2023-09-27

Kernel density estimation of conditional distributions to detect responses in satellite tag data

Hewitt, J

https://pearl.plymouth.ac.uk/handle/10026.1/21732

METHODOLOGY

Kernel density estimation of conditional distributions to detect responses in satellite tag data

Joshua Hewitt[1]*, Alan E. Gelfand[1], Nicola J. Quick[2,5], William R. Cioffi[2,3], Brandon L. Southall[2,3], Stacy L. DeRuiter[4] and Robert S. Schick[2]

Abstract

Background: As levels of anthropogenic noise in the marine environment rise, it is crucial to quantify potential asso-ciated effects on marine mammals. Yet measuring responses is challenging because most species spend the majority of their time submerged. Consequently, much of their sub-surface behavior is difficult or impossible to observe and it can be difficult to determine if—during or following an exposure to sound—an observed dive differs from previously recorded dives. We propose a method for initial assessment of potential behavioral responses observed during con-trolled exposure experiments (CEEs), in which animals are intentionally exposed to anthropogenic sound sources. To identify possible behavioral responses in dive data collected from satellite-linked time–depth recorders, and to inform the selection and parameters for subsequent individual and population-level response analyses, we propose to use kernel density estimates of conditional distributions for quantitative comparison of pre- and post-exposure behavior.

Results: We apply the proposed method to nine Cuvier's beaked whales (*Ziphius cavirostris*) exposed to a lower-amplitude simulation of Mid-Frequency Active Sonar within the context of a CEE. The exploratory procedure provides evidence that exposure to sound causes animals to change their diving behavior. Nearly all animals tended to dive deep immediately following exposure, potentially indicating avoidance behavior. Following the initial deep dive after exposure, the procedure provides evidence that animals either avoided deep dives entirely or initiated deep dives at unusual times relative to their pre-exposure, baseline behavior patterns. The procedure also provides some evidence that animals exposed as a group may tend to respond as a group.

Conclusions: The exploratory approach we propose can identify potential behavioral responses across a range of diving parameters observed during CEEs. The method is particularly useful for analyzing data collected from animals for which neither the baseline, unexposed patterns in dive behavior nor the potential types or duration of behavio-ral responses is well characterized in the literature. The method is able to be applied in settings where little a priori knowledge is known because the statistical analyses employ kernel density estimates of conditional distributions, which are flexible non-parametric techniques. The kernel density estimates allow researchers to initially assess poten-tial behavioral responses without making strong, model-based assumptions about the data.

Keywords: Discrepancy statistics, Kernel density estimator, Tail probability, Pre- and post-data windows, Summary statistics, Time-series, Behavioral response

*Correspondence: joshua.hewitt@duke.edu

[1] Department of Statistical Science, Duke University, Durham, NC, USA
Full list of author information is available at the end of the article

Background

Following several high-profile strandings of marine mammals associated with naval sonar activities [see: 1, 2], researchers have conducted behavioral response

Hewitt *et al. Animal Biotelemetry* (2022) 10:28

Page 2 of 15

studies to better understand how sensitive species respond to these signals [3]. Within these behavioral response studies, researchers monitor animal behavior before, during, and after controlled exposure experiments (CEEs). The pre-exposure behavior data are useful both in describing the baseline behavior for individuals of the species and to evaluate potential changes in behavior as a function of exposure in known conditions, each of which may inform ocean noise exposure criteria and regulations [4]. Both horizontal avoidance from the sound source and changes in diving and foraging patterns are of specific interest, because these response behaviors can be tied to longer-term fitness consequences [5]. While many of these experiments use high-resolution archival tags that remain on the animals for hours up to several days, longer-term datasets with lower resolution may better capture the range of unexposed behavior by leveraging more pre-exposure data [6]. Both high- and low-resolution data remain challenging to analyze for species in which baseline (i.e., pre-exposure) biology and behaviors are not fully characterized [viz., 7, 8].

We propose a novel exploratory data analysis method for response detection that uses kernel density estimates (KDEs) of conditional distributions to analyze behavior summaries across paired time windows. In particular, we demonstrate how to use the method to explore changes in diving patterns derived from depth data, including average depth, ascent/descent rates and vertical distance traveled, and the sequence of deep vs. shallow dives. Properties of these behavior summaries have been studied directly [7, 8] and through the use of various hidden Markov models (HMMs) [9–11], but the impact of response remains difficult to quantify.

The exploratory method we propose quantifies how common the post-exposure behavior is relative to the immediate pre-exposure activity. Comparisons are made separately for each individual, then patterns in changes can be explored across individuals. The definition of "common" is empirical, derived from the distribution of activity between analogous pairs of observation windows from the longer-term baseline dataset. We apply the method to dive data from deep-diving Cuvier's beaked whales (*Ziphius cavirostris*) collected over longer time periods (days to weeks) coincident with controlled exposure experiments [12]. While biologists often study changes in animals following potentially impactful events, the innovation here is the specification and use of a focused screening procedure to uncover evidence of outliers that suggest possible departure from baseline behavior following exposure to sound. The method can be adapted to evaluate possible changes in behavior in other animal groups.

Our exploratory method analyzes behavioral changes in animals whose baseline activity is (a) only partially characterized in the literature, and (b) observed to vary with respect to latent states that may be difficult to explicitly model or estimate. These two factors pose significant analytical challenges. Faced with this challenge, we focus on three reasons why it is important to provide new tools for exploratory analysis or screening of possible behavioral responses in such animals. When baseline behaviors are only partially understood and variable, (i) we do not know if there has been a response; (ii) if a response did occur, we do not know what type of response it might be; and (iii) we do not know when a response occurred and for how long it may have lasted after an exposure. The method we propose helps researchers quickly identify possible outliers from among a large collection of ways behavior could potentially change, which would otherwise be time consuming to manually review.

In the statistics literature, exploratory analysis using descriptive statistics and other data summaries may be suggested as an analytic approach. However, exploratory analyses are customarily designed to motivate or validate standard types of modeling decisions regarding, e.g., choice of distribution or parameters, such as the mean or variance of behaviors, or correlation between behaviors. If we are able to identify a response and its nature, then we might seek to build a model for such a response, e.g., [11] and extensions thereof. Here, our focus is on exploring the data to screen for possible responses.

Controlled exposure experiments

The type of tags deployed in a controlled exposure experiment influences the variety and types of baseline behaviors and responses that can be studied. High-resolution multidimensional data, like DTAGs' sub-second readings of depth, sound levels, and acceleration, are generally only available for relatively short periods of time (i.e., hours to days) due to constraints on battery life, memory, or tag recovery [13, 14]. Satellite-linked tags, by comparison, can record data for much longer periods of time (i.e., weeks to months), but generally at lower resolution. For example, for our study species, tag configuration, and location, satellite-linked tags only provide an average of 6 longitude/latitude locations per day and 12 binned depth readings per hour (i.e., depth is 30-60m). Although lower in resolution than DTAGs, longer-term datasets derived from satellite tags can leverage more pre-exposure data to characterize the range of unexposed behavior [6].

CEE data are challenging to analyze because individual, social, and environmental factors influence both baseline behavior of an animal as well as its likelihood to respond to anthropogenic sound [15]. Extreme deep-divers, like Cuvier's beaked whales, pose additional challenges

Hewitt *et al. Animal Biotelemetry*   (2022) 10:28

Page 3 of 15

because baseline behaviors are not well understood and differ substantially from other diving mammals [7]. A priori, researchers may not know what the nature of the behavioral response might be, the duration of time over which it may occur, or whether there is a lag in the response relative to the exposure. All of these issues make complex behavior and responses inherently difficult to study.

Inferential approaches
Strategies for extracting behavioral response of whales to exposure presented in the literature include, but are not limited to, expert opinion [16, 17], regression-based approaches [18], Mahalanobis distance approaches [19, 20] and an HMM approach [9]. Mahalanobis distance approaches are typically applied to a multivariate data stream consisting of nearly continuous data obtained from a DTAG [13]. First, these data are processed to compute biologically relevant variables, for example fluke stroke rate, from accelerometer and magnetometer data. Then, using this vector time-series, a Mahalanobis distance is calculated and plotted over time in order to assess a behavioral response, i.e., distances that are very large in comparison to pre- and post-exposure behavior [19, 20].

HMMs have been used to study whale diving behavior states, including identification and characterization of foraging states [9, 10, 21]. Time-series models and HMMs describe the temporal evolution of processes, for example, by parametrically modeling the average value of observations and the correlation between successive observations or by modeling the probability of switching between behavioral states and the probability of observing different state-dependent data values or patterns.

We suggest that an HMM approach might be more appropriate as a follow-on to our exploratory screening. We suggest this because, from the satellite tag observations alone, selecting how many latent states there might be or how to structure a transition matrix between states will inevitably introduce subjectivity. Further challenges may arise in specification of the nature of a response (a unique state, altered transition probabilities, or changes in state-dependent distributions). In any event, preliminary analysis to identify the nature of a response could certainly be useful before embarking on such model specification and fitting.

Exploratory approaches
Behavioral ecologists aim to quantify changes in complex systems or behaviors, and our proposed approach allows a structured preliminary investigation to screen a large body of candidate metrics for identifiable changes or differences. We contribute a method that uses kernel density estimates of conditional distributions to identify potentially outlying changes in diving behavior that are observable between pre- and post-exposure time windows. Critically, our method is not a formal hypothesis testing framework, but it is inspired by the way formal hypothesis tests identify unusual patterns in data through tail probabilities of reference distributions. The proposed method's metrics and windows are structured to capture behaviors such that a set of results can be synthesized to both identify whether a response of some type in fact occurred, and to provide insight into explicit post-exposure response.

Ecological processes can include temporal trends and lags that can be challenging to incorporate into parametric, model-based analyses [22]. Studying temporally paired windows accounts for trends at fixed lags, and kernel density estimates can flexibly model empirical patterns in data because they are non-parametric. The approach only offers preliminary analysis because the estimation procedure does not directly share information across lags or metrics. However, more thorough or confirmatory estimation procedures are usually designed to study a small number of specific behaviors at specific time scales, candidates for which we seek to identify here.

The kernel density estimation approach using paired windows is general and can be adapted to study how different animals respond to environmental stressors that are experimentally introduced at known times. The proposed procedure uses nested time windows and a battery of metrics to identify the onset and duration of behavioral responses. The pre-exposure data characterize variability in baseline behavior, serving as a reference against which potential responses are evaluated. We illustrate our procedure through application to nine beaked whale tag records involving known, realistic military sonar signals.

Methods
Data
We used data from nine Cuvier's beaked whales (*Ziphius cavirostris*) that were tagged offshore of Cape Hatteras, NC USA, in 2019 as part of the Atlantic Behavioral Response Study [23]. Each animal was instrumented with a SPLASH10-292 extended depth tag from Wildlife Computers, Inc. (Redmond, WA) in the LIMPET configuration [24], programmed to record depth data, using the time-series function, every 5 min for approximately 2 weeks. The data are illustrated in Fig. 1, which highlights diving data collected immediately before and after exposure for the animal identified internally as ZcTag093_DUML.

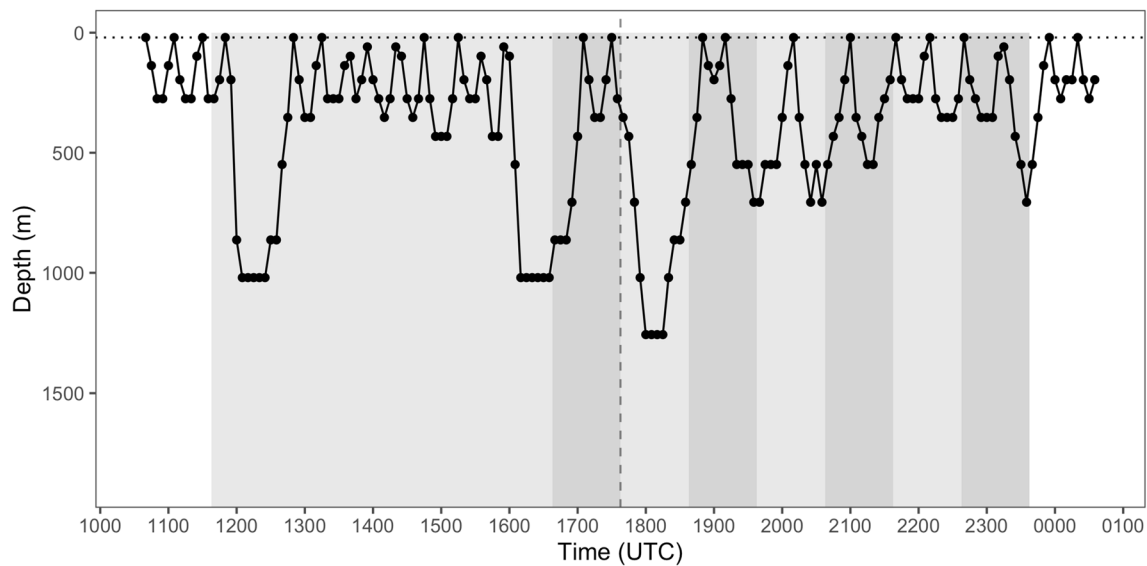Though the tag's depth sensor is accurate to ± 1m, the data returned to the user via satellite telemetry are

Hewitt *et al. Animal Biotelemetry* (2022) 10:28

Page 4 of 15

Fig. 1 Diving behavior surrounding the CEE for ZcTag093_DUML. The vertical dotted line marks the start of the CEE. The grey regions illustrate several combinations of pre- and post-exposure windows analyzed

summarized by low-resolution bins due to bandwidth limitations. For example, an observation recorded at 1,075 m in the water column can be reported as a depth bin centered at 1,091 m with a width of 93.5 m. We base our analysis on the bin centers. Dive data were filtered for potential pressure transducer failures and runaway sensor drift using onboard diagnostics. The animals were exposed to a lower-amplitude simulation of Mid-Frequency Active Sonar (MFAS) from a sound source array [25] deployed off of one of two sport-fishing charter vessels. Table 1 reports the data collection and exposure information for all animals. The field approach in the Atlantic BRS follows typical CEE protocol, of tagging,

pre-exposure baseline, exposure, and post-exposure monitoring [3]. In contrast to previous BRS studies, the use of satellite-linked telemetry tags means that the pre- and post-exposure periods can extend over multiple days (Table 1). We conducted 4 CEEs in 2019, which ranged in duration from 7 to 30 min [23].

Conditional distributions for response behaviors

Each tag records a continuous time-series of depth observations $z_1, \ldots, z_T$. If we assume an exposure begins between times U and $U + 1$, then we refer to $z_1, \ldots, z_U$ as pre-exposure observations, and $z_{U+1}, \ldots, z_T$ as during and post-exposure observations. Table 1 reports

Table 1 Summary of tags and CEEs analyzed in study, including start and end dates for dive data collection, and start date, time and duration for CEE

Tag info.					CEE info.			
ID	Start	End	U	T	ID	Date	Time (UTC)	Duration (min)
ZcTag083_DUML	11-May	21-May	2816	2923	1901	21-May	0904	6.4
ZcTag085_DUML	01-Jun	31-May	3,031	3664	1902	07-Jun	0913	30
ZcTag087_DUML	02-Jun	15-Jun	1451	3620	1902	07-Jun	0913	30
ZcTag088_DUML	02-Jun	15-Jun	1448	3857	1902	07-Jun	0913	30
ZcTag089_DUML	02-Jun	15-Jun	1426	3835	1902	07-Jun	0913	30
ZcTag093_DUML	01-Aug	31-Jul	2071	3683	1903	06-Aug	1137	30
ZcTag095_DUML	12-Aug	25-Aug	2050	3739	1904	19-Aug	1311	30
ZcTag096_DUML	12-Aug	25-Aug	2046	3831	1904	19-Aug	1311	30
ZcTag097_DUML	12-Aug	25-Aug	2041	3826	1904	19-Aug	1311	30

All data were collected in 2019. The variables U and T denote the number of pre-exposure observations used to estimate baseline distributions, and the total number of observations for the tag, respectively

T and U for all animals. The post-exposure window spans a time period during which animals were actively exposed to sonar, potentially responded to sonar after exposure, and potentially returned to baseline behavior. For a selected number of observations $R > 0$ and time lag $L \geq 0$, we screen for changes in diving behavior between the R observations that immediately precede exposure, $w_{pre} = (z_{U-R+1}, \ldots, z_U)$ and R post-exposure observations that begin L time units after exposure, $w_{post} = (z_{U+L+1}, \ldots, z_{U+L+R})$. The grey bands in Fig. 2 illustrate the temporal relationship between w_{pre} and w_{post} when each window is 1h long and separated by a 1 h lag. Our full screening suite uses a series of nested windows to quantify onset and duration of responses. The main window is 0–6 h, with 0–3 h and 3–6 h nested within, and further nesting via the hourly periods 0–1 h through 5–6 h. The 1-h windows are chosen because this covers the typical length of a beaked whale's presumed foraging dive duration—dives over 33 min in duration [7]. The maximum window length of 6h focuses the study on immediate responses to exposure. Summary statistics $s_{k1} = t_k(w_{pre})$ and $s_{k2} = t_k(w_{post})$ characterize the kth diving behavior of interest in w_{pre} and w_{post}. The statistical procedures are described below, and the summary functions $t_k(\cdot)$ are described in "Summary statistics for behaviors" section.

We compare the observed response to a CEE (i.e., relationships between w_{pre} and w_{post}) to pairs of windows in the pre-exposure period, similar to

permutation tests and randomization tests. The data for the kth behavior are the ordered pair of summary statistics (s_{k1}, s_{k2}), and they are evaluated with respect to the conditional distribution $[S_{k2}|S_{k1}]$, where $S_{k1} = t_k(W_1)$ and $S_{k2} = t_k(W_2)$ are summary statistics for a hypothetical pair of windows W_1 and W_2. The windows W_1 and W_2 have the same widths and separation as the paired experimental windows w_1 and w_2, but both W_1 and W_2 follow the distribution of pre-exposure windows.

The pre-exposure data provide $U - 2R - L + 1$ pairs of windows we can use to approximate $[S_{k2}|S_{k1}]$. The method we present below uses each of the $U - 2R - L + 1$ pre-exposure pairs once to approximate $[S_{k2}|S_{k1}]$, similar to how a permutation test generates a reference distribution by enumerating all relevant re-combinations of the source data. Alternatively, if $U - 2R - L + 1$ were too large to enumerate, the reference distribution could be approximated by sub-sampling the pre-exposure pairs. Sub-sampling forms the theoretical basis for randomization tests, as sub-sampling yields a procedure to generate reference distributions that are asymptotically equivalent to permutation tests as the number of sub-samples grows [22, 26, 27].

The conditional distribution $[S_{k2}|S_{k1}]$ is a reference distribution that describes paired behavior patterns (i.e., pre/post relationships) in the absence of sound exposure. It is a function of S_{k1} that quantifies the distribution of behavior that is typically observed in W_2
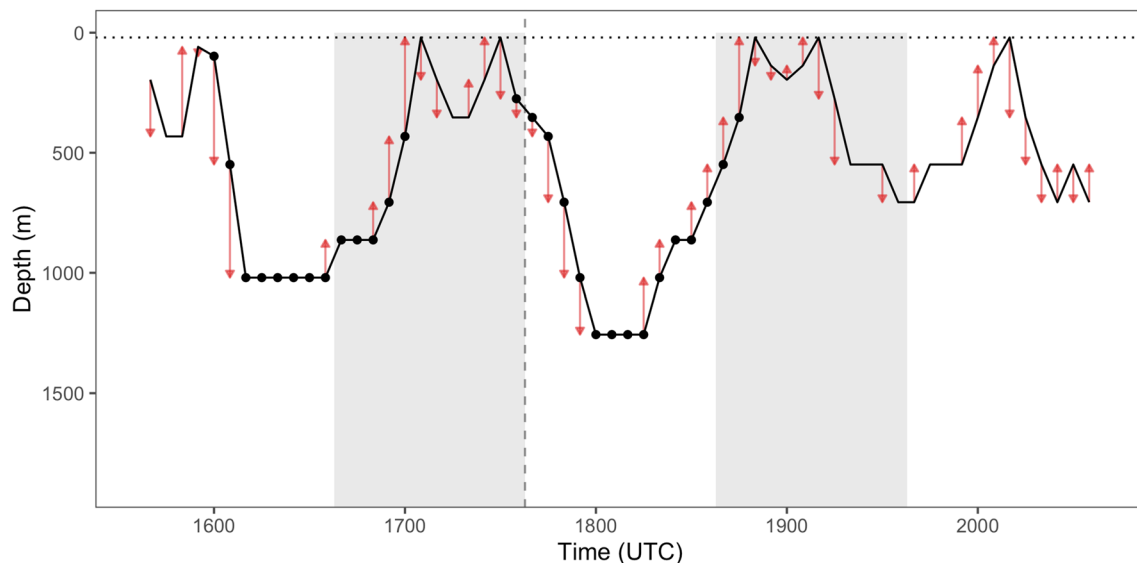
Fig. 2 Diving behavior surrounding the CEE for ZcTag093_DUML. The left grey band highlights a 1-h-long pre-exposure window immediately preceding the exposure time, indicated by the vertical dotted line. The right grey band highlights a 1-h-long post-exposure comparison window that begins 1 hour after exposure. The red arrows visualize the vertical change between observations d_j, and the black dots indicate depths that were part of a deep dive, i.e., observations for which $\ell(z_j) = 1$

Hewitt *et al. Animal Biotelemetry*   (2022) 10:28

Page 6 of 15

when the behavior in the preceding window W_1 is summarized by the value of s_{k1}.

The exploratory procedure we propose uses the tail probabilities of $[S_{k2}|S_{k1}]$ to quantify how unusual the post-exposure behavior s_{k2} is relative to behaviors associated with s_{k1}, as if there were no sound exposure. The lower and upper-tail probabilities of $[S_{k2}|S_{k1}]$ are $P(S_{k2} \leq s_{k2}|S_{k1} = s_{k1})$ and $P(S_{k2} \geq s_{k2}|S_{k1} = s_{k1})$, respectively. Mathematically, the tail probabilities can be computed from the joint density $f_{S_{k1},S_{k2}}(u,v)$ for the pair of summary statistics (S_{k1}, S_{k2}) of random windows W_1 and W_2 via

$$P(S_{k2} \leq s_{k2}|S_{k1} = s_{k1}) = \frac{\int_{-\infty}^{s_{k2}} f_{S_{k1},S_{k2}}(s_{k1},v)dv}{f_{S_{k1}}(s_{k1})}, \tag{1}$$

$$P(S_{k2} \geq s_{k2}|S_{k1} = s_{k1}) = \frac{\int_{s_{k2}}^{\infty} f_{S_{k1},S_{k2}}(s_{k1},v)dv}{f_{S_{k1}}(s_{k1})}, \tag{2}$$

where $f_{S_{k1}}(s_{k1}) = \int_{-\infty}^{\infty} f_{S_{k1},S_{k2}}(s_{k1},v)dv$ is the marginal density for S_{k1} evaluated at s_{k1}. Although the joint density $f_{S_{k1},S_{k2}}(u,v)$ evaluates the likelihood of (S_{k1}, S_{k2}) with $S_{k1} = u$ and $S_{k2} = v$ for any combination of (u, v), its value is only important to us along the vertical line of coordinate pairs that match our pre-exposure conditions, i.e., for pairs in the set $\{(u,v) : u = s_{k1}, v \in \mathbb{R}\}$.

Fig. 3 illustrates the relationship between the joint density $f_{S_{k1},S_{k2}}(u,v)$, the conditional distribution $[S_{k2}|S_{k1}]$, pairs (u, v), and how data will ultimately be analyzed statistically. The figure visualizes the statistical information our exploratory procedure uses to assess possible response with respect to the "time in deep dives" summary statistic (described in "Summary statistics for behaviors" section) computed with no lag separating 6 hour long pre- and post-exposure windows for ZcTag093_DUML. Each point in the plot represents one pair of 6h windows from ZcTag093_DUML's pre-exposure data, used by the KDE described below to approximate $[S_{k2}|S_{k1}]$ from $[S_{k1}, S_{k2}]$ for $k = 3$. The x-axis plots the "time in deep dive" summary statistic for each baseline "pre-exposure" window, and the y-axis plots the same statistic for each baseline "post-exposure" window. The contours represent the KDE's estimate of the baseline distribution $[S_{k1}, S_{k2}]$, which quantifies the relationship between paired windows in the absence of exposure. Vertical cross sections through the contours represent the conditional distribution $[S_{k2}|S_{k1}]$ for specific values of S_{k1}. The vertical cross section the green line highlights represents the distribution of post-exposure behavior we would expect to see for ZcTag093_DUML following exposure if the animal did not respond to the sound source. The
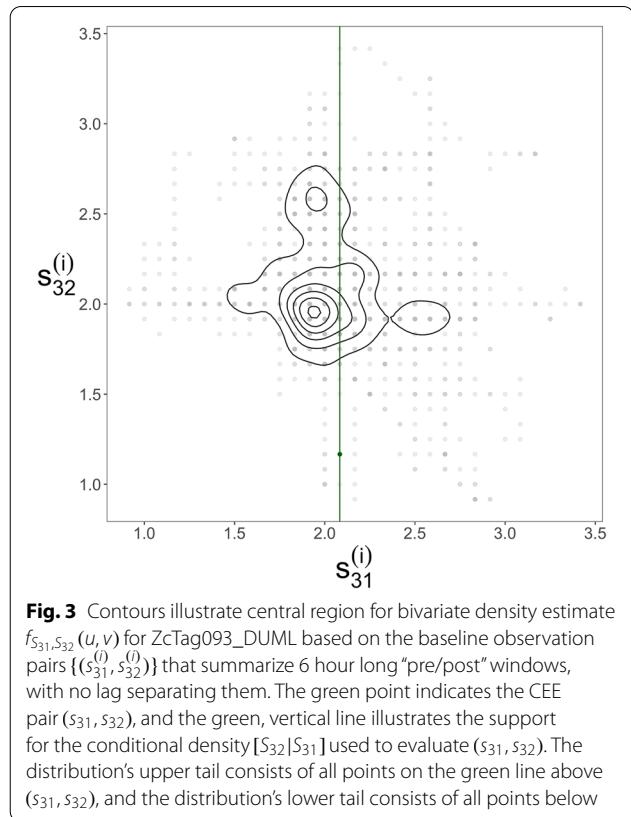

Fig. 3 Contours illustrate central region for bivariate density estimate $f_{S_{31},S_{32}}(u,v)$ for ZcTag093_DUML based on the baseline observation pairs $\{(s_{31}^{(i)}, s_{32}^{(i)})\}$ that summarize 6 hour long "pre/post" windows, with no lag separating them. The green point indicates the CEE pair (s_{31}, s_{32}), and the green, vertical line illustrates the support for the conditional density $[S_{32}|S_{31}]$ used to evaluate (s_{31}, s_{32}). The distribution's upper tail consists of all points on the green line above (s_{31}, s_{32}), and the distribution's lower tail consists of all points below

point on the green line represents the combination of pre- and post-exposure actually observed. Here, we use the left-tail probability to quantify possible response because it is smaller than the right-tail probability. The left-tail probability is the amount of mass that lies on the green line between the y-axis and the green point, and we use the KDE to estimate this value.

In general, the tail probabilities associated with the pair (s_{k1}, s_{k2}) are used to assess whether a response may have occurred. Similar to a p-value from a formal hypothesis test, a small tail probability suggests the animal responded to sonar exposure by deviating from typical behavior patterns. However, since we work outside of a formal hypothesis testing framework, we only interpret small tail probabilities as identifying potentially outlying behavior. A small lower-tail probability (1) suggests the observed post-exposure behavior s_{k2} substantially decreased relative to what is expected following s_{k1} under baseline conditions. Similarly, a small upper-tail probability (2) suggests the post-exposure behavior substantially increased. In either case, behavior changes indicate response to sound exposure because—by the design of CEEs—the tail probabilities are computed with respect to events observed entirely under baseline conditions.

We use a bivariate kernel density estimator $\hat{f}_{S_{k1},S_{k2}}(u,v)$ along with numerical integration to form plug-in estimators of (1) and (2). Kernel density estimators are commonly available in statistical analysis software, for example via the `MASS::kde2d` function in R. For our application, the estimator $\hat{f}_{S_{k1},S_{k2}}(u,v)$ is formed from summary statistics $s_{k1}^{(i)} = t_k(w_{pre}^{(i)})$, $s_{k2}^{(i)} = t_k(w_{post}^{(i)})$ of windows $w_{pre}^{(i)} = (z_{i-R+1}, \ldots, z_i)$ and $w_{post}^{(i)} = (z_{i+L+1}, \ldots, z_{i+L+R})$ for $i \in \mathbb{Z}$, $R \leq i \leq U - R - L$, which lie entirely within the baseline period. The estimator is specified via

$$\hat{f}_{S_{k1},S_{k2}}(u,v) = \frac{1}{U - 2R - L} \sum_{i=R}^{U-R-L} \frac{\phi\left(\left(u - s_{k1}^{(i)}\right)/h_1\right)\phi\left(\left(v - s_{k2}^{(i)}\right)/h_2\right)}{h_1 h_2}, \tag{3}$$

where $\phi(\cdot)$ is the standard normal density kernel and h_1, h_2 are bandwidths for each dimension [28, Chapter 4]. We define each bandwidth via a univariate rule to minimize approximate mean integrated square error with Gaussian kernels [28, eq. 3.28].

We also consider evaluating the summary statistics (s_{k1}, s_{k2}) with respect to the more restricted conditional distribution $[S_{k2}|S_{k1}, G_1]$, where $G_1 = 1$ if the animal is at a deep depth the end of W_1 and $G_1 = 0$ otherwise (i.e., by convention $g_1 = 1$ if $z_U \geq 800$m for the CEE; see [29–31]). We consider the additional conditioning because the distribution of behavior typically observed in W_2 may depend on more than just the summary S_{k1} of behavior in the preceding window W_1. The additional conditioning provided by G_1 specifies $[S_{k2}|S_{k1}, G_1]$ to be a reference distribution describing paired behavior patterns that either end in deep or shallow depth regimes in the absence of sound exposure. As before, estimates for tail probabilities of $[S_{k2}|S_{k1}, G_1]$ use (1) and (2), but with the conditional kernel density estimator $\hat{f}_{S_{k1},S_{k2}|G_1}(u,v)$

instead of $\hat{f}_{S_{k1},S_{k2}}(u,v)$. The conditional density estimator $\hat{f}_{S_{k1},S_{k2}|G_1}(u,v)$ is specified via (3), but uses filtered baseline data in which the pair $(s_{k1}^{(i)}, s_{k2}^{(i)})$ is used only if the diving context of the baseline pair matches the diving context of the experimental pair, i.e., if $g_1^{(i)} = g_1$.

Summary statistics for behaviors

Table 2 describes the $k = 1, \ldots, 10$ basic summary statistics $t_k(w) = \sum_j Y_{jk}$ we use to assess changes between pre- and post-exposure observation windows w_{pre} and w_{post}. Many of these statistics are used in other studies of diving behavior, but we apply them in a modified way. The statistic $t_k(w)$ is formed by aggregating the feature Y_{jk} across the generic observation window $w = (z_1, \ldots, z_R)$. For example, $t_1(w)$ represents the amount of time in w an animal was observed on or near surface depths by defining $Y_{j1} = 1$ if z_j is a sufficiently shallow depth (a user-defined threshold), and 0 otherwise. A small lower-tail probability for $s_{12} = t_1(w_{post})$ suggests an animal spent less time on or near surface depths than is typically observed following $s_{11} = t_1(w_{pre})$. Further exploratory interpretation can be gained by comparing results from other behaviors, for example, to understand if the animal engaged in more deep diving than typical or if the animal spent more time in shallower non-surface depths instead.

Each of the statistics in Table 2 aggregates different characteristics of diving behavior. Consider a generic depth sequence $w = (z_i, \ldots, z_{i+R})$. We say the animal is at a "surface depth" when it is observed to be in the shallowest depth bin (i.e., 0–40 m), and we say the animal is in a "deep dive" when it is observed to be in a

Table 2 Formula and descriptions for non-sequential summary statistics $t_k(w) = \sum_j Y_{jk}$ formed from different choices of Y_{jk}

Order-sensitivity	k	Description of $t_k(w) = \sum_j Y_{jk}$	Y_{jk}
None	1	Time on or near surface	$\mathbf{1}\{z_j \in \mathcal{S}\}$
	2	Average depth	z_j/W
	3	Time in deep dives	$\ell(z_j)$
Pairwise	4	Time spent ascending	$\mathbf{1}\{d_j < 0\}$
	5	Time spent descending	$\mathbf{1}\{d_j > 0\}$
	6	Time spent without vertical movement	$\mathbf{1}\{d_j = 0\}$
	7	Total vertical distance traveled	$|d_j|$
	8	Total ascent distance	$|d_j|\mathbf{1}\{d_j < 0\}$
	9	Total descent distance	$|d_j|\mathbf{1}\{d_j > 0\}$
	10	Total vertical direction changes	$\mathbf{1}\{\mathsf{H}(d_{j+1}) \neq \mathsf{H}(d_j)\}$

The set $\mathcal{S}$ on line 1 represents the range of depths defined to be "surface depths". The indicator function $\mathbf{1}\{\cdot\}$ throughout evaluates to 1 if the argument is true, and 0 otherwise. The Heaviside function $\mathsf{H}(x)$ on line 10 equals 1 if $x > 0$, 1/2 if $x = 0$, and 0 if $x < 0$

Hewitt *et al. Animal Biotelemetry* (2022) 10:28

Page 8 of 15

dive that exceeds 800m [29, 32]. To illustrate, Fig. 2 presents a dive record with all points that constitute a deep dive are labeled. The functions $t_1(w)$ and $t_3(w)$ in Table 2 summarize the number of times a surface depth is observed, and the number of times an animal is observed to be in a deep dive (either the beginning, middle, or end), respectively. Very large (or small) values of the post-exposure statistic suggest deep-diving behavior has substantially decreased (or increased). As described in "Conditional distributions for response behaviors" section, a post-exposure statistic is understood to be "very" large if it has a small tail probability relative to its reference distribution. The function $t_2(w)$ summarizes the average depth at which an animal was observed. The remaining functions $t_4(w), \dots, t_{10}(w)$ summarize vertical depth changes via transformations of the vertical difference between observations $d_j = z_j - z_{j-1}$. To illustrate, Fig. 2 presents a dive record with the differences d_j visualized. Several functions summarize the number of times an animal was observed to be ascending in the water column ($t_4(w)$), descending ($t_5(w)$), or staying at the same depth ($t_6(w)$). Other functions summarize the total vertical distance changes made within the window ($t_7(w)$), or only while ascending ($t_8(w)$) or descending ($t_9(w)$). Lastly, $t_{10}(w)$ summarizes the number of times the animal was observed to change from upward to downward movement (or vice versa).

A quick mathematical study of the statistics in Table 2 demonstrates the statistics can be mostly insensitive to the order in which the data appear. The following theoretical example illustrates the concern. From Table 2, we will always see $t_1(w) = t_1(w_{rev})$ when w_{rev} represents the data analyzed in reverse order $w_{rev} = (z_{i+R}, \dots, z_i)$. We will also always see $t_k(w) = t_k(w')$ for $k = 1, 2, 3$ when w' represents any reshuffling of w. The remaining statistics $t_4(w), \dots, t_{10}(w)$ will only be minimally impacted if pairs of observations are reshuffled. The summary statistics would ideally characterize a window w, so that we would only see $t_k(w) \approx t_k(w')$ if $w \approx w'$. However, the common, high-level summary statistics in Table 2 describe averages and totals, which do not provide such characterization.

We propose to remedy the order-sensitivity issues of the basic statistics in Table 2 by including a modified version in our results. We add the timing-dependent variation $t_k^D(w) = \sum_j j Y_{jk}$ to our exploratory screening suite. The timing-dependent statistic $t_k^D(w)$ modifies $t_k(w)$ by weighting the feature Y_{jk} by its order j in the observation window w. The modified statistics $t_k^D(w)$ are much more sensitive to the sequential nature of diving behavior. Different weighting schemes can be used to increase the sensitivity to changes at different points of an observation

window w, such as replacing j with $1/j$ to weight earlier observations in w greater than later observations.

Tail probabilities for the order-sensitive pair (s_{k1}^D, s_{k2}^D) with $s_{k1}^D = t_k^D(w_{pre})$ and $s_{k2}^D = t_k^D(w_{post})$ should be interpreted alongside results for the basic pair (s_{k1}, s_{k2}). Order-sensitivity makes it more difficult for deviations from baseline behavior to be undetected. We expect it would be more likely to observe a small tail probability for the pair (s_{k1}^D, s_{k2}^D) without seeing a corresponding small tail probability for (s_{k1}, s_{k2}), as opposed to the other way around. If there is only a small tail probability for the pair (s_{k1}^D, s_{k2}^D), we interpret the results as indicating a substantial change in the timing of the activity summarized via $t_k(\cdot)$, but not necessarily the scale of the activity. If there is a small tail probability for (s_{k1}, s_{k2}), we interpret the results as indicating a substantial change in the quantity of the activity summarized via $t_k(\cdot)$. We would simultaneously anticipate the timing of the activity will have changed as well, which would be indicated by a corresponding small tail probability for the pair (s_{k1}^D, s_{k2}^D).

For an individual tag record, we compare a large number of tail probabilities to assess broad behavioral responses in CEEs, which is evocative of the multiple testing problem for formal hypothesis tests. For each window in our screening suite (i.e., each combination of window length R and lag L), we compute a total of 20 tail probabilities for the pairs (s_{k1}, s_{k2}) and (s_{k1}^D, s_{k2}^D) for $k = 1, \dots, 10$. Our screening windows often overlap (i.e., 0–3 h windows overlap 0–6 h windows), so our summary statistics are often correlated. If we were proposing formal hypothesis tests, multiple testing adjustments would not be available for our complex scenarios, which is one reason we present our screening procedure as exploratory. In formal settings, multiple testing adjustments reduce the number of tests that should be rejected, reducing the number of interpretations that can be drawn from a collection of tests. Analogously, we recognize it can be beneficial to not try to interpret all tail probabilities computed for each animal.

To reduce the number of tail probabilities one might interpret from each animal for possible responses, we recommend the Benjamini and Hochberg procedure with $q^* = .05$ for controlling the false discovery rate (FDR) in a collection of independent hypothesis tests [33]. The procedure yields a threshold based on q^*, and individual hypothesis tests are only rejected if their p-value is smaller than the threshold. When all assumptions are met, the procedure guarantees the expected proportion of false-positive rejections is at most q^*. Applied here, the procedure sorts the 180 tail probabilities for each animal and yields an animal-specific threshold for probabilities based on q^* (there are 20 tail probabilities for each of 9 window configurations).

Then, one might only interpret possible behavior changes for the collection of statistics whose tail probability is smaller than the threshold.

Whether one uses the tail probabilities directly or the FDR procedure to screen tail probabilities for possible responses, we can develop an overall impression of the post-exposure behavioral response. Such interpretation would then encourage and inform formal response quantification and modeling, which would allow us to quantify the magnitude of any such change. Our procedure supports the detection of possible responses by identifying a collection of rare behaviors that appear to be outliers relative to pre-exposure data.

Across tag records, we can evaluate the similarity between the tail probabilities for whales i and j by computing the mean-square dissimilarity score $D(i, j)$

$$D(i,j) = \frac{1}{20} \sum_{k=1}^{10} \left(\left(p_{ik} - p_{jk} \right)^2 + \left(p_{ik}^D - p_{jk}^D \right)^2 \right),$$

where p_{ik} and p_{jk} are the left-tail probabilities (1) for the basic statistics in Table 2 for animals i and j, respectively, and p_{ik}^D and p_{jk}^D are left-tail probabilities for the order-sensitive statistics. The score $D(i, j)$ is proportional to the squared Euclidean distance between the vector of tail probabilities for animals i and j. It is sufficient to only focus on left-tail probabilities, because the left-tail and right-tail probabilities always sum to 1 for continuously valued statistics.

Results

We interpret results for summary statistics evaluated with respect to the conditional distributions $[S_{k2}|S_{k1}, G_1]$ discussed in "Conditional distributions for response behaviors" section, which accounts for the influence of diving state at the time of exposure (i.e., deep vs. shallow) on subsequent behaviors. Here, we will focus on the results prior to FDR-screening. The FDR-screened results are presented in the Additional file 1. The types of patterns identified by FDR-screened results are similar to what are presented here, but with fewer "significances" across all animals and with no retained "significances" in the 1–2 h, 2–3 h, ···, 5-6h windows.

The exploratory procedure provides evidence that similar diving behaviors of the study animals change after exposure. At a high level, nearly all animals had small tail probabilities that indicate deep diving activity changed in some manner (i.e., either increase or decrease)

Table 3 Number of animals with either left or right-tail probabilities less than 0.1, conditional on dive state at exposure without order sensitivity $t_k(w)$ (black), and with order sensitivity $t_k^D(w)$ (grey, emphasized)

$t_k(w)$	0-6h	0-3h	3-6h	0-1h	1-2h	2-3h	3-4h	4-5h	5-6h
1. Time near surface	6	5	5	9	2	4	1	1	4
	4	*2*	*3*	*8*	*2*	*4*	*1*	*1*	*5*
2. Average depth		3	3	3				1	4
	1		*3*	*3*				*1*	*3*
3. Time in deep dives	2	5	4	7	3	8	7	8	7
	5	*3*	*5*	*7*	*4*	*8*	*7*	*8*	*8*
4. Time spent ascending	1		2	6	3	1	2	1	4
		1	*2*	*4*	*1*		*1*	*1*	*4*
5. Time spent descending	1	2	2		2		1	1	1
		1	*2*	*1*					*2*
6. Time spent without vertical movement		3	1	4		1	1	2	
		1		*4*					*3*
7. Total vertical distance traveled	5		6				5	7	
	6		*4*	*2*			*4*	*5*	
8. Total ascent distance	6		7	2	1	1	1	1	3
	7	*1*	*6*	*2*		*1*		*1*	*4*
9. Total descent distance	3	1	5				2		
	3	*2*	*2*	*5*			*1*	*1*	
10. Total vertical direction changes	3		2	5		3	3	1	2
	3	*2*	*3*	*2*	*2*	*3*	*3*	*1*	*5*

following exposure (Table 3, row 3). The existence of any response—whether an increase or decrease—is important for marine mammal regulatory applications. Many animals had small tail probabilities both before and after FDR that indicate the amount of time spent near the surface changed as well (Table 3, row 1 before additional FDR screening; after FDR-screening in Additional file 1). Other types of behavior also changed, but with more individual variability (Table 3, remaining rows).

While the deep diving activity of nearly all animals appears to change, the changes indicate similarities and differences between individuals. Animals tended to dive deep immediately following exposure (if they were not exposed while at a deep depth), then either a) avoid subsequent deep dives or b) initiate an additional deep dive at an unusual time. The animals ZcTag093_DUML and ZcTag083_DUML illustrate the two response patterns, respectively. The exploratory procedure indicates the initial deep post-exposure dive was unusual with respect to the pre-exposure behavior for both animals (Tables 4 and 5, rows 1–3, col 0–1 h). Evidence of an initial deep post-exposure dive is consistent with a predator avoidance response hypothesis [34, 35].

The procedure shows individual differences between the animals as well. After the deep dive immediately following exposure, the animal ZcTag093_DUML was not observed to engage in additional deep diving

behavior (Fig. 1). The lack of deep-diving behavior is notable because it is consistent with the hypothesis that ZcTag093_DUML stopped foraging following exposure. Lacking acoustic evidence of actual foraging attempts (echolocation signals), this remains speculative, but the results provide evidence of a response across multiple metrics. The tail probabilities indicate the behavior is unusual with respect to pre-exposure activity due to a substantial decrease in the amount of time ZcTag093_DUML spent in deep dives (Table 4, row 3), a substantially shallower average depth (Table 4, row 2, col 3–6 h), and substantially less vertical distance traveled following exposure (Table 4; row 7; cols 0–6 h, 3–6 h).

By comparison, the animal ZcTag083_DUML did engage in additional deep-diving behavior after the immediate deep dive following exposure (Figs. 4 and 5). But, the tail probabilities indicate ZcTag083_DUML's deep-diving behavior is unusual with respect to pre-exposure activity due to a substantial decrease in the time and distance spent ascending paired with an increase in the quantity and timing of descent distance (Table 5; rows 4, 8, 9; col 3–4 h), followed by a substantial increase in average depth paired with increases in the timing of when vertical distances are traveled (Table 5; rows 2, 7, 8; col 4–5 h). Figures and tables of results for the remaining animals are included in Additional file 1.

Table 4 ZcTag093_DUML, tail probabilities conditional on dive state at exposure without order sensitivity $t_k(w)$ (black), and with order sensitivity $t_k^D(w)$ (grey, emphasized). Probabilities are excluded if greater than 0.1, and are annotated with (−) or (+) to indicate they are for the left or right tail, respectively

$t_k(w)$	0–6h	0–3h	3–6h	0–1h	1–2h	2–3h	3–4h	4–5h	5–6h
1. Time near surface	0.08 (−)			0.00 (−)					
				0.00 (−)					*0.04 (−)*
2. Average depth		0.00 (+)	0.06 (−)	0.00 (+)					
				0.01 (+)					
3. Time in deep dives	0.02 (−)		0.00 (−)	0.00 (+)		0.00 (−)	0.01 (−)	0.00 (−)	0.00 (−)
	0.00 (−)	*0.05 (−)*	*0.00 (−)*	*0.00 (+)*	*0.07 (−)*	*0.00 (−)*	*0.01 (−)*	*0.00 (−)*	*0.01 (−)*
4. Time spent ascending				0.10 (−)			0.01 (+)		0.08 (−)
			0.06 (−)				*0.09 (+)*		*0.08 (−)*
5. Time spent descending									0.02 (+)
			0.04 (+)						*0.01 (+)*
6. Time spent without vertical movement									
7. Total vertical distance traveled	0.07 (−)		0.04 (−)				0.08 (−)		
	0.03 (−)		*0.06 (−)*				*0.03 (−)*		
8. Total ascent distance	0.03 (−)		0.07 (−)						
	0.03 (−)		*0.02 (−)*						*0.06 (−)*
9. Total descent distance			0.08 (−)						
10. Total vertical direction changes	0.03 (−)		0.02 (−)	0.06 (−)			0.01 (−)		0.07 (−)
	0.05 (−)		*0.04 (−)*				*0.07 (−)*		*0.03 (−)*

Hewitt *et al. Animal Biotelemetry*   (2022) 10:28

Page 11 of 15

Table 5 ZcTag083_DUML, tail probabilities conditional on dive state at exposure without order sensitivity $t_k(w)$ (black), and with order sensitivity $t_k^D(w)$ (grey, emphasized). Probabilities are excluded if greater than 0.1, and are annotated with (−) or (+) to indicate they are for the left or right tail, respectively

$t_k(w)$	0–6h	0–3h	3–6h	0–1h	1–2h	2–3h	3–4h	4–5h	5–6h
1. Time near surface				0.01 (−) *0.01 (−)*	0.01 (−) *0.01 (−)*	0.09 (+) *0.06 (+)*			
2. Average depth								0.01 (+)	
3. Time in deep dives	*0.06 (−)*	*0.01 (−)*		0.00 (+) *0.00 (+)*	0.00 (−) *0.00 (−)*	0.00 (−) *0.00 (−)*			0.00 (−) *0.00 (−)*
4. Time spent ascending							0.07 (−)		
5. Time spent descending	0.09 (−)							0.07 (−)	
6. Time spent without vertical movement									
7. Total vertical distance traveled								*0.10 (+)*	
8. Total ascent distance							0.09 (−)	*0.01 (+)*	
9. Total descent distance		*0.07 (−)*		*0.05 (−)*			0.08 (+) *0.07 (+)*		
10. Total vertical direction changes				0.09 (+) *0.05 (+)*					
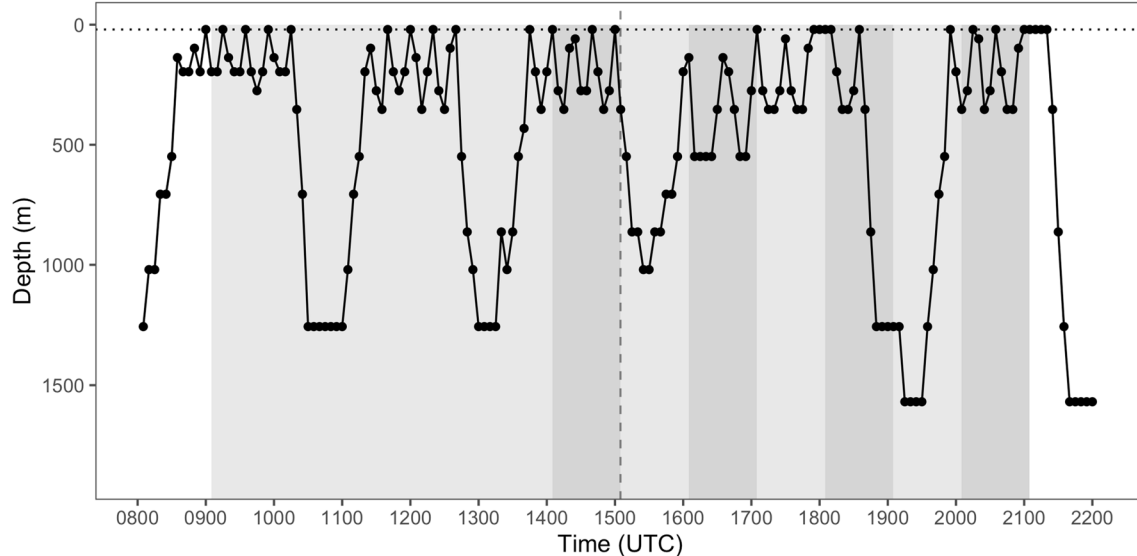
Fig. 4 Diving behavior surrounding the CEE for ZcTag083_DUML. The vertical dotted line marks the start of the CEE. The grey regions illustrate several combinations of pre- and post-exposure windows analyzed

Lastly, the exploratory procedure provides some evidence that animals exposed as a group may tend to respond as a group. The animals ZcTag095_DUML, ZcTag096_DUML, and ZcTag097_DUML were traveling in a group at the time of exposure. The tail probabilities for this group of animals were all much more similar to each other than to other animals (Fig. 6).

Discussion

Here we proposed a simple method to use with satellite tag data to identify possible outliers in diving patterns that suggest changes in animal behavior following exposure to sound. We have illustrated the types of insights
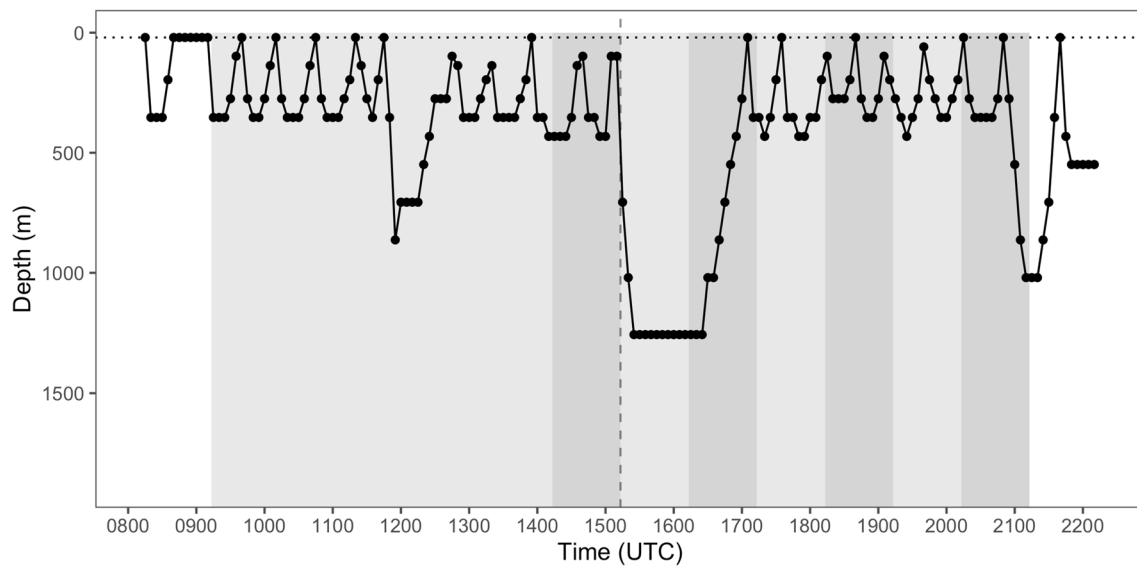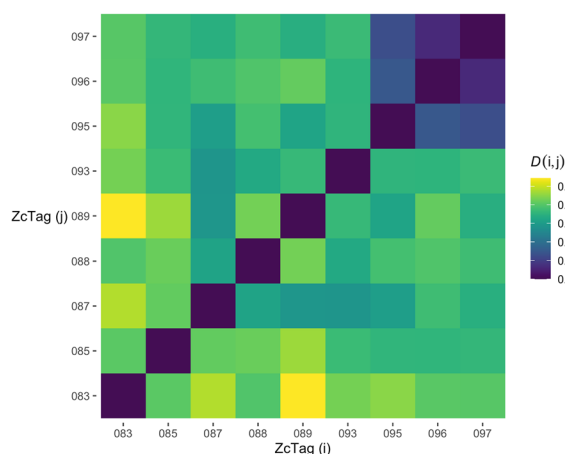
Hewitt *et al. Animal Biotelemetry*   (2022) 10:28

Page 12 of 15

Fig. 5 Diving behavior surrounding the CEE for ZcTag089_DUML. The vertical dotted line marks the start of the CEE. The grey regions illustrate several combinations of pre- and post-exposure windows analyzed

Fig. 6 Dissimilarity scores between pairs of animals in study

it provides on several whales exposed to sonar during controlled exposure experiments. The tool uses kernel density estimates of conditional distributions derived from baseline data to compare post-exposure behavior to pre-exposure behavior. The kernel density estimates are model-free, which allows the method to be applied to biotelemetry data collected from a wide range of animals that are monitored in scenarios where clear interventions lead to a response. In particular, the procedure is ideal for initially analyzing data from cryptic species, such as the beaked whales we work with here. Pre-exposure behavior may be difficult to explicitly model for cryptic species because pre-exposure behavior may only be partially

characterized in the literature. The procedure is also useful for initially analyzing data for non-cryptic species when their post-exposure responses are not yet characterized. Furthermore, window lengths and lags can be tailored to the time scales of important patterns for the animals being studied.

Other model-free methods can be used to assess behavioral responses in CEEs, such as Mahalanobis distances. However, from a distribution-theory perspective, Mahalanobis distances would implicitly assume conditional distributions are normally distributed and linearly related to the overall mean of pre-exposure behavior. Many of our conditional distributions have non-elliptical and potentially multimodal contours, which indicate departures from normality (e.g., Fig. 3).

We propose using basic kernel density estimators in lieu of other non-parametric estimators to keep the method simple, allowing researchers to quickly screen for extremely small tail probabilities that suggest possible responses. While basic kernel density estimators are sufficient for identifying extremely small tail probabilities, hand-tuned models for each summary statistic could potentially model the behavior summaries more accurately, helping to identify tail probabilities for smaller effects. However, simpler strategies may also be beneficial to consider. Basic kernel density estimators use uncorrelated kernel functions, but modifications that use correlated kernel functions have been studied and implemented for use in the ks package for R [36, 37]. Density estimation via multivariate normal mixture models

has also been studied and implemented for use in the `mclust` package for R [38, 39].

The collection of diving behaviors analyzed can be extended with additional summary statistics to study additional types of behavioral responses. We find that even with limited data we are able to quickly identify possible responses to sound. For example, in ZcTag093_DUML we observed a decrease in deep, presumably foraging, dives, which has been noted previously as a possible response in other similar experiments [18, 19]. Without the acoustic data that are available from instruments such as DTAGs [40], we can only speculate as to whether any subsequent deep dives are foraging dives. Nevertheless, this framework offers a useful exploratory data analysis technique to inform subsequent formal modeling, e.g., a process-based model where we estimate transitions into and out of different diving stages [11].

Most animals with possible responses show the rapid onset of a deep dive immediately following exposure due to shifts in summary statistics associated with deep diving (e.g., time spent in deep dives, average depth, time spent descending). We speculate such dives could represent an immediate escape response to avoid the sonar source. If correct, this type of behavior may be especially physiologically costly to the animal owing to the short shallow dive sequence prior to a long dive and a possible lack of sufficient recovery. As above, our approach offers a structured way to screen for possible responses—the results of which can lead to the development of richer, process-based models.

Through our flexible approach to defining and quantifying summary statistics, we have shown how behavior change can manifest differently over different time periods and between animals. Our exploratory approach suggests that subsequent formal models should model response behaviors as a mixture over different response types, such as a complete avoidance of deep-diving behavior, or a change in the timing of deep-diving behavior. Subsequent formal analyses of response behaviors might also attempt to account for similarities in responses between animals due to social behaviors.

One limitation of our approach is that it is necessarily exploratory because of the complex dependence among summary statistics, which precludes formal significance testing. This complex dependence makes it difficult to define formal hypothesis tests and use multiple testing corrections to evaluate potential Type-I errors in the results. Type-I errors are commonly known as "false positive" results. Formal hypothesis testing procedures are formal, in part, because the test's significance level controls the procedure's long-term Type-I error rate. Our method is exploratory, in part, because we do not explicitly control this error rate.

The quality and quantity of pre-exposure data affect the accuracy of our tail probabilities. In applying our approach, longer pre-exposure baseline datasets will more comprehensively capture the variability in behavior patterns. It is easier to separate response behaviors from baseline behaviors when baseline behaviors are better understood. For beaked whales, the satellite-linked telemetry devices we work with provide much longer pre-exposure data collection than other telemetry devices often used in CEEs, such as DTAGs (i.e., days–weeks vs. hours). Although the temporal resolution of the data is much coarser (i.e., 5-min intervals vs. sub-second sampling), the longer collection allows longer-term diving patterns to be better captured.

Another limitation is that the problem we address does not easily lend itself to validation as a formal testing method. While we are motivated by hypothesis testing, the procedure we propose is more similar to an exploratory outlier detection method. If formal testing were the goal, then the accuracy of a method to detect behavioral responses in telemetry data could be assessed by quantifying the method's Type-I and Type-II error rates (i.e., false positive and negative error rates). We provide a basic simulation study in the Additional file 1 to explore the Type-I and Type-II error rates for our exploratory procedure. We approximate the series of depth data using a two-state HMM to simulate alternating sequences of deep and shallow depths. We find that the method generally has reliable control of Type-I error rates, and good power to detect avoidance of deep-diving behavior.

Instead of formal validation, we find indirect evidence that our proposed method has reasonable performance. First, the tables of tail probabilities (Tables 4 and 5, and others in Additional file 1) relate to features that are visibly apparent in plots of diving behavior (Figs. 1 and 4, and others in Additional file 1). Second, the tables of tail probabilities (including Table 3, aggregating results across animals) identify patterns of behavior that suggest exposure to sound disrupts typical diving behavior. Such general findings are consistent with other foundational research conducted using high-resolution archival tags, which have shorter pre-exposure records and more sensors than the satellite tags used here [19, 41].

Conclusion

The exploratory approach we propose can be used to initially assess potential animal responses to external stimuli, such as exposure to anthropogenic sound sources. The approach is useful for analyzing data collected from animals for which standard models and methods have not been established due to a lack of a priori knowledge in the literature. The exploratory approach for response detection offered here suggests application to other

Hewitt *et al. Animal Biotelemetry*   (2022) 10:28

Page 14 of 15

types of response-to-exposure data collected over time because the summary statistics can be customized, and the statistical methods do not require strong distributional assumptions for their application because they are non-parametric. Furthermore, for other species where we have many more exposed records, we may be able to better suggest population-level response to exposure. In addition, we might be able to connect the intensity of exposure, as in [42], with the extent of response. In so doing, we could better understand the quantitative relationship between exposure and response [43].

Abbreviations
CEE: Controlled exposure experiment; FDR: False discovery rate; HMM: Hidden Markov model; KDE: Kernel density estimator/estimate; MFAS: Mid-Frequency Active Sonar.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s40317-022-00299-7.

> **Additional file 1** Additional figures and tables.

Author contributions
JH wrote and ran software to analyze the data, and contributed to the design and writing of this work. AG contributed to the design and writing of this work. NQ and WC contributed to the collection and interpretation of the data, conception of the work, and writing of this work. BS contributed to the collection and interpretation of data and writing of this work. SD contributed to the interpretation of data and writing of this work. RS contributed to the interpretation and analysis of data, and contributed to the design and writing of this work. All authors read and approved the final manuscript.

Availability of data and materials
The datasets generated and/or analyzed during the current study are available in the GitHub repository, https://github.com/jmhewitt/dive_eda_reproducibility.

Declarations

Ethics approval and consent to participate
The data analyzed here, and all research activities carried out, were collected as part of the Atlantic Behavioral Response study under NOAA/NMFS Scientific Research Permit No. 14809 issued to Douglas P. Nowacek, and NOAA General Authorization letter of confirmation No. 16185 issued to Andrew J. Read in accordance with the relevant guidelines and regulations on the ethical use of animals as experimental subjects. The research approach was approved by the Institutional Animal Use and Care Committees (IACUC) of Duke University.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Statistical Science, Duke University, Durham, NC, USA. [2]Nicholas School of the Environment, Duke University, Durham, NC, USA. [3]Southall Environmental Associates, Aptos, CA, USA. [4]Calvin University, Grand Rapids, MI, USA. [5]University of Plymouth, Plymouth, UK.

References
1. D'Amico A, Gisiner RC, Ketten DR, Hammock JA, Johnson C, Tyack PL, Mead J. Beaked whale strandings and naval exercises. Aquat Mamm. 2009;35(4):452–72.
2. Bernaldo de Quirós Y, Fernandez A, Baird RW, Brownell RL, Aguilar de Soto N, Allen D, Arbelo M, Arregui M, Costidis A, Fahlman A, Frantzis A, Gulland FMD, Iñíguez M, Johnson M, Komnenou A, Koopman H, Pabst DA, Roe WD, Sierra E, Tejedor M, Schorr G. Advances in research on the impacts of anti-submarine sonar on beaked whales. Proc R Soc Lond B Biol Sci. 2019;286(1895):20182533.
3. Southall BL, Nowacek DP, Miller PJO, Tyack PL. Experimental field studies to measure behavioral responses of cetaceans to sonar. Endanger Species Res. 2016;31:293–315.
4. Southall BL, Nowacek DP, Bowles AE, Senigaglia V, Bejder L, Tyack PL. Marine mammal noise exposure criteria: assessing the severity of marine mammal behavioral responses to human noise. Aquat Mamm. 2021;47(5):421–64.
5. Pirotta E, Booth CG, Costa DP, Fleishman E, Kraus SD, Lusseau D, Moretti D, New LF, Schick RS, Schwarz LK, Simmons SE, Thomas L, Tyack PL, Weise MJ, Wells RS, Harwood J. Understanding the population consequences of disturbance. Ecol Evol. 2018;8(19):9934–46.
6. Harris CM, Thomas L, Falcone EA, Hildebrand J, Houser D, Kvadsheim PH, Lam F-PA, Miller PJO, Moretti DJ, Read AJ, Slabbekoorn H, Southall BL, Tyack PL, Wartzok D, Janik VM. Marine mammals and sonar: dose-response studies, the risk-disturbance hypothesis and the role of exposure context. J Appl Ecol. 2018;55(1):396–404.
7. Tyack PL, Johnson M, Soto NA, Sturlese A, Madsen PT. Extreme diving of beaked whales. J Exp Biol. 2006;209(21):4238–53.
8. Madsen PT, Aguilar de Soto N, Tyack PL, Johnson M. Beaked whales. Curr Biol. 2014;24(16):728–30. https://doi.org/10.1016/j.cub.2014.06.041.
9. DeRuiter SL, Langrock R, Skirbutas T, Goldbogen JA, Calambokidis J, Friedlaender AS, Southall BL, et al. A multivariate mixed Hidden Markov model for blue whale behaviour and responses to sound exposure. Ann Appl Stat. 2017;11(1):362–92.
10. Langrock R, Marques TA, Baird RW, Thomas L. Modeling the diving behavior of whales: a latent-variable approach with feedback and semi-Markovian components. J Agric Biol Environ Stat. 2014;19(1):82–100.
11. Hewitt J, Schick RS, Gelfand AE. Continuous-Time Discrete-State modeling for deep whale dives. J Agric Biol Environ Stat. 2021;26(2):180–99.
12. Southall B, Baird R, Bowers M, Cioffi W, Harris C, Joseph J, Quick N, Margolina T, Nowacek D, Read A, et al.: Atlantic Behavioral Response Study (BRS)–2017 Annual progress report. Prepared for US Fleet Forces Command. Submitted to Naval Facilities Engineering Command Atlantic, Norfolk, Virginia, under Contract No N62470-15-D-8006, Task Order 50, issued to HDR Inc., Virginia Beach, Virginia (2018)
13. Johnson MP, Tyack PL. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. IEEE J Ocean Eng. 2003;28(1):3–12.

Hewitt *et al. Animal Biotelemetry*   (2022) 10:28

Page 15 of 15

14. Mikkelsen L, Johnson M, Wisniewska DM, van Neer A, Siebert U, Madsen PT, Teilmann J. Long-term sound and movement recording tags to study natural behavior and reaction to ship noise of seals. Ecol Evol. 2019;27:140.

15. Ellison WT, Southall BL, Clark CW, Frankel AS. A new context-based approach to assess marine mammal behavioral responses to anthropogenic sounds. Conserv Biol. 2012;26(1):21–8.

16. Sivle LD, Kvadsheim PH, Curé C, Isojunno S, Wensveen PJ, Lam F-PA, Visser F, Kleivane L, Tyack PL, Harris CM, Miller PJO. Severity of Expert-Identified behavioural responses of humpback whale, minke whale, and northern bottlenose whale to naval sonar. Aquat Mamm. 2015;41(4):469–502.

17. Southall BL, DeRuiter SL, Friedlaender A, Stimpert AK, Goldbogen JA, Hazen E, Casey C, Fregosi S, Cade DE, Allen AN, Harris CM, Schorr G, Moretti D, Guan S, Calambokidis J. Behavioral responses of individual blue whales (*Balaenoptera musculus*) to mid-frequency military sonar. J Exp Biol. 2019. https://doi.org/10.1242/jeb.190637.

18. Falcone EA, Schorr GS, Watwood SL, DeRuiter SL, Zerbini AN, Andrews RD, Morrissey RP, Moretti DJ. Diving behaviour of Cuvier's beaked whales exposed to two types of military sonar. R Soc Open Sci. 2017;4(8):170629.

19. DeRuiter SL, Southall BL, Calambokidis J, Zimmer WM, Sadykova D, Falcone EA, Friedlaender AS, Joseph JE, Moretti D, Schorr GS, et al. First direct measurements of behavioural responses by Cuvier's beaked whales to mid-frequency active sonar. Biol Lett. 2013;9(4):20130223.

20. Antunes R, Kvadsheim PH, Lam FPA, Tyack PL, Thomas L, Wensveen PJ, Miller PJO. High thresholds for avoidance of sonar by free-ranging long-finned pilot whales (*Globicephala melas*). Mar Pollut Bull. 2014;83(1):165–80.

21. Quick NJ, Isojunno S, Sadykova D, Bowers M, Nowacek DP, Read AJ. Hidden Markov models reveal complexity in the diving behaviour of short-finned pilot whales. Sci Rep. 2017;7(1):1–12.

22. Carpenter SR, Frost TM, Heisey D, Kratz TK. Randomized intervention analysis and the interpretation of whole-ecosystem experiments. Ecology. 1989;70(4):1142–52.

23. Southall B, Baird R, Bowers M, Cioffi W, Harris C, Joseph J, Quick N, Margolina T, Nowacek D, Read A, et al.: Atlantic Behavioral Response Study (BRS)–2019 Annual progress report. Prepared for US Fleet Forces Command. Submitted to Naval Facilities Engineering Command Atlantic, Norfolk, Virginia, under Contract No N62470-15-D-8006, Task Order 50, issued to HDR Inc., Virginia Beach, Virginia (2020)

24. Andrews RD, Pitman RL, Ballance LT. Satellite tracking reveals distinct movement patterns for type b and type c killer whales in the southern ross sea, antarctica. Polar Biol. 2008;31(12):1461–8.

25. Southall BL, Moretti D, Abraham B, Calambokidis J, DeRuiter SL, Tyack PL. Marine mammal behavioral response studies in southern California: advances in technology and experimental methods. Mar Technol Soc J. 2012;46(4):48–59.

26. Crowley PH. Resampling methods for computation-intensive data analysis in ecology and evolution. Annu Rev Ecol Syst. 1992;23(1):405–47.

27. Manly BFJ, Navarro Alberto JA. Randomization bootstrap and Monte Carlo methods in Biology. Boca Raton: Taylor & Francis Group; 2020.

28. Silverman B. Density estimation for statistics and data analysis. London: Chapman and Hall; 1986.

29. Shearer JM, Quick NJ, Cioffi WR, Baird RW, Webster DL, Foley HJ, Swaim ZT, Waples DM, Bell JT, Read AJ. Diving behaviour of Cuvier's beaked whales (*Ziphius cavirostris*) off Cape Hatteras, North Carolina. R Soc Open Sci. 2019;6(2): 181728.

30. Baird RW, Webster DL, McSweeney DJ, Ligon AD, Schorr GS, Barlow J. Diving behaviour of Cuvier's (*Ziphius cavirostris*) and Blainville's (*Mesoplodon densirostris*) beaked whales in Hawai'i. Can J Zoo. 2006;84(8):1120–8. https://doi.org/10.1139/z06-095.

31. Baird RW, Webster DL, Schorr GS, McSweeney DJ, Barlow J. Diel variation in beaked whale diving behavior. Marine Mamm Sci. 2008;24(3):630–42. https://doi.org/10.1111/j.1748-7692.2008.00211.x.

32. Baird RW, Webster DL, Schorr GS, McSweeney DJ, Barlow J. Diel variation in beaked whale diving behavior. Marine Mamm Sci. 2008;24(3):630–42.

33. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57(1):289–300.

34. Frid A, Dill L. Human-caused disturbance stimuli as a form of predation risk. Conserv Ecol. 2002. https://doi.org/10.5751/ES-00404-060111.

35. Miller PJO, Isojunno S, Siegal E, Lam F-PA, Kvadsheim PH, Curé C. Behavioral responses to predatory sounds predict sensitivity of cetaceans to anthropogenic noise within a soundscape of fear. Proc Natl Acad Sci. 2022;119(13):2114932119. https://doi.org/10.1073/pnas.2114932119.

36. Duong T. KS: Kernel density estimation and kernel discriminant analysis for multivariate data in R. J Stat Softw. 2007;21(7):1–16.

37. Wand MP, Jones MC. Comparison of smoothing parameterizations in bivariate kernel density estimation. J Am Stat Assoc. 1993;88(422):520–8.

38. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc. 2002;97(458):611–31.

39. Fraley C, Raftery A. Model-based methods of classification: Using the mclust software in chemometrics. J Stat Softw. 2007;18(6):1–13.

40. AguilardeSoto N, Madsen PT, Tyack P, Arranz P, Marrero J, Fais A, Revelli E, Johnson M. No shallow talk: cryptic strategy in the vocal communication of Blainville's beaked whales. Marine Mamm Sci. 2012;28(2):75–92.

41. Quick N, Scott-Hayward L, Sadykova D, Nowacek D, Read A. Effects of a scientific echo sounder on the behavior of short-finned pilot whales (*Globicephala macrorhynchus*). Can J Fish Aquat Sci. 2017;74(5):716–26. https://doi.org/10.1139/cjfas-2016-0293.

42. Schick RS, Bowers M, DeRuiter S, Friedlaender A, Joseph J, Margolina T, Nowacek DP, Southall BL. Accounting for positional uncertainty when modeling received levels for tagged cetaceans exposed to sonar. Aquat Mamm. 2019;45(6):675–90.

43. Tyack PL, Thomas L. Using dose-response functions to improve calculations of the impact of anthropogenic noise. Aquat Conserv. 2019;29(S1):242–53.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.