Final Report

Development of Statistical Methods for Assessing Changes in Whale Vocal Behavior in Response to Mid-Frequency Active Sonar

Submitted to:

Naval Facilities Engineering Command Atlantic under Contract No. N62470-10-D-3011, Task Order 39, issued to HDR, Inc.

Prepared by:

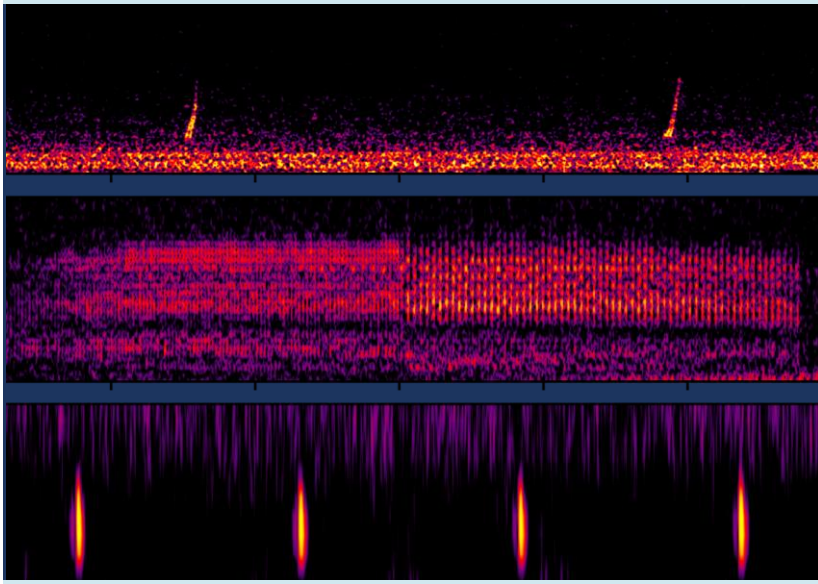
Cornell University

The Cornell Lab of Ornithology

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Centre for Research into Ecological and Environmental Modelling
St. Andrews University
St. Andrews, Fife
Scotland



Submitted by:

HDR

Virginia Beach, VA

20 March 2015

Suggested Citation:

Charif, R.A., C.S. Oedekoven, A. Rahaman, B.J. Estabrook, L. Thomas, and A.N. Rice. 2015. *Development of Statistical Methods for Assessing Changes in Whale Vocal Behavior In Response To Mid-Frequency Active Sonar. Final Report.* Prepared for U.S. Fleet Forces Command. Submitted to Naval Facilities Engineering Command Atlantic, Norfolk, Virginia, under Contract No. N62470-10-3011, Task Order 39, issued to HDR Inc., Virginia Beach, Virginia. 20 March 2015.

Cover Image Source:

Examples of baleen whale sounds targeted in this study. Top: North Atlantic right whale (*Eubalaena glacialis*) upcalls; middle: minke whale (*Balaenoptera acutorostrata*) pulse train; and bottom: fin whale (*Balaenoptera physalus*) 20-Hz notes. See **Figure 3** (page 41).

This project is funded by US Fleet Forces Command and managed by Naval Facilities Engineering Command Atlantic as part of the US Navy's marine species monitoring program.

Executive Summary

Concerns about potential effects of military sonar on cetaceans first arose in the 1990s, with observations of multiple mass strandings of beaked whales at times and places where sonar was known or thought to have been used. Sonar systems implicated in these events included both low-frequency and mid-frequency active sonar (LFAS and MFAS, respectively). Since these initial observations of stranding events, an increasing number of studies have demonstrated a variety of behavioral changes in toothed and baleen whales in response to MFAS both in observational studies during actual military exercises, and in experimental controlled-exposure studies using real and simulated MFAS signals.

In order to further investigate potential changes in behavior in response to MFAS exposure, sounds of marine mammals were recorded in the vicinity of three U.S. Navy training exercises before, during, and after the use of MFAS in Onslow Bay, North Carolina (July 2008), and near Jacksonville, Florida (September–October, and December 2009). The resulting passive acoustic data were analyzed using automated signal-detection software to detect individual sonar transmissions ("pings"), and sounds of North Atlantic right, minke, fin, and sperm whales. In addition, putative right whale "gunshot" sounds that had been detected in an earlier, separate analysis of the Jacksonville recordings were reviewed by two experienced marine mammal acoustic analysts.

Sperm whale click trains were detected on every day of recordings from all three deployments. In all deployments, sperm whale click trains occurred almost continuously during hours of darkness, and rarely during daylight hours, with a few exceptions. Minke whale pulse trains were detected only in the winter Jacksonville deployment. There were no confirmed detections of North Atlantic right whale upcalls or fin whale sounds in any of the three deployments. Most of the impulsive sounds previously identified as right whale gunshot sounds were judged most likely to be from sources other than right whales. The present analysis of the acoustic data provides no compelling evidence of right whale presence in the area during the Jacksonville deployments.

Generalized estimating equations (GEEs) were used to build statistical models predicting the presence or absence of minke and sperm whale vocalizations in 1-minute periods. The model predictions were functions of seven covariates related to the occurrence and timing of sonar pings, and four sonar-independent covariates related to date, time of day, and recording location. GEEs were also used to model changes in the duration of detected minke whale pulse trains using the same set of covariates. Duration models were not applied to the sperm whale data because frequent overlapping of click trains from multiple individuals precluded reliable measurement of durations of discrete vocal events.

For the minke whale presence model, the covariate indicating whether a given minute was *before*, *during*, *between*, or *after* sonar transmissions was retained in the final model. For minke whales, the odds of detecting vocalizations were on average higher in the 24 hours after a sonar exercise compared to the 24 hours before the exercise. However, it is likely that inference on this covariate would have been different for both species if we had applied different criteria for

labelling time periods as *before*, *during*, *between*, or *after* (e.g., using 12-hour rather than 24-hour *before* and *after* periods).

The best fitting presence model for sperm whales contained the factor covariate *Daynight* and the polynomial spline for *Time* providing evidence that during our study the odds of detecting presences of sperm whale vocalizations varied in a diurnal pattern, increasing at night compared to during the day. None of the covariates related to sonar were included in the best-fitting model, suggesting that sonar activity did not significantly affect the occurrence of sperm whale click trains.

For minke whales, the durations of individual detected pulse trains varied in response to sonar activities. The differences consisted of an increase in duration if approximately 40 to 110 sonar pings were detected in the four hours preceding the vocalization and a decrease in duration if approximately 110 to 155 sonar pings were detected in the four hours preceding the vocalization. Although these results indicate that sonar had an effect on the detected duration of minke whale vocalizations during this study, the biological cause or significance of the response observed is unclear. However, the sample size of discrete periods with sonar activity was very low; sonar transmissions were only detected on three days during the only deployment (JAX2) in which minke whale sounds were recorded. Larger sample sizes are needed for stronger inference.

Possible further analyses of these recordings are discussed.

Table of Contents

Acronyms and Abbreviations

dB	decibel(s)
dB re 1 µPa	decibels referenced to 1 microPascal
GEE	generalized estimating equation
GLM	generalized linear model
GLMM	generalized linear mixed model
h	hour(s)
HF	high frequency
Hz	Hertz
JAX	Jacksonville
kHz	kilohertz
km	kilometer(s)
LF	low frequency
LFAS	low-frequency active sonar
m	meter(s)
MARU	marine autonomous recording unit
MFAS	mid-frequency active sonar
NM	nautical mile(s)
RL	received levels
rms	root mean square
USWTR	undersea warfare training range
VIF	variance inflation factors

Glossary of Technical Terms

block

The term 'block' is used in the context of fitting models with generalized estimating equations (GEEs). We group our observations into blocks in the order they were observed assuming that observations are correlated in the same block but not correlated between blocks. The size of the blocks is determined using the *acf* function in R.

bs()

A function of the *splines* R package used for fitting smoothing splines to numeric (continuous or discrete) covariates. For this study, we only fitted polynomial splines as opposed to regression splines.

correlation structure

This term refers to the pattern in correlation assumed within the blocks when fitting GEEs. The *geeglm* function of the *geepack* R package offers several choices of correlation structures. Some choices use a model to describe the change in correlation with increasing lag. We use the default 'independence' which does not use a model to describe the change in correlation but estimates the correlation between the different lags individually.

error structure

This term is often used in the context of fitting regression models. For linear regression models, the error structure is generally assumed to be normal. For generalized models such as GLMs, GAMs, GEEs, etc. other distributions may be used such as the Poisson, binomial, gamma, etc. Error structure refers to the type of distribution that we assume for the errors after fitting the model.

page

A 60-second segment of an audio recording processed by the algorithm used for detecting minke whale pulse trains and sperm whale click trains. The algorithm operates on consecutive non-overlapping pages. Each page was processed independently of previous and successive pages.

sonar exercise

A period of time encompassing all consecutive sonar pings with no gap longer than 48 hours for a given deployment.

1. Background and Objectives

Concerns about potential effects of military sonar on cetaceans first arose in the 1990s, with observations of multiple mass strandings of beaked whales at times and places where sonar was known or thought to have been used (D'Amico et al. 2009, and references therein). Sonar systems implicated in these events included both low-frequency and mid-frequency active sonar (LFAS and MFAS, respectively). LFAS is usually defined as active sonar operating at frequencies < 1 kilohertz (kHz) (D'Amico et al. 2009). MFAS is defined by various sources as operating in either the 1–10 kHz, or 3–14 kHz frequency band (e.g., D'Amico et al., 2009; see Sivle et al., 2012 for different usage of LFAS and MFAS).

Since these initial observations of stranding events, an increasing number of studies have demonstrated a variety of behavioral changes in toothed and baleen whales in response to MFAS, both in observational studies during actual military exercises, and in experimental controlled-exposure studies using real and simulated MFAS signals (see, for example, McCarthy et al., 2011; Tyack et al., 2011; Melcón et al., 2012; Goldbogen et al., 2013; Kuningas et al., 2013; Moretti et al., 2014).

In 2008 and 2009, sounds of marine mammals were recorded in the vicinity of U.S. Navy training exercises before, during, and after periods of MFAS use in Onslow Bay, North Carolina, and near Jacksonville, Florida. This document reports on analysis of these recordings, targeting sounds of North Atlantic right whales, fin whales, minke whales, and sperm whales. The objectives of the analysis were to:

a. Determine whether each of the target whale species was present and vocalizing in the vicinity at the times of the training exercises;

b. Quantitatively describe seasonal and diel patterns of detected vocal activity for each target species;

c. Develop and apply new statistical approaches to determine whether any measurable change in detected vocal activity of target species was associated with the use of MFAS.

In this report, we first describe the methods used to acquire recordings, detect target sounds (including MFAS signals), and characterize their temporal patterns of occurrence. We then describe statistical methods used to detect potential changes in whale vocalizations in relation to sonar activity (**Section 2**). Results from application of these methods are presented in **Section 3**, and discussed in **Section 4**.

The recordings from the Jacksonville deployments have previously undergone a preliminary analysis for the sounds of right, minke and sperm whales by researchers at Bio-Waves, Inc. using different methods (Norris et al., 2012). Except where noted below, detailed comparisons of results from the Bio-Waves analysis and the present one are beyond the scope of this report.

This report is a collaboration between the Bioacoustics Research Program at Cornell University (responsible for collection and processing of acoustic data) and the Centre for Research into

Ecological and Environmental Modelling at St. Andrews University (responsible for development and application of statistical models).

Analysis of the sounds of delphinids recorded during these MFAS exercises is described in a separate report prepared by Bio-Waves, Inc. and the University of St. Andrews (Oswald et al., 2014).

2. Methods

2.1 Data Acquisition

Marine autonomous recording units (MARUs, Clark et al., 2002) were deployed in two areas for periods of time in 2008 and 2009 that encompassed U.S. Navy training exercises that included the use of MFAS. MARUs were deployed with two different recording configurations. "High-frequency" (HF) MARUs recorded continuously with a 32-kHz sample rate, resulting in a nominal recording band of 0-16 kHz. "Low-frequency" (LF) MARUs recorded continuously with a sample rate of 2 kHz, resulting in a nominal recording band of 0–1 kHz. Only HF MARUs were capable of recording MFA sonar signals. Both configurations could record right, fin, and minke whales; sperm whales could be reliably recorded on HF MARUs and in some cases on LF MARUs. Each MARU was tethered approximately 2 m above the seafloor, attached to an anchor by an acoustic release mechanism. At the conclusion of the deployment period, the acoustic release was activated by a signal from the recovery vessel, and the positively buoyant MARU rose to the surface for retrieval. The maximum recording duration for the HF MARUs, limited by the instrument's hard disk storage capacity, was 21 days.

The following sections provide details of the specific deployments.

2.1.1 Onslow Bay

The Onslow Bay research site is located within the U.S. Navy's Cherry Point operating area, approximately 100 kilometers (km) (54 nautical miles [NM]) south of Cape Lookout, North Carolina (**Figure 1**). Seven MARUs were deployed on 6 July 2008, six of which were recovered on 27 July 2008 (**Table 1**), for a total of 20 complete underwater recording days. The missing MARU (Site SB4) did not resurface after several attempts during the recovery process. Depths of individual MARU sites varied from 73 meters (m) to approximately 365 m. MARUs were deployed approximately 12.5 km apart to encompass the target acoustic survey area.

All MARUs deployed at Onslow Bay used the HF recording configuration (above).

2.1.2 Jacksonville

The Jacksonville research site is located on the planned USWTR within the U.S. Navy's Jacksonville (JAX) operating area, approximately 110 km (59 NM) offshore of Jacksonville, Florida (**Figure 2**). Nine MARUs were deployed over the Florida-Hatteras Slope in each of two deployment periods (**Table 2**): 13 September to 8 October 2009 (Deployment 1) and again 4 December 2009 to 7 January 2010 (Deployment 2). Due to the position of the MARUs along the Florida-Hatteras Slope, site depths ranged from 45 m to more than 300 m. The same sites were used for both deployments.

In each of the Jacksonville deployments, six MARUs used the HF recording configuration, and three used the LF configuration, as described above (**Figure 2, Tables 2** and **3**).

2.2 Description of Target Sounds

The four target species of whales were selected based on their known occurrence in the western North Atlantic, and on the lack of published data on their potential responses to sonar. This study focused on the following sound types produced by each species:

- **North Atlantic right whales** (*Eubalaena glacialis*) produce underwater sounds that can be grouped into three broad categories: upcalls (also called contact calls; **Figure 3A**), other tonal calls, and gunshot sounds (Parks & Tyack, 2005). The present study focused on upcalls and gunshots.
    - *Upcalls* are commonly produced by right whales of all age-sex classes and are more stereotyped in structure than other tonal calls. These characteristics have made upcalls the signal of choice for detecting right whale presence in studies using both manual and automated signal-detection approaches (e.g., Mellinger et al., 2007; Urazghildiiev & Clark, 2007; Urazghildiiev et al., 2009; Van Parijs et al., 2009; Clark et al., 2010; Morano et al., 2012a). Upcalls are characterized by a frequency range of 50–400 Hz, a duration of 0.3-1.5 seconds, and broadband source levels up to 155 decibels referenced to 1 micropascal at 1 m (rms dB re 1 µPa at 1 m) (Parks & Tyack, 2005; Trygonis et al., 2013).
    - *Gunshot* sounds are broadband impulsive sounds, produced at typical source levels around 196 decibels (dB) peak-to-peak and 189 dB rms re 1 µPa at 1 m (20 Hz–22 kHz, Parks et al., 2005; see also Trygonis et al., 2013). An individual gunshot can contain one to four pulses separated by < 100 ms (Parks et al., 2005). In the Bay of Fundy, where gunshot sounds have been most extensively studied, available evidence suggests that bouts of repeated gunshots are produced only or primarily by males, and that these sounds function in a reproductive context (Parks et al., 2005; Parks & Tyack, 2005; Parks et al., 2011; Parks et al., 2012). However, in the southeast U.S. calving grounds, isolated gunshots have been recorded from right whale surface active groups when no adult males were present (Trygonis et al., 2013). Gunshots have also been recorded from a female southern right whale (*Eubalaena australis*) on at least one occasion (Clark, 1983; gunshots were originally called “underwater slaps” by Clark).
- **Minke whales** (*Balaenoptera acutorostrata*) in the western North Atlantic produce trains of low frequency pulses (**Figure 3B**), with most energy distributed between 50 and 400 Hz (Winn & Perkins, 1976; Mellinger et al., 2000; Risch et al., 2013). Individual pulse trains typically last between 30 and 60 seconds, with inter-pulse intervals of approximately 0.3–0.7 seconds (Mellinger et al., 2000; Risch et al., 2013). Average source levels (SL_{rms}) for minke whale pulse trains are 164 – 168 dB re 1 µPa at 1 m (Risch et al., 2014b).
- **Fin whale** (*Balaenoptera physalus*) males produce songs consisting of long repetitive series of simple notes, each of which consists of a rapid downsweep between approximately 23 and 18 Hz over approximately 1 second (Watkins et al., 1987; Croll et

al., 2002). These “20-Hz notes,” which were the target signals in the present study, are highly stereotyped in structure and are repeated at very regular intervals within songs that typically last 1–20 minutes (**Figure 3C**). Songs are typically delivered in bouts that may last in excess of 30 hours (h), with pauses between songs lasting from 1 minute up to approximately 2 h (Watkins et al., 1987). Inter-note intervals within songs in the western North Atlantic vary seasonally at a given location (Watkins et al., 1987; Morano et al., 2012b). Source levels for fin whale 20-Hz notes average 189 dB re 1 µPa at 1 m (Sirović et al., 2007; Weirathmueller et al., 2013).

- **Sperm whales** (*Physeter macrocephalus*) produce four different types of broadband clicks: usual or foraging clicks (Møhl et al., 2003), slow clicks (Weilgart & Whitehead, 1988), creaks (Goold & Jones, 1995), and codas (Watkins & Schevill, 1977). The four types of clicks are differentiated by duration of the click trains, dominant frequency, click rate, and inter-click-intervals. This study focused on foraging clicks (**Figure 4**) because past research has found that sperm whales spend a significant time, greater than 72 percent, in foraging dives and produce foraging clicks during approximately 68 percent of the dive cycle (Watwood et al., 2006). Sperm whale foraging clicks are sequences of impulsive signals characterized by a frequency band of 0.2–32 kHz, a dominant frequency around 5 kHz, inter-click interval of 0.025–1.25 seconds, and click duration of 2–24 ms (Backus & Schevill, 1967; Weilgart & Whitehead, 1988). Foraging clicks are also characterized by a high source level, up to 223 dB *re* 1 µPa at 1 m peak equivalent rms (Møhl et al., 2000).
- **MFAS transmissions (“pings”)** consisted of a variety of tonal signals, typically one to several seconds in duration, with varying combinations of constant-frequency, upsweep, and downsweep elements. Pings occurred in two distinct frequency bands, centered around approximately 3.5 and 7 kHz (**Figure 5**). Operating source levels of the sonars in these recordings are not available, but source levels of 223 to 235 dB re 1µPa at 1 m have been reported for recent tactical MFAS systems (see references in D’Amico & Pittenger, 2009) although maximum source levels are classified

2.3 Detection of Target Sounds

Automated detection algorithms were used to find sonar transmissions and sounds of right, minke, fin, and sperm whales in the recordings from both sites. For the Jacksonville deployments, detectors for sonar pings and sperm whale click trains were run only on recordings from the six HF sites, since these signals occur in bands that exceeded the upper frequency limit of the LF recordings.

Mid-frequency active sonar transmissions were detected using the band-limited energy detector algorithm in Raven Pro (Bioacoustics Research Program, 2014). Two detector configurations were used, one for the 2.5–4.4 kHz band, and one for the 6.4–8.7 kHz band. In both configurations events were detected when the in-band power exceeded a local estimate of the background noise level by 6 dB for ≥ 50 percent of the spectrogram frames within a range of durations corresponding to the range of durations of recorded MFAS pings.

To search for right whale upcalls, we used an automated detection algorithm that uses a multistage, hypothesis-testing technique based on the generalized likelihood ratio test (GLRT) detector, spectrogram testing, and feature vector testing (Urazghildiiev et al., 2009). In published tests, this algorithm detected approximately 80 percent of upcalls detected by human analysts (Urazghildiiev et al., 2009).

Minke whale pulse trains and sperm whale click trains were detected using new intensity-based image-processing algorithms. For efficiency in processing, the pulse-train (or click-train) detection algorithm operated on successive non-overlapping 60-second portions (henceforth *pages*) of the audio recordings. For each page, a spectrogram was computed using a Blackman window, 1024-point FFT with 90% overlap, yielding spectrogram hop sizes of 51.2 and 12.8 ms for minke whale and sperm whale recordings (at sample rates of 2 and 8 kHz, respectively). Next, the spectrogram was transformed into a binary image in order to perform segmentation and object detection (Pal & Pal, 1993; Thode et al., 2012; Popescu et al., 2013a). The binary image was obtained by replacing all pixels in the input spectrogram with a value of 1 (white) if they had luminance greater than a threshold γ and replacing all other pixels with a value of 0 (black). The threshold γ was selected using gray-level histogram and intra-class variance minimization (Otsu, 1975). Connected region analysis, also known as connected region labeling (Samet & Tamminen, 1988; Weeks, 1996; Thode et al., 2012; Popescu et al., 2013a; Pourhomayoun et al., 2013), was performed on the resulting binary image in order to remove other objects that were not of interest based on shape, area and angle of orientation. This removed "salt and pepper" noise that arose during image segmentation, as well as other frequency modulated or tonal shapes. The detection was then performed using the number of repeating pulses and, in the case of minke whales only, the inter-pulse intervals to determine the presence of a pulse train. This was implemented by transforming the binary image into a projection function (Popescu et al., 2013a) and finding local maxima. For detection of minke whale pulse trains, three decision rules were applied: (1) the number of local maxima must be between specified minimum and maximum acceptable values, (2) the times between consecutive local maxima must be within a prescribed range, and (3) the number of absent consecutive local maxima must be below some maximum threshold. For detection of sperm whale click trains, rules (2) and (3) were not applied. Although foraging clicks from individual sperm whales tend to be given at regular intervals, these recordings commonly included overlapping click trains from two or more individuals, resulting in high variability in inter-click intervals in the recordings. Applying rules (2) and (3) would have resulted in rejection of most overlapping sperm whale click trains. If all rules for the target species were satisfied, the decision process returned the begin and end times of the detected pulse or click train within the 60 s processing page (Popescu et al., 2013a; Popescu et al. in prep). The algorithm was implemented using the software package Matlab 2013b and the DeLMA application (Dugan et al., 2013; Popescu et al., 2013b). Preliminary estimates based on inspection of the current data set suggest that this algorithm typically detects 70 to 90 percent of the events that are detectable to a human analyst, depending on noise conditions.

In the data analyzed here, individual minke whale pulse trains, with typical durations of 30 to 60 seconds, were easily distinguished from each other and rarely overlapped. Individual events found by the detector algorithm thus corresponded to individual minke pulse trains, as confirmed

in the manual review of detection results (see **Section 2.3.2**, below). However, when individual pulse trains spanned the 1-minute processing page boundaries, a single pulse train could yield two detections. Pulse trains that were split across pages in this way were consolidated in a later post-detection processing step, prior to statistical modeling. When sperm whale click trains occurred in these recordings, click trains from a single individual varied greatly in duration (from a few seconds up to several minutes) and click trains from multiple whales commonly overlapped each other. Consequently, each sperm whale detection corresponds not to a discrete acoustic event (as with minke whales), but to a 1-minute interval with sperm whale clicks present.

To detect fin whale 20-Hz notes, a spectrogram cross-correlation-based data template detector implemented in XBAT (Bioacoustics Research Program, 2011) was applied to the acoustic data from all MARUs in all deployments. The detector uses multiple exemplars of 20-Hz fin whale notes and detects sounds for which the peak spectrogram cross-correlation value exceeds a specified threshold. Preliminary estimates based on inspection of other data sets containing fin whale pulses from the western North Atlantic suggest that this algorithm typically detects between 70 and 90 percent of events detectable to a human analyst. The performance of this algorithm on the current data set could not be assessed because no fin whale 20-Hz notes were observed in these recordings.

2.3.1 Sonar Detector Sensitivity

The sensitivity of the automated sonar ping detector relative to a human analyst was estimated by examining ten randomly selected 10-minute sample intervals from each day of recording on which any sonar events were detected (i.e., a total of 100 minutes per day, or 6.9 percent of all data from each day with any sonar detections). For each 10-minute sample, sonar pings that were missed by the detection algorithm were logged manually. The sensitivity of the detector across all days was then estimated as

$$S = \frac{d}{d + m}$$

where *d* is the total number of pings detected by the algorithm and *m* is the total number of pings found by a human analyst that were missed by the detector.

2.3.2 Review of Automated Detector Results

Acoustic events detected by automated algorithms were subsequently reviewed by human analysts to confirm their source and to eliminate false detections. Detected events were reviewed visually (as spectrograms in Raven Pro), and in some cases aurally, and classified as “true” or “false” by experienced acoustic analysts. This review process eliminated false positive events (“false alarms”) where the detector algorithm mistakenly detected some acoustic event which was not a true sound of interest. All events represented in the data summaries here were thus judged to be true with a high degree of confidence.

Due to limited resources, detection results were not reviewed for all sites of all deployments. Potential sonar detections were reviewed for one central MARU site from each deployment (Onslow Bay Site DB2 and Jacksonville Site 05). Limited sampling of multiple sites during

periods of sonar activity indicated that most sonar pings were detectable on all HF MARUs in a deployment. Although LF MARUs in the Jacksonville deployments were not capable of recording sonar pings, we presume that most pings detectable at nearby HF MARUs would also have been detectable at the LF sites had the recorders at those sites been capable of recording high-frequency signals. For potential whale detections, review effort was prioritized to sites within each deployment judged most likely to yield confirmed detections, based on what is presently known about the distribution and ecology of each species. For minke whales in the Jacksonville data, MARU Site 03 was selected for review because a previous analysis of these recordings had indicated that Site 03 (along with Site 01) had the highest rate of occurrence of minke vocalizations (Norris et al., 2012). For sperm whales, detections were reviewed from the same sites reviewed for sonar detections (Onslow Bay Site DB2 and Jacksonville Site 05). Recordings from the Jacksonville deployments were analyzed first and were used to refine procedures and protocols for working with these data.

2.4 Review of Previously Detected Right Whale Gunshot Sounds

In a previous analysis of the recordings from the two Jacksonville deployments, scientists at Bio-Waves, Inc. found 268 acoustic events containing sounds attributed to right whales. Each *event* identified by Bio-Waves consisted of a series of one or more individual right whale sounds, separated from other right whale sounds by > 10 minutes. All but three of these events contained one or more broadband impulsive sounds identified by Bio-Waves analysts as right whale gunshot sounds (Norris et al., 2012). For the present study, these previously identified gunshot events were reviewed independently by two experienced analysts at the Cornell Bioacoustics Research Program. Each analyst examined all putative right whale gunshot events spectrographically and aurally.

Each event was rated as A (strong match), B (moderate match), C (weak match), or X (not a match) indicating how well, in the judgment of the analyst, the sound matched known right whale gunshots based on published descriptions (Parks et al., 2005; Trygonis et al., 2013) and on visual and aural comparison to recordings of gunshots recorded by S. Parks. Criteria for assigning ratings are listed in **Table 4**.

The A, B, or C designation for each event (which may contain multiple individual sounds) was based on the one putative gunshot sound within the event judged to be most likely a true gunshot sound.

Ratings by the two analysts for each putative gunshot were then combined to yield a two-character combined rating for each event (e.g., AA, BC, XX, etc.), where each character represents the independent rating of one analyst.

2.5 Diel Patterns of Acoustic Activity

To assess possible diel patterns in acoustic activity for each whale species, we examined the mean call detection rates for each of the 24 hours in a day. In order to adjust for overall differences in detection rate from day to day, we calculated a mean-adjusted detection rate for

each hour of call-count data by subtracting the mean number of calls detected per hour across all hours for a given day. Hours with call detection rates that were lower or higher than the mean detection rate for their respective days thus had, respectively, negative or positive mean-adjusted detection rates (Stafford et al., 2005). Mean values across all days for the mean-adjusted call detection rates for each of the 24 hours were then plotted.

2.6 Statistical Modeling of Minke and Sperm Whale Detections

For the statistical modeling of minke whale vocalizations we used detections from MARU Site 03 (LF) of the second deployment at the Jacksonville study area. For modeling sperm whale vocalizations we used data from MARU Site 05 (HF) from each of the first and second deployment of the Jacksonville study area and MARU Site DB2 of the Onslow Bay study area. These data were intended to illustrate the analysis methods. Modeling was restricted to a subset of all recording sites because of resource limitations.

The available data on sonar detections included the begin-time of each detection of individual sonar pings detected with the sonar detectors described above in **Section 2.3**. End-times of individual sonar pings could not be determined reliably due to variable durations of reverberation included in individual detections. Hence, duration of discrete sonar pings could not be measured. Each ping detection record included information on the frequency band in which the sonar ping was detected.

The available data on whale detections included nominal begin- and end-times for each detection of whale vocalizations obtained using the minke and sperm whale detectors described above in **Section 2.3**. However, the interpretation of the begin and end times were different for the two species. For minke whales, begin- and end-times of detections referred to the actual begin- and end-times of individual pulse train detection events because individual pulse trains rarely overlapped each other, and duplicate detections of pulse trains that spanned page boundaries were consolidated in a second processing stage after the initial detection, as described in **Section 2.3**. Because the nominal begin and end times reflected actual beginnings and ends of pulse trains exceeding the detection threshold, durations could be calculated for minke whale detection. For sperm whales, click trains were generally longer than the 1-minute pages in which the detector processed the audio data, and pulse trains from multiple animals commonly overlapped each other, making it impossible in many cases to discern the true begin and end times of individual click trains. Hence, unlike the detections of minke whales, sperm whale detections did not refer to discrete click train events from a single animal. Here, the detector searched for five consecutive clicks that exceeded a certain threshold in a given 1-minute interval. The first time this occurred within a 1-minute processing page was recorded as the begin-time for the detection. The end-time referred to the last time a click exceeded the threshold within the one minute page. If clicks were present at the start of a processing page, the nominal start time reported by the detector simply represented the start of the processing page, not the actual start of a series of clicks, which may have been a continuation of click trains that started on a previous page. Similarly, if clicks continued past the end of a processing page, the nominal end time reported reflected the end of the processing page, not the end of a real biological event, which may have continued past the end of the page. Since the nominal start and end of sperm whale detections do not reliably reflect the start and end times of discrete

biological events, durations were not calculated for sperm whale detections. There was no information available on how many clicks occurred between the begin- and end-times (for more details see **Section 2.6.1**).

The research objective for this study was to develop statistical methods for detecting and quantifying potential changes in vocal behavior of baleen and sperm whales in response to sonar. Using the available data from this study we addressed this with two modeling approaches described in the following sections. For simplicity, we refer to the modeling approaches with the terms presence and duration models. The particular questions that may be addressed with these approaches are:

1. Presence models: does the probability of detecting vocalizations change in the presence of sonar?
2. Duration models: given that the animals are vocalizing, does the duration of the individual vocalization detections change in the presence of sonar?

2.6.1 Defining the Data and the Response Variables

The available acoustic data spanned periods with sonar activities, as well as periods with no sonar activities. In order to detect any potential changes in vocal behavior of minke and sperm whales, we had to define a period during which we assumed sonar activity might have an effect on the vocal behavior, as well as a control period to which we could compare the potentially affected vocal behavior. Several factors were considered here: firstly, we use the term *sonar exercise* to refer to a period of time encompassing all consecutive sonar pings with no gap longer than 48 hours for a given deployment. We further assumed that if sonar had an effect on vocalization, this effect would not only be evident at the same time that sonar pings occur (*during* sonar exercises) but also within short breaks of sonar exercises (*between* sonar activities) and/or after sonar ceased (*after* sonar activities). For the latter, we used a 24-hour period after the last ping of each sonar exercise. This represented a compromise between a conservative guess of how long we expected this potential effect to last and trying to avoid introducing additional variability in vocal behavior due to other factors.

To identify potential changes in the vocal behavior of minke or sperm whales in response to sonar, observations from periods *during*, *between*, or *after* sonar exercises had to be compared to observations from a control period without sonar. Again, we used a 24-hour period before the commencement of each sonar exercise as the control period labeled as *before*. We decided against using more than 24 hours for this control period to avoid introducing additional variability in vocal behavior and to keep it balanced with the 24-hour *after* period.

Hence, for each deployment at each study site, we defined a sonar event to include all the sonar pings occurring consecutively with no gap of 48 hours or longer. When a gap was 48 hours or longer, the subsequent sonar pings were attributed to a different exercise. We included data from the first to the last ping of each exercise as well as the 24 hours surrounding each exercise on either side. Note that we only included data from full days of recording (see **Tables 1, 2, and 3**). With respect to the two modeling approaches, we defined two different data sets,

each with a different response variable, one for the presence models and one for the duration models. These are described in the following section.

For the presence models, we created a continuous data set of consecutive 1-minute segments starting 24 hours before the first sonar exercise for each deployment and ending 24 hours after the last sonar exercise was completed. This consecutive set of 1-minute segments was only interrupted in the case that sonar exercises ceased for 48 hours or more. Then we still included the 24 hours after the preceding sonar exercise and the 24 hours before the next exercise, but left out any additional time between. Thus, data on the presence of whale calls during periods that were more than 24 hours before the start or after the end of a sonar exercise were not used in the models. For each of the 1-minute segments we recorded the presence of whale vocalizations as a binary variable (1 for presence, 0 for absence of vocalizations) and used a binomial error structure for the models.

For the duration models we used the detections of the vocalizations themselves. The response variable was the duration of individual vocalizations measured in seconds to one decimal place precision. In modeling duration, we assumed a gamma error structure (i.e., that vocal duration followed a gamma distribution). For sperm whales, this is not strictly correct, since the gamma distribution is unbounded above, but the maximum duration of sperm whale detections was 60 seconds (even though the actual pulse trains may be longer) because of the 1-minute processing pages used by the detector (see **Section 2.3** for details). Due to this artificial constraint, modeling duration of sperm whale vocalizations was not biologically meaningful using the present data. Hence, we refrained from building duration models for sperm whales. For minke whales, there was no such artificial limit to duration of pulse trains because pulse trains that were initially split into multiple detections by the 1-minute detector processing pages were later consolidated into single detection events, as described in **Section 2.3** (see **Figure 6**). Hence the gamma distribution could appropriately be used to model the error structure in duration models of minke whale pulse trains.

2.6.2 Potential Explanatory Covariates

The covariates included in the analyses are listed in **Table 5**. As all available detections of minke whales were from the same location and the same deployment (the second deployment at the Jacksonville study area), covariate *Site* was not considered for this species. For sperm whales on the other hand, this covariate had three levels: *JAX1*, *JAX2* and *OB2*, referring to the first and second deployment at the Jacksonville study area and the deployment at the Onslow Bay study area.

We included three covariates related to time in the analysis: *Julian date* was measured in number of days since the preceding 31 December; *Time* (time of day) was measured as the number of seconds which had passed since midnight that day; *Daynight* was a two-level factor covariate indicating whether a vocalization occurred between sunrise and sunset (*day*) or between sunset and sunrise (*night*). We used local average times for sunrise and sunset where the averages were taken over the recording periods for each deployment. These were 06:10 and 18:15 for JAX1, 07:09 and 17:22 for JAX2 and 05:07 and 19:15 for OB2.

Sonar2 was set up as a covariate for the presence models, with four different levels that indicated whether a 1-minute segment occurred within the 24 hours before the first sonar ping of a sonar exercise (level *before*), during the occurrence of one or more sonar pings (level *during*), between sonar pings (level *between*) or within the 24 hours after the last sonar ping of an exercise (level *after*). The equivalent covariate for the duration models was *Sonar3* where levels *during* and *between* were combined as one level, i.e., *during/between*. This was done because the relative timing of a vocalization and a sonar ping in the recording depends in part on the unknown distances of the sound sources from the recorder (due to the finite speed of sound). To the extent that the exact timing of a sonar ping might influence the behavior of a whale, the time that matters would presumably be the time at which the ping is received by the whale, which is unknown and would be different from the time when it is received by the MARU. For the presence model this represented less of a problem as here we labeled the 1-minute segments with the different levels for *Sonar2*.

Sonarlag was measured as the number of minutes since the occurrence of the last sonar transmission. Values for this covariate could not be observed for the time before the first sonar exercise of each deployment. Hence, this covariate had to be fitted as an interaction term with an indicator variable which switched *Sonarlag* on or off depending on whether values could be observed or not. Equivalently, covariates pertaining to the number of pings and the average ping interval per 30, 60, 120 or 240 minutes preceding a 1-minute segment or vocalization had to be fitted with an indicator variable switching the covariate off outside the respective periods.

Detector3 and *Detector7.5* were two-level factor covariates indicating the presence of sonar pings detected within the respective frequency bands (see **Table** 5) within a 1-minute segment (presence model) or during a vocalization (duration model).

The detection process of MFAS transmissions described in **Section 2.3** does not allow determining end times of sonar transmissions precisely. Hence, no covariate pertaining to the duration of sonar transmissions was included in the analyses.

2.6.3 Modeling Whale Detections Using Generalized Estimating Equations

Generalized estimating equations (GEEs) are an extension of generalized linear models (GLMs) and, similar to GLMs, allow the specification of different distributions for the response variable such as binomial or gamma. We note that this distribution generally refers to the errors after fitting the model, hence is often referred to in the context of error structure. Both GEEs and GLMs use link functions to relate the response to the covariates. These link functions vary between the type of response. For the presence models we used a binomial error structure with a logit-link function while for the duration models we used a gamma distribution with the identity link function. Generally we recommend using the log-link function with a gamma distribution as it ensures that predicted values remain larger than zero (the duration of a vocalization can never be zero or less). However, for the minke whale duration data we used the identity-link as both observed and fitted values were far from zero and the coefficients are easier to interpret.

However, unlike GLMs, the only information used about this distribution is the mean-to-variance ratio. This makes GEEs more robust to mis-specification of the distribution, allows overdispersion to be readily accommodated, and, of particular use in the current study, allows

modeling of various correlation structures between the errors (i.e., residual not explained by the model) in successive observations. Overdispersion occurs when the variance is greater than assumed under the model. For example, for a binomial GLM, the assumptions include a mean-variance relationship of

$$variance = mean(1-p) = np(1-p),$$

where *p* is the probability of success (i.e., detecting a vocalization, in this case) and *n* is the sample size (number of time periods). Also, model errors are assumed to be uncorrelated. We accommodated potential violation of these assumptions by using GEEs as the model-fitting tool. GEEs estimate a dispersion parameter and, therefore, inflate the expected variance in the case that data are overdispersed.

An alternative approach for accommodating correlation between observations is the use of generalized linear mixed models (GLMMs); however GEEs have the advantage that they allow unbiased estimation of regression coefficients despite possible misspecification of the correlation structure (Ghisletta & Spini, 2004). Hence, GEEs are most useful when the interest lies in the relationship between the response and the explanatory variables, as was the case for this study (as opposed to the correlation structure). For this study, we expected correlation in the observations, regardless of the type of response. Also, GEEs estimate the dispersion parameter and therefore accommodate overdispersed data.

Like GLMs, GEEs require specification of a response variable distribution and a link function, although the distribution is only used to specify the mean-variance relationship (e.g., Ghisletta & Spini, 2004). For the presence model we used a binomial response variable distribution and logit-link function; for the duration models we used a gamma response variable distribution and identity link function.

GEEs may be fitted in the statistical software R (R Core Team 2013) using the *geeglm* function of the *geepack* package (Halekoh et al., 2006). As with the *glm* function used for fitting GLMs, smoothing terms can be added using the *bs* function of the splines package. Using splines allowed for more flexibility in the relationship between the response and the explanatory covariate compared to restricting this relationship to be linear (on the scale of the link function). However, we only allowed for some flexibility by using the default settings of the *bs* function, which fits polynomial splines with three degrees of freedom. A polynomial spline can be thought of as a smoothing function for which the number of maxima and minima depends on the specified number of degrees of freedom. Using three degrees of freedom often generates a smoothing function of the form $\beta_1 x_k + \beta_2 x_k^2 + \beta_3 x_k^3$, where x_k represents the *k*th covariate and the β represent the coefficients associated with the polynomial terms. This function often has one maximum and one minimum. More flexibility could be achieved by fitting regression splines and including knots, which allows the relationship to be more 'wiggly' than a polynomial spline (e.g., using the *mgcv* package, Wood, 2013). This, however, was beyond the scope of this study.

We used the default correlation structure 'independence' where correlated observations were grouped into blocks using a block identifier (argument *id* from the *geeglm* function). Each block

consisted of consecutive observations (1-minute segments for the presence models or vocalizations for the duration models) where the size of the blocks was determined using the *acf* function from the *stats* package in R. This function estimates the autocorrelation between consecutive residuals for various lags. A plot created by the same function displays these estimates by lag including 95 percent confidence intervals around zero for comparison with the estimates. The estimated autocorrelation is 1 for lag 0 and, depending on the type (negative or positive) and amount of correlation, generally decays more or less rapidly with increasing lag. We used the lag at which the absolute value of the correlation between residuals first decayed within confidence bounds around zero autocorrelation. For the independence correlation structure, block size has no effect on the parameter estimates; however, for a given model, standard errors and *p*-values associated with the estimates increase with an increase in block size. Larger *p*-values, in turn, influence which covariates should be retained (see next section on model selection for GEEs).

2.6.4 Model Selection for GEEs

Our methods for selecting the final model included three main steps: (1) stepwise forward selection based on marginal *p*-values; (2) elimination of collinear covariates; and (3) stepwise backwards selection by inspecting 95 percent confidence intervals around partial fit plots. These are detailed in the following section.

Model selection for GEEs remains an area of ongoing research, with no clear-cut best strategy, in particular when models include smoothing terms (as in this study). The often-used QIC (quasi-likelihood under the independence model information criterion) is somewhat equivalent to AIC (Akaike Information Criterion) for GLMs or GLMMs (Pan, 2001). The main difference is that QIC uses a quasi-likelihood as opposed to the likelihood used in AIC. Like AIC, QIC only takes into account the coefficients and the relative fit of these coefficients to the data, but not the autocorrelation of the errors. The autocorrelation is, however, reflected in the *p*-values of the coefficient estimates. In this study, because there was an independence correlation structure for the errors, estimates of the coefficient values remained the same for any given model regardless of choice of block size, while *p*-values of the estimated coefficients increased with increasing block sizes. Hence, using QIC for model selection may lead to retaining covariates in the model with relatively large *p*-values. Because of this, we used *p*-value based forward model selection, where we started with a null model (with no covariates) and added one covariate at a time, testing whether it improved the model. For this test, we used the marginal *p*-values associated with an F-test statistic, which tested whether each covariate in the model was important given that the other covariates were already in the model. For this purpose we used the *getPvalues* function from the R package *MRSea* (Scott-Hayward et al., 2013). Continuous covariates were fitted as smoothing terms first. If the smoothing terms were not significant, we tried adding these as linear terms.

In the case that covariates are collinear, it is possible to retain covariates in the model that have otherwise no effect on the response. If more than one covariate was retained in the model, these were tested for collinearity using variance inflation factors. Collinear variables were eliminated by measuring variance inflation factors (VIF) using the *vif* function from the *car* library in R software. We excluded all covariates that scored VIFs > 10 (Fox & Monette, 1992).

A further step for selecting the best model included a potential backwards step. This step consisted of inspecting partial fit plots for each of the covariates retained in the so far best-fitting model. Partial fit plots were created using a modified version of the *runPartialPlots* function from the *MRSea* package (Scott-Hayward et al., 2013), which uses parametric bootstrapping of model coefficients to create confidence intervals around the partial fit. During this step we eliminated covariates which exhibited 95 percent confidence intervals around their partial fits that were wide enough in the vertical dimension to fit a straight horizontal line within the bounds of the confidence limits through the entire range of observed covariate values.

This page intentionally left blank.

3. Results

3.1 Sonar Detector Sensitivity

A total of ten days yielded sufficient data to estimate the sensitivity of the sonar detection algorithm (3, 6, and 1 days of data from the Onslow Bay, JAX 1, and JAX 2 deployments, respectively). In the 1,000 minutes of audio data that were directly examined (10 samples of 10 minutes on each of 10 days), the detection algorithm found a total of 921 pings. An additional 125 pings found by a human analyst were missed by the detector, yielding an overall estimated detector sensitivity of 88 percent (= 921/[921+125]). Detector sensitivities for individual deployments were as follows: Onslow Bay, 88 percent; JAX 1, 87 percent; JAX 2, 100 percent. Only one day of data from JAX 2 yielded any sonar events in the 10 randomly selected minutes used for performance evaluation.

3.2 Onslow Bay (7 – 26 July 2008)

Of the six MARUs that were successfully retrieved from the Onslow Bay deployment, five yielded 20 complete days of recorded data, as expected given the units' hard drive storage capacity. The MARU at site DB1 stopped recording for unknown reasons after two days. Data from that unit are not included in the analyses here.

Potential detections for each of the four target species and for sonar events were each reviewed for only one or two recording sites, due to limited analysis resources. Sites were selected for each species in the depth regime where that species was considered most likely to occur, based on published information about each species' occurrence in different bathymetric zones. Sonar detections were reviewed for site DB2, which had the lowest mean distance to all other recording sites.

Potential right whale detections were reviewed for recordings from site SB7. Only one event (on 16 July) was judged to be a possible right whale upcall. However, based on the co-occurrence of other similar noise events that were judged to be probably non-biological, and on the absence of any other events resembling upcalls nearby in time, the event was not considered a reliable upcall detection.

Potential fin whale detections were reviewed only for site SB3. No fin whale detections were confirmed.

Potential sperm whale detections were reviewed only for site DB2. A published analysis of these same recordings (Hodge et al., 2013) reported that sperm whale clicks occurred frequently only at site DB2.

Potential minke whale detections were reviewed only for sites SB3 (366 m deep) and DB2 (236 m). No minke whale detections were confirmed.

3.2.1 Sonar: Onslow Bay

Figure 7 shows numbers of confirmed sonar events detected for the entire Onslow Bay recording period, in 30-minute bins. Sustained periods of sonar activity occurred on only two days, 16 and 17 July, about halfway through the entire deployment period. A few brief periods of sonar transmission occurred near the end of the recording period, on 24 and 26 July (**Figure 7**).

3.2.2 Sperm Whales: Onslow Bay

Numerous sperm whale click trains occurred on every day of recording, and were limited almost exclusively to nighttime hours (**Figures 8 and 9**). Over all days, 90 percent of sperm whale detections occurred at night. For individual days, the percentage of sperm whale detections that were at night varied between 64.0 percent and 98.6 percent. Overall the pattern of sperm whale click detections was similar to that reported in a previous published analysis of the same recordings (Hodge et al., 2013)

3.3 Jacksonville Deployment 1 (14 September–4 October 2009)

All nine MARUs in the Jacksonville Deployment 1 successfully recorded the expected amount of data. The six HF MARUs each yielded 20 complete days of recording (14 September–4 October), after which recording stopped (as planned) when the internal hard disks were filled. The three LF MARUs, which use up storage space at 1/16 the rate of the HF recorders, continued to record until their retrieval on 8 October. However, recordings from the LF MARUs after 4 October have not been reviewed, since no data are available on occurrence of sonar after the cessation of HF recording.

Sonar and sperm whale detection events were reviewed only for recordings from site 05 (see **Section 2.3.2**).

Potential right whale upcall detections were reviewed for all nine recording sites. Although a total of five isolated events on three different MARUs were identified as being possible upcalls, all were ultimately rejected because of poor signal-to-noise ratio, proximity to similar non-biological sounds, or absence of other likely upcalls nearby in time. Review of potential right whale gunshots is discussed in **Section 3.3.2** below.

There were no confirmed fin whale detections at any recording site in Deployment 1.

Potential minke whale detections were reviewed for all nine sites. There were no confirmed minke detections in Deployment 1.

3.3.1 Sonar: Jacksonville Deployment 1

Figure 10 shows numbers of confirmed sonar events detected for the entire Jacksonville 1 recording period, in 30-minute bins. Sonar activity was detected on eight of the 20 days analyzed, with most transmissions concentrated primarily in a 4-day period (16–19 September), beginning on the third complete day of recording. During these days, there are gaps of 0.5–5.5 h with no detected sonar activity. Shorter periods of lower-level sonar activity occurred during the first two complete days of recording (14–15 September) and on 1 October (the 18th of 20 complete recording days).

3.3.2 Right Whale gunshots: Jacksonville Deployment 1

A total of 167 putative right whale gunshot events were identified by Bio-Waves, Inc. in Deployment 1 (Norris et al., 2012), varying in duration from 0:00:01 to 1:54:24 (h:mm:ss). Each event was independently rated by two Cornell analysts. Results of this process are summarized in **Figure 11**. The two Cornell analysts agreed in their ratings for 113 events (68 percent), and disagreed by one rating step (e.g., AB or BC) for 52 events (31 percent), and by two steps (BX) for 2 events (1 percent).

Fifteen events (9 percent) received either an AA or AB rating, indicating greatest resemblance to known right whale gunshots. Fifty-six events (34 percent) were rejected by both analysts (XX rating) as being gunshots. The remainder of the events received intermediate ratings (**Figure 11**).

3.3.3 Sperm Whales: Jacksonville Deployment 1

Sperm whale click trains occurred on every day of the deployment (**Figure 12**). On most days, almost all sperm whale detections occurred after sunset and before sunrise (**Figures 12 and 13**). However, a few days deviated from this pattern, with high levels of sperm whale activity over many daylight hours (**Figure 12**). Over the entire deployment, 81.7 percent of sperm whale detections were at night. For individual days, the percentage of sperm whale detections that were at night varied between 45.4 percent and 100 percent.

3.4 Jacksonville Deployment 2 (5–25 December 2009)

All nine MARUs in Jacksonville Deployment 2 successfully recorded the expected amount of data. The six HF MARUs each yielded 21 complete days of recording (5–25 December), as expected based on the storage capacity of their hard drives. The three LF MARUs continued to record until their retrieval on 7 January 2010. However, recordings from the LF MARUs after 25 December have not been reviewed, since no data are available on occurrence of sonar after the cessation of HF recording.

Potential sonar and sperm whale detections from Jacksonville Deployment 2 were reviewed only for site 05 (the same site for which they were reviewed for Deployment 1).

Potential right whale upcall detections were reviewed for eight of the nine recording sites; detections from site 1 (one of the deepest sites, hence judged least likely to yield confirmed right whale detections) were not reviewed. Although 11 isolated events on four different MARUs were identified as being possible upcalls, three were ultimately identified as humpback whale sounds, and the remaining eight events were ultimately rejected because of poor signal-to-noise ratio, proximity to similar non-biological sounds, or absence of other likely upcalls nearby in time. Review of potential right whale gunshots is discussed in **Section 3.4.2** below.

There were no confirmed fin whale detections at any recording site in Deployment 2. Sperm whale and minke whale detections are discussed in **Sections 3.4.3 and 3.4.4** below.

3.4.1 Sonar: Jacksonville Deployment 2

Sonar activity was detected on three of the 21 complete recording days (**Figure 14**). On two of those days, only four or fewer 30-minute bins contained sonar detections. The only day on which sonar persisted for longer than four consecutive bins (2 h) was 10 December, when sonar occurred continuously for 12.5 h.

3.4.2 Right Whale Gunshots: Jacksonville Deployment 2

A total of 101 putative right whale events were identified by Bio-Waves, Inc. in Deployment 2, varying in duration from 0:00:01 to 0:46:55 (h:mm:ss). Results of the independent rating process are summarized in **Figure 15**. The two Cornell analysts agreed in their ratings for 68 events (67 percent), and disagreed by one rating step (e.g., AB or BC) for 30 events (30 percent), and by two steps (AC, BX) for 3 events (3 percent).

Five events (5 percent) received either an AA or AB rating, indicating greatest resemblance to known right whale gunshots. Forty-six events were (46 percent) were rejected by both analysts (XX rating) as being gunshots. The remainder of the events received intermediate ratings (**Figure 15**).

3.4.3 Sperm Whales: Jacksonville Deployment 2

Sperm whale click trains were detected on every day of the deployment (**Figure 16**). As in the Onslow Bay and Jacksonville 1 data sets, there was a strong diel pattern to the occurrence of sperm whale click trains (**Figure 16 and Figure** 17), with 98.8% of all detections occurring at night. For individual days, the percentage of sperm whale detections that were at night varied between 86.4 percent and 100 percent.

3.4.4 Minke Whales: Jacksonville Deployment 2

Potential minke whale detections were reviewed for all nine recording sites. The highest numbers of confirmed minke pulse trains were found at the three deepest sites, with 1,241 to 2,859 confirmed detections, before consolidation of duplicate detections caused by pulse trains spanning page boundaries (see **Section 2.3**). The highest number of detections occurred at site 3. The three mid-depth sites each yielded 308 to 497 total detections. Across the three shallow sites, only one minke pulse train detection was confirmed.

Figure 18 shows numbers of confirmed and consolidated minke detections at site 3 (2,351 in total). Minke whale pulse trains were detected on 20 out of 21 days. The maximum number of 30-min bins with minke whale pulse trains in any day was 47. Inspection of **Figure 18** suggests a trend of increasing numbers of pulse trains detected per day over at least the first half of the 21 complete recording days of the Jacksonville Deployment 2, an impression that is supported by the minke whale presence models discussed in **Section 3.5.1.1** below. This trend may reflect a seasonal increase in vocal activity or migratory movement of minke whales into or through the monitoring area (as suggested by Risch et al., 2014a).

Minke whale call detections showed a weak diel pattern, with lower-than-average call rates during nighttime hours (**Figure 19**) and highly variable rates during daylight hours. This pattern

was in contrast to that observed in late summer and fall in waters off Massachusetts, when minke whale acoustic detections were much higher at night than during the day (Risch et al., 2013).

3.5 Statistical Modeling of Minke and Sperm Whale Detections

Using the time periods defined in **Section 2.6.1** we analysed a total of 8,821 1-minute segments for the minke whale presence models and 32,346 1-minute segments for the sperm whale presence models (**Table 6**). For the duration models, 414 detections of minke whales occurred during periods designated as *before*, *during/between*, or *after* sonar, and were thus included in the models. Duration models were not applied to the sperm whale data because the 1-minute page processing of the detection algorithm precluded measuring the actual durations of sperm whale vocalization periods (see **Section 2.6.2** for more details). The large discrepancy in numbers of 1-minute segments and detections for minke whales is due to the fact that for the former, all 1-minute segments in the periods defined in **Section 2.6.1** are included in the count regardless of whether a minke whale was detected.

3.5.1 Presence Models

Block sizes for fitting the presence models using the GEE approach were determined by assessing **Figure** 20 for each of the two whale species. For minke whales we used 2 1-minute segments as the maximum block size while for sperm whales we used 528 1-minute segments (**Table 2**). For a definition of blocks see **Section 2.6.3**, for a definition of 1-minute segments see **Section 2.6.1**.

The best fitting models were determined with the three step model selection described in **Section 2.6.4**. Parameter estimates of the best fitting models are given in

Table 7. Using a logit-link function, the relationship between the coefficients and the response can be interpreted as the following: the expected odds (i.e., probability *p* of observing a presence, divided by the probability of observing an absence, *1-p*) are expressed as the exponent of the predictor η, e.g., $p/(1-p) = exp(\eta) = exp(ß_0 + ß_1x_1 + \ldots + ß_kx_k)$, where the *ß* terms represent the intercept and coefficients associated with the *k* covariates *x*. This equation can be expressed in terms of the probability *p* of observing a presence where $p = exp(\eta) / (1+exp(\eta))$. The model can be used to estimate the expected probability of a call detection under different scenarios encompassed by the model. For example if covariate *Daynight* was retained in the best model we can calculate how we expect the expected odds $p/(1-p)$ of observing a presence or the expected probability *p* of observing a presence to change for night time detections compared to day time detections.

3.5.1.1 PRESENCE MODEL FOR MINKE WHALES

The final presence model for minke whales contained the factor covariate *Sonar2* and the polynomial spline for covariate *Julian date* (

Table 7). The coefficients as well as the partial fit plot for covariate *Sonar2* (**Figure 21**) indicated that the odds of detecting minke whale vocalizations were higher on average in the 24 hours after a sonar exercise compared to the 24 hours before. The partial fit plots for *Julian date* help interpreting the relationship between the response and this covariate (**Figure 21**). Here, we can infer that within the range of observed dates, the odds of observing presences of vocalizations decreased between 7 and 10 December (Julian dates 341 and 344), and increased between 10 and 16 December (Julian dates 344 to 351, **Figure 21**).

3.5.1.2 PRESENCE MODEL FOR SPERM WHALES

The best fitting presence model for sperm whales contained the factor covariate *Daynight* where the coefficient for level *night* was positive and 95 percent confidence intervals did not overlap 0 (

Table 7 **and Figure 22**). This provided evidence that during our study the odds of observing presences of sperm whale vocalizations increased at night compared to during the day, consistent with patterns described in **Sections 3.2.2, 3.3.3, and 3.4.3**.

Also retained in the best fitting presence model for sperm whales was the polynomial spline for *Time* (

Table 7 **and Figure 22**). The partial fit plot for covariate *Time* provided evidence that the odds of observing presences of sperm whale vocalizations were highest early in the morning and late at night and lowest in the early afternoon (**Figure 22**), consistent with patterns in the data plots shown in **Sections 3.2.2, 3.3.3, and 3.4.3**. The 95 percent confidence intervals surrounding the partial fit were narrow enough to support these findings.

None of the covariates related to sonar were included in the best-fitting model (

Table 7), suggesting that sonar activity did not significantly affect the occurrence of sperm whale click trains.

3.5.1.3 ASSESSING ASSUMPTIONS FOR PRESENCE MODELS

The dispersion parameters for the presence models were estimated to be less than one (

Table 7). The standard errors associated with the dispersion parameters were relatively small for both whale species providing no evidence for overdispersion of the data.

We incorporated a blocking structure to the observations where observations within the same block were allowed to be correlated (

Table 7). Taking these block sizes into account, we expected to see no additional pattern in the residuals on scales larger than these blocks. To assess this, we split the Pearson's residuals into 20 equally sized bins in the order of observation and assessed whether these exhibited a random pattern (**Figure 23**). For both the minke whale and sperm whale presence models, we found no change in variability across the range of observed values and considered the pattern of points random. As the minke whale data originated from one MARU deployment, the order of observation corresponded directly to an increase in date. For the sperm whale data, the data were ordered by deployment ID first and then by date.

3.5.1.4 ASSESSING PRESENCE MODEL FIT

We assessed model fit by visually inspecting the binned observed versus fitted values (**Figure 24**). In a well-fitting binomial model the means of the fitted are close to the means of the observed values resulting in a more or less scattered pattern centered around the red lines shown in the plots. As for our models this seemed to be the case, we concluded that the fit of the model was adequate.

An additional method to assess model fit is to compare the predicted presences and absences against the observed presences and absences for each observation. For this purpose, we generated predicted presences using the fitted values of the best model. If, for a given observation, the fitted value was larger than the overall mean of the fitted values, we attributed a presence to the respective record. In the case that the fitted value was smaller than the overall mean of the fitted values, we attributed an absence to the respective record. **Table 8** lists the number of correct predictions as well as the falsely predicted presences and absences. Overall, the presence models for minke and sperm whales predicted 68 percent and 70 percent, respectively, of all observations correctly.

3.5.2 Duration Models

To determine block sizes for GEEs for the duration models, we evaluated the autocorrelation of model residuals shown in **Figure 25**. We used 1 as the maximum block size for the minke whale duration models.

We fitted the minke whale duration models with a gamma error structure and the identity-link function. Here, the relationship between the response and the explanatory covariates can be expressed as: $y = \beta_0 + \beta_1 x_2 + \beta_2 x_2 + \ldots + \beta_k x_k$ where y represents the response and the β_i are the intercept and the coefficients associated with the explanatory covariates x_i. Hence, for factor terms, a positive coefficient indicates an increase in the response compared to the baseline level of the covariate, while a negative coefficient indicates an expected decrease in the response compared to the baseline level of the covariate (the baseline level of a factor covariate is the level that the other levels are compared to, e.g., level *0* (absence of detections) is the baseline level for covariate *Detector3*).

The best fitting duration model for minke whales contained the factor covariate *Detector3* with a negative coefficient for level *1* referring to presence of detection. Hence, when sonar pings were detected in the respective frequency band of this sonar detector, we expected that the duration of minke whale detections decreased.

Also retained in the final minke whale duration model were the polynomial splines for *Julian date* and *Pingnum.240min* (

Table 9). The partial fit plot for *Julian date* indicated that, on average, the duration of detections increased between 10 and 17 December (**Figure 26**). The partial fit plot for *Pingnum.240min* indicated that the duration of detections changed in relation to the number of sonar pings in the four hours preceding the vocalization. This relationship had a positive slope between approximately 40 to 110 detections of sonar pings in the four hours preceding the vocalization and a negative slope for approximately 110 to 155 sonar pings.

3.5.2.1 ASSESSING ASSUMPTIONS FOR DURATION MODELS

For a gamma GLM, the variance is assumed to equal the squared mean. The estimate of the dispersion parameters for the minke whale duration model was less than one with small standard errors (

Table 9). Hence, there was no evidence for overdispersion of the data. After accommodating correlated errors, we expected Pearson's residuals plotted against the fitted values to show a right-skewed random pattern centered around zero. To assess this, we plotted Pearson's residuals in the order of observation which confirmed the randomness of Pearson's residuals (**Figure 27**). The histogram confirms the right-skewed distribution of Pearson's residuals.

3.5.2.2 ASSESSING DURATION MODEL FIT

We assessed the model fit using the concordance correlation using the *runDiagnostics* function of the *MRSea* R package (Scott-Hayward et al. 2013) and the observed versus fitted plot (**Figure 28**). In the plot, the horizontal alignments of most data points were caused by that the predictor included factor covariates (*Sonar3* and *Detector3*) while the polynomial spline for *Pingnum.240min* only affected a small proportion of the predictions (see **Section 2.6.2**). However, in a well-fitting gamma model, the fitted values are close to the observed values resulting in a more or less scattered right-skewed pattern centered on the red line shown in the plots. For the minke whale duration model, the right-skewedness was evident from the wider scatter of points to the right of the red line compared to the left of the red line. On the other hand, the amount of scatter around the line of perfect fit was quite large as a lot of large observed values were under-predicted and a large number of small observed values were over-predicted. In addition, the concordance correlation was low. This measure always ranges between 0 (poor fit) and 1 (perfect fit). For the minke whale duration model it was 0.11. Hence we conclude that model fit was relatively poor. On the other hand, no additional pattern was evident in the residual plot (**Figure 27**) or the observed vs fitted plot (**Figure 28**) that could be captured by a missing covariate.

4. Discussion

Sperm whale click trains were detected on every day of recordings from all three deployments: Onslow Bay in the summer (July 2008), and Jacksonville in the fall (September–October 2009) and winter (December 2009). Minke whale pulse trains were detected only in the winter Jacksonville deployment. There were no confirmed detections of sounds of North Atlantic right whale upcalls or fin whales in any of the three deployments. Putative right whale gunshot sounds are discussed in a separate section below.

In all three deployments, sperm whale click trains were detected almost exclusively at night, with the exception of a few days during the fall Jacksonville deployment. Minke whale pulse trains occurred at variable rates at all hours of the day during the Jacksonville winter deployment, with a weak trend to slightly lower levels of activity at night. Over the course of the 21 full days of recording, there was an apparent trend of increasing numbers of minke detections per day. This was confirmed by the statistical analyses using GEEs where *Julian date* was one of the two covariates retained in the best fitting presence model for minke whales. Here, we found evidence for a decrease in the odds of detecting minke whale vocalizations between 7 and 10 December and an increase between 10 and 16 December 2009 (**Figure 21**).

In each of the three deployments, sonar activity was concentrated in one to five consecutive days, during which transmissions were detected at rates up to 307 pings per 30 minutes (at Onslow Bay; maximum ping rates in the Jacksonville fall and winter deployments were 137 and 107 pings per 30 minutes, respectively). Recordings from each deployment period also included one to several short (< 6 h) periods of sonar activity separated from other periods by more than two hours.

4.1 Review of Putative Right Whale Gunshot Detections

Of the 267 events identified by analysts at Bio-Waves, Inc. as right whale gunshots in the two Jacksonville deployments, 20 (8 percent) were rated AA or AB by Cornell analysts, indicating the greatest judged resemblance to published descriptions and recordings of known gunshots. Overall, there was a high level of agreement in ratings between the two independent Cornell analysts, with exact agreement in 68% of cases or close agreement (i.e., AB, BC, or CX paired ratings) in an additional 31% of cases. More than one third (102 events, 38 percent) were judged by both Cornell analysts not to be gunshots (i.e., XX rating), in most cases because an alternative source was considered more likely, based on spectrographic and/or aural characteristics.

Based on examination of the individual putative right whale vocalizations identified in the earlier analysis of these recordings (Norris et al., 2012), we do not find compelling evidence in the acoustic data to conclude that right whales were vocalizing in the area during either Jacksonville deployment.

The most notable aspect of these results is the disparity between the initial labeling of events by Bio-Waves analysts and the ratings subsequently assigned by Cornell analysts. Several factors contribute to this lack of consistency:

- **Reliance on expert judgment:** Currently, recognition of right whale gunshot sounds depends on human expert judgment of sound similarity to published descriptions of known gunshots (Parks et al., 2005; Parks et al., 2011; Parks et al., 2012; Trygonis et al., 2013; Matthews et al., 2014). Although some progress has been made recently on quantitative methods for recognizing gunshots (Trygonis et al. 2013; Binder & Hines, 2014), more work is needed on developing and validating automated methods that can reliably discriminate gunshots from other impulsive sounds under a broad range of sound transmission and noise conditions.
- **Diversity of sources of impulsive sounds:** It is not uncommon to find occasional broadband impulsive sounds in MARU recordings, even from times and places where gunshot sounds from right whales can confidently be ruled out. Potential sources for such sounds include impacts of drifting debris against the MARU housing or hydrophone, small marine animals (e.g., fish) that are temporarily trapped in spaces inside the plastic "hard hat" that encloses the MARU's glass housing, sperm whale clicks, and anthropogenic sources such as underwater explosions. Often the sources of specific impulsive sounds cannot be determined or can only be presumed from context. Variation among analysts in experience reviewing recordings from diverse times and places may result in differing judgments about the likely or possible source of a given sound.
- **Importance of aural cues:** Spectrograms often do not contain sufficient information to allow even highly experienced analysts to discriminate visually between impulsive sounds that are easily distinguishable by auditory cues. This observation is consistent with reports that military sonar operators can sometimes discriminate targets from clutter more reliably by listening to echo returns than by observing visual sonar displays (Allen et al., 2011). In the present study, Cornell analysts used both aural and visual cues to assign ratings to putative gunshot sounds. It is unclear whether analysts at Bio-Waves relied exclusively on visual cues in making their determinations (Norris et al., 2012); if so, this difference in methodology may contribute to the discrepancy between the results of the two analysis teams.

4.2 Approaches to Assessing Potential Effects of Sonar on Whales

Studies of potential effects of sonar on the behavior of cetaceans have taken the general approach of comparing one or more behavioral metrics during periods immediately before, during, and after periods of real or simulated sonar transmissions in the vicinity of the animals. Metrics have included presence in an area, assessed by visual and/or passive acoustic monitoring (Kuningas et al., 2013); spatial distribution relative to sound sources, assessed by acoustic localization (Tyack et al., 2011; McCarthy et al., 2011; Moretti et al., 2014); movement patterns, including dive profiles, of tagged individuals (Tyack et al., 2011; Sivle et al., 2012; Miller et al., 2012; DeRuiter et al., 2013; Goldbogen et al., 2013); and vocal behavior of tagged individuals (Tyack et al., 2011; Miller et al., 2012) and local populations (Tyack et al., 2011; McCarthy et al., 2011; Melcón et al., 2012; Moretti et al., 2014).

Perhaps the most important distinction among the studies that have been done thus far is between *observational studies* that have relied on data collected during actual military training operations and *controlled-exposure experiments*. In observational studies, the timing and source levels of sonar transmissions, and the positions of the acoustic sources are all dictated by military training considerations, without regard to studying the responses of marine mammals. In controlled-exposure experiments, these parameters are all under the control of researchers, and can be manipulated to optimize the utility of the resulting data in drawing inferences about behavioral responses to sonar exposure.

In the present study, which used an observational approach during actual sonar training events, the temporal distribution of sonar activity was a matter of chance with respect to the presence and behavior of the target whale species. As a result, the amount of data usable for assessing potential responses of sperm and minke whales to sonar exposure was severely limited by two issues. First, the number of discrete periods with sonar exposure where whale responses could be assessed using a before-during-after design was small and some of the exposure periods included very few pings (see **Figures 7, 10,** and **14**). Second, in the case of minke whales in the JAX 2 deployment the level of whale vocal activity was already extremely low and declining during the 24 hours before sonar transmissions commenced (**Figure 18; Figure 21** lower panel). If whales tend to greatly reduce or cease their vocalizations (as observed, for example with beaked whales at AUTEC, Tyack et al., 2011; McCarthy et al., 2011) such a change can only be detected if the whales are vocally active immediately before the start of sonar transmissions.

4.3 Modeling Approaches Using GEEs

4.3.1 Pros and Cons of the Two Modeling Strategies Using GEEs

When assessing a potential effect of sonar on the vocalization behavior of minke and sperm whales, modeling the presence of vocalizations has the advantage over modeling their duration, in that the former takes into account the amount of time during which no vocalizations were detected. For the duration models, the time periods with no vocalizations do not contribute any information to the model. For sperm whales the duration models were not appropriate to fit as here the duration was limited by an artificial limit imposed by the detection process where 1-minute pages were searched at a time.

A difficulty for fitting the duration models for minke whales was the paucity of observations that were most likely to be affected by sonar activities, i.e., from those periods defined as during or between sonar in **Section 2.6.1**. For minke whales, only 4.4 percent of all detections included in the analyses came from these combined periods (**Figure 18**; note, however, that only minke whale detections that occurred within 24 hours before and after sonar activities were included in the analyses).

With respect to model selection we found that results would have been somewhat ambiguous if only the first two steps from our model selection procedure were included. The first step was forwards selection based on p-values obtained from an F-test statistic (adjusted for overdispersion and correlated data; see **Section 2.6.4**) while the second eliminated collinear

covariates from the model. Step three was a backwards selection step based on inspecting confidence limits of partial fit plots and eliminating those covariates for which no effect of the respective covariate was a plausible outcome. No effect would be case if the relationship between the respective covariate and the response was a constant across the range of observed values of the covariate.

We do emphasize, however, the importance of accounting for correlation and overdispersion in the data as well as assessing collinearity in covariates retained in the model. Accounting for these issues reduces the risk of falsely retaining covariates in the final model and falsely inferring an effect of these. We encountered these potential issues in particular when fitting the sperm whale presence models. Here, due to the strong diel pattern of vocal activity at night only, correlation for model residuals was high and block sizes needed to be large to account for this. Ignoring correlation of model residuals may lead to falsely retaining unimportant covariates in the final model and false inference.

4.3.2 Possible Inference from the Models Fitted with GEEs

We restricted the data we included in the statistical modeling to the 24 hours before and after each sonar exercise as well as the times we defined as during and between sonar pings in **Section 2.6.2**. We believe this represented a conservative compromise between capturing a potential lag in the effect after sonar activities and introducing additional variability in the response (presence or duration of vocalizations). This additional variability may be caused by other factors, such as time or prey availability, which we might not be able to capture with other available covariates. The 24-hour periods are somewhat arbitrary in the sense that they are an a priori estimate of how long potential effects may last. However, for this study we were more interested in whether there is an effect of sonar on vocal behavior rather than examining how long this effect would last. There is evidence that for some whales it may last longer (e.g., McCarthy et al., 2011). Hence, excluding data beyond 24 hours *after* prevented us from deciding whether to increase our periods which we labeled *after* or whether to assume that the effect was no longer present and label these periods as *before*. False decisions for this issue have the potential of diluting the evidence of the effect.

4.3.2.1 INFERENCES FROM PRESENCE MODELS

In the previous section we described the pros and cons of the two modeling approaches for assessing quantitative changes in vocalizations of minke and sperm whales in the presence of sonar. However, one has to keep in mind what kind of inference can be drawn from these models, and the data going into them. The presence models do not explain variability in the proportion of time that animals were vocalizing. They only describe changes in the probability of detecting vocalizing animals. If, for example, covariate *Sonar2* was retained in the best fitting presence model and level *during* was significantly higher compared to the base level *before*, we could only infer that *during* sonar, the proportion of time we detected vocalizations was higher than *before* sonar. We could not directly infer that animals spent a larger proportion of time vocalizing. For the latter, we would need to make the implicit assumption that by looking at the probability of detecting vocalizations on a MARU, we are examining the probability of animals calling. But this would not be appropriate. Alternative explanations could be that, while animals spent the same proportion of time vocalizing, animal density changed, animals redistributed

themselves, altered the source levels of their vocalizations, or changed their orientation relative to the MARU (assuming some directionality to their sounds, e.g., Møhl et al., 2000, Blackwell et al., 2012). As the probability of detecting vocalizations is dependent on received level, these other changes would also have an effect on the proportion of time vocalizing. It is also possible (though unlikely, in our opinion) that background noise (e.g. from fish or invertebrates) diminished in response to sonar, which would have increased the detectability of minke pulse trains. All of these possibilities would result in fewer detections of vocalizations.

4.3.2.2 INFERENCES FROM DURATION MODELS

Changes in the duration of minke whale call detections could be caused by the whales changing the actual duration of their pulse trains, or by animals being further away from the recording device, or by changes in the source levels of their vocalizations during or after sonar. Recorded vocalizations need to exceed a certain signal-to-noise ratio (SNR) in order to be detected. Begin- and end-times of when the signal exceeds this threshold are logged by the detector. Minke whale pulse trains start with relatively low-amplitude pulses, and then gradually increase in amplitude (Risch et al., 2014b). The duration of a pulse train detection event would therefore tend to diminish with increasing distance because more of the early part of the pulse train would fall below the SNR threshold for detection. Hence, shorter duration of detected events could reflect a change in distance rather than an actual change in vocal behavior. Similarly, a reduction in overall source level could reduce the duration of detection events, independent of the actual pulse train duration or the animals' distance from the recorder. Measurement of received levels (RLs, which was outside the scope of this study) could shed light on whether shorter detections were the result of reduced RL at the MARU.

4.3.2.3 SPERM WHALE PRESENCE MODEL

The best fitting presence model for sperm whales contained the factor covariate *Daynight* and the polynomial spline for *Time* providing evidence that during our study the odds of detecting presences of sperm whale vocalizations varied in a diurnal pattern, increasing at night compared to during the day. None of the covariates related to sonar were included in the best-fitting model, suggesting that sonar activity did not significantly affect the occurrence of sperm whale click trains.

We encountered highly correlated data for the 1-minute presence data of sperm whales. This was due in part to the strong diel pattern of vocalization, in which sperm whale click trains occurred nearly continuously during hours of darkness on most nights, and infrequently during daylight hours on most days (**Figures 8, 12, and 16**). Also, sperm whale click trains are often longer than 1 minute due to their long dives and click trains from multiple individuals often overlapped, leading to periods of continuous clicking that may be many minutes long. Because the detector processes 1-minute pages, this can cause a single continuous clicking period (either from one or multiple animals) to yield a succession of positive 1-minute detections. Hence, observing 1 minute with a presence of click trains will likely result in several consecutive presences of click trains. Furthermore, we refrained from fitting duration models for sperm whales as the duration of the individual vocalization events was artificially limited to 60 seconds due to the detection process. We did not consider these artificially shortened durations

biologically meaningful. It would be more biologically meaningful to model the duration of click trains without any artificial bound. Then, a gamma distribution would have been appropriate.

4.3.2.4 MINKE WHALE PRESENCE MODEL

For the minke whale presence model, covariate *Sonar2* (indicating whether a given minute was *before*, *during*, *between*, or *after* sonar transmissions) was retained in the final model. For minke whales, the odds of detecting vocalizations were on average higher in the 24 hours after a sonar event compared to the 24 hours before the event. This factor covariate was used to contrast potential differences in presences of vocalizations in the four defined periods related to sonar exercises. This is different from covariate *Julian Date* which was used to capture potential changes in presences of vocalizations throughout time unrelated to sonar exercises.

It is likely that inference on the covariate *Sonar2* would have been different for either minke whales or sperm whales if we had applied different criteria for labelling time periods as *before*, *during*, *between*, or *after* (see **Section 2.6.1** for a detailed explanation of the strategy used). Had we chosen a different length of control periods, say 12 hours for the *before* and *after* levels, rather than 24 hours, this would not only reduce the amount of data included but also change the labelling for some of the 1-minute segments still retained. For example, with the current labelling using the 24 hour periods, all 1-minute segments from the JAX2 deployment between 18:09 on 8 December and 08:04 on 10 December were labelled as *between* sonar due to the occurrence of sonar on 8 and 10 December (see **Figure 18**). Had we chosen 12 hour periods instead, the 1-minute segments between 18:09 on 8 December and 06:09 on 9 December would have been labelled as *after* and the 1-minute segments between 20:04 on 9 December and 08:04 on 10 December would have been labelled *before*. Also, the 1-minute segments between 06:09 and 20:04 on 9 December would have been excluded from the analyses. In this case, there would have been no minke detections at all during the 12-hour period *before* the sustained high levels of sonar activity on 10 December. Hence, it is likely that inference on this covariate would have changed.

4.3.2.5 MINKE WHALE DURATION MODEL

We identified differences in the duration of call detections in response to sonar activities for the minke whale duration of vocalization model. The differences consisted of an expected increase in duration if approximately 40 to 110 sonar pings were detected in the four hours preceding the vocalization and a decrease in duration if approximately 110 to 155 sonar pings were detected in the four hours preceding the vocalization (captured by covariate *Pingnum.240min,* **Figure 26**). Although the signal in the data was not very strong, we may conclude that our data provided some evidence that sonar had an effect on the detected duration of minke whale vocalizations during this study. The biological cause or significance of the response illustrated in **Figure 26** is unclear. However, the sample size of discrete periods with sonar activity was very low – sonar transmissions were only detected on three days during the JAX2 deployment (**Figure 18**). Larger sample sizes are needed for stronger inference. Alternatively, controlled exposure experiments may allow a wider inference on the vocal responses of the animals to sonar signals.

4.4 Conclusions and Recommendations for Future Work

4.4.1 Use of Gunshot Sounds to Detect Right Whale Presence

Upcalls have long been the signal of choice for detecting the presence of north Atlantic right whales via passive acoustic monitoring because they are produced by all age-sex classes, are highly stereotyped, and are dissimilar from other commonly encountered sounds in the ocean (with the exception of some humpback whale sounds; Urazghildiiev & Clark, 2007; Urazghildiiev et al., 2009; Van Parijs et al., 2009; Clark et al., 2010; Morano et al., 2012a). More recently, gunshot sounds have been proposed as a useful additional means of detecting right whale presence (Van Parijs et al., 2009; Parks et al., 2011; Matthews et al., 2014). The discrepancies reported here between the judgments of two different analysis teams regarding identification of impulsive sounds as gunshots raises concerns about the use of such sounds to diagnose right whale presence in the absence of other supporting evidence such as visual sightings, presence of upcalls, or seasonal patterns of occurrence consistent with right whale biology (e.g., Parks & Tyack, 2005; Trygonis et al., 2013; Matthews et al., 2014).

Most studies of right whale gunshots have been based on close-range recordings of right whales subject to direct visual observation, or via attached acoustic recording tags (Parks et al., 2005; Parks et al., 2011; Parks et al., 2012; Trygonis et al., 2013). In order to support future use of gunshots as diagnostic indicators of right whale presence, further research should focus on acoustic properties of gunshots recorded at much greater distances, as would be expected in passive acoustic monitoring studies. How do the acoustic properties of these signals change with propagation distance, and in particular, how can distant gunshots be reliably distinguished from other impulsive sounds? Future work should also focus on improving guidelines for human analysts and algorithms to support such discrimination under the types of noise and clutter conditions typical of passive acoustic monitoring recordings.

4.4.2 Further Analysis of Existing Recordings

For the minke whale pulse trains recorded in Jacksonville Deployment 2, measurement of received levels (which was beyond the scope of the present study) could be used to further investigate possible responses of these whales to MFAS. Using published data on estimated source levels of minke pulse trains (Risch et al., 2014b) in conjunction with sound propagation models, it may be possible to estimate the distances of calling minke whales from the MARU, and use these distances as an additional response variable for models of minke whale sounds before, during, and after periods of sonar activity. Inference on call abundance may also be possible using these types of methods (e.g., Harris, 2012).

For sperm whales, we did not apply duration models because we could not reliably determine durations of pulse trains from individual whales (as explained in **Section 2.6**). However, an alternative approach would be to compute durations of periods of aggregate clicking activity of all whales that are close enough to the MARU for their clicks to be detected. These aggregate click durations could be computed from the existing detection data, and then duration models could be applied to investigate whether MFAS affected aggregate clicking behavior. Changes in aggregate click train durations could reflect changes in either the duration of individual click

trains, the durations of non-clicking periods between click trains for individuals, the number of clicking animals within detection range, or the orientation (Møhl et al., 2000) of the animals relative to the MARU.

5. Literature Cited

Allen, N., Hines, P. C., & Young, V. W. (2011). Performances of human listeners and an automatic aural classifier in discriminating between sonar target echoes and clutter. *The Journal of the Acoustical Society of America*, *130*, 1287–98.

Backus, R. H., & Schevill, W. E. (1967). Physeter clicks. In K. S. Norris (Ed.), *Whales, Dolphins, and Porpoises* (pp. 510–527). Berkeley, CA, USA: University of California Press.

Binder, C. M., & Hines, P. C. (2014). Automated aural classification used for inter-species discrimination of cetaceans. *The Journal of the Acoustical Society of America*, *135*, 2113–2125.

Bioacoustics Research Program. (2014). Raven Pro: Interactive Sound Analysis Software (Version 1.5) [Computer software].

Bioacoustics Research Program. (2011). XBAT: eXtensible BioAcoustic Tool.

Blackwell, S. B., McDonald, T. L., Kim, K. H., Aerts, L. A. M., Richardson, W. J., et al. (2012). Directionality of bowhead whale calls measured with multiple sensors. *Marine Mammal Science*, *28*, 200–212.

Clark, C. W. (1983). Acoustic Communication and Behavior of the Southern Right Whale (Eubalaena australis). In R. S. Payne (Ed.), *Communication and Behavior of Whales* (pp. 163–198). AAAS.

Clark, C. W., Borsani, J. F., & Notarbartolo-di-Sciara, G. (2002). Vocal activity of fin whales, Balaenoptera physalus, in the Ligurian Sea. *Marine Mammal Science*, *18*, 286–295.

Clark, C. W., Brown, M. W., & Corkeron, P. (2010). Visual and acoustic surveys for North Atlantic right whales, Eubalaena glacialis, in Cape Cod Bay, Massachusetts, 2001-2005: Management implications. *Marine Mammal Science*, *26*, 837–854.

Croll, D. A., Clark, C. W., Acevedo, A., Tershy, B., Flores, S., et al. (2002). Bioacoustics: Only male fin whales sing loud songs. *Nature*, *417*, 809.

D'Amico, A., Gisiner, R. C., Ketten, D. R., Hammock, J. A., Johnson, C., et al. (2009). Beaked whale strandings and naval exercises. *Aquatic Mammals*, *35*, 452–472.

D'Amico, A., & Pittenger, R. (2009). A Brief History of Active Sonar. *Aquatic Mammals*, *35*, 426–434.

DeRuiter, S. L., Southall, B. L., Calambokidis, J., Zimmer, W. M. X., Sadykova, D., et al. (2013). First direct measurements of behavioural responses by Cuvier's beaked whales to mid-frequency active sonar. *Biology Letters*.

Dugan, P. J., Popescu, M., Risch, D., Zollweg, J. A., Mijolajczyk, A., et al. (2013). Computing at scale: A new client server based system for high performance computing for automatic

signal recognition. In *International Workshop on Detection, Classification, Localization and Density Estimation of Marine Mammals using Passive Acoustics* (p. 6: 39.).

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Hournal of the American Staistical Association*, *87*, 178–183.

Ghisletta, P., & Spini, D. (2004). An introduction to generalised estimating equations. *Journal of Educational and Behavioral Studies*, *29(4)*, 421–437.

Goldbogen, J. A., Southall, B. L., DeRuiter, S. L., Calambokidis, J., Friedlaender, A. S., et al. (2013). Blue whales respond to simulated mid-frequency military sonar. *Proceedings of the Royal Society B-Biological Sciences*, *280*.

Goold, J. C., & Jones, S. E. (1995). Time and frequency domain characteristics of sperm whale clicks. *J*, *98*, 1279–1291.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The {R} {P}ackage geepack for {G}eneralized {E}stimating {E}quations. *Journal of Statistical Software*, *15 (2)*, 1–11.

Harris, D. (2012). Estimating whale abundance using sparse hydrophone arrays.

Hodge, L. E. W., Bell, J. T., Kumar, A., & Read, A. J. (2013). The influence of habitat and time of day on the occurrence of odontocete vocalizations in Onslow Bay, North Carolina. *Marine Mammal Science*, *29*, E411–E427.

Kuningas, S., Kvadsheim, P. H., Lam, F.-P. A., & Miller, P. J. O. (2013). Killer whale presence in relation to naval sonar activity and prey abundance in northern Norway. *ICES Journal of Marine Science*, *70*, 1287–1293.

Matthews, L. P., McCordic, J. a., & Parks, S. E. (2014). Remote Acoustic Monitoring of North Atlantic Right Whales (Eubalaena glacialis) Reveals Seasonal and Diel Variations in Acoustic Behavior (M. L. Fine, Ed.). *PLoS ONE*, *9*, e91367.

McCarthy, E., Moretti, D., Thomas, L., DiMarzio, N., Morrissey, R., et al. (2011). Changes in spatial and temporal distribution and vocal behavior of Blainville's beaked whales (Mesoplodon densirostris) during multiship exercises with mid-frequency sonar. *Marine Mammal Science*, *27*, E206–E226.

Melcón, M. L., Cummins, A. J., Kerosky, S. M., Roche, L. K., Wiggins, S. M., et al. (2012). Blue whales respond to anthropogenic noise. *PLoS ONE*, *7*, e32681.

Mellinger, D. K., Carson, C. D., & Clark, C. W. (2000). Characteristics of minke whale (Balaenoptera acutorostrata) pulse trains recorded near Puerto Rico. *Marine Mammal Science*, *16*, 739–756.

Mellinger, D. K., Nieukirk, S. L., Matsumoto, H., Heimlich, S. L., Dziak, R. P., et al. (2007). Seasonal occurrence of North Atlantic right whale (Eubalaena glacialis) vocalizations at two sites on the Scotian Shelf. *Marine Mammal Science*, *23*, 856–867.

Miller, P. J. O., Kvadsheim, P. H., Lam, F. P. A., Wensveen, P. J., Antunes, R., et al. (2012). The Severity of Behavioral Changes Observed During Experimental Exposures of Killer (Orcinus orca), Long-Finned Pilot (Globicephala melas), and Sperm (Physeter macrocephalus) Whales to Naval Sonar. *Aquatic Mammals*, *38*, 362–401.

Møhl, B., Wahlberg, M., Madsen, P. T., Heerfordt, A., & Lund, A. (2003). The monopulsed nature of sperm whale clicks. *The Journal of the Acoustical Society of America*, *114*, 1143.

Møhl, B., Wahlberg, M., Madsen, P. T., Miller, L. a, & Surlykke, A. (2000). Sperm whale clicks: directionality and source level revisited. *Journal of the Acoustical Society of America*, *107*, 638–48.

Morano, J. L., Rice, A. N., Tielens, J. T., Estabrook, B. J., Murray, A., et al. (2012a). Acoustically detected year-round presence of right whales in an urbanized migration corridor. *Conservation Biology*, *26*, 698–707.

Morano, J. L., Salisbury, D. P., Rice, A. N., Conklin, K. L., Falk, K. L., et al. (2012b). Seasonal and geographical patterns of fin whale song in the western North Atlantic Ocean. *The Journal of the Acoustical Society of America*, *132*, 1207–12.

Moretti, D., Thomas, L., Marques, T., Harwood, J., Dilley, A., et al. (2014). A Risk Function for Behavioral Disruption of Blainville's Beaked Whales (Mesoplodon densirostris) from Mid-Frequency Active Sonar. *PloS one*, *9*, e85064.

Norris, T. F., Oswald, J. N., Yack, T. N., & Ferguson, E. L. (2012). *An Analysis of Marine Acoustic Recording Unit (MARU) Data Collected off Jacksonville , Florida in Fall 2009 and Winter 2009-2010*.

Oswald, J. N., Oedekoven, C. S., Yack, T. N., Thomas, L., & Ferguson, E. L. (2014). *Development of statistical methods for examining relationships between odontocete vocal behavior and navy sonar signals: Preliminary report. Submitted to HDR Environmental, Operations and Construction, Inc. Norfolk, Virginia, under contract no. CON-005-43*.

Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, *11*, 23–27.

Pal, N. R., & Pal, S. J. (1993). A review on image segmentation techniques. *Pattern Recogn.*, *26*, 1277–1294.

Pan, W. (2001). Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics*, *57(1)*, 120 – 125.

Van Parijs, S., Clark, C., Sousa-Lima, R., Parks, S., Rankin, S., et al. (2009). Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales. *Marine Ecology Progress Series*, *395*, 21–36.

Parks, S. E., Hamilton, P. K., Kraus, S. D., & Tyack, P. L. (2005). The gunshot sound produced by male North Atlantic right whales (Eubalaena glacialis) and its potential function in reproductive advertisement. *Marine Mammal Science*, *21*, 458–475.

Parks, S. E., Hotchkin, C. F., Cortopassi, K. A., & Clark, C. W. (2012). Characteristics of gunshot sound displays by North Atlantic right whales in the Bay of Fundy. *Journal of the Acoustical Society of America*, *131*, 3173–3179.

Parks, S. E., Searby, A., Célérier, A., Johnson, M. P., Nowacek, D. P., et al. (2011). Sound production behavior of individual North Atlantic right whales: implications for passive acoustic monitoring. *Endangered Species Research*, *15*, 63–76.

Parks, S. E., & Tyack, P. L. (2005). Sound production by North Atlantic right whales (Eubalaena glacialis) in surface active groups. *Journal of the Acoustical Society of America*, *117*, 3297.

Popescu, M., Dugan, P. J., Pourhomayoun, M., Risch, D., Lewis III, H. W., et al. (2013a). Bioacoustical periodic pulse train signal detection and classification using spectrogram intensity binarization and energy projection. In *ICML 2013 Workshop on Machine Learning for Bioacoustics* Atlanta, GA.

Popescu, C. M., Dugan, P. J., Zollweg, J. A., Mijolajczyk, A., & Clark, C. W. (2013b). Large scale detection classification: case study of four examples using an applied high performance distributed computing platform. In *International Workshop on Detection, Classification, Localization, and Density Estimation (DCLDE) of Marine Mammals using Passive Acoustic* (p. 6: 96).

Pourhomayoun, M., Dugan, P., Popescu, M., & Clark, C. (2013). Bioacoustic signal classification based on continuous region processing, grid masking and artificial neural network.

Risch, D., Castellote, M., Clark, C. W., Davis, G. E., Dugan, P. J., et al. (2014a). Seasonal migrations of North Atlantic minke whales : novel insights from large-scale passive acoustic monitoring networks. 1–17.

Risch, D., Clark, C., Dugan, P., Popescu, M., Siebert, U., et al. (2013). Minke whale acoustic behavior and multi-year seasonal and diel vocalization patterns in Massachusetts Bay, USA. *Marine Ecology Progress Series*, *489*, 279–295.

Risch, D., Siebert, U., & Parijs, S. M. Van. (2014b). Individual calling behaviour and movements of North Atlantic minke whales (Balaenoptera acutorostrata). *Behaviour*.

Samet, H., & Tamminen, M. (1988). Efficient component labeling of images of arbitrary dimension represented by linear bintrees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10*, 579–586.

Scott-Hayward, L. A. S., Oedekoven, C. S., Mackenzie, M. L., & Rexstad, E. (2013). *MRSea package (version 0.0.1). Tech report.*

Sirović, A., Hildebrand, J. a, & Wiggins, S. M. (2007). Blue and fin whale call source levels and propagation range in the Southern Ocean. *The Journal of the Acoustical Society of America*, *122*, 1208–15.

Sivle, L. D., Kvadsheim, P. H., Fahlman, A., Lam, F. P. a, Tyack, P. L., et al. (2012). Changes in dive behavior during naval sonar exposure in killer whales, long-finned pilot whales, and sperm whales. *Frontiers in physiology*, *3*, 400.

Stafford, K., Moore, S., & Fox, C. (2005). Diel variation in blue whale calls recorded in the eastern tropical Pacific. *Animal Behaviour*, *69*, 951–958.

Thode, A. M., Kim, K. H., Blackwell, S. B., Greene, C. R., Nations, C. S., et al. (2012). Automated detection and localization of bowhead whale sounds in the presence of seismic airgun surveys. *Journal of the Acoustical Society of America*, *131*, 3726–3747.

Trygonis, V., Gerstein, E., Moir, J., & McCulloch, S. (2013). Vocalization characteristics of North Atlantic right whale surface active groups in the calving habitat, southeastern United States. *The Journal of the Acoustical Society of America*, *134*, 4518.

Tyack, P. L., Zimmer, W. M. X., Moretti, D., Southall, B. L., Claridge, D. E., et al. (2011). Beaked whales respond to simulated and actual Navy sonar. *PLoS ONE*, *6*.

Urazghildiiev, I. R., & Clark, C. W. (2007). Acoustic detection of North Atlantic right whale contact calls using spectrogram-based statistics. *Journal of the Acoustical Society of America*, *122*, 769–776.

Urazghildiiev, I. R., Clark, C. W., Krein, T. P., & Parks, S. E. (2009). Detection and recognition of North Atlantic Right Whale contact calls in the presence of ambient noise. *IEEE Journal of Oceanic Engineering*, *34*, 358–368.

Watkins, W. A., & Schevill, E. (1977). Sperm whale codas. *Journal of the Acoustical Society of America*, *62*, 1485–1490.

Watkins, W. A., Tyack, P., Moore, K. E., & Bird, J. E. (1987). The 20-Hz signals of finback whales (Balaenoptera physalus). *Journal of the Acoustical Society of America*, 1901–1912.

Watwood, S. L., Miller, P. J. O., Johnson, M., Madsen, P. T., & Tyack, P. L. (2006). Deep-diving foraging behaviour of sperm whales (Physeter macrocephalus). *Journal of Animal Ecology*, *75*, 814–25.

Weeks, A. R. (1996). *Fundamentals of electronic image processing*. Bellingham: SPIE Optical Engineering Press.

Weilgart, L., & Whitehead, H. (1988). Distinctive vocalizations from mature male sperm whales (Physeter macrocephalus). *Canadian Journal of Zoology*, *66*, 1931–1937.

Weirathmueller, M. J., Wilcock, W. S. D., & Soule, D. C. (2013). Source levels of fin whale 20 Hz pulses measured in the Northeast Pacific Ocean. *Journal of the Acoustical Society of America*, *133*, 741–749.

Winn, W. E., & Perkins, P. J. (1976). Distribution and sounds of the minke whale, with a review of mysticete sounds. *Cetology*, *19*, 1–12.

Wood, S. N. (2013). *Package "mgcv."*

6. Figures

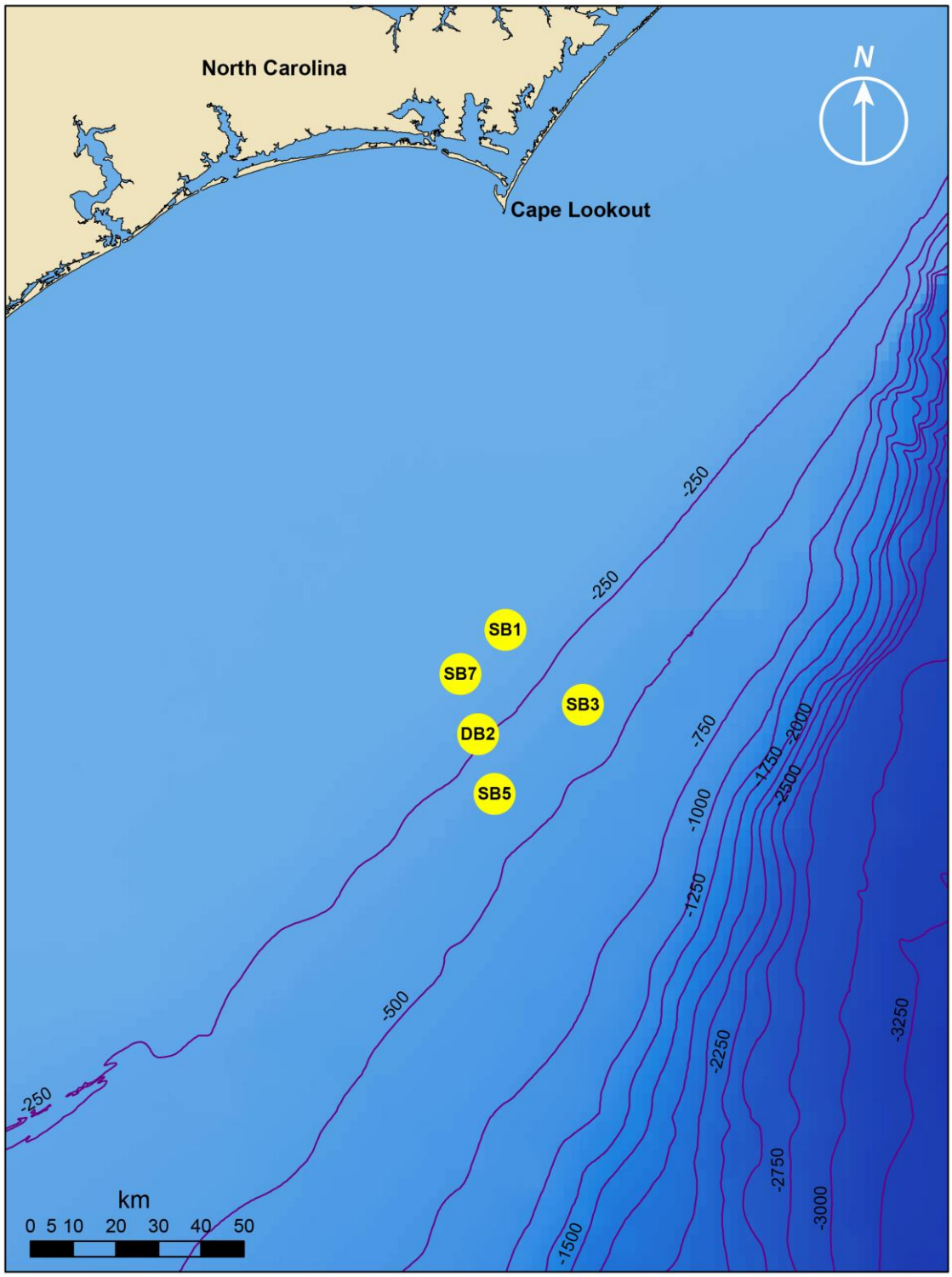


Figure 1. Map of Onslow Bay high-frequency MARU deployment sites. Not shown are deployment sites DB1, where the MARU failed after two days of recording, and SB4, where the MARU was not recovered.

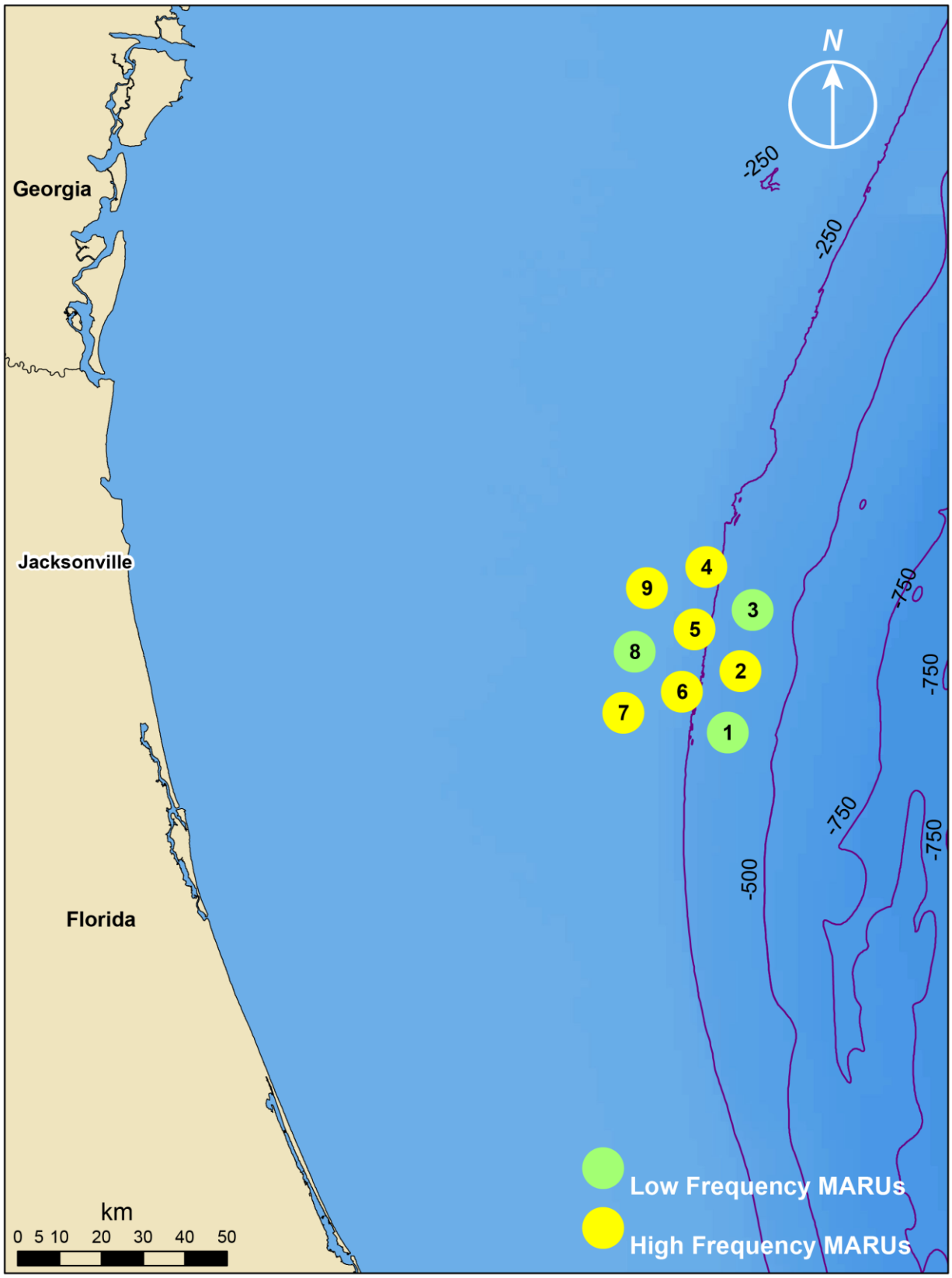


Figure 2. Map of Jacksonville MARU deployment sites.

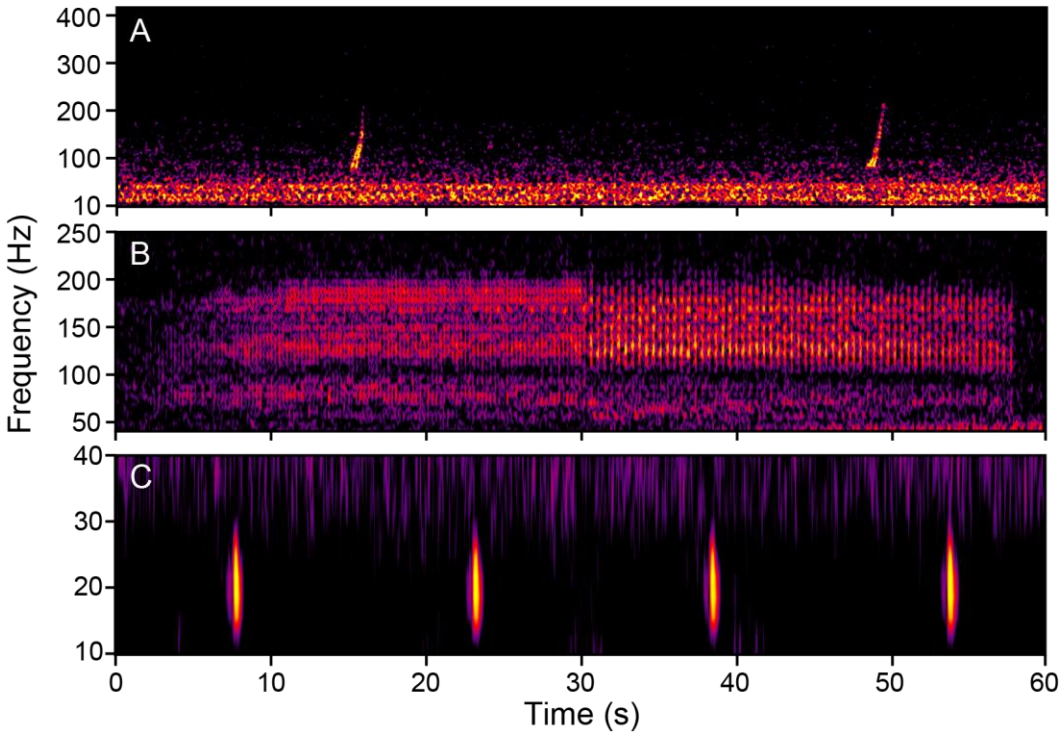


Figure 3. Examples of baleen whale sounds targeted in this study. A: Right whale upcalls, recorded May 2012, south of Nantucket, Massachusetts. B: Minke whale pulse train, recorded in the present study, Jacksonville Deployment 2, 24 December 2009. C: Fin whale 20-Hz notes, recorded December 2011, south of Nantucket.

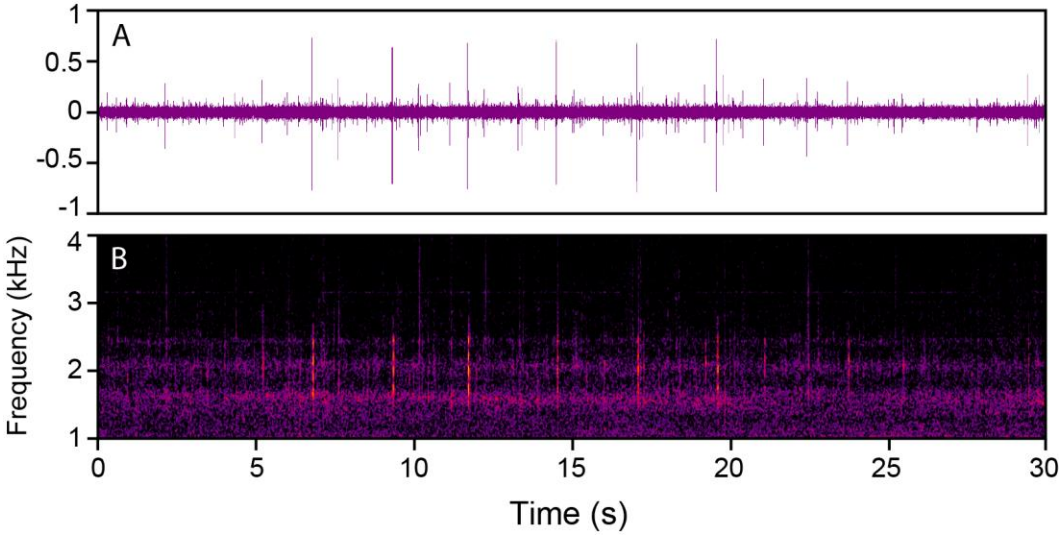


Figure 4. Sperm whale foraging clicks recorded during the present study, Jacksonville Deployment 1, 27 September 2009. In this segment of recording, click trains from at least two sperm whales are evident, with different amplitudes and inter-click intervals. A:

Waveform bandpass filtered at 1 and 4 kHz. Vertical scale is in arbitrary units. B: Spectrogram of the waveform depicted in A.

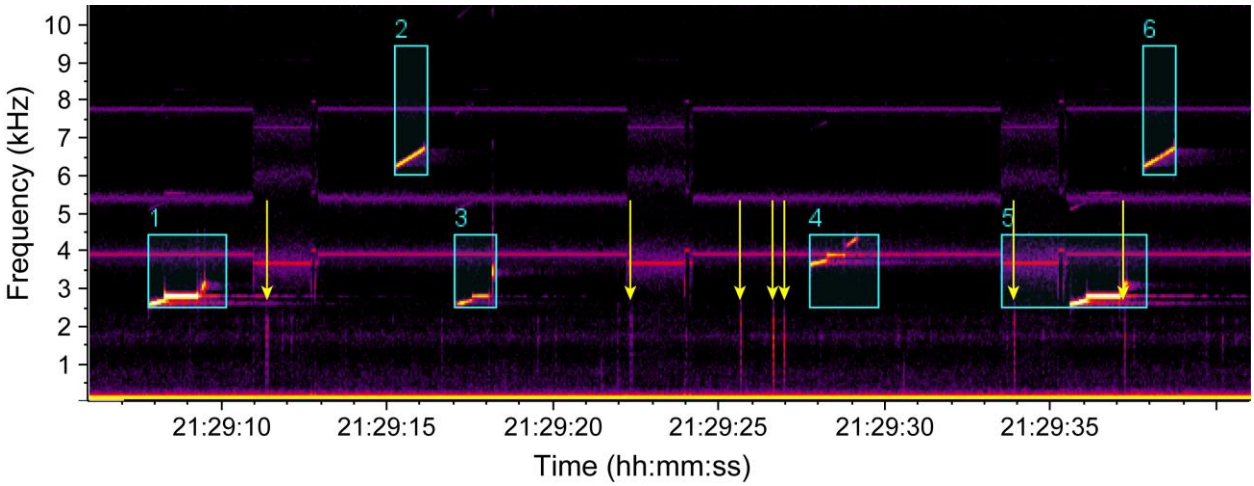


Figure 5. A sequence of sonar pings recorded over 35 seconds at Onslow Bay, July 2008. The numbered rectangles are the event boundaries created by the band-limited energy detector. The continuous line just below 4 kHz, and the brief frequency downshifts at 11-second intervals, are noise artifacts from the MARU's hard disk drive. The faint vertical lines marked by yellow arrows are sperm whale clicks.

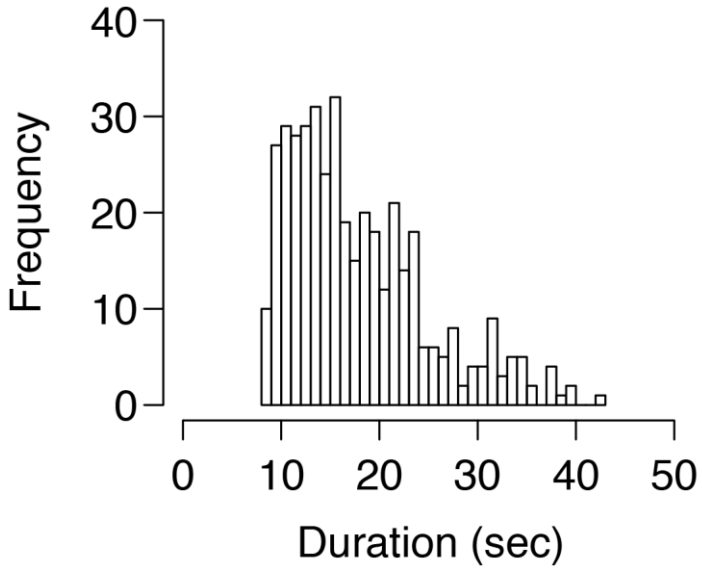


Figure 6. Duration of minke whale vocalizations in seconds. Each vocalization represents an individual pulse train from a single animal.

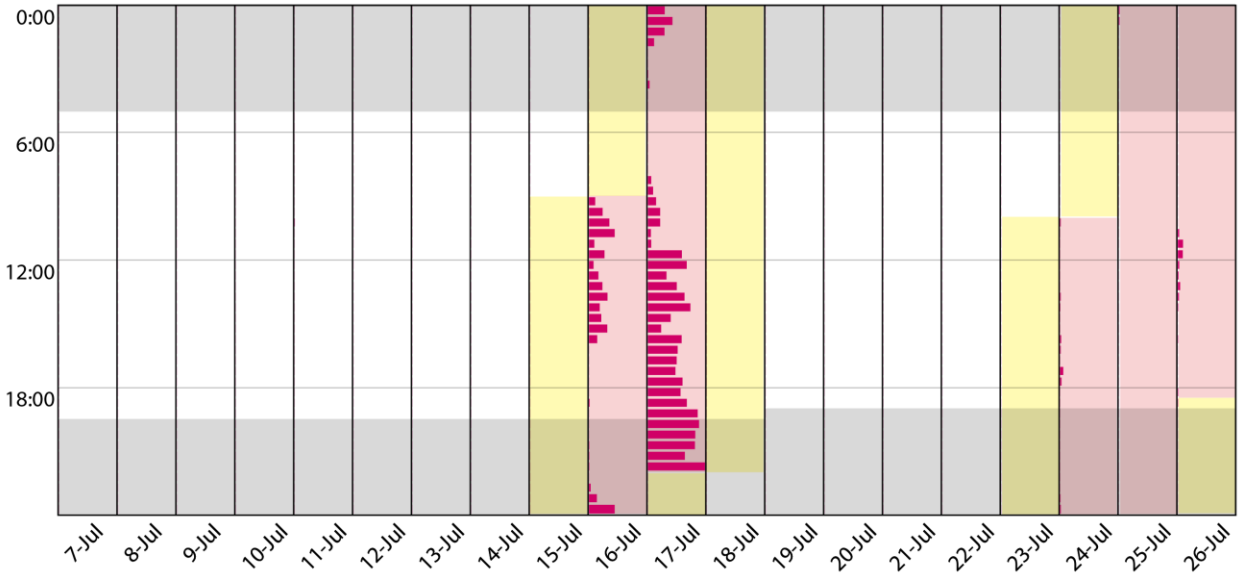


Figure 7. Temporal distribution of sonar transmissions recorded at site DB2, Onslow Bay, in 30-minute bins. Maximum bar height (on 17 July) represents 306 pings. Gray shading indicates bins when the sun was below the horizon for more than half of the bin period. Pink shading indicates periods designated as sonar events for statistical modeling; yellow shading indicates periods designated as *before* and *after* sonar events.

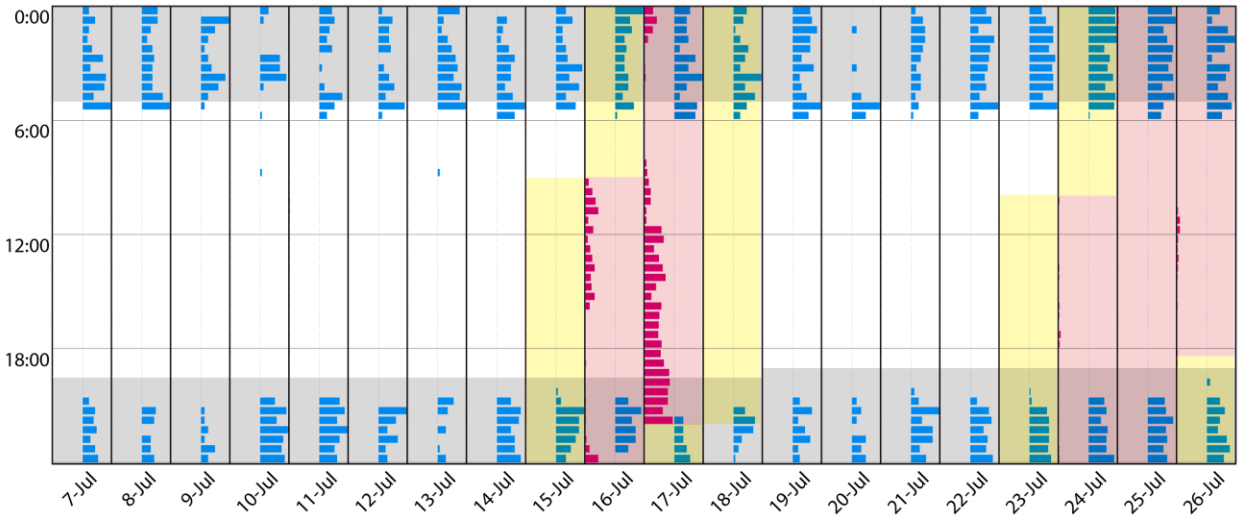


Figure 8. Occurrence of sperm whale click trains and sonar pings detected at recording site DB2, Onslow Bay. For each day, relative numbers of detected sperm whale click trains in each 30-minute bin are indicated by blue bars; number of sonar pings are indicated by red bars. Maximum bar height for sperm whale click trains = 30; maximum sonar = 306 pings. Gray shading indicates bins when the sun was below the horizon for more than half of the bin period. Pink shading indicates periods designated as sonar exercises for statistical modeling; yellow shading indicates periods designated as *before* and *after* sonar exercises.

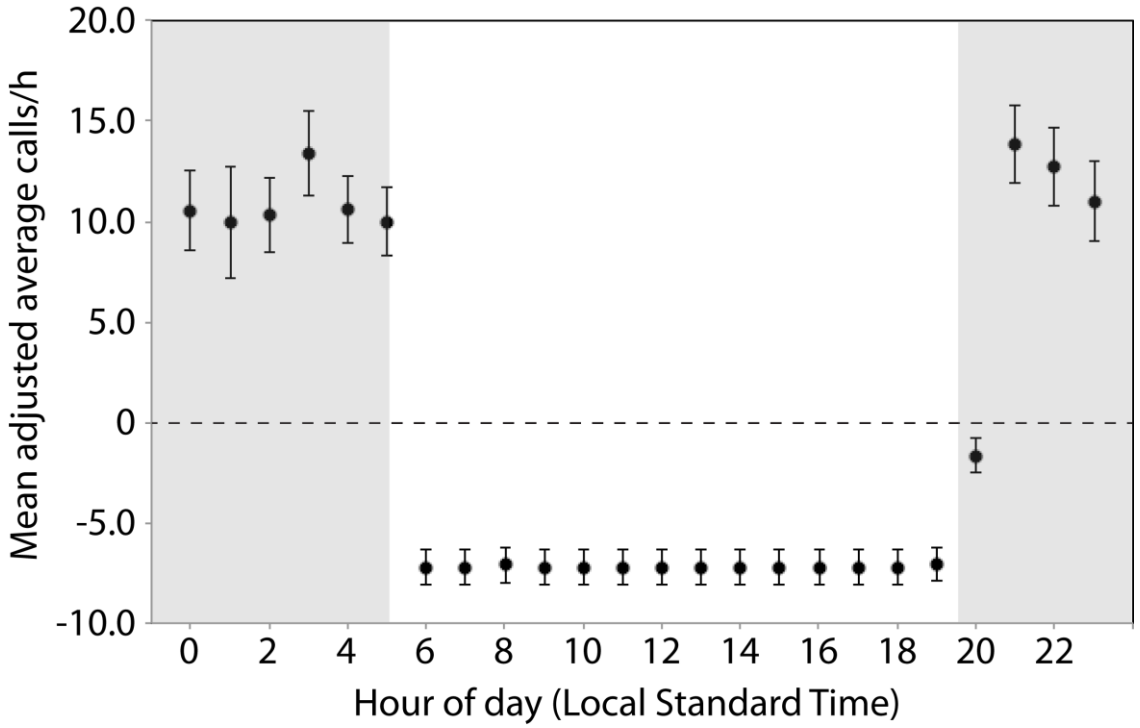


Figure 9. Onslow Bay Site DB2: Mean ± SEM number of sperm whale click trains per hour, adjusted relative to the mean number of click trains per hour for each day (indicated by the dashed line). Standard errors based on *N* = 20 days.

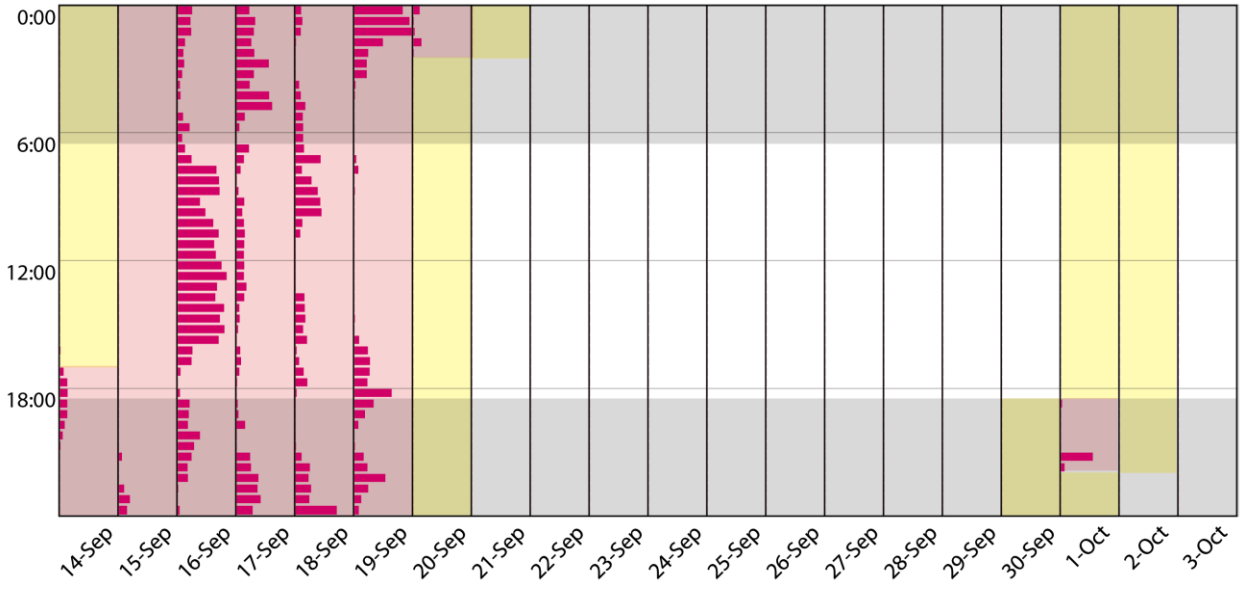


Figure 10. Temporal distribution of sonar transmissions recorded at site 5, Jacksonville Deployment 1, in 30-minute bins. Maximum bar length (as seen on 19 Sep) represents 137 pings. Gray shading indicates bins when the sun was below the horizon for more than half of the bin period. Pink shading indicates periods designated as sonar exercises for statistical modeling; yellow shading indicates periods designated as *before* and *after* sonar exercises.

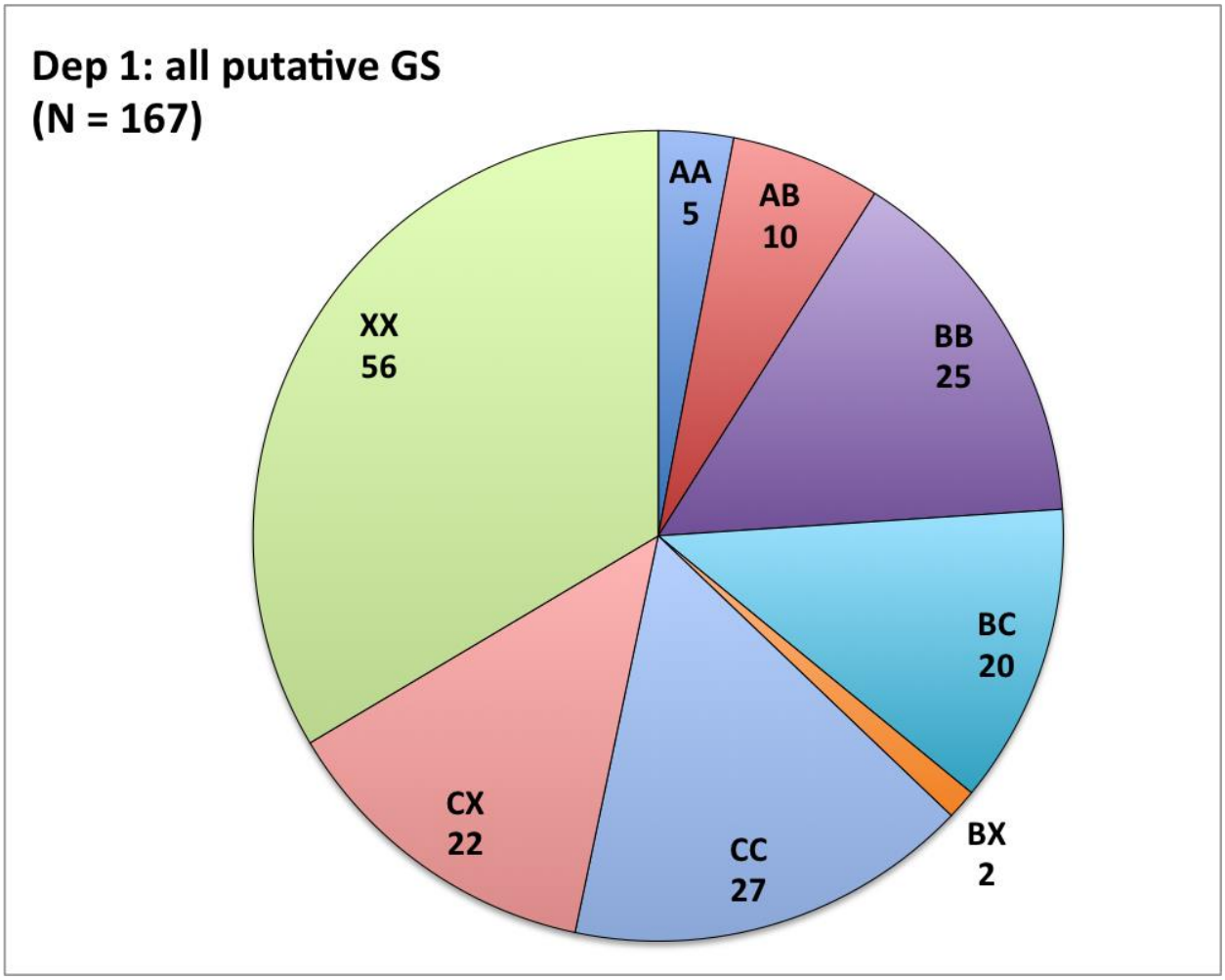


Figure 11. Results of independent review of 167 putative right whale gunshot (GS) sounds in Jacksonville Deployment 1 by two experienced analysts. Each sector represents the proportion of all events that were scored with the codes shown by the two-letter labels, according to the scoring criteria given in Table 4. Numbers below the scoring codes indicate the number of events with the corresponding score.

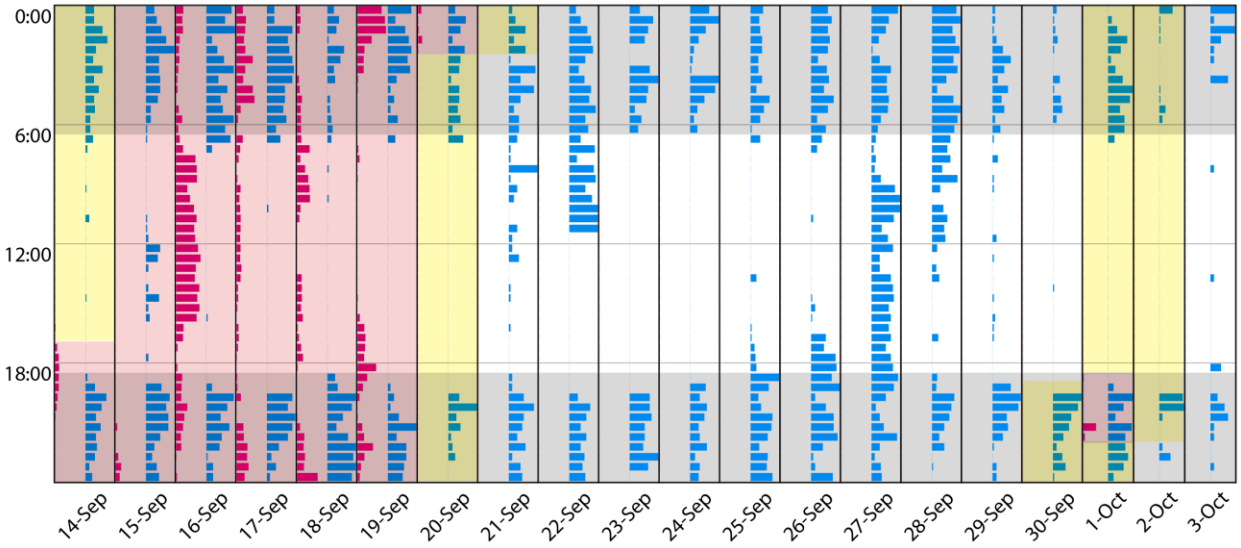


Figure 12. Occurrence of sperm whale click trains and sonar pings, Jacksonville Deployment 1, Site 5. For each day, relative numbers of detected sperm whale click trains in each 30-minute bin are indicated by blue bars; number of sonar pings are indicated by red bars. Maximum bar height for sperm whale click trains = 31; maximum sonar = 137 pings. Gray shading indicates bins when the sun was below the horizon for more than half of the bin period. Pink shading indicates periods designated as sonar exercises for statistical modeling; yellow shading indicates periods designated as *before* and *after* sonar exercises.

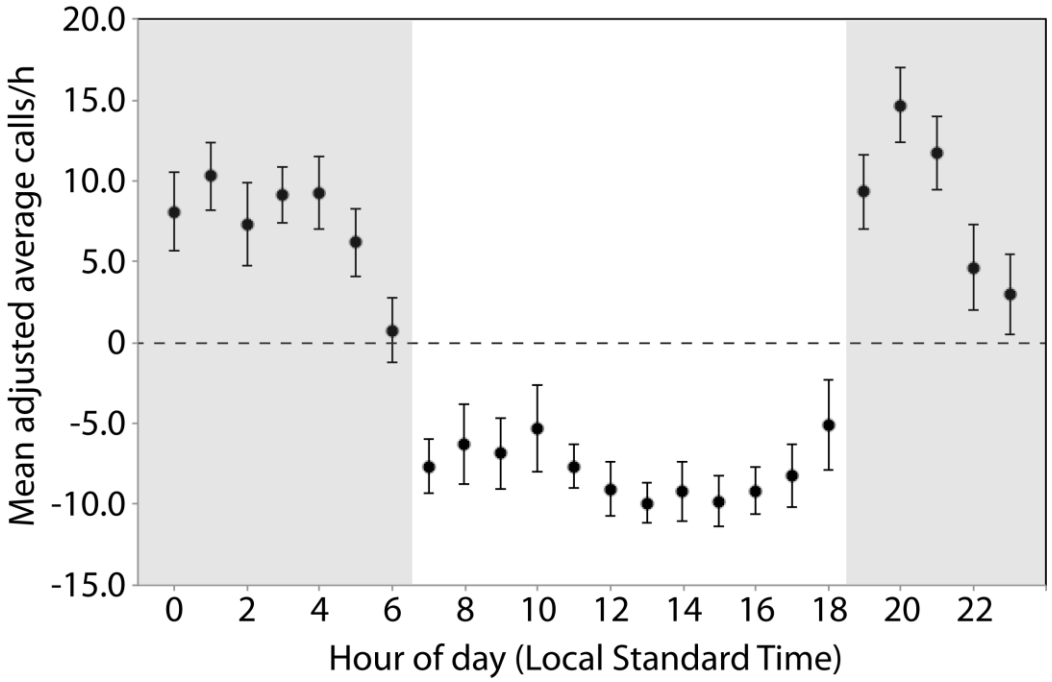


Figure 13. Jacksonville Deployment 1: Mean ± SEM number of sperm whale click trains detected per hour at Site 05, adjusted relative to the mean number of click trains per hour for each day (indicated by the dashed line). Standard errors based on *N* = 21 days.

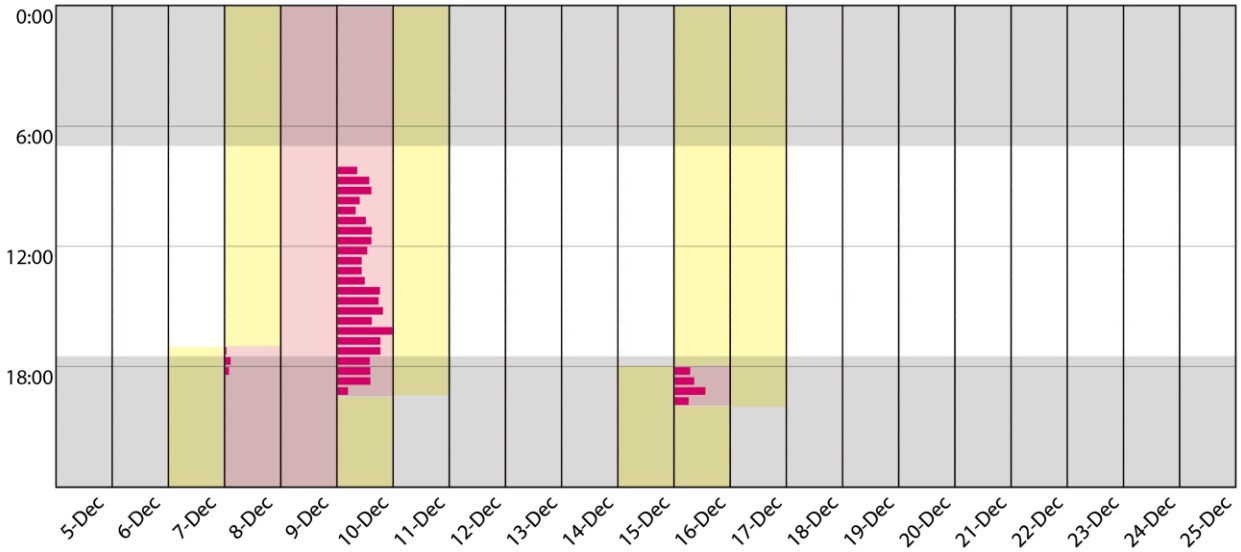


Figure 14. Temporal distribution of sonar transmissions recorded at site S05, Jacksonville Deployment 2, in 30-minute bins. Gray shading indicates bins when the sun was below the horizon for more than half of the bin period. Maximum bar length represents 107 pings. Pink shading indicates periods designated as sonar exercises for statistical modeling; yellow shading indicates periods designated as *before* and *after* sonar exercises.

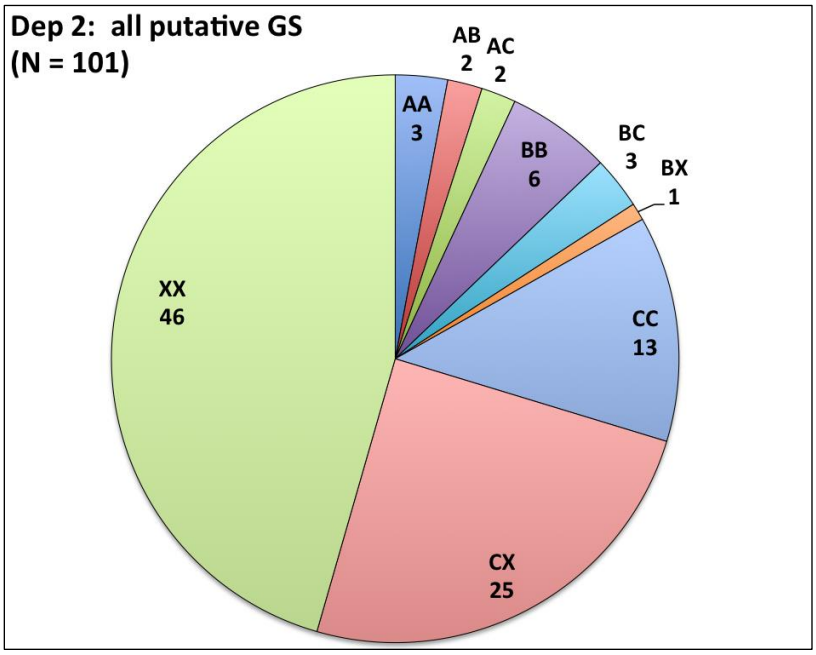


Figure 15. Results of independent review of 101 putative right whale gunshot (GS) sounds in Jacksonville Deployment 2 by two experienced analysts. Each sector represents the proportion of all events that were scored with the codes shown by the two-letter labels, according to the scoring criteria given in Table 4. Numbers below the scoring codes indicate the number of events with the corresponding score.

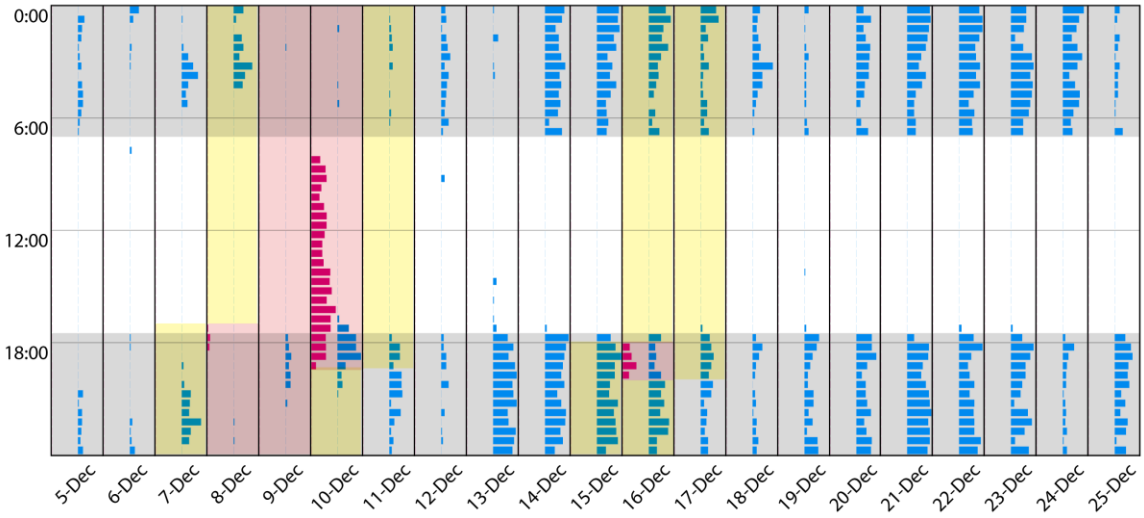


Figure 16. Occurrence of sperm whale click trains and sonar pings detected during Jacksonville Deployment 2 at Site 05. For each day, relative numbers of detected sperm whale click trains in each 30-minute bin are indicated by blue bars; number of sonar pings are indicated by red bars. Gray shading indicates bins when the sun was below the horizon for more than half of the bin period. Maximum bar height for sperm whale click trains = 30; maximum sonar = 107. Pink shading indicates periods designated as sonar exercises for statistical modeling; yellow shading indicates periods designated as *before* and *after* sonar exercises.

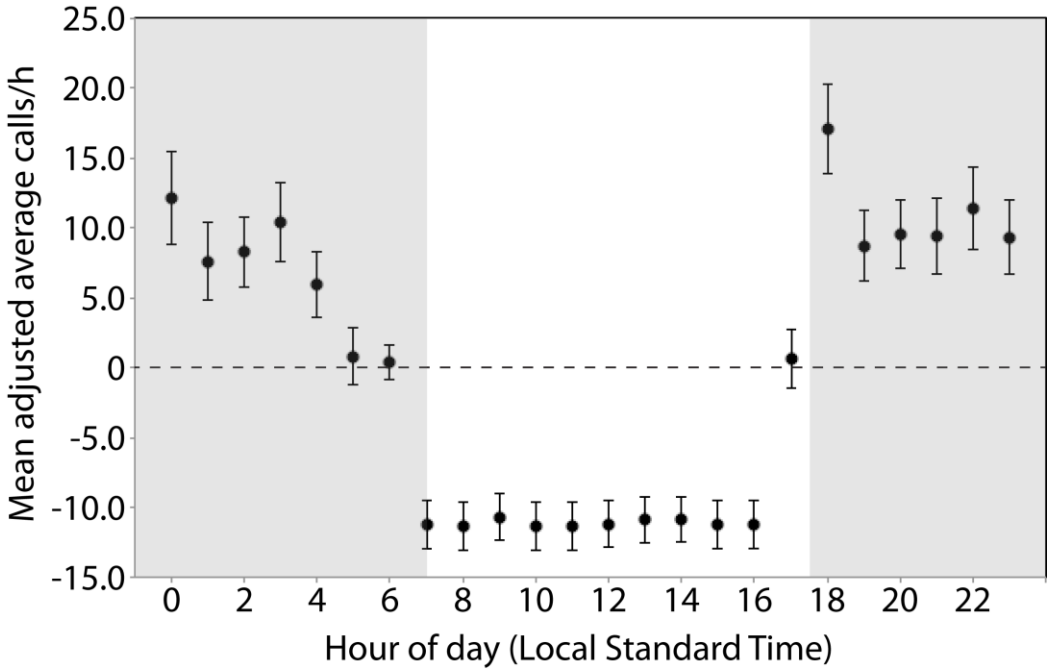


Figure 17. Jacksonville Deployment 2: Mean ± SEM number of sperm whale click trains per hour, adjusted relative to the mean number of click trains per hour for each day (indicated by the dashed line). Standard errors based on *N* = 21 days.

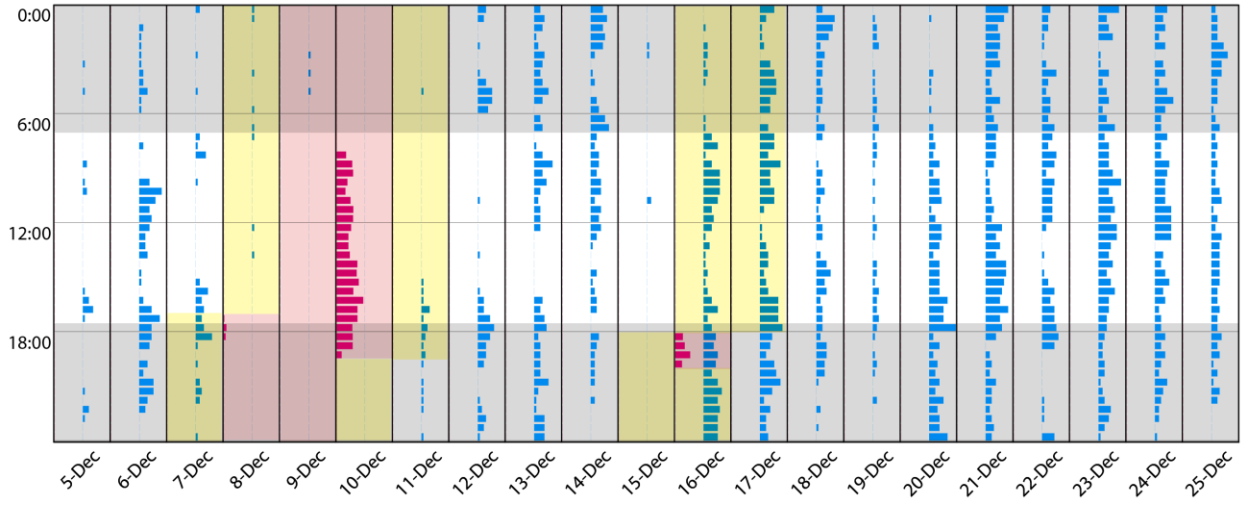


Figure 18. Detections of minke whale pulse trains at Site 3 and sonar pings at Site 5, Jacksonville Deployment 2. For each day, relative numbers of detected minke whale pulse trains in each 30-minute bin are indicated by blue bars; number of sonar pings are indicated by red bars. Maximum bar height for minke whale pulse trains = 13; maximum bar height for sonar = 107 pings. Gray shading indicates bins when the sun was below the horizon for more than half of the bin period. Pink shading indicates periods designated as sonar exercises for statistical modeling; yellow shading indicates periods designated as *before* and *after* sonar exercises.

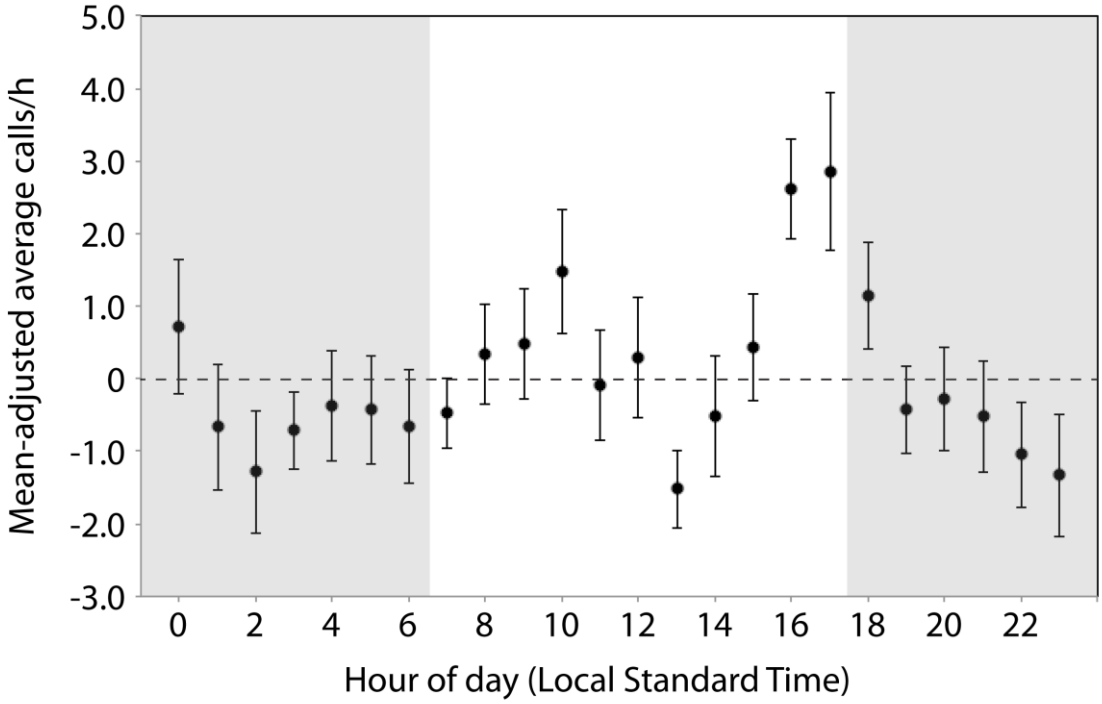


Figure 19. Jacksonville Deployment 2: Mean ± SEM number of minke calls per hour, adjusted relative to the mean number of calls per hour for each day (indicated by the dashed line). Standard errors were based on *N* = 21 days.

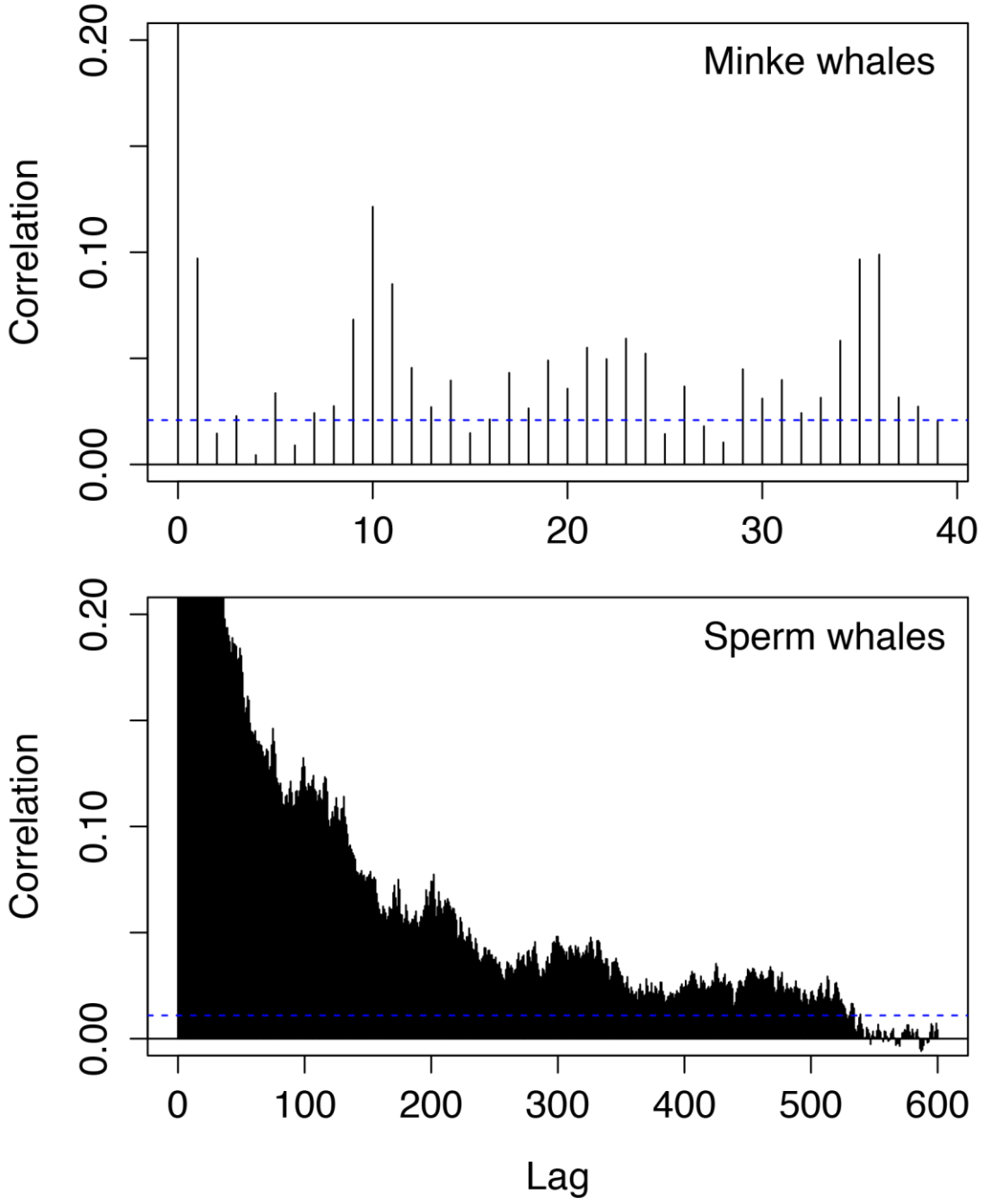


Figure 20. Autocorrelation of Pearson's residuals from presence models for minke (top) and sperm whales (bottom), including 95 percent confidence intervals around zero autocorrelation (blue dashed line). Lag is in units of 1-minute intervals. The y-axis was limited to 0.2 for illustrative purposes.

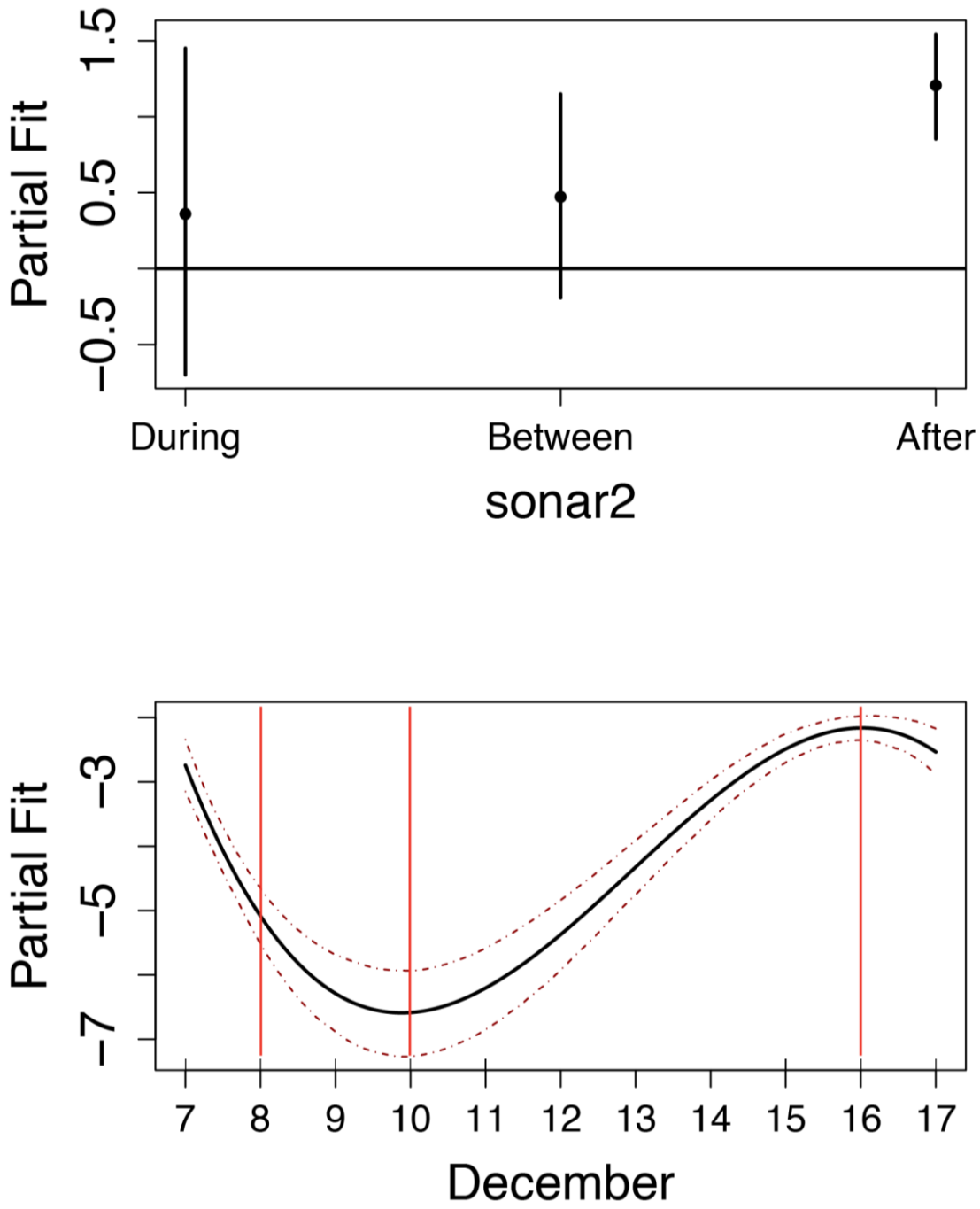


Figure 21. Partial fit plots for the best fitting presence model for minke whales (note that the partial fit is given on the scale of the logit-link function). For the factor covariate *Sonar2* (upper panel) the vertical bars are 95 percent confidence intervals for the respective coefficients. For *Julian date* (lower panel), the dashed lines are 95 percent confidence intervals around the partial fit and the vertical red lines indicate dates on which sonar transmissions occurred. Tick marks along the x-axis demarcate the locations of the observed values of the covariate. Data on whale call presence during periods more than 24 hours before the start or after the end of a sonar even were not used in the models.

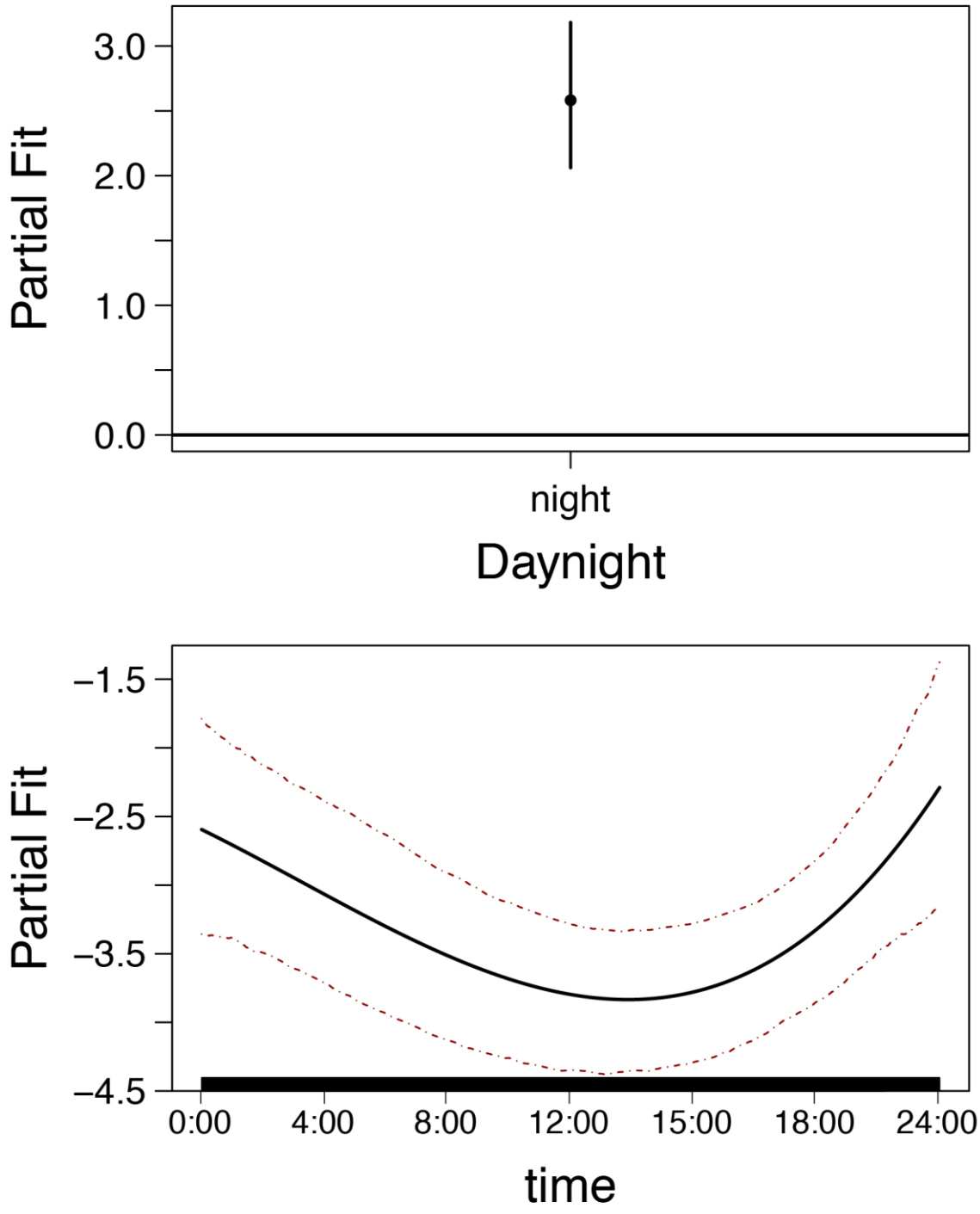


Figure 22. Partial fit plots for the best fitting presence model for sperm whales (note that the partial fit is given on the scale of the logit-link function). For factor covariate *Daynight* the vertical bar represents the 95 percent confidence intervals for the respective coefficient. For time, the dashed lines are 95 percent confidence intervals around the partial fit. Tick marks along the x-axis indicate the observed values of the respective covariate. For factor covariate *Daynight* the base level was *day*.

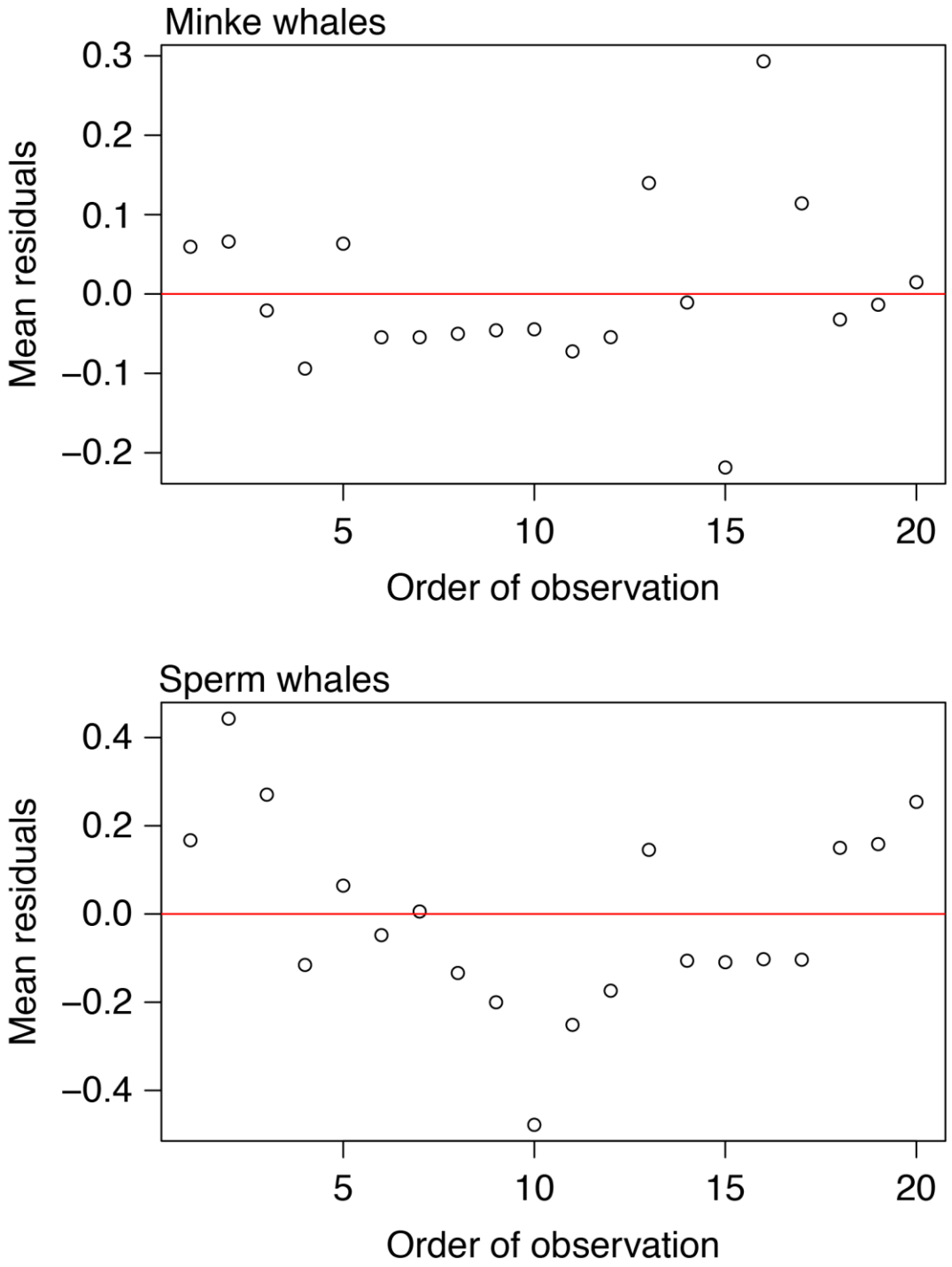


Figure 23. Means of binned fitted values versus means of corresponding residuals from presence models for minke (top) and sperm whale (bottom) detections. Binning occurred by splitting the residuals into 20 equally sized bins in the order of observation.

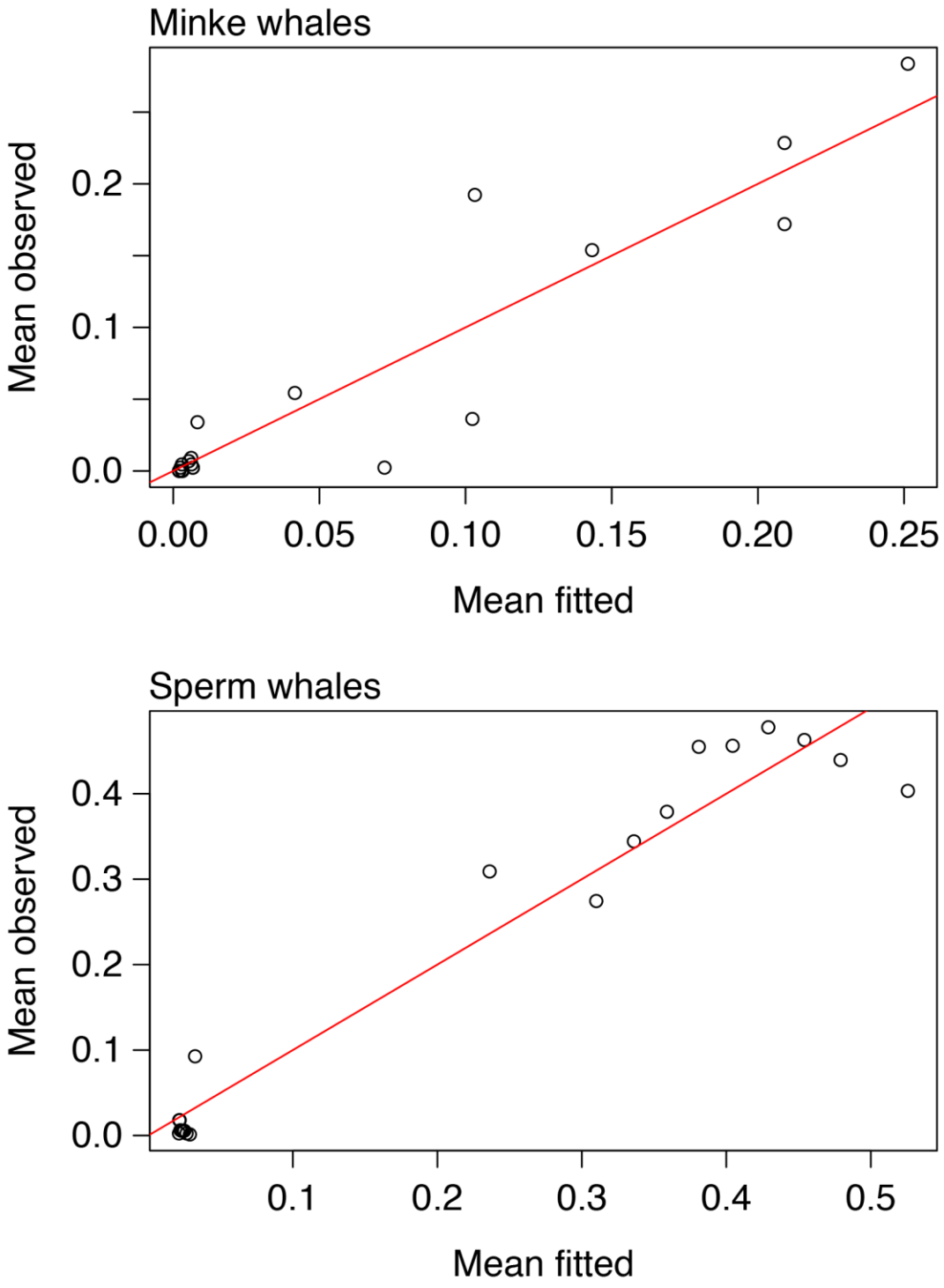


Figure 24. Mean observed versus mean fitted values from presence of vocalizations models for minke and sperm whales. Note that observations and fitted values were combined into 20 equally sized bins in ascending order of fitted values for which the means were calculated. Red lines represent a perfect fit for the models.

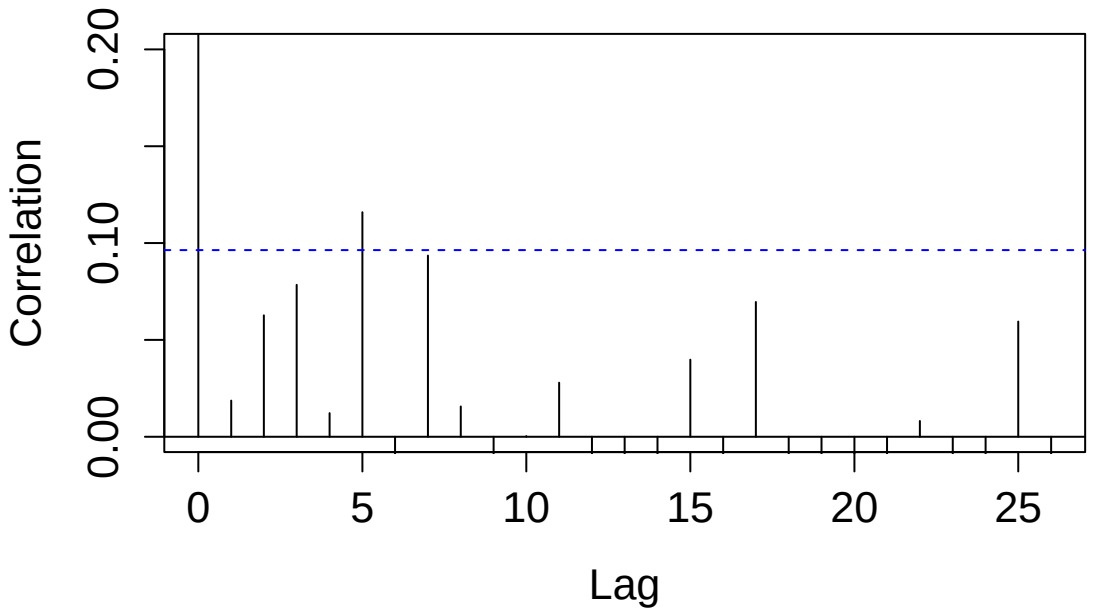


Figure 25. Autocorrelation of Pearson's residuals from duration models for minke whales including 95 percent confidence intervals around zero autocorrelation (blue dashed line). The y-axis was limited to 0.2 for illustrative purposes.

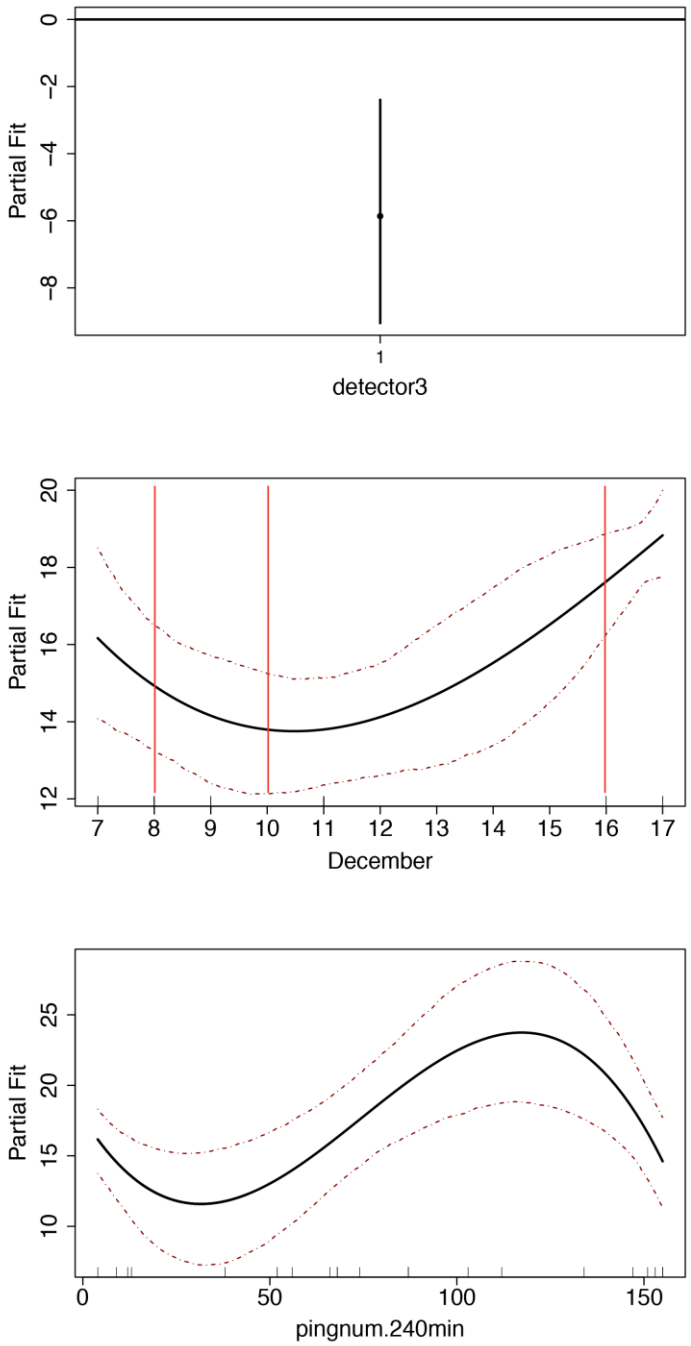


Figure 26. Partial fit plot for the best fitting duration model for minke whales (note that the partial fit is given on the scale of the identity-link function). For the factor covariate *detector3* (top panel) the vertical bars are 95 percent confidence intervals for the respective coefficients. For the smoothing terms *Julian date* (middle panel) and *Pingnum.240min* (bottom panel), the dashed lines are 95 percent confidence intervals around the partial fit. The base level for covariate *Detector3* was *0* corresponding to absence of detections of sonar pings in the 3.5 kHz band with the detector (as opposed to *1*, corresponding to presence of sonar detections).

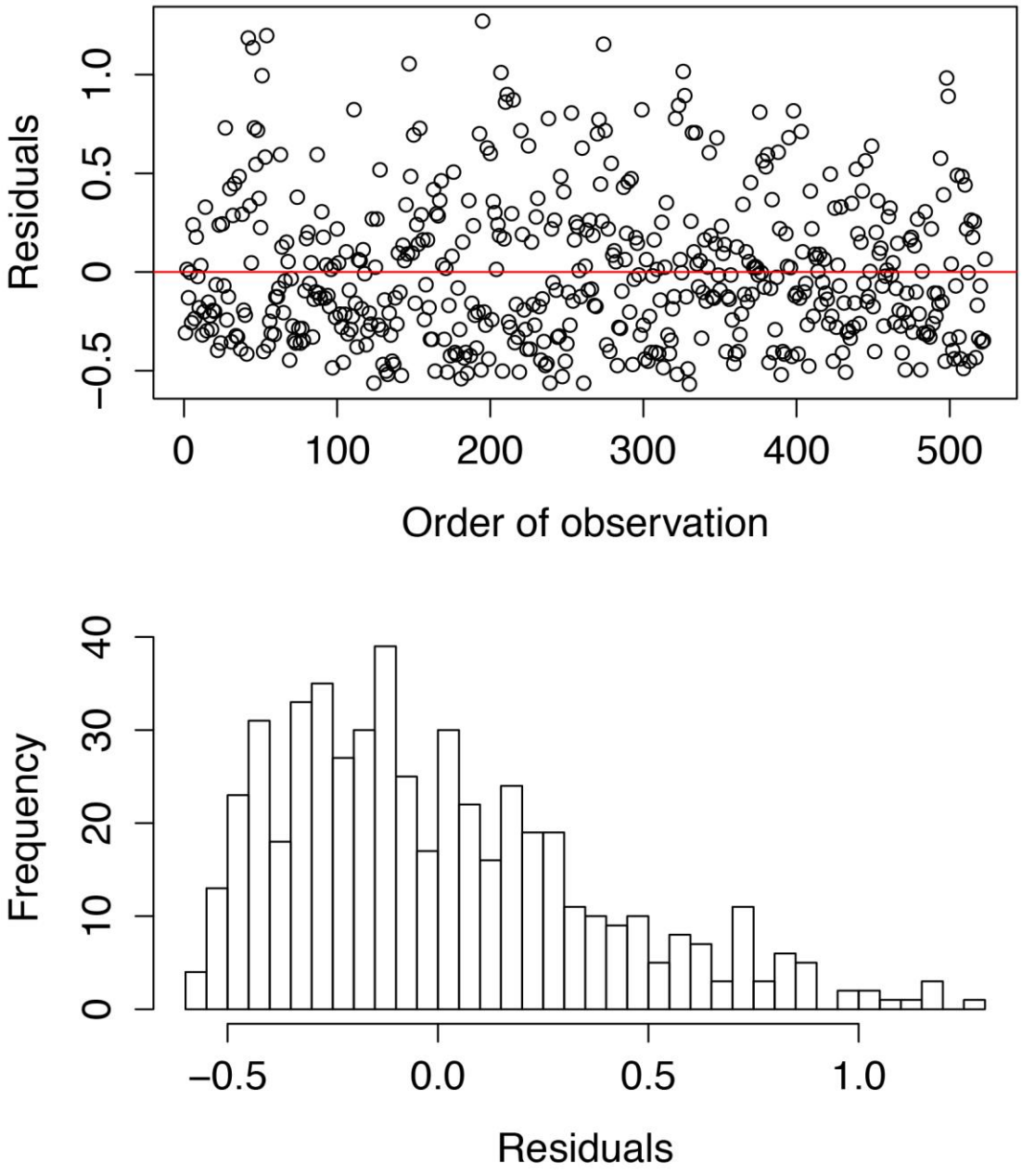


Figure 27. Pearson's residuals plotted in order of observation and histogram of Pearson's residuals from duration model for minke whales.

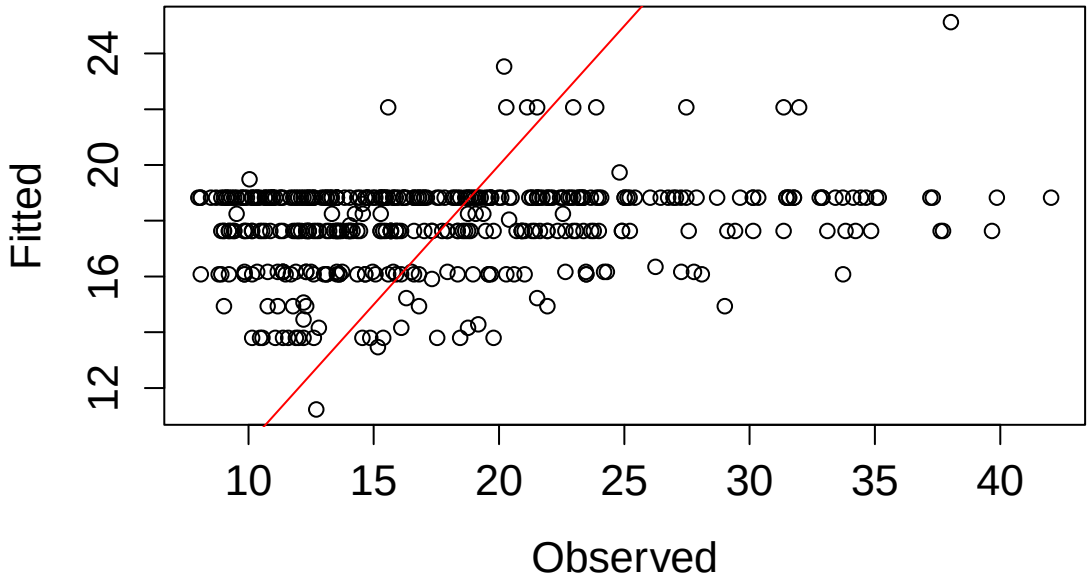


Figure 28. Observed vs fitted duration of vocalization from duration models for minke whales. The red lines indicate a perfect fit of the model to the observed data.

This page intentionally left blank.

7. Tables

Table 1. Summary of MARU deployment information for Onslow Bay Deployment. The MARU at Site DB1 stopped recording for unknown reasons after 2 days. The MARU at Site SB4 was not recovered. Due to limited analysis resources, sonar and whale detection data are presented only from Site DB2.

Site ID	Sampling Rate	Depth (m)	Latitude	Longitude	First Full Day of Recording	Last Full Day of Recording	Total Days
DB1	32 kHz	229	33˚45.1659' N	76˚29.8259' W	07-Jul-08	08-Jul-08	2
DB2*	32 kHz	236	33˚40.4544' N	76˚35.3824' W	07-Jul-08	26-Jul-08	20
SB1	32 kHz	64	33˚51.4066' N	76˚31.8920' W	07-Jul-08	26-Jul-08	20
SB3	32 kHz	≈365	33˚43.5459' N	76˚22.1315' W	07-Jul-08	26-Jul-08	20
SB4	32 kHz	305+	33˚38.7709' N	76˚27.8068 W	07-Jul-08	NA	NA
SB5	32 kHz	≈365	33˚34.1638' N	76˚33.3090' W	07-Jul-08	26-Jul-08	20
SB7	32 kHz	73	33˚46.7647' N	76˚37.5837' W	07-Jul-08	26-Jul-08	20

Table 2. Summary of MARU site information for JAX Deployment 1. The fathometer onboard the deployment vessel could not report depths > 305 m. Due to limited analysis resources, sonar and sperm whale detection data are presented only for Site 5.

Site ID	Sampling Rate	Depth (m)	Latitude	Longitude	First Full Day of Recording	Last Full Day of Recording	Days Analyzed
1	02 kHz	305+	30° 03.015' N	80° 06.575' W	14-Sep-09	8-Oct-09*	20
2	32 kHz	305+	30° 09.867' N	80° 04.966' W	14-Sep-09	3-Oct-09	20
3	02 kHz	305+	30° 16.686' N	80° 03.361' W	14-Sep-09	8-Oct-09*	20
4	32 kHz	168	30° 21.435' N	80° 09.331' W	14-Sep-09	3-Oct-09	20
5	32 kHz	201	30° 14.505' N	80° 10.879' W	14-Sep-09	3-Oct-09	20
6	32 kHz	192	30° 07.594' N	80° 12.486' W	14-Sep-09	3-Oct-09	20
7	32 kHz	45	30° 05.218' N	80° 20.055' W	14-Sep-09	3-Oct-09	20
8	02 kHz	46	30° 12.052' N	80° 18.585' W	14-Sep-09	8-Oct-09*	20
9	32 kHz	45	30° 19.092' N	80° 17.010' W	14-Sep-09	3-Oct-09	20

* Although the LF (2 kHz) MARUs recorded through 8 Oct, recordings were analyzed only through 3 Oct, the last complete day of HF recording.

Table 3. Summary of MARU site information for JAX Deployment 2. Deployment sites are the same as for Deployment 1 (Table 2); only the dates differ. Due to limited analysis resources, sonar and sperm whale detection data are presented only for Site 5, and minke whale detections only for Site 3.

Site ID	Sampling Rate	Depth (m)	Latitude	Longitude	First Full Day of Recording	Last Full Day of Recording	Days Analyzed
1	2 kHz	305+	30° 03.015' N	80° 06.575' W	5-Dec-09	7-Jan-10*	21
2	32 kHz	305+	30° 09.867' N	80° 04.966' W	5-Dec-09	25-Dec-09	21
3	2 kHz	305+	30° 16.686' N	80° 03.361' W	5-Dec-09	7-Jan-10*	21
4	32 kHz	168	30° 21.435' N	80° 09.331' W	5-Dec-09	25-Dec-09	21
5	32 kHz	201	30° 14.505' N	80° 10.879' W	5-Dec-09	25-Dec-09	21
6	32 kHz	192	30° 07.594' N	80° 12.486' W	5-Dec-09	25-Dec-09	21
7	32 kHz	45	30° 05.218' N	80° 20.055' W	5-Dec-09	25-Dec-09	21
8	2 kHz	46	30° 12.052' N	80° 18.585' W	5-Dec-09	7-Jan-10*	21
9	32 kHz	45	30° 19.092' N	80° 17.010' W	5-Dec-09	25-Dec-09	21

* Although the LF (2 kHz) MARUs recorded through 7 Jan, recordings were analyzed only through 25 Dec, the last day of HF recording.

Table 4. Rating scheme for evaluating putative North Atlantic right whale (NARW) gunshot (GS) sounds.

Rating	Interpretation	Criteria
A	Strong match; not distinguishable from known gunshots	• Broadband pulse as described by Parks et al., 2005) and Trygonis et al., 2013); • ≥ 2 pulses visible; • High SNR (>≈ 10 dB); • Aural quality similar to exemplars recorded by S. Parks.
B	Moderate match	• Broadband impulsive sound; • Low SNR and/or overlapping other sounds, and/or • Frequency spectrum deviates from expectation and/or • Multi-pulse structure not clearly evident;
C	Weak but possible match; clearly different from known GS, but cannot exclude NARW as source.	• Broadband impulsive sound; • Low SNR and/or overlapping other sounds, and/or • Frequency spectrum deviates from expectation and/or • Duration longer than expected and/or • No multi-pulse structure
X	Reject as GS	• No broadband impulsive sound found in the event, or • Impulsive sound aurally similar to alternative source (e.g., sperm whale click, impact of drifting debris on MARU, anthropogenic source).

Table 5. Covariates included in the analyses

Covariate	Description	Unit
Site	Combination of location and deployment	--
Time	Time of day	Seconds
JD	Julian Date	Days
Daynight	Factor covariate: *day* or *night*	--
Sonar2	Factor covariate for presence models: temporal relation to sonar exercises *(before, during, between* or *after*)	--
Sonar3	Factor covariate for duration models: temporal relation to sonar exercises (*before*, *during/ between* and *after*)	--
Sonarlag	Time passed since last sonar ping	Minutes
Number of pings in 30, 60, 120 or 240 minutes (labeled e.g., Pingnum.30min)	Numbers of pings occurring in the respective period preceding the 1-minute segment (presence models) or start of vocalization (duration models)	Number of pings
Average ping interval in 30, 60, 120 or 240 minutes (labeled e.g., Pingint.30min)	Average time lag between pings in the respective period preceding the 1 minute segment (presence models) or start of vocalization (duration models)	Seconds
Detector3	Factor covariate: presence of sonar ping detections within the 2.5 – 4.4 kHz band	--
Detector7.5	Factor covariate: presence of sonar ping detections within the 6.4 – 8.7 kHz band	--

Table 6. Number of 1-minute segments and number of vocalizations used for the presence and duration models, respectively, given for each species and deployment. No values are shown for detections for sperm whales because duration models were not applicable to the sperm whale detection data.

Deployment	Species	1-minute segments (presence models)	Detections (duration model)
JAX 1	Minke whale	0	0
JAX 2	Minke whale	8,821	414
OB 2	Minke whale	0	0
JAX 1	Sperm whale	13,281	--
JAX 2	Sperm whale	8,821	--
OB 2	Sperm whale	10,244	--

Table 7. Presence models for minke and sperm whales: maximum likelihood estimates (MLE) of parameters on the logit-link scale and standard errors (SE) from best-fitting models. For factor terms, we list the level for the coefficient. For the smoothing terms (indicated with *bs()*), the three coefficients refer to the β associated with the polynomial term in the overall contribution of covariate x_k to the predictor: $\beta_1 x_k + \beta_2 x_k^2 + \beta_3 x_k^3$.

	Minke Whales		Sperm Whales	
	MLE	SE	MLE	SE
Max. block sizes	2		528	
Linear and factor terms				
Intercept	-2.74	0.21***	-2.58	0.43***
Daynight level: night	--	--	2.58	0.30***
Sonar2 level: during	0.36	0.55	--	--
Sonar2 level: between	0.47	0.35	--	--
Sonar2 level: after	1.21	0.18***	--	--
Smoothing terms				
bs(time)1	--	--	-0.88	0.85
bs(time)2	--	--	-2.43	0.848***
bs(time)3	--	--	0.31	0.28.
bs(JulianDate)1	-9.96	1.02***	--	--
bs(JulianDate)2	2.87	0.49***	--	--
bs(JulianDate)3	0.21	0.29	--	--
Additional Parameters				
Dispersion parameters	0.94	1.71	0.91	0.55

.P < 0.10
*P < 0.05
** P < 0.01
*** P < 0.001

Table 8-Observed versus predicted presences (1) and absences (0). Numbers in black are proportions of correct predictions, numbers in red and blue represent proportions of falsely predicted absences or presences, respectively.

Predicted	Observed			
	Minke Whales		Sperm Whales	
	absent	present	absent	present
absent	0.62	0.00	0.50	0.01
present	0.32	0.06	0.30	0.20

Table 9. Duration models for minke whales: parameter estimates (MLE) on the identity-link scale and standard errors (SE) from best-fitting models. For factor terms, we list the level for the coefficient. For the smoothing terms (indicated with *bs()*), the three coefficients refer to the β associated with the polynomial term in the overall contribution of covariate x_k to the predictor: $\beta_1 x_k + \beta_2 x_k^2 + \beta_3 x_k^3$. Note that covariate *Pingnum.240min* was constructed so that it only affected the model for those observations where the covariate could be observed (see Section 2.6.2 for details).

	Minke Whales	
	MLE	SE
Max. block sizes	1	
Linear and factor terms		
Intercept	16.17	1.15***
Detector3) level: 1 (presence)	-5.86	1.76***
Smoothing terms		
bs(JulianDate)1	-4.94	3.40
bs(JulianDate)2	-1.41	3.46
bs(JulianDate)3	2.66	1.29*
bs(Pingnum.240min) 1	-18.20	5.75**
bs(Pingnum.240min) 2	25.58	6.92***
bs(Pingnum.240min) 3	-1.56	1.10
Additional parameters		
Dispersion parameters (SE)	0.15	0.01

.P < 0.10
*P < 0.05
** P < 0.01
*** P < 0.001

This page intentionally left blank.