JASA ARTICLE

North Atlantic right whale detection and localisation using deep learning, spectrogram cross-correlation, and nonlinear Bayesian inversion

Romina A. S. Gehrmann,[1,a)] Oliver S. Kirsebom,[2] Bruno Padovese,[3] Fabio Frazao,[3] Graham A. Warner,[4] Kamden P. D. Thebeau,[1] David R. Barclay,[5] and Sean Pecknold[1]

[1]*Defence Research and Development Canada, Dartmouth, Nova Scotia, Canada*

[2]*Open Ocean Robotics, Victoria, British Columbia, Canada; School of Environmental Science, Simon Fraser University, Burnaby, British Columbia, Canada; Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada*

[3]*School of Environmental Science, Simon Fraser University, Burnaby, British Columbia, Canada; Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada*

[4]*Previously at JASCO Applied Sciences, Victoria, British Columbia, Canada*

[5]*Faculty of Science, Dalhousie University, Halifax, Nova Scotia, Canada*

ABSTRACT:

We present a processing routine for marine mammal call detection and localisation, using North Atlantic right whale (NARW) upcalls and moans passively recorded by a sonobuoy grid in July 2018 in the southern Gulf of St. Lawrence, Canada. Right whales are critically endangered and at risk from ship strikes and fishing gear entanglement. We demonstrate the effectiveness of passive acoustic monitoring followed by a three-step analysis as a mitigation tool. The first step detects NARW calls using a publicly available deep learning model, trained on open acoustic datasets, adapted and optimised for the sonobuoy dataset using a smaller adaptation set. The second step applies spectrogram correlation for contact association and time difference of arrival estimates, identifying the same call across multiple sonobuoys. The third step estimates call location via nonlinear Bayesian inversion, assuming direct call propagation, yielding source positions with uncertainty estimates. This three-step analysis effectively detects and localises NARW upcalls, especially from within or near the sonobuoy grid. It enhances information compared to visual surveys and provides realistic uncertainty estimates, supporting its role in conservation and management efforts. https://doi.org/10.1121/10.0042758

(Received 12 August 2025; revised 3 December 2025; accepted 7 January 2026; published online 11 March 2026)

[Editor: Stan E. Dosso] Pages: 2123–2138

I. INTRODUCTION

Localising North Atlantic right whales (NARW) in the Gulf of St. Lawrence (GSL) has become an important task toward preservation of the critically endangered species. Historically, NARW migrate along the western North Atlantic between Florida, United States, and Newfoundland, Canada. Since 2015, NARW have been regularly observed foraging in the GSL. The shift in foraging behavior is climate-driven due to warming of the North Atlantic and changes in ocean circulation altering prey availability. NARW, since then, have suffered from elevated rates of anthropological-caused mortality, such as ship strikes and entanglement in fishing gear (Crowe *et al.*, 2021; Davies and Brillant, 2019; Davis *et al.*, 2017; Meyer-Gutbrod *et al.*, 2021, 2023). The government of Canada has established dynamic vessel and fishery management zones to preserve the NARW population, balancing economic and environmental interests (Minister of Transport, Chrystia Freeland, 2025). The successful implementation of these dynamic zones therefore depends on high spatial and temporal resolution in NARW distribution and their uncertainties (e.g., Ross *et al.*, 2021; Indeck *et al.*, 2025).

Passive acoustic monitoring (PAM) to monitor marine mammal presence and density based on their vocalisation habits has been widely established (e.g., Zimmer, 2011). PAM is more cost effective than aerial and ship-borne surveillance, less impacted by adverse weather conditions, and suitable for continuous monitoring during day and night. PAM has been successfully implemented for NARW detection on acoustic data from moored hydrophones as well as real-time implementations on underwater gliders (e.g., Baumgartner *et al.*, 2013; Simard *et al.*, 2019; Davis *et al.*, 2017).

Our study presents a workflow establishing the presence of NARW and their location and associated uncertainties using acoustic data from two sets of sonobuoy deployments in 2018 in a dynamic vessel and fishery management zone. This workflow combines a set of state-of-the-art methods: detection of NARW upcalls by adapting a pre-trained deep learning (DL) model, cross-spectrogram cross correlation

[a)]Email: Romina.Gehrmann@forces.gc.ca

for contact association, and Bayesian inversion for source location and uncertainty estimates.

A. Introduction to the study area

The focus of our study is the Transport Canada-defined restricted area located south of Gaspe Peninsula (Fig. 1). Location and size of the restricted area are based on historical data of NARW aggregations, and restrictions for vessels to avoid the area or reduce speed are triggered by near real-time detections (Transport Canada, 2024). The restricted area overlaps with the Shediac Valley, an area of 1530 km^2, which is rich in biodiversity and a variety of fish species, such as the Atlantic cod seek refuge, feed, nurse, and spawn in the area. It is a suitable NARW foraging habitat all year round with a larger food (mostly zooplankton *Calanus finmarchicus*) abundance from June to August (Plourde *et al.*, 2024). Meyer-Gutbrod *et al.* (2023) note, however, that NARW are most likely driven to forage in the southern GSL by decreased prey availability in previously consistent summer foraging habitats in the Gulf of Maine and Scotian Shelf, and alert about the general decline of the food source as an additional stressor on NARW.

B. Introduction to the dataset

The underwater acoustic dataset used in this study is from two sets of sonobuoy deployments in a core NARW foraging habitat in the southern GSL on 30 and 31 July 2018, here referred to as day 1 and day 2, respectively. Each day 32 sonobuoys were deployed in a grid at about 8 km apart (see Fig. 1) recording up to 6 h of underwater acoustic data each. In addition, visual surveys for the presence of NARW were performed on board the survey vessel (both days) and from an airplane (on day 2). A Slocum ocean glider acquired conductivity, temperature, and pressure/depth (CTD) data within the sonobuoy array.

The sonobuoys were deployed from a Royal Canadian Air Force aircraft. Upon impact with water, sonobuoys inflate a surface float with a radio transmitter for communication and release a hydrophone to a pre-specified depth, here, ∼27.4 m below sea level. The acoustic data were telemetered back to the aircraft through a radio link in multiplexed format and recorded. After the survey, the data were demultiplexed and saved in .wav format. Note that the data made available for the Detection, Classification, Localisation, and Density Estimation (DCLDE) 2024 workshop were first downsampled to 8000 Hz but later re-extracted and downsampled to 7350 Hz. We use the latter version of the dataset. The original dataset had recording times with second precision. We use a matched filter to re-create the recording times with respect to a single sonobuoy using the radio frequency (RF) signal present on all sonobuoys at the same time resulting in ms precision.

On day 1 and day 2 sonobuoy type AN/SSQ-53F and 53D3 were used, respectively. The main difference between the types is that coordinates for the 53D3 buoys are only known for the time when the buoy hits the water, whereas the 53 F buoys transmitted regular updates during the deployment (von Benda-Beckmann *et al.*, 2022).

Sonobuoy arrays have unique advantages and disadvantages. One advantage is that they are time-synchronised, which allows using the time difference of arrival methodology for target localisation. The disadvantage is that sonobuoy deployments are time-limited to a few hours. They are easy to deploy, however, in a hydrophone grid of your choosing and offer a near-real time data acquisition and analysis, which make them suitable for short-term

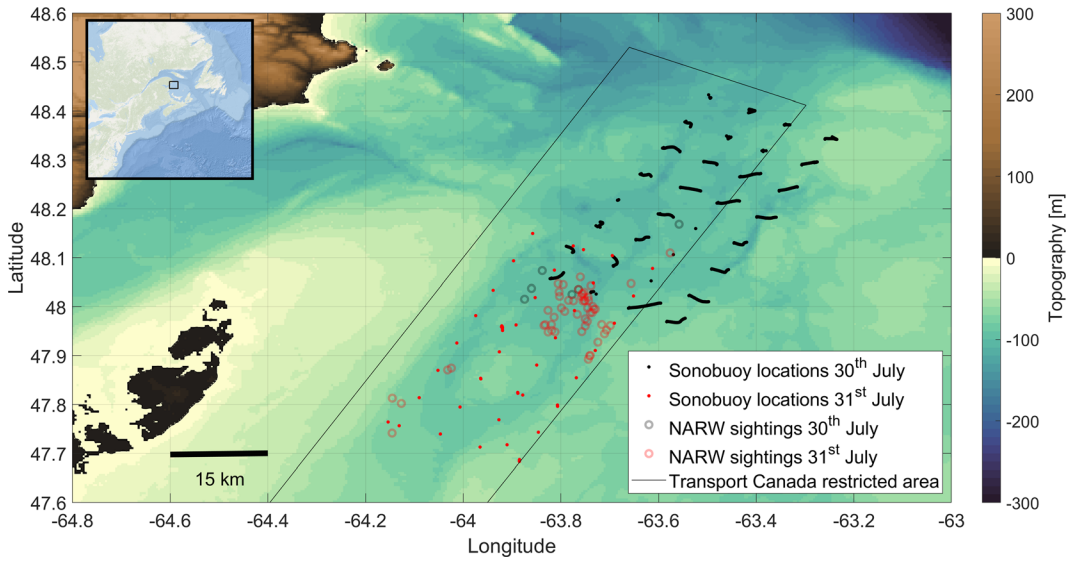

FIG. 1. Topography of experiment area in the southern Gulf of St. Lawrence, southeast of Gaspe Peninsula (GEBCO Compilation Group, 2024, GEBCO 2024 Grid, doi:10.5285/1c44ce99-0a0d-5f4f-e063-7086abc0ea0f). Sonobuoy locations (black dots: 30 July 2018, showing the drift of the buoys over time; red dots: 31 July 2018) and NARW sightings (black circles: 30 July 2018; red circles: 31 July 2018) within the Transport Canada defined restricted area near the Shediac valley (Transport Canada, 2024). The inlay in the top left corner is the ESRI oceans map showing the eastern part of North America and the survey area (black rectangle).

https://doi.org/10.1121/10.0042758

experiments with medium to high frequency resolution for underwater signals.

The data were made publicly available as part of the Detection, Classification, Localisation, and Density Estimation of Marine Mammals workshop in 2024. The true location of the calling whales, however, remained unknown (von Benda-Beckmann *et al.*, 2022). Hence, we compare our source location results to visual observations.

C. NARW vocalisations

The acoustic dataset was manually annotated for four different right whale call types, upcalls, gunshots, moans and screams, as well as minke whale pulse trains, low-frequency sounds by blue, fin, or sei whales, blows by unknown species, and generally unknown signals. The total annotated NARW vocalisations are below ∼600 on day 1 and below ∼1300 on day 2 (von Benda-Beckmann *et al.*, 2022). In this paper we focus on the NARW low-frequency upcalls and moans.

Upcalls are frequency-modulated upsweeps between about 50 and 200 Hz. Higher frequency harmonics up to 3140 Hz are often attenuated by the time the signal reaches the receivers (Matthews and Parks, 2021; Parks *et al.*, 2011). The annotated upcalls are up to 1.5 s long [Fig. 2(a)].

Moans are generally in the range of 50–500 Hz and vary in amplitude and frequency modulations. NARW have been observed to produce moans facultatively, within the top 10 m of the water and at a higher rate when in larger groups (Matthews *et al.*, 2001). Annotated moans can be up to 6 s long and some show initial downsweep characteristics that level out after 1–2 s [Fig. 2(b)]. In some cases we observe higher-frequency harmonics.

D. Introduction to DL methods for whale call detection

Computational bioacoustics has significantly benefited from the advent of DL to address and automate the analysis of large acoustic datasets that were previously unmanageable due to their sheer volume and complexity (Stowell, 2022). Traditional methods of detecting and classifying marine mammal vocalisations were time-intensive, often relying on hand-engineered features and manual analysis (Baumgartner and Mussoline, 2011; Binder and Hines, 2012). Deep Neural Networks (DNNs) have shifted this paradigm, allowing researchers to leverage powerful models that can implicitly learn complex patterns directly from the acoustic time series or from straightforward transformations, such as spectrograms, drastically reducing the level of feature engineering required by other methods (e.g., Li *et al.*, 2020; Kirsebom *et al.*, 2020).

A DNN consists of multiple interconnected layers that sequentially learn hierarchical representations of input features, from low-level components, such as frequencies and time segments, to more abstract representations of specific calls or environmental noises (Goodfellow *et al.*, 2016). Convolutional Neural Networks (CNNs), a subtype of DNNs, have been particularly impactful in bioacoustic

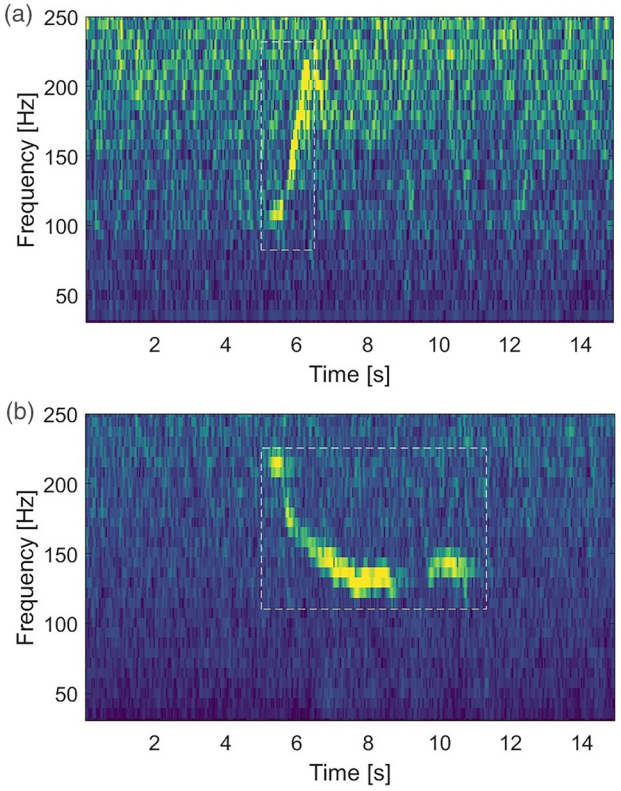

FIG. 2. Example of a spectrogram representation of a NARW upcall (a) and of a NARW moan (b). Sound pressure spectrum level (dB re μPa^2/Hz) increases from dark blue to yellow.

applications due to their proficiency in extracting spatial and temporal patterns when combined with spectrogram representations of the audio.

For NARW classification and detection, the upcall vocalisation [Fig. 2(a)] is widely recognised as a primary acoustic indicator of presence (Parks and Tyack, 2005; Parks *et al.*, 2011). This vocalisation is produced by both sexes across various behavioral contexts, thus making the upcall an ideal target for automated detection due to its distinct and consistent acoustic signature (Kirsebom *et al.*, 2020; Padovese *et al.*, 2023; Shiu *et al.*, 2020). For this study, we utilise a pre-trained DNN that specifically trained to detect and classify NARW upcalls (Padovese *et al.*, 2023), based on an open-source, Python-based toolbox, Ketos, (Kirsebom *et al.*, 2021) designed for marine bioacoustics. We choose this tool over other detection algorithms due to the data-driven adaptation ability. We evaluate the model's performance against the annotations. Note that the spectrogram cross correlation and localisation analysis described below are also applied to upcalls, which were annotated but missed by the detector, and moans, which were only manually annotated.

E. Introduction to time difference of arrival (TDOA) analysis and spectrogram cross correlation approaches

When a signal is detected in multiple, spatially separated receivers, the TDOAs can be used to constrain the

location of the source. TDOA geometric localisation is a well-established approach in underwater acoustics since the 1960s (Watkins and Schevill, 1972). Due to the nonlinearity of the problem and measurement uncertainties, the source location is often inferred using least squares or maximum likelihood algorithms (e.g., Knapp and Carter, 1976; Chan *et al.*, 2006). The accuracy of these algorithms depends on the precision of the receiver locations and the synchronisation of internal clocks. The accuracy with which the sound propagation can be modelled also plays into the accuracy of the source localisation.

Depending on the number of receivers in the array, the array geometry, and the acoustic source position relative to the array, the source may be localised in one, two, or all three spatial dimensions with varying levels of accuracy using TDOA (e.g., Thode, 2005). In general, at least three receivers are required to localise the source in two spatial dimensions, whereas four receivers are required to localise the source in all three dimensions. The technique generally requires that the source and the receivers are in line of sight (e.g., Li *et al.*, 2016).

Determining the time of arrival of a NARW upcall with sufficient precision to localise the calling whale is a non-trivial task, even for the wide-aperture array considered in the present work. Sound propagation effects cause calls to be received differently at different locations, which makes the determination of the arrival time an ambiguous task. The finite (∼1.0 s) duration of the NARW upcall easily causes errors of over 100 ms, a sizable error considering that the maximum TDOA between any two adjacent receivers separated by 10 km is ∼7 s.

One commonly used method to reduce systematic errors and obtain a robust TDOA estimate involves computing the cross correlation of the acoustic signals measured at the receivers. The computation may be performed either in the time or frequency domain. Clark and Ellison (2000), for example, estimate the TDOA of bowhead whale calls using 3–5 hydrophone arrays with 1–4.5 km spacings. They used both spectrogram and waveform correlation and point out the improved time resolution using waveform correlation. The benefit of the spectrogram cross correlation is that it displays the call's time-varying frequency modulated structure and that this structure is more robust against propagation effects (such as interference nulls at the receiver). Simard and Roy (2008) find using spectrograms generally easier for TDOA estimation than using a filtered time series due to noise interference. Tiemann *et al.* (2004) also compared the waveform vs spectrogram correlation and found that the localisations using spectrogram correlations yielded a better match with visual observations. Helble *et al.* (2015) point out the benefit of several second-long sequences when applying cross correlation of humpback whale songs, which improve the TDOA results. NARW upcalls, however, may not occur in comparable sequences.

Following an approach adopted in multiple previous studies (Fucile *et al.*, 2006; Hendricks *et al.*, 2019; Janik *et al.*, 2000; Ramaswamy *et al.*, 2001), we compute the cross correlation in the frequency domain. This approach has the advantage of being computationally more efficient and allowing us to apply certain noise-suppression steps, described in detail below. These methods are essential to building an automated system capable of localising NARW in near real time in a realistic deployment scenario with limited resources for *in situ* data processing at the receiver and constrained bandwidth for data transmission. The method is also more suitable for large-aperture arrays, such as the one considered here, where uncertainties introduced by sound propagation effects dominate. Spectrogram cross correlation has been successfully applied not only to TDOA estimation but also to detect calls in the first place (e.g., Mellinger and Clark, 2000, found that the method did only slightly worse compared to neural networks, although the comparison was done against a shallow, fully connected neural network, i.e., from before the advent of modern deep CNNs).

F. Introduction to nonlinear Bayesian inversion approaches

Localising marine mammals using the TDOAs can be solved with deterministic inversion techniques. However, the solution to the inverse problem is typically non-unique and there may be several models for which the predicted data fit the observed data equally well within the data error (e.g., Tarantola, 2005). Bayesian inversion is a statistical approach that uses probability to estimate which possible solutions are most likely, using the measured data, the data error, and the prior knowledge about the system. Instead of producing a single "best-fit" answer, it provides a range of possible solutions and their associated probabilities. This means it not only locates the marine mammal but also quantifies how uncertain that estimate is. Therefore, it is important to evaluate data errors that are comprised of measurement errors due to ambient noise and instrumentation limitations, and systematic errors due to approximations in the physical theory implementation, as well as errors based on the choice of parametrisation of the model.

The underlying physical system is often nonlinear, which can be overcome with a linearised approximation for the case at hand. However, linearisation generally does not allow a rigorous uncertainty estimation. Instead, we implement a nonlinear Bayesian inversion (based on work by Warner *et al.*, 2017; Dosso and Wilmut, 2008), which estimates a posterior probability density for all unknown or uncertain parameters. The model and uncertainty estimated from Bayesian inversion need to be understood as the most appropriate solution for the observed data given the specific parametrisation (e.g., Gelman *et al.*, 2008). In this study, parametrisation beyond the unknown source location includes unknown receiver location differences to the latest location update, or original deployment position, a time delay between individual buoys, and the average sound speed.

Bayesian inversion is generally more time intensive than deterministic inversions, but our approach takes advantage of the short computational effort required for the

TDOA forward model (Spiesberger, 2005). The technique performs relatively well compared to inversions that use more accurate forward models (Warner *et al.*, 2016), which take the modal-dispersion in a shallow-water environment into account.

G. An integrated approach to NARW acoustic localisation

Our paper addresses the challenge of localising NARW using underwater acoustic data from a sonobuoy grid. We use a pre-trained DL model to automate the detection of NARW upcall. Through transfer learning, we adapt the model to the target dataset, demonstrating that a small set of training samples is sufficient to achieve a significant improvement in performance.

We describe and implement an algorithm for estimating TDOAs based on spectrogram cross correlation functions, which utilises compressed (binary) spectrograms. The reduced data size generally (although not tested here) allows for efficient data transfer from the individual hydrophones to a central or cloud-based server.

Finally, we discuss a Bayesian inversion approach to estimating the source location, addressing uncertainties in the receiver parameters, and assessing the source location uncertainty, which can be substantial, especially when the sound source is outside the array. The inversion algorithm is more expensive computationally than the detection and TDOA estimation steps but enables quantitative estimates of the location uncertainties.

In the absence of ground truth—in the form of known locations of vocalising whales—we validate our methodology by comparing our source localisation estimates to visual observations using the Mahalanobis distance.

We argue (but do not demonstrate) that our solution can be implemented in a practical setting with realistic power budgets, computational resources, and data transmission bandwidths, providing localisation estimates in near real time. Call detection and spectrogram computation and compression can occur locally using individual hydrophone data, and source localisation through TDOA analysis and Bayesian inversion can take place on central server or in the cloud.

Although our solution was developed for a sonobuoy dataset, we argue that it is applicable to a broader range of hydrophone array deployments, the key requirements being well-synchronised hydrophones with known positions and capacity for local data processing and data transmission.

II. METHODS

A. DL model for upcall detection

For the task of detecting and classifying NARW upcalls, we utilised a pre-trained model available in the NARW detection tool (Padovese *et al.*, 2023). The tool employs a conventional ResNet-18 architecture (He *et al.*, 2016), a residual neural network consisting of 18 layers that utilise skip connections to facilitate the learning of deeper features without performance degradation. ResNet architectures have been widely adopted in bioacoustic classification tasks due to their relative ease of implementation and robust performance across various species and vocalisation types (Stowell, 2022). The network was trained on over 50 000 NARW upcalls clips, collected from hydrophones stationed at various locations. Additional details on the tool and model's implementation can be found in Padovese *et al.* (2023) and its associated GitLab repository.[1]

The ResNet-18 model operates on 3-s magnitude spectrograms with a frequency range of 0–500 Hz to classify audio segments as either containing a NARW upcall or not. To generate spectrograms compatible with the ResNet model, audio files from the DCLDE 2024 dataset were first downsampled to 1000 Hz and then pre-processed. For each 3-s segment, a magnitude spectrogram was computed using a Hamming window with a duration of 0.256 s (256 samples) and a step size of 0.032 s, resulting in an 87.5% overlap between consecutive windows. This configuration produced a two-dimensional spectrogram with dimensions of 94 time frames and 129 frequency bins. Finally, the spectrograms were standardised individually by subtracting its mean and dividing by its standard deviation to ensure consistent scaling across all inputs (Stowell, 2022). The preprocessing approach, along with the chosen time-frequency configuation, follows similar, though not identical, methodological conventions established in previous NARW upcall detection studies (e.g., Kirsebom *et al.*, 2020; Shiu *et al.*, 2020).

The detection is performed twice. In an initial run the original model is used without modifications. Three performance metrics are estimated, recall, precision, and number of false positives per hour. Recall is calculated by dividing the true positives by the sum of the true positives and the false negatives. Precision is calculated by dividing the true positives by the sum of the true positives and the false positives (Saito and Rehmsmeier, 2015). Precision decreases with the number of false detections and is smaller when there are initially less true detections. It might therefore not be the best indicator to decide if the detector results are sufficient. A better indicator could be the number of false positive per hour. Shiu *et al.* (2020), for example, use a threshold of 20 false positives per hour to evaluate their detector success.

After the initial run, we inspect the detections and annotate them as hard negatives (background) if they were false positives or positives (upcalls) if they were true positives. We then adapt the NARW upcall model with 1992 background (hard negatives) and 194 upcalls (true positives) from the sonobuoy dataset as described in Padovese *et al.* (2023). The detection is then performed a second time using the adapted model and followed by metrics evaluation. Note: The decision to adapt the model to the sonobuoy dataset and run it a second time resulted from evaluating the first run. It may, however, be sufficient to run it only one time for other datasets that are more similar to the original training dataset.

B. Source localisation from TDOA

The TDOA of the call is based on a simple direct-path propagation model. A source signal, such as the whale call, at time t_S, at the coordinates x_S, and y_S is received at receiver i with the coordinates x_i and y_i at time t_i. The time from the source to the receiver is the path length divided by the average speed of sound through the water c. Including a potential time delay of the receiver clock Δt_{Ri}, the arrival time becomes

$$t_i = t_S + \frac{\sqrt{(x_S - x_i)^2 + (y_S - y_i)^2}}{c} + \Delta t_{Ri}. \quad (1)$$

Subtracting the time from the source to the receiver for two different receivers i and j, the TDOA between two receivers becomes $\Delta t_{ij} = t_i - t_j$, eliminating absolute measures such as t_S, and introducing Δt_{RiRj}, the relative time delay between two receivers' internal clocks.

With $3 + 2N + (N - 1)$ parameters for N receivers, but only N-1 independent TDOAs, the problem becomes over-parameterised, which may lead to overfitting the data and therefore fitting the data error, as well as overestimating parameter uncertainties. Note that additional data points from pairing different receivers results in data points that are correlated with each other and do not offer additional information. Fortunately, receiver locations are relatively well-known (most recent Global Positioning System (GPS) fix from the aircraft on day 1, or drop off point on day 2) and our prior is a uniform distribution of a relative drift in position. Additionally, the internal sonobuoy time delays are to be assumed minimal, and we choose a uniform prior with 2 s maximum delay.

In the present case, with the receivers all found at the same depth (27.4 m below sea level), the depth of the acoustic source is essentially unconstrained by TDOA, whereas the source's latitudinal and longitudinal position may be determined with varying levels of uncertainty.

C. TDOA estimation using spectrogram cross correlation

The technique presented in this section aims to enable near real-time processing in a realistic deployment scenario where computational resources and transmission bandwidth are limited.

1. Spectrogram computation and data compression

In the following, we describe the steps adopted for computing the spectrograms used as input for the TDOA analysis. Magnitude spectrograms are computed using a window size of 256 ms and step size of 12.8 ms (95% overlap) using a Hamming window function and converted to decibel scale (Fig. 3, left).

Next, a number of cuts and noise suppression methods are applied to the spectrogram with the dual purpose of enhancing the cross correlation signal and reducing the amount of transmitted data. Row and column medians are subtracted to suppress narrow-band, tonal noise and broad-band, impulsive noise, respectively (Fig. 3, middle). Frequency bins below 60 Hz and above 260 Hz are discarded. A fixed (configurable) threshold of 8 dB is applied, assigning the value 0 to pixels below this threshold and the value 1 to pixels above this threshold. All isolated, positive (1) pixels were subsequently converted to negative (0) pixels (Fig. 3, right).

This last step reduces the data size by a factor of 32 because every pixel in the spectrogram is represented by a binary value (0 or 1) instead of a 32-bit precision float. With the spectrogram dimensions and resolution used in this study, the data size is reduced from 134 kb to 4.2 kb. Depending on the spectral structure of the background noise, the data size can in some cases be further reduced by converting from a matrix representation of 0 s and 1 s to a point-cloud representation in which only the positions of the positive pixels are stored.

The 8-dB threshold used here was found to provide a good balance between upcall sensitivity and data compression. A systematic exploration and optimisation of this threshold parameter as well as the window and step size used in the spectrogram computation is beyond the scope of the present study.

2. Event construction and cross correlation analysis

A NARW acoustic event is triggered when an upcall is detected by the DL model in any of the receivers, with a score and signal-to-noise ratio (SNR) above fixed (configurable) threshold values. A simple algorithm (Kirsebom *et al.*, 2020) is used for estimating the SNR of the call. In this study, we use a score threshold of 0.5 and SNR threshold of 8 dB.

Given the upcall detection time, t_i, in the primary (triggering) receiver i, a "trigger window" is computed for all other receivers as $(t_i - r_{ij}/c - \Delta t; t_i + r_{ij}/c + \Delta t)$, where r_{ij} is the straight-path distance between receivers i and j, c is the speed of sound, assumed here to have the constant and uniform value $c = 1,480$ m/s within the water volume sampled by the receiver array, and $\Delta t = 3$ s is the uncertainty on the detection time, given by the width of the input spectrogram of the DL model. Any upcall detected by the DL model within the trigger window at other receivers is considered part of the same acoustic event. Note that multiple upcalls may occur in the same receiver within the same trigger window. In the present study, the same score and SNR thresholds were used to accept/reject the primary/triggering signal and the secondary signals.

For each detected upcall, we extract a 9.0-s wide segment (the 3.0-s wide window flagged by the DL model, plus 3.0 s on either side) and compute the compressed/binary spectrogram representation described above. For every pair of detections, a and b, we slide the central 3.0-s window of a across the full 9.0-s window of b, at every step computing

https://doi.org/10.1121/10.0042758

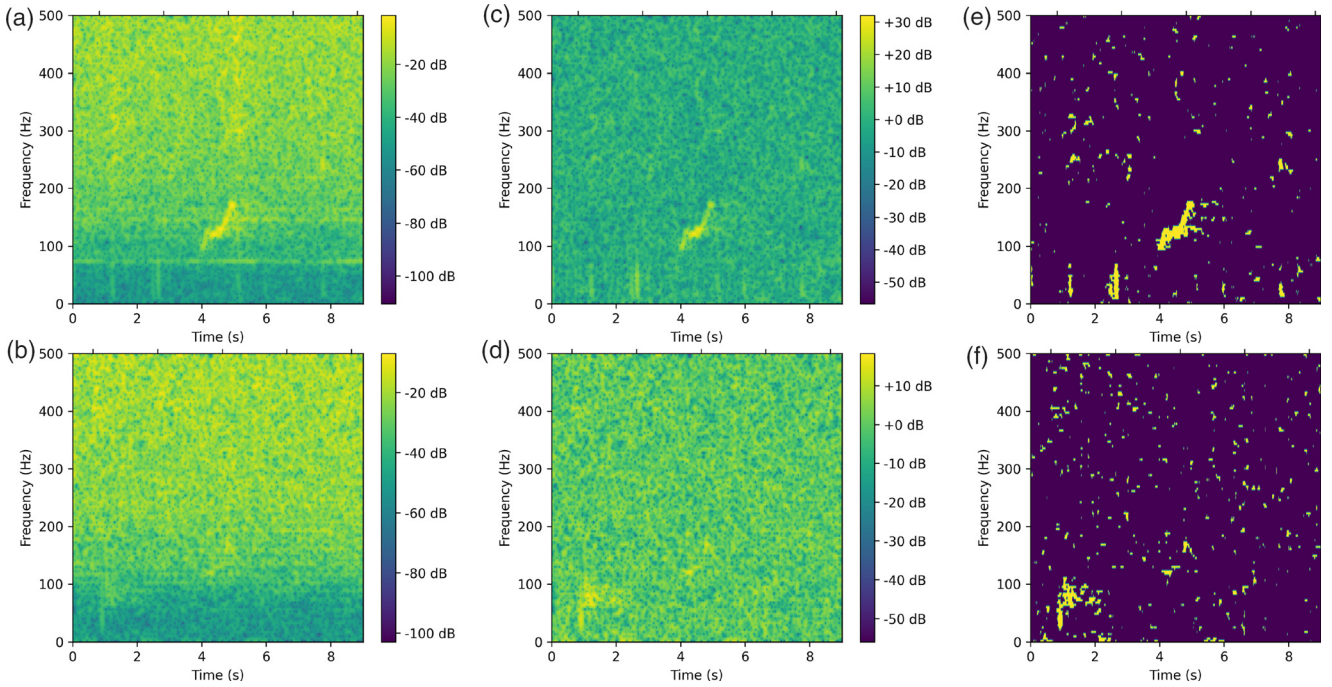

FIG. 3. Two examples of spectrogram processing results for upcalls with estimated SNRs of 21 dB (top row) and 7.6 dB (bottom row). From left to right: raw spectrogram (dB, uncalibrated), denoised spectrogram (dB, uncalibrated), and binary spectrogram. The high-SNR upcall (top) was detected in receiver 40 at 17:43:14.2 on 31 July 2018, whereas the low-SNR upcall (bottom) was detected in receiver 97 at 17:43:20.4 on 31 July 2018.

the cross correlation of the two spectrograms. The time shift yielding the largest cross correlation is then used as the TDOA (Fig. 4).

This approach to estimating the TDOA is expected to work well when the upcall has similar spectral structure in the two receivers (i.e., sound propagation effects have not altered the upcall substantially) and is reasonably free of overlapping transient noise. When these assumptions are not fulfilled, the spectrogram cross correlation approach may not yield an accurate TDOA estimate. We use the ratio of the maximum and the median cross correlation value as a metric to automatically detect poor-quality cross correlation functions. In this study, a ratio of 10 or greater was required for a cross correlation function to be accepted. See Fig. 4 for a comparison between an accepted (left) and rejected (right) instance.

In cases where multiple detections occur in the same receiver, the detection with the largest, aggregate cross correlation (summed over the detections in the other receivers) is selected and the others discarded. For an event with N detections in separate receivers, we thus obtain $N^2 - N$ TDOA values. Using a χ^2 minimisation routine, we use this to constrain the $N - 1$ independent TDOAs, obtaining best-fit values and estimates of statistical uncertainties σ. These form the inputs for the source localisation routine, discussed below.

We stress that these computational steps take place on a single machine/server, receiving the compressed 9.0-s spectrograms from the individual receivers, and additionally allow further processing for a near-real time deterministic estimate of the source location.

D. Source localisation and uncertainties with nonlinear Bayesian inversion

The nonlinear Bayesian inversion enables the estimation of unknown parameters and their uncertainty during the post-survey analysis.

1. Likelihood function and error assumption

The data error σ_k for the observed TDOA data d_k at each data point k is estimated during the spectrogram cross correlation process. It does not include systematic errors and is assumed to be Gaussian-distributed. The likelihood function therefore becomes

$$L(\mathbf{m}) = \frac{\exp\left[-0.5\sum_{k=1}^{N-1}\left(\frac{d_k - d_k(\mathbf{m})}{\sigma_k}\right)^2\right]}{(2\pi)^{\frac{N-1}{2}}\prod_{k=1}^{N-1}\sigma_k}, \quad (2)$$

where $d_k(\mathbf{m})$ is the predicted TDOA data for model parameter vector $\mathbf{m}$.

2. Nonlinear Bayesian inversion

In the Bayesian framework, the data $\mathbf{d}$ and the unknown model parameters $\mathbf{m}$ are random variables, each with specific probability densities (e.g., Gelman *et al.*, 2008). The general solution to the Bayesian inverse problem is the posterior probability density $P(\mathbf{m}|\mathbf{d})$, PPD, of a

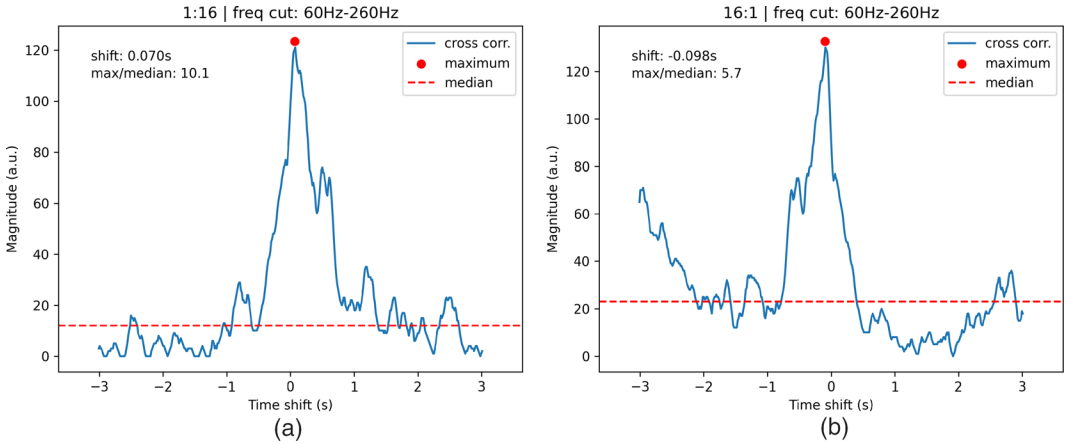

FIG. 4. Correlation function of the two upcalls shown in Fig. 3. The time difference of arrival is estimated from the maximum of the correlation function. Correlation functions are rejected when the ratio of the maximum and the median is below a set threshold. In this study, a threshold of 10.0 was used, implying that the correlation function to the left was accepted, whereas the one to the right was rejected.

model given the observed data, and given a known parameterisation of the model Bayes' rule defines

$$P(\mathbf{m}|\mathbf{d}) = \frac{P(\mathbf{d}|\mathbf{m})P(\mathbf{m})}{P(\mathbf{d})}, \tag{3}$$

where $P(\mathbf{d})$ is the probability of the data, $P(\mathbf{d}|\mathbf{m})$ is the likelihood function (introduced above), and $P(\mathbf{m})$ is the prior knowledge about the model parameters independent of the data. The prior probability densities are chosen to be uniformly distributed.

The PPD can be sampled numerically. To avoid sampling the whole parameter space, including areas with low PPDs (as is the case for a large fraction of the model space), a Markov-chain Monte Carlo (MCMC) method with the Metropolis-Hastings (MH) acceptance criterion is applied (Hastings, 1970; Metropolis *et al.*, 1953). The technique used in this paper additionally addresses correlation between model parameters and samples in the independent eigenparameter space instead (Dosso and Wilmut, 2008). Parameters are rotated using the eigenvalue decomposition of the model covariance matrix, which is estimated from all previous MCMC samples. The rotated parameters are perturbed using a Cauchy distribution and then rotated back into their original space. Additionally, we avoid being caught in local minima and to explore potential multimodal distributions by implementing parallel tempering (Dosso *et al.*, 2012), which relaxes the likelihood in the MH criterion in parallel MCMC chains, allowing the exploration of a larger model space. The parallel chains exchange models with the main chain, however, with the original likelihood function. Additionally, we use a chain-thinning approach that only accepts every fifth accepted sample to reduce the correlation between samples and represent the PPD more efficiently (Warner *et al.*, 2017). The sampling is completed after 1 00 000 accepted samples. The computation requires between a few minutes and several hours depending on problem complexity, when run on a single central processing core (2.6 GHz) with 32 Gbits of random access memory.

For the Bayesian inversion, we use a starting model estimated with a hybrid optimisation technique comprising a transition from a global to a local search (Dosso *et al.*, 2001). We allow source and relative receiver locations as well as sound speed and receiver clock drift as free parameters. The prior bounds for the receiver deviations are about −10 to 10 m from the last measured GPS buoy location on day 1 and up to −1000 to 1000 m on day 2. The clock drift, which is essentially a relative delay time between the buoys, is allowed to vary by −1 to 1 s. Even though the relative delay times between the buoys has ms precision, early inversion results indicate that models with larger delay times are preferred, potentially also accounting for systematic errors, such as not considering the three-dimensional variability of the ocean sound–speed profile, which can affect measured travel times of acoustic arrivals. The source location can be up to 80 km away from the centre of the grid.

3. Mahalanobis distance between visual and acoustic observations

The Mahalanobis distance D measures how far a point $\mathbf{x}$ lies from the mean μ of a multivariate distribution while taking into account the covariance structure (here, two-dimensional),

$$D(\mathbf{x}) = \sqrt{(\mathbf{x}-\mu)C^{-1}(\mathbf{x}-\mu)}. \tag{4}$$

It generalises the concept of the standard score (distance in standard deviations) to multiple correlated dimensions. Hence, a small number supports the hypothesis that the visual observation is part of the posterior probability distribution of the source location and therefore a match between visual and acoustic observations (e.g., Johnson and Wichern, 2023). We choose a threshold of 3 to consider a visual observation to match with the acoustic one.

https://doi.org/10.1121/10.0042758

III. RESULTS

A. DL model for upcall detection

The original DL model for NARW upcalls (Padovese *et al.*, 2023) achieves a recall above 50% and precision of about 10% for a score threshold of 0.5. It achieves a number of false positives per hour below 10 (Fig. 5, top). After the model adaptation and rerun of the detection, the recall improves slightly to 58%, and the precision increases to almost 80%. The number of false positives per hour reduces to about 0.3 (Fig. 5, bottom). Although the number of false positives per hour is significantly reduced, just under 50% of the annotated calls (which were annotated by an expert human annotator) are missed by the adapted model.

B. Source localisation and uncertainties with nonlinear Bayesian inversion

1. Call localisations on day 1

During the first day of the experiment, the sonobuoys were of type AN/SSQ-53F and provided frequent updates of their position. The prior bounds for the deviation from the last position were therefore set to 10 m. The sonobuoys drifted on average about 0.2 m/s with the ocean currents and on average 2 km from their original position. The source position, as expected, is less well-constrained when the call originates from outside the sonobuoy grid (Fig. 6) than when it is received by receivers from multiple directions toward the source of the call (Fig. 7).

The average sound speed profile is not well-constrained, but marginal probability densities overlap with the observed sound speed profiles (Fig. 8).

The marginal probability density for the time delays between the buoys overlap within the error bars.

NARW low-frequency vocalisations originated in the northwest, southwest, and east of the grid (Fig. 9). The source location to the northwest is not as well-known because the calls originated at least 10 km away from the buoy grid (Fig. 6). A train of low-frequency moans is located close to the grid to the east. The moans occurred during a 3.5 h time frame with well-constrained source locations that seem to move away from the grid. The train covers about 9 km between the first and the last moan.

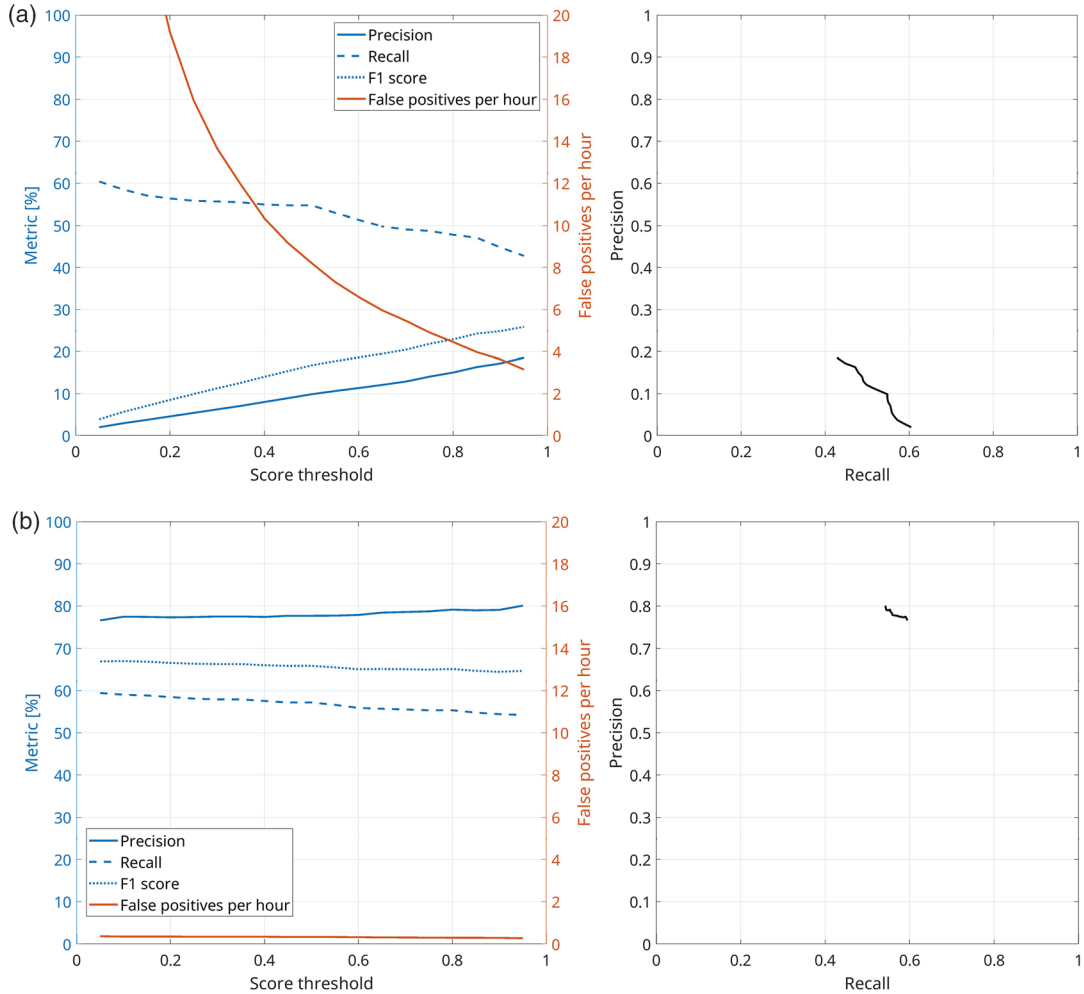

FIG. 5. DL results metrics precision, recall, F1 score, and number of false positives per hour over the score threshold for the original model (top) and the adapted model (bottom).

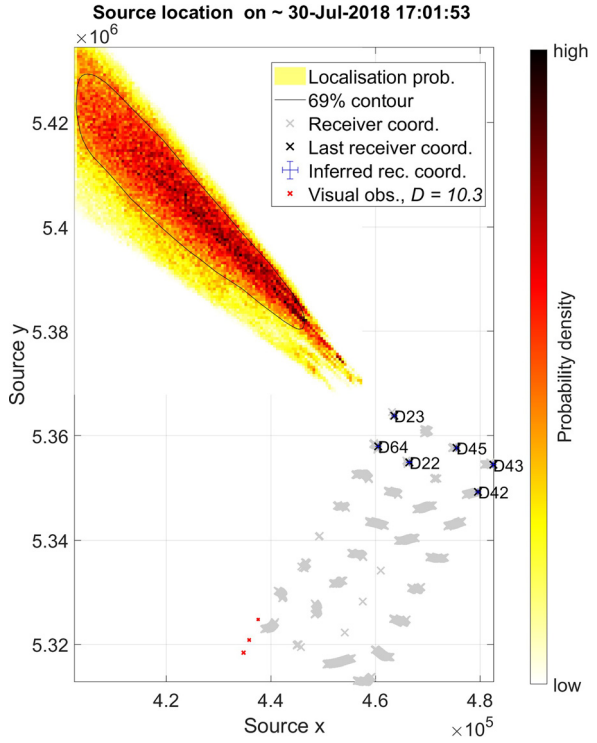

FIG. 6. Marginal probability density distribution of the source location of an upcall initiated around 17:02 UTC on 30 July 2018. The 69% credibility interval is extracted to compare to other calls in Fig. 9. The receiver coordinates over time (grey x) are overlain by the last receiver location before the call (black x) and the inferred receiver locations and their uncertainties (blue cross). Visual observations within an 8-h time frame are located in the southwest of the grid and with a minimum Mahalanobis distance $D =$ 10.3 are likely not related to the acoustic detection.

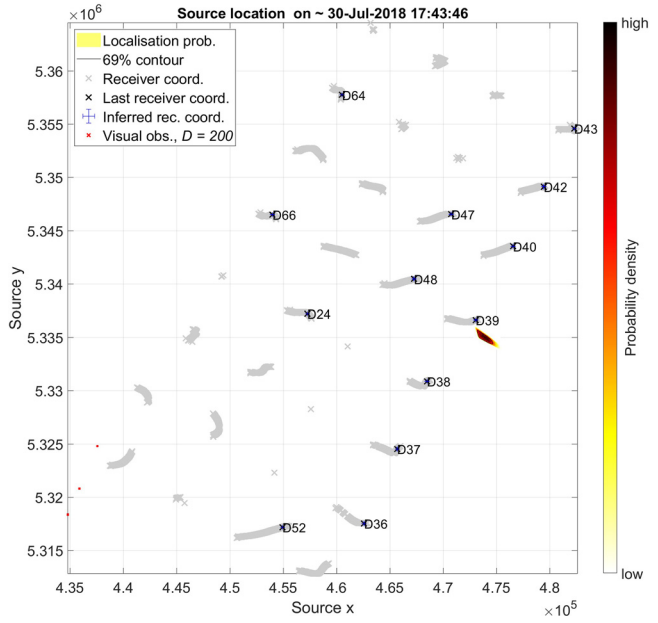

FIG. 7. Marginal probability density distributions of the source location of a moan initiated around 17:44 UTC on 30 July 2018. The receiver coordinates over time (grey x) are overlain by the last receiver location before the call (black x) and the inferred receiver locations and their uncertainties (blue cross). Visual observations within an 8-h time frame are located in the southwest of the grid and with a minimum Mahalanobis distance $D =$ 200 are likely not related to the acoustic detection.

2. Call localisations on day 2

Call locations are better constrained when originated within the grid. For example, an upcall around 17:43 UTC originated close to buoy D56 and is located in the general area of visual observations of right whales on day 2 (Fig. 10). The upcall was recorded on nineteen buoys resulting in 59 free parameters. The receiver locations are less well-constrained than for day 1 due to the unknown drift of the sonobuoys after dropping into the ocean (marginal probability densities for receiver locations are represented by error bars in Fig. 10).

Most of the calls on day 2 originated from the southwest outside the grid and are not well-constrained (Fig. 11).

Several visual NARW observations were made in the upper middle of the sonobuoy grid and were within a 30 km radius north of calls localised from PAM data.

3. Mahalanobis distance between visual and acoustic observations

On day 1, only a few observations were made from a vessel between 15:50 UTC to 22:50 UTC to the southwest. No aerial surveys were possible and visibility was restricted due to weather. Only one acoustic detection that is located in the southwest (light purple line at about 47.9° N, 64° W on Fig. 9) could be related to a visual observation with a Mahalanobis distance $D = 2.7$. The detections are about 1.5 h apart. The acoustic detections in the northwest and east (Fig. 9) cannot be related to visual observations.

On day 2, we localise 13 calls in the southwest corner of the sonobuoy grid between 15:15 UTC and 18:00 UTC. The one to five visual observations with Mahalanobis distances $D < 3$ were observed between 15:20 UTC and 16:30 UTC (Fig. 12). Three acoustic observations in the centre of the grid between 17:43 UTC and 17:46 UTC have Mahalanobis distances $D < 2$ for visual observations between 16:40 UTC and 17:00 UTC (Fig. 12). Thirty visual observations in the centre and northeast of the grid between 11:55 UTC and 18:15 UTC have Mahalanobis distances $D < 3$ (15 have $D < 2$) for one acoustic call location in the northeast (15:24 UTC). The nearest visual observation was observed at 13:48 UTC (yellow cross-at about 48.1° N and 63.57° W) with $D = 1.3$. At least 11 visual observations in the centre do not match with marginal probability densities from acoustic observations using our methods.

IV. DISCUSSION

A. DL model

The pre-trained NARW upcall DL model is straightforward to use and already offers a >50% recall compared to the human-annotated calls. The number of false positives per hour is below 10, which is relatively low compared to the threshold of 20 after which detector results can be deemed useless (concluded by Shiu *et al.*, 2020). Using the upcalls from the model requires manually eliminating the

https://doi.org/10.1121/10.0042758

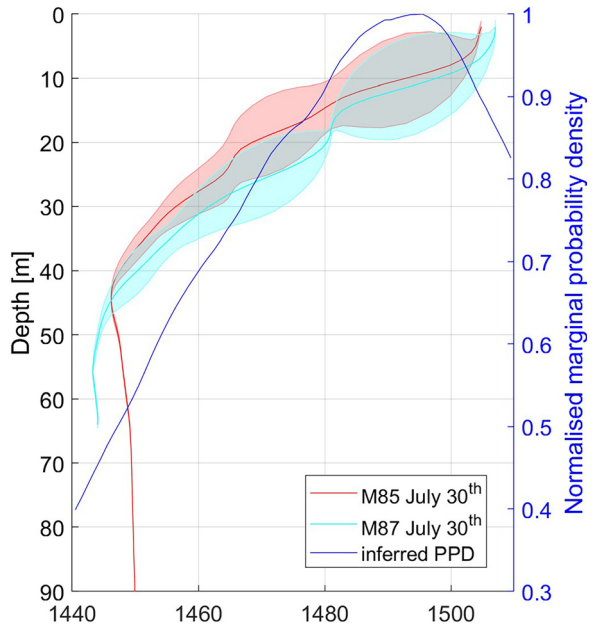

FIG. 8. Observed sound speed in m/s calculated from CTD measurements from an underwater glider over depth (left axis). The shaded areas are standard deviations over several glider dives for one day. The normalised marginal probability density for the inferred (for the call location shown in Fig. 6) sound speed is overlain (right axis).

false positives. However, after adapting the model with 1992 hard negatives and 194 upcalls, we are able to reduce the number of false positives per hour to about 0.3. The annotation is more time-efficient compared to manually annotating the whole underwater acoustic dataset, as the model effectively filters out a large portion of irrelevant data, significantly reducing the workload.

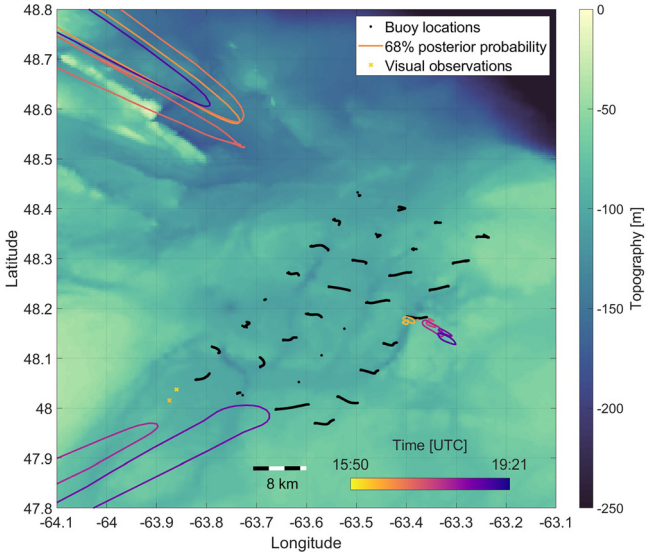

FIG. 9. Localisation of NARW vocalisations on 30 July 2018 between 12:50 and 16:21 local time. Localisations encompass the 68% probability margin for low-frequency sounds, including upcalls and moans (solid line). Whale localisations are indicated in the northwest, southwest, and close to the sonobuoy grid in the east. Visual observations (crosses) were made in the southwest. Regularly updated buoy locations are shown as black dots. [Bathymetry source: GEBCO Compilation Group (2024) GEBCO 2024 Grid (doi:10.5285/1c44ce99-0a0d-5f4f-e063-7086abc0ea0f).]

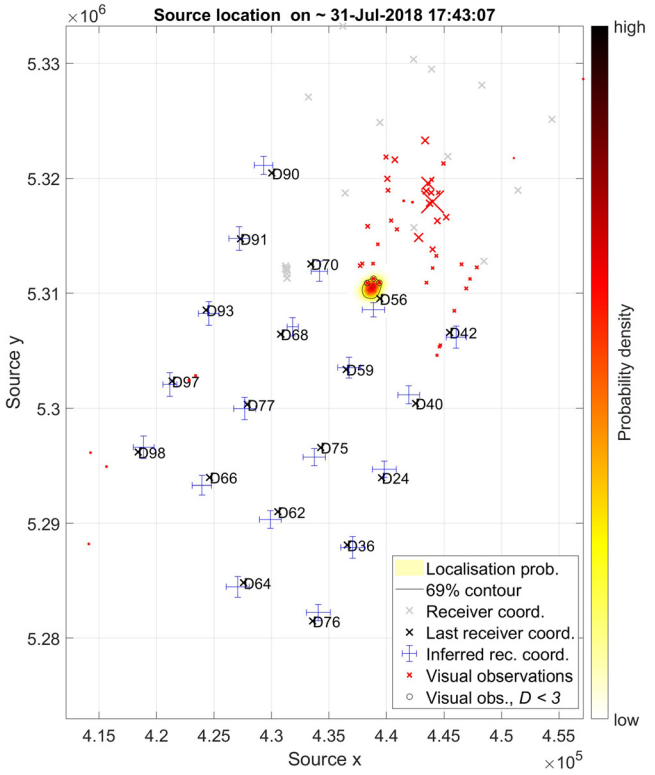

FIG. 10. Marginal probability density for upcall source location that was initiated around 17:43 UTC on 31 July 2018. The 69% credibility interval is extracted to compare to other calls in Fig. 11. Visual observations (red crosses) were made in the northeast within the sonobuoy grid. The receiver coordinates over time (grey x) are overlain by the last receiver location (black x) at the initial drop time and the inferred receiver locations and their uncertainties (blue crosses). Visual observations that were made in a 4-h window are shown as red crosses. The larger the crosses, the closer the time was to when the call was recorded. When the Mahalanobis distance D is smaller than 3, observations are marked with black circles.

During this step, we find additional positives that the DL model detected compared to the manually annotated upcalls (about 3% of the new total number of upcalls). Annotating is a subjective process, and it can happen that calls are overlooked even by experienced annotators. Additionally, the purpose of the annotations for this dataset was localisation and not detector performance (von Benda-Beckmann *et al.*, 2022).

The adaptation of the model was performed using an *F1* score as loss function. The relatively small variation in precision and recall across score thresholds (Fig. 5) is a known artifact of models trained with an *F1* score (or other non-decomposable) loss functions. Such losses encourage a stable operating point that maximises the harmonic mean of precision and recall rather than producing well-calibrated probability outputs. As a result, the model tends to output scores clustered near the extremes (close to 0 or 1), yielding limited sensitivity to threshold changes. Although unconventional, this pattern reflects that the optimisation target corresponds to consistent performance at the default score threshold of 0.5.

Although just under 50% of the annotated calls are missed by the adapted model, it is able to significantly

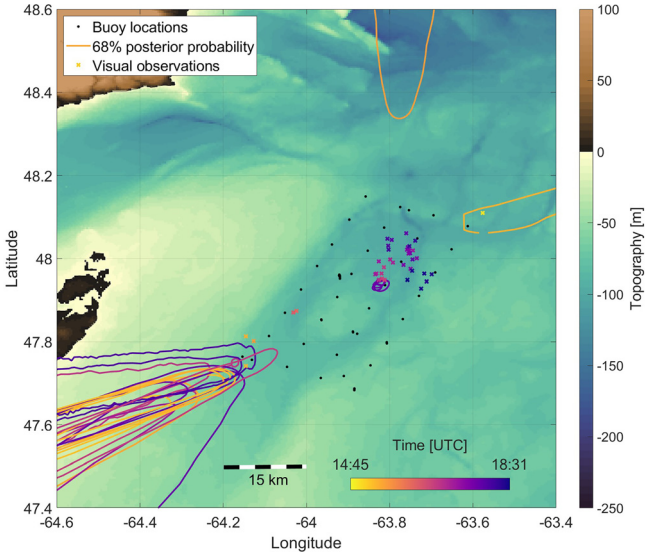

FIG. 11. Localisation of NARW vocalisations on 31 July 2018 between 11:45 and 15:21 local time. Localisations encompass the 68% probability margin for low-frequency sounds, including upcalls and moans (solid line). Whale localisations are indicated in the north, northeast, southwest, and inside the sonobuoy grid in the east. Visual observations (crosses) were made in the southwest and inside the grid. Drop buoy locations are shown as black dots. [Bathymetry source: GEBCO Compilation Group (2024) GEBCO 2024 Grid (doi:10.5285/1c44ce99-0a0d-5f4f-e063-7086abc0ea0f).]

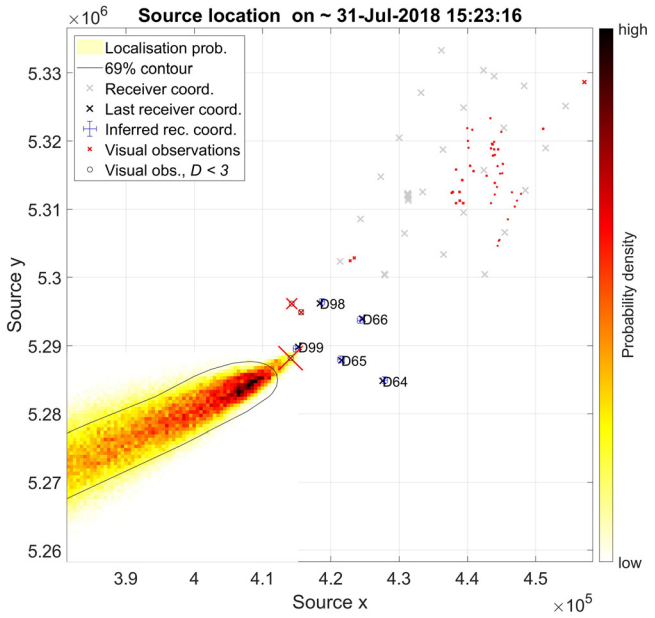

FIG. 12. Marginal probability density for upcall source location that was initiated around 15:23 UTC on 31 July 2018. The 69% credibility interval is extracted to compare to other calls in Fig. 11. The receiver coordinates over time (grey x) are overlain by the last receiver location (black x) at the initial drop time and the inferred receiver locations and their uncertainties (blue crosses). Visual observations that were made 4 h before or after the acoustic observation are shown as red crosses. The larger the crosses, the closer the time was to when the call was recorded. When the Mahalanobis distance D is smaller than 3, observations are marked with black circles. The closest visual observation has a Mahalanobis distance of 2.

reduce the false positives. Reasons for the model to not perform as well on the sonobuoy dataset can be that the sonobuoy dataset has a rather distinct kind of noise signature and other features compared to the original data the model was trained on. The background noise profile of the sonobuoys differs to the training datasets' noise profiles due to the intermittent RF communication, as well as the depth and nature of the system's deployment. One advantage of the sonobuoys that compensates for missed detections is the large number and relatively small distance between sonobuoys compared to most moored arrays, which offers a level of repetition/redundancy. If detections are recorded on at least three sonobuoys, a localisation based on the TDOA method can be performed.

Human activity and other biological sounds as well as other environmental factors, i.e., breaking waves, wind, and rain can reduce the SNR of the call. Variations in sound propagation further influence how a vocalisation is perceived, as it may be altered by reverberations, diffraction, attenuation, and reflections. Additionally, although transfer learning does not require a large dataset, the 194 upcalls used here for adapting the model might just be too small. Ways to improve the recall that were not the focus of this study could be to increase the number of upcalls by augmentation (e.g., time shifts, Padovese *et al.*, 2021).

When considering a new dataset that has not been annotated yet, running the NARW upcall model as described above is a reasonable first step to detect NARW upcalls for further localisation analysis. The number of detections can intrinsically improve when applying the spectrogram cross correlation, which is an independent way to detect the same calls on different receivers.

The adapted model can now be used for underwater acoustic data recorded with sonobuoys. It, however, does only consider data in one specific shallow water environment and two specific kinds of sonobuoys, and might require additional adaptation for different underwater environments and equipment. The DL framework is available as open software and comes with a tutorial for running and adapting it (Padovese *et al.*, 2023).

B. TDOA-based source localisation

1. Practical Implementation

To implement our solution in a practical setting, the following requirements must be fulfilled: The hydrophones must be well-synchronised (millisecond precision) and have known positions, which can be achieved with GPS. Call detection and spectrogram computation must take place locally at the hydrophone in real time. This can be achieved by equipping the hydrophone with a small computer. For example, a Raspberry Pi would be able to compute spectrograms and run our DL model in real time using only a few watts on average. Last, the hydrophones must be able to transmit the compressed and timestamped spectrograms to the "outside world" with little or no latency. This could, e.g., be achieved with hydrophones tethered to surface buoys or uncrewed surface vehicles (USVs) communicating via satellite link (e.g., Iridium). An implicit assumption is that the buoy/USV is capable of powering these devices for

the duration of the deployment (e.g., by a combination of batteries and solar power).

The TDOA estimation and source localisation would take place on a central server or in the cloud. TDOAs are determined with a fast and fully automated cross correlation algorithm, utilising the compressed (binary) spectrograms to reduce the amount of data that needs to be transferred from the individual hydrophones.

Our current implementation of the Bayesian inversion algorithm has not been optimised for speed and, as a result, is comparably slow to the other two methods. However, we foresee that the computational time can be drastically reduced through improved algorithm engineering and high-performance computing infrastructure, allowing the full processing pipeline to run in near real time with a latency of minutes, rather than hours, as required for dynamic management purposes.

2. Spectrogram cross correlation

NARW upcalls and moans are finite signals and vary in intensity and duration. Questions arise such as: Should the time of maximum sound intensity be used or the time at which the intensity first exceeds some arbitrary signal-to-noise threshold relative to the ambient background level, or rather take advantage of the spectral shape and duration using correlation algorithms? Spectrogram cross correlation methods to estimate TDOAs are a robust choice given the variability of the calls (e.g., Simard *et al.*, 2008; Tiemann *et al.*, 2004). One source of error that contributes to the TDOA uncertainty is the time resolution of the digitised signal or of a spectrogram used in cross correlation. Other sources of error are the uncertainties in the exact position of hydrophones or in the speed of sound at the field site, and the use of a two-dimensional array in a three-dimensional environment (Janik *et al.*, 2000).

In our case, the measurement errors for the TDOAs are estimated from the spectrogram cross correlation technique for each buoy and assumed to be Gaussian distributed. Instrumentation limitations comprise time delays between the recording at the sonobuoy and the transmission to the aircraft, and gaps of buoy location GPS updates (especially on day 2 of the experiment). The physical theory is approximated with a direct acoustic wave propagation ignoring propagation loss, modal dispersion, or seabed parameters, which become more important in shallower water and lower frequencies. We are also simplifying the sound speed profile by assuming one average sound speed for the whole water column. The time window chosen for the cross correlation, however, seems to account for errors in the sound propagation assumption.

Helble *et al.* (2015) point out that setting a threshold for the TDOA estimation implies a trade-off between the accuracy of the TDOA readings (and therefore the resulting localisation) and the number of accepted TDOAs (i.e., call instances). They test their threshold for the signal strength using simulated call sequences and varying levels of ambient noise. A detailed evaluation of the chosen thresholds is beyond the scope of this article. In our case, it is preferable to use as many call instances as possible to the relatively small number of NARW vocalisations. However, that may result in incorrect TDOA estimations, which can be caused by noisy signals, non-direct paths, and overlapping calls. We therefore exclude outliers, and additionally, the Bayesian approach proved to sample inefficiently for incorrect TDOAs.

C. Discussion on localisation results using nonlinear Bayesian inversion

Nonlinear Bayesian inversion enables estimating the source location and unknown experiment parameters. Especially on day 2, when the receiver locations were known with about 1 km uncertainty, the Bayesian approach incorporated the uncertainty, whereas deterministic approaches, such as time offset and receiver locations, would not be able to.

Marginal probability densities for time delay between buoys usually are within the −1 and 1 s range. The sonobuoys, however, are expected to have a much smaller delay between buoys of the same type. The delay might therefore potentially account for other factors, such as arrival time that depends on environmental factors, bathymetry, etc, that varies for different source-receiver paths. We also only invert for an average sound speed and ignore the sound speed profile. Results, however, show that the average sound speed is within the limits of the measured sound speed profile.

On the second day, receiver locations may be as different to their drop positions (on average ∼2 km over up to 6 h). The inversion results present little sensitivity to the position, and the marginal probability densities often cover the area between the prior bounds (−1 to 1 km). The inversion results for the receiver location could be improved in the future using different calls at the same time that were recorded on the same buoys to improve the data to unknown parameters ratio. Unfortunately, however, there are not many calls that fit these requirements due to the facultative calling behavior of the whales.

The presented work uses an average sound speed profile and a simple direct path assumption for the location estimation. Using a more sophisticated propagation model for the water depth and frequency range and including topography and source depth might improve the results (e.g., Spiesberger and Fristrup, 1990).

We analysed not only the detected but also the remaining annotated upcalls and moans. Therefore, it is possible to compare our localisation results to other approaches from the DCLDE 2024 workshop in the future.

D. Discussion on call occurrences used for localisation

Localisations are possible for 14 upcall and moan events on day 1 and 21 events on day 2. These events are

associated with 122 of 282 annotated calls on day 1 and 129 of 557 annotated calls on day 2. Nine and 21 calls, respectively, that were found during the cross correlation could not be assigned to an annotation. Reasons for not being able to utilise all the calls that were annotated could be that they were not assigned to an event with more than two calls during the cross correlation process due to their SNR, the chosen thresholds for spectrogram reduction, and the threshold for accepting cross correlation peaks as TDOA. The explanation could also be that the calls origins were significantly outside the sonobuoy grid and only recorded on a couple of buoys. The sonobuoys have presented unique advantages and disadvantages. Calls inside the grid and near the grid were easy to localise within a small radius due to multiple sonobuoys recording the same call. The TDOA algorithm is most sensitive to timing errors. Therefore, the time synchronisation between the buoys across several km proves to be an important asset. Sonobuoy deployments are also quite flexible. During the survey, the operators were able to adjust the deployment location for the second day to move closer to where the whales were being observed. Main challenges are that the second day recording did not have updated locations for the buoys and that the buoys are short-lived and cannot be used to analyse call behaviour for larger than a 6-h time series.

E. Discussion on NARW calling behaviour

On day 1, acoustic and visual observations can potentially be matched for one visual observation with a Mahalanobis distance of 2, whereas localisations from acoustic detections in the northwest and east do not match with visual observations. That is likely caused by weather restricting visual surveys on that day to the southwest of the grid. Therefore, we cannot make any conclusive correlations between visual observations and inferred call locations from PAM data in this case.

On day 2, 13 calls that are localised in the southwest of the grid can be matched with one to five visual observations. The acoustic and visual observations overlap in time, but acoustic observations still occur 1.5 h after the last visual observation. Therefore, PAM data successfully complement the visual observations by extending the time frame of NARW observations.

Three calls that are localised in the centre of the grid match with visual observations that were made one hour earlier. However, over 40 visual observations were made to the northeast of the acoustic observation with Mahalanobis distances larger than 3 that are likely not related. About 30 of these, however, have Mahalanobis distances smaller than 3 for a call that originated in the northeast. The Mahalanobis distances are likely small due to the large uncertainty of the call location. However, at least 11 visual observations do not match with acoustic observations, suggesting that the NARW did not necessarily produce upcalls or moans when observed, but rather sounded facultatively and possibly when submerged (as also found by Matthews *et al.*, 2001). In this case, the visual observations successfully complement the acoustic observations by detecting NARW when they cannot be detected acoustically.

V. CONCLUSIONS

The presented integrated processing solution for marine mammal calls on passive acoustic data effectively detects and locates NARW upcalls in the GSL.

A publicly available DL model is successfully adapted to the sonobuoy dataset, which required a much smaller dataset size than the original training dataset and therefore minimal manual annotation efforts.

The spectrogram cross correlation technique is a fully automated approach designed to be suitable for deployment in power and bandwidth constrained environments, and allows for source localisation and overall confidence estimates.

The nonlinear Bayesian inversion allows for estimates not only of the source location but also of relative receiver locations, instrument clock drift, and average sound speed. The simple forward model allows for the inversion to run relatively fast (minutes up to a few hours) dependent on the number of parameters.

Implementing these techniques, we are able to constrain the position of the NARW when calling (Figs. 9 and 11) and highlight the importance of passive acoustic monitoring in addition to visual observations due to its larger range and ability to monitor 24/7 for vocalising whales, including when submerged, in all weather conditions.

ACKNOWLEDGMENTS

This dataset acquisition was supported by the research consortium consisting of Dalhousie University, Fisheries and Oceans Canada, Defence Research and Development Canada, the Acoustic Data Analysis Centre, the National Oceanic and Atmospheric Administration, Northeast Fisheries Science Center, the New England Aquarium, Anderson Cabot Center for Ocean Life, and the Canadian Whale Institute. The acoustic data were collected by the Royal Canadian Airforce. We thank JASCO Applied Sciences for sharing the TDOA localisation algorithm for the Bayesian inversion. We would like to thank Hansen Johnson for the manual annotations. We thank Alexander von Benda-Beckmann and Carolyn Binder for fruitful discussions. Last, but not least, we thank two anonymous reviewers for their constructive comments.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available within the article von Benda-Beckmann *et al.* (2022). The data are also available from the corresponding author upon reasonable request.

[1]GitLab repository: https://git-dev.cs.dal.ca/meridian/NARW_detection_tool

Baumgartner, M. F., Fratantoni, D. M., Hurst, T. P., Brown, M. W., Cole, T. V. N., Van Parijs, S. M., and Johnson, M. (**2013**). "Real-time reporting of baleen whale passive acoustic detections from ocean gliders," J. Acoust. Soc. Am. **134**(3), 1814–1823.

Baumgartner, M. F., and Mussoline, S. E. (**2011**). "A generalized baleen whale call detection and classification system," J. Acoust. Soc. Am. **129**(5), 2889–2902.

Binder, C. M., and Hines, P. (**2012**). "Applying automatic aural classification to cetacean vocalizations," Proc. Mtgs. Acoust. **17**(1), 070029.

Chan, Y.-T., Yau Chin Hang, H., and Ching, P.-c. (**2006**). "Exact and approximate maximum likelihood localization algorithms," IEEE Trans. Veh. Technol. **55**(1), 10–16.

Clark, C. W., and Ellison, W. T. (**2000**). "Calibration and comparison of the acoustic location methods used during the spring migration of the bowhead whale, *Balaena mysticetus*, off Pt. Barrow, Alaska, 1984–1993," J. Acoust. Soc. Am. **107**(6), 3509–3517.

Crowe, L. M., Brown, M. W., Corkeron, P. J., Hamilton, P. K., Ramp, C., Ratelle, S., Vanderlaan, A. S., and Cole, T. V. (**2021**). "In plane sight: A mark-recapture analysis of North Atlantic right whales in the Gulf of St. Lawrence," Endang. Species Res. **46**, 227–251.

Davies, K. T., and Brillant, S. W. (**2019**). "Mass human-caused mortality spurs federal action to protect endangered North Atlantic right whales in Canada," Mar. Policy **104**, 157–162.

Davis, G. E., Baumgartner, M. F., Bonnell, J. M., Bell, J., Berchok, C., Bort Thornton, J., Brault, S., Buchanan, G., Charif, R. A., Cholewiak, D., Clark, C. W., Corkeron, P., Delarue, J., Dudzinski, K., Hatch, L., Hildebrand, J., Hodge, L., Klinck, H., Kraus, S., Martin, B., Mellinger, D. K., Moors-Murphy, H., Nieukirk, S., Nowacek, D. P., Parks, S., Read, A. J., Rice, A. N., Risch, D., Širović, A., Soldevilla, M., Stafford, K., Stanistreet, J. E., Summers, E., Todd, S., Warde, A., and Van Parijs, S. M. (**2017**). "Long-term passive acoustic recordings track the changing distribution of North Atlantic right whales (*Eubalaena glacialis*) from 2004 to 2014," Sci. Rep. **7**(1), 13460.

Dosso, S. E., Holland, C. W., and Sambridge, M. (**2012**). "Parallel tempering for strongly nonlinear geoacoustic inversion," J. Acoust. Soc. Am. **132**(5), 3030–3040.

Dosso, S. E., and Wilmut, M. J. (**2008**). "Uncertainty estimation in simultaneous bayesian tracking and environmental inversion," J. Acoust. Soc. Am. **124**(1), 82–97.

Dosso, S. E., Wilmut, M. J., and Lapinski, A.-L. (**2001**). "An adaptive-hybrid algorithm for geoacoustic inversion," IEEE J. Oceanic Eng. **26**(3), 324–336.

Fucile, P. D., Singer, R. C., Baumgartner, M., and Ball, K. (**2006**). "A self contained recorder for acoustic observations from AUV's," in *Oceans 2006*, 18–21 September 2006, Boston, pp. 1–4.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (**2008**). *Bayesian Data Analysis*, 2nd ed. (Chapman and Hall, London).

Goodfellow, I., Bengio, Y., and Courville, A. (**2016**). *Deep Learning* (MIT Press), http://www.deeplearningbook.org (Last viewed 24 February 2026).

Hastings, W. K. (**1970**). "Monte Carlo sampling methods using Markov chains and their applications," Biometrika **57**(1), 97–109.

He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 27–30 June 2016, Las Vegas, NV, pp. 770–778.

Helble, T. A., Ierley, G. R., D'Spain, G. L., and Martin, S. W. (**2015**). "Automated acoustic localization and call association for vocalizing humpback whales on the Navy's Pacific Missile Range Facility," J. Acoust. Soc. Am. **137**(1), 11–21.

Hendricks, B., Wray, J. L., Keen, E. M., Alidina, H. M., Gulliver, T. A., and Picard, C. R. (**2019**). "Automated localization of whales in coastal fjords," J. Acoust. Soc. Am. **146**(6), 4672–4686.

Indeck, K. L., Baumgartner, M. F., Lecavalier, L., Whoriskey, F., and Davies, K. T. A. (**2025**). "Passive acoustic gliders are effective monitoring tools for dynamic management plans aimed at mitigating whale-vessel strikes," Conserv. Lett. **18**, e13102.

Janik, V. M., Van Parijs, S. M., and Thompson, P. M. (**2000**). "A two-dimensional acoustic localization system for marine mammals," Mar. Mammal Sci. **16**(2), 437–447.

Johnson, R., and Wichern, D. (**2023**). *Applied Multivariate Statistical Analysis*, 6th ed. (Pearson, London).

Kirsebom, O. S., Frazao, F., Padovese, B., Sakib, S., and Matwin, S. (**2021**). "Ketos—A deep learning package for creating acoustic detectors and classifiers," J. Acoust. Soc. Am. **150**(4), A164.

Kirsebom, O. S., Frazao, F., Simard, Y., Roy, N., Matwin, S., and Giard, S. (**2020**). "Performance of a deep neural network at detecting North Atlantic right whale upcalls," J. Acoust. Soc. Am. **147**(4), 2636–2646.

Knapp, C., and Carter, G. (**1976**). "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech. Signal Process. **24**(4), 320–327.

Li, P., Liu, X., Palmer, K. J., Fleishman, E., Gillespie, D., Nosal, E.-M., Shiu, Y., Klinck, H., Cholewiak, D., Helble, T., and Roch, M. A. (**2020**). "Learning deep models from synthetic data for extracting dolphin whistle contours," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 19–24 July 2020, Glasgow, UK, pp. 1–10.

Li, X., Deng, Z. D., Rauchenstein, L. T., and Carlson, T. J. (**2016**). "Contributed review: Source-localization algorithms and applications using time of arrival and time difference of arrival measurements," Rev. Sci. Instrum. **87**(4), 041502.

Matthews, J., Brown, S., Gillespie, D., Johnson, M., McLanaghan, R., Moscrop, A., Nowacek, D., Leaper, R., Lewis, T., and Tyack, P. (**2001**). "Vocalisation rates of the North Atlantic right whale (*Eubalaena glacialis*)," J. Cetacean Res. Manage. **3**(3), 271–282.

Matthews, L. P., and Parks, S. E. (**2021**). "An overview of North Atlantic right whale acoustic behavior, hearing capabilities, and responses to sound," Mar. Pollut. Bull. **173**, 113043.

Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," J. Acoust. Soc. Am. **107**(6), 3518–3529.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (**1953**). "Equation of state calculations by fast computing machines," J. Chem. Phys. **21**(6), 1087–1092.

Meyer-Gutbrod, E. L., Davies, K. T. A., Johnson, C. L., Plourde, S., Sorochan, K. A., Kenney, R. D., Ramp, C., Gosselin, J.-F., Lawson, J. W., and Greene, C. H. (**2023**). "Redefining North Atlantic right whale habitat-use patterns under climate change," Limnol. Oceanogr. **68**(S1), S71–S86.

Meyer-Gutbrod, E. L., Greene, C. H., Davies, K. T., and Johns, D. G. (**2021**). "Ocean regime shift is driving collapse of the North Atlantic right whale population," Oceanography **34**(3), 22–31.

Minister of Transport, Chrystia Freeland (**2025**). "Interim Order No. 3 for the Protection of North Atlantic Right Whales (*Eubalaena glacialis*) in the Gulf of St. Lawrence, 2025," Technical Report, https://tc.canada.ca/en/ministerial-orders-interim-orders-directives-directions-response-letters/interim-order-no-3-protection-north-atlantic-right-whales-eubalaena-glacialis-gulf-st-lawrence-2025 (Last viewed 24 February 2026).

Padovese, B., Frazao, F., Kirsebom, O. S., and Matwin, S. (**2021**). "Data augmentation for the classification of North Atlantic right whales upcalls," J. Acoust. Soc. Am. **149**(4), 2520–2530.

Padovese, B., Kirsebom, O. S., Frazao, F., Evers, C. H., Beslin, W. A., Theriault, J., and Matwin, S. (**2023**). "Adapting deep learning models to new acoustic environments—A case study on the North Atlantic right whale upcall," Ecol. Inf. **77**, 102169.

Parks, S., Searby, A., Célérier, A., Johnson, M., Nowacek, D., and Tyack, P. (**2011**). "Sound production behavior of individual North Atlantic right whales: Implications for passive acoustic monitoring," Endang. Species Res. **15**(1), 63–76.

Parks, S. E., and Tyack, P. L. (**2005**). "Sound production by North Atlantic right whales (*Eubalaena glacialis*) in surface active groups," J. Acoust. Soc. Am. **117**(5), 3297–3306.

Plourde, S., Lehoux, C., Roberts, J. J., Johnson, C. L., Record, N., Pepin, P., Orphanides, C., Schick, R. S., Walsh, H. J., and Ross, C. H. (**2024**). "Describing the seasonal and spatial distribution of *Calanus* prey and North Atlantic right whale potential foraging habitats in Canadian waters using species distribution models," Research Document No. 2024/039, Fisheries and Oceans Canada, Canadian Science Advisory Secretariat.

Ramaswamy, B., Potty, G., and Miller, J. (**2001**). "A marine mammal acoustic detection and localization algorithm using spectrogram image correlation," in *MTS/IEEE Oceans 2001. An Ocean Odyssey. Conference Proceedings (IEEE Cat. No.01CH37295)*, 5–8 November 2001, Honolulu, HI, Vol. 4, pp. 2354–2358.

Ross, C. H., Pendleton, D. E., Tupper, B., Brickman, D., Zani, M. A., Mayo, C. A., and Record, N. R. (**2021**). "Projecting regions of North Atlantic right whale, *Eubalaena glacialis*, habitat suitability in the Gulf of Maine for the year 2050," Elementa: Sci. Anthropocene **9**(1), 00058.

Saito, T., and Rehmsmeier, M. (**2015**). "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS One **10**(3), e0118432.

Shiu, Y., Palmer, K., Roch, M. A., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gillespie, D., and Klinck, H. (**2020**). "Deep neural networks for automated detection of marine mammal species," Sci. Rep. **10**(1), 607.

Simard, Y., and Roy, N. (**2008**). "Detection and localization of blue and fin whales from large-aperture autonomous hydrophone arrays: A case study from the St. Lawrence estuary," Can. Acoust. **36**(1), 104–110.

Simard, Y., Roy, N., and Gervaise, C. (**2008**). "Passive acoustic detection and localization of whales: Effects of shipping noise in Saguenay-St. Lawrence Marine Park," J. Acoust. Soc. Am. **123**(6), 4109–4117.

Simard, Y., Roy, N., Giard, S., and Aulanier, F. (**2019**). "North Atlantic right whale shift to the Gulf of St. Lawrence in 2015, revealed by long-term passive acoustics," Endang. Species Res. **40**, 271–284.

Spiesberger, J. L. (**2005**). "Probability distributions for locations of calling animals, receivers, sound speeds, winds, and data from travel time differences," J. Acoust. Soc. Am. **118**(3), 1790–1800

Spiesberger, J. L., and Fristrup, K. M. (**1990**). "Passive localization of calling animals and sensing of their acoustic environment using acoustic tomography," Am. Naturalist **135**(1), 107–153.

Stowell, D. (**2022**). "Computational bioacoustics with deep learning: A review and roadmap," PeerJ **10**, e13152.

Tarantola, A. (**2005**). *Inverse Problem Theory and Methods for Model Parameter Estimation* (Society for Industrial and Applied Mathematics, Philadelphia).

Thode, A. (**2005**). "Three-dimensional passive acoustic tracking of sperm whales (*Physeter macrocephalus*) in ray-refracting environments," J. Acoust. Soc. Am. **118**(6), 3575–3584.

Tiemann, C. O., Porter, M. B., and Frazer, L. N. (**2004**). "Localization of marine mammals near Hawaii using an acoustic propagation model," J. Acoust. Soc. Am. **115**(6), 2834–2843.

Transport Canada (**2024**). "Protecting the North Atlantic right whale: Speed restriction measures in the Gulf of St. Lawrence."

von Benda-Beckmann, A., Binder, C., Johnson, H., MacDonnell, J., Theriault, J., Vanderlaan, A., and Thomson, D. (**2022**). "Northern right whale sonobuoy localisation dataset for the DCLDE workshop," (TNO—Applied Scientific Research, The Hague, The Netherlands).

Warner, G. A., Dosso, S. E., and Hannay, D. E. (**2017**). "Bowhead whale localization using time-difference-of-arrival data from asynchronous recorders," J. Acoust. Soc. Am. **141**(3), 1921–1935.

Warner, G. A., Dosso, S. E., Hannay, D. E., and Dettmer, J. (**2016**). "Bowhead whale localization using asynchronous hydrophones in the Chukchi Sea," J. Acoust. Soc. Am. **140**(1), 20–34.

Watkins, W. A., and Schevill, W. E. (**1972**). "Sound source location by arrival-times on a non-rigid three-dimensional hydrophone array," Deep Sea Res. Oceanographic Abstracts **19**(10), 691–706.

Zimmer, W. M. (**2011**). *Passive Acoustic Monitoring of Cetaceans* (Cambridge University Press, Cambridge).