Polar Biology*, 40(11):2225–2238, 2017.

L. Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

N. Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

K. Revfeim. An initial model of the relationship between rainfall events and daily rainfalls. *Journal of Hydrology*, 75(1-4):357–364, 1984.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

F. Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

S. M. Ross. *Introduction to probability models*. Academic press, 2014.

S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

A. M. Schäfer and H. G. Zimmermann. Recurrent neural networks are universal approximators. In *International Conference on Artificial Neural Networks*, pages 632–640. Springer, 2006.

J. Schmidhuber. Who invented backpropagation? *More[DL2]*, 2014.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117, 2015.

G. Sheffield and J. George. Diet and prey. In *The bowhead whale: Balaena mysticetus: Biology and human interactions*, pages 429–455. Elsevier, 2021.

H. Shono. Application of the tweedie distribution to zero-catch data in cpue analysis. *Fisheries Research*, 93(1-2):154–162, 2008.

N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 2021.

H. Siegstad and M.-P. Heide-Jørgensen. Ice entrapments of narwhals (Monodon monoceros) and white whales (Delphinapterus leucas) in greenland. *Meddeleser om Grønland Bioscience*, 39:151–160, 1994.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

O. Smith, S. Adelfang, and J. Tubbs. A bivariate gamma probability distribution with application to gust modeling. *Nasa technical memorandum*, 82483, 1982.

G. K. Smyth and B. Jørgensen. Fitting Tweedie's compound poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin: The Journal of the IAA*, 32(1): 143–157, 2002.

K. M. Stafford, K. L. Laidre, and M. P. Heide-Jørgensen. First acoustic recordings of narwhals (Monodon monoceros) in winter. *Marine Mammal Science*, 28(2):E197–E207, 2012.

R. Sutton. Two problems with back propagation and other steepest descent learning procedures for networks. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society, 1986*, pages 823–832, 1986.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

L. R. Taylor. Aggregation, variance and the mean. *Nature*, 189(4766):732–735, 1961.

D. H. Terman and E. M. Izhikevich. State space. *Scholarpedia*, 3(3):1924, 2008. doi: 10.4249/scholarpedia.1924. revision #137545.

O. M. Tervo, S. B. Blackwell, S. Ditlevsen, A. S. Conrad, A. L. Samson, E. Garde, R. G. Hansen, and H.-J. Mads Peter. Narwhals react to ship noise and airgun pulses embedded in background noise. *Biology letters*, 17(11):20210220, 2021.

C. Thompson. Homogeneity analysis of rainfall series: an application of the use of a realistic rainfall model. *Journal of climatology*, 4(6):609–619, 1984.

M. Tweedie. The regression of the sample variance on the sample mean. *Journal of the London Mathematical Society*, 1(1):22–28, 1946.

M. V. Valueva, N. Nagornov, P. A. Lyakhov, G. V. Valuev, and N. I. Chervyakov. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation*, 177:232–243, 2020.

O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. M. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell, et al. Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2, 2019.

M. A. Wani, T. M. Khoshgoftaar, and V. Palade. *Deep Learning Applications, Volume 2*. Springer, 2020.

D. Wetzel, J. Reynolds, P. III, G. Givens, E. Pulster, and J. George. Age estimation for bowhead whales, balaena mysticetus, using aspartic acid racemization with enhanced hydrolysis and derivatization procedures. *Journal of Cetacean Research and Management*, 17:9–14, 2017.

T. M. Williams, S. R. Noren, and M. Glenn. Extreme physiological adaptations as predictors of climate-change sensitivity in the narwhal, Monodon monoceros. *Marine Mammal Science*, 27(2):334–349, 2011.

T. M. Williams, S. B. Blackwell, B. Richter, M.-H. S. Sinding, and M. P. Heide-Jørgensen. Paradoxical escape responses by narwhals (Monodon monoceros). *Science*, 358(6368): 1328–1331, 2017.

D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

K. Yamaguchi, K. Sakamoto, T. Akabane, and Y. Fujimoto. A neural network for speaker-independent isolated word recognition. In *ICSLP*, 1990.

S. Yue, T. B. Ouarda, and B. Bobée. A review of bivariate gamma distributions for hydrological application. *Journal of Hydrology*, 246(1-4):1–18, 2001.

M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

D.-X. Zhou. Universality of deep convolutional neural networks. *Applied and computational harmonic analysis*, 48(2):787–794, 2020.

W. Zucchini, I. L. MacDonald, and R. Langrock. *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC, 2016.

Chapter 6

Paper I

Joint work with

Mads Peter Heide-Jørgensen and Susanne Ditlevsen

This chapter is based on the published article: Manh Cuong Ngo, Mads Peter Heide-Jørgensen and Susanne Ditlevsen. *Understanding narwhal diving behaviour using Hidden Markov Models with dependent state distributions and long range dependence.* PLoS Computational Biology, 15(3): e1006425, 2019.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Understanding narwhal diving behaviour using Hidden Markov Models with dependent state distributions and long range dependence

Manh Cuong Ngô[1,2], **Mads Peter Heide-Jørgensen**[1,3], **Susanne Ditlevsen**[2]*

1 Greenland Institute of Natural Resources, Nuuk, Greenland, **2** Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark, **3** Greenland Institute of Natural Resources, c/o Greenland Representation, Copenhagen, Denmark

* susanne@math.ku.dk

OPEN ACCESS

Citation: Ngô MC, Heide-Jørgensen MP, Ditlevsen S (2019) Understanding narwhal diving behaviour using Hidden Markov Models with dependent state distributions and long range dependence. PLoS Comput Biol 15(3): e1006425. https://doi.org/10.1371/journal.pcbi.1006425

Editor: Bard G. Ermentrout, University of Pittsburgh, UNITED STATES

Received: August 2, 2018

Accepted: January 28, 2019

Published: March 14, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: MPHJ received funding from the Greenland Institute of Natural Resources (www.natur.gl); the Danish Cooperation for the Environment in the Arctic (http://mst.dk/kemi/kemikalier/arktis/dancea-miljoestoette-til-arktis/) and the Carlsberg Foundation, grant number 2013_01_0289 and CF14-0169 (www.carlsbergfondet.dk/da). SD received funding from

Abstract

Diving behaviour of narwhals is still largely unknown. We use Hidden Markov models (HMMs) to describe the diving behaviour of a narwhal and fit the models to a three-dimensional response vector of maximum dive depth, duration of dives and post-dive surface time of 8,609 dives measured in East Greenland over 83 days, an extraordinarily long and rich data set. Narwhal diving patterns have not been analysed like this before, but in studies of other whale species, response variables have been assumed independent. We extend the existing models to allow for dependence between state distributions, and show that the dependence has an impact on the conclusions drawn about the diving behaviour. We try several HMMs with 2, 3 or 4 states, and with independent and dependent log-normal and gamma distributions, respectively, and different covariates to characterize dive patterns. In particular, diurnal patterns in diving behaviour is inferred, by using periodic B-splines with boundary knots in 0 and 24 hours.

Author summary

Narwhals live in pristine environments. However, the increase in average temperatures in the Arctic and the concomitant loss of summer sea ice, as well as increased human activities, such as ship traffic and mineral exploration leading to increased noise pollution, are changing the environment, and therefore probably also the behavior and well-being of the narwhal. Here, we use probabilistic models to unravel the diving and feeding behavior of a male narwhal, tagged in East Greenland in 2013, and followed for more than two months. The goal is to gain knowledge of the whales' normal behavior, to be able to later detect possible changes in behavior due to climatic changes and human influences. We find that the narwhal uses around two thirds of its time searching for food, it typically feeds during deep dives (more than 350m), and it can have extended periods, up to 3 days, without feeding activity.

PLOS COMPUTATIONAL BIOLOGY

University of Copenhagen Excellence Programme for Interdisciplinary Research (https://research.ku.dk/strengths/excellence-programmes/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The narwhal (*Monodon monoceros*) primarily inhabit cold waters of the Atlantic sector of the Arctic, with the largest abundances found in East and West Greenland and in the Canadian High Arctic [1]. The narwhal is one of the deepest diving cetaceans with the maximum exceeding 1800*m* [2], and it comes third only to Cuvier's beaked whale *(Ziphius cavirostris)* (2992*m*) [3] and sperm whale *(Physeter macrocephalus)* (2035*m*) [4]. Narwhals dive to forage, and their diet consists of few prey species including Greenland halibut (*Reinhardtius hippoglossoides*), polar cod (*Boreogadus saida*), capelin (*Ammodytes villosus*) and squids (*Gonatus sp*.) [5, 6]. Narwhals depend on acoustics for sensing their environment, navigating and capturing prey at depth [7]. Anthropogenic factors like underwater noise are a concern for a species that, with decreasing sea ice coverage, is increasingly exposed to underwater noise from shipping and seismic exploration [8]. It is therefore important to understand and quantitatively describe the diving activities of narwhals, by robust statistical methods, to ensure the long-term conservation of one of the most specialized species in the North Atlantic.

The first step is to understand the diving patterns of narwhals under natural conditions, which we address in this study. Diving behaviour is however cryptic since it includes both physiological constraints, energetic demands and habitat and environmental regimes. Modelling of the observed diving behaviour is one way of gaining insight to the overall diving patterns, and changes in model parameters is a way to compare and estimate quantitatively changes in diving behavior or differences between individuals.

We apply multivariate Hidden Markov Models (HMMs) with covariates [9], to describe the diving dynamics in the vertical dimension of an individual narwhal. These types of models for similar diving data of Blainville's beaked whales (*Mesoplodon densirostris*) were first introduced in [10]. A HMM assumes an underlying unobserved process, which governs the dynamics of the observed variables. The assumption is that the observed behaviour in a dive will depend on the present state, and introduces autocorrelation in the model [9]. These HMMs have been used for modelling animal movement by taking into account the correlation over time between different movement patterns, mainly in two horizontal dimensions (see, e.g., [11–13]), and recently, in one vertical dimension [10, 14], possibly including further information on vertical movements. In this study, we use vertical depth data, and the three response variables are the maximum depth reached in a dive, the duration of a dive, and the post-dive surface time before initiating a new dive.

In all previous studies, *contemporaneous conditional independence* was assumed, meaning that the state dependent processes are independent given the underlying state. This is a strong and often also an unrealistic assumption, since deeper dives will typically take longer. Even when conditioning the dive to be either shallow, medium or deep, a positive correlation is still expected, beyond the correlation implied by the hidden states. DeRuiter et al. [14] argued for the assumption of conditional independence because unless a multivariate normal distribution can be assumed, there is usually no simple candidate multivariate distribution to specify the correlation structure. This is partly due to some of their response variables being discrete. In this study, we will relax the assumption of conditional independence, taking advantage of the continuity of the response variables. They are all restricted to be positive and with right skewed distributions. Previous studies have therefore used conditionally independent gamma distributions for these variables. Here, we will assume dependent log-normal distributions, such that their log-transforms follow a multivariate normal distribution. We also do the analysis with the standard choice of the gamma distributions with both dependence and independence, as well as the independent log-normal distributions, and compare the results.

Covariates were included in [10, 13, 14], appearing in the transition probabilities between hidden states, whereas no covariates were included in [15]. Here we include covariates in all elements of the transition matrix, trying out different covariate process models and select the optimal model by the Akaike Information Criterion (AIC). We consider two covariates related to the recent deep dives performed by the narwhal. Dives can reach $> 1800m$, and deeper dives are assumed to be related to feeding [2]. We define a *deep dive* as a dive to a depth of at least $350m$. One covariate is the time passed since the last deep dive, which was also used in [10]. The hypothesis is that the longer the time passed since last deep dive, the higher the narwhal's propensity for initiating a deep dive will be. Another covariate counts the number of consecutive deep dives that the narwhal has performed. The hypothesis is that the more dives in a row and more time spent at great depths, the higher the narwhal's propensity for changing diving pattern to shallower depth or near-surface travelling. By introducing such history dependent covariates, the model allows a longer dependence structure than the one implied by the Markov property. These models with dependencies between observables caused by the underlying state, as well as including feedback from the observed process, were introduced in [10] to model Blainville's beaked whale. The last covariate is time of day at initiation of the dive, modelled by a periodic B-spline with boundary knots in 0 and 24 hours. Diurnal effects on marine mammal diving patterns are difficult to estimate in this type of models because the time series are typically too short. Here, we analyse a data set of a tagged narwhal that is extraordinarily long, nearly three months, making this inference possible. Normally, such time series are on the order of hours or days. However, we only have data from a single whale, and results might not generalize.

Materials and methods

Ethics statement

Permission for capturing, handling, and tagging of narwhals was provided by the Government of Greenland (Case ID 2010–035453, document number 429 926).

Data

We analyse the time series of depth measurements of a mature male narwhal (420 cm, estimated mass 950 kg) tagged in East Greenland from August 13th until November 6th 2013. The tag (a satellite linked time depth recorder, the Mk10 time-depth recorder from Wildlife Computers, Redmond, WA, USA) was attached to the whale and retrieved one year later with 1994.83 hours of dive data (approximately 83 days and 2 hours), see [16]. In this time interval the narwhal performed 8,609 dives to depths of at least $20m$. Depth was measured every second at a resolution of $0.5m$, and preprocessed before analysis by summarizing in three variables within each dive to describe the behaviour: maximum depth (MD), dive duration (DT), and post-dive surface time (PD), as also used in [14]. A dive was scored every time the depth record went deeper than $20m$ (i.e., about four to six body lengths) to exclude brief shallow submersions between respirations, otherwise it is considered time spent at the surface, summarized in the variable PD. This threshold was chosen in order to avoid creating too many shallow dives near the surface, see [17]. We use a custom-written procedure in C++ combining with R [18] via Rcpp [19]. The dives are found by locating all zero depth measurements. If there is at least one depth measurement of at least $20m$ between two consecutive measurements of $0m$, this is classified as a dive. Otherwise an interval between two $0m$ measurements is classified as part of the post-dive time after the last dive. For each identified dive, the largest depth measurement is defined as the maximum depth of the dive, and the dive duration is the time difference

between the two $0m$ measurements. The surface and dive durations also enter in the model as part of the covariate counting the time since last deep dive.

In this study, the observed response variable, denoted by $\boldsymbol{X}_t$, is three-dimensional, describing the diving behaviour related to each dive, where t indicates the dive number, $t = 1, 2, \ldots, T$. The first response variable, $X_{1,t}$, is MD reached in dive number t. The second response variable, $X_{2,t}$, is DT of dive number t. The third response variable, $X_{3,t}$, is PD after dive t. We assume that the diving behaviour depends on an underlying unobserved process, which we denote by C_t, $t = 1, 2, \ldots$, with a number m of unobserved behavioural states, $C_t \in \{1, \ldots, m\}$, which govern the dynamics of the observed variables. The assumption is that the distributions of the observed MD, DT and PD of dive number t depend on the state.

Hidden Markov Model

An m-dimensional hidden Markov model assumes that the distribution of the p-dimensional response vector $\boldsymbol{X}_t$ depends on a hidden state C_t, where $\{C_t: t = 1, 2, \ldots\}$ is an unobserved underlying process satisfying the Markov property:

$$P(C_t = j \mid C_{t-1} = i, \ldots, C_1 = l) = P(C_t = j \mid C_{t-1} = i),$$

where $C_t \in \{1, \ldots, m\}$ for $t = 2, 3, \ldots$. Denote the state transition probabilities at time t by $\omega_{ij}(t)$, $i, j = 1, \ldots, m$, where $\omega_{ij}(t) = P(C_{t+1} = j | C_t = i)$. The transition probability matrix $\boldsymbol{\Omega}(t)$ is then

$$\boldsymbol{\Omega}(t) = \begin{bmatrix} \omega_{11}(t) & \cdots & \omega_{1m}(t) \\ \vdots & \ddots & \vdots \\ \omega_{m1}(t) & \cdots & \omega_{mm}(t) \end{bmatrix} \tag{1}$$

where $\omega_{ij}(t) \geq 0$ and $\sum_{j=1}^{m} \omega_{ij}(t) = 1$. Here, we let $\omega_{ij}(t)$ depend on t to allow time varying covariates to affect the transition probabilities, see Section Covariates. The distribution of $\boldsymbol{X}_t$ is conditionally independent of everything else given C_t:

$$f(\boldsymbol{X}_t | \boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_1, C_t, C_{t-1}, \ldots, C_1) = f(\boldsymbol{X}_t | C_t), \; t = 1, 2, \ldots \tag{2}$$

where f denotes a probability density function, i.e., the distribution of $\boldsymbol{X}_t$ depends only on the current state C_t and not on previous states or observations. The model is illustrated in Fig 1.

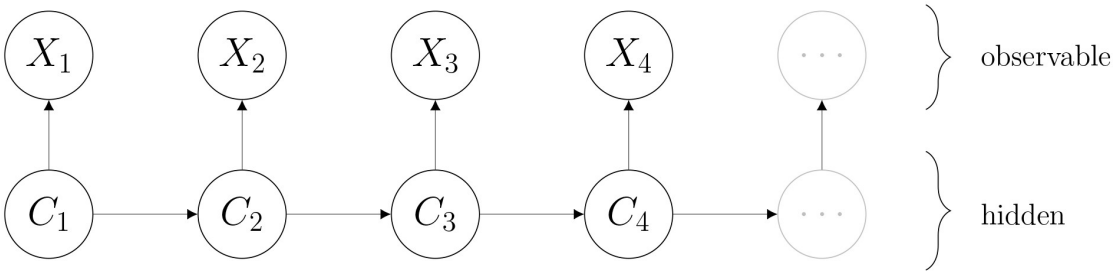

Fig 1. Hidden Markov Model. The hidden states C_t represent behavioural states that influence the distribution of the observed variables X_t.

https://doi.org/10.1371/journal.pcbi.1006425.g001

PLOS COMPUTATIONAL BIOLOGY

State dependent distributions

The state-dependent distributions are the probability density functions of $\boldsymbol{X}_t$ associated with state i. Under the *contemporaneous conditional independence* assumption, the p different components of the response vector $\boldsymbol{X}_t$ are assumed independent given the hidden state, and the probability density can be decomposed as

$$f(\boldsymbol{X}_t \mid C_t = i) = f_i(\boldsymbol{X}_t) = \prod_{k=1}^{p} f_{i,k}(X_{k,t}), \tag{3}$$

where $X_{k,t}$ is the kth observed component of $\boldsymbol{X}_t$. Here we have $p = 3$, the components being MD, DT and PD. Thus, $\boldsymbol{X}_t = (X_{MD,t}, X_{DT,t}, X_{PD,t})^T$, where ^{T} denotes transposition. *Contemporaneous conditional independence* implies that the state dependent processes $X_{MD,t}$, $X_{DT,t}$ and $X_{PD,t}$ are independent given the underlying state C_t. This assumption has been used in [14] and [15] because in general, there is no simple way to address the correlation between variables within states, and the dependence induced by the Markov chain is often sufficient to fit the data. However, in this paper, we will relax this assumption, and let f_i be a joint distribution function, allowing for dependent coordinates, which for our data turn out to improve the fit considerably.

All three response variables are positive right-skewed variables, so natural candidates for $f_{i,k}$ are gamma distributions, as used in [14] and [15], or log-normal distributions, i.e., the logarithm of the response variables follow a 3-dimensional normal distribution. Here, we will try four different distributions. The first candidate is independent gamma distributions, to compare with the usual approach. The gamma distribution is parametrized by shape parameter μ and scale parameter σ, with mean $\mu\sigma$ and variance $\mu\sigma^2$, and the state dependent probability density functions are given by

$$f_i(\boldsymbol{X}_t) = \prod_{k\in\{MD,DT,PT\}} f_{i,k}(X_{k,t}) = \prod_{k\in\{MD,DT,PT\}} \Gamma(\mu_i^k)^{-1}(\sigma_i^k)^{-\mu_i^k} X_{k,t}^{\mu_i^k-1} e^{-\frac{X_{k,t}}{\sigma_i^k}}, \tag{4}$$

for $i = 1, \ldots, m$.

We will also assume dependent gamma distributions [20] and both independent and correlated log-normal distributions, such that $\log \boldsymbol{X}_t$ is multivariate normal, where $\log \boldsymbol{X}_t = (\log X_{MD,t}, \log X_{DT,t}, \log X_{PT,t})^T$, taking advantage of the computational convenience of the normal distribution. The log-normal distribution is parametrized by log-mean μ and log-variance σ^2. Thus, given $C_t = i$ and k, the mean and variance of $\log \boldsymbol{X}_{k,t}$ is μ_i^k and $(\sigma_i^k)^2$, and the mean and variance of $\boldsymbol{X}_{k,t}$ is $\exp(\mu_i^k + (\sigma_i^k)^2/2)$ and $(\exp((\sigma_i^k)^2) - 1)\exp(2\mu_i^k + (\sigma_i^k)^2)$. The log-correlation between responses k_1 and k_2, for $k_1, k_2 \in \{MD, DT, PT\}$ is denoted by $\rho_i^{k_1,k_2}$. The correlation between components k_1 and k_2 is $(\exp(\rho_i \sigma_i^{k_1} \sigma_i^{k_2}) - 1)/\sqrt{(\exp((\sigma_i^{k_1})^2) - 1)(\exp((\sigma_i^{k_2})^2) - 1)}$, where $(\sigma_i^{k_1})^2$ and $(\sigma_i^{k_2})^2$ are the log-variances of k_1 and k_2, respectively. The correlation is approximately equal to the log-correlation $\rho_i^{k_1,k_2}$ when $(\sigma_i^{k_1})^2$ and $(\sigma_i^{k_2})^2$ are small. Thus, the state dependent probability density functions are given by

$$f_i(\boldsymbol{X}_t) = \frac{1}{(2\pi)^{3/2}\sqrt{|\Sigma_i|}\cdot\prod_{k\in\{MD,DT,PT\}}\log \boldsymbol{X}_{k,t}} \exp\left(-\frac{1}{2}(\log \boldsymbol{X}_t - \mu_i)^\top \Sigma_i^{-1}(\log \boldsymbol{X}_t - \mu_i)\right), \tag{5}$$

PLOS COMPUTATIONAL BIOLOGY

where $|\cdot|$ denotes the determinant of a matrix, $\mu_i = (\mu_i^{MD}, \mu_i^{DT}, \mu_i^{PD})^T$,

$$\Sigma_i = \begin{bmatrix} (\sigma_i^{MD})^2 & \rho_i^{MD,DT}\sigma_i^{MD}\sigma_i^{DT} & \rho_i^{MD,PD}\sigma_i^{MD}\sigma_i^{PD} \\ \rho_i^{MD,DT}\sigma_i^{MD}\sigma_i^{DT} & (\sigma_i^{DT})^2 & \rho_i^{DT,PD}\sigma_i^{DT}\sigma_i^{PD} \\ \rho_i^{MD,PD}\sigma_i^{MD}\sigma_i^{PD} & \rho_i^{DT,PD}\sigma_i^{DT}\sigma_i^{PD} & (\sigma_i^{PD})^2 \end{bmatrix}$$

and $\rho_i^{k_1,k_2} = 0$ in the independent case.

Covariates

To allow for a longer memory in the model beyond the autocorrelation induced by the hidden process, we incorporate feedback mechanisms by letting the state transition probabilities depend on the history. We consider two covariates related to the recent deep dives performed by the narwhal. One covariate is the continuous variable τ_t, defined as time passed since the last deep dive before dive number t, where a *deep dive* is defined as a dive to a depth of at least $350m$. Maximum depths are bimodal, and the value is chosen as a lower threshold of the deeper dives. Note that this definition is only used to define the covariates, and is not related to the decoding of states. The other covariate is the discrete variable d_t taking non-negative integer values, counting the number of consecutive deep dives that the narwhal has performed before dive number t. Thus, covariate τ_t measures physical time since last deep dive, whereas covariate d_t counts number of deep dives in a row, independently of time passed. Finally, we consider the covariate of the hour of the day at which the dive is initiated. More specifically, we define the covariate processes $\mathcal{T}_t$, the time since the last deep dive, D_t, the number of consecutive deep dives up to dive number t, and H_t, the hour of initiation of dive t, and denote the measured covariates by τ_t, d_t and h_t. Thus, the short term memory is modelled by the hidden states, and the long term memory is modelled by modulation of the transition probabilities as a function of past dynamics. The model is illustrated in Fig 2. Fig 3 illustrates the response variables and the three covariates for 60 consecutive dives.

The covariates enter the transition probabilities $\omega_{ij}(t) = \omega_{ij}(\eta_{ij}(t))$ in Eq (1) through a *predictor*, $\eta_{ij}(t)$, see Eq (7) below. We consider several models. If there are no covariates for a given predictor, then $\eta_{ij}(t) = \eta_{ij}$ does not depend on t. In S1 Table in the Supporting Information, all the covariate models that were fitted are listed, where α_{ij}, β_{ij}, γ_{ij}, δ_{ij}, θ_{ij} and ζ_{ij} are real parameters. Covariates d_t and τ_t were incorporated as natural cubic splines with three degrees of freedom. The effect of time of day is modelled by a periodic B-spline with three degrees of freedom, with boundary knots in 0 and 24 hours.

The likelihood function and optimization

The likelihood L_T of $x_1, x_2, \ldots, x_T$, where x_t is the observation of $\boldsymbol{X}_t$, assumed to be generated by an m-state HMM, can in general be computed recursively in only $O(Tm^2)$ operations by the forward algorithm [9]. The likelihood is expressed as

$$L_T = \delta\mathcal{P}(x_1)\boldsymbol{\Omega}(\tau_1, d_1, h_1)\mathcal{P}(x_2)\cdots\boldsymbol{\Omega}(\tau_{T-1}, d_{T-1}, h_{T-1})\mathcal{P}(x_T)\mathbf{1}, \tag{6}$$

where $\mathcal{P}(x_t) = \mathrm{diag}(f_1(x_t), \ldots, f_m(x_t))$ is a diagonal matrix with diagonal elements $f_i(x_t)$ given in Eq (4) when the gamma distribution is used, or Eq (5) when the log-normal distribution is used, $\boldsymbol{\Omega}$ is given by Eq (1) and $\mathbf{1} \in \mathbb{R}^m$ is a column vector of ones. The initial state distribution is denoted by δ, which is an m-dimensional row vector; $\delta_i = P(C_1 = i)$. For δ, we choose the uniform distribution, $\delta_i = 1/m$. Alternatively, it can be estimated, but there is no need for this extra computational effort, since our dataset is large and the influence of δ will be negligible.

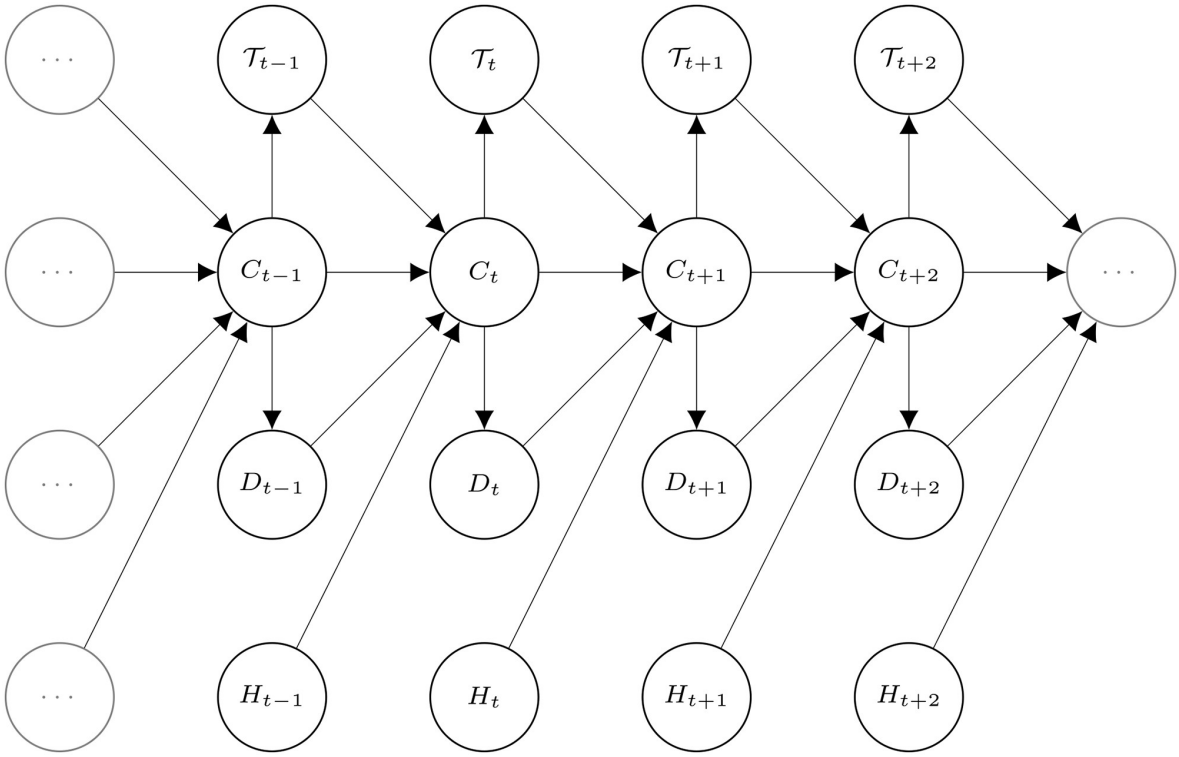

Fig 2. Hidden Markov Model with feedback processes. The transition probabilities between hidden states C_t depends on the observed covariate processes $\mathcal{T}_t$, D_t and H_t.

https://doi.org/10.1371/journal.pcbi.1006425.g002

To test this hypothesis, we repeated the optimization with the optimized parameters as initial condition, only changing the distribution of δ to the decoded distribution at time 1. This did not change the estimates. Furthermore, δ has no particular biological relevance.

The transition parameters in Eq (1) are constrained to be between 0 and 1 with row sums equal to 1, and thus, even if there are m^2 entries, there are only $m \cdot (m-1)$ free parameters. To obtain an unconstrained optimization problem, we reparametrise to working parameters, as also done in [13–15], see also [9], by defining

$$\omega_{ij}(t) = \frac{\exp(\eta_{ij}(t))}{\sum_{j=1}^{m} \exp(\eta_{ij}(t))} \tag{7}$$

where $\eta_{ij}(t)$ is the predictor for dive t for $1 \leq i, j \leq 3$, $i \neq j$, and $\eta_{ii} = 0$ for $i = 1, 2, 3$. This assures positive entries and that rows sum to 1.

We used the direct numerical Newton-Raphson algorithm `nlm` (`optim` in case `nlm` failed) in R [18] to estimate the parameters of the model by maximizing the log-likelihood, $\mathcal{L}_T := \log L_T$, where L_T is given in Eq (6). The procedure `ns` from the package `splines` (version 3.5.0) was used to calculate the natural cubic splines. The procedure `pbs` from the package `pbs` (version 1.1) was used to calculate the periodic splines.

Using a combination of R and Rcpp [19] for calculating the log-likelihood function $\mathcal{L}_T$ improved the runtime considerably. To mitigate the problem of local maxima, we ran the optimization algorithm up to a thousand times with different starting values for the parameters.

PLOS COMPUTATIONAL BIOLOGY

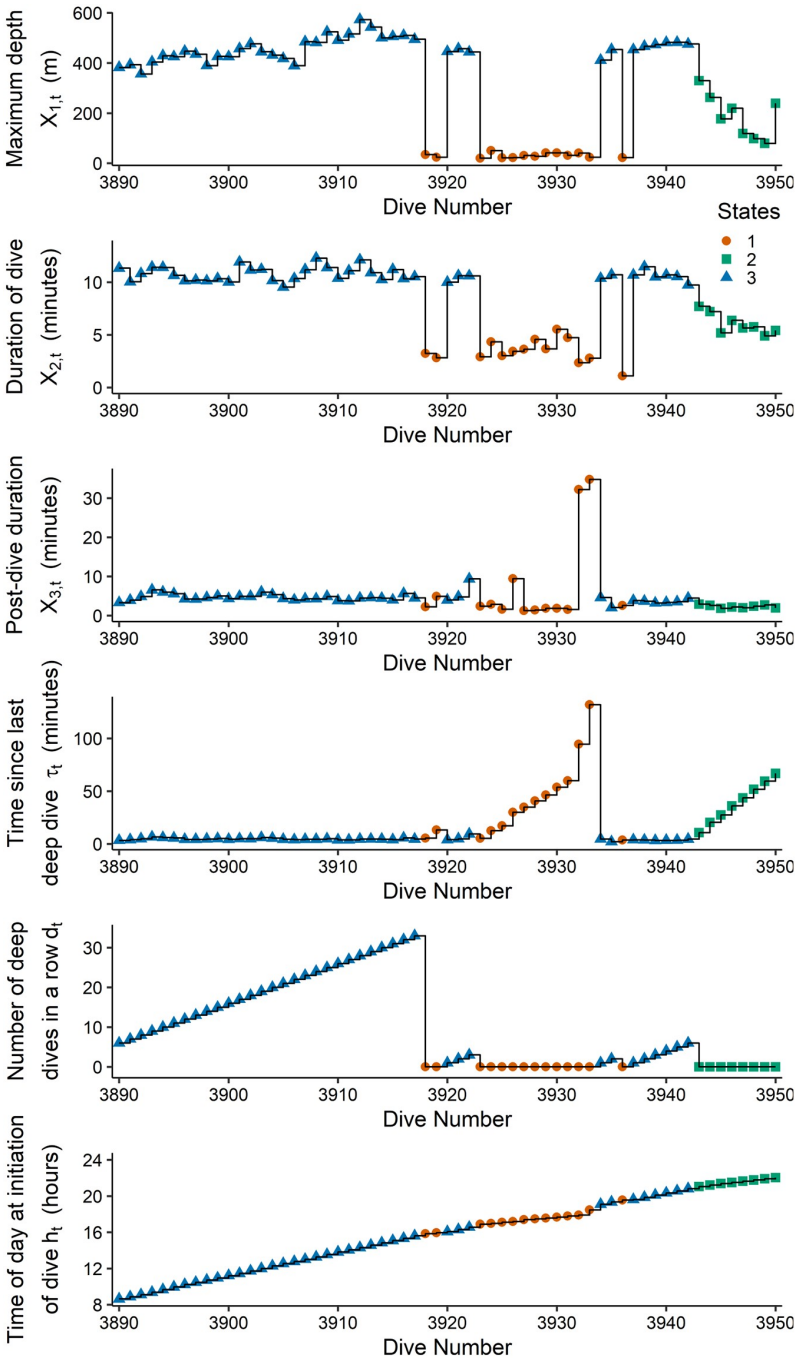

Fig 3. Response variables and covariate processes. Time series plot of maximum depth (MD), duration of dive (DT), and post-dive duration (PD) from dive number 3890 to 3950 and the covariate processes counting the time since last deep dive (τ_t), number of deep dives in a row (d_t), and the hour at initiation of dive (h_t). The symbols indicate the decoded hidden states from a model fitted to a dependent log-normal distribution (Model 1).

https://doi.org/10.1371/journal.pcbi.1006425.g003

PLOS COMPUTATIONAL BIOLOGY

The starting values were chosen as follows. For the parameters of the state-dependent distributions, an independent mixture model was fitted to the response distributions, and the estimated parameters were used as initial conditions. In the correlated models, the correlation parameter between MD and DT was initiated at the empirical correlation in the data set. The parameters of the covariates were varied in a regular grid together with the jittering procedure used in [14], such that they looped through 0 to ±5 in steps of 1 for α_{ij}, β_{ij} and γ_{ij}. The final result was chosen as the one giving the maximum log-likelihood.

The best model fit was evaluated by AIC. Once the optimal model was selected and parameters of the model were estimated, it was of interest to decode the most likely state sequence $c_1^*, \ldots, c_T^*$. The Viterbi algorithm [9, 21] was used to estimate the hidden states given the observed depths and durations:

$$(c_1^*, \ldots, c_T^*) = \underset{(c_1,\ldots,c_T)\in\{1,\ldots,m\}}{\operatorname{argmax}} \Pr(C_1 = c_1, \ldots, C_T = c_T \mid x_0, \ldots, x_T).$$

Results

The data set covers 1,995 hours (∼ 83 days) with $T = 8,609$ dives, and is extraordinarily long, and thus provides a unique opportunity to obtain detailed information on diving behaviour. An example of the data is shown in Fig 4. Such data are usually only on the order of a couple of days or less, for example, the time series of short-finned pilot whales *(Globicephala macrorhynchus)* analysed in [15] cover up to 18 hours and 64 dives, whereas the time series of blue whales *(Balaenoptera musculus)* analysed in [14] cover up to 6 hours and 67 dives, and Langrock et al. [12] analyses 79 hours of a single Blainville's beaked whale. Detailed diving data of narwhals are available for up to 33 hours [6] or up to one week [7]. However, here we only have data from a single narwhal limiting the generalizability of the analysis.

The first week of tagging, the narwhal also had the temperature of the stomach measured, see [22]. A temperature drop indicates that a prey has entered the stomach. The red parts in Fig 4 indicate temperature drops. These typically happen during deep dives, and support the assumption that deep dives are related to foraging. This is also supported by the findings in [7], where buzzes, related to foraging, are typically produced when the whales are at 200–600m.

The variable MD takes values between 20 and 910.5m, DT takes values between 33 seconds and 28 minutes, and PD takes values between 1 second and 209.7 minutes. Fig 5 shows

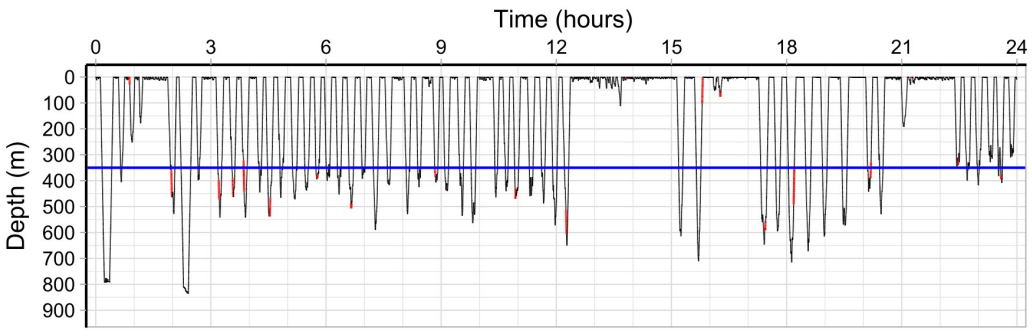

Fig 4. Diving data. Representative part of the narwhal diving data, covering 24 hours of dives on August 15th 2013. The red parts are where a lower temperature in the stomach has been registered, indicating that the narwhal has swallowed a prey. The blue line indicates a depth of 350m, the threshold for a *deep dive* used in the definition of the covariates.

https://doi.org/10.1371/journal.pcbi.1006425.g004

PLOS COMPUTATIONAL BIOLOGY

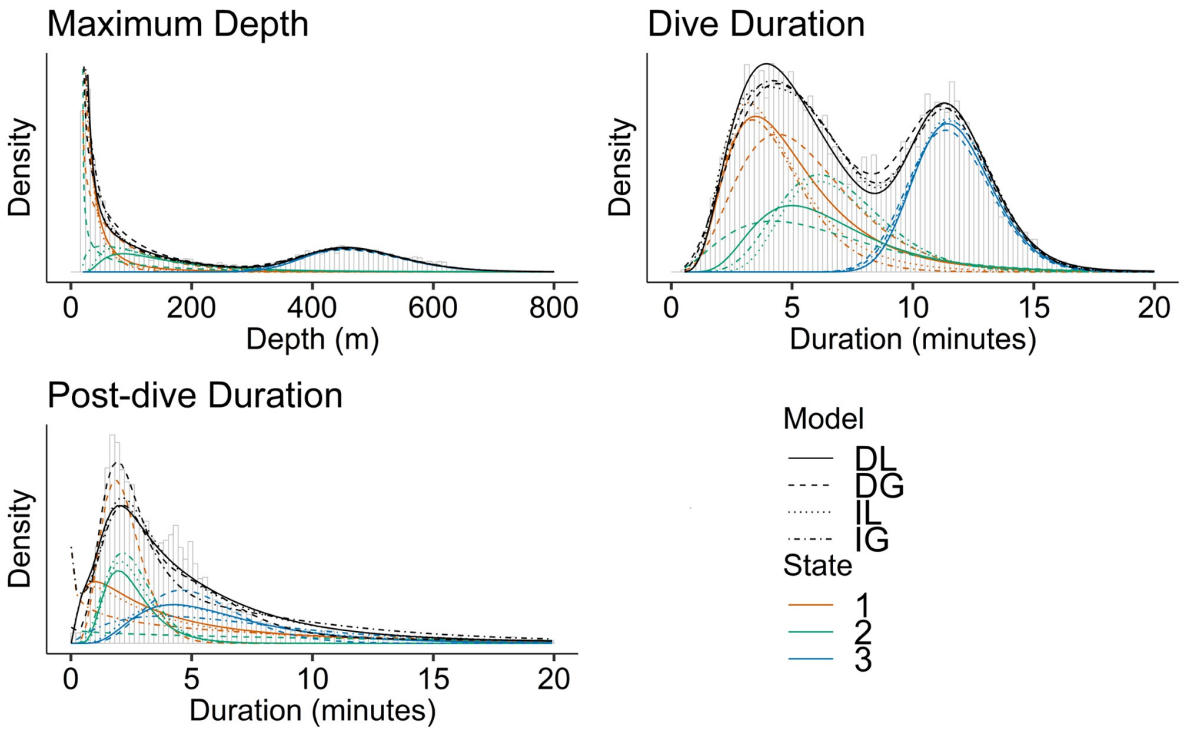

Fig 5. Model fit. Histograms of response variables MD, DT and PD. The fit of Model 1 is indicated with black curves, for dependent lognormal (DL), independent lognormal (IL), dependent gamma (DG) and independent gamma (IG). The distribution of the fitted states are indicated with colours as given in the legend. State 1 corresponds to near surface, state 2 medium depths, and state 3 large depths.

https://doi.org/10.1371/journal.pcbi.1006425.g005

histograms of the three response variables. Maximum depths are bimodal and typically either less than 200*m* or between 400 and 600*m*. This was used to select the threshold of 350*m* to define a deep dive. The value is chosen as a lower threshold of the deeper dives. We furthermore tried different values between 250 and 450*m* in steps of 50*m*. The results only changed very little within this range, and thus, the analysis is robust to the choice of threshold.

To choose the number of states *m*, we optimized models with each of the four state distributions for $m = 2, 3$ and 4 states, including all covariates. Since the gamma model is computationally very expensive, and furthermore does not provide a better fit, we only ran the gamma models for $m = 2$ and 3. Typical runtimes are given in Table 1. The runtimes vary over many orders of magnitudes. For all state distributions, the 4-state model takes on the order of hours to run, which makes it infeasible, since for each covariate model, many repetitions from different starting conditions have to be run, and the number of needed repetitions explode as the number of parameters increase. Moreover, the 4-state model did not improve qq-plots, as shown later. The 3-state correlated gamma model is also very slow and not feasible to use if many covariate models should be explored. In general, the log-normal model is much faster than the gamma model, and the computational cost of including dependence is small. It is not obvious if a 2 or a 3-state model should be chosen. However, the runtimes for the 3-state model are acceptable, and based on both qq-plots and AIC values presented below, the 3-state HMM is preferred. Thus, similar to the blue whales data analysed in [14], our narwhal data suggest three distinct states. Pohle et al. [23] recommended against using more than four states in biological modelling like this, in order to avoid the complexity of the correspondence

Table 1. Complexity of models. Runtimes and number of variables for different state distributions and for 2, 3 and 4 states for covariate model 1. Runtimes are on Intel Xeon E5-2697v2 @ 2.7 GHz.

	No. of variables	Range of runtime	Average of runtime
Correlated log-normal			
2-state	28	0.9–3 (min)	1.9 (min)
3-state	63	2.25–15 (min)	7.3 (min)
4-state	112	1–2.4 (hrs)	1.6 (hrs)
Correlated gamma			
2-state	28	1.14–9.30 (min)	5.65 (min)
3-state	63	17.25–86.81 (min)	54.88 (min)
Independent log-normal			
2-state	26	0.11–3.66 (min)	1.28 (min)
3-state	60	3–12.56 (min)	5.7 (min)
4-state	108	1–3 (hrs)	1.88 (hrs)
Independent Gamma			
2-state	26	1.58–3.57 (min)	2.43 (min)
3-state	60	11.81–26.35 (min)	15.53 (min)

https://doi.org/10.1371/journal.pcbi.1006425.t001

between states of the model and the biological phenomenon. DeRuiter et al. [14] suggested three states for their data, even if a formal model selection procedure would point to a more complex model, because models with more underlying states might obscure patterns in the data and provide less insight in the underlying biological process, even if they might perform better in terms of forecasting. Biological knowledge should guide the choice of number of states. They also argue that model misspecifications, such as too inflexible state dependent distributions, variations over time, missing covariate information or outliers might cause model selection criteria to favour models with more complex structures than warranted. Therefore, we choose the 3-state HMM. The algorithm allocates labels arbitrarily, so to compare across models we relabelled the states, such that state 1 represents the shortest and shallowest dives, which we interpret as near-surface travelling, social activities and resting, state 2 represents medium long and deep dives, which we identify with a feeding state for prey located at medium depths, and state 3 represents the deepest and longest dives, which we identify with a feeding state for prey located at deep depths.

The empirical correlations between response variables in the full data set are small for MD and PD (0.046), and for DT and PD (0.042), only the correlation between MD and DT is significant (0.86). If the data set is split into three subsets according to MD, namely for MD between 20 and 50 m, for MD between 50 and 350 m, and for MD above 350 m, these results still hold. All correlations involving PD in all groups are less than 0.11 in absolute values, whereas the correlations between MD and DT are 0.27, 0.58 and 0.41, respectively. We therefore only assumed dependence between MD and DT. This improved convergence and runtime. To check that this assumption is reasonable, covariate model 1 with 3 states was fitted to the fully correlated log-normal model, and all estimated correlations with PD were smaller than 0.14, except for state 2, where it was around 0.5. The other estimates did not change compared to a model with only correlation between MD and DT.

We tried a total of 14 covariate models, listed in S1 Table in the Supporting Information. Here, we only include the best model based on the AIC criteria (model 1), and 3 more models for illustration (Table 2).

Model 1 has diurnal effects on all transition probabilities, and nonlinear effects of τ_t and d_t on some of the transition probabilities. The covariate d_t counts number of deep dives in a row,

PLOS COMPUTATIONAL BIOLOGY

Table 2. Different models for covariate effects on the transition probabilities between behavioural states. The predictors η_{ij} relate to the transition probabilities as given in Eq (7). The spline effects of hour are denoted by $H_{ij}^t = \sum_k \delta_{ij}^{(k)} h_k^t$, of τ_t by $T_{ij}^t = \sum_k \theta_{ij}^{(k)} s_k^t$, and of d_t by $D_{ij}^t = \sum_k \zeta_{ij}^{(k)} d_k^t$ for $k = 1, 2, 3$ and $i, j = 1, 2, 3$; $i \neq j$. A list of all explored models can be found in S1 Table in the Supporting Information.

	Predictors in the transition probabilities					
Model	$\eta_{12}(t)$	$\eta_{13}(t)$	$\eta_{21}(t)$	$\eta_{23}(t)$	$\eta_{31}(t)$	$\eta_{32}(t)$
1	$\alpha_{00} + T_{12}^t + H_{12}^t$	$\alpha_{01} + T_{13}^t + H_{13}^t$	$\beta_{00} + T_{21}^t + H_{21}^t$	$\beta_{01} + T_{23}^t + H_{23}^t$	$\gamma_{00} + D_{31}^t + H_{31}^t$	$\gamma_{01} + D_{32}^t + H_{32}^t$
2	$\alpha_{00} + H_{12}^t$	$\alpha_{01} + H_{13}^t$	$\beta_{00} + H_{21}^t$	$\beta_{01} + H_{23}^t$	$\gamma_{00} + H_{31}^t$	$\gamma_{01} + H_{32}^t$
3	$\alpha_{00} + T_{12}^t$	$\alpha_{01} + T_{13}^t$	$\beta_{00} + T_{21}^t$	$\beta_{01} + T_{23}^t$	$\gamma_{00} + D_{31}^t + H_{31}^t$	$\gamma_{01} + D_{32}^t + H_{32}^t$
4	$\alpha_{00} + T_{12}^t$	$\alpha_{01} + T_{13}^t$	$\beta_{00} + T_{21}^t$	$\beta_{01} + T_{23}^t$	$\gamma_{00} + D_{31}^t$	$\gamma_{01} + D_{32}^t$

https://doi.org/10.1371/journal.pcbi.1006425.t002

and is therefore around 0 when not in state 3. This covariate therefore carries no information unless in state 3, and only enters in η_{31} and η_{32}. Likewise, τ_t is expected to be around 0 when in state 3, and therefore only enters η_{ij} for $i = 1$ or 2. Model 2 only has diurnal effects. Model 3 has effects of the dive covariates, but only diurnal effects in state 3. Finally, model 4 has only dive effects and no diurnal effects.

Table 3 lists the model selection results from the optimization. We use AIC to select the best model, which is highlighted in bold. The correlated log-normal model is clearly preferred above the other models, with huge AIC differences. The dependent models are clearly preferred above the independent models, and the log-normal distribution is clearly preferred above the gamma distribution. Models with ΔAIC larger than 10 have essentially no support in the data compared to the best model [24]. Model 1 is the best among the tested models for all state distribution models, which balance accuracy and complexity of the model. The marginal fit of covariate model 1 is illustrated in Fig 5 for the four state distributions, where the black curves provide the overall distributions of the three response variables, as well as the distributions within each state. The fits look convincing for MD and DT, whereas the models capture the bimodality of PD less well. Note that the splitting into states 1 and 2 depends on the state distributions, whereas the distributions of state 3 are approximately the same for all state distributions. Thus, the classification of behavioral states will depend on the chosen state distribution mainly for small and medium dives.

To check the fit of the model beyond what is presented in Fig 5, we calculated the pseudo-residuals [9] and made qq-plots (Fig 6) for the correlated log-normal model with $m = 2$, 3 and 4 states. The other state distributions give similar qq-plots, and are therefore omitted. A slight improvement is observed when passing from 2 to 3 states, in particular for PD. The fit does not improve when passing from 3 to 4 states. The fit is acceptable for MD and DT, maybe except for a too small lower tail for the MD. This is probably due to the threshold of a depth of

Table 3. Model selection results. Differences in AIC values, ΔAIC = AIC—AIC_{min}, between the different models with 3 hidden states, where AIC_{min} is the value of the model with the lowest AIC. The best fit is given by the minimum AIC. For all the tested state distributions, covariate model 1 was preferred, and for all covariate models, the dependent log-normal state distribution was preferred. Because the runtimes for the correlated gamma model are high, only Model 1 was fitted. The best model is highlighted in bold. *np*: number of parameters.

	Independent Gamma distribution		**Independent Log-normal distribution**		**Correlated Gamma distribution**		**Correlated Log-normal distribution**	
Model	***np***	**ΔAIC**	***np***	**ΔAIC**	***np***	**ΔAIC**	***np***	**ΔAIC**
1	60	5050.5	60	2309.1	63	1901.9	**63**	**0**
2	42	5386.9	42	2652.4	45	-	45	256.0
3	48	5096.5	48	2353.2	51	-	51	34.3
4	42	5194.4	42	2451.9	45	-	45	166.9

https://doi.org/10.1371/journal.pcbi.1006425.t003

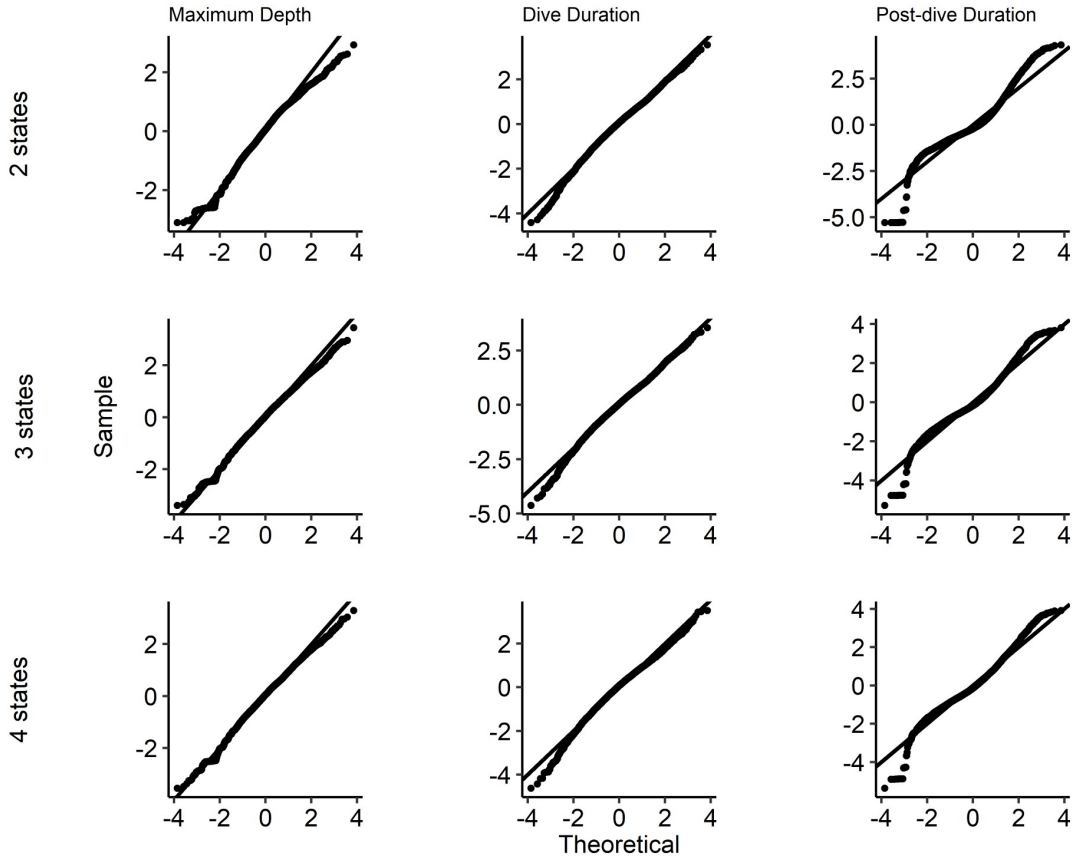

Fig 6. Quantile-quantile residual plots. QQ-plots of forecast pseudo-residuals from covariate model 1 with correlated log-normal state distribution.

https://doi.org/10.1371/journal.pcbi.1006425.g006

$20m$ in the definition of a dive. The PD is less well fitted, especially in the lower tail, which could also be partly due to the cut-off threshold of $20m$ in the definition of PD. It is acceptable for 3 and 4 states.

Fig 7 illustrates the estimated covariate effects for the optimal model, the correlated log-normal state distributions with covariate model 1. Parameter estimates and confidence intervals can be found in S2 and S3 Tables in the Supporting Information.

The covariate τ_t indicates the time passed since last deep dive. We expect that τ_t has impacts on states 1 and 2, but not on state 3 (which is the case for the selected model). In the left panel of Fig 7A the effect of τ_t is illustrated. The transition probabilities do not seem to depend much on τ_t, except for the probability of changing from state 1 to state 3. The probability is higher for small values of τ_t, and decreasing fast towards 0 for larger values. This is not what was expected, but might reflect the following. When short time has passed since last deep dive, it was probably also a short time since the whale was in state 3. Thus, it reflects that the whale is still in an overall behavioral state 3, but just had a short break in state 1. This phenomenon can be seen in Fig 8 where the state decoding is shown for 12 representative hours. It is seen that after (at least) six dives in state 3, the whale changes to a few shallow dives for a short time, and then continues with another three dives in state 3. When a little longer time passes, the whale has effectively stopped diving deep, and the probability of a change to state 3 becomes smaller.

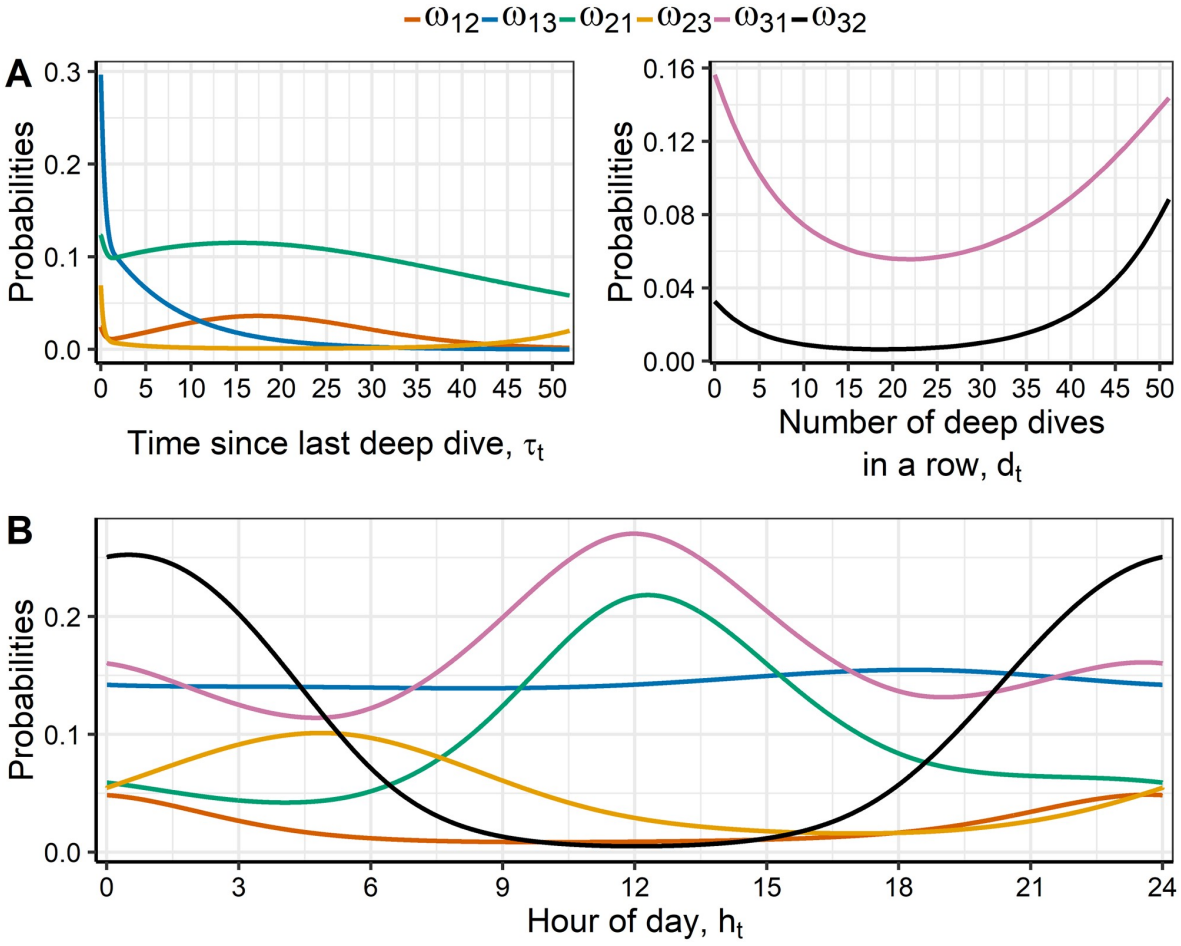

Fig 7. Covariate effects. A: Transition probabilities between behavioural states depending on covariates related to deep dives of correlated log-normal model 1, at approximately 12 pm. B: Transition probabilities depending on diurnal effects in model 1 with correlated log-normal state distributions, calculated for $\tau_t = 0.58$ and $d_t = 0$ (the medians).

https://doi.org/10.1371/journal.pcbi.1006425.g007

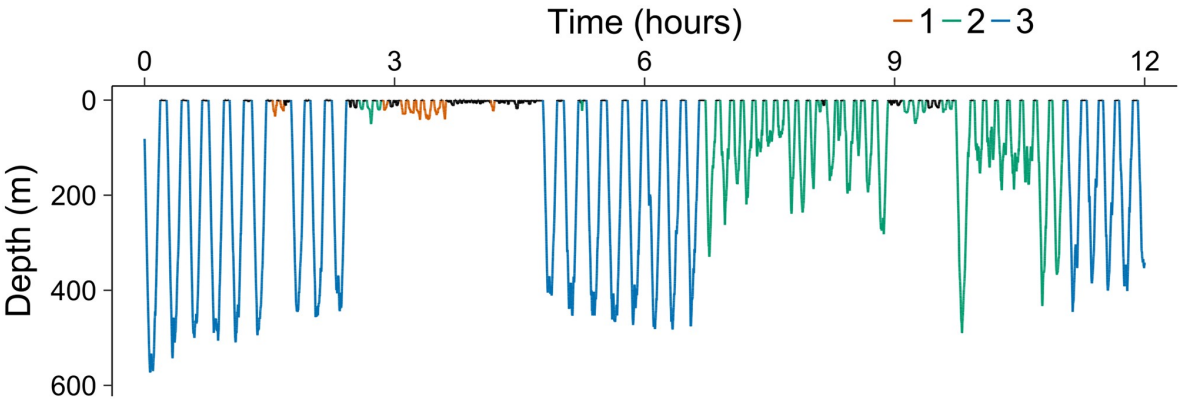

Fig 8. State decoding close-up. The estimated hidden state per dive for 12 hours of the data, starting on 22 September 2013 at 14:18:39.

https://doi.org/10.1371/journal.pcbi.1006425.g008

Then, when long time has passed, we expect the transition probability to increase, which is not what is estimated. However, there are few large observations of τ_t: 75% of the values are below 2.8 hours, and 90% are below 7.8 hours. Therefore, the estimates of covariate effects for large values are unreliable. The effect of d_t is illustrated in the right panel of Fig 7A. As expected, for values above 20 dives in a row, the probabilities to exit state 3 increase with increasing d_t. However, the data is sparse for large values of d_t and estimates might not be trusted: more than half are 0, 75% are 2 or smaller, and 90% are 8 or lower. The probability of changing to state 1 is much higher than the probability of changing to state 2 after a period in state 3.

Fig 7B shows the diurnal effects on the transition probabilities. Changing from state 3 to 2 has highest probability around midnight, whereas changing from state 2 to 3 has highest probability around 6 am. Changing to state 1 has highest probability around noon. The transition probabilities from state 1 do not depend much on diurnal effects.

Table 4 lists the estimated means and standard deviations of the four state distributions. Means and standard deviations of maximum depth are estimated larger for both state 1 and state 2 with the correlated models compared to the independent models, whereas all models estimate mean and variances approximately the same for state 3. Thus, taking into account the dependence between the two state variables reveals more variable diving patterns (i.e., larger variance within states), unless the narwhal is doing deep dives in state 3, where the need for regular breathing do not allow the whale to make detours. In general, the distributions of the response variables within states change depending on the assumed state distributions, and whether correlation is accounted for or not. To understand the classification of behavioural states provided by the HMM, we also added the empirical measures from the data decomposed into three subsets according to maximum depth: state 1 defined as dives between 20 and 50m, state 2 defined as dives between 50 and 350m, and state 3 for dives of more than 350m. This shows that none of the HMMs classifies the dives only according to depth, since these empirical measures differ from all the estimated distributions. Thus, the HMMs might reveal more complex behavioural states than given by the diving depths.

The Viterbi algorithm classifies each dive to one of the three hidden states. The classification depends on the model, but all models roughly group dives according to maximum depth. One goal of comparing models is to access if conclusions on diving behaviour expressed through the decoded classes of the dives differ between models. If they all classify the same, it does not matter which model we use, maybe except for the estimation of covariate effects. If the classification differ from model to model, it is important to choose the statistically best model, measured from AIC, qq-plots, runtimes and biological interpretability.

Fig 9 shows the decoded hidden states for Model 1 with dependent log-normal state distribution. The correlated log-normal model estimates that the narwhal spends around 43.7% of its dives, corresponding to 28.8% of the time in State 1, which encompasses dives down to 793m of durations up to 28 minutes. This is a large value for the surface state, but it is only the extreme tail of the distribution, and is represented by a single dive. It reflects that the log-normal distribution has heavier tails than the gamma distribution, and that the behavioural states are more complex than what can be explained only by maximum depth. Of the time spent in state 1, only 15.9% of the time is spent diving, the rest of the time the whale is at the surface. The narwhal spends around 22.4% of its dives, corresponding to 19.2% of the time, in medium depths of between 22.5m and 836m and durations between 0.8 and 21.3 minutes. Also here, a few deep dives are decoded as belonging to state 2. Of the time spent in state 2, 10.6% of the time is spent diving, the rest of the time the whale is at the surface. Finally, 33.9% of dives, corresponding to 52.1% of the time, are spent in state 3 at depths between 243m and 910.5m and durations between 7.2 and 19.5 minutes. Of the time spent in state 3, 28.9% of the time is spent diving, the rest of the time the whale is at the surface. Fig 8 illustrates a close-up of the

Table 4. Summary measures of Model 1 with 3 states. Means and standard deviations based on correlated Log-normal, correlated Gamma, independent Log-normal and independent Gamma distribution. MD: Maximum Depth; DT: Diving Time; PD: Post-Dive duration. E: mean; SD: standard deviation; $Corr_1$: Correlation between MD and and DT. $Corr_2$: Correlation between MD and and PD. $Corr_3$: Correlation between DT and and PD. The empirical distribution is the empirical measures in three subgroups of the data classified according to MD, state 1: MD between 20 and 50 m, state 2: MD between 50 and 350 m, state 3: MD above 350 m.

	State 1	State 2	State 3
Correlated Log-normal distribution			
E_{MD}	51.04	174.19	479.29
SD_{MD}	57.54	109.09	81.36
E_{DT}	5.05	6.54	11.79
SD_{DT}	2.61	2.52	1.65
E_{PD}	7.56	2.58	6.93
SD_{PD}	14.85	1.23	7.45
$Corr_1$	0.56	0.81	0.46
Correlated Gamma distribution			
E_{MD}	88.46	112.37	471.81
SD_{MD}	78.60	153.96	83.03
E_{DT}	5.50	5.95	11.60
SD_{DT}	2.49	3.48	1.72
E_{PD}	2.19	16.03	5.36
SD_{PD}	0.87	20.43	2.29
$Corr_1$	0.59	0.80	0.53
Independent Log-normal distribution			
E_{MD}	42.68	150.29	477.37
SD_{MD}	34.53	89.87	83.50
E_{DT}	4.37	7.00	11.81
SD_{DT}	2.11	2.07	1.70
E_{PD}	7.64	2.66	7.18
SD_{PD}	14.77	1.25	8.99
Independent Gamma distribution			
E_{MD}	39.47	133.14	474.87
SD_{MD}	25.86	87.27	85.80
E_{DT}	4.10	6.85	11.77
SD_{DT}	1.81	2.23	1.72
E_{PD}	8.15	2.55	7.34
SD_{PD}	15.13	1.12	9.65
Empirical distribution			
E_{MD}	30.87	143.52	484.67
SD_{MD}	8.45	86.16	77.14
E_{DT}	4.25	6.73	11.83
SD_{DT}	2.04	2.43	1.72
E_{PD}	7.07	4.43	7.18
SD_{PD}	13.99	8.45	9.44
$Corr_1$	0.27	0.58	0.41
$Corr_2$	-0.11	0.07	0.05
$Corr_3$	-0.01	0.08	0.06

https://doi.org/10.1371/journal.pcbi.1006425.t004

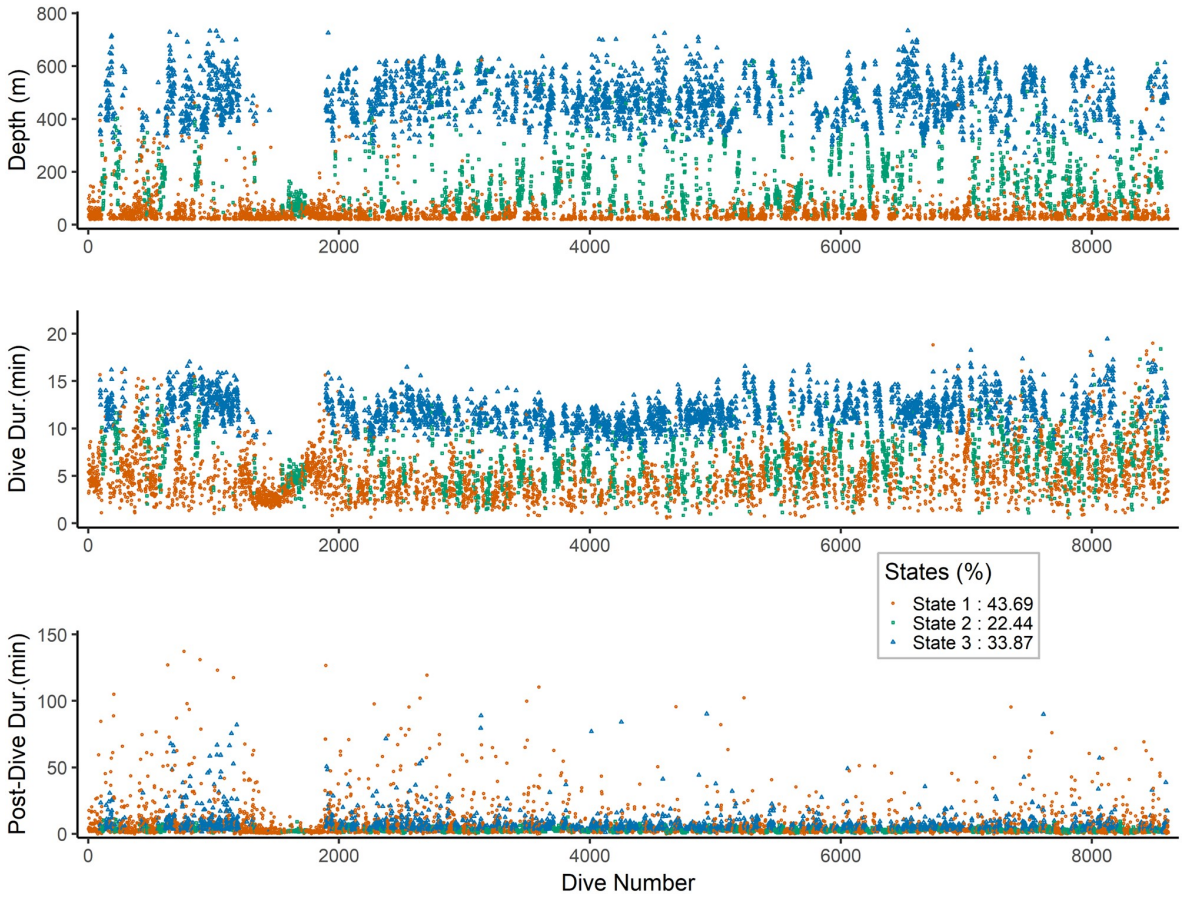

Fig 9. State decoding. The estimated hidden state per dive for each of the three observed variables under covariate model 1 and state distribution the correlated log-normal. The longest pause of no deep dives starts from the 1345th dive until the 1894th dive, and it lasts approximately 2 days and 17.5 hours.

https://doi.org/10.1371/journal.pcbi.1006425.g009

decoding of dives for an example period of 12 hours. The correlated model thus decodes a few of the deep dives as pertaining to states 1 and 2, probably because of these dives taking longer time than the deep dives decoded as state 3.

Apparently the whale could stay in state 1 and 2 for long periods (> 24 hours) without transiting to state 3, and it even showed a pause of almost 3 days without deep dives, see Fig 9 for dives 1345-1894. This indicates that feeding occurs infrequently and that narwhals at least during summer and fall may have extended periods without feeding activity (see also [6]). However, the median of these pauses without state 3 dives was 44 minutes and the mean was 2 hours.

Discussion

In this study, we investigate different multivariate HMMs with covariate effects for modelling the diving activity of a narwhal in the vertical dimension in the water column. Although narwhals show relatively little behavioural plasticity [6, 7, 16], the present analysis is based on a sample of only one individual and there is therefore obvious limits to how far reaching conclusions that can be drawn from the diving behaviour of this individual. However, the value in

the present analysis is the extraordinarily long data set and it is therefore also useful for examining the application of HMM methods as a tool for analyzing ontogenetic diving activity. The value of the sample includes the option for describing diurnal patterns in diving behaviour, during the fall migration.

We extend the existing HMMs for diving behaviour of marine mammals to allow for dependence between state distributions, and show that the dependence has some impact on the conclusions drawn about the diving behaviour. We find that statistically the correlated model outperforms the independent model, that the log-normal model outperforms the gamma model, and more importantly, conclusions on the diving behaviour differ between the models. The main differences are that the correlated models estimate more variable state distributions of MD and DT compared to the uncorrelated models. Thus, a major biological insight from the analysis of the correlated model is that variability is larger in behavioural states 1 and 2, but not in state 3. In the dependent log-normal model 56.3% of the dives are for feeding, compared to 60.5% in the independent log-normal model, under the assumption that states 2 and 3 in fact are representing feeding states in both models. Even if it is only a proportion of the dives that are not for feeding, it can be assumed that it is approximately the same proportion for the correlated and the independent models, and it is still a relatively large proportion of the diving effort that is allocated to feeding activities. This provides an important ecological insight that is useful when comparing feeding activities for whales inhabiting different ocean parts with different prey availability. Finally, ignoring the dependence between response variables leads to wrongly estimated standard deviations on parameter estimates, and thus confidence intervals are no longer valid.

The correlations between the post-dive duration and diving depth and duration are found to be vanishing. However, the post-dive response variable probably covers different behaviours that can not be distinguished from this data, such as recovering from a deep dive, resting between bouts of dives, social activities, travelling, etc.

Direct observations of feeding events were limited to the first week of the diving data but the depths where feeding events were detected served as a valid proxy for the depth threshold between behavioural state 2 and state 3. The observation that feeding events involve deep dives ($\geq 350m$) is also supported by studies of the buzzing activity during dives to different depths for narwhals travelling in the same area and time of the year as the whale included in this study [7].

Transition from state 1 to presumed feeding activity is more likely to be to state 3 with deep dives, and rarely goes to state 2 from state 1. Diving activity in state 3 usually last for a series of dives (5-10) perhaps indicating that specific layers of prey is being detected and explored for a series of dives before the whale needs to spend an extended period at the surface. The post dive time is typically around 6.9 minutes after a state 3 dive, whereas it is typically only 2.6 minutes after a state 2 dive. The whale probably needs to spend more time at the surface to recover from nitrogen tissue tension following a longer breath-hold diving activity. Williams et al. (2011) [25] calculated that the oxygen stores in tissues from narwhals of similar size as the one in this study would support dives of less than 20 min and that energy saving during gliding on descent might increase this calculated aerobic dive limit to up to 24 min. The deep dives in state 3 in this study seem to be in good agreement with these physiological limitations.

Even though detailed dive information supplemented by data on feeding events have been available for this analysis it may still not be adequate for describing the important drivers of diving behaviour. Both physiological constrains and reproductive state as well as environmental conditions may influence the diving activity to an extent that cannot be fully discerned in HMM analysis of dive series. For logistical reasons it is very difficult if not impossible to obtain information on all factors that affect the diving behaviour. However, the analysis of dive series

provides a minimal insight into the integrated effect of the various factors driving the diving behaviour and the major advantage of the HMM analysis probably relies in the objective inter- and intra-specific comparison of diving activity. This study demonstrated the usefulness of HMMs for gaining insight to the hidden structures of dive patterns, something that is difficult to achieve with traditional statistics. It will be important to apply HMM techniques to larger data sets of diving activity from several whales to estimate how effective HMMs are for providing broader ecological insight to energetics and multispecies effects of whale predation.

Supporting information

S1 Table. Different models for covariate effects on the transition probabilities between behavioural states. The predictors η_{ij} relate to the transition probabilities.
(PDF)

S2 Table. Estimates of the model parameters of the state distributions and their 95% confidence intervals in model 1 for correlated log-normal distribution. In state i, μ_i and σ_i are the log-mean and log-standard deviation of the correlated log-normal distribution. Index MD stands for Maximum Depth, DT stands for Dive Duration and PD stands for Post-Dive time. The depth is measured in meters, and time in seconds. The confidence intervals were computed from the Hessian of the negative log-likelihood function, i.e., based on the inverse of the observed Fisher information.
(PDF)

S3 Table. Estimates of the model parameters of covariate effects and their 95% confidence intervals in model 1 for correlated log-normal distribution. The spline effects of hour are denoted by $H_{ij}^t = \sum_k \delta_{ij}^{(k)} h_k^t$, of τ_t by $T_{ij}^t = \sum_k \theta_{ij}^{(k)} s_k^t$, and of d_t by $D_{ij}^t = \sum_k \zeta_{ij}^{(k)} d_k^t$ for $k = 1, 2, 3$ and $i, j = 1, 2, 3$; $i \neq j$.
(PDF)

S1 Data. Data analyzed in the paper. Data columns are: `DiveNumber`: Number of dive; `Date`: Date of dive; `StartTime`: Start time in hh:mm:ss of dive; `MaxDepth`: Maximum depth reach in dive in meters; `Duration`: Duration of dive in minutes; `PostDiveDur`: Duration of time spent in the surface (above 20 m) after the dive in minutes.
(ZIP)

Author Contributions

Conceptualization: Manh Cuong Ngô, Mads Peter Heide-Jørgensen, Susanne Ditlevsen.

Data curation: Mads Peter Heide-Jørgensen.

Formal analysis: Manh Cuong Ngô.

Funding acquisition: Mads Peter Heide-Jørgensen.

Investigation: Manh Cuong Ngô, Susanne Ditlevsen.

Methodology: Manh Cuong Ngô, Susanne Ditlevsen.

Resources: Mads Peter Heide-Jørgensen.

Software: Manh Cuong Ngô.

Supervision: Mads Peter Heide-Jørgensen, Susanne Ditlevsen.

Visualization: Manh Cuong Ngô.

Writing – original draft: Manh Cuong Ngô, Mads Peter Heide-Jørgensen, Susanne Ditlevsen.

Writing – review & editing: Manh Cuong Ngô, Mads Peter Heide-Jørgensen, Susanne Ditlevsen.

References

1. NAMMCO. Report of the NAMMCO Global Review of Monodontids. 13-16 March 2017, Hillerød, Denmark; 2018.
2. Heide-Jørgensen MP. Narwhal *Monodon monoceros*. In: Perrin WF and Wursig B and Thewissen JGM, editor. Encyclopedia of Marine Mammals, 2nd Edition; 2009. pp. 754–758.
3. Schorr GS, Falcone EA, Moretti DJ, Andrews RD. First long-term behavioral records from Cuvier's beaked whales (*Ziphius cavirostris*) reveal record-breaking dives. PLoS ONE. 2014. https://doi.org/10.1371/journal.pone.0092633 PMID: 24670984
4. Watkins WA, Daher MA, Fristrup KM, Howald TJ, Di Sciara GN. Sperm Whales Tagged with Transponders and Tracked Underwater by Sona. Marine Mammal Science. 1993;(1):55–67. https://doi.org/10.1111/j.1748-7692.1993.tb00426.x
5. Heide-Jørgensen MP, Dietz R, Leatherwood S. A note on the diet of narwhals (*Monodon monoceros*) in Inglefield Bredning (NW Greenland). Meddr Grønland, Biosci. 1994; 39:213–216.
6. Laidre K, Heide-Jørgensen M. Winter feeding intensity of narwhals (*Monodon monoceros*). Marine Mammal Science. 2005; 21(1):45–57. https://doi.org/10.1111/j.1748-7692.2005.tb01207.x
7. Blackwell S, Tervo O, Conrad A, Sinding MHR, Ditlevsen S, Heide-Jørgensen MP. Spatial and temporal patterns of sound production in East Greenland narwhals. PLoS ONE. 2018; 13(6) https://doi.org/10.1371/journal.pone.0198295
8. Reeves RR, Ewins PJ, Agbayani S, Heide-Jørgensen MP, Kovacs KM, Lydersen C, et al. Distribution of endemic cetaceans in relation to hydrocarbon development and commercial shipping in a warming Arctic. Marine Policy. 2014; 44:375–389. https://doi.org/10.1016/j.marpol.2013.10.005
9. Zucchini W, MacDonald IL, Langrock R. Hidden Markov Models for Time Series: An Introduction using R. 2nd ed. Chapman & Hall/CRC, FL, Boca Raton; 2016.
10. Langrock R, Marques TA, Baird RW, Thomas L. Modeling the Diving Behavior of Whales: A Latent-Variable Approach with Feedback and Semi-Markovian Components. Journal of Agricultural Biological and Environmental Statistics. 2014; 19(1):82–100. https://doi.org/10.1007/s13253-013-0158-6
11. Patterson TA, Basson M, Bravington MV, Gunn JS. Classifying movement behaviour in relation to environmental conditions using hidden Markov models. Journal of Animal Ecology. 2009; 78(6):1113–1123. https://doi.org/10.1111/j.1365-2656.2009.01583.x PMID: 19563470
12. Langrock R, King R, Matthiopoulos J, Thomas L, Fortin D, Morales JM. Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. Ecology. 2012; 93(11):2336–2342. https://doi.org/10.1890/11-2241.1 PMID: 23236905
13. Michelot T, Langrock R, Bestley S, Jonsen ID, Photopoulou T, Patterson TA. Estimation and simulation of foraging trips in land-based marine predators. Ecology. 2017; 98(7):1932–1944. https://doi.org/10.1002/ecy.1880 PMID: 28470722
14. DeRuiter SL, Langrock R, Skirbutas T, Goldbogen JA, Calambokidis J, Friedlaender AS, et al. A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. Annals of Applied Statistics. 2017; 11(1):362–392. https://doi.org/10.1214/16-AOAS1008
15. Quick NJ, Isojunno S, Sadykova D, Bowers M, Nowacek DP, Read AJ. Hidden Markov models reveal complexity in the diving behaviour of short-finned pilot whales. Scientific Reports. 2017; 7. https://doi.org/10.1038/srep45765
16. Heide-Jørgensen MP, Nielsen NH, Hansen RG, Schmidt HC, Blackwell SB, Jørgensen OA. The predictable narwhal: satellite tracking shows behavioural similarities between isolated subpopulations. Journal of Zoology. 2015; 297(1):54–65. https://doi.org/10.1111/jzo.12257
17. Aguilar Soto N, Johnson MP, Madsen PT, Díaz F, Domínguez I, Brito A, et al. Cheetahs of the deep sea: deep foraging sprints in short-finned pilot whales off Tenerife (Canary Islands). Journal of Animal Ecology. 2008; 77(5):936–947. https://doi.org/10.1111/j.1365-2656.2008.01393.x PMID: 18444999
18. R Core Team. R: A Language and Environment for Statistical Computing; 2017. Available from: https://www.R-project.org/.
19. Eddelbuettel D, Francois R. Rcpp: Seamless R and C plus plus Integration. Journal of Statistical Software. 2011; 40(8):1–18. https://doi.org/10.18637/jss.v040.i08

20. Moran P. Statistical Inference with Bivariate Gamma Distributions. Biometrika. 1969; 56(3):627–634. https://doi.org/10.1093/biomet/56.3.627

21. Forney G. Viterbi algorithm. Proceedings of the IEEE. 1973; 61(3):268–278. https://doi.org/10.1109/PROC.1973.9030

22. Heide-Jørgensen MP, Nielsen NH, Hansen RG, Blackwell SB. Stomach temperature of narwhals (*Monodon monoceros*) during feeding events. Animal Biotelemetry. 2014; 2(1):9. https://doi.org/10.1186/2050-3385-2-9

23. Pohle J, Langrock R, van Beest FM, Schmidt NM. Selecting the Number of States in Hidden Markov Models: Pragmatic Solutions Illustrated Using Animal Movement. Journal of Agricultural Biological and Environmental Statistics. 2017; 22(3):270–293. https://doi.org/10.1007/s13253-017-0283-8

24. Burnham KM, Anderson DR. Model Selection and Multimodel Inference. 2nd ed. Springer, New York; 2002.

25. Williams TM, Noren SR, Glenn M. Extreme physiological adaptations as predictors of climate-change sensitivity in the narwhal, *Monodon Monoceros*. Marine Mammal Science. 2011; 27:334–349. https://doi.org/10.1111/j.1748-7692.2010.00408.x

Table S1. Different models for covariate effects on the transition probabilities between behavioural states. The predictors η_{ij} relate to the transition probabilities.

Model	$\eta_{12}^{(t)}$	$\eta_{13}^{(t)}$	$\eta_{21}^{(t)}$	$\eta_{23}^{(t)}$	$\eta_{31}^{(t)}$	$\eta_{32}^{(t)}$
1	$\alpha_{00}+\alpha_{10}\tau_t+\alpha_{20}\tau_t^2+\sum_i \delta_{i0}h_i$	$\alpha_{01}+\alpha_{11}\tau_t+\alpha_{21}\tau_t^2+\sum_i \delta_{i1}h_i$	$\beta_{00}+\beta_{10}\tau_t+\beta_{20}\tau_t^2+\sum_i \theta_{i0}h_i$	$\beta_{01}+\beta_{11}\tau_t+\beta_{21}\tau_t^2+\sum_i \theta_{i1}h_i$	$\gamma_{00}+\gamma_{10}d_t+\gamma_{20}d_t^2+\sum_i \zeta_{i0}h_i$	$\gamma_{01}+\gamma_{11}d_t+\gamma_{21}d_t^2+\sum_i \zeta_{i1}h_i$
2	$\alpha_{00}+\sum_i \delta_{i0}h_i$	$\alpha_{01}+\sum_i \delta_{i1}h_i$	$\beta_{00}+\sum_i \theta_{i0}h_i$	$\beta_{01}+\sum_i \theta_{i1}h_i$	$\gamma_{00}+\sum_i \zeta_{i0}h_i$	$\gamma_{01}+\sum_i \zeta_{i1}h_i$
3	$\alpha_{00}+\alpha_{10}\tau_t+\alpha_{20}\tau_t^2$	$\alpha_{01}+\alpha_{11}\tau_t+\alpha_{21}\tau_t^2$	$\beta_{00}+\beta_{10}\tau_t+\beta_{20}\tau_t^2$	$\beta_{01}+\beta_{11}\tau_t+\beta_{21}\tau_t^2$	$\gamma_{00}+\gamma_{10}d_t+\gamma_{20}d_t^2+\sum_i \zeta_{i0}h_i$	$\gamma_{01}+\gamma_{11}d_t+\gamma_{21}d_t^2+\sum_i \zeta_{i1}h_i$
4	$\alpha_{00}+\alpha_{10}\tau_t+\alpha_{20}\tau_t^2$	$\alpha_{01}+\alpha_{11}\tau_t+\alpha_{21}\tau_t^2$	$\beta_{00}+\beta_{10}\tau_t+\beta_{20}\tau_t^2$	$\beta_{01}+\beta_{11}\tau_t+\beta_{21}\tau_t^2$	$\gamma_{00}+\gamma_{10}d_t+\gamma_{20}d_t^2$	$\gamma_{01}+\gamma_{11}d_t+\gamma_{21}d_t^2$
5	α_{00}	α_{01}	β_{00}	β_{01}	γ_{00}	γ_{01}
6	$\alpha_{00}+\alpha_{10}\tau_t$	$\alpha_{01}+\alpha_{11}\tau_t$	$\beta_{00}+\beta_{10}\tau_t$	$\beta_{01}+\beta_{11}\tau_t$	$\gamma_{00}+\gamma_{10}b_t$	$\gamma_{01}+\gamma_{11}b_t$
7	$\alpha_{00}+\alpha_{10}\tau_t$	$\alpha_{01}+\alpha_{11}\tau_t$	$\beta_{00}+\beta_{10}\tau_t$	$\beta_{01}+\beta_{11}\tau_t$	$\gamma_{00}+\gamma_{10}d_t$	$\gamma_{01}+\gamma_{11}d_t$
8	$\alpha_{00}+\alpha_{10}\tau_t+\alpha_{20}\tau_t^2$	$\alpha_{01}+\alpha_{11}\tau_t+\alpha_{21}\tau_t^2$	β_{00}	$\beta_{01}+\beta_{11}\tau_t+\beta_{21}\tau_t^2$	$\gamma_{00}+\gamma_{10}d_t+\gamma_{20}d_t^2$	$\gamma_{01}+\gamma_{11}d_t+\gamma_{21}d_t^2$
9	α_{00}	$\alpha_{01}+\alpha_{11}\tau_t+\alpha_{21}\tau_t^2$	β_{00}	$\beta_{01}+\beta_{11}\tau_t+\beta_{21}\tau_t^2$	$\gamma_{00}+\gamma_{10}d_t+\gamma_{20}d_t^2$	$\gamma_{01}+\gamma_{11}d_t+\gamma_{21}d_t^2$
10	$\alpha_{00}+\alpha_{10}\tau_t$	$\alpha_{01}+\alpha_{11}\tau_t+\alpha_{21}\tau_t^2$	$\beta_{00}+\beta_{10}\tau_t$	$\beta_{01}+\beta_{11}\tau_t+\beta_{21}\tau_t^2$	$\gamma_{00}+\gamma_{10}d_t+\gamma_{20}d_t^2$	$\gamma_{01}+\gamma_{11}d_t+\gamma_{21}d_t^2$
11	$\alpha_{10}\tau_t+\alpha_{20}\tau_t^2+\sum_i \delta_{i0}h_i$	$\alpha_{11}\tau_t+\alpha_{21}\tau_t^2+\sum_i \delta_{i1}h_i$	$\beta_{10}\tau_t+\beta_{20}\tau_t^2+\sum_i \theta_{i0}h_i$	$\beta_{11}\tau_t+\beta_{21}\tau_t^2+\sum_i \theta_{i1}h_i$	$\gamma_{10}d_t+\gamma_{20}d_t^2+\sum_i \zeta_{i0}h_i$	$\gamma_{11}d_t+\gamma_{21}d_t^2+\sum_i \zeta_{i1}h_i$
12	$\alpha_{00}+\alpha_{10}\tau_t+\alpha_{20}\tau_t^2$	$\alpha_{01}+\alpha_{11}\tau_t+\alpha_{21}\tau_t^2+\sum_i \delta_{i1}h_i$	$\beta_{00}+\beta_{10}\tau_t+\beta_{20}\tau_t^2$	$\beta_{01}+\beta_{11}\tau_t+\beta_{21}\tau_t^2+\sum_i \theta_{i1}h_i$	$\gamma_{00}+\gamma_{10}d_t+\gamma_{20}d_t^2+\sum_i \zeta_{i0}h_i$	$\gamma_{01}+\gamma_{11}d_t+\gamma_{21}d_t^2+\sum_i \zeta_{i1}h_i$
13	α_{00}	$\alpha_{01}+\alpha_{11}\tau_t+\alpha_{21}\tau_t^2$	β_{00}	$\beta_{01}+\beta_{11}\tau_t+\beta_{21}\tau_t^2$	$\gamma_{00}+\gamma_{10}d_t+\gamma_{20}d_t^2+\sum_i \zeta_{i0}h_i$	$\gamma_{01}+\gamma_{11}d_t+\gamma_{21}d_t^2+\sum_i \zeta_{i0}h_i$
14	$\alpha_{00}+\sum_i \delta_{i0}h_i$	$\alpha_{01}+\alpha_{11}\tau_t+\alpha_{21}\tau_t^2+\sum_i \delta_{i1}h_i$	$\beta_{00}+\sum_i \theta_{i0}h_i$	$\beta_{01}+\beta_{11}\tau_t+\beta_{21}\tau_t^2\sum_i \theta_{i1}h_i$	$\gamma_{00}+\gamma_{10}d_t+\gamma_{20}d_t^2$	$\gamma_{01}+\gamma_{11}d_t+\gamma_{21}d_t^2+\sum_i \zeta_{i1}h_i$

Table S2. Estimates of the model parameters of the state distributions and their 95% confidence intervals in model 1 for correlated Log-normal distribution. In state i, μ_i and σ_i are the log-mean and log-standard deviation of the correlated log-normal distribution. Index MD stands for Maximum Depth, DT stands for Dive Duration and PD stands for Post-Dive time. The depth is measured in meters, and time in seconds. The confidence intervals were computed from the Hessian of the negative log-likelihood function, i.e., based on the inverse of the observed Fisher information.

Correlated log-normal distribution		
	Estimate	95% CI
μ_1^{MD}	2.61	$[2.56, 2.66]$
μ_2^{MD}	4.78	$[4.73, 4.84]$
μ_3^{MD}	6.11	$[6.11, 6.12]$
σ_1^{MD}	1.36	$[1.33, 1.39]$
σ_2^{MD}	0.77	$[0.72, 0.81]$
σ_3^{MD}	0.18	$[0.17, 0.19]$
μ_1^{DT}	1.50	$[1.48, 1.52]$
μ_2^{DT}	1.80	$[1.77, 1.82]$
μ_3^{DT}	2.46	$[2.45, 2.46]$
σ_1^{DT}	0.50	$[0.49, 0.51]$
σ_2^{DT}	0.43	$[0.41, 0.46]$
σ_3^{DT}	0.14	$[0.14, 0.14]$
μ_1^{PD}	1.26	$[1.22, 1.3]$
μ_2^{PD}	0.86	$[0.83, 0.88]$
μ_3^{PD}	1.73	$[1.71, 1.75]$
σ_1^{PD}	1.13	$[1.10, 1.15]$
σ_2^{PD}	0.43	$[0.41, 0.45]$
σ_3^{PD}	0.53	$[0.52, 0.55]$
ρ_1	0.56	$[0.53, 0.58]$
ρ_2	0.81	$[0.78, 0.83]$
ρ_3	0.46	$[0.43, 0.50]$

Table S3. Estimates of the model parameters of covariate effects and their 95% confidence intervals in model 1 for correlated Log-normal distribution. The spline effects of hour are denoted by $H_{ij}^t = \sum_k \delta_{ij}^{(k)} h_k^t$, of τ_t by $T_{ij}^t = \sum_k \theta_{ij}^{(k)} s_k^t$, and of d_t by $D_{ij}^t = \sum_k \zeta_{ij}^{(k)} d_k^t$ for $k = 1, 2, 3$ and $i, j = 1, 2, 3$; $i \neq j$.

Correlated log-normal distribution		
	Estimate	95% CI
α_{00}	-3.82	$[-5.30, -2.34]$
α_{01}	-0.97	$[-1.57, -0.37]$
β_{00}	0.08	$[-1.18, 1.33]$
β_{01}	-1.74	$[-3.22, -0.263]$
γ_{00}	1.88	$[1.10, 2.66]$
γ_{01}	-3.08	$[-5.15, -1.02]$
$\theta_{12}^{(1)}$	1.95	$[0.11, 3.80]$
$\theta_{12}^{(2)}$	-1.71	$[-3.77, 0.36]$
$\theta_{12}^{(3)}$	-2.48	$[-5.48, 0.52]$
$\theta_{13}^{(1)}$	-3.11	$[-4.39, -1.82]$
$\theta_{13}^{(2)}$	-6.60	$[-8.28, -4.92]$
$\theta_{13}^{(3)}$	-7.08	$[-10.5, -3.61]$
$\theta_{21}^{(1)}$	0.19	$[-1.58, 1.97]$
$\theta_{21}^{(2)}$	-0.57	$[-2.39, 1.25]$
$\theta_{21}^{(3)}$	-0.51	$[-4.10, 3.08]$
$\theta_{23}^{(1)}$	-5.04	$[-8.82, -1.26]$
$\theta_{23}^{(2)}$	-5.39	$[-7.13, -3.66]$
$\theta_{23}^{(3)}$	-0.48	$[-1.93, 0.98]$
$\zeta_{31}^{(1)}$	-2.83	$[-3.75, -1.92]$
$\zeta_{31}^{(2)}$	-5.35	$[-6.61, -4.1]$
$\zeta_{31}^{(3)}$	-0.84	$[-2.51, 0.83]$
$\zeta_{32}^{(1)}$	-4.83	$[-6.38, -3.28]$
$\zeta_{32}^{(2)}$	-8.80	$[-10.30, -7.28]$
$\zeta_{32}^{(3)}$	-0.38	$[-2.76, 2.00]$
$\delta_{12}^{(1)}$	0.41	$[-1.78, 2.60]$
$\delta_{12}^{(2)}$	2.53	$[1.33, 3.73]$
$\delta_{12}^{(3)}$	-0.06	$[-2.34, 2.22]$
$\delta_{13}^{(1)}$	0.21	$[-0.82, 1.23]$
$\delta_{13}^{(2)}$	0.08	$[-0.42, 0.58]$
$\delta_{13}^{(3)}$	0.02	$[-1.06, 1.10]$
$\delta_{21}^{(1)}$	-2.77	$[-4.88, -0.66]$
$\delta_{21}^{(2)}$	-2.29	$[-3.34, -1.25]$
$\delta_{21}^{(3)}$	-3.46	$[-5.37, -1.55]$
$\delta_{23}^{(1)}$	-1.24	$[-3.48, 1.01]$
$\delta_{23}^{(2)}$	0.63	$[-0.93, 2.19]$
$\delta_{23}^{(3)}$	1.54	$[-0.37, 3.45]$
$\delta_{31}^{(1)}$	-2.32	$[-3.41, -1.23]$
$\delta_{31}^{(2)}$	-0.60	$[-1.14, -0.06]$
$\delta_{31}^{(3)}$	-2.49	$[-3.59, -1.39]$
$\delta_{32}^{(1)}$	4.45	$[0.99, 7.90]$
$\delta_{32}^{(2)}$	6.28	$[4.05, 8.51]$
$\delta_{32}^{(3)}$	4.94	$[1.47, 8.41]$

Chapter 7

Paper II

Joint work with

Raghavendra Selvan, Outi Tervo, Mads Peter Heide-Jørgensen, and Susanne Ditlevsen

This chapter is based on the published article: Manh Cuong Ngo, Raghavendra Selvan, Outi Tervo, Mads Peter Heide-Jørgensen, Susanne Ditlevsen. *Detection of foraging behavior from accelerometer data using U-Net type convolutional networks.* Ecological Informatics., 62, 101275, 2021.

Ecological Informatics 62 (2021) 101275

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Detection of foraging behavior from accelerometer data using U-Net type convolutional networks

Mạnh Cường Ngô [a,c,*], Raghavendra Selvan [b,d], Outi Tervo [c], Mads Peter Heide-Jørgensen [c], Susanne Ditlevsen [a,*]

[a] Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark
[b] Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100 Copenhagen Ø, Denmark
[c] Greenland Institute of Natural Resources, Strandgade 91, 2, DK-1401 Copenhagen K, Denmark
[d] Department of Neuroscience, University of Copenhagen, Blegdamsvej 3, 2200 Copenhagen Ø, Denmark

ARTICLE INFO

Keywords:
Buzz
Accelerometer data
Narwhals (*Monodon monoceros*)
East Greenland
Convolutional neural network
U-net
Random forest

ABSTRACT

Narwhal (*Monodon monoceros*) is one of the most elusive marine mammals, due to its isolated habitat in the Arctic region. Tagging is a technology that has the potential to explore the activities of this species, where behavioral information can be collected from instrumented individuals. This includes accelerometer data, diving and acoustic data as well as GPS positioning. An essential element in understanding the ecological role of toothed whales is to characterize their feeding behavior and estimate the amount of food consumption. Buzzes are sounds emitted by toothed whales that are related directly to the foraging behaviors. It is therefore of interest to measure or estimate the rate of buzzing to estimate prey intake. The main goal of this paper is to find a way to detect prey capture attempts directly from accelerometer data, and thus be able to estimate food consumption without the need for the more demanding acoustic data. We develop three automated buzz detection methods based on accelerometer and depth data solely. We use a dataset from five narwhals instrumented in East Greenland in 2018 to train, validate and test a logistic regression model and the state-of-the art machine learning algorithms random forest and deep learning, using the buzzes detected from acoustic data as the ground truth. The deep learning algorithm performed best among the tested methods. We conclude that reliable buzz detectors can be derived from high-frequency-sampling, back-mounted accelerometer tags, thus providing an alternative tool for studies of foraging ecology of marine mammals in their natural environments. We also compare buzz detection with certain movement patterns, such as sudden changes in acceleration (jerks), found in other marine mammal species for estimating prey capture. We find that narwhals do not seem to make big jerks when foraging and conclude that their hunting patterns in that respect might differ from other marine mammals.

1. Introduction

The narwhal (*Monodon monoceros*) is a high-Arctic cetacean known for its characteristic tusk (Graham et al., 2020). It is among the deepest diving cetaceans and can dive to depths of more than 1800 m (Heide-Jørgensen, 2009). Narwhals dive to forage, and their main prey includes Greenland halibut (*Reinhardtius hippoglossoides*), polar cod (*Boreogadus saida*), capelin (*Ammodytes villosus*) and squids (*Gonatus* sp.) (Heide-Jørgensen et al., 1994; Laidre and Heide-Jørgensen, 2005). In disphotic and aphotic zones, they need to use acoustics to explore their environment and locate prey, i.e., echolocation, by producing short-duration sounds (clicks) and listening for echoes reflected from surrounding objects (Berta et al., 2015); and *buzzes*, a series of clicks with short inter-click-interval (below 50 milliseconds) (Blackwell et al., 2018). The clicks are used for orientation, and buzzes mark the final phase of a potential prey capture event for several cetacean species, including sperm whales, porpoises, and beaked whales (DeRuiter et al., 2009; Johnson et al., 2004; Miller et al., 2004), so we hypothesize the same for narwhals. How frequently they forage and successfully catch a prey is largely unknown due to difficult environmental and logistical conditions in the Arctic that complicate direct studies of prey intake. We have therefore used tagging technologies to collect behavioral data to elucidate the feeding behavior. The movements of five whales were studied during summer in Scoresby Sound of East Greenland in 2018. Facing

* Corresponding authors at: Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark.
E-mail addresses: nxp418@ku.dk, susanne@math.ku.dk (M.C. Ngô).

https://doi.org/10.1016/j.ecoinf.2021.101275
Received 9 December 2020; Received in revised form 18 February 2021; Accepted 18 February 2021
Available online 12 March 2021

strong climate changes in the Arctic, many species in this region are under threat. Understanding foraging behavior of narwhals helps us to understand the conflicts between their food intake, changes in their habitat and the increasing level of anthropogenic activities, e.g., fisheries and shipping, in the effort to conserve this unique cetacean.

Accelerometer data and acoustic data are widely used in marine mammal science to understand behavior of whales (Nowacek et al., 2016). Accelerometer data are collected by tri-axial accelerometers that combine two components: the static acceleration due to gravity and the dynamic acceleration due to the motion of the whale, along three axes: surge (longitudinal *X*-axis); sway (*Y*-axis); and heave (vertical *Z*-axis) orientations (Shepard et al., 2008; Wilson et al., 2008). Acoustic data are recordings of vibration of the medium around the recording devices caused by acoustic radiation (Swanson, 2008). Currently, acoustic data is the best way to estimate potential successful prey capture due to the assumption that a buzz is the sound narwhals make just before attempting to catch their targets. Hence, we assume that the whales have specific movement pattern around the time a buzz occurs. Such a pattern was discovered in harbor seals (*Phoca vitulina*) (Ydesen et al., 2014), and harbor porpoises (*Phocoena phocoena*) (Wisniewska et al., 2016), where they made big jerks, i.e., sudden movements, before catching prey. If jerks or other specific movement patterns around prey capture can be identified accelerometer data can be used to quantify prey capture events.

One of the crucial differences between acceleration and acoustic signals in biotelemetry is that acceleration is an easier parameter to collect. Due to the relatively low sampling frequencies of $<$500 Hz, accelerometer data collection can be achieved by less memory and less battery power enabling longer deployments. Accelerometers are therefore also small-in-size enabling their use in various applications for a wide range of species. Furthermore, the small size of instruments decreases drag and other negative effects on the individual carrying the tag. Due to the general high sampling rate of acoustic data, acoustics are currently used only in animal-borne archival applications, where the data are stored onboard the instrument. Retrieval of archival instruments can, however, be challenging in habitats such as the polar regions, where ice and extreme seasonal variation of light constrain research to the summer months. While accelerometer data are currently stored onboard, detecting behaviors from preprocessed accelerometer data would be an important step for developing satellite-linked biotelemetry applications, because accelerometer data can be compressed to what is suitable for satellite transmission. That would allow data upload directly from instruments mounted on animals and would extent the temporal and spatial range of behavioral and ecological research.

Analyzing accelerometer and acoustic data to detect behavioral patterns of whales is of considerable interest, e.g., for sperm whale (Fais et al., 2016) and whale/dolphin (Hillman et al., 2003). The bulk of these analyses are performed by engineering features obtained by transforming the acquired data into more intuitive features of the underlying dynamics, then processing them with a suitable prediction algorithm such as logistic regression, support vector machines or random forest. However, feature engineering is a difficult art that requires lots of expert knowledge and is time consuming (Ng, 2015), especially when the data is noisy, vast and not already well understood. Since the 2000s, the deep learning era has had many breakthroughs from computer vision to natural language understanding, using huge amount of input data without (much) feature engineering like in traditional machine learning approaches (Alsheikh et al., 2015; Goodfellow et al., 2016). Convolutional Neural Networks (CNN) are among the most widely used deep learning architectures (Farabet et al., 2013; Krizhevsky et al., 2012; Szegedy et al., 2015; Tompson et al., 2015). Unlike in the classical machine learning methods, where one needs to design filters by traditional engineering, CNN can "learn" features directly from the data by its huge number of parameters.

Given the huge amount of data acquired from animal-borne instruments sampling at high frequency, we have chosen to explore CNN-based methods for developing robust techniques for detecting buzzes from accelerometer data. Analyzing time series data at multiple resolutions allows capturing useful temporal correlations and renders it suitable for modelling behavior of narwhals. Our baseline machine learning approach is random forest, which has been used extensively in human accelerometer datasets (e.g., see Bayat et al., 2014; Kwapisz et al., 2010) as well as in animal studies (Shepard et al., 2008; Wang, 2019; Wilson et al., 2008). Furthermore, we compare our CNN model with logistic regression. We hypothesize that there exists some hidden movement pattern during or around the buzzes. Therefore, the main goal in this work is to develop feeding-behavior detection models based on narwhal's accelerometer data using machine learning algorithms, including traditional ones like random forest, advanced ones like deep learning, or logistic regression. Furthermore, we will also investigate if jerks are correlated with buzz events.

2. Material and methods

2.1. Ethics statement

Permission for capturing, handling, and tagging of narwhals was provided by the Government of Greenland (Case ID 2010 $\pm$ 035453, document number 429926). The project was reviewed and approved by the IACUC of the University of Copenhagen (June 17th, 2015). Access and permits to use land facilities in Scoresby Sound were provided by the Government of Greenland. No protected species were sampled.

2.2. Data

Five male narwhals were instrumented with Acousonde™recorders model 3B (length of 22.8 cm and weight of 360 g, at www.acousonde.com), sampling acoustic data (25,811 Hz), 3-axis acceleration (100 Hz), 3-axis orientation/ magnetometer (10 Hz), pressure/depth (10 Hz), light (10 Hz), and temperature (10 Hz), and backpack satellite transmitters (www.wildlifecomputers.org) for FastLoc GPS data. The narwhales were captured and tagged in Scoresby Sound, East Greenland in 2018 (Fig. 1), for details of the tagging methods see (Heide-Jørgensen et al., 2015). In brief, the Acousonde's were attached to the dorsal ridge with suction cups, two 1-mm nylon lines and magnesium links. After at most eight days, they were detached from the whales by the corrosion between magnesium links and sea water. Then they were picked up by local hunters localizing them thanks to their position's signals from Argos transmitter and VHS transmitter (ATS Telemetry) attached to the Acousonde. The tags were designed to assure that their weights and their shapes were less than 3% of the whales' weights and frontal area.

Narwhal acoustic signals can reliably be detected using a relatively low sampling rate (Blackwell et al., 2018) and the deployments in this work used continuous sampling at 25,811 Hz (16 bit-resolution). In addition, a tri-axial accelerometer sampled the movements of the whales at 100 Hz and a pressure transducer sampled the depth of the whale at 10 Hz. The data were initially collected to analyze the effects of noise from seismic exploration on the natural lives of narwhals in East Greenland. To avoid interference with altered behavior it was decided to only include data collected before the seismic exposure. In addition, we removed the first 24 h in all five data sets to eliminate a possible influence from the capturing and tagging (Tervo et al., 2021). The data set of the five narwhals is large with a total length of 121.8 h of 100 Hz accelerometer data, and of acoustic data of 10 Hz, with a total of 2615 buzzes whose total length is 1 h 31 min 56.8 s (Table 1).

Most of the sound files from the Acousondes have a fixed length of 30 min, except the last file of each whale. They were examined manually by two analysts in MTViewer (a custom-written program for analysis of Acousonde data, W.C. Burgess, personal communication) for continuous click trains produced by the whale. A custom-written buzz detector (Matlab, The MathWorks, Inc., Natick, MA, USA) was used to identify buzzes made by the whales, then the positive detections were verified

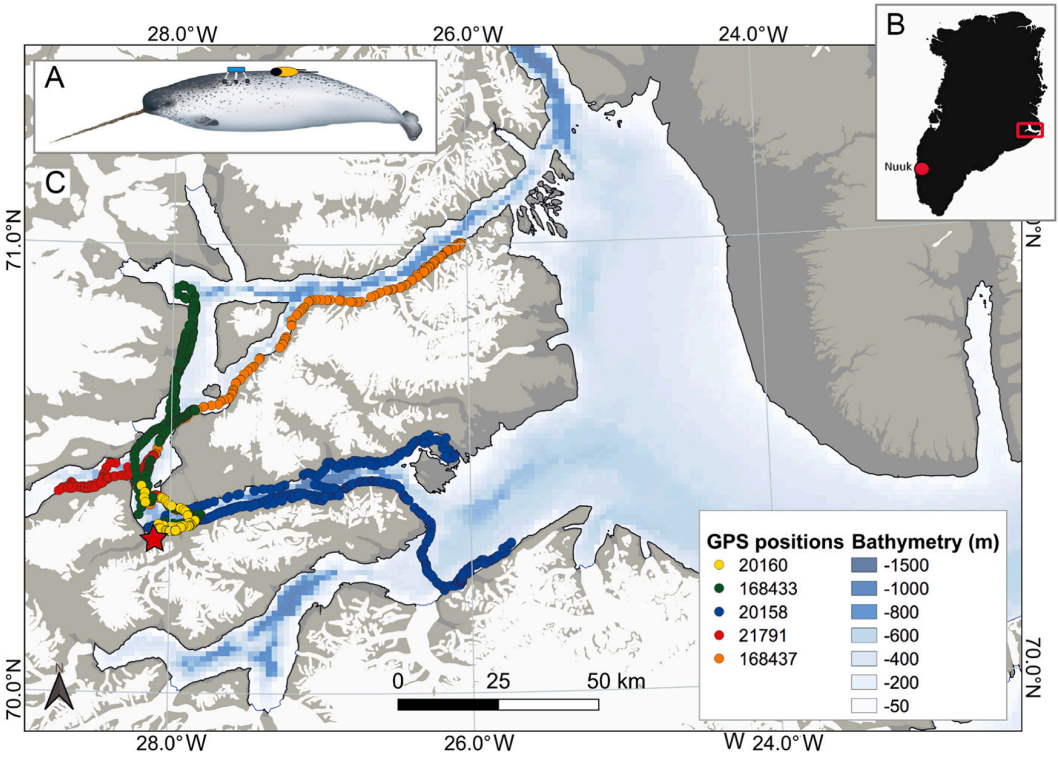

Fig. 1. Diagram of the placement of GPS saddle-back tag (light blue) and Acousonde TM behavioral tag (orange) on a narwhal (A), map of Greenland showing the Scoresby Sound fjord (red box) in East Greenland (B) and a zoomed in map of the study area with the location of the field site, Hjørnedal (marked with a red star), where tagging of narwhals took place (C). The tracks of the five male narwhals used in this study are shown as hourly mean GPS positions. Illustration of a narwhal by Uko Gortner. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Data lengths and number of buzzes of five narwhals.

Whale ID	20,158	20,160	21,791	168,433	168,437
Data length (hours)	38.66	9.53	31.89	9.15	32.56
No. of Buzz	494	197	836	208	880

manually by experienced manual analysts. Each file's acoustic signal was reviewed visually, then the analyst listened to the first five seconds of each buzz to estimate the background noise. The data set was too big to listen to in its entirety, hence only the positive buzz detections periods were examined. The lengths of buzzes varied from 0.4 s to 6.7 s. The positive rates of buzz labels, i.e., the sum of the lengths of buzzes over the total length of the data, of each whale were: 1.37% for narwhal 21,791, 0.73% for narwhal 20,158, 1.77% for narwhal 168,437, 1.12% for narwhal 20,160, and 1.40% for narwhal 168,433. Note that these lengths were estimated by an automatic detector, and not as accurate if manually recorded. The buzz resolution was 10 Hz; hence it was expanded to 100 Hz to fit the resolution of the accelerometer data. The true dynamic component of accelerometer data is difficult to extract without gyroscopes or speedometers and is unknown for narwhals, so we used the raw data to let the algorithms explore it directly. We included depth as a feature since the buzz distribution strongly depends on depth (Blackwell et al., 2018).

We tested whether there was an association between buzzes and jerks. A jerk is defined by the norm of the differences of the acceleration of each axis. We define an RMS jerk to be the root-mean-square (RMS) of the three jerk values over a window of 200 milliseconds, i.e., over a total of $3 \times 20 = 60$ data points (Ydesen et al., 2014). The calculation of RMS over a window was used for smoothing the accelerometer data and attenuate the impact of clipping, i.e., when the signal is larger than the detection threshold (Ydesen et al., 2014). In each window, we defined a buzz to happen if the whale was buzzing at least half of the duration of the window. We defined a big RMS jerk to happen if the jerk's RMS was above pre-defined thresholds defined below. We defined a window to be a positive if there was a big RMS jerk (negative if no big RMS jerk), and a true positive if there was both a big RMS jerk and a buzz, and likewise for false positives and true/false negatives. We calculated the precision and the recall for different thresholds for each whale, where the precision is defined as the ratio of the number of true positives over the sum of true positives and false positives, and the recall is defined as the ratio of the number of true positives over the sum of true positives and false negatives. The thresholds were chosen between 0 and 166,000 mG/s, slightly larger than the maximum value of RMS jerks measured in the data. The thresholds were evaluated in steps of 2000 mG/s, where $1\ G = 9.81\ m/s^2$. We calculated precision and recall for instantaneous big RMS jerks (at the same time as the buzz), as well as for delays of 0.2, 0.4, 0.6, 0.8 and 1 s, respectively, to check if the big RMS jerks happen at a fixed time after the buzz.

We defined a dive as a continuous period during which the maximum depth is at least 20 m, while 10 m was chosen as the onset and the end of a dive (Fig. 2). We chose 10 m due to possible wrong zero-offsets of the dive (Luque and Fried, 2011). A dive was separated into three phases: the descending phase, the bottom phase, and the ascending phase, apart from the surface phase. The bottom phase was defined as the period at which the whale spent at or below 75% of the maximum depth of the dive (Tervo et al., 2021). The descent phase was defined as the period between the onset of the dive and the onset of the bottom phase, and the ascent phase was defined as the period between the end of the bottom phase and the end of the dive.

The phase of a dive is important for the buzzing activity, since the whales typically forage and buzz at the bottom of the dive, and buzz less

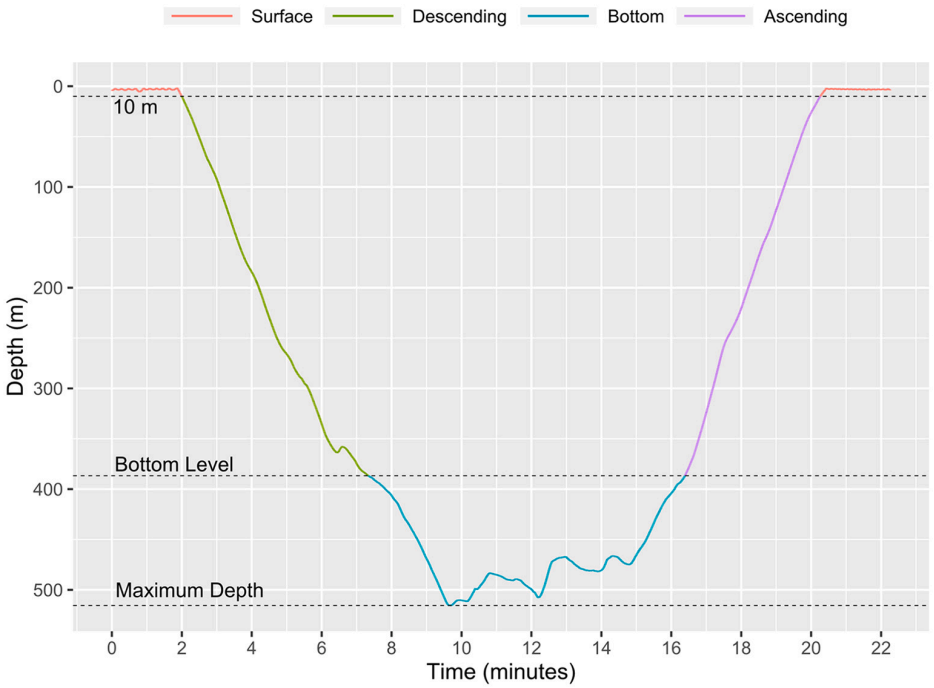

Fig. 2. Example of a dive from narwhal 168,433. The dashed line "Bottom Level" is the depth threshold (75% of the Maximum Depth) for the bottom phase. Red curves, green curve, cyan curve, and violet curve indicate surface, descending, bottom, and ascending phases, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

during descend and ascend. The categorical feature describing the four dive phases was included by one-hot encoding where each phase is transformed into a binary vector: surface as (1,0,0,0), descending as (0,1,0,0), bottom as (0,0,1,0), and ascending as (0,0,0,1). Thus, we have five features: the three accelerometer axes A_X, A_Y, A_Z, the depth, and the diving phase. An example of a subsample of the record of narwhal 21,791 shows the features during the bottom phase of a dive together with the response variable of buzzing, where one encodes that a buzz is

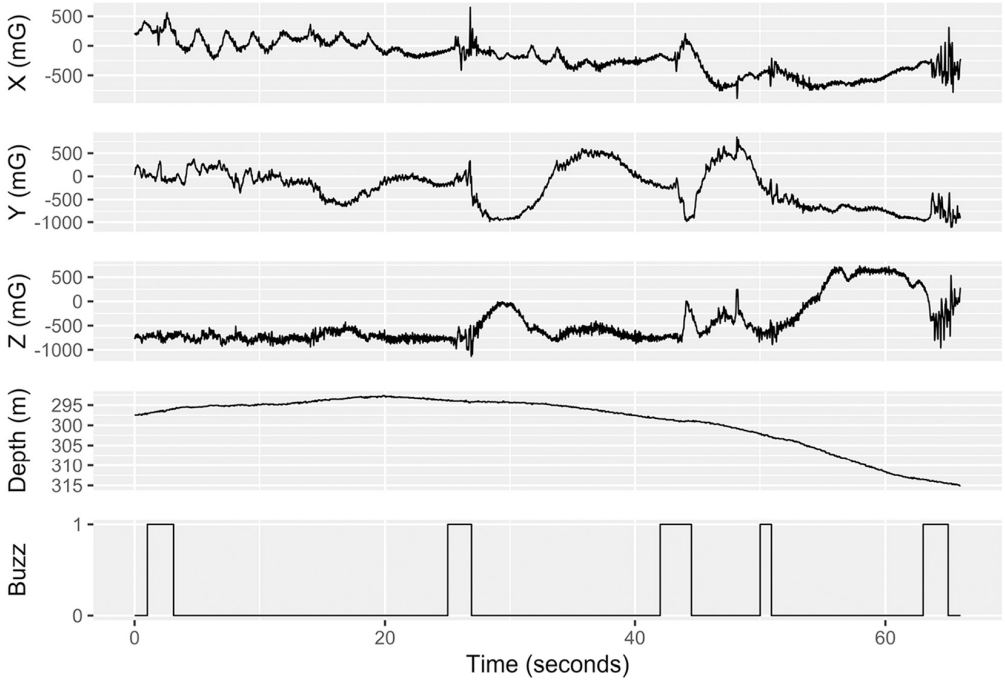

Fig. 3. Example of the record of narwhal 21,791 showing the time evolution of the four features and the response variable Buzz, during the bottom phase of a dive. The panels show the tri-axial accelerometer data, the depth, and the buzzes (1 is presence, and 0 is absence). Note the increased variability in the accelerometer data during some of the buzzes.

happening (Fig. 3).

2.3. Supervised U-Net

Recently, Perslev et al. (2019) used a U-Net encoder-decoder architecture, a specific design using CNN as the base (Ronneberger et al., 2015), for multidimensional time series, called U-Time. The U-Net originally was designed for image segmentation tasks (Ronneberger et al., 2015). It uses an encoder-decoder type architecture as shown in Fig. 4. U-Net encodes input data to feature maps at multiple resolutions, by applying convolution layers followed by downsampling layers (using max-pooling) in the encoder. The sequence of steps in the encoder, also called the contracting path, allow the convolution layers to learn useful features at different resolutions of the data. Then the decoder up-samples such encoded features through an up-sampling layer, then concatenates with the corresponding feature maps from the encoder through skip connections (Drozdzal et al., 2016). It helps the decoder to have detailed information in the earlier stages from the contracting path, which is lost due to pooling layers in the encoder. Moreover, skip connections make the model easier to optimize in practice (He et al., 2016). The output of the decoder in a U-Net makes predictions at full resolution in the output data. Perslev et al. (2019) have shown that their U-Net model, the U-Time, has obtained similar performance as Recurrent Neural Networks (RNN) (Williams et al., 1986), the default choice for time series data, while RNN is harder to train. Therefore, following their work, we used U-Time architecture as the deep learning model for detecting buzzes from the input data.

2.4. U-Net implementation

We implemented the U-Time/U-Net model (Perslev et al., 2019; Ronneberger et al., 2015) in Python 3.6.9 using PyTorch 1.6.0 (Paszke et al., 2017) on Google Collaboratory with NVIDIA P100 of 16 GB of RAM (Google, 2020). We used the same architecture as (Perslev et al., 2019), except that we replaced the last activation function `softmax` with `sigmoid`, since the classification problem is binary.

2.5. Optimization objective

The dataset is very imbalanced (buzzes occur only rarely in the dataset). This makes the standard machine learning algorithms, including deep learning, perform poorly. Accuracy is the default loss used by most algorithms, where wrongly predicted zeros and ones are penalized the same. However, when most labels are the same (in the dataset, more than 98% of the responses are zeros), the prediction of all being zero (i.e., concentrating on the majority) will lead to a high accuracy, but be uninformative for the problem (Visa and Ralescu, 2005).

We therefore used the Dice loss (*DL*) (Smith et al., 2020), which is a loss designed for highly imbalanced data, where wrongly predicted ones are penalized more than wrongly predicted zeros, defined by

$$DL = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i}$$

where p_i is the predicted probability of a buzz at time i, and g_i is the ground truth (the observed buzz or not) at time i and takes values 0 or 1 for $i = 1, \ldots, N$.

Let $\alpha = \frac{1}{N} \times \sum_{i=1}^{N} g_i$ be the proportion of ones in the data set, where $0 < \alpha \ll 1$ since the dataset is imbalanced. If the model predicts all 0's, i.e. $p_i = 0$ for all $=1, \ldots, N$, then $DL = 1$ because $p_i g_i = 0$ for all $i = 1, \ldots, N$. If the model predicts all 1's, i.e. $p_i = 1$ for all $i = 1, \ldots, N$, then $DL > 1 - 2\alpha \approx 1$ since $\alpha \ll 1$. On the other hand, if the model predicts all correctly, then DL $= 0$. Therefore, Dice loss penalizes effectively if the model predicts all 0's or all 1's.

2.6. Model selection

We divided the data set into training, validation and test sets following a ratio of 60:20:20 chronologically for each of the five whales, then combined the training, validation, and test sets from each whale. We also performed *cross validation* on each of the whales, i.e., trained on three whales, validated on one, and used the last one for testing, to evaluate how well the trained model generalized to a new dataset. The Dice loss was computed between the model predictions and the ground truth of the training data. The validation data is used to avoid overfitting. The difference between the model's prediction and the ground truth was measured by the Dice loss. Stochastic gradient-based optimization algorithms were used on the training data to update the model parameters gradually by iterations, denoted epochs. The validation set was used to select the epoch at which the validation loss, the loss measured on the validation set, was minimal, before the models overfitted. The trained models were then evaluated independently on the test sets to avoid data leakage, a phenomenon where there is some information leakage from validation sets or test sets into training sets (Kaufman et al., 2011).

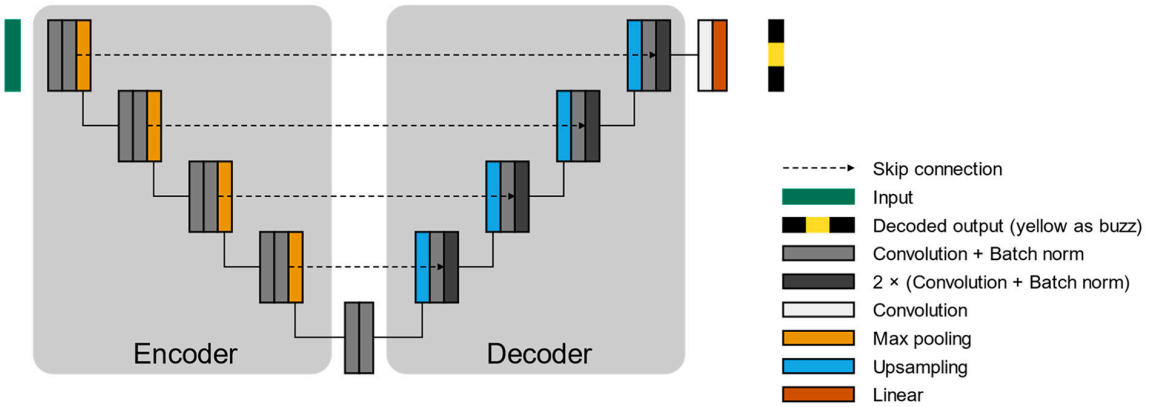

Fig. 4. Structure of a U-Time network. The dashed lines are skip connections. The continuous lines connect the encoder and the decoder. The inputs are time series (accelerometer and depth data), while the yellow interval indicates the duration when buzz happens. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.7. Optimization

We did hyper-parameter search for batch size and the number of convolution filters at the first convolutional layer of the U-Net. Batch size, related to mini-batch gradient descent, is primarily used to smooth the gradients, and can be parallelized. Convolution filters, or filter banks (whose parameters need to be learnt), are used to transform the input to feature maps. The number of hidden units varied between two, four, eight, and sixteen, while batch sizes varied between two, four, eight, and sixteen during our preliminary experiments. We used Adam (Kingma and Ba, 2015) with different learning rates between 0.01, 5×10^{-3}, 10^{-3}, and 5×10^{-4}. Smaller learning rates did not help to make the model converge after trial and error. We ran up to 301 epochs but did early stopping if there were no improvements after 150 epochs since the best epoch, i.e., the epoch at which the loss function is minimum. We decided to tune the best hyperparameters on only one whale, the data from narwhal 168,433, since it was computationally expensive.

2.8. Random forest and logistic regression implementation

There are no hyper parameters in logistic regression, and therefore no need for a validation set. We therefore divided the data set into training and test sets following a ratio of 80:20 for each of the five whales. We implemented logistic regression (Harrell, 2015) and random forest (Boehmke and Greenwell, 2019; Breiman, 2001; Ho, 1995) as our baseline methods to compare with U-Net. We used the default hyper-parameter setting of random forest in Scikit-learn (Pedregosa et al., 2011), which works well in most cases (Probst et al., 2019). We implemented random forest with balanced subsample, i.e., for each tree we assigned greater weights to the minority class (here the positive class) based on its bootstrap sample (Chen et al., 2004). To improve the learning of these models, manual feature extraction should be done. Selecting the right features is a difficult and intricate task, however, these two traditional methods are more robust and easier to interpret (Warmerdam, 2018).

Feature extraction was done for both logistic regression and random forest, following Bayat et al. (2014). The data were divided into successive windows consisting of 100 consecutive data points, i.e., one second, to compress information into features. Each window shared an overlap of 50 data points with the next window, to assure that no specific pattern was broken due to the edges of these windows (Fig. 5). We used twenty-seven features:

- Mean, standard deviation (STD), root mean square (RMS) and MinMax (the difference between maximum and minimum within a window) of accelerometer components A_X, A_Y, A_Z along three axes X, Y, Z, as well as the mean depth (13 features).
- STD, RMS and MinMax of the magnitude of the acceleration $A_m = \sqrt{A_X^2 + A_Y^2 + A_Z^2}$ (three features).
- Number of peaks, elapsed time between consecutive local peaks of accelerometer components A_X, A_Y, A_Z along three axes X, Y, Z, as well as the variance of the number of peaks of A_X, A_Y, A_Z (seven features).
- Correlations between A_X and A_Y, A_Y and A_Z, A_Z and A_X (three features).
- Dive phase, encoded by one-hot encoding.

A window was marked as positive if more than 50% of its corresponding output values belonged to a buzz, and negative otherwise (Fig. 5). We had 6348 positive windows out of 526,086 windows against 27 features, enough for robust maximum likelihood estimation of the logistic regression model following the *one in ten* rule (Harrell, 2015; Peduzzi et al., 2006). We also tried to test whether logistic regression worked better when detecting only the start of buzzes instead of the whole length of the buzzes, however, the performance was worse (results not shown).

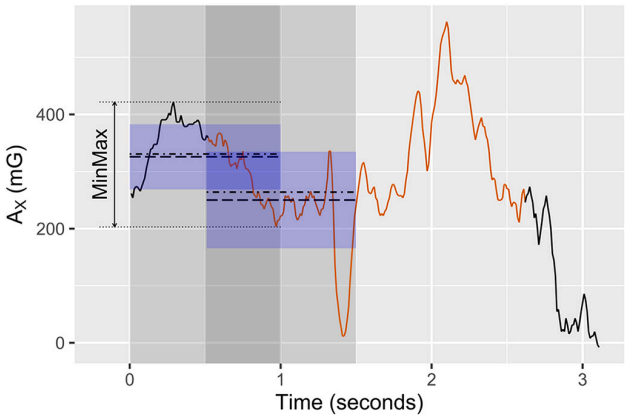

Fig. 5. Illustration of feature extraction from the accelerometer of the X-axis. The dark shaded area indicates where two windows of size of 1-s have 50% overlap. Orange part of the curve indicates the duration of a buzz. Dashed lines are the means of each window, while shaded violet areas indicate +/− one standard deviation. Dot-dashed lines indicate RMS's of each window. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Results

3.1. Machine learning models

Model hyperparameters of the U-Net models were tuned using a smaller dataset from a single narwhal 168,433. The best U-Net models were those with four hidden units. The features of the data for the U-Net model were accelerometer components A_X, A_Y, A_Z, the depth data and the diving phase. The model for all five whales having the best validation loss is presented in Fig. 6. Dice loss was smallest at the 224th epoch of the validation set, at which the parameters were chosen for U-Net models.

We evaluated the U-Net models on the five test sets from each whale as well as the entire test set in Fig. 7. It shows the proportion of correct predictions of buzzes of the models, where we define a correct prediction with some slack: the percentage overlap between predicted and true buzzes. An example for the cases of 50% percentage overlap and 1 s

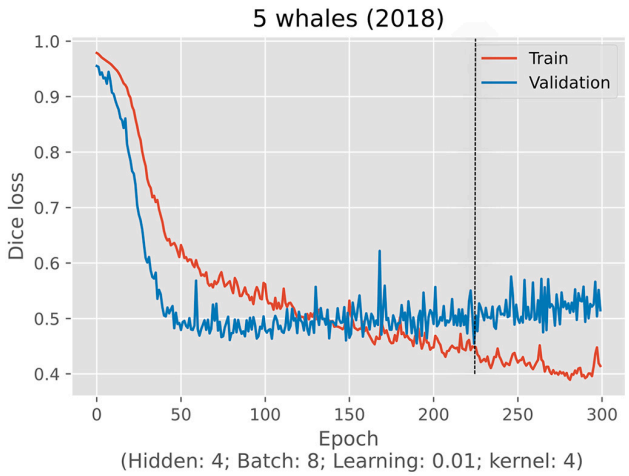

Fig. 6. The training and validation loss when the number of hidden units at first convolution layer are four of the best models with respect to hyper-parameters tuning.

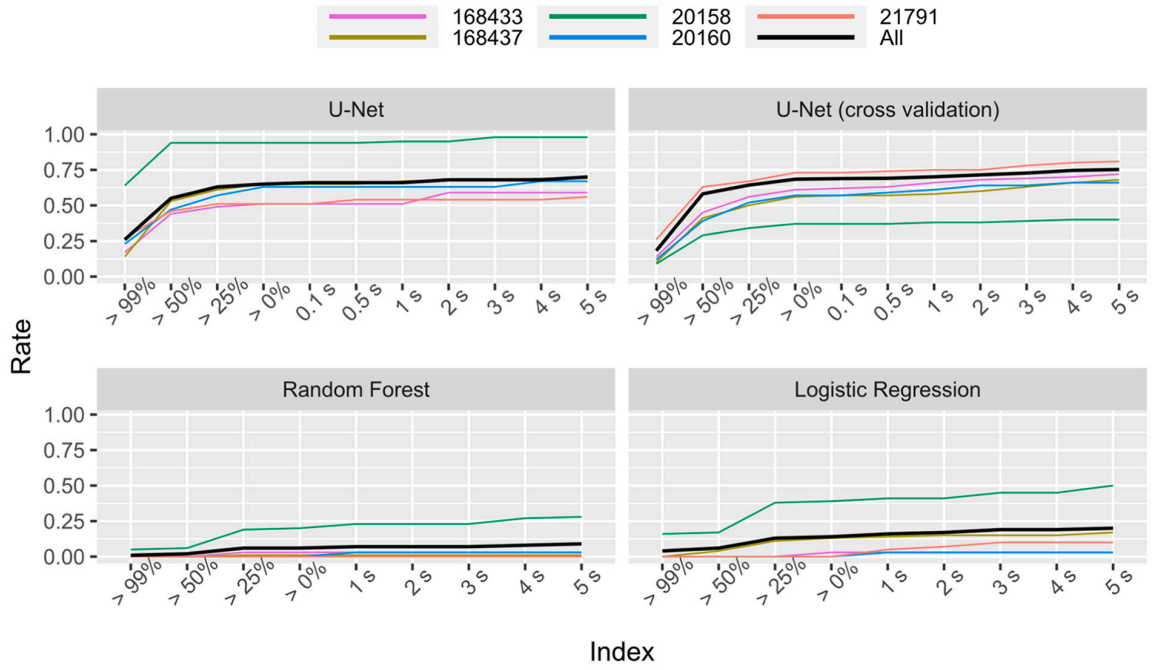

Fig. 7. The proportion of correct predictions of buzzes of the four models: U-Net, U-Net (cross validation), random forest, and logistic regression. On X-axis is shown: the first four indices of the overlap between predicted and true buzzes; the next indices show the proportion of predicted buzzes with a maximum distance to its nearest true buzz smaller than 0.1 s, 0.5 s (only for U-Net), and 1 s, 2 s, 3 s, 4 s and 5 s for all the models. The colored lines are the values for each whale, the black line is for all five whales together.

distance between predicted and true buzzes is shown in Fig. 8. The proportion of correct predictions of buzzes with a maximum distance to its nearest true buzz smaller than 0.1 s and 0.5 s (only for U-Net, since the other two models, random forest and logistic regression, only have a resolution of one second), and 1 s, 2 s, 3 s, 4 s, 5 s were calculated for all the models (Fig. 7).

The U-Net with cross validation performed similarly on each whale compared to the U-Net trained and validated on data from all whales, except for narwhal 20,158, probably because it had a lower buzzing rate than the other whales. This is reassuring since the algorithm then could possibly generalize well to the narwhal population. There were almost no differences between 500, 1000, and 2000 trees for random forest (results not shown), so we chose the one with 2000 trees. This random forest model predicted poorly on the raw data and even worse with lowpass filtered data of 0.25 Hz (results not shown). Finally, logistic regression models predicted better than random forest, but not as good as the U-Net.

If the focus of the analysis is to identify foraging dives, i.e., those dives where buzzes occur, and in that case, how many buzzes the whale emits during that dive, the results improved the classification of dives with feeding activities. We evaluated whether the methods could distinguish between foraging dives with buzzes and exploring dives without buzzes, as well as whether the number of buzzes in each dive could be predicted, even if the exact timing of the buzzes were wrong. Technically, we evaluated the models by counting the number of consecutive series of ones (i.e., buzzing events). There were 456 dives in total, among them 152 were foraging dives, i.e., having buzzes (33.3%). The number of predicted buzzes against the number of true buzzes within each dive is plotted in Fig. 9, for U-Net, random forest and logistic regression. Only the U-Net model distinguished well between dives with buzzes and those without buzzes. For the U-Net model, the number of true negatives were 39, false positives were 3, false negatives were 0, and true positives were 23, thus, for identifying foraging/non-foraging dives, the precision was 88% and the recall was 100%. Furthermore, it captured the trend of the number of buzzes within dives well, even if slightly overestimated. The random forest model correctly classified the non-foraging dives, however, for foraging dives it always underestimated the number of buzzes, often as zero. Logistic regression predicted slightly better, but it also underestimated the number of buzzes in each dive and estimated too many non-foraging dives.

Fig. 8. Examples of the definition of partial correct prediction: a) 50% overlap, and b) distance of one second between ground truth (orange) and prediction (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In Fig. 10, we compared the differences of the number of buzzes with the time spent buzzing per dive for foraging dives. The U-Net model tended to predict more buzzes than the ground truth, while random forest and logistic regression models predicted less than the ground truth, as also shown in Fig. 11. The same pattern emerged for the buzz lengths. All in all, the U-Net performed best on both predictions of number of buzzes as well as on the time spent buzzing per dive (length of buzzing period).

An example of the predictions compared to the ground truth from the three models is shown in Fig. 11. It illustrates the dive with most buzzes, which is a dive from narwhal 20,158, showing clearly, that the U-Net model predicted best with many overlaps between predictions and ground truth, while logistic regression came second.

3.2. Jerk analysis

Fig. 12 shows the precision and recall of RMS jerks at different

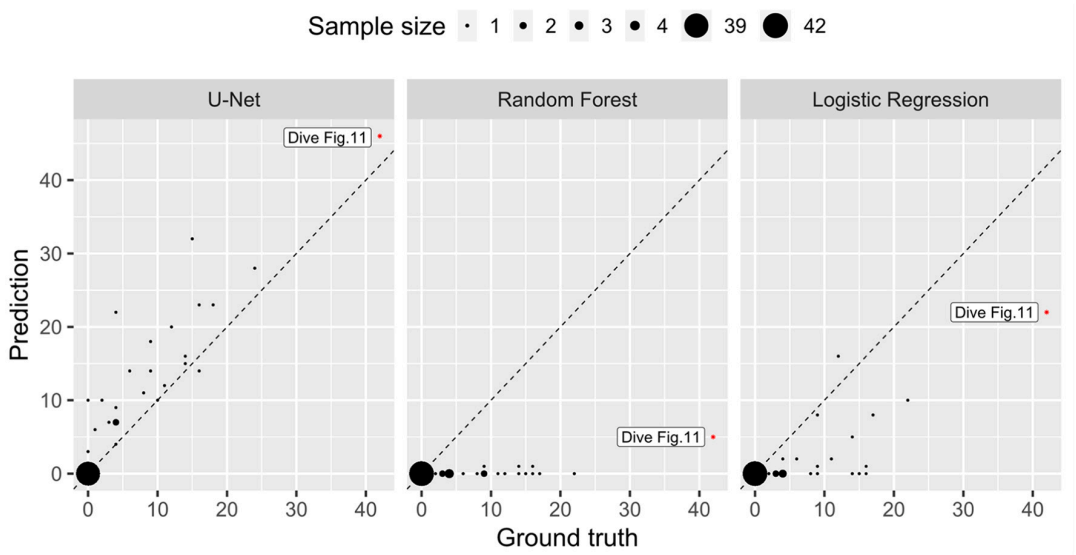

Fig. 9. Scatter plot of the number of predicted buzzes against the number of true buzzes per dive for U-Net, random forest, and logistic regression. Size of the dots indicate the number of points. The dashed line is the identity line. The red points indicate the dive illustrated in Fig. 11. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

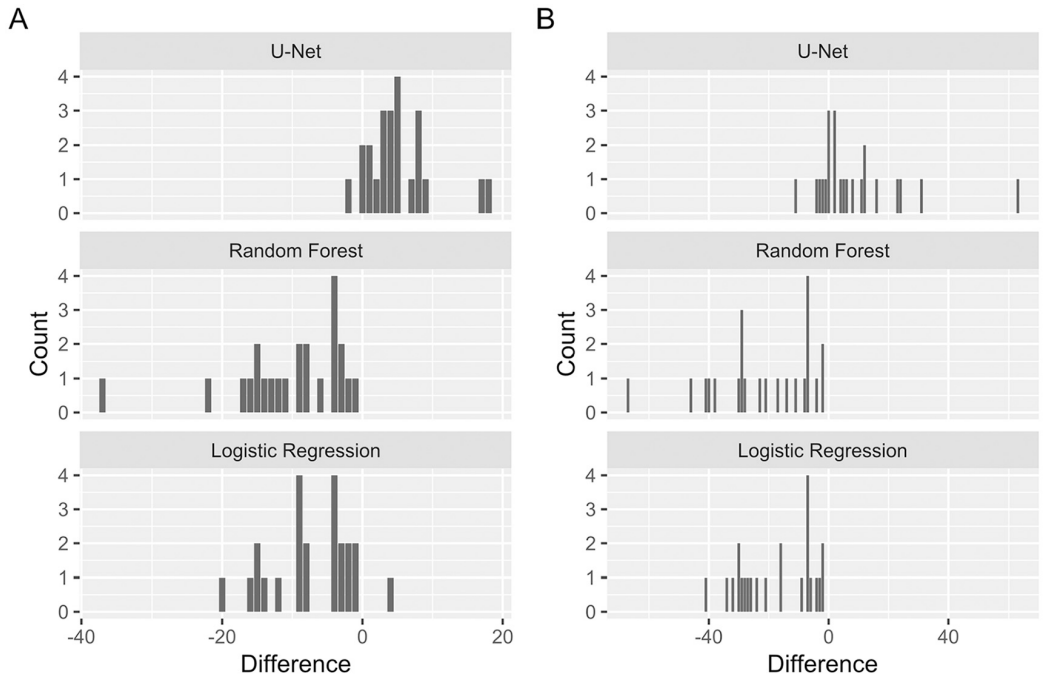

Fig. 10. Histograms of A) the difference between the number of buzzes from the predictions and the ground truth, and B) the difference between the sum of the lengths of the buzzes per dive for U-Net, random forest, and logistic regression.

thresholds. The precision of prediction of buzzes from big RMS jerks is low, less than 0.25, for thresholds less than 12,500 mG/s. It increases for larger thresholds; the precision for narwhals 168,437 and 168,433 even reach 1 for some thresholds, but the true positives and the recalls decrease extremely fast to close to zero. Note that for a threshold of zero, the precision equals the proportion of ones in the data, and the recall equals one. Additional attempts with a delayed jerk within 1 s (0.2, 0.4, 0.6, 0.8, and 1 s) after the buzzes can be found in Supplementary Material. Fig. 13 shows an example trace, with several high RMS peaks without any buzz activity, while there are a few high RMS peaks close to the buzzes. We therefore conclude that jerks are not a suitable criterion for detecting buzzes and prey capture events.

4. Conclusions

In this study, we investigated if some special movement patterns present around the times of buzzes from free ranging narwhals can be detected by machine learning methods. Our results show that the U-Net

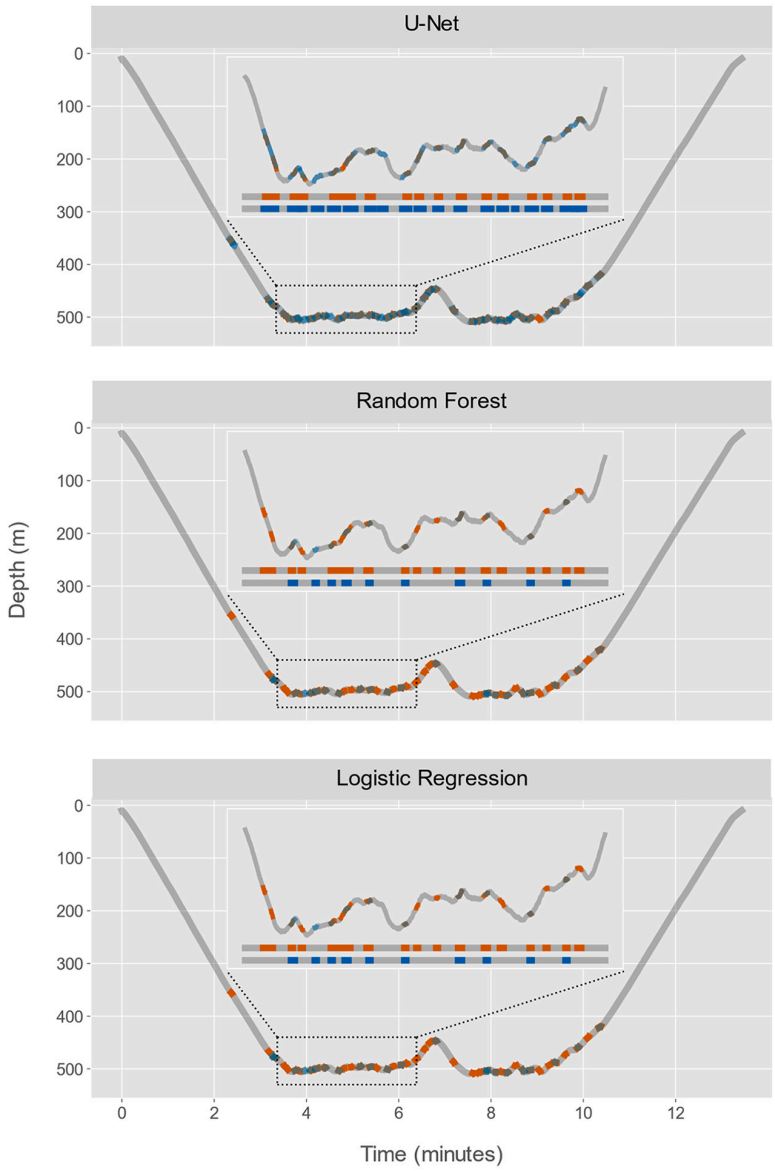

Fig. 11. Example of a dive of the record of narwhal 20,158 with ground truth buzz (orange) and prediction buzz (blue). The dark grey shadings on the depth lines indicate the overlaps between predictions and ground truths. The first panel illustrates the U-Net model, the second illustrates the random forest, and the third illustrates logistic regression. The smaller panel within each panel show a zoomed-in section of the time-depth series marked with a dashed line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

can be used to detect buzzes from accelerometer data. We also examined whether the narwhals make big jerks around buzzes, which have been found in a previous study of captive harbor seals (Ydesen et al., 2014). They used a triaxial accelerometer to collect head- and jaw mounted accelerometer data in prey capturing attempts. It worked well also in a wild environment for harbor porpoises (Wisniewska et al., 2016), as well as sperm whales (Fais et al., 2016). In our study, the tags are positioned on the back of narwhals, so they may not detect more subtle head- and jaw-jerks, but major body movements towards targeted prey during buzzing events should be detectable. We frequently identified acceleration peaks in the narwhal data that did not follow a buzzing event, and analysis showed that both the precision and recall were poor. Moreover, we tested on free ranging narwhals rather than captive specimens, so the variances are larger. The narwhals may engage in many different movement activities that imply quick movements, so false positives might be high if only big RMS jerks are used as a criterion. We therefore conclude that big RMS jerks are not trustworthy indicators for detection of buzz events in narwhals. From an anatomical perspective the absence of teeth in narwhal jaws also makes raptorial feeding less likely and suggests that narwhals ingest prey by buccal suction feeding. The narwhals that were instrumented for this study feed on squids in the water column that presumably are slow moving and easy to capture and ingest. Other narwhal populations feed on halibut that may require a more raptorial capturing approach and rapid movements during the buzz phase of the prey strike.

With the improvement of tagging technologies, more data especially from accelerometer instruments can be collected and tools like machine learning for big data analysis might contribute enormously to the understanding of marine predators. We have demonstrated an application of deep learning, with U-Net, to accelerometer and depth data for detection of buzzes in narwhals. The performance of U-Net was superior to random forest, the baseline method of tabular dataset, which failed to detect the buzzes. We used the Dice loss function, which is suitable for an imbalanced dataset. The trained model can be used to make predictions or facilitate the training process on new datasets, called transfer learning (Pratt, 1993). It distinguished well between foraging dives with buzzes and exploring dives without buzzes, much better than random forest and logistic regression. Its buzz predictions were much closer to

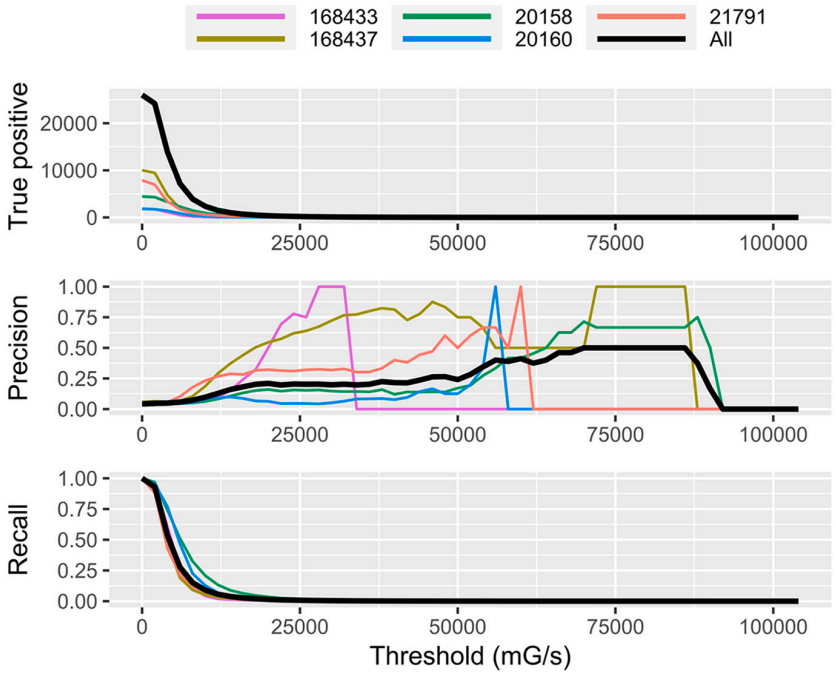

Fig. 12. Precision and recall of RMS jerks for different thresholds for predicting buzzes.

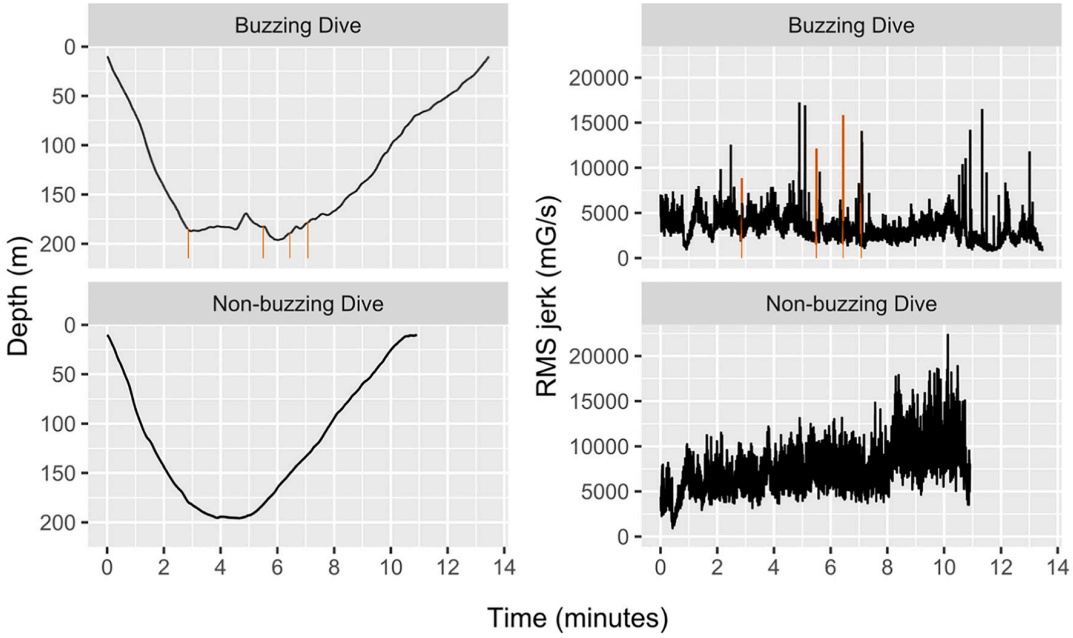

Fig. 13. Example of two dives from narwhal 21,791. The upper two panels show the time-depth series and RMS jerk of a buzzing dive; the lower two panels show the time-depth series and RMS jerks of a non-buzzing dive. The orange lines indicate presence of buzzes, in the upper right panel the RMS jerks are colored orange when buzzing.

the ground truth than the predictions from the two other models.

Finding the right features for random forest or logistic regression is particularly hard in new applications. Thus, we cannot definitively conclude that U-Net, or more general deep learning, is superior without further research. A simple method like logistic regression performed better than random forest, although worse than U-Net. Furthermore, the determination of the right hyper-parameters for the U-Net is computationally expensive. The performance of logistic regression might be improved by more careful feature selection such as including correlation of buzzes. Logistic regressions have an advantage of being much simpler and much more transparent than the U-Net or deep learning in general.

Although our study shows positive results on the use of U-Net models, there are several limitations that require more analysis. For

example, deep learning methods, in general, are not transparent because of the huge number of parameters. The general way of learning is by trial-and-error, i.e., to test different hyper-parameters and/or loss functions. It creates an enormous training time, as well as a high carbon footprint (Anthony et al., 2020), and makes it vulnerable to spurious findings due to the lack of transparency. Combining signal processing techniques and more transparent statistical/machine learning methods, may help to understand when and why the methods work (Forde and Paganini, 2019; Succi and Coveney, 2019). However, our results provide some evidence that deep learning provides a valuable tool compared to other machine learning or statistical methods. The supervised machine learning approaches in this study could be extended to any other marine mammals' datasets.

Declaration of Competing Interest

None.

Acknowledgments

We would like to thank Susanna Blackwell and Alexander Conrad for supporting of data labelling of narwhal buzzing activity. SD was supported by Independent Research Fund Denmark, case: 9040-00215B, MCN was funded by Greenland Research Council and OT and MPHJ were funded by the Greenland Institute of Natural Resources. This study is part of the Northeast Greenland Environmental Study Program which is a collaboration between DCE – Danish Centre for Environment and Energy at Aarhus University, the Greenland Institute of Natural Resources, and the Environmental Agency for Mineral Resource Activities of the Government of Greenland.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2021.101275. Code is available at https://github.com/kirimaru-jp/buzz-accel.

References

Alsheikh, M.A., Selim, A., Niyato, D., Doyle, L., Lin, S., Tan, H.-P., 2015. Deep activity recognition models with triaxial accelerometers. In: The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 8–13. https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12627.

Anthony, L.F., Kanding, B., Selvan, R., 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. In: ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems. URL. https://arxiv.org/abs/2007.03051.

Bayat, A., Pomplun, M., Tran, D.A., 2014. A study on human activity recognition using accelerometer data from smartphones. Proc. Comput. Sci. 34, 450–457. https://doi.org/10.1016/j.procs.2014.07.009.

Berta, A., Sumich, J.L., Kovacs, K.M., 2015. Sound production for communication, echolocation, and prey capture. In: Berta, I.A., Sumich, J.L., Kovacs, K.M. (Eds.), Marine Mammals: Evolutionary Biology, Third edition. Academic Press, pp. 345–395. https://doi.org/10.1016/C2011-0-07338-6. ISBN: 9780123972576.

Blackwell, S., Tervo, O., Conrad, A., Sinding, M., Hansen, R., Ditlevsen, S., Heide-Jørgensen, M., 2018. Spatial and temporal patterns of sound production in East Greenland narwhals. PLoS One 13 (6), e0198295. https://doi.org/10.1371/journal.pone.0198295.

Boehmke, B., Greenwell, B.M., 2019. Hands-on Machine Learning with R. Chapman and Hall/CRC. https://doi.org/10.1201/9780367816377. ISBN: 9780367816377.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, C., Liaw, A., Breiman, L., 2004. Using random forest to learn. In: Report Number: 666. University of California, Berkeley. URL. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf.

DeRuiter, S., Bahr, A., Blanchet, M.-A., Hansen, S.F., Kristensen, J.H., Madsen, P., Wahlberg, M., 2009. Acoustic behaviour of echolocating porpoises during prey capture. J. Exp. Biol. 212, 3100–3107. https://doi.org/10.1242/jeb.030825.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C., 2016. The importance of skip connections in biomedical image segmentation. In: International Workshop on Deep Learning in Medical Image Analysis, pp. 179–187. https://doi.org/10.1007/978-3-319-46976-8_19.

Fais, A., Johnson, M., Wilson, M., Aguilar Soto, N., Madsen, P.T., 2016. Sperm whale predator-prey interactions involve chasing and buzzing, but no acoustic stunning. Sci. Rep. 6, 28562. https://doi.org/10.1038/srep28562.

Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2013. Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1915–1929. https://doi.org/10.1109/TPAMI.2012.231.

Forde, J.Z., Paganini, M., 2019. The scientific method in the science of machine learning. In: ICLR 2019 Debugging Machine Learning Models Workshop. URL. https://arxiv.org/abs/1904.10922.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. ISBN: 9780262035613. URL. https://mitpress.mit.edu/books/deep-learning.

Google, 2020. Colaboratory. Retrieved from Google Research. https://research.google.com/colaboratory/faq.html.

Graham, Z.A., Garde, E., Heide-Jørgensen, M.P., Palaoro, A.V., 2020. The longer the better: evidence that narwhal tusks are sexually selected. Biol. Lett. 16, 20190950. https://doi.org/10.1098/rsbl.2019.0950.

Harrell, F., 2015. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7. ISBN: 9783319194240.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 770–778. https://doi.org/10.1109/CVPR.2016.90.

Heide-Jørgensen, 2009. Narwhal *Monodon monoceros*. I e. In: Perrin, W.F., Wursig, B., JGM, Thewissen (Eds.), Encyclopedia of Marine Mammals, 2nd edition, pp. 754–758. ISBN: 9780080919935. https://doi.org/10.1016/B978-0-12-373553-9.X0001-6.

Heide-Jørgensen, M., Dietz, R., Leatherwood, S., 1994. A note on the diet of narwhals (Monodon monoceros) in Inglefield Bredning (NW Greenland). Monogr. Greenl. Biosci. 39, 213–216. URL. https://www.mtp.dk/details.asp?eln=201434.

Heide-Jørgensen, M., Nielsen, N., Hansen, R., Schmidt, H., Blackwell, S., Jørgensen, O., 2015. The predictable narwhal: satellite tracking shows behavioural similarities between isolated subpopulations. J. Zool. 297, 54–65. https://doi.org/10.1111/jzo.12257.

Hillman, G., Wursig, B., Gailey, G., Kehtarnavaz, N., Drobyshevsky, A., Araabi, B., Tagare, H., 2003. Computer-assisted photo-identification of individual marine vertebrates: a multi-species system. Aquat. Mamm. 29, 117–123. https://doi.org/10.1578/016754203101023960.

Ho, T.K., 1995. Random decision forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, 1, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994.

Johnson, M., Madsen, P., Zimmer, W., Aguilar de Soto, N.T., 2004. Beaked whales echolocate on prey. Proc. R. Soc. B Biol. Sci. 271, 383–386. https://doi.org/10.1098/rsbl.2004.0208.

Kaufman, S., Rosset, S., Perlich, C., 2011. Leakage in data mining: Formulation, detection, and avoidance. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 6, pp. 556–563. https://doi.org/10.1145/2020408.2020496.

Kingma, D., Ba, J., 2015. Adam: a method for stochastic optimization. In: International Conference on Learning Representations, pp. 1–15. URL. https://arxiv.org/abs/1412.6980v9.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Proces. Syst. 60, 1097–1105. https://doi.org/10.1145/3065386.

Kwapisz, J.R., Weiss, G.M., Moore, S.A., 2010. Cell phone-based biometric identification. Fourth IEEE International Conference on Biometrics: Theory Applicationsand Systems (BTAS). https://doi.org/10.1109/BTAS.2010.5634532.

Laidre, K., Heide-Jørgensen, M., 2005. Winter feeding intensity of narwhals (*Monodon monoceros*). Mar. Mammal Sci. 21, 45–57. https://doi.org/10.1111/j.1748-7692.2005.tb01207.x.

Luque, S.P., Fried, R., 2011. Recursive filtering for zero offset correction of diving depth time series with GNU R package diveMove. PLoS One 6, e15850. https://doi.org/10.1371/journal.pone.0015850.

Miller, P.J., Johnson, M.P., Tyack, P.L., 2004. Sperm whale behaviour indicates the use of echolocation click buzzes "creaks" in prey capture. Proc. R. Soc. B Biol. Sci. 271, 2239–2247. https://doi.org/10.1098/rspb.2004.2863.

Ng, A., 2015. Deep Learning. In: GPU Technology Conference. URL. https://video.ibm.com/recorded/60113824.

Nowacek, D.P., Christiansen, F., Bejder, L., Goldbogen, J.A., Friedlaender, A.S., 2016. Studying cetacean behaviour: new technological approaches and conservation applications. Anim. Behav. 120, 235–244. https://doi.org/10.1016/j.anbehav.2016.07.019.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lerer, A., 2017. Automatic differentiation in PyTorch. In: NIPS 2017 Autodiff Workshop: The Future of Gradient-Based Machine Learning Software and Techniques. URL. https://openreview.net/pdf?id=BJJsrmfCZ.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. https://doi.org/10.5555/1953048.2078195.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R., 2006. A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. 49, 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3.

Perslev, M., Jensen, M.H., Darkner, S., Jennum, P.J., Igel, C., 2019. U-time: a fully convolutional network for time series segmentation applied to sleep staging. Adv. Neural Inform. Process. Syst. (NeurIPS) 32, 4415–4426. URL. https://arxiv.org/abs/1910.11162.

Pratt, L., 1993. Discriminability-based transfer between neural networks. In: NIPS Conference: Advances in Neural Information Processing Systems, pp. 204–211. https://doi.org/10.5555/645753.668046.

Probst, P., Boulesteix, A.-L., Bischl, B., 2019. Tunability: importance of hyperparameters of machine learning algorithms. J. Machine Learning Res. 20, 1–32. URL. https://www.jmlr.org/papers/volume20/18-444/18-444.pdf.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 MICCAI 2015, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Shepard, E., Wilson, R., Quintana, F., Laich, A., Liebsch, N., Albareda, D., Myers, A., 2008. Identification of animal movement patterns using tri-axial accelerometry. Endanger. Species Res. 10, 47–60. https://doi.org/10.3354/esr00084.

Smith, A.G., Petersen, J., Selvan, R., Rasmussen, C.R., 2020. Segmentation of roots in soil with U-Net. Plant Methods 16, 13 (2020). https://doi.org/10.1186/s13007-020-0563-0.

Succi, S., Coveney, P.V., 2019. Big data: the end of the scientific method? Phil. Trans. R. Soc. A 377, 20180145. https://doi.org/10.1098/rsta.2018.0145.

Swanson, D.C., 2008. Acoustic data acquisition. In: Havelock, I.D., Kuwano, S., Vorländer, M. (Eds.), Handbook of Signal Processing in Acoustics. Springer New York., New York, NY, pp. 17–32. https://doi.org/10.1007/978-0-387-30441-0. ISBN: 9780387776989.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., A. R, 2015. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

Tervo, O.M., Ditlevsen, S., Ngô, M.C., Nielsen, N.H., Blackwell, S.B., Williams, T., Heide-Jørgensen, M.P., 2021. Hunting by the stroke: how foraging drives diving behavior and swimming biomechanics of East-Greenland narwhals (Monodon monoceros). Front. Mar. Sci. 7, 1244–1261. https://doi.org/10.3389/fmars.2020.596469.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C., 2015. Efficient object localization using convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 648–656. https://doi.org/10.1109/CVPR.2015.7298664.

Visa, S., Ralescu, A., 2005. Issues in mining imbalanced data sets-a review paper. In: Proceedings of the sixteen midwest artificial intelligence and cognitive science conference, pp. 67–73. URL. https://eecs.ceas.uc.edu/~ralescal/PAPERS/VRMaics2005.pdf.

Wang, G., 2019. Machine learning for inferring animal behavior from location and movement data. Ecol. Inform. 49, 69–76. https://doi.org/10.1016/j.ecoinf.2018.12.002.

Warmerdam, V.D., 2018. Winning with Simple, Even Linear, Models. PyData, London. URL. https://youtu.be/68ABAU_V8qI.

Williams, R.J., Hinton, G.E., Rumelhart, D.E., 1986. Learning representations by back-propagating errors. Nature 323, 533–536. https://doi.org/10.1038/323533a0.

Wilson, R., Shepard, E., Liebsch, N., 2008. Prying into the intimate details of animal lives: use of a daily diary on animals. Endanger. Species Res. 4, 123–137. https://doi.org/10.3354/esr00064.

Wisniewska, D.M., Johnson, M., Teilmann, J., Rojano-Doñate, L., Shearer, J., Sveegaard, S., Madsen, P.T., 2016. Ultra-high foraging rates of harbor porpoises make them vulnerable to anthropogenic disturbance. Curr. Biol. 26, 1441–1446. https://doi.org/10.1016/j.cub.2016.03.069.

Ydesen, K., Wisniewska, D., Hansen, J., Beedholm, K., Johnson, M., Madsen, P., 2014. What a jerk: prey engulfment revealed by high-rate, super-cranial accelerometry on a harbour seal (*Phoca vitulina*). J. Exp. Biol. 217, 2239–2243. https://doi.org/10.1242/jeb.100016.

Chapter 8

Paper III

Joint work with

Susanne Ditlevsen and Mads Peter Heide-Jørgensen

Sea surface temperatures drive the movements of bowhead whales

Abstract

Arctic cetaceans are under threats of global warming due to rapid warming water. Specially, for bowhead whale, endemic species of this area, it is very challenging because they only stay in Arctic year around. 84 bowhead whales in Baffin Bay – East Greenland are tagged during 11-year period between 2001 and 2011 to help us understand the effects of global warming on this species. With this long dataset, the main goal of this paper is to investigate the effect of sea surface temperature (SST) on their distribution. We use high resolution of daily positions and daily SST. We develop seasonal models based on Tweedie generalize linear models to model the duration that bowhead whales spend in each season during the period of 2001-2011. Our study confirms the previous research that bowhead whales prefer spending time in colder water, hence more warming water coming will force them to move further north, hence reduce their habitats.

8.1 Introduction

The bowhead whale (*Balaena mysticetus*), also called the Greenland right whale, is a baleen whale endemic to Arctic and sub-Arctic waters. It has several features that makes it well adapted to a life in cold and ice-covered waters, e.g. extremely thick blubber layer exceeding 40 cm in adult whales, a thick epidermis or skin, and a low body core temperature of 33.8°C (George et al., 1994; Haldiman and Tarpley, 1993). It reaches body lengths of up to 17-19 m, with an estimated body mass of up to 100 tons (George et al., 2021). Sexual maturity is reached late in life ($>$ 18 years for females and $>$ 25 years for males, Tarpley et al., 2021), with three-year intervals between pregnancies and it is believed to reach the oldest age ($>$ 200yrs) of any mammals (George et al., 2021). It is mainly distributed in the high Arctic with three stocks known as Bering-Chukchi-Beaufort Seas, East Canada-West Greenland (ECWG), and the East Greenland-Svalbard-Barents Sea stock. A small relict stock persists in the Okhotsk Sea (Givens & Heide-Jørgensen, 2021). Due to its distribution in the Arctic, it is believed to be sensitive to the ongoing warming that is amplified in the Arctic with rapid reduction in sea ice and increasing sea

surface temperatures (Alexander et al., 2018; IPCC, 2013). Recent climate changes have impacted the movements and habitats of many species' endemic to the Arctic, especially 11 species of marine mammal, including the bowhead whale (Perrin et al., 2009; Kovacs et al., 2011; Laidre et al., 2015; Citta et al., 2021). Cetaceans are, unlike seals and walrus, not directly dependent on sea ice and sometimes sea ice appears as an obstacle for the movements of the whales, but sea ice may still be important for governing the trophic cascade that eventually creates the concentrations of prey items that whales, and especially filter feeders like bowhead whales, are so heavily dependent on.

Bowhead whales feed on zooplankton and especially Calanus species seem to be important prey items for bowhead whales in the ECWG stock (Heide-Jørgensen et al., 2012; Pomerleau et al., 2017; Fortune et al., 2019). Climate change with loss of sea ice and warming of Arctic waters may change the availability of prey that bowhead whales can target. Some prey species may move north in response to ocean warming or competition from more southern species entering the Arctic (Michel et al., 2012). Also, the density of prey items and the timing of the pelagic phase of zooplankton may change with reduction of sea ice that dictates the onset of primary production that the zooplankton depends on (Hansen et al., 2003).

Bowhead whales are physiologically adapted to year-round presence in cold water and the blubber insulation and lack of dorsal fin prevents heat dump during excessive exercise activities (Hokkanen, 1990). The restrictive options for heat dump are also the reason why bowhead whales are among the slowest of the baleen whales (Geogre et al., 2020). They have limited capacity to make rapid movements over short or long distances in response to increased water temperatures or new predators and anthropogenic activities invading pristine Arctic waters.

We summarise the water circulation in Baffin Bay as described in (Hansen et al., 2020). The seasonal movements of bowhead whales in Baffin Bay and the Canadian Arctic Archipelago are impacted by two currents: the West Greenland current (WGC) and the Baffin current (BC). The WGC is a mixed current of the two waters: the medium deep (200 – 1000 m) and warm water of Irminger Current, and the shallow and cold East Greenland current (ECG). At the Davis Strait, WGC divides into two branches. The first branch turns west to join the Labrador Outer Current that flows to the south. The second branch continues flowing to the north until $\sim 75^{\circ}$ North where it turns west. Here, it joins the current from the polar basin that is flowing south through Nares Strait and the other channels in the Canadian Arctic Archipelago, to create the south-going Baffin Bay current.

It is notoriously difficult to study the trophic cascade in the remote Arctic waters where bowhead whales roam and the whales are in many cases a better indicator of the underlying processes although a full mechanistic understanding behind the processes require more targeted studies. Here we analyse the reaction of bowhead whales to changes in sea temperatures, and use the whales as a proxy for the underlying changes in the trophic cascade. We use a large time series of satellite tracking of bowhead whales in West Greenland and Canada, collected between 2001 to 2011, using Tweedie GLMs (Jørgensen, 1987; Dunn & Smith, 2018). We exploit the power of Graphical Processing Unit (GPU) computing, allowing us to fit the model on a much finer temporal and spatial dataset than in previous studies (Chambault et al., 2018), using the classical second-order Newton method *iteratively reweighted least squares* (IRLS).

8.2 Material & Methods

8.2.1 Whale distributions and tagging area

We analyse a data set collected from 84 bowhead whales in Disko Bay, Greenland that were tagged with ARGOS satellite transmitters during 2001-2011. All the tags were made by Wildlife Computers (`https://wildlifecomputers.com`). Details of the tagging methods are provided in Heide-Jørgensen et al. (2003, 2006). Briefly, to tag the whales, small boats with a length of 6 m were deployed on days with good weather (calm sea and good visibility). Once the whales were spotted, the boats approached them close enough to tag them using an 8-m long fiberglass or a pneumatic gun (Heide-Jørgensen et al., 2001). If the tagging failed because the whales started diving, the boats spread out and searched for the whales again, then the procedure was repeated until the whales were tagged. Skin samples were also collected for genetic studies and sex identification (Heide-Jørgensen et al., 2013). By comparing the length of the whales to the size of the boats, the length of the whales was estimated. The duration of the tracking of the individual whales varied between years, from a couple of months to a couple of years: May-June 2001, May-November 2002, May-December 2003, April-October 2005, April-September 2006, April-November 2008, April 2009 - July 2011 (Figure 8.1).

8.2.2 Environmental data

Our study focused on the impact of temperature on the distribution of bowhead whales, so the main predictor was the sea surface temperature (SST). These remote sensing environmental data were measured by the satellite system of Copernicus Climate Change Service (`https://climate.copernicus.eu/`). All the temperatures lower than -1.7°C were replaced uniform randomly by values between -1.7°C and -1.8°C, because the freezing temperature of sea water is within this range (Overland et al., 1986; Overland, 1990). The data was distributed through the geographic grid system made of meridians and lines of latitude to create a grid of squares, denoted cells, of size $0.083° \times 0.083°$, or roughly 10×10 km.

8.2.3 Location filtering

Whale position data were collected through the ARGOS system. However, non-Gaussian errors always exist in these data sets together with the influence of other environmental factors (Jonsen et al., 2005; Patterson et al., 2008). The method of Albertsen et al. (2015) was applied to correct these errors. Positions of the whale were allocated to the same cells as in the environmental data of size 10×10 km. The whale positions were often scarce (approximately 131 minutes between each position on average), so we interpolated positions at 5 min intervals from the linear trajectory between corrected positions to estimate the duration the whales spent in each cell. Interpolated on-land positions were removed based on the General Bathymetric Chart of the Oceans (GEBCO) database (`http://www.gebco.net/`).

8.2.4 Habitat modelling

We modelled the duration of the time spent by the whales at each cell per day. Our hypothesis was that the longer the time whales spent in a given cell, the larger the prob-

ability that the whale had encountered suitable environmental conditions (i.e. SST). We only included cells where at least one whale appeared at some time interval during the 10 years of observations, and we considered a global spatio-temporal model to capture the preferable environmental conditions for a 10-year period. We fitted four different seasonal models for four seasons spring (January-March), summer (April-June), autumn (July-September), and winter (October-December) to avoid the effect of whale seasonal migration. We used a compound Poisson-gamma distribution, which is a special case of a Tweedie distribution, to model the continuous positive duration of bowhead whales at each cell when they appeared there, and with zero observations at the days they did not. The compound Poisson-gamma distribution allows for exact zeros but is otherwise continuous. Let Y be a response variable following a compound Poisson-gamma distribution. Assume that the mean $E(Y) = \mu > 0$, and the variance $\mathrm{Var}(Y) = \phi\mu^p$, then $Y \sim \mathrm{Tweedie}_p(\mu, \phi)$ where p and $\phi > 0$ are variance power and dispersion parameters, respectively. The compound Poisson-gamma distribution has $1 < p < 2$.

The response variable is the duration the whales spent at each cell at each time point, regressed on SST within each cell, and an offset with the total number of observed whales at the given time point. The interaction term between SST and cell corrects for all time-invariant unmeasured confounders at each cell, such as depth, distance to coast, or bowhead whale specific site-fidelities. We included an offset with the number of tagged whales n_t at day t, because the number of tagged whales was not constant over time.

The regression provided estimates of cell-specific intercepts and temperature effects (the slopes). The averages of the slopes and intercepts reveal the overall effect of temperature on the choice of locations of the whales. In the generalized linear regression, the log-link was used. The model can be written in R as

```
glm(formula = duration ~ cell-1 + cell:SST,
    family = tweedie(link.power = 0), offset = n)
```

using the package `statmod` (Giner & Smith, 2016) and `tweedie` (Dunn, 2017). However, we did not use R but our own Python code in the models.

The common way to fit the model is with maximum likelihood estimation (MLE) for a series of fixed values of the variance power parameter p, and the AICs are extracted. Then the (interpolated) value of p that minimizes the AIC is chosen (Dunn & Smith, 2018). Due to the interaction of more than 14,500 cells in the spring model, we have more than 29,000 parameters. With such large number of parameters, the loss function is highly complex with numerous local minima, so comparing different model AICs may not give an accurate assessment. A simpler way is to estimate p based on the assumption $\mathrm{Var}(Y) = \phi\mu^p$: hence p was found by the linear regression $\log(\mathrm{Var}(Y_c)) = \log(\phi) + p\log(\mu_c)$ within each cell for every cell c (Dunn & Smith, 2018), separately for each season. The values of p of our spring, summer, autumn, and winter models were estimated to 1.69, 1.74, 1.80, and 1.80, respectively (Table 8.1). All of these values satisfy the assumption of Poisson-gamma model, $1 < p < 2$.

8.2.5 Implementation

The dataset was too big to use `glm.fit` in R. Therefore, we implemented the GLM model in Python 3.8 using CuPy with NVIDIA Quadro P4000 of 8 GB of RAM. The sparsity of the design matrix due to the interaction of SST with each cell was exploited

for GPU computing based on CUDA (Nickolls et al., 2007). The matrix-matrix/matrix-vector multiplication was sped up at least 300 times to CPU computing using SciPy (Fatahalian et al., 2004).

8.2.6 Optimization

The classical IRLS algorithm for GLM fitting was used (Green, 1984). The CUDA platform design prefers single-precision arithmetic (32 bits) to double-precision arithmetic (64 bits), hence it was much faster to use float format to double format as well as it costs only 50% of constrained GPU memory. However, the GLM model used the log-link, therefore a gradient explosion can easily happen if the gradient becomes too big at some step, similar to the same phenomenon in deep learning (Pascanu et al., 2013). Therefore, we used double-precision arithmetic. Nevertheless, gradient explosion still happened sometimes. To solve this issue, we used the step-halving method of `glm2` (Marschner, 2011). The idea is to halve the IRLS step size whenever it leads to gradient explosion. It is also used to assure that the deviance decreases after each step, similar to the Armijo condition (Armijo, 1966).

8.3 Results

There were 29 to 5466 locations per whale, with tracking durations ranging from 4.25 to 489.29 days (Figure 8.1). The total travelling distance estimated linearly from the corrected Argos's data ranged from 179 to 16,581 km. In 2001, only 2 months of data were collected but more data were collected in the subsequent years, including the whole period between April 2009 and July 2011.

The average sea surface temperature of June in our study area showed a general spatial pattern during 2001-2011 (Figure 8.2). Warm water was found in Davis Strait, along West Greenland up to Disko Bay, and in the North Water. Cold water was found throughout the Canadian Arctic Archipelago and Baffin Bay was dominated by a large pool of cold water. Compared to 2001, the warm water from the south was found increasingly further north between 2002 and 2010.

To assess the temperature trend in the bowhead whale habitat, we estimated the slope of each cell using linear regression (Figure 8.3). Positive slopes, i.e. increasing trends, were detected in most cells (more than 95% of slopes > 0.01), except some small areas outside Disko Bay and north of Baffin Bay. The largest increases were observed in Disko Bay, the north and the south of Baffin Bay and along its east coast. The water in the central part of Baffin Bay had the lowest temperature increases, partly because of the mixing of warm water from Southwest Greenland and the cold polar water from the north. The average slope was 0.04 (Figure 8.3B). This corresponds to an average increase of 0.44 degrees over the 11 year's study period, and an increase of more than one degree in the most affected areas. We also randomly selected 10 cells where the whales appeared and noticed a slow increasing trend in all 10 cells (Figure 8.3C).

We ran the Poisson-gamma model for all the cells where the bowhead whales appeared at some point. The average coefficients were negative in all four seasonal models, implying that the whales prefer colder waters (Figure 8.4A and Table 8.1). The intercepts indicate the log of the average durations per day in each cell at 0°C. Some slopes had estimated large negative values in all seasons, from -75.11 to -56.24, probably due to a few outliers within these cells. There were many zeros in the data ($> 91\%$ of zeros in each cell), thus,

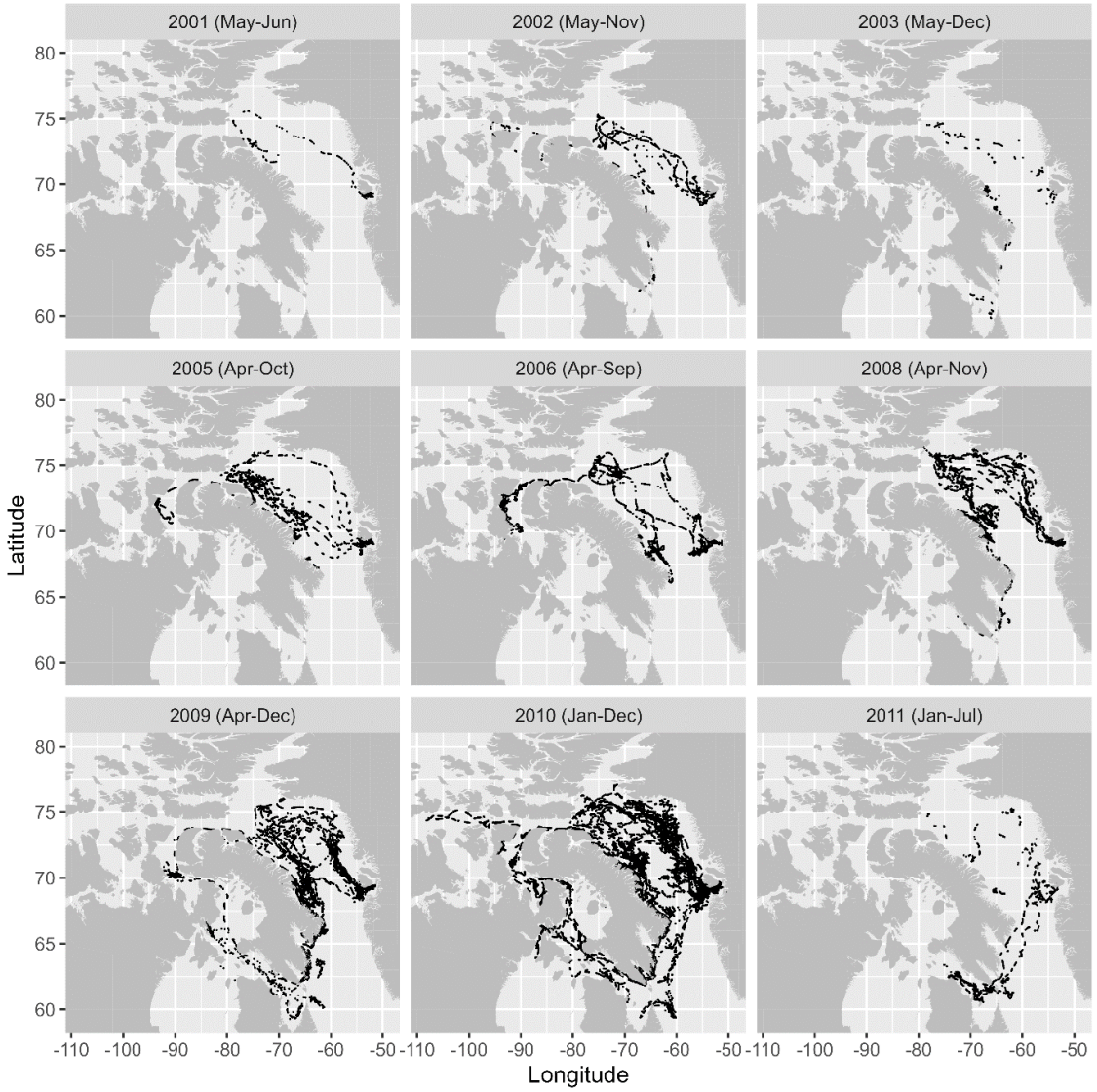

Figure 8.1: Positions of tagged bowhead whales over the years: 2001— 2003, 2005, 2006, 2008— 2011.

the model might need to be extended to zero-inflated Tweedie models to deal with the imbalanced data (Zhou et al., 2020). The estimated values of p were similar in all seasonal models (between 1.69 and 1.80), stating that the relationships between mean and variance were stable between seasons. The average slopes were large, suggesting that the whales spent from 2.7 (in autumn) to 66.7 times (in summer) longer at a cell if it is 0.5°C colder (Table 8.1).

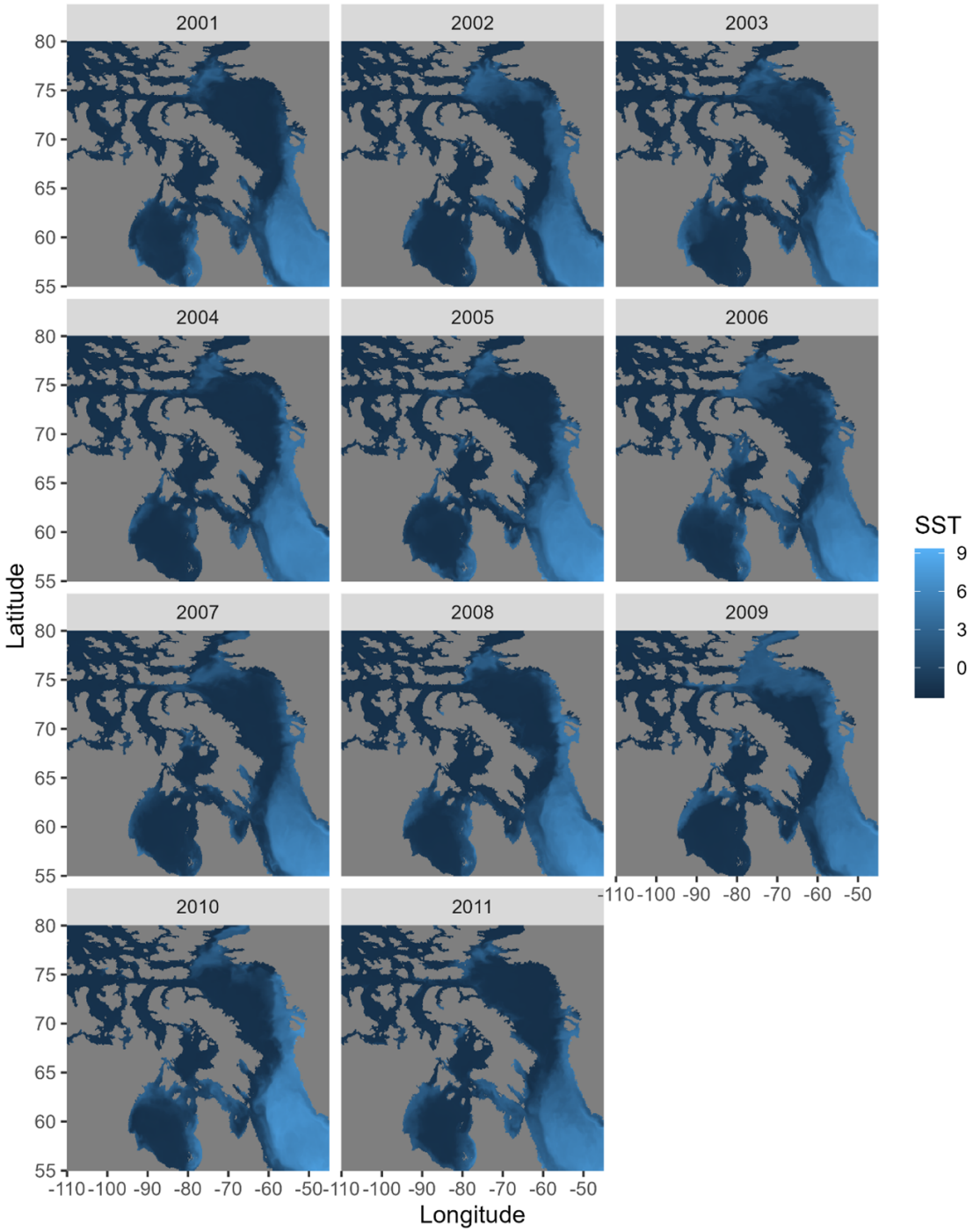

Figure 8.2: Average sea surface temperature of June during 2001-2011 in Baffin Bay and the Canadian Arctic Archipelago.

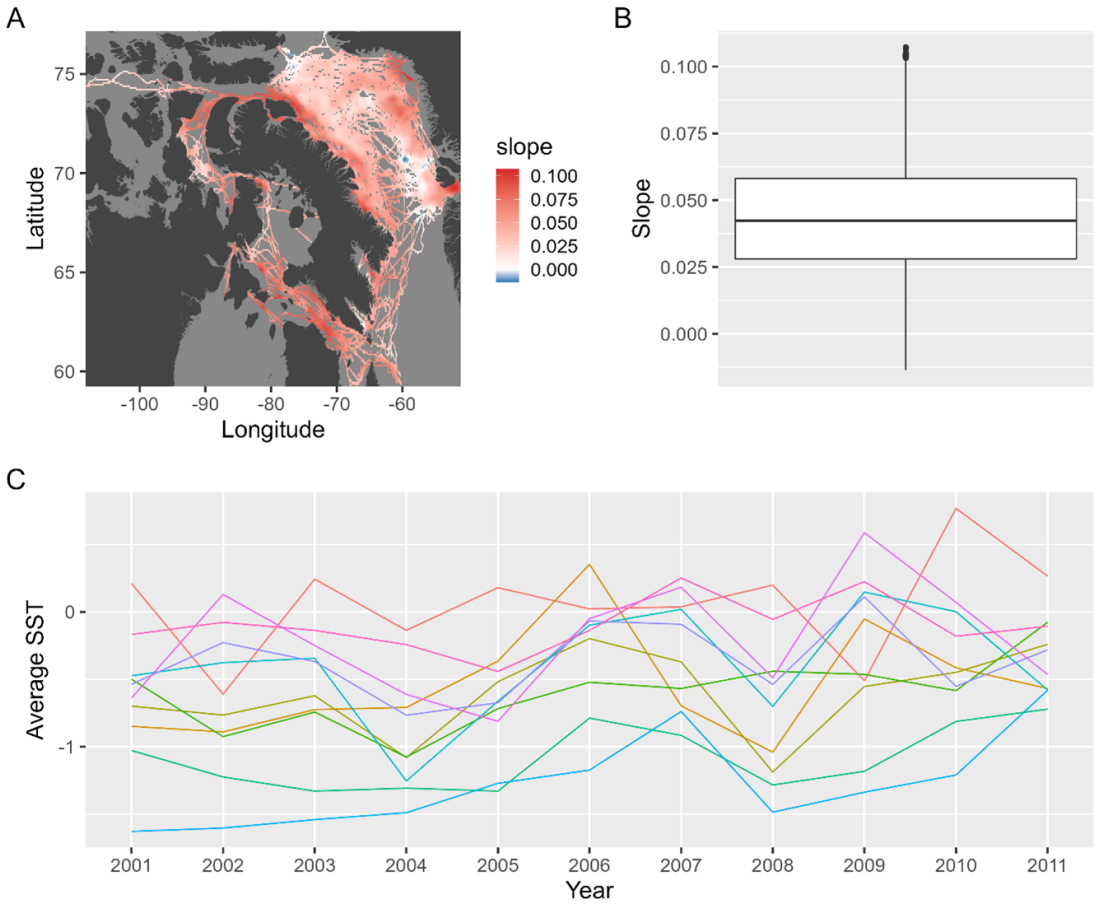

Figure 8.3: Changes in sea surface temperature in the study area after 2000. A) Spatial distribution of slopes, B) Boxplot of slopes: centerline indicates 50th quantile; the bottom and the top of box indicate 25th and 75th quantiles respectively; and dots represent outliers, and C) Yearly average temperature of 10 randomly selected cells where the whales appeared.

	Intercept	Slope	p
Spring	-56.24	-4.51	1.69
Summer	-75.11	-8.40	1.74
Autumn	-62.72	-2.03	1.80
Winter	-66.03	-3.91	1.80

Table 8.1: Slopes, intercepts, and values of the power parameter p of the four seasonal models. The slopes are the average SST effects. The intercepts indicate the log of the average durations per day in each cell at 0°C. Negative slopes imply that the whales prefer colder waters.

We also calculated the weighted mean of slopes and intercepts of each seasonal model, with weights the inverse of the variances of the slope estimates and intercept estimates. It had weighted average slopes between -1 and 1 and average intercepts between -7.7 and -5.6 (results not shown).

The slope estimates in the 95% central part of the distribution (between the 2.5th and the 97.5th percentiles) were between -23.8 and 7.8. In general, there is a small trend toward negative slopes in all seasons (Table 8.2 and Figure 8.5).

Season	Number of slopes	Negative slopes	Positive slopes	Average slope
Spring	13838	7576	6193	-0.75
Summer	10486	5030	5419	-0.25
Autumn	4157	2162	1980	-0.57
Winter	1766	931	806	-0.74

Table 8.2: Statistics of slope estimates inside the 2.5th-97.5th percentiles in the four seasonal models.

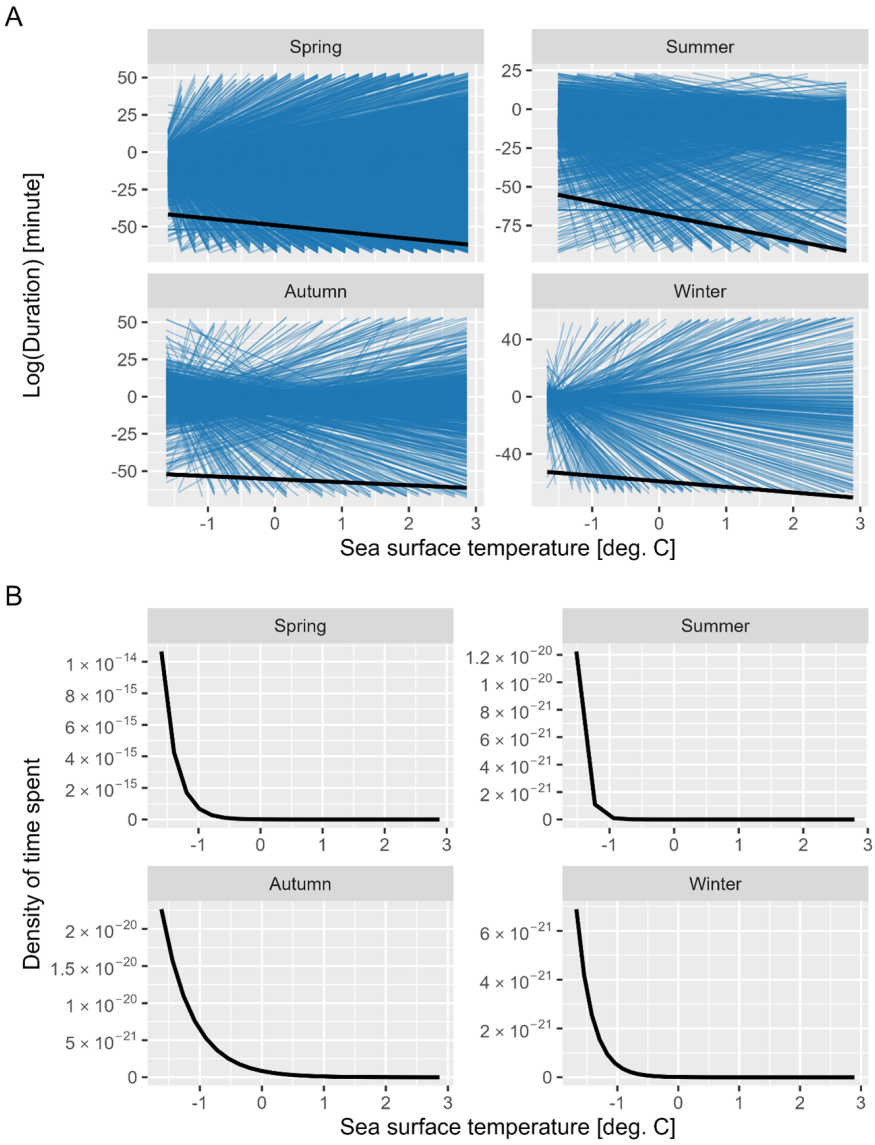

Figure 8.4: A) Regression lines for each cell (in blue). The black lines show the average regression lines (average SST effect) with negative average SST coefficients in all four seasonal models (Table 8.1). Negative slopes imply that the whales prefer colder waters. B) The average curves on the original scales are shown with predicted density of the fraction of time spent at different temperatures summed over all cells.

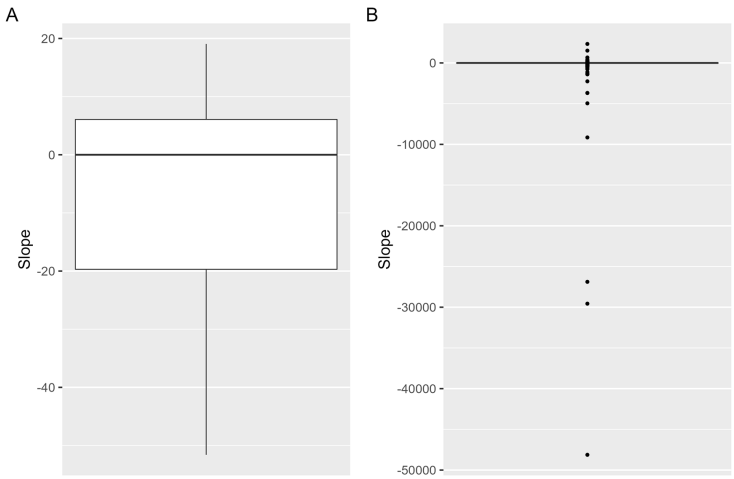

Figure 8.5: A) Boxplot of slope estimates without outliers (only datapoints in the range of 1th and 99th percentiles): centerline indicates median; the bottom and the top of box indicate 3th and 97th percentiles, respectively; lower and upper extremes indicate 1th and 99th percentiles, respectively; B) boxplot including outliers (all estimates).

There were more extreme values of negative slope estimates than for the positive estimates (Figure 8.6). Most slope estimates are around 0 in all models. In winter and spring, there were more cells with low average temperature (< 0°C in spring and < -1°C in summer). In summer and autumn, more cells had higher average temperature, $1.5 - 3$°C in summer and $0 - 1$°C in autumn. However, in autumn there were also cells having average temperature of around -1.5°C, possibly because the temperature is lower at the end of autumn.

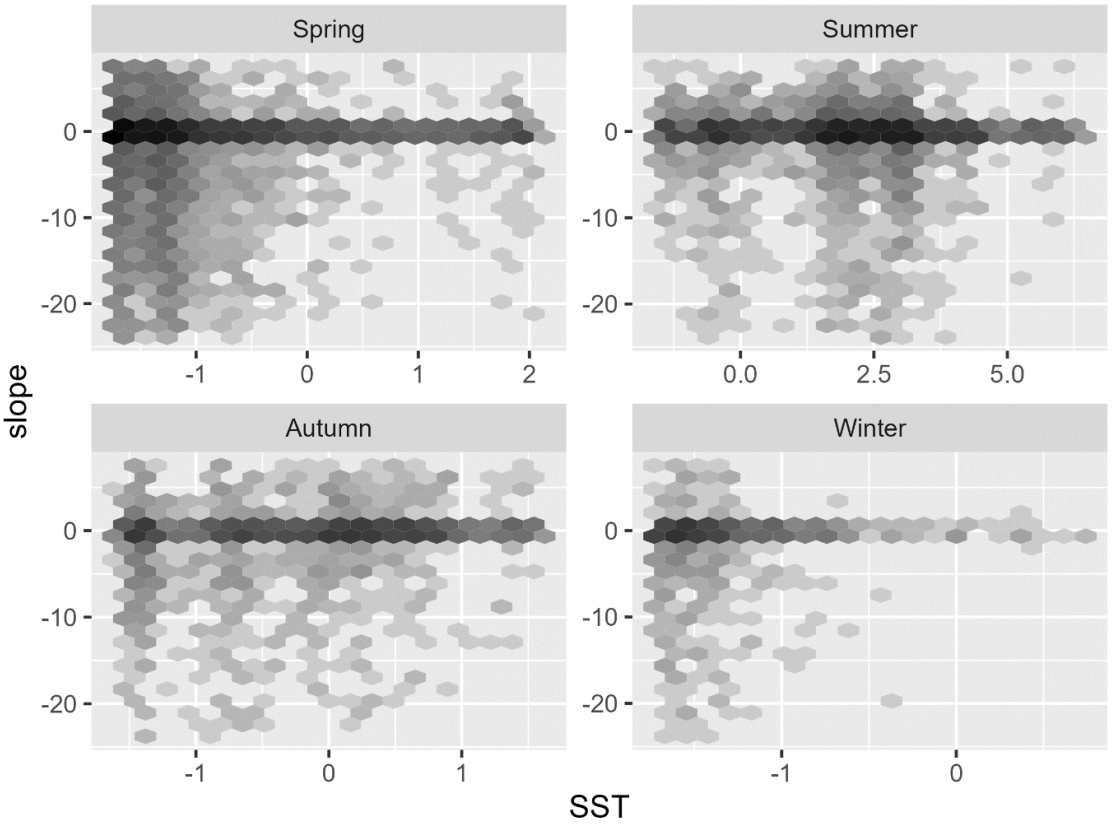

Figure 8.6: Estimated slopes inside the 2.5th-97.5th percentiles against the seasonal average temperature of cells where whales appeared, illustrated by a hexagon plot. The darker the color of a hexagon, the more data points inside it.

8.4 Discussion & Conclusions

In this study, we investigated the relationship between sea surface temperature and movements of bowhead whales. The study was based on a large dataset with up to 9.6 million datapoints. We showed that a GLM can be suitable for such large data sets without the need to switch to more complex models. To deal with a high number of zeros, Poisson-gamma models were used, which are special cases of Tweedie GLM models, that are not as well-known models as Poisson GLM or Gamma GLM. GPU computing allowed us to fit the model with a high number of parameters using Newton method IRLS.

The averages of estimated slopes were negative in all four seasons, indicating that the bowhead whales preferred colder areas. The medians were extremely close to 0. The average slope was much smaller in summer than other seasons, suggesting that the whales are more driven to search for colder water in summer when they are in habitats more influenced by warm Atlantic water, than in the other seasons.

The tagging data showed that at the end of spring, the whale started migrating along West Greenland to northern Baffin Bay, the east coast of Baffin Island and inside the Canadian Arctic Archipelago. What triggers the departure from Disko Bay remains unknown, but the spring influx of warm water to Disko Bay could be involved because during cold conditions in winter and spring bowhead whales are feeding actively in Disko Bay (Laidre et al. 2007, Heide-Jørgensen et al. 2012). This influx of warm water in spring occurs at 3-400m depth (Madsen et al., 2001). It is however not possible to assess the temperature at these depths at the seasonal resolution of cells applied to this study. For this study, only remote sensing of the surface temperature provided the necessary resolution.

The study confirmed the overall observation that the temperature is negatively correlated with bowhead whale distribution. This agrees with the findings in Chambault et al. (2018) that found that bowhead whales targeted a narrow range of SSTs from –0.5 to 2°C. A main difference is that Chambault et al. used a much larger grid, 0.5×0.5 decimal degree (approximately 60×60 km) and monthly data instead of daily data for predicting habitat suitability. The larger dataset allowed to capture more of the variance, and our model also included the interaction term between SST and cell, allowing the correction for all time-invariant unmeasured confounders at each cell, such as depth, distance to coast, or bowhead whale specific site-fidelities. The SST data showed that the temperature of sea water has increased slowly in most of the area in our study between 2001-2011 (more than 95% of the cell temperature increases exceeded 0.01 degrees/year). It confirmed the finding of Alexander et al. (2018) based on simulations that strong warming has been happening in the Arctic. This could also be observed from the changes of the temporal migration patterns of bowhead whales. Based on reports from the whaling stations it was calculated that the mean departure date for bowhead whales in Disko Bay was around 5 June (interquartile range 20 May to 14 June) for the period 1780–1837 (Eschricht and Reinhardt, 1861). The whales tracked in the present study departed 1.5–3 weeks earlier (Laidre and Heide-Jørgensen, 2012) and it is likely due to the warmer waters compared to the earlier period that coincided with the little ice age (Mann, 2003).

Our work can be extended to overcome some limitations. Note that our model is not sophisticated enough to include movements, so we can only conclude about selection of cells. A high negative slope means that the whales preferred to stay in colder areas, but it does not inform us about how they moved. The correlation coefficients with the neighbouring cells in the past could show us the whale's movement, e.g., by using

Generalized Linear Autoregressive Moving Average Models (Dunsmuir, 2015).

Chambault et al. (2018) showed that GAMs provide more flexible ways to interpret the data than GLMs. Our analysis can be extended easily to a GAM using smooth functions on SST. Other important sea ice related predictors such as sea ice concentration could also be included. However, sea ice and SST are highly correlated and Chambault et al. (2018) found SST to be a more important predictor of bowhead whale distribution, partly because of lack of sea ice during the summer months. Another limitation is the linear interpolation method we have used to estimate the duration of bowhead whales at each cell. Finally, the huge number of zeros in the data are not well captured by the standard Tweedie model. One way to improve the model is to extend to zero-inflated Tweedie models (Zhou et al., 2020).

Acknowledgements

We would like to thank Jonas Peters for fruitful discussions. This study was funded by the Greenland Institute of Natural Resources. The study was conducted under the general permission from the Greenland Government to the Greenland Institute of Natural Resources for tagging baleen whales. SD was supported by Independent Research Fund Denmark, case: 9040-00215B and the Novo Nordisk Foundation NNF20OC0062958, MCN was funded by Greenland Research Council and MPHJ were funded by the Greenland Institute of Natural Resources.

References

Albertsen, C. M., Whoriskey, K., Yurkowski, D., Nielsen, A., & Flemming, J. M. (2015). Fast fitting of non-Gaussian state-space models to animal movement data via Template Model Builder.

Alexander, M. A., Scott, J. D., Friedland, K. D., Mills, K. E., Nye, J. A., Pershing, A. J., ... & Carmack, E. C. (2018). Projected sea surface temperatures over the 21st century: Changes in the mean, variability, and extremes for large marine ecosystem regions of Northern Oceans. Elementa: Science of the Anthropocene, 6.

Armijo, L. (1966). Minimization of functions having Lipschitz continuous first partial derivatives. Pacific Journal of mathematics, 16(1), 1-3.

Bonat, W. H., & Kokonendji, C. C. (2017). Flexible Tweedie regression models for continuous data. Journal of Statistical Computation and Simulation, 87(11), 2138-2152.

Chambault, P., Albertsen, C. M., Patterson, T. A., Hansen, R. G., Tervo, O., Laidre, K. L., & Heide-Jørgensen, M. P. (2018). Sea surface temperature predicts the movements of an Arctic cetacean: the bowhead whale. Scientific Reports, 8(1), 1-12.

Christiansen, R., Baumann, M., Kuemmerle, T., Mahecha, M. D., & Peters, J. (2021). Towards causal inference for spatio-temporal data: Conflict and forest loss in Colombia. Journal of the American Statistical Association, 1-28.

Citta, J. J., Olnes, J., Okkonen, S. R., Quakenbush, L., George, J. C., Maslowski, W., ... & Heide-Jørgensen, M. P. (2021). Influence of oceanography on bowhead whale (Balaena mysticetus) foraging in the Chukchi Sea as inferred from animal-borne instrumentation. Continental Shelf Research, 104434.

Dunn, P.K. (2017). Tweedie: Evaluation of Tweedie Exponential Family Models. R package version 2.3.0.

Dunn, P. K., & Smyth, G. K. (2018). Generalized linear models with examples in R (p. 562). New York: Springer.

Dunsmuir, W. T. (2015). Generalized linear autoregressive moving average models. Handbook of discrete-valued time series, 51-76.

Eschricht, D. F., and Reinhardt, J. 1861. Om nordhvalen (Balaena mysticetus L.) navnlig med hensyn til dens udbredning i fortiden og nutiden og til dens ydre og indre saerkjender. K. Danske Videnskabernes Selskabs Skrifter, Series 5, Naturvidenskabelig og Mathematisk Afdeling, 5: 433–590 (in Danish).

Fatahalian, K., Sugerman, J., & Hanrahan, P. (2004). Understanding the efficiency of GPU algorithms for matrix-matrix multiplication. In Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware (pp. 133-137).

Fortune, S.S.M.E., S. H. Ferguson, A. W. Trites, B. LeBlanc, V. LeMay, Justine M. Hudson, M. F. Baumgartner. 2020. Seasonal diving and foraging behaviour of Eastern Canada-West Greenland bowhead whales. Mar Ecol Prog Ser, Vol. 643: 197–217.

George, J. C., Philo, L. M., Hazard, K., Withrow, D., Carroll, G. M., & Suydam, R. (1994). Frequency of Killer Whale (Orcinus orcd) Attacks and Ship Collisions Based on Scarring on Bowhead Whales (Balaena mysticetus) of the Bering-Chukchi-Beaufort Seas Stock. Arctic, 247-255.

George, J.C., J.G.M. Thewissen, A. Von Duyke, G.A. Breed, R. Suydam, T.L. Sformo, B.T. Person, H.K. Brower Jr. 2021. Life history, growth and from. In The Bowhead Whale: Balaena Mysticetus: Biology and Human Interactions, eds. J.C. George J.G.M. Thewissen, 87-115.

J. C. George et al. The bowhead whale: Balaena mysticetus: Biology and human interactions. Academic Press, 2020.

Giner G, Smyth GK (2016). statmod: probability calculations for the inverse Gaussian distribution. R Journal, 8(1), 339-351.

Givens, G.H. and Heide-Jørgensen, M.P., 2021. Abundance. In The Bowhead Whale: Balaena Mysticetus: Biology and Human Interactions, eds. J.C. George J.G.M. Thewissen, 77-86

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. Journal of the Royal Statistical Society: Series B (Methodological), 46(2), 149-170.

Haldiman, J.T., and Tarpley, R.T. (1993). Anatomy and physiology. In "The Bowhead Whale", (J.J. Burns, J.J. Montague, and C.J. Cowles, Eds), Special publication No. 2 of the Society of Marine Mammalogy.

Hansen, A.S., T. G Nielsen, H. Levinsen, S. D. Madsen, T. F. Thingstad, B. W. Hansen. 2003. Impact of changing ice cover on pelagic productivity and food web structure in Disko Bay, West Greenland: a dynamic model approach. Deep-Sea Research I 50: 171–187.

Hansen, K. E., Giraudeau, J., Wacker, L., Pearce, C., & Seidenkrantz, M. S. (2020). Reconstruction of Holocene oceanographic conditions in eastern Baffin Bay. Climate of the Past, 16(3), 1075-1095.

Heide-Jørgensen, M. P., Kleivane, L., ØIen, N., Laidre, K. L., & Jensen, M. V. (2001). A new technique for deploying satellite transmitters on baleen whales: Tracking a blue whale (Balaenoptera musculus) in the North Atlantic. Marine Mammal Science, 17(4), 949-954.

Heide-Jørgensen, M.P., K. L. Laidre, Ø. Wiig, M.V. Jensen, L. Dueck, H.C. Schmidt and R. C. Hobbs. 2003. From Greenland to Canada in ten days: Tracks of bowhead

whales, Balaena mysticetus, across Baffin Bay. Arctic 56: 21-31.

Heide-Jørgensen, M.P., Laidre, K.L., Jensen, M.V., Dueck, L. and L. D. Postma, L.D. 2006. Dissolving stock discreteness with satellite tracking: Bowhead whales in Baffin Bay. Marine Mammal Science, 22(1): 34-45.

Heide-Jørgensen, M.P., E. Garde, N.H. Nielsen and O. N. Andersen. 2012. Biological data from the hunt of bowhead whales in West Greenland 2009 and 2010. Journal of Cetacean Research and Management 12(3): 329-333.

Heide-Jørgensen, M. P., Richard, P. R., Dietz, R., & Laidre, K. L. (2013). A metapopulation model for Canadian and West Greenland narwhals. Animal Conservation, 16(3), 331-343.

Heide-Jørgensen, M.P., Hansen, R.G., O.V. Shpak. 2021. Distribution, migrations, and ecology of the Atlantic and the Okhotsk Sea populations. In The Bowhead Whale: Balaena Mysticetus: Biology and Human Interactions, eds. J.C. George J.G.M. Thewissen, 57-76

Hokkanen, J. E. I. (1990). Temperature regulation of marine mammals. Journal of Theoretical Biology, 145(4), 465-485.

IPCC (2013). Summary for policymakers. In: Stocker TF, Qin D, Plattner GK, Tignor M and others (eds) Climate Change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge

Jonsen, I., Bestley, S., Wotherspoon, S., Sumner, M., & Flemming, J. M. (2015). bsam: Bayesian state-space models for animal movement. R package. R Foundation for Statistical Computing version 0.43, 1.

Jørgensen, B. (1987). Exponential dispersion models. Journal of the Royal Statistical Society: Series B (Methodological), 49(2), 127-145.

Kovacs, K. M., Lydersen, C., Overland, J. E., & Moore, S. E. (2011). Impacts of changing sea-ice conditions on Arctic marine mammals. Marine Biodiversity, 41(1), 181-194.

Laidre, K.L., M. P. Heide-Jørgensen and T.G. Nielsen. 2007. Role of bowhead whale as predator in West Greenland. Marine Ecology Progress Series 346: 285-297.

Laidre K.L. and M.P. Heide-Jørgensen. 2012. Springtime partitioning of Disko Bay, West Greenland by Arctic and sub-Arctic baleen whales. ICES Journal of Marine Science.

Laidre, Kristin L., Harry Stern, Kit M. Kovacs, Lloyd Lowry, Sue E. Moore, Eric V. Regehr, Steven H. Ferguson et al. Arctic marine mammal population status, sea ice habitat loss, and conservation recommendations for the 21st century. Conservation Biology 29, no. 3 (2015): 724-737.

Lindsay, R., and Schweiger, A. (2015). Arctic sea ice thickness loss determined using subsurface, aircraft, and satellite observations, The Cryosphere, 9, 269–283, doi:10.5194/tc-9-269-2015, 2015.

Madsen, S. D., Nielsen, T. G., & Hansen, B. W. (2001). Annual population development and production by Calanus Wnmarchicus, C. glacialis and C. hyperboreus in Disko Bay, western Greenland. Mar Biol, 139(1), 75-83.

Mann, Michael (2003). Little Ice Age. In Michael C MacCracken; John S Perry (eds.). Encyclopedia of Global Environmental Change, Volume 1, The Earth System: Physical and Chemical Dimensions of Global Environmental Change. John Wiley & Sons. Retrieved 17 November 2012.

Marschner, I. C. glm2: Fitting generalized linear models with convergence problems. The R Journal. 2011; 3 (2): 12-15. NVIDIA. CUDA C++ programming guide. NVIDIA,

Aug (2020).

Michel, C., B. Bluhm, V. Gallucci, A.J. Gaston, F.J.L. Gordillo, R. Gradinger, R. Hopcroft, N. Jensen, T. Mustonen, A. Niemi & T.G. Nielsen (2012) Biodiversity of Arctic marine ecosystems and responses to climate change, Biodiversity, 13:3-4, 200-214.

Nickolls, J., Buck, I., Garland, M., & Skadron, K. (2008). Scalable parallel programming with cuda: Is cuda the parallel programming model that application developers have been waiting for. Queue, 6(2), 40-53.

Overland, J. E., Pease, C. H., Preisendorfer, R. W., & Comiskey, A. L. (1986). Prediction of vessel icing. Journal of climate and applied meteorology, 25(12), 1793-1806.

Overland, J. E. (1990). Prediction of vessel icing for near-freezing sea temperatures. Weather and forecasting, 5(1), 62-77.

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., & Matthiopoulos, J. (2008). State–space models of individual animal movement. Trends in ecology & evolution, 23(2), 87-94.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In International conference on machine learning (pp. 1310-1318). PMLR.

Perrin, W. F., Würsig, B. & Thewissen, J. G. M. Encyclopedia of Marine Mammals. (Academic Press, 2009).

Pomerleau, C., Heide-Jørgensen, M. P., Ferguson, S. H., Stern, H. L., Høyer, J. L., & Stern, G. A. (2017). Reconstructing variability in West Greenland ocean biogeochemistry and bowhead whale (Balaena mysticetus) food web structure using amino acid isotope ratios. Polar Biology, 40(11), 2225-2238.

Tarpley, R.J., D.J. Hillmann, J.C. George, J.G.M. Thewissen. 2021. Female and male reproduction. In The Bowhead Whale: Balaena Mysticetus: Biology and Human Interactions, eds. J.C. George J.G.M. Thewissen, 185-211.

Zhou, H., Qian, W., & Yang, Y. (2020). Tweedie gradient boosting for extremely unbalanced zero-inflated data. Communications in Statistics-Simulation and Computation, 1-23.