JASA ARTICLE

# Automating multi-target tracking of singing humpback whales recorded with vector sensors

Pina Gruden,[1,a] Junsu Jang,[2] Anke Kügler,[3,4] Thomas Kropfreiter,[2] Ludovic Tenorio-Hallé,[2] Marc O. Lammers,[5] Aaron Thode,[2] and Florian Meyer[2]

[1]*Cooperative Institute for Marine and Atmospheric Research, Research Corporation of the University of Hawai'i, Honolulu, Hawaii 96822, USA*

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

[3]*Marine Biology Graduate Program, University of Hawai'i at Mānoa, Honolulu, Hawaii 96822, USA*

[4]*Bioacoustics and Behavioral Ecology Lab, Syracuse University, Syracuse, New York 13244, USA*

[5]*Hawaiian Islands Humpback Whale National Marine Sanctuary, Kihei, Hawaii 96753, USA*

**ABSTRACT:**

Passive acoustic monitoring is widely used for detection and localization of marine mammals. Typically, pressure sensors are used, although several studies utilized acoustic vector sensors (AVSs), that measure acoustic pressure and particle velocity and can estimate azimuths to acoustic sources. The AVSs can localize sources using a reduced number of sensors and do not require precise time synchronization between sensors. However, when multiple animals are calling concurrently, automated tracking of individual sources still poses a challenge, and manual methods are typically employed to link together sequences of measurements from a given source. This paper extends the method previously reported by Tenorio-Hallé, Thode, Lammers, Conrad, and Kim [J. Acoust. Soc. Am. **151**(1), 126–137 (2022)] by employing and comparing two fully-automated approaches for azimuthal tracking based on the AVS data. One approach is based on random finite set statistics and the other on message passing algorithms, but both approaches utilize the underlying Bayesian statistical framework. The proposed methods are tested on several days of AVS data obtained off the coast of Maui and results show that both approaches successfully and efficiently track multiple singing humpback whales. The proposed methods thus made it possible to develop a fully-automated AVS tracking approach applicable to all species of baleen whales. © *2023 Acoustical Society of America*.

https://doi.org/10.1121/10.0021972

(Received 2 June 2023; revised 22 September 2023; accepted 5 October 2023; published online 24 October 2023)

[Editor: Karim G. Sabra] Pages: 2579–2593



## I. INTRODUCTION

Passive acoustic monitoring (PAM) has become a standard technology for detecting the presence of marine mammals for both research and mitigation purposes. The majority of current applications use non-directional single hydrophones (underwater acoustic pressure sensors) to detect the presence of marine mammal sounds within a defined "detection range" of the sensor. There are also, however, increasing numbers of PAM deployments that attempt more precise sound localization by combining measurements from multiple spatially-distributed sensors to provide one-dimensional (1D), two-dimensional (2D), or even three-dimensional estimates of a call's location, to aid population density estimation and behavioral studies.[1–4]

Most localization methods are based on measurements of the time difference of arrival (TDOA) of the same sound between various pairs of hydrophones.[5–8] More specifically, most acoustic tracking of baleen whales in two or three dimensions measures the relative arrival time of a specific call on multiple hydrophones spaced anywhere from hundreds of meters to kilometers apart, typically via cross correlation methods. If one assumes that waterborne sound speed is spatially homogeneous, the TDOA between each sensor pair defines a locus of geometric points that fits a hyperbola in case of two dimensions or a hyperboloid in case of three dimensions. The intersection of all resulting hyperbolas/hyperboloids identifies the position of sound origin. This approach requires precise time-synchronization between monitoring hydrophones, which is a common issue with these widely-spaced arrays.

Over the past two decades, an alternative localization strategy using directional sensors has been demonstrated on bowhead,[9,10] fin,[11] and most recently, humpback whales.[12] Directional acoustic sensors, also known as "acoustic vector sensors" (AVS), can estimate the azimuth of an acoustic signal by measuring both acoustic pressure and particle velocity at a single point. They typically have one or more directional particle velocity channels that are orthogonal to each other, so that the relative signal amplitudes on the directional channels can yield an azimuthal estimate. Sets of azimuthal measurements between at least two AVS can then triangulate a source position, an advantage compared to localization with hydrophones where at least three sensors

[a]Electronic mail: pgruden@hawaii.edu

are required. This approach does not require that acoustic data be precisely time-synchronized between hydrophones to the degree that TDOA methods do.

Regardless of which localization approach is used, when relatively few animals are acoustically active, it is straightforward to match measurements between sensors that are associated with the same sound and produce a localization estimate directly from a single call. A sequence of localizations can then be linked together across time to form a continuous track of an animal's two or three-dimensional path (localize-then-track).[3] Unfortunately, as the number of animals calling concurrently increases, this "data association" problem of assigning distributed measurements to specific sources becomes much more difficult. Under these circumstances it can be more convenient to link a temporal sequence of measurements (e.g., TDOAs or azimuthal estimates) together to form a "measurement trajectory" or "track," and then link measurement tracks between sensors that are identified as belonging to the same source, by identifying similar features between the measurement trajectories (track-then-localize).[6,8,12,13] These features can include the time-evolution of the measurements or the time-frequency structure of the signals. The challenge with the track-then-localize approach is devising an algorithm that correctly links sequences of measurements together and then matches the resulting measurement trajectories between sensors. Additional challenges arise from false alarms due to clutter (measurements not originating from target signals) and missed detections (signals from an animal present, but no measurement collected).

Several track-then-localize methods have been proposed for addressing the data association challenges of multiple calling animals, where calls are difficult to classify or associate.[6,8,12–14] Reference 12, in particular describes a compact and straightforward algorithm for generating azimuthal measurement trajectories from multiple concurrently-singing humpback whales using AVSs. The algorithm also demonstrated how azimuthal measurement tracks, or "azitracks," between vector sensors could be associated together correctly over time. An appealing feature of the vector sensor approach is that most steps, from pre-processing raw data to final localization, are easily automated and require relatively few parameters and only relatively crude time-synchronization between sensors. However, in Ref. 12, the azitracks of the singing humpback whales needed to be manually traced before applying the final automated step that links these tracks across instruments.

This paper demonstrates and compares two fully-automated approaches for azimuthal tracking, thus making possible a fully-automated vector sensor tracking algorithm based on Ref. 12 that can be applied to all known species of baleen whales. One approach is a "set-type" approach based on random finite sets (RFSs) and the probability hypothesis density (PHD) filter;[8] the other graph-based approach is based on a "vector-type" formulation, specifically on message passing or belief propagation (BP) algorithms.[15]

Section II provides essential background on two tracking approaches, and Sec. III discusses the specific dataset, data pre-processing, parameter selection, and evaluation metrics used to apply and compare the methods, the results of which are shown in Sec. IV. Finally, Secs. V and VI provide discussion and conclusions about the relative performance of the tracking approaches, along with other PAM scenarios where these tracking algorithms are applicable.

## II. TRACKING APPROACH

Target tracking approaches are typically model based and rely on the following key elements: a model that governs the system dynamics (state-transition model), a model that relates measurements to target states (measurement model), models for the noise processes (probabilistic description of the disturbances on both state-transition and measurement models), prior information, and track maintenance procedures.[16,17]

In this work, two tracking approaches for extracting whale azimuthal trajectories from vector sensor data are considered. Both approaches use the same underlying Bayesian statistical modeling framework, which is outlined in Sec. II A. Briefly, the vector sensor data are processed into sets of measurements (azimuths), which are due to the sound emitted by an unknown and time-varying number of humpback whales and other compact sources (such as boats), as well as spurious measurements (clutter). Both approaches then proceed through a series of evenly-spaced time steps. At each time step a predicted probability density function (pdf) is formed, based on both prior knowledge (i.e., posterior pdf from a previous time step) and information from the state-transition model (which statistically describes how whale azimuths develop). Then this pdf is updated to form the posterior target pdf, based on both the azimuthal measurements received at that time step and a measurement model that statistically describes the measurement generating process. This posterior target distribution is the fundamental quantity that allows computing an estimate for the number of whales and their location in azimuth. In addition, the posterior target distribution makes it possible to extract uncertainty information for estimates. Further details of the fundamental models underpinning the two approaches are given in Sec. II A.

An exact computation of the posterior target distribution is impractical because it would require integrating multidimensional functions; an enormous computational burden since the computational complexity would scale exponentially with time. The two tracking approaches in this work differ in the exact strategy they take in approximating the posterior target distributions. At each time step, the RFS approach and the corresponding PHD filter (outlined in Sec. II B) computes an approximation of the joint posterior multitarget pdf, while the graph-based approach and the corresponding BP algorithm (outlined in Sec. II C) provide approximations of the marginal posterior pdfs for the individual targets. Additionally, the RFS approach operates on

https://doi.org/10.1121/10.0021972

sets of unordered target states and measurements, and the graph-based approach uses ordered target states and measurements.

### A. Fundamental models and assumptions

We formulate the problem based on the following assumptions. At each time step $k$, the sensor data has been pre-processed into a set of points (measurements, $z_k$) by thresholding.[16] Targets (entities to be tracked, in this case, baleen whales) are represented by target states ($\boldsymbol{x}_k$), vectors that embed all information about the targets at time $k$. What we can observe about the targets that are present is encapsulated in the measurements, and each target can produce only a single measurement at a given time. The relationship between targets and measurements is not one-to-one: targets might be present, but due to detection uncertainty no measurement is collected for a variety of reasons, ranging from low signal-to-noise ratio (SNR) to changes in acoustic behavior. Conversely, measurements might arise from random fluctuations or other processes unrelated to any target, defined here as "clutter" or "false alarms." Therefore, the multi-target tracker receives a random number of measurements at each time $k$ due to the combined effects of detection uncertainty and false alarms.[16] Moreover, it is not known which measurement originates from which target, or whether the received measurement is a false alarm; a problem known as measurement origin uncertainty.[15,16] Therefore, at each time step, the multi-target state $\boldsymbol{X}_k$ and the multi-target measurement $\boldsymbol{Z}_k$ consist of a random number of unordered single target states ($\boldsymbol{x}_k$) and measurements ($z_k$).

Most target tracking approaches are formulated within the Bayesian filtering framework and consist of prediction and update steps.[16] As such, they require the definition of four underlying models, which will be discussed in more detail shortly. Briefly, one of these is a state-transition model, which models the evolution of azimuthal trajectories resulting from whales. Another is a measurement model, which defines the relationship between predicted whale azimuthal trajectories and its associated azimuthal measurements. Moreover, a clutter model, which specifies the generation of false alarms (i.e., azimuthal measurements that do not originate from whales) is required. Finally, one is required to know how and how often new azimuthal trajectories are formed, specified with a birth model. Further details on these models are as follows.

This work defines the target state as consisting of two variables: azimuth $\varphi$ and azimuth rate $\dot{\varphi}$:

$$\boldsymbol{x} = [\varphi, \dot{\varphi}]^T, \tag{1}$$

where $[\cdot]^T$ denotes the transpose. As will be seen shortly, the azimuth is the only directly observable variable, while the azimuth rate is a "hidden" variable.

The target states evolve independently and follow a nearly-constant velocity state-transition model,

$$\boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{x}_{k-1} + \boldsymbol{n}_k = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} \boldsymbol{x}_{k-1} + \boldsymbol{n}_k, \tag{2}$$

where subscripts $k$ and $k-1$ denote the current and previous time steps, and $\Delta$ denotes the time interval between consecutive time steps. For the humpback whale dataset studied here, $\Delta = 60$ s. The system noise, $\boldsymbol{n}_k$, is a Gaussian white noise with a covariance matrix $\boldsymbol{Q}$, which is typically defined as[17]

$$\boldsymbol{Q} = \begin{bmatrix} \frac{1}{3}\Delta^3 & \frac{1}{2}\Delta^2 \\ \frac{1}{2}\Delta^2 & \Delta \end{bmatrix} \sigma_{\mathrm{v}}^2, \tag{3}$$

where $\sigma_{\mathrm{v}}$ is the standard deviation of the system noise.

Each existing target has survival probability $p_{\mathrm{S}}$, i.e., a probability of target surviving from one step to another, and $p_{\mathrm{S}}$ is assumed to be independent of the state. For the humpback whale dataset studied here, $p_{\mathrm{S}}$ is set high ($p_{\mathrm{S}} = 0.95$), since humpback whales tend to sing for prolonged periods of time.[18]

At each time step, new targets can also be generated based on the measurements, a process defined by the birth rate parameter $\nu_{\mathrm{b}}$. Newborn targets become persistent targets in the next time step. Note that new targets can arise from new animals swimming into the area, existing ones starting/resuming singing after a prolonged silence, or a previously-singing whale increasing the SNR of its song.

Azimuthal measurements are related to the target states through the following measurement model:

$$z_k = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{\eta}_k = \begin{bmatrix} 1 & 0 \end{bmatrix} \boldsymbol{x}_k + \boldsymbol{\eta}_k, \tag{4}$$

where $\eta_k$ is the measurement noise assumed to be independent Gaussian white noise with a standard deviation $\sigma_{\mathrm{r}}$. Note, we assume that the azimuthal rate cannot be measured from the data. Also, note, in the RFS approach the amplitude of the track measurement is also incorporated (Sec. II B).

Each measurement either originates from a target or from clutter. A target is detected (and thus generates a measurement) with a probability of detection $p_{\mathrm{D}}$. The number of clutter measurements at time $k$ is Poisson distributed with clutter rate $r_{\mathrm{c}}$. It is independent across the sensors and of the number of targets, as well as target-generated measurements. Clutter azimuths are assumed to be uniformly distributed across the surveillance region (i.e., between 0° and 360°) at each time $k$,

$$c_k = \mathcal{U}[0, 360], \tag{5}$$

with $\mathcal{U}[a, b]$ denoting a uniform distribution between parameters $a$ and $b$.

Section III D explains how the parameters for these models are determined from a subset of the measurement data.

### B. RFS-based multi-target tracking: The PHD filter

In this approach, the multi-target states and multi-target measurements are described as RFS, i.e., sets with a random number of unordered elements. The sources' appearance and disappearance, as well as clutter and missed detections, are incorporated in the problem formulation and cast into the Bayesian framework.[19] A detailed overview of this approach can be found in Refs. 16 and 19.

One of the filters within the RFS framework is the PHD filter, which propagates the first-order moment (also known as the PHD) of the multi-target posterior distribution at each time step and is a tractable approximation of the multi-target Bayes filter.[20] Under the linear Gaussian assumptions for the underlying models and noise processes, a closed form solution to the PHD filter, the Gaussian mixture probability hypothesis density (GM-PHD),[21] can be derived. The PHD function in the GM-PHD filter is approximated with a mixture of weighted Gaussian components that are recursively propagated through time via a two stage prediction and update process. At every time step, new targets are introduced through the birth model, target states are estimated from the posterior PHD, and pruning and merging techniques are employed to reduce computational cost.[21] Improvements to the PHD filter have been proposed that extend the measurement model to incorporate additional features for better distinction between targets and clutter while reducing the bias in the number of estimated targets whenever new targets are initiated based on measurements.[8,22,23]

This work employs an extended version of the GM-PHD filter (GM-PHD-SA) filter developed from Ref. 8, which uses separate updates for persistent and newborn targets (S), and measured signal amplitude (A) as an additional feature in the measurements. Therefore, in addition to the measured azimuth [Eq. (4)] the associated amplitude from the thresholded azimuthal histogram display (AHD) is also noted and incorporated into the models, as detailed in Secs. III B and III C in Ref. 8.

#### 1. Prediction step

Briefly, in the prediction step of the GM-PHD-SA, the predicted PHDs for persistent, $v_{k|k-1,p}(\cdot)$, and newborn, $v_{k|k-1,b}(\cdot)$, targets at time $k$ consist of weighted Gaussian components, $J_{k-1}$ persistent components surviving from the previous time step, and $N_b$ newborn components. Note, the subscripts $k|k-1$ and $k|k$ denote predicted and updated quantities. The prediction step is thus[8,21,23]

$$v_{k|k-1,p}(\boldsymbol{x}) = \sum_{i=1}^{J_{k-1}} w_{k|k-1}^{(i)} \mathcal{N}\left(x; \boldsymbol{m}_{k|k-1}^{(i)}, \boldsymbol{P}_{k|k-1}^{(i)}\right), \tag{6}$$

$$v_{k|k-1,b}(\boldsymbol{x}) = \sum_{i=1}^{N_b} w_{k,b}^{(i)} \mathcal{N}\left(x; \boldsymbol{m}_{k,b}^{(i)}, \boldsymbol{P}_{k,b}^{(i)}\right), \tag{7}$$

where $\mathcal{N}(\cdot; \boldsymbol{m}, \boldsymbol{P})$ denotes a Gaussian density with mean $\boldsymbol{m}$ and covariance $\boldsymbol{P}$. The weights of the persistent targets, $w_{k|k-1}^{(i)}$, depend on the probability that the target survived from the previous time step: $w_{k|k-1}^{(i)} = p_S w_{k-1}^{(i)}$ and sum to the number of the targets present at time $k$. The means $\boldsymbol{m}_{k|k-1}^{(i)}$ and covariances $\boldsymbol{P}_{k|k-1}^{(i)}$ of persistent targets are predicted with a Kalman filter.[21,24]

The weights of the newborn targets, $w_{k,b}^{(i)}$, are proportional to the measured amplitudes and sum to the expected number of newborn targets (i.e., birth rate), $\nu_b$. The means of the newborn targets, $\boldsymbol{m}_{k,b}^{(i)}$, are based on the measurements, and their covariances $\boldsymbol{P}_{k,b}^{(i)}$ are proportional to $\boldsymbol{Q}$ in Eq. (3). See Sec. III B 5 in Ref. 8 for more details on newborn target generation.

The identity of Gaussian components (and thus target identity) is maintained by appending a label to each newborn Gaussian component and propagating it through time.[25,26]

#### 2. Update step

In the update step for the GM-PHD-SA filter, the persistent and newborn targets are updated separately. The updated PHDs for persistent, $v_{k|k,p}(\cdot)$, and newborn, $v_{k|k,b}(\cdot)$, targets also consist of weighted Gaussian components,[8]

$$\begin{aligned} v_{k|k,p}(\boldsymbol{x}) &= [1 - p_D] v_{k|k-1,p}(\boldsymbol{x}) \\ &\quad + \sum_{\tilde{z} \in \tilde{\boldsymbol{Z}}_k} \sum_{j=1}^{J_{k-1}} w_{k|k}^{(j)}(\tilde{z}) \mathcal{N}\left(x; \boldsymbol{m}_{k|k}^{(j)}(\tilde{z}), \boldsymbol{P}_{k|k}^{(j)}\right), \\ v_{k|k,b}(\boldsymbol{x}) &= \sum_{\tilde{z} \in \tilde{\boldsymbol{Z}}_k} \sum_{i=1}^{N_b} w_{k|k,b}^{(i)}(\tilde{z}) \mathcal{N}\left(x; \boldsymbol{m}_{k|k,b}^{(i)}(\tilde{z}), \boldsymbol{P}_{k|k,b}^{(i)}\right), \end{aligned} \tag{8}$$

where $[1 - p_D]$ represents a probability of a missed detection, and $\tilde{z}$ denotes a measurement with the added amplitude feature $a$, in a multi-target measurement set $\tilde{\boldsymbol{Z}}_k$. The means $\boldsymbol{m}_{k|k}^{(j)}$ and covariances $\boldsymbol{P}_{k|k}^{(j)}$ of the updated mixture for persistent targets are calculated from $v_{k|k-1,p}(\boldsymbol{x})$ with the Kalman filter update,[21,24] and the updated weights $w_{k|k}^{(j)}(\tilde{z})$ are calculated according to Ref. 8,

$$w_{k|k}^{(j)}(\tilde{z}) = \frac{w_{k|k-1}^{(j)} g_a(a) g_p^{(j)}(z|\boldsymbol{x})}{\mathcal{L}(\tilde{z})}, \tag{9}$$

$$\begin{aligned} \mathcal{L}(\tilde{z}) &= r_c c_k c_a(a) + g_a(a) \sum_{l=1}^{N_b} w_{k,b}^{(l)} g_b^{(l)}(z|\boldsymbol{x}) \\ &\quad + g_a(a) \sum_{l=1}^{J_{k-1}} w_{k|k-1}^{(l)} g_p^{(l)}(z|\boldsymbol{x}), \end{aligned} \tag{10}$$

where $g_a(a)$ is the target amplitude likelihood function, $c_a(a)$ is the clutter amplitude likelihood function, and $g_p(\cdot)$ and $g_b(\cdot)$ are the persistent and newborn target likelihood functions based on azimuthal measurements, respectively.

The means $\boldsymbol{m}_{k|k,b}^{(i)}(\tilde{z})$ and covariances $\boldsymbol{P}_{k|k,b}^{(i)}$ of the newborn targets are updated from $v_{k|k-1,b}(\boldsymbol{x})$ with the Kalman

https://doi.org/10.1121/10.0021972

filter update,[24] and the updated newborn weights $w_{k|k,b}^{(i)}$ are calculated according to Ref. 8,

$$w_{k|k,b}^{(i)}(\tilde{z}) = \frac{w_{k,b}^{(i)} g_a(a) g_b^{(j)}(z|\boldsymbol{x})}{\mathcal{L}(\tilde{z})}. \tag{11}$$

To maintain computational efficiency, additional parameters are defined to limit the number of Gaussian components (potential targets): the pruning threshold $T_\mathrm{r}$, merging threshold $U$, and maximum allowed number of Gaussian components $J_\mathrm{max}$.[21] At the end of each iteration the target states are estimated from the updated PHDs, by taking all components whose weights exceeded the weight threshold $w_\mathrm{th}$.[21] In this study, we set $T_\mathrm{r} = 0.001$, $U = 4$, $J_\mathrm{max} = 100$, and $w_\mathrm{th} = 0.1$. In addition, some of the parameters for the GM-PHD-SA filter were determined based on the training data and are listed in Table II.

### C. Graph-based multi-target tracking

In what follows, we present the graph-based multi-target tracking approach.[15] Here, the target states and measurements are described as random vectors, i.e., they are ordered. Probabilistic data association is used to associate measurements to target states and a binary existence variable introduced for each target state indicates whether a particular target exists. A graphical model makes it possible to exploit statistical independence assumptions, which lead to factored forms of posterior probability distributions, to reduce the computational complexity of filtering and probabilistic data association. A detailed overview of this approach can be found in Refs. 15, 27, and 28.

#### 1. Object state model and factor graph

The time-varying number of targets is taken into account by the introduction of potential targets (PT). Each of the $j \in \{1, \ldots, j_k\}$ PTs is represented by a PT state defined as $\boldsymbol{y}_k^{(j)} \triangleq [\boldsymbol{x}_k^{(j)\mathrm{T}}\ r_k^{(j)}]^\mathrm{T}$, where $\boldsymbol{x}_k^{(j)}$ is the kinematic single-target state introduced in Eq. (2) and $r_k^{(j)} \in \{0, 1\}$ is the existence variable modeling the existence/nonexistence of the PT $j$. For each PT state $\boldsymbol{y}_{k-1}^{(j)}$, $j \in \{1, \ldots, j_{k-1}\}$ at time $k-1$, there is one "legacy" PT state $\underline{\boldsymbol{y}}_k^{(j)}$ at time $k$. In addition, at time $k$, $m_k$ new PT states are introduced, i.e., $\bar{\boldsymbol{y}}_k^{(m)} \triangleq [\bar{\boldsymbol{x}}_k^{(m)\mathrm{T}}\ \bar{r}_k^{(m)}]^\mathrm{T}$, $m \in \{1, \ldots, m_k\}$. The total number of (legacy and new) PT states at time $k$ is $j_k = j_{k-1} + m_k$ and the joint PT state at time $k$ is denoted as $\boldsymbol{y}_k \triangleq [\boldsymbol{y}_k^{(1)\mathrm{T}} \cdots \boldsymbol{y}_k^{(j_k)\mathrm{T}}]^\mathrm{T}$. Since the measurements are subject to data association (DA) uncertainty, it is unknown which measurement originated from which PT, and a measurement may also arise from clutter. The DA in the graph-based approach is modeled by the "object-oriented" DA vector $\boldsymbol{a}_k = [a_k^{(1)} \cdots a_k^{(j_{k-1})}]^\mathrm{T}$ and the "measurement-oriented" DA vector $\boldsymbol{b}_k = [b_k^{(1)} \cdots b_k^{(m_k)}]^\mathrm{T}$. Here, the quantities $a_k^{(j)} \in \{0, \ldots, m_k\}$ and $b_k^{(m)} \in \{0, \ldots, j_{k-1}\}$ describe the potential associations between PTs and measurements, and thus quantity the strength of the links between targets and measurements. Note that the description of the DA in terms of $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$ is redundant, but this symmetric arrangement of the links enables more efficient graph-based multi-target tracking, as seen in the following.

Given the state-transition model described in Eq. (2) and the definition of the PT target state noted previously, the joint posterior pdf of $\boldsymbol{y}_{0:k}$, $\boldsymbol{a}_{1:k}$, and $\boldsymbol{b}_{1:k}$ can be arranged as[15]

$$\begin{aligned} f(\boldsymbol{y}_{0:k}, \boldsymbol{a}_{1:k}, \boldsymbol{b}_{1:k} | \boldsymbol{z}_{1:k}) \\ \propto \left( \prod_{j''=1}^{j_0} f(\boldsymbol{y}_0^{(j'')}) \right) \prod_{k'=1}^{k} \left( \prod_{j'=1}^{j_{k'-1}} f(\underline{\boldsymbol{y}}_{k'}^{(j')} | \boldsymbol{y}_{k'-1}^{(j')}) \right) \\ \times \prod_{m=1}^{m_{k'}} v(\bar{\boldsymbol{x}}_{k'}^{(m)}, \bar{r}_{k'}^{(m)}, b_{k'}^{(m)}; \boldsymbol{z}_{k'}^{(m)}) \\ \times \left( \prod_{j=1}^{j_{k'-1}} q(\underline{\boldsymbol{x}}_{k'}^{(j)}, \underline{r}_{k'}^{(j)}, a_{k'}^{(j)}; \boldsymbol{z}_{k'}) \prod_{m'=1}^{m_{k'}} \Psi_{j,m'}(a_{k'}^{(j)}, b_{k'}^{(m')}) \right). \end{aligned} \tag{12}$$

Here, $f(\boldsymbol{y}_0^{(j)})$ is the prior distribution of PT $j \in \{1, \ldots, j_0\}$, $f(\underline{\boldsymbol{y}}_k^{(j)} | \boldsymbol{y}_{k-1}^{(j)})$ describes the temporal evolution of PT $j \in \{1, \ldots, j_{k-1}\}$ from time step $k-1$ to time step $k$ and is based on the state transition model in Eq. (2). Functions $v(\bar{\boldsymbol{x}}_k^{(m)}, \bar{r}_k^{(m)}, b_k^{(m)}; \boldsymbol{z}_{k'}^{(m)})$ and $q(\underline{\boldsymbol{x}}_k^{(j)}, \underline{r}_k^{(j)}, a_k^{(j)}; \boldsymbol{z}_k)$ describe the detection process of newborn targets $m \in \{1, \ldots, m_k\}$ and legacy targets $j \in \{1, \ldots, j_{k-1}\}$, respectively, as discussed in Sec. II A. In addition, function $\Psi_{j,m}(a_k^{(j)}, b_k^{(m)})$ checks if measurement $m$ was originated by PT $j$.[29] Further explanations and exact definitions of these functions can be found in Sec. VIII G of Ref. 15.

The direct evaluation of the joint posterior in Eq. (12) is computationally intractable; however, note that the posterior is the product of numerous factors. It can thus be represented by the "factor graph" in Fig. 1, which is a graphical model that represents the factorization structures of probability distributions.[30,31] More precisely, in Fig. 1, the random variables and the pdfs/ probability mass functions (pmfs) involved in $f(\boldsymbol{y}_{0:k}, \boldsymbol{a}_{1:k}, \boldsymbol{b}_{1:k} | \boldsymbol{z}_{1:k})$ are represented by circles and squares, respectively. A circle is connected with a square if the random variable represented by that circle is involved in the corresponding pdf/pmf represented by the corresponding square.

#### 2. Problem formulation and selected message passing operations

In the following, we utilize the factorization structure of the joint posterior pdf $f(\boldsymbol{y}_{0:k}, \boldsymbol{a}_{1:k}, \boldsymbol{b}_{1:k} | \boldsymbol{z}_{1:k})$ in Eq. (12), which is visualized in the factor graph of Fig. 1, to perform object detection and state estimation at time $k \geq 1$ based on all measurements $\boldsymbol{z}_{1:k}$ collected up to time $k$. More precisely, object detection is performed by comparing the posterior existence probability of PT $j \in \{1, \ldots, j_k\}$, i.e., $p(r_k^{(j)} = 1 | \boldsymbol{z}_{1:k})$, with a threshold $P_\mathrm{th}$. All PTs with posterior

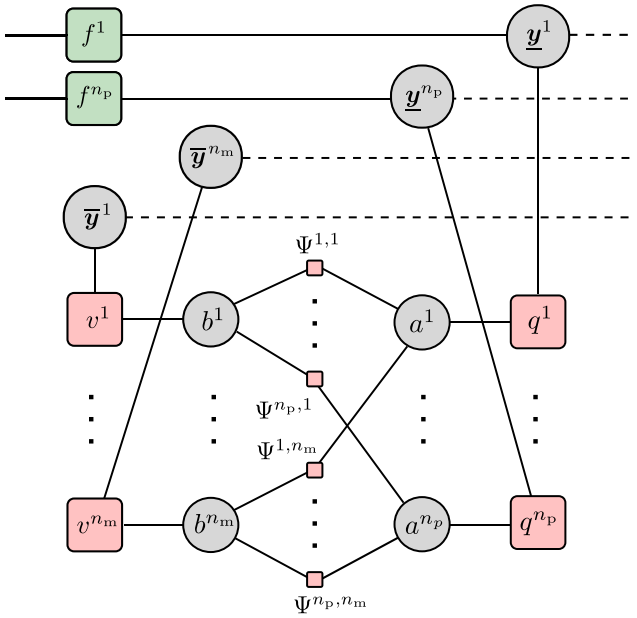



FIG. 1. (Color online) Factor graph representing the joint posterior pdf $f(\boldsymbol{x}_{0:k}, \boldsymbol{r}_{0:k}, \boldsymbol{a}_{1:k}, \boldsymbol{b}_{1:k}|\boldsymbol{z}_{1:k})$ in Eq. (12) for one time step $k$.

existence probabilities that exceed this threshold are declared to exist. For existent PTs, state estimation is performed by calculating the minimum-mean squared errors estimate[32] as

$$\hat{\boldsymbol{x}}_k^{(j)} \triangleq \int \boldsymbol{x}_k^{(j)} f(\boldsymbol{x}_k^{(j)}|r_k^{(j)} = 1, \boldsymbol{z}_{1:k}) \mathrm{d}\boldsymbol{x}_k^{(j)}, \tag{13}$$

where $f(\boldsymbol{x}_k^{(j)}|r_k^{(j)} = 1, \boldsymbol{z}_{1:k}) = f(\boldsymbol{x}_k^{(j)}, r_k^{(j)} = 1|\boldsymbol{z}_{1:k}) / p(r_k^{(j)} = 1|\boldsymbol{z}_{1:k}) = f(\boldsymbol{x}_k^{(j)}, r_k^{(j)} = 1|\boldsymbol{z}_{1:k}) / \int f(\boldsymbol{x}_k^{(j)}, r_k^{(j)} = 1|\boldsymbol{z}_{1:k}) \mathrm{d}\boldsymbol{x}_k^{(j)}$.

Object detection and state estimation require the marginal posterior pdfs $f(\boldsymbol{x}_k^{(j)}, r_k^{(j)}|\boldsymbol{z}_{1:k}) \triangleq f(\boldsymbol{y}_k^{(j)}|\boldsymbol{z}_{1:k}), \quad j \in \{1, \ldots, j_k\}$. However, calculating $f(\boldsymbol{x}_k^{(j)}, r_k^{(j)}|\boldsymbol{z}_{1:k})$ by direct marginalization (integration) of Eq. (12) is infeasible due to the high dimensionality of $\boldsymbol{y}_{1:k}$, $\boldsymbol{a}_{1:k}$, and $\boldsymbol{b}_{1:k}$. As in Refs. 15 and 27, we consider an approximate calculation by performing "message passing," also known as "BP," on the factor graph in Fig. 1. More precisely, in the BP approach local operations at the individual graph nodes are performed and their results—called messages—are exchanged along the graph edges. This makes it possible to efficiently calculate so-called beliefs $\tilde{f}(\boldsymbol{x}_k^{(j)}, r_k^{(j)})$, $j \in \{1, \ldots, j_k\}$ that accurately approximate the marginal posterior pdfs $f(\boldsymbol{x}_k^{(j)}, r_k^{(j)}|\boldsymbol{z}_{1:k}), j \in \{1, \ldots, j_k\}$ required for the object detection and state estimation formulas [e.g., Eq. (13)].

In the final implementation of our graph-based tracking algorithm, each spatial pdf $f(\boldsymbol{x}_k^{(j)}|r_k^{(j)} = 1, \boldsymbol{z}_{1:k}), j \in \{1, \ldots, j_k\}$ is represented by 300 particles (see Ref. 27 for details). To keep computational complexity low, at each time $k$, a suboptimal pruning step has to be performed: any PT with a probability of existence $p(r_k^{(j)} = 1|\boldsymbol{z}^{(1:n)})$ below a threshold $P_{\mathrm{pr}}$ is removed from the state space. In this work, we set $P_{\mathrm{th}} = 0.75$ and $P_{\mathrm{pr}} = 10^{-5}$. The values of further system parameters are obtained by training from annotated data and are summarized in Table II.

## III. METHODS

### A. Equipment, deployment geometry, and timelines

The dataset used to test these automated tracking algorithms is an expanded version of the one used in Ref. 12. Between March and July 2020, three Directional Autonomous Seafloor Acoustic Recorders (DASARs)[9] were deployed along the south-facing coast of western Maui, approximately 3 km apart, to detect and track singing humpback whales within the Hawaiian Islands Humpback Whale National Marine Sanctuary. The northernmost DASAR was designated "A," the middle DASAR "B," and southernmost DASAR "C."

Each DASAR is equipped with a vector sensor, similar to those used in disposable 53-series Directional Frequency Analysis and Recording (DIFAR) sonobuoys.[33] DASARs contain an omnidirectional acoustic pressure sensor (sensitivity of −149 dB re V/ 1 $\mu$Pa) and two horizontal directional sensors capable of measuring the north-south and east-west components of acoustic particle velocity.[9,12] The DASARs were deployed by rope from a small vessel in about 20 m of water. Accurate triangulated distances up to 9 km range were expected (i.e., about three times the baseline distance between DASARs). The orientation of the DIFAR sensor on each DASAR was determined after deployment by having the small vessel trace a 100-m radius circle six times around each DASAR location: three clockwise and three counterclockwise. A brute-force inversion then yielded both the DIFAR orientation and the clock offset between the internal DASAR clock and GPS-UTM time. Additional descriptions of the equipment and deployment are provided in Ref. 12.

In this study thirteen days in April were used to train and test the proposed tracking schemes (Table I), including April 18, the original 24-h dataset used by Ref. 12.

### B. Computing azimuthal tracks

The Ref. 12 provides a detailed procedure for generating azimuthal tracks, or azitracks, from the AVS data, but the process is summarized here. Estimating the azimuth of

TABLE I. Dataset used in this study. Each Encounter comprises a full 24 h period. The number of annotated azitracks are reported for each DASAR (C/B/A) respectively. Days used for testing are indicated by *.

| Encounter Date | Encounter ID | N annotated azitracks |
|---|---|---|
| 04042020* | 1 | 82 / 73 / 74 |
| 04062020 | 2 | 97 / 99 / 106 |
| 04102020 | 3 | 84 / 71/ 84 |
| 04142020 | 4 | 69 / 68 / 87 |
| 04152020 | 5 | 35 / 23 / 28 |
| 04162020 | 6 | 42 / 46 / 55 |
| 04182020* | 7 | 24 / 20 / 17 |
| 04232020* | 8 | 20 / 24 / 28 |
| 04242020 | 9 | 13 / 13/ 22 |
| 04262020 | 10 | 12 / 11/ 10 |
| 04282020 | 11 | 20 / 17 / 15 |
| 04292020 | 12 | 21 / 19 / 21 |
| 04302020 | 13 | 24 / 21/ 19 |

https://doi.org/10.1121/10.0021972

sounds measured on each individual DASAR begins by computing an estimate of the acoustic vector active intensity. During the deployment, each of the three channels (pressure, north-south velocity, and east-west velocity) is low-pass filtered and sampled at 1 kHz before being stored on disk. During post-recovery analysis, each digital time series is then converted into a complex spectrogram using the short-time Fast Fourier Transform (SFFT) with a 1024-point analysis window (∼1 s). For each time-frequency bin in the SFFT output, the in-phase product of the acoustic pressure and particle velocity is computed along each axis. This quantity, called the "active intensity" in the literature,[34] measures the component of the acoustic field that is actively transporting energy through the measurement point. The ratio of the active intensity components along both axes yields the four-quadrant tangent of the azimuth of the dominant signal present at that time-frequency bin. A surface that displays the estimated azimuth for every time-frequency bin of the SFFT has been dubbed an "azigram" in previous work.[35] Azigrams produce a convenient time-frequency visualization of the dominant directionality of the ambient sound field. Furthermore, a histogram of the azigram values, constructed from a 60-s window and sampled across frequencies between 10 and 450 Hz, provides a means for estimating the number of sources and their azimuths by identifying local peaks in the histogram. When subsequent 60-s azigram histograms are stacked against each other, the resulting 24-h image, dubbed an AHD, creates ridges that identify persistent sources over time. Tracing these ridges generates the azitracks for individual sources on that DASAR. A metric called the normalized transport velocity (NTV)[36] can be used to retain only those time-frequency estimates that arise from azimuthally compact sources like whales and boats. The azitracks in this paper were generated over 24-h intervals using 1024-point FFT windows, 60-s histograms with 2° resolution, and filtered using a 0.75 minimum threshold for the acoustic transport velocity.

### C. Ground truth data

Two trained analysts manually traced all visible azitracks for each of the three DASARs over each 24-h period from midnight to midnight local standard time using AHDs and a custom MATLAB script.[12] Tracing was done conservatively; unless they could be unambiguously separated, azitracks were logged as separate events whenever two or more tracks crossed, or individual tracks showed large gaps.

On days with large numbers of singing whales, it was often convenient to trace azitracks viewing only three hours of data at a time. The resulting manually-traced azitracks were then "stitched" together to make continuous tracks over the 24 h window. Tracks that existed within 15 min of either the beginning or end of the three-hour window are extrapolated to the start or end time of the window. They were then assigned to the extrapolated tracks from other contiguous segments from earlier and later three-hour windows using the Hungarian algorithm.[37]

Noise signals from nearby ships and boats sometimes dominated the acoustic signal, masking the presence of humpback song. Azitracks were flagged as arising from vessels if their azimuthal change exceeded 10° $\text{min}^{-1}$ and were removed from evaluation. Azitracks shorter than 20 min were also eliminated, resulting in a finalized manually-annotated "ground truth" azitrack database (Table I).

#### 1. Reducing azitrack biases

Despite the use of an inversion to estimate the DASAR orientation on the ocean floor, biases in the DASAR orientations still potentially existed. Here, we briefly review the process in which 2D tracks are generated from azitracks following Ref. 12, and then present a method for correcting potential biases in the DASAR orientation by localizing a subset of azitracks.

Azitracks between a pair of DASARs are matched based on the similarity of the time-frequency structure of the songs they are tracking. Only the pairs that overlap for at least 15 min in time are considered for matching. At each DASAR, a binary spectrogram image of a tracked humpback song is created by applying a directional threshold to a 60-s azigram. The similarity of two potentially linked azitracks is scored based on the maximum value of the cross correlation between two binary images between DASARs. The cross correlation scores were computed for every 60 s long time segments using an azimuthal threshold of 15° degrees. Correctly associated tracks from each pair of DASARs would yield a high median score, as illustrated by the confusion matrix in Fig. 2. In particular, the azitracks between each pair of DASARs with the highest score in each column of the confusion matrix is considered to have originated from the same humpback whale, i.e., a match. However, if the matched azitrack score was below a threshold of 5, the matched azitracks are discarded.

Azitracks from two pairs of DASARs, i.e., DASARs A, B and DASARs B, C, are associated and scored. Three azitracks are considered to be from the same whale if its track in DASAR B yields high median scores in both confusion matrices. More technical details of the matching process can be found in Sec. IV E of Ref. 12.

The 2D whale tracks are then generated by triangulating the matched azitracks from at least two DASARs.[38] Each 2D track is a link of a whale's position in terms of latitude and longitude localized by triangulating every minute. It starts when there are at least two matched azitracks to triangulate from and stops when there are less than two matched azitracks. When three DASARs are used to triangulate a position, a localization confidence ellipse with 90% confidence can be drawn following Ref. 9.

The size of the 90% confidence ellipse can be used to correct potential misaligned orientation biases of the DASARs. First, a normalized "uncertainty metric" is defined as the ratio of the equivalent spherical radius of the confidence ellipse of 2D localization to the distance of the whale from DASAR B. Twenty spatial 2D whale tracks from encounter 7 (Table I) were then selected for optimization.

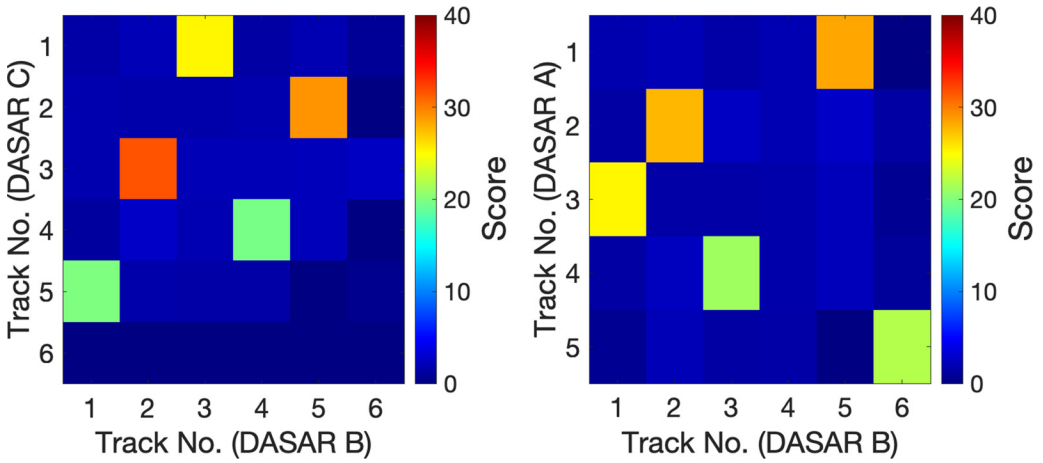



FIG. 2. (Color online) The confusion matrix from matching DASAR BC (left) and DASAR AB (right). The color represents the median scores of the compared azitracks over a six-hour window from the manually tracked data in April 18th, 2020. The azitracks from a pair of DASARs are considered a match if they yielded the highest score in a column of the confusion matrix and a score above a threshold score of 5. By inspecting the track number of DASAR B with high score in both score boards, one can determine whether azitracks in all three DASARs have matched. For example, Track No. 2 in DASAR B matches to Track No. 3 and 2 on DASARs C and A, respectively. Azitracks that are matched between at least two DASARs are triangulated to generate the 2D whale track.

The reference orientation of each DASAR was systematically adjusted in 0.1° increments, and the error ellipses and corresponding uncertainty metrics for all 20 tracks were recomputed. The combination of orientation adjustments for all three DASARs that minimized the mean uncertainty metric for the 20 tracks was determined to be the optimal orientation adjustment, and these values were used when computing the final azitrack traces.

### D. Learning filters' parameters

The models used by the tracking filters (in Sec. II A) require specification of parameters which are typically either known *a priori* or learned from training data. In this work, filtering parameters learned from training data include the probability of detection ($p_D$), clutter rate ($r_c$), birth rate ($\nu_b$), system, and measurement noise deviations ($\sigma_v$ and $\sigma_r$, respectively).

In order to learn these parameters, the ground truth data (see Sec. III C) were divided into training and test sets. Out of 13 days (encounters) of the ground truth data that contained singing humpback whales, three encounters were randomly selected to be the test data (one encounter with higher numbers of whale azitracks, and two with lower numbers, Table I), and the rest were used as the training data to learn the tracking parameters.

Since each DASAR was deployed at a different location that had different noise characteristics (e.g., DASAR A was close to a local harbor), it was decided to compute parameters for each DASAR separately.

To compute the parameter values from the training set, the manually-annotated azitracks were first associated with the measurements obtained from the AHDs in the training data. Measurements were then extracted from an AHD for each time step by selecting local peaks in that AHD above a threshold of 0.1. The association was performed using the Munkres Hungarian algorithm and a 5° cutoff distance between measured and annotated azimuthal angles.

Extracted measurements that were not assigned to any hand annotated azitrack were considered false alarms. Thus, the number of false alarms could be obtained per time step, and the clutter rate ($r_c$) was computed for each DASAR as the mean or the variance of the number of false alarms (whichever was bigger) across the time steps. Note the number of false alarms is dependent on the selected threshold used to extract peaks from AHD: if a lower threshold is used, the clutter rate is higher.

The birth rate (i.e., the expected number of the newborn target in a given time step, $\nu_b$) was computed as the mean or the variance of the number of newborn azitracks (whichever was bigger) across the time steps.

The probability of detection per time step, $p_D$, was computed for each time step as the number of associated measurements (i.e., detections) divided by the number of annotated azitracks in that time step. The overall probability of detection was taken as the mean across time steps.

The measurement noise deviation, $\sigma_r$ [Eq. (4)], was computed as the mean difference between annotated azitracks and measurements across all time steps. The system noise deviation, $\sigma_v$ [Eq. (3)], was computed as a median of the maximum second order derivatives of azimuth (which can be interpreted as source acceleration in the azimuthal space) across the annotated azitracks.

The derived parameter values varied, depending on the days used in the training data (Fig. 3), and some parameters seemed correlated with the number of azitracks in that encounter (e.g., birth rate), a result that is expected. However, when testing the algorithms on the test data, the final values for each parameter were computed as a mean value across the training dataset and are shown in Table II.

### E. Performance evaluation using the generalized optimum subpattern assignment metric (GOSPA)

This study conducts two types of comparisons between the manually-annotated and automatically extracted

https://doi.org/10.1121/10.0021972

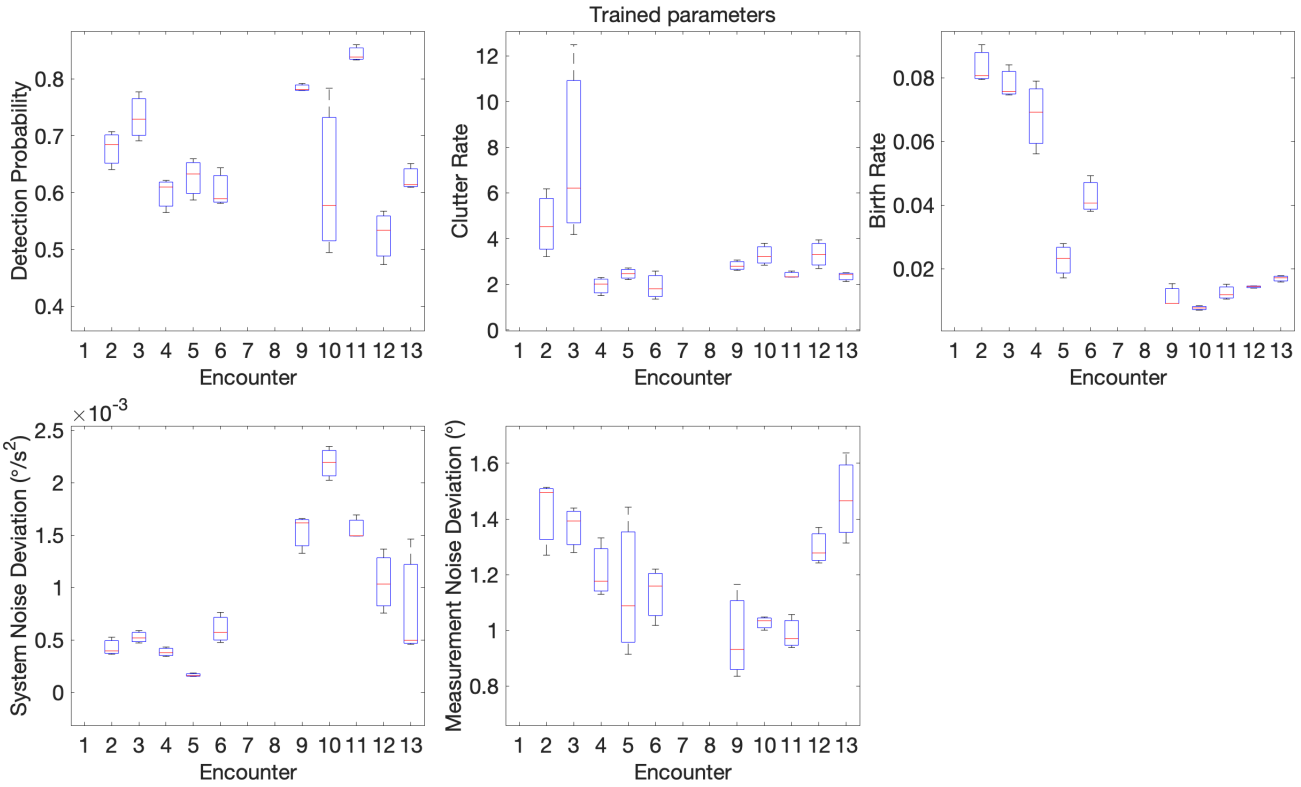



FIG. 3. (Color online) Trained parameters per encounter. Note that encounters 1, 7, and 8 are used for testing, and are thus not displayed. For each encounter a given parameter value is shown as a box-plot of all three DASAR values with the median value indicated as a solid line. The bottom and top edges of the box indicate 25th and 75th percentiles, respectively, and the whiskers extend to most extreme data points not considered outliers.

azitracks. A qualitative comparison simply derives the full two-dimensional tracks of singing whales and compares the resulting maps side by side. A more quantitative metric is also used to rigorously evaluate the performance.

Evaluating and comparing the performance of multi-target tracking methods is complicated by the fact that the number of estimated target states can be larger or smaller than true number of targets. Often performance assessment is based on practically relevant concepts such as localization error for correctly detected targets, and the number of missed or erroneously estimated targets. This approach is appealing, but the resulting errors are typically defined in a rather *ad hoc* manner, which makes it hard to reproduce or compare performance results.

An alternative approach for defining estimation errors in a target tracking scenario is to consider both ground truth and estimates as two sets with potentially different numbers of elements. The estimation error can then be seen as the "distance" between these sets, with distance being defined by a metric that satisfies the properties of non-negativity, definiteness, symmetry, and triangle inequality. A particularly appealing metric within this framework is the GOSPA,[39] which merges localization errors for properly detected targets with penalties for false and missed targets, creating a single performance metric. In addition to satisfying all the fundamental requirements of a metric, the GOSPA can be subdivided into separate error contributions from localization errors and mislabeled tracks, an important consideration for practical applications.

TABLE II. Summary of trained parameters used in BP and GM-PHD-SA filters for azimuthal tracking. $r_c$ denotes the clutter rate; $\nu_b$ denotes the birth rate (expected number of newborn targets); $p_D$ denotes the probability of detection; $\sigma_r$ and $\sigma_v$ denote measurement and system noise deviations, respectively.

| DASAR | $r_c$ | $\nu_b$ | $p_D$ | $\sigma_r$ [°] | $\sigma_v$ [°/s$^2$] |
|---|---|---|---|---|---|
| A | 2.99 | 0.04 | 0.64 | 1.30 | $8.6 \times 10^{-4}$ |
| B | 3.06 | 0.03 | 0.67 | 1.15 | $8.9 \times 10^{-4}$ |
| C | 3.78 | 0.04 | 0.68 | 1.16 | $10 \times 10^{-4}$ |

Before defining the GOSPA metric, we need to introduce the notion of an "assignment set" $\vartheta$. An assignment set $\vartheta \in \{1,\ldots,I\} \times \{1,\ldots,J\}$ between two sets with $I$ and $J$ elements has the property that $(i,j),(i,j') \in \vartheta$ implies $j = j'$ and that $(i,j),(i',j) \in \vartheta$ implies $i = i'$. The assignment $\vartheta$ introduces a generalized intersection of the two sets. For a certain choice of parameters (see Ref. 39 for details), the GOSPA for true target states $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{I}$ and estimated target states $\mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^{J}$, can now be expressed as

$$D(\mathcal{X},\mathcal{Y}) = \min_{\vartheta \in \Theta_{|X||Y|}} \left[ \sum_{(i,j)\in\vartheta} d(\mathbf{x}_i,\mathbf{y}_j) + \frac{c}{2}(|\mathcal{X}| + |\mathcal{Y}| - 2|\vartheta|) \right], \tag{14}$$

where $\Theta_{|X||Y|}$ is the set of all possible combinations of indexes, $c > 0$ is a cut-off parameter, and $d(\boldsymbol{x}_i,\boldsymbol{y}_j)$ is an

arbitrary "inner metric" defined for vectors, e.g., the 2-norm. The GOSPA thus includes a component that represents the localization error between true target states and estimated target states. In the considered tracking scenario—without ground truth information and in the azimuth domain—this component represents the angular error between manually annotated and automatically estimated 1D azimuth values. As an inner metric, we use the 1D Euclidean distance, i.e., $|x_i - y_j|$.

From expression in Eq. (14), it can be seen how the GOSPA combines angular errors for properly detected targets with the penalty for false and missed targets. The localization error is represented by the first term. Here, the properly detected targets and their estimates are assigned according to $\vartheta$. Targets that remain unassigned are either false or missed targets and are penalized by $c/2$. In particular, $|X| - |\vartheta|$ and $|Y| - |\vartheta|$ represent the number of missed and false targets, respectively. Thus, according to the term $c(|X| + |Y| - 2|\vartheta|)/2$, any missed or false target results in a cost of $c/2$. In the results that follow $c$ is set to 5°.

Note that GOSPA in Eq. (14) is computed per each time step, and total GOSPA for each encounter is computed as a mean across all time steps.

## IV. RESULTS

The two tracking approaches discussed in Secs. II B and II C were applied to the test data (Table I) that contained two days with smaller numbers of azitracks (singing whales), and one day with a larger number of azitracks. The GOSPA metric was computed for all encounters (Figs. 4–6), and run times noted for both approaches (Table III).

The performance of the algorithms varied depending on the encounter. The total GOSPA error was higher for both filters when the number of annotated azitracks was higher: encounter 1 had on average 76 azitracks across the DASARs (Table I) and had higher total GOSPA error (Fig. 4) compared to encounters 7 and 8 where there were on average 20 and 24 azitracks (Table I), respectively, and the total error was lower (Figs. 5 and 6). The total GOSPA error was comparable between the two filters in encounters 1 and 7 (Figs. 4 and 5), but was higher for the PHD filter in encounter 8 compared to the BP filter (Fig. 6). A closer examination of the automated tracks revealed that a couple of vessels with a slower rate of change in azimuths were present in the course of 2-h period. While both methods detected parts of these vessel tracks, the PHD detected more of the vessel tracks in small fragments resulting in a higher average false target error which in turn resulted in a higher average overall error. Individual contributions of localization errors, missed targets, and false alarms to the overall total GOSPA error metric were also examined.

Individual components of GOSPA reflected the trend observed for total GOSPA metric: the errors were higher in encounter 1 where the number of azitracks was higher, compared to encounters 7 and 8 which had a smaller number of azitracks. For all encounters, the biggest contribution to the overall GOSPA error for both filters came from the missed target error component of the GOSPA metric (bottom left subplot on Figs. 4–6). Note, for example, a missed target error of 2.5° signifies that on average 1 target was missed across all time steps [Eq. (14)]. When the numbers of azitracks were lower in encounters 7 and 8 there was on average about 1 missed target (i.e., the average error of about



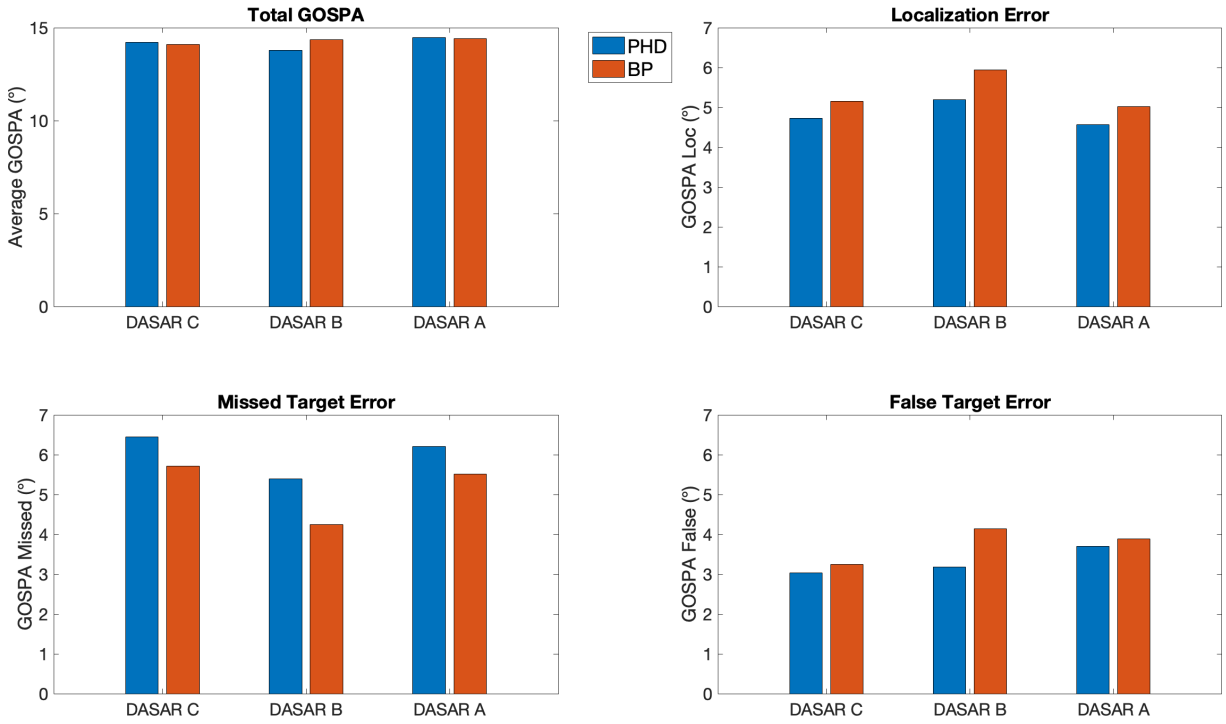



FIG. 4. (Color online) GOSPA error metric and its components for encounter 1, a day with larger numbers of annotated azitracks.

https://doi.org/10.1121/10.0021972

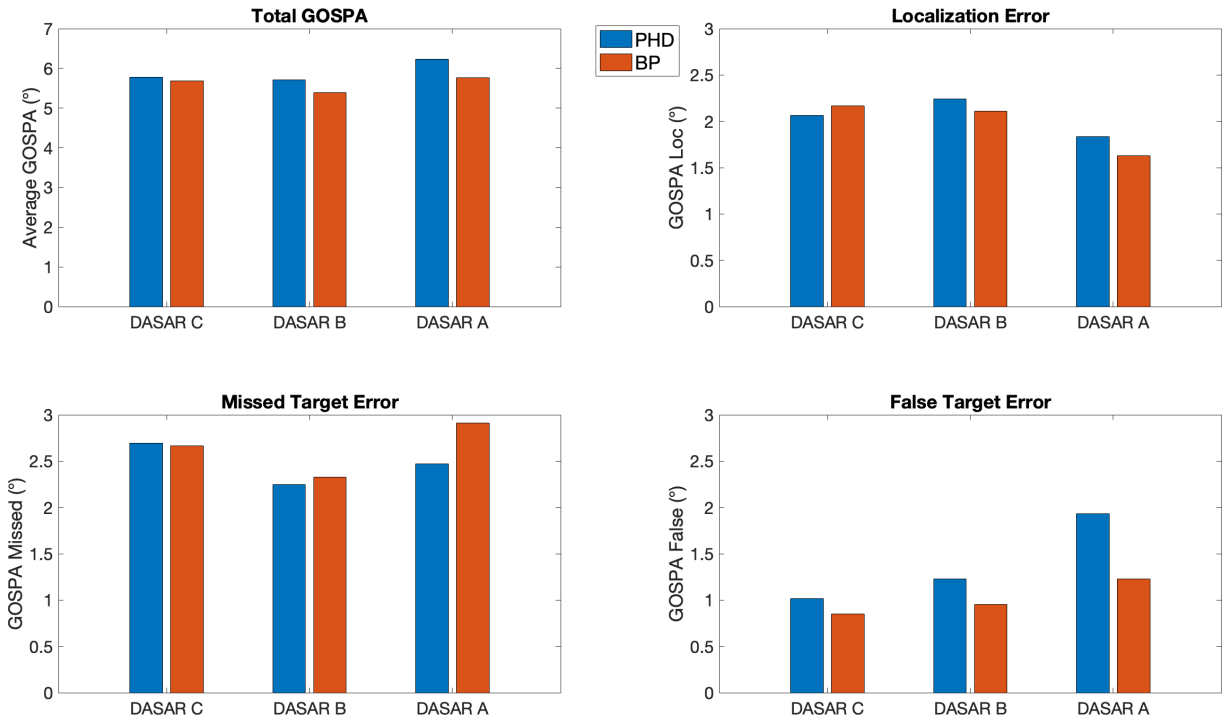



FIG. 5. (Color online) GOSPA error metric and its components for encounter 7, a day with lower numbers of annotated azitracks.

2.5°) for that 24-h time period (Figs. 5 and 6), and when numbers of azitracks were higher there were on average about two missed targets (Fig. 4). False target error was low, on average there was less than 1 false alarm (i.e., average error less than 2.5°) in encounters 7 and 8 (Figs. 5 and 6), and less than two false alarms in encounter 1 (Fig. 4). The sum of all localization errors averaged over time was about 5° in encounter 1 (Fig. 4), less than 2.5° in encounter 7 (Fig. 5) and less than 1.5° in encounter 8 (Fig. 6).

The run times of both filters for each 24-h encounter in the test data were noted (Table III), and run times of the PHD filter were faster compared to the run times of the BP filter. The number of azitracks in the encounter did not appear to influence the run times in the PHD filter whereas



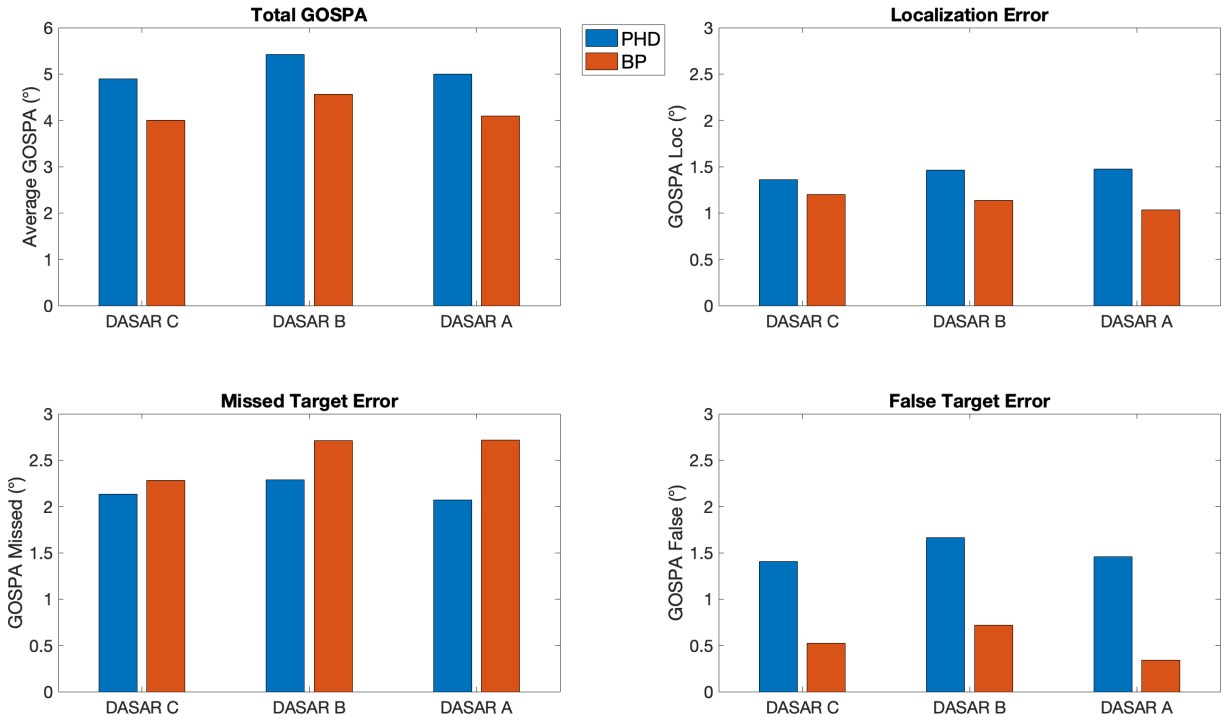



FIG. 6. (Color online) GOSPA error metric and its components for encounter 8, a day with lower numbers of annotated azitracks.

TABLE III. Run times of the PHD and BP filters for each 24-h encounter in the test data (implemented in MATLAB, Release R2022a, on a Mac, Os X, processor 2.7 GHz and 8 GB RAM). The run times and number of annotated azitracks are reported for each DASAR (C/B/A), respectively.

| Encounter | Encounter ID | N annotated azitracks | PHD run time (s) | BP run time (s) |
|---|---|---|---|---|
| 04042020 | 1 | 82 / 73 / 74 | 4.8 / 5.2 / 5.2 | 54.1 / 104.1 / 111.2 |
| 04182020 | 7 | 24 / 20 / 17 | 3.6 / 3.3 / 3.3 | 14.4 / 16.8 / 17.6 |
| 04232020 | 8 | 20 / 24 / 28 | 3.2 / 3.2 / 3.1 | 10.7 / 13.8 / 8.5 |

BP run times were longer when more azitracks were present in a given encounter (Table III). The difference in the run times is due to the BP filter using computationally more intensive particle-based implementation, whereas the PHD filter uses computationally less intensive Gaussian mixture implementation.

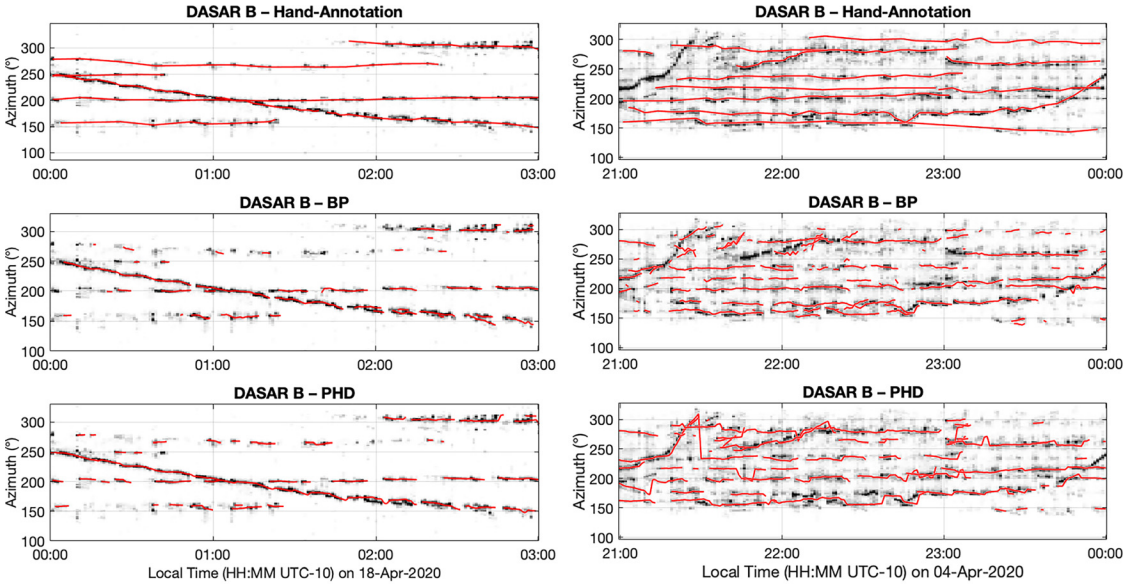
Examples of automatically extracted azitracks, based on a 3-h long excerpt from a day with lower (encounter 7) and a day with higher (encounter 1) numbers of annotated azitracks, are shown in Fig. 7. Additionally, a 2-h long period with higher boat activity is shown in Fig. 8. The boat tracks are typically characterized by a fast rate of change of azimuth over a short period of time, and the filters do not extract the tracks when the rate of change is fast (Fig. 8). However, when the rate of change of azimuth in the boat track is low (e.g., boat track between 15:30 and 16:00 on Fig. 8), then the filters track those parts of the track. Further examples of extracted tracks for all days in the test data are given in the supplementary material.

Next, to further evaluate the automatically extracted azitracks from the test data (Table I), 2D spatial tracks were generated from both the manually annotated and automatically extracted azimuthal tracks following the method described in Sec. III C 1. Figure 9 presents examples of 3 h-long 2D tracks using the data measured on April 18th (encounter 7), the sample example used in Ref. 12. The bias corrections of 6.5°, –4.7°, and −7.6° for DASAR A, B, and C were applied, respectively. Since the biases were not applied in Ref. 12, the tracks in Ref. 12 are different from our tracking results [Fig. 9(a)].

## V. DISCUSSION

Two methods to fully automate the tracking of multiple animals with vector sensors were proposed and investigated in this work. Both methods successfully and accurately tracked multiple singing humpback whales with an array of three-vector sensors.

The biggest contribution to the error metric came from the missed target error, which occurred due to track fragmentation in the automated methods. Track fragmentation is a common challenge in automated bio-acoustic tracking[8] where mobile sources emit intermittent vocalizations. GOSPA metric measures error on an individual time step basis, and a missed target is counted every time step during which an estimate of the target is not made compared to a ground truth. The manual analyst (ground truth) connected all parts of the track, even through periods of prolonged silence (when no measurements were available to the filters). On the other hand, both of the automated approaches tracked through shorter periods of silence during the time




(a) 18 April 2020 (encounter 7) (b) 4 April 2020 (encounter 1)

FIG. 7. (Color online) 3-h excerpts from a day with (a) lower and (b) higher numbers of annotated azitracks. On each, azigrams with overlayed manually extracted tracks [top], automatically extracted tracks with BP [middle] and PHD [bottom] filters are shown. Tracks are shown as solid lines.

https://doi.org/10.1121/10.0021972

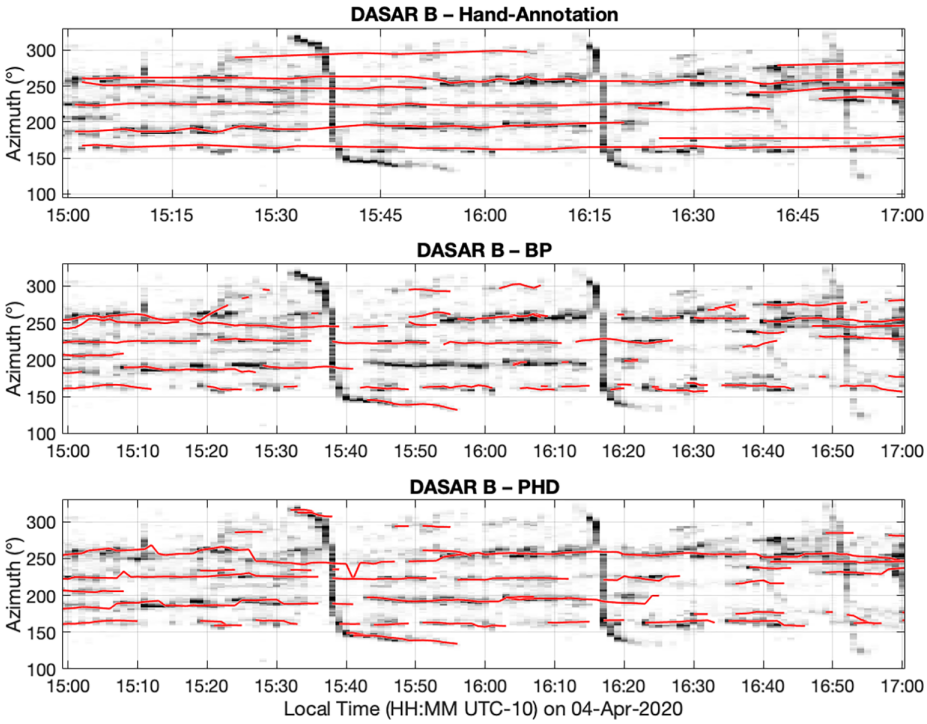



FIG. 8. (Color online) 2-h long example with 3 boat tracks present in addition to humpback whales. Boat tracks show as dark lines with a fast rate of change of azimuth on the azigram. The annotated and automatically extracted tracks of humpback whales are shown as solid lines.

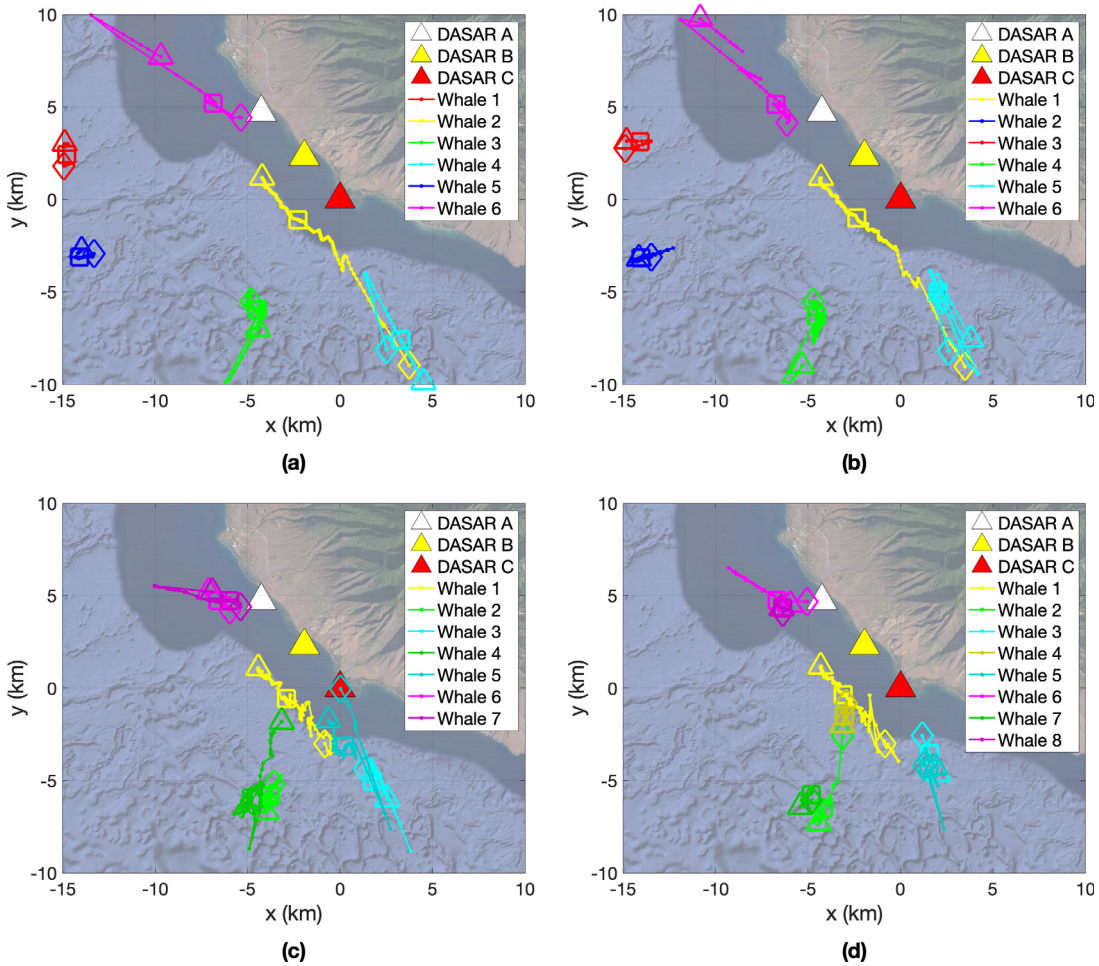



FIG. 9. (Color online) 2D tracks generated from the computed azitracks with (a) hand annotation presented in Ref. 12, (b) hand annotation from another operator, (c) BP and (d) PHD. The triangle, square, and diamond indicate the beginning, median, and end of each track, respectively.

periods when the animal was singing but stopped tracking when longer periods of silence occurred (e.g., when the animal stopped to breathe).

The error was bigger when there was a higher animal track density compared to when there were fewer tracks in the encounter. In the present work, the parameter values were trained as a mean across encounters in the training data; however, it was seen that certain parameters that model clutter (false alarm rate) and new target generation (birth rate) were different between encounters that had higher and lower numbers of tracks. In future work two sets of parameters could be considered- parameters trained for the height of the humpback whale season when many whales sing concurrently, and parameters trained for the beginning/end of the whale season when fewer animals are singing.

Boat tracks sometimes resulted in false tracks, especially when the azimuthal rate of change was low, and thus these tracks resembled those produced by singing whales (Fig. 8). Azimuthal tracking is species-agnostic, i.e., the expected target movement is determined by the state-transition model, and thus any acoustic source with azimuthal changes (i.e., swimming speeds and distances to the sensor) similar to humpback whales would be tracked. A possible solution would be to incorporate additional features or a classifier to resolve this.

While both filters had similar performance, it should be noted that the PHD filter used an amplitude feature while BP did not. Previous research showed a significant improvement in PHD filter's performance when using an amplitude feature compared to the filter without it.[8,22] Using the amplitude feature in the PHD filter helps to improve the cardinality estimates of the filter (i.e., the information about the number of targets in the observation region), which are known to have high variance when number of targets increases.[22] It is likely that addition of the amplitude feature to BP filter would improve its performance,[40] and future studies should consider the development of the BP filter with an amplitude feature for the azigram tracking application. Moreover, since the amplitude model used in the PHD filter was developed based on a different application, a development of a new model that better fits the current data would likely improve the performance of both filters.

In this study, the track-then-localize scheme was used, and although it was not the focus of this study, the 2D spatial tracks were computed and visually compared between manually annotated and automatically extracted tracks. Since the 2D track construction used DASAR B as the main instrument and imposed a minimum 15 min criteria for the tracks to match between the sensor pairs before being considered for the 2D track construction, the results are different when obtained from manually extracted tracks compared to automatically extracted tracks.

The applied 2D track extraction approach is not designed for fragmented tracks of a single whale track. In automatically extracted tracks, fragmentation of individual whale tracks occurs during longer periods of silence. If the fragments are shorter than the 15 min threshold or do not have a concurrent at least 15 min long track on at least two sensors, the track is not considered for 2D spatial track construction. For example, whales 1 and 5 in Fig. 9(a) and 2 and 3 in Fig. 9(b), which were produced from manually annotated tracks, are detected in short fragments by both automated methods [Fig. 7(a)], and, thus do not result in 2D spatial tracks in Figs. 9(c) and 9(d). Furthermore, when there are multiple fragments of the same track automatically extracted, the resulting spatial 2D tracks show as two separate whales in the automated methods—e.g., whale 2 and 4 in Fig. 9(c) and whale 2 and 7 as well as whale 6 and 8 in Fig. 9(d). Finally, fragments of a longer track are matched with a false positive track, yielding a false positive 2D track—e.g., whale 5 and 7 in Fig. 9(c) and whale 4 in Fig. 9(d). Note also that the whale tracks towards the endfire of the linearly arranged vector sensors are less reliable since a small shift in azimuthal tracks will lead to a substantial change in the 2D tracks.

Therefore, future studies should consider a different approach to extracting 2D spatial tracks that would allow for track fragmentation or a second step where 2D tracks would be joined. An alternative approach for matching azimuthal tracks from the same source across multiple DASARs would be to treat the automated azimuthal tracking discussed in this paper as a de-clutter step, then apply either of the two methods again to track in the 2D space.

The run-time of both methods was fast (typically under a minute for 24-h period), with the PHD having shorter run-times compared to BP. The significantly higher run times of BP compared to PHD are related to the fact that we are using a particle-based implementation of BP while PHD relies on a computationally less intensive Gaussian mixture implementation. The particle-based implementation of BP has been developed for multi-target tracking problems with highly nonlinear state-transition models,[27] whereas for humpback whale azitracks a much simpler state-transition model can be used, which the PHD filter takes advantage of.

With some further improvements, the automated tracking methods could lead to a new way of enumerating acoustically humpback whales in high density areas. Being able to accurately count and localize singing humpback whales when chorusing levels are high would lead to new insights into how singing whales distribute themselves and how density dependent singing behavior is among humpback whales. In turn, this would help further advance efforts to establish the relationship between chorusing behavior and overall whale abundance.[41]

## VI. CONCLUSIONS

In this work, two methods to fully automate the tracking of baleen whales with an array of vector sensors were proposed. Both methods are based on the Bayesian statistical framework, but take different approaches to how the posterior distributions are approximated. Both methods were

applied to a dataset in Maui and successfully tracked multiple singing humpback whales.

## SUPPLEMENTARY MATERIAL

See the supplementary material for manually and automatically extracted tracks for all days in test data.

## ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation (NSF) under CAREER Award No. 2146261.

## AUTHOR DECLARATIONS

### Conflict of Interest

Authors have no conflicts of interest to disclose.

## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

[1]L. Thomas and T. A. Marques, "Passive acoustic monitoring for estimating animal density," Acoust. Today **8**(3), 35–44 (2012).
[2]S. W. Martin, C. R. Martin, B. M. Matsuyama, and E. E. Henderson, "Minke whales (*Balaenoptera acutorostrata*) respond to navy training," J. Acoust. Soc. Am. **137**(5), 2533–2541 (2015).
[3]T. A. Helble, E. E. Henderson, G. R. Ierley, and S. W. Martin, "Swim track kinematics and calling behavior attributed to Bryde's whales on the Navy's Pacific Missile Range Facility," J. Acoust. Soc. Am. **140**(6), 4170–4177 (2016).
[4]E. E. Henderson, T. A. Helble, G. Ierley, and S. Martin, "Identifying behavioral states and habitat use of acoustically tracked humpback whales in Hawaii," Mar. Mammal Sci. **34**(3), 701–717 (2018).
[5]A. Thode, "Three-dimensional passive acoustic tracking of sperm whales (*Physeter macrocephalus*) in ray-refracting environments," J. Acoust. Soc. Am. **118**(6), 3575–3584 (2005).
[6]E.-M. Nosal, "Methods for tracking multiple marine mammals with wide-baseline passive acoustic arrays," J. Acoust. Soc. Am. **134**(3), 2383–2392 (2013).
[7]A. I. DeAngelis, R. Valtierra, S. M. Van Parijs, and D. Cholewiak, "Using multipath reflections to obtain dive depths of beaked whales from a towed hydrophone array," J. Acoust. Soc. Am. **142**(2), 1078–1087 (2017).
[8]P. Gruden, E.-M. Nosal, and E. Oleson, "Tracking time differences of arrivals of multiple sound sources in the presence of clutter and missed detections," J. Acoust. Soc. Am. **150**(5), 3399–3416 (2021).
[9]C. R. Greene, M. W. McLennan, R. G. Norman, T. L. McDonald, R. S. Jakubczak, and W. J. Richardson, "Directional frequency and recording (DIFAR) sensors in seafloor recorders to locate calling bowhead whales during their fall migration," J. Acoust. Soc. Am. **116**(2), 799–813 (2004).
[10]A. M. Thode, K. H. Kim, S. B. Blackwell, C. R. Greene, Jr, C. S. Nations, T. L. McDonald, and A. M. Macrander, "Automated detection and localization of bowhead whale sounds in the presence of seismic airgun surveys," J. Acoust. Soc. Am. **131**(5), 3726–3747 (2012).
[11]D. V. Harris, J. L. Miksis-Olds, J. A. Vernon, and L. Thomas, "Fin whale density and distribution estimation using acoustic bearings derived from sparse arrays," J. Acoust. Soc. Am. **143**, 2980–2993 (2018).
[12]L. Tenorio-Hallé, A. M. Thode, M. O. Lammers, A. S. Conrad, and K. H. Kim, "Multi-target 2D tracking method for singing humpback whales using vector sensors," J. Acoust. Soc. Am. **151**(1), 126–137 (2022).
[13]J. Jang, F. Meyer, E. R. Snyder, S. M. Wiggins, S. Baumann-Pickering, and J. A. Hildebrand, "Bayesian detection and tracking of odontocetes in 3-D from their echolocation clicks," J. Acoust. Soc. Am. **153**(5), 2690–2705 (2023).
[14]P. M. Baggenstoss, "A multi-hypothesis tracker for clicking whales," J. Acoust. Soc. Am. **137**(5), 2552–2562 (2015).
[15]F. Meyer, T. Kropfreiter, J. L. Williams, R. A. Lau, F. Hlawatsch, P. Braca, and M. Z. Win, "Message passing algorithms for scalable multitarget tracking," Proc. IEEE **106**(2), 221–259 (2018).
[16]B.-N. Vo, M. Mallick, Y. Bar-Shalom, S. Coraluppi, R. Osborne III, R. Mahler, and B.-T. Vo, "Multitarget tracking," in *Wiley Encyclopedia of Electrical and Electronics Engineering* (John Wiley and Sons, Inc., New York, 2015), pp. 1–23.
[17]Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation* (John Wiley & Sons Inc., New York, 2001).
[18]R. S. Payne and S. McVay, "Songs of humpback whales: Humpbacks emit sounds in long, predictable patterns ranging over frequencies audible to humans," Science **173**, 585–597 (1971).
[19]R. P. Mahler, *Statistical Multisource-Multitarget Information Fusion* (Artech House, Inc., Norwood, MA, 2007).
[20]R. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," IEEE Trans. Aerosp. Electron. Syst. **39**(4), 1152–1178 (2003).
[21]B.-N. Vo and W.-K. Ma, "The Gaussian mixture probability hypothesis density filter," IEEE Trans. Signal Process. **54**(11), 4091–4104 (2006).
[22]D. Clark, B. Ristic, B.-N. Vo, and B. T. Vo, "Bayesian multi-object filtering with amplitude feature likelihood for unknown object SNR," IEEE Trans. Signal Process. **58**(1), 26–37 (2010).
[23]B. Ristic, D. Clark, B.-N. Vo, and B.-T. Vo, "Adaptive target birth intensity for PHD and CPHD filters," IEEE Trans. Aerosp. Electron. Syst. **48**(2), 1656–1668 (2012).
[24]S. M. Bozic, *Digital and Kalman Filtering* (Edward Arnold Ltd., London, UK, 1979).
[25]K. Panta, D. E. Clark, and B.-N. Vo, "Data association and track management for the Gaussian mixture probability hypothesis density filter," IEEE Trans. Aerosp. Electron. Syst. **45**(3), 1003–1016 (2009).
[26]P. Gruden and P. R. White, "Automated tracking of dolphin whistles using Gaussian mixture probability hypothesis density filters," J. Acoust. Soc. Am. **140**(3), 1981–1991 (2016).
[27]F. Meyer, P. Braca, P. Willett, and F. Hlawatsch, "A scalable algorithm for tracking an unknown number of targets using multiple sensors," IEEE Trans. Signal Process. **65**(13), 3478–3493 (2017).
[28]F. Meyer and J. L. Williams, "Scalable detection and tracking of geometric extended objects," IEEE Trans. Signal Process. **69**, 6283–6298 (2021).
[29]J. L. Williams and R. Lau, "Approximate evaluation of marginal association probabilities with belief propagation," IEEE Trans. Aerosp. Electron. Syst. **50**(4), 2942–2959 (2014).
[30]F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," IEEE Trans. Inf. Theory **47**(2), 498–519 (2001).
[31]D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, MA, 2009).
[32]H. V. Poor, *An Introduction to Signal Detection and Estimation* (Springer, New York, 1994).
[33]R. A. Holler, "The evolution of the sonobuoy from World War II to the Cold War," U.S. Navy J. Underwater Acoust. **62**, 322–347 (2014).
[34]J. A. Mann III, J. Tichy, and A. J. Romano, "Instantaneous and time-averaged energy transfer in acoustic fields," J. Acoust. Soc. Am. **82**(1), 17–30 (1987).
[35]A. M. Thode, T. Sakai, J. Michalec, S. Rankin, M. S. Soldevilla, B. Martin, and K. H. Kim, "Displaying bioacoustic directional information from sonobuoys using 'azigrams,' " J. Acoust. Soc. Am. **146**(1), 95–102 (2019).
[36]G. L. D'Spain, W. S. Hodgkiss, and G. L. Edmonds, "Energetics of the deep ocean's infrasonic sound field," J. Acoust. Soc. Am. **89**(3), 1134–1158 (1991).
[37]H. W. Kuhn, "The Hungarian method for the assignment problem," Naval Res. Logistics Quart. **2**(1–2), 83–97 (1955).
[38]R. V. Lenth, "On finding the source of a signal," Technometrics **23**(2), 149–154 (1981).
[39]A. S. Rahmathullah, A. F. Garcia-Fernandez, and L. Svensson, "Generalized optimal sub-pattern assignment metric," in *Proceedings of the 2017 20th International Conference on Information Fusion (FUSION)*, Xi'an, China (July 10–13, 2017), pp. 1–8.
[40]X. Li, E. Leitinger, A. Venus, and F. Tufvesson, "Sequential detection and estimation of multipath channel parameters using belief propagation," IEEE Trans. Wireless Commun. **21**(10), 8385–8402 (2022).
[41]A. Kügler, M. O. Lammers, E. J. Zang, and A. A. Pack, "Male humpback whale chorusing in Hawai'i and its relationship with whale abundance and density," Front. Mar. Sci. **8**, 735664 (2021).