Simulating cetacean responses to sonar exposure within a Bayesian hierarchical modelling framework

Technical report

Phil Bouchet, Catriona Harris, Len Thomas

20 April, 2020

Contents

1 Preamble

In this report, we present a framework for simulating responses of cetaceans to various military sonar exposure contexts using Bayesian hierarchical modelling. This work was motivated by the need to assess the utility of different types of animal-attached biotelemetry tags in improving our understanding of dose–response relationships (Schick *et al.* 2019). Specifically, we used a Monte Carlo approach to conduct a sensitivity analysis of the effects of uncertainty in acoustic dose measurements (i.e. received sound levels) on the probability of behavioural response. Accompanying R code is available and fully described in a sister document (see Bouchet *et al.* 2020 for details).

2 Introduction

Sound plays a critical role in the lives of cetaceans, and many species of whales, dolphins, and porpoises are sensitive to the adverse effects of chronic and acute exposure to anthropogenic underwater noise (Weilgart 2007; Williams *et al.* 2015; Erbe *et al.* 2019). For instance, elevated noise levels (e.g. in areas of dense vessel traffic) have the potential to impair animal communication ('auditory masking'; Erbe *et al.* 2016; Cholewiak *et al.* 2018), disrupt movement and diving behaviours, elicit physiological stress, and/or cause displacements from preferred habitats (DeRuiter *et al.* 2013), ultimately interfering with key life functions such as foraging, mating, nursing, or resting, with knock-on repercussions on individual fitness, energy expenditure, and survival (Tyack 2008; Erbe *et al.* 2018; Wensveen *et al.* 2019). In recognition of man-made noise as an emerging threat to wildlife, an increasing number of calls have been made to strengthen management and mitigation frameworks for sound-producing activities (Dolman *et al.* 2011; Dolman and Jasny 2015). In the United States, the Marine Mammal Protection Act of 1972 (MMPA, 16 U.S.C. 1361 et seq.) regulates the 'take' (i.e. defined as the harassment, hunting, capture, or killing) of marine mammals by U.S.-based organisations anywhere around the globe, including areas beyond national jurisdiction (i.e. on the high seas). The U.S. Navy is legally bound to comply with the MMPA and other U.S. Federal laws (e.g. the Endangered Species Act ESA 16 U.S.C.1531 et seq.) pertaining to protected marine taxa, and is thus required to determine the potential effects of Systems Command military readiness training exercises on cetaceans, especially where those involve the use of tactical high-powered sonar technology and the deployment of explosives/munitions.

Of particular concern are the impulsive sounds produced by active sonars operating in the lower (LFAS, ~0.1-2 kHz) and mid-frequency bands (MFAS, 3–8 kHz) (Falcone *et al.* 2017). LFAS and MFAS systems were developed in the 1950s for anti-submarine detection and naval warfare (D'Amico and Pittenger 2009; de Quirós *et al.* 2019), and their use has recently been implicated in a number of atypical mass strandings largely involving deep-diving pelagic whales from the *Ziphiidae* family, such as Cuvier's beaked whales (*Ziphius cavirostris*) (Cox *et al.* 2006; Angela D'Amico and Mead 2009; Filadelfo *et al.* 2009; Fernández *et al.* 2012; Simonis *et al.* 2020; see also Parsons 2017 for a recent review). In the last two decades, recurring reports of such mortality events prompted a series of coordinated international research efforts aimed at quantifying probabilities of response to both simulated and actual naval sonar sources under controlled experimental exposure conditions (Southall *et al.* 2016; Harris *et al.* 2016). These behavioural response studies (BRSs) have catalysed significant advances in our understanding of the short-term impacts of specific acoustic doses on animals (Harris *et al.* 2016; Harris *et al.* 2018), highlighting substantial variability in the nature, magnitude, and consequences of observed responses within and between individuals and populations (e.g. DeRuiter *et al.* 2013; Goldbogen2013; Friedlaender *et al.* 2016; Southall *et al.* 2019).

In BRSs, whale behaviour is typically monitored using animal-borne bio-logging tags, with additional information sometimes derived from opportunistic visual observations or passive acoustics (e.g. Berga *et al.* 2019; von Benda-Beckmann *et al.* 2019). The onerous costs of running at-sea BRS experiments, which often exceed many hundreds of thousands of dollars for a single field season (Harris *et al.* 2016), provide a strong impetus for integrating different sampling approaches to maximise data collection opportunities over a range of complementary spatio-temporal scales. As such, a rising number of studies simultaneously deploy short-term (ca. hours), high-resolution, archival digital tags (DTAGs; Johnson and Tyack 2003) and medium to long-term (ca. days to months), coarse-resolution, position and depth-transmitting satellite tags (e.g. Tyack *et al.* 2011; Wensveen *et al.* 2019; Schick *et al.* 2019). Suction cup DTAGS incorporate multiple sensors, including a hydrophone (sampling rate up to 192 kHz), a pressure sensor, triaxial accelerometers and magnetometers, and an embedded VHF transmitter, which enable fine-scale diving behaviour (i.e. orientation, depth, and speed) to be captured in three dimensions synchronously with the recording of audio data (Tyson *et al.* 2012; Laplanche *et al.* 2015). Although DTAGs offer detailed insights into dynamic activity states, their limited sampling duration precludes assessments of long-term baseline behaviours, both prior to and following noise disturbance (Schick *et al.* 2019). By contrast, implantable satellite tags allow the animals' horizontal movements to be captured over much wider spatial and temporal domains (Schorr *et al.* 2014), yet most modern instruments lack on-board hydrophones and therefore cannot obtain direct sound measurements. Furthermore, satellite tags programmed to transmit via the Argos system (https://www.argos-system.org/) can suffer from substantial positional errors that may introduce large uncertainties in estimates of acoustic dose (often exceeding 50 dB re 1μPa rms in range) (Schick *et al.* 2019; von Benda-Beckmann *et al.* 2019).

Such discrepancies in data quality and resolution between the two types of tags raise important questions with regards to the optimisation of field protocols in BRSs (Harris *et al.* 2018). To inform optimal choices of tag configurations, we conducted a Bayesian simulation exercise designed to explore how uncertainty in the measurements of received sound levels made on both digital and satellite tags may affect inference of dose–response relationships. Specifically, we simulated behavioural response data from virtual whales exposed to military sonar and fitted with different tags, and investigated the sample sizes and accuracy required to estimate dose–response functions (also referred to as risk functions, Moretti *et al.* 2014) with an acceptable degree of confidence. The use of computational Bayesian methods for model fitting in ecology has increased in recent decades (Clark 2005; Beaumont 2010; Dorazio 2016). Here, Bayesian analysis

offers a natural mechanism for estimating the parameters of potentially complex hierarchical models within a single framework that is robust to the small sample sizes often encountered in BRS research, and can provide measures of parameter uncertainty that are directly interpretable in probabilistic terms (Parent and Rivot 2012; Antunes *et al.* 2014). This is crucial for making appropriate predictions of responsiveness during real-world naval exercises (Harris *et al.* 2018).

3 Original model

This work expands on the Bayesian hierarchical dose–response model presented by Miller *et al.* (2014) (Figure 1). We only summarise the model briefly here, and refer the reader to the original publication for full details.

The model assumes that for any sonar exposure session, each individual whale i has a response threshold that is a function of (1) the typical average response threshold of all whales, μ; (2) two contextual covariates (namely, exposure history and frequency of the sonar signal/stimulus), in addition to random between-whale (ϕ^2) and within-whale, between-session variation (σ^2). The full model consists of both a **process** model, which describes the underlying factors driving the true thresholds of exposure for each session, and an **observation** model, which links the true thresholds to the observed values, measured with error (δ^2).

The **process** model is interpreted as follows. Let t_{ij} be the true, unknown threshold of exposure that elicits a behavioural response for the i_{th} whale in the j_{th} exposure session. We assume that this threshold follows a truncated normal distribution such that:

$$t_{ij} \sim TN(\mu_{ij}, \sigma^2, L, U) \tag{1}$$

where σ^2 is the within-animal between-session variance in threshold, and L and U are lower and upper limits to the threshold. Let us also assume that the expected threshold μ_{ij} for the i_{th} whale in the j_{th} exposure session is a function of the expected threshold for that whale, μ_i, as well as whether the animal has been previously exposed to sonar, and the frequency band of the sonar signal used. This gives:

$$\mu_{ij} \sim \mu_i + \alpha\, I(\text{exposed})_{ij} + \beta\, I(\text{MFAS})_{ij} \tag{2}$$

Here, α is a parameter governing the effect of exposure history on threshold, and $I(\text{exposed})_{ij}$ is an indicator function to which a value of 0 is assigned during the first exposure session, and a value of 1 thereafter. Likewise β represents the effect of MFAS relative to LFAS, with $I(\text{MFAS})_{ij}$ taking the value 1 if the exposure session was with MFAS, and 0 otherwise. Lastly, we assume that the expected threshold for each whale μ_i follows a truncated normal distribution:

$$\mu_i \sim TN(\mu, \phi^2, L, U) \tag{3}$$

where μ is the mean threshold for all whales, ϕ is the between-whale variance in threshold, and L and U are defined as above.

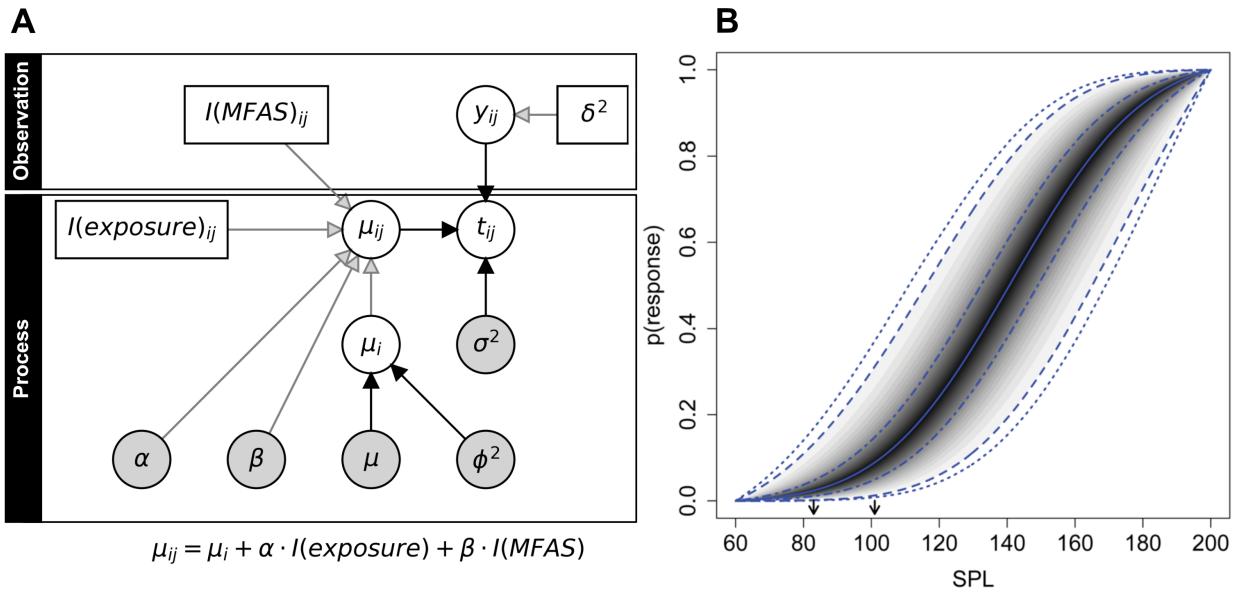

Figure 1: Bayesian hierarchical dose–response model used by Miller *et al.* (2014) in their analysis of killer whale (*Orcinus orca*) dose–escalation data. **(A)** Directed acyclic graph showing the model structure. Model variables are denoted by circles, whereas constants are represented by boxes. Variables monitored for posterior inference are shaded in grey. Black and grey arrows indicate stochastic and deterministic relationships, respectively. **(B)** Posterior dose–response curve showing the probability of onset of avoidance against received sound pressure level (SPL, dB re μPa), for the same species. The solid central line represents the posterior mean, followed by 50%, 95%, and 99% credible interval lines.

The **observation model** allows the inclusion of uncertainty in simulated threshold values. Here, we assumed that measurements on tags followed a normal distribution i.e.

$$y_{ij} \sim N(t_{ij}, \delta^2) \tag{4}$$

Note that Miller *et al.* (2014) set the standard deviation δ to 2.5 dB, giving a 95% density interval for the threshold of $\pm$ 5.0 dB around the point estimate.

4 Simulations

4.1 Scenarios

We considered four scenarios, each a variant of the original Miller *et al.* (2014) model (Figure 2). Scenarios differed in the nature and complexity of their observation and process model components, as described below. Simulation plans and directed acyclic graphs (DAGs) are reported in Appendix A and Appendix B), respectively, at the end of this document.

Scenario 1 is a reduced version of Miller *et al.* (2014), whereby individuals are exposed to sonar only once and fitted with the same tag type. As such, the within-whale variance, ϕ, and the between-whale variance, σ, are combined into a single parameter representing the overall variance in threshold, ω. There are no

covariates affecting the probability of response in this scenario. We tested an array of realistic sample sizes, from N = 5 to N = 40, and increasing levels of measurement error from δ = 2.5 dB re 1μPa to 35 dB re 1μPa. The lower bound reflects typical errors observed on DTAGs (Isojunno & Wensveen, personal communication), while the upper bound is consistent with estimates from sattelite-tagged whales (Schick *et al.* 2019; von Benda-Beckmann *et al.* 2019; Joyce *et al.* 2020).

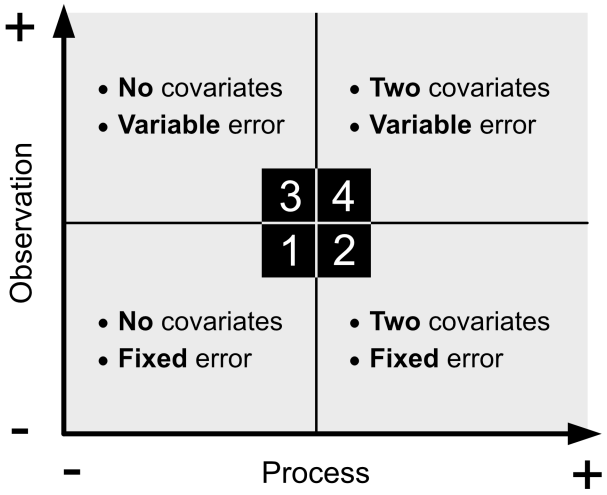

Figure 2: Visual summary of the four simulation scenarios considered. The complexity of process and observation model components goes from low (-) to high (+) along each axis, and reflects the inclusion/omission of covariates affecting the response thresholds (process model) and errors in measurements of the dose being treated as constants for all animals or as tag-specific variables. Covariates include exposure history and sonar signal type. Scenario IDs are shown in the central black boxes.

Scenario 2 is identical to the original Miller *et al.* (2014) model, and was implemented with a range of measurement errors between 2.5 dB re 1μPa and 35 dB re 1μPa, as in scenario 1. To keep the average expected threshold μ_{ij} centred in the simulations, we treated signal type as a relative effect between animals exposed to MFAS vs. LFAS, assuming that the former exacerbated sensitivity to sonar exposure. As such, the β parameter was coded in as an effect size, such that for β = 20 dB re 1μPa, the corresponding coefficient values for MFAS and LFAS were -10 and +10 dB re 1μPa, respectively.

Scenario 3 mimicks scenario 1, but includes a more complex observation model that accommodates tag-specific measurement errors. Indeed, the positions of tagged individuals are recorded with varying precision — typically higher for animals fitted with DTAGs that are also concurrently followed by visual observers at the surface, and far lower for satellite-tagged animals monitored through the Argos system (Costa *et al.* 2010). Argos relies on the Doppler shift between a polar-orbiting satellite and the animal to communicate positional information (Schick *et al.* 2019). For tagged individuals to be detected and geolocalised, a sufficient number of satellites must be available when animals are at the surface. This presents a significant challenge for deep-diving cetaceans such as beaked whales, which only come up to breathe for short periods of time, thus reducing the likelihood of successful data uplinks to overhead receivers. In practice, observations are still recorded with uncertainty, even when links are successful (Schick *et al.* 2019). Prior to 2008, each position was assigned an ordinal location quality code (e.g., 3, 2, 1, 0, A, B, and Z), with typically substantially higher errors in longitude than in latitude (Vincent *et al.* 2002), such that true errors around calculated positions are better represented by 2-dimensional anisotropic ellipses than by 1-dimensional circles (McClintock *et al.*

2015). Following 2008, Argos has therefore been supplying error ellipses with each location, whereby each ellipse has three components — namely its semi-major axis (M), semi-minor axis (m), and orientation (c). Taken together, these define a bivariate normal distribution of geolocation error, with larger ellipses being associated with higher positional uncertainty (McClintock *et al.* 2015). Accounting for this uncertainty is critical in making fair assessments of variance in received levels, and thus in quantifying dose–response relationships effectively (Schick *et al.* 2019). To address this, we first estimated the coordinates of each virtual whale on the (x, y) plane at the time of exposure, based on its simulated true response threshold t_{ij} and a simple inverse-square circular transmission loss model (Figure 3). Note that sound absorption is frequency-dependent, and here, we assumed an absorption coefficient of 0.185 dB re 1 μPa per km, which corresponds to a 3 kHz signal under normal sea conditions (Miller *et al.* 2014). Coordinates were obtained relative to the noise source, using a random angle sampled from a uniform distribution $U(0, 360)$. Next, we created one plausible realisation of an Argos ellipse for each animal by randomly sampling a vector of ellipse parameters $\theta_{ij} = (M_{ij}, m_{ij}, c_{ij})$ from an existing dataset on tagged Cuvier's beaked whales (*Ziphius cavirostris*) collected as part of the Atlantic BRS (Schick *et al.* 2019). We then generated 10,000 candidate locations within this ellipse, and calculated the acoustic dose received at each of these locations based on the aforementioned transmission loss model. We took the standard deviation of the resulting values as a reasonable estimate of the measurement uncertainty associated with each satellite-tagged whale. Note that the Atlantic BRS also targeted short-finned pilot whales (*Globicephala macrorhynchus*). Given the unique diving behaviours of each species, the R code has been set up so that the above calculations can be run on either (see the argument species.argos in the function run.scenario(); Bouchet *et al.* 2020). By contrast, we fixed δ to a constant value of 2.5 dB re 1 μPa for whales fitted with DTAGs, as the accuracy of these instruments is unlikely to vary appreciably. At each sample size, animals were randomly chosen and fitted with either type of tag, according to a pre-determined ratio (digital vs. satellite) which we varied from 0% to 100% in 20% increments (Appendix A).

Scenario 4 merges scenarios 2 and 3 insofar as individuals are repeatedly exposed and fitted with different types of tags. Here, response thresholds are also dependent on signal type (MFAS vs LFAS) and exposure history.

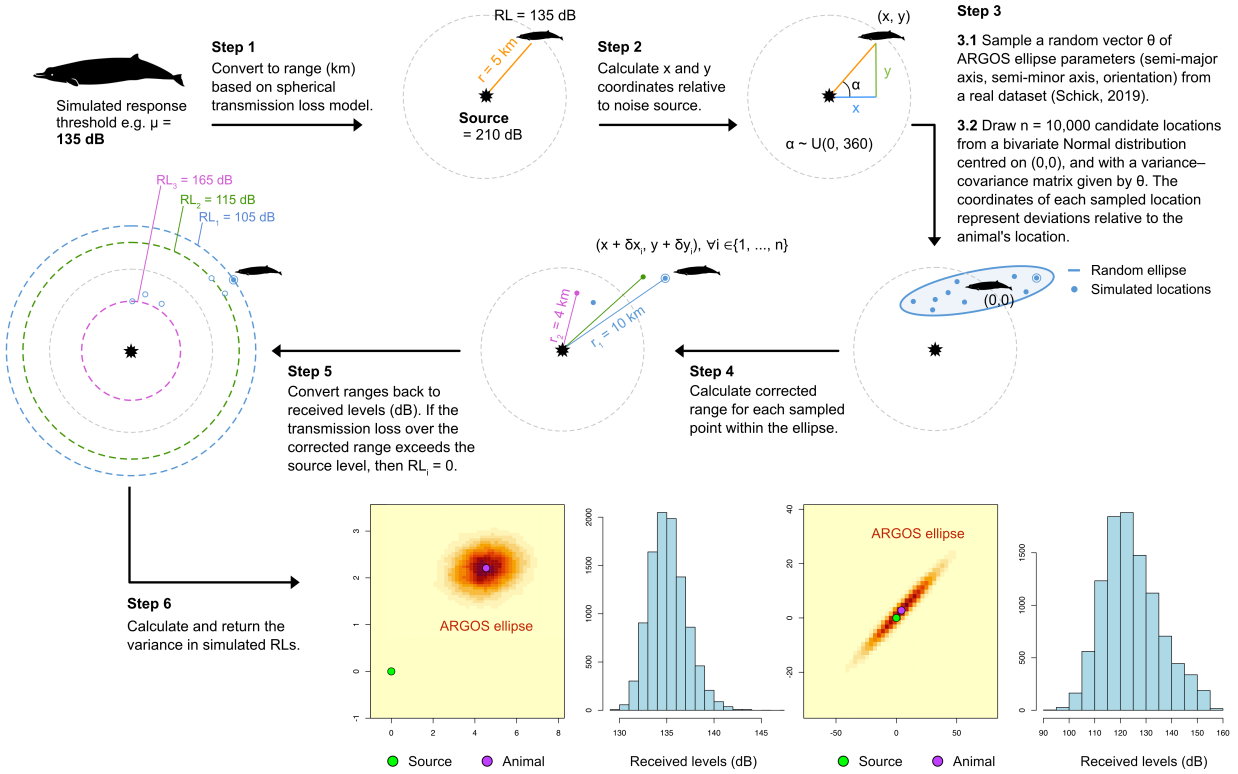

Figure 3: Visual summary of the approach taken to estimate uncertainty in received levels after accounting for positional errors inherent to Argos-linked satellite tags.

4.2 Data and parameters

Table 1 below provides a summary of simulation inputs and parameter values.

In brief, the simulated data were generated so as to reflect the number of trials, sonar frequencies, and orders of exposures observed in real-world BRSs (Miller *et al.* 2014). We focused on Cuvier's beaked whales (*Ziphius cavirostris*) as a model species, considering a single exposure session in scenarios 1 and 3 and three repeated trials in scenarios 2 and 4. The source level of the sonar signal was held constant at 210 dB re 1 μPa in all simulations (Tyack *et al.* 2011; DeRuiter *et al.* 2013; Antunes *et al.* 2014; Stimpert *et al.* 2014). The true underlying mean response threshold (for all whales) was taken as μ = 150 dB re 1 μPa based on Moretti *et al.* (2014), and is also broadly consistent with the values used in Miller *et al.* (2014), Schick *et al.* (2019), and other studies. By default, between-animal and within-animal, between-session variation were set to ϕ = 20 dB re 1 μPa and σ = 25 dB re 1 μPa, respectively. As variances add, the overall combined variation ω was equal to $\sqrt{\phi^2 + \sigma^2} \approx$ 30 dB re 1 μPa.

It is common for some individuals to exhibit no response across the range of doses experienced during an escalation session (Antunes *et al.* 2014; Harris *et al.* 2015). The resulting data are right-censored, and indicate that the animals' response thresholds exceed the maximum dose received by some unknown amount (Klein and Moeschberger 2003). It is critical to include these data in any analysis, as they are informative about the nature of dose–response relationships (Harris *et al.* 2015). Note, however, that there is limited empirical evidence of behavioural effects at very high sound levels nearing 200 dB re 1μPa (or above), and that the value used as an upper boundary for right-censoring may have an overwhelming influence on

posterior inference if not chosen appropriately (Wensveen 2016). Here, we simulated right-censoring by making a random draw from a uniform distribution $R_c \sim U(190, 200)$ for each exposed animal (Tyack *et al.* 2011; DeRuiter *et al.* 2013; Antunes *et al.* 2014; Stimpert *et al.* 2014). In practice, if the simulated response threshold for an animal (i.e. μ_i in scenarios 1 and 3, and t_{ij} in scenarios 3 and 4; see Appendix B) was greater than the associated realised maximum dose Rc, then that threshold was assumed to be equally likely between Rc and the upper limit U (set equal to the source level, as all whales would be expected to respond if located right over the source).

Note that, for simplicity, we did not implement Gibbs Variable Selection (GVS) (Tenan *et al.* 2014; Miller *et al.* 2014). Rather, in scenarios 2 and 4, we assessed the ability of the models to discriminate covariate effects by examining the posterior distributions of the relevant coefficients, α and β. If the corresponding 95% credible intervals included zero, then we deemed the model unable to detect an effect.

Table 1: Parameter values used in bayesian simulations. The last two parameters (i.e. animal density and number of bins) relate to the effective response range. See section 4.5 for details.

Parameter	Definition	Value	Unit	Scenario
Ns	Number of simulations	500		All
N	Sample size (number of whales)	5, ..., 40		All
Nt	Number of trials (exposure sessions)	3		2 + 4
δ	Uncertainty (sd) in dose measurements	2.5, ..., 35	dB	1 + 2
P(SAT)	Proportion of whales fitted with satellite tags	0, ..., 100	%	3 + 4
DTAG.sd	Uncertainty (sd) in dose measurements on DTAGs	2.5	dB	3 + 4
SL	Level of the noise source	210	dB	All
Sp	Study species	Z. cavirostris		All
μ	Mean response threshold for all whales	150	dB	All
ω	Combined variation in threshold (sd)	30	dB	1 + 3
ϕ	Between-whale variation (sd)	20	dB	2 + 4
σ	Within-whale, between-session variation (sd)	25	dB	2 + 4
α	Effect of exposure history on response threshold	8	dB	2 + 4
β	Effect of MFAS vs. LFAS on response threshold	20	dB	2 + 4
Rc	Right-truncation limit (right-censored data)	190 to 200	dB	All
D	Animal density	1	Animal/km^2	All
Nb	Number of bins used to calculate the ERR	500		All

4.3 Priors

Prior distributions are required on all top-level random variables in the hierarchical model (Laplanche *et al.* 2015) (shown as grey circles in Figure 1). We followed Miller *et al.* (2014) and largely chose diffuse uniform priors (Table 2), with the exception of the two coefficients governing the respective effects of exposure history (α) and sonar frequency band (β), for which we assumed normal distributions centred on zero. Note that this choice equates to a prior belief of no effect. Our expectation therefore was that as sample size would increase and errors decrease, posterior estimates of α and β would move away from zero and closer to the truth, and that their associated credible intervals would shrink. Importantly, priors were specified so as to constrain model parameters within biologically plausible bounds. For instance, μ could take any value between 60 and 210 dB re 1μPa with equal probability, under the conservative assumption that any noise below the lower limit would be barely audible above ambient (Pacini *et al.* 2011; Schick *et al.* 2019), and that all animals would respond at or above the upper limit (Antunes *et al.* 2014; Miller *et al.* 2014). Variance parameters for the observation model were assumed known, and hence did not require priors.

Table 2: Summary of prior values for each parameter, expressed in dB re 1 μPa. Lower and upper limits are reported for uniform distributions (U), and mean and standard deviations are reported for normal distributions (N). True underlying values are shown in the right-most column.

Scenario	Variable	Prior
All	μ	U(60, 200)
1 + 3	ω	U(0, 40)
2 + 4	ϕ	U(0, 30)
2 + 4	σ	U(0, 30)
2 + 4	α	N(0, 10)
2 + 4	β	N(0, 10)

4.4 Model fitting

Models were fitted using a Markov Chain Monte Carlo (MCMC) algorithm, implemented in the software JAGS via the rjags library (Plummer 2019) in R v3.6.0 (R Core Team 2019). Model parameters were estimated based on 10,000 posterior samples, taken after a variable burn-in (i.e. as the number of samples required to achieve convergence differed between scenarios and simulation conditions; see Table 3). Each parameter was initialised using arbitrary starting values. In all models, MCMC runs consisted of three Markov chains, which were assessed for convergence using functions from the coda (Plummer *et al.* 2019) and bayesplot (Gabry and Mahr 2019) packages. This was done both by visual inspection of trace plots, and by ensuring that the scale reduction factor, or Gelman-Rubin statistic ($\hat{R}$), was < 1.1 (Kruschke 2015). We fitted models to 500 simulated datasets, running the R code in parallel on multiple cores to increase execution speed (Bouchet *et al.* 2020).

Table 3: Number of Markov Chain Monte Carlo (MCMC) iterations discarded as initial burn-in in each scenario. δ = Uncertainty in measurements of the acoustic dose (sd, in dB re 1μPa). P(SAT) = Proportion of animals fitted with satellite tags (in %).

		Scenarios			
δ	P(SAT)	1	2	3	4
2.5	-	10,000	60,000	-	-
5	-	10,000	60,000	-	-
10	-	40,000	90,000	-	-
15	-	75,000	125,000	-	-
20	-	75,000	125,000	-	-
25	-	100,000	150,000	-	-
30	-	100,000	150,000	-	-
35	-	100,000	150,000	-	-
-	0	-	-	50,000	100,000
-	20	-	-	50,000	100,000
-	40	-	-	75,000	125,000
-	60	-	-	75,000	125,000
-	80	-	-	125,000	175,000
-	100	-	-	125,000	175,000

4.5 Model outputs

Reliable assessments of sonar impacts cannot be made without knowledge of animal density patterns and sound transmission properties, for a given exposure context. In particular, single-value step function thresholds commonly used in 'traditional' impact assessments have been shown to grossly underestimate the numbers of animals affected by a given sound disturbance (Tyack and Thomas 2019). Because of this, we calculated the effective response range (or effective response radius, ERR) from the posterior estimates of model parameters in each simulation. Drawing from concepts rooted in distance sampling theory, the ERR is a novel metric of impact defined as the distance beyond which as many animals are expected to respond as do not respond within it (see Tyack and Thomas 2019 for a technical explanation). We assumed the same simple inverse-square circular transmission loss model as previously described, and performed calculations within 500 bins placed over a maximum range given by the distance at which received levels drop below 60 dB re 1μPa (i.e. such that the probability of response is effectively zero), which in this case equalled ca. 240 km.

We compared the posterior distributions of μ, ω, ϕ, σ, α, β and the ERR with their 'true' underlying values, focusing on three key diagnostics: **(1) precision**, expressed as the average width of the parameters' posterior credible intervals (CIw, in dB re 1μPa or km); **(2) accuracy**, measured as the average absolute percent mean bias (PMB, in %); and **(3) identifiability**, as captured by the average prior posterior overlap (PPO), obtained using the MCMCtrace function in package MCMCvis (Youngflesh 2018). Checking the PPO is useful for identifying parameter-redundant models, i.e. models in which the prior for a parameter simply dictates its posterior distribution, and the data have little if any bearing on the results. A 35% guideline for overlap has been suggested as an indicator of weak parameter identifiability, meaning that when PPO < 35%, the data may be informative enough to overcome the influence of the prior (Gimenez *et al.* 2009). Note that, as a derived quantity, the ERR has no prior and therefore no PPO.

Lastly, we also computed dose–response curves from each model in the same way as Miller *et al.* (2014), and created plots of the associated posterior median and posterior credible interval lines for a range of chosen quantiles (from 5% to 95%, in 5% increments).

5 Results

For both clarity and brevity, only key results are presented below. The full set of dose–response curves and summary plots for each scenario is available from the authors upon request.

5.1 Scenario 1

Dose–response curves indicated a strong interaction between sample size (N) and measurement errors (δ) in this scenario (Figure 4), with the uncertainty around estimated relationships rapidly inflating as N decreased and δ increased. All else being equal, changes in posterior uncertainty were more pronounced at the lowest sample sizes. The curves obtained for sample sizes of N = 5 to 10 and δ = 15 to 20 dB re 1μPa were qualitatively similar to the one derived by Miller *et al.* (2014) for killer whales (*Orcinus orca*). The model appeared able to estimate μ with negligible bias, however ω was consistently under-estimated when sample sizes were low and errors high, a pattern mirrored in estimates of the ERR (Figures 5 and 6). For a given sample size, precision increased (up to two-fold) for all parameters as errors were lowered to a level commensurate with that of DTAGs (δ = 2.5 dB re 1μPa). μ was found to be identifiable (PPO < 35%) in all but the most extreme conditions (i.e. lowest N and highest δ), highlighting the relevance of the information contained in the simulated data for predicting the population-level threshold of response. However, considerable overlap between prior and posterior was observed for ω, particularly when N < 15 (irrespective of the level of measurement error) (Figure 6).

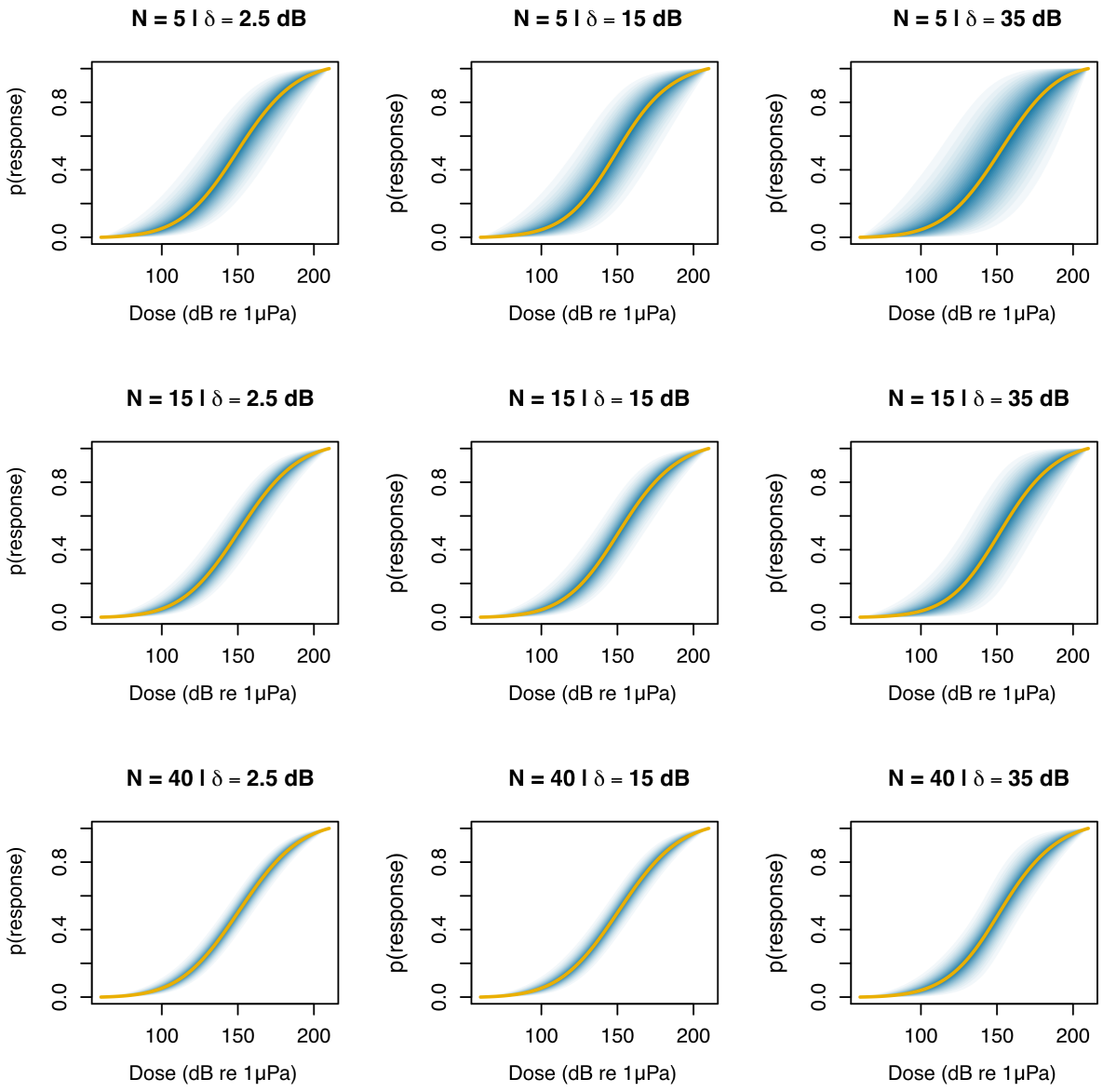

Figure 4: Example dose–response curves estimated under scenario 1 for a range of sample sizes (N) and errors in dose measurements (δ). The solid line represents the average posterior median across Ns = 500 simulations, followed by the average 5%, 10%, 15% … and 95% credible intervals in darker to lighter shades of blue.

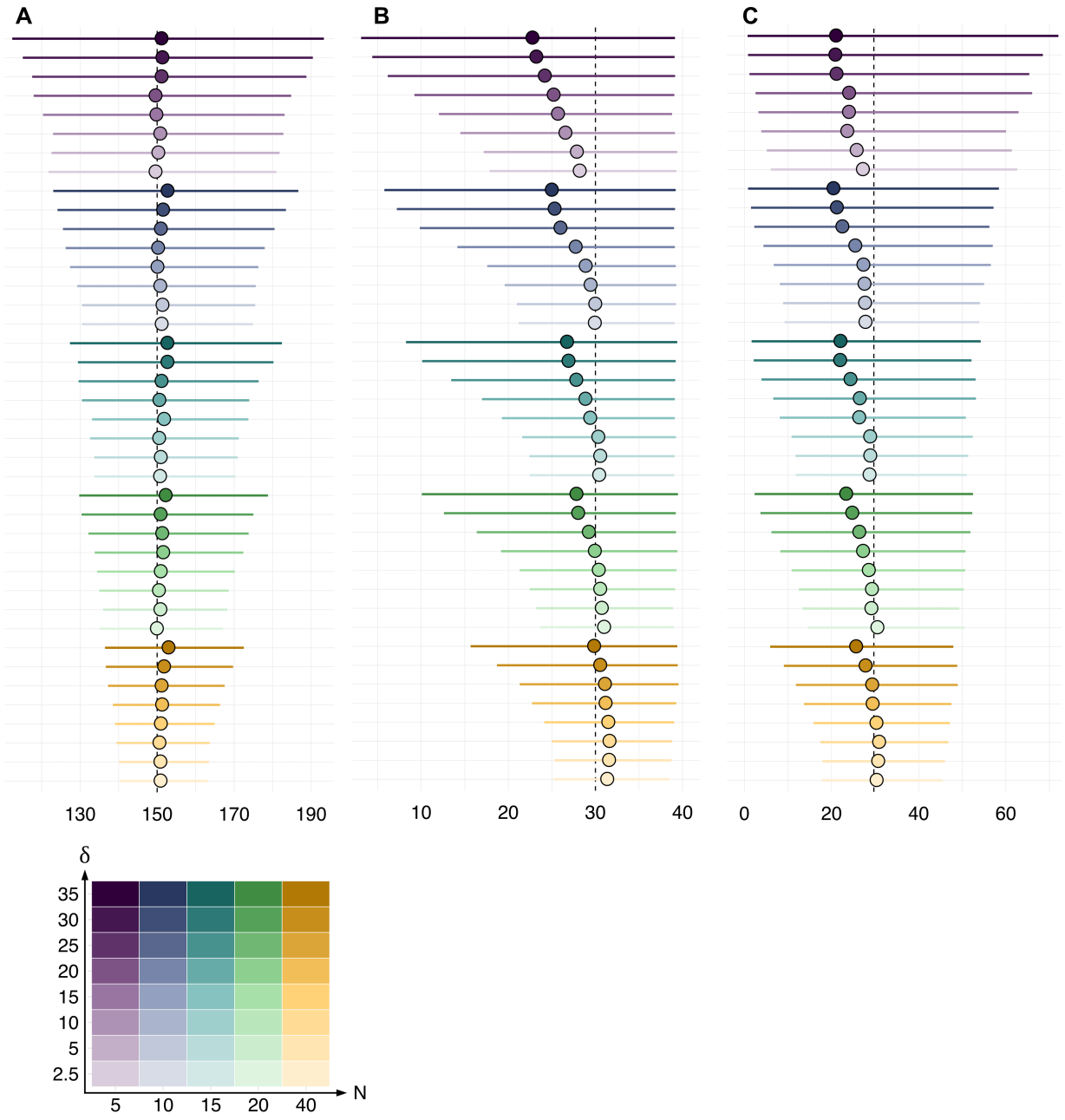

Figure 5: Summary of posterior inference for key model parameters under scenario 1. Circles and bars denote, respectively, the average posterior median and average posterior credible intervals across Ns = 500 simulations, for combinations of sample sizes (N) and observation errors (δ, standard deviation in measurements of the acoustic dose). Parameters are as follows: **(A)** mean response threshold for all whales, μ, **(B)** overall (between and within-whale) variation in response threshold, ω, and **(C)** effective response range (ERR). X-axis scales are expressed in dB re 1μPa (A,B) and km (C), respectively. Dashed lines represent true underlying values for each parameter.

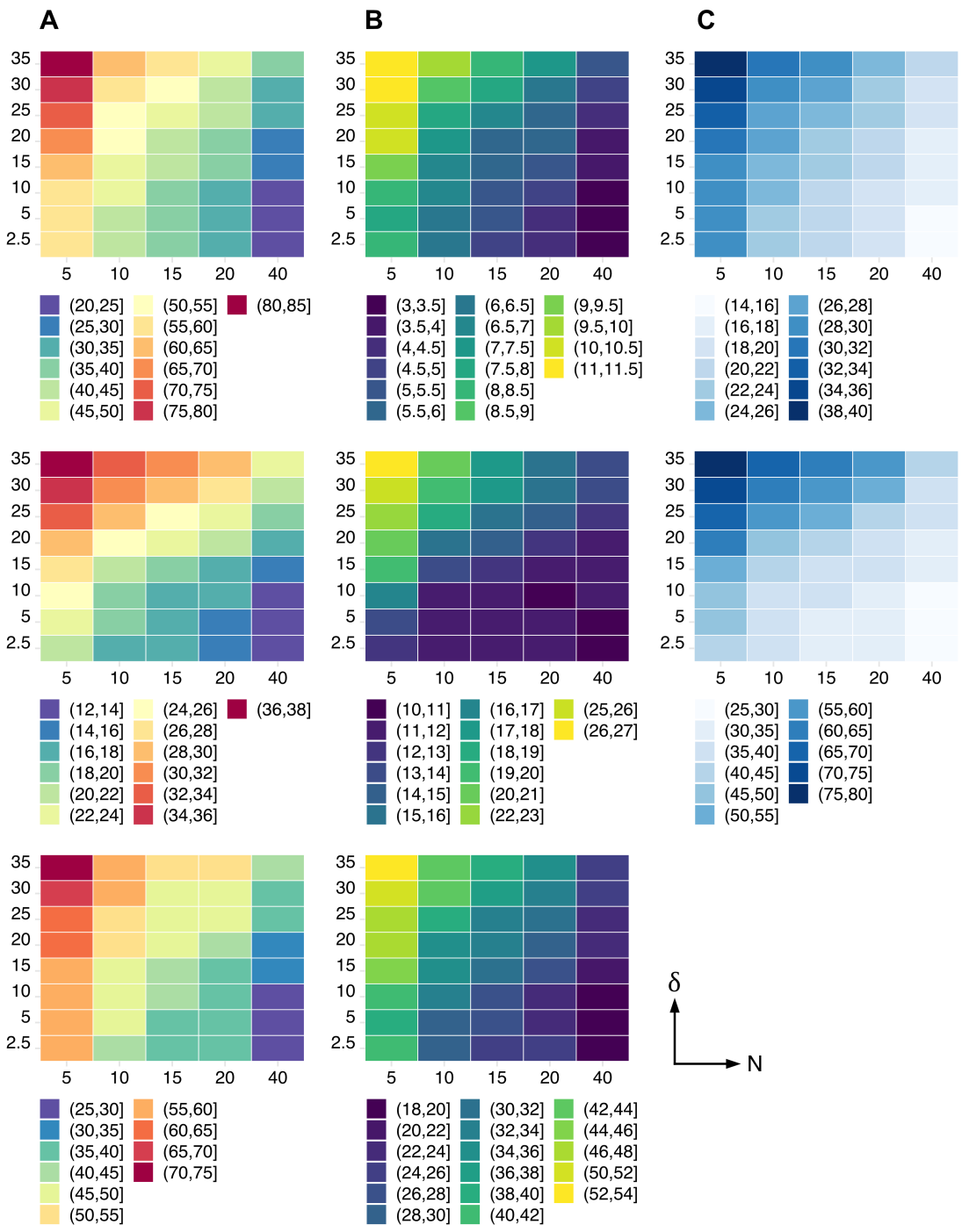

Figure 6: Precision, accuracy, and identifiability of posterior estimates of the mean response threshold for all whales (μ) **(top)**, the combined (between-whale and within-whale) variation in threshold (ω) **(middle)**, and the effective response range (ERR) **(bottom)**, under scenario 1. Columns show **(A)** the average width of credible intervals (CIw) across simulations, expressed in dB re 1μPa for μ and ω and in km for ERR, **(B)** the average absolute percent mean bias (PMB, in %), defined as (posterior mean-truth)/truth, and **(C)** the prior posterior overlap (PPO, in %). PPO values above 35% indicate that parameters may be non-identifiable (Gimenez *et al.* 2009).

5.2 Scenario 2

Similarly to scenario 1, an interaction between sample size (N) and measurement errors (δ) was apparent in this case (Figure 7), with larger uncertainty in estimated relationships with decreasing N and increasing δ. The model was able to estimate ϕ with negligible bias across most samples sizes, as well as β when the number of simulated whales exceeded N = 20. All other posterior parameter estimates deviated from their true underlying values to varying degrees, with larger bias consistently observed in simulations with high measurement errors. Again, the effective response range (ERR) was under-estimated in all but the best simulation conditions (i.e. high sample sizes and low errors) (Figures 8 & 9). The model was unable to detect an effect of previous exposure history on expected response thresholds (α), however the simulated effect of sonar signal frequency (β) was successfully identified in all runs at N = 40, and when measurement errors were limited to a maximum of SD = 15 and 5 dB re 1μPa at N = 20 and N = 15, respectively. This aligns with observations of high overlap (>35%) between prior and posterior for α, whereas β was identifiable in more than half of test conditions (Figure 9).

5.3 Scenario 3

Sample size was found to be the main driver of variability around estimated dose–response relationships in this scenario, such that curves remained virtually identical as the ratio of digital vs. satellite tags changed, but became more uncertain as N decreased (Figure 10). Accordingly, credible intervals for all parameters were large with small datasets (e.g. N = 5), and gradually decreased as a higher number of virtual whales were tagged (Figure 11). Losses in accuracy, precision, and identifiability with decreasing N were more pronounced than in scenario 1 (with an equivalent underlying process model) (Figure 12). μ was always identifiable (PPO < 35%), and ω consistently so for N > 15.

5.4 Scenario 4

Consistent with scenario 3 (i.e. where the same observation model was used), dose–response curves remained largely similar for a given sample size, irrespective of the number of satellite tags deployed (Figure 13). Credible intervals for all parameters were large with small datasets (e.g. N = 5), and decreased as a higher number of virtual whales were tagged (Figure 14). No effect of previous exposure history could be identified (i.e. the posterior credible intervals for α always included zero), yet the influence of sonar frequency (β) could be detected with sample sizes as low at N = 15, in line with β being identifiable in two thirds of test conditions (Figure 15). The effective response range (ERR) was estimated with limited bias, and with lowest uncertainty at highest sample sizes.

Note that in the previous two scenarios, we emulated the deployment of digital and satellite tags by modifying the level of measurement error (δ) allowed in the observation model. This was done by treating δ as a constant for all whales regardless of exposure, with low δ values (e.g. 2.5 dB re1μPa) taken to be representative of the uncertainty associated with DTAGs, and high values corresponding to uplinks of consistently poor quality (i.e. high geolocational error) from satellite tags (e.g. 35 dB re1μPa). Under these conditions, measurement uncertainty played a strong role in dictating dose–response relationships (see sections 5.1 and 5.2). This is an extereme case, however, as positional estimates derived from satellite tags vary in quality, meaning that some measurements of the acoustic dose experienced by satellite-tagged animals can still

be made with reasonable precision (Figure 16). Here, the dominant influence of sample size over tag choice in scenarios 3 and 4 can be explained by the limited range of variation in received levels estimated from satellite tags under real-world settings, conditional on the sound propagation model described previously (average 25%, 50% and 75% quantiles of SD(dose) = 8.3, 11.6, and 20.2 dB re1μPa respectively).

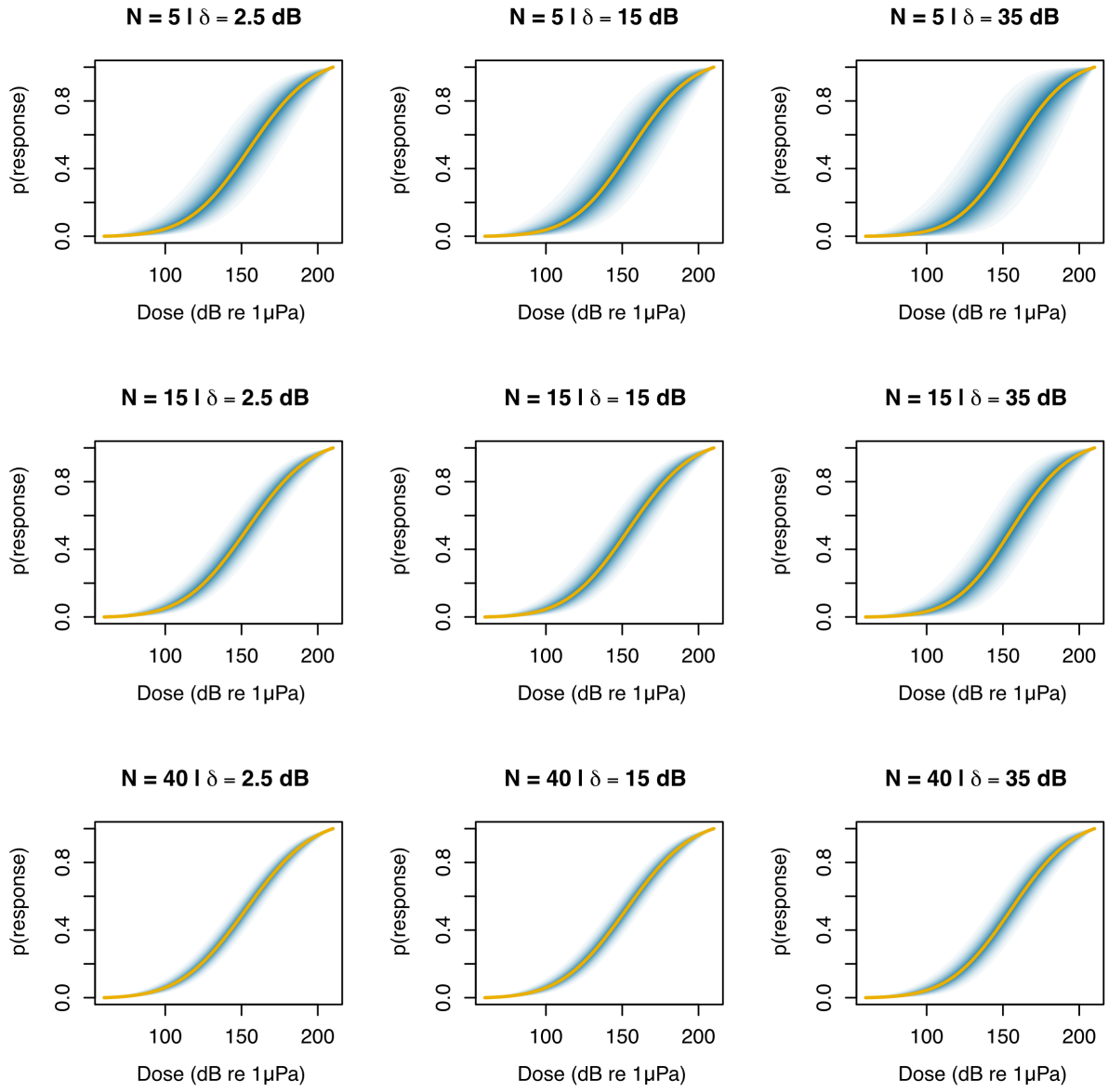

Figure 7: Example dose–response curves estimated under scenario 2 for a range of sample sizes (N) and errors in dose measurements (δ). The solid line represents the average posterior median across Ns = 500 simulations, followed by the average 5%, 10%, 15% … and 95% credible intervals in darker to lighter shades of blue.

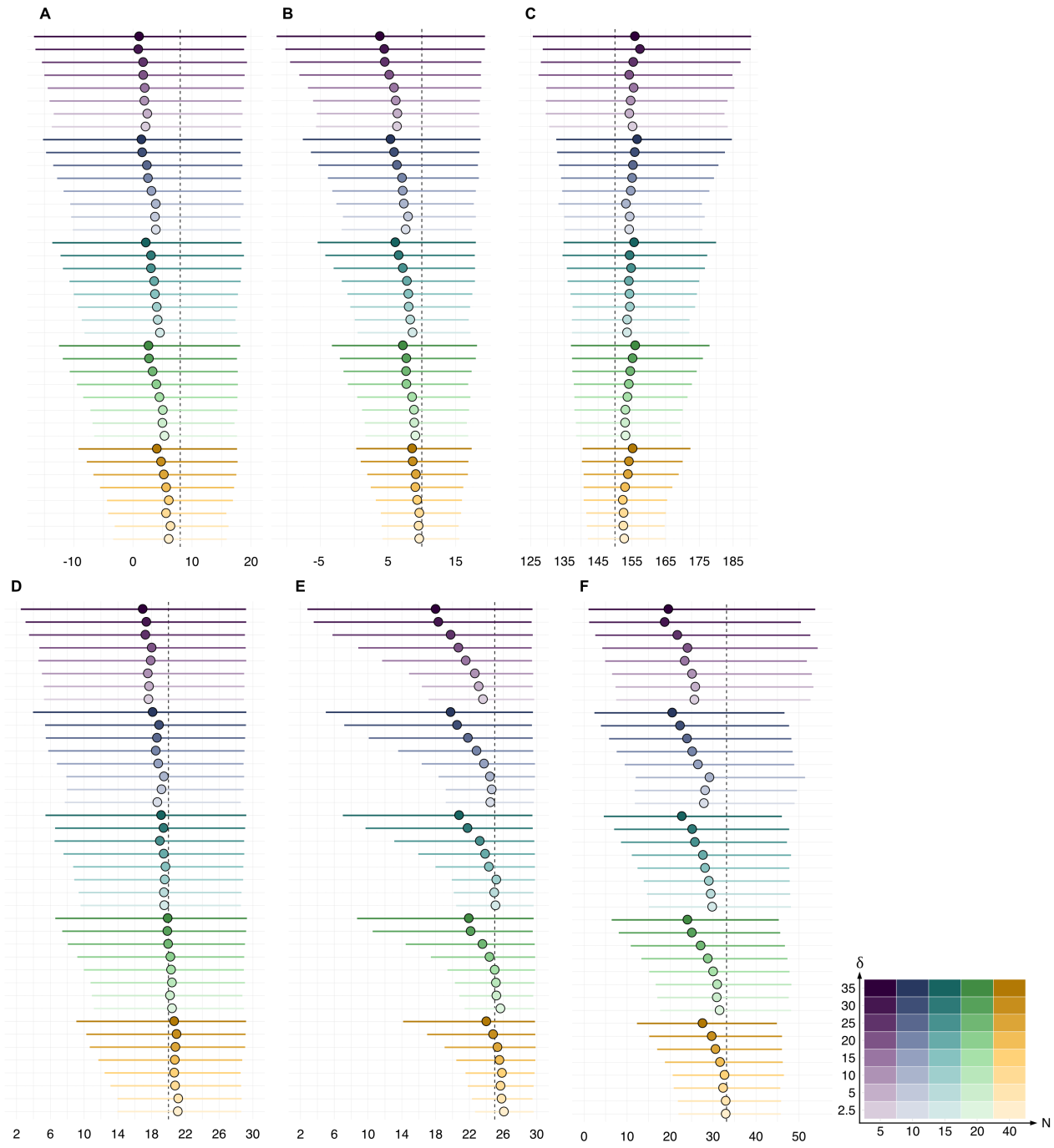

Figure 8: Summary of posterior inference for key model parameters under scenario 2. Circles and bars denote, respectively, the average posterior median and average posterior credible intervals across Ns = 500 simulations, for combinations of sample sizes (N) and observation errors (δ, standard deviation in measurements of the acoustic dose). Parameters are as follows: **(A)** effect of exposure history on response threshold, α, **(B)** effect of sonar frequency on response threshold, β, **(C)** mean response threshold for all whales, μ, **(D)** between-whale variance in response threshold, ϕ, **(E)** within-whale, between-session variance in response threshold, σ, and **(F)** effective response range (ERR). X-axis scales are expressed in dB re 1μPa (A to E) and km (F), respectively. Dashed lines represent true underlying values for each parameter.

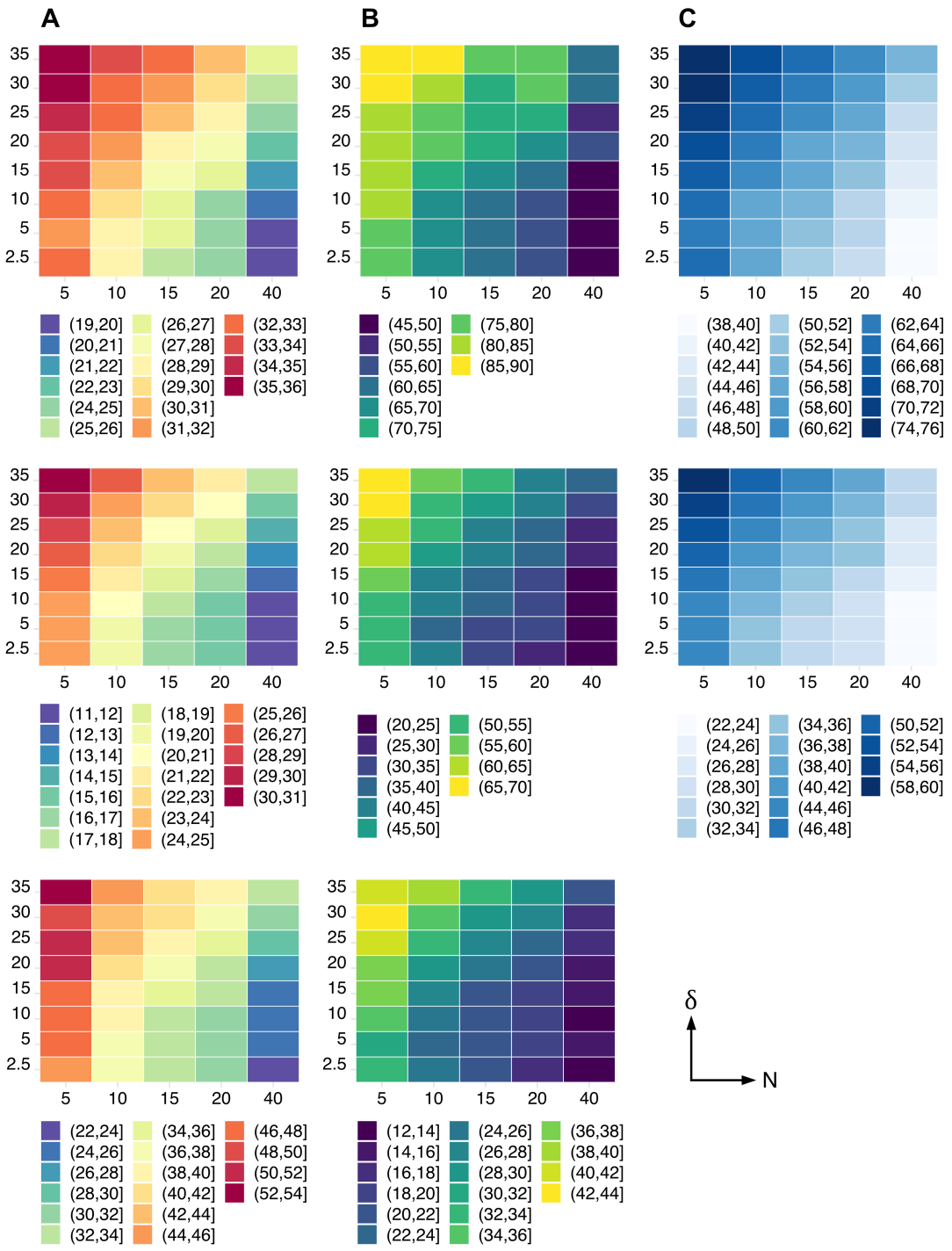

Figure 9: Precision, accuracy, and identifiability of posterior estimates of the effect of exposure history (α) **(top)**, the effect of sonar frequency (ω) **(middle)**, and the effective response range (ERR) **(bottom)**, under scenario 2. Columns show **(A)** the average width of credible intervals (CIw) across simulations, expressed in dB re 1μPa for α and β and in km for ERR, **(B)** the average absolute percent mean bias (PMB, in %), defined as (posterior mean-truth)/truth, and **(C)** the prior posterior overlap (PPO, in %). PPO values above 35% indicate that parameters may be non-identifiable (Gimenez *et al.* 2009).

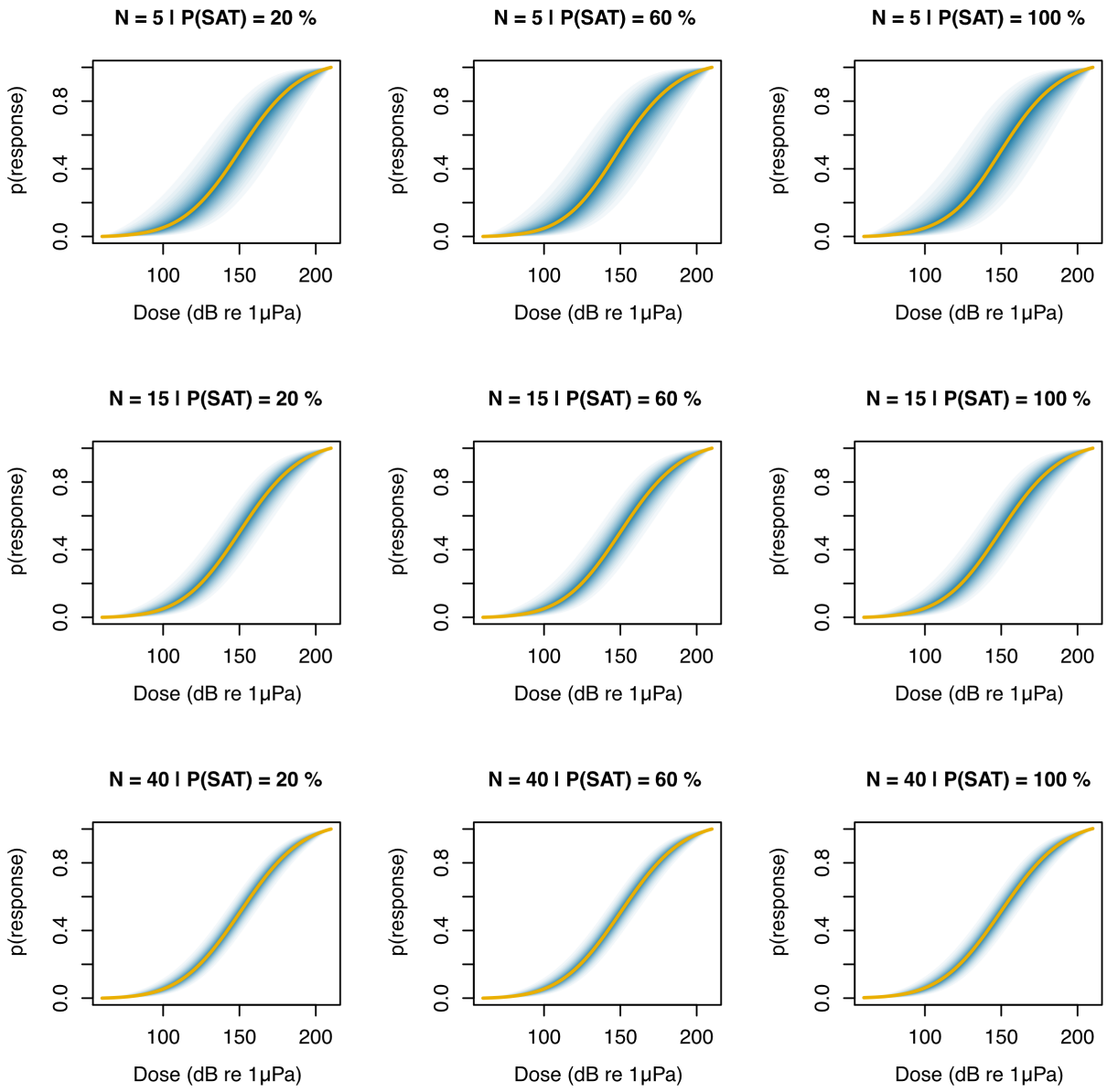

Figure 10: Example dose–response curves estimated under scenario 3 for a range of sample sizes (N) and proportions of animals fitted with satellite tags (P(SAT)). The solid line represents the average posterior median across Ns = 500 simulations, followed by the average 5%, 10%, 15% … and 95% credible intervals in darker to lighter shades of blue.

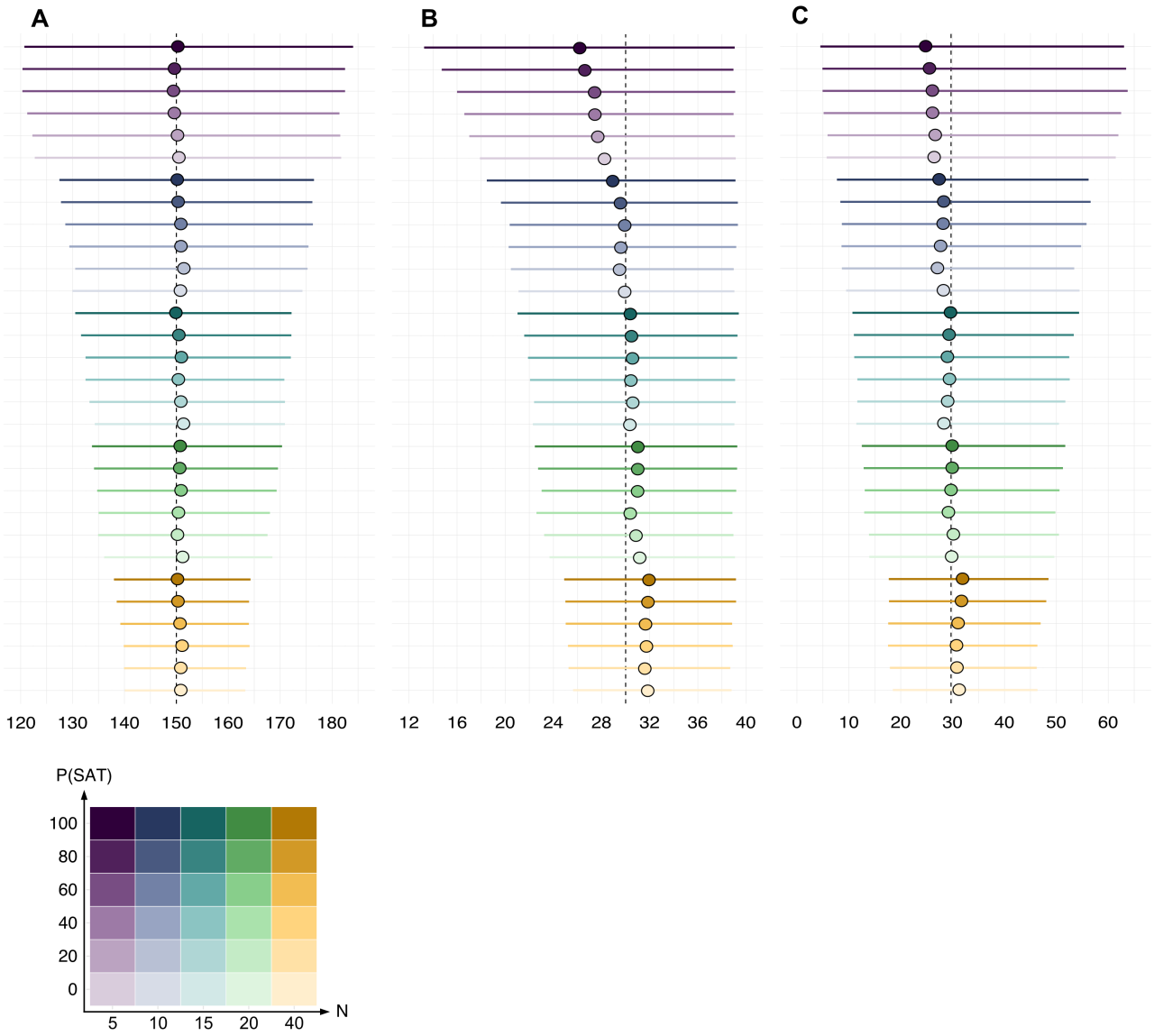

Figure 11: Summary of posterior inference for key model parameters under scenario 3. Circles and bars denote, respectively, the average posterior median and average posterior credible intervals across Ns = 500 simulations, for combinations of sample sizes (N) and proportions of animals fitted with satellite tags (P(SAT), in %). Parameters are as follows: **(A)** mean response threshold for all whales, μ, **(B)** overall (between and within-whale) variation in response threshold, ω, and **(C)** effective response range (ERR). X-axis scales are expressed in dB re 1μPa (A,B) and km (C), respectively. Dashed lines represent true underlying values for each parameter.

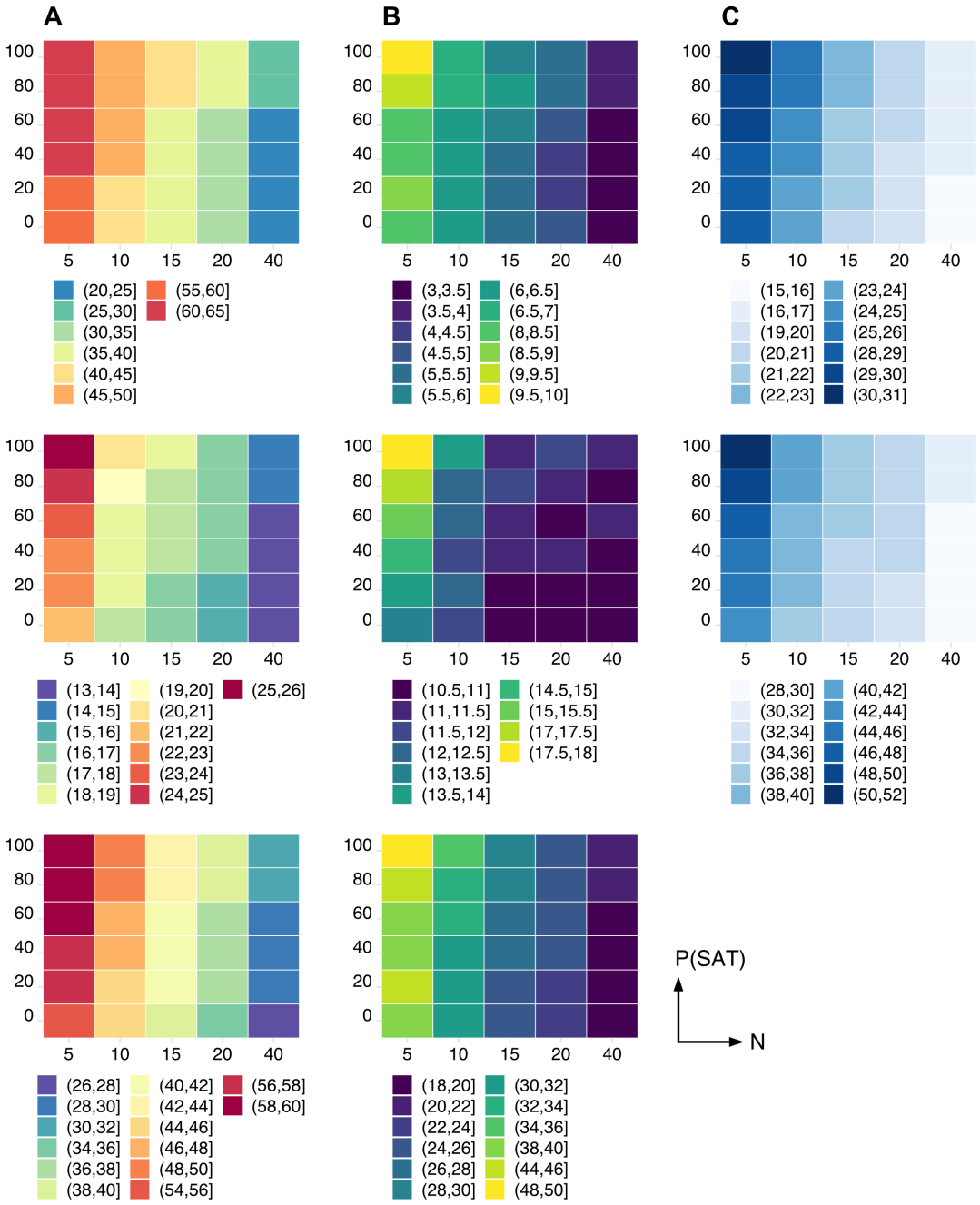

Figure 12: Precision, accuracy, and identifiability of posterior estimates of the mean response threshold for all whales (μ) **(top)**, the combined (between-whale and within-whale) variation in threshold (ω) **(middle)**, and the effective response range (ERR) **(bottom)**, under scenario 3. Columns show **(A)** the average width of credible intervals (CIw) across simulations, expressed in dB re 1μPa for μ and ω and in km for ERR, **(B)** the average absolute percent mean bias (PMB, in %), defined as (posterior mean-truth)/truth, and **(C)** the prior posterior overlap (PPO, in %). PPO values above 35% indicate that parameters may be non-identifiable (Gimenez *et al.* 2009).

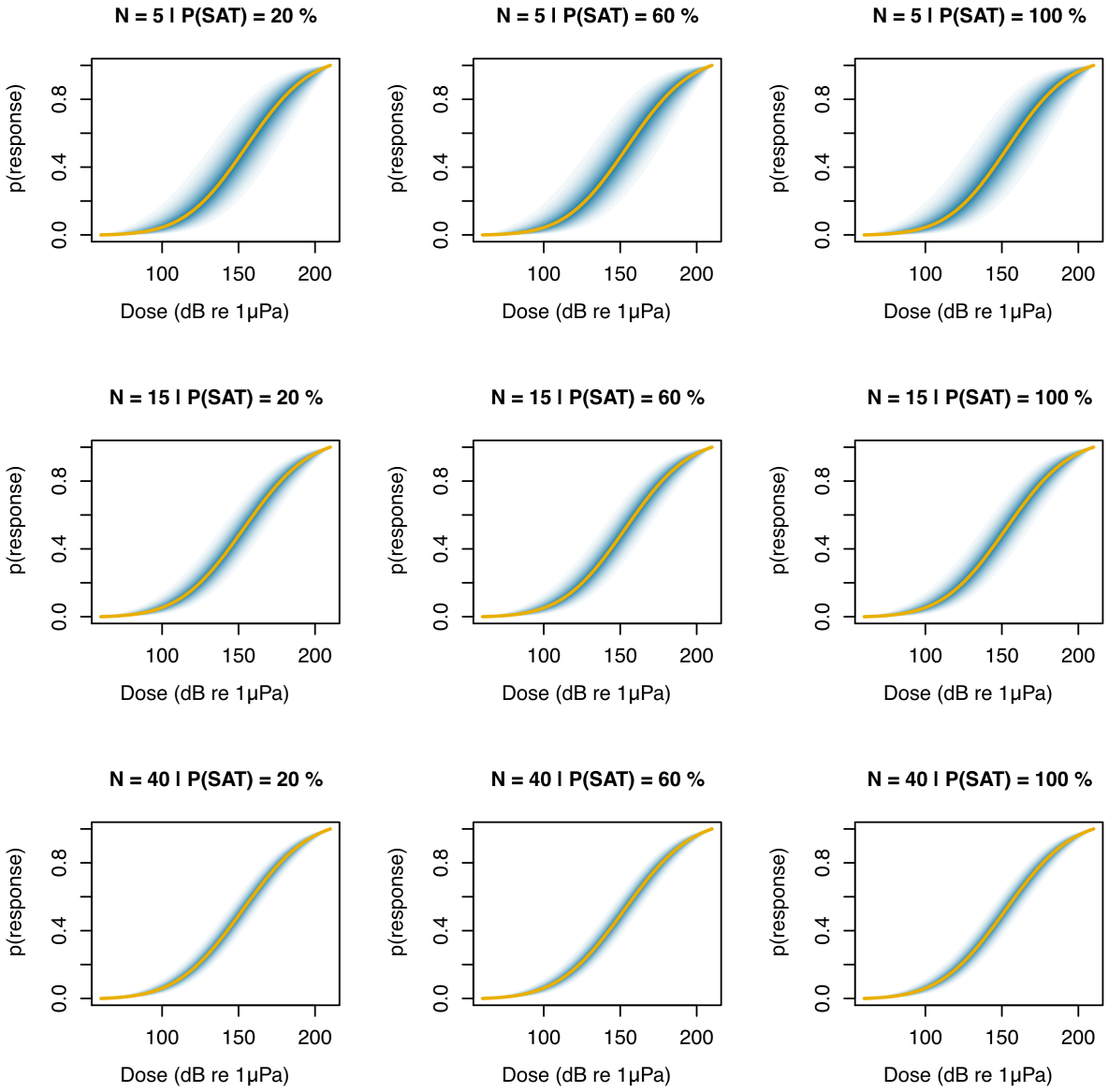

Figure 13: Example dose–response curves estimated under scenario 4 for a range of sample sizes (N) and proportions of animals fitted with satellite tags (P(SAT)). The solid line represents the average posterior median across Ns = 500 simulations, followed by the average 5%, 10%, 15% … and 95% credible intervals in darker to lighter shades of blue.

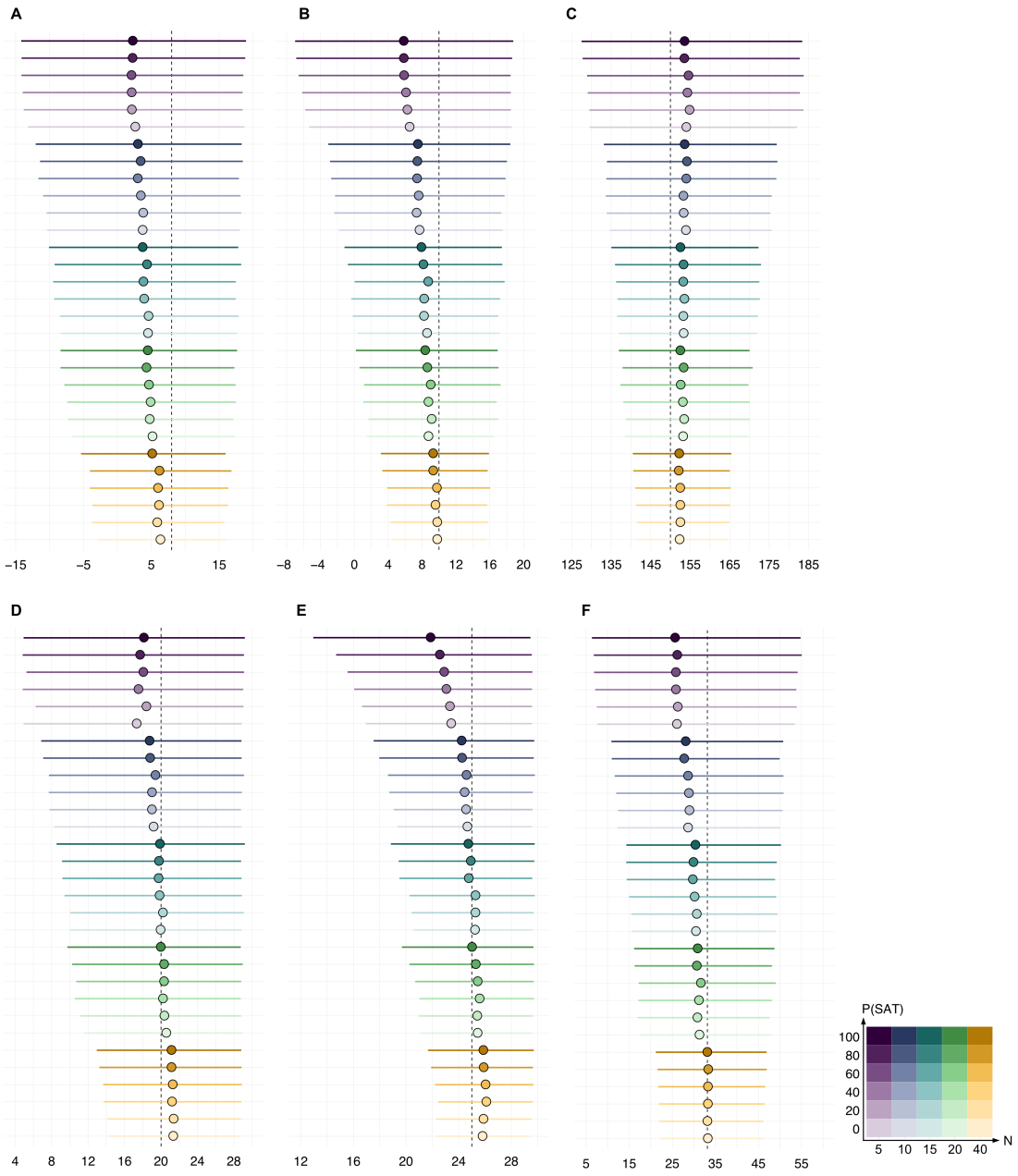

Figure 14: Summary of posterior inference for key model parameters under scenario 4. Circles and bars denote, respectively, the average posterior median and average posterior credible intervals across Ns = 500 simulations, for combinations of sample sizes (N) and proportions of animals fitted with satellite tags (P(SAT), in %). Parameters are as follows: **(A)** effect of exposure history on response threshold, α, **(B)** effect of sonar frequency on response threshold, β, **(C)** mean response threshold for all whales, μ, **(D)** between-whale variance in response threshold, ϕ, **(E)** within-whale, between-session variance in response threshold, σ, and **(F)** effective response range (ERR). X-axis scales are expressed in dB re 1μPa (A to E) and km (F), respectively. Dashed lines represent true underlying values for each parameter.

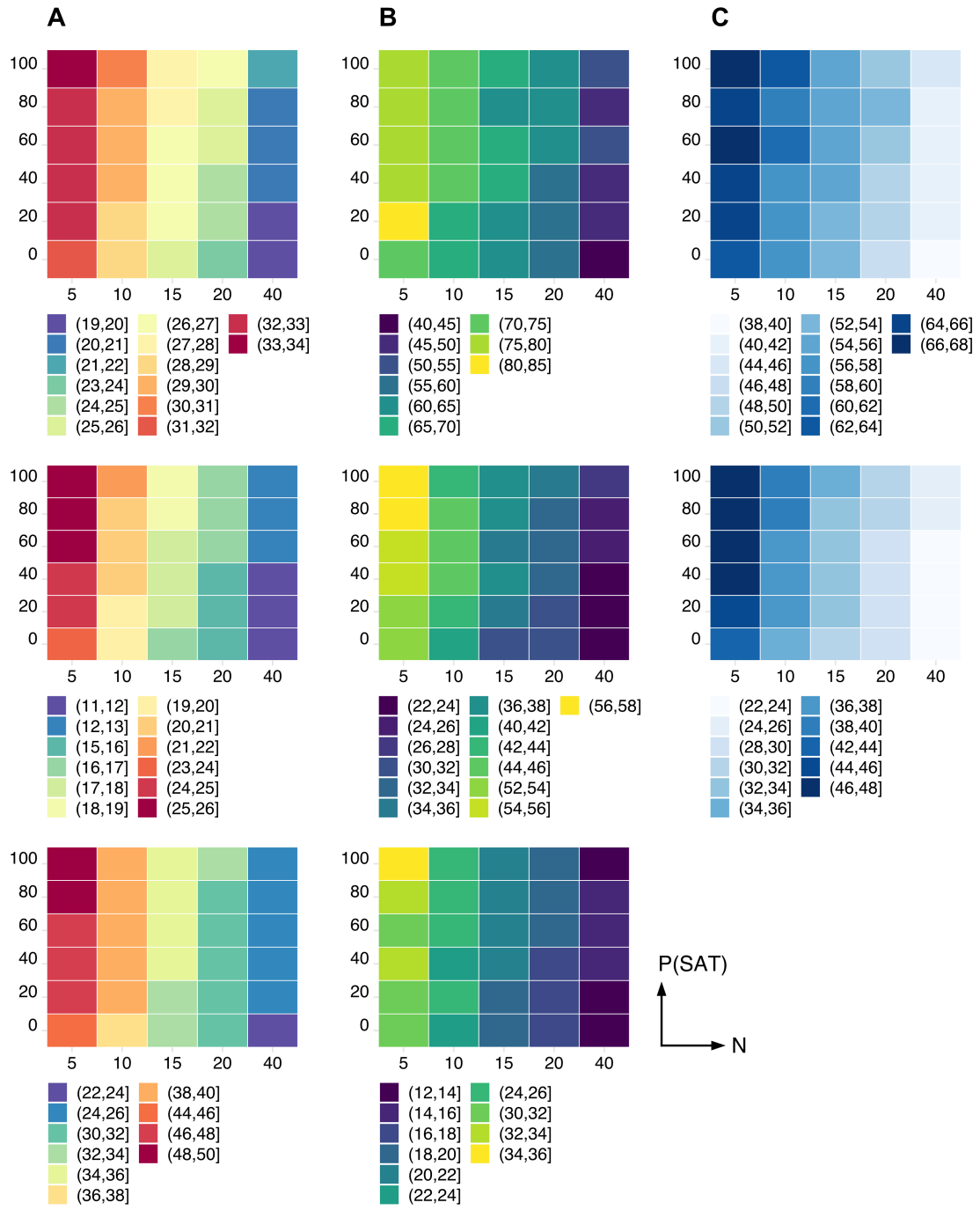

Figure 15: Precision, accuracy, and identifiability of posterior estimates of the effect of exposure history (α) **(top)**, the effect of sonar frequency (ω) **(middle)**, and the effective response range (ERR) **(bottom)**, under scenario 4. Columns show **(A)** the average width of credible intervals (CIw) across simulations, expressed in dB re 1μPa for μ and ω and in km for ERR, **(B)** the average absolute percent mean bias (PMB, in %), defined as (posterior mean-truth)/truth, and **(C)** the prior posterior overlap (PPO, in %). PPO values above 35% indicate that parameters may be non-identifiable (Gimenez *et al.* 2009).

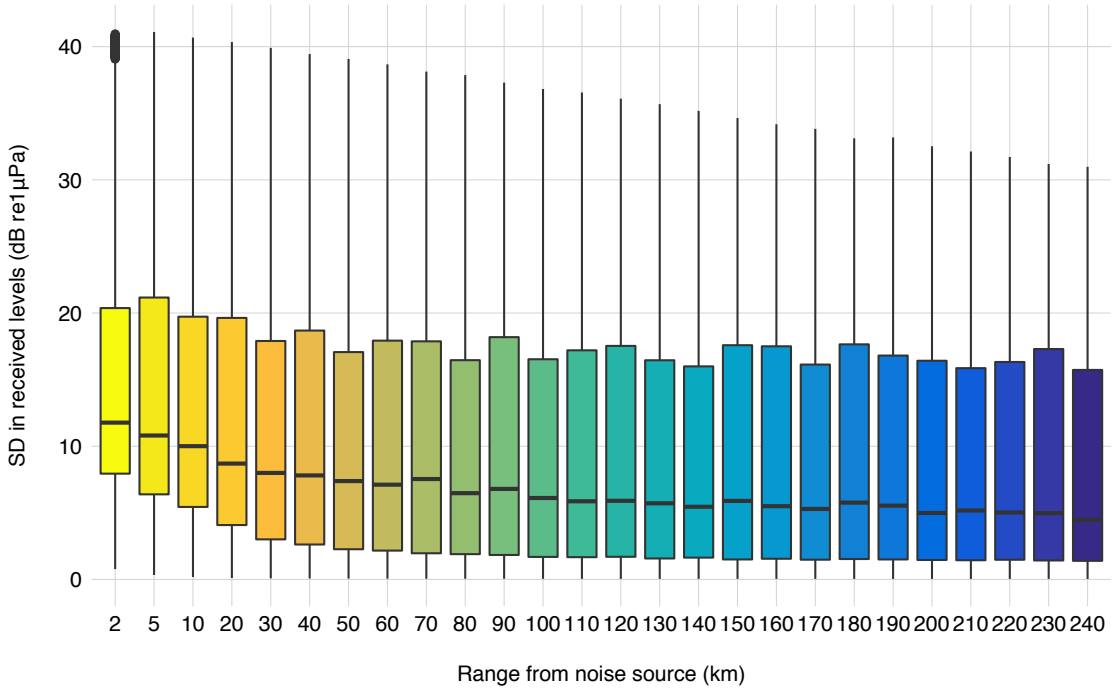

Figure 16: Variation in the estimated acoustic dose experienced by 2,000 simulated animals fitted with Argos-linked satellite tags at incremental distances from the sonar noise source. Values are expressed as the standard deviation (SD) in received levels across 10,000 candidate locations sampled within a plausible error ellipse around each individual, assuming an inverse-square circular transmission loss model (see Figure 3 for details).

6 Key messages

- The diagrams presented herein facilitate direct comparisons of experimental conditions (i.e. combinations of N x δ and N x P(SAT)), such that questions surrounding the optimisation of field protocols in BRSs can be readily answered.
- For instance, a similar level of confidence in estimates of the effective response range can be obtained from a sample of N = 10 whales with low measurement errors (δ = 2.5 dB re 1μPa, emulating the deployment of DTAGs) as from a sample of N = 40 whales with high measurement errors (δ = 35 dB re 1μPa, emulating satellite tags with consistently poor uplinks) (scenario 1).
- The larger the sample size, the larger the observation error (or the proportion of satellite tags) that is permissible to make inference of dose–response relationships with a desired level of certainty.
- Similar patterns of decreased precision, accuracy, and identifiability at smaller sample sizes and larger errors are observed across scenarios.
- The variance in response threshold is consistently more difficult to estimate than the mean, with posterior distributions undifferentiable from their associated priors in many cases.
- Conditional on the use of a circulation sound propagation model, substantial improvements to esti-

mates of variance can however be obtained with larger sample sizes. This is critical, as these parameters are key to determining dose–response relationships and therefore quantifying impact.

- When the observation model allows for tag-specific measurement errors (scenarios 3 and 4), sample size becomes the main driver of uncertainty in dose–response relationships, as satellite tags still allow measurements of received levels with relatively high precision.
- The effect of exposure history on expected response thresholds was seldom detected, whereas that of sonar frequency could be correctly identified in most cases when $N > 15$.
- The data presented in this report reflect the specific parameter values used in the simulations. Although these were chosen to be as representative of current BRSs as possible, it is important that users consider these results in light of their own study designs and experimental needs.

7 Conclusion

We demonstrated substantial improvements in the ability to estimate dose–response relationships and disturbance impacts at larger sample sizes, largely independently of the types/combinations of biotelemetry tags deployed on simulated whales. Despite heterogeneous measurement error, this suggests that previous efforts at deploying satellite tags have been valuable at furthering our understanding of cetacean responses to sonar, and should be pursued into the future.

8 Future work

A draft manuscript summarising the results of this work is currently being prepared for publication.

Bouchet PJ, Harris CM, Thomas L. (In prep). Optimising tagging programmes for understanding cetacean responses to military sonar exposure. Target journal: **Biology Letters** (https://royalsocietypublishing.org/journal/rsbl).

9 Acknowledgements

The research reported here was financially supported by the US Navy Living Marine Resources Programme (LMR) Contract No. N3943018C2080. We thank Rob Shick and the Atlantic BRS project for providing the Argos ellipse data required in scenarios 3 and 4. Support for that project is provided by Naval Facilities Engineering Command Atlantic under Contract No. N62470-15-D-8006, Task Order 18F4036, Issued to HDR, Inc.

10 Data availability

The R code used in the analyses is fully described and available at: https://github.com/pjbouchet/doublemocha_sim

11 References

Angela D'Amico DRK Robert C. Gisiner, Mead J (2009). Beaked whale strandings and naval exercises. *Aquatic Mammals* **35**, 452–472. DOI: 10.1578/AM.35.4.2009.452

Antunes R, Kvadsheim PH, Lam FPA, Tyack PL, Thomas L, Wensveen PJ, Miller PJO (2014). High thresholds for avoidance of sonar by free-ranging long-finned pilot whales (*Globicephala melas*). *Marine Pollution Bulletin* **83**, 165–180. DOI: 10.1016/j.marpolbul.2014.03.056

Beaumont MA (2010). Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics* **41**, 379–406. DOI: 10.1146/annurev-ecolsys-102209-144621

Berga AS, Trickey JS, Rice A, Širović A, Roch MA, Paxton CG, Oedekoven CS, Wiggins SM, Hildebrand J, Thomas L, Baumann-Pickering S (2019). Potential impact of mid-frequency active sonar on whales from passive acoustic monitoring data. *Journal of the Acoustical Society of America* **146**, 2939. DOI: 10.1121/1.5137205

Bouchet PJ, Harris C, Thomas L (2020). Simulating cetacean responses to sonar exposure within a Bayesian hierarchical modelling framework - R code description. Double MOCHA Report, University of St Andrews, 20 p.

Cholewiak D, Clark CW, Ponirakis D, Frankel A, Hatch LT, Risch D, Stanistreet JE, Thompson M, Vu E, Van Parijs SM (2018). Communicating amidst the noise: Modeling the aggregate influence of ambient and vessel noise on baleen whale communication space in a national marine sanctuary. *Endangered Species Research* **36**, 59–75. DOI: 10.3354/esr00875

Clark JS (2005). Why environmental scientists are becoming bayesians. *Ecology Letters* **8**, 2–14. DOI: 10.1111/j.1461-0248.2004.00702.x

Costa DP, Robinson PW, Arnould JPY, Harrison A-L, Simmons SE, Hassrick JL, Hoskins AJ, Kirkman SP, Oosthuizen H, Villegas-Amtmann S, Crocker DE (2010). Accuracy of ARGOS locations of pinnipeds at-sea estimated using Fastloc GPS. *PLoS One* **5**, e8677.

Cox T. M., Ragen TJ, Read AJ, Vos E, Baird RW, Balcomb K, Barlow J, Caldwell J, Cranford T, Crum L, D'Amico A, D'Spain G, Fernández A, Finneran J, Gentry R, Gerth W, Gulland F, Hildebrand J, Houser D, Hullar T, Jepson PD, Ketten D, MacLeod CD, Miller P, Moore S, Mountain D, Palka D, Ponganis P, Rommel S, Rowles T, Taylor B, Tyack P, Wartzok D, Gisiner R, Mead J, Benner L (2006). Understanding the impacts of anthropogenic sound on beaked whales. *Journal of Cetacean Research and Management* **7**, 177–187.

D'Amico A, Pittenger R (2009). A brief history of active sonar. *Aquatic Mammals* **35**, 426–434. DOI: 10.1578/AM.35.4.2009.426

de Quirós YB, Fernandez A, Baird RW, Brownell RL Jr., Soto NA de, Allen D, Arbelo M, Arregui M, Costidis A, Fahlman A, Frantzis A, Gulland FMD, Iñíguez M, Johnson M, Komnenou A, Koopman H, Pabst DA, Roe WD, Sierra E, Tejedor M, Schorr G (2019). Advances in research on the impacts of anti-submarine sonar on beaked whales. *Proceedings of the Royal Society B: Biological Sciences* **286**, 20182533. DOI: 10.1098/rspb.2018.2533

DeRuiter SL, Southall BL, Calambokidis J, Zimmer WMX, Sadykova D, Falcone EA, Friedlaender AS, Joseph JE, Moretti D, Schorr GS, Thomas L, Tyack PL (2013). First direct measurements of behavioural responses by Cuvier's beaked whales to mid-frequency active sonar. *Biology Letters* **9**, 20130223. DOI: 10.1098/rsbl.2013.0223

Dolman SJ, Evans PGH, Notarbartolo-di-Sciara G, Frisch H (2011). Active sonar, beaked whales and European regional policy. *Marine Pollution Bulletin* **63**, 27–34. DOI: 10.1016/j.marpolbul.2010.03.034

Dolman SJ, Jasny M (2015). Evolution of marine noise pollution management. *Aquatic Mammals* **41**, 357–374. DOI: http://dx.doi.org/10.1578/AM.41.4.2015.357

Dorazio RM (2016). Bayesian data analysis in population ecology: Motivations, methods, and benefits. *Population Ecology* **58**, 31–44. DOI: 10.1007/s10144-015-0503-4

Erbe C, Dunlop R, Dolman S (2018). Effects of noise on marine mammals. *In*: 'Effects of anthropogenic noise on animals', pp. 277–309 (Eds H. Slabbekoorn, R. J. Dooling, A. N. Popper, and R. R. Fay). Springer.

Erbe C, Marley SA, Schoeman RP, Smith JN, Trigg LE, Embling CB (2019). The effects of ship noise on marine mammals: A review. *Frontiers in Marine Science* **6**, 606. DOI: 10.3389/fmars.2019.00606

Erbe C, Reichmuth C, Cunningham K, Lucke K, Dooling R (2016). Communication masking in marine mammals: A review and research strategy. *Marine Pollution Bulletin* **103**, 15–38. DOI: 10.1016/j.marpolbul.2015.12.007

Falcone EA, Schorr GS, Watwood SL, DeRuiter SL, Zerbini AN, Andrews RD, Morrissey RP, Moretti DJ (2017). Diving behaviour of Cuvier's beaked whales exposed to two types of military sonar. *Royal Society Open Science* **4**, 170629. DOI: 10.1098/rsos.170629

Fernández A, Sierra E, Martín V, Méndez M, Sacchinni S, Quirós YB de, Andrada M, Rivero M, Quesada O, Tejedor M (2012). Last atypical beaked whales mass stranding in the Canary Islands (July, 2004). *Journal of Marine Science: Research & Development* **2**, 1–3. DOI: 10.4172/2155-9910.1000107

Filadelfo R, Mintz J, Ketten DR (2009). Correlating military sonar use with beaked whale mass strandings: What do the historical data show? *Aquatic Mammals* **35**, 435–444. DOI: 10.1578/AM.35.4.2009.435

Friedlaender AS, Hazen EL, Goldbogen JA, Stimpert AK, Calambokidis J, Southall BL (2016). Prey-mediated behavioral responses of feeding blue whales in controlled sound exposure experiments. *Ecological Applications* **26**, 1075–1085. DOI: 10.1002/15-0783

Gabry J, Mahr T (2019). Bayesplot: Plotting for bayesian models. R package version 1.7.0. Available at: https://cran.r-project.org/web/packages/bayesplot/index.html

Gimenez O, Morgan BJT, Brooks SP (2009). Weak identifiability in models for mark-recapture-recovery data. *In*: 'Modeling demographic processes in marked populations', pp. 1055–1067 (Eds D. L. Thomson, E. G. Cooch, and M. J. Conroy). Springer US. DOI: 10.1007/978-0-387-78151-8_48

Harris CM, Sadykova D, DeRuiter SL, Tyack PL, Miller PJO, Kvadsheim PH, Lam FPA, Thomas L (2015). Dose response severity functions for acoustic disturbance in cetaceans using recurrent event survival analysis. *Ecosphere* **6**, art236–14. DOI: 10.1890/ES15-00242.1

Harris CM, Thomas L, Falcone EA, Hildebrand J, Houser D, Kvadsheim PH, Lam F-PA, Miller PJO, Moretti DJ, Read AJ, Slabbekoorn H, Southall BL, Tyack PL, Wartzok D, Janik VM (2018). Marine mammals and sonar: Dose-response studies, the risk-disturbance hypothesis and the role of exposure context. *Journal of Applied Ecology* **55**, 396–404. DOI: 10.1111/1365-2664.12955

Harris CM, Thomas L, Sadykova D, DeRuiter SL, Tyack PL, Southall BL, Read AJ, Miller PJO (2016). The challenges of analyzing behavioral response study data: An overview of the MOCHA (Multi-study OCean Acoustics Human Effects Analysis) project. *In*: 'The effects of noise on aquatic life II', pp. 399–407 (Eds A. N. Popper and A. Hawkins). Springer, New York, NY.

Johnson MP, Tyack PL (2003). A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE Journal of Oceanic Engineering* **28**, 3–12. DOI: 10.1109/JOE.2002.808212

Joyce TW, Durban JW, Claridge DE, Dunn CA, Hickmott LS, Fearnbach H, Dolan K, Moretti D (2020). Behavioral responses of satellite tracked Blainville's beaked whales (*Mesoplodon densirostris*) to mid-frequency active sonar. *Marine Mammal Science* **36**, 29–46. DOI: 10.1111/mms.12624

King R, Morgan B, Gimenez O, Brooks S (2009). Bayesian analysis for population ecology. CRC Press, 456 p.

Klein JP, Moeschberger ML (2003). Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag, New York, 536 p. DOI: 10.1007/978-0-387-21645-4

Kruschke JK (2015). Doing Bayesian Data Analysis: A Tutorial with R and BUGS 2nd ed. Academic Press, Oxford, 759 p.

Laplanche C, Marques TA, Thomas L (2015). Tracking marine mammals in 3D using electronic tag data. *Methods in Ecology and Evolution* **6**, 987–996. DOI: 10.1111/2041-210X.12373

McClintock BT, London JM, Cameron MF, Boveng PL (2015). Modelling animal movement using the Argos satellite telemetry location error ellipse. *Methods in Ecology and Evolution* **6**, 266–277. DOI: 10.1111/2041-210X.12311

Miller PJO, Antunes RN, Wensveen PJ, Samarra FIP, Alves AC, Tyack PL, Kvadsheim PH, Kleivane L, Lam F-PA, Ainslie MA, Thomas L (2014). Dose-response relationships for the onset of avoidance of sonar by free-ranging killer whales. *Journal of the Acoustical Society of America* **135**, 975–993. DOI: 10.1121/1.4861346

Moretti D, Thomas L, Marques T, Harwood J, Dilley A, Neales B, Shaffer J, McCarthy E, New L, Jarvis S, Morrissey R (2014). A risk function for behavioral disruption of Blainville's beaked whales (Mesoplodon densirostris) from mid-frequency active sonar. *PLoS One* **9**, e85064. DOI: 10.1371/journal.pone.0085064

Pacini AF, Nachtigall PE, Quintos CT, Schofield TD, Look DA, Levine GA, Turner JP (2011). Audiogram of a stranded Blainville's beaked whale (*Mesoplodon densirostris*) measured using auditory evoked potentials. *Journal of Experimental Biology* **214**, 2409–2415. DOI: 10.1242/jeb.054338

Parent E, Rivot E (2012). Introduction to hierarchical bayesian modeling for ecological data. Chapman-Hall CRC, 427 p.

Parsons ECM (2017). Impacts of navy sonar on whales and dolphins: Now beyond a smoking gun? *Frontiers in Marine Science* **4**, 295. DOI: 10.3389/fmars.2017.00295

Plummer M (2019). Rjags: Bayesian graphical models using mcmc. R package version 4-9. Available at: https://CRAN.R-project.org/package=rjags

Plummer M, Best N, Cowles K, Vines K (2019). CODA: Convergence diagnosis and output analysis for mcmc. R package version 19-3. Available at: https://cran.r-project.org/web/packages/coda/index.html

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/

Schick RS, Bowers M, DeRuiter S, Friedlaender A, Joseph J, Margolina T, Nowacek DP, Southall BL (2019). Accounting for positional uncertainty when modeling received levels for tagged cetaceans exposed to sonar. *Aquatic Mammals* **45**, 675–690. DOI: 10.1578/AM.45.6.2019.675

Schorr GS, Falcone EA, Moretti DJ, Andrews RD (2014). First long-term behavioral records from Cuvier's beaked whales (*Ziphius cavirostris*) reveal record-breaking dives. *PLoS One* **9**, e92633. DOI: 10.1371/journal.pone.0092633

Simonis AE, BrownellRobert L Jr., Thayre BJ, Trickey JS, Oleson EM, Huntington R, Baumann-Pickering S (2020). Co-occurrence of beaked whale strandings and naval sonar in the Mariana Islands, Western Pacific. *Proceedings of the Royal Society B* **287**, 20200070. DOI: 10.1098/rspb.2020.0070

Southall BL, DeRuiter SL, Friedlaender A, Stimpert AK, Goldbogen JA, Hazen E, Casey C, Fregosi S, Cade DE, Allen AN, Harris CM, Schorr G, Moretti D, Guan S, Calambokidis J (2019). Behavioral responses of individual blue whales (*Balaenoptera musculus*) to mid-frequency military sonar. *Journal of Experimental Biology* **222**, jeb190637. DOI: 10.1242/jeb.190637

Southall BL, Nowacek DP, Miller PJO, Tyack PL (2016). Experimental field studies to measure behavioral responses of cetaceans to sonar. *Endangered Species Research* **31**, 293–315. DOI: 10.3354/esr00764

Stimpert AK, DeRuiter SL, Southall BL, Moretti DJ, Falcone EA, Goldbogen JA, Friedlaender A, Schorr GS, Calambokidis J (2014). Acoustic and foraging behavior of a Baird's beaked whale, *Berardius bairdii*, exposed to simulated sonar. *Scientific Reports* **4**, 1–8. DOI: 10.1038/srep07031

Tenan S, O'Hara RB, Hendriks I, Tavecchia G (2014). Bayesian model selection: The steepest mountain to climb. *Ecological Modelling* **283**, 62–69. DOI: 10.1016/j.ecolmodel.2014.03.017

Tyack PL (2008). Implications for marine mammals of large-scale changes in the marine acoustic environment. *Journal of Mammalogy* **89**, 549–558. DOI: 10.1644/07-MAMM-S-307R.1

Tyack PL, Thomas L (2019). Using dose–response functions to improve calculations of the impact of anthropogenic noise. *Aquatic Conservation: Marine and Freshwater Ecosystems* **29**, 242–253. DOI: 10.1002/aqc.3149

Tyack PL, Zimmer WMX, Moretti D, Southall BL, Claridge DE, Durban JW, Clark CW, D'Amico A, DiMarzio N, Jarvis S, McCarthy E, Morrissey R, Ward J, Boyd IL (2011). Beaked whales respond to simulated and actual Navy sonar. *PLoS One* **6**, e17009. DOI: 10.1371/journal.pone.0017009

Tyson RB, Friedlaender AS, Ware C, Stimpert AK, Nowacek DP (2012). Synchronous mother and calf foraging behaviour in humpback whales *Megaptera novaeangliae*: Insights from multi-sensor suction cup tags. *Marine Ecology Progress Series* **457**, 209–220. DOI: 10.3354/meps09708

Vincent C, McConnell BJ, Ridoux V, Fedak MA (2002). Assessment of ARGOS location accuracy from satellite tags deployed on captive gray seals. *Marine Mammal Science* **18**, 156–166. DOI: 10.1111/j.1748-7692.2002.tb01025.x

von Benda-Beckmann AM, Wensveen PJ, Prior M, Ainslie MA, Hansen RR, Isojunno S, Lam FPA, Kvadsheim PH, Miller PJO (2019). Predicting acoustic dose associated with marine mammal behavioural responses to sound as detected with fixed acoustic recorders and satellite tags. *Journal of the Acoustical Society of America* **145**, 1401. DOI: 10.1121/1.5093543

Weilgart LS (2007). The impacts of anthropogenic ocean noise on cetaceans and implications for management. *Canadian Journal of Zoology* **85**, 1091–1116. DOI: 10.1139/Z07-101

Wensveen P (2016). Detecting, asessing, and mitigating the effect of naval sonar on cetaceans. University of St Andrews.

Wensveen PJ, Isojunno S, Hansen RR, Benda-Beckmann AM von, Kleivane L, IJsselmuide S van, Lam F-PA, Kvadsheim PH, DeRuiter SL, CurC, Narazaki T, Tyack PL, Miller PJO (2019). Northern bottlenose whales in a pristine environment respond strongly to close and distant navy sonar signals. *Proceedings of the Royal Society B* **286**, 20182592. DOI: 10.1098/rspb.2018.2592

Williams R, Wright AJ, Ashe E, Blight LK, Bruintjes R, Canessa R, Clark CW, Cullis-Suzuki S, Dakin DT, Erbe C, Hammond PS, Merchant ND, O'Hara PD, Purser J, Radford AN, Simpson SD, Thomas L, Wale MA (2015). Impacts of anthropogenic noise on marine life: Publication patterns, new discoveries, and future directions in research and management. *Ocean & Coastal Management* **115**, 17–24. DOI: 10.1016/j.ocecoaman.2015.05.021

Youngflesh C (2018). MCMCvis: Tools to visualize, manipulate, and summarize mcmc output. *Journal of Open Source Software* **3**, 640. DOI: 10.21105/joss.00640

A Appendix A – Simulation plans

Tabular summaries of the simulation plans for each scenario are given below. **sim:** Unique simulation identification number. **N:** Sample size (number of animals). **Ns:** Number of simulations. **N(DTAG):** Number of animals fitted with digital tags. **N(SAT):** Number of animals fitted with Argos satellite tags. **P(SAT):** Proportion of animals fitted with Argos satellite tags. See Table 1 for details on model parameters.

A.1 Scenario 1

sim	N	μ	ω	δ	Covariates	Ns
1	5	165	30	2.5	None	500
2	5	165	30	5.0	None	500
3	5	165	30	10.0	None	500
4	5	165	30	15.0	None	500
5	5	165	30	20.0	None	500
6	5	165	30	25.0	None	500
7	5	165	30	30.0	None	500
8	5	165	30	35.0	None	500
9	10	165	30	2.5	None	500
10	10	165	30	5.0	None	500
11	10	165	30	10.0	None	500
12	10	165	30	15.0	None	500
13	10	165	30	20.0	None	500
14	10	165	30	25.0	None	500
15	10	165	30	30.0	None	500
16	10	165	30	35.0	None	500
17	15	165	30	2.5	None	500
18	15	165	30	5.0	None	500
19	15	165	30	10.0	None	500
20	15	165	30	15.0	None	500
21	15	165	30	20.0	None	500
22	15	165	30	25.0	None	500
23	15	165	30	30.0	None	500
24	15	165	30	35.0	None	500
25	20	165	30	2.5	None	500
26	20	165	30	5.0	None	500
27	20	165	30	10.0	None	500
28	20	165	30	15.0	None	500
29	20	165	30	20.0	None	500
30	20	165	30	25.0	None	500
31	20	165	30	30.0	None	500
32	20	165	30	35.0	None	500
33	40	165	30	2.5	None	500
34	40	165	30	5.0	None	500
35	40	165	30	10.0	None	500
36	40	165	30	15.0	None	500
37	40	165	30	20.0	None	500
38	40	165	30	25.0	None	500
39	40	165	30	30.0	None	500
40	40	165	30	35.0	None	500

A.2 Scenario 2

sim	N	μ	ϕ	σ	δ	β	α	Covariates	Ns
41	5	165	20	25	2.5	20	8	Frequency + exposure	500
42	5	165	20	25	5.0	20	8	Frequency + exposure	500
43	5	165	20	25	10.0	20	8	Frequency + exposure	500
44	5	165	20	25	15.0	20	8	Frequency + exposure	500
45	5	165	20	25	20.0	20	8	Frequency + exposure	500
46	5	165	20	25	25.0	20	8	Frequency + exposure	500
47	5	165	20	25	30.0	20	8	Frequency + exposure	500
48	5	165	20	25	35.0	20	8	Frequency + exposure	500
49	10	165	20	25	2.5	20	8	Frequency + exposure	500
50	10	165	20	25	5.0	20	8	Frequency + exposure	500
51	10	165	20	25	10.0	20	8	Frequency + exposure	500
52	10	165	20	25	15.0	20	8	Frequency + exposure	500
53	10	165	20	25	20.0	20	8	Frequency + exposure	500
54	10	165	20	25	25.0	20	8	Frequency + exposure	500
55	10	165	20	25	30.0	20	8	Frequency + exposure	500
56	10	165	20	25	35.0	20	8	Frequency + exposure	500
57	15	165	20	25	2.5	20	8	Frequency + exposure	500
58	15	165	20	25	5.0	20	8	Frequency + exposure	500
59	15	165	20	25	10.0	20	8	Frequency + exposure	500
60	15	165	20	25	15.0	20	8	Frequency + exposure	500
61	15	165	20	25	20.0	20	8	Frequency + exposure	500
62	15	165	20	25	25.0	20	8	Frequency + exposure	500
63	15	165	20	25	30.0	20	8	Frequency + exposure	500
64	15	165	20	25	35.0	20	8	Frequency + exposure	500
65	20	165	20	25	2.5	20	8	Frequency + exposure	500
66	20	165	20	25	5.0	20	8	Frequency + exposure	500
67	20	165	20	25	10.0	20	8	Frequency + exposure	500
68	20	165	20	25	15.0	20	8	Frequency + exposure	500
69	20	165	20	25	20.0	20	8	Frequency + exposure	500
70	20	165	20	25	25.0	20	8	Frequency + exposure	500
71	20	165	20	25	30.0	20	8	Frequency + exposure	500
72	20	165	20	25	35.0	20	8	Frequency + exposure	500
73	40	165	20	25	2.5	20	8	Frequency + exposure	500
74	40	165	20	25	5.0	20	8	Frequency + exposure	500
75	40	165	20	25	10.0	20	8	Frequency + exposure	500
76	40	165	20	25	15.0	20	8	Frequency + exposure	500
77	40	165	20	25	20.0	20	8	Frequency + exposure	500
78	40	165	20	25	25.0	20	8	Frequency + exposure	500
79	40	165	20	25	30.0	20	8	Frequency + exposure	500
80	40	165	20	25	35.0	20	8	Frequency + exposure	500

A.3 Scenario 3

sim	N	N(DTAG)	N(SAT)	P(SAT)	μ	ω	Covariates	Ns
41	5	0	5	100	165	30	None	500
42	5	1	4	80	165	30	None	500
43	5	2	3	60	165	30	None	500
44	5	3	2	40	165	30	None	500
45	5	4	1	20	165	30	None	500
46	5	5	0	0	165	30	None	500
47	10	0	10	100	165	30	None	500
48	10	2	8	80	165	30	None	500
49	10	4	6	60	165	30	None	500
50	10	6	4	40	165	30	None	500
51	10	8	2	20	165	30	None	500
52	10	10	0	0	165	30	None	500
53	15	0	15	100	165	30	None	500
54	15	3	12	80	165	30	None	500
55	15	6	9	60	165	30	None	500
56	15	9	6	40	165	30	None	500
57	15	12	3	20	165	30	None	500
58	15	15	0	0	165	30	None	500
59	20	0	20	100	165	30	None	500
60	20	4	16	80	165	30	None	500
61	20	8	12	60	165	30	None	500
62	20	12	8	40	165	30	None	500
63	20	16	4	20	165	30	None	500
64	20	20	0	0	165	30	None	500
65	40	0	40	100	165	30	None	500
66	40	8	32	80	165	30	None	500
67	40	16	24	60	165	30	None	500
68	40	24	16	40	165	30	None	500
69	40	32	8	20	165	30	None	500
70	40	40	0	0	165	30	None	500

A.4 Scenario 4

sim	N	N(DTAG)	N(SAT)	P(SAT)	μ	σ	ϕ	β	α	Covariates	Ns
91	5	0	5	100	165	25	20	20	8	Frequency + exposure	500
92	5	1	4	80	165	25	20	20	8	Frequency + exposure	500
93	5	2	3	60	165	25	20	20	8	Frequency + exposure	500
94	5	3	2	40	165	25	20	20	8	Frequency + exposure	500
95	5	4	1	20	165	25	20	20	8	Frequency + exposure	500
96	5	5	0	0	165	25	20	20	8	Frequency + exposure	500
97	10	0	10	100	165	25	20	20	8	Frequency + exposure	500
98	10	2	8	80	165	25	20	20	8	Frequency + exposure	500
99	10	4	6	60	165	25	20	20	8	Frequency + exposure	500
100	10	6	4	40	165	25	20	20	8	Frequency + exposure	500
101	10	8	2	20	165	25	20	20	8	Frequency + exposure	500
102	10	10	0	0	165	25	20	20	8	Frequency + exposure	500
103	15	0	15	100	165	25	20	20	8	Frequency + exposure	500
104	15	3	12	80	165	25	20	20	8	Frequency + exposure	500
105	15	6	9	60	165	25	20	20	8	Frequency + exposure	500
106	15	9	6	40	165	25	20	20	8	Frequency + exposure	500
107	15	12	3	20	165	25	20	20	8	Frequency + exposure	500
108	15	15	0	0	165	25	20	20	8	Frequency + exposure	500
109	20	0	20	100	165	25	20	20	8	Frequency + exposure	500
110	20	4	16	80	165	25	20	20	8	Frequency + exposure	500
111	20	8	12	60	165	25	20	20	8	Frequency + exposure	500
112	20	12	8	40	165	25	20	20	8	Frequency + exposure	500
113	20	16	4	20	165	25	20	20	8	Frequency + exposure	500
114	20	20	0	0	165	25	20	20	8	Frequency + exposure	500
115	40	0	40	100	165	25	20	20	8	Frequency + exposure	500
116	40	8	32	80	165	25	20	20	8	Frequency + exposure	500
117	40	16	24	60	165	25	20	20	8	Frequency + exposure	500
118	40	24	16	40	165	25	20	20	8	Frequency + exposure	500
119	40	32	8	20	165	25	20	20	8	Frequency + exposure	500
120	40	40	0	0	165	25	20	20	8	Frequency + exposure	500

B Appendix B – Directed acyclic graphs (DAGs)

Bayesian models can be viewed as networks of components, some of which are known and many unknown. Directed acyclic graphs (DAGs) are diagrammatical representations of these models, often used to summarise and communicate complex hierarchical structures (King *et al.* 2009). Within a DAG, the data, model parameters, and their corresponding prior distributions, are all represented as graphical nodes, inter-linked in a way that appropriately captures the directional relationships between variables.

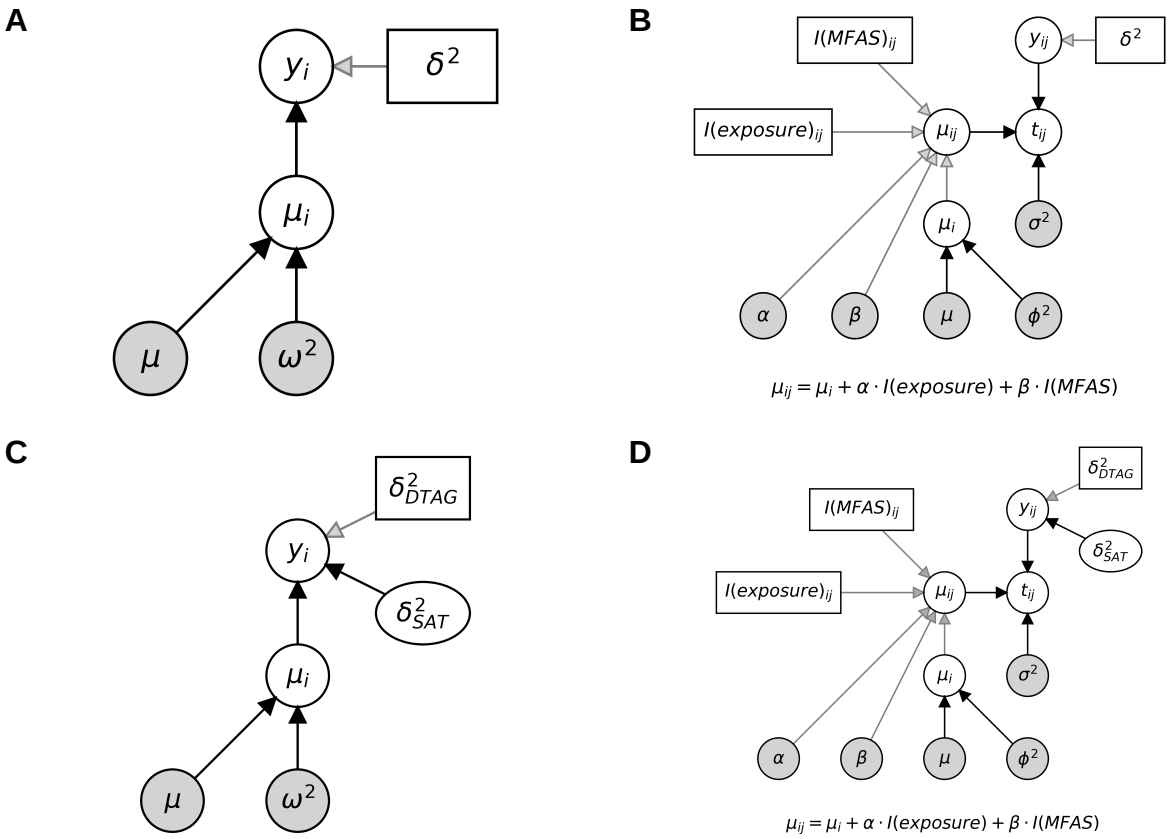

Figure B1: Directed acyclic graphs showing the structure of the four candidate hierarchical Bayesian models available under the simulation framework. **(A)** Scenario 1: no covariates, fixed uncertainty in dose measurements. **(B)** Scenario 2: probability of response affected by signal type (MFAS vs LFAS) and exposure history, fixed uncertainty in dose. **(C)** Scenario 3: no covariates, uncertainty in dose measurements varies by tag type. **(D)** Scenario 4: probability of response affected by signal type (MFAS vs LFAS) and exposure history, uncertainty in dose measurements varies by tag type. Model variables are represented by circles and constants by boxes. Grey circles denote variables monitored for posterior inference. Black and grey arrows represent stochastic and deterministic relationships, respectively.