Final Report

Development of Statistical Methods for Examining Relationships Between Odontocete Vocal Behavior and Navy Sonar Signals

Submitted to:

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

NAVFAC
Naval Facilities Engineering Command

Prepared by:

Bio•Waves
Incorporated

Bio-Waves, Inc.
364 2nd Street, Suite #3
Encinitas, CA 92024
(760) 452-2575

and

CREEM
Centre for Research into
Ecological and Environmental Modelling

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

Submitted by:

HDR

Virginia Beach, VA

30 March 2015

Suggested Citation:

Oswald, J.N., C.S. Oedekoven, T.M. Yack, R. Langrock, L. Thomas, E. Ferguson, and T. Norris. 2015. *Development of Statistical Methods for Examining Relationships Between Odontocete Vocal Behavior and Navy Sonar Signals. 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. 30 March 2015.

Cover Photo Credits:

Atlantic spotted dolphins *(Stenella frontalis)* taken by Heather Foley, Duke University. Photo taken under NOAA Permit No. 16185.

Work conducted under contracts between Bio-Waves, Inc. and HDR, Inc., and between University of St Andrews and HDR, Inc.

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

Acoustic recordings were made off Jacksonville, Florida (JAX) and Onslow Bay (OB), North Carolina, using seafloor-deployed Marine Acoustic Recording Units (MARUs) to provide data to examine marine mammal vocal behavior before, during and after mid-frequency active (MFA) sonar (henceforth, sonar) exercises. In a preliminary analysis of the MARU data recorded off Jacksonville, the probabilities of detecting calls produced by marine mammals in the presence and absence of sonar were calculated for several species of cetaceans and two broader taxonomic categories: 'delphinids' and 'blackfish' (including pilot whales, false killer whales and melon-headed whales; Norris et al. 2012). The results of the probability analysis for these taxonomic categories were indeterminate, and as such, it was decided that a more in-depth analysis that examined vocal effects of sonar on a species-by-species basis with more sophisticated statistical methods was needed. The study presented here is intended to address this need. There were three primary objectives for this study (listed chronologically): 1) classify delphinid detections to the lowest taxonomic group possible using the Real-time Odontocete Call Classification Algorithm (ROCCA, Oswald 2013), 2) detect and characterize MFA sonar in MARU recordings, and 3) develop robust statistical methods that can be used to evaluate differences in delphinid whistling behavior in response to MFA sonar.

The first step in this analysis was to identify delphinid acoustic encounters to be used as inputs in the species classification analysis. A delphinid acoustic encounter was defined as a continuous portion of a recording with no more than a 30 minute (30 min) gap between sounds produced by delphinids. Delphinid sounds were defined as one of three mutually exclusive types: 1) whistles, 2) clicks or 3) burst pulses. ROCCA was used to extract and measure whistle features and to classify encounters containing whistles to species.

The second step in this study involved the detection and characterization of MFA sonar events. This was conducted using the program SonarFinder (Bio-Waves 2013). SonarFinder is a Matlab-based program designed to automatically detect sonar pings and measure their acoustic features. Sonar events were defined as a series of sonar pings with no more than a 30 min gap between individual pings.

A total of 1,259 delphinid acoustic encounters were logged from MARU recordings made at JAX and OB (deployments 1 and 2). Of these, 313 contained whistles with sufficient quality (i.e., > 3 dB SNR and did not overlap significantly with other whistle contours) to include in the ROCCA analysis. All encounters that were included in the ROCCA analysis were classified as either striped dolphin (*Stenella coeruleoalba*), short-beaked common dolphin (*Delphinus delphis*), or short-finned pilot whale (*Globicephala macrorhynchus*). No encounters were classified as bottlenose (*Tursiops truncatus*) or Atlantic spotted dolphins (*Stenella frontalis*). This result was unexpected, given that bottlenose and Atlantic spotted dolphins were the two species most frequently encountered during recent aerial and vessel-based line-transect visual surveys off Florida and North Carolina (DoN 2008, Halpin et al. 2009, Hodge et al. 2013). It is likely that many of the encounters classified as striped and short-beaked common dolphins were misclassifications. Although the ROCCA classifier performed well when it was ground-truthed using visually validated test data from the Northwest Atlantic Ocean (86 percent of acoustic encounters were correctly classified, n = 131; Oswald 2013), some bottlenose dolphin and Atlantic spotted dolphin encounters in the test data were misclassified (30 percent and 10 percent, respectively). Both bottlenose and Atlantic spotted dolphins were most commonly misclassified as striped dolphins, which may partially explain the large number of delphinid encounters classified as striped dolphins. Misclassifications may have been due to the classifier being trained using data recorded using a towed hydrophone array close to the sea surface, whereas the MARU recordings were made using stationary recorders moored at depth. It is possible that the classifier is using features of the signal that are susceptible to propagation or instrumentation effects (i.e., differences in whistle features detected were caused by the sound propagation or the recorders, which may have affected classification results. An additional issue that may result in misclassifications is the fact that ROCCA's Atlantic classifier only includes five species of whistling delphinids that are known to occur in the Northwest Atlantic Ocean. At least seven additional species that occur in this region are known to produce whistles (Palka 2012, Waring et al. 2012). If any of the delphinid vocalization encounters included in our analysis contained whistles from any of these species, they would be misclassified as one of the five species in the classifier, and so some of the encounters classified as striped and common dolphins could actually have been another species.

The MFA sonar analysis was conducted using SonarFinder software. This analysis resulted in the detection of 58 sonar events comprising 421.2 hours consisting of 31,826 sonar pings for JAX deployment 1. For OB, 72 sonar events comprising 158.5 hours and 30,403 sonar pings were detected. JAX deployment 2 was not included in the analysis because results of Norris et al. (2012) indicated that sonar events during this deployment were relatively infrequent, brief and/or sporadic.

The third step, statistical analysis, was divided into two approaches. The first approach used generalized estimating equations (GEEs), and the other used hidden Markov models (HMM). For the GEE approach, we defined three possible response variables, which in turn addressed three different research questions. These questions were as follows:

1. Does the probability of detecting delphinid acoustic encounters change in the presence of sonar?
2. Does the probability of detecting whistles, clicks or buzzes within a delphinid acoustic encounter change in the presence of sonar?
3. Given that a whistle encounter is detected, do whistle characteristics change in the presence of sonar?

The three corresponding response variables used in the three models were:

1. Presence (or absence) of acoustic encounters in 1 min segments;
2. Presence of signal type (e.g., whistles, clicks or buzzes) within an acoustic encounter (which we refer to as signal type models for short); and
3. Response intensity (constructed by combining multiple whistle characteristics into Mahalanobis distances; DeRuiter et al. 2013).

Each response variable was related to explanatory covariates. GEEs were used as the model fitting tool to accommodate potential over-dispersion in the data and correlation in the model errors. A three-step model selection procedure was used to obtain the best-fitting models for each approach. Due to potentially confounding differences in responses among species and species groups, we built separate models for pilot whale acoustic detections and for the combined detections from the remaining delphinid species (abbreviated as DEUO in the following, including common dolphins, striped dolphins and unidentified odontocetes). We found that covariates pertaining to sonar were retained in the best fitting signal type models for the DEUO species group. Covariates pertaining to sonar were also retained in the best fitting whistle characteristics models for the DEUO species group. However, we did not explore all potentially important covariates with respect to sonar. None of our covariates included a cumulative effect (e.g., the number of sonar pings in the two hours preceding a 1 min segment for the presence models or sound exposure levels of sonar). Additional analyses are necessary before these cumulative affect covariates can be included.

In the HMM-based modeling approach, the time series of acoustic encounters (response variable type 1 above) is assumed to be generated by a doubly stochastic process that switches between two different states, corresponding to acoustically active and more silent phases. HMMs naturally account for the multiphasic nature of the time series, with long periods without any acoustic encounters being recorded, occasionally interspersed with shorter periods that contain at least some acoustic encounters. In contrast to GEEs, in which case the correlation in the residuals is treated as a nuisance (i.e., a feature of the model that is not the focus of inference, but that needs to be accounted for, often in the simplest way possible - see, e.g., Basu 1977), HMMs attempt to explicitly model the correlation pattern, at the cost of increased computational complexity. By building separate models for pilot whales and for other delphinids, we investigated the effect of sonar-related covariates on the state-switching dynamics. For pilot whale HMMs, very few vocalizations (and hence also state transitions) occurred during the observation period. As a result, the estimation was numerically unstable in terms of local maxima of the likelihood. Furthermore, no clear pattern was found in the Akaike Information Criterion (AIC) values for the fitted models, likely due to the limited amount of information contained in these time series. For pilot whales in the JAX study area, the model with the covariate pertaining to the standard deviation of the ping interval (i.e. the *SDEV ping interval* covariate) was favored by the AIC, whereas in the OB study area the model without any covariates was favored. For the DEUO species group, the model with the *Sonar* covariate affecting the state transition probabilities was deemed best by the AIC, for both the JAX and OB study areas.

The results of this work provide a framework for developing tools that can be used in many locations and situations to understand the potential effects of MFA sonar on marine mammal acoustic behavior.

This page intentionally left blank.

Table of Contents

Appendix

A Variables Measured by VOCCA

Acronyms and Abbreviations

ACF	Auto-Correlation Function
AIC	Akaike Information Criterion
dB	decibels
DEUO	striped dolphins, common dolphins and unidentified odontocetes combined
HMM	hidden Markov model
Hz	hertz
GEE	Generalized Estimating Equation
GLM	Generalized Linear Model
JAX	Jacksonville
kHz	kilohertz
LTSA	Long-term Spectral Average
m	meter(s)
MARU	Marine Autonomous Recording Unit
MFA	mid-frequency active
min	minute(s)
OB	Onslow Bay
QIC	quasi-likelihood information criterion
ROCCA	Real-time Odontocete Call Classification Algorithm
SNR	signal-to-noise ratio
t.p.m.	transition probability matrix
UO	unidentified odontocete
VIF	Variance Inflation Factor
μs	microseconds

1. Introduction

Passive acoustic monitoring using autonomous recorders deployed on the seafloor is an effective method for long-term monitoring of marine mammals (Mellinger et al. 2007, Van Parijs et al. 2009). Autonomous recorders have been used to investigate the distribution, occurrence, and acoustic behaviors of a variety of marine mammals in diverse habitats and geographic locations (Clark et al. 2002, Clark and Clapham 2004, Baumgartner et al. 2008, Johnston et al. 2008, Sousa-Lima et al. 2013). Recently, researchers have analyzed autonomous recorder data to investigate the effects of anthropogenic noise, such as seismic airguns and sonar, on the calling behavior of baleen whales (Nieukirk et al. 2004, Di Iorio and Clark 2010, Castellote et al. 2012, Melcón et al. 2012, Risch et al. 2012). However, there have been only a few studies focused on using remote or autonomous recorder data to examine the effects of anthropogenic noise such as sonar on the acoustic behaviors of odontocetes (McCarthy et al. 2011, Tyack et al. 2011).

Changes in vocal behavior in response to anthropogenic noise have been studied in several species of odontocetes. For example, beluga whales (*Delphinapterus leucas*), killer whales (*Orcinus orca*) and Pacific humpback dolphins (*Sousa chinensis*) have been documented to change call rates and time-frequency characteristics of their calls in response to vessel noise (Au et al. 1985, Lesage et al. 1999, Van Parijs and Corkeron 2001, Foote et al. 2004). Much less is known about the behavioral responses of odontocetes to mid-frequency active (MFA) sonar. Rendell and Gordon (1999) reported that long-finned pilot whales (*Globicephala melas*) increased whistling rates during and after exposure to military sonar signals. DeRuiter et al. (2013) analyzed acoustic data collected from DTAGs during controlled-exposure experiments using playbacks of MFA sonar and found that false killer whales (*Pseudorca crassidens*) and melon-headed whales (*Peponocephala electra*) increased whistling rates and appeared to mimic MFA signals after exposures. Based on these results, DeRuiter et al. (2013) suggested that vocal responses may be one of the dominant types of response to acoustic stimuli for delphinids. This may be because these highly social cetaceans rely on group awareness and group defense to alert others to perceived threats by communicating in the form of whistles and other acoustic signals.

In a study designed to examine marine mammal vocal behavior before, during and after MFA sonar exercises, acoustic recordings were made off Jacksonville, Florida (JAX), and Onslow Bay, North Carolina (OB), using seafloor deployed Marine Acoustic Recording Units (MARUs). In a preliminary analysis of the MARU data recorded off JAX (Norris et al. 2012), the probabilities of detecting sounds produced by marine mammals in the presence and absence of sonar were calculated for several species of cetaceans (minke whales, *Balaenoptera acutorostrata;* sperm whales, *Physeter macrocephalus;* and delphinids). Delphinid species were analyzed as a single group instead of as individual species because most whistles from odontocetes are difficult to classify to species level without conducting detailed analysis using trained classification algorithms (Oswald et al. 2013). In the Norris et al. (2012) study, the results of the probability analysis for these taxonomic categories were indeterminate, likely because the analysis was classified to species-groups (e.g., delphinids) rather than the species level. For example, if one species responded to sonar by increasing its level of acoustic activity and another responded by decreasing its activity, then the overall response would have indicated little or no change in acoustic behavior. Therefore, it is important to identify acoustic detections to the lowest taxonomic level possible before investigating the effects of sounds on acoustic behavior patterns. In this study, we used the dolphin whistle classification algorithm 'Real-time Odontocete Call Classification Algorithm' (ROCCA; Oswald et al. 2013) to classify dolphin whistles before additional analyses were conducted.

Calculating the probability of detecting sounds produced by marine mammals in the presence and absence of sonar has the potential to reveal changes in vocal behavior that occur in response to sonar. It is important to recognize, however, that differences in the probabilities of detecting an acoustic encounter do not necessarily imply changes in acoustic behavior. For instance, a decrease in the probability of detecting an encounter may mean that animals are less acoustically active. It could mean that animals have moved out of the area and their sounds are no longer detectable. It could also be a result of increased background noise. Hence, more detailed analyses are necessary in order to understand responses and the causes of these responses. Currently, no widely used or generally accepted statistical methods are available for examining these types of issues regarding detection of bioacoustics encounters. The main goal of this project, a collaboration between Bio-Waves, Inc., and St. Andrews University, is to develop statistical methods that can be used to analyze acoustic behaviors of delphinids before, during and after MFA sonar events.

There were several objectives in this effort. The first objective was to classify delphinid detections to the lowest taxonomic level possible using the Atlantic delphinid whistle classifier version of ROCCA (Oswald 2013). This classification analysis allowed the relationships between vocal behavior and the presence of sonar to be examined on a species-by-species (or species-group, as sample size and classification results allowed) basis. The second objective of this project was to detect MFA sonar events and characterize individual sonar pings. This was accomplished using an automated detection and measurement algorithm called SonarFinder (Bio-Waves 2013). The outputs of ROCCA and SonarFinder were then provided to researchers at St. Andrews University who performed several statistical analyses.

The third objective of this study was to investigate potential statistical analyses to address the following questions:

1. Does the probability of detecting delphinid acoustic encounters change during or after periods in which MFA sonar is being broadcast?
2. Does the probability of detecting whistles, clicks or buzzes within a delphinid acoustic encounter change in during or after periods in which MFA sonar is being broadcast?
3. Given that a whistle encounter is detected, do whistle characteristics change during or after periods in which MFA sonar is being broadcast?

In this report, we describe the several different analytical methods that can be used to address these questions, discuss the limits of possible inference, and provide examples of the types of answers that can be provided by analyzing the JAX and OB MARU datasets as an example. The results of this work provide a framework for developing tools that can be used in many locations and situations to understand the potential effects of MFA sonar on marine mammal acoustic behavior.

2. Methods

2.1 Recording Methods

Acoustic data were recorded using non-synchronized, sparsely distributed sets of MARUs deployed off JAX and OB (**Figure 1**). The MARUs were configured to record using two sampling rates: 32 kilohertz (kHz) (referred to as 32-kHz MARUs) and 2 kHz (referred to as 2-kHz MARUs). Because of the relatively high frequencies (typically above 10 kHz) produced by whistling delphinids, only data from the 32-kHz MARUs were used in this analysis. There were two deployments of six 32-kHz MARUs each off JAX, using the same locations in both deployments. The first deployment was in fall (13 September–4 October 2009), and the second deployment occurred in winter (4–26 December 2009). Two of the six MARUs were deployed in shallow water (44–46-meter [m] depth), three were deployed in water of medium depth (183 m), and one was deployed in deep water (305 m). In addition to the JAX deployments, there was a deployment of five MARUs in OB in the summer (6–27 July 2008). Two of the five OB MARUs were deployed in shallow water (64–73 m), one was deployed at medium depth (236 m), and two were deployed in deep water (366 m).

2.2 Logging Delphinid Acoustic Encounters

A delphinid acoustic encounter was defined as a continuous portion of a recording with no more than a 30 min gap between sounds produced by delphinids. When the interval between the detections of sounds exceeded 30 min, a new encounter was delineated. Signals were defined as one of three mutually exclusive sound types: 1) whistles, 2) clicks or 3) burst pulses. Whistles were defined as tonal (i.e., narrow-band) sounds that usually included some level of frequency modulation. Clicks were defined as broad-band impulsive sounds of relatively short duration (< 500 microseconds [μs]). Burst pulses (or buzzes) were defined as broadband click trains with very short inter-click intervals, so that the individual clicks were not discernible to the human ear.

For the purposes of statistical analyses, delphinid acoustic encounters that occurred within 24 hours of a sonar event were logged in greater detail (sub-logged) using a 1 min definition for the gap between events in addition to the 30 min time definition. In this case, a new sub-event was defined when at least 1 min elapsed between sequential delphinid sounds. This finer-resolution allowed for a more detailed analysis of the temporal relationship between delphinid acoustic encounters and the presence of MFA sonar. The start and end times, type of sound (i.e., whistles, clicks, or buzzes), and sonar event condition (i.e., 24 hours before sonar, during sonar, 24 hours after sonar, or between sonar events) were documented for each encounter and sub-event. Delphinid encounters were logged using the 1 min resolution for JAX deployment 1 and OB data. The JAX deployment 2 data were not included in the statistical analysis due to the brief and sporadic nature of sonar events in these recordings.

Acoustic data were reviewed using Triton software (Wiggins 2007) to create long-term spectral averages (LTSAs). When a delphinid acoustic encounter was detected, the start and end times for that encounter were logged in Triton and saved to a Microsoft Excel spreadsheet. Because there were no visual observations associated with the MARU recordings, it was not possible to validate that acoustic encounters were independent (i.e., it was not possible to differentiate when one group, or school, of acoustically active animals left the area being monitored and another entered).

2.3 Whistle Classification

We used the ROCCA module in the software package PAMGuard (www.pamguard.org; Gillespie et al. 2008, Oswald et al. 2013) to classify each delphinid acoustic encounter to species. Only encounters that contained at least 10 whistles with moderate to good signal-to-noise ratios (SNRs) (i.e., at least 3 decibels [dB]) were included in the analyses. Signal-to-noise was computed using Ishmael software (Mellinger 2001), a Matlab-based acoustic analysis software program. If an encounter contained more than 30 whistles, an algorithm written in R software (Version 2.15; The R Foundation for Statistical Computing 2012) was used to randomly select 30 whistles for analysis. Each whistle included in the analysis was automatically saved by ROCCA as an individual audio file in .wav format. To extract time-frequency contours from whistles, the analyst traced contours on ROCCA's spectrographic display using a computer touch-pad. ROCCA automatically measured 50 variables from each extracted contour, including duration, frequencies (e.g., minimum, maximum, beginning, ending, and at various points along the whistle), slopes, and variables describing shape of the whistles (e.g., number of inflection points and steps; see **Appendix A** and Barkley et al. 2011 for a complete list and description of variables measured).

A random-forest classifier within ROCCA was used to identify individual whistles and delphinid acoustic encounters to species. A random forest is a collection of decision trees grown using binary partitioning of the data. Each binary partition of the data is based on the value of one feature (in this case, a whistle feature; Breiman 2001). The goal for each split is to divide the data into two nodes, each as homogeneous as possible (i.e., containing whistles from the smallest number of species possible). Randomness is introduced into the tree-growing process by examining a random subsample of all of the features at each node. The feature that produces the most homogeneous split is chosen at each partition. When whistle features are analyzed using a random forest, each of the trees in the forest produces a species classification. Classifications are then tallied over all trees and the whistle is classified as the species that received the highest proportion of classifications. In addition to classifying individual whistles, delphinid acoustic encounters were classified based on the number of tree classifications for each species, summed over all of the whistles that were analyzed for that encounter.

The number of tree classifications for the predicted species was used as a measure of the certainty of the classification. It was assumed that if a greater percentage of trees classified the whistle as a particular species, then that classification had a higher degree of certainty. Based on this assumption, a 'strong whistle threshold' was defined. If the percentage of trees that classified the whistle as a particular species was greater than this strong whistle threshold, the whistle was considered strongly classified, or simply 'strong' (Oswald et al. 2011). If the percentage of trees that classified the whistle as a particular species did not exceed the strong whistle threshold, then the classification was considered unreliable and the whistle was labeled as 'ambiguous.' If all of the whistles within a single acoustic encounter were labeled as ambiguous, then that acoustic encounter was also classified as ambiguous. In general, as strong whistle threshold increases, correct classification scores also increase (Oswald et al. 2013). However, as the strong whistle threshold increases, the number of whistles that can be classified decreases, eventually resulting in entire acoustic encounters being classified as ambiguous. In this study, a strong whistle threshold of 60 percent was used in order to maximize correct classification scores while minimizing the number of encounters that were labeled as ambiguous.

The random-forest model used to analyze the MARU data was a two-stage model trained using whistles recorded from single-species schools in the Northwest Atlantic Ocean. A two-stage model was used because it resulted in much higher correct classification scores than a one-stage model that classified whistles directly to species (Oswald 2013). Five species were included in the model: bottlenose dolphins (*Tursiops truncatus)*, short-beaked common dolphins (*Delphinus delphis*), striped dolphins (*Stenella coeruleoalba*), Atlantic spotted dolphins (*S. frontalis*) and short-finned pilot whales (*Globicephala macrorhynchus*). The two-stage model first classified whistles to one of three categories: small delphinids (including common and striped dolphins), medium-sized delphinids (including bottlenose and spotted dolphins) and pilot whales. Whistles within each category were then classified to species in stage two (**Figure 2**).

The recordings used to train the random-forest classifier had varying sampling rates from the MARU data. Some of the training data had a sampling rate of 48 kHz and other training data had a sampling rate of 198 kHz, whereas the MARU data had a sampling rate of 32 kHz. The MARU sampling rate resulted in an effective bandwidth of 16 kHz, which is not sufficient to capture the entire fundamental frequency contour of some whistles (Oswald et al. 2004). Using data that contains frequencies that are not captured in the MARU recordings in the classifier training set could affect classification results. To address this issue, any whistle contour that extended above 16 kHz was removed from the training dataset so that the classifier also was constrained to an effective ‘bandwidth’ of 16 kHz. When the 16 kHz bandwidth classification model was evaluated using a test dataset of visually validated recordings, 76 percent of whistles (n = 638) and 81 percent of encounters (n = 113) were correctly classified (**Table 1)**.

1.1 Mid-frequency Active Sonar

Analysis of MFA sonar was conducted using the software program SonarFinder (Bio-Waves 2013). This Matlab-based program was designed to automatically detect sonar pings and measure acoustic variables that characterize them. SonarFinder was used to detect individual sonar pings within sonar events that were initially detected manually by bioacoustic analysts using Triton software (Wiggins 2007) in a previous study (Norris et al. 2012). An MFA sonar event was defined as a series of sonar pings with no longer than 30 min elapsing between individual pings. Each MFA sonar event was given a unique identification number containing the sample rate, deployment number, site and event number. The .wav files for each event were placed into separate folders for later batch processing using SonarFinder.

SonarFinder uses three stages to detect potential sonar pings. Each stage requires user-defined thresholds. In the first stage, potential pings are detected based on a user-defined amplitude threshold. In the second stage, clipped signals (i.e., recorded signals that result in distortions of the true signal when the amplitude exceeds the dynamic range of the recording system) are detected and eliminated. In the third stage, SonarFinder narrows the remaining detections by eliminating some of the false positives (the remaining false positive detections were removed by manually inspecting the spectrogram of each candidate detection). See Bio-Waves (2013) for more detail on SonarFinder's three detection stages. Thresholds for the three SonarFinder stages were determined individually for each site and deployment. The three thresholds were adjusted to maximize true detections and minimize missed detections. Thresholds were selected so that no more than 20 percent of pings were missed by SonarFinder.

Once thresholds had been determined for each site and deployment, SonarFinder was used to batch process all MFA sonar events in the JAX deployment 1 and OB datasets. SonarFinder was not run on JAX deployment 2 because the sonar events in this deployment were considered too short and sporadic to be useful in the statistical analysis. SonarFinder was used to automatically detect and measure features from each ping. These measurements were saved to a Microsoft Excel spreadsheet.

The features documented by SonarFinder included: date/time when the ping occurred, and measures of frequency bandwidth, peak frequency, minimum frequency, maximum frequency and sound pressure level (Bio-Waves 2013). Each ping was categorized according to its peak frequency and duration. Frequency labels included: "Type 1" (for pings < 3,999 Hertz [Hz]), "Type 2" (for pings between 4,000 Hz and 6,999 Hz) and "Type 3" (for pings > 7,000 Hz). Duration labels included: "Short" (for pings < 1.49 seconds in duration), "Medium" (for pings between 1.50 and 3.99 seconds in duration), and "Long" (for pings > 4.00 seconds in duration). SonarFinder also averaged ping measurements over all of the pings in a single MFA sonar event and measured average inter-ping interval and ping repetition rate for each event. These measurements were output to a second Microsoft Excel spreadsheet.

3. Results

3.1 Delphinid Acoustic Encounters

A total of 1,259 delphinid acoustic encounters were logged from JAX (deployments 1 and 2) and OB (**Table 2**). The greater number of encounters was logged from JAX deployment 1 (n = 550) and fewer encounters were logged from OB (n = 265). For JAX deployment 1 and OB, acoustic encounters logged at 30 min resolution were sub-logged into 1,738 and 933 1 min resolution encounters, respectively. JAX deployment 2 was not sub-logged because these data were not used in the statistical analyses.

3.2 Whistle Classification

3.2.1 OB

A total of 100 delphinid acoustic encounters were analyzed using ROCCA. Numerous encounters (n = 165) were not included in the ROCCA analysis because they contained few or no whistles, or because the whistles were not of sufficient quality (i.e., < 3 dB SNR or overlapped significantly with other whistle contours) for contour extraction. All encounters analyzed were classified as either short-beaked common dolphin (n = 48), striped dolphin (n = 37), or short-finned pilot whale (n = 15) (**Figure 3**). No encounters were classified as bottlenose dolphin or Atlantic spotted dolphin.

3.2.2 JAX Deployment 1

A total of 158 delphinid acoustic encounters were analyzed using ROCCA. A large number of encounters (n = 392) were not included in the ROCCA analysis because they contained few or no whistles, or the whistles were not of sufficient quality for contour extraction. All encounters that were included in the ROCCA analysis were classified as either striped dolphin (n = 74), short-beaked common dolphin (n = 54), or short-finned pilot whale (n = 30; **Figure 4**). No encounters were classified as bottlenose or Atlantic spotted dolphins.

3.2.3 JAX Deployment 2

A total of 55 delphinid acoustic encounters were analyzed using ROCCA. Many (n = 389) were excluded from the ROCCA analysis because they contained few or no whistles, or the whistles were not of sufficient quality for contour extraction. All encounters that were included in the ROCCA analysis were classified as either striped dolphin (n = 21), short-finned pilot whale (n = 18), or short-beaked common dolphin (n = 16; **Figure 5**). No encounters were classified as bottlenose or Atlantic spotted dolphins.

3.3 Mid-frequency Active Sonar

3.3.1 OB

A total of 72 sonar events consisting of 158.5 hours were detected by SonarFinder. The mean duration of sonar events was 2.2 hours, with a standard deviation of 3.3 hours. There were 30,403 sonar pings detected during these events. Sonar events were plotted with delphinid acoustic encounters, both for all delphinid acoustic encounters combined and by individual species (ROCCA-classified encounters) to indicate when sonar and delphinid acoustic encounters overlapped (**Figures 6-9**).

3.3.2 JAX Deployment 1

A total of 58 sonar events comprising 421.2 hours were detected by Sonar Finder. The mean duration of sonar events was 7.3 hours, with a standard deviation of 11.3 hours. There were 31,826 sonar pings detected during these events. Sonar events were plotted with delphinid acoustic encounters, both for all delphinid acoustic encounters combined and by individual species (ROCCA-classified encounters) to show when sonar and delphinid acoustic encounters overlapped (**Figures 10-13**).

3.3.3 JAX Deployment 2

During a previous analysis (Norris et al. 2012), 63 sonar events comprising 95.5 hours of sonar were logged using Triton for JAX deployment 2. Sonar events were plotted with delphinid acoustic encounters, both for all delphinid acoustic encounters combined and by individual species (ROCCA-classified encounters) to show when sonar and delphinid acoustic encounters overlapped (**Figures 14–17**).

4. Discussion

4.1 Species Identification of Delphinid Vocalization Encounters

All delphinid vocalization encounters that were classified to species using ROCCA were classified into one of only three species: short-finned pilot whales (20 percent), striped dolphins (42 percent) or short-beaked common dolphins (38 percent; **Figures 3-5**). Both striped dolphins and short-finned pilot whales were expected to occur in the study areas and generally occur beyond the continental shelf (CETAP 1982, Au and Perryman 1985, Selzer and Payne 1988, Olson and Reilly 2002, Payne and Heinemann 1993), although sightings of striped dolphins in these areas have been extremely rare (CETAP 1982, Waring et al. 2013). Short-beaked common dolphins, however, generally are distributed along the shelf break from Cape Hatteras to Nova Scotia and are considered very rare south of Cape Hatteras (CETAP 1982, Selzer and Payne 1988, Gaskin 1992, Waring et al. 2013). As such, the relatively high number of encounters classified as striped or common dolphins was unexpected. Another unexpected result was that no encounters were classified as bottlenose or Atlantic spotted dolphins. Bottlenose and Atlantic spotted dolphins were the two species most frequently observed during recent aerial and boat-based line-transect visual surveys off Florida and North Carolina (DoN 2008, Halpin et al. 2009, Hodge et al. 2013). Therefore, we expected that a large proportion of delphinid acoustic encounters would be classified as either Bottlenose or Atlantic spotted dolphins.

There are several possible explanations for these results. First, it is important to note that the acoustic encounters that were classified using ROCCA represent less than half of the total delphinid acoustic encounters (40, 14 and 38 percent of encounters for JAX deployment 1, 2, and OB, respectively). Although short-beaked common dolphins and striped dolphins made up the majority of the classifications, they did not necessarily represent the majority of the total encounters. The delphinid encounters that were excluded in the classification analysis were omitted because they either contained few or no whistles, or the SNR was too low for reliable whistle contour extraction. The paucity of whistles did not necessarily represent a lack of whistling by the dolphins. The dolphins may have simply been too far away from the recorders to obtain good signals. Poor quality whistles may also be caused by propagation effects (such as a ‘surface duct,’ which can trap sounds near the ocean surface), as dolphins tend to spend a significant proportion of their time in the surface waters and the MARUs were located on the sea floor at depths ranging from 44 m to 366 m. The lack of whistles may also reflect the behavior of the dolphins. For example, the habitat in which the MARUs were deployed might be used more for foraging than for socializing, and the predominance of clicks versus whistles in the recordings may reflect this.

In addition, the majority of encounters included in the classification analysis were recorded on deep (305–366 m) or medium-depth (183–236 m) deployments (e.g., 70 percent of encounters for JAX, 86 percent of encounters for OB). Almost all sightings of Atlantic spotted dolphins in Onslow Bay have occurred over the shelf (Hodge et al. 2013), so the deep and medium-depth MARU deployments may be located outside the expected range for this species (although they would still be expected on the shallow MARUs). In addition, a more diverse assemblage of delphinid species has been observed beyond the shelf break in the OB study area (Hodge et al. 2013). Short-beaked common and striped dolphins may be more common in the JAX and OB study areas than previously thought based on visual observations.

Finally, it is likely that some of the encounters classified as striped and short-beaked common dolphins were misclassifications. Although the ROCCA classifier performed well when it was ground-truthed using visually validated test data from the Northwest Atlantic Ocean (81 percent of acoustic encounters were correctly classified, n = 113; Oswald 2013), both the training and testing data were collected near the surface (< 20 m depth) using towed hydrophone arrays (Oswald 2013). It is possible that the classifier is using features that are susceptible to propagation effects or instrumentation effects (i.e., differences in whistle features caused by the sound propagation). For example, one of the most important variables in the classifier was duration (Oswald 2013). This variable is susceptible to propagation effects because signals can be ‘smeared’ in time via multi-path signal arrivals for low-frequency components, or shortened due to attenuation for high-frequency components of signals. If whistles recorded at depth have different characteristics than those recorded at the surface, this could influence the performance of classifiers. Efforts should be made to collect visual data and recordings at the surface in conjunction with autonomous recordings. These data could be used to ground-truth the performance of ROCCA’s Atlantic classifier on data collected at different depths and gain insight to why so many encounters were classified as striped and common dolphins in this study.

Perhaps the most important consideration is that the Atlantic version of the ROCCA classifier used in this study includes only five delphinid species that occur in the northwest Atlantic Ocean and are known to whistle. Whistling species such as pantropical spotted dolphin (*Stenella attenuata),* rough-toothed dolphin (*Steno bredanensis*), Clymene dolphin (*Stenella clymene),* long-finned pilot whale (*Globicephala melas),* false killer whale, Risso’s dolphin (*Grampus griseus*) and Atlantic white-sided dolphin (*Lagenorhynchus acutus)* are known to occur in the northwest Atlantic Ocean (Palka 2012, Waring et al. 2012) but are not currently included in the ROCCA classifier because the available data were insufficient to reliably train the classifier at the time the Atlantic classifier was developed (Oswald 2013). If any of the delphinid vocalization encounters included in our analysis contained whistles from any of these species, they would be misclassified as one of the five species in the classifier and so some of the encounters classified as striped and common dolphins may actually have been another species.

4.2 SonarFinder

Ideally, an automated detection method should find only the signals it was intended to detect (true positives) and reject those it was not (true negatives) without missing signals (false negative) nor incorrectly identifying signals (false positives). But in reality the sensitivity of the detector is configured to optimize the trade-off among true positives, false positives and false negatives, where optimization is project dependent (Mellinger et al. 2007). The desired outcome of this tradeoff will influence the parameterization of automated detectors, because in some contexts it is more important that true signals are not missed (e.g., during real-time monitoring for impact mitigation), whereas in other situations it is more important to eliminate most false detections even if some true signals are missed (e.g., when using automated detection methods for density estimation). We used the Matlab program SonarFinder to provide data on the occurrence and characteristics of sonar to allow an assessment of delphinid vocal behavior in relation to sonar. As such, it was important to maximize true positive detections while minimizing missed detections. To accomplish this goal, SonarFinder’s performance was optimized to accurately detect 80 percent of medium-intensity pings (SNR >2 dB), which we decided was the optimal tradeoff for minimizing missed detections and producing a manageable number of false positives. Although this approach resulted in a small number of missed pings, there were many false positive detections. These false positives were caused primarily by ‘self-noise’ due to hard drive ‘spin-up’ within the individual MARU devices, which occurred as frequently as every 11 seconds for some periods. This spin-up noise is a frequency-modulated signal that is similar in some aspects (e.g., the frequency band) to sonar signals. A few false positive detections also occurred due to boat noise and biological noise, but these were less common than the hard-drive spin-up false positives. False positives were removed by manual review of all sonar detections by data analysts, a time-consuming process. Changing the parameters within SonarFinder to further reduce the number of missed pings would have resulted in an unreasonable number of false positives that would have had to be reviewed manually by data analysts. Future efforts should focus on obtaining ‘clean’ (i.e., high-fidelity) recordings without sources of instrumentation noise, especially in the band of concern.

This page intentionally left blank.

5. Modelling Approach Using Generalized Estimating Equations

5.1 Methods

Statistical analyses were conducted on delphinid acoustic sub-event data recorded during sonar exercises (see definition in the next section), as well as the periods 24 hours before and after these exercises for JAX deployment 1 and the OB deployment (**Table 3**).

5.1.1 Defining Control Periods

To identify potential changes in delphinid vocal behavior in response to MFA sonar, observations from periods during or after sonar were compared to observations from a control period before sonar. We used a 24-hour period before the commencement of each sonar exercise as the control period. Every acoustic sub-event occurring during these periods was labeled as ‘before.’ This 24-hour period represented a compromise between trying to capture a potentially ongoing effect after sonar and to avoid introducing additional variability by extending too far beyond the time of the sonar exercise, while keeping a balance between the periods included before and after. For each site, a sonar exercise was defined to include all the sonar pings occurring consecutively with no gap of longer than 48 hours. When a gap was longer than 48 hours, the subsequent sonar pings were attributed to a different exercise. Acoustic sub-events were labeled as ‘during’ when sonar pings were recorded simultaneously at any time during the acoustic sub-event. Acoustic sub-events were labeled as ‘between’ when they occurred between sonar pings within an exercise. Acoustic sub-events occurring in the 24 hours directly after a sonar exercise were labeled as ‘after’. These definitions of ‘before’, ‘during’, ‘between’ and ‘after’ are illustrated in **Figure 18**.

5.1.2 Defining the Response Variable

The research questions of interest for this study were:

1. Does the probability of detecting delphinid acoustic encounters change during or after periods in which MFA sonar is being broadcast?
2. Does the probability of detecting whistles, clicks or buzzes within a delphinid acoustic encounter change during or after periods in which MFA sonar is being broadcast?
3. Given that a whistle encounter is detected, do whistle characteristics change during or after periods in which MFA sonar is being broadcast?

To investigate these questions, we used a modeling approach where a response variable was related to explanatory covariates to describe and investigate the pattern in the response. Although the available covariates were the same for each question, the type of response varied among them. These are detailed in the following sections.

5.1.2.1 1 MIN PRESENCE OF ACOUSTIC ENCOUNTERS MODELS

In the following, we refer to this type of model as the presence model for simplicity. For the presence models we created data records for each site, which consisted of 1 min effort segments encompassing the 24 hours before a sonar exercise (n = 1440 segments), the time during a sonar exercise (variable number of 1 min segments) and the 24 hours after a sonar exercise (n = 1440 segments). For each 1 min segment, the detection or non-detection of an acoustic encounter was recorded as a binary variable (1 for presence, 0 for absence). Hence, we take the output of the detector as a proxy for acoustic presence or absence. In addition, each effort segment was labeled as one of four conditions (‘before,’ ‘during,’ ‘between’ or ‘after’) according to when it occurred in relation to the sonar exercises at the same site, using the definitions for these from Section 5.1.1. Those effort segments that corresponded with periods of the sonar exercise were either labeled as ‘during’ if any sonar pings had been recorded during the respective 1 min segment and as ‘between’ if not. For this analysis we assumed a binomial error structure and modeled the predictor with factor, linear and smooth terms via a logit-link function.

To eliminate potentially confounding issues in responses among different species, we conducted the analysis separately for two different encounters: 1) pilot whales and 2) striped dolphins, common dolphins and unidentified odontocetes combined (which we refer to hereafter as the DEUO species encounter). There are potentially confounding issues with respect to combining multiple species in a classification. One of these is that one species might react to sonar differently (e.g. by vocalizing less frequently) than another species (e.g. that might react by vocalizing more frequently). The overall effect might then not be detectable (i.e. because they offset each other) when combining whistles from these two species in a single analysis. However, it was not possible to analyze striped and common dolphins separately because the classifier could not reliably distinguish between these two species.

5.1.2.2 PRESENCE-OF-SIGNAL-TYPE-GIVEN-ACOUSTIC-ENCOUNTER (PSTGAE) MODELS

For the PSTGAE models, we analyzed the delphinid acoustic sub-events, where each sub-event represented a separate observation, regardless of length of the sub-event. Of the 2,673 sub-events included in this analysis, 57.5 percent were less than 1 min, and 71.7 percent were less than 2 min. However, some were as long as 2 hours, with 0.6 percent being longer than 1 hour. For each of the acoustic signal types—whistles, clicks and buzzes—the response was then defined as a binary variable similar to the presence models in the previous section (1 or 0 to indicate presence or absence, respectively, of the respective signal type within each acoustic sub-event). As for the presence models, we conducted each analysis for two different encounters to eliminate potentially confounding issues in responses among different species. These encounters were: 1) pilot whales and 2) and the DEUO species encounter.

5.1.2.3 WHISTLE-CHARACTERISTICS MODELS

For this approach, individual whistles represented the observation units for analysis. These were extracted from those detections that were classified to species using ROCCA (see **Section 2.3** above for details). In total, we analyzed 2,904 whistles—670 from pilot whales, 1,178 from common dolphins and 1,056 from striped dolphins. We analyzed the whistles from pilot whale detections separately from striped and common dolphins because the identification of pilot whale encounters using ROCCA has a small error rate (see above, **Section 2.3**). As identification of common and striped dolphins had a relatively large error rate, we lumped whistles from these two species together for the analysis.

The whistle characteristics included in these analyses were: maximum and minimum and the standard deviation of frequency, duration, mean slope of the whistle, mean of the positive slope values in the whistle contour, mean of the negative slope values in the whistle contour, percent positive slope and percent zero slope. Other whistle characteristics, including mean frequency and percent negative slope, were excluded due to collinearity with other characteristics. The observed values of the whistle characteristics were normalized by subtracting the mean value and dividing by the standard deviation for each vector of observed whistle characteristics (using the scale function of the base package in R). For each whistle, the information from these characteristics was combined into one response variable using Mahalanobis distances, D_{Mi} (DeRuiter et al. 2013). These were calculated for the *i*th whistle using:

$$D_{Mi}(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^T \boldsymbol{S}_c^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)} \, ,$$

where $\boldsymbol{x}_i$ represents a vector of observed characteristics for the *i*th whistle, $\boldsymbol{\mu}_c$ represents a vector of means for each whistle characteristic obtained from a set of control whistles (see below for a definition of control whistles) and $\boldsymbol{S}_c$ is the covariance matrix of the control whistles). (T is a mathematical symbol for transposing a vector.)

For defining the control whistles, two main strategies may be followed. For one strategy, all whistles that occurred in the 24 hours before a sonar exercise are included in the control group. Each whistle, including those from ‘before,’ ‘during,’ ‘between’ and ‘after’ sonar, is then compared to this control group. An alternative strategy is to define the control whistles individually for each whistle as a set of *n* whistles that are preceding the respective whistle. The benefit of the second strategy is that it may reveal sharper contrasts in the case of a short-term change. However, the difficulty for this strategy is choosing *n* such that it is biologically meaningful. Here, we applied the first strategy, and used these Mahalanobis distances as proxies for potential response intensity following the example of DeRuiter et al. (2013); the distances were then related to explanatory covariates.

5.1.3 Explanatory Covariates

All covariates that were available for the analyses are listed in **Table 4**. The covariate *Sonar* was a factor covariate with four levels: ‘before,’ ‘during,’ ‘between’ and ‘after’ a sonar exercise, where level ‘before’ represented the base level to which the other levels were compared. The covariate *Anysonar* was a binary variable which was set to 1 for all records where *Sonar* equaled the condition ‘during;’ otherwise it was set to 0. *Anysonar* was considered only if *Sonar* was not significant. If *Sonar* was significant, *Anysonar* only entered the model as an indicator variable in an interaction term with covariates pertaining to sonar measurements. As an indicator variable it switched on the respective covariate related to sonar measurements for those observations where *Sonar* equaled ‘during’ or else set them to zero in the case that the covariate could not be observed. This is different from setting the covariate to zero itself as it influences the covariance matrix of the model.

Sonarlag was used to represent the time lag since the last recorded sonar ping. When using 1 min segments for the presence models, this covariate was measured in 1 min segments; otherwise it was measured in seconds. This covariate entered the model as an interaction with an indicator variable, *I(Sonarlag)*, which switched *Sonarlag* off for those effort segments/acoustic sub-events before the first recorded sonar ping at each site (i.e., when it could not be determined how much time had passed since the last sonar occurrence before the deployment of the MARUs).

The covariate *Site* was included to determine if the potential differences in any of the responses were due to differences in the MARU sites (e.g., different ocean depth, propagation conditions, differences in the instruments, or different time of the year as recordings at the different sites were made during different times of the year—see **Table 3**). The covariate *Location* was only tested if *Site* was not significant in the model.

Several types of sonar pings were classified during this study (see **Section 2.4**). Because it was possible, even likely, that more than one sonar type occurred in any given 1 min effort segment, the presence/absence of each type was recorded for each sonar event. Two other covariates related to individual sonar pings were included, sound pressure level and peak frequency. All other covariates related to sonar from **Table 4** represent summary statistics of sonar events (see Section 2.4 for definition of sonar event).

5.1.4 Modeling Approach Using Generalized Estimating Equations

Generalized estimating equations (GEEs) are an extension of generalized linear models (GLMs) that, similar to GLMs, allow the specification of different distributions for the error structure (e.g., binomial or Poisson). In contrast to GLMs, GEEs also allow modeling of various correlation structures for the errors. In this respect GEEs have the advantage over mixed-effect models because they allow unbiased estimation of regression coefficients despite possible mis-specification of the correlation structure (Ghisletta and Spini 2004). Hence, GEEs are most useful when the main interest lies in examining the relationship between the response and the explanatory variables (as opposed to the correlation structure), as was the case for this study. For this study, we expected correlation in the observations, regardless of the type of response. Another advantage of GEEs is that they estimate the dispersion parameter and therefore accommodate overdispersed data.

GEEs may be fitted in R using the *geeglm* function of the *geepack* package (Halekoh et al. 2006). As with the *glm* function, smoothing terms can be added using the *bs* function of the splines package. Using splines allows 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 limited the flexibility of the relationship in our models by using the default settings of the *bs* function which fits polynomial splines with only three degrees of freedom. A polynomial spline can be thought of as a smooth 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 smooth function with one maximum and one minimum. More flexibility could be achieved by including knots which allow the relationship to be more ‘wiggly’ than a polynomial spline. This, however, was beyond the scope of this study.

For the MARU dataset, we used the default correlation structure ‘independence’ where correlated observations were grouped using a block identifier (argument *id* from the *geeglm* function). Each block consisted of consecutive observations. These were the consecutive 1 min segments, consecutive delphinid acoustic sub-events, or consecutive whistles, respectively, for the presence models, PSTGAE models and the whistle-characteristics models, where the size of the blocks was determined using the *acf* function from the *stats* package. This function estimates the autocorrelation between consecutive residuals for various lags. A plot created by the same function displays these estimates by lags, 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 Pearson’s residuals first decayed within confidence bounds around zero autocorrelation. For the independence correlation structure, group 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 group size. Larger *p*-values, in turn, influence which covariates should be retained (see **Section 5.1.4.1**).

5.1.4.1 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 steps are explained in greater detail below.

Model selection for GEEs remains an area of ongoing research with no clear-cut best strategy, particularly when models include smooth terms (such as occurred in this study). The often-used quasi-likelihood information criterion (QIC) under the independence model is considered somewhat equivalent to AIC (Akaike Information Criterion) for GLMs or generalized linear mixed models (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 our 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 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 smooth terms first. If the smooth terms were not significant, we tried adding these as linear terms.

If covariates are collinear, it is possible to retain covariates in the model that otherwise have 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 and Monette 1992).

A further step for selecting the best model included a potential backwards step, which included inspecting partial fit plots for each of the covariates retained in the best fitting model so far. Partial fit plots were created using a modified version of the *runPartialPlots* function from the *MRSea* package, which uses parametric bootstrapping of model coefficients to create confidence intervals around the partial fit. During this step we eliminated covariates that exhibited 95 percent confidence intervals around their partial fits that were wide enough to fit a straight horizontal line within the bounds of the confidence limits through the entire range of observed covariate values. This was an indication that the respective covariate potentially had no effect on the response.

5.2 Results

5.2.1 Presence Models

Sizes for the blocks of the GEEs were identified for the presence models for each species group by visually inspecting the ACF (Auto-Correlation Function) plots in **Figure 19**. The group sizes were 150 and 1,050 for the pilot whale and the DEUO species group, respectively (**Table 5**).

Using a logit-link function, the relationship between the coefficients of **Table 5** and the response can be interpreted as follows: the expected odds (i.e., probability of success p divided by the probability of failure, 1-p) are expressed as the exponent of the predictor, e.g., $log(p/(1-p)) = \beta_0 + \beta_1 x_1 + \dots \beta_k x_k$, where the β_0 represents the intercept and the β_k are the coefficients associated with covariates x_k, respectively. The partial fit plots in **Figures 20** and **21** indicate the relationship between the log of the odds, $log(p/(1-p))$, and the respective covariate. For polynomial splines fitted here, the effect of the three terms are summed which can best be traced with the partial fit plots (**Figures 20 and 21**) rather than looking at the coefficients from **Table 5**.

For the presence models, the covariates retained in the best fitting model included the factor covariate *Site* for both the pilot whale and the DEUO species group models (**Table 5, Figures 20** and **21**). Factor covariates are generally fitted by first defining a base level (usually the first level in numerical or alphabetical order). This base level forms part of the intercept estimate against which other levels are contrasted. For example, for covariate *Site*, the level 2 was absorbed in the intercept. The coefficients of the remaining levels for this covariate represent how these levels contrast against the intercept. For the pilot whale models only levels 154 and 161 were significantly different from the base level. For the DEUO species group, the levels 4, 7, 154 and 161 were significantly different from the base line.

For the DEUO species group, the polynomial spline for *Time* was also retained in the best fitting final model. The partial fit plot for covariate *Time* in the DEUO model indicated that vocalization rates were higher early and late in the day, corresponding to night-time hours, and reached a low in the late afternoon (*Time* was measured in seconds, hence 60,000 seconds corresponded to approximately 4:40 pm) (**Figure 21**).

5.2.1.1 ASSESSING PRESENCE MODEL ASSUMPTIONS

For a binomial GLM, model assumptions include a mean-variance relationship of

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

where p is the probability of success and n is the sample size. Also, model errors are assumed to be uncorrelated. We accommodated potential violation of these assumptions by using GEEs as the model- fitting tool. For the binomial models, the estimates of the scale parameter were near one for both the pilot whale and the DEUO species group models giving indication that the data were likely not overdispersed (**Table 5**).

As described in **Section 5.1.4**, we incorporated a blocking structure to the observations where observations within the same block were allowed to be correlated (see **Table 5**). Taking these block sizes into account, we expected to see no additional pattern in the residuals on scales larger than these blocks. When plotted in order of observation, we expected to see a random pattern of standardized residuals around zero with a constant mean and variance across the range of observed values. To assess whether the distribution of the residuals remained constant across the range of observations, we split the residuals into 20 equally sized bins in the order of observation (i.e., ordered by *Site* first and then by date). For each of these bins, we calculated the mean of the residuals (**Figure 22**). For both the pilot whale and the DEUO species group model, the pattern of residual means across the range of observed values was random with no increase or decrease in variance.

5.2.1.2 ASSESSING PRESENCE MODEL FIT

To assess model fit, we split the fitted values into 20 equally sized bins similar to the previous section, however, this time in ascending order of fitted values. We then calculated the means of the fitted values per bin and plotted these against the mean of the corresponding observed values in **Figure 23**. We expected to see a random pattern around the line of perfect fit. This was the case for both the pilot whale and the DEUO species group models.

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 6** lists the number of correct predictions as well as the falsely predicted presences and absences. Overall, the presence models for pilot whales and the DEUO species group predicted 55 percent and 53 percent, respectively, of all observations correctly. A large percentage of predictions, however, were false positives (44 percent and 41 percent for the pilot whales and DEUO species group, respectively). The large percentage of falsely predicted presences for pilot whales was due to predicting with one factor covariate (i.e., *Site*). For the pilot whale model, 99 percent of all 1 min segments were observed absences. Due to converting those predictions into presences where the fitted value was larger than the overall mean of the fitted values, all 1 min segments from a given site received the same predicted value. This was 1 (presence) for those levels of *Site* that had larger coefficients (levels 2, 4, 7, 9 and 159) and 0 (absence) for the levels of *Site* that had smaller coefficients (see **Table 5** and **Figure 20**). We conclude that our model was able to pick up some variability between the different sites. None of the other available covariates, however, were able to describe the finer-scale forces that drive the underlying pattern.

The pilot whale results were similar to those for the DEUO species group model. For the DEUO model, 94 percent of all 1 min segments were observed absences. For this model, the polynomial spline for *Time* was retained in the final model in addition to the factor covariate *Site*. This improved the predictive power of the model; however, a large percentage of false predictions remained.

5.2.2 PSTGAE Models

We ran three separate PSTGAE models for both the pilot whales and the DEUO species group. These will be referred to as whistle, click or buzz models in the following. Block sizes for the correlation structures of the GEEs were determined using the ACF plots shown in **Figure 24**. These block sizes were one, one and two for the whistle, clicks and buzzes models for pilot whales (**Table 7**). Block sizes were six, six and seven for each of the whistle, click and buzz models, respectively, for the DEUO species group.

As for the presence models from the previous section, a logit-link function was used to relate the explanatory covariates to the response. For pilot whales, each best-fitting model contained the factor covariate *Site* indicating that for each of the signal types the respective presences varied (**Table 10**). Additional covariates retained in the whistle model for pilot whales were the factor covariate *Presence of clicks* and the polynomial spline for *Time*. The negative coefficient for level ‘presence’ of the covariate *Presence of clicks* indicated that when clicks were present within a given delphinid acoustic encounter, the odds of observing presences of whistles were lower (**Table 7** and **Figure 25**). The polynomial spline for *Time* indicated that between midnight (time = 0) and approximately 8:20 am (time = 30,000) the odds of detecting whistles within a delphinid acoustic encounter were increasing and remained relatively high throughout the remainder of the day.

Additional covariates retained in the best fitting click model for pilot whales were the factor covariates *Presence of whistles* (with a negative coefficient for level ‘presence’) and *Presence of buzzes* (with a positive coefficient for level ‘presence’) (**Table 7** and **Figure 26**). Presences of whistles within an acoustic sub-event had a negative effect on the odds of observing clicks while presences of buzzes had a positive effect on the odds of observing clicks.

An additional covariate retained in the best fitting buzz model for pilot whales was the factor covariate *Presence of clicks* with a positive coefficient for level ‘presence’ (**Table 7** and **Figure 27**).

For the DEUO species group, *Site* was again retained in all PSTGAE models, indicating that for each signal type the odds of observing the respective signal type within a delphinid acoustic encounter varied (**Table 8** and **Figures 28 - 30**). Additional covariates retained in the whistle model were the factor covariates *Sonar* and *Presence of clicks*. The significantly positive coefficient for level ‘during’ of covariate *Sonar* indicated that the odds of observing whistles within a delphinid acoustic encounter were higher during sonar events compared to before sonar events. The coefficients of the other two levels, ‘between’ and ‘after’, were not significant. The negative coefficient for level ‘presence’ of the *Presence of clicks* covariate indicated that the odds of detecting whistles decreased when clicks were present in a delphinid acoustic encounter.

Four additional factor covariates were retained in the best click model for the DEUO species group. These were *Presence of whistles* (with a negative coefficient for level ‘presence’) as well as *Presence of buzzes*, *Presence of Type 1-short sonar pings* and *Presence of Type 3-medium sonar pings*—each with positive coefficients for level ‘presence’ (the latter two covariates pertain to the type of sonar ping and its duration—see Section 2.4 for details). Again, negative coefficients indicate a decrease in the odds of detecting clicks within a delphinid acoustic sub-event while positive coefficients indicate an increase in the odds.

For the buzz models for the DEUO species group, additional covariates retained in the best model were the two factor covariates *Sonar* and *Presence of clicks* (**Table 8** and **Figure 30**). For *Sonar*, all coefficients were significantly positive, indicating an increase in the odds of observing buzzes within a delphinid acoustic sub-event for each level ‘during’, ‘between’ and ‘after’ sonar events compared to the 24 hours before. Again, the positive coefficient for level ‘presence’ of covariate *Presence of clicks* indicated an increase in the odds of observing buzzes when clicks were present within the delphinid acoustic encounter.

5.2.2.1 ASSESSING PSTGAE MODEL ASSUMPTIONS

As was done for the presence models (see **Section 5.2.1**), we accommodated potential over-dispersion by using GEEs as the model fitting tool. No over-dispersion was evident for five of the PSTGAE models. For the pilot whale models, scale parameter estimates were < 1 for the whistle model (0.22, SE=0.09) and buzzes model (0.84, SE=1.49). However, for the presence of clicks model, the estimate was 3.32 (SE=114.0, **Table 7**). For the DEUO species group models, the estimates of the scale parameter were all near 1: 1.02 (SE=1.35) for the whistles model, 1.07 (SE=0.54) for the clicks model and 0.98 (SE=0.72) for the buzzes model (**Table 8**).

We also assessed whether the distribution of residuals remained constant across the observed values by splitting the residuals into 20 equally sized bins in order of observation (i.e., ordered by *Site* first, then by date). The mean of the residuals were plotted against the means of the respective fitted values for each bin in **Figure 31**. Here we expected to see a random pattern which was the case for all models, except the whistle model for pilot whales, where variability in mean residuals decreased throughout the observations.

5.2.2.2 ASSESSING PSTGAE MODEL FIT

To assess model fit, we plotted the means of fitted values split into 20 equally sized bins in ascending order against the means of the corresponding observed values for each bin in **Figure 32**. Again, we expected a random pattern with a tight fit around the line of perfect fit (red line in **Figure 32**). This was the case for each of the models.

As was done for the presence models, we also assessed model fit by comparing predictions with observed presences and absences. Again, we converted the fitted values (which correspond to the estimated probability of success) into presences and absences by attributing a predicted presence to the respective observation if the fitted value was larger than the overall mean of fitted values and predicted absences otherwise. For pilot whales, 91 percent, 88 percent and 60 percent of all predictions were correct, respectively, for the whistle, click and buzz models (**Table 9**). For the DEUO species group, 80 percent, 77 percent and 74 percent of predictions were correct, respectively, for the whistle, click and buzz models (**Table 9**).

5.2.3 Whistle Characteristics Models

For the whistle characteristics models, we first divided the whistles into those from pilot whales and those belonging to common and striped dolphins. We then calculated the Mahalanobis distances for each whistle where we used the 24 hours before as the control period. These distances were related to explanatory covariates using GEEs as the model-fitting tool. Here, we only included *Site*, *Time*, *Sonar* and *Sonarlag* in the model selection. Block sizes for both the pilot whale and the common/striped dolphin models were determined as 3 and 12, respectively, using **Figure 33**.

For the whistle characteristics models, we used a gamma error structure with an identity link function. Using the identity link, the response M_D was related to the explanatory covariates x_k via $M_D = ß_0 + ß_1 x_1 + \dots\ ß_k x_k$ where the $ß_k$ term represents the coefficients. Again, both the pilot whale and the common/striped dolphin models contained the covariate *Site* in the best-fitting model (**Table 10** and **Figures 34** and **35**). An additional covariate retained in the best model for the common/striped dolphins was *Sonar*. A significantly positive coefficient for a factor level from **Table 10** indicated an expected increase in M_D, while a significantly negative coefficient indicated an expected decrease in M_D. We conclude that during and after the emission of sonar, M_D increased, i.e., whistle characteristics changed causing an increase in the respective M_D. Further analyses are necessary to determine in which manner these characteristics changed.

5.2.3.1 ASSESSING WHISTLE CHARACTERISTICS MODEL ASSUMPTION

For a gamma GLM, the variance is assumed to equal the squared mean. The estimates of the dispersion parameters for the pilot whale model were < 1, with small standard errors (0.34, SE=0.07, **Table 10**). Hence, there was no evidence for over-dispersion of the pilot whale M_D data. For the common/striped dolphins, the scale parameter was > 1 (4.42, SE=2.16), however, this over-dispersion was accounted for in our model.

For a gamma model, we expected to see a random pattern of Pearson's residuals with a right-skewed distribution across the entire range of observed values. The random distribution of residuals was the case for the pilot whale data (**Figure 36**). For the common/striped dolphins, this was also the case; however, we observed a spike in residual values which was caused by 11 residual values (i.e., < 0.5 percent of all residuals) that were larger than 10. For both models, the distribution of residuals was right skewed which was evident from the histograms of residuals.

5.2.3.2 ASSESSING WHISTLE CHARACTERISTICS MODEL FIT

We assessed the model fit using observed versus fitted plots (**Figure 37**). The horizontal lines in **Figure 37** were caused by predicting with factor covariates, which limits the number of unique predicted values according to the levels of factor covariates. We had 10 unique predicted values for the pilot whale model and 25 for the common/striped dolphin model. For each of these, we expected a random right-skewed pattern centered around the red lines shown in the plots. To inspect whether this was the case, we created histograms for the observed values corresponding to each unique fitted value (not shown) and added means and medians of observed values for each unique fitted value to **Figure 37**. For our models, the distribution of observed values was right-skewed (confirmed by the histograms and evident from the means being larger than the medians, **Figure 37**). However, large observed values were often under-predicted for both whistle characteristics models, in particular for the common/striped dolphins. Here, the largest discrepancy was in the small proportion of observed values that were larger than 10 which were all predicted to be approximately 4.

5.3 Discussion

5.3.1 Pros and Cons of Modeling Strategies

The three types of models we developed for analyzing the acoustic sub-event data were the presence-of-acoustic-encounter-models, the presence-of-signal-type-given-acoustic-encounter models and the whistle-characteristics models. Modeling the presence of acoustic encounters in 1 min segments as a binary response has the advantage over modeling length of acoustic sub-events (or other possible measures) in, for example, a gamma model, that it takes into account the amount of time during which no acoustic encounters were detected. For length-of-vocalization models, the time periods with no vocalizations do not contribute any information to the model. Our presence models also have the advantage over modeling some measure of vocalization rate available from the acoustic sub-event data (e.g., number of seconds during which vocalizations were detected within 10 min segments) in a Poisson model, because the length of the segments create an artificial limit on possible values for model residuals, resulting in violations of model assumptions. This issue could be avoided by increasing the size of the segments so that they are longer than the longest delphinid acoustic encounter. However, as some delphinid acoustic sub-events from the present study lasted up to 2 hours, this would result in the loss of temporal resolution of the delphinid acoustic encounter data.

Another potential way to avoid the issue of creating an artificial limit on possible values for model residuals would be to model the proportion of time within predefined segments (e.g., 10 min segments) during which delphinid acoustic sub-events were recorded in a binomial model. However, for either the vocalization-rate models or the proportion-of-time models, the current format of the data is not appropriate. Individual delphinid acoustic sub-events were defined to encompass all consecutive vocalizations with gaps no longer than 60 seconds. No information was available on the actual time that vocalizations were detected within a sub-event. As a result, a 1 min delphinid acoustic sub-event may consist of two whistles that were 60 seconds apart or of continuous vocalizations by multiple individuals. This is an additional reason why our 1 min presence models were more appropriate than vocalization-rate or proportion-of-time models.

We encountered difficulty in determining the best way to fit covariates that were related to measurements of sonar pings. Generally, we included these covariates as interaction terms with the covariate *Anysonar,* which acted as an indicator variable, switching covariates on or off depending on whether sonar pings coincided with any time of the observation. This method implied that these covariates only had the potential for affecting the immediate vocalization behavior of delphinids. Hence, inference on long-term responses of delphinid vocal behavior to any of these measurements individually was not possible. The only covariate that incorporated a lag in potential response was *Sonarlag*, i.e., the time lag since the last detected sonar ping. Again, we fitted this covariate as an interaction term with an indicator variable which switched the covariate off for those records where this covariate could not be observed (i.e., the first 24 hours before the first sonar ping for each site). Although this is not an ideal approach, it represents an improvement over assigning an artificial large value to those observations where the values could not be observed. The improvement lies in that using artificially large values creates a large gap between these and observed values. This may, in turn, create a large amount of uncertainty in the partial fit for this covariate, in particular when fitting polynomial splines. However, *Sonarlag* was not retained in any of the best-fitting models.

An additional problem was encountered with fitting smooth functions in our models. If our model selection strategy had only included the first two steps described in **Section 5.1.4.1** (i.e., marginal *p*-values and elimination of collinear covariates), we would have retained covariates as polynomial splines in the model for which no-effect was a plausible answer. Generally, the objective of model selection is to retain only important covariates in the model. Using *p*-values from an *F*-test statistic, important covariates were identified as those that explained some of the residual variability in the data given that the other covariates were in the model. The residual variability that was explained by the respective covariate was illustrated in the partial fit plots. Generally, the further the partial fit is from a straight horizontal line, the more variability the covariate explains. Confidence intervals around the partial fits give a range of plausible values for the relationship between the covariate and the response. After model selection step 2, for some of the smooths of our models, the confidence limits were wide enough in the vertical dimension across the range of observed covariate values that it was possible to fit a straight line through them. This indicated that one possible function describing the relationship between the covariate and the response was a constant (i.e., no effect). Hence, we included the third step in our model selection procedure (i.e., eliminating those covariates for which no-effect was plausible). Inference from the partial fit plots for factor covariates was more straightforward. For factor covariates, confidence intervals for a level that do not include zero as a plausible value provide evidence for a significant difference between the respective level and the baseline.

One may argue that for the presence-of-signal-type-given-acoustic-event models we should include an offset term to account for varying durations of acoustic encounters. However, for these models it is difficult to determine what this offset term should be and how to define the relationship between the response and the offset term. One option is to include the length of the acoustic sub-event as an offset term. Another option is to include the time since the start of the delphinid acoustic encounter until the time of the first occurrence of the respective signal type as the offset term. In this case, this relationship may be strongly correlated with what type of signal initiated the new acoustic encounter. In either case, the fitted values of the model (i.e., the expected probability of detecting the respective signal type) are bound by 1, whereas the length of the offset for either case does not have any length restrictions and can potentially take any value). As the relationship between the probability of observing the respective signal type and the offset term therefore clearly cannot be linear, an additional parameter that describes this relationship would need to be estimated.

Further, we realize that for these analyses, it was not ideal to combine multiple delphinid species into one species group (i.e., the DEUO species group) as they may have different responses to sonar with respect to any of the three responses (presence of acoustic encounter, presence of signal type given acoustic encounter and whistle characteristics). However, as identification of all delphinid species other than pilot whales using ROCCA has higher uncertainty, we concluded that combining these species (common dolphins, striped dolphins and unidentified odontocetes) into one analysis might be the best option. There may be confounding issues when trying to detect a change in vocal behavior using this strategy. Some species may respond to sonar by not vocalizing while others may increase their vocal activity during sonar. In addition, responses in delphinid acoustic behaviors to sonar are very likely influenced by behavioral context, an animal’s previous experience with sonar, and the animal’s motivation and habituation (Weilgart 2007, Ellison et al. 2011). For example, foraging dolphins may have a very different response to sonar than resting or socializing dolphins. Information on behavioral context was not available for these data, but would be a valuable addition to data collection protocols in the future. This would require concurrent visual observations or tag data (e.g. multi-sensor data-loggers). Finally, dolphins may change the characteristics of their whistles during sonar. As species identification using ROCCA is based on these characteristics, a change in whistle characteristics due to sonar may lead to an increase in mis-identification as a result of sonar. In order to make inference for changes in whistle characteristics for individual delphinid species, certainty of species identification is needed for acoustic encounters for each period with respect to sonar (before, during, between and after).

5.3.2 Possible Inference from Models

In the previous section we described the pros and cons of the three different modeling approaches. However, one has to also be aware of the types of inference that can be drawn from these models. The presence-of-acoustic encounters models, for example, do not explain variability in the proportion of time that animals were acoustically active. Rather, they only describe changes in the probability of detecting acoustically active animals (which entails the probability of animals being acoustically active but also the probability that these signals are received by the detector, etc.), and only at a maximum resolution of 1 min. If, for example, covariate *Sonar* was retained in the best-fitting proportion model and the level ‘during’ of this covariate was significantly higher compared to the base level of ‘before’, we could only infer that during sonar, the proportion of time during which vocalizations were detected 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 recording, that we are examining the probability of animals vocalizing. But this is far from axiomatic. Alternative explanations include the possibility that if animals spent the same proportion of time vocalizing, animal density may have changed or animals may have re-distributed with respect to the recorder locations. As the probability of detecting vocalizations is dependent on range and propagation conditions (see **Section 5.3.3**), this would also have an effect on the proportion of time vocalizing. Other properties, including source levels of vocalizations or orientations of animals, may change as a result of sonar occurrence. All of these possibilities could result in fewer detections of acoustic signals. Detection of any of the signal types within a delphinid acoustic encounter may be affected in a similar manner.

5.3.3 Why Not Model Potentially Changing Numbers of Delphinids in Relation to Sonar Occurrence?

Using passive acoustic monitoring devices such as MARUs has the advantage of providing the capability of collecting large amounts of data at a relatively low cost. However, several difficulties exist when analyzing data obtained from single-hydrophone passive-acoustic monitoring devices. In particular, inference related to the number of delphinids detected by the device, be it via an estimate of density or abundance, is limited. Detection probabilities generally decay with increasing distance of vocalizing delphinid schools from the hydrophone (e.g., Helble et al. 2013, Küsel et al. 2011). This decay in detection probabilities may also vary among different devices due to varying technical properties of the devices or different sound propagation properties or background noise levels near the device locations. If we were able to measure or estimate the distances to the vocalizing dolphin schools (e.g., by using sound-propagation models, or localizing detections), we could apply distance sampling methods to estimate density of delphinid vocalization cues around the hydrophone locations. Alternatively, if at each study site the MARUs were located near enough to each other so that the same vocalization could be captured at more than one hydrophone, spatially explicit capture-recapture methods could be applied to estimate density of vocalizations (Borchers 2010, Marques et al. 2012, Martin et al. 2013, Borchers et al. submitted). To convert estimates of vocalization density into estimates of dolphin density requires additional information to estimate vocalization rates and average school sizes, which was not available for our study. Therefore, inference on potentially varying dolphin densities at the study sites in relation to sonar activities was beyond the scope of this study.

5.3.4 Results on Effects of MFA Sonar on Delphinid Vocalization Behavior

The best-fitting presence-of-acoustic-encounter models for either pilot whales or the DEUO species group did not retain any covariates pertaining to sonar in the final model. Here, however, the percentage of observed presences was very small (1 percent and 6 percent of all 1 min segments for the pilot whale and the DEUO species group model, respectively) and therefore, predictive power of our models was relatively poor (see **Section 5.2.1.2**).

For our signal-type-given-acoustic-encounter models, predictive power was generally better compared to the presence-of-vocalization models (see **Section 5.2.2.2**). For this type of model, only the DEUO species group models retained covariates related to sonar. In this model, presence of whistles given vocalization and presence of buzzes given vocalization contained the covariate *Sonar*. These models provided evidence that the expected odds of observing whistles within a vocalization encounter were higher during the emission of sonar pings compared to the 24 hours before sonar. Also, the odds of observing buzzes within a vocalization encounter were higher during, between and in the 24 hours after sonar compared to the 24 hours before sonar. Furthermore, we found evidence that for the DEUO species group, the odds of observing clicks in an acoustic encounter increased during the presence of Type 1-short and Type 3-medium pings. Similarly, the odds of observing buzzes within an acoustic encounter increased during the presence of Type 3-medium sonar pings. In addition, we found evidence that, whistle characteristics of common/striped dolphins changed during the emission of sonar and in the 24 hours after sonar when compared to 24 hours before sonar. Further analyses are needed to identify which characteristics changed and in which manner. For pilot whales no change in whistle characteristics in relation to sonar was evident.

However, we did not explore all potentially important covariates with respect to sonar. None of our covariates included a cumulative effect (i.e., a quantitative measure that summarizes the energy or level of sonar in the time directly preceding the vocalization). An example for such a covariate is the number of sonar pings in the two hours preceding a 1 min segment for the presence models or sound exposure levels of sonar. Additional analyses are necessary before these can be included.

This page intentionally left blank.

6. Analysis Using Hidden Markov Models

6.1 Introduction and Motivation for the Consideration of HMMs

We now summarize the results of fitting HMMs to the time series of acoustic sub-events of pilot whales and of the DEUO species group, with the aim of identifying potential effects of sonar exposure on the animals' acoustic behaviors. The full data set is described in detail in **Section 5.1.2.1**, and here we fit the HMMs to the series of delphinid acoustic encounters in time intervals of length 60 seconds.

A total of 20 time series were considered: six for both pilot whales and the DUEO species group from the JAX MARUs, and four for both pilot whales and the DUEO species group from the OB MARUs. There were long periods without any vocalizations present. Those periods were occasionally interspersed with shorter periods that contained at least a few vocalizations. This pattern can be modeled using stochastic mixtures, with one mixture component corresponding to periods with no activity (i.e., absence of any vocalization) and the other mixture component corresponding to periods with activity (i.e., positive probability of detecting vocalizations), with a random mechanism selecting which of the two components is active at any time. Furthermore, there was persistence in both the non-activity and the activity periods, as confirmed by the sample autocorrelation functions (shown for pilot whales in JAX in **Figure 38**). Overall, this pattern motivates the use of dependent stochastic mixtures such as HMMs to model these data. These stochastic mixtures account both for the multiphasic nature of the time series (in terms of alternating non-active and active periods) and for the serial correlation in the occurrence of the different phases.

6.2 A basic HMM for Modeling Vocalization Frequency

6.2.1 Model Formulation

An HMM is a stochastic time-series model involving two stochastic processes, only one of which is observed. The unobserved 'state' (or 'component') process is an N-state Markov chain (i.e., a stochastic process that takes values in $\{1, \ldots, N\}$ and satisfies the memoryless property, such that conditional on the present value of the process, future values are independent of past values). The observed time series, typically referred to as the 'state-dependent' process, is such that its values are assumed to be generated by one of N component distributions, with the underlying Markov chain selecting which component distribution is active at any time. Such a model is a dependent mixture, with each state corresponding to one mixture component (here corresponding to different levels of vocal activity) and the Markov chain selecting the states and hence inducing dependence. In the following, we denote the observable state-dependent stochastic process by X_t and the underlying unobservable N-state Markov chain by S_t, and focus on the two-state case $(N = 2)$. We assume a basic dependence structure where, given the current state of S_t, the variable X_t is conditionally independent from previous and future observations and states, and where the Markov chain is of first order. **Figure 39** displays the dependence structure of such a basic HMM in a directed graph.

Assuming homogeneity of the Markov chain, we summarize the probabilities of transitions between the different states in the 2×2 transition probability matrix (t.p.m.)

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix},$$

where $\gamma_{ij} = \Pr(S_{t+1} = j | S_t = i)$. The initial state probabilities are summarized in the row vector $\boldsymbol{\delta}$, where $\delta_i = \Pr(S_1 = i)$.

Conditional on the current underlying state, it is assumed that the binary observation X_t, which indicates if a vocalization is detected at time t (in which case $X_t = 1$), follows a Bernoulli distribution, with probability of detecting a vocalization varying across states.

That is, we assume that

$$X_t |\, S_t = i \sim \text{Bernoulli}(\pi_i),$$

so that

$$\Pr(\text{vocalization recorded}|S_t = i) = 1 - \Pr(\text{no vocalization recorded}|S_t = i) = \pi_i,$$

with $\pi_1 \neq \pi_2$ in general.

This simple HMM was fitted separately to:

- the six time series of delphinid acoustic encounters of pilot whales in JAX
- the four time series of delphinid acoustic encounters of pilot whales in OB
- the six time series of delphinid acoustic encounters of the DUEO species group in JAX
- the four time series of delphinid acoustic encounters of the DUEO species group in OB.

In each of these four cases, the five model parameters were assumed to be common to all (either four or six) time series. This is a potentially unrealistic, yet necessary, assumption given the very small numbers of state transitions observed per time series (the individual time series simply do not contain sufficient information in order to fit HMMs separately to each series). The consideration of models with random effects is not feasible either, given both a) the small number of component series and b) the computational complexity of HMMs incorporating random effects (Schliehe-Diecks et al. 2012). For the non-hierarchical models described above, model fitting was performed via a numerical maximization of the log-likelihood, as described in Zucchini and MacDonald (2009).

6.2.2 Results of Fitting the Baseline HMM to the Pilot Whale and DUEO Species Group Vocalization Data

For the four different combinations of species and study area, the t.p.m.s were estimated as

$$\mathbf{\Gamma} = \begin{pmatrix} 0.9992 & 0.0008 \\ 0.0504 & 0.9496 \end{pmatrix} \quad (\text{Jacksonville} - \text{pilot whales}),$$

$$\mathbf{\Gamma} = \begin{pmatrix} 0.9998 & 0.0002 \\ 0.0180 & 0.9820 \end{pmatrix} \quad (\text{Onslow Bay} - \text{pilot whales}),$$

$$\mathbf{\Gamma} = \begin{pmatrix} 0.9948 & 0.0052 \\ 0.0531 & 0.9469 \end{pmatrix} \ \text{(Jacksonville} - \text{DEUO)} \ \text{and}$$

$$\mathbf{\Gamma} = \begin{pmatrix} 0.9977 & 0.0023 \\ 0.0361 & 0.9639 \end{pmatrix} \ \text{(Onslow Bay} - \text{DEUO).}$$

According to the fitted models, the times spent in state 1 are substantially longer, on average, for pilot whales than for the DEUO species group. In each of the four cases, the first state is the 'silent' state and the second state is the 'vocally active' state (i.e., producing whistles, clicks or buzzes), with the state-dependent vocalization probabilities estimated as

$$(\pi_1, \pi_2) = (0.000, 0.803) \ \text{(Jacksonville} - \text{pilot whales),}$$

$$(\pi_1, \pi_2) = (0.000, 0.874) \ \text{(Onslow Bay} - \text{pilot whales),}$$

$$(\pi_1, \pi_2) = (0.001, 0.799) \ \text{(Jacksonville} - \text{DEUO)} \ \text{and}$$

$$(\pi_1, \pi_2) = (0.002, 0.815) \ \text{(Onslow Bay} - \text{DEUO).}$$

To illustrate the modeling approach and potential sources of lack of fit, **Figure 40** panel (a) graphically represents one of the time series of pilot whale vocalizations recorded in the OB study area, and the bottom panel (b) depicts a time series simulated from the model fitted to the OB pilot whale data. Overall, the model seemed to capture the observed general pattern of vocalizations well, exhibiting very long 'non-active' periods that were occasionally interspersed with very short vocally active periods. In the top panel (a) in **Figure 41**, one of the time series of DEUO vocalizations collected in the JAX study area is provided as an example, with the bottom panel (b) showing a time series simulated from the model fitted to the JAX DEUO data. In this case, we still obtained roughly the same total number of vocalizations for simulated and real data, but the model clearly failed to adequately capture the durations of 'vocally active' periods. This is due to the heavy tail of the distribution of these durations found for the empirical data, a feature that cannot be captured by basic HMMs due to the implicit assumption of geometric distributions for the state dwell times (i.e., the duration of time the Markov chain spends in a state before switching to a different state). So-called hidden semi-Markov models (Langrock and Zucchini 2011) allow for the specification of any distribution on the positive integers for the state dwell times (e.g., a shifted negative binomial), and hence constitute an obvious alternative class of models that are likely to substantially improve the fit in this respect. However, in the given scenario it is computationally infeasible to fit such models, due to the fine time resolution considered and the resulting very large support of the dwell-time distributions.

6.3 HMMs Incorporating Sonar-related Covariates

6.3.1 Model Formulation

In this section we extend the model by allowing sonar-related covariates to affect state-switching probabilities. This was done in order to investigate, for example, if sonar exposure led to a decreased probability of the animals occupying the 'vocally active' state. For any given series of covariates, $z_1, z_2, z_3, \ldots$, the model described in **Section 6.2.1** was modified by letting the state transition probabilities depend on the covariate values, as follows:

$$\gamma_{12}^{t} = \Pr(S_{t+1} = 2 | S_t = 1) = \text{logit}^{-1}(\beta_{1,0} + \beta_{1,1} z_t)$$

$$\gamma_{21}^{t} = \Pr(S_{t+1} = 1 | S_t = 2) = \text{logit}^{-1}(\beta_{2,0} + \beta_{2,1} z_t)$$

The diagonal elements of the t.p.m. are then obtained as $\gamma_{11}^{t} = 1 - \gamma_{12}^{t}$ and as $\gamma_{22}^{t} = 1 - \gamma_{21}^{t}$. We note that in this model, the transition probabilities depend on time since the covariates vary over time. We considered several covariates, one at a time, in separate models: *Sonar*, *Anysonar*, *Sonarlag*, *Mean repetition rate* and *SDEV ping interval.* In the case of *Sonar*, which is a categorical covariate, the above equations were modified in the standard way. It should be noted here that since model fitting is much more computer-intensive for HMMs than it is for GEEs, we did not consider all possible covariates listed in **Section 5.1.3**. Model fitting was again performed via numerical likelihood maximization.

6.3.2 Model Fitting Results and Model Selection

Table 11 lists the values of the Akaike Information Criterion (AIC) for the different models considered, for each of the four different combinations of species and study area. For pilot whales, very few vocalizations (and hence also state transitions) occurred during the observation period (cf. **Figure 40**). This makes it very difficult to make inference on the factors driving the state-switching dynamics. In particular, the estimation was numerically unstable in terms of local maxima of the likelihood. Furthermore, no clear pattern was found in the AIC values for the fitted models, likely due to the limited amount of information contained in these time series. For pilot whales in the JAX study area, the model with the *SDEV ping interval* covariate was favored by the AIC, whereas in the OB study area the model without any covariates was favored.

In view of the caveats of the data collected on pilot whale vocalizations, we discuss in more detail only the results obtained for the DEUO species group vocalization models. For the DEUO species group (excluding pilot whales), the model with the *Sonar* covariate affecting the state transition probabilities was deemed best by the AIC, for both the JAX and the OB study areas. Moreover, according to the AIC this was the only covariate of those considered that affected the state-switching dynamics.

For the DEUO species in the JAX study area, the t.p.m.s of the model selected by the AIC, associated with the four different levels of the categorical covariate *Sonar*, were

$$\mathbf{\Gamma}(\text{before}) = \begin{pmatrix} 0.997 & 0.003 \\ 0.047 & 0.953 \end{pmatrix},$$

$$\mathbf{\Gamma}(\text{during}) = \begin{pmatrix} 0.994 & 0.006 \\ 0.048 & 0.952 \end{pmatrix},$$

$$\mathbf{\Gamma}(\text{between}) = \begin{pmatrix} 0.995 & 0.005 \\ 0.063 & 0.937 \end{pmatrix} \text{ and}$$

$$\mathbf{\Gamma}(\text{after}) = \begin{pmatrix} 0.993 & 0.007 \\ 0.052 & 0.948 \end{pmatrix}.$$

The associated stationary (or steady-state) distributions, which indicated the expected proportion of time spent in the two states (for a given covariate level), were given by

$$\boldsymbol{\delta}(\text{before}) = (0.933, 0.067),$$

$$\boldsymbol{\delta}(\text{during}) = (0.894, 0.105),$$

$$\boldsymbol{\delta}(\text{between}) = (0.922, 0.078) \text{ and}$$

$$\boldsymbol{\delta}(\text{after}) = (0.887, 0.113).$$

In both states, the persistence was highest before the sonar exposure. In other words, the fitted model indicated that sonar exposure leads to a higher probability of transitioning between the states, in both directions (i.e., from ‘vocally active’ to ‘silent’ and vice versa). The probability of leaving the ‘vocally active’ state was highest in the ‘between’ periods. In the ‘during’ and ‘after’ periods, animals spent more time being vocally active than in the ‘before’ and ‘between’ periods, according to the fitted model.

For the OB study area, the t.p.m.s of the selected model, associated with the four different levels of the covariate *Sonar*, were

$$\boldsymbol{\Gamma}(\text{before}) = \begin{pmatrix} 0.998 & 0.002 \\ 0.059 & 0.941 \end{pmatrix},$$

$$\boldsymbol{\Gamma}(\text{during}) = \begin{pmatrix} 0.998 & 0.002 \\ 0.036 & 0.964 \end{pmatrix},$$

$$\boldsymbol{\Gamma}(\text{between}) = \begin{pmatrix} 0.996 & 0.004 \\ 0.029 & 0.971 \end{pmatrix} \text{ and}$$

$$\boldsymbol{\Gamma}(\text{after}) = \begin{pmatrix} 0.999 & 0.001 \\ 0.064 & 0.936 \end{pmatrix}.$$

The associated stationary distributions were given by

$$\boldsymbol{\delta}(\text{before}) = (0.964, 0.036),$$

$$\boldsymbol{\delta}(\text{during}) = (0.939, 0.061),$$

$$\boldsymbol{\delta}(\text{between}) = (0.871, 0.129) \text{ and}$$

$$\boldsymbol{\delta}(\text{after}) = (0.982, 0.018).$$

In this case, the probability of leaving the ‘vocally active’ state, and of occupying the ‘silent’ state, was highest in the ‘after’ periods. The model further indicated that in the ‘between’ periods more time was spent in the ‘vocally active’ state than during any other phase of the sonar exposure experiment.

6.4 Discussion

For the DEUO species group, there was some indication that the categorical covariate *Sonar* affects the state-switching dynamics, and as a result, the number of vocalizations detected. This could be an indication that sonar exposure alters the behavioral dynamics. However, the pattern identified by the HMM regarding the exact effect of *Sonar* on the state process was not consistent across the two study areas, JAX and OB. In the case of the pilot whales, the data set did not allow for a detailed investigation of the effect of sonar exposure on vocalization activity, since there were too few vocalizations present in the data. Because these results do not draw a clear picture, and in particular do not provide a clear outcome of the effects of sonar on vocal activity, we recommend conducting extensive simulation studies in order to investigate how much data would be required to accurately infer the effect of sonar using HMMs or related stochastic processes.

6.4.1 Comparison with GEE Approach

In order to compare whether inference from the HMM approach was similar to that from the GEE approach, we conducted additional GEE analyses where the data from the different study areas (JAX and OB) were analyzed separately and where model selection was limited to those covariates used in the HMM analyses. Due to time constraints, we could not use all the available covariates listed **Section 5.1.3** for the HMM analyses. This section solely serves the purpose of determining whether the same covariates would be retained in the best fitting presence of vocalization models using GEEs as were retained in the best fitting HMM under the same circumstances. Hence, we do not elaborate on the details of the GEE modelling results.

For this purpose, we applied the GEE methods described in **Section 5.1** to the data structured as in this section (i.e., in four different analyses): 1) pilot whales at the JAX location; 2) pilot whales at the OB location; 3) DEUO at the JAX location and 4) DEUO at the OB location. This is different from the data format for **Section 5** where all locations were analyzed together for both pilot whales and the DEUO species group. For these GEE analyses, we applied the three-step model-selection procedure described in **Section 5.1.4.1** limited to the same covariates listed in **Table 11** (*Anysonar, Sonar, Sonarlag, Mean rep. rate* and *SDEV ping interval*). For analyses 1–3, the same covariates were retained in the best fitting GEE analyses as in the HMM analyses (**Table 11**). The best fitting GEE model for analysis 1 contained the covariate *SDEV ping interval*. The best fitting model for analysis 2 contained no covariates. The best fitting GEE model for analysis 3 contained the covariate *Sonar*. The inference from the GEE model on this covariate was similar to that from the HMM. From the GEE model, we inferred that the odds of observing presences of vocalizations were highest for levels ‘after’ and ‘during’ of covariate *Sonar* and lowest for level ‘before’. From the HMM we inferred, that in the ‘during’ and ‘after’ periods, animals spent more time being vocally active than in the ‘before’ and ‘between’ periods. However, for analysis 4, a discrepancy in the inference between the GEE and HMM approaches existed in that the best fitting GEE model contained no covariate while the corresponding HMM contained covariate *Sonar*.

One of the potential benefits of the HMM approach compared to the GEE approach is that the HMM approach explicitly models the correlation pattern, thus acknowledging the time-series nature of the data, whereas in a GEE approach the correlation is treated as nuisance. As a consequence, HMMs allow for more detailed inference on the dynamics of the process observed, and in many cases will lead to a better understanding of the data-generating system. Furthermore, while information criteria can be used for selecting the best HMM from a suite of candidate models, model selection methods are an area of ongoing research for GEEs. On the other hand, GEEs are less time consuming and more user-friendly to fit due to the availability of model-fitting functions in e.g., the *geepack* R package.

7. Conclusions and Recommendations for Future Work

The statistical analyses in this project address three main questions: 1) Does the probability of detecting delphinid acoustic encounters change during or after periods in which MFA sonar is being broadcast?; 2) Does the probability of detecting whistles, clicks or buzzes within a delphinid acoustic encounter change during or after periods in which MFA sonar is being broadcast?; and 3) Given that a whistle encounter is detected, do whistle characteristics change during or after periods in which MFA sonar is being broadcast? For each of these research questions we developed models fitted with GEEs to accommodate potential over-dispersion and correlated errors. In addition, we fitted HMMs to address research question 1 (potential changes in the probability of detecting vocalizations in the presence of sonar). HMMs have the benefit over GEEs in that they model the dynamics of switching between different states (as opposed to only modeling the expected state). Inference on which covariates were retained in the final models was similar between the two approaches. An additional potential benefit of the HMM approach compared to the GEE approach is that the HMM approach explicitly models the correlation pattern, thus acknowledging the time-series nature of the data, whereas in a GEE approach the correlation is treated as nuisance. Furthermore, while information criteria can be used for selecting the best HMM from a suite of candidate models, model-selection methods are an area of ongoing research for GEEs. We propose a three-step model-selection procedure to determine the best fitting GEE, which is a combination of determining important covariates based on marginal p-values, eliminating collinear covariates and inspecting partial fit plots (including 95 percent confidence intervals) for plausible no-effect of each of the respective covariates.

The results presented in this report provide a good foundation for the further development of statistical methods for evaluating changes in delphinid acoustic behavior in response to MFA sonar, however larger sample sizes and the inclusion of covariates representing a cumulative effect of sonar on vocalization behaviors (e.g., the number of sonar pings detected in certain time periods preceding a 1 min vocalization encounter) are necessary to make these analyses more robust. A larger number of both independent sonar events and delphinid acoustic encounters should be analyzed in order to improve statistical power. For our study, the predictive power of the presence models was relatively poor: less than 60% of all predictions were correct (**Table 6**). With only two possible outcomes – 1 for presence or 0 for absence —we expect to obtain 50 percent correct predictions by chance alone. The low proportion of correct predictions may be due to small percentage of observed presences (1 percent and 6 percent for the pilot whale and the DEUO species group models, respectively) and/or that the available covariates did not capture the underlying pattern. Similar problems were encountered when fitting the HMM to the presence of vocalization data in particular for the pilot whales, where larger proportions of presences are needed for more reliable modeling of the underlying processes. The predictive power improved for the signal-type-given-acoustic-encounter models. Here, the proportion of correct predictions by the fitted model was between 60 and 91 percent (**Table 9**).

The low predictive power of the models due to low sample sizes made it difficult to draw conclusions regarding the effects of sonar on acoustic behavior. Additional data that could be included in future analyses include archival acoustic recordings made with a variety of autonomous recorders such as MARUs, High-frequency Recording Packages (HARPs) and Ecological Acoustic Recorders (EARs). Many of these recordings contain sonar events and would be a valuable supplement to the analyses conducted here. We also recommend conducting simulation studies in order to investigate how much data would be required to accurately infer the effect of sonar using HMMs or related stochastic processes.

For the statistical analyses, the data provided were limited to the resolution of delphinid acoustic sub-events. Each sub-event contained vocalizations that were separated by no more than 60 seconds. Start times of individual vocalizations or other quantitative measures of vocalizations within the encounters (e.g., number of vocalizations per encounter) were not logged. More detailed logging that includes start times of individual sounds or total length of whistle fragments (Gillespie et al. 2012) would allow a more fine-scale or quantitative analysis of changes in vocal behavior and may reveal patterns that are not apparent when analyzing encounters that are binned.

Furthermore, it would be valuable to have the ability to examine responses to MFAS on a species-by-species basis. In order to do this, the performance of the classifier must be evaluated using recordings made at depth. The relatively poor classification results reported here were unexpected, based the high correct classification scores that occurred for similar whistle classification work conducted by Oswald (2013). Future work should examine whether classifications reported here were accurate and, if not, why. The classifier used in this analysis was trained using acoustic data recorded using towed hydrophone arrays at the sea surface. It is likely that propagation through the water column affects whistle variables extracted from deep-water recordings, and this could negatively impact classifier performance. The magnitude of these effects should be evaluated and if necessary, new classifiers should be developed specifically for seafloor-mounted recorder data. To accomplish these goals, vessel surveys should co-occur with autonomous recorder deployments. These vessel surveys should incorporate both visual effort and acoustic recordings made near the sea-surface. These data could be used to ground-truth classification results from autonomous-recorder data as well as examine changes in whistle structure with depth. The effects of recording depth on whistle features could also be examined with playback studies and propagation modelling work.

Classifier performance could also be improved by adding more species to the classification algorithm. Currently, ROCCA’s Atlantic classifier contains five species commonly encountered in some areas the Northwest Atlantic, but several species that produce whistles and are known to occur in the Northwest Atlantic (e.g., rough-toothed dolphin, false killer whale, etc.) are not included in the classifier due to sample size limitations. The addition of these species would improve the performance of ROCCA when applied to whistles recorded in this region. Effort should be made to collect visually validated, single species recordings of both the species that are not currently included in ROCCA and those that are currently included in ROCCA. Increasing the training data sample size for all species in the classifier would allow the classifier to capture more of the variability in the data and will likely lead to more accurate classification results. Archival, visually validated acoustic recordings exist that could be used towards accomplishing these tasks. In addition, whistles from each time period related to sonar (i.e., the 24 hours before, during, between and the 24 hours after sonar) should be included in training classifiers. This necessity became evident from our whistle-characteristics models, which showed that characteristics of whistles changed during and in the 24 hours after the emission of sonar compared to the control period, i.e., the 24 hours before sonar (see **Section 5.3.4**). More detailed examination of whistle characteristics should be conducted to determine which characteristics changed, in which manner they changed, how quickly changes occurred after the onset of sonar, and how long the changes persisted.

Visual observation, tagging and localization capabilities would be valuable additions to the methods presented here. Currently, it is unknown whether increases in the probability of detection of whistles and buzzes in the presence of sonar are due to increases in the number of animals present, due to increases in the acoustic activity of the animals in the area, or due to some combination of these. Localization capabilities would allow us to track individuals as well as estimate density of delphinid vocalization cues around hydrophones. In addition, if hydrophones were located near enough to each other so that the same vocalization could be captured at more than one hydrophone, spatially explicit capture-recapture methods could also be used to estimate density of vocalizations. The ability to track animals, both visually and acoustically, will provide a means for better understanding their responses to MFA sonar.

Finally, the analyses reported here did not include any attempt to account for or investigate the behavioral context of the animals during the periods of sound exposure. The behavioral state, motivation and experience level of animals exposed to sound is known to have an effect on how marine mammals respond to sound (Ellison et al. 2011). Behavioral data from individuals or groups was not available for the MARU recordings analyzed here. The addition of visual observations, tagging, and/or localization and tracking capabilities would allow valuable context information to be collected during future efforts and would allow for a more comprehensive and context dependent evaluation of how and why acoustic behavior changes in response to MFA sonar.

This page intentionally left blank.

8. Acknowledgements

We would like to thank Shannon Coates, Talia Dominello, Kerry Dunleavy, Cory Hom-Weaver and Robyn Walker for many hours spent analyzing whistle and sonar data. Thank you to Lynne Hodge for providing information and guidance on the OB data. Thank you to Dr. Monique Mackenzie for statistical advice pertaining to GEEs. We thank Naval Facilities Engineering Command, Atlantic (NAVFAC LANT) for providing funding for this analysis. We are grateful for the logistic support and advice from Joel Bell and Anurag Kumar (NAVFAC LANT) and from Dan Engelhaupt (HDR).

This page intentionally left blank.

9. References

Au, D.W.K. and W.L. Perryman. 1985. Dolphin habitats in the eastern tropical Pacific. *Fishery Bulletin* 83(4):623-643.

Au, W.W., D.A. Carder, R.H. Penner, and B.L. Scronce. 1985. Demonstration of adaptation in beluga whale echolocation signals. *Journal of the Acoustical Society of America* 77:726–730.

Barkley, Y., J.N. Oswald, J.V. Carretta, S. Rankin, A. Rudd, and M.O. Lammers. 2011. Comparison of real-time and post-cruise acoustic species identification of dolphin whistles using ROCCA (Real-time Odontocete Call Classification Algorithm). NOAA Technical Memorandum NMFS-SWFSC-473. National Marine Fisheries Service, La Jolla, California. 29 pp.

Basu, D. 1977. On the elimination of nuisance parameters. *Journal of the American Statistical Association* 77:355-366.

Baumgartner, M.F., L. Freitag, J. Partan, K.R. Ball, and P.E. Kenneth. 2008. Tracking large marine predators in three dimensions: The Real-Time Acoustic Tracking System. *IEEE Journal of Oceanic Engineering* 33(2):146-157.

Bio-Waves, Inc. 2013. *SonarFinder User's Manual*. Submitted to HDR, Inc., MSA # CON-005-4394-009, Subproject # 164744-KB10. Bio-Waves, Inc., Encinitas, California. 30 pp.

Borchers, D. 2010. A non-technical overview of spatially explicit capture-recapture models. *Journal of Ornithology* 152(2):435-444.

Borchers, D. L., B.C. Stevenson, D. Kidney, L. Thomas, and T.A. Marques. (Submitted). A unifying model for capture-recapture and distance sampling. *Journal of the American Statistical Association.* DOI: 10.1080/01621459.2014.893884

Breiman, L. 2001 Random forests. *Machine*. *Learning* 45:5-32.

Castellote, M., C.W. Clark, and M.O. Lammers. 2012. Acoustic and behavioral changes by fin whales (*Balaenoptera physalus*) in response to shipping and airgun noise. *Biological Conservation* 147(1):115-122.

CETAP (Cetacean and Turtle Assessment Program, University of Rhode Island). 1982. A Characterization of Marine Mammals and Turtles in the Mid- and North Atlantic Areas of the U. S. Outer Continental Shelf, Final Report. Contract AA551-CT8-48. Bureau of Land Management, Washington, D.C. 586 pp.

Clark, C.W., J.F. Borsani, and G. Notarbartolo-di-Sciara. 2002. Vocal activity of fin whales, *Balaenoptera physalus*, in the Ligurian Sea. *Marine Mammal Science* 18(1):286-295.

Clark, C.W. and P.J. Clapham. 2004. Acoustic monitoring on a humpback whale (*Megaptera novaeangliae*) feeding ground shows continual singing into late spring. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 271:1051-1057.

DeRuiter, S.L., I.L. Boyd, D.E. Claridge, C.W. Clark, B.L. Southall, and P.L. Tyack. 2013. Delphinid whistle production and call matching during playback of simulated military sonar. *Marine Mammal Science* 29:46-59*.*

Di Iorio, L., and C.W. Clark. 2010. Exposure to seismic survey alters blue whale acoustic communication. *Biology Letters* 6(1):51-54.

DoN (Department of the Navy). 2008. *Marine Resources Assessment Update for the Charleston/ Jacksonville Operating Area*. Naval Facilities Engineering Command, Atlantic Division; Norfolk, Virginia. Contract number N62470-02-D-9997, Task Order Number 0056. Prepared by Geo-Marine, Inc., Hampton, Virginia.

Ellison, W.T., B.L. Southall, C.W. Clark, and A.S. Frankel. 2011. A new context-based approach to assess marine mammal behavioral responses to anthropogenic sounds. *Conservation Biology* 26:21-28.

Foote, A.D., R.W. Osborne, and A.R. Hoelzel. 2004. Whale-call response to masking boat noise. *Nature (London)* 428:910.

Fox, J., and G. Monette. 1992. Generalized collinearity diagnostics. *Journal of the Acoustical Society of America* 87:178-183.

Gaskin, D.E. 1992. Status of the common dolphin, *Delphinus delphis*, in Canada. *Canadian Field-Naturalist* 106(1):55-63.

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

Gillespie, D., J. Gordon, R. McHugh, D. McLaren, D.K. Mellinger, P. Redmond, A. Thode, P. Trinder, and D. Xiao. 2008. PAMGUARD: Semiautomated, open-source software for real-time acoustic detection and localization of cetaceans. *Journal of the Acoustical Society of America* 30(5):54-62.

Gillespie, D., M. Caillat, J. Gordon and P. White (2012). Automatic detection and classification of odontocete whistles. *Journal of the Acoustical Society of America* 134(3):2427-2437.

Halekoh, U., S. Højsgaard, and J. Yan. 2006. The R package geepack for Generalized Estimating Equations. *Journal of Statistical Software* 15(2):1-11.

Halpin, P. N., A. J. Read, B. D. Best, K. D. Hyrenbach, E. Fujioka, M. S. Coyne, L. B. Crowder, S. A. Freeman, and C. Spoerri. 2009. OBIS-SEAMAP: The world data center for marine mammal, sea bird, and sea turtle distributions. *Oceanography* 22:104-115.

Helble, T. A., G.L. D'Spain, G.S. Campbell, and J.A. Hildebrand. 2013. Calibrating passive acoustic monitoring: Correcting humpback whale call detections for site-specific and time-dependent environmental characteristics. *Journal of the American Statistical Association* 134(5):EL400-406. doi:10.1121/1.4822319.

Hodge, L. E. W., J.T. Bell, A. Kumar, and A.J. Read. 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. doi: 10.1111/mms.12006

Johnston, D.W., M. McDonald, J. Polovina, R. Domokos, S. Wiggins, and J. Hildebrand. 2008. Temporal patterns in the acoustic signals of beaked whales at Cross Seamount. *Biology Letters* 4:208-211.

Küsel, E. T., D.K. Mellinger, L. Thomas, T.A. Marques, D. Moretti, and J. Ward. 2011. Cetacean population density estimation from single fixed sensors using passive acoustics. *Journal of the American Statistical Association* 129(6):3610-3622.

Langrock, R., and W. Zucchini. 2011. Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics and Data Analysis* 55:715-724.

Lesage, V., C. Barrette, M.C.S. Kingsley, and B. Sjare. 1999. The effect of vessel noise on the vocal behavior of belugas in the St. Lawrence River Estuary, Canada. *Marine Mammal Science* 15:65–84.

Marques, T. A., L. Thomas, S.W. Martin, D.K. Mellinger, S. Jarvis, R.P. Morrissey, C. Ciminello, and N. DiMarzio. 2012. Spatially explicit capture recapture methods to estimate minke whale abundance from data collected at bottom mounted hydrophones. *Journal of Ornithology* 152(Supplement 2): S445-S455.

Martin, S.W., T.A. Marques, L. Thomas, R.P. Morrissey, S. Jarvis, N. DiMarzio, D. Moretti and D. Mellinger. 2013. Estimating minke whale *(Balenoptera acutorostrata*) boing sound density using passive acoustic sensors. *Marine Mammal Science* 29:142-158.

McCarthy, E., D. Moretti, L. Thomas, N. DiMarzio, R. Morrissey, S. Jarvis, J. Ward, A. Izzi, and A. Dilley. 2011. Changes in spatial and temporal distribution and vocal behavior of Blainville's beaked whales (*Mesoplodon densirostris*) during multi-ship exercises with mid-frequency sonar. *Marine Mammal Science* 27:E206-E226.

Melcón, M.L., A.J. Cummins, S.M. Kerosky, L.K. Roche, S.M. Wiggins, and J.A. Hildebrand. 2012. Blue whales respond to anthropogenic noise. *PLoS ONE* 7(2):e32681. doi:10.1371/journal.pone.0032681.

Mellinger, David K., 2001. Ishmael 1.0 User's Guide. NOAA Technical Memorandum OAR PMEL-120, available from NOAA/PMEL/OERD, 2115 SE OSU Drive, Newport, Oregon 97365.

Mellinger, D.K., K.M. Stafford, S.E. Moore, R.P. Dziak, and H. Matsumoto. 2007. An overview of fixed passive acoustic observation methods for cetaceans. *Oceanography* 20(4):36-45.

Nieukirk, S.L., K.M. Stafford, D.K. Mellinger, R.P. Dziak, and C.G. Fox. 2004. Low-frequency whale and seismic airgun sounds recorded in the mid-Atlantic Ocean*. Journal of the Acoustical Society of America* 115(1):1832-1843.

Norris, T.F., J.O. Oswald, T.M. Yack, and E.L. Ferguson. 2012. *An Analysis of Marine Acoustic Recording Unit (MARU) Data Collected off Jacksonville, Florida in Fall 2009 and Winter 2009-2010*. Draft Report. Submitted to Naval Facilities Engineering Command (NAVFAC) Atlantic, Norfolk, Virginia, under Contract No. N62470-10-D-3011, Task Order 021, issued to HDR Inc., Norfolk, Virginia. Prepared by Bio-Waves Inc., Encinitas, California. 138 pp.

Olson, P.A. and S.B. Reilly. 2002. Pilot whales *Globicephala melas* and *G. macrorhynchus*. Pages 898-903 in Perrin, W.F., B. Würsig, and J.G.M. Thewissen, eds. *Encyclopedia of Marine ammals*. Academic Press, San Diego, California.

Oswald, J.N., S. Rankin, and J. Barlow. 2004. The effect of recording and analysis bandwidth on acoustic identification of delphinid species. *Journal of the Acoustical Society of America* 116:3178-3185.

Oswald, J.N. 2013. *Development of a Classifier for the Acoustic Identification of Delphinid Species in the Northwest Atlantic Ocean.* Final Report. Submitted to HDR Environmental, Operations and Construction, Inc. Norfolk, Virginia under Contract No. CON005-4394-009, Subproject 164744, Task Order 003, Agreement#105067. Prepared by Bio-Waves, Inc., Encinitas, California. 58 pp.

Oswald, J.N., J.V. Carretta, M. Oswald, S. Rankin and W.W.L. Au. 2011. Seeing the species through the trees: Using random forest classification trees to identify species-specific whistle types*. Journal of the Acoustical Society of America* 129:2639.

Oswald, J.N., S. Rankin, J. Barlow, M. Oswald and M.O. Lammers. 2013. Real-time Call Classification Algorithm (ROCCA): software for species identification of delphinid whistles. Pages 245-266 in O. Adam, and F. Samaran, eds. *Detection, Classification and Localization of Marine Mammals using Passive Acoustics, 2003-2013: 10 years of International Research.* DIRAC NGO, Paris, France.

Palka, D. 2012. Cetacean abundance estimates in U.S. Northwestern Atlantic Ocean waters. Northeast Fisheries Science Center Reference Document 12-29. National Marine Fisheries Service, Woods Hole, Massachusetts. 37 pp.

Pan, W. 2001. Akaike's Information Criterion in generalized estimating equations. *Biometrics* 57(1):120-125.

Payne, P.M., and D.W. Heinemann. 1993. The distribution of pilot whales (*Globicephala* spp.) in shelf/shelf-edge and slope waters of the northeastern United States, 1978-1988. *Reports of the International Whaling Commission* (Special Issue 14):51-68.

Rendell, L.E., and J.C.D. Gordon. 1999. Vocal response of long-finned pilot whales (*Globicephala melas*) to military sonar in the Ligurian Sea. *Marine Mammal Science* 15:198-204.

Risch, D., P.J. Corkeron, W.T. Ellison, and S.M. Van Parijs. 2012. Changes in humpback whale song occurrence in response to an acoustic source 200 km away. *PLoS ONE* 7(1):e29741. doi:10.1371/journal.pone.0029741.

Schliehe-Diecks, S., P.M. Kappeler, and R. Langrock. 2012. On the application of mixed hidden Markov models to multiple behavioural time series. *Interface Focus* 2:180–189.

Scott-Hayward, L.A.S., C.S. Oedekoven, M.L. Mackenzie, and E. Rexstad. 2013. MRSea package (version 0.0.1). Technical Report. CREEM. University of St. Andrews.

Selzer, L.A., and P.M. Payne. 1988. The distribution of white-sided (*Lagenorhynchus acutus*) and common dolphins (*Delphinus delphis*) vs. environmental features of the continental shelf of the northeastern United States. *Marine Mammal Science* 4:141-153.

Sousa-Lima, R.S., T.F. Norris, J.N. Oswald, and D.P. Fernandes. 2013. A review and inventory of fixed autonomous recorders for passive acoustic monitoring of marine mammals. *Aquatic Mammals* 39:23-53.

Tyack, P.L., W.M.X. Zimmer, D. Moretti, B.L. Southall, D.E. Claridge, J.W. Durban, C.W. Clark, A. D'Amico, N. DiMarzio, S. Jarvis, E. McCarthy, R. Morrissey, J. Ward and I.L. Boyd. 2011. Beaked whales respond to simulated and actual Navy sonar. *PloS ONE* 6(3):e17009. doi:17010.11371/journal.pone.0017009.

Van Parijs, S.M., C.W. Clark, R.S. Sousa-Lima, S. Parks, S. Rankin, D. Risch, D., and I.C. van Opzeeland. 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.

Van Parijs, S.M., and P.J. Corkeron. 2001. Boat traffic affects the acoustic behaviour of Pacific humpback dolphins, *Sousa chinensis*. *Journal of the Marine Biological Association of the United Kingdom* 81:533-538.

Waring, G.T., E. Josephson, K. Maze-Foley, and P.E. Rosel. 2012. U.S. Atlantic and Gulf of Mexico marine mammal stock assessments – 2011. NOAA Technical Memorandum NMFS-NE-221. National Marine Fisheries Service, Woods Hole, Massachusetts.

Weilgart, L.S. 2007. The impacts of anthropogenic noise on cetaceans and implications for management. *Canadian Journal of Zoology* 85:1091-1116.

Wiggins, S. 2007. Triton (Version 1.80) [Acoustic Processing Software]. Scripps Institution of Oceanography, UC San Diego, La Jolla, California. Retrieved August 1, 2011. Available from www.cetus.ucsd.edu.

Zucchini, W., and I.L. MacDonald. 2009. *Hidden Markov Models for Time Series: An Introduction Using R*. CRC Press, Boca Raton, Florida.

This page intentionally left blank.

10. Figures

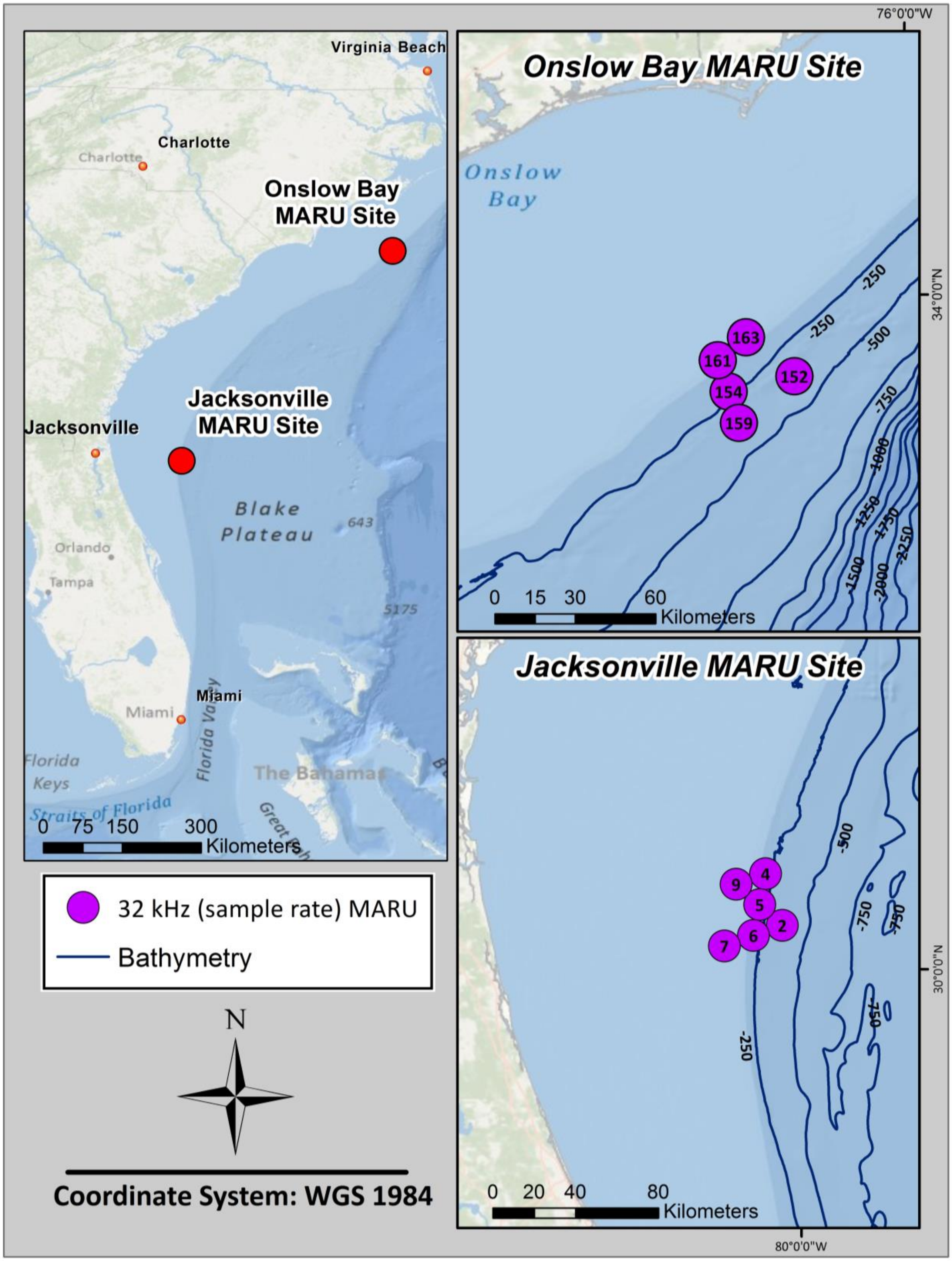

Figure 1. Map of MARUs off Jacksonville, Florida, and Onslow Bay, North Carolina.

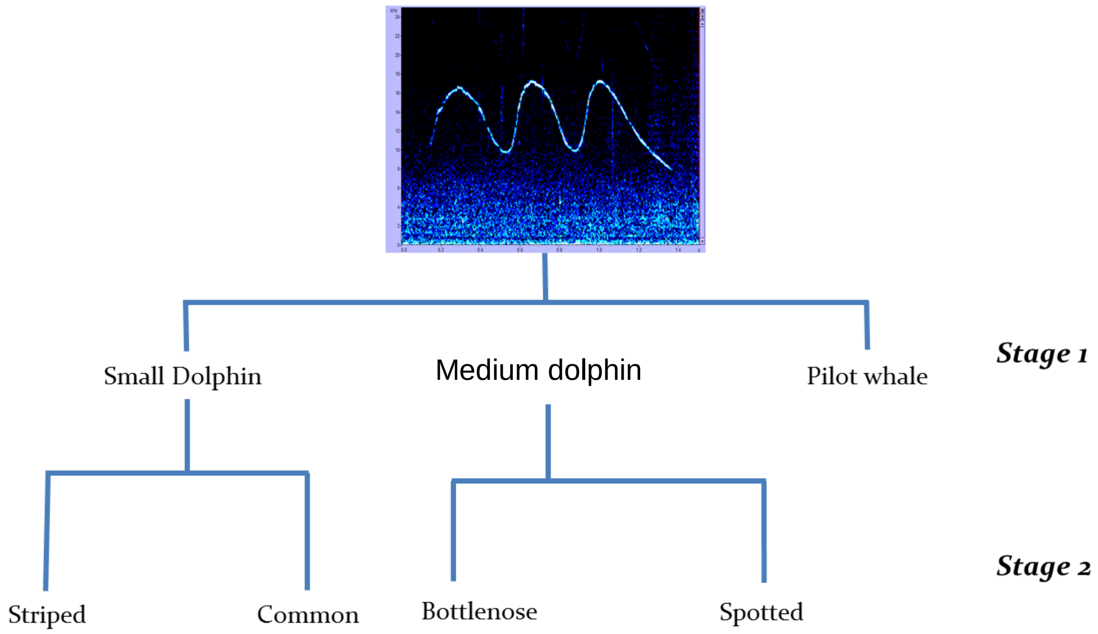

Figure 2. Schematic diagram of two-stage random forest. In stage one; whistles are classified as small dolphins (e.g., common and striped dolphins), medium dolphins (e.g., bottlenose and spotted dolphins) or pilot whales. Whistles are then classified to species in stage two.

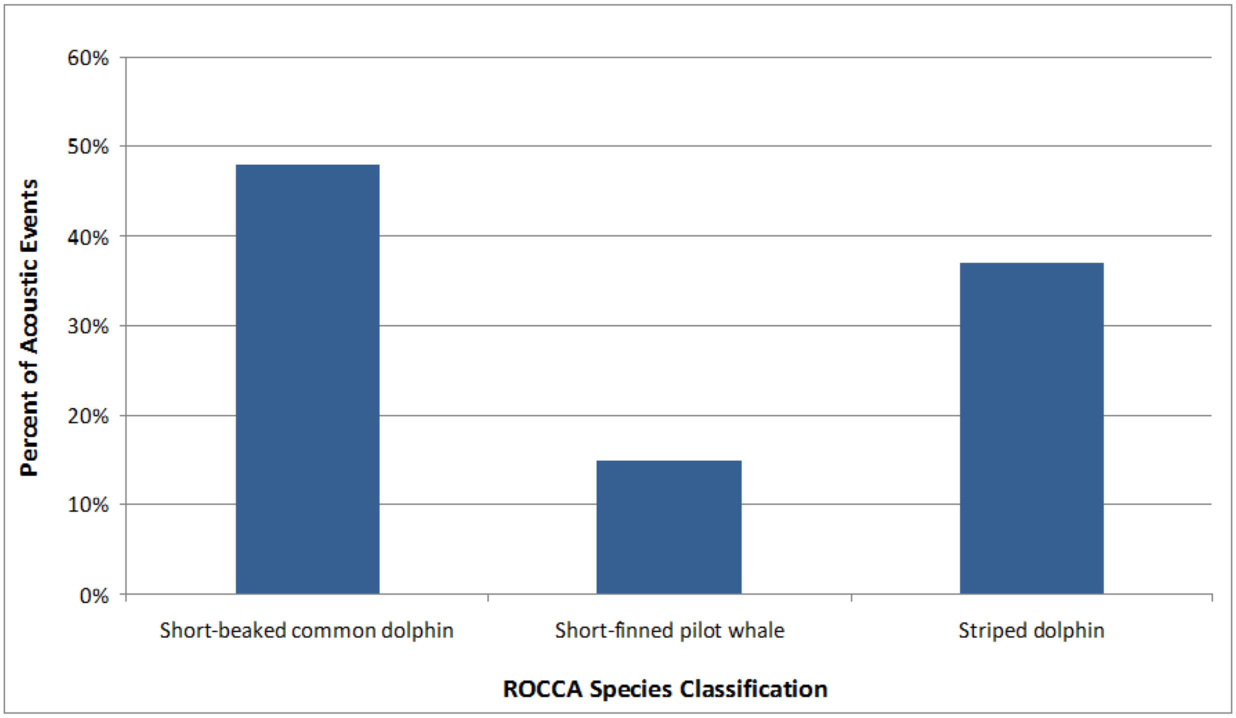

Figure 3. Percentage of delphinid acoustic encounters (n = 100) from OB MARU data classified as short-beaked common dolphin, short-finned pilot whale or striped dolphin using a random-forest classifier in PAMGuard's ROCCA whistle classification module. Note that no encounters were classified as bottlenose or Atlantic spotted dolphins.

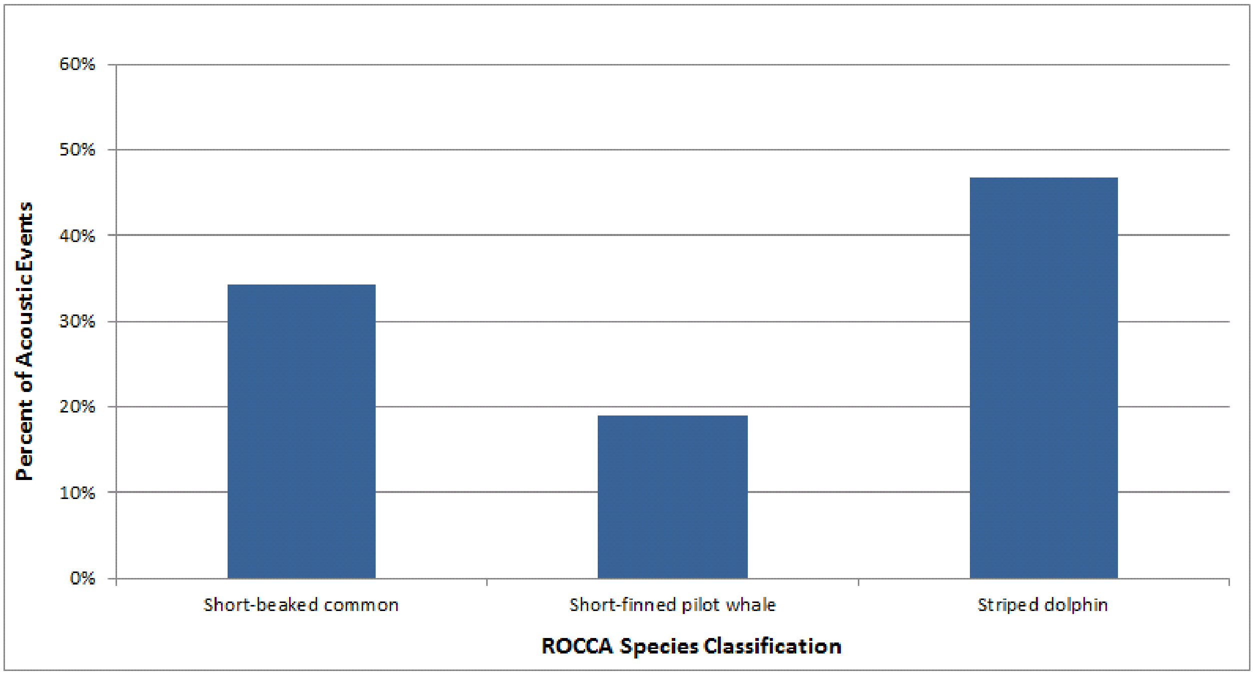

Figure 4. Percentage of delphinid acoustic encounters (n = 158) from JAX deployment 1 MARU data classified as short-beaked common dolphin, short-finned pilot whale or striped dolphin using a random-forest classifier in PAMGuard's ROCCA whistle classification module. Note that no encounters were classified as bottlenose or Atlantic spotted dolphins.

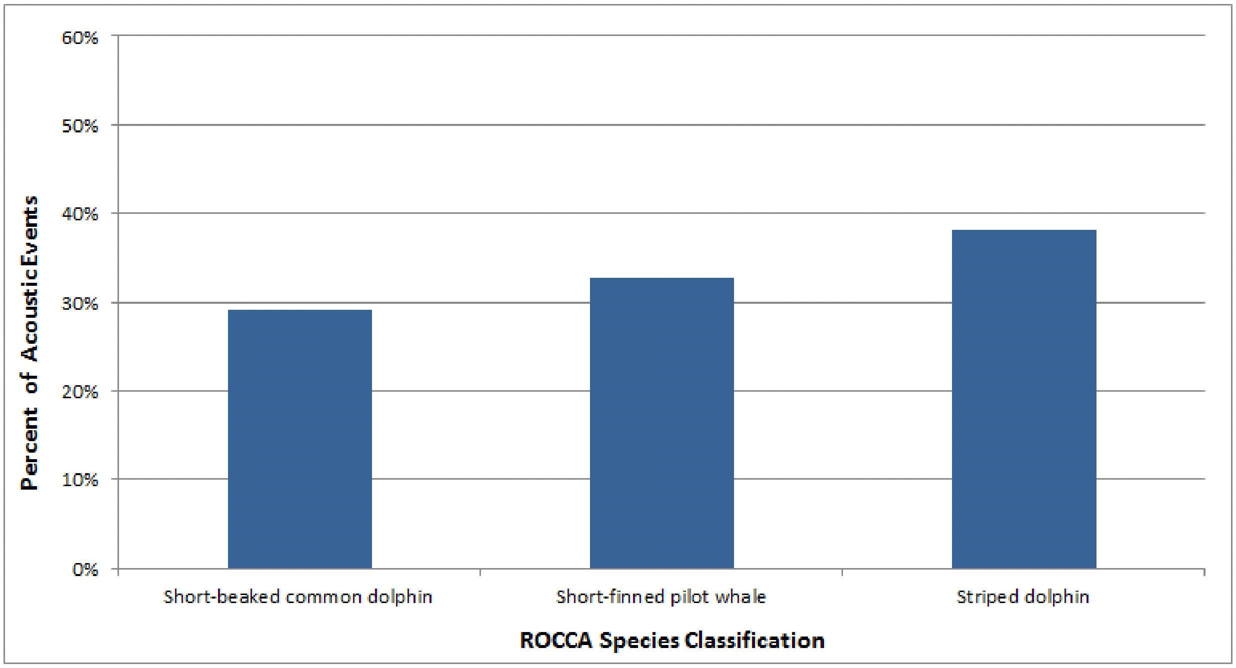

Figure 5. Percentage of delphinid acoustic encounters (n = 55) from JAX deployment 2 MARU data classified as short-beaked common dolphin, short-finned pilot whale or striped dolphin using a random-forest classifier in PAMGuard's ROCCA whistle classification module. Note that no encounters were classified as bottlenose or Atlantic spotted dolphins.

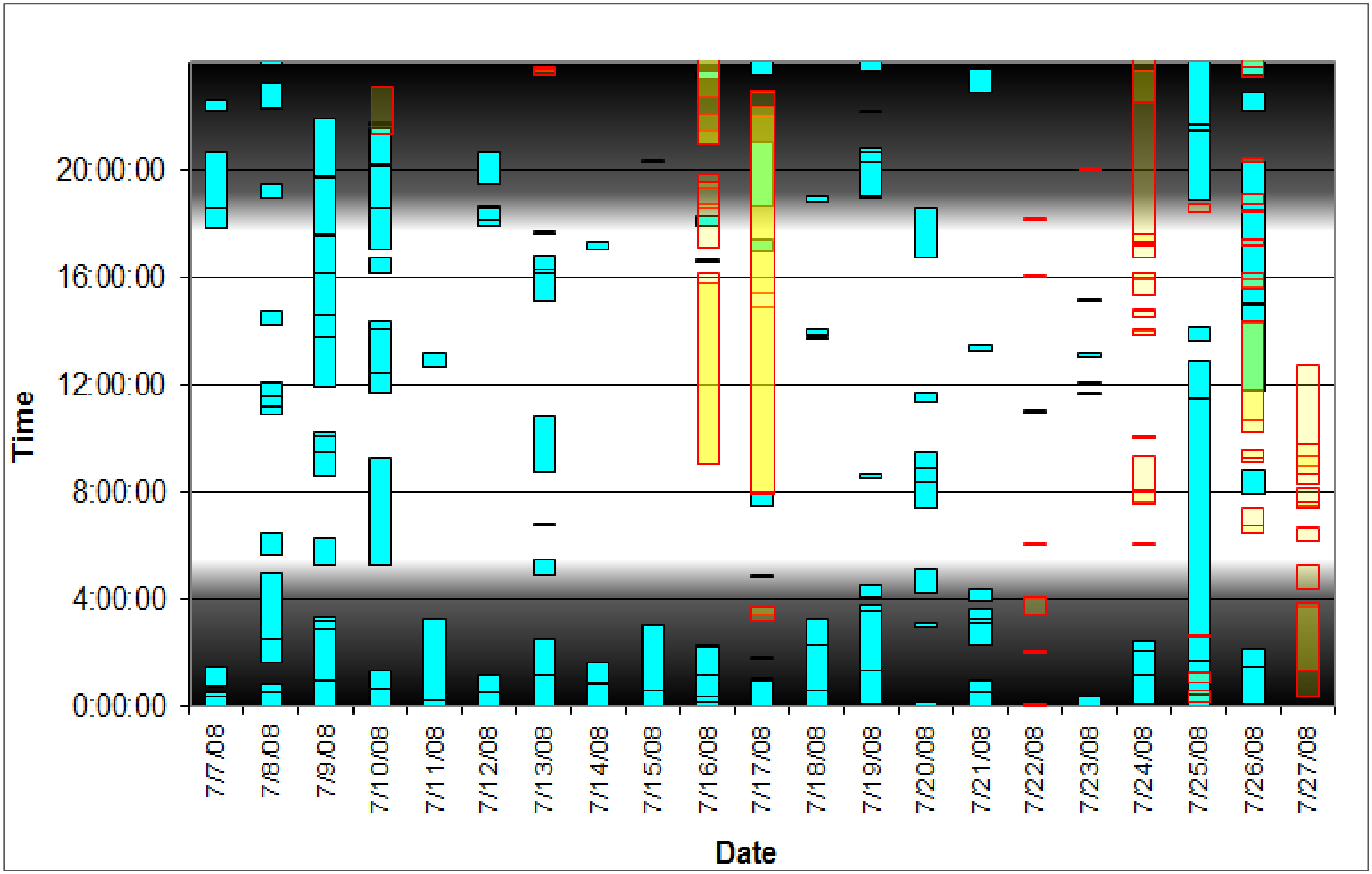

Figure 6. All delphinid acoustic encounters and sonar events recorded with the OB MARUs, by time of day (y-axis) and date (x-axis). Delphinid acoustic encounters are shown in teal. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., a sonar event occurring at multiple MARU sites). Black and white shading represents average periods of daylight (white) and darkness (black) for the deployment period.

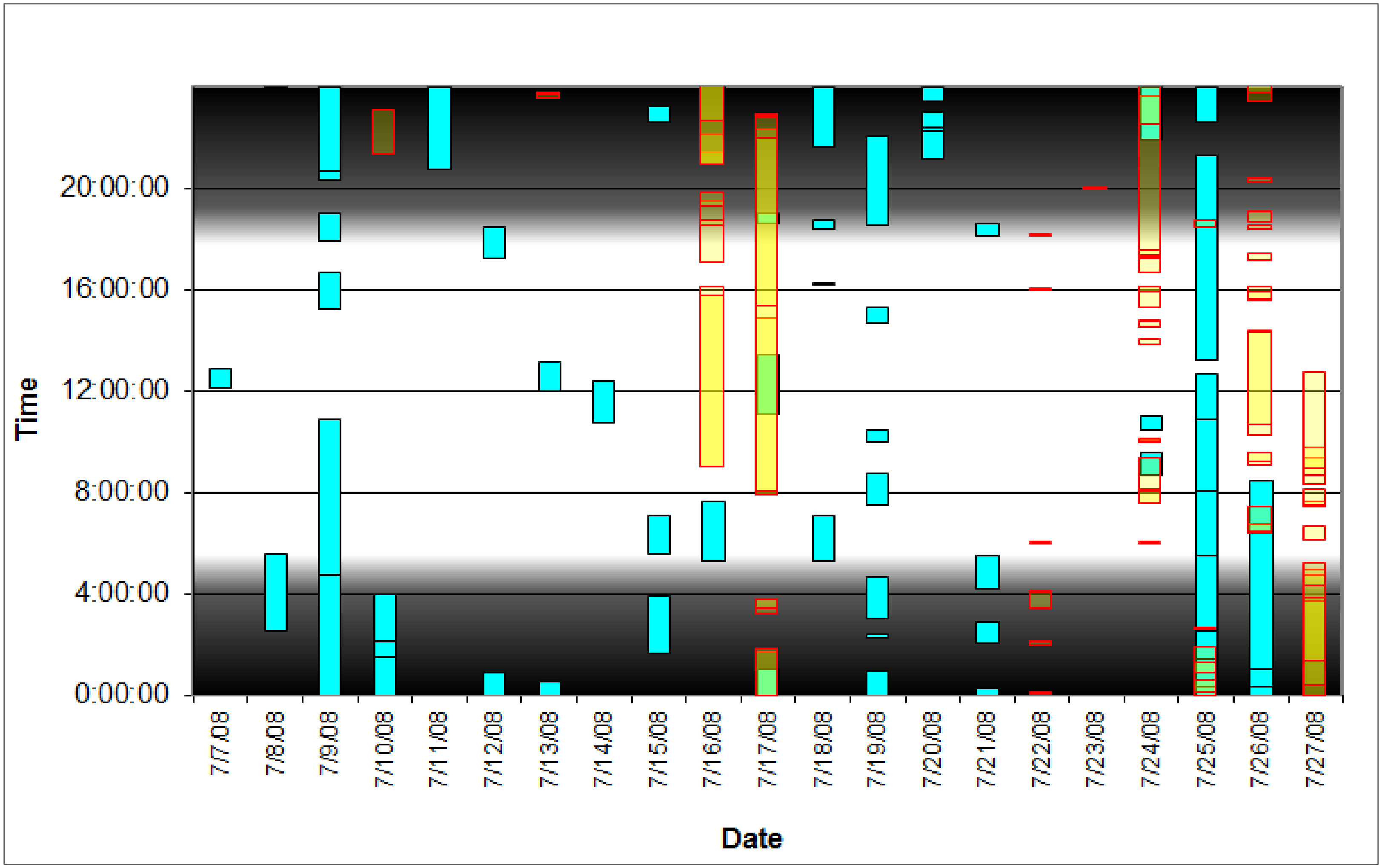

Figure 7. Short-beaked common dolphin whistle encounters and sonar events recorded with the OB MARUs, by time of day (y-axis) and date (x-axis). Whistle encounters are shown in. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Black and white shading represents average daylight (white) and darkness (black) for the deployment period.

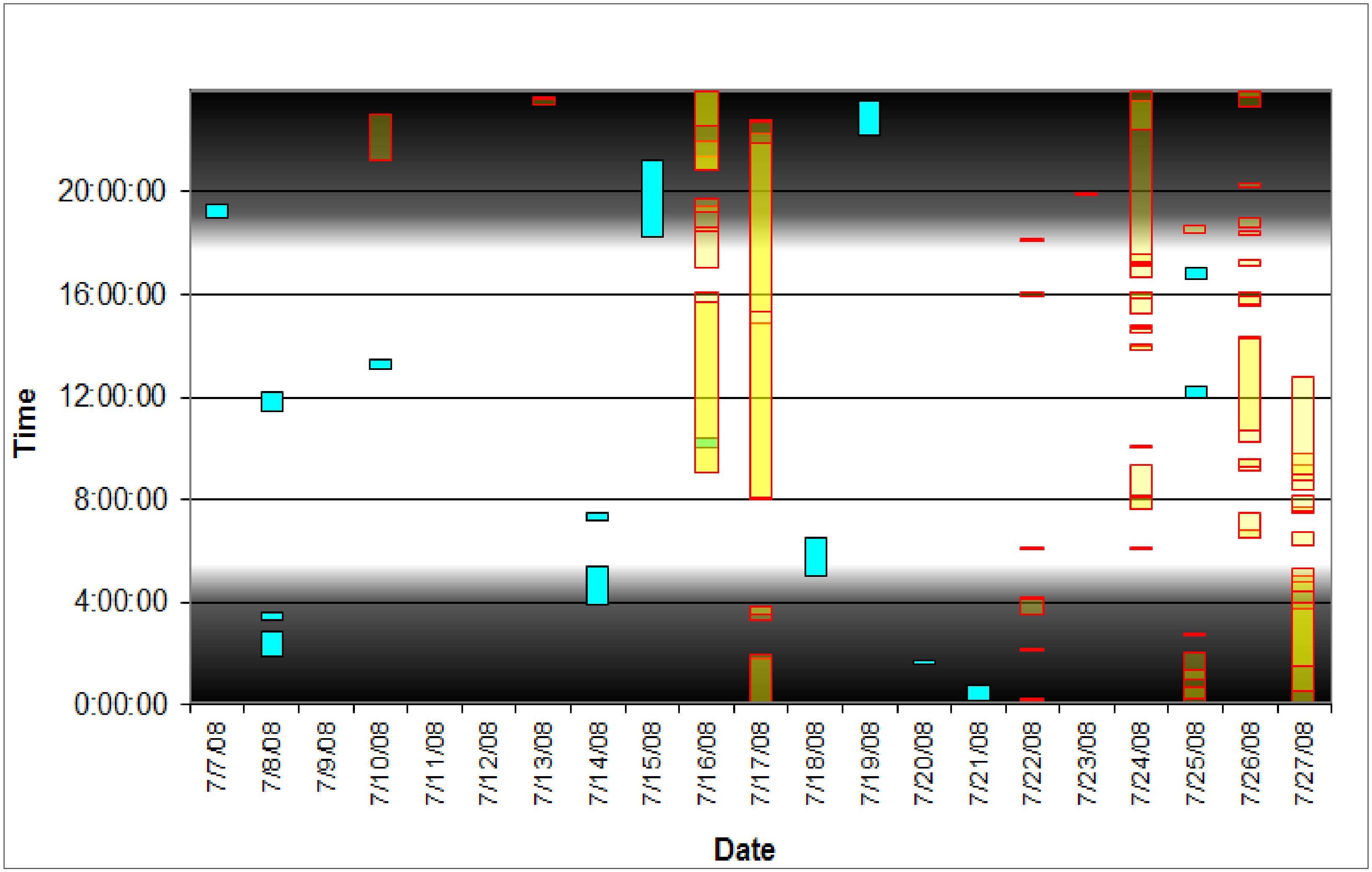

Figure 8. Short-finned pilot whale whistle encounters and sonar events recorded with the OB MARUs, by time of day (y-axis) and date (x-axis). Whistle encounters are shown in teal. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Black and white shading represents average daylight (white) and darkness (black) for the deployment period.

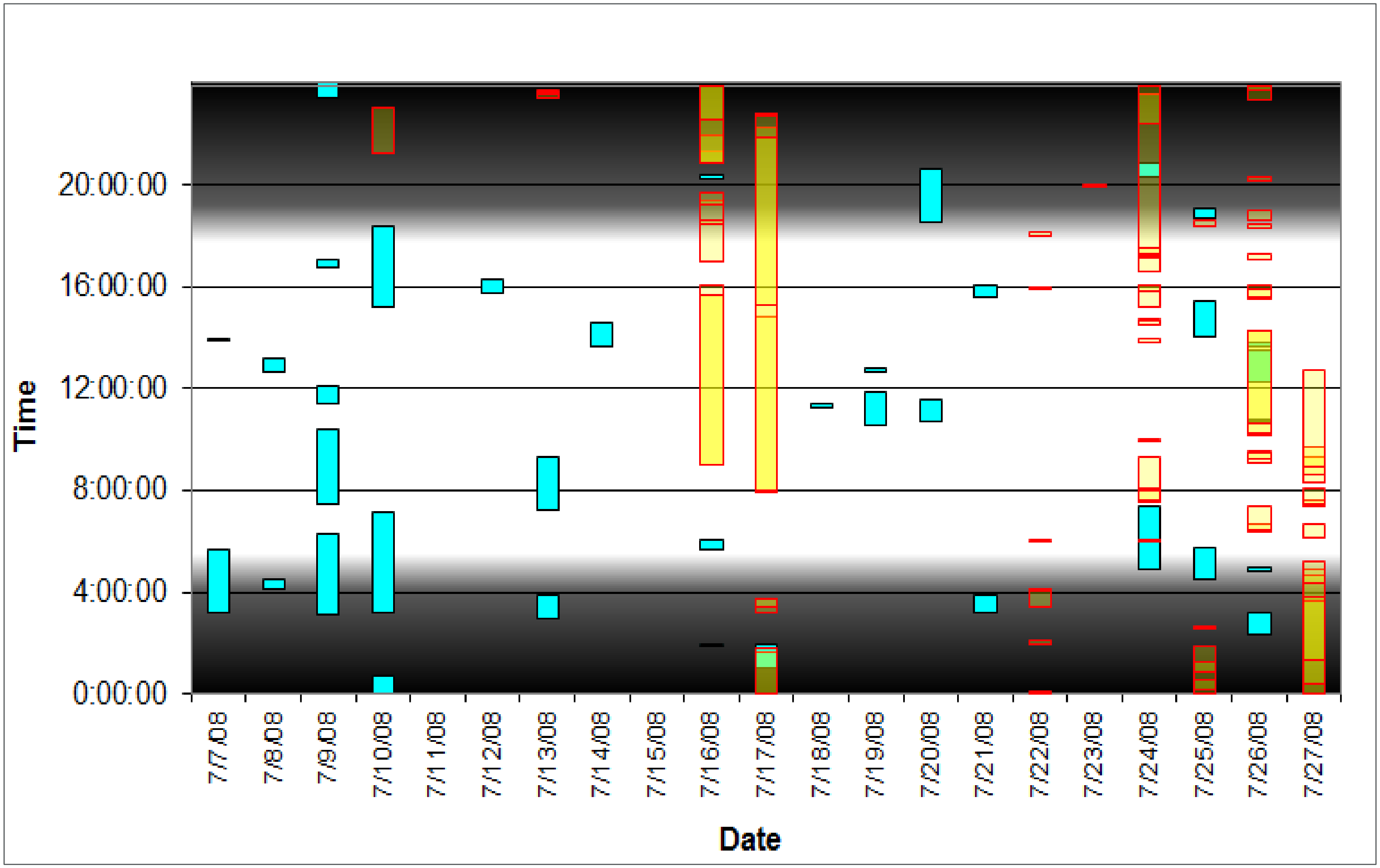

Figure 9. Striped dolphin whistle encounters and sonar events recorded with the OB MARUs, by time of day (y-axis) and date (x-axis). Whistle encounters are shown in teal. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Black and white shading represents average daylight (white) and darkness (black) for the deployment period.

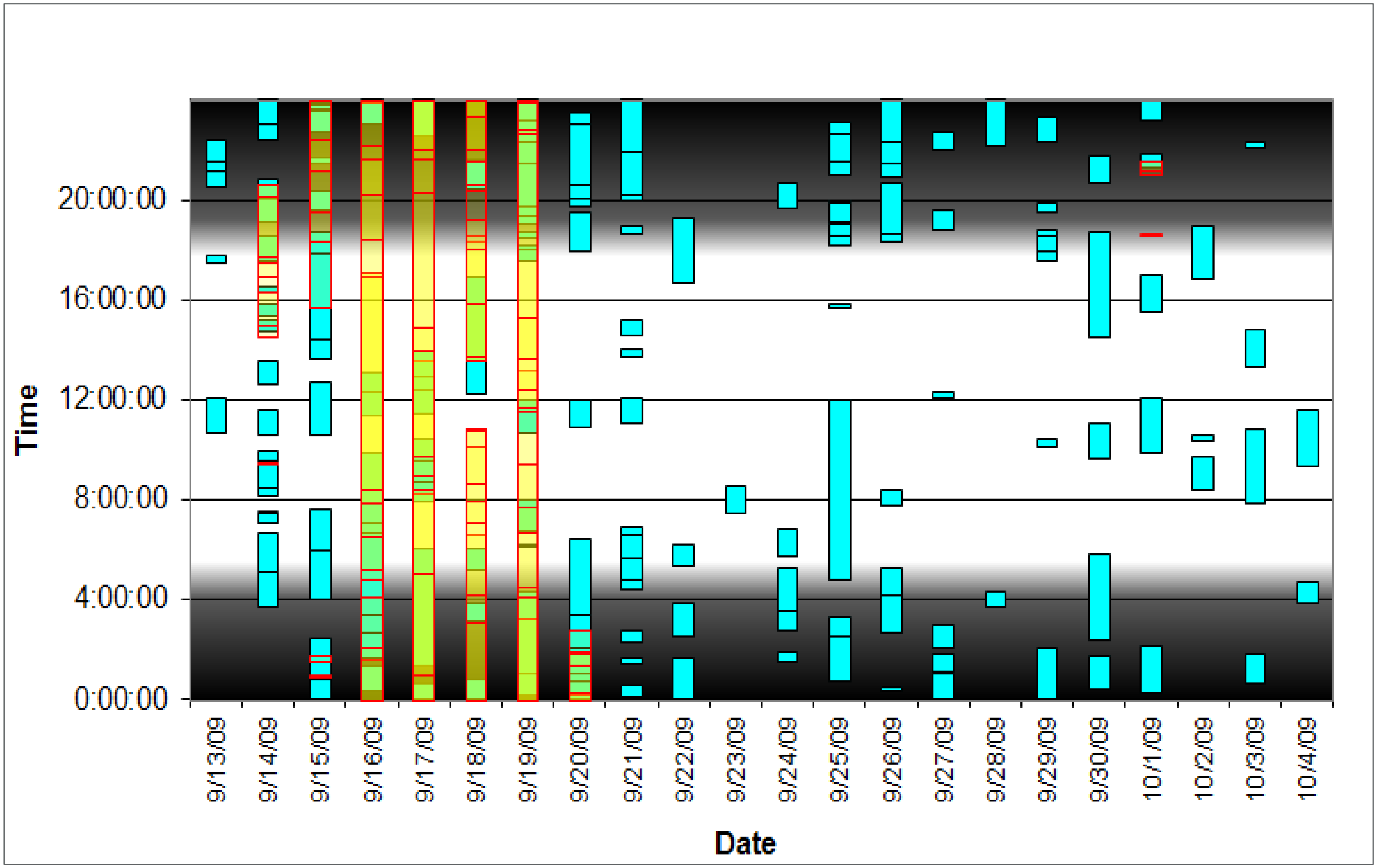

Figure 10. All delphinid acoustic encounters and sonar events recorded during JAX MARU deployment 1, by time of day (y-axis) and date (x-axis). Whistle encounters are shown in teal. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Shading represents average daylight (white) and darkness (black) for the deployment period.

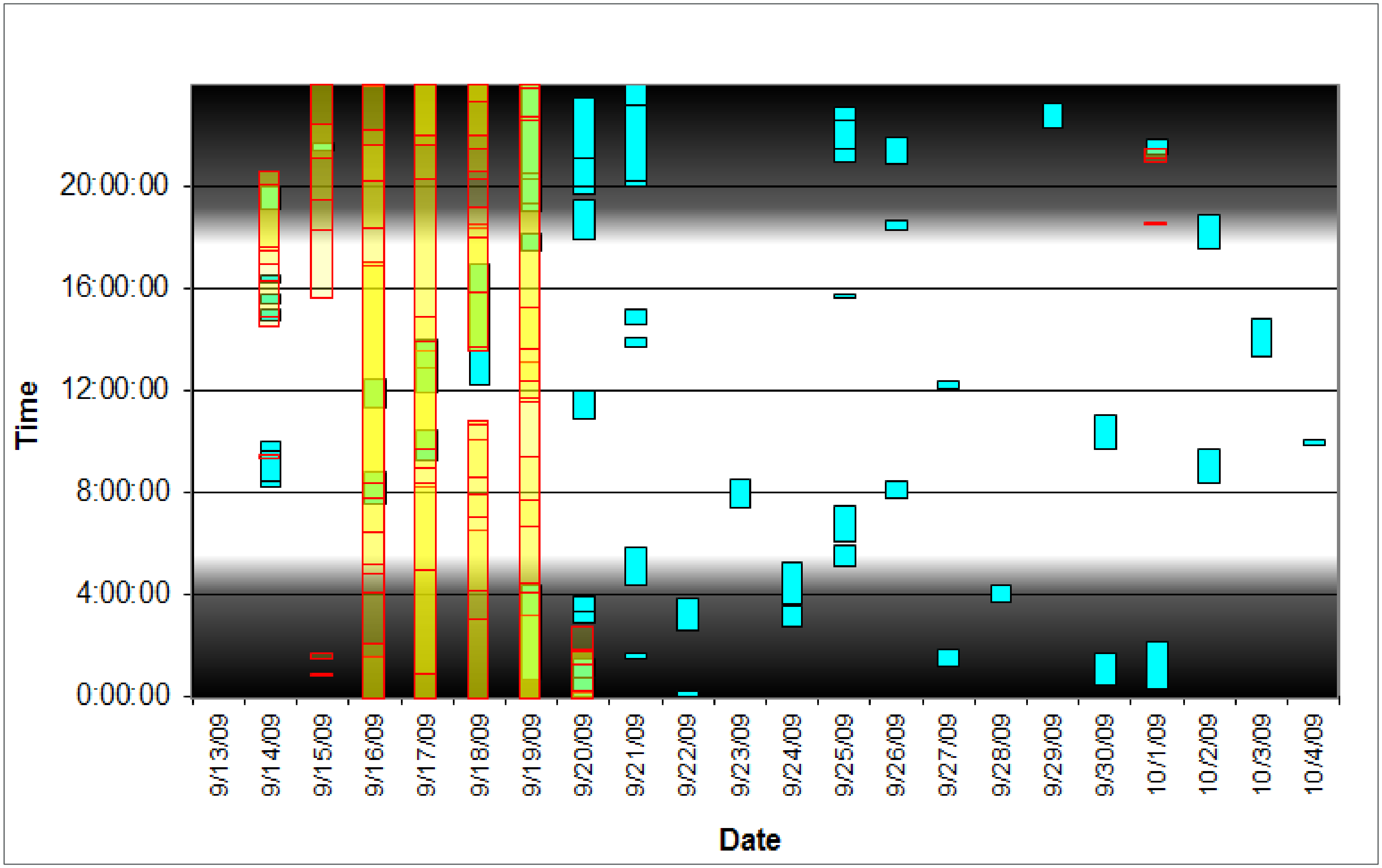

Figure 11. Short-beaked common dolphin whistle encounters and sonar events recorded during JAX MARU deployment 1, by time of day (y-axis) and date (x-axis). Whistle encounters are shown in teal. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Black and white shading represents average daylight (white) and darkness (black) for the deployment period.

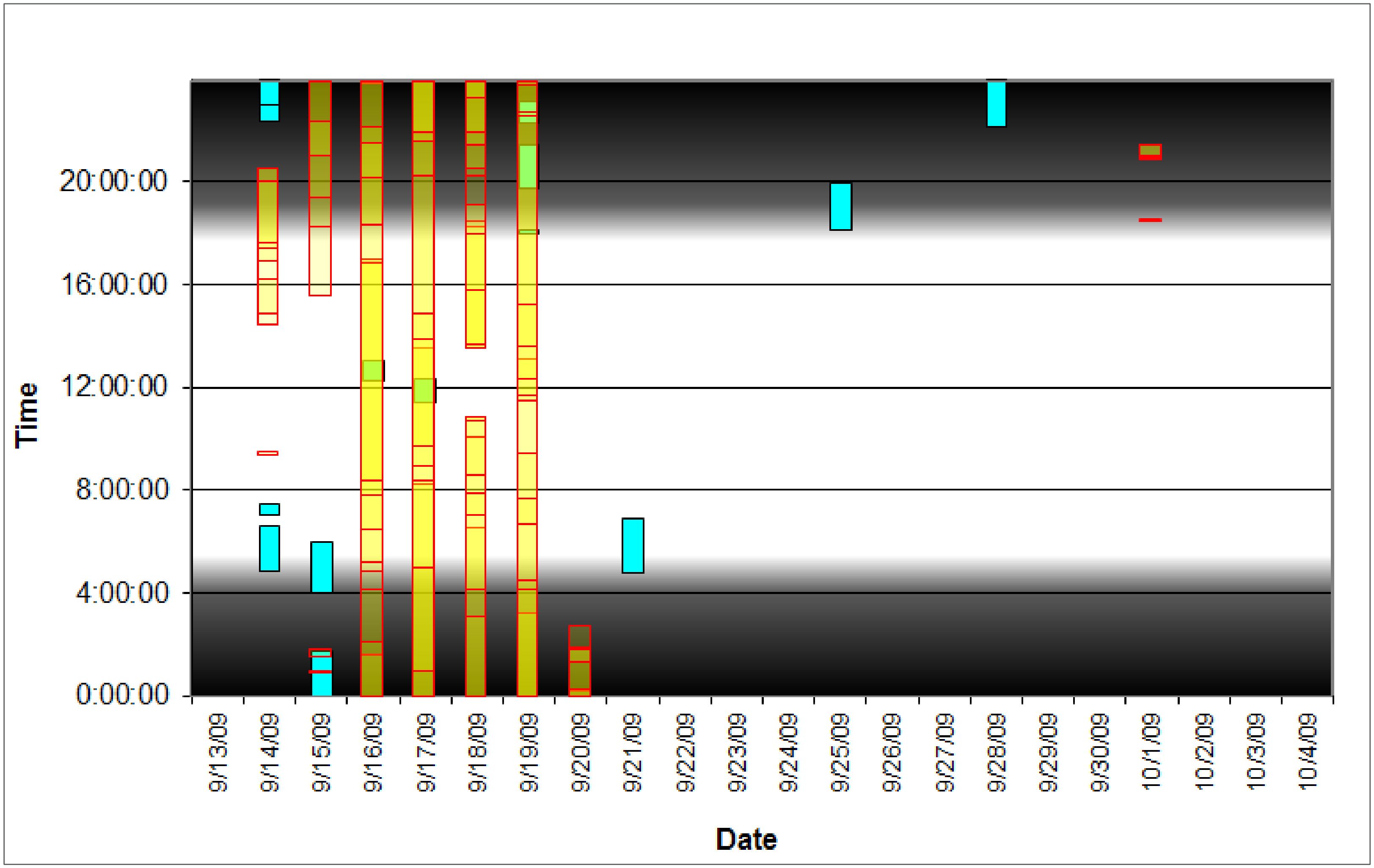

Figure 12. Short-finned pilot whale whistle encounters and sonar events recorded during JAX MARU deployment 1, by time of day (y-axis) and date (x-axis). Whistle encounters are shown in teal. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Black and white shading represents average daylight (white) and darkness (black) for the deployment period.

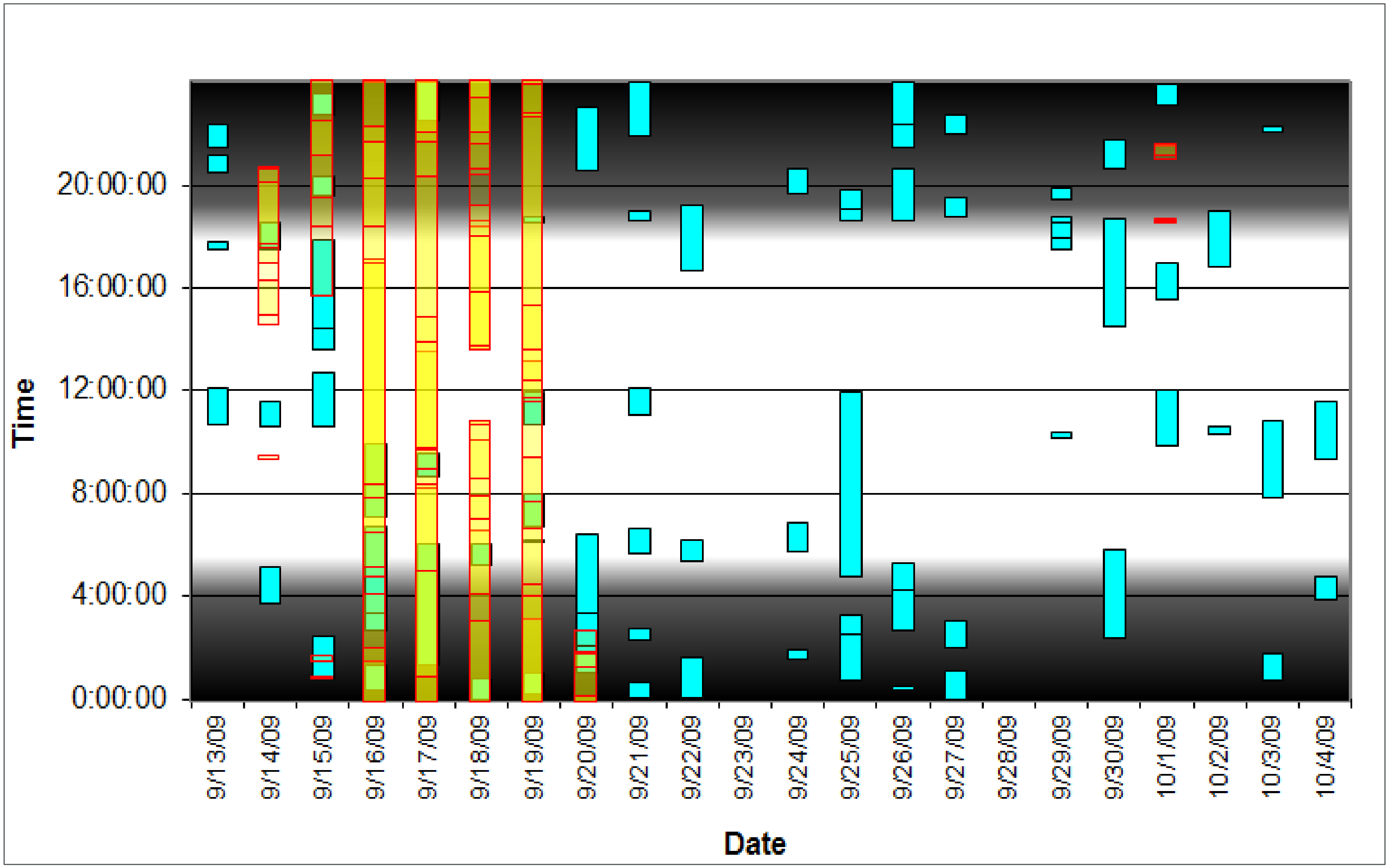

Figure 13. Striped dolphin whistle encounters and sonar events recorded during JAX MARU deployment 1, by time of day (y-axis) and date (x-axis). Whistle encounters are shown in teal. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Black and white shading represents average daylight (white) and darkness (black) for the deployment period.

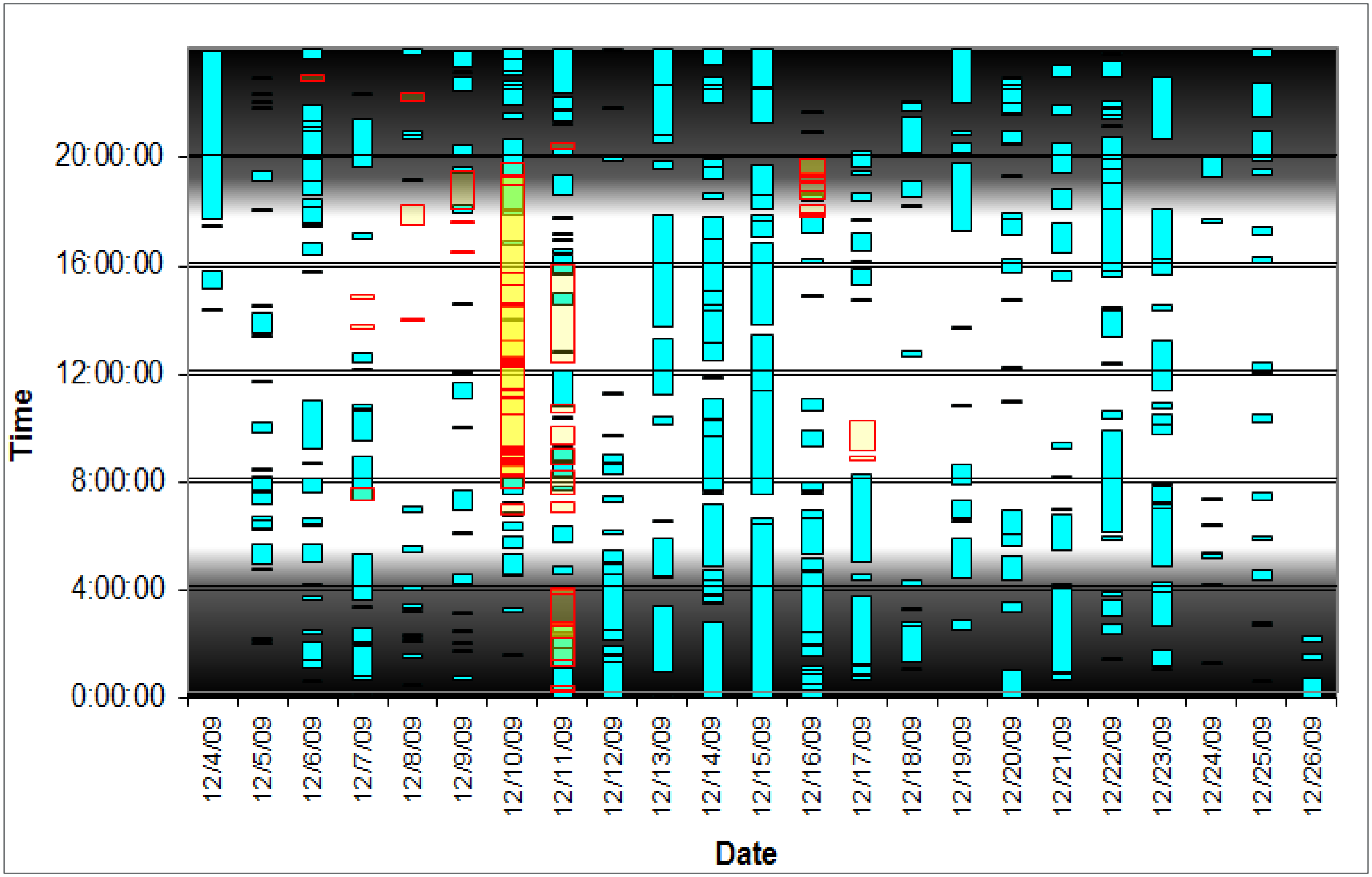

Figure 14. All delphinid acoustic encounters and sonar events recorded during the JAX MARU deployment 2, by time of day (y-axis) and date (x-axis). Delphinid acoustic encounters are shown in teal. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Black and white shading represents average daylight (white) and darkness (black) for the deployment period.

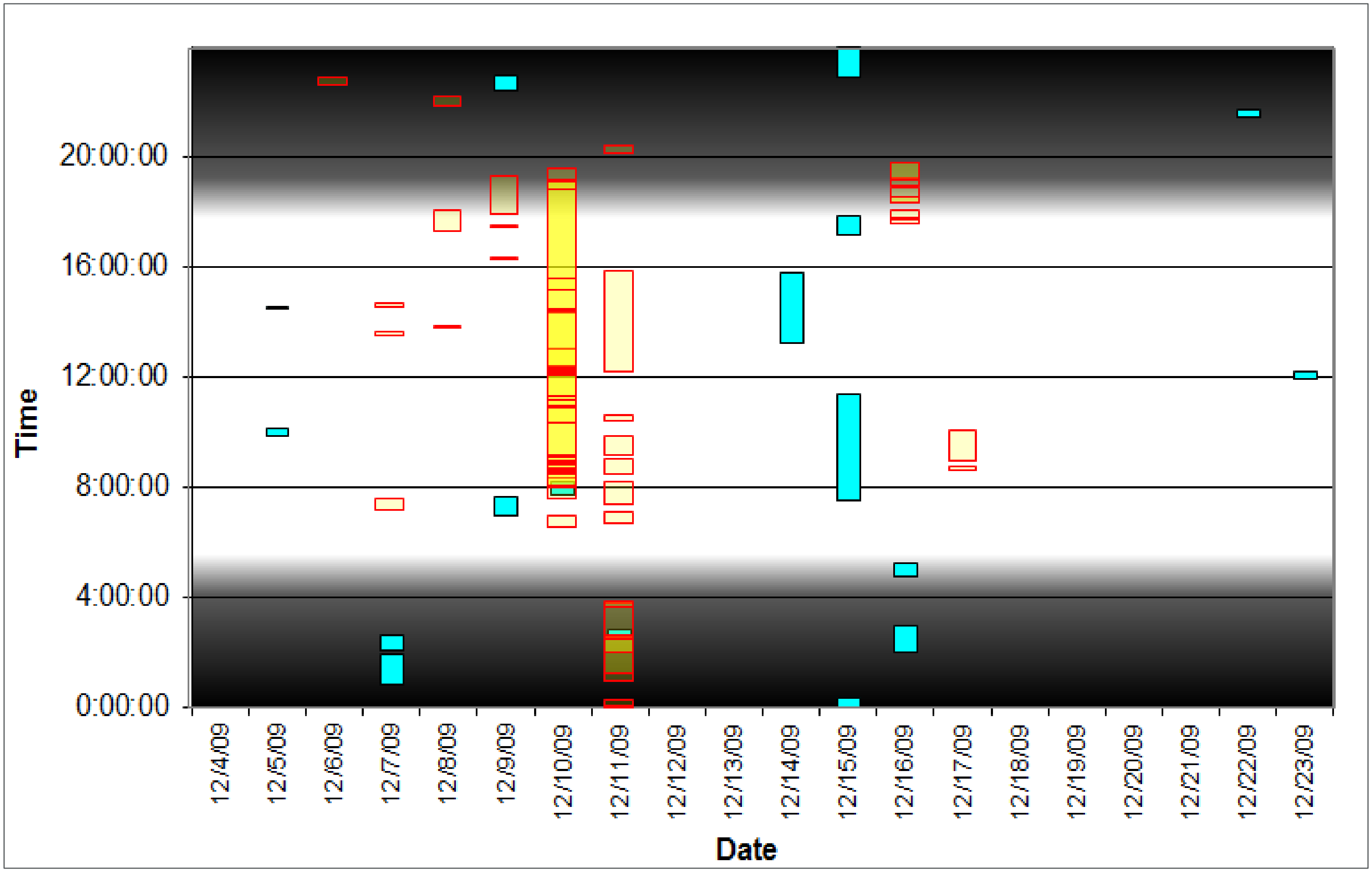

Figure 15. Short-beaked common dolphin whistle encounters and sonar events recorded during JAX MARU deployment 2, by time of day (y-axis) and date (x-axis). Whistle encounters are shown in teal. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Black and white shading represents average daylight (white) and darkness (black) for the deployment period.

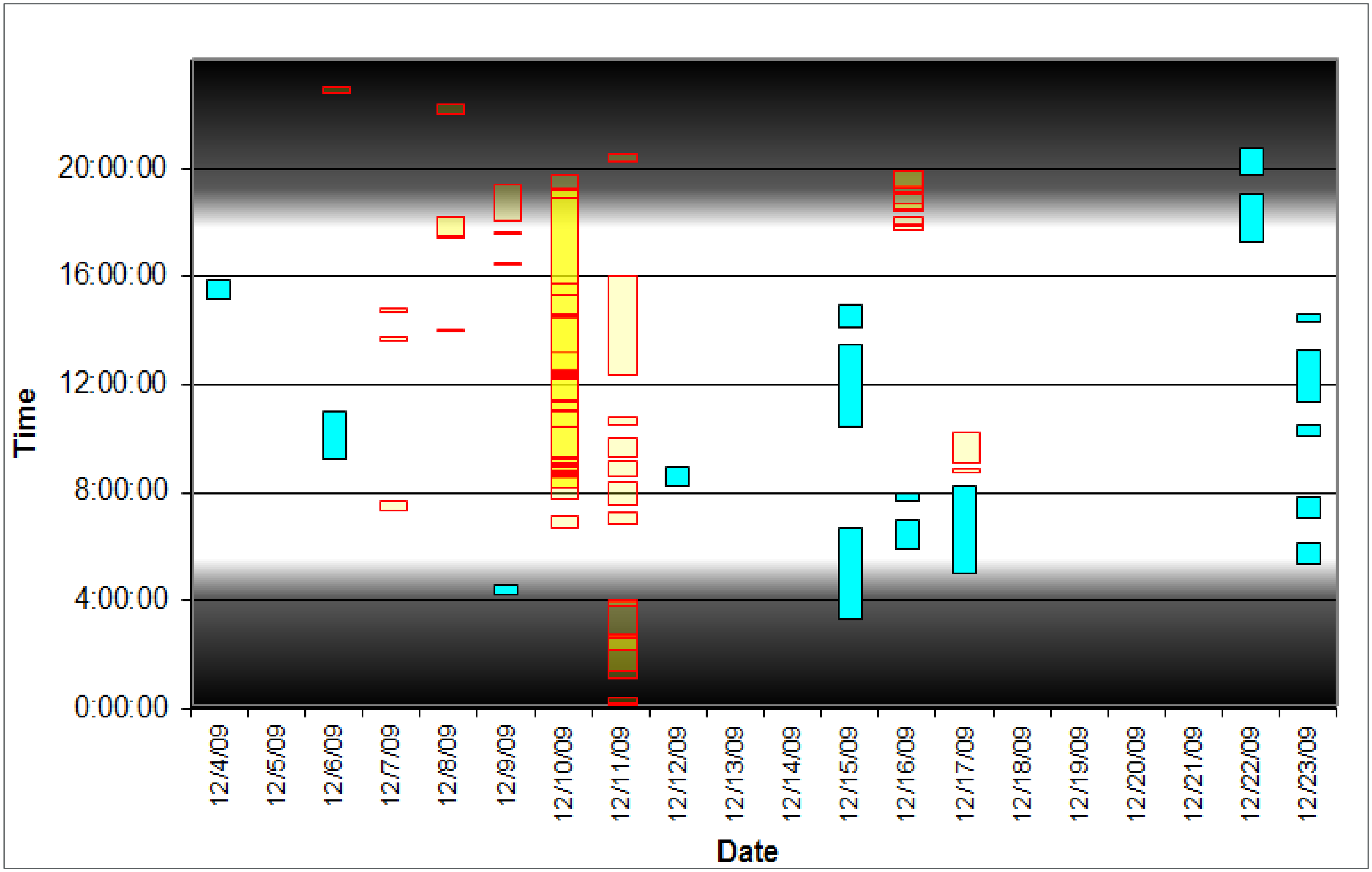

Figure 16. Short-finned pilot whale whistle encounters and sonar events recorded during JAX MARU deployment 2, by time of day (y-axis) and date (x-axis). Whistle encounters are shown in. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Black and white shading represents average daylight (white) and darkness (black) for the deployment period.

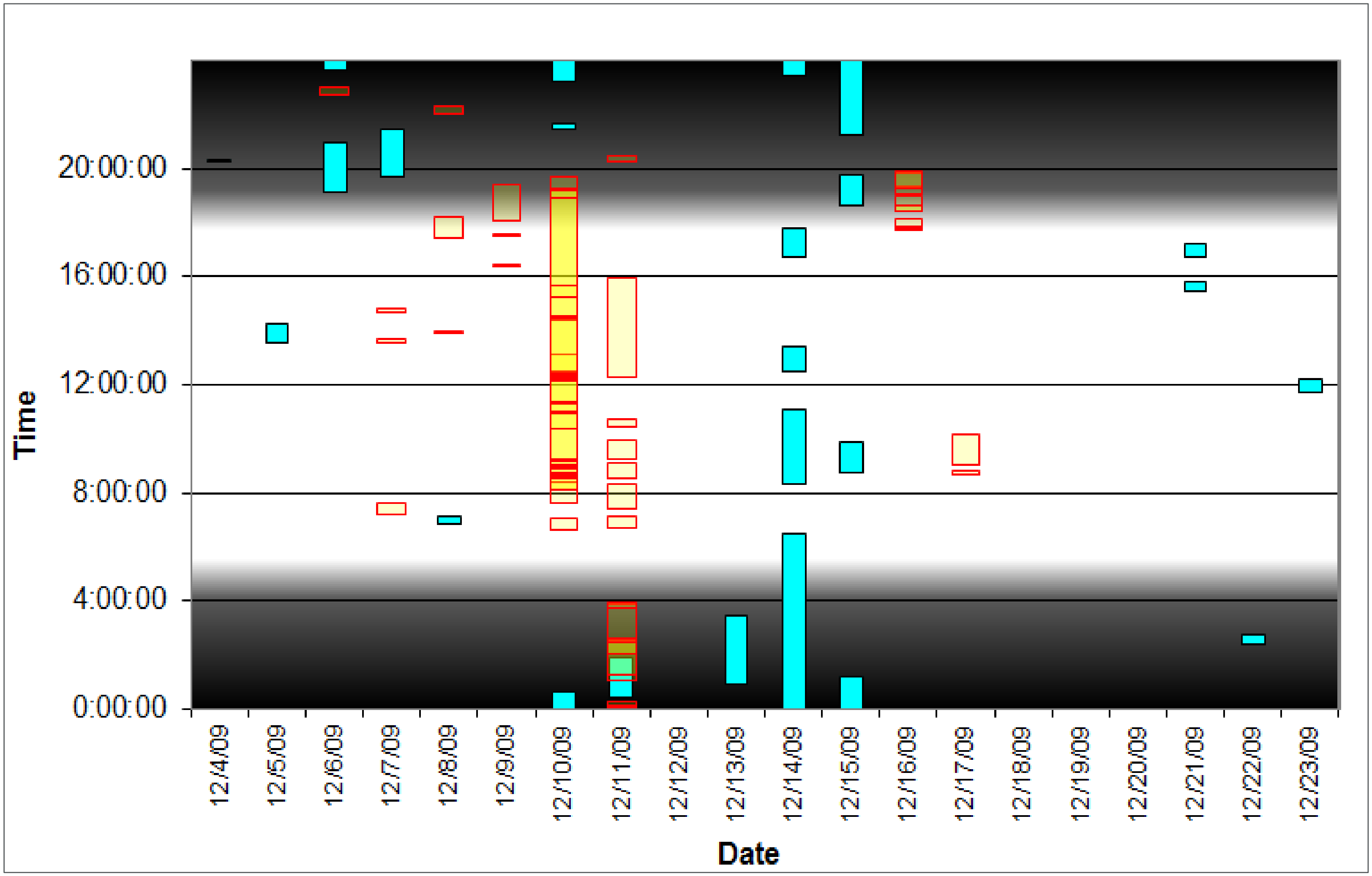

Figure 17. Striped dolphin whistle encounters and sonar events recorded during JAX MARU deployment 2, by time of day (y-axis) and date (x-axis). Whistle encounters are shown in teal. Sonar events are shown in yellow. Bars with overlapping colors are representative of event overlap (i.e., an event occurring at multiple MARU sites). Black and white shading represents average daylight (white) and darkness (black) for the deployment period.

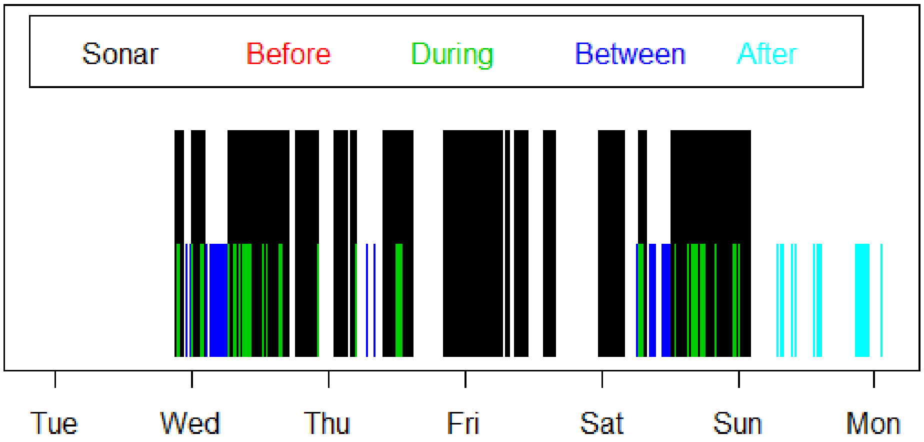

Figure 18. Occurrence of acoustic sub-events color-coded in relation to their occurrence of sonar pings (black) for the JAX site 7. All sonar pings belonged to the same sonar exercise as no gaps in sonar pings longer than 48 hours occurred during the recording at this site.

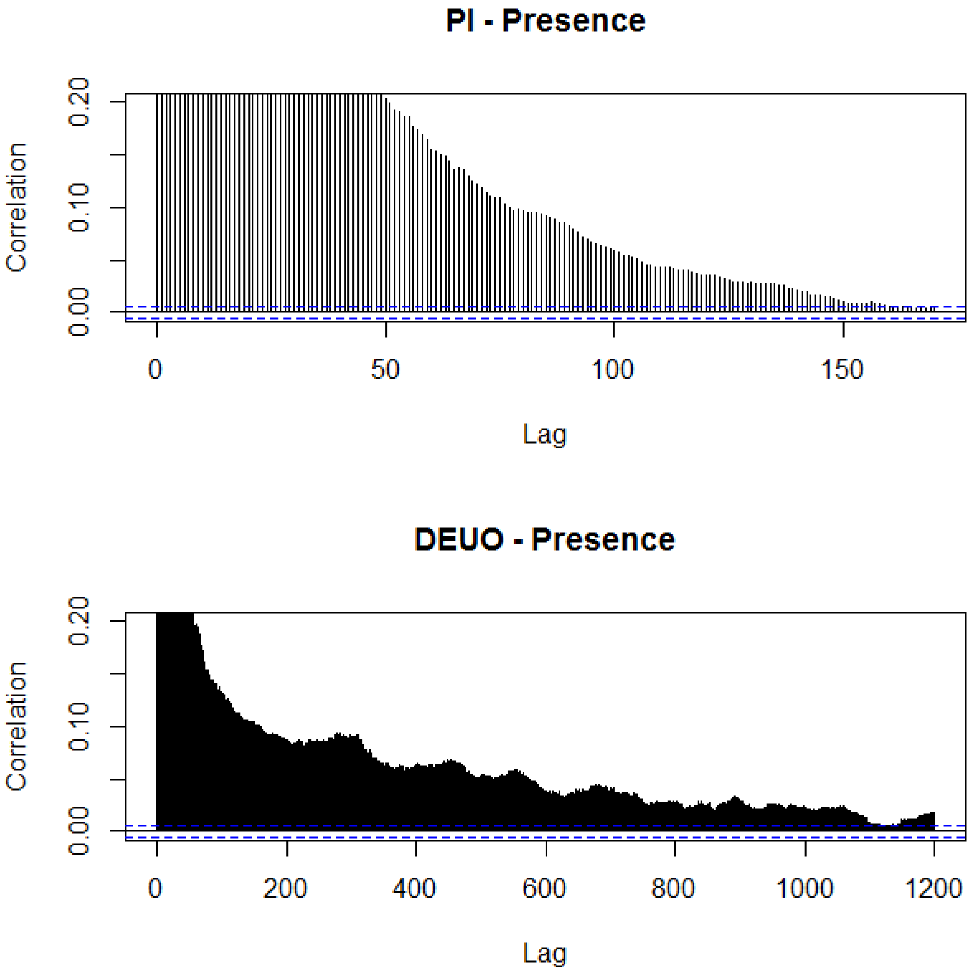

Figure 19. Autocorrelation of Pearson's residuals from presence models for pilot whales (upper) and the DEUO species group (lower) including 95 percent confidence intervals around zero autocorrelation (blue dashed line). Note that the y-axes were limited to 0.2 for illustrative purposes.

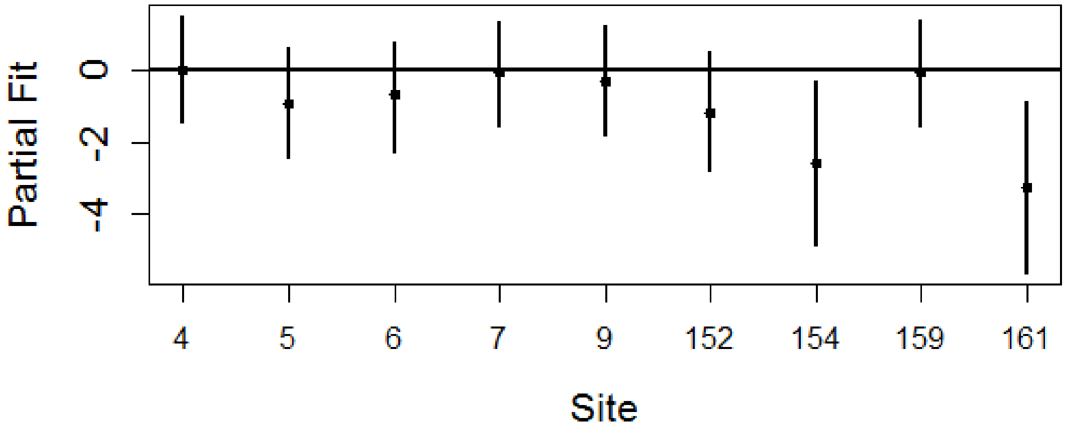

Figure 20. Partial fit plots for presence model for pilot whales (note that the partial fit is given on the scale of the logit-link function). The covariate retained in the final model was *Site*.

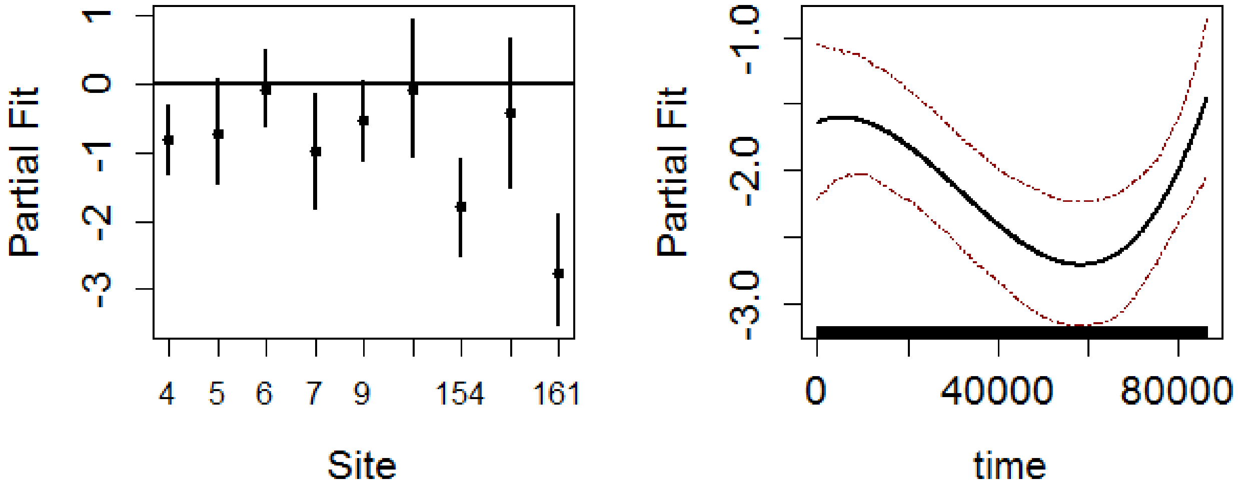

Figure 21. Partial fit plots for presence model for DEUO species group (note that the partial fit is given on the scale of the logit-link function). Covariates in the final model were *Site* and *Time*. For the smooth term, tick marks on the upper side of the x-axis indicate the observed values for the covariate.

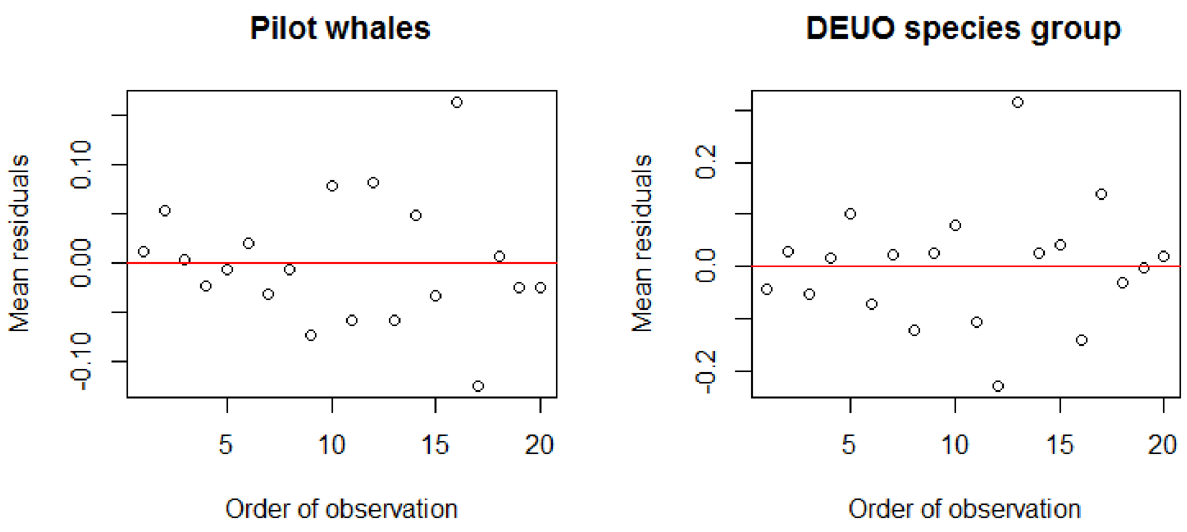

Figure 22. Means of binned fitted values versus means of corresponding residuals from presence models for pilot whales (left) and the DEUO species group (right). Binning occurred by splitting the fitted values into 20 equally sized bins in ascending order of observation.

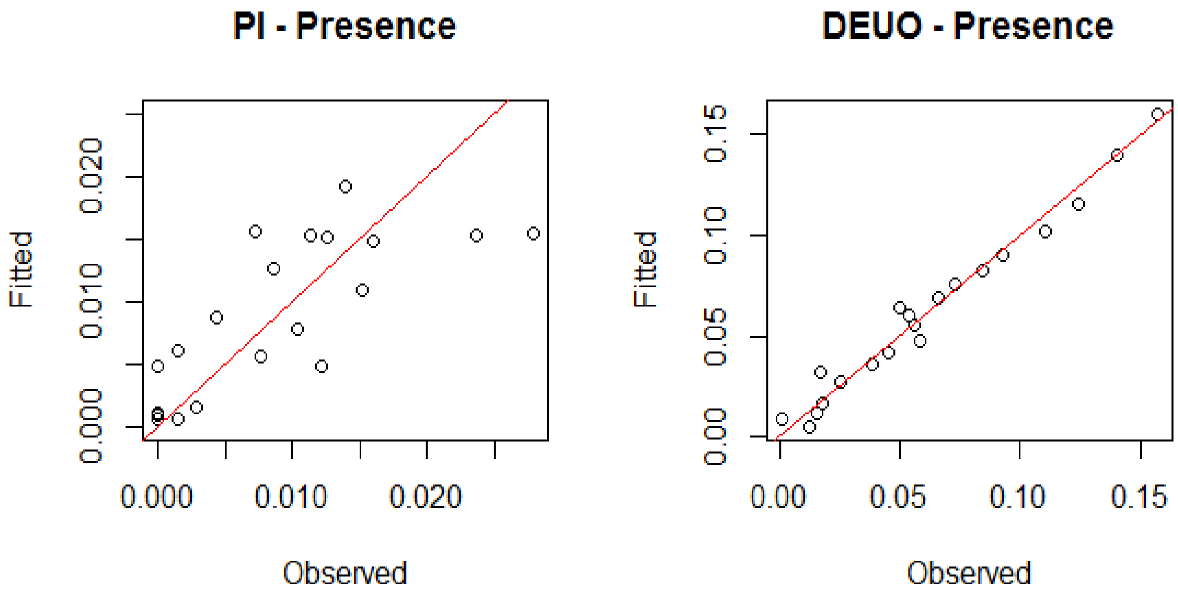

Figure 23. Mean observed versus mean fitted values from presence-of-vocalizations models for pilot whale (left) and DEUO species group (right). Note that observations and fitted values were combined into 20 equally sized bins in ascending order of fitted values for which the mean was calculated. The red lines indicate a perfect fit of the model to the observed data.

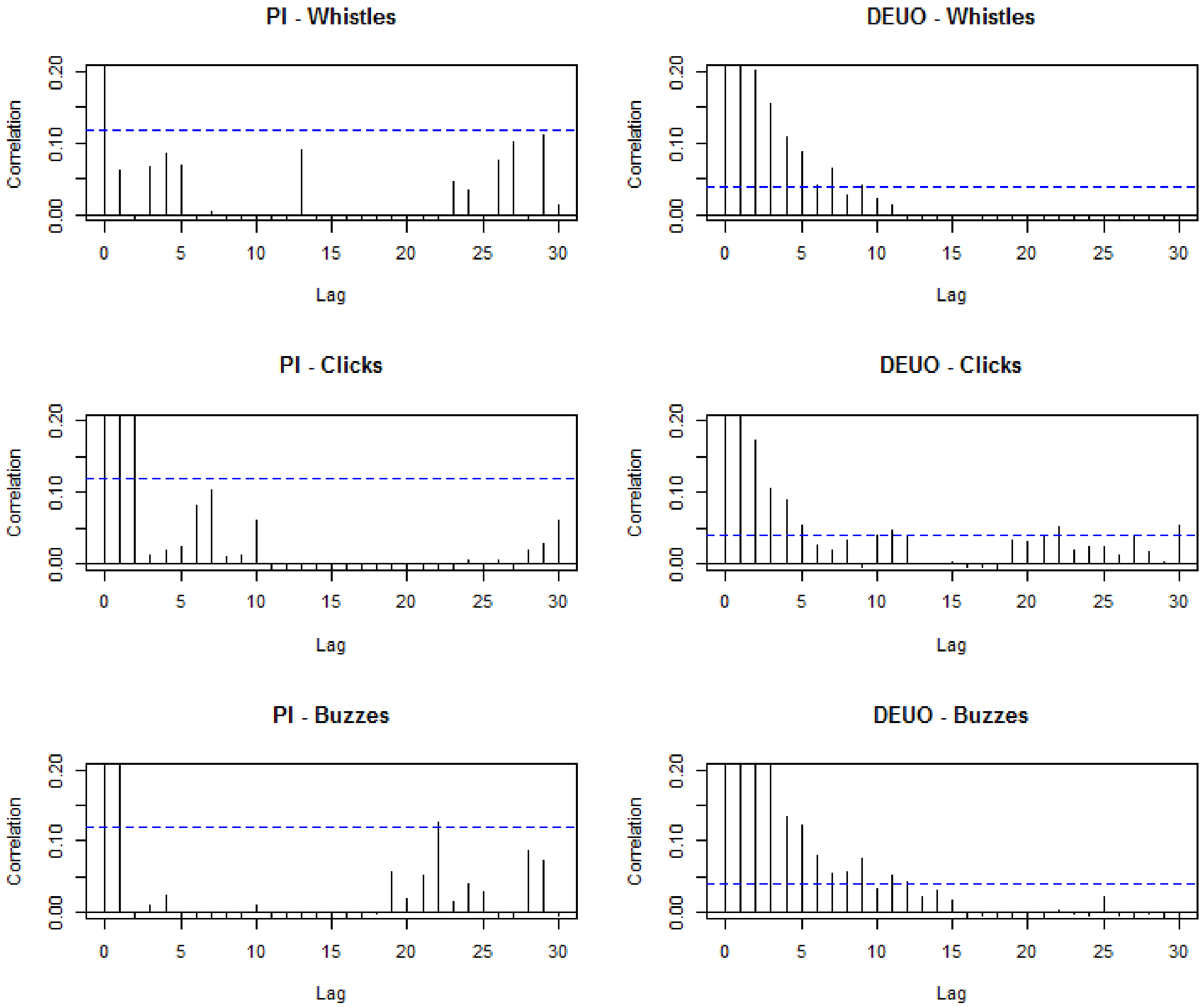

Figure 24. Autocorrelation of Pearson's residuals from presence of vocalization type given vocalization models including 95 percent confidence intervals around zero autocorrelation for pilot whales (left panels) and the DEUO species group (right panels). Vocalization types were whistles (top), clicks (middle) and buzzes (bottom).

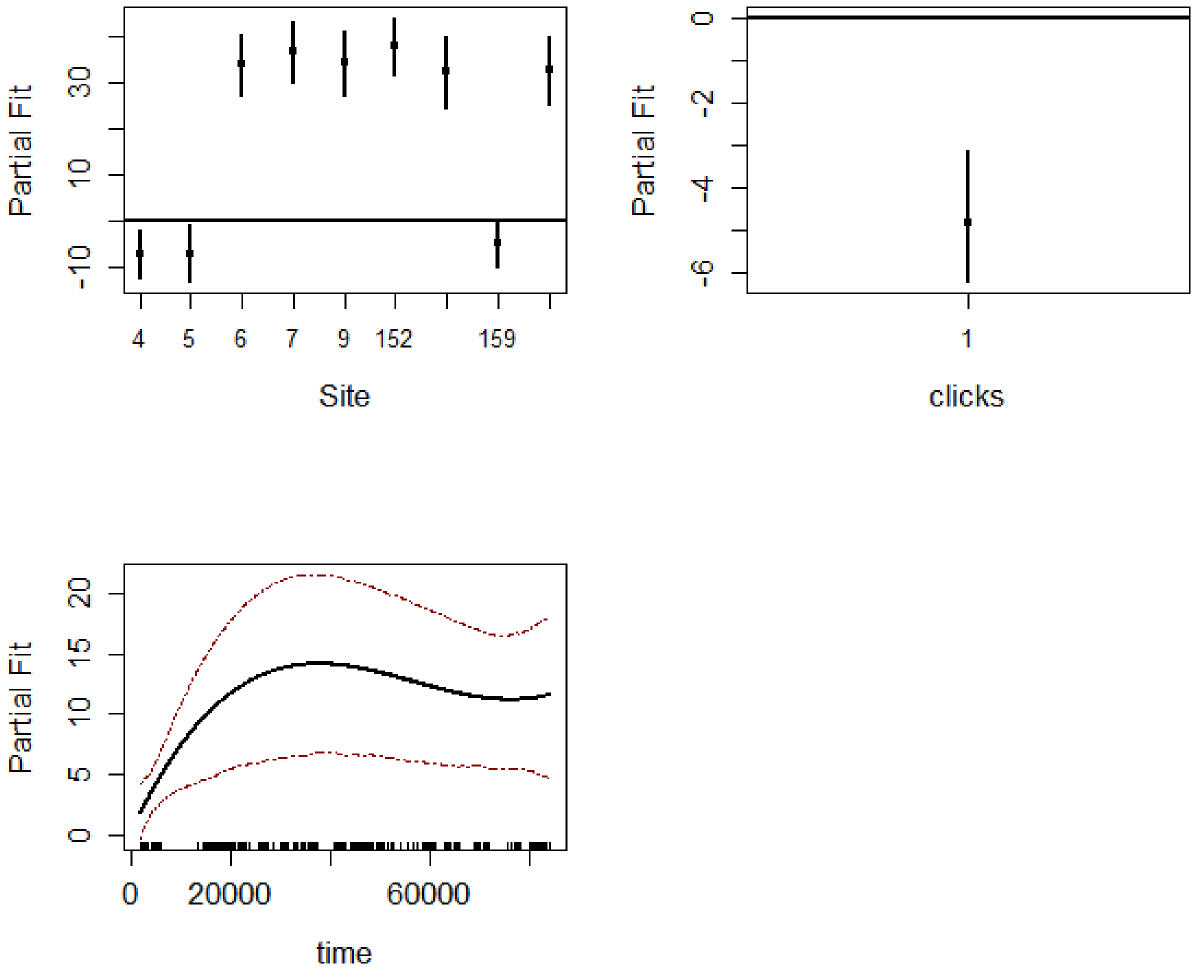

Figure 25. Partial fit for each model term for presence-of-whistles-given-acoustic-encounter models for pilot whales (note that the partial fit is given on the scale of the logit-link function). Covariates retained in the best model were *Site*, *Presence of clicks* and the polynomial smooth for *Time*. For covariate *Time*, tick marks on the upper side of the x-axis indicate the observed values for the covariate.

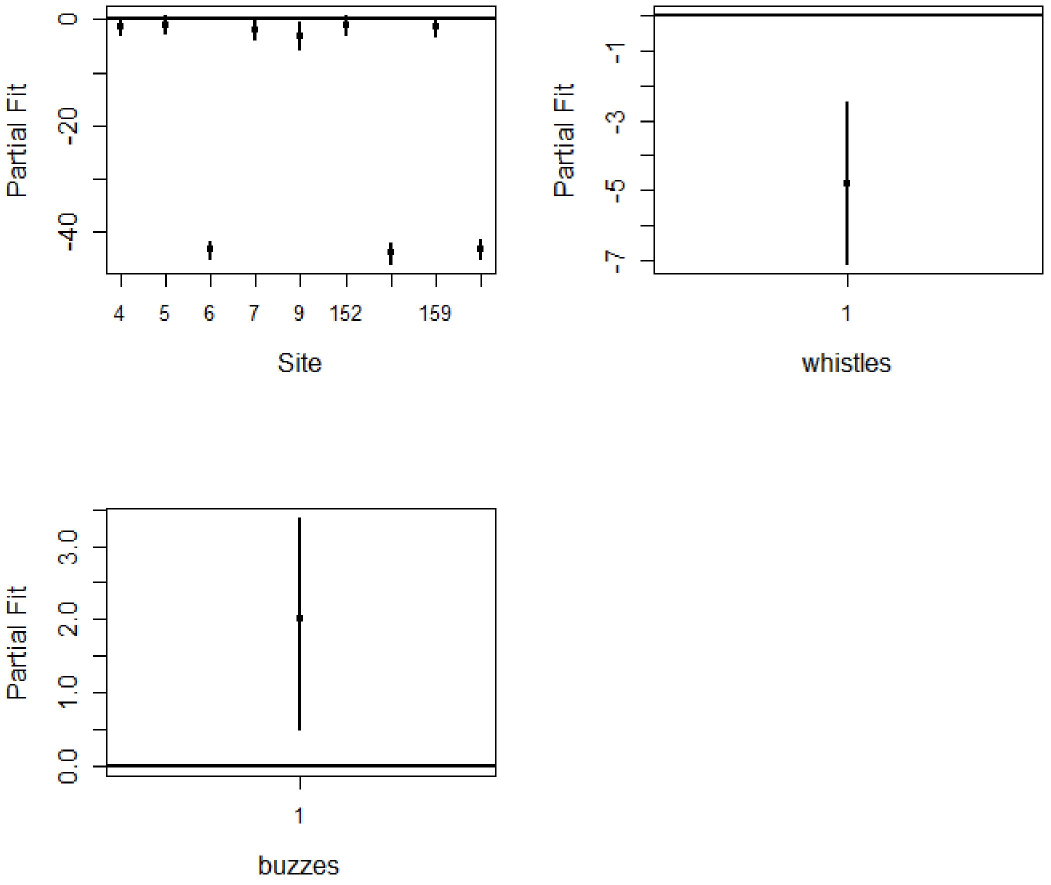

Figure 26. Partial fit for each model term for presence-of-clicks-given-acoustic-encounters models for pilot whales (note that the partial fit is given on the scale of the logit-link function). Covariates retained in the best model were *Site*, *Presence of whistles* and *Presence of buzzes*.

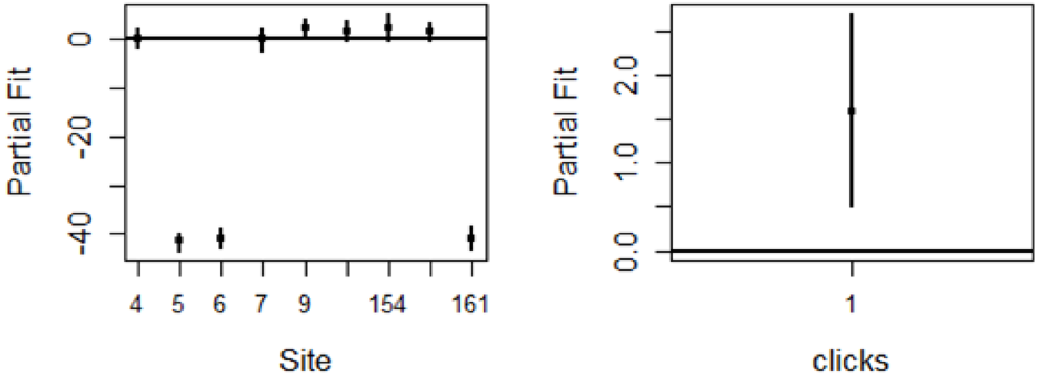

Figure 27. Partial fit for each model term for presence-of-buzzes-given-acoustic-encounter models for pilot whales (note that the partial fit is given on the scale of the logit-link function). Covariates retained in the best model were *Site* and *Presence of clicks.*

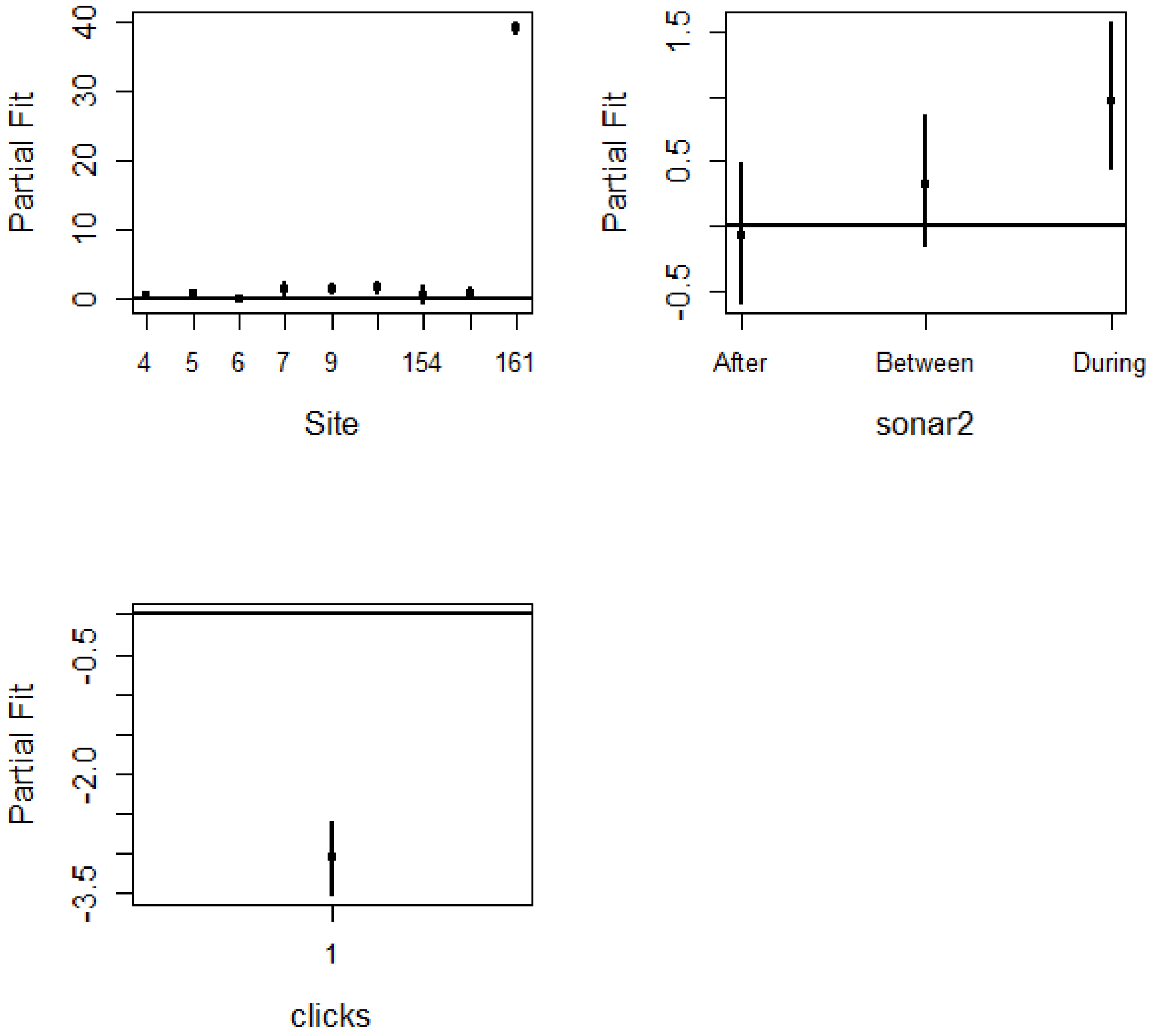

Figure 28. Partial fit for each model term for presence-of-whistles-given-acoustic-encounter models for the DEUO species group (note that the partial fit is given on the scale of the logit-link function). Covariates retained in the best model were *Site, Sonar* and *Presence of clicks.*

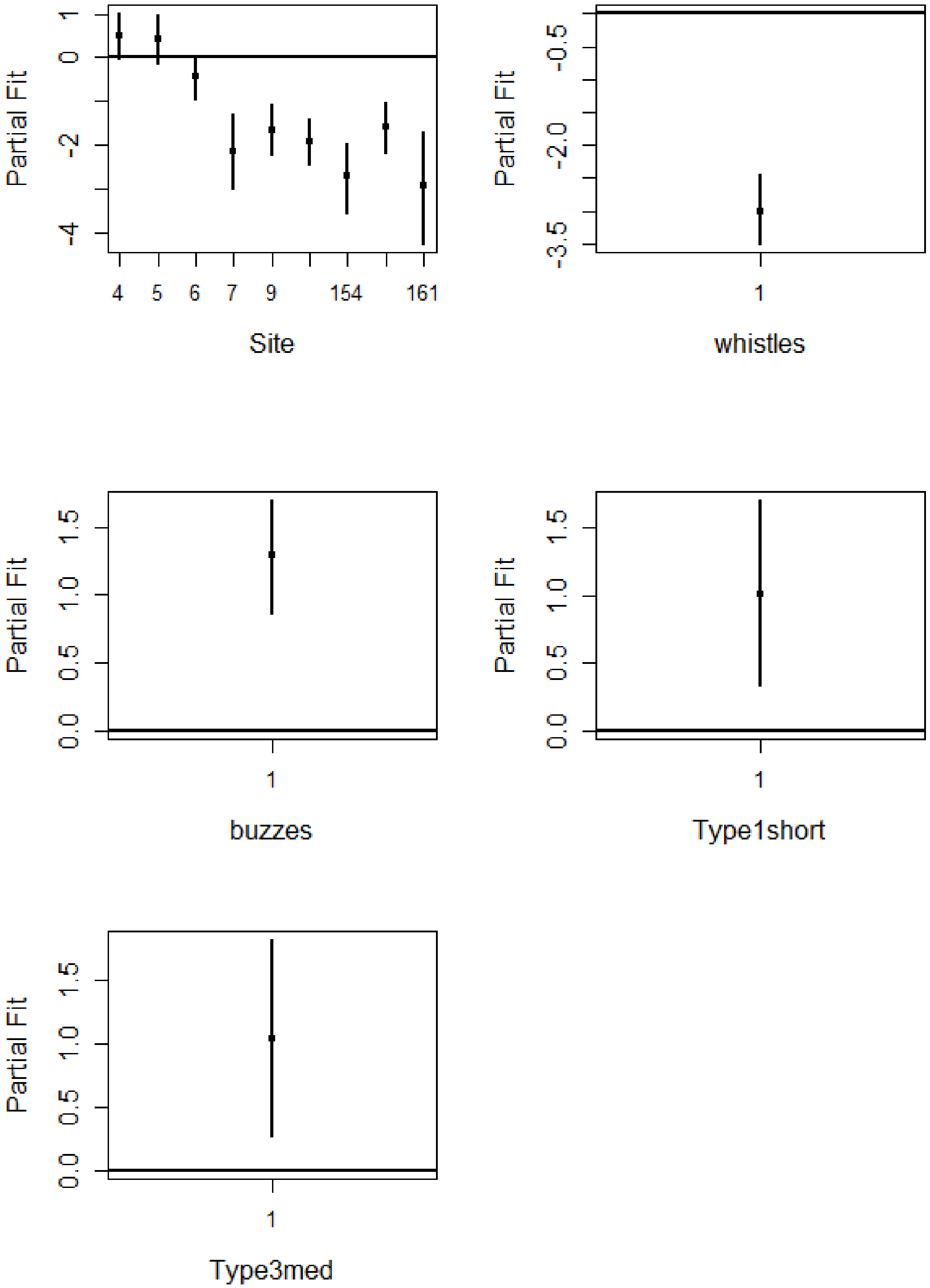

Figure 29. Partial fit for each model term for presence-of-clicks-given-acoustic-encounter models for the DEUO species group (note that the partial fit is given on the scale of the logit-link function). Covariates retained in the best models were *Presence of buzzes*, *Presence of Type 1-short sonar pings* and *Presence of Type 3-medium sonar pings*.

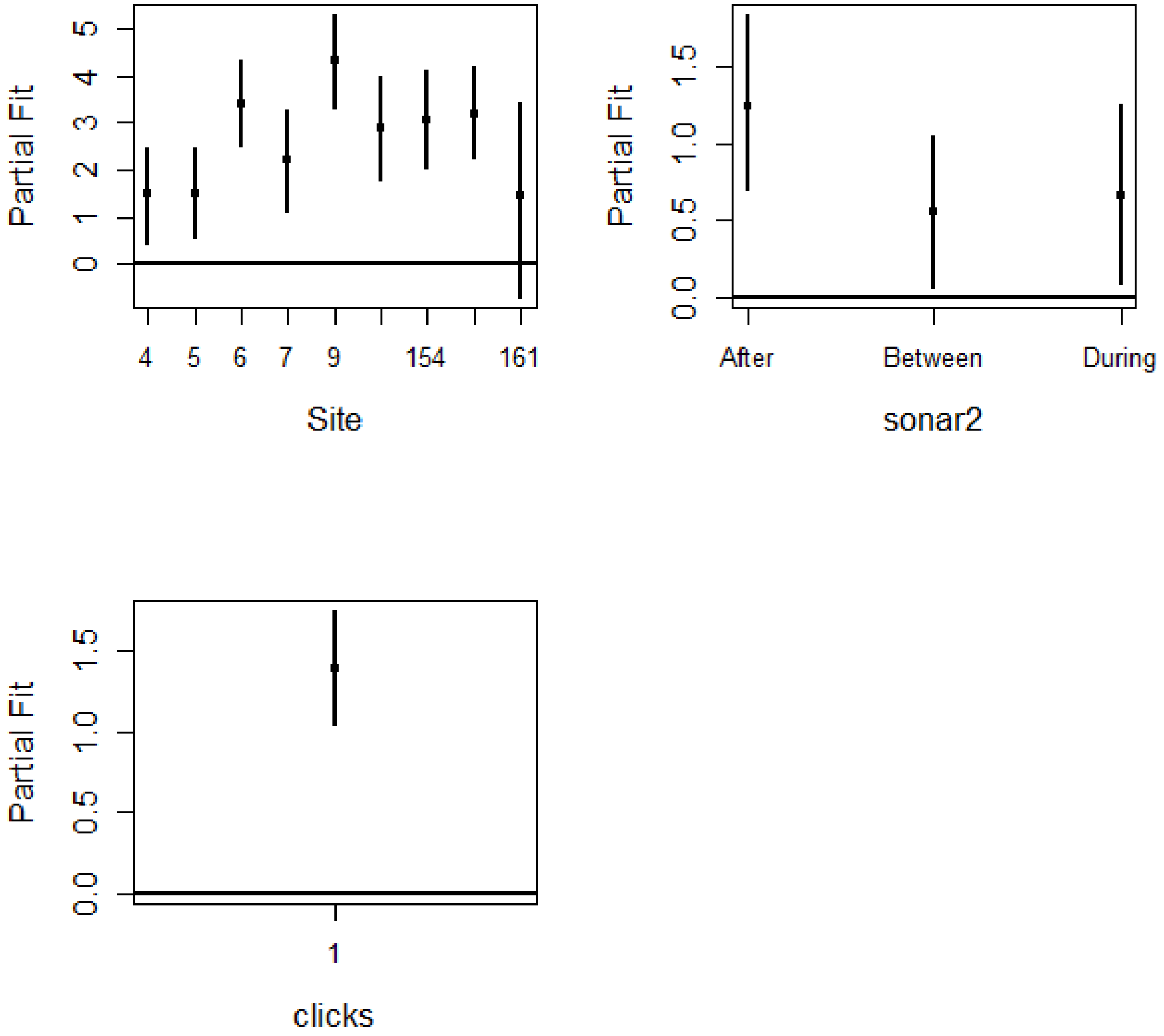

Figure 30. Partial fit for each model term for presence-of-buzzes-given-acoustic-encounter models for the DEUO species group (note that the partial fit is given on the scale of the logit-link function). Covariates retained in the best model were *Site*, *Sonar* and *Presence of clicks.*

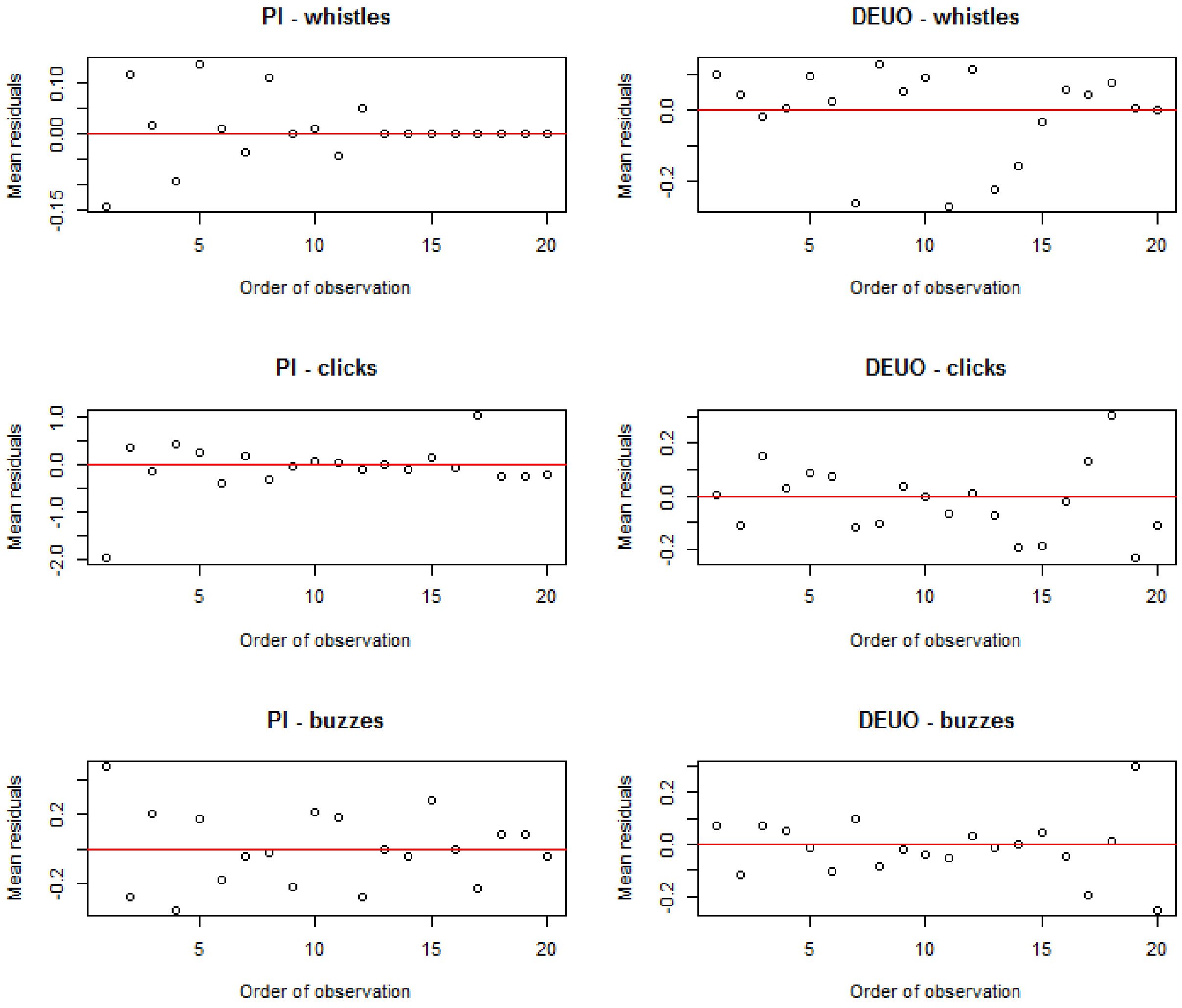

Figure 31. Means of binned Person residuals in order of observation for presence-of-signal-type-given-acoustic-encounter models: pilot whales (left panels) and the DEUO species group (right panels). types were whistles (top), clicks (middle) and buzzes (bottom).

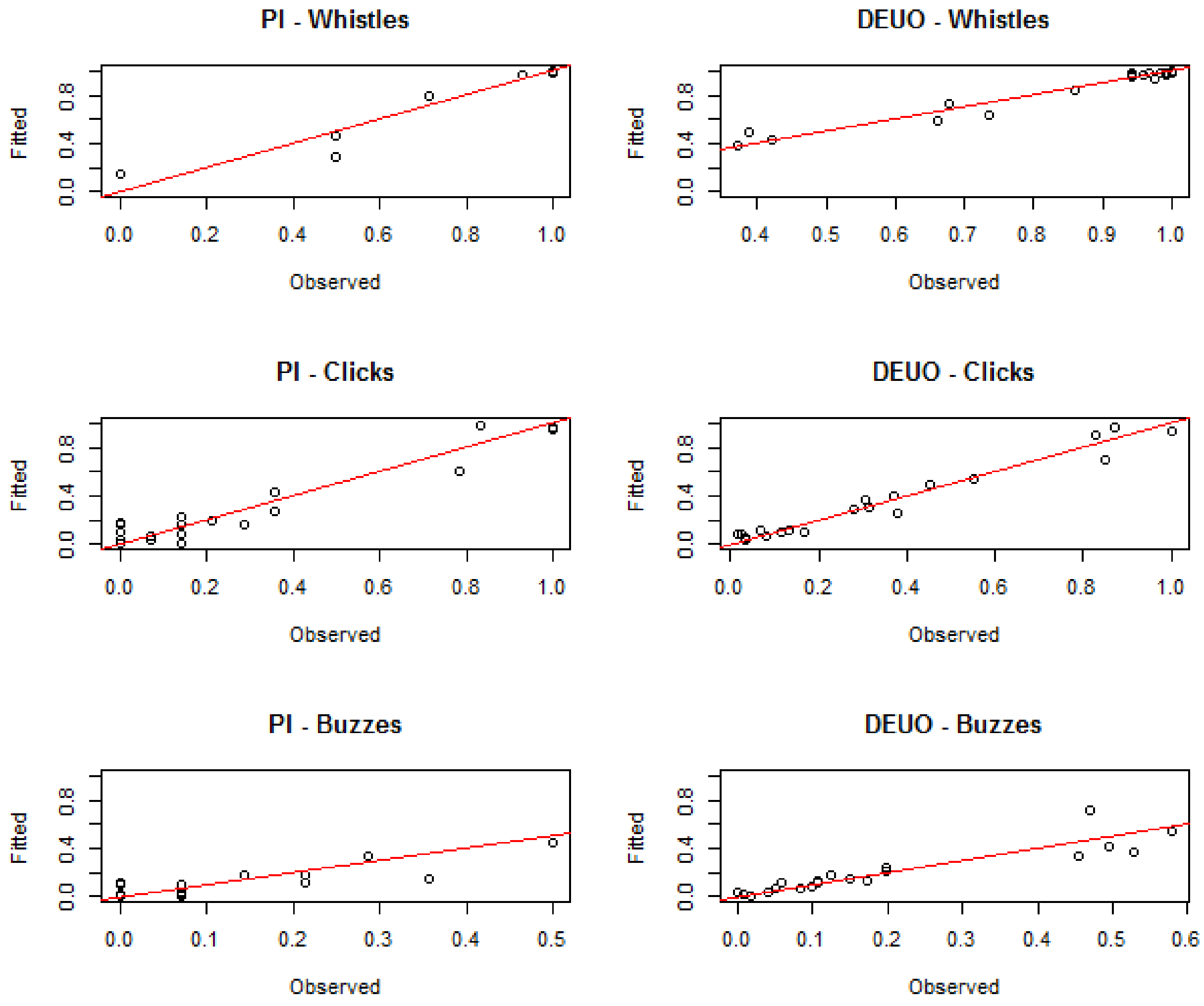

Figure 32. Mean observed versus mean fitted values in ascending order of fitted values for which the mean was calculated from presence-of-signal-type-given-acoustic-encounter models: pilot whales (left panels) and the DEUO species group (right panels). types were whistles (top), clicks (middle) and buzzes (bottom). The red lines indicate a perfect fit of the model to the observed data.

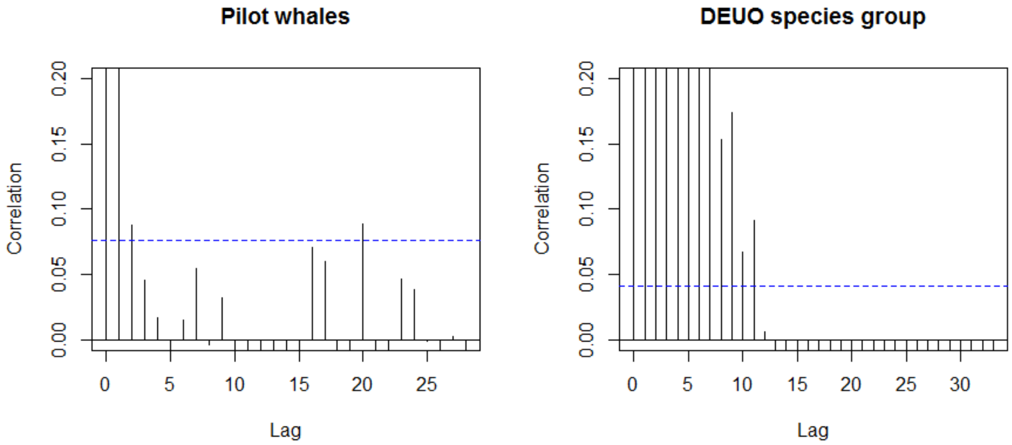

Figure 33. Autocorrelation of Pearson's residuals from whistle characteristics models for pilot whales (left) and common/striped dolphins (right) including 95 percent confidence intervals around zero autocorrelation (blue dashed line). Note that the y-axis was limited to 0.2 for illustrative purposes.

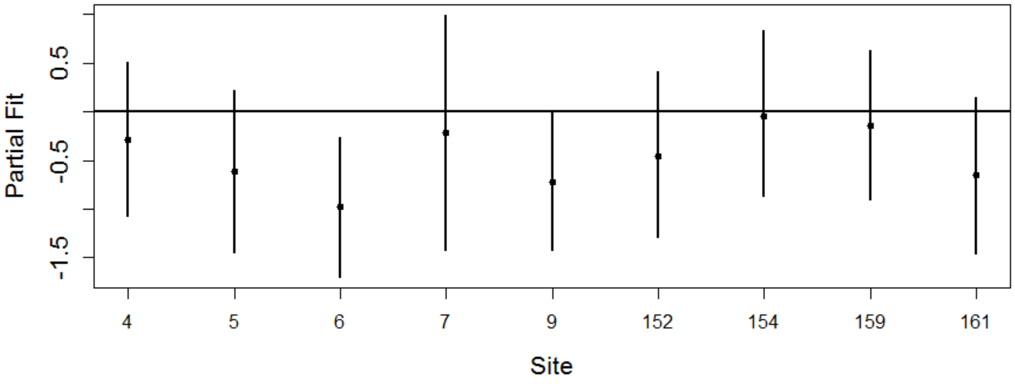

Figure 34. Partial fit plots for whistle characteristics model for pilot whales (note that the partial fit is given on the scale of the identity-link function). The covariate retained in the final model was *Site*.

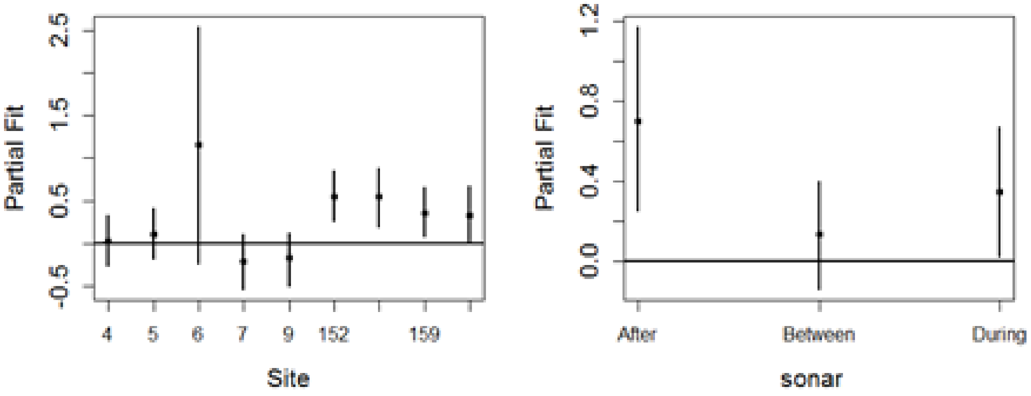

Figure 35. Partial fit plots for whistle-characteristics model for the common/striped dolphins (note that the partial fit is given on the scale of the identity-link function). The covariates retained in the final model were *Site* and *Sonar*.

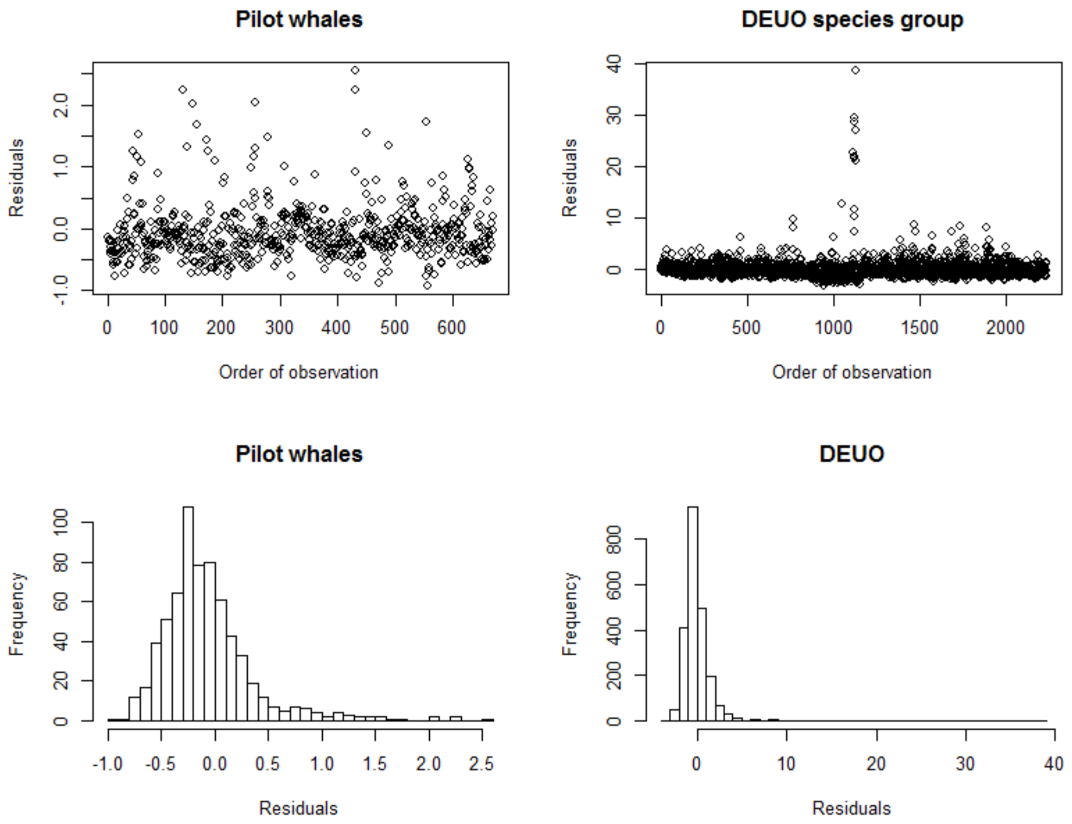

Figure 36. Residuals plots (top) and histograms (bottom) from whistle-characteristics models for pilot whales (left) and common/striped dolphins (right).

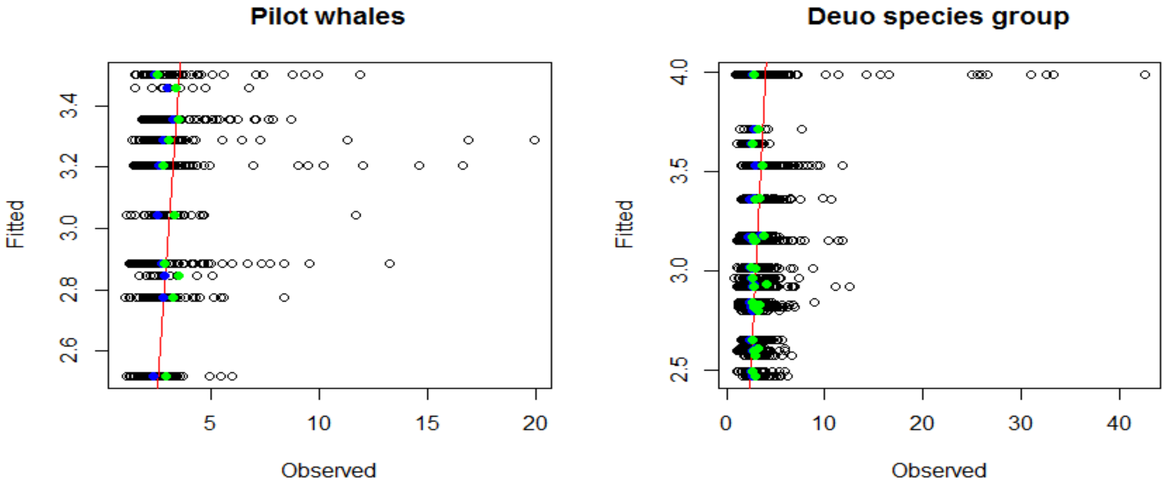

Figure 37. Fitted values from whistle characteristics models for pilot whales (left) and the common/striped dolphins (right). The red lines indicate a perfect fit of the model to the observed data. Blue and green points respectively represent the median and mean of the observed values corresponding to the unique fitted values.

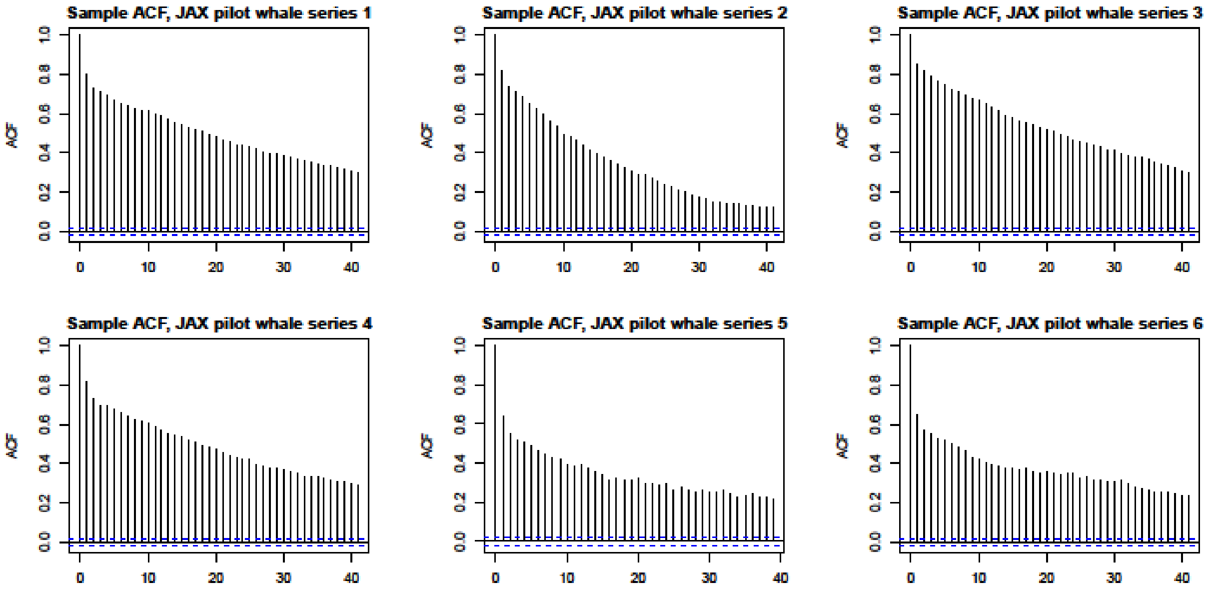

Figure 38. Sample autocorrelation functions for the time series of vocalization encounters recorded for pilot whales at sites 2,4,5,6,7,9 (JAX).

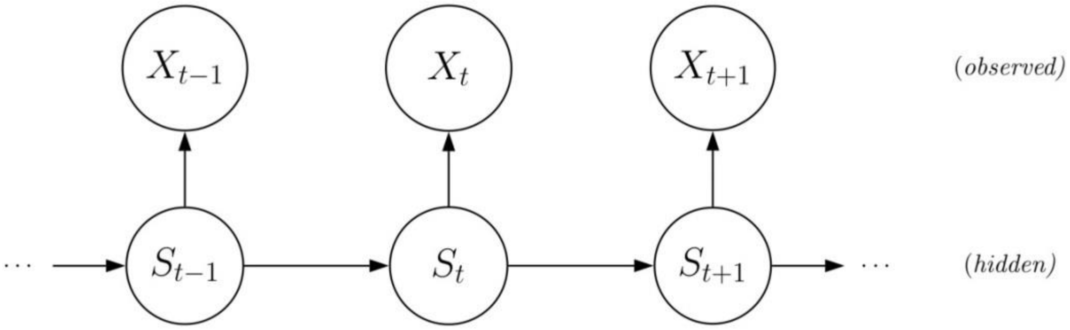

Figure 39. Dependence structure of a basic HMM. In the given application to vocalization data, this means that it is assumed that the state (either ‘silent’ or ‘vocally active’) at time t depends on the state at time t-1, and that whether or not a vocalization is detected at time t depends on the current underlying state at time t.

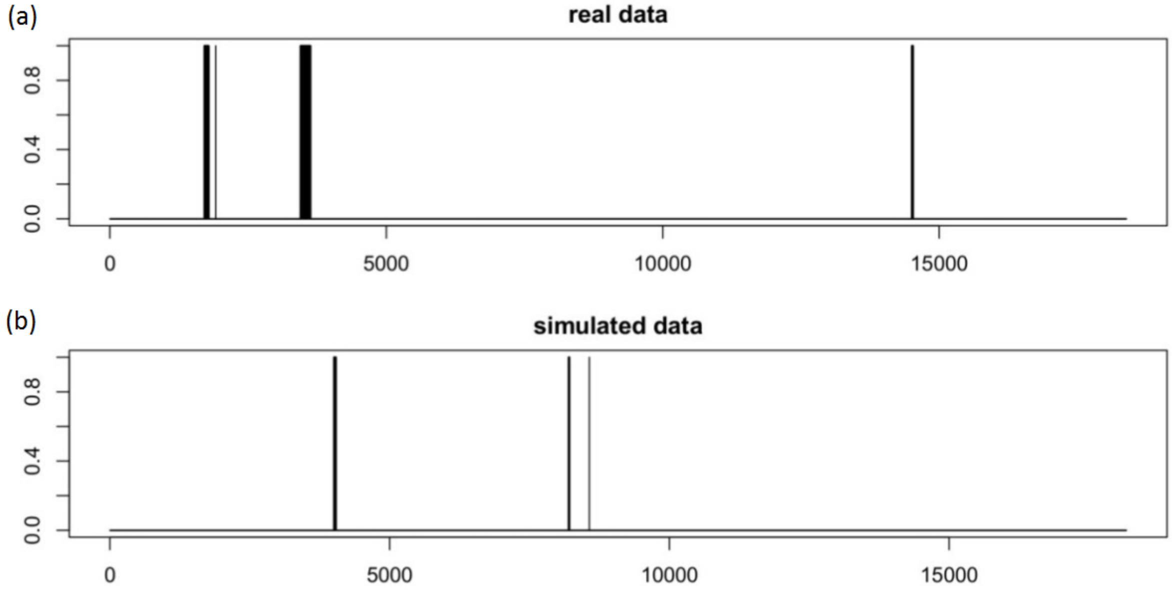

Figure 40. Real and simulated data for pilot whales at OB; only one of the observed series (Site 159) is shown. Horizontal axes indicate the time (with one unit corresponding to 60 seconds), and vertical axes indicate if a vocalization was detected (with the value 1 corresponding to a detection).

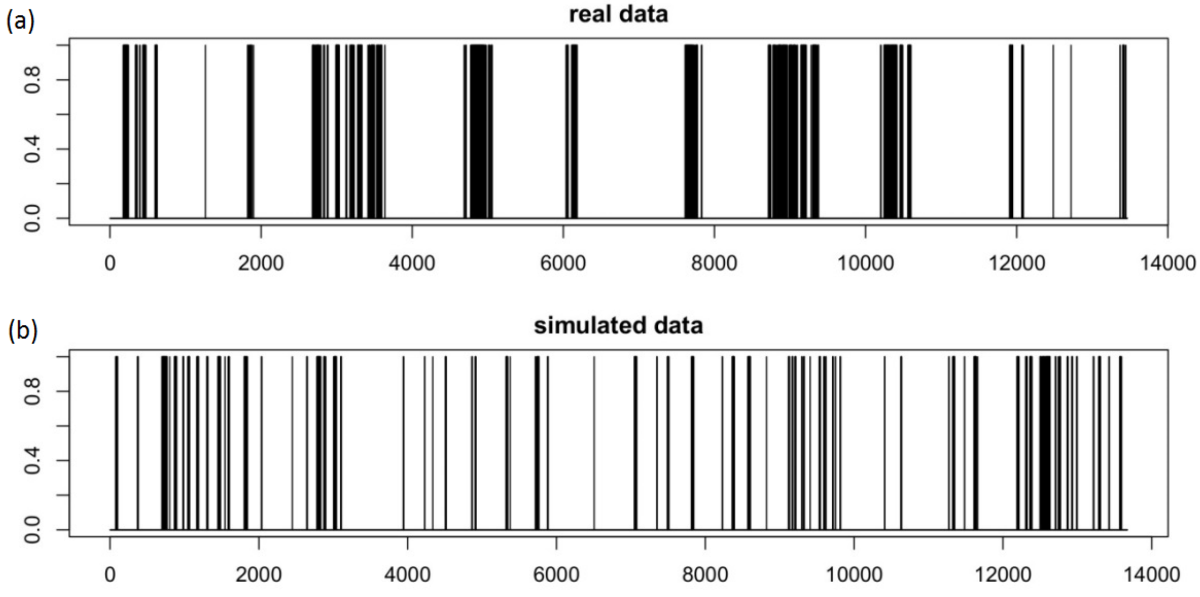

Figure 41. Real and simulated data for DEUO at JAX; only one of the observed series (Site 6) is shown. Horizontal axes indicate the time (with one unit corresponding to 60 seconds), and vertical axes indicate if a vocalization was detected (with the value 1 corresponding to a detection).

11. Tables

Table 1. Confusion matrix for the random-forest classifier model used to identify whistles recorded on the JAX and OB MARUs to species. The percent of encounters in the test dataset classified correctly is in bold. Overall, 81 percent of encounters were correctly classified. Sample size (n) is the number of encounters in the test dataset.

Actual species	Percent classified as					*n*
	Short-beaked common dolphin	Short-finned pilot whale	Striped dolphin	Atlantic spotted dolphin	Bottlenose dolphin	
Short-beaked common dolphin	**75**	0	0	13	13	8
Short-finned pilot whale	0	**81**	13	0	6	16
Striped dolphin	0	0	**73**	9	18	11
Atlantic spotted dolphin	0	6	0	**94**	0	31
Bottlenose dolphin	0	13	2	2	**83**	47

Table 2. Number of delphinid acoustic encounters and sub-events for JAX and OB deployments. Acoustic encounters are defined as a series of whistles, clicks and/or buzzes with no more than 30 min between vocalizations. Sub-events are a series of whistles, clicks and/or buzzes with no more than 1 min between vocalizations.

Deployment	Number of Encounters (30 min resolution)	Number of Sub-Events (1 min resolution)
OB	265	933
JAX deployment 1	550	1,738
JAX deployment 2	444	NA
Total	**1,259**	**2,671**

Table 3. Dates of all MFA sonar events recorded by six MARUs deployed off JAX deployment 1 and five MARUs deployed off OB. For a visual display of the times of the sonar events refer to Figures 6–13.

Site	Dates
JAX 2	14–20 Sept. 2009, 1 Oct 2009
JAX 4	14, 16–20 Sept. 2009, 1 Oct. 2009
JAX 5	14–20 Sept. 2009, 1 Oct 2009
JAX 6	14–20 Sept. 2009, 1 Oct 2009
JAX 7	15–20 Sept. 2009
JAX 9	14, 16–19 Sept. 2009, 1 Oct 2009
OB 152	13, 16, 17, 22, 24–27 July 2008
OB 154	13, 16, 17, 22, 24–27 July 2008
OB 159	13, 16, 17, 22, 24–27 July 2008
OB 161	10, 16, 17, 22, 24, 27 July 2008

Table 4. Covariates available for analyses.

Covariate	Description	Unit
Sonar	Relation to sonar exercise	--
Anysonar	1 for Sonar = During, else 0	--
Time	Time of day	Seconds
Site	Site numbers	--
Sonarlag	Time lag since last sonar	Number of 1 min segments or seconds
Location	JAX or OB	--
Type and duration of sonar ping	See text Section 2.4	--
Sound pressure level		dB re 1 µPa
Peak frequency	--	Hz
Length of sonar event	--	Minutes
Number of sonar detections	--	--
Mean ping interval	--	Seconds
SDEV ping interval	--	Seconds
Mean repetition rate	--	Pings/second
SDEV repetition rate	--	Pings/second
Mean peak frequency	--	Hz
SDEV peak frequency	--	Hz
Mean minimum frequency	--	Hz
SDEV minimum frequency	--	Hz
Mean maximum frequency	--	Hz
SDEV maximum frequency	--	Hz
Mean bandwidth	--	Hz
SDEV bandwidth	--	Hz

Table 5. Models for the presence of vocalizations: parameter estimates (MLE) on the logit-link scale and standard errors (SE) from best fitting models with significance codes related to p-values (1 ' ' 0.1 '.' 0.05 '*' 0.01 '' 0.001 '***' 0). The interaction terms with *Anysonar* allowed setting the respective covariate value to 0 for those segments other than 'During' a sonar exercise. DEUO = striped dolphins, common dolphins and unidentified odontocetes combined.**

	Pilot Whale		**DEUO**	
Max. block size	150		1050	
Linear and factor terms	MLE	SE	MLE	SE
Intercept	-4.14	0.61***	-1.64	0.30***
Site 4	0.02	0.79	-0.81	0.27**
Site 5	-0.90	0.79	-0.72	0.40
Site 6	-0.68	0.78	-0.07	0.30
Site 7	-0.06	0.78	-0.99	0.44*
Site 9	-0.27	0.81	-0.54	0.29
Site 152	-1.17	0.88	-0.08	0.53
Site 154	-2.62	1.17*	-1.78	0.35***
Site 159	-0.03	0.79	-0.41	0.56
Site 161	-3.27	1.17**	-2.76	0.43***
Smooth terms				
bs(time)1	--	--	0.39	0.66
bs(time)2	--	--	-2.72	0.74***
bs(time)3	--	--	0.18	0.30
Addition. parameters				
Scale parameters (SE)	1.00	6.07	1.03	0.90

Table 6. Observed versus predicted presences (1) and absences (0) for the pilot whale and the DEUO species group presence of acoustic encounters models. Numbers in black are correct predictions, numbers in red and blue represent falsely predicted absences or presences, respectively.

	Observed			
	Pilot whale		**DEUO**	
Predicted	0	1	0	1
0	0.55	0.00	0.53	0.02
1	0.44	0.01	0.41	0.05

Table 7. Models for the presence of signal type (whistles, clicks or buzzes) given acoustic encounter for pilot whales: parameter estimates (PE) on the logit-link scale and standard errors (SE) from best fitting models with significance codes related to p-values from (1 ’ ’ 0.1 ’.’ 0.05 ’*’ 0.01 ’’ 0.001 ’***’ 0).**

	Whistles		**Clicks**		**Buzzes**	
Max. group size	1		3		2	
Linear and factor terms	PE	SE	PE	SE	PE	SE
Intercept	1.74	1.13	4.49	1.43**	-3.79	1.04***
Site 4	-7.09	2.68**	-1.31	0.67*	0.20	1.03
Site 5	-7.04	3.16*	-1.06	0.75	41.46	0.97***
Site 6	34.01	3.33***	-43.27	0.56***	40.78	1.08***
Site 7	36.81	3.24***	-2.07	0.84*	0.09	1.36
Site 9	34.49	3.51***	-3.19	1.04**	2.30	1.11
Site 152	38.01	3.26***	-1.09	0.75	1.73	1.06
Site 154	32.39	3.85***	-43.96	0.70***	2.40	1.39.
Site 159	-4.92	2.40*	-1.37	0.72	1.67	0.99.
Site 161	32.83	3.60***	-43.27	0.64***	40.78	1.16***
Presence of whistles	--	--	-4.79	1.19***	--	--
Presence of clicks	-4.84	0.84***	--	--	1.58	0.60**
Presence of buzzes	--	--	2.00	0.62**	--	--
Smooth terms						
bs(time)1	22.73	8.91*	--	--	--	--
bs(time)2	6.77	4.35	--	--	--	--
bs(time)3	9.89	3.87*	--	--	--	--
Additional parameters						
Scale parameter	0.22	0.09	3.32	114	0.84	1.49

Table 8. Models for the presence of signal type (whistles, clicks or buzzes) given acoustic encounter for the DEUO species group including striped dolphin, short-beaked common dolphin and unidentified odontocetes: parameter estimates (PE, on the logit-link scale) and standard errors (SE) from best fitting models with significance codes related to p-values from (1 ’ ’ 0.1 ’.’ 0.05 ’*’ 0.01 ’’ 0.001 ’***’ 0).**

	Whistles		**Clicks**		**Buzzes**	
Max. group size	6		6		7	
Linear and factor terms	PE	SE	PE	SE	PE	SE
Intercept	2.57	0.30***	2.53	0.27***	-5.39	0.43***
Sonar During	0.53	0.25*	--	--	1.13	0.30***
Sonar Between	0.39	0.33	--	--	0.34	0.34
Sonar After	-0.01	0.27	--	--	0.74	0.29*
Site 4	0.42	0.28	0.37	0.26.	1.28	0.42**
Site 5	0.62	0.34.	0.32	0.29	1.06	0.39**
Site 6	0.07	0.26	-0.50	0.26.	3.16	0.37***
Site 7	1.32	0.62*	-2.27	0.46***	2.07	0.43***
Site 9	1.42	0.36***	-1.75	0.30***	4.21	0.43***
Site 152	1.60	0.50**	-1.98	0.28***	2.90	0.47***
Site 154	0.58	0.60	-2.70	0.41***	3.10	0.44***
Site 159	0.90	0.43*	-1.59	0.30***	3.08	0.40***
Site 161	39.22	0.41***	-2.98	0.66***	1.49	1.00
Presence of whistles	--	--	-3.00	0.27***	--	--
Presence of clicks	-2.98	0.23***	--	--	1.46	0.19***
Presence of buzzes	--	--	1.25	0.21***	--	--
Type 1 - short	--	--	0.49	0.24**	--	--
Type 3 medium	--	--	0.78	0.23*	0.69	0.23**
Additional parameters						
Scale parameter	1.00	1.24	1.08	0.61	0.94	0.29

Table 9. Observed versus predicted presences (1) and absences (0) for the pilot whale and the DEUO species group presence-of-signal-type-given-acoustic encounter models. Numbers in black are correct predictions, numbers in red and blue represent falsely predicted absences or presences, respectively.

	Observed					
	Whistles		**Clicks**		**Buzzes**	
Pilot whales						
Predicted	**0**	**1**	**0**	**1**	**0**	**1**
0	0.12	0.09	0.69	0.07	0.51	0.01
1	0.00	0.79	0.05	0.19	0.39	0.09
DEUO species group						
0	0.15	0.19	0.51	0.08	0.59	0.05
1	0.02	0.65	0.15	0.26	0.22	0.15

Table 10. Models for the whistle characteristics: parameter estimates (PE, on the identity-link scale) and standard errors (SE) from best fitting models with significance codes related to p-values (1 ’ ’ 0.1 ’.’ 0.05 ’*’ 0.01 ’’ 0.001 ’***’ 0). DEUO = species group including striped dolphins, common dolphins and unidentified odontocetes.**

	Pilot whale		**DEUO**	
Max. block size	3		12	
Linear and factor terms	PE	SE	PE	SE
Intercept	3.50	0.34***	2.47	0.14***
Sonar During	--	--	0.35	0.17*
Sonar Between	--	--	0.14	0.14
Sonar After	--	--	0.70	0.23**
Site 4	-0.29	0.40	0.02	0.15
Site 5	-0.61	0.41	0.10	0.15
Site 6	-0.98	0.36**	1.17	0.72
Site 7	-0.21	0.61	-0.22	0.17
Site 9	-0.72	0.37*	-0.17	0.16
Site 152	-0.46	0.43	0.54	0.16***
Site 154	-0.04	0.45	0.54	0.18**
Site 159	-0.14	0.38	0.36	0.15*
Site 161	-0.66	0.42	0.33	0.17.
Addition. parameters				
Scale parameters	0.34	0.07	4.42	2.16

Table 11. AIC values for the different models considered. Lowest AIC values are underlined and bold. The DEUO species group refers to delphinid species, not including pilot whales.

	Model Without Covariates	**Model With *Anysonar***	**Model With *Sonar***	**Model With *Sonarlag***	**Model With *Mean Rep. Rate***	**Model With *SDEV Ping Interval***
Pilot whales, Jacksonville	2417.36	2420.99	2397.84	2418.86	2416.82	**<u>2379.62</u>**
Pilot whales, Onslow Bay	**<u>667.40</u>**	670.74	675.67	670.75	670.75	670.68
DEUO, Jacksonville	14381.58	14383.87	**<u>14375.67</u>**	14383.05	14384.80	14382.58
DEUO, Onslow Bay	8094.76	8098.75	**<u>8049.28</u>**	8098.52	8098.09	8096.36

This page intentionally left blank.

A

Variables Measured by VOCCA

This page intentionally left blank.

Appendix A: Variables Measured by ROCCA

Variable	Explanation
Begsweep	slope of the beginning sweep (1 = positive, -1 = negative, 0 = zero)
Begup	binary variable: 1 = beginning slope is positive, 0 = beginning slope is negative
Begdwn	binary variable: 1 = beginning slope is negative, 0 = beginning slope is positive
Endsweep	slope of the end sweep (1 = positive, -1 = negative, = 0 zero)
Endup	binary variable: 1 = ending slope is positive, 0 = ending slope is negative
Enddwn	binary variable: 1 = ending slope is negative, 0 = ending slope is positive
Beg	beginning frequency (Hertz [Hz])
End	ending frequency (Hz)
Min	minimum frequency (Hz)
Dur	duration (seconds)
Range	maximum frequency - minimum frequency (Hz)
Max	maximum frequency (Hz)
mean freq	mean frequency (Hz)
median freq	median frequency (Hz)
std freq	standard deviation of the frequency (Hz)
Spread	difference between the 75th and the 25th percentiles of the frequency
quart freq	frequency at one-quarter of the duration (Hz)
half freq	frequency at one-half of the duration (Hz)
Threequart	frequency at three-quarters of the duration (Hz)
Centerfreq	(minimum frequency + (maximum frequency-minimum frequency))/2
rel bw	relative bandwidth: (maximum frequency - minimum frequency)/center frequency
Maxmin	maximum frequency/minimum frequency
Begend	beginning frequency/end frequency
Cofm	coefficient of frequency modulation (COFM): take 20 frequency measurements equally spaced in time, then subtract each frequency value from the one before it. COFM is the sum of the absolute values of these differences, all divided by 10,000
tot step	number of steps (10 percent or greater increase or decrease in frequency over two contour points)
tot inflect	number of inflection points (changes from positive to negative or negative to positive slope)
max delta	maximum time between inflection points
min delta	minimum time between inflection points
maxmin delta	maximum delta/minimum delta
mean delta	mean time between inflection points
std delta	standard deviation of the time between inflection points
median delta	median of the time between inflection points
mean slope	overall mean slope
mean pos slope	mean positive slope
mean neg slope	mean negative slope
mean absslope	mean absolute value of the slope

Variable	Explanation
Posneg	mean positive slope/mean negative slope
perc up	percent of the whistle that has a positive slope
perc dwn	percent of the whistle that has a negative slope
perc flt	percent of the whistle that has zero slope
up dwn	number of inflection points that go from positive slope to negative slope
dwn up	number of inflection points that go from negative slope to positive slope
up flt	number of times the slope changes from positive to zero
dwn flt	number of times the slope changes from negative to zero
flt dwn	number of times the slope changes from zero to negative
flt up	number of times the slope changes from zero to positive
step up	number of steps that have increasing frequency
step dwn	number of steps that have decreasing frequency
step.dur	number of steps/duration
inflect.dur	number of inflection points/duration