esa

ECOSPHERE

Sperm whale response to tag boat presence: biologically informed hidden state models quantify lost feeding opportunities

SAANA ISOJUNNO† AND PATRICK J. O. MILLER

School of Biology, University of St Andrews, Bute Building, St Andrews, Fife KY16 9TS United Kingdom

Citation: Isojunno, S., and P. J. O. Miller. 2015. Sperm whale response to tag boat presence: biologically informed hidden state models quantify lost feeding opportunities. Ecosphere 6(1):6. http://dx.doi.org/10.1890/ES14-00130.1

Abstract. Animal-attached sensors provide invaluable data to describe behavior of cryptic species, such as cetaceans, and are increasingly used to assess anthropogenic disturbance effects. Tag deployment and handling may itself alter the behavior of study animals and there is a need to assess if and when behavior recovers to an undisturbed level. Not all behavioral changes have fitness consequences, and our goal is to derive metrics that can be linked to fitness implications, such as time and energy allocation to different functional behaviors. Here we detail an approach that incorporates biological knowledge and multiple streams of tag-recorded data in a hidden state-switching model to estimate time series of functional behavioral st ates for 12 sperm whales off Norway. Foraging, recovery and resting states were specified in the hidden state model by state-dependent likelihood structures. Comparison of hidden state models revealed a parsimonious set of input time series, and supported the inclusion of a less informed 'silent active' state. There was a high agreement between state estimates and expert classifications. We then used the estimated states in time series models to test three hypotheses for behavioral change during suction-cup tag deployment procedures: change in behavioral states, change in prey capture attempts and locomotion cost, given behavioral state. Sperm whales spent 34% less time at the sea surface and 60% more time in non-foraging silent active state in the presence of the tag boat ("tagging period" 0.1–2.8 h) than during post-tagging baseline period (1.8–20.8 h). No comparable pre-tagging baseline data were available. Nevertheless, time-decaying models of tagging effects were not retained in model selection, indicating a short-term effect that ceased immediately after the tagging period. We did not find changes in energetic proxies, given behavioral state, however changes in functional state budget indicate costs in terms of lost feeding opportunities and recovery time at surface. These results are useful to quantitatively identify data periods that should not be considered baseline behavior within tag recordings. This functional state approach proves effective to quantify disturbance in terms of time and energy allocation that is based upon general principles that can be applied to other species and biologging applications.

Key words: Bayesian; DTAG; functional state; northern Norway; *Physeter microcephalus*; research effects, state-dependent likelihood; state-switching model; suction-cup tag attachment; time-series model.

Received 24 April 2014; revised 30 July 2014; accepted 5 August 2014; final version received 10 December 2014; **published** 21 January 2015. Corresponding Editor: D. P. C. Peters.

† **E-mail:** si66@st-andrews.ac.uk

Introduction

Animal-attached sensors have become an important means to monitor individual behavior for a wide range of species and habitats in the wild. With technological advances in miniaturization, resolution and longevity of bio-logging sensors and transmitters, there is scope for a more integrated understanding of how individual behavior and physiology interact with their environment and anthropogenic stressors (Cooke et al. 2004, Johnson et al. 2009). As such, biologging science can provide first clues of individual-level mechanisms that could drive anthropogenic impacts on populations (Cooke et al. 2004, Tyack 2009, Berger-Tal et al. 2011, Miller et al. 2012). For population-level inferences to be reliably made, it is important to consider how representative the tagged individuals' baseline behavior (such as time spent foraging) or response to stimuli (such as probability of avoidance) are of the wild, non-tagged population of conservation interest. Evaluation of possible effects of bio-logging experimental procedures is therefore important when considering how representative tag data might be to the entire population ("measurement affects performance"; Wilson et al. 1986, Miller et al. 2009).

Research effects of biologging studies comprise both the effects elicited by the tag deployment procedures, such as approach, physical contact or capture (hereafter collectively termed as 'handling') and the presence of the device itself upon the animal. Documented tagging and marking effects range from injury, physiological stress and behavioral changes to reproductive success and survival rates (Murray and Fuller 2000, Godfrey and Bryant 2003, Barron et al. 2010, Walker et al. 2012). The relative importance of handling and device effects depends upon their relative invasiveness, duration and repetition that may allow for habituation or sensitization. The effects of tag presence are of particular concern for flying and swimming species that may be more sensitive to alterations to their streamlining, such as tag-induced drag (Bannasch et al. 1994, Barron et al. 2010, Hazekamp et al. 2010), and subsequent increases in transport costs (Wilson et al. 1986, Ropert-Coudert et al. 2000, Wilson and McMahon 2006, Fossette et al. 2007). These effects are reduced by use of relatively smaller and more aero- and hydrodynamic tag shapes (e.g., Bannasch et al. 1994). Locomotion costs can also be expected to increase if the tag significantly increases the mechanical loading (weight), buoyancy or center of gravity of an individual (Wilson et al. 1986). Tag attachment method (e.g., harness vs. glue) may also impair movement (Barron et al. 2010), but also have more subtle physiological effects, such as changes in the distribution of animal surface temperature (McCafferty et al. 2007).

In marine mammals, most studies have reported short-term behavioral effects of tagging with little evidence of impacts on survival (McMahon et al. 2008, Walker et al. 2012). While extensive research on tagging effects have helped to guide deployment practices and tag development (e.g., Fossette et al. 2007), generalizing the device-specific and mostly qualitative results to different species and constantly evolving telemetry set ups is challenging (Murray and Fuller 2000). Not only are tagging effects likely to depend upon specific handling procedures and tag design but also individual (age, sex, condition) and behavioral and environmental context (e.g., nursing, prey availability) (Murray and Fuller 2000, Walker et al. 2012). Reliable estimation of tagging effects therefore requires case-by-case assessment. However, with limited availability and cost of alternative study platforms, tagging studies are rarely able to empirically cross-validate tag data with data from a 'pre-tagging' period or data from non-tagged individuals (Murray and Fuller 2000, Godfrey and Bryant 2003, Walker et al. 2012). Most studies therefore assume that tagging has negligible or no influence on parameters of interest after some cut-off recovery time since handling ('baseline' period; Murray and Fuller 2000, Godfrey and Bryant 2003, but see definition of baseline period based upon affected dive parameters in Miller et al. 2009).

An alternative and quantitative approach is to compare tagged individual behavior between different available 'doses' of tagging procedures, such as varying tag size (Wilson et al. 1986) or handling intensity (Engelhard et al. 2002). Such an approach could be used to back-calculate true population parameters (Wilson et al. 1986, Wilson and McMahon 2006). For example, Ropert-Coudert et al. (2007) compared diving and movement behavior of Adelie penguins

between two different tag sizes to extrapolate effects on penguins with tags of negligible size. Based upon their results on tagged individuals, the authors were able to predict that non-tagged penguins would maintain similar energy expenditure than tagged animals but be able to swim faster, dive deeper, and range farther in pursuit of prey. Similarly, data can be compared within each tag record under the assumption that handling effects are strongest at the time of attachment and decrease afterwards. For example, Miller et al. (2009) found that the first dive after tagging was shorter than subsequent dives of sperm whales. Such a 'during-after' comparison can reduce confounding individual variability, but does assume that tag records are long enough to allow at least partial recovery.

In this paper, we develop and apply a novel approach to quantitatively assess the effects of suction-cup tag deployment procedures ('handling') on sperm whales for which no pre-tagging control was available. Our goal was to compare whale behavior in the presence vs. absence of the tag boat, and to evaluate different models of recovery from effects due to tag attachment and tag boat presence. We evaluate three classes of possible behavioral effects: (1) change in behavioral time-budget, (2) reduction in prey capture attempts (proxy for foraging success), given behavioral state and (3) increase in movement cost, given behavioral state.

To obtain behavioral states for hypothesis testing, we used multiple streams of tag data in a hidden state-switching model to estimate biologically informed states and their uncertainty. As well as classification of sperm whale behavior, our goal was to formulate 'functional' states that could be generalized to other species and used to assess a range of disturbance stimuli. Conceptually, our analytical approach follows the movement ecology paradigm (Nathan et al. 2008) and functional state framework (Isojunno and Miller 2014). Functional state decomposes behavioral time series into behavioral states as units of 'effort' that are associated with a goal or set of goals, combining both the ultimate and proximate drivers of behavior (Nathan et al. 2008, Isojunno and Miller 2014). For example, these goals could be mating, information, breathing or shelter. The achievement of these goals can be measured using currencies (e.g., prey capture for feeding goal) or proxy indicators of the currency (e.g., terminal echolocation as an indicator of foraging success) and expressed as success or cost rate within each functional state (Isojunno and Miller 2014). The states capture mean differences across different behavioral states of the currencies of interest.

We used adult male sperm whales in a subarctic foraging ground in Northern Norway as a relatively simple model system where individuals spend most of their time solitary and feeding (Teloni et al. 2008, Oliveira and Wahlberg 2013). Sperm whales perform deep (200–1000 m) and long (30–60 min) echolocation-based foraging dives (Watwood et al. 2006), facing trade-offs between time spent foraging at depth and recovering oxygen stores at the sea surface (Boyd 1997). These trade-offs formed the conceptual basis for our functional state model for sperm whales. We considered two bio-energetic currencies, foraging success and movement cost, that vary across the foraging dive cycle (surfacing, descending transit, layer-restricted search, ascending transit). Terminal echolocation buzzes (Miller et al. 2004) and dynamic body acceleration (ODBA; Halsey et al. 2009) were quantified as proxies for prey capture attempts (~foraging success) and locomotion cost, respectively. Besides foraging dive cycles, we also expected sperm whales to spend time in shallower dives for other purposes, such as resting or 'silent active' swimming. Sperm whale resting dives occur in consecutive bouts of variable duration, are typically shallower than foraging dives, and are stereotypically characterized by a vertical 'head-up' or 'head-down' posture (Miller et al. 2008). Non-foraging but active behaviors are also described for sperm whales (Miller et al. 2008), and likely reflect social or anti-predatory functions (Curé et al. 2013). We were able to test how many non-foraging functional states are utilized by sperm whales by comparing models with five (foraging states + resting) versus six states (+ active non-foraging) (Fig. 1).

Methods

We first estimate time series of functional states and then use the resulting state classification to test for behavioral disturbances likely linked to individual fitness. Behavioral states were esti-

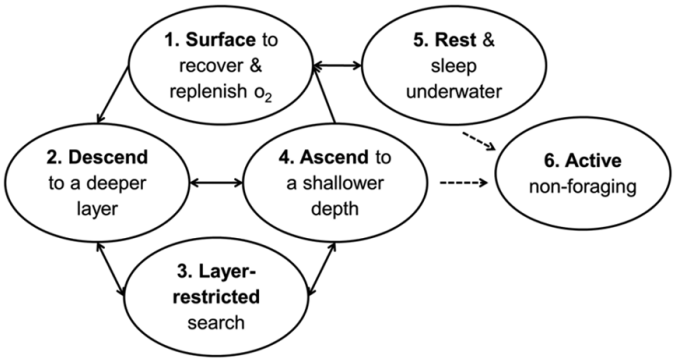

Fig. 1. We specified five or six functional states for sperm whales in their foraging ground: (1) surfacing, oxygen replenishment and physiological recovery at the surface; (2) descending transit, transiting to a deeper prey layer; (3) layer restricted search (LRS), searching at a prey layer; (4) ascending transit, transiting to a shallower depth or the surface; (5) resting and sleep underwater and (6) active non-foraging, which could encompass multiple functions. States 1–4 are considered to be functional states for foraging. Solid arrows show transitions that were expected to be likely and dashed arrows highlight the uncertainty related to the transition probability to and from state 6. These expectations and uncertainties were incorporated in the model as respective informative and uniform priors for the transition probabilities (Appendix A: Fig. A1).

mated in a hidden state model in order to formalize our prior expectations of functional behavior (surfacing, transiting, layer-restricted search, resting, and other 'silent active') and utilize multiple input data time-series. The state estimates and uncertainty were next used as data in a second analysis step that tested for time or energetic costs of tag deployment procedures with different models of recovery from disturbed to post-tagging baseline behavior.

Data

Data were collected for 12 individual sperm whales tagged with an audio and movement-recording bio-logging device (Dtag; Johnson et al. 2009). Four whales were tagged in 2005 (Teloni et al. 2008) and eight whales were tagged in 2008–2010 (Miller et al. 2012) near Lofoten Islands in Northern Norway. Sperm whales were localized at sea visually and acoustically by monitoring their echolocation clicks with a towed hydrophone array. The protocol included initial observations at 200–1000 m from a main observation vessel (MS Stronstad, 29 m). A smaller tag boat (rigid-hulled inflatable boat or similar) was launched to approach each whale and deploy tags with a pole that varied in length each year of research (Table 1).

Tag data were processed to calculate depth as well as whale-frame acceleration and magnetometer data which was converted to pitch, roll and heading time-series (Miller et al. 2012). Time-series data from the tag was down-sampled to one sample per minute to reduce computational time and concentrate analysis efforts on dive phase scale rather than fine-scale behavior, such as thrusting strokes. Depth was sampled at the start of each 1-min interval, while mean pitch and 'overall dynamic body acceleration' (ODBA) were calculated over the entire 1-min interval. ODBA was calculated as the sum over each minute of the two-norm of high-pass filtered acceleration (finite impulse response filter, cut-on frequency 0.05 Hz). To account for effects of tag position on ODBA, ODBA values for each whale were divided (normalized) by its median value and then multiplied by the median ODBA across whales. Surface periods were detected using a depth threshold of 2 m for accepting a dive, and a threshold of 1 m for reaching the surface. Time (min) since the last surface period was calculated for the start of each 1-min interval (minFromSurf).

Audio data (stereo at 96 kHz) were monitored aurally and visually using spectrograms for echolocation click trains (regular and buzz clicks) and marked for their start time and duration in each record. The presence or absence of these

Table 1. Summary of tag records.

Tag id, pole length (m)	Sample duration (h)			Time in pre-detected behavioral states (%)					
	Total	Tagging analysis	Tag-boat	Surface period	Bottom phase	Dives 1–4	Dives 5–7	Dives 8–9	Dives 10–11
sw05_196a, 15	21.32	21.32	0.50	29.7	52.0	91.6	5.8	0.0	2.7
sw05_199a, 15	18.07	0.00	0.00	18.6	57.8	100.0	0.0	0.0	0.0
sw05_199b, 15	13.82	0.00	0.00	22.9	46.1	82.9	6.6	10.5	0.0
sw05_199c, 15	13.38	0.00	0.00	24.3	18.1	55.9	6.7	34.9	2.4
sw08_152a, 5	8.65	4.60	2.83	16.2	no data	70.0	24.2	0.0	5.8
sw09_141a, 9	15.28	3.83	0.82	20.7	no data	42.4	28.4	7.6	21.6
sw09_142a, 9	14.77	2.98	0.23	21.0	no data	59.8	13.9	13.1	13.2
sw09_153a, 9	8.53	8.53	0.12	17.6	61.7	100.0	0.0	0.0	0.0
sw09_160a, 9	14.78	3.47	0.22	17.4	no data	94.7	2.9	0.0	2.4
sw10_147a, 12	15.77	15.77	0.93	30.7	27.7	71.8	3.6	24.0	0.6
sw10_149a, 12	16.13	14.15	1.80	21.9	51.8	95.9	0.0	0.0	4.1
sw10_150a, 12	14.87	12.97	0.78	25.9	30.5	93.2	4.5	0.0	2.3
Total	175.37	87.62	8.23	266.8	345.6	958.3	96.6	90.1	55.0

Notes: Total sample duration (h) refers to data that was used to fit hidden state models, while tagging analysis show durations of data retained for tagging and post-tagging datasets (see text). Tag boat shows the total number of hours that the boat remained near the whale after tag deployment. Expert classified dives are given as dives with clicking and usual dive profile (1–4), dives with clicking and unusual dive profile (5–7), dives without clicking and drifting behavior (8–9) and dives without clicking and silent active swimming (10–11) (Appendix C: Table C4).

aurally monitored clicks in each 1-min interval was used in the hidden state models in conjunction with the depth and accelerometer data. Other types of clicks (slow clicks, codas) were not included in the analysis.

Six whales were subject in a controlled exposure experiment that included up to five 20–30 min exposure sessions (Miller et al. 2012, Curé et al. 2013). Two of these six whales were exposed to just two sessions, followed by a secondary suction-cup tag deployment 1.2 h after all experiments ended. All data from all 12 individual sperm whales were used to parameterize the hidden state model, but non-tagging baseline periods excluded all exposure sessions and post-exposure periods. For tagging effects analysis, tag handling periods were defined as the time period between tag deployment and recovery of tag boat to the main research vessel or movement of the tag boat (>1 km) from the tagged whale.

A calibration data set of behavioral states was used to compare with the hidden model state estimates. "Bottom phases" were defined by the period between the first positive and the last negative pitch in a dive for nine whales (Miller et al. 2004; Table 1). Dive types were classified by consensus of three experts, including the authors and Dr. Stacy DeRuiter. The resulting consensus comprised 11 dive types (Appendix C: Table C4).

Hidden state model

Our state-switching model for sperm whale behavior consisted of four functional foraging states and either one or two additional states for non-foraging related behavior (Fig. 1). Alternative model structures were considered to assess how many states (five or six; Fig. 1) and which combinations of input data (depth, clicking, minFromSurf, ODBA and/or pitch) should be included to classify the behavioral time series most effectively. Each model consisted of a five-by-five or six-by-six state transition probability matrix and state-dependent likelihoods for the input data.

Depth was modeled as a random walk Gaussian variable with a state-specific mean and variance (Langrock et al. 2013, Photopoulou 2013):

$$d_t \sim N(d_{t-1} + \pi_{s_{t-1}}, \sigma^2_{s_{t-1}}) \quad (1)$$

where d_t denotes depth at time step t and s denotes the hidden state at time step $t - 1$. Descent and ascent states were modeled as a directional random walk ('bias' parameter π_s estimated $\neq 0$), and all other states a non-directional random walk ($\pi_s = 0$). A separate variance for depth changes ($\sigma^2{}_s$) was estimated for each state. A step function was used to constrain predicted depths to be >0 m.

To relax the Markov assumption that state transitions depend only upon the previous time

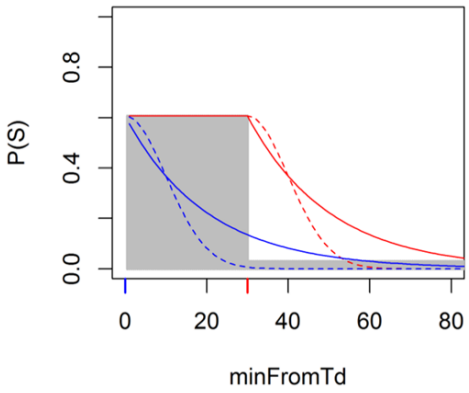

Fig. 2. Illustration of the log-linear model probability of state transition $\log(P(s)) = \alpha + \beta \times x$ with five different hypotheses for tagging dose. Blue and red tick marks on x-axis show Tagging period, with start of tagging data in blue and end of Tagging in red. The first hypothesis for dose was a presence/absence effect of tag boat, Tagging, shown as shaded gray. Four time-decaying explanatory variables were tested for hypotheses of recovery from either tag deployment (blue; minFromTd and minFromTd^2) or end of Tagging period (red; minFromTagging and minFromTagging^2). The variables were calculated as linear or squared time since tag deployment or Tagging, representing either exponential (dashed lines $f(x)$; minFromTd and minFromTagging) or exponential with delayed (dashed lines $f(x^2)$; minFromTd^2 and minFromTagging^2) speed of recovery. In this illustration example, the intercept α was set at -0.5 and coefficient β at -0.005.

step, all models allowed the probability of surfacing at time t to increase with decreasing depth at time $t - 1$ in a multinomial logistic regression (see Langrock et al. 2013 for a similar formulation of feed-back in transition probabilities). minFromSurf (x_1) was an additional covariate in the regression for the probability of transition to LRS (state 3). The linear predictor for the probability of state s at time t was therefore:

$$f(P(s_t)) = \beta_{0,s_{t-1},s_t} + \beta_{1,s_t} d_{t-1} + \beta_{2,s_t} x_{1,t} \quad (2)$$

where intercept β_0 was specific to a state-transition, coefficient β_1 was associated with transitions to surface (state 1), and β_{2,S_t} associated with staying in LRS (state 3). The coefficients were fixed at zero for other transitions, i.e., when $s_t \neq 1$, then β_1 was set to zero, and when $s_t \neq 3$ and $s_{t-1} \neq 3$, then β_2 was set to zero.

The presence/ absence of clicking (c) was estimated a state-specific probability ($c_t \sim$ Bernoulli($\gamma_{s[t]}$)). ODBA (o) was similarly modeled as a Gamma distributed variable with state-dependent shape and rate parameters ($o_t \sim$ Gamma($\varphi_{s[t]}$, $\omega_{s[t]}$)).

The absolute value of the pitch angle p was modeled in a logistic Beta regression (Ferrari and Cribari-Neto 2004) so that within mobile states (i.e., not surfacing or resting), pitch was related to vertical step length in a linear predictor:

$$g(p_t) = \alpha_{0,s_t} + \alpha_{1,s_t} |d_{t-1} - d_t|. \quad (3)$$

Here, the coefficient for vertical step $\alpha_{1,s[t]}$ was specific to each state so that all mobile states were estimated a single coefficient which was fixed at zero for surface and resting. Pitch during surfacing and resting were estimated state-dependent means ($\alpha_{0,1}$, $\alpha_{0,5}$), while mobile states were assigned a common intercept.

The joint likelihood for the full model (all five data streams) was the product of their conditionally independent likelihoods (for a similar formulation, see McClintock et al. 2013):

$$l(\pi_s, \sigma^2, \beta, \gamma, \varphi, \omega, \alpha, \tau, s|\theta) = \prod_{t=1}^{T} l(\pi_t|\theta, s_t) l(\sigma_t^2|\theta, s_t) l(\beta_t|\theta, s_t) l(\gamma_t|\theta, s_t) l(\varphi_t|\theta, s_t) l(\omega_t|\theta, s_t) l(\alpha_t|\theta, s_t) l(\tau_t|\theta, s_t) l(s_t|\theta, s_{t-1}) \quad (4)$$

where θ denotes the included set of state-dependent parameters. The full model (all four model components) had 54 estimable parameters in addition to the hidden states that were estimated for each data point.

After initial inspection of model performance, one additional parameter was introduced for the seven best DIC model structures. In the forementioned models, we had assumed a time-constant average step length within each state by estimating a state-specific σ^2_s. Inspection of the data revealed that step lengths increased as a function of the depth during foraging dives (dives consisting of only descent, LRS, and ascent). The observed relationship appeared to be linear when depth was square root transformed (see Fig. 4, middle panel). We therefore specified a time-varying σ^2_s for LRS state and time-varying drift for descent and ascent states by setting:

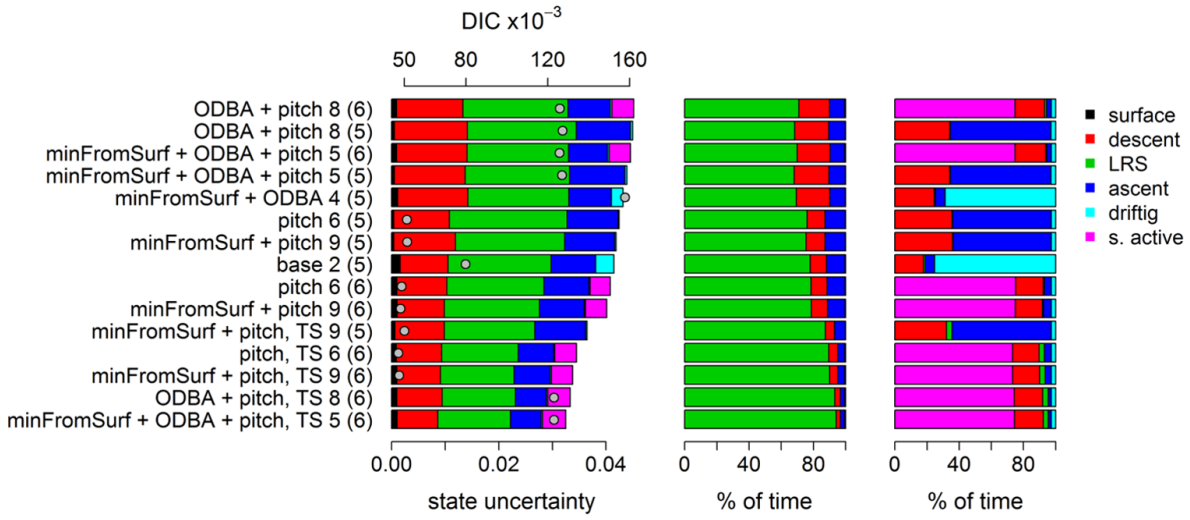

Fig. 3. Hidden state model selection. Model structure numbers are given before number of states in brackets. 'Base' structure here refers to depth + clicking that were included in all of the converged set of models; 'TS' refers to time-varying step length models. Left panel: overall state uncertainty for each model (total bar width) with contributing states color-coded. Overall state uncertainty was calculated for each model as the total proportion of posterior samples that were not the most prevalent state. Gray circles show DIC (from Table 2). Middle panel: percentage of time estimated in each state during pre-classified bottom phases. Contributing states are color-coded so that green shows sensitivity of layer-restricted search to pre-classified bottom phases. Right panel: percentage of time estimated in each state during expert classified silent active dives. Models in all panels are shown in ascending order for overall state uncertainty.

$$\begin{aligned}\sigma^2_{3,t} &= \overline{\sigma}^2_3 + \mu\sqrt{d_{t-1}} \\ \pi^2_{2,t} &= \overline{\pi}_2 + \mu\sqrt{d_{t-1}} \\ \pi^2_{4,t} &= \overline{\pi}_4 + \mu\sqrt{d_{t-1}}.\end{aligned} \tag{5}$$

Here $\overline{\sigma}^2_s$ and $\overline{\pi}_s$ are the time-constant intercepts for variance and drift for the random walk, $\sigma^2{}_{s,t}$ and $\pi_{s,t}$ are the respective time-varying parameters, and μ the increase in step length for every square root unit increase in depth. We therefore specified that the relationship between step length and depth was constant across the three foraging states (descent, LRS, ascent). For an exhaustive list of model parameters, see Appendix A.

In order to incorporate prior information on whale behavior, a Bayesian approach was taken to parameterize the models. A Gibbs sampling algorithm was used to sample from the joint posterior distribution of the model. We used freely available jags software (2003) within r (coda package, Plummer 2003 and R2jags package Su and Yajima 2012). Descent and ascent rate were specified with informative priors using Gamma distribution with a mean and variance parameter from literature (Watwood et al. 2006). A lower mean and variance for ODBA was used to construct a Gamma prior for resting. Probability of clicking was also informed, with a higher mean for foraging states (descent, LRS, ascent). Pitch regression coefficients had uninformative priors with no parameter difference between states except that the coefficient for vertical step was fixed at zero for surface and resting as explained above. Uniform (uninformative) priors were specified for most transition probabilities (state-specific intercepts). Coefficients for the probability of transition to surface and LRS were assigned uninformative normal priors. The probability of transition to surface was constrained to be negative by truncating its prior distribution. See Appendix A for a comprehensive list (illustration in Appendix D), and example model scripts and data in the Supplement.

All models were sampled in three independent chains, each with an initial 16,000 iterations. Model convergence was assessed at this point,

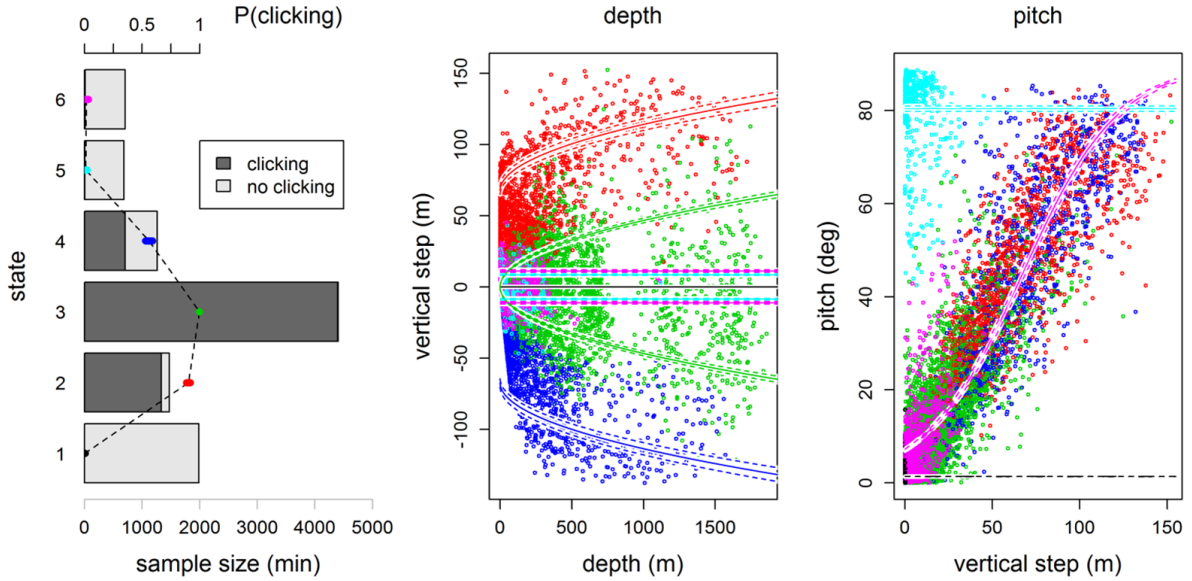

Fig. 4. Characteristics of the selected functional state model. Left panel: sample size and posterior 95% quantile for probability of clicking by functional state (Fig. 1). The total numbers of states with and without clicking are given on the bottom gray x-axis, and the posterior estimate for the probability of clicking on the top black x-axis. Middle panel: vertical steps (m/min) predicted as a function of depth (m). Posterior mean steps as a function of depth were predicted based on the posterior mean (solid lines) and 95% quantile (dashed lines) for the random walk parameters σ^2_s, μ, π_2 and π_4 (Appendix D: Table D1). Predictions for each state are color-coded; vertical step predictions for descents (red) and ascents (blue) include drift (bias, π) and are slightly asymmetric around zero because descent and ascent drift were estimated separately as π_2 and π_4 in the model. Vertical step predictions for states 1 (surfacing, in black), 3 (LRS, in green), 5 (resting, in indigo) and 6 (active-silent, in pink) did not include drift (i.e., not signed) but for illustration, are overlaid here symmetrically with observations both above and below zero. Right panel: Absolute value of pitch (deg) predicted as a function of vertical step length (m/min) (right), each overlaid with observed data. Pitch values were predicted based on the posterior mean (solid lines) and 95% quantile (dashed lines) values the pitch regression intercept $\alpha_{0,s}$ and coefficient for depth $\alpha_{1,s}$ (Appendix D: Table D1).

and a subset of models that were deemed to reach convergence in terms of state classification were updated a further 20,000 times. Initial values were set manually for all state parameters (Appendix A: Table A1). Brooks-Gelman-Rubin diagnostic (BGR; Brooks and Gelman 1998, Gelman et al. 2003) was used to assess model convergence, which was rejected based on its poorest converging parameter (BGR estimate $\leq$ 1.05). Detailed methods and results for model convergence can be found in Appendix B.

Four criteria were used to rank models that were deemed to have converged: (1) goodness of fit relative to model complexity (deviance information criterion DIC), (2) uncertainty in state classification, and (3) sensitivity and specificity to pre-classified bottom phases and (4) comparison to pre-classified resting and silent-active dives. Detailed methods for model selection can be found in Appendix C.

Use of state classifications for assessment of tagging effects

We used the top-ranked time series estimate of hidden states as data, and their uncertainty as weights, in a second analysis step that tested the effects of tagging on three response variables: (1) estimated activity state ($\sim$ multinomial, proxy for functional state), (2) presence/absence of buzzing ($\sim$ Bernoulli, proxy for foraging success), and (3) overall dynamic body acceleration (ODBA $\sim$ Gamma, proxy for locomotion cost).

Probability of state, given previous state, was modeled by including previous state (prevState) as factor baseline covariate. State was used as a factor baseline covariate in models for ODBA and buzz. We also allowed for mean differences in all three response variables across individuals by including tag id (whale) of the record as a factor covariate. The binomial model for buzz was fitted to a subset of data that only included foraging states (descent, LRS and ascent). No buzzing was observed in the non-foraging states (surfacing, drifting or silent active), so estimating standard errors for their coefficients would have not been possible.

Candidate exposure covariates were assessed for inclusion using model selection, and were designed to test between different hypotheses of the time-course of possible behavioral responses to tag deployment procedures (Fig. 2). Presence/absence of tag boat was included either as a main effect (Tagging), or interaction with year (Tagging : year) or pole length (Tagging : poleL) to assess any differences in level of response across years or as a function of pole length, respectively.

We chose a maximum likelihood framework for fitting these models for ease of model selection using AIC. Multinomial log-linear regression models were fit using function multinom in r library nnet, while binomial (logit link) and Gaussian (identity link) regression models were fit using function glm in r library stats. Multinomial models were weighted by the posterior probability of the state estimate, thus accounting for the uncertainty in state estimation. AIC unit difference of $\Delta AIC < -2$ was considered support for candidate tagging covariates compared to the baseline models for each response variable (state $\sim$ prevState + whale, ODBA $\sim$ state + whale, and buzz $\sim$ state + whale). All tagging effects models included the baseline covariates and up to two tagging-related explanatory variables. To avoid spurious relationships, only one of the four time-decay covariates (minFromTd, minFromTd^2, minFromTagging, or minFromTagging^2) were included in any one model, and were not included in the same model with Tagging:Year.

The lowest AIC models were diagnosed for influential individuals and data, goodness of fit, distributional assumptions, and serial correlation in residuals (Appendix D). Models that were diagnosed with serial correlation of residuals were re-fit within a generalized estimating equation (GEE) in SAS 9.3 (procedure 'genmod'). In multinomial models, the state that appeared to change most in response to tagging was used as a binomial response variable in the GEE. Any tagging effects were re-assessed using the empirical standard error estimates that do not assume any particular working correlation within the GEE, but account for the smaller effective sample size of correlated data within clusters. Small empirical standard errors (estimates $> 2 \times$ SE) and significant type 3-tests ($p < 0.05$) were considered as support for candidate covariates. GEE models included whale as a cluster variable rather than an explanatory variable, and therefore explicitly estimated the parameters of the model for the group of whales rather than separately for each individual.

Results

Data

A total of 175.37 hours of DTAG data were analyzed, an average of 14.6 hours of data recorded per whale (Table 1). All data from the 12 deployments were used to parameterize the hidden state model. Data for tagging effects analysis included nine DTAG deployments (87.62 h) from the time of first tag-on to the first experiment or end of the full tag record. For two of these whales, we also included the period between the start of secondary tag deployment until the end of the full DTAG record. Three whales (sw05_199a-c) were excluded completely due to incidental exposures to unidentified sonar at the beginning of the tag records (0–3 hours from tag deployment) (Table 1).

Hidden state model selection

Based on their state classification convergence at 6k–16k iterations, eight fixed-step length (FS) models were rejected and 10 accepted for further updates. All six-state FS models that did not include pitch failed to converge in terms of state classification, suggesting that pitch was important in discriminating between resting (state 5) and active-silent state (state 6). In the 10 FS models selected to be updated, all parameters converged adequately (BGR estimate < 1.05) after 16,000 iterations.

Table 2. Deviance, effective number of parameters (p_v) and deviance information criterion (DIC, based on p_v) for the 15 converged models in the last 10,000 iterations.

Model (no. states)	Structure	Deviance	p_v	DIC
6 TS (6)	depth + clicking + pitch	42654.9	4369.5	47024.4
9 TS (6)	depth + clicking + minFromSurf + pitch	42767.0	4665.1	47432.1
9 FS (6)	depth + clicking + minFromSurf + pitch	43478.9	4593.8	48072.7
6 FS (6)	depth + clicking + pitch	43446.7	5122.8	48569.5
9 TS (5)	depth + clicking + minFromSurf + pitch	46541.5	3393.7	49935.1
6 FS (5)	depth + clicking + pitch	47463.5	3654.5	51118.1
9 FS (5)	depth + clicking + minFromSurf + pitch	47505.7	3852.6	51358.3
2 FS (5)	depth + clicking	74128.5	5614.6	79743.2
8 TS (6)	depth + clicking + ODBA + pitch	118344.3	4658.2	123002.5
5 TS (6)	depth + clicking + minFromSurf + ODBA + pitch	118495.1	4543.1	123038.1
5 FS (6)	depth + clicking + minFromSurf + ODBA + pitch	119157.2	6587.2	125744.4
8 FS (6)	depth + clicking + ODBA + pitch	119091.0	6682.1	125773.1
5 FS (5)	depth + clicking + minFromSurf + ODBA + pitch	123024.8	3913.2	126938.0
8 FS (5)	depth + clicking + ODBA + pitch	122975.6	4254.8	127230.4
4 FS (5)	depth + clicking + minFromSurf + ODBA	151908.6	5713.6	157622.3

Notes: Models labelled 'FS' (fixed step length) estimated a constant step length within each state; models labelled 'TS' (time-varying step length) allowed vertical step to increase as a function of depth during foraging states. See Appendix C for calculation of p_v.

As described in the methods, the seven lowest DIC model structures were also fitted with time-varying step length (TS) during foraging states descent, LRS and ascent (Table 2). Of the seven TS models, two models with five states (models 2 and 6) failed to converge in terms of state classification; the remaining five TS models and their parameters appeared to converge sufficiently. The converged set of models improved within-chain correlation of all posterior transition probabilities from state 3 to states 2–6 compared to the same models without TS (Fig. B3).

Six-state and TS models outperformed respective five-state and FS models, both in terms of lower DIC, lower state uncertainty and higher sensitivity to pre-classified bottom phases (Fig. 3). Six-state models estimated most of the time in expert classified 'silent active swimming' dives as silent active state, and models with pitch were further able to discriminate between expert classified drifting and silent active swimming dives (Fig. 3). When vertical step was allowed to vary with depth (TS models), inclusion of ODBA appeared to somewhat improve overall state certainty and sensitivity to pre-classified bottom phases (Fig. 3).

Minimum DIC was obtained for model structures 6 (base + pitch) and 9 (base + pitch + minFromSurf) both within five- and six-state models. However, both uncertainty in state classification and sensitivity to pre-classified bottom phases ranked three models slightly above the lowest DIC model (six states, TS and pitch): full six-state TS model, and six-state TS models pitch + minFromSurf and pitch + ODBA (Fig. 3). Including ODBA in the best DIC model with pitch changed only 2.8% of its state estimates, a magnitude similar to their overall state uncertainty (~3%), and had only small contribution on the state classifications of the full TS model (Appendix C: Fig. C6). In the interests of model parsimony therefore, we selected against ODBA in the hidden state model. Including minFromSurf in the best DIC model changed the state classification even less, by 0.6%. Without minFromSurf, TS model posterior samples, transition probabilities from state 3 in particular, had a greater (>400) effective sample size. Therefore, it was the lowest DIC model 6 (base + pitch) with six states and time-varying step length that was selected for interpretation and further analyses of tagging effects.

Description of selected hidden state model

The posterior distributions of the selected hidden state model were consistent with our prior expectation of behavior. A high probability of clicking was estimated for the foraging states (posterior means for descent: 0.90, layer-restricted search: 0.99, and ascent: 0.56) while a low probability of clicking was estimated for surface, resting and silent active states (<0.02). Descent and ascent rates overall were very similar when accounting for their variability and effects of

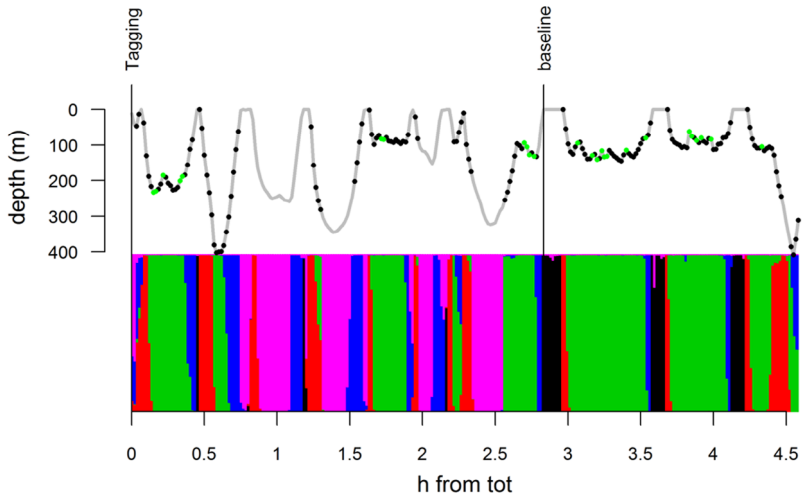

Fig. 5. Time series of state budget and dive profile for whale sw08_152a during the Tagging and post tagging baseline period. X-axis shows time since tag-on time (tot). Bottom graph shows posterior probabilities for each state (color-coded as in Fig. 3). Top graph shows 1-minute depth data (gray) overlaid with presence/absence of clicking (black) and presence/absence of buzzing (green).

depth (Fig. 4). During foraging states (descent, LRS and ascent), step length was estimated to increase by 1.47 (SD 0.02) m/min for every unit increase in square root transformed depth. The posterior mean absolute value of pitch was 1.3 (SD 45.6) degrees during surfacing and 80.5 (SD 45.8) degrees during resting. See Appendix D for complete description of the selected model.

Effects of tagging

When the tag boat remained near the whales in tagging operations (n = 8.1 h), the whales spent no time resting, and across individuals, an average of 1.6× more time in the silent active state (10.1%, SD = 13.2) and less time surfacing (12.4%, SD = 10.2) compared to baseline periods when the boat was recovered (n = 79.4 h; 60% increase from 6.3%, SD = 15.1 and 34% decrease from 18.8%, SD = 5.5, respectively) (Figs. 5 and 6).

The most prolonged tagging period was for whale sw08_152a that was approached by the tag boat for 2.8 hours after tag attachment attempting to photograph the whale (Fig. 5). During those 2.8 hours, the whale spent only 1.8% of the time in surfacing state 1, and 31.4% of the time in silent-active state. With most of the silent-active state comprised silent diving, the whale spent only 12.3% of its time near surface (<10 m). This compared to 12.7–27.8% of time spent in the surface state across the deployments during post-tagging (Appendix E: Table E3). Immediately after the tag boat left the whale, it spent eight minutes in the surfacing state, which was the longest period the whale spent in the surfacing state during the entire DTAG record (post-tagging individual average surface duration was 6.8 min SD 2.0).

The lowest AIC model for state transitions included prevState + whale + Tagging, which improved the baseline model prevState + whale by 9.5 AIC units (Fig. 7). Tagging covariate was also supported by a likelihood ratio test between the two models (df = 5, p = 0.002, function anova.nnet()). The model estimated 86.5% of the post-tagging baseline states and 79.2% of the Tagging period states correctly. The model fit best to LRS and drifting states (92.7% and 88.5% correct predictions, respectively), and worst to silent active state (64.8%) (Appendix E: Fig. E1).

The binomial GEE model for silent active state with prevState and Tagging as covariates and whale as a cluster variable improved the QIC of the baseline GEE model by 28.9 units (Table 3). The GEE model with Tagging estimated the odds of silent active state to increase by a factor of 3.70 [95% CI 1.3, 10.1] during the tagging period (Appendix E: Table E5), slightly greater but

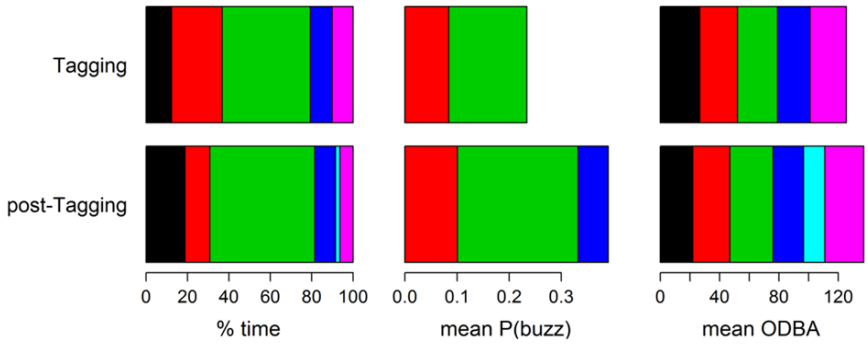

Fig. 6. Behavioral time budgets (left) and proxies of foraging success (probability of buzzing, center) and locomotion cost (mean ODBA, right) averaged across individuals during the Tagging condition (top, 8.1 h) and post Tagging baseline (bottom, 79.4 h). Each state is color-coded as in Fig. 3.

within the confidence intervals (2 × SE) of the respective coefficient estimate from the multinomial model (Appendix E: Table E4). The GEE model results confirmed that the detected change in state transitions of the multinomial model was not merely a by-product of serial correlation. We detected both positive and negative residual correlation in the best multinomial model (Appendix E: Fig. E2).

Probability of buzzing was highly variable within and across individuals, but the individual average for foraging states was somewhat lower during the tagging period (descent state: 8.4% SD = 14.4; LRS state: 15.0%, SD = 11.1; ascent state: 0.0%) than in the post-tagging baseline (descent state: 10.1 % SD = 8.6; LRS state: 23.1% SD = 14.6; ascent state: 5.8% SD = 4.5) (Fig. 6). There was no consistent increase in ODBA during the tagging period compared to baseline across states. Only surface and ascent states had slightly greater individual average ODBA during Tagging (surface state: 26.7 SD = 7.4; ascent state: 22.0 SD = 5.5) than post-tagging (surface state: 21.9 SD = 3.5; ascent state: 20.6, SD = 3.5) (Fig. 6).

In the AIC model selection, there was little support for an overall tagging effect on probability of buzzing, given the foraging states (descent, LRS and ascent states) and whale as a factor covariate. Tagging improved the baseline model state + whale by only 0.74 AIC units (Fig.

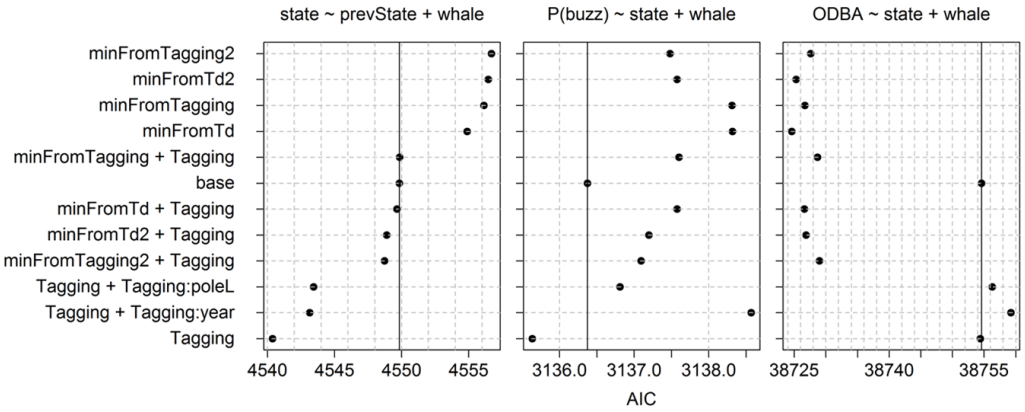

Fig. 7. AIC model selection for tagging effects on state transitions (left) probability of buzzing (middle) and ODBA (right). Baseline models are shown on top of each Fig. and candidate Tagging covariate combinations on the left. Black solid circles show AIC for each model and vertical line AIC value for the baseline model. Models were considered to have performed better than the baseline model if their AIC was at least two units lower (horizontal grid length).

Table 3. The lowest AIC models (generalized linear models; no random effects) and the corresponding GEE models with QIC (with whale as a random effect).

Response variable	Explanatory variables	Random effect	AIC/QIC	ΔAIC/ΔQIC
state	prevState + whale		4549.84	0.00
state	prevState + whale + Tagging		4540.38	−9.45
state 6	prevState	whale	1020.14	0.00
state 6	prevState + Tagging	whale	991.21	−28.92
buzz	state + whale		3136.38	0.00
buzz	state + whale + Tagging		3135.63	−0.74
buzz	state	whale	3253.02	0.00
buzz	state + Tagging	whale	3246.63	−6.39
ODBA	state + whale		38754.62	0.00
ODBA	state + whale + minFromTd		38724.62	−30.00
ODBA	state + whale + Tagging		38754.39	−0.23
ODBA	state	whale	191513.45	0.00
ODBA	state + minFromTd	whale	192057.18	+543.73
ODBA	state + Tagging	whale	191531.97	+18.52

7). The coefficient estimate for Tagging was small (−0.32, SE = 0.20) with no evidence that it was different from zero ($z = -1.61$, $p = 0.107$).

The best AIC model for ODBA included state + whale + minFromTd. minFromTd improved the baseline model by 30.0 AIC units (Fig. 7), however, the estimated effect was very small (−0.18 decrease in mean ODBA for every hour). When fitted within a GEE which accounts for serial correlation in the data, neither minFromTd, Tagging nor Tagging:state were supported with respective QIC increases of 543.7, 18.5 and 1353.22 units compared to the base model (Table 3). We therefore concluded that there was no evidence for a change in ODBA as a function of time since tag-on time or tag boat presence.

Discussion

In this study, we were able to estimate a time-series of functional behavioral states that most fully captured the variability in diverse data streams recorded by an animal-attached movement and sound recording tag. Our results demonstrate that a 'silent active' state can be identified despite lack of a prior functional description for this state, and that including this state along with a defined resting state improved the functional state models for behavior of Norwegian sperm whales. We then used the output of the model to evaluate three possible behavioral responses to tagging procedures: (1) change in behavioral time-budget, (2) reduction in prey capture attempts, given behavioral state and (3) increase in movement cost, given behavioral state. We demonstrate that whales spent more time in the silent active state when the tag boat was present, and that a simple present versus absent response explained the data better than time-decaying models of behavioral response. This enables quantitative determination of post-handling periods that should be excluded to retain periods more likely to reflect baseline behavior.

Hidden state models

The hidden state models were able to estimate both very stereotyped states (surfacing, resting) and states with highly variable data signatures (layer-restricted search LRS, other non-foraging) (Fig. 4). Although there was uncertainty in formal model selection in this Bayesian framework, different hidden state models arrived at similar state classifications, which all agreed well with expert classifications (Fig. 3; Appendix C: Tables C1 and C2). The hidden state models succeeded to identify and characterize states that could be interpreted in terms of functional behaviors previously documented for sperm whale foraging. The accepted hidden state model included six states, time-varying vertical step length for foraging states (descent, LRS and ascent), clicking, and log-linear relationships between vertical step and the absolute value of pitch. This model had the lowest DIC score, and was selected as the most parsimonious model amongst the models with the lowest uncertainty and agreement with expert opinion (Fig. 3).

Allowing step length to increase with depth (TS models) improved the DIC, state uncertainty,

and sensitivity to pre-classified bottom phases compared to models with a simple random walk with fixed step length (FS models) (Fig. 3). TS models also appeared to capture an active foraging mode better than FS models that estimated the highest average ODBA during descent rather than LRS. We do not postulate that sperm whales are intrinsically more mobile at depth, but rather that the time-varying formulation for step length was more flexible by accepting a wider distribution of step lengths for LRS, and was subsequently able to more fully capture an active foraging mode. Such high variability in step length across foraging phases could be expected when prey layers vary in vertical thickness, quality and/or prey species that in turn influence whales' hunting and searching strategy.

Functional time budget of foraging male sperm whales

Layer-restricted search (LRS) was estimated as the most prevalent state in the post-tagging data (47.5% of all data, and 51.2% of all foraging states 1–4), consistent with the high proportion of time spent in foraging and high diving efficiencies (foraging phase duration: dive cycle duration) reported for sperm whales both at high- and low latitudes (Jaquet et al. 2000, Watwood et al. 2006). Unlike studies using bottom phase (defined by the first descent and final ascent of a dive) or foraging phase (defined by the first and last buzz of a dive) (Watwood et al. 2006) alone as a measure of foraging time, we were able to estimate multiple foraging phases within a dive (e.g., Appendix D: Fig. D3b). Thirty of 119 (25.2%) "usual" foraging dives (expert dive types 1–4 in Appendix C: Table C4) contained more than one foraging (LRS) phase.

There was strong support for a sixth 'silent active' state, with six-state models outperforming five-state models in terms of higher overall posterior probability of states (Fig. 3) and a better fit of state-dependent likelihoods to the data (Appendix C: Fig. C3). Furthermore, there was high concordance between the state 6 estimates and expert classified "silent active" dives (Fig. 3). Nevertheless, state transitions appeared to be relatively weak predictors of state 6 compared to other states (Appendix C: Fig. C5), with wide posterior credible intervals for the transition probability of staying in state 6 (Appendix D: Fig. D1). However, variability related to state transitions is expected as state 6 likely encompassed several non-foraging behaviors that may have been associated with different functional behaviors and contexts, such as socializing, avoiding the tag boat near surface, or horizontal transit. Future work with larger datasets could consider the potential to divide state 6 into more specific functional states.

State 5 (resting/drifting) was estimated for 3.8% of post-tagging baseline data, most of which coincided with expert classified drifting dives based upon the description of this behavior by Miller et al. (2008). For two whales (sw05_196a and sw10_150a) state 5 also identified drifting to the sea surface that occurred at the end of foraging dives (max depth 306 m; Appendix D: Fig. D3a). Drifting had a very distinct data signature featuring little vertical movement yet nearly vertical pitch (posterior mean and 95% CRI was estimated for step length as 8.5 [7.9, 9.0] m, and for pitch as 80.5 [79.8, 81.0] degrees), consistent with stereotyped vertical posture drift-dives documented for sperm whales world-wide (Miller et al. 2008).

Effects of tagging

Using the estimated states and uncertainty to assess tagging effects, we found that sperm whales increased time in non-foraging silent active state (Table 3, Figs. 6–7). Within each behavioral state, we found no evidence of changes in a proxy for locomotion cost (ODBA) or a proxy for foraging success (probability of buzzing) (Table 3, Figs. 6–7). Sw08_152a was re-approached by the tag boat for the longest duration (2.8 h), and during this time, spent 31.4% of time in silent active state compared to no time in this state in the post-tagging period that consisted of three foraging dives and longer surface periods (Fig. 5). These results indicate an evasion or vigilance reaction to the tag boat that disrupted behavior, rather than a direct reduction in prey capture attempts or change in locomotion cost within behavioral states. No longer-term effects could be detected on the time scale of each tag record (~15–20 h in duration), suggesting that whales recovered to a post-tagging level of behavior relatively soon after the tag boat left.

We were not able to collect comparable data during the pre-tagging period, and therefore could not establish with certainty that behavior was resumed to a completely undisturbed (non-tagged) level. However, two whales (sw10_149a and sw10_150a) were re-approached for a secondary tag deployment and had a response profile consistent to whales that were tagged only once. Both whales spent time in the silent active state near the sea surface during first and second tagging periods, while full foraging dive cycles were resumed soon after the tag boat left the whales. Within both tag records, the first and second post-tagging deployment periods consisted of near identical time and depth profile of foraging (descent, LRS and ascent states) with little apparent effect of the presence of a secondary tag (Appendix D: Fig. D3k–l). We conclude that the presence of tags alone on the animal was likely to have little effect on whale behavior compared to tag deployment procedures ('handling'). Indeed the DTAG only weighs 300 g, which is less than 0.01% of an adult sperm whale mass (14–50 t). Little is known about the effects of suction-cup tag attachment, however tags typically detach if a sperm whale performs a breach, indicating that an uncomfortable tag can be removed by the whale (Johnson et al. 2009). Nevertheless, multiple tagging of the same individual appears a promising approach to decompose the effects of handling vs. tag presence on the animal, as well any sensitization or habituation to tag deployment procedures. Future studies could address these effects with a larger sample size of secondary tag deployments and include tags of varying sizes.

Although we did not find evidence for changes in energetic proxies within states, the increased probability of non-foraging silent active behavior and reduced time at surface suggests an energetic cost of tag boat presence. Miller et al. (2009) found similar short-term changes in sperm whale foraging behavior during the first dive of the tag record but not subsequent dives. These changes included reduced buzz and pitching rates during the bottom phase, and shorter dive duration compared to the subsequent dive. However, the presence/absence of tag boat was a more important predictor of effects than time since tag-on time, suggesting a lack of a specific cut-off period after tag attachment. This result is expected when the tagging procedure, including re-approach for photo-identification, varies across tagging occasions. In such cases it is important to collect detailed data on tagging effort to describe the 'dose' of handling, such as tag boat distance to the whale, with focus on recording the intensity and duration of approach both before and after a successful tag attachment.

Methods considerations

Although we did find evidence of short-term tag boat effects (tagging period duration ranged between 0.1 and 2.8 h), the sensitivity of our test for subtle longer-term effects was likely to be limited due to the relatively small number of tags ($n = 9$) and high variability in state budgets and buzz rates across the tag records. Variability in the tagging procedure across years and different tagging crews was also likely to affect the probability and level of the individual responses. We did not find evidence for a different level of response to shorter pole length (Fig. 7), however due to the small sample size we could not account for other factors that could have been equally important, such as tag boat handling and targeting tag placement near the head vs. the back of the animal. Tagging protocol in 2005 and 2008 aimed to place tags at the anterior end of the head, while in 2009–2010 whales were typically approached from behind and tags were placed on the back of the whale anterior to the dorsal fin.

Tagging periods after tag attachment ranged from just 6 minutes up to 2.8 hours (1.1–61.4% of tag records). Our time-series approach explicitly modeled this unbalanced sample, and we also contrasted results from GLMs that estimated individual level differences with GEEs that estimated individual average and between-individual variability in the response data. Nevertheless, it was possible that a few individuals that responded strongly to the presence of tag boat were influential in the estimation of a population effect. We tested the influence of individuals by re-fitting the baseline and the tagging effects GLM:s for state without each individual, and found that excluding either sw08_152a or sw10_150a lowered the AIC difference below our ΔAIC threshold of -2 (Appendix E: Fig. E4). sw08_152a was exposed to the longest tagging

period (2.8 h), whereas sw10_150a was approached by the tag boat twice for shorter periods, including secondary tag deployment. Both sw08_152a and sw10_150a spent the longest durations in state 6 (an average of 6.5 and 4.0 minutes, respectively) compared to any other whale during the tagging period, and were not estimated to return to state 6 in the baseline period. Therefore, had we not sampled these two individuals, we would have not been able to estimate a tag boat effect overall.

As one of the first attempts at multivariate hidden state modeling of individual behavior, we simplified the hidden state model structure by assuming mostly Markov transitions, no individual effects or spatial memory for prey layers. Despite the relatively simple process model, a sufficiently strong signal in the input data allowed for robust state classification and estimation of time budgets that were highly variable across individuals. A more realistic (complex) process model would be required if disturbance was incorporated and tested within the hidden state model. For example, a hidden state model with individual as a random effect could estimate population-average effects by incorporating tag boat presence as an explanatory variable for buzzing within each state.

Implications and future steps

The functional state approach appears to be able to effectively estimate behavioral disturbances that can be linked to individual fitness. We showed that after tag deployment, whales can remain vigilant to the presence of tag boat and thus trade off foraging time for perceived risk at surface. During-after comparisons of functional states and currency proxies were influenced by individuals that were exposed to tag boat repeatedly or for extended durations, highlighting the importance of consistent deployment procedures and minimizing handling time. Nevertheless, we succeeded to estimate a cut-off point (tag boat recovery) after which whales were likely to have returned to a post-tagging level of behavior, and recommend before-during comparison of behavior where pre-tagging baseline data is not available. Our results also lend support for the exclusion of handling periods to better capture post-tagging baseline behavior. However, comparable pre-tagging data would be needed to quantify tagging effect as a deviation from the non-tagged population of interest. An optimal design would monitor behavior during all phases of tag deployment (before approach, during tag deployment procedures, during attachment, and after the device is detached), and from a platform that minimized research effects. For cetaceans such as sperm whales that use biosonar to locate prey, remote visual and passive acoustic tracking could be used to monitor foraging and movement before-during-after tag deployment, as well as complement fine-scale on-board acoustic and orientation data during the tag record.

Our concept model and hidden state approach was based upon first principles of searching behavior (transiting vs. encamped search) and central-place foraging (surface vs. diving) that are transferrable across species, and we show that not all such functional states need strict definition a priori to be estimated. The hidden state model also incorporated species-specific expectations of behavior (echolocation, drifting posture), combined multiple sources of data to estimate biologically interpretable states and parameters (such as descent rate), and allowed modeling of currency proxies within the relevant behavioral contexts. Unlike expert classifications, hidden state modeling is automated and quantifies uncertainty. As well as for offline-analysis, these are also desirable properties for an on-board data compression algorithm, and state-estimation of fine-resolution archival tags could guide the development and use of such algorithms in data-relaying tags that aim to collect data for months or even years. With recent advances in deviance-based selection for Bayesian mixture models (e.g., Plummer 2003), there is also more scope to incorporate and test a range of disturbance effects as explanatory variables within hidden state models, rather than as a second AIC-based analysis step. Behavioral context is increasingly highlighted as the key to understanding the fitness trade-offs of behavioral decision making in response to anthropogenic stimuli (Gill et al. 2001, Beale 2007) and within such flexible hierarchical estimation frameworks, could be explicitly modeled by conditioning disturbance effects by behavioral state.

Acknowledgments

We thank Mark Johnson, Peter Madsen, and all 3S (Sea mammals, Sonar, Safety) team members for efforts on the field data collection and access. We would like to thank 3S principal investigators Petter Kvadsheim, Frans-Peter Lam and Peter Tyack especially for enabling this research along with 3S partners and funders (UK Ministry of Defence, U.S. Office of Naval Research, and World Wildlife Fund, Norway). We also owe the team in Centre for Research into Ecological and Environmental Modelling (CREEM, St Andrews) including Len Thomas, Catriona Harris, Stacy DeRuiter, Dina Sadykova, Roland Langrock and Tiago Marques a big thanks for analytical advice.

Literature Cited

Bannasch, R., R. P. Wilson, and B. Culik. 1994. Hydrodynamic aspects of design and attachment of a back-mounted device in penguins. Journal of Experimental Biology 194:83–96.

Barron, D. G., J. D. Brawn, and P. J. Weatherhead. 2010. Meta-analysis of transmitter effects on avian behaviour and ecology. Methods in Ecology and Evolution 1:180–187.

Beale, C. M. 2007. The behavioral ecology of disturbance responses. International Journal of Comparative Psychology 20:111–120.

Berger-Tal, O., T. Polak, A. Oron, and Y. Lubin. 2011. Integrating animal behavior and conservation biology: a conceptual framework. Behavioral Ecology 22:236–239.

Boyd, I. L. 1997. The behavioural and physiological ecology of diving. Trends in Ecology & Evolution 12:213–217.

Brooks, S. P., and A. Gelman. 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7:434–455.

Cooke, S. J., S. G. Hinch, M. Wikelski, R. D. Andrews, L. J. Kuchel, T. G. Wolcott, and P. J. Butler. 2004. Biotelemetry: a mechanistic approach to ecology. Trends in Ecology and Evolution 19:334–343.

Curé, C., R. Antunes, A. C. Alves, F. Visser, and P. H. Kvadsheim. 2013. Responses of male sperm whales (*Physeter macrocephalus*) to killer whale sounds: implications for anti-predator strategies. Scientific Reports 3:1–7.

Engelhard, G. H., A. J. Hall, S. M. J. M. Brasseur, and P. J. H. Reijnders. 2002. Blood chemistry in southern elephant seal mothers and pups during lactation reveals no effect of handling. Comparative Biochemistry and Physiology Part A 133:367–378.

Ferrari, S. L. P., and F. Cribari-Neto. 2004. Beta regression for modelling rates and proportions. Journal of Applied Statistics 31:799–815.

Fossette, S., H. Corbel, P. Gaspar, Y. Le Maho, and J.-Y. Georges. 2007. An alternative technique for the long-term satellite tracking of leatherback turtles. Endangered Species Research 3:33–41.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. Bayesian data analysis. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Gill, J. A., K. Norris, and W. J. Sutherland. 2001. Why behavioural responses may not reflect the population consequences of human disturbance. Biological Conservation 97:265–268.

Glas, A. S., J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt. 2003. The diagnostic odds ratio: a single indicator of test performance. Journal of Clinical Epidemiology 56:1129–1135.

Godfrey, J. D., and D. M. Bryant. 2003. Effects of radio transmitters: Review of recent radio-tracking studies. Science for Conservation 214:83–95.

Halsey, L. G., E. L. C. Shepard, F. Quintana, A. Gomez Laich, J. A. Green, and R. P. Wilson. 2009. The relationship between oxygen consumption and body acceleration in a range of species. Comparative Biochemistry and Physiology, Part A 152:197–202.

Hazekamp, A. A. H., R. Mayer, and N. Osinga. 2010. Flow simulation along a seal: the impact of an external device. European Journal of Wildlife Research 56:131–140.

Isojunno, S., and P. J. O. Miller. 2014. Hidden Markov models capture behavioral responses to suction-cup tag deployment: a functional state approach to behavioral context. Effects of Noise on Aquatic Life. Springer, New York, New York, USA.

Jaquet, N., S. Dawson, and E. Slooten. 2000. Seasonal distribution and diving behaviour of male sperm whales off Kaikoura: foraging implications. Canadian Journal of Zoology 78:407–419.

Johnson, M. P., N. A. Soto, and P. T. Madsen. 2009. Studying the behaviour and sensory ecology of marine mammals using acoustic recording tags: a review. Marine Ecology Progress Series 395:55–73.

Langrock, R., T. A. Marques, R. W. Baird, and L. Thomas. 2013. Modeling the diving behavior of whales: a latent-variable approach with feedback and semi-Markovian components. Journal of Agricultural, Biological, and Environmental Statistics 19:82–100.

Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. 2013. The BUGS book: A practical introduction to Bayesian analysis. Chapman & Hall/CRC, Boca Raton, Florida, USA.

McCafferty, D. J., J. Currie, and C. E. Sparling. 2007. The effect of instrument attachment on the surface temperature of juvenile grey seals (*Halichoerus grypus*) as measured by infrared thermography. Deep Sea Research II 54:424–436.

McClintock, B. T., D. J. F. Russell, J. Matthiopoulos, and R. King. 2013. Combining individual animal movement and ancillary biotelemetry data to investigate population-level activity budgets. Ecology 94:838–849.

McMahon, C. R., I. C. Field, C. J. A. Bradshaw, G. C. White, and M. A. Hindell. 2008. Tracking and data–logging devices attached to elephant seals do not affect individual mass gain or survival. Journal of Experimental Marine Biology and Ecology 360:71–77.

Miller, P. J. O., K. Aoki, L. E. Rendell, and M. Amano. 2008. Stereotypical resting behavior of the sperm whale. Current Biology 18:21–23.

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 behavior of sperm whales in the Gulf of Mexico. Deep Sea Research I 56:1168–1181.

Miller, P. J. O., M. P. Johnson, and P. L. Tyack. 2004. Sperm whale behaviour indicates the use of echolocation click buzzes "creaks" in prey capture. Proceeding of the Royal Society of London B 217:2239–2247.

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

Murray, D. L., and M. R. Fuller. 2000. A critical review of the effects of marking on the biology of vertebrates. Pages 15–64 *in* L. Boitani and T. K. Fuller, editors. Research techniques in animal ecology. Columbia University Press, New York, New York, USA.

Nathan, R., W. M. Getz, E. Revilla, M. Holyoak, R. Kadmon, D. Saltz, and P. E. Smouse. 2008. A movement ecology paradigm for unifying organismal movement research. Proceedings of the National Academy of Sciences 105:19052–19059.

Oliveira, C., and M. Wahlberg. 2013. The function of male sperm whale slow clicks in a high latitude habitat: Communication, echolocation, or prey debilitation? Journal of the Acoustical Society of America 133:3135–3144.

Photopoulou, T. 2012. Diving and depth use in seals: inferences from telemetry data using regression and random walk movement. Dissertation. University of St Andrews, Fife, UK.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *In* Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Austrian Association for Statistical Computing, Vienna, Austria.

Ropert-Coudert, Y., C.-A. Bost, Y. Handrich, R. M. Bevan, P. J. Butler, A. J. Woakes, and Y. Maho. 2000. Impact of externally attached loggers on the diving behaviour of the king penguin. Physiological and Biochemical Zoology 73:438–445.

Ropert-Coudert, Y., N. Knott, A. Chiaradia, and A. Kato. 2007. How do different data logger sizes and attachment positions affect the diving behaviour of little penguins? Deep Sea Research I 54:415–423.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B 64:583–639.

Su, Y., and M. Yajima. 2012. R2jags: A package for running jags from R. http://CRAN. R-project. org/package= R2jags

Teloni, V., J. P. Mark, M. J. O. Patrick, and M. T. Peter. 2008. Shallow food for deep divers: Dynamic foraging behavior of male sperm whales in a high latitude habitat. Journal of Experimental Marine Biology and Ecology 354:119–131.

Tyack, P. 2009. Acoustic playback experiments to study behavioral responses of free-ranging marine animals to anthropogenic sound. Marine Ecology Progress Series 395:187–200.

Viterbi, A. J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. Information Theory, IEEE Transactions on 13:260–269.

Walker, K. A., A. W. Trites, M. Haulena, and D. M. Weary. 2012. A review of the effects of different marking and tagging techniques on marine mammals. Wildlife Research 39:15–30.

Watwood, S. L., P. J. O. Miller, M. Johnson, P. T. Madsen, and P. L. Tyack. 2006. Deep-diving foraging behaviour of sperm whales (*Physeter macrocephalus*). Journal of Animal Ecology 75:814–825.

Wilson, R. P., W. S. Grant, and D. C. Duffy. 1986. Recording devices on free-ranging marine animals: does measurement affect foraging performance? Ecology 67:1091–1093.

Wilson, R. P., and C. R. McMahon. 2006. Measuring devices on wild animals: what constitutes acceptable practice? Frontiers in Ecology and the Environment 4:147–154.

SUPPLEMENTAL MATERIAL

APPENDIX A

Hidden state model specification

Table A1. Estimable parameters in full model with prior and initial value specification.

Parameter			Initial values		
Jags name [state]	Symbol	Prior distribution	Chain 1	Chain 2	Chain 3
T.beta11[1]	β_{1,S_t}	Gaussian(0, 0.1)T(−15, −0.2)	−0.2	−2.3	−5.8
T.beta31[3]	β_{2,S_t}	Gaussian(0, 0.1)T(,1.0E−6)	−0.3	−0.03	−4.02
tau[1]	σ^2_1	Gamma (1, 1)	0.001	4	1
tau[2]	σ^2_2	Gamma(2.0736, 0.0023)	1500	900	450
tau[3]	σ^2_3	Gamma(2.0736, 0.0023)	300/100	700/200	150/93
tau[4]	σ^2_4	Gamma(2.0736, 0.0023)	1500	900	450
tau[5]	σ^2_5	Gamma(2.0736, 0.0023)	60	24	93
tau[6]	σ^2_6	Gamma(2.0736, 0.0023)	300	700	150
tau.beta0[2, 3, 4]	μ	Gamma(3, 2)	0.1	2	0.5
drift.t[2]	π_2	Gamma(9.7344, 0.1248)	60	132	45
drift.t[4]	π_4	Gamma(9.7344, 0.1248)	60	132	45
RC.beta0[1]	γ_1	Beta(1, 10)	0.004	0.08	0.03
RC.beta0[2]	γ_2	Beta(2, 1)	0.59	0.98	0.76
RC.beta0[3]	γ_3	Beta(2, 1)	0.59	0.98	0.76
RC.beta0[4]	γ_4	Beta(2, 1)	0.59	0.98	0.76
RC.beta0[5]	γ_5	Beta(1, 10)	0.004	0.08	0.03
RC.beta0[6]	γ_6	Beta(1, 10)	0.004	0.08	0.03
a.mean[1]	φ_S/ω_1	Gamma(5, 0.2)	18	10	30
a.mean[2]	φ_S/ω_2	Gamma(5, 0.2)	18	10	30
a.mean[3]	φ_S/ω_3	Gamma(5, 0.2)	18	10	30
a.mean[4]	φ_S/ω_4	Gamma(5, 0.2)	18	10	30
a.mean[5]	φ_S/ω_5	Gamma(1, 1)	9	5	15
a.mean[6]	φ_S/ω_6	Gamma(5, 0.2)	18	10	30
a.rate[1]	ω_1	Gamma(3, 0.5)	1.5	0.1	1
a.rate[2]	ω_2	Gamma(3, 0.5)	1.5	0.1	1
a.rate[3]	ω_3	Gamma(3, 0.5)	1.5	0.1	1
a.rate[4]	ω_4	Gamma(3, 0.5)	1.5	0.1	1
a.rate[5]	ω_5	Gamma(1, 2)	1.5	0.1	1
a.rate[6]	ω_6	Gamma(3, 0.5)	1.5	0.1	1
p.beta0[1]	$\alpha_{0,1}$	Gaussian(0, 0.1)	−3	−1	−5
p.beta0[2, 3, 4, 6]	α_{0,S_t}	Gaussian(0, 0.1)	−3	−1	−5
p.beta0[5]	$\alpha_{0,5}$	Gaussian(0, 0.1)	2	1	3
p.gamma[1]	τ_1	Gamma(1, 0.1)T(1.0E-6,)	20	7	70
p.gamma[2, 3, 4, 6]	τ_S	Gamma(1, 0.1)T(1.0E-6,)	20	7	70
p.gamma[5]	τ_5	Gamma(1, 0.1)T(1.0E-6,)	20	7	70
p.beta1[2, 3, 4, 6]	α_{1,S_t}	Gaussian(0, 1.0E-6)	0.04	0.0031	0.0928

Notes: Parameter names are given as in jags code in the Supplement, with number in square brackets referring to the state (1–6) that the parameter is associated with. Symbols refer to the notation in the main text. Gamma distribution for ODBA was parameterized in terms of mean and a rate parameter, so that mean = shape/rate. Parameters for prior distributions were specified and are given here as in JAGS Version 3.2.0 user manual (mean and precision for Gaussian, shape and rate parameters for Gamma, and first and second shape parameters for Beta distribution). T(x1, x2) shows prior truncation with lower limit x1 and/or upper limit x2 (a single limit indicates one-sided truncation). Initial values were fixed, i.e., not generated randomly for each chain. The values were chosen to represent the prior distribution in an over-dispersed fashion. For models that allowed vertical step to vary with depth, the initial value for tau[3] was lowered because it was introduced in the model as an intercept. The initial values for models with fixed step length and varying step length are shown before and after slash (/).

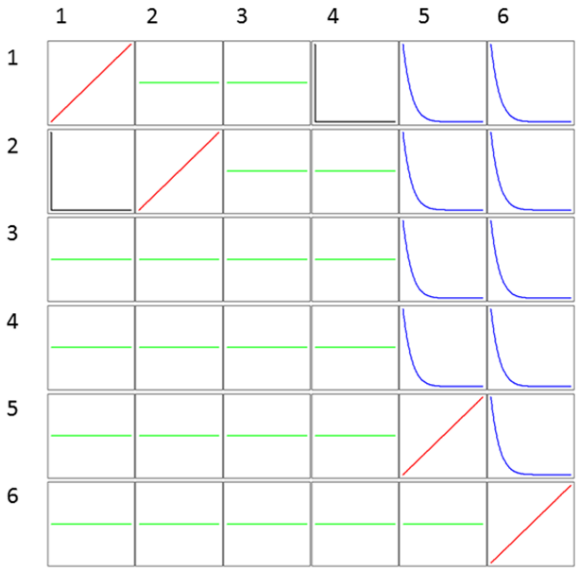

Fig. A1. Beta prior densities (y-axes) illustrated for transition probability intercepts (x-axes). Rows show state (1–6) at time t, and columns state (1–6) at time t + 1. Green shows transitions with uninformative prior density Beta(shape1 = 1, shape2 = 1), and other colors show different types of informative Beta priors: from state 2 to 1, and state 1 to 4 (~Beta(1, 1e+06)); from states 1–4 to state 5, and states1–5 to 6 (~Beta(1, 10)); and staying in states 1, 2, 5 and 6 (~Beta(2, 1)).

Appendix B

Hidden state model convergence

Methods.—All models were sampled in 3 independent chains, with an initial 16 000 iterations each. Initial values were set manually for all state parameters (Table A1).The first 6k iterations were discarded for adaptation and burn-in. The remaining 10k iterations were down-sampled (thinned) to reduce autocorrelation in the analyzed samples, and monitored for convergence of state classification to accept a subset of models for a further 20k iterations (Fig. B1). A lower down-sampling rate was used initially (every 18th and 6th sample for iterations 6–12k and 12–16k) to explore serial correlation in chains, while a factor of 50 was used to down-sample the additional (20k) iterations.

The model assessment after 16k iterations was done to reduce computation time on models that were deemed unlikely to converge at all. Two criteria were set for rejecting a model for further updates: (1) for at least two states, state proportions were so diverged that the samples did not overlap, (2) the state proportions appeared stationary across the iterations (6–16k) (Fig. B1). Such models were considered to be poor representations of the data and were not included in further model selection. The remaining models were assessed visually for parameter convergence, stationarity and serial correlation of the chains across all iterations (6–36k). Brooks-Gelman-Rubin (BGR; Brooks and Gelman 1998, Gelman et al. 2003) diagnostics were calculated separately for iterations 16–36k and 26–36k using a 95% credible interval in function gelman.diag (r package 'coda'). Model convergence was rejected based on its poorest converging parameter. All convergence assessment accounted for down-sampling rate. The convergence assessment procedure was repeated for models with variable step length, but with more iterations (48k) and a down-sampling rate at 18.

An accurate estimate of the posterior distribution requires a sufficient number of independent samples. Conventionally, an effective sample size

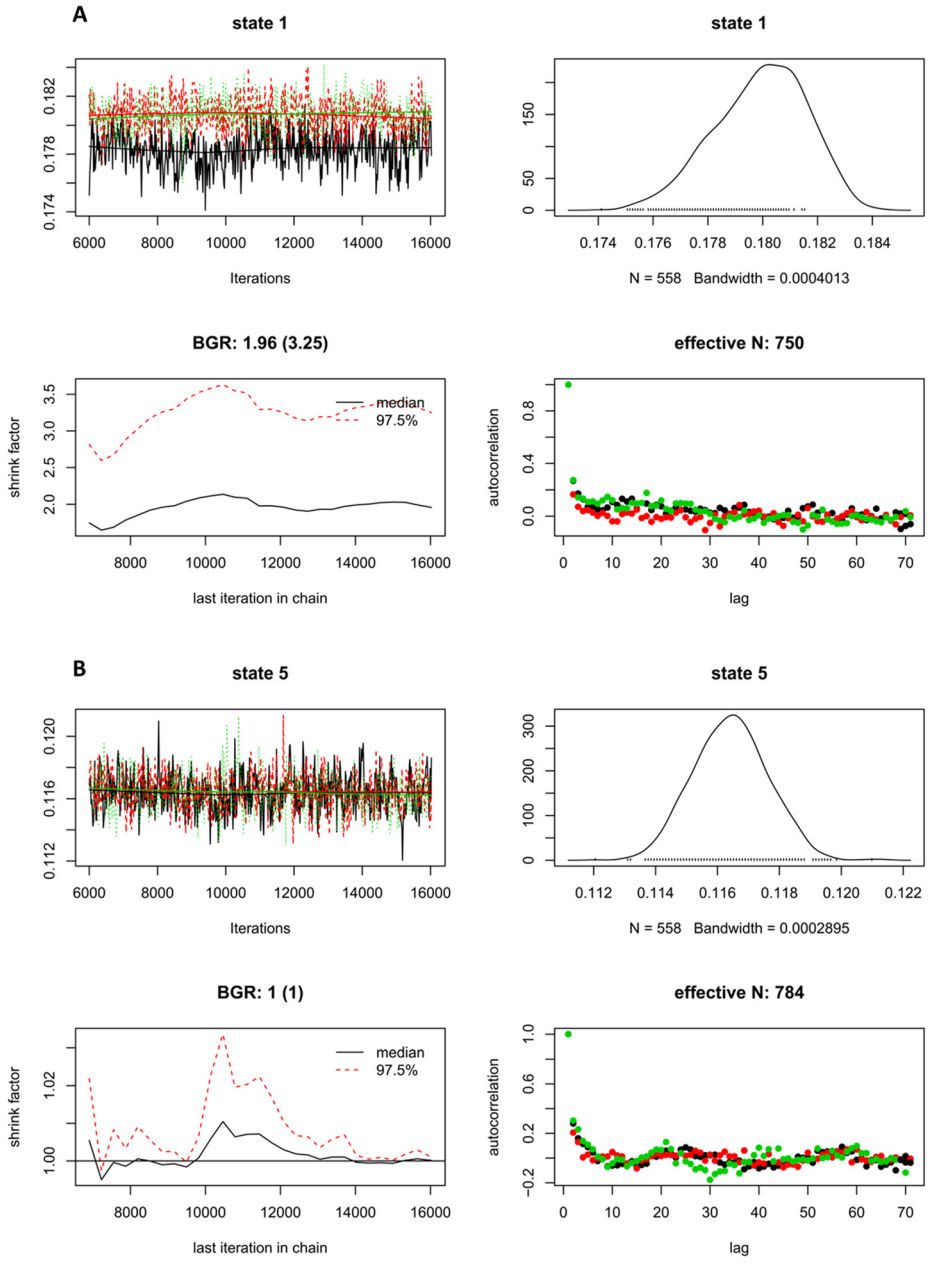

Fig. B1. Example output used to visually assess convergence of state classifications. Convergence of state classification was approximated by total proportion of states in the posterior sample. Top left panel shows proportions in each chain (color-coded) as a function of iteration history (x-axis). Top right panel shows posterior density distribution pooled across chains. Bottom left panel shows shrink factor (Brooks-Gelman-Rubin Diagnostic, BGR) as a function of the upper limit of samples that were used in the calculation of the diagnostic (function gelman.plot in package coda in r). The BGR value in the title was calculated for the whole posterior sample, with 95% confidence interval in the brackets. Bottom right plot shows autocorrelation for each chain as a function of lag; the total effective sample size for the samples is shown in the title (function effectiveSize in package coda in r). The examples show model 1 with six states (a) that was rejected based on its poor convergence of state proportions, and model 2 with five states (b) that appeared to converge adequately.

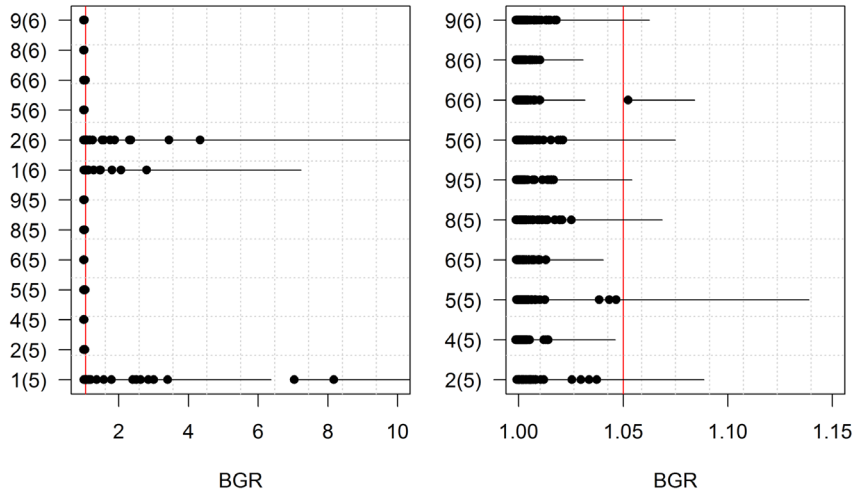

Fig. B2. Brooks-Gelman-Rubin (BGR) estimate (solid circles) and 95% CI (lines) for each fixed step length model (y-axis, number of states in brackets) at iterations 16–36k. BGR estimate of 1.05 was used as an acceptance threshold for parameter convergence. Left-hand panel includes three models that were rejected based on their poor state classification convergence (1(5), 1(6) and 2(6)) at 6–16k iterations. These models were updated to check that the rejection based on state classification was conservative. As expected, model parameters remained poorly converged. Right-hand panel only includes models that were accepted for state classification convergence.

of 400 is used (Lunn et al. 2013). Posterior summary statistics were therefore based on the number of iterations that contained at least 400 independent samples for each parameter and state proportion in the converged set (function effectiveSize in package coda).

Results.—Based on their state classification convergence at 6k-16k iterations, eight FS models were rejected and 10 accepted for further updates (Fig. B2). No five-state model state classification converged unless they included clicking (Model 2), however otherwise it was less clear how five-state model structures contributed to state convergence at 6–16k iterations. While 'clicking with minFromSurf' or 'ODBA' models (3 and 7) failed to converge, clicking did converge with 'minFromSurf + ODBA' (Model 4). In addition, convergence of state classification was satisfactory for all five-state models that included pitch (models 5, 6, 8 and 9). Near-surface behaviors appeared to be the most challenging to converge into states 1, 5 and/or 6 across model structures. The base model (Model 1, depth) and Model 3 (clicking + minFromSurf) indicated a failure to classify these behaviors both with and without the sixth state: in all four models, the coefficient for the probability of surfacing (T.beta11) and precision for depth during states 1 and 5 were divergent. Similarly, it appeared that ODBA alone could not discriminate between active and non-active behaviors near surface, such as shallow resting, unless the model was limited to

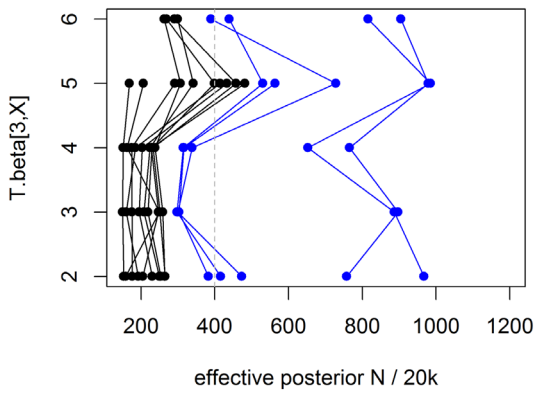

Fig. B3. Effective posterior sample size (x-axis) for transition probabilities from state 3 to states 2–6 (y-axis) in fixed-step length (FS) models (black lines) versus time-varying step length (TS) models (blue). Effective posterior sample sizes are given per 20 k iterations; for effective size analysis, posterior samples were thinned by a factor of 50 for FS models and by a factor of 54 for TS models.

five states and the duration of submergence was accounted for (minFromSurf). For both five- and six-state Model 7, state 1 and 5 parameters for depth precision (tau) and state 5 and 6 parameters for ODBA (a.mean, a.rate) diverged.

In the 10 updated FS models, the Brooks-Gelman-Ruben (BGR) estimates were less than or equal to 1.05 for all parameters and state proportions at 16–36k. However, in seven of the 10 FS models, the upper 95% confidence intervals of BGR estimates exceeded the threshold (max 1.14) (Fig. B2). Most of the values (15/20) that were over the threshold were related to transition probabilities from state 3 to states 2–6 that were also more serially correlated (i.e., mixed slower) than other parameters, indicating that within-chain correlation increased uncertainty in parameter convergence.

While a sufficient number of effective posterior samples (>>400) could be achieved for all other parameters at 16–36k iterations, transition probabilities from state 3 had effective sample sizes less than 400 in all FS models (min 150). We did not find any improvement in the chains when minFromSurf was included as a covariate for the probability of staying in state 3, indicating the slow mixing of state 3 transition probabilities was not due to inaccuracies of the 1st order Markov assumption. Nevertheless, even the most serially correlated posteriors appeared stationary (no trend) across the iterations. We therefore used a wide sample window for all posterior summary statistics (16–36k) to improve accuracy. However, the converged set of time-varying step length models appeared to improve within-chain correlation compared to FS models, at least for the transition probabilities from state 3 (Fig. B3).

Appendix C

Hidden state model selection

Methods.—For measuring goodness of fit relative to model complexity, we used deviance information criterion (DIC). The DIC is an extension of Akaike's Information Criterion (AIC) and is particularly useful for models that have been fitted outside of a maximum likelihood (ML) framework. Similarly to the AIC, the DIC is based upon both model fit and model complexity (Spiegelhalter et al. 2002, Lunn et al. 2013):

$$\mathrm{DIC} = \overline{D} + p_D \tag{C.1}$$

where $\overline{D}$ is the posterior mean deviance of the model, and p_D is the effective number of model parameters. Jags calculates the deviance as the sum of the deviances of all observed random variables defined in the model (i.e., "stochastic nodes" in BUGS terminology), and p_D as the difference between the expected deviance $\overline{D}$ and the deviance evaluated at the posterior means ($D(\overline{\theta})$; Spiegelhalter et al. 2002). However, p_D cannot be evaluated for discrete parameters such as hidden states (Lunn et al. 2013). We used an alternative measure of effective number of parameters p_v instead, which is invariant to reparameterization but assumes that the information in the likelihood dominates that of the prior (Gelman et al. 2003, Lunn et al. 2013):

$$p_v = \mathrm{var}(D)/2. \tag{C.2}$$

We also assessed how well the posterior state-dependent likelihoods fitted to data. The joint probability density of data ("emission probability") and probability of state transitions were calculated for each model given the posterior parameter samples. The emission and transition probabilities were calculated for each data point as per model specification, but ignoring prior distributions. Transition probability matrix was updated at each time step to incorporate the linear predictor with data on depth and time since surfacing (minFromSurf). The "emission only" prediction was calculated by selecting the state that maximized the sum of the log-likelihoods for data (i.e., in the full model, the likelihood for depth, clicking, pitch and ODBA). The "Viterbi sequence" accounted for both emission and transition probabilities by calculating the likelihood for the entire sequence using the Viterbi algorithm (Viterbi 1967, see Supplement script). The predicted states were then compared to the posterior state estimates to assess the contribution of state-dependent likelihoods vs. transition probabilities in the state classification. The two estimates are expected to differ because the posterior state-dependent likelihoods (data) should not always support

Table C1. Estimated time budget for each fixed step (FS) and time-varying step (TS) length model.

Model	Percentage of time in each state					
	1	2	3	4	5	6
FS 2(5)	19.95	15.83	37.84	14.77	11.61	
FS 4(5)	19.05	20.17	35.02	14.27	11.5	
FS 5(5)	19.14	21.2	34.38	18.72	6.56	
FS 6(5)	19.15	18.34	36.55	19.38	6.58	
FS 8(5)	19.13	20.89	34.58	18.84	6.56	
FS 9(5)	19.16	18.74	36.47	19.06	6.58	
FS 5(6)	19	19.27	35.24	13.19	6.46	6.83
FS 6(6)	18.97	15.66	37.74	14.23	6.5	6.9
FS 8(6)	18.99	18.75	35.5	13.42	6.46	6.88
FS 9(6)	18.99	15.61	37.79	14.15	6.5	6.96
TS 9(5)	19.1	15.8	40.7	17.8	6.6	
TS 5(6)	19	13.5	43.6	10.6	6.5	6.9
TS 6(6)	18.9	14	41.9	12	6.5	6.7
TS 8(6)	18.9	13.8	43.1	10.8	6.5	6.9
TS 9(6)	18.9	13.8	42.1	11.9	6.5	6.7

Note: Blank cells are unavailable state 6 estimates in five-state models.

the expected states based on the sequence of states, e.g., a data point resembling drifting (state 5) in the middle of an estimated layer-restricted search phase.

To measure the contribution of data in a given state classification, we re-calculated the most likely states based on a sub-set of emission probabilities from the full model. The full model was chosen in order to compare the contributions of all data streams. The predicted states based on emission probabilities were compared to the model's state estimates, and the percentage of correct predictions for each state was contrasted across the sub-sets.

Layer-restricted search (LRS) state estimates were compared to pre-classified bottom phases and drifting state and silent active state estimates to expert classification of dives (Table C4) to assess their concordance to existing methods of behavioral classification. Unlike LRS state, bottom phases were limited to a single phase within each dive that started and ended with changes in descend and ascend pitch. Therefore a higher sensitivity of LRS state to bottom phases could have also indicated a classification that was less sensitive to multi-layered dives.

Table C2. Pair-wise similarity in state classification between the converged set of models

Model no.	Fixed step length models										Variable step length models				
	2(5)	4(5)	5(5)	6(5)	8(5)	9(5)	5(6)	6(6)	8(6)	9(6)	9(5)	5(6)	6(6)	8(6)	9(6)
2(5)	**37.8**	92.3	86.9	91.1	87.1	91.0	86.5	91.3	87.0	90.8	86.2	84.0	85.4	84.3	85.2
4(5)	87.1	**35.0**	93.6	89.4	93.2	89.7	92.4	87.8	91.9	87.9	85.3	82.4	83.3	82.5	83.2
5(5)	86.8	97.4	**34.4**	94.3	99.0	94.4	91.7	87.6	91.3	87.6	90.1	81.9	83.0	82.1	82.9
6(5)	96.3	87.2	87.9	**36.6**	94.5	99.1	87.7	91.7	88.1	91.4	93.1	84.5	86.0	84.7	85.9
8(5)	87.3	96.5	97.8	88.6	**34.6**	94.3	91.5	87.8	91.6	87.6	90.3	82.1	83.2	82.3	83.1
9(5)	95.9	87.9	88.5	98.0	88.5	**36.5**	88.1	91.6	88.1	91.6	92.6	84.4	85.9	84.6	85.8
5(6)	87.2	98.6	96.3	87.0	95.9	87.9	**35.2**	94.5	99.0	94.6	83.7	89.2	89.7	89.3	89.6
6(6)	98.5	87.3	86.8	96.3	87.4	96.1	87.6	**37.7**	95.0	99.2	87.0	91.9	93.2	92.1	93.0
8(6)	88.2	97.2	95.5	87.7	96.5	87.7	97.6	88.5	**35.5**	94.8	83.9	89.3	90.0	89.6	89.8
9(6)	97.3	87.7	86.8	95.7	87.1	96.1	88.1	98.1	88.3	**37.8**	87.0	91.9	93.2	92.1	93.1
9(5)	85.0	78.0	77.6	84.5	78.1	84.0	78.0	85.4	78.4	85.6	**40.7**	89.4	91.6	89.7	91.5
5(6)	82.4	76.2	74.8	80.6	75.4	80.4	76.4	82.8	76.8	82.9	91.5	**43.6**	97.0	99.3	97.2
6(6)	84.8	77.3	76.2	83.0	76.8	82.8	77.5	85.2	78.2	85.3	95.9	94.0	**41.9**	97.3	99.4
8(6)	83.0	76.3	75.0	80.9	75.6	80.7	76.5	83.2	77.1	83.2	92.0	98.5	94.8	**43.1**	97.3
9(6)	84.3	77.3	76.1	82.9	76.6	82.7	77.5	84.9	77.9	85.2	95.7	94.6	98.7	94.7	**42.1**

Notes: The number of model structures is followed by the number of states in parentheses. Percentages of the time series of state estimates that agreed between the two models are given above the diagonal. The diagonal (values in boldface) shows the total proportion of states estimated as LRS state for each model. Percentages of LRS state estimates (one or both models estimated LRS state) that agreed between the two models are given below the diagonal.

Table C3. State uncertainty (probability that the estimated state was not the true state) in each model.

Model	Data average and 95% quantile probability (×100)					
	1	2	3	4	5	6
FS 2(5)	0.8	5.65	5.08	5.62	2.95	
	(2.26)	(36.1)	(32.67)	(36.2)	(19.23)	
FS 4(5)	0.58	6.52	5.38	5.56	1.88	
	(0.75)	(40.32)	(33.75)	(35.67)	(11.36)	
FS 5(5)	0.26	6.24	5.67	5.5	0.6	
	(0.08)	(39.75)	(34.5)	(35.88)	(0.05)	
FS 6(5)	0.21	5.66	6	4.9	0.36	
	(0.17)	(35.25)	(36.33)	(33.53)	(0)	
FS 8(5)	0.26	6.53	5.88	5.36	0.65	
	(0.08)	(40.03)	(34.42)	(36.97)	(0.08)	
FS 9(5)	0.21	6.13	5.59	4.9	0.38	
	(0.17)	(38.92)	(34.25)	(32.78)	(0)	
FS 5(6)	0.48	6.83	5.37	5.53	0.47	5.83
	(0.92)	(42.61)	(33.39)	(35.8)	(0.17)	(34.59)
FS 6(6)	0.51	5.97	4.8	5.89	0.33	5.44
	(1.42)	(37.4)	(30.92)	(38.93)	(0.08)	(32.92)
FS 8(6)	0.49	6.6	5.52	5.89	0.48	5.94
	(1)	(40.13)	(34.08)	(40.35)	(0.17)	(35.56)
FS 9(6)	0.55	5.65	4.69	5.9	0.32	5.73
	(1.42)	(37.06)	(30.83)	(39.37)	(0.08)	(34.16)
TS 9(5)	0.33	5.83	4.15	5.32	0.40	
	(0.17)	(37.67)	(29.66)	(34.35)	(0.02)	
TS 5(6)	0.53	5.63	3.12	5.37	0.42	6.30
	(0.90)	(35.66)	(24.33)	(33.26)	(0.15)	(34.53)
TS 6(6)	0.46	6.03	3.42	5.49	0.35	6.01
	(1.19)	(37.04)	(26.66)	(34.90)	(0.12)	(34.31)
TS 8(6)	0.50	6.13	3.18	5.34	0.43	6.06
	(0.85)	(36.92)	(24.34)	(35.03)	(0.20)	(33.56)
TS 9(6)	0.52	5.87	3.26	5.63	0.41	5.92
	(1.20)	(37.38)	(23.90)	(35.03)	(0.14)	(33.96)

Notes: the given probabilities were calculated as the average posterior proportion of states that were not the most prevalent state within each state estimate (i.e., accounting for the prevalence of state estimates in data, unlike Fig. C2). Blank cells are unavailable state 6 estimates in five-state models.

Measures of accuracy and diagnostic odds ratio (DOR, Glas et al. 2003) were used to compare LRS state to pre-classified bottom phases. The sensitivity of LRS state estimates to pre-classified bottom phases was calculated as the total proportion of LRS state estimates within bottom phases. The specificity of LRS state estimates was calculated as the proportion of non-LRS state estimates within the whole time-series that was not classified as outside bottom phases.

DOR combines sensitivity and specificity into one discriminatory test performance diagnostic (Glas et al. 2003):

$$\mathrm{DOR} = \frac{\text{sensitivity}}{1-\text{sensitivity}} \Big/ \frac{1-\text{specificity}}{\text{specificity}}. \quad \text{(C.3)}$$

Thus, DOR was the ratio of the odds of LRS state in a bottom phases to the odds of LRS state outside the bottom phase. The higher the value, the better the estimated state could discriminate between the human classified states. Standard errors were calculated for the logarithm of DOR that follows approximately a normal distribution (Glas et al. 2003). Sensitivity and specificity of the LRS state classification to bottom phases were calculated based on the proportion of bottom phases ('true conditions') and intervals between bottom phases ('false conditions') that were estimated as LRS state. The number of bottom phases were therefore accounted for in SE(log(-DOR)).

Results.—Six-state models had both lower posterior mean deviance and DIC than their respective five-state model structures. Conversely, posterior mean deviances were higher for models with ODBA despite increased model complexity (Table 3). The effective number of parameters was estimated small for models with smaller deviance (model structures 6 and 9) and higher for models with higher deviance (model structures 4, 5 and 8) both by p_D and p_v (Table 3,

Table C4. Expert dive classification.

Category 1	Category 2	Dive no.	Description
A) Dives with clicking	Usual dive profile	1	Shallow (<300 m in max depth)
		2	Mid-depth dives
		3	Deep dives (>1000 m in max depth)
		4	Multi-layer dives; whale spends time at several depth layers, or there is an excursion of more than about 2x the depth extent of the main layer
	Unusual dive profile	5	Clicking but not buzzes, unusual shape with few inflections and smoother shape in dive profile
		6	Some clicking on descent, but clicking ceases and whale drifts during ascent
		7	Some clicking on descent, but clicking ceases and whale swims actively during ascent
B) Dives without clicking	Pitch and ODBA indicate drifting behavior	8	Shallow (<20 m)
		9	Deep (>20 m)
	Silent active swimming	10	Shallow (<20 m)
		11	Deep (>20 m)

Notes: Dive types were divided into two main categories based on the presence of echolocation clicks. Sub-categories were determined by the shape of the dive profile and active swimming versus drifting.

Fig. C1). Therefore, deviance and DIC arrived at a similar ranking of models.

Time series of state estimates were calculated for each model as the most prevalent state in the posterior sample at iterations 16–36k for fixed step length (FS) models and samples 24–48k for time-varying step length (TS) models. The state estimates from all the 15 models agreed on 77.9% of data, yielding similar time budgets (Tables C1 and C2). Six-state model classifications were more consistent within FS models (93.9%) and within TS models (93.9%) than across (86.7%). TS models estimated the highest proportions of data as LRS state than any FS model (>40%; Table C2).

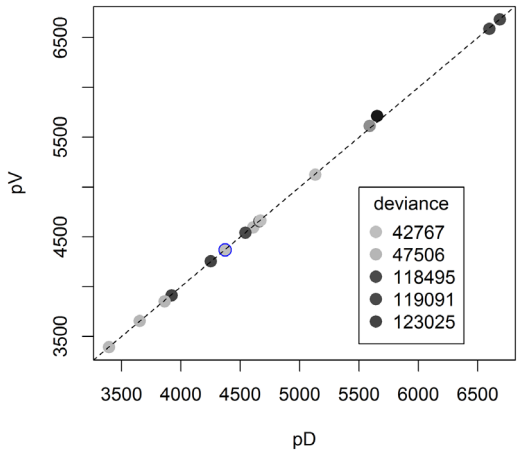

Fig. C1. Comparison of the two estimates for the effective number of parameters for each converged model (n = 15): *pD* (returned by jags, difference between the posterior mean and the deviance evaluated at posterior means) and *pV* (variance of posterior deviance divided by two). The lowest deviance model is circled in blue.

'Overall state uncertainty' was designed to measure the average residual or overall lack of support for the estimated sequence of states. Overall state uncertainty ranged between 3.3-4.5% of samples across all models. Allowing for step length to increase with depth improved the overall state certainty in all converged model structures (Fig. C2). TS models 5 (full model) and 8 (pitch + ODBA) had the lowest overall state uncertainty (3.25% and 3.33%).

Based on emission probability of data alone (depth, clicking, pitch, ODBA), the models' discriminatory power broadly mirrored that of their overall state uncertainty (Figs. C2 and C3), indicating that any lack of support for the most prevalent states was driven by the state-dependent likelihoods. Emission probabilities predicted a posterior average of 89.1–93.45% of state estimates across models. When accounting for transition probabilities (Viterbi algorithm), the models' ability to discriminate states was improved and less variable between models (97.8–

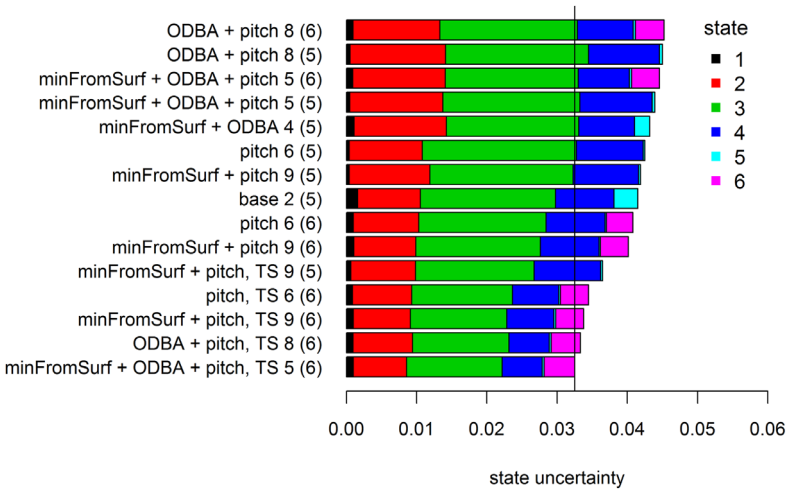

Fig. C2. Overall state uncertainty for each model (total bar width) with contributing states (states 1–6 color-coded from left to right). Overall state uncertainty was calculated for each model as the total proportion of posterior samples that were not the most prevalent state. Models are shown in ascending order of overall uncertainty with the lowest (best) values at the bottom. Model structure numbers are given before number of states in brackets. 'Base' structure here refers to depth + clicking that were included in all of the converged set of models; 'TS' refers to time-varying step length models.

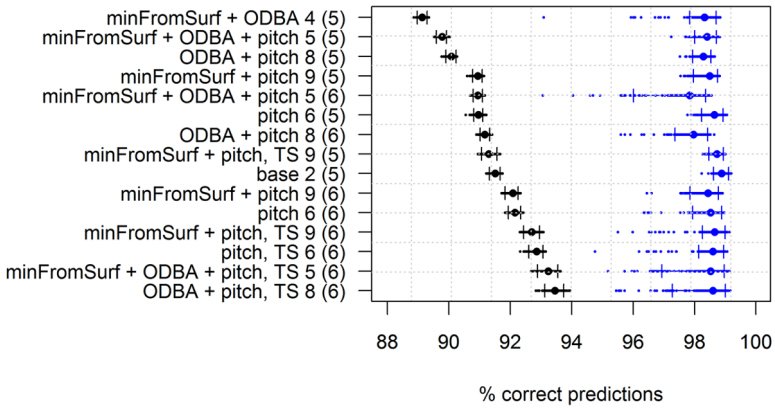

Fig. C3. Percentage of state estimates predicted correctly when the predicted state was based on emission probability alone (black) and on both emission and transition probabilities (Viterbi algorithm, blue). Percentages for each posterior sample (small dots), mean percentage (large dots) and 95% quantiles (intervals) are shown. Models are shown in ascending order of average percentage of correct predictions based on emission alone, with the highest (best) values at the bottom. Emission probabilities were calculated as the sum of the state-dependent posterior log densities for each data stream. The predicted state in a time step, given an emission, was found by maximising the emission probability across the states. Transition probabilities were calculated based on the posterior transition matrix and linear predictor at each time step for each model ('TS': time-varying time step length models). The most likely path was found by Viterbi algorithm that minimizes both the emission and transition probabilities across sequences of states. See Supplement for r script.

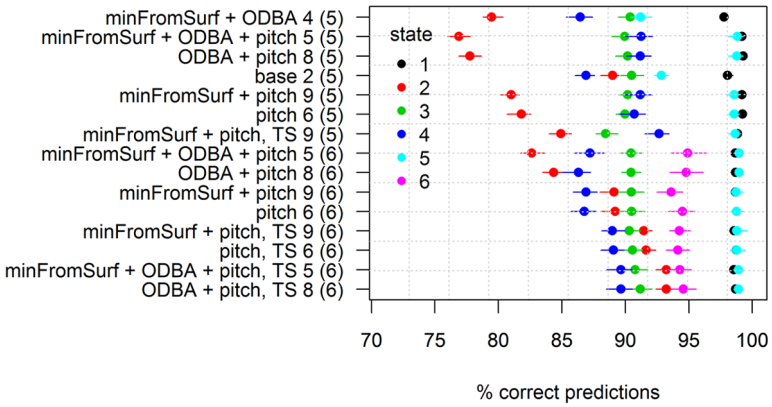

Fig. C4. The average (solid circle) and 95 quantile for the percentage of correct predictions for each state estimate when predicted by emission probability alone. Emission probabilities were calculated as the sum of the state-dependent posterior log densities for each data stream. The predicted state in a time step, given an emission, was found by maximising the emission probability across the states. 'TS' refers to time-varying step length models.

98.9%). Viterbi algorithm improved the state predictions only 7.0% on average, highlighting how variable and relatively little the (mostly) Markov state-transitions contributed to the state classification (Fig. C3).

Surface and drifting states had the lowest average state uncertainties (0.44% and 0.70%) and descent and ascent states the highest across all models (6.08% and 5.48%) (Table C3, Fig. C4). Silent active state had similarly high average uncertainty both within FS and TS models (5.74% and 6.07%, Table C3). Although the emission probabilities predicted silent active state better than the foraging states 2–4 (Fig. C4), silent active state was predicted worse than any other state when accounting for transition probabilities (Fig. C5). However, excluding the sixth state from the model appeared to decrease the contribution of state-dependent likelihoods in descent state estimation (Fig. C4). The state-dependent likelihoods for TS models were further better able to discriminate descent and ascent states than FS models (Fig. C4). The overall lower state uncertainty of TS models therefore appeared to be driven by the foraging states (descent, LRS and ascent).

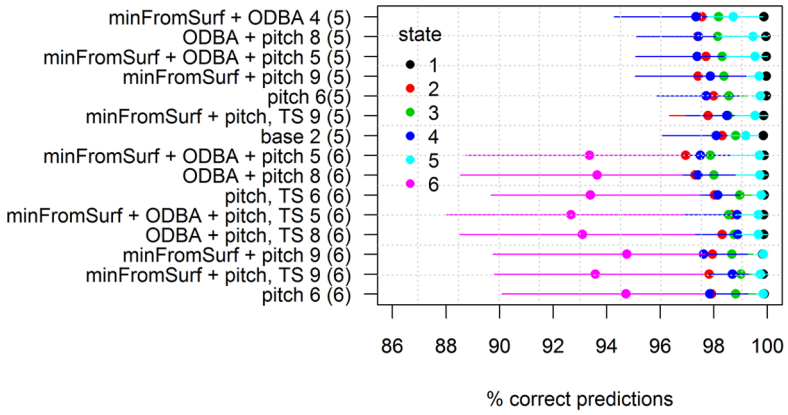

Fig. C5. The average (solid circle) and 95 quantile for the percentage of correct predictions for each state estimate when predicted by the Viterbi algorithm. 'TS' refers to time-varying step length models.

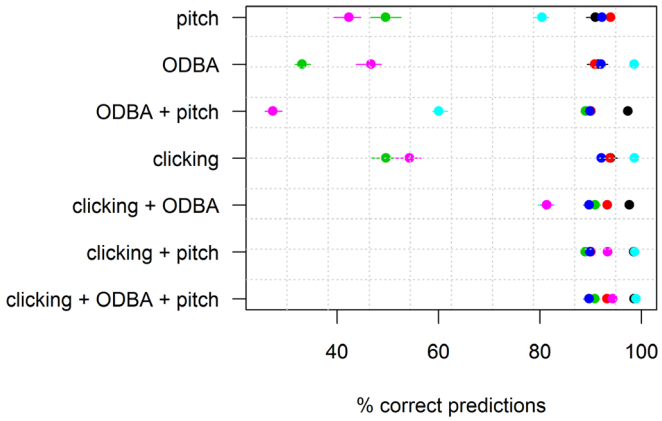

Fig. C6. The average (solid circle) and 95 quantile for the percentage of correct predictions for each state estimate when predicted by a sub-set of state dependent likelihoods for each data (y-axis) in the full time-varying step length (TS) model 5 with six states.

Contribution of data was measured for the state classification of the full TS model structure 5 with six states. Compared to the full set of likelihoods, the percentage of correct predictions decreased most for LRS state and silent active state when clicking was excluded from the predictions. In contrast, removing ODBA changed the percentage of correct predictions least (Fig. C6).

There were only small differences in the estimated time budgets for expert classified (Table C4) drifting dives between the models (Fig. C7). All models estimated drifting dives to consist more than an average of 76% (76.3–82.9%) of time in drifting state, and at least an average of 13% (12.9–18.1%) of time in state 2 (descending). Five-state models with pitch estimated drifting dives to also contain ascending (3.9–4.4.6%) while six-state models estimated silent active (silent active state, 7.0–7.9%) (Fig. C7).

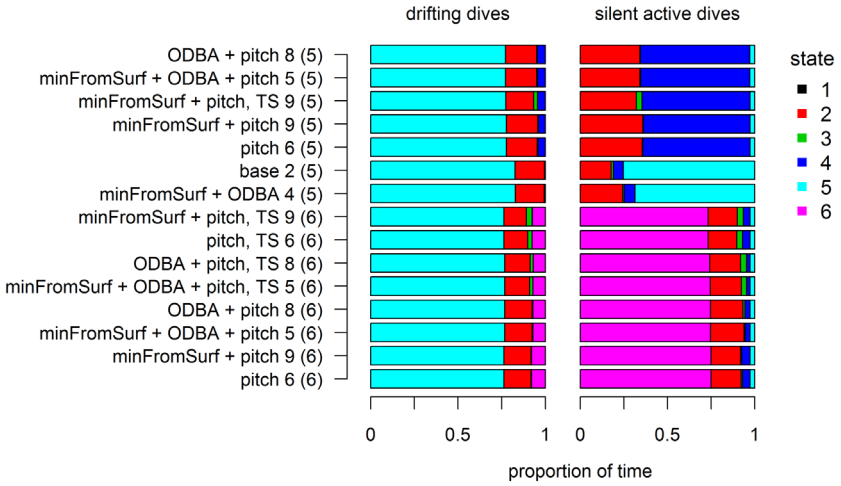

Fig. C7. Comparing state estimates of the converged set of models to expert classified drifting dives and silent active dives. Bar plots show time budget of state estimates (color-coded) within each type of dive. Models are shown in ascending order of state proportion for silent active dives, then by drifting dives, with the highest values at the bottom. 'TS' refers to time-varying step length models.

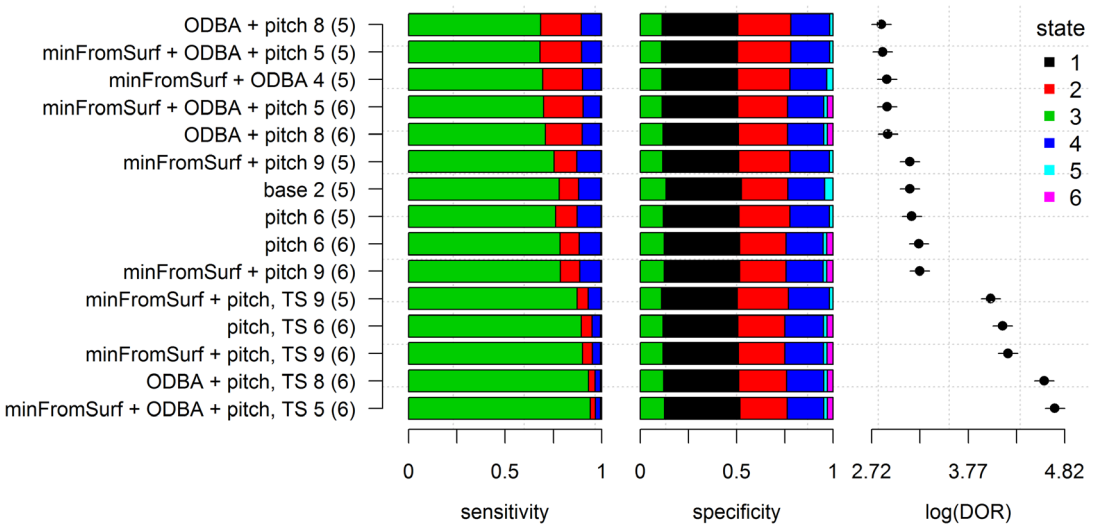

Fig. C8. Comparing LRS state estimates of the converged set of models (y-axis) to pre-classified bottom phases. Bar plots show proportion of state estimates (color-coded) in pre-classified bottom phases (left) and intervals between pre-classified bottom phases (middle). Green bar widths therefore represent sensitivity and specificity of LRS state estimates to the bottom phases. Logarithm of the diagnostic odds ratios (DOR) are shown in the far right figure, with 95% Gaussian confidence intervals. Sample size for standard errors were based on number of bottom phases (n = 171). Models are shown in ascending order of DOR with the highest (best) values at the bottom. Model structure numbers are given before state number in brackets. 'Base' structure here refers to depth + clicking that were included in all of the converged set of models; 'TS' refers to time-varying step length models.

Expert classified silent active swimming dives had more variable time budgets across models than drifting dives. Five-state models without pitch estimated these dives to consist mostly of drifting state (Model 2 average: 75.2%, Model 4 average: 68.7%) while five-state models with pitch estimated these dives to consist mostly of ascend (state 4 averages 61.2–62.6%) (Fig. C7). Six-state classifications were more consistent, with ~75% in silent active state and ~3% in drifting state.

LRS state estimates of model structure 6 and 9 with six states were most sensitive to the pre-classified bottom phases (0.79 and 0.78, respectively; Fig. C8). With little differences in specificity between the models, also the diagnostic odds ratio (DOR) selected for these two models as the best match for pre-classified bottom phases. In terms of the 95% confidence intervals for DOR, all FS models appeared to be significantly poorer classifiers of bottom phases than TS models (Fig. C8).

APPENDIX D

Posterior estimates from the selected hidden state model

Table D1. Posterior distribution estimates for transition probabilities.

				Quantiles (%)				
Parameter[state]	Mean	SD	SE	2.50	25.00	50.00	75.00	97.50
T.beta0[1, 1]	2.53	0.62	11.65	1.38	2.10	2.49	2.92	3.76
T.beta0[1, 2]	−2.34	0.13	2.32	−2.63	−2.43	−2.34	−2.25	−2.10
T.beta0[1, 3]	−7.90	1.28	20.27	−11.17	−8.54	−7.68	−7.00	−6.08
T.beta0[1, 4]	−14.42	1.31	22.75	−17.50	−15.11	−14.20	−13.51	−12.55
T.beta0[1, 5]	−5.93	0.47	7.43	−6.98	−6.21	−5.90	−5.61	−5.11
T.beta0[1, 6]	−2.85	0.15	2.60	−3.16	−2.95	−2.85	−2.76	−2.58
T.beta0[2, 1]	−14.39	1.30	20.53	−17.58	−15.05	−14.20	−13.46	−12.50
T.beta0[2, 2]	2.18	1.36	29.13	−0.02	1.25	2.02	2.90	5.39
T.beta0[2, 3]	−1.37	0.25	6.72	−2.00	−1.48	−1.31	−1.21	−1.05
T.beta0[2, 4]	−6.11	0.95	15.62	−8.43	−6.55	−5.96	−5.45	−4.79
T.beta0[2, 5]	−6.04	0.61	11.47	−7.42	−6.40	−5.98	−5.61	−5.02
T.beta0[2, 6]	−4.20	0.31	6.14	−4.89	−4.39	−4.18	−3.99	−3.67
T.beta0[3, 1]	.33	1.27	20.48	−0.77	0.49	1.17	2.04	4.20
T.beta0[3, 2]	−4.60	0.27	7.46	−5.24	−4.75	−4.57	−4.42	−4.17
T.beta0[3, 3]	2.05	1.32	32.04	−0.14	1.12	1.91	2.83	5.09
T.beta0[3, 4]	−2.99	0.22	6.80	−3.58	−3.09	−2.93	−2.83	−2.71
T.beta0[3, 5]	−6.69	0.47	8.80	−7.73	−6.98	−6.65	−6.37	−5.89
T.beta0[3, 6]	−5.44	0.32	7.93	−6.16	−5.64	−5.41	−5.21	−4.89
T.beta0[4, 1]	2.11	0.83	15.89	0.55	1.55	2.07	2.61	3.83
T.beta0[4, 2]	−8.80	0.80	15.38	−10.66	−9.25	−8.73	−8.25	−7.46
T.beta0[4, 3]	−6.84	0.46	10.85	−7.81	−7.14	−6.81	−6.52	−6.02
T.beta0[4, 4]	−3.14	0.41	9.69	−4.01	−3.40	−3.11	−2.85	−2.43
T.beta0[4, 5]	−9.14	0.84	16.36	−11.07	−9.62	−9.05	−8.56	−7.76
T.beta0[4, 6]	−7.46	0.51	10.83	−8.51	−7.80	−7.43	−7.11	−6.54
T.beta0[5, 1]	2.27	1.36	21.48	0.04	1.36	2.13	3.02	5.44
T.beta0[5, 2]	−7.17	1.27	20.45	−10.21	−7.84	−6.98	−6.28	−5.17
T.beta0[5, 3]	−5.37	0.58	9.37	−6.63	−5.71	−5.32	−4.97	−4.38
T.beta0[5, 4]	−4.28	0.43	7.12	−5.20	−4.55	−4.24	−3.97	−3.51
T.beta0[5, 5]	2.19	1.32	20.53	0.22	1.26	1.99	2.92	5.36
T.beta0[5, 6]	−2.95	0.26	4.16	−3.49	−3.11	−2.93	−2.77	−2.49
T.beta0[6, 1]	2.35	0.70	12.30	1.09	1.85	2.32	2.80	3.86
T.beta0[6, 2]	−2.65	0.32	7.26	−3.28	−2.86	−2.65	−2.43	−2.03
T.beta0[6, 3]	−4.24	0.40	7.89	−5.08	−4.49	−4.22	−3.97	−3.47
T.beta0[6, 4]	−4.19	0.38	7.92	−4.97	−4.43	−4.17	−3.92	−3.47
T.beta0[6, 5]	−3.24	0.32	6.69	−3.88	−3.45	−3.22	−3.02	−2.62
T.beta0[6, 6]	−0.20	0.52	12.33	−1.09	−0.55	−0.24	0.10	0.96
T.beta11[1]	−8.37	1.18	21.49	−10.79	−9.13	−8.31	−7.54	−6.24
tau[1]	0.010	0.000	0.006	0.010	0.010	0.010	0.011	0.011
tau[2]	677.9	28.2	485.8	623.8	658.7	677.4	696.8	734.6
tau[3]	0.045	0.058	1.047	0.002	0.012	0.027	0.057	0.203
tau[4]	443.5	21.8	423.3	402.5	428.9	442.3	457.6	488.1
tau[5]	72.41	5.01	88.64	62.96	69.07	72.20	75.69	82.75
tau[6]	125.0	10.5	215.8	105.6	118.0	124.7	131.8	146.9
tau.beta0	1.469	0.023	0.432	1.423	1.453	1.470	1.485	1.514
drift[2]	41.61	0.83	13.70	39.98	41.06	41.62	42.17	43.18
drift[4]	46.03	0.93	16.76	44.22	45.43	46.02	46.64	47.90
RC.beta0[1]	0.002	0.001	0.026	0.000	0.001	0.002	0.003	0.005
RC.beta0[2]	0.902	0.009	0.138	0.885	0.896	0.902	0.907	0.918
RC.beta0[3]	0.997	0.001	0.017	0.995	0.996	0.997	0.998	0.999
RC.beta0[4]	0.559	0.016	0.278	0.528	0.548	0.559	0.569	0.589
RC.beta0[5]	0.012	0.005	0.088	0.005	0.008	0.011	0.014	0.022
RC.beta0[6]	0.015	0.008	0.212	0.003	0.010	0.015	0.021	0.033
p.beta0[1]	−4.198	0.026	0.413	−4.248	−4.216	−4.197	−4.180	−4.147
p.beta0[2, 3, 4, 6]	−2.458	0.013	0.211	−2.484	−2.467	−2.458	−2.449	−2.432
p.beta0[5]	2.132	0.035	0.569	2.062	2.109	2.131	2.156	2.198
p.gamma[1]	51.27	1.92	29.25	47.61	49.95	51.22	52.56	55.02

Table D1. Continued.

Parameter[state]	Mean	SD	SE	Quantiles (%)				
				2.50	25.00	50.00	75.00	97.50
p.gamma[2, 3, 4, 6]	12.27	0.21	3.40	11.87	12.13	12.27	12.41	12.68
p.gamma[5]	14.67	0.94	15.05	12.89	14.01	14.67	15.26	16.58
p.beta1[2, 3, 4, 6]	0.037	0.000	0.004	0.036	0.036	0.037	0.037	0.037

Notes: Posterior summary statistics were calculated using 'summary.mcmc' function in package 'coda' in r. Time-series standard error (SE) was based on an estimate of the spectral density at zero. Parameter names are given as in BUGS code (Supplement), with numbers in square brackets referring to the state(s) (1–6) with which the parameter is associated.

Table D2. Descriptive dive cycle (diving + post surfacing) statistics for each state estimate within expert classified dive types.

Dive type	N	Duration (min)	Max depth (m)	Percentage of dive cycle time in each state						Total percentage of dives with multiple states in each state					
				1	2	3	4	5	6	1	2	3	4	5	6
1	72	34.3	215.1	21.6	9.8	55.5	6.7	1.3	5.1	0.0	9.7	13.9	5.6	0.0	2.8
				19.1	7.7	24.6	5.3	4.2	14.1						
2	40	33.7	503.7	23.0	19.3	35.9	17.9	0.0	3.9	0.0	30.0	45.0	15.0	0.0	2.5
				8.9	9.6	13.4	6.1	0.0	11.6						
3	3	37.3	842.4	11.4	10.4	31.7	13.2	0.0	33.3	0.0	0.0	0.0	0.0	0.0	0.0
				10.1	12.4	28.3	11.5	0.0	57.7						
4	4	20.8	210.7	31.0	9.0	19.5	9.8	0.0	30.6	0.0	25.0	50.0	25.0	0.0	0.0
				33.0	13.7	25.0	11.6	0.0	47.5						
6	2	37.0	78.7	66.0	2.8	7.5	0.0	17.9	5.7	0.0	0.0	0.0	0.0	0.0	0.0
				34.6	4.0	10.7	0.0	25.3	5.4						
7	2	25.5	28.9	30.9	0.0	0.0	0.0	27.6	41.5	0.0	0.0	0.0	0.0	50.0	50.0
				5.1	0.0	0.0	0.0	39.0	44.1						
8	6	33.0	43.2	35.4	0.0	0.0	0.0	59.0	5.6	0.0	0.0	0.0	0.0	33.3	83.3
				5.3	0.0	0.0	0.0	4.8	1.4						
10	1	23.0	51.9	21.7	0.0	0.0	0.0	0.0	78.3	0.0	0.0	0.0	0.0	0.0	0.0
11	1	18.0	37.3	16.7	0.0	0.0	0.0	0.0	83.3	0.0	0.0	0.0	0.0	0.0	0.0

Notes: Expert classified dive types (rows, 1–11) are as shown in Table C4. Only data from post-tagging baseline period are included. Duration and maximum depth are means. Percentage of dive cycle time values are mean with SD below.

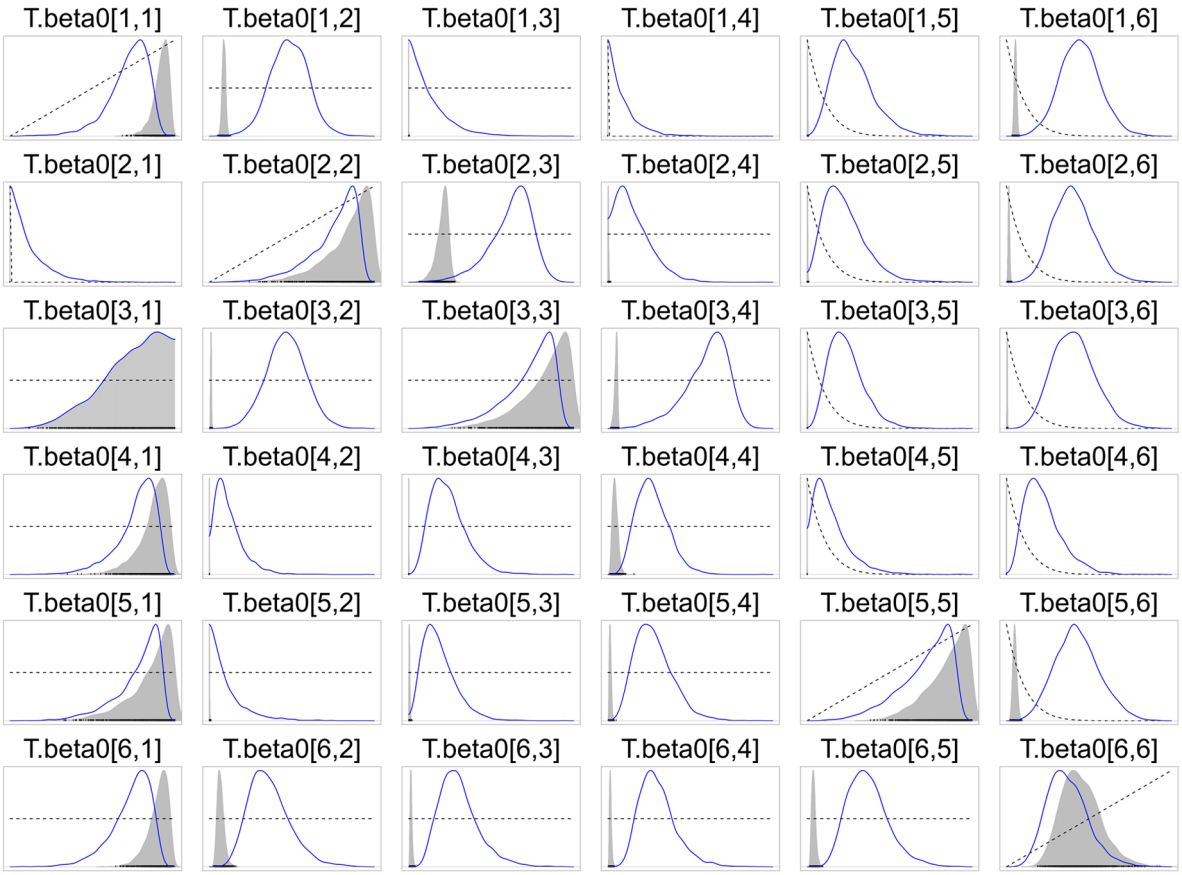

Fig. D1. Prior and posterior densities for the best model (Model 5 with 6 states and time-varying step length). Posterior densities are shown both at range [0, 1] (shaded gray) and zoomed in to the posterior range (blue). Prior densities are shown as dashed line at [0, 1] range. Thinned posterior samples (1334 out of 20k iterations) are given as rug plot at the bottom of each graph at the scale of the prior.

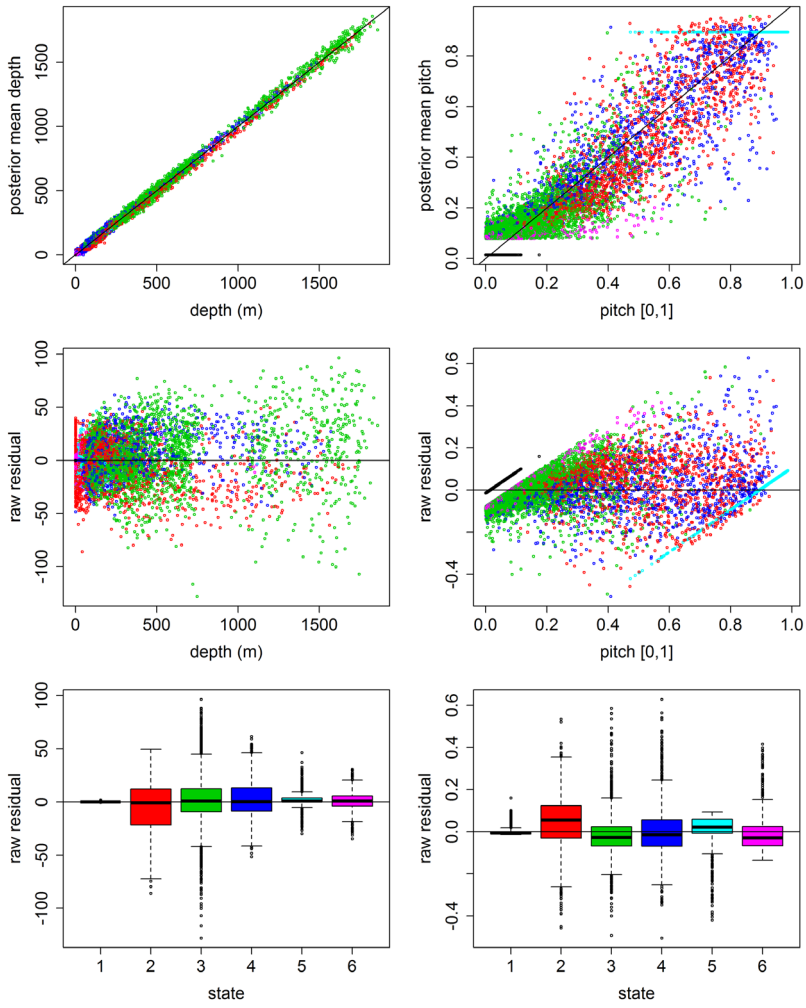

Fig. D2. Residual plots for depth and pitch in the best hidden state model. Raw residuals were calculated as posterior mean ("fitted") minus observed value for each time step. In order to obtain sign for the posterior predicted depth, we used the posterior mean of samples that were monitored for the random walk in the model (parameter 'mu' in Supplement Jags script). For pitch, posterior mean was calculated based on the posterior means of the beta regression coefficients p.beta0 and p.beta1, as well as the observed vertical step length.

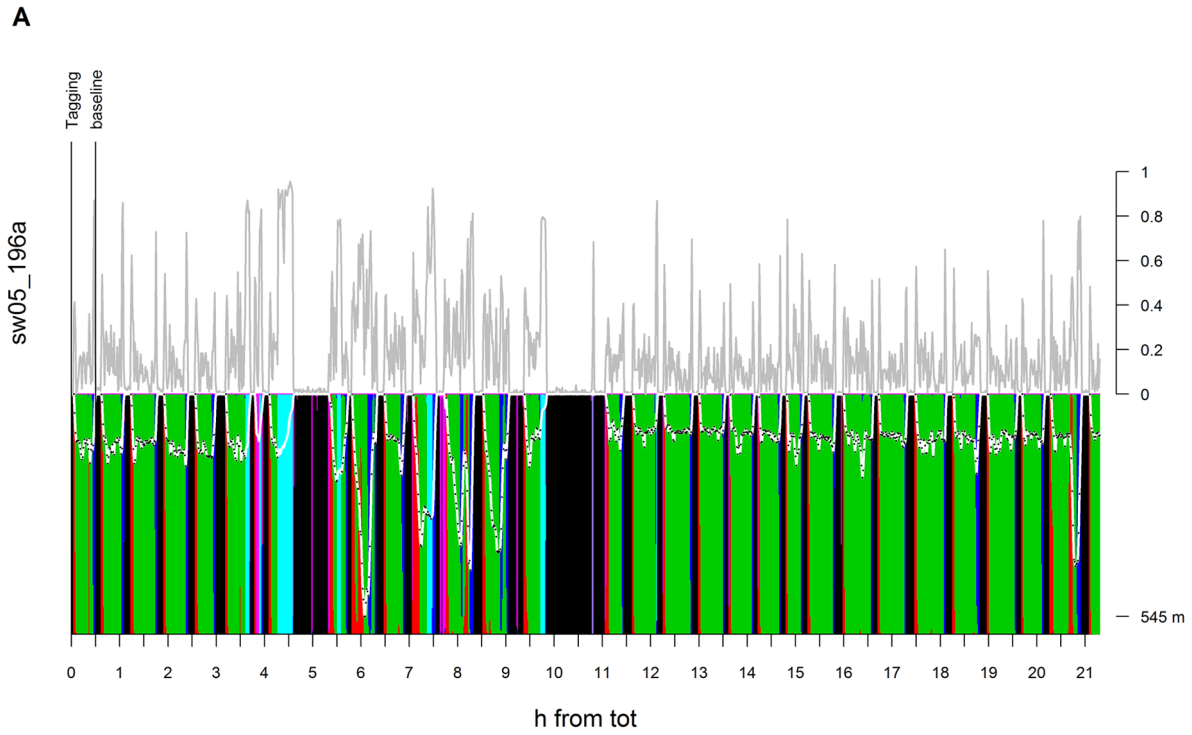

Fig. D3. Time series of posterior state probabilities for each individual (color-coded), overlaid with dive profile (white) and presence/absence of clicking (black dots) at bottom half of the graph. The top half of the graph shows absolute value for pitch scaled to 0–1 (gray), and for sonar exposures, sound exposure levels (black; dB re 1 $\mu Pa^2 s$). Sonar pings from an unidentified source are given as yellow vertical lines. Black vertical lines labelled at top x-axis show received sound exposure level (SEL).

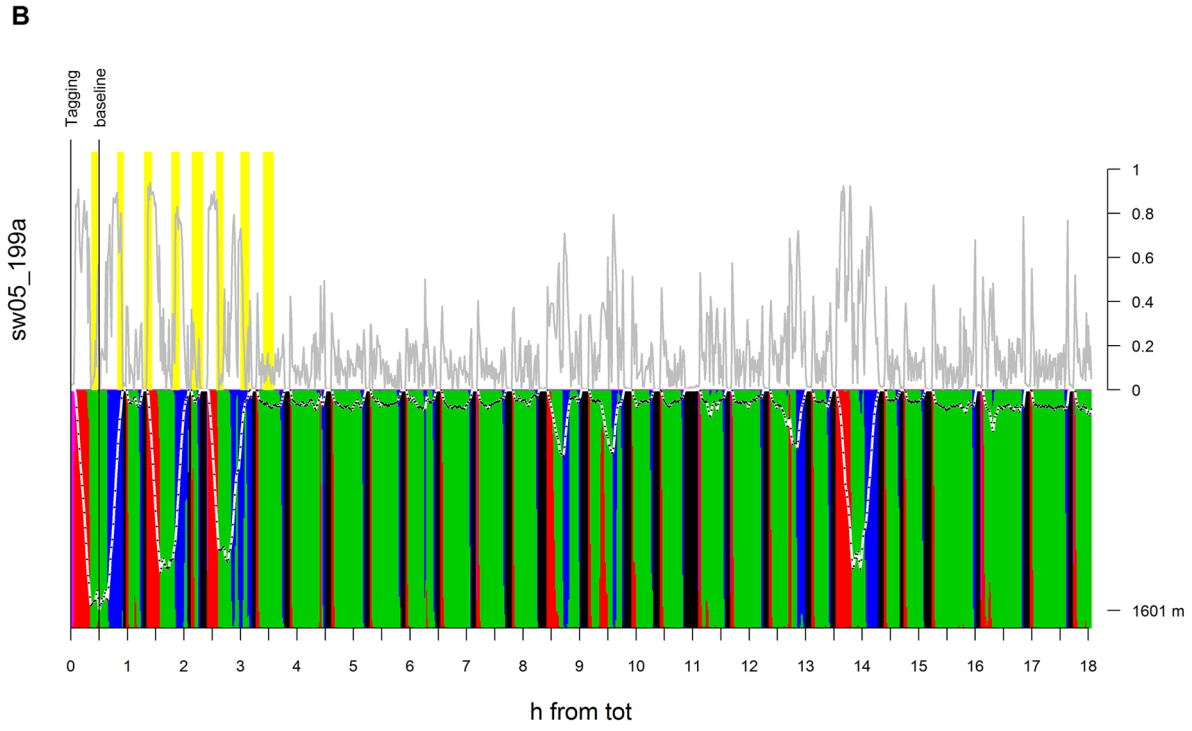

Fig. D3. Continued.

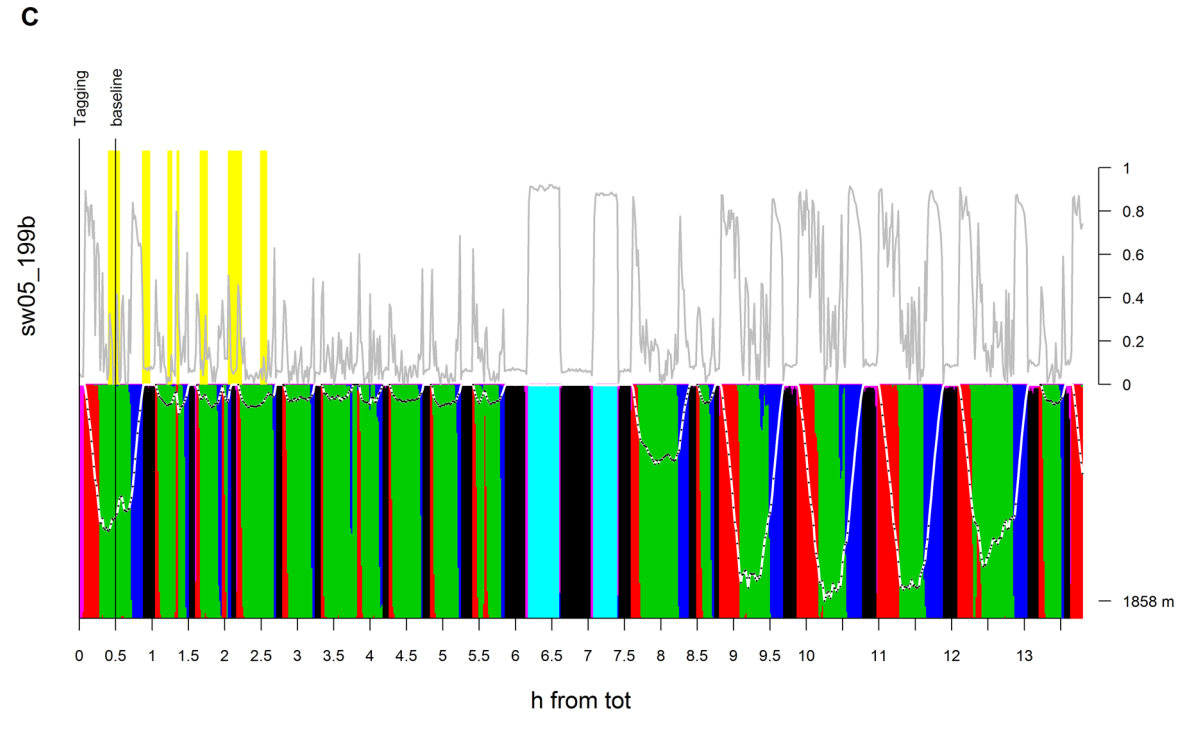

Fig. D3. Continued.

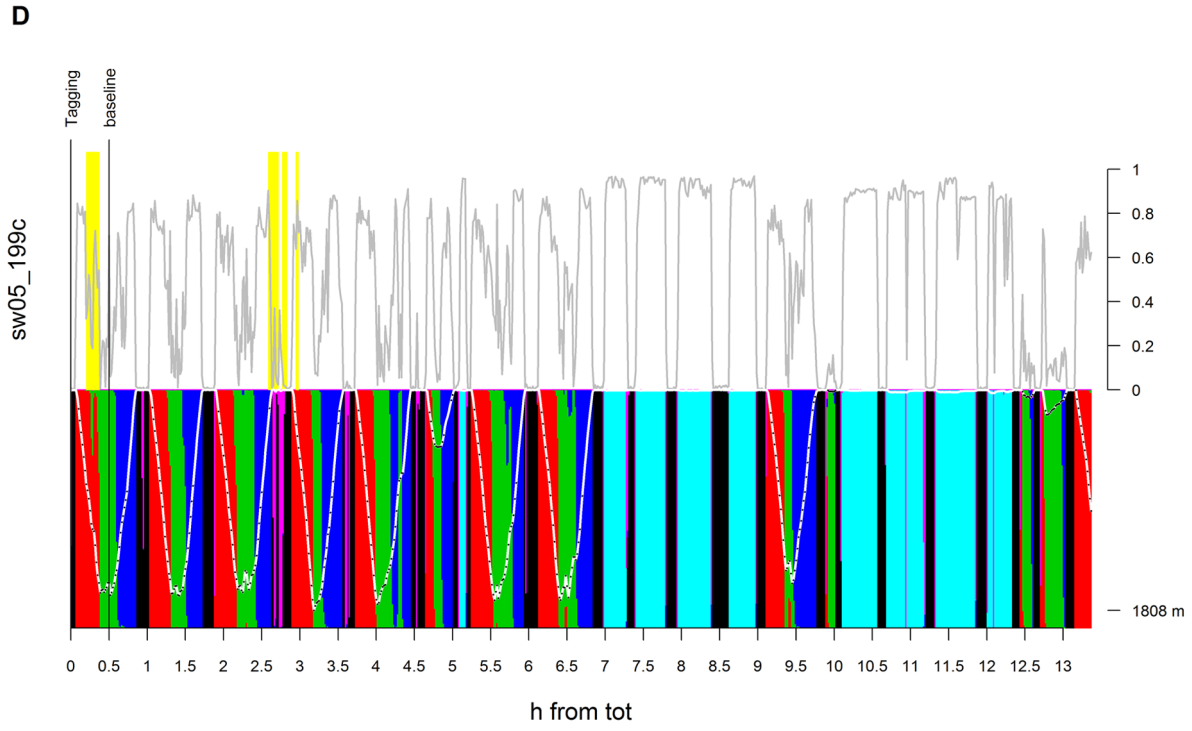

Fig. D3. Continued.

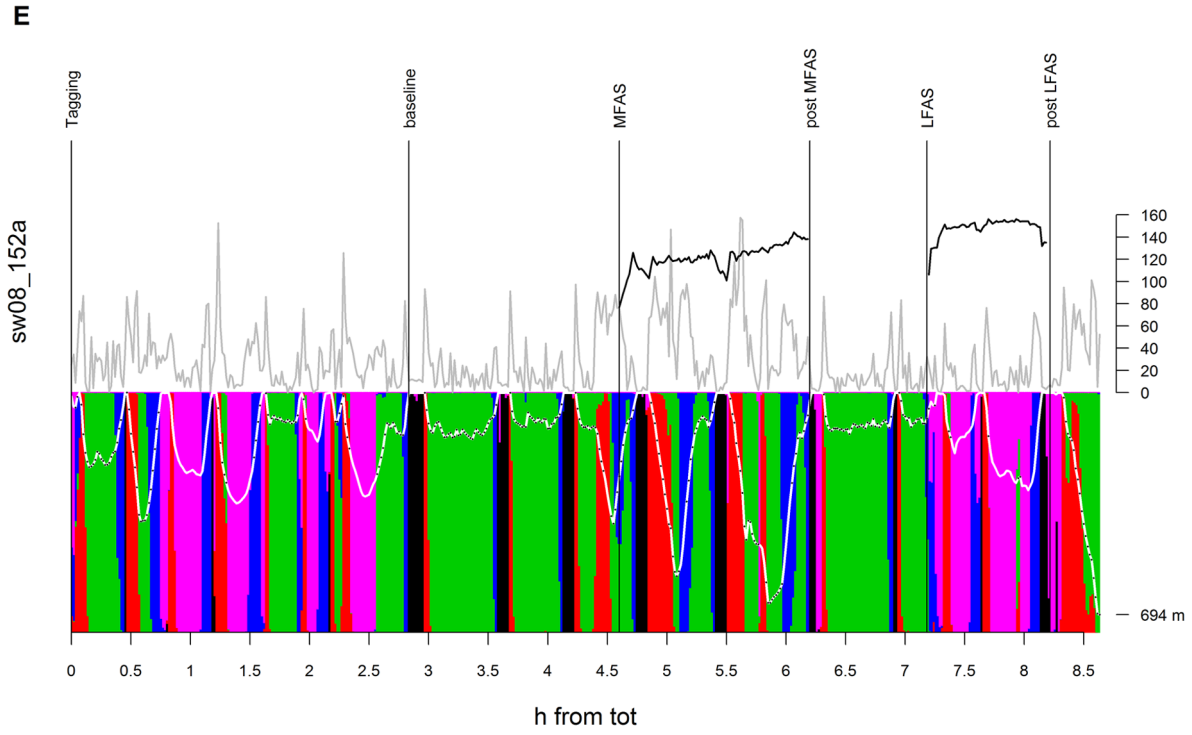

Fig. D3. Continued.

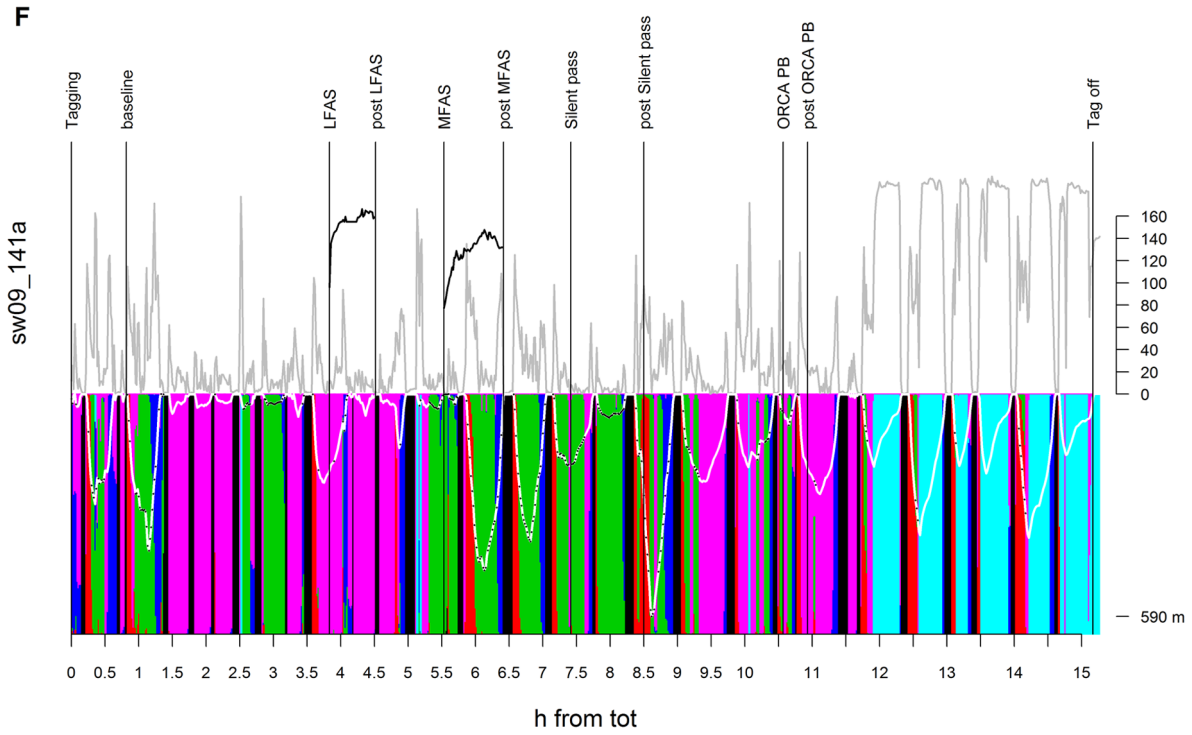

Fig. D3. Continued.

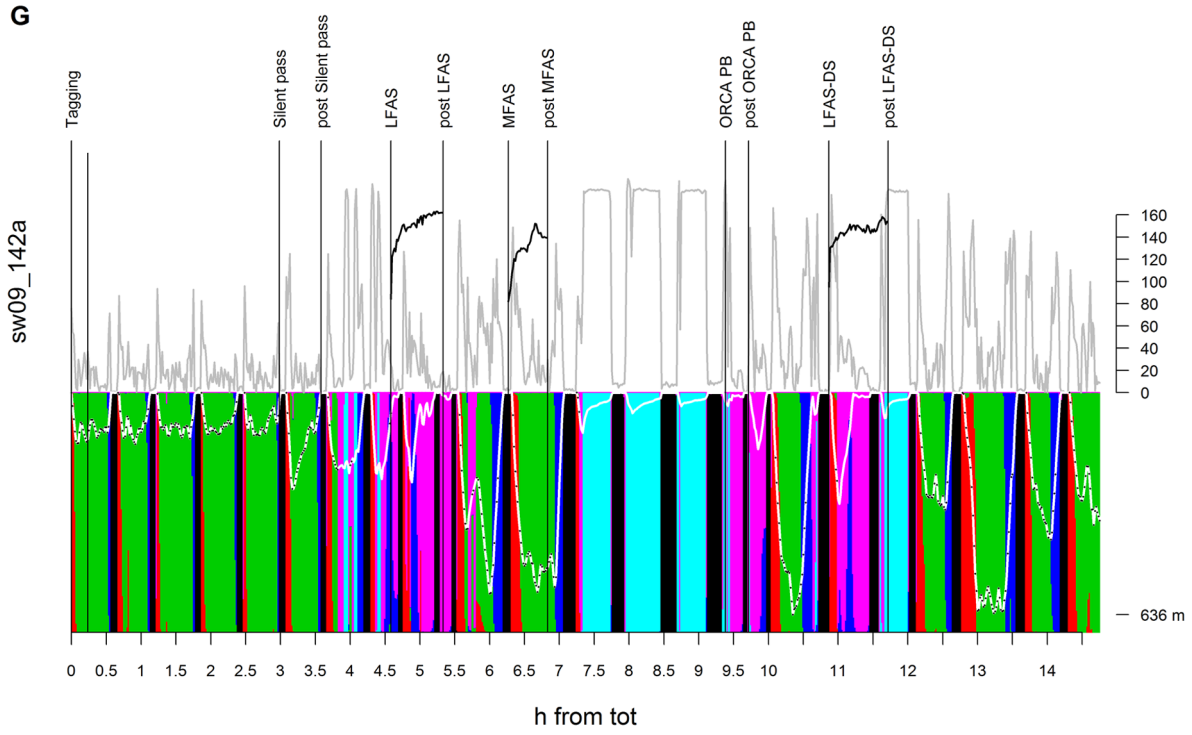

Fig. D3. Continued.

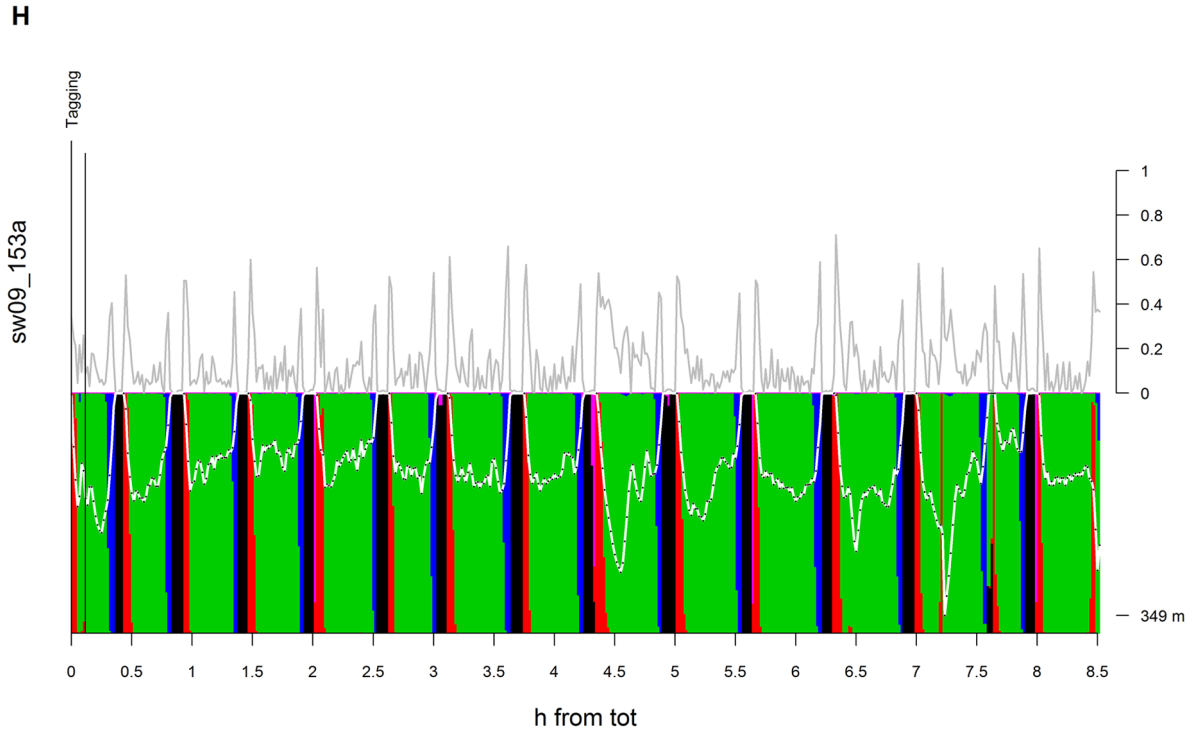

Fig. D3. Continued.

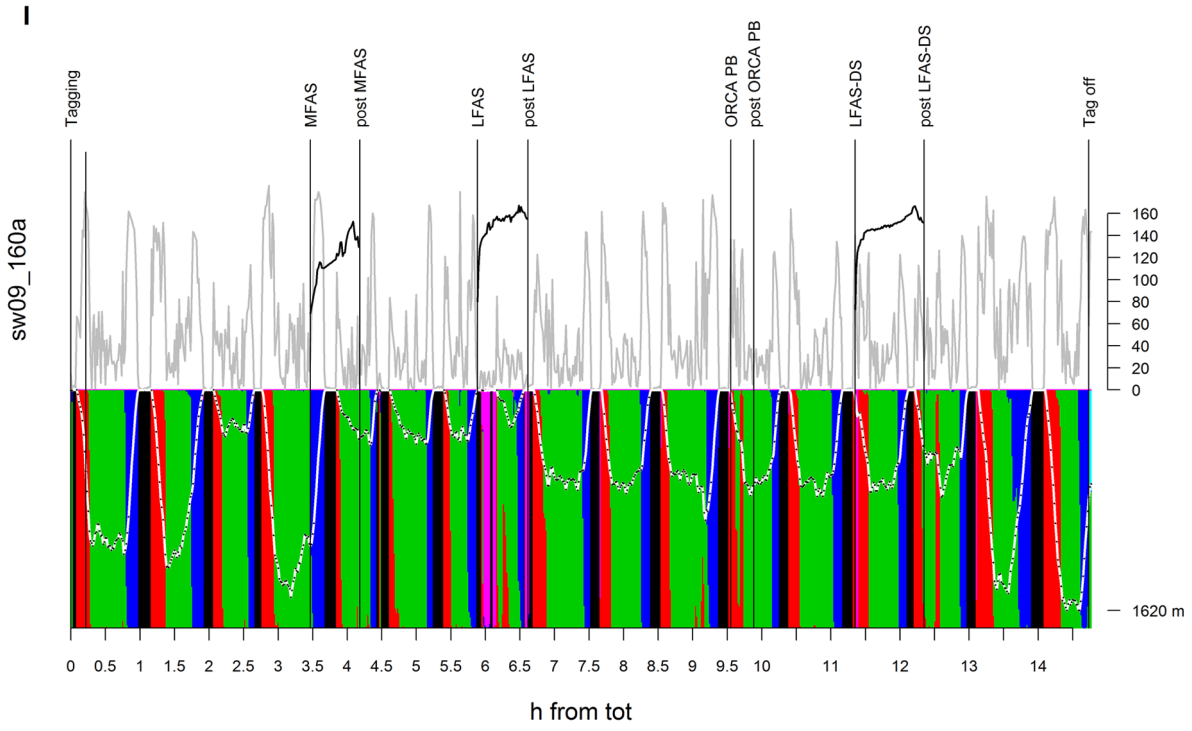

Fig. D3. Continued.

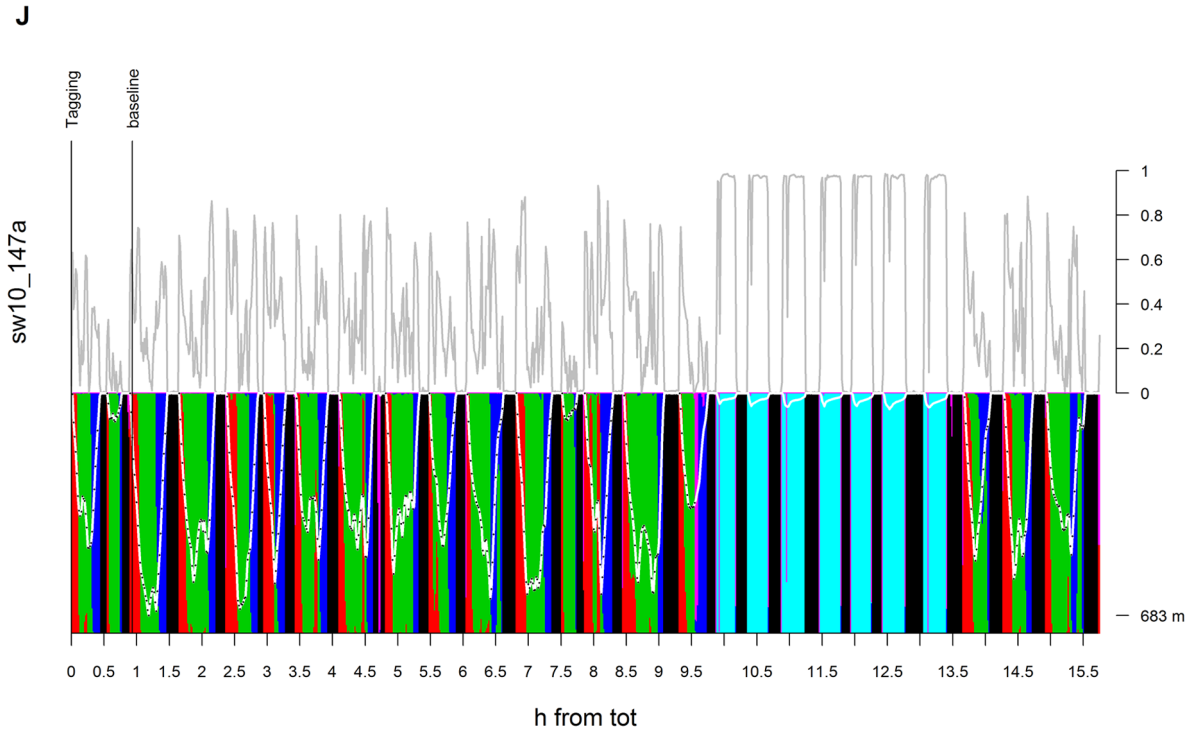

Fig. D3. Continued.

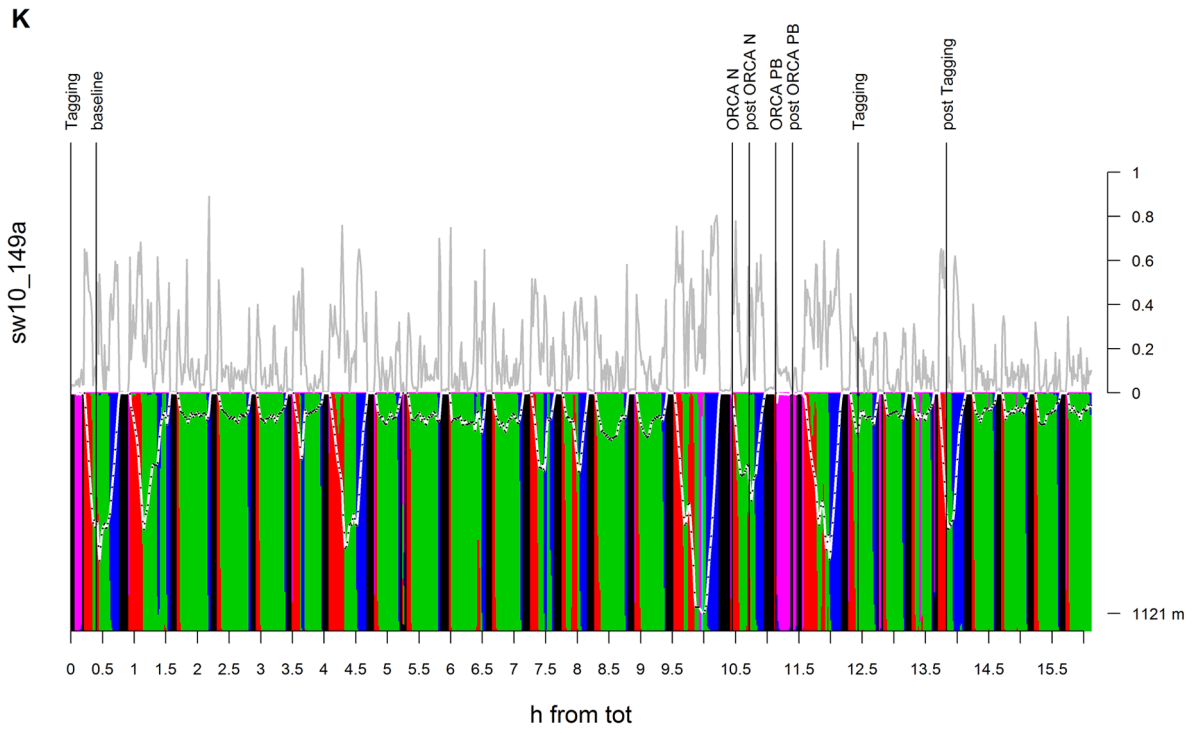

Fig. D3. Continued.

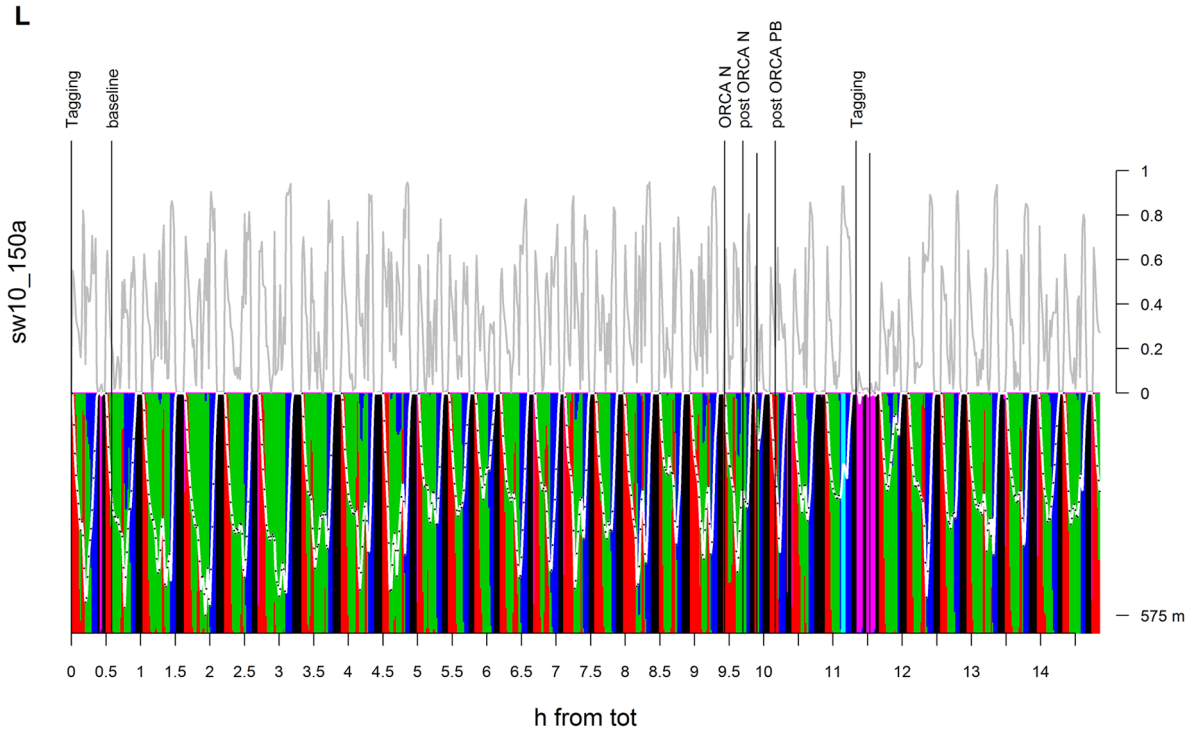

Fig. D3. Continued.

Appendix E

Tagging effects

Table E1. Individual average state duration and state depth during tagging and post-tagging.

	Tagging period					Post-tagging baseline				
		Duration (min)		Depth (m)			Duration (min)		Depth (m)	
State	N tags (states)	Mean	SD	Mean	SD	N tags (states)	Mean	SD	Mean	SD
1	6 (18)	3.2	1.8	0.22	0.35	9 (135)	6.8	2.0	0.08	0.07
2	6 (17)	3.1	1.2	136.3	87.6	9 (143)	4.6	2.5	167.0	119.9
3	6 (21)	7.8	3.5	230.2	72.1	9 (151)	16.8	6.4	283.9	234.4
4	5 (15)	4.5	1.2	111.5	45.5	9 (123)	4.3	2.3	113.0	92.2
5	0					3 (19)	5.5	5.5	49.9	43.0
6	4 (19)	4.0	1.8	62.1	54.1	7 (68)	2.4	2.9	30.4	50.2

Notes: Only complete states that started and ended within tagging or post-tagging are included in these statistics. Sample size (N) is given both per tag and number of states.

Table E2. Individual average buzz presence and mean ODBA during tagging and post-tagging.

	Tagging period					Post-tagging baseline				
		BUZZ %		ODBA			BUZZ %		ODBA	
State	N tags	Mean	SD	Mean	SD	N tags	Mean	SD	Mean	SD
1	7	0.0	0.0	26.7	7.4	9	0.0	0.0	21.9	3.5
2	9	8.4	14.4	25.4	6.9	9	10.1	8.6	25.1	3.6
3	9	15.0	11.1	27.0	4.7	9	23.1	14.6	29.1	4.5
4	6	0.0	0.0	22.0	5.5	9	5.8	4.5	20.6	3.5
5	0					4	0.0	0.0	14.3	7.2
6	4	0.0	0.0	24.5	5.5	7	0.0	0.0	26.2	2.3

Table E3. Time spent in each state during the tagging period and post-tagging period for each individual.

			Time in state (h)					
Tag ID	Data	Percentage of time	1	2	3	4	5	6
sw05_196a	tagging	0.5	6.9	10.3	72.4	10.3	0	0
	post	20.8	26.8	8	52.6	6.6	3.8	2.2
sw08_152a	tagging	2.8	1.8	16.6	32	18.3	0	31.4
	post	1.8	18.9	12.3	63.2	5.7	0	0
sw09_141a	tagging	0.8	16.7	8.3	20.8	27.1	0	27.1
	post	3	18.2	4.4	24.9	5	1.1	46.4
sw09_142a	tagging	0.2	0	23.1	76.9	0	0	0
	post	2.8	12.7	6.7	73.9	5.5	0	1.2
sw09_153a	tagging	0.1	0	33.3	66.7	0	0	0
	post	8.4	13.9	8.9	67.9	8.5	0	0.8
sw09_160a	tagging	0.2	25	66.7	8.3	0	0	0
	post	3.3	12.8	18.5	54.4	14.4	0	0
sw10_147a	tagging	0.9	23.6	18.2	40	18.2	0	0
	post	14.8	27.8	13.1	28.2	13.1	14.8	2.9
sw10_149a	tagging	1.8	17.8	19.6	43.9	7.5	0	11.2
	post	12.4	17.8	14	54	12.6	0.1	1.5
sw10_150a	tagging	0.8	19.6	23.9	21.7	13	0	21.7
	post	12.2	20.5	22.4	37.2	18.1	0	1.8

Table E4. Coefficient estimates from the best multinomial model (state ~ prevState + whale + Tagging).

	State 2		State 3		State 4		State 5		State 6	
Parameter	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
intercept	−2.68	0.19	−17.91	0.25	−30.93	0.29	−5.93	0.75	−3.29	0.25
prevState 2	27.85	0.27	42.26	0.25	50.12	0.60	−14.60	0.00	23.69	0.41
prevState 3	4.31	0.62	24.46	0.51	34.11	0.53	6.93	1.02	3.41	0.73
prevState 4	−2.16	0.81	15.30	0.35	31.70	0.27	1.58	1.26	−1.17	0.60
prevState 5	−13.47	0.00	18.10	0.70	29.76	0.87	9.49	0.92	4.43	0.69
prevState 6	2.69	0.29	16.17	0.47	28.93	0.45	5.54	0.80	3.63	0.27
sw08_152a	0.88	0.46	0.45	0.50	0.71	0.49	−7.21	81.20	1.72	0.47
sw09_141a	−0.69	0.48	−0.38	0.51	0.22	0.47	−1.25	0.86	1.95	0.35
sw09_142a	0.41	0.60	0.39	0.64	0.05	0.64	−7.46	50.12	0.09	0.89
sw09_153a	0.79	0.34	0.62	0.38	0.31	0.36	−13.97	0.00	−0.14	0.63
w09_160a	1.16	0.44	0.33	0.49	0.70	0.45	−6.83	48.66	−12.35	0.00
sw10_147a	0.16	0.26	−0.36	0.28	0.59	0.27	0.68	0.41	−0.05	0.34
sw10_149a	0.59	0.27	0.12	0.30	0.45	0.28	−1.88	1.24	0.25	0.37
sw10_150a	0.68	0.26	−0.19	0.29	0.76	0.27	−21.72	0.00	0.41	0.36
Tagging	0.30	0.31	0.03	0.34	0.28	0.33	−15.68	0.00	1.21	0.31

Table E5. Coefficient estimates for the GEE model (state6 ~ prevState + Tagging).

Parameter	Estimate, e^x		95% CI		Sequential Wald test		
					df	χ^2	p
intercept	0.23	1.26	−1.02	1.47			
prevState 1	−3.32	0.04	−4.22	−2.41	5	1409.9	<0.001
prevState 2	−5.29	0.01	−6.11	−4.47			
prevState 3	−6.36	0.00	−7.21	−5.51			
prevState 4	−5.67	0.00	−6.96	−4.38			
prevState 5	−3.02	0.05	−4.32	−1.72			
Tagging	1.31	3.70	0.30	2.31	1	6.52	0.011

Notes: Type 3 Wald tests are shown for each explanatory variable (prevState and Tagging). Mean estimates are given both in link scale and e^x transformed.

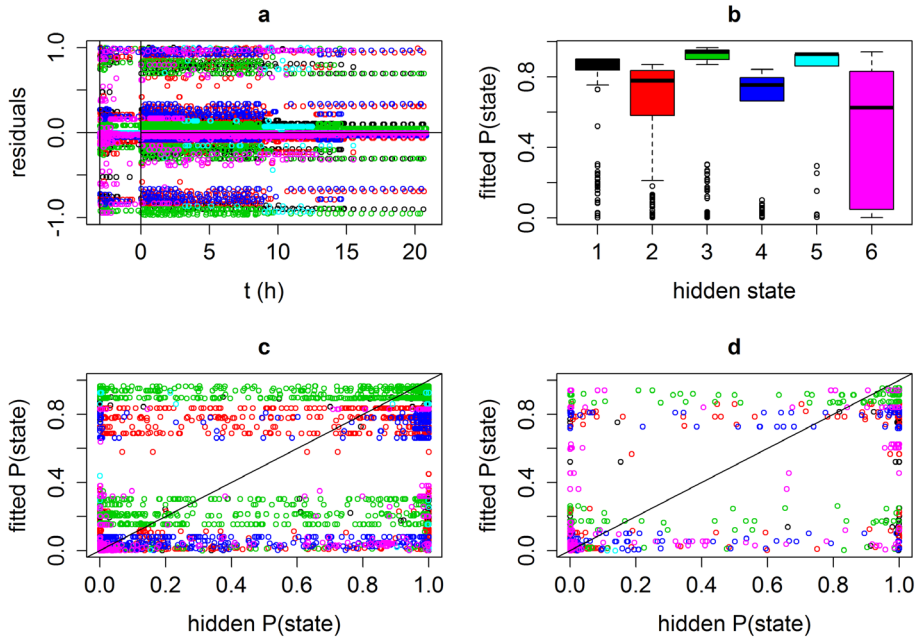

Fig. E1. Residuals and fitted values for the best multinomial model (state ~ prevState + whale + Tagging). (a) All state-dependent raw residuals for the post-tagging baseline data as a function of time since the end of tagging period (vertical line at 0 hours), and for the tagging condition as a function of time since tag deployment (vertical line at −3 hours); (b) fitted probabilities for each hidden state (i.e., observed state vs. fitted probability in the multinomial model); (c–d) fitted probabilities by the multinomial model as a function of the posterior probability of each state. Panel (c) shows data for post-tagging baseline, while panel (d) shows data for the tagging period.

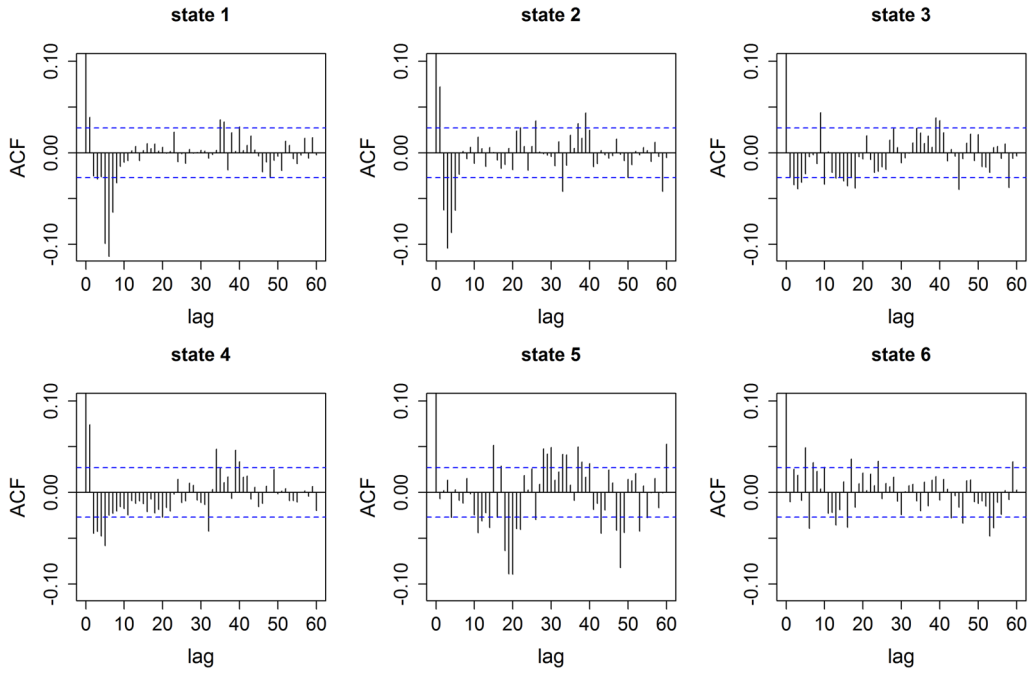

Fig. E2. Autocorrelation (y-axis) as a function of lag (x-axis) for the state-specific residuals of the best multinomial model (state $\sim$ prevState + whale + Tagging).

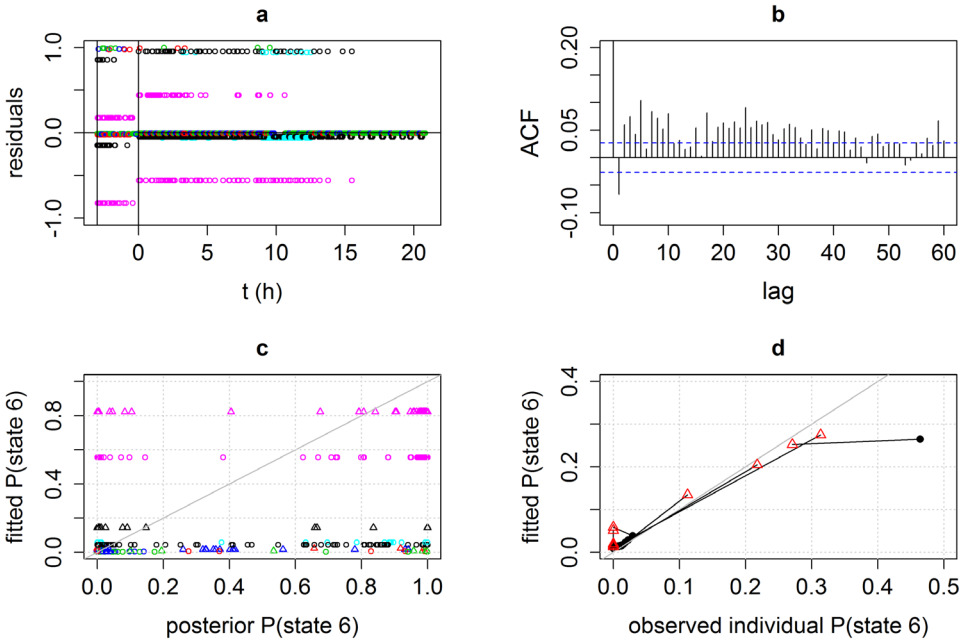

Fig. E3. Residuals and fitted values for the GEE model (state6 ~ prevState + Tagging). (a) Raw binomial residuals for the post-tagging baseline data as a function of time since the end of the tagging period (vertical line at 0 hours), and for the Tagging condition as a function of time since tag deployment (vertical line at −3 hours), colored by previous state (pink = state 6). A positive residual indicates state 6 in the data, and a smaller positive value indicates that the GEE fitted a higher probability. (b) Autocorrelation of the raw residuals (y-axis) as a function of lag (x-axis). (c) Fitted probability of state 6 as a function of posterior probability of state 6. Circles and triangles show data from post-tagging baseline and tagging periods, respectively, colored by previous state. (d) Individual average fitted probability vs. average observed presence of state 6 within the post-tagging baseline (black solid circles) and Tagging (red triangles) periods. Segments join data from the same individual.

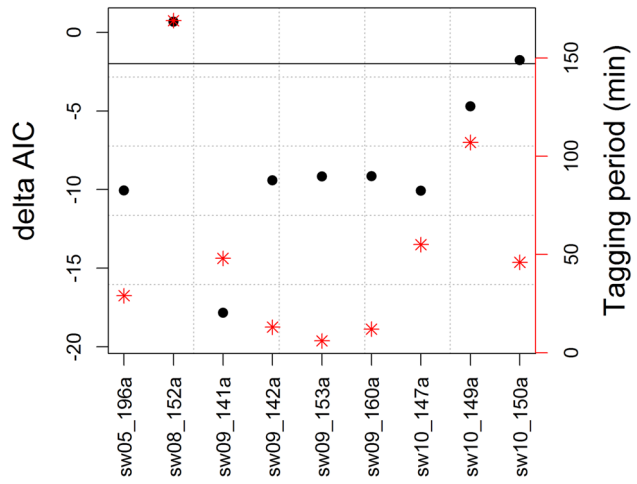

Fig. E4. The best tagging effects model for state (prevState + whale + Tagging) re-fit without each whale (x-axis) and checked for AIC against the baseline model (prevState + whale). Right y-axis (red) shows duration of the tagging period in minutes. Horizontal line shows our cut-off ΔAIC = −2.

Supplement

R scripts for fitting the base- and full model structures to sample data (*Ecological Archives*, http://dx.doi.org/10.1890/ES14-00130.1.sm).