Title

Decadal-scale frequency shift of migrating bowhead whale calls in the shallow Beaufort Sea

Permalink
https://escholarship.org/uc/item/4rc0p3jz

Journal
The Journal of the Acoustical Society of America, 142(3)

ISSN
0001-4966

Authors
Thode, Aaron M
Blackwell, Susanna B
Conrad, Alexander S
et al.

Publication Date
2017-09-01

DOI
10.1121/1.5001064

Peer reviewed

Decadal-scale frequency shift of migrating bowhead whale calls in the shallow Beaufort Sea

Aaron M. Thode, Susanna B. Blackwell, Alexander S. Conrad, Katherine H. Kim, and A. Michael Macrander

Citation: The Journal of the Acoustical Society of America **142**, 1482 (2017); doi: 10.1121/1.5001064

View online: http://dx.doi.org/10.1121/1.5001064

View Table of Contents: http://asa.scitation.org/toc/jas/142/3

Published by the Acoustical Society of America

Decadal-scale frequency shift of migrating bowhead whale calls in the shallow Beaufort Sea

Aaron M. Thode,[1,a)] Susanna B. Blackwell,[2] Alexander S. Conrad,[2] Katherine H. Kim,[2] and A. Michael Macrander[3]

[1]*Marine Physical Laboratory, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093-0238, USA*

[2]*Greeneridge Sciences, Inc., 90 Arnold Place, Suite D, Santa Barbara, California 93117, USA*

[3]*Shell Exploration and Production Company, 3601 C Street, Anchorage, Alaska 99503, USA*

(Received 19 March 2017; revised 9 August 2017; accepted 17 August 2017; published online 21 September 2017)

Automated and manual acoustic localizations of bowhead whale calls in the Beaufort Sea were used to estimate the minimum frequency attained by their highly variable FM-modulated call repertoire during seven westerly fall migrations. Analyses of 13 355 manual and 100 009 automated call localizations found that between 2008 and 2014 the proportion of calls that dipped below 75 Hz increased from 27% to 41%, shifting the mean value of the minimum frequency distribution from 94 to 84 Hz. Multivariate regression analyses using both generalized linear models and generalized estimating equations found that this frequency shift persisted even when accounting for ten other factors, including calling depth, call range, call type, noise level, signal-to-noise ratio, local water depth (site), airgun activity, and call spatial density. No single call type was responsible for the observed shift, but so-called "complex" calls experienced larger percentage downward shifts. By contrast, the call source level distribution remained stable over the same period. The observed frequency shift also could not be explained by migration corridor shifts, relative changes in call detectability between different frequency bands, long-term degradation in the automated airgun detector, physiological growth in the population, or behavioral responses to increasing population density (estimated via call density). © *2017 Acoustical Society of America*.

[http://dx.doi.org/10.1121/1.5001064]

[RAD] Pages: 1482–1502

I. INTRODUCTION

After summering in the eastern Beaufort Sea, the Bering-Chukchi-Beaufort (BCB) population of bowhead whales (*Balaena mysticetus*) begins its autumn westward migration in late August. Unlike the spring migration, the autumn migration takes place relatively close to the northern shores of Alaska (Moore and Reeves, 1993). During their travels the animals produce a wide variety of frequency-modulated (FM) and other signals that defy simple classification into specific call types (Ljungblad *et al.*, 1982; Clark and Johnson, 1984; Cummings and Holliday, 1987; Moore *et al.*, 2006; Blackwell *et al.*, 2007), but past work has roughly divided calls between "simple" FM calls and so-called "complex" calls (Blackwell *et al.*, 2007). These calls are distinct from so-called bowhead "song" produced during the winter season at more southern latitudes (Stafford *et al.*, 2008; Delarue *et al.*, 2009; Tervo *et al.*, 2009; Tervo *et al.*, 2011). While bowhead song appears to serve a reproductive purpose, the functional purpose of the call repertoire used during the migration remains largely unknown.

Between 2007 and 2014, the Shell Exploration and Production Company (SEPCO) commissioned Greeneridge Sciences Inc. (GSI) to deploy at least 35 seafloor acoustic recorders, divided among five sites in the coastal Beaufort Sea (Fig. 1). Over that period, hundreds of thousands of bowhead whale calls were recorded during each fall migration season. The motivation behind the effort was to evaluate the potential impact of airgun and other industrial sounds on bowhead whale behavior during their westward fall migration in the relatively shallow arctic waters off Alaska (Blackwell *et al.*, 2015).

The scale of the dataset, combined with a need for timely analysis, motivated the development of methods for automatically detecting, classifying, and localizing bowhead whale sounds, while exploiting the directional localization capabilities of the DASAR packages (Thode *et al.*, 2012). The results of the automated analysis have previously been used to track seismic airgun activity around the Beaufort Sea (Thode *et al.*, 2010), to determine that bowhead whales change their sound production rates in response to both nearby and distant airguns (Blackwell *et al.*, 2015), and to establish source levels and calling depth distributions of the migrating population (Thode *et al.*, 2016).

Here this 7-year automatically analyzed dataset, combined with supplemental manual analysis, is used to measure trends in the frequency content of the call repertoire. Section II describes the equipment used, along with the automated and manual call detection procedures, detailing specific efforts to ensure that seismic airgun signals were not mistakenly labeled as whale calls. The section also describes the construction and evaluation of statistical multivariate regression

[a)]Electronic mail: athode@ucsd.edu

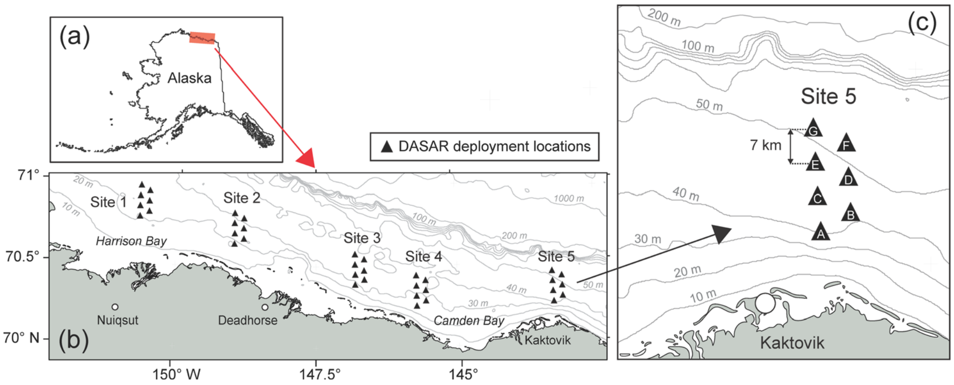

FIG. 1. (Color online) Location and bathymetry of DASAR deployments at Site 5, near Kaktovik, AK.

models for call frequencies. Section III uses descriptive statistics and regression models to observe a 7-year downward trend in the mean minimum frequency of the calls, a trend that persists even when ten other potential factors are incorporated into the regression model, including acoustic propagation factors and behavioral factors like call type. Section IV examines various possibilities for this observed 7-year shift, including shifts in propagation conditions (such as shifts in calling depth or distance offshore), relative differences in detection rates arising from differential changes in ambient noise levels across different frequency bands, increasing contamination by airgun signals arising from degradation in the automated detector, changes in the relative proportion of call types used by the population, physiological growth of the population, and behavioral responses to increasing population densities.

II. METHODS

A. Equipment and deployment configuration

The acoustic data for this study were recorded on Directional Autonomous Seafloor Acoustic Recorders (DASARs, model C) (Greene *et al.*, 2004), autonomous acoustic recording packages equipped with an omnidirectional acoustic pressure sensor (sensitivity of -149 dB re V/ 1 μPa) and two horizontal directional sensors capable of measuring the north-south and east-west components of acoustic particle velocity. This arrangement permits the azimuth of bowhead whale sounds to be measured from individual DASARs. Each time series is sampled at 1 kHz, but has a maximum usable acoustic frequency of 450 Hz, due to antialiasing filter rolloff. DASAR bearing precision, derived by comparing the active acoustic intensity measured along orthogonal directions, has been found empirically to be between 15° and 20° for signals with signal-to-noise ratios (SNR) of 5 dB or less, and between 1° and 2° for signals with SNR greater than 10 dB. Coincident bearings to calls detected on different DASARs are combined via triangulation to yield two-dimensional call positions, from which the range of each call to every DASAR can be estimated (Greene *et al.*, 2004).

From August to October 2007 to 2014, between 35 and 40 DASARs were deployed across a 280 km swath off the Alaskan North Slope, on the continental shelf in water depths between 20 and 53 m. The deployments were grouped into “Sites,” labeled 1–5 traveling from west to east (Fig. 1). Most sites contained seven DASARs deployed in a triangular grid with 7 km separation, and labeled “A” to “G” from south to north (Fig. 1). The analysis here focuses on data collected at both Sites 3 and 5, as these sites were located relatively distant from local seismic exploration activities (in 2008 and 2010). The analysis rejects data from the first year of the study, 2007, when a different type of sensor was used for the DASAR measurements.

B. Automated detection and localization

Each year bowhead whale calls in the raw acoustic data were post-processed by a six-stage automated detection and localization program, which has been extensively described and evaluated elsewhere (Thode *et al.*, 2012; Thode *et al.*, 2016). The algorithm basically conducts extensive preprocessing of the signals to remove regular pulses produced by seismic airgun surveys, and then uses image processing to extract 25 descriptive features from noise-equalized spectrograms, including the minimum and maximum frequency attained by a FM sweep. It then applies two cascaded feed-forward neural networks to winnow candidate detections based on these features. The rest of the algorithm associates call detections between DASARs to permit triangulation, as previously described in Greene *et al.* (2004) and Thode *et al.* (2012).

Certain relevant aspects of this algorithm are reviewed here, anticipating later discussion about whether errors and degradation in the automated detector’s performance are factors in the observed long-term trends presented in Sec. III. Specific topics of relevance include how airgun signals are identified, how the minimum frequency obtained by an FM sweep is measured, how the neural networks were trained to recognize calls, and how the automated analysis linked individual DASAR detections together to generate location estimates.

One early portion of the automated procedure attempts to identify and remove distant seismic airgun signals from further consideration. These signals have bandwidths and signal structure similar to bowhead whale FM downsweeps. Propagation effects can distort their time-frequency structure into a variety of range-dependent patterns with bandwidths very similar to those of low-frequency bowhead whale calls, which can make identifying and isolating these signals challenging. The automated procedure identifies these signals by flagging regularly timed events arriving from consistent azimuths. For each "target" event initially detected, the azimuths of 200 preceding and succeeding events are examined for those that lie with 15° of the target event and occur within 800 s of the target event. This set of surviving detections is then scrutinized for patterns of regular intervals, or "interval sequences," that lie between 5 and 42 s, a span that covers typical airgun firing intervals between 10 and 20 s. A given detection before or after the target event is assigned to an interval sequence if its arrival time fits within ±0.5 s of an arrival time predicted by a trial interval. If at least 12 out of 20 trial interval timeslots surrounding the target event match detections with appropriate bearings, then that trial interval is assigned to the target event, which is then relabeled as a "candidate" airgun event. The process is then sequentially repeated for all events. A second pass is then made through all candidate airgun events that have been assigned intervals. For each candidate detection, at least 7 out of 20 candidate detections preceding or succeeding it must share an interval that lies within 0.4 s of the given candidate's interval value, in order for the candidate detection to be classified as an airgun. This last step is performed to reduce removals of valid bowhead whale calls during times of heavy bowhead whale calling activity, when bursts of calls could arrive from similar azimuths, but at irregular timing intervals. The relatively ad hoc parameter values listed above permitted identification of airgun signals from up to four surveys simultaneously, but the algorithm was not foolproof, so long-term deterioration of this portion of the algorithm is a valid concern that will be addressed in Sec. IV.

An image-processing stage extracts descriptive signal features by applying various image processing techniques to spectrograms to identify complete time-frequency signal contours, and not just fragments of the most intense signal portions of the signal. Additional steps are used to assign harmonics and overtones to the fundamental. Various features are then extracted from the resulting time-frequency signal structures, emphasizing the lowest-frequency, or "fundamental," portion of the signal. The minimum frequency attained by the call's fundamental FM component (Fig. 2) is defined as the call's "minimum frequency" f_{min} for the rest of this paper.

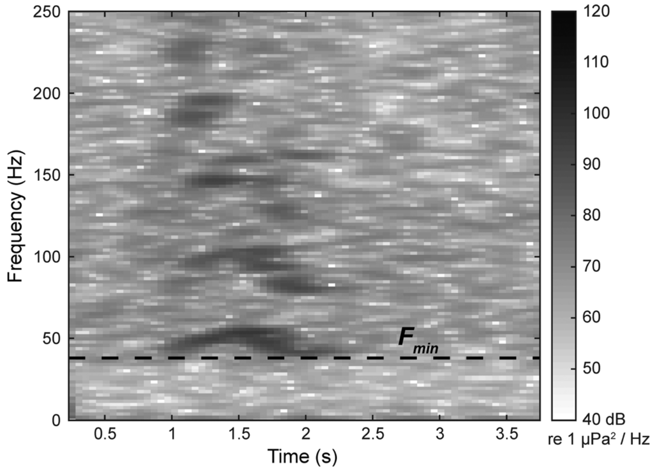

FIG. 2. Definition of the minimum frequency of a bowhead whale call, f_{min}.

The neural networks required training data provided by manual analyses in order to adjust the network weights and output thresholds. The training data were obtained by running the first three algorithm stages on subsets of acoustic data from 2008 and 2009 that had been reviewed manually for bowhead whale calls. A comparison of the automated results with the manual data divided the automated results into appropriate "whale" and "non-whale" classes, producing the training sets. Once the networks had been trained and their weights fixed, all stages were then applied to the complete 7-year acoustic data sets. Because the networks were trained only on data from the first two years of the program, some degradation in performance over time was anticipated, and thus supplementary manual analysis was conducted for each year in a manner discussed in Sec. II C. These later analyses were not used to retrain the networks, but to provide independent quality checks on the automated result statistics. Using the manual data from later years to retrain the networks would have blended the manual and automated datasets together, eliminating the ability to evaluate side-by-side comparisons between them. As will be seen in Sec. III B 1, the data sample sizes involved were so large that applying additional restrictive criteria was successful in combating neural network deterioration without compromising the statistical power of the manual data analyses.

Associating multiple detections on individual DASARs with a single call was challenging in this environment, because low-frequency whale and airgun signals dispersed over a few kilometers range, such that the signal structure of a given call differed between different DASARs, and simple cross-correlation techniques could not be used to match calls. A modified form of spectrogram correlation (Mellinger and Clark, 2000; Thode *et al.*, 2012) was employed to match similar images between DASAR receivers. Were some airgun signals to be mistakenly identified as whale calls, and then mistakenly matched with true whale calls on other DASARs, then an airgun signal could conceivably be linked to a position close to a DASAR. Section II D below details steps used to check that such inadvertent admissions had not occurred.

C. Manual analysis

As discussed in Sec. IV B, during each year of the project between 5 and 8 non-contiguous days of a deployment were selected for detailed manual analysis, with the intention of performing quality assurance checks of the automated analysis, which was always performed on the complete season. A team of roughly one to two dozen trained analysts used custom-written MATLAB software to review calls within the chosen dates. Each whale call was identified and classified by examining spectrograms of the acoustic data, 1 min at a time, and listening to recordings of each call or suspected call. The same minute recorded at each DASAR location was shown as a series of spectrograms on the analyst's screen. For each DASAR, an analyst drew a bounding box around every occurrence of a particular call. The bounding boxes allowed the software to calculate parameters such as the call's duration (width of the bounding box) and bandwidth (height of the bounding box), minimum frequency detected, received root-mean-square (rms) and sound exposure levels, SNR, and bearing, the last of which was used to triangulate the call position. Once all the calls included in the minute had been marked, the analyst moved on to the next minute of data. The lead analyst, who was the same person throughout all the years of the study, performed regular checks to maintain consistency among analysts.

The manual analysis also classified various frequency-modulated (FM) bowhead calls into "simple" FM modulated calls that displayed at most one inflection point (e.g., "upsweeps," "downsweeps," "constant," "n-shaped," and "u-shaped"), as well as a "complex" call category that incorporated any call that was not a simple FM sweep, including "pulsed sounds, squeals, growl-type sounds with abundant harmonic content, and combinations of two or more simple segments" (Blackwell *et al.*, 2007). As described later, these manual classifications were used to determine the statistical effect, if any, that call type had on the observed minimum frequency f_{min}.

A weakness in the manual analysis was that consistent procedures were never defined for treating signal harmonics. As a result some analysts drew bounding boxes around multiple harmonics, while others simply drew a box around the fundamental call component. Thus, when comparing automated to manual analysis, the metric of choice used here is the minimum frequency f_{min} attained by a call, since its selection criteria were consistent for both automated and manual analysis approaches. An additional weakness in the manual analysis is that the analyst software did not display spectrogram content below 50 Hz, creating biased estimates of f_{min} for low-frequency calls.

D. Data culling and quality assurance

The raw bowhead whale localization results produced by both types of analyses (automated and manual) were further culled to reduce any impact of potentially misidentified airgun signals or incomplete whale call detections on the observed call frequency distributions. In order to be included in the final analysis, a given bowhead call had to satisfy the following criteria, similar to those applied in Thode *et al.* (2016):

(1) The call had to be detected on three or more DASARs at ranges less than 50 km from every DASAR, and the call's localized range to the closest DASAR had to be less than a threshold value R_{min}. Had the algorithm accidentally localized distant airgun signals, the resulting localization ranges would have easily exceeded 50 km, and thus this thresholding step removed most possibilities of flagging airgun localizations as whale calls. Two values of R_{min}—2 and 15 km—were applied, values identical to those employed in Thode *et al.* (2016). The 2 km threshold was selected because previous analysis had shown that DASARs are effective in detecting calls made at ranges of 2 km or less, regardless of ambient noise conditions (Blackwell *et al.*, 2013; Blackwell *et al.*, 2015). The 15 km threshold, which roughly corresponds to the distance spanned by three in-line DASARs, was selected to increase the sample size available, as well as check to what degree the minimum frequency distribution of the call repertoire was influenced by R_{min}.

(2) A call's 90% confidence localization ellipse (Greene *et al.*, 2004) had to have the equivalent area of a 150 m-radius circle, or a mean radius less than ∼7% of the range to the nearest DASAR for the 2 km scenario.

(3) The call's f_{min} feature had to lie between 20 and 170 Hz, in order to remove the potential impact of call directivity on the analysis. A 170 Hz signal has an 8.5 m wavelength in arctic water (1450 m/s sound speed), roughly the length of a bowhead whale, so from a physical viewpoint frequencies below 170 Hz can be reasonably expected to be omnidirectional.

(4) The call's estimated source level (in terms of sound exposure level, or SEL) had to be within 6 dB of the source level computed from any other DASAR detecting the same call, a metric dubbed the "discrepancy" in Thode *et al.* (2016). This procedure provides a safeguard against the possibility that the automated algorithm captured only a fragment of a call, thus potentially missing the minimum frequency f_{min}. Had only a fragment of a call been captured on one or more DASARs, then the estimated SEL source level would vary between the DASARs.

Calls that survived the above criteria were assigned the value of f_{min} detected on the DASAR closest to the call's position.

E. Ambient noise analysis

As discussed in Sec. IV B, one possible explanation for long-term changes in the frequency content of the whale call repertoire is that the frequency-dependent ambient noise spectrum is evolving over time, changing the relative amount of masking (detectability) occurring for calls at different frequencies. To investigate this possibility, ambient noise properties were analyzed two ways.

The first approach created percentile distributions from continuous ambient noise data samples within two different frequency bands across all seasons. A set of 13 Fast-Fourier Transform (FFT) sample spectra were generated from a 2-s data segment (1 kHz sampling rate), using 512-point samples overlapped by 75%. The FFTs were converted into units of power spectral density (PSD), and then averaged together to produce one PSD estimate (periodogram) for each 2-s audio sample. The next PSD estimate retained the latter half of the original FFT samples, and then used one second of new data to generate the new PSD estimate. Thus, once an entire season was processed, a PSD estimate was produced every second.

For a given frequency, the corresponding PSD levels were then extracted and converted into a cumulative probability distribution for each season, from which percentiles could be derived. By using values from the 75th percentile and lower, potential contamination from whales and airguns (which only influenced the top 10% of PSD samples) could be removed from the ambient noise estimate. Section IV B examines percentile trends of ambient levels over all seven seasons, and estimates to what degree changes in relative ambient levels between two different frequencies could have shifted the observed distributions of f_{min} across multiple seasons.

The second approach to noise analysis, discussed in detail below, used short noise samples collected over a fixed bandwidth, just before times when whale calls were detected. These samples were then used in a regression analysis to determine the association between f_{min}, noise level, and signal-to-noise ratio.

F. Statistical regression

The large sample sizes yielded by both the manual and automated analyses ensured that most simple statistical hypothesis tests (e.g., Student's *t*, Kolmogorov-Smirnov, analysis of variance) consistently rejected the null hypothesis that the distribution or mean of f_{min} remained constant over time. Thus, more rigorous statistical multiparameter regression analyses on f_{min} were conducted using both generalized linear models (GLM) and generalized estimating equations (GEE; Dobson and Barnett, 2008) in order to examine effect sizes, confidence intervals, and the degree of interdependence between samples. Data from Sites 3 and 5 were analyzed both separately and together, but only the combined site analysis is presented here. Manual and automated detections using both $R_{min} = 15$ km and the more restricted dataset with $R_{min} = 2$ km were analyzed to check for modeling consistency, yielding a total of four datasets to which a given regression model was applied (i.e., $R_{min} = 2$, 15 km; manual or automated analysis).

Several potential explanatory variables were examined that could impact the measured f_{min}. Before conducting the regression, all predictor variables were examined for linear independence using both Pearson's correlation coefficients and variance inflation factors (VIF). Twelve different variables were examined:

(a) Year (Year, 1 = 2008, 2 = 2009, etc.) was treated as a continuous, and not a categorical variable, in order to estimate the f_{min} shift rate.

(b) Whale calling depth (CallingDepth, m) was computed using the methods described in Thode *et al.* (2016). Source depth has a large impact on frequency-dependent acoustic propagation in a shallow-water waveguide, so including this factor was a necessity.

(c) Whale range from closest DASAR (Range, km) was also tested, as different frequency components attenuate with range at different rates.

(d) Sample discrepancy (Discrepancy, dB re 1 μPa^2-s), as defined in Thode *et al.* (2016), was included to test f_{min} as a function of data quality. As discussed above, all datasets were restricted so that this factor could not exceed 6 dB.

(e) The median power spectral density (PSD) between 40 and 60 Hz was defined as NoiseBand1 (dB re 1 μPa^2/Hz). The 75 to 125 Hz band (NoiseBand2) was also examined. The full-bandwidth rms sound pressure level (SPL) was also computed across both bandwidths, to determine to what degree the choice of noise metric affected the regression result. As it turns out, the statistical analysis for both bandwidths and both metrics yielded similar results, so only the PSD of NoiseBand1 is discussed in detail below.

(f) The signal-to-noise ratio (SNR, dB), of the call sample was measured at the closest DASAR.

(g) Site (Site), a categorical variable, was assigned as a proxy for water depth and other geographical differences in acoustic propagation factors. For example, the median DASAR depth at Site 5 was 52 m vs 38 m for Site 3.

(h) Airgun, another categorical variable, was assigned a value of *true* whenever a given call was recorded during times when distant seismic airgun activity was detected, as reported by the airgun detection algorithm. An airgun survey was judged to be present if 15 airgun pulses were detected within 10 min on the DASAR closest to the call (e.g., 50% of airgun pulses from a survey with 20 s intervals).

(i) The Universal Transverse Mercator (UTM) Northing (Northing, km) was tested, a proxy for both latitude and offshore water depth, since the continental shelf monotonically deepens with increasing latitude [Fig. 1(c)]. This factor was included to account for the possibility that deeper waters would shift the optimum propagation frequency. The mean northing of all calls sampled at a given site was subtracted from each northing sample in order to reduce the effect of Site on this variable.

(j) A previous publication on long-term changes in the frequency content of blue whale calls (McDonald *et al.*, 2009) proposed that increases in animal density could prompt behavioral decreases in call frequency. Independent information on animal density was not available for this study; however, call localization rates were used as a proxy for population density, assuming that call rates detected on a DASAR would be similar to those detected by whales within a certain distance of the nearest sensor.

Thus, two potential proxies, CallRate and CallDensity, were defined and tested. CallRate is simply the raw number of localized whale calls detected per minute over a 75-min interval centered on a given call sample, regardless of the distance of the call from the DASAR. The 75-min interval was chosen based on an autocorrelation analysis of f_{min}, which found that samples measured 75 min apart had f_{min} correlation coefficients that fell below 0.1. This result was interpreted as representing a timescale over which a given migrating animal or group of animals could be localized by the array, and would thus be the timescale over which a calling animal would detect call rates similar to those measured on the DASARs. Call detections measured within this window were used to compute CallRate regardless of their value of R_{min}, discrepancy, localization success, or localization precision.

CallDensity, by contrast, was computed over the same time interval, but used only calls localized within the same value of R_{min} used to construct the dataset. CallDensity thus serves as a more accurate measure of the true underlying call density surrounding the DASAR in terms of calls per unit area, while CallRate is intended to measure the raw number of calls *perceived* per unit time by a calling animal. CallRate is much more sensitive to ambient noise levels than CallDensity; however, CallDensity would still be expected to be somewhat sensitive to noise levels for datasets where $R_{min} = 15$ km, since previous work has shown that call detection rates decrease at ranges greater than 2 km from a DASAR.

(k) Finally, for the manual data analysis a final categorical variable CallType was used to distinguish between simple modulated FM sweeps (type "A"), "constant" tonal calls (type "B"), and "complex" calls (type "C"), which have been noted as having lower frequency content in previous studies (Blackwell *et al.*, 2007). A preliminary data exploration of the frequency content of various call types suggested that constant "tonal" calls should be lumped into a separate "B" category from all the other simple FM call types, the rest of which were then lumped together into the same baseline "A" category.

CallingDepth and Discrepancy were obtained using full normal-mode propagation models derived from geoacoustic inversions of selected calls (Thode *et al.*, 2016). NoiseBand1, NoiseBand2, and SNR were measured by extracting a noise sample from the closest DASAR to a given call's location, using a time window that lasted between 1 and 0.5 s from the start of the measured call, and then using a 512 pt FFT window with 90% overlap to extract 10 PSD estimates. A 0.5 s gap was set between the end of the noise sample and start of the call in order to avoid inadvertently incorporating call components into the noise sample. Using noise samples measured 0.5 to 1 s *after* the end of the call was found to yield no significant difference in the analysis. While both NoiseBand measurements quantified the median power spectral density over a fixed bandwidth (40–60 Hz and 75–125 Hz), the SNR measurement was only computed over the same bandwidth as the following call. The SNR measurement was thus directly related to the detectability of the call, while the NoiseBand measurements were included to test whether potential relationships exist between the general noise background and f_{min}.

To evaluate the degree to which adjacent data measurements are correlated, the autocorrelation function was computed between the measured f_{min} values, the various explanatory variables, and model residuals. A threshold normalized autocorrelation value of 0.1 was used to determine block sizes for a first-order autoregressive (AR) covariance structure for a GEE with a normal distribution. In effect, the autocorrelation was used to estimate the average number of correlated samples taken from one "subject" (i.e., the number of samples effectively obtained from one individual or group swimming through a site). The AR structure fit the observed correlation structure better than an equicorrelated structure. The resulting GEE regression coefficients and their confidence intervals were then compared to a normal GLM that assumed complete sample independence. The GEE and GLM models yielded similar regression coefficients and confidence intervals, so most of the discussion that follows uses the GLM results. The GEE models were more applicable when examining ambient noise level trends, since noise levels changed slowly with time and were thus often highly correlated between adjacent samples.

Four levels of regression models were tested on each data set. Model 1 simply performed a least-squares linear fit as a function of Year only, using both the normal GLM and GEE; model 2 incorporated linear terms from up to 11 (automated detection) or up to 12 (manual detection) predictor variables using both the GLM and GEE; model 3 further incorporated first-order interactive terms between the predictor variables, using the GLM framework only; while model 4 was a "full" model that permitted both pure and mixed terms up to the quartic level (fourth-power). For models 2–4, individual terms were added whenever inclusion of the term lowered the Bayes Information Criterion (BIC). The confidence bands generated from the final prediction slices use simultaneous bounds to ensure that the entire curve has a 95% confidence interval of lying between the bands (which generates a much wider confidence interval than non-simultaneous bounds, which just present the confidence intervals around single observations).

III. RESULTS

A. Data sample sizes

Figure 3(b) shows that over 1.6×10^6 calls were automatically localized over 7 years at Sites 3 and 5 combined. For a subset of days over that same period, 264 576 calls were manually localized at the same sites [Fig. 3(a)], with nearly half of detected calls obtained between 2012 and 2014, the last three years of the study. Over that same period at both sites, over 1.6×10^6 airgun signals were flagged and removed by the automated procedure [Fig. 3(g)].

Once the additional culling steps in Sec. II D were applied to the raw results, the sample sizes dropped considerably. Figures 3(c) and 3(d) indicate that 13 355 calls (5%) from the manually analyzed dataset and 100 009 calls (6%) from the original automated dataset remained when R_{min} was set to 15 km, and that 2585 (<1%) and 15 704 (<1%) calls remained when R_{min} was set to 2 km [Figs. 3(e) and 3(f)]. These reduced datasets changed the relative proportions contributed by different years in the study, but still provided significant samples from all years, with the possible exception of 2011, which had unusually low numbers of calls. Note that the subsets of days manually analyzed varied from year to year, so the yearly distributions of the manual and automated analyses at a given R_{min} are not expected to match.

B. Descriptive statistics

This section plots various distributions of call source level, f_{min}, ambient noise levels, and airgun signals, providing some context for the more formal statistical regressions reported in Sec. III C.

1. Source level distribution

Plotting the source level distribution vs year provides useful insight about the degree to which the neural network

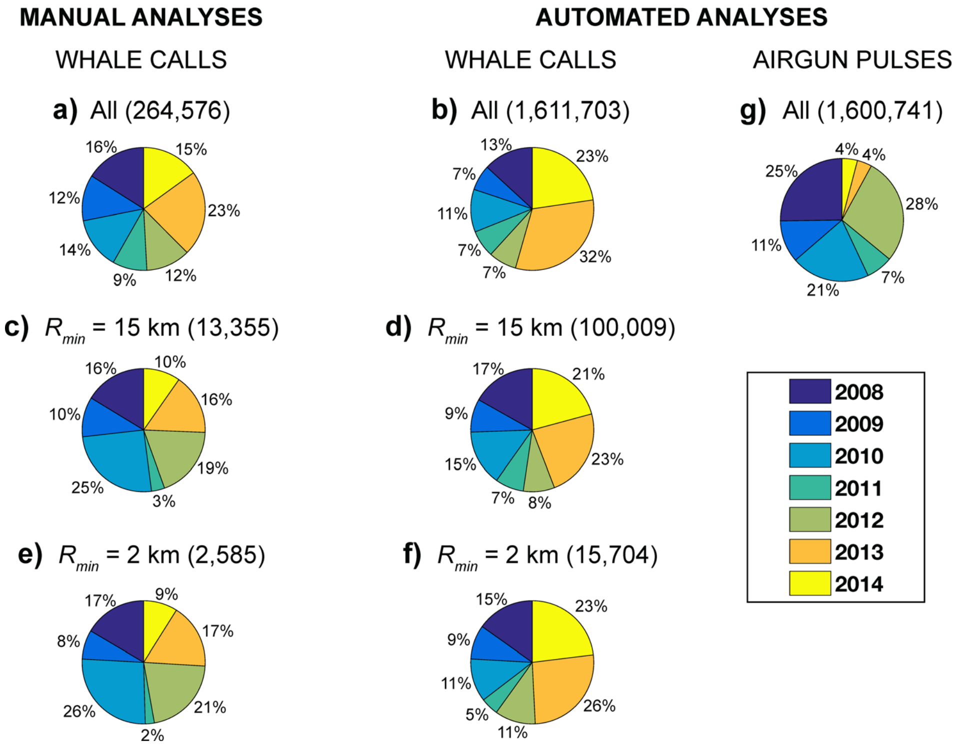

FIG. 3. (Color online) Pie charts showing breakdown of call samples among years of the study. Manual analyses are displayed in left column, automated analyses—whale calls and airgun pulses—are displayed on the right. The number of samples used are shown in parentheses. (a) All manually detected and (b) automatically detected calls obtained from Sites 3 and 5, with discrepancies <6 dB; (c) manually detected and (d) automatically detected calls culled using the criteria listed in Sec. II D, with $R_{min} = 15$; (e) manually detected and (f) automatically detected results with $R_{min} = 2$ km and discrepancies <6 dB; (g) automated detections of airgun pulses.

performance deteriorated over the life of the study, and how this deterioration can be compensated for with more stringent data culling. This analysis is also a necessary preliminary step for estimating how long-term shifts in ambient noise levels could influence call detectability (Sec. IV B).

Figure 4 arranges call source level distributions by deployment year. [Note that Thode *et al.* (2016) displayed source level distributions for all years combined.] The dashed (yellow) lines are distributions derived after applying Criteria 1–3 to the data (not checking for source level discrepancy),

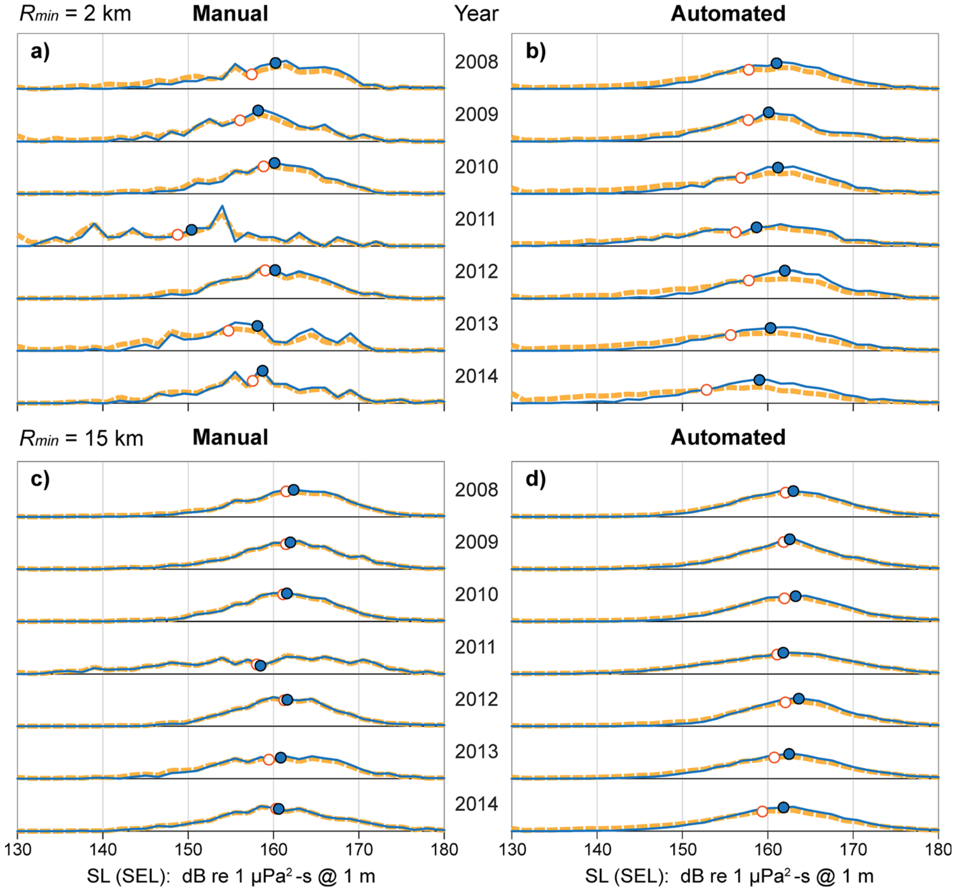

FIG. 4. (Color online) Long-term trends in source level distribution. Source level distributions are shown (a) from manually analyzed data, $R_{min} = 2$ km; (b) from automated analysis, $R_{min} = 2$ km; (c) from manual analysis, $R_{min} = 15$ km; (d) from automated analysis, $R_{min} = 15$ km. The dashed (yellow) lines contain all data samples, and the solid line shows results after excluding source level estimates with discrepancies greater than 6 dB. The circles on each distribution show the mean value of the *x* axis parameter (i.e., the "center of mass") of the distribution.

while the solid (blue) lines represent samples further winnowed by Criteria 4, such that calls associated with source level discrepancies greater than 6 dB are excluded. The open and solid circles represent the mean dB-scale values, or "center of masses" of the respective complete and winnowed distributions.

Figure 4(b), which shows automated source level distributions for calls generated less than 2 km from a DASAR, illustrates why data with large source level discrepancies were winnowed. When looking at the distributions where discrepancies have not been culled (yellow lines, open circle), one sees how the mean source level seems to decrease over time (−6.5 dB over 7 years), a pattern also visible to a lesser extent in Fig. 4(d), where $R_{min} = 15$ km. As mentioned in Sec. II D, high discrepancy values arise whenever automated samples capture only fragments of calls on at least one DASAR. Criteria 4 in Sec. II D, which restricts the permissible discrepancy to values less than 6 dB, reduces these "substandard" samples, and the resulting final distributions in both Fig. 4(b) and 4(d) (solid circles; blue lines) show the resulting mean source levels appearing relatively stable across years, with the possible exception of 2012, the year in which Shell performed exploratory drilling in the Beaufort Sea, between Sites 3 and 4. This stability in source level structure is consistent with the associated manual analysis results, which also display a stable mean source level at R_{min} values of both 2 and 15 km [Figs. 4(a) and 4(c)] across all years except 2011.

The relative increase in high-discrepancy estimates over time [as seen by the growing difference between the open and closed circles in Fig. 4(b)] provides evidence that the neural networks used by the automated algorithm, which were trained on 2008 and 2009 data, are gradually degrading in performance over time, a sign that some aspect of the bowhead whale call repertoire is gradually diverging from the training dataset. However, Fig. 4 also shows how one can compensate for this degradation by rejecting high-discrepancy call samples from the sample distributions.

In the detailed statistical analyses that follow, only samples with discrepancies less than 6 dB are retained. Enforcing this criterion reduced the final sample size by only 15%, a much less stringent culling than what the localization restrictions of Criteria 1 and 2 required.

2. Minimum frequency distribution

Figure 5 uses 8 Hz-wide frequency bins to produce histograms of the 7-year trend in the minimum frequency (f_{min}) distributions of the call samples analyzed using the four data sets. A long-term trend is visible in these distributions: the mean value of f_{min} decreases over 7 years, due to a relative increase in the fraction of calls descending below 75 Hz. The direction and magnitude of this shift is robust to the type of analysis applied: it is independent of R_{min}, whether manual or automated analysis is used, and whether high-discrepancy samples are retained or excluded. The fact that the shift in mean f_{min} is independent of R_{min} [Figs. 5(a) and 5(b) vs 5(c) and 5(d)] indicates that the observed distribution is characteristic of the call repertoire generated by the animals, and has not been modified by frequency-dependent acoustic

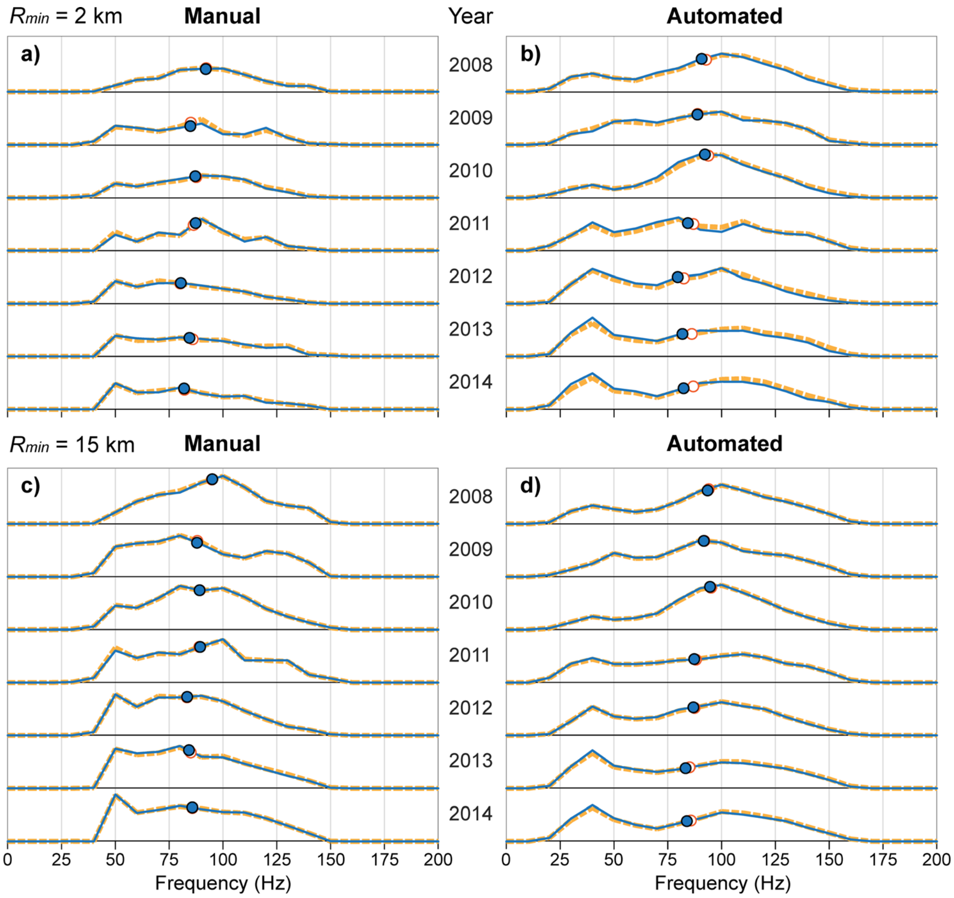

FIG. 5. (Color online) Long-term trends in minimum frequency f_{min} of call distributions, using a format identical to Fig. 4, and using 8 Hz frequency bins. Note the prominent increase in relative numbers of calls below 75 Hz over time.

propagation effects. These qualitative observations are borne out by the formal regression analyses in Sec. III C.

The clearest temporal changes in the observed distribution are visible in Fig. 5(d), which uses the largest sample size (100 009, automated analysis, $R_{min} = 15$ km), and which shows a 10.5 Hz downward shift in the mean f_{min} over the life of the study. At the start of the study, calls with f_{min} below 75 Hz (labeled from this point on as "low-frequency calls") comprised 27% of the total, but by the last 2 years the low-frequency calls comprised 41% of the total. Similar shifts are visible for the other analyses. Figure 6 contains example spectrograms of these low-frequency calls, with four examples randomly selected from each of the 7 years of the study. All calls shown were generated less than 2 km range from the DASAR. A significant fraction of the calls have several harmonics above the 30–50 Hz fundamental, extending past 100 Hz.

3. Airgun signals

Figure 7 displays the frequency, bearing, and timing interval distributions of identified airgun signals on a yearly basis, as compared with the distributions of automatically classified bowhead whale calls, for $R_{min} = 15$ km.

The left column of Fig. 7, which shows the frequency distributions of airgun signals and whale calls, reveals how the peak frequency of detected airgun signals tends to reside between 25 and 75 Hz, the same frequency region that is registering increased numbers of calls in Fig. 7(d) [which reproduces Fig. 5(d) for $R_{min} = 15$ km]. The middle column of Fig. 7 plots the bearings (azimuths) of both airguns and whale calls, with the azimuths of the latter category measured from the DASAR closest to a call's position. Airgun bearings are clearly dominated by certain directions, reflecting the fact that most surveys detected are distant from Sites 3 and 5 and are thus confined to a relatively small angular sector throughout a season. By contrast, whale bearings measured from the closest DASAR show much wider angular spreads around locations mostly to the east and west of the site, although 2010 shows heavy concentrations of calls to the south of the DASARs, indicating a southern shift in the migration route during that year. The last column displays the mean intervals between detections (evaluated over 800 s), and shows how the intervals between adjacent whale localizations are irregular and generally greater than the much more regular 10 to 20-s intervals of the airguns. The similarity between airgun and bowhead whale frequencies justifies the inclusion of airgun activity as a predictor variable in the regression analysis, but the lack of correspondence between their respective bearing and interval distributions indicates that the effect of airgun presence should not be significant.

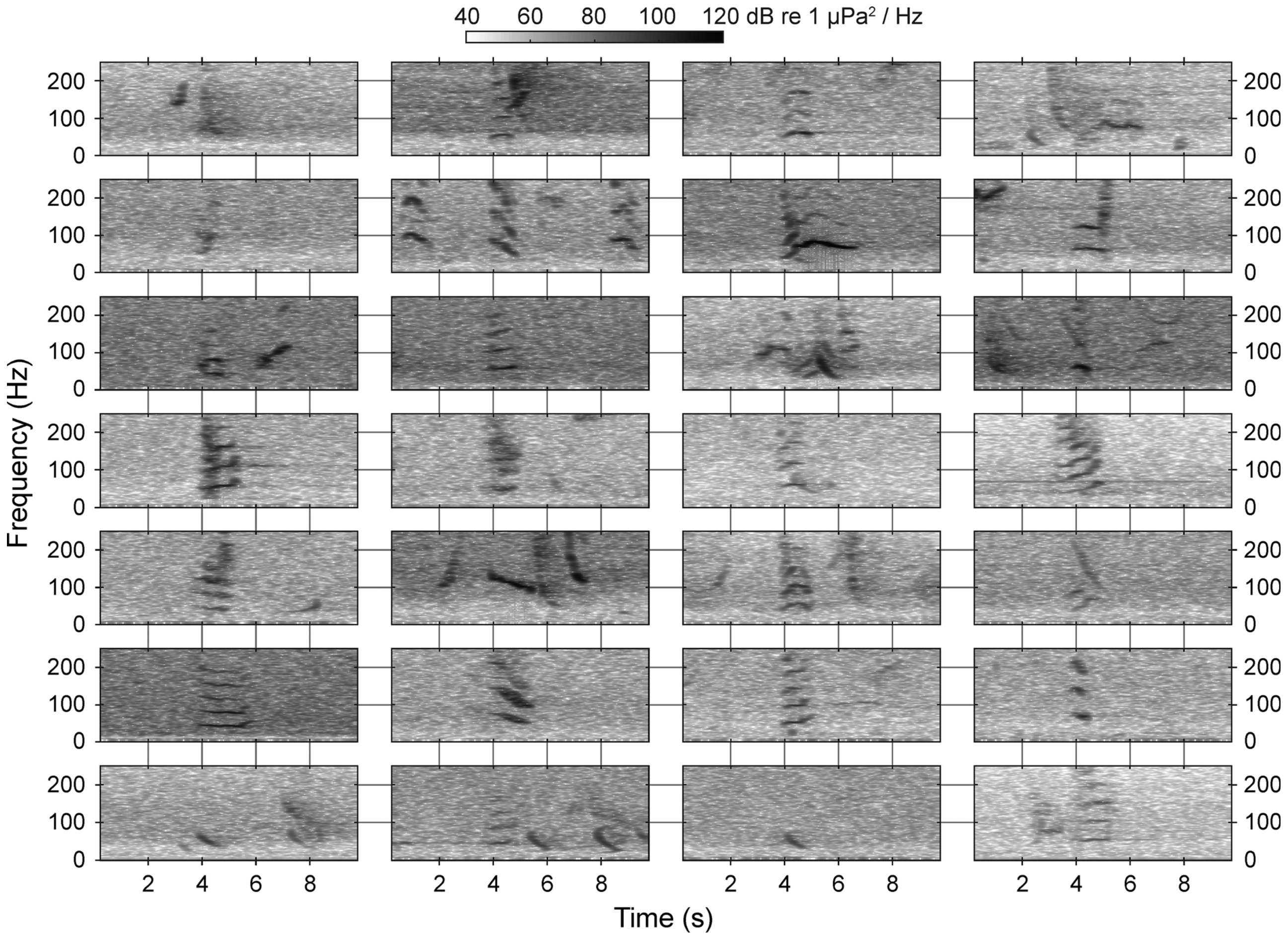

FIG. 6. Examples of low-frequency ($f_{min} < 75$ Hz) calls that became more prevalent with time. Each row represents a year of the study, and each column represents a randomly picked sample from that year. FFT sample size is 512 pts, 90% overlap, sampling rate 1 kHz.

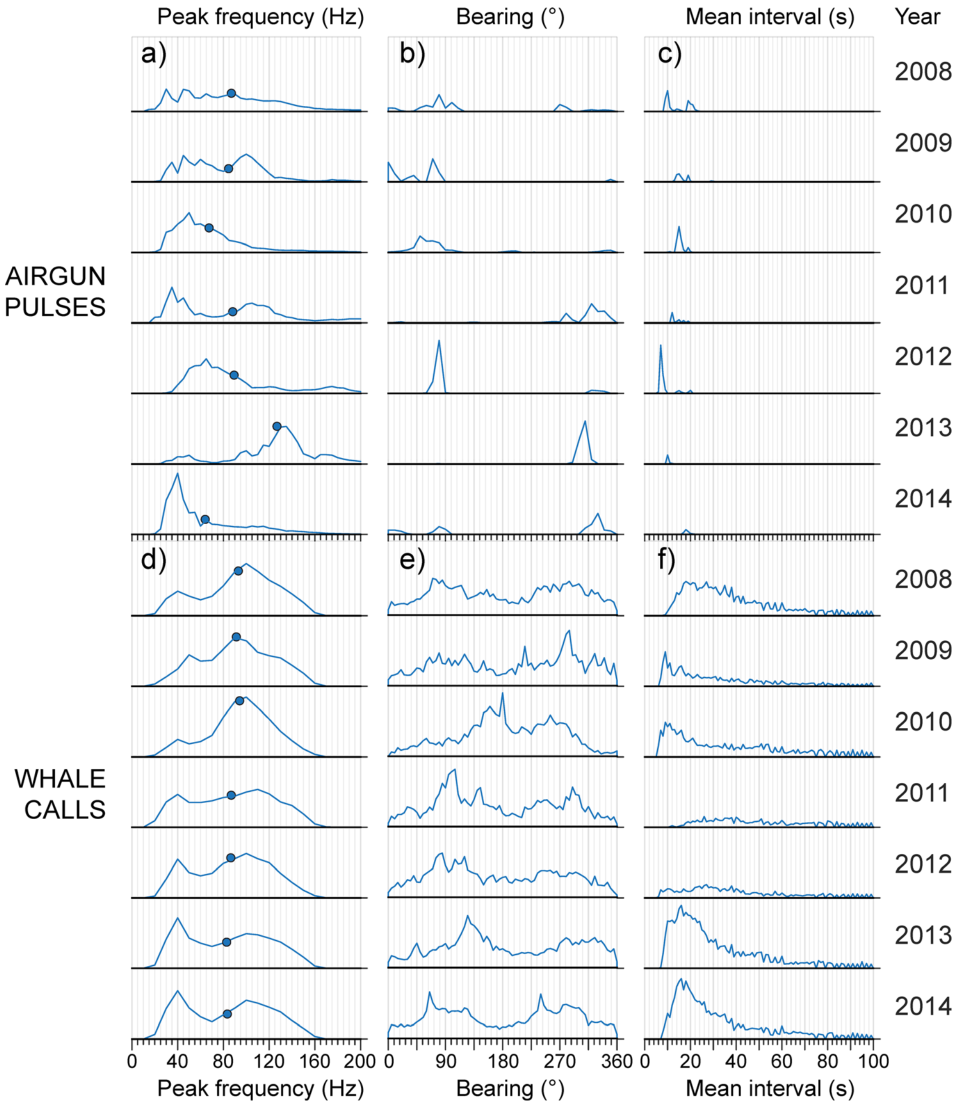

FIG. 7. (Color online) Comparison between automated airgun signal detections (top) and automated bowhead whale detections (bottom, $R_{min} = 15$ km). Left column shows frequency content of airguns (a) and whale calls (d), middle column shows bearings to airguns (b) and whale calls (e), as measured from measured DASAR (with 0° representing true north), and right column shows intervals between two airgun pulses (c) or whale calls (f) detected on the same DASAR.

4. Ambient noise levels

In most shallow water environments ambient noise levels increase with decreasing frequency, due in large part to the ubiquitous presence of commercial shipping, the dominant source for low-frequency noise worldwide (Wenz, 1962; McDonald *et al.*, 2006). However, in the Beaufort Sea, shipping traffic is very light, leaving wind-driven breaking waves as the dominant noise-production mechanism. Propagation modeling of sound at Site 5 at 50 m water depth, using environmental models derived from geoacoustic inversions of whale sounds (Abadi *et al.*, 2014), reveals that acoustic propagation is more efficient for shallow wind-generated sources at higher frequencies; below 75 Hz sound attenuates more quickly with range. The combined result of these factors is that wind-generated ambient noise levels would be expected to be lower below 75 Hz than they are above 100 Hz.

Figures 8(a) and 8(b) confirm this expectation by plotting various percentiles (between the 25th and 75th percentiles) of the seasonal ambient noise power spectral density at 40 Hz and 103 Hz, frequencies that represent the low and high-frequency peaks of the bimodal call frequency distribution

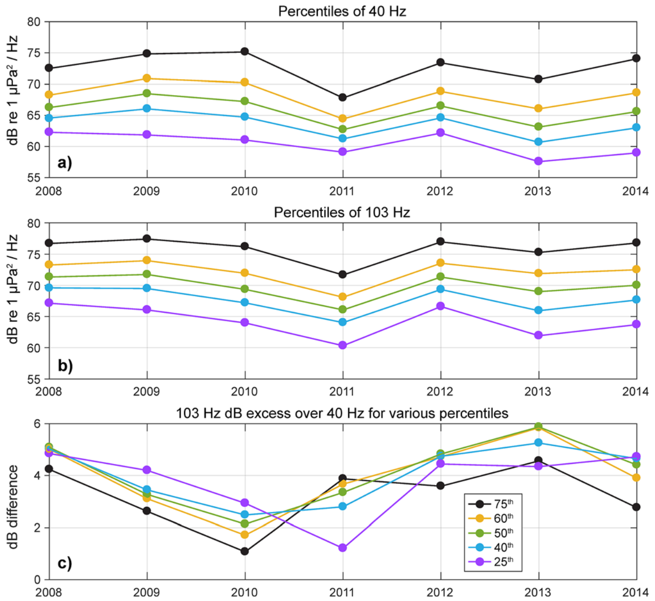

FIG. 8. (Color online) Changes in ambient noise levels over years. (a) Various percentiles of the power spectral density levels at 40 Hz vs year, ranging from the 25th to the 75th percentile; (b) same as previous, but showing 103 Hz component; (c) dB difference between 103 and 40 Hz for various percentiles.

visible in Fig. 5(d). These data are derived from all acoustic data, whether whale calls were present or not, using the techniques in Sec. II E. Ambient levels are always higher at 103 Hz. Figure 8(c) plots the relative difference in dB level between the two frequencies for each season. The highest median ambient noise levels at both frequencies occur in 2009, while the lowest levels occur in 2011. However, Fig. 8(c) also shows that during the first 3 years of the study (2008–2010) the relative noise difference between the two frequency bands decreases, but then in the last years of the study (2011–2014) the 103 Hz band becomes increasingly noisy relative to 40 Hz, culminating in a 6 dB difference in the 50th percentiles (and most other percentiles) in 2013. Section IV B will examine whether a 6 dB ambient noise increase at higher frequencies could have sufficiently suppressed detection rates of higher frequency calls to cause an apparent shift in the f_{min} distribution.

C. Regression analysis

Tables I and II present the predictor coefficient estimates, associated confidence intervals (CI), and R^2 fit for models 1–3. The p-values from hypothesis tests that a given regression coefficient is actually zero were generally minuscule and are only reported if they were greater than 10^{-5}. In the results reported below, the CI is provided in brackets.

1. Models 1 and 2

Model 1, a simple linear regression of f_{min} with Year, found annual decreases of the mean f_{min} on the order of -1.6 [-2.1, -1.1] to -1.94 [-2.04, -1.85] Hz per year, depending on which of the four datasets is used. These rates translate into a 9.6 to 11.6 Hz shift in mean f_{min} over the life of the study, replicating the graphical observations of Fig. 5. The R^2 value of the fit was small, between 0.01 and 0.02,

TABLE I. GLM (plain) and GEE (*italics*) coefficient estimates for Models 1 (Year only) and 2 (linear combination of predictors), applied to the four datasets. All estimated values have p-values less than 10^{-5} unless otherwise noted. α is the best-fit variance parameter for the AR(1) matrix structure for the GEE. Quantities in brackets are confidence intervals. The absence of numbers in a particular grid cell indicates that a factor was not significant for that particular dataset.

1	2	3	4	5
Model 1	*Manual analysis,*	*Automated analysis,*	*Manual analysis,*	*Automated analysis,*
	$R_{min} = 2$ *km*	$R_{min} = 2$ *km*	$R_{min} = 15$ *km*	$R_{min} = 15$ *km*
F_{min} vs Year	−1.6 [−2.1 −1.1]	−1.6 [−1.9 −1.4]	−1.7 [−1.9 −1.5]	−1.94 [−2.04 −1.85]
	−1.6 [−2.4 −0.8]	*−1.6 [−2.1 −1.2]*	*−1.7 [−2.1 −1.3]*	*−2.0 [−2.1 −1.8]*
	$\alpha = 0.18$, $R^2 = 0.017$	$\alpha = 0.26$, $R^2 = 0.011$	$\alpha = 0.24$, $R^2 = 0.020$	$\alpha = 0.26$, $R^2 = 0.015$
NoiseBand1 vs Year	—	*−0.49 [−0.77 −0.2]*	*−0.71 [−1.04 −0.37]*	*−0.51 [−0.64 −0.38]*
(40–60 Hz, *GEE only*)		$\alpha = 0.81$	$\alpha = 0.78$	$\alpha = 0.84$
NoiseBand2 vs Year	*−0.30 [−0.62 0.02]*	*−0.34 [−0.45 −0.23]*	*−0.41 [−0.69 −0.14]*	*−0.33 [−0.45 −0.21]*
(75–125 Hz, *GEE only*)	$\alpha = 0.68$	$\alpha = 0.63$	$\alpha = 0.71$	$\alpha = 0.84$
Model 2	$\alpha = 0.17$, $R^2 = 0.2$	$\alpha = 0.23$, $R^2 = 0.074$	$\alpha = 0.22$, $R^2 = 0.20$	$\alpha = 0.24$, $R^2 = 0.089$
Year	−1.8 [−2.3 −1.4]	−1.4 [−1.7 −1.1]	−1.3 [−1.5 −1.1]	−1.4 [−1.5 −1.3]
	−1.9 [−2.7 −1.1]	*−1.4 [−1.9 −0.9]*	*−1.5 [−1.9 −1.1]*	*−1.4 [−1.6 −1.2]*
CallingDepth (m)	—	−0.3 [−0.39 −0.28]	−0.15 [−0.19 −0.11]	−0.27 [−0.29 −0.25]
	−0.13 [−0.2 −0.03]	*−0.3 [−0.35 −0.23]*	*−0.14 [−0.19 −0.10]*	*−0.24 [−0.26 −0.22]*
Range (km)	—	—	0.77 [0.59 0.95]	1.8 [1.7 1.9]
			1.04 [0.76 1.32]	*1.8 [1.7 1.9]*
Discrepancy (dB)	1.2 [0.60 1.8]	2.0 [1.7 2.4]	1.3 [1.0 1.6]	2.0 [1.8 2.1]
	1.2 [0.61 1.8]	*1.9 [1.5 2.3]*	*1.27 [0.94 1.60]*	*1.7 [1.6 1.9]*
NoiseBand1 (dB re 1 μPa rms)	*0.2 [0.01 0.32]*, p = 0.03	0.89 [0.82 0.95]	0.46 [0.40 0.52]	1.1 [1.07 1.13]
(40–60 Hz)		*0.83 [0.71 0.96]*	*0.53 [0.44 0.62]*	*1.0 [0.98 1.1]*
NoiseBand2 (dB re 1 μPa rms)	—	0.81 [0.74 0.89]	0.50 [0.43 0.58]	1.05 [1.02 1.08]
(75–125 Hz)		*0.75 [0.61 0.90]*	*0.64 [0.51 0.77]*	*0.98 [1.18 2.02]*
SNR (dB)	—	0.54 [0.44 0.65]	0.14 [0.08 0.20]	0.73 [0.68 0.78]
		0.52 [0.38 0.66]	*0.21 [0.13 0.30]*	*0.69 [0.62 0.76]*
CallRate (calls / min)	—	—	—	−0.58 [−0.72 −0.45]
	0.42 [0.*14 0.71*] p = 0.004	—	—	*−0.63 [−0.87 −0.38]*
CallDensity (calls / min)	—	−11.5 [−14.2 −8.8]	—	−3.8 [−4.1 −3.4]
	3.84 [1.82 5.87]	*−11.7 [−16.0 −7.4]*	*0.69 [0.07 1.3]*	*−3.8 [−4.3 −3.3]*
			($p = 0.03$)	
Site 5[a]	3.0 [1.3 4.7]	3.4 [2.3 4.5]	4.8 [4.1 5.6]	2.4 [2.0 2.8]
	4.0 [1.1 6.6]	*3.2 [1.3 5.1]*	*5.1 [3.5 6.6]*	*2.4 [1.6 3.2]*
CallType B[a]	6.3 [4.4 8.3]	—	6.4 [5.5 7.3]	—
		—		—
CallType C[a]	−25.9 [−28.3 −23.5]	—	−24.7 [−25.9 −23.5]	—
		—		—
Airgun[a]	—	−2.9 [−4.1 −1.7]	—	−1.0 [−1.5 −0.53]
	—	*−2.8 [−4.7 −0.80]*	—	*−1.0 [−1.8 −0.03]*, $p = 0.02$

[a] Categorical variable.

TABLE II. GLM terms from Model 3 (interactive terms). Model 3 for $R_{min} = 2$ km manual analysis is identical to the corresponding Model 2, so is not shown below. ΔBIC is the change in the Bayes Information Criterion (BIC) caused by removal of a particular term, with positive values indicating that the term's removal decreases the BIC. ΔR^2 is the change in R^2 fit caused by the removal of the term (multiplied by 1000).

Term	*Estimate*	*Lower CI*	*Upper CI*	ΔBIC	*1000**ΔR^2
$R_{min} = 2$ km; Automated Analysis, 14 terms, $R^2 = 0.206$					
Year	−2.33	−2.88	−1.78	−60.21	4.06
CallingDepth	−0.15	−0.24	−0.06	−2.02	0.68
Discrepancy	−0.36	−1.15	0.44	8.88	0.05
NoiseBand1	−0.14	−0.39	0.10	8.30	0.08
SNR	−4.29	−5.22	−3.36	−71.86	4.74
CallRate	−42.15	−55.81	−28.49	−26.90	2.12
Site 5	9.20	6.62	11.77	−39.37	2.85
Airgun	22.80	12.59	33.01	−9.51	1.11
Year:Discrepancy	0.52	0.36	0.68	−33.15	2.49
Year:CallDensity	−3.58	−5.07	−2.08	−12.42	1.28
CallingDepth:Site 5	−0.28	−0.39	−0.17	−15.68	1.47
NoiseBand1:SNR	0.07	0.05	0.08	−82.60	5.36
NoiseBand1:Airgun	−0.39	−0.54	−0.24	−16.09	1.49
SNR:CallDensity	2.24	1.63	2.85	−41.97	3.00
$R_{min} = 15$ km; Manual Analysis, 20 terms, $R^2 = 0.213$					
Year	7.13	4.59	9.67	−20.75	1.79
CallingDepth	−0.15	−0.19	−0.11	−39.91	2.92
Range	0.07	−0.30	0.45	9.35	0.01
Discrepancy	5.57	2.48	8.66	−3.04	0.74
NoiseBand1	0.62	0.41	0.82	−25.37	2.06
SNR	−1.38	−1.96	−0.81	−12.64	1.31
Site 5	−11.99	−19.33	−4.66	−0.80	0.61
CallType B	−9.56	−18.66	−0.46	7.69	0.67
CallType C	−18.80	−30.86	−6.73	7.69	0.67
Year:Range	0.19	0.10	0.27	−8.02	1.04
Year:NoiseBand1	−0.12	−0.15	−0.08	−35.26	2.65
Year:SNR	−0.06	−0.09	−0.03	−7.78	1.02
Year:CallType B	−1.37	−1.84	−0.90	−23.61	2.52
Year:CallType C	−1.39	−2.05	−0.74	−23.61	2.52
Discrepancy: NoiseBand1	−0.08	−0.12	−0.04	−4.86	0.85
Discrepancy:SNR	0.09	0.04	0.13	−5.54	0.89
NoiseBand1:SNR	0.02	0.01	0.03	−22.02	1.86
NoiseBand1:Site 5	0.24	0.13	0.34	−9.23	1.11
NoiseBand1:CallType B	0.28	0.15	0.40	−1.33	1.20
NoiseBand1:CallType C	0.01	−0.15	0.18	−1.33	1.20
Site 5:CallType B	3.58	1.83	5.34	−7.59	1.57
Site 5:CallType C	−2.73	−5.14	−0.33	−7.59	1.57
$R_{min} = 15$ km; Automated Analysis, 25 terms, $R^2 = 0.10$					
Year	−12.63	−13.92	−11.34	−357.43	3.32
CallingDepth	−0.21	−0.29	−0.14	−21.25	0.29
Range	0.14	−0.19	0.46	10.84	0.01
Discrepancy	−0.96	−1.40	−0.53	−7.48	0.17
NoiseBand1	−0.06	−0.20	0.08	10.82	0.01
SNR	−2.56	−3.04	−2.07	−93.56	0.94
CallDensity	−4.27	−5.17	−3.37	−74.35	0.77
Site 5	10.63	9.36	11.90	−257.24	2.42
Airgun	8.62	3.52	13.72	0.55	0.10
Year:CallingDepth	−0.02	−0.03	−0.01	−0.80	0.11
Year:Range	0.16	0.12	0.21	−39.58	0.46
Year:Discrepancy	0.45	0.38	0.52	−168.52	1.62
Year:NoiseBand1	0.12	0.10	0.13	−199.61	1.90
Year:SNR	0.13	0.11	0.16	−104.01	1.04
Year:CallDensity	−0.39	−0.58	−0.19	−3.91	0.14
Year:Airgun	−0.38	−0.59	−0.16	−0.34	0.11
CallingDepth:Range	0.04	0.03	0.05	−79.41	0.82
CallingDepth:Discrepancy	0.05	0.03	0.06	−37.34	0.44
CallingDepth:Site 5	−0.42	−0.46	−0.37	−372.90	3.46
Range:CallDensity	0.40	0.20	0.60	−4.53	0.14
Range:Site 5	−0.33	−0.50	−0.16	−2.84	0.13
Range:Airgun	0.39	0.19	0.59	−3.36	0.13
NoiseBand1:SNR	0.04	0.03	0.05	−120.12	1.18
NoiseBand1:Airgun	−0.19	−0.26	−0.12	−16.62	0.25
Site 5:Airgun	3.94	3.04	4.83	−63.27	0.67

reflecting the fact that the f_{min} depends on many other factors than just Year.

Autocorrelation of f_{min} measurements found that adjacent samples have correlation coefficients of around 0.2, with correlations dropping to 0.1 or less after 20 to 50 offset samples, the latter covering roughly 75-min. Thus when models 1 and 2 were run with a GEE, the block size was set to 50, and the best-fit intra-class correlation coefficient α for the autoregressive model was found to be between 0.18 and 0.26 (as would be expected from the autocorrelation results). The relatively low value of α indicated that the localization samples became decorrelated relatively quickly over time, resulting in predictor coefficient values and CIs for Year that are very similar to the GLM regression (which assumes samples are uncorrelated). We interpret this result to mean that the time scale over which migrating individual whales (or associated individuals with similar acoustic behavior) were resampled was short compared to the duration of a season.

The next two rows of Table I show the result of regressing NoiseBand1 and NoiseBand2 against Year using the GEE framework, to determine whether long-term trends in background noise levels exist. Noise samples collected from each call localization were much more highly correlated with adjacent samples ($\alpha \sim 0.63$ to 0.84), indicating that the timescale over which background noise conditions evolved was longer than the residence time of calling animals within the tracking region. Using the largest dataset ($R_{min} = 15$ km; automated), the regression found that over seven seasons the NoiseBand1 median PSD decreased 3.1 [2.3, 3.8] dB and NoiseBand2 decreased 2.0 [1.3, 2.7] dB re 1 μPa^2/Hz. Thus the mean difference in ambient noise levels between the two bands changed by only 1 to 2 dB over 7 years, consistent with the net change in ambient noise levels over seven seasons shown in Fig. 8(c).

The other prediction factors were relatively uncorrelated with each other, with correlation coefficients generally less than 0.15. Some exceptions emerged: the Airgun indicator did have higher correlations with Site (0.23) and Year (0.44). This is not surprising, since airgun activity only occurred in certain places and in certain years. Also unsurprising was that CallRate was negatively correlated with NoiseBand1 (−0.34), as higher noise levels would be expected to reduce the total number of localized calls detected. Regardless of these specific relationships, the maximum variance inflation factor (VIF) never exceeded 1.3, with typical values around 1.05, so the predictor variables were effectively linearly independent.

When these additional factors were incorporated into model 2 (linear terms in the GLM and GEE; Table I), the Year coefficient and its CI remained similar to model 1, but other factors were also found to have non-zero coefficient values: CallingDepth (GEE only), Discrepancy, NoiseBand1, and Site were factors in all four datasets. CallType, in particular, was a major factor in predicting f_{min} for manually analyzed datasets, with type C (complex calls) having minimum frequencies 25 Hz [24, 26] lower than the type A (simple FM) category, and type B calls (constant FM) generally having a f_{min} value 6.4 [5.5, 7.3] Hz greater than type A. R^2 increased to 0.2 for the manual analysis and 0.09 for the automated results. The manual results have a higher R^2 primarily due to the inclusion of CallType: without this factor the manual R^2 fit drops to 0.07. Range and SNR are also significant regression factors for automated datasets, and (to a lesser extent) the manual datasets.

Several factors were revealed to have minor to insignificant impacts on f_{min}. For example, Northing did not yield a coefficient significantly different from zero for any dataset or model and is not discussed further. In addition, measuring either NoiseBand1 or NoiseBand2 in terms of SPL instead of PSD made little difference in the results presented here, and thus more complex regression models used only NoiseBand1 in terms of PSD. While Airgun presence was not a significant factor in the manual datasets, it was a small factor in the automated datasets, with a −1 [−1.5, −0.5] to −2.9 [−4.1, −1.7] Hz shift in f_{min} when airguns were present. However, the large CI for Airgun and the relatively high p-value (0.02) indicated a relatively minor effect.

CallDensity always appeared as a significant factor across the four datasets, but CallRate was significant only in a couple of datasets. However, the magnitude and even the

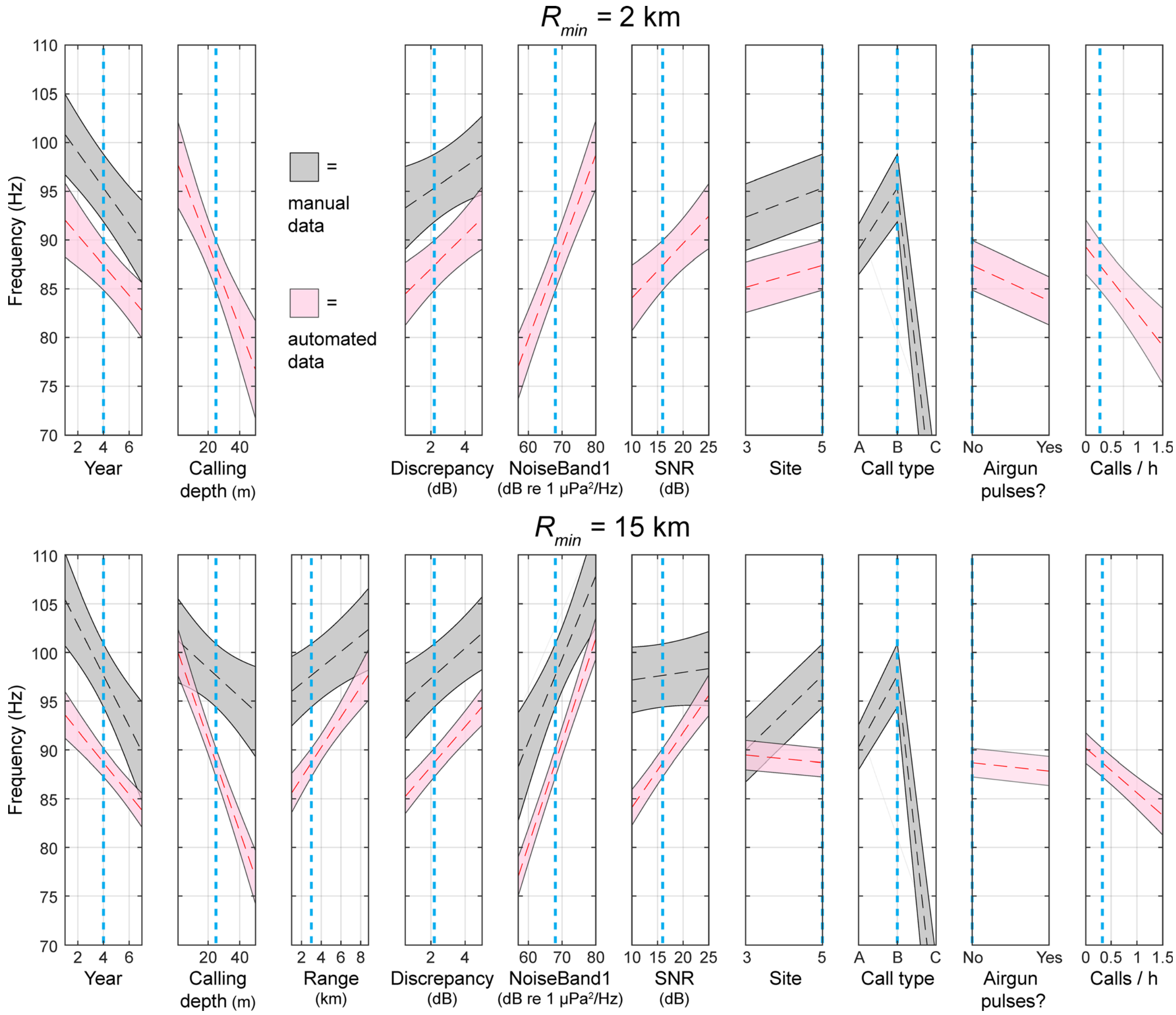

FIG. 9. (Color online) Plots of prediction slices for model 3, a GLM with interactive linear terms, using data from both Sites 3 and 5. Top row: $R_{min} = 2$ km; Bottom row: $R_{min} = 15$ km. Gray: manual data; pink: automated data. Dashed lines show mean values, and shaded regions are 95% simultaneous prediction bounds. The absence of a curve on a particular plot means that the factor was not significant for that dataset. Vertical dashed (blue) lines indicate fixed values used for computing prediction slices in adjacent plots on the same row.

sign of these regression coefficients fluctuated depending on the type of analysis and the value of R_{min}used. CallRate tended not to be significant in more complex regression models and so later discussion will only discuss CallDensity.

2. Models 3 (interactive terms) and 4 (nonlinear interactive terms)

Model 3, where predictor variables were allowed to have first-order interactions, was the same as model 2 for the smallest dataset (manual; R_{min} 2 km). Yet Table II shows that significant interactions existed between variables in the other datasets, yielding between 10 to 25 terms for the final linear predictor. For example, in the automated $R_{min} = 2$ km dataset the Year coefficient was influenced by Discrepancy and CallDensity. The larger the value of Discrepancy, the lower the yearly decrease, and the larger the value of CallDensity, the faster the yearly decrease. Table II shows that 5 to 7 factors influenced the Year coefficient in the larger $R_{min} = 15$ km datasets. The ΔBIC column entries for these factors indicate that the interactions between Year:NoiseBand1 and Year:CallType contributed the most important modifications to the trend with Year. For manual data, both B and C call types decreased 1.4 Hz faster per year than simple FM calls. Also, for manual data, a 10 dB increase in NoiseBand1 caused f_{min} to decrease faster by 1 Hz a year, but for automated data, a 10 dB *decrease* in NoiseBand1 had the same impact.

Figure 9 uses the coefficients in Table II to plot linear prediction slices for each variable in model 3, while holding the other variables fixed at their median values in the data sample (dashed blue lines). Thus, even though model 3 incorporated interactions between variables, Fig. 9 isolates the effects of each variable when others are held fixed. The fixed values for each variable are as follows: Year: 4; CallingDepth: 25 m; Range: 1 km (for $R_{min} = 2$ km) or 3 km (for $R_{min} = 15$ km); Discrepancy: 2.2 dB; NoiseBand: 68 dB re 1 μPa^2/Hz; SNR: 16 dB; Site: 5; CallType: B; Airgun: *false*; and CallDensity: 0.1 calls/min ($R_{min} = 2$ km) or 0.33 calls/min ($R_{min} = 15$ km). The span shown for each factor lies between the 5th and 95th percentile of the largest (100 009 sample) dataset. The 95% simultaneous confidence bands for each slice are also displayed. The manual distribution had a higher f_{min} value than the automated dataset because the manual analyst software did not permit f_{min} to be selected below 50 Hz. Figure 9 also confirms that Site and Airgun had only small effects on f_{min}, because the span of the confidence bounds exceeds the shift in the mean f_{min} caused by changing between these categorical variables. As a result the trends visible in Fig. 9 remain the same as those displayed if Site is switched to Site 3. The dependence of f_{min} on other factors will be addressed in more detail in Sec. IV.

As CallType has the greatest impact on the predicted f_{min}, Fig. 10(a) breaks out the impact of this factor on the manually analyzed prediction curves, with each curve representing a different call type. The figure clearly shows how the B (constant call; yellow shading) and C (complex call; green shading) call types are 6 Hz higher and 20–30 Hz lower, respectively, than call type A (simple call; purple shading). All three call types showed decreasing f_{min} across the seasons, but the B and C call types showed up to a 20 Hz decrease over seven seasons, while A decreased by only about 10 Hz over the same interval. Figure 10(b) shows that type A dominated the logged calls (overall, 60% of total, vs 27% for B and 12% for C), with no trend in call type composition over time. The dominance of call type A explains why the 10 Hz downward shift of f_{min} shown in Fig. 5(d) was similar to the shift displayed by call type A alone.

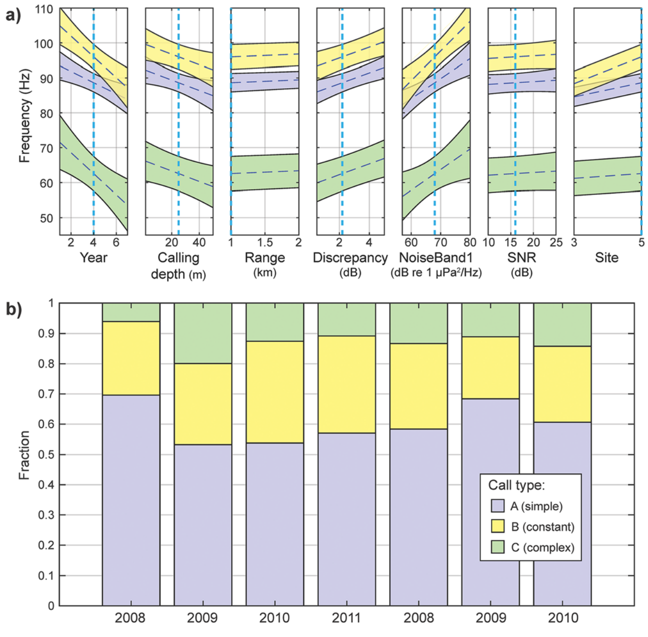

FIG. 10. (Color online) (a) Effect of manually analyzed call type on f_{min} and other regressive model factors, $R_{min} = 15$ km. Purple: Call type A; yellow: Call type B; green: Call type C. (b) Relative proportions of the three call types as a function of year.

Model 4, which permitted up to fourth-order polynomial terms along with various lower-order combinations of interactions, yielded many terms that reduce the model BIC: 27 terms for the automated, $R_{min} = 2$ km dataset; 28 terms for the manual, $R_{min} = 15$ km dataset; and 50 terms for the automated, $R_{min} = 15$ km dataset. The R^2 values for the various models reach 0.21 and 0.13 for the manual and automated 2 km R_{min} datasets, and 0.24 and 0.14 for the manual and automated 15 km R_{min} datasets. As with all previous models, the better fits to the manual data primarily arise from the inclusion of CallType as a predictive factor. The overall dependencies of f_{min} on the predictor variables are similar between models 3 and 4, and so the terms and plots for model 4 are only provided in the supplementary material.[1]

IV. DISCUSSION

Figure 5 shows that the distribution of f_{min} values was quite broad. Even the most complex statistical regression model (model 4) accounted for just 14% to 24% of this variance, despite employing nine predictive factors and up to 50 terms. The models clearly demonstrate that f_{min} has substantial dependencies on many factors related to acoustic propagation and potentially behavior. Some factors, like Site, Airgun, and Northing, were either found to be insignificant or found to have effect sizes smaller than the confidence bounds for the prediction curves. Others, like NoiseBand1, CallingDepth, and CallType, had strong and complex relationships with f_{min}. In particular, CallType, which was only available for the manual analyses, seemed to be a crucial factor in predicting f_{min} (e.g., Fig. 9); for example, whenever CallType was removed as a predictor variable from models 3 and 4, the R^2 values dropped from 0.21 to 0.08 and from 0.24 to 0.12, respectively.

Despite the presence of the factors mentioned above, every statistical model found a significant regression coefficient between Year and f_{min}. Over 7 years, the shift in predicted mean value (dashed lines in shaded regions of Fig. 9) matched the mean shift visible in Fig. 5 and exceeded the confidence bounds assigned to the prediction curve (Fig. 9, leftmost column). The 7-season shift retained roughly the same direction and magnitude—a decrease of about 10 Hz over 7 years—regardless of the model used, the dataset applied, or the number of predictive factors included (e.g., Tables I and II).

Before discussing potential behavioral or physiological explanations for these observed long-term frequency shifts, we first examine three potential non-biological explanations for this evolution: long-term changes in propagation factors (including the location of the migration corridor), long-term shifts in relative ambient noise levels, and increasing misclassification of airgun signals as whale calls. We then examine specific biological explanations, including a population-wide shift in the relative use of call types, the addition of a new call type, physiological growth in the population, and finally, a behavioral response to increasing call spatial densities.

A. Possible explanations of frequency shift arising from acoustic propagation factors

The bimodal structure shown in Fig. 5(d), along with the presence of many harmonics in Fig. 6, raises the question as to whether the bowhead population call repertoire is not actually shifting in frequency; perhaps the relative detectability of a low-frequency fundamental is increasing with time. Possible acoustic propagation mechanisms for increased low-frequency detectability include shifts in the population calling depths or the mean distance from a DASAR, a shift of the migration route northward into deeper waters that are more favorable to low-frequency propagation, or changes in the sound speed profile over multiple years. It was for this reason that the factors Range, CallingDepth, and Northing were included in the statistical regression models, along with Site, which served as a proxy for the differing average water depths of the sites. Site had only a minor effect on f_{min} that fell well within the confidence bounds, while Northing, a proxy for both offshore distance and water depth, did not yield a prediction coefficient significantly different from zero ($p = 0.68$) and so was listed in neither the Tables nor the prediction plots.

Range was only a factor for the $R_{min} = 15$ km data sets, and Fig. 9 shows that the effect size of Range on the manual analysis was smaller than the curve's confidence bounds. However, Range had a clear effect on the frequency detected by the automated analysis, with larger ranges resulting in a higher detected mean f_{min}. The results are consistent with the fact that low-frequency attenuation (i.e., below 75 Hz) is greater in this shallow-water environment than "higher" frequencies (i.e., 100 Hz and above). Even though the propagation modeling used to estimate calling depth incorporated frequency-dependent effects, that modeling cannot recover low-frequency signal components that have fallen below background noise levels.

Regardless, Fig. 9 indicates that the multi-year downward shift of f_{min} is similar for both $R_{min} = 15$ km datasets, and still exists whenever Range is held at a fixed value. The effect of Range on the yearly frequency shift is also relatively small; for example, Table II (model 3, manual analysis) displays a Year:Range interactive term of 0.19, so doubling the fixed localization Range from 1 to 2 km generates an additional 1.3 Hz (7 years * 0.19 * 1 km) shift over 7 years.

Figure 9 also shows a significant effect size for CallingDepth, in that shallower calling depths are generally associated with higher-frequency detections. For example, shifting a call from 10 m to 25 m depth is associated with a downward shift in f_{min} by 10 Hz at Site 5. The effect due to calling depth is smaller (8 Hz shift) at the shallower Site 3. As discussed in Thode *et al.* (2016), these depth dependencies likely arise from propagation effects: for example, 28 m turns out to be the optimum depth for propagating sounds lower than 100 Hz at Site 5, so at a given fixed range and source level, lower-frequency signals are more likely to be detected when generated at that depth. Similar calculations show that a call generated at 10 m depth has an effective transmission loss power law (to 5 km range) of 15logR at 80 Hz, but only 13.5logR at 100 Hz. Thus, a 100 Hz component for a 10-m deep call would be enhanced 5 dB relative to an equivalently intense 80 Hz component over 3 km range, the most likely propagation distance represented in the dataset, and, therefore, one would expect 100 Hz components to be detected more easily than 80 Hz components for 10-m deep calls, all other factors being constant. Thus, the relationship between f_{min} and CallingDepth is interpreted to arise from relative detectability of lower-frequency calls arising from propagation effects. Despite these effects, the 7-season shift in f_{min} persists regardless of the fixed value of CallingDepth chosen.

Numerical simulations of sound propagation at Site 5 found that changes in sound speed profiles only had a minor impact on a 100-Hz signal in a 50-m waveguide; long-range absorption under those circumstances were dominated by the seafloor sediment composition, which presumably remains stable over decadal scales. The details of the sound speed profile only influenced acoustic propagation above 150 Hz, a frequency range excluded from the datasets due to Criteria 3 in Sec. II D. Thus, potential changes in the temperature profile over time were judged to be an unlikely cause of the f_{min} shift.

In summary, while range and calling depth are shown to influence the detected f_{min} in the dataset, the predictive models demonstrate that they do not explain the multi-season shift.

B. Possible explanations of frequency shift arising from ambient noise trends

Here we discuss two analyses of the hypothesis that the observed frequency shifts in call repertoire arise from improvements in relative detectability of low-frequency calls vs higher-frequency calls, due to long-term shifts in relative ambient noise levels.

A formal regression analysis found that the predicted mean f_{min} is a strong function of NoiseBand1 in both manual and automated analysis with $R_{min} = 15$ km, but only for the automated analysis when $R_{min} = 2$ km. A 10 dB increase in median PSD led to a 5 to 20 Hz *increase* in f_{min} over what is plotted in Figs. 9 and 10, depending on the model and dataset. These functional relationships still hold whenever NoiseBand2 was used instead.

One possible interpretation of this relationship is that background noise levels decrease with increasing frequency, so an overall increase in ambient noise levels would tend to mask lower frequencies. However, Fig. 8 shows that noise levels were generally *higher* at higher frequencies, a result that can be checked by measuring the "spectral tilt," or slope of the PSD curve, across various bandwidths. Between 40 and 60 Hz, this slope was 0.14 dB/Hz [−0.27 0.54], so the noise level *increased* nearly 3 dB over this band. Between 75 and 125 Hz, the bandwidth over which most f_{min} values were measured, the slope was −0.01 dB/Hz [−0.14 0.12], a relatively small −0.5 dB change over the bandwidth.

Our preferred interpretation of this relationship is that bowhead calls tended to have lower received levels at f_{min} when compared with received levels at other frequencies in the middle of the bandwidth, a situation that perhaps arises from frequency-dependent differences in propagation. As noise levels rose across the frequency spectrum, the lowest-frequency (and weaker) call components were masked first, shifting the detected f_{min} upward. Whatever the reason behind the dependence of f_{min} on NoiseBand, whenever it was held fixed at typical background noise levels (50th percentile) of 68 dB re 1 μPa^2/Hz, all four datasets still predicted the multi-season f_{min} decrease (Figs. 9 and 10), suggesting that shifts in background noise levels were not responsible for the 7-season shift.

This conclusion can be checked by estimating what fraction of bowhead calls would have been masked by an ambient noise increase. Estimating long-term changes in ambient noise was a subtle matter, as year-to-year shifts were highly variable (Fig. 8). A simple GEE statistical regression of NoiseLevel1 and NoiseLevel2 (median PSD) against Year found PSD values decreasing over time for both bands; using the values from the largest dataset, over 7 years, the 40–60 Hz band decreased −3.1 dB [−3.8 −2.3], while the higher band decreased −2.0 dB [−2.7 −1.3]. Consequently, the lower frequency band became relatively quieter by only ~1 dB. However, when seasonal PSD percentiles are plotted directly in Fig. 8 one sees that the long-term relative ambient noise difference between a representative "high frequency" (103 Hz) and "low frequency" (40 Hz) band was more complex than a simple linear shift; while the overall shift across 7 years was small, the 103 Hz PSD actually increased 4 dB relative to the 40 Hz PSD over the last 4 years of the study.

To estimate how relative changes in ambient noise levels would impact relative detection rates for low- and high-frequency calls, we exploited the fact that the source level distribution of the population seemed consistent across time (Fig. 4), and that the spatial distribution of the migration corridor did not change (based on a plot of the distributions of Northing). We restricted the analysis to a single site, Site 5, as ambient noise levels were more closely related across a site than between sites.

Picking the year with the lowest 103 Hz ambient noise levels [2011; Fig. 8(b)], the received levels of localized high-frequency bowhead whale calls from Site 5 for that season were then plotted as a cumulative distribution in Fig. 11, where the received level was measured on the DASAR closest to the call's location. A detection threshold was then estimated by taking the PSD of noise levels at 103 Hz for that year, adding $10\log_{10}(20\,\text{Hz})$ to account for typical bowhead FM call bandwidth and then converting PSD into units of sound pressure level (SPL); 6 dB was then added to this value, to account for an estimated detection signal-to-noise ratio (SNR) of 6 dB for the automated algorithm, producing a 2011 detection threshold of 85 dB re 1 μPa in Fig. 11.

This threshold detection level would shift 6 dB higher, to 91 dB re 1 μPa, during the season with the highest ambient noise levels (2009; Fig. 8). If one assumes that the received level distribution from 2009 would have been similar to that in 2011 (since the overall source level distributions are steady over time), we can estimate what fraction of calls detected in 2011 would *not* have been detected had noise levels been 6 dB greater, and find that about 10% of the 2011 call sample would have been lost with a 6 dB increase (Fig. 11). A corresponding analysis for the 40 Hz band finds that basically all calls in 2011 would still have been detected in 2009 (since noise levels are generally lower in that band). A relative change of 6 dB in noise levels between the bands would thus be expected to reduce the high-frequency call count by ~10%. Table III estimates what impact this noise masking would have had on the Site 5 f_{min} distribution via column 5, which reduces the sample size of calls above 75 Hz (column 4) by 10%, and then recomputes how the proportions of the f_{min} distribution would

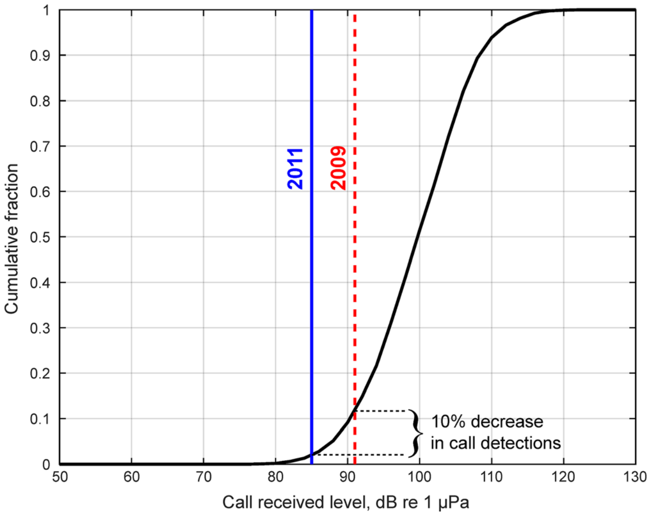

FIG. 11. (Color online) Cumulative distribution of received levels (rms sound pressure level) on high-frequency ($f_{min} > 75$ Hz) bowhead whale calls detected on closest DASAR during the 2011 season. Solid (blue) line shows detection threshold estimated from median ambient noise levels in 2011; dashed (red) line shows estimated detection threshold using median 2009 noise levels, which were 6 dB greater than 2011.

TABLE III. Breakdown of call distribution for Site 5. Column 2 shows total number of calls used in the analysis; Column 3 shows the number (percent) of calls with minimum frequencies below 75 Hz; Column 4 shows the number (percent) of calls with minimum frequencies above 75 Hz; Column 5 indicates how the relative percentage of high-frequency calls would change if the number of high-frequency samples is reduced by 10% (per Fig. 11).

1	2	3	4	5
Year	*Automated calls within 15 km of a DASAR, Site 5 only, discrepancy < 6 dB*	*Automated calls, f_{min} < 75 Hz (percent of total)*	*Automated calls, f_{min} > 75 Hz (percent of total)*	*Automated calls, f_{min} > 75 Hz,* reduced by 10% (percent of total)
2008	9361	2512 (27)	6849 (73)	6164 (71)
2009	6132	1469 (24)	4663 (76)	4196 (74)
2010	9022	2153 (24)	6869 (76)	6182 (74)
2011	3957	1345 (34)	2612 (66)	2350 (64)
2012	4851	1534 (32)	3317 (68)	2985 (66)
2013	9517	4068 (43)	5449 (57)	4904 (55)
2014	9396	3757 (40)	5639 (60)	5075 (58)
Total	52 236	16 838 (32)	35 398 (68)	31 858 (65)

change. The result is that the proportion of high-frequency calls would decrease by only 2%–3% for a differential 6 dB ambient noise level increase between the frequency bands. This detection-related shift is insufficient to explain the magnitude of the observed reduction in the proportion of Site 5 high-frequency calls from 73% to 60% between 2008 and 2014, a reduction of 13 percentage points. (Had the animals shifted their source level distribution in response to changes in ambient noise background, then there would not have been any change in relative detectability at all.)

In summary, neither the statistical regressions (Fig. 9) nor the masking calculations (Fig. 11) support the hypothesis of increased low-frequency call detectability arising from changes in relative ambient noise levels.

C. The potential impact of airgun signal contamination

A final potential non-biological explanation for the relative increase in low-frequency f_{min} calls over time is that the automated algorithm was gradually permitting more airgun signals to become misclassified over time, signals that tended to have frequency content lying between 25 and 50 Hz [Fig. 7(a)]. While applying Criteria 1 of Sec. II D to the data would remove many of these misclassified signals (by restricting any localizations to the immediate vicinity of the DASARs) the possibility exists that misclassified airgun signals on one DASAR were associated with true whale calls on another DASAR, yielding positions that would lie within the threshold R_{min}. If this false position lay closest to the DASAR with the misclassified airgun signal, then the f_{min} sample would have been assigned to an airgun signal.

This concern was addressed two ways. The straightforward counterargument is that the inclusion of the categorical predictor variable Airgun had little to no impact on the measured f_{min} (Figs. 9; Tables I and II) and had little impact on the predicted trend of f_{min} with Year when Airgun was set to a fixed value, as is shown by the relatively small value of the Year:Airgun term (−0.38 Hz) in Table II, model 3 (15 km, automated). The fact that the Airgun coefficient was nonzero for automated datasets suggests that there was some misclassification of the automated samples, but not nearly enough to affect the long-term trend.

These statistical conclusions are consistent with a visual examination of the center column of Fig. 7, which compares the bearings of identified airgun signals with those from the culled whale call data sets. Were large numbers of airgun signals inadvertently being incorporated as whale calls, then the bearings of these false "whale calls," as measured from the closest DASAR, would substantially overlap the bearings of identified airgun signals. Figures 7(b) and 7(e) show that the bearing distributions of whale calls do not display local maxima that match the highly concentrated angular distributions of the airgun signals for each season, with the possible exception of 2012 and 2014 (between 60° and 90°) and to a lesser extent 2008. A similar argument could be made for the interval distributions in Fig. 7(c) and 7(f). Figure 3(g) also reveals that airgun signals decrease substantially during the last 2 years of the study, even though those years exhibit the biggest increase in low f_{min} calls [Fig. 5(d)]. Thus changing airgun misclassification rates can be discounted as an explanation for the long-term frequency trend.

D. Physiological and behavioral explanations

From this point forward we assume that the observed frequency shift represents an actual frequency shift in the sounds being produced by the whale population. We examine three physiological/behavioral hypotheses for this shift: a change in call repertoire, long-term physiological growth in the population, and a behavioral response to increasing population densities.

1. Change in call repertoire

One explanation for the observed population-scale shift in f_{min} is that the bowhead population's migrating call repertoire is changing, either by shifting the relative proportion of call types generated, or by introducing a new call type.

Bowhead and humpback whale sounds are known to evolve over time. For example, both species produce highly stereotyped and repetitive sequences of FM-modulated sweeps, or "songs," which gradually evolve over the course of a season and across years (Payne and McVay, 1971; Payne and Payne, 1985; Würsig and Clark, 1993; Noad *et al.*, 2000; Tervo *et al.*, 2009; Tervo *et al.*, 2011). We note, however, that the frequency content, sequencing, and seasonal timing of the calls analyzed here are inconsistent with what is known about bowhead song. Non-song social calls of humpback whales (Rekdahl *et al.*, 2013) have also been found to shift frequency range (frequency span) over a 10-year period. Unfortunately, the direction and magnitude of this shift was not noted, and "the measured call parameters generally showed no clear trend over time, [displaying] considerable within-call type variability."

Figure 10(a) clearly shows that different call types, particularly the C call type, display different f_{min} values, a result consistent with the Blackwell *et al.* (2007) observations that "complex" call types (type C) have lower frequency content. A relative increase in the proportion of C-type calls would tend to shift f_{min} downward. However, Fig. 10(b) demonstrates that the total proportion of C-type calls (as measured using $R_{min} = 15$ km) is not increasing consistently over time; instead, the C-type proportion fluctuates between 6% and 20%. Furthermore, Fig. 10(a) shows that *all* call types are decreasing in minimum frequency over time, although some call types are decreasing faster than others.

Could a new call type be entering the repertoire? For example, numerous low-frequency calls shown in Fig. 6 look similar to spectrograms of so-called "pulse tone" calls published by Clark and Johnson (1984). In that manuscript, the authors wrote that relatively few sounds of this type were detected, and "detailed analyses of these harmonically rich calls revealed that they were narrow band pulses with pulse repetition rates between 30 and 75 pulses/s." However, close examination of the low-frequency call samples in Fig. 6 found no pulse trains; instead, the skewed waveforms actually observed indicate that all call harmonics shared a common zero crossing and were thus tightly locked in phase. Furthermore, were a new call type to have emerged it would likely have been assigned to the generic C call type, and Fig. 10(b) already eliminates the possibility of a trend in call repertoire proportions. We thus conclude that a simple change in call repertoire is insufficient to explain the observed 7-season shift.

2. Physiological growth in the population

Another potential hypothesis proposes that the mean size of individual bowhead whales is increasing over time, enhancing their ability to generate lower frequency sounds. The resulting shift in observed f_{min} could then arise as an inadvertent byproduct of sound production by a physiologically growing population. When discussing long-term shifts in blue whale call frequencies, McDonald *et al.* (2009) considered this hypothesis unlikely, arguing that any reduction in mean individual size resulting from commercial whaling would have recovered within a decade or so after the end of widespread whaling.

Figure 10(a) suggests a quantitative test of this "growth" hypothesis. To date every physical resonator or oscillator likely to be involved in baleen whale sound production (Aroyan *et al.*, 2000; Bass and Clark, 2003; Adam *et al.*, 2013)—resonating tubes, Helmholtz resonators, simple harmonic oscillation of the larynx, radiation from an acoustic monopole—displays a power law relationship between the physical scale of the mechanism and the frequency generated: $f \sim CL^b$, where f is the output resonant frequency, b and C are fixed constants, and L is some representative dimension for the object. For example, in a resonating tube the resonant frequencies are inversely proportional to tube length L, hence $b = -1$. Similarly, the resonant frequency of a Helmholtz resonator also scales inversely with the resonator dimension (provided that all dimensions of the resonator grow proportionately), as does the resonant frequency of a gas-filled sphere and a pulsating sphere maintaining constant acoustic radiation intensity. A feature of these power law relationships is that

$$\Delta f/f = b(\Delta L/L). \tag{1}$$

Equation (1) states that if an acoustic source changes size by a certain percentage, the resulting percentage frequency shift in all of its signal components should be the same, regardless of the particular frequency examined. Thus if the 100 Hz component of a sound source changes by −10% to 90 Hz, due to an increase in source size, then a 50 Hz component co-generated by the same source should also display a −10% shift to 45 Hz.

Figure 10(a) shows how the three grouped bowhead call types span different frequencies, with the f_{min} of the C call type nearly 20 Hz lower than the baseline A call type and nearly 35 Hz lower than the B call type. One sees that each call type shifts f_{min} by a different amount and percentage over 6 years: −15 Hz (−22%) for the C type, −7 Hz (−7%) for the A type, and −15 Hz (−14%) for the B call type. The percentage change of these calls' f_{min} differ, whereas they should all be similar if all frequency shifts were byproducts in the sound production mechanism's growth. We therefore interpret Fig. 10(a) as arguing against population-wide physiological growth being responsible for the shift. This argument, however, relies on the assumptions that the different call types are produced by the same mechanism, and that this mechanism obeys a power law. More accurate and detailed modeling of the mysticete sound production mechanism might reveal more complex relationships than Eq. (1).

3. Response to increasing population density

As the number of vocalizing animals in a region increases, the frequency structure of individual calls can shift. Multiple papers have found that male frogs lower the dominant frequency of their vocalizations in the presence of other males (Lopez *et al.*, 1988), although it remains uncertain whether this shift arises from an attempt to falsely signal a larger size (Bee *et al.*, 2000), or to truthfully advertise a willingness to fight (Wagner, 1992). Male-to-male competition for female attention has also been proposed as a mechanism behind long-term frequency shifts in baleen whale calls. McDonald *et al.* (2009) hypothesized that long-term increases in population density might spawn changes in blue whale call frequency, arguing that as population density increases, the source level required to communicate between nearby animals decreases, allowing males an opportunity to generate lower

lower-frequency sounds with the same amount of physiological effort in order to exploit sexual selection preferences of females. The theory was formulated assuming deep-water propagation conditions that are independent of frequency, which is not true here, and also assumes that mate attraction is a primary function of the calls, which may not be the case for bowhead whales. Keeping these caveats in mind, it is possible to check two specific predictions of this theory. First, is the source level distribution shifting slightly lower as the mean f_{min} decreases? Second, is call density a predictor for the mean f_{min} in the regression analysis?

Aroyan *et al.* (2000), on which the theory of McDonald *et al.* (2009) is based, describes the relationship between source intensity P_0, sphere volume V, and frequency (f) of a pulsating sphere:

$$V = \frac{P_0}{\rho \pi f^2}. \tag{2a}$$

(Note that this equation also exhibits a power law between frequency and dimension, with $b = -3/2$.) From this expression one can derive the following prediction between changes in source level SL and changes in frequency shift:

$$\Delta SL = 20(\Delta f / f), \tag{2b}$$

which expresses the dB change in source level arising from a fractional change in call frequency. A 10% decrease in mean call frequency should correspond with a 2 dB *decrease* in mean source level. When the statistical regression procedures described in Sec. II F are applied to the source level distributions in Fig. 4, one finds that Year is either a non-significant factor in the manually analyzed data (ΔSL = 0) or that the mean source level in the automated datasets are *increasing* slightly with Year (0.11 dB/yr, [0.09 0.13]), regardless of the value of R_{min} chosen. The specific mechanism proposed by McDonald *et al.* (2009) is thus either incorrect or inappropriate when applied to the bowhead whale migration.

However, the detailed regression model in Sec. II F did define two predictor variables, CallRate and CallDensity, to specifically examine whether f_{min} is related to population density. This approach must assume that population density is related to call density, and that call rates or densities estimated on a DASAR are representative of what is encountered by whales within a distance R_{min} of the sensor. As discussed previously, CallRate measured raw call detection rates, while CallDensity estimated true call rate densities with a radius R_{min}. Both variables are derived quantities that can depend on other predictor variables like NoiseBand and CallingDepth. The existence of a non-zero regression coefficient between CallDensity and f_{min} is relatively uninformative for the linear-only terms in model 2 (Table I), especially since the sign and magnitude of the regression coefficient differs substantially between the four datasets.

The results for model 3, however, show more consistent evidence for a potential relationship between CallDensity and f_{min}. No manual analysis dataset produces any strong relationship, but both automated datasets ($R_{min} = 2$ and 15 km) produce similar predictions, in that a doubling in calling density shifts the predicted mean f_{min} downward by 3 to 5 Hz, with confidence bounds small enough to suggest that the effect size is real. The exact dependence is also a function of Year, with later years producing a faster decrease in f_{min} for a given CallDensity shift (i.e., the interactive term Year:CallDensity has a value of -0.39 for the 15 km dataset in Table II). Table II also shows that Range also interacts with CallDensity. By contrast, the use of CallRate yields no significant relationship. The fact that CallDensity is a more reliable predictor is intriguing, in that would confirm that animals are able to estimate the relative proximity of detected calls, and thus estimate the local density of animals, rather than just raw call detection rates. Model 4 (quintic terms) yields similar relationships to model 3, as well as higher-order interactions between NoiseBand1 and CallDensity.

We conclude that some evidence exists that f_{min} is influenced by local call density; however, Fig. 9 shows that the effect, if it does exist, is independent of the long-term frequency trend. Specifically, whenever CallDensity is held fixed in the predictive model, the observed dependence between minimum frequency and Year remains, so changes in call density over time cannot account for the observed 7-season shift in f_{min}.

V. CONCLUSION

The minimum frequency f_{min} attained by a bowhead call forms a distribution that is evolving over time. Specifically, between 2008 and 2014, the proportion of calls with f_{min} values below 75 Hz increased from 27% to 41%, shifting the mean value of f_{min} observed for the population by 10.5 Hz over seven seasons.

Long-term downward shifts in the frequency content of calls have been noted in blue whale populations, but in contrast to the results reported here, they involved shifts in the fundamental frequency of a single call type on an intra-annual or inter-annual basis (McDonald *et al.*, 2009; Gavrilov *et al.*, 2012; Miller *et al.*, 2014).

This multi-year shift cannot be explained by changes in propagation conditions, ambient noise levels, or changes in automated detector performance, although a small portion of the observed shift can be explained by relative changes in call detectability arising from differential changes in ambient noise levels at different frequencies. This shift also cannot be explained by changes in the relative proportion of different call types, and hypotheses based on physiological growth cannot explain why different call types would experience different percentage shifts. Increases in call density surrounding a DASAR were related to decreases in call frequency, but only for automated analyses, and the effect was independent of the multiyear frequency shift.

We conclude that the observed frequency shift is a population-scale behavioral change that involves more than one call type, but cannot determine whether the change is a random fluctuation in the repertoire, or a response to some evolving external condition. The relatively low values of R^2 in the regression models suggest that the minimum frequency displays a large degree of individual variation, and/or some long-term explanatory factor has not been included.

For example, if long-term ambient noise levels were increasing over a decadal-time scale, due to expanding ice-free areas in the summer, we would expect the noise increase to be larger at higher frequencies (above 100 Hz) than below 75 Hz, for the reasons presented in Sec. IV B. It would thus be reasonable for the bowhead whale population to shift to calls with lower frequency content in order to enhance call detectability, as has been noted for right whale populations (Parks *et al.*, 2007; Parks *et al.*, 2009; Parks *et al.*, 2012; Parks *et al.*, 2016). The time span of the data set is insufficiently long, however, to distinguish whether the observed 7-season shift is a simple decadal-scale random fluctuation in the repertoire, or is a long-term response to changing external factors such as open-water ambient noise levels.

We recommend that future studies on this subject conduct concurrent automated and manual statistical analyses, to enable independent side-by-side comparisons of these datasets. While some manual data might be used to retrain or update an automated detector, at least some manual data should be withheld from the training or validation of the algorithms in order to preserve the ability to conduct separate statistical analyses. Both manual and automated datasets have their quirks and weaknesses, so we feel that the ability to generate similar statistical results independently from two datasets provides a more robust analysis than an analysis of a single "blended" dataset.

ACKNOWLEDGMENTS

The authors wish to thank SEPCO for permission to use this dataset, as well as the crews of the *Norseman II*, *Alpha Helix*, and *Westward Wind*, particularly deck chief Scotty Hameister, for consistently safe and professional deployments and retrievals of the DASARs over the project life. Sheyna Wisdom and her colleagues at Fairweather LLC provided high-quality logistic support for the fieldwork. Bob Norman of Greeneridge provided helpful support on DASAR hardware and recording details. Chris Clark provided useful insights on past historical work on the bowhead call repertoire, Peter Narins provided useful references for Sec. IV D 3, and two anonymous reviewers enormously increased the quality of the paper. Finally, none of this work would have existed had Charles R. Greene, Jr. not had the vision to deploy DASARs off the Beaufort coast nearly two decades ago and to recruit all the authors to the project.

[1] See supplementary material at http://dx.doi.org/10.1121/1.5001064 for the terms and plots for model 4.

Abadi, S. H., A. M. Thode, Blackwell, S. B., and Dowling, D. R. (**2014**). "Comparison of three methods of ranging bowhead whale calls in a shallow-water waveguide," J. Acoust. Soc. Am. **136**, 130–144.

Adam, O., Cazau, D., Gandilhon, N., Fabre, B., Laitman, J. T., and Reidenberg, J. S. (**2013**). "New acoustic model for humpback whale sound production," Appl. Acoust. **74**, 1182–1190.

Aroyan, J. L., McDonald, M. A., Webb, S. C., Hildebrand, J. A., Clark, D., Laitman, J. T., and Reidenberg, J. S. (**2000**). "Acoustic models of sound production and propagation," Springer Handb. Audit. Res. **12**, 409–469.

Bass, A. H., and Clark, C. W. (**2003**). "The physical acoustics of underwater sound communication," in *Acoustic Communication* (Springer, New York), Vol. 16, pp. 15–64.

Bee, M. A., Perrill, S. A., and Owen, P. C. (**2000**). "Male green frogs lower the pitch of acoustic signals in defense of territories: A possible dishonest signal of size?," Behav. Ecol. **11**, 169–177.

Blackwell, S. B., Nations, C. S., McDonald, T. L., Greene, C. R., Jr., Thode, A. M., Guerra, M., and Macrander, A. M. (**2013**). "Effects of airgun sounds on bowhead whale calling rates in the Alaskan Beaufort Sea," Mar. Mamm. Sci. **29**, E342–E365.

Blackwell, S. B., Nations, C. S., McDonald, T. L., Thode, A. M., Mathias, D., Kim, K. H., Greene, C. R., Jr., and Macrander, A. M. (**2015**). "The effects of airgun sounds on bowhead whale calling rates: Evidence for two behavioral thresholds," PLoS One **10**, e0125720.

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**, 255–270.

Clark, C. W., and Johnson, J. H. (**1984**). "The sounds of the bowhead whale, Balaena mysticetus, during the spring migrations of 1979 and 1980," Can. J. Zool. **62**, 1436–1441.

Cummings, W. C., and Holliday, D. V. (**1987**). "Sounds and source levels from bowhead whales off Pt. Barrow, Alaska," J. Acoust. Soc. Am. **82**, 814–821.

Delarue, J., Laurinolli, M., and Martin, B. (**2009**). "Bowhead whale (*Balaena mysticetus*) songs in the Chukchi Sea between October 2007 and May 2008," J. Acoust. Soc. Am. **126**, 3319–3328.

Dobson, A. J., and Barnett, A. (**2008**). *An Introduction to Generalized Linear Models* (CRC Press, New York).

Gavrilov, A. N., McCauley, R. D., and Gedamke, J. (**2012**). "Steady inter and intra-annual decrease in the vocalization frequency of Antarctic blue whales," J. Acoust. Soc. Am. **131**, 4476–4480.

Greene, C. R., 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**, 799–813.

Ljungblad, D. K., Thompson, P. O., and Moore, S. E. (**1982**). "Underwater sounds recorded from migrating bowhead whales, *Balaena mysticetus*, in 1979," J. Acoust. Soc. Am. **71**, 477–482.

Lopez, P. T., Narins, P. M., Lewis, E. R., and Moore, S. W. (**1988**). "Acoustically induced call modification in the white-lipped frog, *Leptodactylus albilabris*," Anim. Behav. **36**, 1295–1308.

McDonald, M. A., Hildebrand, J. A., and Mesnick, S. (**2009**). "Worldwide decline in tonal frequencies of blue whale songs," Endang. Spec. Res. **9**, 13–21.

McDonald, M. A., Hildebrand, J. A., and Wiggins, S. M. (**2006**). "Increases in deep ocean anibient noise in the northeast pacific west of San Nicolas Island, California," J. Acoust. Soc. Am. **120**, 711–718.

Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," J. Acoust. Soc. Am. **107**, 3518–3529.

Miller, B. S., Leaper, R., Calderan, S., and Gedamke, J. (**2014**). "Red shift, blue shift: Investigating Doppler shifts, blubber thickness, and migration as explanations of seasonal variation in the tonality of Antarctic Blue Whale Song," PLoS One **9**, E107740.

Moore, S. E., and Reeves, R. R. (**1993**). "Distribution and movement," in *The Bowhead Whale*, edited by J. Burns, J. Montague, and C. Cowles (Allen Press, Lawrence, KS), pp. 313–386.

Moore, S. E., Stafford, K. M., Mellinger, D. K., and Hildebrand, J. A. (**2006**). "Listening for large whales in the offshore waters of Alaska," Bioscience **56**, 49–55.

Noad, M. J., Cato, D. H., Bryden, M. M., Jenner, M.-N., and Jenner, C. S. (**2000**). "Cultural revolution in whale songs," Nature **400**, 537.

Parks, S. E., Clark, C. W., and Tyack, P. L. (**2007**). "Short- and long-term changes in right whale calling behavior: The potential effects of noise on acoustic communication," J. Acoust. Soc. Am. **122**, 3725–3731.

Parks, S. E., Groch, K., Flores, P., Sousa-Lima, R., and Urazghildiiev, I. R. (**2016**). "Humans, fish, and whales: How right whales modify calling behavior in response to shifting background noise conditions," in *Effects of Noise on Aquatic Life II*, edited by A. N. Popper and A. Hawkins (Springer Science and Business Media, New York), pp. 809–813.

Parks, S. E., Johnson, M. P., Nowacek, D. P., and Tyack, P. L. (**2012**). "Changes in vocal behavior of North Atlantic right whales in increased noise," in *Effects of Noise on Aquatic Life*, edited by A. N. Popper and A. Hawkins (Springer Science and Business Media, New York), pp. 317–320.

Parks, S. E., Urazghildiiev, I., and Clark, C. W. (**2009**). "Variability in ambient noise levels and call parameters of North Atlantic right whales in three habitat areas," J. Acoust. Soc. Am. **125**, 1230–1239.

Payne, K., and Payne, R. (**1985**). "Large scale changes over 19 years in songs of humpback whales in Bermuda," A. Tierpsychol. **68**, 89–114.

Payne, R., and McVay, S. (**1971**). "Songs of humpback whales," Science **173**, 585–597.

Rekdahl, M. L., Dunlop, R. A., Noad, M. J., and Goldizen, A. W. (**2013**). "Temporal stability and change in the social call repertoire of migrating humpback whales," J. Acoust. Soc. Am. **133**, 1785–1795.

Stafford, K. M., Moore, S. E., Laidre, K. L., and Heide-Jørgensen, M. P. (**2008**). "Bowhead whale springtime song off West Greenland," J. Acoust. Soc. Am. **124**, 3315–3323.

Tervo, O. M., Parks, S. E., Christoffersen, M. F., Miller, L. A., and Kristensen, R. M. (**2011**). "Annual changes in the winter song of bowhead whales (*Balaena mysticetus*) in Disko Bay, Western Greenland," Mar. Mamm. Sci. **27**, E241–E252.

Tervo, O. M., Parks, S. E., and Miller, L. A. (**2009**). "Seasonal changes in the vocal behavior of bowhead whales (*Balaena mysticetus*) in Disko Bay, Western-Greenland," J. Acoust. Soc. Am. **126**, 1570–1580.

Thode, A. M., Blackwell, S. B., Seger, K. D., Conrad, A. S., Kim, K. H., and Macrander, A. M. (**2016**). "Source level and calling depth distributions of migrating bowhead whale calls in the shallow Beaufort Sea," J. Acoust. Soc. Am. **140**, 4288–4297.

Thode, A. M., Kim, K. H., Blackwell, S. B., Greene, C. R., and Macrander, M. A. (**2012**). "Automated detection and localization of bowhead whale sounds in the presence of seismic airgun surveys," J. Acoust. Soc. Am. **131**, 3726–3747.

Thode, A. M., Kim, K., Greene, C. R., and Roth, E. H. (**2010**). "Long range transmission loss of broadband seismic pulses in the Arctic under ice-free conditions," J. Acoust. Soc. Am. **128**, EL181–EL187.

Wagner, W. E. (**1992**). "Deceptive or honest signalling of fighting ability? A test of alternative hypotheses for the function of changes in call dominant frequency by male cricket frogs," Anim. Behav. **44**, 449–462.

Wenz, G. M. (**1962**). "Acoustic ambient noise in the ocean: Spectra and sources," J. Acoust. Soc. Am. **334**, 1936–1956.

Würsig, B., and Clark, C. (**1993**). "Behavior," in *The Bowhead Whale*, edited by J. J. Burns, J. J. Montague, and C. J. Cowles (Society of Marine Mammalogy, Lawrence, KS), pp. 157–199.