Detecting changes in the distribution of calling bowhead whales exposed to fluctuating anthropogenic sounds

TRENT L. MCDONALD*, W. JOHN RICHARDSON+, CHARLES R. GREENE, JR.#, SUSANNA B. BLACKWELL#, CHRISTOPHER S. NATIONS*, RYAN M. NIELSON* AND BILL STREEVER†

Contact e-mail: tmcdonald@west-inc.com

ABSTRACT

This paper describes an analysis approach designed to detect the effects of fluctuating anthropogenic underwater sound on the distribution of calling bowhead whales (*Balaena mysticetus*) during migration. The anthropogenic sounds in this case were associated with an offshore oil production island (Northstar Island) in the Beaufort Sea northwest of Prudhoe Bay, Alaska, but the method has wider applicability. In autumn, bowhead whales migrate westward at varying distances offshore where some are exposed to Northstar sounds. Anthropogenic effects, if present, were hypothesised to be most pronounced in the southern (proximal) part of the migration corridor. Underwater sound levels were measured continuously *ca*. 500m from Northstar, and locations of calling whales were determined by a seafloor array of directional acoustical recorders. Weighted quantile regression related the 5th quantile of offshore call distance to anthropogenic sounds and other covariates. Case weights were inversely proportional to both probability of detection and location uncertainty. Due to potential dependencies in call locations, block permutation of uncorrelated whale call clusters was used to assign significance levels to coefficients in the quantile regression model. Statistical model selection was used to determine the anthropogenic sound measures most correlated with the 5th quantile of offshore call distances, after allowing for natural within-season variation quantified by day–night changes, distance of the call east or west of Northstar, and date. Data used to illustrate the method were collected over 29 days in September 2003 and included 25,176 bowhead calls. The estimated offshore distance of the 5th quantile call was 0.67km (95% confidence interval 0.31 to 1.05km) farther offshore when tones associated with Northstar were recorded in the 10–450Hz band during the 15 minutes just prior to each call. The method has been applied successfully to similar data collected near Northstar in other years, and may be useful in other studies that simultaneously collect data on animal locations and fluctuating stimuli.

KEYWORDS: ACOUSTICS; ARCTIC; BOWHEAD WHALE; MIGRATION; MODELLING; MONITORING; MOVEMENTS; NOISE; SURVEY-ACOUSTIC

INTRODUCTION

In autumn each year, bowhead whales (*Balaena mysticetus*) migrate west-northwest along the north coast of Alaska enroute to their over-wintering habitat in the Bering Sea (Moore, 2000; Moore and Reeves, 1993; Treacy *et al.*, 2006). In early 2000, an oil production island named Northstar was constructed in 12m of water *ca* 10km offshore and 20km west of Prudhoe Bay, Alaska, in the Beaufort Sea (Fig. 1). In a typical year, most bowheads travel westward more than 10km seaward of Northstar (Moore, 2000; Moore and Reeves, 1993; Treacy *et al.*, 2006), but occasionally bowheads have been observed <1km from Northstar. A whale within several kilometres of the island could be exposed to underwater industrial sounds, especially during periods of high island sound production or low ambient noise conditions (Blackwell and Greene, 2006). This raises concerns because underwater sound emanating from various other industrial activities (such as ship operations, marine seismic surveys, and offshore drilling) is known to displace some migrating whales (Richardson *et al.*, 1995).

Given both the bowhead's protected status under various environmental regulations, including the US Marine Mammal Protection Act and a local ordinance designed to address concerns of subsistence whale hunters in the Inupiat community, a monitoring study at Northstar was required. The overall objective of this monitoring study was to assess bowhead whale responses to sounds associated with Northstar activities. Previous measurements of underwater sounds near oil industry activities have shown that sound levels associated with activities on gravel islands are lower than those associated with drillships, dredges and seismic surveys to which bowhead whales sometimes react (Richardson *et al.*, 1995). The monitoring study at Northstar was designed to detect responses that heretofore would have been considered subtle.

Previous studies of whale deflection around anthropogenic sound sources have often focused on detecting deflection of individuals (Croll *et al.*, 2001; Malme and Miles, 1985; Richardson *et al.*, 1985; 1995). Some of these studies tracked individual whales, usually by visual means, as they passed a sound source, or as a sound source passed the whales. By comparing tracks with and without exposure to anthropogenic sounds, or by considering received sound levels, these studies sought to assess deflection. Other studies have used aerial surveys to look for locally-reduced animal densities near a sound source (e.g. Mobley, 2005; Richardson *et al.*, 1999). In fact, aerial surveys of waters surrounding the future Northstar site were conducted prior to 2000. However, in both types of studies sample sizes near the sound source were usually limited to (at most) tens of individuals due to difficulties sighting or following individual whales, inability to observe visually at night, weather limitations, etc. A power

* *Western EcoSystems Technology, Inc., 2003 Central Ave., Cheyenne, WY 82001, USA.*
+ *LGL Ltd., environmental research associates, 22 Fisher St., POB 280, King City, Ont. L7B 1A6, Canada.*
# *Greeneridge Sciences, Inc., 6160-C Wallace Becknell Rd., Santa Barbara, CA 93117, USA.*
† *BP Exploration (Alaska) Inc., 900 East Benson Blvd., P.O. Box 196612, Anchorage, AK 99519-6612, USA.*

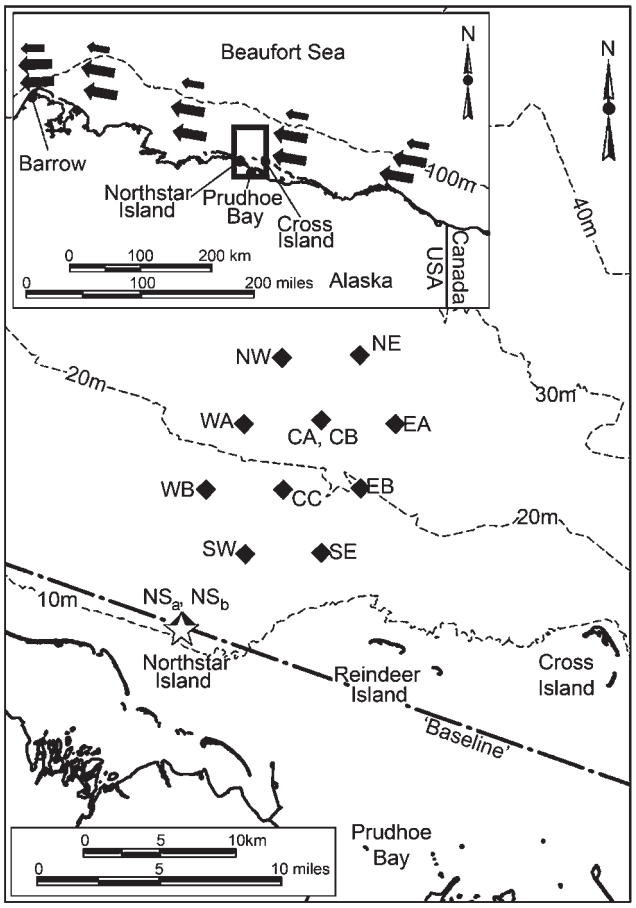

Fig. 1. Study area location in northern Alaska (box within inset), showing the main autumn migration corridor of bowhead whales (thick arrows); 100m depth contour is near the shelf-break. Detail shows DASAR locations (diamonds) within the study area northeast of the Northstar oil development. Two DASARs were located at the CA–CB location. Dashed WNW–ESE line through Northstar is the baseline from which perpendicular 'offshore distances' were measured.

analysis of prior aerial survey data from the Northstar area (T. McDonald, WEST, unpublished) indicated that sighting rates of >15 individuals per 1,000km of flight would be required to detect a 50% reduction in whale groups within a circle of radius 10 miles surrounding Northstar with 50% probability in 3 years. Because historical sighting rates in the general vicinity were lower than 15 individuals per 1,000km unless the migration corridor was unusually close to shore (Miller *et al.*, 1996) and 50% power to detect a 50% reduction was insufficient, it was apparent that the typical sample sizes in these types of studies would not yield the required level of sensitivity at Northstar.

The alternative approach used here focuses on call locations. This approach takes advantage of the fact that bowheads call frequently during both spring and autumn migration (Clark *et al.*, 1986; Moore *et al.*, 1989), and that these calls can be localised using directional hydrophones (Greene *et al.*, 2004). By associating changes in the location or shape of a spatial distribution of calls with changes in anthropogenic sound, certain types of disturbance effects can be investigated. Such an approach can take advantage of continuous acoustic monitoring, account for natural variation in the call distribution, and yield large sample sizes.

There are, however, three challenges associated with an approach based on call locations. The first applies to all studies that utilise call locations, while the other two apply to disturbance studies whose objectives are similar to those of this study. The first challenge is that call locations close in space and time are potentially dependent on one another when single whales call more than once or react to other whales, so individual calls are not the appropriate sampling units. Here, this dependency was addressed by applying a block permutation method (Lahiri, 2003). The second challenge is that anthropogenic noise could affect calling rate as well as whale location, which introduces an interpretational challenge that can be difficult to address if estimates of deflection *per se* are sought. This is not a problem if the study's objective is to detect and quantify disturbance-related changes in the distribution of calls, notwithstanding whether such changes arise from changes in whale location or calling rate or some combination of these and other causes. In other words, if anthropogenic noise causes the calling rate (or proportion of whales that call) to vary, changes in the spatial distribution of calls would be evident, but it would be impossible to determine from calls alone whether those changes are due to changes in calling behaviour or to physical displacement of whales or some other cause. In such cases, as here, results should be clearly understood to apply to calling whales rather than to all whales. The third challenge is, disturbance effects are expected to be most pronounced in animals nearest the sound source, and usually to diminish with increasing distance. Depending on industrial sound level at various positions across the width of the migration corridor, whales near the middle of the corridor may not be disturbed while those at the proximal edge may be affected. In other words, the proximal edge of a call distribution may shift but the centre may not. This challenge can be overcome by focusing on one or more quantiles in the call distribution. Recall that the xth quantile of a univariate distribution (here, offshore distance) is a value below which x% of the observations occur, and above which (100–x)% occur ('quantile' is synonymous with 'percentile', e.g. the median is the 50th quantile).

The specific objective of this paper is to develop a statistical approach suitable for the situation where effects of underwater sounds on call locations are of concern. The study at Northstar was the motivation for this approach. At Northstar, call locations were determined at times with varying levels of anthropogenic sounds measured near the island (~450m away). The approach then quantified the relationship between emitted anthropogenic sound (predictor variable) and the 5th quantile of offshore distances for local calls (dependent variable), after adjustment for other factors (covariates).

METHODS

In this study, there were two key types of data: whale call locations offshore of Northstar and underwater sound levels near the island. Call locations were estimated using data from 11 Directional Autonomous Seafloor Acoustic Recorders (DASARs) placed on the seafloor. Underwater sound levels from Northstar were monitored via hydrophones near the island. Previous publications have detailed the field methods, data collection, and data analysis through the call localisation stage (Blackwell and Greene, 2006; Blackwell *et al.*, 2007;

Greene *et al.*, 2004). Those papers describe DASAR design, construction, deployment, field calibration and retrieval, analysis of near-island sound recordings, call extraction and localisation, and general characteristics of the bowhead calls and migration corridor. The parts of those papers describing methods used to localise calls and measure Northstar sound levels are summarised in the next two subsections. The third subsection below describes the statistical approach used to relate the 5th quantile of offshore whale call distances to anthropogenic sounds and covariates. This paper focuses on methodology, and the methods are illustrated using data from one year (2003) of a longer study (Richardson *et al.*, In prep).

Whale call localisation

In 2003, whale calls were recorded continuously from 19:15 local time on 29 August to 04:39 on 28 September using an array of 11 DASARs deployed 6–21km offshore of Northstar Island (Fig. 1). The area where the DASARS were deployed was within the southern part of the bowhead migration corridor, although historically there has been substantial annual variation in that corridor (Moore, 2000; Treacy *et al.*, 2006). The main bowhead migration season typically extends from around 1 September into mid-October (Moore and Reeves, 1993). Because of deteriorating weather and concerns that they might become irretrievable under pack ice, the DASARs are retrieved as soon as possible after 25 September each year. Retrieval occurred on 28 September in 2003.

Each DASAR receiving a call provided a directional bearing to the call, with some uncertainty (Greene *et al.*, 2004). Calls were localisable when two or (preferably) more DASARs provided intersecting bearings for the same call. Precise DASAR orientations were determined by projecting calibration sounds from known (via GPS) locations around each DASAR. Calibration sounds were played at precisely known times on five dates (approx. weekly) during the 2003 field season. These data were used not only to calibrate each DASAR's orientation but also to correct for slight drift in each DASAR's internal clock. After correcting for clock drift, times of calls were determined to an accuracy of 1–2 sec, which was adequate to assess whether a call received at several DASARs represented a single call or multiple calls. DASARs provided reliable acoustic data up to 450Hz. Most of the energy in the great majority of bowhead calls is below 450Hz. The one exceptional call type ('high' calls) is rare and associated with complex calls, which contain energy below 450Hz (Würsig and Clark, 1993). Therefore, data up to 450Hz were deemed adequate for localisations.

The Huber robust location estimator was applied to triangulate call locations based on the intersection(s) of bearings from multiple DASARs (Greene *et al.*, 2004; Lenth, 1981). The Huber estimator down-weighted the occasional outlying bearing and yielded a location solution more often than alternative techniques. Calls could have been detected by only one DASAR, or missed completely, if the call was weak, occurred far from the DASAR array, or occurred during times when background levels of underwater sound (mainly due to wind and wave action) were high. Even when calls were received by ≥2 DASARs they occasionally did not produce a location estimate because estimated bearings either did not cross or were too disparate to allow the Huber estimator to converge.

For each estimated call location, a 90% confidence ellipse was calculated using methods in Lenth (1981). These methods were based on the number of DASARs that received the call, the geometry of all pair-wise bearing intersections, disparity of intersections, and inherent variation estimated from calibration data for each DASAR (Greene *et al.*, 2004).

Offshore distances were computed as perpendicular distances from the call's estimated location to a 'baseline' oriented 108° to 288° True (dashed WNW-ESE line in Fig. 1), through Northstar Island and parallel to the general trend of the coast. The Discussion section provides justification for using this measure of offshore distance and information about the lack of sensitivity of results to changes in orientation of the baseline.

Calibration sounds projected near the DASAR array, along with boat noise from the associated vessel, may have temporarily affected whale positions or calling behaviour. Because primary interest was in the effects of operations associated with Northstar itself, periods when the calibration boat was >2km north of Northstar Island, and periods within 2hr after the boat returned to waters <2km north of Northstar, were excluded from analysis. Two hours was chosen based on typical durations of avoidance reactions to boats (usually ½–1hr, Richardson *et al.*, 1985; Richardson and Malme, 1993), plus a 1–1½hr allowance for displacement and behavioural effects to subside. This provision resulted in exclusion of 8% (57.3hr of 705.4hr) of the 2003 field season and 1,506 localised calls.

Near-island sounds

Underwater sounds produced on the island and by associated vessels were measured 460m or 550m seaward (north) of the northern edge of the island either by a cabled hydrophone prior to its destruction by storm surge (31 August to 16 September 2003) or by a spare DASAR (18–28 September 2003). Both sensors were positioned just above the sea floor in water 12–13m deep (Blackwell and Greene, 2006). From the near-island recordings, sound spectral densities were determined for 1min periods every 4.37min, or ~330 times per day. These spectral densities were used to determine broadband (10–450Hz) and one-third octave band levels for each 1min sampling period. Totals of 5,262 and 3,232 1min samples were obtained from the cabled hydrophone and near-island DASAR, respectively. Because anthropogenic sound was not measured on 16–18 September 2003, *ca.* 2,827 calls recorded during this period were excluded from the analysis.

Near-island sounds received 460m and 550m north of the island were partly from industrial activities on the island, partly from vessels supporting Northstar activities, partly from wind and wave action (Blackwell and Greene, 2006), and partly from other sources. In 2003, broadband (10–450Hz) levels of underwater sound at this location ranged from 90.4 to 136.8 dB re 1 μPa and averaged 103.4dB. To measure sounds associated with industrial activities at Northstar, near-island sounds were quantified via the following five 'Industrial Sound Indices' (ISIs). These measures were later summarised and combined over varying time periods preceding each call (see Table 1) for inclusion as anthropogenic covariates in the analysis, as listed below.

(1) Sounds in five contiguous ⅓ octave frequency bands, centred at 31.5, 40, 50, 63 and 80Hz and spanning the 28–90Hz range, were predominantly associated with industrial activities at Northstar (Blackwell and Greene, 2006). However, some natural (e.g. wind and wave action) and non-Northstar (e.g. non-Northstar boats) sound did occur in these bands. The *isi*5 variable was defined to be the sum of the mean-square sound pressures in these five ⅓-octave bands, expressed in dB re 1 µPa. In 2003, this five-band ISI (28–90Hz) ranged from 84.5 to 128.8 dB re 1 µPa and averaged 97.6 dB (on average, 5.7dB less than the broadband (10–450Hz) sound pressure level).

(2) The near-island recording included prominent and recurrent tonal sounds in the 10–450Hz range at specific frequencies associated with industry activities (Blackwell and Greene, 2006). Tones occurred, for example, when engines generated sounds at constant frequencies. The 0–1 indicator variable *isi.tone.pres* was defined to be true if, during a 1min sample, sound spectral density in any 1Hz band was >5dB above the average spectral density in the four adjacent 1Hz bands (two below, two above, excluding the band being tested). In 2003, tones were present during 57% of the recorded 1min periods.

(3) The *isi.tone* variable quantified the strength of tones identified by the *isi.tone.pres* measure. When tones were not present in a 1min sample, *isi.tone* was 0. When tones were present, the strength of individual tones was mean-square sound pressure in the 1.7Hz wide Fourier analysis bin (centred on integer Hz) containing the tone minus the average mean-square sound pressure in the four adjacent bins (background noise). The strength of all tones in the 1min sample was the sum of tone strength (on mean-squared sound pressure scale) over all bands defined to have tones (see (2) above). When they occurred, tone strength ranged from 64.5 to 130 dB re 1 µPa and averaged 86.95 dB.

(4) Vessels routinely visited the island throughout the season, producing both tonal and non-tonal underwater sound. Vessel sounds tended to occur as transients lasting minutes to tens of minutes (Blackwell and Greene, 2006). The 0–1 indicator variable *isi.trans.pres* was true if, for a 1min sample, sound pressure (dB) in the 28–90Hz range was >5dB above sound pressures in these same bands averaged over the previous and subsequent 2h (i.e. a 4h moving average, excluding the 1min sample in question). In 2003, transients occurred during 10% of the recorded 1min periods.

(5) The *isi.trans* variable quantified the strength of transients identified by *isi.trans.pres*. Strength of transients in a 1min sample was 0 if no transients were present. When a transient was present, *isi.trans* was the difference between sound pressure (dB) in the 28–90Hz range for the 1min sample containing the transient minus that in the 4h moving average in these frequencies that was used to identify the transient. By construction, the minimum strength of transients was 5dB. When they occurred, transient strength averaged 10.2 dB re 1 µPa above the 4h moving average, and ranged to a maximum of 28.9 dB re 1 µPa above the moving average.

Analysis methods

This section describes estimation of a quantile regression relationship (Koenker, 2004; 2005; Koenker and Bassett, 1978; Koenker and Machoda, 1999; Koenker and Xiao, 2002) between the 5th quantile of offshore distances and anthropogenic sound after adjusting for certain environmental covariates. Conceptually, the quantile regression estimated a semi-linear model with functional form.

$$Q_5\,(y|x) = \beta_0 + f(\textit{non-industry variables}) + g(\textit{industry variables})$$

where $Q_5(y|x)$ was the 5th quantile of offshore distance given the values of all explanatory variables, *f(non-industry variables)* was a smooth function of naturally occurring exogenous variables that might be expected to influence offshore distance to calls, and *g(industry variables)* was a linear function of anthropogenic sound levels measured ~500m from Northstar (i.e. the ISIs). The remainder of this section describes exclusion criteria for calls, call weighting factors, model selection, computation of significance levels via block permutation, and estimation of anthropogenic effect size under various anthropogenic sound scenarios.

Call exclusion

Calls that occurred during times of high ambient (background) noise, e.g. during times of high wind and wave action, were more difficult to detect than calls occurring at other times. In particular, this caused calls originating outside the array and at large distances offshore to be underrepresented in the data during times of high background noise (Greene *et al.*, 2004). This bias in sampling, if not addressed, could have caused an apparent positive offshore displacement during low ambient noise periods, and conversely could have hidden a positive offshore displacement during high ambient noise times.

To eliminate this bias, an approach analogous to multiple observer distance sampling (Alpízar-Jara and Pollock, 1996; Buckland *et al.*, 2004, chapter 6; Good *et al.*, 2007) was adopted to estimate the probability of detecting and localising calls. Based on logistic regression models (see Appendix 1), calls were excluded if (for the circumstances of the particular call) the estimated probability of detecting and localising a call was below an arbitrary cutpoint, which was set at <10% (see next paragraph). The net effect was that calls within or close to the DASAR array (generally <10km from its centre) were included unless background sound levels exceeded ~108 dB re 1 µPa, as occurred during large storm events. Calls that occurred far from the array (e.g. >60km from array centre) were generally excluded regardless of background sound levels due to attenuation of the call's strength. In between, calls were included or excluded based on distance from the array, background sound at the time, and whether the call was east or west of the array (Appendix 1). The logistic regressions estimated that a call's probability of detection and localisation decreased as the call's distance from the array increased, or as background sound level increased, or both. Also, calls occurring east of the array were detected and localised with

slightly higher probability than calls west of the array (Appendix 1). The average location where probability of detection and localisation dropped below 50% was ~30km east, west and north of the centre of the DASAR array.

The call exclusion cutpoint was set at 10% for three reasons. First, a 10% cutpoint retained the vast majority of the detected calls (*ca* 90% were retained). Second, it was reasoned that most bias in the sample was represented by calls with small (<10%) probability of detection and localisation because, for every such call detected, the theorem of Horvitz and Thompson (1952) would indicate that >10 similar calls were missed. This contrasts with the fact that over 50% of the calls detected and localised were obtained in situations when detection and localisation probability was estimated to be >95%. Statistical theory implies that ≤0.053 similar calls were missed for every call detected with probability >95%. Third, during line transect studies that exclude distant sightings for similar reasons, a common criterion for exclusion is probability of detection less than 10% to 15% (Buckland *et al.*, 2001).

To assess whether exclusion of calls with <10% probability of detection affected the results, a sensitivity analysis was run. Following the full quantile regression analysis, the requirement that probability of detection and localisation be >10% was dropped and the entire analysis was re-run using all localised calls. For 2003 (and 2002), effects in the top ('best-fitting') quantile regression models were exactly the same whether or not the '<10% probability' calls were included. In those years, the direction, general magnitude and significance levels of coefficients in the two models were also the same. However, for 2001 and 2004, inclusion of low probability calls, primarily those with estimated locations >100km from the array, destabilised the estimation methods to the point that the quantile regression routine would not converge. The main analyses reported here exclude calls from situations with probability of detection and localisation <10%, which should reduce biases and allows the same procedure to be applied to all years.

Localisation uncertainty weights

Uncertainty in offshore distance measurements differed by several orders of magnitude among calls. To account for this, a weighting factor derived from the size of the 90% confidence ellipse for the call location was used in all quantile regressions. These weights were calculated as the reciprocal of error ellipse diameter along the 18º–198º axis, which was perpendicular to the 'baseline' that ran through Northstar roughly parallel to shore.

A small number of calls (~2%) were localised via 3 or more DASAR bearings that intersected at nearly a single point. When this happened, the estimated error ellipse was unrealistically small (e.g. <10m^2) given the uncertainties in individual bearings (Greene *et al.*, 2004). To keep these few calls from dominating the results, we replaced all confidence ellipse diameters less than the 2nd percentile of confidence ellipse diameters with the value of the 2nd percentile.

Detection probability weights

After excluding '<10% probability' calls, the remaining calls were in situations where detection and localisation probability was 10–99%. Detection and localisation probability was >95% for the vast majority of calls within 1–2km of the DASAR array perimeter, and was lower for most of those farther away. To account for differential probabilities of inclusion for calls remaining in the analysis, the quantile regression analysis included a weighting factor that was inversely proportional to the probability of detection and localisation under the circumstances of that call. This weighting factor was the Horvitz-Thompson (HT) weight for each call (see Buckland *et al.*, 2004, p. 9; Horvitz and Thompson, 1952; Särndal *et al.*, 1992, p. 43). HT weights have been used in similar situations (e.g. distance sampling) for the same purpose.

Because HT weights were estimated with statistical error, the overall quantile regression analysis was again re-run after the primary analysis was complete, this time without HT weights, to assess whether use of HT weights affected the results. Results were very similar with and without HT weights (see Results). The lack of sensitivity of results to HT weights was not surprising because location uncertainty weights were also included in the analyses. Location uncertainty weights dominated because they decreased faster than HT weights as distance from the array centre increased.

Model selection

Selection of variables for inclusion in the quantile regression model occurred in two stages. First, a reasonable model for *f(non-industry variables)* containing natural exogenous variables was determined. This 'natural variation' model explained as much variation in offshore distance as possible, given the available predictor variables. Then, models for *g(industry variables)* were added to the natural variation model and the additional predictive strength of the industrial sound variables was assessed. The remainder of this section is a description of these steps.

To start, quantile regression was used to identify the combination of four available non-industry variables (Table 2) that best predicts the 5th quantile of offshore distances. Backward stepwise elimination was used to select variables in the natural variation model. Starting with all four terms in the model, terms were successively removed if their *P*-values were greater than alpha-to-exit = 0.20. Between eliminations, previously deleted terms were restored if their *P*-values diminished below alpha-to-enter = 0.20. Elimination stopped when *P*-values for all terms in the model were below alpha-to-exit = 0.20. *P*-values were computed via block permutation, as described below.

Among the non-industry variables considered in step one, day of the year (0 = 31 August, 1 = 1 September, 2 = 2 September, etc.) and uprange distance (east–west distance of call along axis parallel to baseline) were fitted as nine degree-of-freedom smoothing splines (i.e. variables *dayofyear.smu* and *uprange.smu* in Table 2). This allowed estimation of non-linear and high order polynomial relationships between these variables and the 5th quantile of offshore distance. The degree of smoothness (number of 'anchors' or df) in both splines was chosen by generalised cross validation (Gu and Wahba, 1991; Gu and Xiang, 2001; Wood, 2004) in generalised additive models (Hastie and Tibshirani, 1990) relating mean offshore distance to day of year or uprange distance only.

Step two of model estimation started with the best fitting 'non-industry' model from step one, and successively evaluated 49 candidate models containing industrial sound indices arising from seven forms of anthropogenic sound (Table 1) crossed with seven possible averaging times for each ISI. Multiple ISI averaging times were considered because there was no *a priori* basis upon which to predict the most appropriate interval, from a bowhead whale's perspective, over which to average the sound measurements. In picking the range of averaging times to consider, it was reasoned that the 1min sound measurement closest in time to the call was unlikely to be adequate because disturbance effects, if present, would likely last longer than the 4.37min interval between successive sound measurements. In addition, if changes in the distribution of calling whales arise mostly from changes in location (displacement), responses of whales to Northstar sound would take considerably longer to develop than 4.37min. Typical swimming speed for a bowhead during autumn migration is 4–5km/h (Koski *et al.*, 2002), so whales take a few hours to travel through the area where the DASAR array could reliably detect and locate their calls. Likewise, averaging times greater than 2–3h were not likely to be adequate because a whale could receive and respond to multiple auditory events in such a long time interval. During pilot analyses after each year's data became available, averaging times of 5–160min were considered. From these analyses, it appeared that a Northstar effect, if present, would be strongest for averaging times between 15 and 120min. For the analyses reported here, the following seven averaging times were used: 15, 30, 45, 60, 70, 90 and 120min. ISI variables averaged over different time periods were not considered in the same model due to high correlation amongst them.

At the end of step two, the resulting set of 49 fitted models was ranked based on amount of variation explained. The model explaining the highest proportion of residual variation was selected as the best fitting model among those tested for the year in question. Akaike's Information Criterion (AIC) (Burnham and Anderson, 2004) was not used to rank competing models because AIC is a function of the maximised value of a statistical likelihood, and quantile regression is non-parametric so no statistical likelihood is defined. Following model selection, the significance of terms in the best model was determined by block permutation (described next).

Significance levels

Two difficulties prevent straightforward computation of significance for terms in the quantile regression models used here. First, the statistical properties of quantile regression parameters are not mathematically tractable (Bilias *et al.*, 2000; Hahn, 1995; Horowitz, 1998). This prevents use of a tabulated statistical distribution (such as the t or F distribution). Second, offshore distances were not independent of one another. For example, a particularly vocal whale could yield tens of calls but only one distinct measurement of offshore distance. Or, whales at multiple offshore distances could be calling in response to one another. This lack of independence prevented use of individual calls as the basis for statistical replication.

Given these complications, block permutation (Lahiri, 2003) was used to establish statistical significance levels. Block permutation is closely related to block bootstrap

Table 1

Industrial sound variables and models considered for inclusion in the quantile regression of offshore distances. In total, 49 models reflecting *a priori* notions of anthropogenic sound were considered: seven models × seven sound averaging times (XX = 15, 30, 45, 60, 70, 90 and 120min).

Model	Description
isi5.XX	Variable *isi5.XX* = sound level (in dB re 1 µPa) within the five 1/3rd octaves spanning 28–90Hz, averaged over the 1min samples within XX min immediately prior to the call. This model fit a linear relationship between *isi5.XX* and the 5th quantile of offshore distance.
isi.tone.pres.XX	Variable *isi.tone.pres.XX* =1 when at least one tone (>5dB above levels at neighbouring frequencies) was present at 10–450Hz in the nearshore sound record during XX min immediately prior to the call. *Isi.tone.pres.XX* = 0 when no tone was present during any 1min sampling times in the XX min period. This model estimated the average amount by which the 5th quantile of offshore distance increased or decreased when industrial tones were present prior to the call.
isi.trans.pres.XX	Variable *isi.trans.pres.XX* = 1 when at least one transient (>5dB above 4h running average background level) was present in the 28–90Hz band nearshore sound record during XX min immediately prior to the call. *Isi.trans.pres.XX* = 0 when no transient was present. This model estimated the average amount by which the 5th quantile of offshore distance increased or decreased when transient sounds of an industrial nature were present prior to the call.
isi.tone.pres.XX + *isi.tone.XX*	Variable *isi.tone.XX* = average strength of tones (on mean-square sound pressure scale) over the 1min samples defined to have tones within sample XX min immediately prior to the call. *Isi.tone.XX* = 0 when no tones were present during any 1min sample within XX min prior to a call. Strength of tone in a 1min sample was mean-square sound pressure in a 1.7Hz wide Fourier analysis bin (centred on integer Hz) minus average mean-square sound pressure in 4 adjacent bins (background noise). This interaction model fitted no relationship between *isi.tone.XX* and offshore distance when no tones were present, and a linear relationship when tones were present.
isi.trans.pres.XX + *isi.trans.XX*	Variable *isi.trans.XX* = sum of mean-square sound pressures of transient strength in all 1min samples defined to contain transients within XX min immediately prior to a call, converted to dB re 1 µPa. Transient strength was difference between sound pressure (dB) in the 28–90Hz band of a 1min sample containing the transient and a centred 4h moving average of sound pressure in the 28–90Hz band. *Isi.trans.XX* = 0 when no transients were present during XX min prior to a call. This interaction model fitted no relationship between *isi.trans.XX* when no transients were present, and a linear relationship when transients were present.
isi.tone.pres.XX + *isi.tone.pres.XX* * *uprange.smu*	This model fitted separate smoothed curves relating uprange distance and 5th quantile of offshore distance for times when tones were present in the previous XX min vs. not present.
isi.trans.pres.XX + *isi.trans.pres.XX* * *uprange.smu*	This model fitted separate smoothed curves relating uprange distance and 5th quantile of offshore distance for times when transients were present in the previous XX min vs. not present.

Table 2
Natural, or non-sound, variables considered for inclusion in the *f(non-industry variables)* portion of the quantile regression models.

Variable	Degrees of freedom	Description
sunlight	1	Day/night indicator: *Sunlight* = 1 if sun was above the horizon; *sunlight* = 0 if sun was below the horizon. Local sunrise and sunset times for Prudhoe Bay, AK, obtained from *http://www.sunrisesunset.com.*
upstream	1	East/west indicator: *Upstream* = 1 if location was on or east of a line extending through DASAR CB (Fig. 1) and Northstar (i.e. uprange distance >0). *Upstream* = 0 if location was west of this line (i.e. uprange distance <0).
uprange.smu	9	Smoothed (via B-spline) function of east-west distance along baseline, in meters. Computed based on distance from Northstar to the point on the baseline closest to the call, with call locations east and west of Northstar coded as positive and negative values, respectively. B-splines allowed estimation of piecewise cubic polynomials between nine 'anchors' (or 'knots', seven internal, two at extremes) spaced evenly from the lowest to the highest observed values of uprange distance. Number of 'anchors' was chosen by generalised cross validation (Wood, 2004) in a generalised additive model relating offshore distance to this variable.
dayofyear.smu	9	Smoothed (via B-spline) function of day of the year, coded as 1 September = 1, 2 September = 2, etc. Otherwise calculated as for *uprange.smu*.

methods (Fitzenberger, 1997; Lahiri, 2003), which have an established history of application in quantile regression for confidence interval construction. In this case, block permutation was used to establish the null distribution of the drop-in-dispersion F statistics (Cade and Richards, 2006) and confidence limits for coefficients in both the 'natural' and 'industrial' quantile regression models. Details of the drop-in-dispersion F test and derivation of confidence limits via block permutation appear in Appendix 2.

To apply block permutation, 'blocks' composed of calls belonging to independent groups of whales must be identified. Here, however, neither individuals nor pods could be identified, let alone pods that might be in communication with one another. Instead, uncorrelated 'blocks' of calls were sought and serve equally well in the method. Uncorrelated blocks of calls were constructed by the hierarchical clustering procedure described in Appendix 3. This procedure grouped calls close in space and time until the centroid locations and average arrival times of calls within groups were uncorrelated, as measured by Mantel's test.

All estimation and significance testing was performed using the R programming language (R Development Core Team, 2005) augmented with packages *quantreg, mgcv*, and *splines* (*http://cran.r-project.org/web/packages/*). *Quantreg* (version 3.85) performs quantile regression using a linear programming approach (Koenker and D'Orey, 1987). The *splines* package was used to compute B-spline orthogonal base transformations of the date and uprange distance variables. The *mgcv* package computed a generalised cross validation estimate for the number of knots (or df) in the B-spline transformations, which in turn determined their smoothness.

RESULTS

A total of 45,622 calls were received by the DASAR array during the 29.4-day recording period in 2003. Of these, 8,778 were received by only one DASAR (preventing triangulation) and 3,907 others could not be localised because the bearings involved were too disparate, leaving 32,937 localised calls. Of these, 1,506 were excluded because they were localised during times when this project's research boat was servicing the array, 3,428 were excluded because probability of detection in the prevailing circumstances was <10%, and 2,827 were excluded because corresponding measurements of industrial sounds (ISI) were missing (i.e. between the time when the cabled hydrophone was lost and installation of a DASAR near Northstar). This left 25,176 call locations in the quantile regression analysis. Fig. 2 shows estimated locations of most localised whale calls, excluding those estimated to be beyond the mapped area.

Calls were detected in 'pulses', both in time and in space, during each year of this study (Blackwell *et al.*, 2007). This was evident in plots of offshore distances as a function of date (Fig. 3). For example, most calls were detected 10km and farther offshore on 7 September 2003, but six days later, on 13 September, numerous calls were detected very close to shore (Fig. 3). Clustering of calls in time and space is consistent with numerous observations by both Inupiat whalers and researchers (Blackwell *et al.*, 2007, pp.260, 264). For purposes of statistical analysis, the 25,176 calls were grouped into 3,000 clusters (Appendix 3).

Considering non-anthropogenic variables only, the best-fitting quantile regression model contained *upstream* (β = –415m, 95% CI –740m to –96m, $P = 0.023$), *dayofyear.smu* ($P = 0.001$) and *uprange.smu* ($P = 0.001$). None of the coefficients in this model changed substantially when anthropogenic sound variables were added. For consistency and brevity, we focus on the models containing both natural variables and anthropogenic sound variables, and do not report the 18 coefficients for *dayofyear.smu* and *uprange.smu* in the natural variation model.

Anthropogenic sound quantified in 49 ways was added to the best natural model and the resulting models were ranked according to the amount of variation they explained. The top 25 of these 49 models are summarised in Table 3. For 2003, no single anthropogenic sound model stood out from others among the top 21 models in explaining variation in the 5th quantile of offshore distance. The proportion of variation explained by the 21st-ranked model was only 3.8% less than that for the top ranked model (Table 3). These top 21 models included all three single-variable measures of sound averaged over all seven averaging times that were considered. All these single-variable measures of sound were similarly effective in predicting the southern portion of the call distribution, and similar conclusions might be expected from any of these models. The sound measure coefficient in each of the top 21 models was positive, indicating that, regardless of the sound measure or assumed averaging time, the southernmost calling whales tended to be farther offshore when industrial sounds increased.

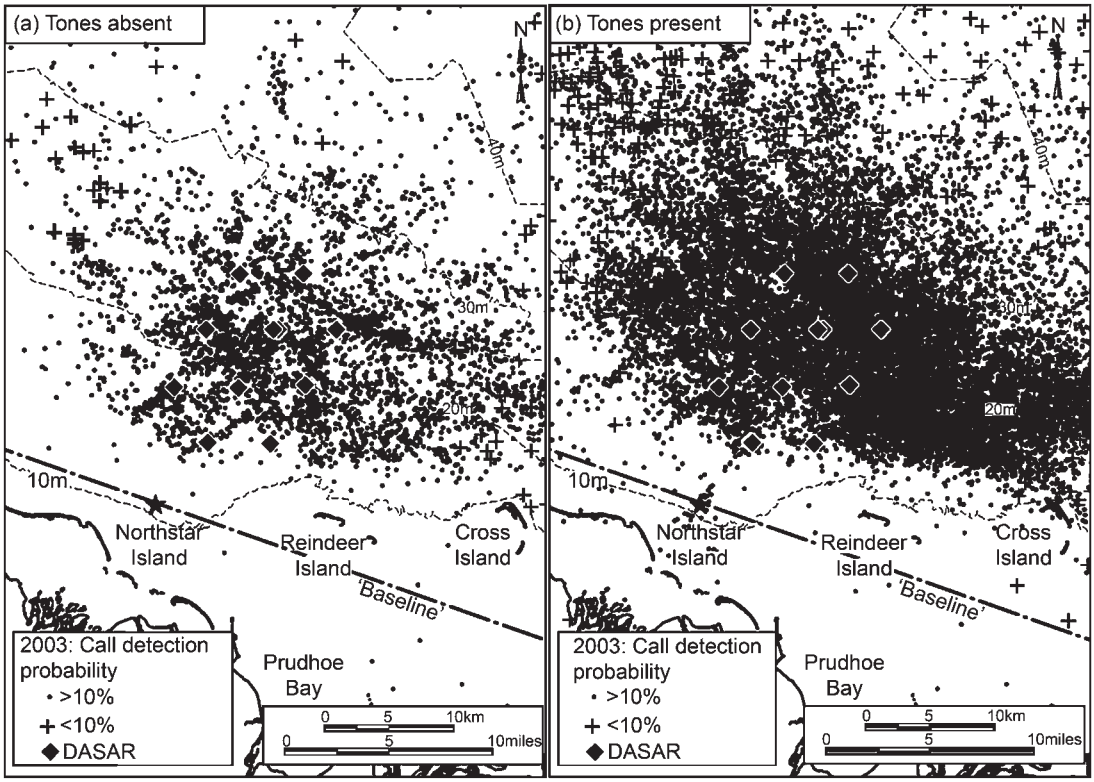

Fig. 2. Maps of estimated whale-call locations in 2003. Whale calls are distinguished according to the absence (a) or presence (b) of prominent tones near Northstar in the 15min period preceding each call [see Table 1 for definition of *tones present*]. Calls detected in situations where the probability of detection and localisation was <10% are also distinguished.

Although the top 21 models all had similar predictive abilities, the remainder of this section focuses on the 'best' predictor model because model averaging (Burnham and Anderson, 2004) is not possible without a likelihood-based criterion of model fit. The best-fitting (top) model allowed for *upstream, dayofyear.smu, uprange.smu* and *isi.tone.pres.15*. Coefficient estimates and confidence intervals for this model appear in Table 4. Each of these effects is described below.

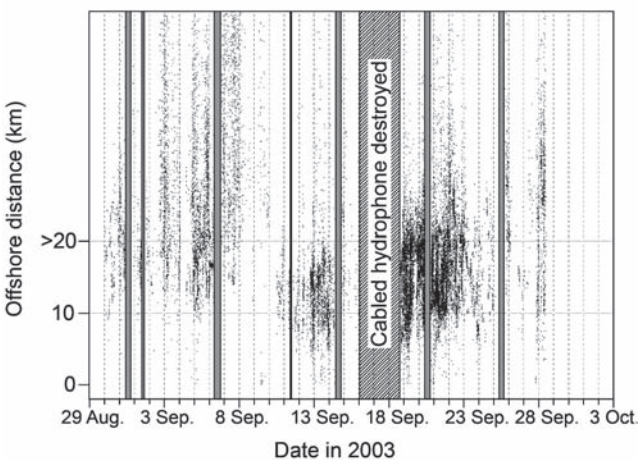

Fig. 3. Offshore distance for every detected whale call estimated to be within 50km of shore vs. date during the 2003 study period. All localised calls are included, regardless of probability of detection. Distances >20km offshore have large uncertainties and should only be used as an index of the frequency of whale calls far from Northstar. Shaded vertical segments delimit time slots when our vessel was in the DASAR array; whale calls during those periods were not analysed and were excluded from the graph. Calls arriving on 16, 17 and most of 18 September could not be used because storm surge destroyed the near-island recording equipment. Date labels appear at the start of each day (00:00 AkDT).

The seasonal variable *dayofyear.smu* included in the best model indicated that the normal (in the absence of prominent tones) southern edge of the distribution of whale calls varied substantially throughout the season. The 5th quantile of the offshore distances of calling whales ranged over time from 3.9 to 10.3km offshore of Northstar (Fig. 4a) when underwater sound near Northstar did not include any prominent industrial tones. The dates in 2003 when (in the absence of prominent tones) the 5th quantile achieved those minimum and maximum offshore distances were 16 and 4 September, respectively. These 5th-quantile offshore distances were determined at the study's centreline – a straight line through Northstar and DASAR CB.

The effects quantified by *uprange.smu* and *upstream* were both significant ($P \leq 0.001$) in the best fitting model. In general, the southern edge of the distribution of bowhead calls, as estimated by these two effects, was approximately parallel to the baseline and to the broad-scale trend of the coast within ~25km east and ~10km west of Northstar (Fig. 4). However, the overall trend of the 5th quantile deviated farther offshore ~10km downstream (west) of Northstar when compared to upstream of Northstar. The *upstream* effect estimated the 5th quantile to be 0.75km farther offshore west of Northstar than east (95% CI = 0.44 to 1.1km, Table 4). Relatively few call locations with high location accuracy were obtained >10–15km west of Northstar, so the 5th quantile estimates in this region were necessarily estimated with less precision than those within and nearer to the DASAR array.

Presence of a tone within 15min preceding the call (i.e. *isi.tone.pres.15*) was statistically significant at $P = 0.006$ in the best fitting model for bowhead call locations in 2003.

Table 3

Industrial effects in the top twenty-five 5th quantile regression models as ranked by proportion of variation explained. All models also included natural factors *upstream, dayofyear.smu,* and *uprange.smu* (see Table 2). Here, F_{effect} is the drop in dispersion statistic of Cade and Richards (2006) measuring the proportion of residual variation explained by adding the industrial sound term to the model. $\%\Delta F_{effect} = 100(\max(F_{effect})–F_{effect})/\max(F_{effect})$. *df* = number of coefficients estimated in the anthropogenic portion of the model.

Model	*df*	F_{effect}	$\% \Delta F_{effect}$
isi.tone.pres.15	1	3,875.1	0.0
isi.tone.pres.30	1	3,842.6	0.8
isi5.90	1	3,790.0	2.2
isi5.15	1	3,786.1	2.3
isi5.45	1	3,783.8	2.4
isi5.30	1	3,782.7	2.4
isi5.70	1	3,782.0	2.4
isi5.60	1	3,779.9	2.5
isi.tone.pres.45	1	3,778.7	2.5
isi.trans.pres.90	1	3,775.1	2.6
isi5.120	1	3,770.6	2.7
isi.tone.pres.70	1	3,761.9	2.9
isi.tone.pres.90	1	3,758.6	3.0
isi.tone.pres.60	1	3,758.1	3.0
isi.trans.pres.70	1	3,753.2	3.1
isi.trans.pres.15	1	3,750.2	3.2
isi.trans.pres.30	1	3,749.3	3.2
isi.trans.pres.60	1	3,746.8	3.3
isi.trans.pres.45	1	3,746.5	3.3
isi.tone.pres.120	1	3,737.7	3.5
isi.trans.pres.120	1	3,726.1	3.8
isi.tone.pres.30 + *isi.tone.30*	2	1,943.5	49.8
isi.tone.pres.15 + *isi.tone.15*	2	1,939.6	49.9
isi.tone.pres.90 + *isi.tone.90*	2	1,938.9	50.0
isi.tone.pres.70 + *isi.tone.70*	2	1,927.0	50.3

Table 4

Coefficients and 95% confidence intervals for effects in the best fitting 5th quantile regression model for data collected in 2003 relating offshore distance to natural and anthropogenic sound variables. Units of *upstream* and *isi.tone.pres.15* coefficients are metres. Coefficients for *dayofyear.smu* and *uprange.smu* are unitless due to B-spline transformation of these variables.

	2003		
Term	Coefficient	Low 95%	Upper 95%
Background Model			
(Intercept)	16,529.4	16,529.4	16,529.4
upstream	–749.7	–1,064.0	–439.7
dayofyear.smu.1	2,519.4	–2,119.6	7,317.2
dayofyear.smu.2	4,778.3	875.6	8,797.4
dayofyear.smu.3	–5,016.3	–9,066.4	–1,011.8
dayofyear.smu.4	–2,986.8	–6,212.7	86.9
dayofyear.smu.5	–179.2	–3,650.6	3,122.4
dayofyear.smu.6	–191.9	–3,244.8	2,872.8
dayofyear.smu.7	–592.2	–4,883.5	3,446.4
dayofyear.smu.8	–3,682.5	–9,033.1	2,472.1
dayofyear.smu.9	–1,052.1	–4,934.6	3,856.6
uprange.smu.1	689.3	–7,284.1	6,851.1
uprange.smu.2	–9,550.9	–14,446.6	–6,615.2
uprange.smu.3	–7,599.1	–13,093.0	–3,867.8
uprange.smu.4	–8,996.9	–14,277.9	–5,719.4
uprange.smu.5	–5,270.8	–10,565.2	–1,767.6
uprange.smu.6	–7,193.2	–12,578.6	–3,853.2
uprange.smu.7	–2,894.5	–8,425.8	900.0
uprange.smu.8	–10,020.6	–16,088.2	–4,778.5
uprange.smu.9	–14,029.2	–21,189.7	–7,179.4
Anthropogenic Model			
isi.tone.pres.15	666.9	309.9	1,053.9

The positive coefficient for *isi.tone.pres.15* indicated that the 5th quantile of offshore distance tended to be 0.67km (95% CI 0.31 to 1.05km) farther offshore when tones were present (Table 4). Fig. 4b plots the estimated 5th quantile of offshore distance for times with and without tones on a typical day of the season (21 September 2003). For comparison, with *isi.tone.pres.15* removed from the model, the predicted intersection of the 5th quantile and the centreline changed an average of 0.55km each day. Thus, the estimated anthropogenic effect (0.67km) was approximately equal to natural average daily changes, and was small when compared to the natural range of the 5th quantile (6.45km) observed during the entire season (Fig. 4b vs. 4a, Table 4). A similarly small but statistically significant anthropogenic effect was found in the autumns of 2001, 2002 and 2004 at times when levels of underwater sound near Northstar were elevated (Richardson *et al.*, In prep).

DISCUSSION

The primary goal of this analysis was to demonstrate a statistical method appropriate for detecting and quantifying effects of a specific source of anthropogenic sound on a distribution of calling whales measured via acoustic localisation. Data from a single year (2003) of a 4-year study focusing on Northstar Island in the Beaufort Sea are used to demonstrate the method. Results from all four years of the project, and a discussion of the biological implications of those results will appear elsewhere (Richardson *et al.*, In prep). Statistical issues surrounding the analysis are discussed here, while biological interpretation, importance, and management implications are addressed in the other paper. The statistical issues here centre on assumptions made during analysis and whether the analysis incorrectly detected an effect that was not actually present.

Throughout this discussion, it should be kept in mind that this was an observational study. An important assumption of observational studies is that all major sources of variation or disturbance are known, adequately measured, and correctly included in the appropriate models. This assumption becomes increasingly difficult to justify as the number of potential anthropogenic or natural effects increases. If nuisance variation is not adequately modelled, establishing the validity of primary effects can be difficult or impossible. Likewise, if multiple anthropogenic effects act cumulatively or interactively, quantifying the combination of factors that influence the primary response (here, the call distribution) may be difficult and never fully satisfactory. If either nuisance variation or an anthropogenic effect is not adequately modelled, the specific methods used here may not be adequate or may break down entirely. In this study, there is reason to believe that nuisance variation and anthropogenic effects were adequately modelled, as outlined below.

Overall design

In many studies designed to detect impacts of human activities, data from a reference or control area are compared to those from an impacted area both before and after the supposed impact (McDonald *et al.*, 2000). Such designs are efficient for detecting anthropogenic effects, but are difficult

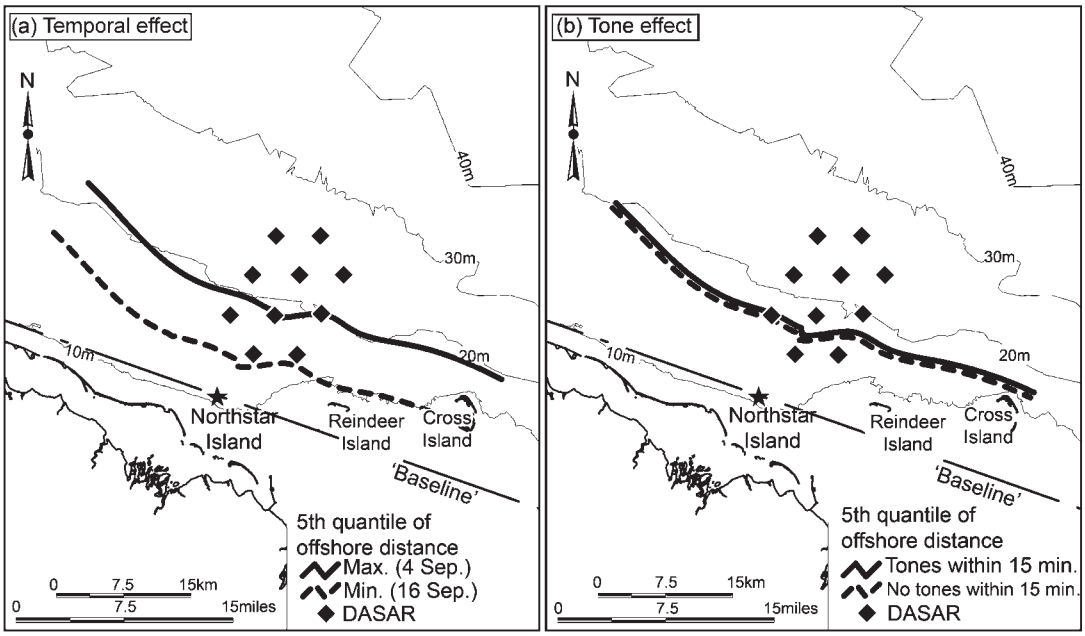

Fig. 4. Estimated 5th quantiles of offshore distance to whale calls at times in 2003 when (a) those quantiles were at their minimum and maximum distances offshore and (b) prominent anthropogenic tones were present vs. absent near Northstar. For (a), 5th quantiles were predicted by the final quantile regression model (Table 4) assuming no industrial sound effect (i.e. *isi.tone.pres.*15 = 0). For (b), 5th quantiles were predicted by the final quartile regression model for a typical day (*21 September 2003*; *dayofyear* = 21) with and without prominent underwater tones during the 15min interval immediately preceding the call. In (b), estimated displacement with tones = 0.67km (95% CI = 0.31 to 1.10km).

to apply in an observational setting such as this. In Northstar's case, no nearby location was entirely appropriate as a reference area due to varying physical conditions and varying amounts of human activity along the coast. These human activities included boat traffic, oil exploration, oil production and subsistence hunting. In addition, no comparable 'before Northstar' data were available. Consequently, the 'reference' condition used here consisted of times when less anthropogenic sound was being emitted from Northstar, rather than reference areas.

The overall design assumed that a dose-response relationship existed between whale behaviour and anthropogenic sound. In particular, the analysis assumed that a whale receiving enough anthropogenic sound would change its calling behaviour or offshore distance in a way that would affect the distribution of calls. The design also assumed that at least some whales in the southern (proximal) part of the migration corridor would detect Northstar sound at times when elevated levels of Northstar sound were measurable ~500m from Northstar. This was an appropriate assumption because, at times, Northstar-related vessel sound is detectable above background sound levels at distances as far as 27km offshore (Blackwell and Greene, 2006).

Interpretation of the response

Whether or not there is much physical displacement, it was recognised *a priori* that exposure of bowhead whales to Northstar sound might affect some aspect(s) of bowhead calling behaviour such as calling rate or source level of calls. Call types, frequencies, durations and received levels were logged during this study, and possible noise-induced changes in calling are being investigated (Blackwell *et al*., unpublished data). In the meantime, the present analysis does not attempt to differentiate actual displacement from effects on calling behaviour. A noise-related change in the distribution of bowhead calls represents a disturbance effect on some aspect(s) of bowhead whale behaviour regardless of the mechanism. Here, the two most likely mechanisms causing change in the call distribution are shifts in the physical distribution and changes in calling behaviour. However, Blackwell *et al*. (2012) found that bowhead calls were directional, thereby admitting the possibility that orientation of the individual could play a role in affecting the distribution of detectable calls. Regardless of the mechanism, identifying the presence and general magnitude of an anthropogenic effect is a useful step. Subsequent research should seek to identify the specific aspect(s) of behaviour that are subject to noise-induced effects.

Offshore distances

The responses analysed here were offshore distances, defined as the perpendicular distances of calling whales from a 'baseline' oriented parallel to the broad-scale alignment of the coast and of bowhead migration in autumn (Figs 1–4). However, bowhead headings in autumn are quite variable (e.g. Würsig *et al.*, 2002), leaving open the possibility that the average direction of travel could differ from the baseline orientation by as much as ±10º. To test whether choice of baseline orientation might have affected results, the baseline's orientation was changed by –10°, –5°, +5° and +10° (positive = counter-clockwise) and the significance of all terms in the best fitting model was re-computed. Under all four rotations of the baseline, all terms in the best fitting model remained significant at $P \leq 0.011$. Relative to its location when tones were absent, the 5th quantile of offshore distance with tones present was estimated to be displaced by 0.69, 0.68, 0.63 and 0.62km for rotations of –10º, –5º, +5º and +10º, respectively, as compared with 0.67km for 0º

rotation. The 95% CIs were 0.26 to 1.07km for –10°, 0.28 to 1.04km for –5°, 0.28 to 1.00km for +5°, and 0.26 to 0.97km for +10°, vs. 0.31 to 1.05km for 0º. Thus, the results are robust in relation to plausible changes in baseline orientation.

Radial distances

An obvious alternative to offshore distance as a response was the radial distance of calls from Northstar. Offshore distances were analysed because undisturbed bowheads are generally thought to migrate parallel to the coast, and offshore distances were thought to provide a more powerful and sensitive measure of displacement in this case. However, the predominant whale activity (i.e. migrating or feeding or milling, etc.) and the nature of the sound source (e.g. stationary or mobile) may affect which measure is most sensitive in other studies. Here, the question of sensitivity was largely moot because offshore distances were quite similar to radial distances in the relatively narrow east-west region where calls were located with sufficient precision to receive substantial weight in the analysis. Indeed, when the final quantile regression model was applied to radial distances, change in the 5th quantile of radial distance was 0.51km (95% CI = 0.22 to 0.81) when tones were present. These results are essentially identical to those for offshore distances (i.e. 0.67km, 95% CI = 0.31 to 1.05).

Linearity of effects

The relationship between offshore distance and certain natural covariates could not be assumed linear, and was estimated (within the quantile regression) via trend-following techniques, i.e. B-splines. However, for simplicity the relationship between offshore distance and anthropogenic sound levels was assumed to be either linear or discontinuous (i.e. on-off). This assumption was made because the goal was detection of any anthropogenic effects, not detailed characterisation of its functional form. Fitting a linear relationship between offshore distance and anthropogenic sound levels will detect changes under a wide variety of potential non-linear relationships. For example, a linear model should detect change if whales displace a fixed distance or stop calling altogether in response to levels of anthropogenic sound above some threshold (i.e. a step or threshold effect).

It is possible that both linear and discontinuous terms might fail to detect certain complex non-linear relationships. For example, there might be attraction or increased calling rate as sound level increased from low to moderate, but avoidance or reduced calling rate at the highest sound levels. Thus, in other studies it might be desirable to consider non-linear functions of sound level. In this study, it is conceivable that a complex effect could be missed by fitting a linear or on-off relationship. However, we detected an apparent relationship, and it is inconceivable how that could occur fallaciously by assuming a linear or on-off relationship.

Permutation blocks

Many passing whales were expected to emit a number of calls in succession, and this would cause statistical dependency in the locations (and resulting offshore distances) of individual calls. Because independent whales or independent whale pods could not be distinguished by their calls, it was necessary to find a proxy for dependent groups that would neither over- nor under-estimate the strength and statistical significance of anthropogenic sound effects. Hierarchical cluster analysis was used to group call locations in time and space until there was no measurable autocorrelation between cluster centroids (Appendix 3). These clusters were then used to assess significance during all quantile regressions. This technique allowed for interdependence of offshore distances within the identified clusters, but treated separate clusters as uncorrelated.

The clusters no doubt incorporated calls from single whales calling repeatedly, calls from different whales within pods, and calls from whales within different pods that sometimes were in acoustic contact with one another. In other words, the cluster analysis could have identified either more or fewer uncorrelated groups of whales than actually existed. If too few clusters were identified, i.e. if two or more independent groups of whales were sometimes unnecessarily combined into one cluster, the power of the study to detect anthropogenic sound effects would be reduced. If too many clusters were identified, i.e. if an interdependent group of calls was sometimes split into two or more clusters, the risk of incorrectly rejecting the null hypothesis could be larger than the nominal significance level (here 5%). Given that the current analysis provided evidence of an effect of anthropogenic sound on the offshore distribution of calls, the concern here is that too many clusters might have been identified. If so, sample size was overrepresented and the apparent effect of Northstar sound on offshore distances might have been false (spurious).

In actuality, the structure of some clusters suggests that the number of groups was lower than necessary. A small number of clusters spanned extremely long time periods (on the order of a week), and most of these clusters were small and centred far from the DASAR array. It is implausible that all whales within such ‘groups’ were somehow interdependent. Such clusters probably included multiple uncorrelated groups and should have been split into two or more clusters.

Nonetheless, it was of interest whether the apparent Northstar effect would disappear if fewer clusters were used in the block permutation procedure. To investigate this, clusters were sorted based on average arrival time of all calls in the cluster, and pairs of temporally adjacent clusters were amalgamated to reduce the number of clusters by 50%. The best-fitting model was then re-estimated and significance levels were re-computed. The significance level of all non-industrial terms in the best model remained <0.001 when half the number of clusters were used, and the significance of *isi.tone.pres.15* changed from $P = 0.006$ under the original clustering to $P = 0.009$ with ½ the number of clusters – within the error range of the permutation method. The estimate of displacement when anthropogenic tones were present within 15min preceding a call was unchanged, with slight variation in its confidence interval due to the stochastic nature of the permutation test (estimate = 0.67km with 95% CI 0.26 to 1.04km). Thus, even though the long time spans in some call clusters indicate that too few clusters may have been used in the main analysis, the results are robust in relation to uncertainty about the most appropriate number of clusters to use for block permutation.

Sound averaging time

Prior to data collection, there was no specific basis on which to predict the averaging time most relevant to bowhead whales (see Analysis Methods/Model Selection, much earlier). Based partly on a pilot analysis, the current analysis considered averaging times of 15, 30, 45, 60, 70, 90 and 120min preceding the call in question. A broader exploratory analysis seeking alternative acoustic measures and averaging times most strongly related to changes in the distribution of calls might be interesting, but further exploratory analysis was not undertaken here. The present analysis identified a measure of sound that was significantly correlated with offshore distances (*isi.tone.pres.15*; Table 4), along with other acoustic measures that were (for 2003) almost as closely related (Table 3). Further exploration and testing of the data would have exacerbated concerns about multiple testing issues and the possibility of a spurious result (see 'Overfitting and data mining?', below).

Model selection

The approach taken here allows, insofar as the data permit, for effects of natural environmental factors on the southern edge (5th quantile) of the distribution of calls offshore of Northstar. The method for selecting the best-fitting quantile regression model first incorporated a combination of non-industrial variables, and then assessed the ability of anthropogenic sound variables to explain remaining variation. This approach agrees with the usual ANOVA testing philosophy wherein the significance of the factor of primary concern (here anthropogenic sound) is assessed after accounting for variation explained by other factors. Allowance for the effects of natural covariates is expected to increase the power to detect and characterise the factor of main interest. However, with natural variables being fitted first, anthropogenic effects might appear insignificant if they were correlated with natural variables. For example, if industrial sound levels were higher during daylight than during night, and whales actually responded to industrial noise, adding *sunlight* to the model first could have masked the industrial effect. In this case, industrial sound levels and *sunlight* would be confounded. When a variable of interest is confounded with one or more other variables in an observational (uncontrolled) study, it is impossible to separate their effects by any analysis technique. Fortunately, in this study, anthropogenic sound measures showed no large correlations with natural variables that would indicate significant confounding or deleterious effects on interpretation. All model coefficients remained stable regardless of which other effects were included in the model.

Alternative model selection procedures might perform step-wise selection over all variables, not just natural ones, or might include more interactions among variables. This study incorporated a logical, constrained (non-open-ended) and repeatable model selection procedure that arrived at a useful model for detecting and characterising anthropogenic sound effects on the distribution of whale calls. An alternative model selection procedure utilising the same set of covariates might give a slightly different or refined picture of anthropogenic sound effects in 2003. However, given the similarities in goodness-of-fit for the 21 best-fitting models (Table 3), defensible alternative models utilising these measures of sound would almost certainly confirm the presence of a response to anthropogenic sound.

Overfitting and data mining?

As this analysis procedure was developed and refined, there was discussion of multiple comparison issues, possible overfitting, and the increased likelihood of spurious effects when data are 'mined' for significant effects. This issue is directly related to the 'experiment-wise' alpha level of the study and to the idea that, with $\alpha = 0.05$, we might expect 2–3 seemingly-significant tests among 49 (the number of anthropogenic sound models considered) by chance alone. Historically, these topics have been a source of much discussion in the statistics literature (see Hochberg and Tamhane, 1987; Saville, 1990; Tukey, 1994). One point of view is represented by Saville (1990) who argued that all testing procedures designed to protect experiment-wise significance levels are inconsistent except the unrestricted least significant difference (LSD) procedure (or multiple *t* test). Other researchers control multiple comparison problems by testing only a constrained set of hypotheses defined *a priori* (Burnham and Anderson, 2002; 2004). Others argue that a correction similar to Bonferroni's (Miller, 1981; Steel *et al.*, 1996) should be done whether or not hypotheses were defined *a priori.* Many would argue that all results, however derived, are unconfirmed until replicated by independent studies. Here, multiple testing problems were controlled by testing a constrained set of hypotheses, in large part defined *a priori*, about anthropogenic sound effects. However, the sound hypotheses tested were not strictly *a priori* because analysis procedures evolved over an extended period of data collection, preliminary analysis and peer review. In addition, pilot analyses were used to confirm that sound averaging times in the 15 to 120min range were reasonable.

The key question is whether the identified effects and model are real and likely to be replicated in subsequent studies. The authors offer the following five arguments that results of this study are robust and will be substantiated in future.

(1) It made sense *a priori* that some combination of the sound averaging times and anthropogenic sound measures would be related to displacement or changes in whale calling behaviour in the southern part of the migration corridor, if either were occurring.

(2) Previous disturbance studies, corroborated by pilot analyses of Northstar data, indicated that sound averaging times within the range considered here were reasonable.

(3) Several similar combinations of averaging time and sound measure were strongly related to offshore distances (Table 3); the chances that all these relationships were spurious are low.

(4) Results are robust to revision of several key analysis decisions (i.e. inclusion probability weights, baseline orientation, identified clusters, and radial distances).

(5) Separate applications of this method to data from 3 additional autumn migration seasons (2001, 2002, 2004) has found anthropogenic sound effects each year, although the specific measure of sound most closely associated with the effect was different each year (Richardson *et al*., In prep.).

Ultimately, verification (or otherwise) of a disturbance effect on the distribution of calling bowhead whales that receive relatively low levels of anthropogenic sound will come through additional data collection and replication. To help ensure future studies have similar or better power to detect the same sized effects, we recommend that (1) future studies focus on the most sensitive (proximal) portion of the spatial distribution (the southern edge of the migration corridor in this study), (2) additional covariates be considered where relevant, (3) whale identities be distinguished if possible, and (4) average calling rates for the population or (ideally) for individual whales be estimated if possible. If a future study has similar or higher power and is not confounded by the effects of additional factors (e.g. additional disturbance sources), it is reasonable to believe that the results described here will stand. If so, further work would be needed to determine whether the change in distribution of calling whales reflects a change in location of the whales, a change in calling behaviour, or both.

ACKNOWLEDGEMENTS

This work was funded by BP as part of their monitoring efforts at Northstar Island. The authors thank (in alphabetical order) Bill Burgess, Wilson Cullor, Ted Elliott, Allison Erickson, Tia Farmer, Ray Jakubczak, Jonah Leavitt, Bill McLennan, Bob Norman, Dave Trudgen, Mike Williams and Anne Wright for their help with the research and/or manuscript. The authors also thank Lisanne Aerts, Tom Albert, Richard Anderson-Sprechter, Robyn Angliss, Chris Clark, William Ellison, Craig George, Geoff Givens, Ken Hollingshead, John Kelley, Bryan Manly, Robert Suydam and Judith Zeh for their reviews and comments on various aspects of this work.

REFERENCES

Alpízar-Jara, R. and Pollock, K.H. 1996. A combination line transect and capture-recapture sampling model for multiple observers in aerial surveys. *Environ. Ecol. Stat.* 3(4): 311–27.

Bilias, Y., Chen, S. and Ying, Z. 2000. Simple resampling methods for censored regression quantiles. *J. Econometrics* 99(2): 373–86.

Blackwell, S.B. and Greene, C.R., Jr. 2006. Sounds from an oil production island in the Beaufort Sea in summer: characteristics and contribution of vessels. *J. Acoust. Soc. Am.* 119(1): 182–96.

Blackwell, S.B., Richardson, W.J., Greene, C.R., Jr. and Streever, B. 2007. Bowhead whale (*Balaena mysticetus*) migration and calling behaviour in the Alaskan Beaufort Sea, autumn 2001–04: an acoustic localization study. *Arctic* 60(3): 255–70.

Blackwell, S.B., McDonald, T.L., Kim, K.H., Aerts, L.A.M., Richardson, W.J., Greene, C.R., Jr. and Streever, B. 2012. Directionality of bowhead whale calls measured with multiple sensors. *Mar. Mammal Sci.* 28(1): 200–212.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. 2001. *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*. Oxford University Press, Oxford, UK. 432pp.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. 2004. *Advanced Distance Sampling: Estimating Abundance of Biological Populations*. Oxford University Press, Oxford, UK. 416pp.

Burnham, K.P. and Anderson, D.R. 2002. *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach*. 2nd ed. Springer-Verlag, New York. 488pp.

Burnham, K.P. and Anderson, D.R. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.* 33(2): 261–304.

Cade, B.S. and Richards, J.D. 2006. A permutation test for quantile regression. *J. Agric. Biol. Environ. Stat.* 11(1): 106–26.

Clark, C.W., Ellison, W.T. and Beeman, K. 1986. A preliminary account of the acoustic study conducted during the 1985 spring bowhead whale, *Balaena mysticetus*, migration off Point Barrow, Alaska. *Rep. int. Whal. Commn* 36: 311–16.

Conover, W.J. 1999. *Practical Nonparametric Statistics*. 3rd ed. John Wiley & Sons, New York. 592pp.

Croll, D.A., Clark, C.W., Calambokidis, J., Ellison, W.T. and Tershy, B.R. 2001. Effect of anthropogenic low-frequency noise on the foraging ecology of *Balaenoptera* whales. *Anim. Conserv.* 4(1): 13–27.

Fitzenberger, B. 1997. The moving blocks bootstrap and robust inference for linear least squares and quantile regressions. *J. Econometrics* 82(2): 235–87.

Good, R.E., Nielson, R.M., Sawyer, H. and McDonald, L.L. 2007. A population estimate for golden eagles in the western United States. *J. Wildl. Manage.* 71(2): 395–402.

Greene, C.R., Jr., McLennan, M.W., Norman, R.G., McDonald, T.L., Jakubczak, R.S. and Richardson, W.J. 2004. Directional frequency and recording (DIFAR) sensors in seafloor recorders to locate calling bowhead whales during their fall migration. *J. Acoust. Soc. Am.* 116(2): 799–813.

Gu, C. and Wahba, G. 1991. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Stat. Comput.* 12(2): 383–98.

Gu, C. and Xiang, D. 2001. Cross-validating non-Gaussian data: generalized approximate cross-validation revisited. *J. Comp. Graph. Stat.* 10(3): 581–92.

Hahn, J. 1995. Bootstrapping quantile regression estimators. *Economet. Theor.* 11(1): 105–21.

Hall, P. 1992. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York. 372pp.

Hastie, T.J. and Tibshirani, R.J. 1990. *Generalized Additive Models*, *Monographs on Statistics and Applied Probability. No. 43*. Chapman and Hall, London. 335pp.

Hochberg, Y. and Tamhane, A.C. 1987. *Multiple Comparison Procedures*. Wiley, New York. 450pp.

Horowitz, J. 1998. Bootstrap methods for median regression models. *Econometrica* 66(6): 1327–51.

Horvitz, D.G. and Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47(260): 663–85.

Koenker, R.W. 2004. Quantile regression for longitudinal data. *J. Multivar. Analysis* 91(1): 74–89.

Koenker, R.W. 2005. *Quantile Regression* Cambridge University Press, Cambridge. 349pp.

Koenker, R.W. and Bassett, G., Jr. 1978. Regression quantiles. *Econometrica* 46(1): 33–50.

Koenker, R.W. and D'Orey, V. 1987. Computing regression quantiles. *Appl. Stat.* 36(3): 383–93.

Koenker, R.W. and Machoda, J.A.F. 1999. Goodness of fit and related inference process for quantile regression. *J. Am. Stat. Assoc.* 94(448): 1,296–310.

Koenker, R.W. and Xiao, Z. 2002. Inference on the quantile regression process. *Econometrica* 70(4): 1583–612.

Koski, W.R., Thomas, T.A., Miller, G.W., Elliott, R.E., Davis, R.A. and Richardson, W.J. 2002. Rates of movement and residence times of bowhead whales in the Beaufort Sea and Amundsen Gulf during summer and autumn. pp.11-1 to 11-41. *In*: Richardson, W.J. and Thomson, D.H. (eds). *Bowhead Whale Feeding in the Eastern Alaskan Beaufort Sea: Update of Scientific and Traditional Information*. OCS Study 2002–012. US Minerals Management Service, Anchorage, AK and Herndon, VA. 697pp. [Available at: *http://www.boem.gov/.*].

Lahiri, S.N. 2003. *Resampling Methods for Dependent Data*. Springer, New York. 374pp.

Legendre, P. and Legendre, L. 1998. *Numerical Ecology*. Elsevier, New York. 853pp.

Lenth, R.V. 1981. On finding the source of a signal. *Technometrics* 23(2): 149–54.

Malme, C.I. and Miles, P.R. 1985. Behavioral responses of marine mammals (gray whales) to seismic discharges. pp.353–86. *In*: Greene, G.D., Engelhardt, F.R. and Paterson, R.J. (eds). *Proceedings of the Workshop on Effects of Explosives Use in the Marine Environment, 29–31 January 1985, Halifax*. Technical Report No. 5. Canada Oil and Gas Lands Administration, Environmental Protection Branch, Ottawa. 398pp.

Manly, B.F.J. 2005. *Multivariate Statistical Methods: a Primer*. Chapman and Hall, Boca Raton. 214pp.

Manly, B.F.J. 2007. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Chapman and Hall, Boca Raton. 455pp.
McDonald, T.L., Erickson, W.P. and McDonald, L.L. 2000. Analysis of count data from before-after control-impact studies. *J. Agric. Biol. Environ. Stat.* 5(3): 262–79.
Miller, G.W., Elliott, R.E. and Richardson, W.J. 1996. Marine mammal distribution, numbers and movements. pp.3–72. *In: Northstar Marine Mammal Monitoring Program, 1995: baseline surveys and retrospective analyses of marine mammal and ambient noise data from the central Alaskan Beaufort Sea*. LGL Rep 2101–2. LGL Ltd., King City, Ontario and Greeneridge Sciences, Inc., Santa Barbara, CA, for BP Exploration (Alaska) Inc., Anchorage, AK. 104pp. Available from Arctic Institute of North America Library, University of Calgary, Alb.
Miller, R.G., Jr. 1981. *Simultaneous Statistical Inference*. Springer-Verlag, New York. 299pp.
Mobley, J.R., Jr. 2005. Assessing responses of humpback whales to North Pacific Acoustic Laboratory (NPAL) transmissions: results of 2001–2003 aerial surveys north of Kauai. *J. Acoust. Soc. Am.* 117(3): 1666–73.
Moore, S.E. 2000. Variability of cetacean distribution and habitat selection in the Alaskan Arctic, autumn 1982–91. *Arctic* 53(4): 448–60.
Moore, S.E. and Reeves, R.R. 1993. Distribution and movement. pp.313–86. *In*: Burns, J.J., Montague, J.J. and Cowles, C.J. (eds). *The Bowhead Whale*. Special Publication No. 2. Society for Marine Mammalogy, Lawrence, KS. 787pp.
Moore, S.E., Bennett, J.C. and Ljungblad, D.K. 1989. Use of passive acoustics in conjunction with aerial surveys to monitor the fall bowhead whale (*Balaena mysticetus*) migration. *Rep. int. Whal. Commn* 39: 291–95.
Oksanen, J., Kindt, R. and O'Hara, R.B. 2005. *Vegan: community ecology package* (Vers. 1.6–10). Computer software to simulate vegetation data. [Available from: *http://cc.oulu.fi/~jarioksa/*].
R Development Core Team. 2005. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0 *http://www.R-project.org.*
Richardson, W.J. and Malme, C.I. 1993. Man-made noise and behavioural responses. pp.631–700. *In*: Burns, J.J., Montague, J.J. and Cowles, C.J. (eds). *The Bowhead Whale*. The Society for Marine Mammalogy, Lawrence, Kansas. 787pp.
Richardson, W.J., Fraker, M.A., Würsig, B. and Wells, R.S. 1985. Behaviour of bowhead whales *Balaena mysticetus* summering in the Beaufort Sea: reactions to industrial activities. *Biol. Conserv.* 32(3): 195–230.
Richardson, W.J., Greene Jr, C.R., Malme, C.I. and Thomson, D.H. 1995. *Marine Mammals and Noise*. Academic Press, San Diego. 576pp.
Richardson, W.J., Miller, G.W. and Greene, C.R., Jr. 1999. Displacement of migrating bowhead whales by sounds from seismic surveys in shallow waters of the Beaufort Sea. *J. Acoust. Soc. Am.* 106(4, Pt.2): 2281.
Richardson, W.J., McDonald, T.L., Greene, C.R., Jr., Blackwell, S.B. and Streever, B. In prep. Distribution of calling bowhead whales near an oil production island with variable underwater sound, 2001–2004. [Contact author for details].
Särndal, C.E., Swensson, B. and Wretman, J. 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York. 694pp.
Saville, D.J. 1990. Multiple comparison procedures: the practical solution. *Am. Stat.* 44(2): 174–80.
Steel, R.G.D., Torrie, J.H. and Dickey, D.A. 1996. *Principles and Procedures of Statistics: a Biometrical Approach*. McGraw-Hill, New York. 672pp.
Treacy, S.D., Gleason, J.S. and Cowles, C.J. 2006. Offshore distances of bowhead whales (*Balaena mysticetus*) observed during fall in the Beaufort Sea, 1982–2000: an alternative interpretation. *Arctic* 59(1): 83–90.
Tukey, J.W. 1994. *The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948–1983*. Chapman and Hall, New York. 300pp.
Ward, J.H., Jr. 1963. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58(301): 236–44.
Wood, S.N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.* 99(467): 673–86.
Würsig, B. and Clark, C.W. 1993. Behavior. pp.157–99. *In*: Burns, J.J., Montague, J.J. and Cowles, C.J. (eds). *The Bowhead Whale.* The Society for Marine Mammalogy, Lawrence, Kansas. 787pp.
Würsig, B., Koski, W.R., Thomas, T.A. and Richardson, W.J. 2002. Activities and behavior of bowhead whales in the eastern Beaufort Sea during late summer and autumn. pp.12–1 to 12–38. *In*: Richardson, W.J. and Thomson, D.H. (eds). *Bowhead Whale Feeding in the Eastern Alaskan Beaufort Sea: Update of Scientific and Traditional Information*. OCS Study MMS 2002-012; LGL Rep. TA2196-7. Rep. from LGL Ltd, King City, Ontario, for US Minerals Management Service, Anchorage, Alaska and Herndon, Virginia, USA. 420pp. [Available from *http://www.boem.gov*].

Date received: April 2009
Date accepted: January 2011

Appendix 1

DETERMINING PROBABILITY OF DETECTION AND LOCALISATION

Methods

Probability of detecting and localising a whale call was estimated using two logistic regression analyses. First, a logistic regression function was estimated to model probability of detection by two or more DASARs as a declining function of distance from the centre of the DASAR array, measured background sound level at the time, and whether the source was east or west of the array. Other variables considered for inclusion were distance uprange parallel to the baseline, distance offshore perpendicular to the baseline, and non-linear (quadratic and log) transformations of these two distances. Calibration sounds projected from known locations and received (or not received) at various DASARs (Greene *et al.*, 2004) were used to estimate coefficients of this regression. Second, another logistic regression estimated probability of localising a call given that it was detected by 2+ DASARs. (Detection by multiple DASARs did not guarantee a location estimate; inability to localise occurred primarily when bearings were highly disparate and non-crossing.) Variables considered for inclusion in the second regression were measured background sound level at the time, call type, number of DASARs detecting the call, time of day, average low frequency of the call, average high frequency, average duration (log transformed), average signal level, average signal-to-noise ratio, mean direction of bearings, dispersion among bearings, and the proportion of bearing intersections (out of $n(n - 1) / 2$ possible intersections, arcsin transformed). All calls received by 2+ DASARs, and whether or not each yielded a location estimate, were used to estimate coefficients of the second regression.

Background sound levels used in both logistic models were measured at the DASAR farthest from Northstar (NE; Fig. 1). Northstar sounds were on occasion received at DASAR NE, but were less likely to propagate to that location than to closer DASARs, and were weaker at NE. A measure of total underwater sound at NE, both natural and anthropogenic, was acceptable as a proxy for background sound in these analyses because anthropogenic sounds recorded at NE were intermittent and (when detected) weak. Calls were in fact often recorded and localised during times of strong industrial sound.

Variable selection for both logistic regressions was conducted by forward selection using the AIC criterion (Burnham and Anderson, 2004), i.e. terms were added to the model one-at-a-time until the AIC statistic increased. Following variable selection via stepwise AIC, a generalised

linear mixed model identical to the first regression was estimated to check for and account for potential dependencies in the detection of calibration sounds (the independence of localisations given detection, an assumption of the second regression, was clearly reasonable). Dependencies in detections at different DASARs could have been introduced by un-modelled environmental conditions or by the human operators processing the calls. The generalised linear mixed models were estimated using the SAS GLIMMIX macro. Error structures examined were compound symmetric, spatial power, spatial Gaussian, and spatial exponential errors, along with the uncorrelated (independent) structure. Fit of all the generalised mixed linear models was assessed using AIC, and the one with lowest AIC was deemed best.

Given probability of detection by a single DASAR, modelled via the first logistic regression, probability of detection by 2 or more DASARs was computed as 1 – Pr(detection by 1 DASAR) – Pr(detection by 0 DASARs). Probability of detection by 0 DASARs was computed as

$$\text{P(detection by 0 DASARs)} = (1-p_{i1})\times(1-p_{i2})\times\ldots\times(1-p_{in}),$$

where p_{ij} was the modelled probability of detecting call i on DASAR j. Probability of detection by exactly 1 DASAR was computed as

$$\begin{aligned}\text{P(detection by 1 DASAR)} = {} & p_{i1}\times(1-p_{i2})\times(1-p_{i3})\times\ldots\times(1-p_{in})+\\ & (1-p_{i1})\times p_{i2}\times(1-p_{i3})\times\ldots\times(1-p_{in})+\\ & \vdots\\ & (1-p_{i1})\times(1-p_{i2})\times(1-p_{i3})\times\ldots\times p_{in}.\end{aligned}$$

Given probability of detection by 2 or more DASARs and probability of obtaining a localisation given detection by 2+ DASARs, modelled via the second logistic regression, the joint probability of detecting and localising a call was computed as the product: Pr(detection and localisation) = P(detection at 2+ DASARs) P[(obtain a location)|(detection at 2+ DASARs)].

Results

The final forward step-wise logistic regression for probability of detection by a single DASAR was

$$\ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = 63.947 - 5.395ddl - 0.373ambsound - 5.950upstream + 0.798ddl\times upstream + 0.023ddl\times ambsound - 0.013ambsound\times upstream.$$

where *ddl* was the natural logarithm of distance to DASAR, *ambsound* was ambient noise level, and *upstream* was the indicator variable for whether the call originated to the east or west of the DASAR array. When non-independent error structures were allowed in this model (using GLIMMIX), the model assuming independence had lowest AIC. Even under high ambient-noise conditions, this model predicted high overall probabilities of detection by 2+ DASARs (> 70%) inside the DASAR array. When ambient sound levels were lower, predicted probability of detection by 2+ DASARs was higher in all areas. Although *upstream* was involved in two interaction terms, the coefficients of these interaction terms were small relative to other coefficients. Given this, the relatively large negative coefficient of *upstream* itself caused the estimated probability of detecting a call to the southeast of the centreline to be lower than that to the northwest.

Locations were not obtained for ~4.4% of calls detected by 2+ DASARs. Among calls detected by exactly 2 DASARs, the non-location rate was slightly higher at 7.5% of calls. The final forward step-wise logistic regression model for the probability of obtaining a location given detection by 2+ DASARs was

$$\ln\left(\frac{\pi_{l|d}}{1-\pi_{l|d}}\right) = -2.817 - 0.001xbar + 0.380ybar + 5.903intp - 0.535dur + 0.014s2n$$

where $\pi_{l|d}$ was the conditional probability of obtaining a location, the pair (*xbar, ybar*) was mean bearing direction, *intp* was proportion of bearing intersections (arcsin transformed), *dur* was call duration (logarithmic transformed), and *s2n* was average signal-to-noise ratio among the DASARs detecting the call. Most of the model's explanatory power was achieved through *intp*. AIC of the univariate model containing *intp* alone was very similar to AIC with all included variables (AIC = 7,217 with *intp* alone vs. 7,178 with all variables). All other univariate models were very poor predictors (AIC > 17,000 for all).

Appendix 2

DROP IN DISPERSION TEST VIA BLOCK PERMUTATION

The drop in dispersion test proceeded as follows. The quantile regression residual for the i-th call in a model containing all effects under consideration was defined as

$$r_i^f = y_i - \sum_{j=0}^{p}\hat{\beta}_j^f x_{ij}$$

where y_i was offshore distance of the i-th call, x_{ij} was the value of the j-th covariate for the i-th call, $\hat{\beta}_j^f$ was the estimated j-th covariate's quantile regression coefficient in the full model, and p was the total number of (non-intercept) covariates in the full model. Dispersion of the full model was

$$D_f = \sum_{i=1}^{n} w_i r_i^f(\tau - I(r_i^f < 0))$$

(Cade and Richards, 2006, Eqn. 2.1) where τ was the quantile of interest (i.e. 0.05), n was the total number of calls in the data base, w_i was the weighting factor inversely proportional to the i^{th} call's error ellipse width and probability of inclusion, and $I(r_i^f < 0)$ was an indicator function equalling one if $r_i^f < 0$ and 0 if $r_i^f \geq 0$. Estimates of the coefficients $\hat{\beta}_j^f$, were obtained by minimising D_f, so dispersion of the full model was simply the obtained minimum. The drop in dispersion test statistic was computed by dropping the covariate being tested from

the full model, recomputing coefficients and residuals of the now reduced quantile regression model (labelled $\hat{\beta}_j^r$ and r_i^r), and then recomputing the reduced value of dispersion D_r. The drop in dispersion test statistic was then

$$F_{effect} = \frac{D_r - D_f}{D_f}$$

(Cade and Richards, 2006, Eqn. 2.1).

To compute significance levels, the distribution of F_{effect} under the null hypothesis of no effect (i.e. when H_0: $\hat{\beta}_j^f = 0$ was true) was required. Following standard permutation testing methods (Manly, 2007), the null distribution of F_{effect} was constructed using random block permutations of the original data as follows. A large number (999) of null data sets with $\hat{\beta}_j^f$ exactly zero were obtained by randomly permuting blocks of partial residuals r_i^r, where blocks were defined by the hierarchical cluster analysis (Appendix 3), and associating them with un-permuted values of the explanatory variables. This permutation broke any association between responses and explanatory variables and assured that $\hat{\beta}_j^f = 0$ in every permuted data set, yet preserved any correlation of residuals that existed within the clusters. The full and reduced models were re-fitted to the randomly permuted residuals and F_{effect} for the term being considered was recomputed. The distribution of these 999 F_{effect} values plus the original F_{effect} represented the distribution of F_{effect} under the null hypothesis of no relationship. Significance of the term being considered was the number of F_{effect} greater or equal to the original F_{effect} out of 1,000, divided by 1,000.

Ninety-five percent confidence intervals for coefficients of g(*industry variables*) in the best fitting model were computed using Hall's percentile method (Hall, 1992, p.36; Manly, 2007, p.48). This method approximated the distribution of true errors in β_j, i.e., $\varepsilon = \hat{\beta}_j - \beta_j$, by the distribution of coefficients, $\hat{\beta}_j^*$, obtained by fitting the best model to randomly permuted blocks of residuals. Both the distribution of $\hat{\beta}_j^*$ and ε had zero means, and by construction, variation in the distribution of $\hat{\beta}_j^*$ approximated the variation in ε. To compute the confidence interval for β_j, the percentiles ε_L and ε_H were computed from the distribution of 999 coefficients obtained by block permutation such that

$$\Pr(\varepsilon_L < \hat{\beta}_j^*) = \alpha / 2$$

and

$$\Pr(\hat{\beta}_j^* < \varepsilon_H) = 1 - \alpha / 2 ,$$

where $\alpha = 0.05$. Assuming the distribution $\hat{\beta}_j^*$ of is a good approximation to the distribution of ε,

$$\Pr(\varepsilon_L < \hat{\beta}_j - \beta_j < \varepsilon_H) \approx 1 - \alpha$$

so the $100(1 - \alpha)$% confidence interval for β_j was

$$\hat{\beta}_j - \varepsilon_H < \beta_j < \hat{\beta}_j - \varepsilon_L .$$

Similarly, the $100(1 - \alpha)$% CI for displacement of quantiles when sound was above ambient was

$$\hat{D} - d_H < D < \hat{D} - d_L$$

where d_L and d_H were computed as the $(\alpha/2)$-th and $(1 - \alpha/2)$-th percentiles of displacements computed from the 999 sets of coefficients obtained via block permutation.

Appendix 3

HIERARCHICAL CLUSTER ANALYSIS TO DETERMINE BLOCKS

Hierarchical, agglomerative clustering (Manly, 2005) was performed to group whale call localisations within a given autumn until cluster centroids were uncorrelated in time and space. Clustering started with N clusters, where each localisation was its own cluster, and cycled through a total of $N–1$ iterations during which 2 clusters were merged to form a new cluster. During each iteration, Ward's algorithm (Ward, 1963) was used to determine which clusters were merged. At each iteration, space–time correlation among cluster centroids was calculated using Mantel's procedure (Legendre and Legendre, 1998), and agglomeration stopped when the Mantel Statistic was small and negative. The largest number of clusters with a negative correlation in space and time was chosen as the final clustering.

Mantel's procedure calculated the Spearman rank correlation coefficient (Conover, 1999) between corresponding elements of a $N \times N$ spatial difference matrix and an $N \times N$ temporal difference matrix. Unfortunately, it was not feasible to compute Mantel's statistics on more than ~6,400 clusters due to the large size of these matrices. When the number of calls was >6,400, a contiguous (in time) block of 5,000 clusters was randomly selected, Mantel's statistic was computed, and the average Mantel Statistic from 100 such randomly chosen (with replacement) blocks was used as the measure of correlation that stopped cluster agglomeration. All space–time coordinates were standardised prior to clustering by subtracting their mean and dividing by standard deviation (Manly, 2005).

Despite sub-sampling to compute Mantel's statistics, Ward's method could not be applied to data sets larger than ~6,400 observations (i.e. 2002–2004). Clustering was therefore performed separately on subsets of locations, where the subsets were chosen based on 90% error polygon size. To choose subsets, all localisations in a year were sorted based on 90% error polygon size, and contiguous blocks of 6,400 locations were taken as the subsets. As a check that sub-setting was not introducing correlation among calls in different subsets, the between-cluster and average within-cluster space-time correlations among all clusters in all subsets were calculated and observed to be a small negative number.

Clustering was accomplished using the contributed package CLUSTER (*http://cran.r-project.org/src/contrib/Descriptions/cluster.html*) and the *R* statistical software package (R Development Core Team, 2005). Computation of Mantel's statistic was accomplished in R using the contributed package VEGAN (Oksanen *et al.*, 2005).

In 2003, average space-time correlation prior to clustering was $\bar{r} = 0.163$. The 25,176 whale call localisations considered in 2003 were grouped into 3,000 clusters. The final between-cluster space-time correlation was –0.025, with average within-cluster space-time correlation of 0.089 (standard deviation = 0.48). The median distance in time between two localisations within a cluster was 13.9h.