esa ECOSPHERE

# Dose response severity functions for acoustic disturbance in cetaceans using recurrent event survival analysis

C. M. Harris,[1,†] D. Sadykova,[1] S. L. DeRuiter,[1] P. L. Tyack,[2] P. J. O. Miller,[2] P. H. Kvadsheim,[3]
F. P. A. Lam,[4] and L. Thomas[1]

[1]*Centre for Research into Ecological and Environmental Modelling, Buchanan Gardens,*
*University of St Andrews, St Andrews, Fife KY16 9LZ Scotland, United Kingdom*
[2]*Sea Mammal Research Unit, Scottish Oceans Institute, University of St Andrews, St Andrews,*
*Fife KY16 8LB Scotland, United Kingdom*
[3]*Norwegian Defence Research Establishment (FFI), Maritime Systems, N-3191 Horten, Norway*
[4]*Acoustics & Sonar Research Group, Netherlands Organization for Applied Scientific Research (TNO),*
*Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands*

**Citation:** Harris, C. M., D. Sadykova, S. L. DeRuiter, P. L. Tyack, P. J. O. Miller, P. H. Kvadsheim, F. P. A. Lam, and L. Thomas. 2015. Dose response severity functions for acoustic disturbance in cetaceans using recurrent event survival analysis. Ecosphere 6(11):236. http://dx.doi.org/10.1890/ES15-00242.1

**Abstract.** Behavioral response studies (BRSs) aim to enhance our understanding of the behavior changes made by animals in response to specific exposure levels of different stimuli, often presented in an increasing dosage. Here, we focus on BRSs that aim to understand behavioral responses of free-ranging whales and dolphins to manmade acoustic signals (although the methods are applicable more generally). One desired outcome of these studies is dose-response functions relevant to different species, signals and contexts. We adapted and applied recurrent event survival analysis (Cox proportional hazard models) to data from the 3S BRS project, where multiple behavioral responses of different severities had been observed per experimental exposure and per individual based upon expert scoring. We included species, signal type, exposure number and behavioral state prior to exposure as potential covariates. The best model included all main effect terms, with the exception of exposure number, as well as two interaction terms. The interactions between signal and behavioral state, and between species and behavioral state highlighted that the sensitivity of animals to different signal types (a 6–7 kHz upsweep sonar signal [MFAS] or a 1–2 kHz upsweep sonar signal [LFAS]) depended on their behavioral state (feeding or non-feeding), and this differed across species. Of the three species included in this analysis (sperm whale [*Physeter macrocephalus*], killer whale [*Orcinus orca*] and long-finned pilot whale [*Globicephala melas*]), killer whales were consistently the most likely to exhibit behavioral responses to naval sonar exposure. We conclude that recurrent event survival analysis provides an effective framework for fitting dose-response severity functions to data from behavioral response studies. It can provide outputs that can help government and industry to evaluate the potential impacts of anthropogenic sound production in the ocean.

**Key words:** behavioral response; cetaceans; controlled exposure experiment; Cox proportional hazards model; *Globicephala melas*; *Orcinus orca*; *Physeter macrocephalus*; response intensity; sonar.

**Received** 23 April 2015; revised 4 June 2015; accepted 8 June 2015; **published** 20 November 2015. Corresponding Editor: D. P. C. Peters.

† **E-mail:** catriona.harris@st-andrews.ac.uk

## Introduction

Many marine mammals rely on sound for foraging, maintaining group cohesion, navigation, finding mates and avoiding predators. Hence, they may be profoundly affected by the introduction of anthropogenic noise into the marine environment. Examples of potentially harmful noise sources include vessel noise and active acoustic devices such as naval sonar or airguns used for seismic prospecting (Richardson et al. 1995, DeRuiter 2010). Potential adverse effects of those sounds include reduction or cessation of feeding (Miller et al. 2009, Goldbogen et al. 2013), strong avoidance responses (Tyack et al. 2011, DeRuiter et al. 2013, Miller et al. 2014), and stranding (D'Amico et al. 2009). Strong avoidance responses may exclude animals from important habitats or result in separation of dependent offspring and mother (Miller et al. 2012).

Controlled exposure experiments (CEEs) are behavioral response studies (BRSs) that follow an experimental design and are an important approach for studying the short-term responses of animals to specific doses of potential stressors. A growing number of these studies have been carried out in recent years on a number of different cetacean species, specifically looking at different acoustic stimuli (Kvadsheim et al. 2011, 2012, 2014, Miller et al. 2011, Tyack et al. 2011, Southall et al. 2012, Dunlop et al. 2013). Together, these studies are increasing our understanding of species differences in sensitivity to sound, and the importance of context in influencing how individuals respond. These types of studies are not unique to the marine environment and have followed the example of many terrestrial studies that have investigated the behavioral responses of a range of species experimentally exposed to human-induced disturbance (for example, Rocky Mountain elk [Preisler et al. 2006], grassland raptors [Holmes et al. 1993], waterbirds [Klein 1993] and ungulates [see Stankowich 2008 for review]). The common objective across many of these studies has been to determine the relationship between the dose of a stressor (which can be represented by many different metrics) and response.

CEEs on cetaceans are costly to undertake. Many species of interest occur at low density, or are hard to monitor (for example if they are difficult to locate or track, either at the surface or underwater), and experiments can only take place in good weather conditions and when other interfering noise sources are not present. Because of these factors, the sample sizes associated with CEEs are usually very low: typically fewer than 10 exposures per species per field season, and sometimes substantially fewer (e.g., Kvadsheim et al. 2014).

In a CEE, the focal species is selected based upon research need and the focal animal is the individual that becomes the focus of the study. The behavior of the focal animal is monitored using visual observations, passive acoustics, animal-borne tags or a combination of these. After pre-exposure observations, the focal whale is exposed to a stimulus, such as a potentially disturbing sound or control sound, and its response is monitored. In many of the experiments the dose of sound increases over the duration of the exposure and therefore they can be thought of as dose-escalation studies. The increasing dose is achieved either by increasing the level of the source, or by approach of the vessel, which can increase the level of sound received at the animal by transmitting the sound stimulus while approaching the focal animal. To understand the influence of this experimental design on responsiveness, no-sonar control exposures are also conducted where the vessel approaches in the same manner but no sounds are transmitted (e.g., Miller et al. 2011). Various measurements are recorded before, during and after exposure periods, including location of the focal animal through time, vocal behavior, underwater orientation, and behavior observed at the surface. Care is taken to ensure that the experimental animals are not injured or stressed, for example by having mitigation protocols in place that result in shut-down if animals approach too close or display an extreme response.

Following data collection, typically the first stage of analysis is to assemble and visualize the observational data for each individual to determine whether or not it responded, gauge the magnitude of any response and relate the onset of response to sound exposure level ("dose"). There is ongoing development of quantitative methods for identifying responses or "change-points" in behavior (for example, see DeRuiter et al. 2013, Goldbogen et al. 2013, Antunes et al. 2014; http://www.creem.st-and.ac.uk/mocha/); however qualitative methods have also been used effectively. The process of examining visual observation and animal-borne tag records and subsequent identification of putative responses to exposure stimuli by panels of experts is described in detail for the 3S BRS project in Miller et al. (2012). Miller et al. (2012) note that the results from such scoring exercises (herein referred to as expert scoring) are one interpretation of the experiment outcome, and whilst they believe that they identified most responses that occurred during the experiments, there is the possibility that some behavioral changes that were scored as responses may not have been in response to the sonar. Therefore, although there is potential for bias, it is likely that the outputs from expert scoring are precautionary, which is the preferred stance from a policy perspective (Miller et al. 2012). The output from expert scoring is a detailed record of all behavioral changes that are likely to be responses to the stimuli for each exposure session with details of the time the response started, the corresponding exposure level (measured as sound pressure level [SPL] and cumulative sound exposure level [SELcum]) at the point of response, and an assessment of the severity of the response (see Miller et al. 2012: Appendix A). Since the dose metrics relate to the sonar sound, the responses identified during the no-sonar control experiments have zero dose associated with them. Each response is attributed a score between 0 and 9 which describes the severity of the response of the animal. This scoring is derived from the severity scale described in Southall et al. (2007), and modified by Miller et al. (2012), and ranges from no effect (0), effects not likely to influence vital rates (severity of 1–3), effects that could affect vital rates (severity of 4–6), to effects that are thought likely to influence vital rates (severity of 7–9).

The next stage is to combine the results from individual exposures to estimate the likelihood of response as a function of exposure intensity (dose), and perhaps behavioral or environmental context. There are two things to consider in our approach to this analysis. First, there may be multiple responses per exposure and multiple exposures per individual animal. Second, there may be individuals that showed no response across the range of doses they experienced over an exposure session. To account for the latter, a framework is needed that allows inclusion of right-censored data where it is assumed a response will occur at some point above the maximum dose received during the exposure session, but it is unknown by how much. Right-censored data are informative and should be included in any analysis (Klein and Moeschberger 2003). Miller et al. (2014) and Antunes et al. (2014) describe a Bayesian hierarchical approach that allows the inclusion of censored data, accounts for multiple exposures per individual, and estimates response variability between and within whales. This approach, however, only allows for analysis of one response per individual per exposure session. In their papers, Miller et al. (2014) and Antunes et al. (2014) focused on the threshold for the first avoidance response observed in each exposure session.

Our aim was to find a framework for producing dose-response severity functions that would allow us to consider all of the observed responses per exposure session together, to account for censored data, and to acknowledge the non-independence of responses made by the same individual.

CEEs are similar in many ways to clinical trials, which seek to identify the toxicity of a drug by administering different doses of the drug and assessing responses, and also to medical studies where the objective is to study how long it takes until certain events occur. For this reason we turned to the medical literature to seek appropriate analysis methods for this ecological question. We found that recurrent event survival analysis (Kleinbaum and Klein 2005) is often used to address questions and data similar to ours. This approach is used in medical studies to assess time to events such as tumor occurrence (where the recurrent events are the same), or different disease symptoms (where recurrent events are different; Kleinbaum and Klein 2005). It is also designed to accept censored data, since an individual may leave a study prior to the end, or not display a symptom by the end of a follow-up period. The use of survival analysis as a framework to model time-to-event data has been primarily utilized for modeling time-to-death, or time-to-symptom expression in the medical and epidemiological domains; however there are documented examples of its application within ecology. Muenchow (1986) advocated the use of such models in ecology and proposed a number of ecological questions that could be phrased in terms of "time until an event occurs." Examples include flower visitation events by insects (Muenchow 1986), time of fish passage in rivers (Zabel et al. 2014), tree mortality (Woodall et al. 2005) and duration of tarantula fighting (Moya-Loraño and Wise 2000). Most of these authors note the novelty of the application of survival analysis within their specific fields. One particular area of expansion of its application has been in plant pathology (see Scherm and Ojiambo 2004 for summary) and it is here that we found the only example of the use of recurrent event survival analysis within ecology (Thomson and Copes 2009). We were particularly interested in the variant of these recurrent event survival models called the marginal stratified Cox proportional hazards model. In the marginal approach each event is considered as a separate process (i.e., there is no condition on events being progressive, such as in a disease where symptoms have to occur in a certain order), and different response events can represent different response types that may occur in the same subject (Kleinbaum and Klein 2005).

Here we demonstrate and evaluate the application of recurrent event survival analysis to develop dose-response severity functions in the context of cetacean CEEs, using data on killer whales, long-finned pilot whales and sperm whales from the 3S BRS project (Miller et al. 2011).

## Methods

### Data

The expert scored response data were provided by the 3S BRS project, which includes data on killer whales, long-finned pilot whales and sperm whales exposed to three different sonar signals, as well as killer whale playbacks and no-sonar control sessions (all data used are publicly available; see Miller et al. [2011] for full description of project and controlled exposure experiments, Miller et al. [2012] for description of expert scoring, and Appendix A of Miller et al. [2012] for descriptions of scored responses). Here we focus only on the exposures to two sonar signals that were conducted as dose-escalation experiments where the vessel approached the focal animal during exposure to increase the received sound level at the animal: a 6–7 kHz upsweep signal (MFAS) and a 1–2 kHz upsweep signal (LFAS). These signals were the primary focus of the study and, as such, the data come from a balanced study design where the order in which signal types were presented was alternated from one experiment to the next. The data comprise 27 exposure sessions across 14 different individuals (four sperm whales, four killer whales and six long-finned pilot whales). No-sonar control exposures were not included in this analysis because the sonar dose is zero for all identified responses, and other metrics of sound level such as ship noise related to vessel approach have not been measured.

Behavioral responses during each exposure event were identified by expert scoring (Miller et al. 2012) and each identified response event was assigned a severity score by the experts according to the severity scale outlined in Miller et al. (2012). The responses were also aligned with the corresponding levels of the received sonar signals recorded on the animal-borne tag (DTAG; Johnson and Tyack 2003). Although the severity scale ranges from 0 (no response) to 9 (severe response), we did not have enough data across the scale to fit separate exposure-response functions for each of the nine scores. Therefore, we binned the data into three levels; 1 = mild (severity scores 1–3), 2 = moderate (severity scores 4–6), 3 = severe (severity scores 7–9). Instead of including 0 (no response) as a separate level, we included data from these sessions as right-censored (described below). We identified the first occurrence of each response level within each exposure session for inclusion in the model.

Recurrent event survival analysis is generally used to assess time-to-event; however we were interested in acoustic dose-to-event, and so we replaced time with received acoustic energy in the form of cumulative sound exposure level (SELcum). SELcum is a measure of sound energy integrated over the duration of the exposure, which increases monotonically throughout each sonar exposure session. In the

3S dose-escalation experimental design, the values of SELcum were highly correlated with the maximum received sound pressure level (Miller et al. 2014).

In the marginal variant of the stratified Cox proportional hazards model, each individual is considered to be at risk of all response levels. Therefore for each response level the "start time," or in our case initial dose (SELcum), was the same, i.e., the first dose of that particular exposure session. If all three response levels were observed, in ascending order of severity, then the dose relating to each response event was allocated accordingly. If, however, a moderate or severe response was observed without a preceding mild or moderate response, then the dose allocated to the lower level was the same as that observed for the higher level. Similarly, if a mild response was observed at a higher SELcum than a moderate response, then we replaced the observed SELcum for the mild response with the value for the moderate response. This data structure does not imply that responses have to occur in a progressive manner, with mild responses occurring before moderate, etc. Rather, this approach assumes that observing one more severe response means that the equivalent of all less-severe responses has also occurred, simultaneously if not previously. This is a reasonable assumption in a broad context of response severity and we have therefore structured the data set to reflect this. See Thomson and Copes (2009) for an example of similar data structuring.

In the case of no response across all levels within an exposure session, then each level was allocated the cumulative received level (SELcum) at the end of the exposure session, and the data were labelled as censored. Similarly if only mild and/or moderate response events were observed then a censored value was allocated to the higher severity response level.

### Model specification

The data comprise $K = 14$ individuals and $N = 27$ exposure sessions in total, with up to three exposures per whale. Severity level was denoted as $S = 1, 2, 3$ and $X_{Skn}$ denoted the value of a vector of covariates $\boldsymbol{X}$ for individual $k$ ($k = 1,\ldots, K$) at exposure session $n$ ($n = 1,\ldots, N$) with respect to the $S$th stratum (severity level). Then a marginal stratified Cox proportional hazard regression model can be defined as Eq. 1.

$$h_S(\text{SELcum}|X) = h_{0S}(\text{SELcum})\exp(X^T\boldsymbol{\beta}), \quad S = 1, ..., 3 \tag{1}$$

where $h_S(\text{SELcum}|\boldsymbol{X})$ is the stratum-dependent hazard function, $h_{0S}(\text{SELcum})$ is the stratum-dependent baseline hazard and $\boldsymbol{\beta}$ is the stratum-dependent vector of regression coefficients, which are estimated by the method of maximum partial likelihood estimation (Cox 1975, Therneau and Grambsch 2000). The hazard function in this case gives the probability of a response occurring at a given unit of SELcum, given that the individual has not responded up to that point (Kleinbaum and Klein 2005). We assumed that the observations were clustered (and therefore correlated) within individuals and that there was independence between individuals (clusters). The standard errors of the model estimates were corrected for the correlations within the clusters using a grouped jackknife procedure (Therneau and Grambsch 2000: Section 8.2.1).

The covariates considered were species (killer whale, sperm whale, long-finned pilot whale), signal (MFAS, LFAS), exposure number (1–3) and behavioral state in the pre-exposure period (feeding, non-feeding). All were specified as factor covariates with the exception of exposure number. It was unclear whether exposure number should be included as a continuous covariate, and therefore we fitted models where it was included as a continuous covariate, an ordered categorical covariate or factor covariate. Behavioral state, feeding or non-feeding, was determined by the diving and vocalizing behavior of the animals prior to exposure (based on the behavioral description in Miller et al. [2011]) and feeding was assigned if any feeding behavior was observed during any part of the pre-exposure period. We considered all combinations of covariates up to, and including, all first-order interaction terms and carried out backwards selection from the full model, dropping the covariate with the highest $p$-value (from a Wald $\chi^2$ test) at each iteration until all remaining covariates had $p$-values less than 0.05. For the best fitting model, we tested that the proportional hazards assumption and the no-interaction assumption were both met (Kleinbaum and Klein 2005). The proportional hazards assumption requires that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time (or SELcum in this case), and the no-interaction assumption requires that β coefficients do not vary across severity categories (Kleinbaum and Klein 2005). To test the proportional hazards assumption, we carried out a $\chi^2$ test to determine if the slope of the scaled Schoenfeld residuals differed significantly from zero (Grambsch and Therneau 1994). Schoenfield residuals relate to the difference between an individual's covariate value when there is a response event and the weighted average of the covariate values for the other individuals still at risk at the relevant SELcum. The weights are each individual's hazard (Kleinbaum and Klein 2005). We tested the no-interaction assumption by comparing a model with no interactions and a model where each covariate interacts with the stratum indicator (severity level) using a likelihood ratio test. The fit of the best model was evaluated using the Cox-Snell residuals, which are the estimated cumulative hazards for individuals at their response (or censoring) times (or corresponding SELcum in this case). If a model fits the data well, then the cumulative hazard function conditional on the covariate vector should have an exponential distribution with a hazard rate of one (i.e., the estimated cumulative hazard of the Cox-Snell residuals should look like a 45-degree line).

All statistical analyses were carried out in R version 3.0.2 (R Core Team 2015) using the Survival library (Therneau 2015), and in SAS software version 9.3 (SAS Institute, Cary, North Carolina, USA). We used SAS to carry out the model selection procedure because our model included factor covariates. SAS model output provides $p$-values for factor level comparisons, but also $p$-values that relate to the contribution of the factor to model fit. The latter is not readily available in R, but is required for backwards selection.

## Results

### *Model selection and testing assumptions*

The selected model included signal, species and behavioral state as well as an interaction term between species and behavioral state, and between signal and behavioral state (Table 1). Note that we have no data on sperm whales in a non-feeding behavioral state and therefore we could not make any inference or predictions about non-feeding sperm whales. There was no significant effect of exposure number (no order effect) when included as a continuous, ordered categorical or factor covariate. Fig. 1 shows that the model fitted reasonably well to the data: at high values of cumulative hazard rate the data points lie above the 45 degree line, however, it is in the tail of such functions where variability due to estimation uncertainty is the greatest and so these deviations are not of major concern (Box-Steffensmeier and Jones 2004).

The best-fitting model met both the proportional hazards assumption (global $p$-value from $\chi^2$ test $= 0.067$) and the no-interaction assumption (the model with interactions between covariates and stratum was not significantly better than the model without these interaction terms; $p = 0.184$). As both assumptions were met, no remedial action was required.

### *Biological interpretation*

In this section we provide an overview of the model output from a biological perspective, but we advise caution applying the results from this case study. Though the data are unique, the sample sizes for each species were small and, with the inclusion of explanatory covariates, we are describing outputs which, in some cases, result from a sample size of one individual (e.g., feeding killer whales exposed to LFAS). In addition, we have not included no-sonar control exposures in this analysis and so there is no direct evaluation of the relative contributions of vessel approach and sonar exposure in the probability of response.

From the model we can produce dose-response functions for the three different severity levels averaged across all covariates and accounting for censored data (Fig. 2). All dose-response functions were generated using the survfit function within the Survival library in R, which can be used to produce survival curves based on a fitted model. We can see from Fig. 2A that the probability of a mild level 1 response increased steadily from a SELcum of 87 dB re 1 $\mu Pa^2 s$ through to 168 dB re 1 $\mu Pa^2 s$ at which point the probability of response (P-response) was equal to

Table 1. Maximum likelihood parameter estimates, standard errors, *p*-values and hazard ratios with 95% CIs for the best fitting model.

| Parameter | df | Parameter estimate | SE | *p*-value | Hazard ratio | 95% hazard ratio confidence limits |
|---|---|---|---|---|---|---|
| Signal (MFAS) | 1 | −1.64 | 0.52 | 0.018 | 0.19 | 0.05, 0.75 |
| Species (long-finned pilot) | 1 | −2.41 | 0.74 | 2.8e−07 | 0.09 | 0.04, 0.22 |
| Species (sperm) | 1 | −2.11 | 0.61 | 4.3e−05 | 0.12 | 0.04, 0.33 |
| Behavioral state (nonfeed) | 1 | −4.24 | 0.84 | 1.7e−05 | 0.01 | 0.002, 0.10 |
| Species:behavioral state (long-finned pilot nonfeed) | 1 | 1.25 | 0.84 | 0.07 | 3.49 | 1.25, 9.74 |
| Species:behavioral state (sperm nonfeed) | 1 | NA | NA | NA | NA | NA, NA |
| Signal:behavioral state (MFAS nonfeed) | 1 | 3.80 | 0.81 | 0.001 | 44.92 | 4.32, 467.5 |

*Notes:* Values are given for each level of the factor covariates and are relative to the reference level. The reference factor levels were LFAS, killer whale and feeding. NA indicates no data.

1. The 50% probability of response (P-response = 0.5) related to a SELcum of 154 dB re 1 μPa$^2$s. The shape of the response function for moderate level 2 responses was very similar but shifted to the right slightly, with a P-response = 0.5 at a SELcum of 157 dB re 1 μPa$^2$s (Fig. 2B). In both cases the 95% confidence intervals suddenly increased in size at around 165 dB re 1 μPa$^2$s, so much so that the intervals span the entire range of probabilities (0 to 1; see *Discussion*). The model predicted very low probabilities of a severity level 3 response across the range of observed SELcum (Fig. 2C).

The same patterns, with respect to the effect of covariates, exist across all severity levels (the no-interaction assumption was met indicating that the β coefficients do not vary across the severity levels) and so we only show the results for severity level 2 here (Figs. 3 and 4; for the other levels see Appendix: Figs. A1–A4). We do not make any predictions for non-feeding sperm whales.

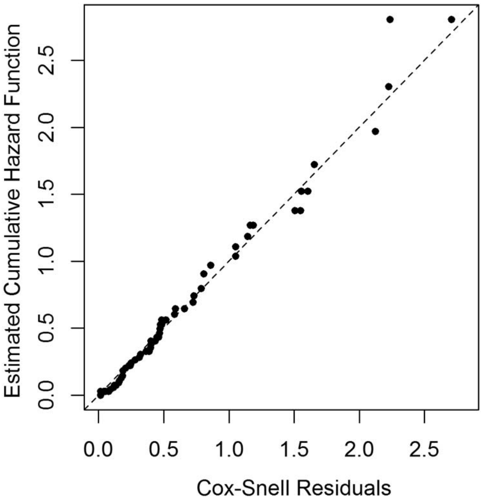

Fig. 1. Cox-Snell residuals from the selected model. The dashed line represents slope = 1.

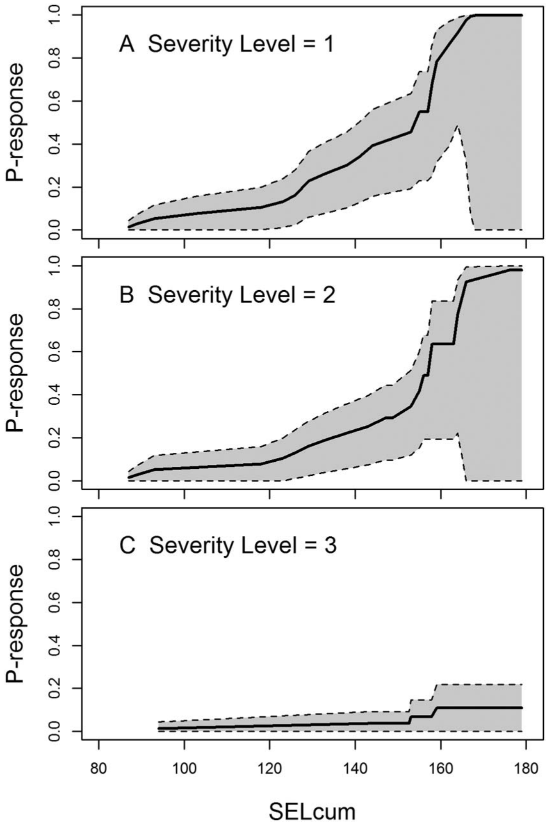

Fig. 2. The probability of a response averaged across all covariates (species, signal type and behavioral state) versus received acoustic energy (SELcum (dB re 1 μPa$^2$s)) for (A) severity level = 1 (mild), (B) severity level = 2 (moderate) and (C) severity level = 3 (severe). The dashed lines represent 95% confidence intervals.

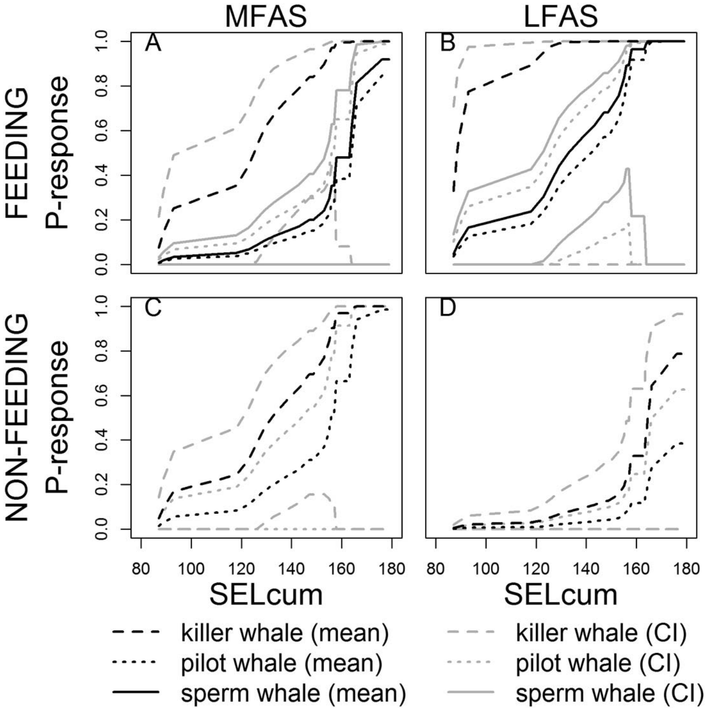

Fig. 3. The probability of a response occurring in killer whales (dashed line), long-finned pilot whales (dotted line) and sperm whales (solid line) versus received acoustic energy (SELcum (dB re 1 μPa$^2$s)) for severity level = 2 when signal = MFAS (A and C) and LFAS (B and D) and behavioral state = feeding (A and B) and non-feeding (C and D). Mean probabilities are all shown in black, while 95% confidence intervals are shown in grey. Similar plots for severity levels 1 and 3 are shown in Appendix: Figs. A1 and A2.

Comparing the sensitivity of all three species, we see that killer whales have a higher probability of responding at lower SELcum than long-finned pilot whales and sperm whales across all signal types and behavioral states (Fig. 3). There was little difference in the probability of response of long-finned pilot whales and sperm whales.

There were differences in the way species responded depending on their behavior state, hence the significant interaction term between these covariates. For example, long-finned pilot whales were predicted to respond to MFAS at slightly lower SELcum when non-feeding compared to feeding, but the opposite was true for LFAS where they were predicted to respond at much higher SELcum when non-feeding compared to feeding. By contrast, killer whales had a lower probability of responding to both signal types when in a non-feeding state compared to feeding, although the difference was much more marked for the LFAS signal (Fig. 3).

Regarding sensitivity within species, sperm whales were predicted to respond to LFAS at lower SELcum compared to MFAS (Fig. 3, data only for feeding sperm whales). Killer whales and long-finned pilot whales were also predicted to respond to LFAS signals at lower SELcum when feeding compared to MFAS (Fig. 4) whilst the opposite is predicted when in a non-feeding state (Fig. 4). This explains the significant interaction term between signal and behavioral state, which was due to a large difference in sensitivity to LFAS across behavioral states rather than a difference in sensitivity to MFAS. However, as noted earlier there is only one datum each for feeding killer whales exposed to LFAS and feeding long-finned pilot whales exposed to LFAS, and therefore we need to limit our inference from these results.

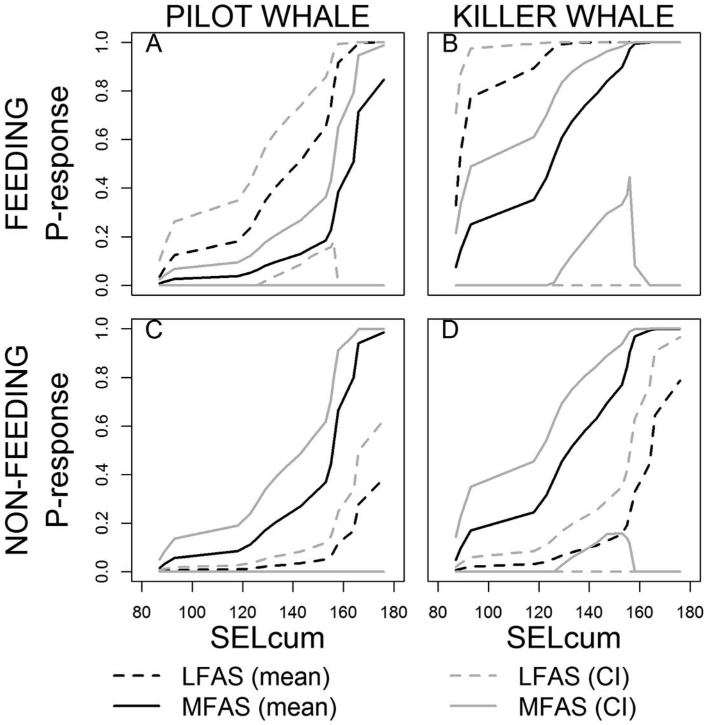

Fig. 4. The probability of a response occurring during LFAS (dashed line) and MFAS (solid line) exposures versus received acoustic energy (SELcum (dB re 1 μPa$^2$s)) for severity level = 2 for long-finned pilot whales (A and C) and killer whales (B and D) and when behavioral state = feeding (A and B) and non-feeding (C and D). Mean probabilities are all shown in black, while 95% confidence intervals are shown in grey. Similar plots for severity levels 1 and 3 are shown in Appendix: Figs. A3 and A4.

## Discussion

### *Recurrent event survival analysis as applied to data from acoustic exposure experiments: pros and cons*

The use of recurrent event survival analysis within the context of cetacean CEEs has allowed us to generate dose-response severity functions from an integrated multi-signal, multi-species model whilst dealing appropriately with censored data. The model was effective in estimating model parameters despite low sample size.

Recurrent event survival analysis has enabled us to expand upon previous exposure-response analysis of CEE data (Antunes et al. 2014, Miller et al. 2014), allowing us to generate functions for responses of different types whilst acknowledging that multiple responses may have been observed within one exposure session and by one individual across multiple exposure sessions. Here we chose to model responses of different severity levels, but an alternative would be to divide the data into behavioral categories instead and model the onset of different response types (e.g., vocal response, dive response, avoidance response). The latter would allow us to determine whether different contexts result in different types of responses as well as different severity levels.

One of the main motivations for searching the medical literature for an analysis method was to find a framework that would accommodate censored data. In experiments such as BRS CEEs it is important to include the data from exposures when animals did not respond, to avoid negative bias whereby we might predict animals, on average, to respond at lower doses than we observed. In this study, the main visible consequence of including the censored data was the very large confidence intervals seen in Fig. 2. This is due to the nature of these data, which were collected as part of a dose-escalation study whereby nearly all of the data points at SELcum above 165dB were censored data points representing animals that did not respond to the maximum sound level received. The preponderance of censored data for high SELcum values leads to very large uncertainty over this part of the function.

The output of the model can be viewed either as dose-response functions for each severity level averaged across all covariates (Fig. 2), or as dose-response functions for particular combinations of covariates (Figs. 3 and 4). The output from our case study data set shows little difference in the functions for the mild and moderate severity levels, which implies a similar probability of observing a mild or moderate response for a given dose. However, it should be noted that the similarity in these functions is partly an artifact of our assumption that if a moderate or severe response was observed, but not a mild response, then all responses at least as severe as that response have also occurred at the same received level. In this data set, this assumption resulted in 19 instances (out of a possible 31) where a mild response was allocated the received level observed for a moderate response. The effect of this assumption would be reduced in data sets where more mild responses were observed. Alternatively, we could remove the need for such an assumption if we chose to model different types of response (e.g., vocal, movement) rather than different intensities of response. However, this assumption predominantly affects the function for the mild level 1 severity responses, which may be of least interest from a regulatory point of view.

Miller et al. (2014) and Antunes et al. (2014) modelled the onset of the first avoidance/horizontal movement response observed for each exposure session using a Bayesian hierarchical model which fitted the observed thresholds to an assumed, underlying dose-response model. The advantage of the Bayesian framework over the survival analysis approach is that previous knowledge could be used to derive initial estimates (priors) of dose-response, which can be useful when faced with such small sample sizes. Miller et al. (2014) used simulation studies to demonstrate that the Bayesian approach was able to recover an underlying dose-response function with limited bias based upon the size of the data set they were able to obtain in the field, providing reassurance that the posterior estimates of the dose-response model were not overly influenced by the uninformative priors used in that case, but rather reflected trends in the data. In addition, the priors and the underlying form of the model constrain the resulting exposure-response function to conform to the expected shape for such functions. The Cox proportional hazards model has no such constraints and therefore there is the possibility of extremely wide confidence intervals for regions of acoustic dose where data are sparse, or exposure-response functions with unexpected shapes. We did not experience the latter problem with this case study data set but it is something that may need to be addressed in future iterations. The advantage of the survival analysis approach over the Bayesian hierarchical model is the ability to model multiple response events within one framework rather than carrying out multiple, independent analyses on each response type.

The proportional hazards assumption that there is a constant hazard ratio across time (SELcum) underpins the use of Cox proportional hazards models, and if this assumption is violated then the use of these models is inappropriate. In our case study the assumption was met, but only marginally. It is worth noting that there are options to consider if the assumption is violated (e.g., see Kleinbaum and Klein [2005] for details), and we discuss these briefly here. The first option would be to stratify by the covariate that is causing the violation. In our case we had already chosen to stratify by severity level to address our particular question of interest and so it would not have been viable to stratify by another covariate. The second option would be to add an interaction term between the covariate causing violation and SELcum, or a function of SELcum, which is referred to as an extended Cox model (Kleinbaum and Klein 2005).

### *Case study results*

From a biological perspective, the novelty of the model fitted here is the inclusion of a contextual variable (pre-exposure behavioral state), and the combining of species in one model. Both Miller et al. (2014) and Antunes et al. (2014) discussed the importance of context in explaining the remaining within- and between-whale variability in their models, but neither study modelled the effect of contextual variables beyond sonar frequency and exposure order. This is understandable given the data, as in both cases they were fitting to data from a single species and therefore sample size precluded the inclusion of more than a few covariates. Even when we combined the data from three species, our sample size was fairly limited, and by including a contextual variable and first-order interaction terms we have pushed the data to its limits. When we divide the data by species, signal type and behavioral state this results in some categories with a sample size of one. We are therefore cautious in our biological interpretation of the model output from this case study. However, there are similarities in the results between our study and previous analyses of these same data using different methodologies (Miller et al. 2012, 2014, Antunes et al. 2014), which gives us confidence in the application of the model.

It is also worth noting that while we show results from backwards model selection based on *p*-values, there are alternative model selection options. We also explored the use of AIC-based model selection with the same data set and the selected model was the same using both approaches, however the AIC method resulted in quite a few models with AIC values close to the selected model and so, given this awareness of competing models, we further emphasize caution when interpreting the biological results.

Our results suggest that pre-exposure behavioral state (feeding or non-feeding) can make a difference to the way in which an individual responds to exposure and that this differs depending upon species and signal type. This agrees with research on bowhead whales which were observed to react differently to seismic airgun sounds depending on whether they were feeding or migrating (Richardson et al. 1986, 1999, Miller et al. 2005), and on blue whales which were found to have different levels of responsiveness depending on their diving and foraging behavior (Goldbogen et al. 2013). The importance of context has long been recognized (Southall et al. 2007, Ellison et al. 2012, Miller et al. 2012) and here we have only managed to investigate one aspect of this. However, as sample size increases, we anticipate that it will be possible to include other relevant contextual variables, for example relating to social settings.

Combining the data from multiple species into one model has allowed us to borrow strength across the individual data sets and increase the sample size, providing an opportunity to incorporate more explanatory variables as described above. It has also allowed us to compare directly the responses of each species within one model and determine whether the descriptive differences resulting from Miller et al. (2014) and Antunes et al. (2014) whereby killer whales were deemed more sensitive than pilot whales is a quantitatively significant difference (Table 1). Direct comparison between species is complicated by the interaction of their responses with context, but the killer whale is significantly more sensitive than pilot whales and sperm whales in all situations, except when exposed to LFAS whilst in a non-feeding state. We believe there is significant scope to expand the multi-species approach using this framework.

### *Future work*

We have demonstrated the utility of this framework using the 3S BRS data on killer whales, long-finned pilot whales and sperm whales as a case study. When further expert-scored data become available, we anticipate using the framework as a tool for meta-analysis across multiple BRS projects. Despite the anticipated increase in sample size, there will undoubtedly be new analytical challenges relating to an increase in the number of covariates, and levels within covariates.

A significant challenge will be the inclusion of no-sonar control exposures, where the source vessel approaches the animal without sonar transmissions. Although it is desirable to include the data from these experimental sessions in our analysis to better understand the contribution of the vessel approach to responsiveness, they have been excluded to date because of difficulties in defining an appropriate dose metric for inclusion in analysis. The sonar dose is zero and not escalating, and other, potentially relevant sound metrics such as ship propulsion noise, have not been measured. Expert scored response data have been published for the no-sonar control exposures (Miller et al. 2012, Sivle et al. 2015) and results showed that these exposures had both the fewest responses scored per session and the lowest proportion of sessions with maximum severity scores of 4 or greater. This implies that there was a lower probability of individuals responding to the approaching vessel than to sonar exposure, and that when they did respond it was to a lesser extent. However, without including data from both sonar exposures and no-sonar control exposures in one dose-response analysis framework we cannot quantitatively evaluate the features of the sonar exposure that are driving a response, in particular the role of the approaching vessel. Miller et al. (2014) also excluded the data from no-sonar control exposures in their Bayesian dose-response analysis; they believed that it was unlikely that the responses identified by expert scoring were a result of ship propulsion noise alone, but that in some cases the approaching vessel may have had an effect on overall responsiveness. One possibility for the future may be to investigate different dose metrics, such as whale-vessel range, which, unlike received sound level, is a measure that is available for both sonar and no-sonar control exposures.

### *Conclusions*

The analogous nature of the scientific questions, experimental approaches, and data from BRS CEEs and medical studies has led to a novel application and extension of recurrent event survival analysis within ecology. Replacement of time with received acoustic energy allowed us to produce event curves relative to a relevant metric (sound dose) and there are likely many other metrics in ecology to which this extension may be applicable. We believe the framework is an effective analytical tool for fitting dose-response severity functions, a key output of BRSs that are much needed by regulatory communities.

## Acknowledgments

This work was financially supported by the U. S. Office of Naval Research grant N00014-12-1-0204, under the project "Multi-study Ocean acoustics Human effects Analysis" (MOCHA). We thank the MOCHA working group for input and feedback at various stages of analysis development. Although the MOCHA project itself is funded by the US Office of Naval Research, we rely on the participation of the working group members whose time and efforts are supported by a wide range of institutes and funders. P. L. Tyack received funding from the MASTS pooling initiative (The Marine Alliance for Science and Technology for Scotland) and their support is gratefully acknowledged. MASTS is funded by the Scottish Funding Council (grant reference HR09011) and contributing institutions. The case study data were provided by the 3S project, which was funded by the U.S. Office of Naval Research, the Norwegian Ministry of Defense, the Netherlands Ministry of Defense, and WWF Norway. We thank an anonymous reviewer for providing constructive feedback on a previous draft of this manuscript.

## Literature Cited

Antunes, A., P. H. Kvadsheim, F. P. A. Lam, P. L. Tyack, L. Thomas, P. J. Wensveen, and P. J. O. Miller. 2014. High thresholds for avoidance of sonar by free-ranging long-finned pilot whales (*Globicephala melas*). Marine Pollution Bulletin 83(1):165–180.

Box-Steffensmeier, J. M., and B. S. Jones. 2004. Event history modeling: a guide for social scientists (analytical methods for social research). Cambridge University Press, Cambridge, UK.

Cox, D. R. 1975. Partial likelihood. Biometrika 62:269–76.

D'Amico, A., R. C. Gisiner, D. R. Ketten, J. A. Hammock, C. Johnson, P. L. Tyack, and J. Mead. 2009. Beaked whale strandings and naval exercises. Aquatic Mammals 35:452–472.

DeRuiter, S. L. 2010. Marine animal acoustics. Pages 425–474 *in* X. Lurton, editor. An introduction to underwater acoustics. Praxis, Chichester, UK.

DeRuiter, S. L. et al. 2013. First direct measurements of behavioural responses by Cuvier's beaked whales to mid-frequency active (MFA) sonar. Biology Letters 9:20130223.

Dunlop, R. A., M. J. Noad, D. H. Cato, E. Kniest, P. J. O. Miller, J. N. Smith, and M. D. Stokes. 2013. Multivariate analysis of behavioural response experiments in humpback whales (*Megaptera novaeangliae*). Journal of Experimental Biology 216:759–770.

Ellison, W. T., B. L. Southall, C. W. Clark, and A. S. Frankel. 2012. A new context-based approach to assess marine mammal behavioral responses to anthropogenic sound. Conservation Biology 26:21–28.

Goldbogen, J. A. et al. 2013. Blue whales respond to simulated mid-frequency military sonar. Proceedings of the Royal Society B 280(1765):20130657.

Grambsch, P., and T. Therneau. 1994. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81:515–526.

Holmes, T. L., R. L. Knight, L. Stegall, and G. R. Craig. 1993. Responses of wintering grassland raptors to human disturbance. Wildlife Society Bulletin 21:461–468.

Johnson, M. P., and P. L. Tyack. 2003. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. IEEE Journal of Oceanic Engineering 28:3–12.

Klein, M. L. 1993. Waterbird behavioral responses to human disturbance. Wildlife Society Bulletin 21:31–39.

Klein, J. P., and M. L. Moeschberger. 2003. Survival analysis: techniques for censored and truncated data. Second edition. Springer, Berlin, Germany.

Kleinbaum, D. G., and M. Klein. 2005. Survival analysis: a self-learning text. Third edition. Springer, Berlin, Germany.

Kvadsheim, P. et al. 2011. Behavioural response studies of cetaceans to naval sonar signals in Norwegian waters –3S-2011 cruise report. http://rapporter.ffi.no/rapporter/2011/01289.pdf

Kvadsheim, P. et al. 2014. Behavioural responses of cetaceans to naval sonar signals—the 3S-2013 cruise report. http://rapporter.ffi.no/rapporter/2014/00752.pdf

Kvadsheim, P., F. P. Lam, P. J. Miller, P. Wensveen, F. Visser, L. Doksæter, L. Kleivane, C. Curé, P. Ensor, S. van Ijsselmuide, and R. Dekeling. 2012. Behavioural response studies of cetaceans to naval sonar signals in Norwegian waters –3S-2012 cruise report. http://rapporter.ffi.no/rapporter/2012/02058.pdf

Miller, G. W., V. D. Moulton, R. A. Davis, M. Holst, P. Millman, A. MacGillivray, and D. Hannay. 2005. Monitoring seismic effects on marine mammals—southeastern Beaufort Sea, 2001-2002. Pages 511–542 *in* S. L. Armsworthy, P. J. Cranford, and K. Lee, editors. Offshore oil and gas environmental effects monitoring: approaches and technologies. Battelle Press, Columbus, Ohio, USA.

Miller, P. J., R. Antunes, A. C. Alves., P. H. Kvadsheim, L. Kleivane, N. Nordlund, F. P. A. Lam, S. van Ijsselmuide, F. Visser, and P. L. Tyack. 2011. The 3S experiments: studying the behavioural effects of naval sonar on killer whales (*Orcinus orca*), sperm whales (*Physeter macrocephalus*), and long-finned long-finned pilot whales (*Globicephala melas*) in Norwegian waters. Scottish Oceans Institute Technical Report SOI-2011- 001. http://soi.st-andrews.ac.uk/documents/424.pdf

Miller, P. J. O., R. N. Antunes, P. J. Wensveen, F. I. P. Samarra, A. C. Alves, P. H. Kvadsheim, L. Kleivane, F. P. A. Lam, M. A. Ainslie, P. L. Tyack, and L. Thomas. 2014. Dose-response relationships for the onset of avoidance of sonar by free-ranging killer whales. Journal of the Acoustical Society of America 135:975.

Miller, P. J. O., M. P. Johnson, P. T. Madsen, N. Biassoni, M. Quero, and P. L. Tyack. 2009. Using at-sea experiments to study the effects of airguns on the foraging behaviour of sperm whales in the Gulf of Mexico. Deep-Sea Research I 56:1168–1181.

Miller, P. J. O., P. H. Kvadsheim, F. P. A. Lam, P. J. Wensveen, R. Antunes, A. C. Alves, F. Visser, L. Kleivane, P. L. Tyack, and L. Doksæter. 2012. The severity of behavioral changes observed during experimental exposures of killer (*Orcinus orca*), long-finned pilot (*Globicephala melas*), and sperm (*Physeter macrocephalus*) whales to naval sonar. Aquatic Mammals 38(4):362–401.

Moya-Loraño, J., and D. H. Wise. 2000. Survival regression analysis: a powerful tool for evaluating fighting and assessment. Animal Behaviour 60:307–313.

Muenchow, G. 1986. Ecological use of failure time analysis. Ecology 67(1):246–250.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.r-project.org

Preisler, H. K., A. A. Ager, and M. J. Wisdom. 2006. Statistical methods for analyzing responses of wildlife to human disturbance. Journal of Applied Ecology 43:164–172.

Richardson, W. J., C. R. Greene, Jr., C. I. Malme, and D. Thompson. 1995. Marine mammals and noise. Academic Press, San Diego, California, USA.

Richardson, W. J., G. W. Miller, and C. R. Greene, Jr. 1999. Displacement of migrating bowhead whales by sounds from seismic surveys in shallow waters of the Beaufort Sea. Journal of the Acoustical Society of America 106:2281.

Richardson, W. J., B. Würsig, and C. R. Greene, Jr. 1986. Reactions of bowhead whales, *Balaena mysticetus*, to seismic exploration in the Canadian Beaufort Sea. Journal of the Acoustical Society of America 79:1117–1128.

Scherm, H., and P. S. Ojiambo. 2004. Applications of survival analysis in botanical epidemiology. Phytophathology 94:1022–1026.

Sivle, L. D., P. H. Kvadsheim, C. Curé, S. Isojunno, P. J. Wensveen, F. P. A. Lam, F. Visser, L. Kleivane, P. L. Tyack, C. M. Harris, and P. J. O. Miller. 2015. Severity of expert-identified behavioural responses of humpback whale, minke whale and northern bottlenose whale to naval sonar. Aquatic Mammals 41(4).

Southall, B. L. et al. 2007. Marine mammal noise exposure criteria: initial scientific recommendations. Aquatic Mammals 33:411–521.

Southall, B. L., D. Moretti, B. Abraham, J. Calambokidis, S. L. DeRuiter, and P. L. Tyack. 2012. Marine mammal behavioral response studies in southern California: advances in technology and experimental methods. Marine Technology Society Journal 46:48–59.

Stankowich, T. 2008. Ungulate flight responses to human disturbance: a review and meta-analysis. Biological Conservation 141:2159–2173.

Therneau, T. M., and P. M. Grambsch. 2000. Modeling survival data: extending the Cox model. Springer, New York, New York, USA.

Therneau, T. 2015. A package for survival analysis in S. R package version 2.38. http://CRAN.R-project.org/package=survival

Thomson, J. L., and W. E. Copes. 2009. Modeling disease progression of Camellia twig blight using a recurrent event model. Phytopathology 99(4):378–384.

Tyack, P. L. et al. 2011. Beaked whales respond to simulated and actual navy sonar. PLoS ONE 6(3):e17009.

Woodall, C. W., P. L. Grambsch, and W. Thomas. 2005. Applying survival analysis to a large-scale forest inventory for assessment of tree mortality in Minnesota. Ecological Modeling 189:199–208.

Zabel, R. W., B. J. Burke, M. L. Moser, and C. C. Caudill. 2014. Modeling temporal phenomena in variable environments with parametric models: an application to migrating salmon. Ecological Modeling 273:23–30.

## Supplemental Material

### Ecological Archives

The Appendix is available online: http://dx.doi.org/10.1890/ES15-00242.1.sm