Final Report

Manual and Automated Atlantic Whistle Classifiers: Improvement, Testing and Application

Submitted to:

Naval Facilities Engineering Command Atlantic under
HDR Environmental, Operations and Construction, Inc.
Contract No. N62470-10-D-3011, CTO 021

Prepared by:

Bio-Waves, Inc.
364 2nd Street, Suite #3
Encinitas, CA 92024
(760) 452-2575

Submitted by:

Virginia Beach, VA

November 2016

Suggested Citation:

Oswald, J.N., R. Walker, C. Hom-Weaver, and T.F. Norris. 2016. Manual and automated Atlantic whistle classifiers: improvement, testing and application. Draft Final report. Submitted to HDR Environmental, Operations and Construction, Inc., Norfolk, Virginia, under Contract No. N62470-10-D-3011 CTO 021. Prepared by Bio-Waves, Inc., Encinitas, California.

Cover Photo:

Atlantic spotted dolphins *(Stenella frontalis)* taken by Heather Foley, Duke University. Photo taken under National Oceanic and Atmospheric Administration Permit No. 16185.

Work conducted under following contract between Bio-Waves, Inc. and HDR, Inc.
MSA #: CON-005-4394-009
Subproject #164744, Agreement #105067, CTO 021, Task 004.

This project is funded by US Fleet Forces Command and managed by Naval Facilities Engineering Command Atlantic as part of the US Navy's marine species monitoring program.

Executive Summary

In 2013, two random-forest classifiers were developed to identify the whistles of five species of odontocetes recorded in the northwestern Atlantic Ocean (bottlenose dolphin, *Tursiops truncatus*; Atlantic spotted dolphin, *Stenella frontalis*; striped dolphin, *S. coeruleoalba*; short-beaked common dolphin, *Delphinus delphis*; and short-finned pilot whale, *Globicephala macrorhynchus*; Oswald 2013). One of these classifiers (the manual classifier) was trained and tested using whistle contours that were detected and extracted using manual methods in the Real-time Odontocete Call Classification Algorithm (ROCCA) module (Oswald et al. 2013) within the acoustic data processing software platform, PAMGuard (Gillespie et al. 2008). The second classifier (the automated classifier) was trained and tested using whistles detected and extracted automatically using the Whistle and Moan Detector (WMD) module in PAMGuard. When both of these classifiers were tested using four-fold cross validation, 86 percent and 91 percent of encounters were correctly classified for the manual classifier and the automated classifier, respectively. Since the initial development of these classifiers, additional visual and acoustic shipboard surveys have occurred in the northwest Atlantic Ocean through the Atlantic Marine Assessment Program for Protected Species (AMAPPS) 2013. This new data allowed biologists to test the performance of the existing classifiers, add new species to the classifiers, and use the manual classifier to identify AMAPPS 2013 acoustic encounters that did not have visual confirmation of species identity.

The AMAPPS 2013 survey was a visual and acoustic line-transect marine mammal abundance survey that was conducted from 1 July to 15 September 2013 by the Southeast Fisheries Science Center (SEFSC) and the Northeast Fisheries Science Center (NEFSC). These surveys covered waters of the northern Atlantic continental shelf-break, from the 100-meter depth contour to the edge of the Exclusive Economic Zone, and ranged from South Carolina in the south to the southern tip of Nova Scotia, Canada, in the north. Passive acoustic data were collected using towed hydrophone arrays (NEFSC and SEFSC 2013).

Whistles were detected and extracted manually from the AMAPPS 2013 passive acoustic dataset using the Raven Pro Software (Version 1.4; Bioacoustics Research Program 2011), and the PAMGuard ROCCA module. They were also detected and extracted automatically using PAMGuard's WMD module. Acoustic encounters with visual confirmation of species identity were classified using both the manual and automated classifier approaches. Recordings for four out of the five species that were included in the Atlantic classifiers (with the exception of short-finned pilot whales) were available. Small sample sizes made it difficult to evaluate classifier performance for short-beaked common dolphins (n=3 acoustic encounters) and striped dolphins (n=2 acoustic encounters). The manual classifier performed relatively well, correctly classifying 77 percent of Atlantic spotted dolphin encounters (n=13) and 93 percent of bottlenose dolphin encounters (n=28). The automated classifier misclassified every encounter as pilot whales. This is likely due to the fact that different versions of PAMGuard were used to detect and extract whistles for the Atlantic classifier training dataset and the AMAPPS 2013 test dataset. Determining the exact cause of these discrepancies within PAMGuard was beyond the scope of this project but should be investigated further.

Addition of the AMAPPS 2013 dataset allowed three species (Risso's dolphin, *Grampus griseus*; rough-toothed dolphin, *Steno bredanensis*; and Clymene dolphin, *Stenella clymene*) to be added to the manual classifier. The new classifier had an overall correct classification score of 55 percent and individual species correct classification scores that ranged from 3 percent for Risso's dolphins to 84 percent for short-finned pilot whales. Although the correct classification score for Risso's dolphin was very low, this species was included in the classifier because few encounters from other species were misclassified as Risso's dolphin, and when a classifier that did not include Risso's dolphin was trained, correct classification scores for the other species were similar to correct classification results using a classifier with Risso's dolphin included. Including Risso's dolphin in the classifier provides the potential for Risso's dolphin encounters to be classified without significantly reducing correct classification scores for other species. Adding other information such as echolocation click measurements may increase correct classification scores for Risso's dolphin and other species, and this is being pursued by Bio-Waves, Inc. in a separate project sponsored by the Office of Naval Research (ONR) and the United States Navy Living Marine Resources Program.

This new eight-species classifier was used to identify AMAPPS 2013 encounters that did not have visual confirmation of species identity. Most (18 out of 20) of these encounters were classified as striped dolphins. The remaining two encounters were classified as Clymene dolphin and short-finned pilot whale. Striped dolphins were one of the most commonly detected small cetaceans during the northern leg of the AMAPPS 2013 survey, both visually and acoustically (NEFSC and SEFSC 2013). All but two of the non-sighted acoustic encounters that were classified as striped dolphins were north of 36°N and offshore of the continental shelf, which is where all of the visual detections of striped dolphins also occurred. Based on their locations, we believe that the two southernmost non-sighted acoustic encounters that were classified as striped dolphins may have been misclassifications. This suggests that geographic location may be another variable that could be useful for improving classification success of the classifiers. This possibility is currently being investigated by Bio-Waves, Inc. in the ONR-sponsored project mentioned above.

The results of this study provide valuable information on the performance of two whistle classifiers for delphinid species in the northwestern Atlantic Ocean. Although the manual classifier requires more time and effort for the detection and extraction of whistles, it proved to be more generalizable to novel datasets than the automated classifier and also allowed identification of acoustic encounters that did not have associated visual observations (i.e., non-sighted encounters). The ability to identify non-sighted encounters allows a more complete understanding of species distribution as well as providing information about which species are more difficult to detect using visual methods. These results highlight the complementary nature of visual and acoustic methods, which if used together allow more and improved information about the distribution of marine mammals to be collected from vessel-based surveys.

Table of Contents

Appendix

Figures

Tables

Abbreviations and Acronyms

AMAPPS	Atlantic Marine Assessment Program for Protected Species
dB	decibel(s)
DTAG	digital acoustic recording tag
kHz	kilohertz
LMR	Living Marine Resources program
m	meters
NEFSC	Northeast Fisheries Science Center
ONR	Office of Naval Research
PAM	passive acoustic monitoring
ROCCA	Real-time Odontocete Call Classification Algorithm
SD	standard deviation
SEFSC	Southeast Fisheries Science Center
V/µPa	volt(s) per 1 microPascal
WMD	Whistle and Moan Detector

This page intentionally left blank.

1. Introduction

In recent decades, passive acoustic monitoring (PAM) has been adopted as an effective method for obtaining information about the occurrence, distribution, and behavior of marine mammals (Mellinger and Barlow 2003); however, this type of monitoring generates huge volumes of data. In order for the data generated from PAM to be effectively used, they need to be efficiently analyzed and accurately interpreted. Researchers have been turning to bio-acoustic analysis software for the detection and classification of marine mammal sounds from digital acoustic recordings. Due to the high variability of sounds both within and among species, identification of marine mammal species based on sounds can be challenging. This is especially true for delphinid whistles, which are among the most variable of calls for any group of cetaceans.

Early delphinid whistle classifiers focused on time-frequency characteristics measured from spectrograms and classification algorithms such as discriminant-function analysis and classification-tree analysis (e.g., Steiner 1981, Fristrup and Watkins 1993, Wang et al. 1995, Matthews et al. 1999, Rendell et al. 1999, Oswald et al. 2003). More recently, other classification algorithms such as Gaussian mixture models (Roch et al. 2007), Hidden Markov models (Brown and Smaragdis 2009) and random forests (Oswald et al. 2013) have been used, with varying degrees of success.

The Real-time Odontocete Call Classification Algorithm (ROCCA) is one of a few classifiers that are readily available for marine mammal researchers, conservationists, and resource managers (Oswald et al. 2013). At present, ROCCA is available as a module within PAMGuard, an open-source software platform that is freely available to the public for the recording, processing, and analysis of bioacoustic data (www.pamguard.org; Gillespie et al. 2008). Initially, ROCCA contained a random-forest classifier that was developed for whistles from eight different species of delphinids occurring in the tropical Pacific Ocean (Oswald et al. 2013), but in October 2013, Bio-Waves, Inc. completed development of two additional whistle classifiers (Oswald 2013). These two classifiers each included five species of delphinids (bottlenose dolphins, *Tursiops truncatus*; Atlantic spotted dolphins, *Stenella frontalis*; striped dolphins, *S. coeruleoalba*; short-beaked common dolphins, *Delphinus delphis*; and short-finned pilot whales, *Globicephala macrorhynchus*) recorded in the northwest Atlantic Ocean. The first classifier (the manual classifier) was trained and tested using whistles detected and extracted using manual methods in ROCCA. The second classifier (the automated classifier) was trained and tested using whistles detected and extracted using the fully automated Whistle and Moan Detector (WMD), a module in PAMGuard. Both the manual and the automated classifiers identify individual whistles to species. Encounters (groups of whistles produced by a single school of dolphins) are then identified based on the combined classification results for all of the whistles in each encounter (Oswald 2013).

ROCCA's Atlantic classifiers are random-forest classifiers that use a two-stage approach, where whistles are first classified to broad species categories (e.g., large delphinids, *Stenella* species) in stage one and then to species within those categories in stage two. This approach resulted in more accurate classification scores than previous single-stage random-forest analyses. Overall, 66 percent (manual classifier) and 68 percent (automated classifier) of encounters were

correctly classified using a single-stage random forest, and 86 percent (manual classifier) and 91 percent (automated classifier) of encounters were correctly classified using a two-stage approach (Oswald 2013). Because of the high correct classification scores obtained using test data, the Atlantic classifier has the potential to be an important element in the marine mammal acoustic signal-processing toolbox. It is therefore important to continue developing and using this classifier.

During the months of July through September 2013 the Northeast Fisheries Science Center (NEFSC) and the Southeast Fisheries Science Center (SEFSC) conducted a combined visual and acoustic survey for marine mammals in the northwestern Atlantic Ocean (Atlantic Marine Assessment Program for Protected Species—AMAPPS). In this study, passive acoustic data collected during the AMAPPS 2013 cruise were used to test, improve, and utilize the ROCCA Atlantic classifiers.

The main goals of this study were as follows:

1. **Continue development of the Atlantic classifiers by adding AMAPPS 2013 recordings of single-species schools to the training datasets.** ROCCA's Atlantic classifier currently includes five species; however, archival recordings are available from SEFSC and NEFSC for at least five additional species (spotted dolphins, *Stenella attenuata*; rough-toothed dolphins, *Steno bredanensis*; Clymene dolphins, *Stenella clymene*; false killer whales, *Pseudorca crassidens*; Risso's dolphins, *Grampus griseus*). During the initial development of the Atlantic classifier, there were not enough data available to include these species, but single-species, visually validated recordings collected during the AMAPPS 2013 survey provide enough data to allow the addition of some of these species to the classifier.
2. **Test and ground-truth the Atlantic classifiers using whistles recorded during visually validated acoustic recordings from the AMAPPS 2013 cruise to provide a more complete understanding of how the classifiers perform on novel data.**
3. **Use the Atlantic classifier to identify schools that were detected acoustically but did not have visual confirmation of species identity.** The identification of schools that were detected acoustically but did not have visual confirmation of species identity will provide a more complete understanding of species occurrence and distribution in the AMAPPS study area.

The methods used to obtain these goals are presented and discussed below.

2. Statement of Navy Relevance

The northwestern Atlantic Ocean contains several regions that are important marine areas for U.S. Navy training and testing. In compliance with the Marine Mammal Protection Act, the United States Navy is required to monitor and assess the impact of training and testing activities on marine mammals during such exercises. However, species-specific information for delphinids is often not available for these regions. PAM is being used extensively to collect information regarding marine mammal occurrence, distribution, and behavior in naval exercise areas; however, it is currently not possible to identify many delphinids to species based on acoustic data alone. It is important to be able to identify species for multiple reasons. Naval activities such as sonar and ship noise may have negative impacts on marine mammals and different species may react in different ways. Knowledge about which species are present is required to assess species-specific responses to naval activities, which will in turn improve resource management plans and allow for more effective mitigation measures. Development of efficient and accurate tools for detection and classification of sounds produced by marine mammals will reduce the need for the Navy to train human operators and analysts and will significantly reduce the amount of time needed to analyze recordings made using towed hydrophone arrays or seafloor-mounted recorders. In general, the ability to analyze passive acoustic data more efficiently will reduce costs and allow the Navy to examine results and make decisions in a more effective and timely manner.

This page intentionally left blank.

3. Methods

3.1 Data Collection

Data used to test, improve, and implement the ROCCA Atlantic classifier were collected during shipboard line-transect marine mammal abundance surveys (AMAPPS) conducted by SEFSC and NEFSC from 1 July to 15 September 2013 on the National Oceanic and Atmospheric Administration ships *Henry B. Bigelow* (NEFSC) and *Gordon Gunter* (SEFSC). These surveys covered northern Atlantic continental shelf-break waters, from the 100-meter (m) depth contour to the edge of the Exclusive Economic Zone. The NEFSC survey took place from 1 July to 19 August and ranged from South Carolina to Massachusetts. The SEFSC survey took place from 13 July to 15 September and ranged from South Carolina to Virginia (**Figure 1**). Acoustic recordings were collected with an array of oil-filled hydrophones towed approximately 300 m behind the research vessel at a depth of 8 to 12 m. The NEFSC array consisted of six APC International hydrophone elements (model 42-1021) and two Reson hydrophone elements (model TC 4013). The SEFSC array consisted of three APC International elements (model 42-1021) and two Reson elements (model TC 4013). The APC hydrophones have a sensitivity of -212 decibels (dB) referenced to 1 volt per microPascal (V/µPa) with a flat frequency response (±4 dB) from 1 to 45 kilohertz (kHz) and a custom-built pre-amplifier gain of 45 dB above 5 kHz. The Reson hydrophones have a sensitivity of -212 dB re 1 V/µPa with a flat frequency response (±2 dB) from 5 to 160 kHz with a custom-built pre-amplifier gain of 50 dB above 5 kHz. Acoustic data were recorded to computer hard drives with sampling rates of 192 kHz and 500 kHz with a 1-kHz high-pass filter using PAMGuard software (Gillespie et al. 2008). Acoustically active delphinid schools were localized using target motion analysis methods in Ishmael software (Mellinger 2001), and this information was used to match acoustic detections with visual detections in real-time for species identification.

3.2 Whistle Measurement

Both manual and automated whistle detection, extraction and classification methods were used to analyze 1) whistles from recordings of single-species delphinid schools that had visual confirmation of species identity, and 2) whistles from recordings that did not have associated visual observations. Recordings from sighted or non-sighted schools will henceforth be referred to as acoustic encounters.

3.2.1 Manual Detection and Contour Extraction

Recordings from each acoustic encounter included in the analysis were first examined aurally and visually using Raven Pro: Interactive Sound Analysis Software (Version 1.4; Bioacoustics Research Program 2011). Analysts selected whistles for measurement that had clear start and end times, and for which the entire time-frequency contour was visible on the spectrogram. Overlapping whistles were included only if each contour could be traced unambiguously. Analysts randomly selected up to 50 whistles from each encounter and saved each as an individual .wav file. A maximum of 50 whistles was selected from each encounter to avoid over-sampling of groups or individuals.

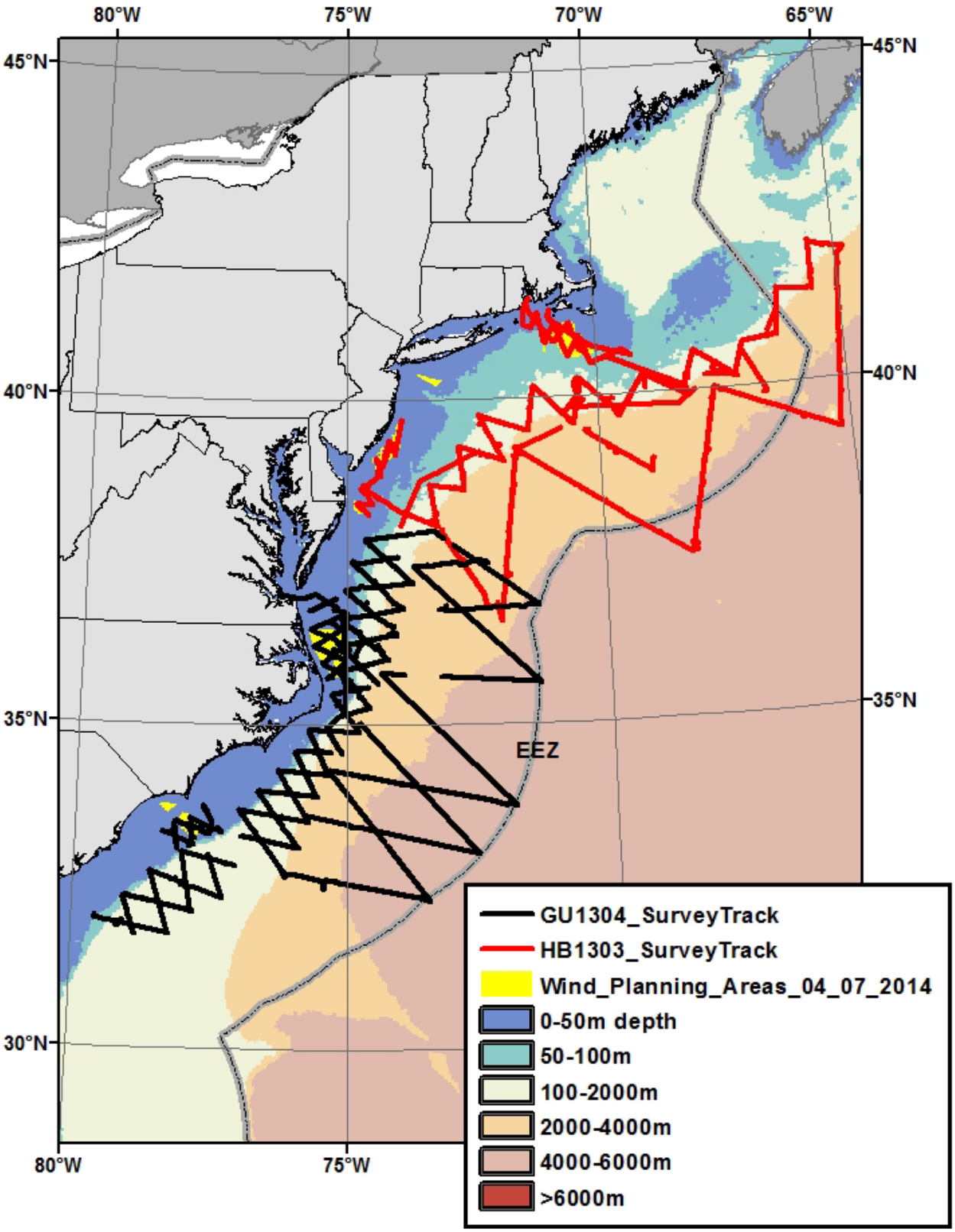

Figure 1. Tracklines completed during the summer (July–September) 2013 AMAPPS shipboard line-transect surveys (figure from NEFSC and SEFSC 2013). GU1304_Survey track (black line) was for the Research Vessel ***Gordon Gunther*** and the HB1303_Survey track (red line) was for the Research Vessel ***Henry B Bigelow***.

To extract time-frequency contours from selected whistles, analysts traced contours on ROCCA's spectrographic display using a computer touch-pad. ROCCA automatically extracted time-frequency contours based on whistle traces and then measured 50 variables from each extracted contour. Measured variables included: duration, frequencies (e.g., minimum, maximum, beginning, ending, and at various points along the whistle), slopes, and variables describing shape of the whistles (e.g., number of inflection points and steps; see Barkley et al. 2011 for a complete list and description of variables measured).

3.2.2 Automated Detection and Contour Extraction

Recordings from each acoustic encounter were also analyzed using automated methods. Automated whistle detection and contour extraction were performed using the WMD module within PAMGuard (Gillespie et al. 2008). The WMD automatically detects and extracts whistle contours by searching for and connecting sequential spectral peaks within a user-specified frequency band. In order to be considered a true whistle detection, a spectral peak must occur within certain user-defined parameters relating to its amplitude and frequency in relation to other candidate spectral peaks detected in the time-slices directly before and after the peak in question. For each acoustic encounter, parameters within the WMD module were adjusted manually via a graphical user interface to maximize accuracy of contour extraction and minimize false detections. The WMD automatically passed extracted contours to the ROCCA module for measurement of whistle variables, as described above.

3.3 Random-Forest Classification Analysis

Whistle contours were classified to species using random-forest classifiers. A random forest is a collection of decision trees grown using binary partitioning of the data. Each binary partition of the data is based on the value of one whistle variable (Breiman 2001). An excellent graphical depiction of the process is presented in Nguyen et al (2013). Randomness is introduced into the tree-growing process by examining a random subsample of all of the variables at each node. The variable that produces the most homogeneous split is chosen at each partition. When whistle variables are run through a random forest, each of the trees in the forest produces a species classification. Each tree can be considered one vote for a given species classification. Votes are then tallied over all trees and the final whistle classification is based on the species with the most votes. In addition to classifying individual whistles, acoustic encounters are classified based on the number of tree classifications for each species, summed over all of the whistles that were analyzed for that encounter.

Two random forest classifiers were used to analyze the AMAPPS 2013 data. These classifiers were trained and tested in a previous effort (Oswald 2013) using acoustic data recorded from single-species schools that had visual confirmation of species identity and were independent of the AMAPPS 2013 data. Acoustic recordings used in these training and testing datasets were made using towed hydrophone arrays during vessel-based visual and acoustic line transect surveys conducted by SEFSC, NEFSC, and Duke University. The surveys took place off the Atlantic coast of the United States between central Florida and Georges Bank (in the Gulf of Maine). Duke University researchers also provided acoustic data recorded with Digital Acoustic Recording Tags (DTAGs, Johnson and Tyack 2003) attached to short-finned pilot whales.

Recordings from DTAGs were used only if the tagged animal was part of a single-species school of short-finned pilot whales and if there were no other species that whistle sighted within 3 nautical miles. A more detailed description of the data and classifiers can be found in Oswald (2013). One of the classifiers was trained using variables measured from whistles extracted using ROCCA's manual method and the other was trained using variables measured from whistle contours extracted automatically using the WMD. Both classifiers were tested using two-fold cross-validation. In two-fold cross-validation, the training dataset is randomly divided in two, with whistles from the same encounter kept together in the same dataset. One dataset is used to train the model, while the other is used to test the model. The datasets are then swapped so that each is used once for both training and testing the model. This procedure was repeated 10 times in order to produce means and standard deviations for confusion matrices.

Both classifiers included five species: short-beaked common dolphin, short-finned pilot whale, striped dolphin, Atlantic spotted dolphin, and bottlenose dolphin. Both classifiers were also two-stage classifiers, where whistle contours were first classified to broad species-groups in stage one and then classified to species within those species-groups in stage two. For the manual classifier, whistles were classified as small dolphins (short-beaked common and striped) or large dolphins (Atlantic spotted, bottlenose, and short-finned pilot whales) in stage one. For the automated classifier, stage one consisted of a pilot whale versus dolphin (short-beaked common, Atlantic spotted, striped, and bottlenose) classifier. When the manual classifier was tested using two-fold cross-validation, 86 percent of encounters (n=131) were classified correctly (**Table 1**), and when the automated classifier was tested using two-fold cross-validation, 91 percent of encounters (n=117) were classified correctly (**Table 2**; Oswald 2013).

Table 1. Confusion matrix for two-stage classifier trained using manually detected and extracted whistles and used to classify encounters recorded during the AMAPPS 2013 survey. The percentage of encounters correctly classified for each species is in bold with standard deviations (SD) in parentheses. Overall, 86 percent (SD=2.5 percent) of encounters were correctly classified (from Oswald 2013).

Actual species	% Classified as					
	Short-beaked common dolphin	**Short-finned pilot whale**	**Striped dolphin**	**Atlantic spotted dolphin**	**Bottlenose dolphin**	***n***
Short-beaked common dolphin	**84.6 (5.7)**	0 (0)	0 (0)	11 (0)	4.4 (5.7)	9
Short-finned pilot whale	0 (0)	**94.6 (1.9)**	5.4 (1.9)	0 (0)	0 (0)	16
Striped dolphin	0 (0)	0 (0)	**91.1 (2.8)**	8.0 (0)	0.8 (2.5)	12
Atlantic spotted dolphin	0.3 (0.9)	2.5 (2.6)	4.8 (3.1)	**90 (6.6)**	2.5 (2.6)	37
Bottlenose dolphin	5.4 (2.2)	2.7 (1.6)	14.7 (4.2)	7.2 (2.9)	**70 (4.3)**	57

Table 2. Confusion matrix for two-stage classifier trained using automatically detected and extracted whistles and used to classify encounters recorded during the AMAPPS 2013 survey. The percentage of encounters correctly classified for each species is in bold, with SD in parentheses. Overall, 91 percent (SD=2.5 percent) of encounters were correctly classified (from Oswald 2013).

Actual species	% Classified as					
	Short-beaked common dolphin	**Short-finned pilot whale**	**Striped dolphin**	**Atlantic spotted dolphin**	**Bottlenose dolphin**	***n***
Short-beaked common dolphin	**95.2 (6.2)**	5.2 (6.7)	0 (0)	0 (0)	0 (0)	8
Short-finned pilot whale	0 (0)	**95.2 (2.5)**	0 (0)	4.8 (2.5)	0 (0)	17
Striped dolphin	0 (0)	12.1 (7.8)	**87.9 (7.8)**	0 (0)	0 (0)	11
Atlantic spotted dolphin	0 (0)	8.9 (5.3)	0 (0)	**89.7 (5.8)**	1.2 (1.5)	32
Bottlenose dolphin	0.8 (1.4)	5.6 (2.5)	0.4 (0.8)	4.1 (2.0)	**88.9 (2.2)**	49

Acoustic encounters that had visual confirmation of species identity were classified using the manual and automated classifiers to test the performance of the classifiers on a novel dataset. Acoustic encounters that did not have visual confirmation of species identity were classified using the manual classifier. Only encounters that were at least 3 nautical miles away from visual or acoustic detections of other whistling species were included in the analysis to reduce the probability of obtaining mixed species group recordings.

3.3.1 Development of New Classifiers

All whistles measured manually from AMAPPS 2013 recordings that had visual confirmation of species identity were added to the manual classifier training dataset and new classifiers were trained and tested. To test the classifiers, the data were first subsampled so that sample sizes were equal for all species. This avoided any one species dominating the data and skewing the results. The subsampled dataset was then randomly divided into four subsets, with whistles from the same encounter kept together in the same subset. Three subsets were then used to train the random forest, while one subset was used to test the model. This was repeated for all possible combinations of the four subsets, so that each whistle was used as a training whistle and as a testing whistle. The results from all four subsets were compiled into a confusion matrix, and the entire process was repeated 100 times to obtain means and standard deviations for the confusion matrix.

Several random-forest models were trained and tested. In the first, whistle contours were classified directly to species. Subsequent models were two-stage models, where whistles were classified to species group (e.g., large delphinid, small delphinid) in stage one and then classified to species within each species group in stage two. Species groupings that were tested in stage one include: pilot whales/rough-toothed dolphins, pilot whales/Risso's dolphins, striped/Clymene/common dolphins, and bottlenose/spotted dolphins, among others. The

classifier that produced the highest correct classification scores was used to identify AMAPPS 2013 acoustic encounters that did not have visual confirmation of species identity. This classifier will be added to ROCCA's classifier toolbox in PAMGuard.

3.3.2 Certainty Scores

Each encounter classification from the sighted and non-sighted AMAPPS 2013 datasets was assigned a certainty score on a scale of one to five. This score reflected the degree of confidence that could be placed in the classification, with a score of one being the least confident and a score of five being the most confident. Certainty scores were based on the following criteria:

1. At least five whistles were included in the classification decision
2. At least 35 percent of trees voted for the predicted species for the five-species classifier and 25 percent voted for the predicted species for the eight-species classifier (the classifier containing additional species from the AMAPPS 2013 data, see **Section 4.3.1**). These values were chosen relative to the percentage of trees that would be expected to vote for the predicted species by chance alone (20 percent for a five-species classifier and 12.5 percent for an eight-species classifier). If the percentage of trees voting for the predicted species is substantially higher than would be expected by chance, it is likely that the school is being classified based on real differences in the whistles.
3. No other species had a similar percentage of tree votes (within 5 percent).

Classifications were assigned the following certainty scores:

1: None of the above conditions met

2: Condition 2 and/or 3 met, condition 1 not met

3: Only condition 1 met

4: Conditions 1 and 2 or 3 met

5: Conditions 1, 2, and 3 met.

4. Results

4.1 Whistle Measurements

ROCCA's manual methods was used to measure 2,416 whistles from 64 encounters during the AMAPPS 2013 cruises that had visual confirmation of species identity (**Table 3**). These whistles were added to the original Atlantic classifier dataset, which consisted of 174 encounters and 3,525 whistles (**Table 4**). The WMD was used to automatically detect and measure 54,037 whistles from 62 encounters during the AMAPPS 2013 cruises that had visual confirmation of species identity (**Table 3**). An additional 441 whistles were measured manually using ROCCA from 20 acoustic encounters that did not have visual confirmation of species identity (**Table 3**).

Table 3. Number of acoustic encounters and whistles with visual confirmation of visual identity measured manually and using automated methods from the AMAPPS 2013 dataset. Unidentified encounters are those that did not have visual confirmation of species identity.

Species	Manual		Automated	
	Number of Encounters	Number of Whistles	Number of Encounters	Number of Whistles
Short-beaked common dolphin	3	59	3	179
Risso's dolphin	2	50	2	283
Pantropical spotted dolphin	2	61	2	2,560
Rough-toothed dolphin	4	176	4	2,832
Striped dolphin	2	72	2	348
Clymene dolphin	3	136	3	15,956
Atlantic spotted dolphin	13	510	13	5,745
Bottlenose dolphin	35	1,350	35	25,900
Unidentified	20	441	n/a	n/a
Total	**84**	**2,855**	**62**	**54,037**

Table 4. Number of encounters and whistles measured manually and using automated methods, for the original Atlantic classifier dataset (Oswald 2013), before adding new species.

Species	Manual		Automated	
	Number of Encounters	Number of Whistles	Number of Encounters	Number of Whistles
Short-beaked Common Dolphin	9	249	18	1,952
Risso's Dolphin	10	119	8	160
Short-Finned Pilot Whale	16	256	17	25,697
False Killer Whale	2	70	2	255
Pantropical spotted dolphin	1	3	NA	NA
Rough-toothed Dolphin	3	98	3	510
Striped Dolphin	12	293	11	2,011
Clymene Dolphin	2	99	2	87
Atlantic Spotted Dolphin	45	706	40	6,520
Bottlenose Dolphin	74	1,632	66	4,187
Total	**174**	**3,525**	**167**	**41,379**

4.2 Manual Classifier

4.2.1 Classification of Sighted Encounters

All whistles manually extracted from AMAPPS 2013 encounters that had visual confirmation of species identity and that were produced by a species that is included in the original Atlantic classifier were analyzed with the manual classifier to test the performance of this classifier with a novel dataset. A total of 53 acoustic encounters were included in this analysis. Overall, 73.6 percent of these encounters were classified to the correct species (**Table 5**). Certainty scores for these classifications ranged from one to five (**Table 6**). When only encounters with high certainty scores (4 and 5) were considered, overall correct classification increased to 87.8 percent (**Table 7**).

4.2.2 Manual Classifier Modifications

Whistles from the AMAPPS 2013 recordings measured using ROCCA's manual method (**Table 3**) were added to the original Atlantic classifier training dataset (**Table 4**). The addition of these whistles increased the sample size for every species and allowed three species to be added to the existing classifier, including Risso's dolphin, Clymene dolphin, and rough-toothed dolphin. A new random-forest classifier was trained using this larger dataset. A one-stage model and several two-stage classification models were tested. The model that produced the highest correct classification scores was a two-stage model with one species-group (rough-toothed/Risso's dolphins) and six individual species in stage one. In stage two, whistles in the rough-toothed/Risso's group were classified to species (either rough-toothed or Risso's dolphin). When this two-stage classifier was tested, 55.4 percent of encounters were correctly classified to species. This is significantly greater (Fisher's exact test, $p<0.0001$) than the 12.5 percent correct classification score that would be expected by chance alone for eight species. For individual species, correct classification scores ranged from 3.2 percent for Risso's dolphin to 84.2 percent for pilot whales (**Table 8**). When a classifier that did not include Risso's dolphin was trained, individual species correct classification scores were similar to the correct classification scores when Risso's dolphin was included (**Table 9**). Overall, 66.0 percent of encounters were correctly classified. This is significantly greater (Fisher's exact test, $p<0.0001$) than the 14.3 percent correct classification score that would be expected by chance alone for seven species.

Table 5. Classification results for AMAPPS 2013 encounters with visual confirmation of species identity, based on whistles measured using ROCCA's manual method. Confusion matrix shows percentage of encounters that were classified as each species. Percent of encounters correctly classified are in bold on the diagonal. Overall, 73.6 percent of encounters were correctly classified.

Actual Species	Percent classified as					
	Short-beaked common dolphin	Striped dolphin	Short-finned pilot whale	Atlantic spotted dolphin	Bottlenose dolphin	*n*
Short-beaked common dolphin	**33.3**	0.0	66.7	0.0	0.0	3
Striped dolphin	0.0	**50.0**	0.0	50.0	0.0	2
Short-finned pilot whale	n/a	n/a	**n/a**	n/a	n/a	0
Atlantic spotted dolphin	0.0	0.0	15.3	**69.4**	15.3	13
Bottlenose dolphin	2.9	2.9	8.6	5.6	**80**	35

Table 6. Classification results and certainty scores for AMAPPS 2013 encounters that had visual confirmation of species identity. Results include the number of whistles measured using ROCCA's manual method, the actual species identification based on visual observations, the species identification based on the whistle classifier, and the certainty score for the acoustic classification.

Encounter ID number	Number of whistles	Actual species	Classified as	Certainty score
284_Dd	1	Short-beaked common	Short-beaked common	2
285_Dd	9	Short-beaked common	Short-finned pilot whale	2
286_Dd	42	Short-beaked common	Short-finned pilot whale	3
306_Sc	26	Striped	Striped	3
307_Sc	13	Striped	Atlantic spotted	3
292_Sf	25	Atlantic spotted	Short-finned pilot whale	4
293_Sf	24	Atlantic spotted	Short-finned pilot whale	4
291_Sf	6	Atlantic spotted	Atlantic spotted	5
310_Sf	31	Atlantic spotted	Atlantic spotted	5
311_Sf	20	Atlantic spotted	Atlantic spotted	5
312_Sf	30	Atlantic spotted	Atlantic spotted	5
313_Sf	7	Atlantic spotted	Atlantic spotted	5
314_Sf	25	Atlantic spotted	Atlantic spotted	5
315_Sf	29	Atlantic spotted	Atlantic spotted	4
318_Sf	20	Atlantic spotted	Atlantic spotted	5
319_Sf	27	Atlantic spotted	Atlantic spotted	5
316_Sf	5	Atlantic spotted	Bottlenose	5
321_Sf	1	Atlantic spotted	Bottlenose	2
326_Tt	2	Bottlenose	Short-beaked common	2
337_Tt	11	Bottlenose	Short-finned pilot whale	5

Encounter ID number	Number of whistles	Actual species	Classified as	Certainty score
341_Tt	2	Bottlenose	Short-finned pilot whale	2
296_Tt	23	Bottlenose	Striped	5
328_Tt	7	Bottlenose	Bottlenose	3
294_Tt	4	Bottlenose	Bottlenose	2
295_Tt	28	Bottlenose	Bottlenose	5
324_Tt	1	Bottlenose	Atlantic spotted	2
325_Tt	22	Bottlenose	Bottlenose	5
327_Tt	11	Bottlenose	Bottlenose	5
329_Tt	25	Bottlenose	Bottlenose	5
330_Tt	49	Bottlenose	Bottlenose	4
332_Tt	12	Bottlenose	Bottlenose	4
333_Tt	27	Bottlenose	Short-finned pilot whale	1
334_Tt	27	Bottlenose	Bottlenose	5
335_Tt	36	Bottlenose	Bottlenose	5
338_Tt	38	Bottlenose	Bottlenose	5
339_Tt	24	Bottlenose	Bottlenose	5
340_Tt	11	Bottlenose	Bottlenose	5
342_Tt	15	Bottlenose	Bottlenose	5
343_Tt	28	Bottlenose	Bottlenose	5
346_Tt	8	Bottlenose	Bottlenose	5
347_Tt	11	Bottlenose	Bottlenose	5
348_Tt	29	Bottlenose	Bottlenose	5
349_Tt	35	Bottlenose	Bottlenose	5
351_Tt	27	Bottlenose	Bottlenose	5
352_Tt	19	Bottlenose	Bottlenose	5
353_Tt	30	Bottlenose	Bottlenose	5
354_Tt	20	Bottlenose	Atlantic spotted	3
355_Tt	23	Bottlenose	Bottlenose	5
356_Tt	30	Bottlenose	Bottlenose	5
357_Tt	24	Bottlenose	Bottlenose	5
358_Tt	21	Bottlenose	Bottlenose	5
359_Tt	11	Bottlenose	Bottlenose	5
360_Tt	30	Bottlenose	Bottlenose	5

Table 7. Classification results with certainty scores of 4 or 5 for AMAPPS 2013 encounters with visual confirmation of species identity, based on whistles measured using ROCCA's manual method. The confusion matrix shows the percentage of encounters that were classified as each species. The percentages of encounters correctly classified are in bold on the diagonal. Overall, 87.8 percent of encounters were correctly classified. None of the short-beaked common dolphin or striped dolphin encounters had a certainty score greater than 3.

Actual Species	Percentage classified as					
	Short-beaked common dolphin	Striped dolphin	Short-finned pilot whale	Atlantic spotted dolphin	Bottlenose dolphin	*n*
Short-beaked common dolphin	n/a	n/a	n/a	n/a	n/a	0
Striped dolphin	n/a	n/a	n/a	n/a	n/a	0
Short-finned pilot whale	n/a	n/a	n/a	n/a	n/a	0
Atlantic spotted dolphin	0.0	0.0	15.4	**76.9**	7.7	13
Bottlenose dolphin	3.6	3.6	0.0	0.0	**92.8**	28

Table 8. Confusion matrix for the new manual Atlantic classifier, including additional species from AMAPPS 2013 data. The percentage of encounters classified as each species, with SD in parenthesis, is given based on 100 iterations of dividing data into training and testing datasets. The percentages of encounters correctly classified are in bold on the diagonal.

Actual Species	Percentage classified as								
	Short-beaked common dolphin	Risso's dolphin	Short-finned pilot whale	Clymene dolphin	Striped dolphin	Rough-toothed dolphin	Atlantic spotted dolphin	Bottle-nose dolphin	*n*
Short-beaked common dolphin	**77.6 (8.1)**	0.3 (1.9)	0 (0)	0 (0)	3.1 (5.4)	8.2 (7.4)	6.6 (6.6)	4.2 (5.8)	12
Risso's dolphin	12.4 (7.4)	**3.2 (4.7)**	22.9 (8.5)	10.9 (6.3)	19.9 (9.8)	1.8 (3.8)	14.2 (9.1)	14.4 (7.5)	12
Short-finned pilot whale	0 (0)	0.9 (2.4)	**84.2 (6.3)**	0.2 (1.2)	5.5 (3.6)	6.3 (5.8)	2.1 (3.0)	0.7 (2.2)	16
Clymene dolphin	0 (0)	3.5 (8.7)	0 (0)	**47.2 (9.3)**	21.7 (8.4)	24.5 (3.5)	0 (0)	3.0 (8.2)	5
Striped dolphin	27.3 (8.6)	2.8 (4.4)	3.7 (4.4)	5.9 (6.2)	**44.6 (10.2)**	3.1 (4.3)	8.2 (6.9)	4.2 (5.4)	14
Rough-toothed dolphin	0 (0)	0.2 (2.0)	25.8 (16.4)	4.0 (8.0)	0 (0)	**70.0 (18.1)**	0 (0)	0 (0)	7
Atlantic spotted dolphin	3.7 (2.6)	4.4 (3.2)	5.1 (2.7)	3.3 (2.7)	3.1 (2.3)	9.3 (3.7)	**61.2 (7.6)**	9.6 (4.2)	59
Bottle-nose dolphin	12.6 (3.5)	3.4 (2.4)	2.8 (2.1)	9.4 (3.2)	4.7 (2.2)	3.6 (2.0)	11.0 (2.8)	**52.4 (4.7)**	109

Table 9. Confusion matrix for the new manual Atlantic classifier, including additional species from AMAPPS 2013 data but not including Risso's dolphin. The percentage of encounters classified as each species, with SD in parenthesis, is given based on 100 iterations of dividing data into training and testing datasets. The percentages of encounters correctly classified are in bold on the diagonal.

Actual Species	Short-beaked common dolphin	Short-finned pilot whale	Clymene dolphin	Striped dolphin	Rough-toothed dolphin	Atlantic spotted dolphin	Bottlenose dolphin	*n*
Short-beaked common dolphin	**78.4 (8.4)**	0 (0)	0 (0)	4.1 (5.8)	0.5 (2.4)	11.7 (3.8)	5.3 (5.7)	9
Short-finned pilot whale	0 (0)	**85.2 (5.4)**	0.1 (0.1)	6.0 (3.8)	4.7 (4.9)	2.7 (3.3)	1.3 (2.5)	16
Clymene dolphin	0 (0)	0 (0)	**51.0 (14.2)**	24.7 (2.5)	20.2 (9.9)	0 (0)	4.0 (9.2)	4
Striped dolphin	29.1 (8.2)	2.9 (4.3)	5.3 (7.0)	**47.6 (11.4)**	1.6 (3.3)	9.9 (5.9)	3.5 (4.8)	13
Rough-toothed dolphin	0 (0)	18.0 (10.8)	2.8 (7.0)	0 (0)	**79.2 (14.5)**	0 (0)	0 (0)	5
Atlantic spotted dolphin	3.9 (2.3)	5.4 (3.1)	2.7 (2.4)	3.7 (2.6)	7.4 (3.8)	**67.0 (7.0)**	9.8 (4.3)	44
Bottlenose dolphin	12.7 (3.5)	2.6 (1.7)	10.6 (3.3)	5.9 (2.7)	2.5 (1.7)	11.7 (3.4)	**53.9 (5.3)**	73

4.2.3 Classification of Encounters Without Visual Confirmation of Species Identity

Whistles were measured from 20 acoustic encounters that did not have visual confirmation of species identity collected from the AMAPPS 2013 cruises. Some of these dolphins were not seen at all and some had associated visual observations, but species identification was not possible due to elusive animal behavior, poor sighting conditions, or other factors. Due to time constraints, a maximum of 30 whistles were measured using ROCCA's manual method from each encounter. These whistles were classified using the new Atlantic classifier that included eight species (**Section 4.2.2**, **Table 8**). Eighteen of the twenty encounters were classified as striped dolphins (**Table 10**), which was one of the most commonly sighted delphinid species during the northern legs of the AMAPPS 2013 survey (NEFSC and SEFSC 2013). Most of the non-sighted acoustic encounters included in this analysis occurred in the northern part of the study area, over and offshore of the continental slope (**Figure 2**). The locations of these encounters are consistent with the locations of visual observations of striped dolphins during the northern legs of the survey (**Figure 3**).

Table 10. Classification results for whistles measured manually from AMAPPS 2013 encounters that did not have visual confirmation of species identity. Results include the number of whistles included in the analysis, the species that the encounters was classified as by ROCCA's manual classifier and the certainty score for the acoustic classification.

Encounter ID	Number of whistles	Classified as	Certainty score
GU1304_UD_001	25	Striped dolphin	5
GU1304_UD_002	30	Striped dolphin	4
GU1304_UD_003	30	Striped dolphin	5
GU1304_UD_005	30	Striped dolphin	5
GU1304_UD_006	2	Striped dolphin	2
GU1304_UD_007	4	Clymene dolphin	2
HB1303_UD_008	25	Short-finned pilot whale	5
HB1303_UD_009	14	Striped dolphin	5
HB1303_UD_010	10	Striped dolphin	5
HB1303_UD_011	17	Striped dolphin	5
HB1303_UD_012	29	Striped dolphin	5
HB1303_UD_013	30	Striped dolphin	5
HB1303_UD_014	30	Striped dolphin	5
HB1303_UD_015	30	Striped dolphin	5
HB1303_UD_016	30	Striped dolphin	5
HB1303_UD_017	30	Striped dolphin	5
HB1303_UD_018	11	Striped dolphin	5
HB1303_UD_019	19	Striped dolphin	5
HB1303_UD_020	15	Striped dolphin	5
HB1303_UD_021	30	Striped dolphin	5

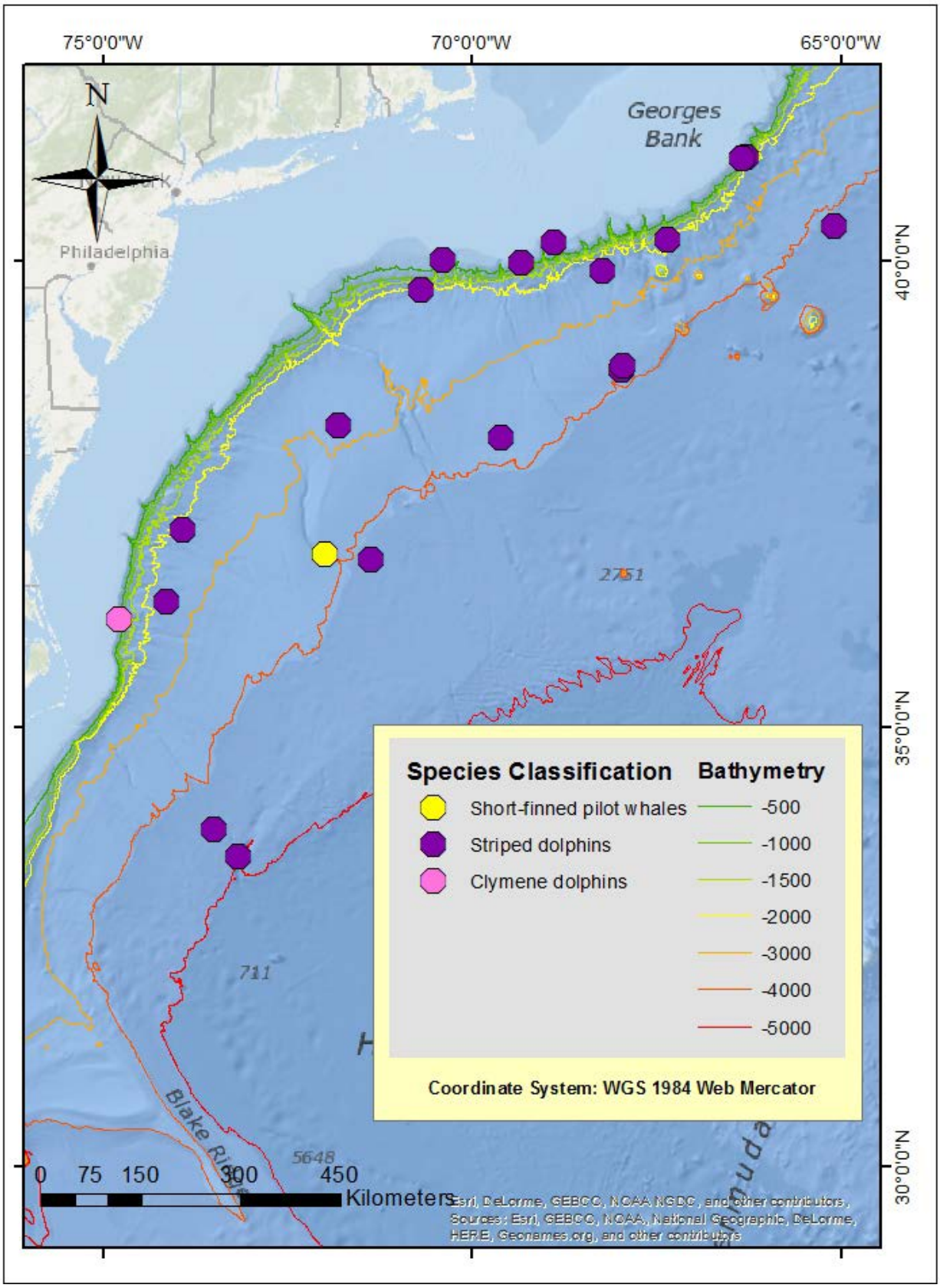

Figure 2. Locations and acoustic classifications for acoustic detections that did not have associated visual observations.

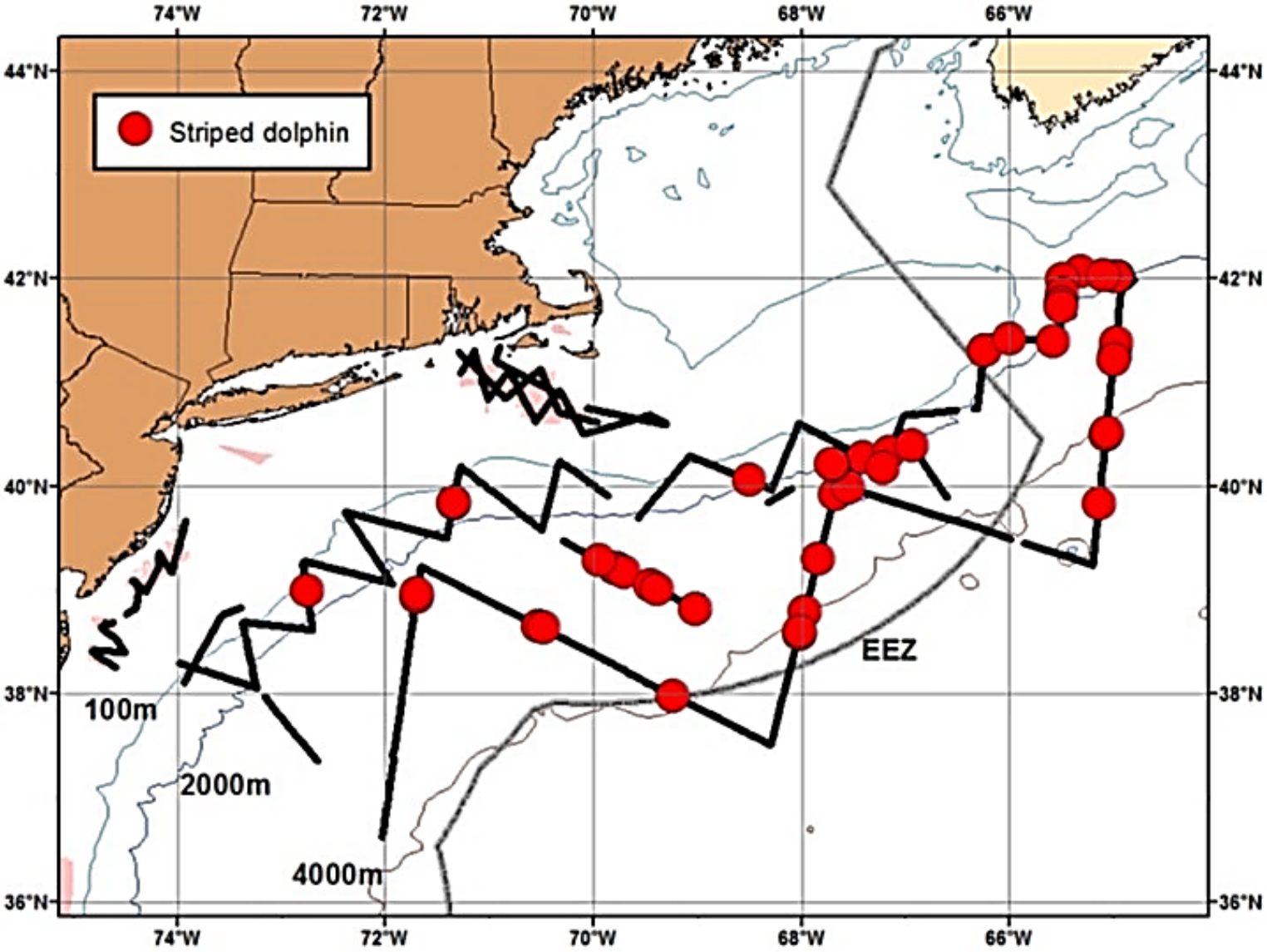

Figure 3. Location of visual sightings of striped dolphin school during the northern legs of the AMAPPS 2013 survey (from NEFSC and SEFSC 2013)

4.3 Automated Classifier

4.3.1 Classification of Sighted Encounters

When the 53 AMAPPS 2013 encounters that had visual confirmation of species identity were analyzed using PAMGuard's automated WMD, every encounter was classified as short-finned pilot whale. One of the outputs of a random-forest analysis is the Gini Variable Importance Index. This index provides a relative measure of the degree to which each variable contributes to the classifier (Breiman et al. 1984). Comparisons of the six variables that are most important to the classifier (duration, maximum frequency, center frequency, mean frequency, mean positive slope, absolute value of the slope; Oswald 2013) showed that almost all variables were significantly different for every species between the original classifier training dataset and the AMAPPS 2013 dataset (Mann-Whitney U test, $p<0.0001$, **Appendix A**). The only exceptions to this were maximum frequency for Atlantic spotted dolphins (Mann-Whitney U test, $p=0.57$) and center frequency (Mann-Whitney U test, $p=0.28$) and mean frequency (Mann-Whitney U test, $p=0.87$) for striped dolphins. Because of these differences, combining the datasets to train a new classifier was not possible and it was not appropriate to use this classifier to identify encounters that did not have associated visual observations.

5. Discussion

The goals of this study were to test, improve, and utilize two whistle classifiers for delphinid species in the northwestern Atlantic Ocean. These goals were met for the manual Atlantic classifier. When this classifier was tested by classifying encounters that had visual confirmation of species identity, correct classification scores were relatively high for Atlantic spotted dolphins and bottlenose dolphins. The correct classification score for Atlantic spotted dolphins (69 percent for all encounters) was somewhat lower than the 90 percent based on original tests of the Atlantic classifier for this species (**Tables 1 and 5**). However, when encounters with certainty scores of 4 or 5 were considered, 10 of 13 encounters (77 percent) were classified correctly. The encounters that were misclassified were misclassified as either short-finned pilot whale or bottlenose dolphin, which is what would be expected based on the original Atlantic classifier confusion matrix in Oswald (2013). The correct classification score for bottlenose dolphins (80 percent for all encounters, 93 percent for encounters with certainty scores of 4 or 5) was higher than the 70 percent based on earlier tests of the Atlantic classifier. Patterns of misclassification were similar to what was expected based on Oswald (2013), with most misclassifications being classified as short-finned pilot whales and Atlantic spotted dolphins when considering all encounters. When only considering classifications with high certainty scores, one encounter was misclassified as short-beaked common dolphin and one encounter was misclassified as striped dolphin. Based on the original Atlantic classifier confusion matrix, the probability of misclassification as these two species is lower than short-finned pilot whales and Atlantic spotted dolphins, but is still expected.

Low sample sizes made it difficult to evaluate classifier performance for short-beaked common and striped dolphins. For both of these species, correct classification scores were lower than expected based on original tests of the Atlantic classifier (**Tables 1 and 5**); however, it is difficult to generalize based on samples sizes of only two and three encounters for striped and short-beaked common dolphins, respectively. In addition, none of these encounters had certainty scores greater than three, and so biologists could not be confident of either the correct classifications or the misclassifications for these species. Short-finned pilot whale encounters were not available for this analysis; therefore, the biologists were not able to evaluate the performance of the classifier with this species. More rigorous testing of the classifier with additional recordings that have visual confirmation of species identity is necessary in order to gain an in-depth understanding of its performance on novel data sets.

When whistles from the AMAPPS 2013 dataset were added to the Atlantic classifier dataset, it was possible to add three species to the classifier (Risso's dolphin, Clymene dolphin, and rough-toothed dolphin). The overall correct classification score for the resulting classifier (55 percent) was lower than for the original Atlantic classifier (91 percent), partially because of the greater number of species included in the classifier, and partially due to the fact that the correct classification score for Risso's dolphin whistles was very low (3 percent; **Table 8**). While the correct classification score for Risso's dolphins was low, few encounters recorded from other species were misclassified as Risso's dolphins. In addition, when a classifier that did not include Risso's dolphins was trained, correct classification scores for the other species were similar to the correct classification results when Risso's dolphin was included. These results suggest that

it is reasonable to include Risso's dolphin in the classifier, as it provides the potential for Risso's dolphin encounters to be classified as such without significantly reducing correct classification scores for other species.

Correct classification scores for Risso's dolphins and other species may be improved by adding information from other sources to the classifier. For example, it has been shown that the echolocation clicks produced by Risso's dolphins have species characteristics that allow them to be identified with a high degree of accuracy (Soldevilla et al. 2008, Roch et al. 2011). Adding information from clicks may allow species to be discerned with a higher degree of accuracy than is possible based on whistles alone. Bio-Waves, Inc. is exploring this possibility in another project (funded by the Office of Naval Research, ONR, and the Living Marine Resources Program, LMR). Preliminary results suggest that combining information from whistles, clicks, and other sources can improve the performance of classifiers for odontocete species.

Most (18 out of 20) AMAPPS 2013 acoustic encounters that did not have visual confirmation of species identity were classified as striped dolphins. Although striped dolphins are uncommon in the southern portion of the study area, they were the one of the most commonly detected small cetacean species both visually and acoustically during the northern legs of this survey (NEFSC and SEFSC 2013). The geographic locations of these non-sighted encounters corresponded with locations of visual encounters with this species (**Figures 2 and 3**). Most were north of 36°N and offshore of the continental shelf, which is where all of the detections occurred during both the NEFSC and SEFSC legs of the survey. Based strictly on their locations, the two southern-most non-sighted acoustic encounters that were classified as striped dolphins may be misclassifications. One of these encounters (GU1304_006) was classified based only on two whistles, so its classification is uncertain. The other encounter (GU1304_UD_005) had a high certainty score, as 33 percent of trees classified the encounter as striped dolphin and no other species had a similar percentage of tree votes. The species with the second highest percentage of tree votes (short-beaked common dolphins) had 21 percent of tree votes and may be a more plausible classification than striped dolphin given the location of this encounter. Geographic location is another variable that may be useful for improving classification success and this has been shown to be the case in preliminary tests of the new classifiers being developed by Bio-Waves, Inc. in the previously mentioned ONR/LMR-funded project.

The 20 non-sighted acoustic encounters analyzed here represent a small subsample of the non-sighted acoustic encounters recorded during the AMAPPS 2013 survey. The encounters included here were chosen based on distance from other encounters of whistling species. Only encounters that were at least 3 nautical miles from other whistling groups were included in the analysis. Although this does not ensure that the recordings contained whistles produced by a single species, it does reduce the probability of multiple species being present in the recordings. One of the limitations of the ROCCA classifier is that it does not have the capability to identify schools as multi-species. If more than one species is present in the recording, the encounter will be identified as only the species with the greatest number of tree votes. It is possible that some of the encounters analyzed here contained more than one species. It may be possible to identify multi-species schools by further analysis of the distribution of tree votes among species or some other method, but this is something that requires further exploration. If it becomes possible to identify multi-species groups, it will not be necessary to select encounters that are separated as

widely from other whistling schools and it will be possible to gain a more complete understanding of the distribution of species and of what kinds of groups are not seen by visual observers.

When the automated classifier was used to identify encounters that had visual confirmation of species identity, all classifications were incorrect. Every encounter was classified as pilot whale. In order to rule out the possibility that this was caused by an error in the classification software, we re-classified the encounters using a classifier trained using the same dataset, but with different random-forest code written in the R programming language (R Developmental Core Team 2016), by an independent programmer (G. Alongi). This analysis gave identical results to the ROCCA classifications, which suggests that the issue lies in the training or testing data. This hypothesis is supported by the variable comparisons presented in **Section 4.3.1** and **Appendix A**. Although some amount of within-species variability is expected, the significant differences shown in these figures suggest something other than within-species variability. The variables that were compared are variables that have high rankings on the Gini Variable Importance Index and so they have a significant effect on the performance of the classifier. The main difference between the two datasets is the version of PAMGuard that was used to automatically detect and extract whistle contours. The Atlantic training data were processed using PAMGuardBeta_1_13_02 and PAMGuardBeta_1_13_03. The AMAPPS 2013 data were processed using a later version of PAMGuard, PAMGuard_SMRU_1_13_05d. Changes made to the later version of PAMGuard may have inadvertently affected the performance of the WMD and led to the differences in the measured variables that are evident in the figures in **Appendix A.** As part of a different project, a subset of the Atlantic classifier training data were re-measured using PAMGuard_SMRU_1_13_05d. When these re-measured whistles were used to train a classifier and analyze the AMAPPS 2013 data, the classification results were different. It is currently unclear whether the different classification results are due to differences in versions of PAMGuard or due to the fact that only a portion of the Atlantic training dataset was used. Unfortunately, it was not possible to re-measure the entire AMAPPS 2013 dataset within the scope of this project, however this would be a first step in determining the source of the differences between datasets and classification results. More investigation is necessary to determine whether the different classification results are due to differences in the versions of PAMGuard, differences in the WMD settings used, and/or due to the fact that only a subset of the training data were used. This issue needs to be resolved before datasets can be combined to train new classifiers.

This page intentionally left blank.

6. Summary and Conclusions

This study provides valuable information on the performance of two whistle classifiers for delphinid species in the northwestern Atlantic Ocean. The manual classifier performed as expected for the two species that could be tested. Sample sizes were too low to allow the evaluation of the performance of the classifier for the other three species in the classifier and so a priority should be placed on obtaining visually validated acoustic recordings for these species. When species were added to the manual classifier, correct classification scores decreased, highlighting the need to add additional information to the classifier feature vectors. However, when the new classifier was used to identify encounters that did not have visual confirmation of species identity, results were consistent with what was expected based on species distribution in the area.

The automated classifier performed well with the original Atlantic testing and training datasets, as described in Oswald (2013). However, it classified every visually validated AMAPPS 2013 encounter incorrectly as pilot whales. Comparisons of the Atlantic classifier and AMAPPS 2013 datasets suggest that the version of PAMGuard used to analyze the data may have a significant effect on contour extraction and measurement and therefore on classifier performance. Determining the source of the discrepancy within PAMGuard was beyond the scope of this project, but is a topic that should be pursued. In the meantime, it will be crucial to be consistent with PAMGuard versions when using the Atlantic classifier.

While the manual classifier requires more time and effort for the detection and extraction whistles, it proved to be more generalizable to novel datasets than did the automated classifier. The manual classifier allowed identification of acoustic encounters that did not have associated visual observations. The ability to identify non-sighted schools allows a more complete understanding of species distribution as well as providing insight to species that are more difficult to detect using visual methods. These capabilities highlight the complementary nature of visual and acoustic methods and allow more information to be gleaned from shipboard surveys.

This page intentionally left blank.

7. Acknowledgements

We would like to thank Danielle Cholewiak (NEFSC) and Melissa Soldevilla (SEFSC) for generously providing AMAPPS 2013 data, as well as earlier data for the Atlantic classifier and analysis advice throughout the project. We also thank Andrew Read and Lynne Hodge (Duke University) for providing data for the Atlantic classifier training dataset. We are grateful to Gabriela Alongi, Shannon Coates, and Kerry Dunleavy for their assistance with data analysis. We thank U.S. Fleet Forces Command for providing funding for this analysis under the U.S. Navy's Marine Species Monitoring Program. Project management and technical review was provided by Naval Facilities Engineering Command Atlantic (NAVFAC LANT). We are especially grateful for the logistic support and advice from Joel Bell (NAVFAC LANT) and from Michael Richlen and Dan Engelhaupt (HDR).

This page intentionally left blank.

8. References

Barkley, Y., J.N. Oswald, J.V. Carretta, S. Rankin, A. Rudd, and M.O. Lammers. 2011. Comparison of real-time and post-cruise acoustic species identification of dolphin whistles using ROCCA (Real-time Odontocete Call Classification Algorithm). NOAA Technical Memorandum NOAA-TM-NMFS-SWFSC-473. National Marine Fisheries Service, La Jolla, CA. 29 pp.

Bioacoustics Research Program. 2011. Raven Pro: Interactive sound analysis software (Version 1.4) [Computer software]. The Cornell Lab of Ornithology, Ithaca, NY. Available from http://www.birds.cornell.edu/raven.

Breiman, L. 2001. Random forests. *Machine Learning* 45: 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

Brown, J.C., and P. Smaragdis. 2009. Hidden Markov and Guassian mixture models for automatic call classification. *Journal of the Acoustical Society of America* 125: EL221–EL224.

Gillespie, D., J. Gordon, R. McHugh, D. McLaren, D.K. Mellinger, P. Redmond, A. Thode, P. Trinder, and X.Y. Deng. 2008. PAMGUARD: Semiautomated, open-source software for real-time acoustic detection and localization of cetaceans. Pages 54–62 in: Institute of Acoustics. *Conference on Underwater Noise Measurement, Impact and Mitigation 2008*. Proceedings of the Institute of Acoustics, volume 30, part 5. Institute of Acoustics, St. Albans, UK.

Fristrup, K.M., and W.A. Watkins. 1993. Marine animal sound classification. Technical Report No. WHOI-94-13. Woods Hole Oceanographic Institute, MA. 32 pp.

Johnson, M.P., and P.L. Tyack. 2003. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE Journal of Oceanic Engineering* 28: 3–12.

Matthews, J.N., L.E. Rendell, J.C.D. Gordon, and D.W. MacDonald. 1999. A review of frequency and time parameters of cetacean tonal calls. *Bioacoustics* 10: 47-71.

Mellinger, D.K. 2001. Ishmael 1.0 user's guide. NOAA Technical Memorandum OAR PMEL-120. NOAA/PMEL/OERD, Newport, OR. http://www.pmel.noaa.gov/pubs/PDF/mell2434/mell2434.pdf.

Mellinger, D.K., and J. Barlow. 2003. Future directions for marine mammal acoustic surveys: stock assessment and habitat use. Workshop held in La Jolla, CA, 20-22 November 2002. NOAA/PMEL Contribution No. 2557. NOAA/PMEL, Seattle, WA. 45 pp.

Northeast Fisheries Science Center (NEFSC), and Southeast Fisheries Science Center (SEFSC). 2013. Annual report of a comprehensive assessment of marine mammal, marine turtle, and seabird abundance and spatial distribution in US waters of the western North Atlantic Ocean. Available from http://www.nefsc.noaa.gov/psb/AMAPPS/docs/NMFS_AMAPPS_2013_annual_report_FINAL3.pdf

Nguyen, C., Y. Wang, and H.N. Nguyen. 2013. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering* 06: 551-560.

Oswald, J.N. 2013. Development of a classifier for the acoustic identification of delphinid species in the northwest Atlantic Ocean*.* Final Report. Submitted to HDR Environmental, Operations and Construction, Inc. Norfolk, Virginia under Contract No. CON005-4394-009, Subproject 164744, Task Order 003, Agreement # 105067. Prepared by Bio-Waves, Inc., Encinitas, California.

Oswald, J.N., J. Barlow, and T.F. Norris. 2003. Acoustic identification of nine delphinid species in the eastern tropical Pacific Ocean. *Marine Mammal Science* 19: 20–37.

Oswald, J.N., J.V. Carretta, M. Oswald, S. Rankin and W.W.L. Au. 2011. Seeing the species through the trees: Using random forest classification trees to identify species-specific whistle types*. Journal of the Acoustical Society of America* 129: 2639.

Oswald, J.N., S. Rankin, J. Barlow, M. Oswald and M.O. Lammers. 2013. Real-time Call Classification Algorithm (ROCCA): software for species identification of delphinid whistles. Pages 245–266 in: O. Adam, and F. Samaran, eds. *Detection, Classification and Localization of Marine Mammals using Passive Acoustics, 2003-2013: 10 years of International Research.* DIRAC NGO, Paris, France.

R Developmental Core Team. 2016. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rendell, L.E., J.N. Matthews, A. Gill, J.C.D. Gordon, and D.W. MacDonald. 1999. Quantitative analysis of tonal calls from five odontocete species, examining interspecific and intraspecific variation. *Journal of Zoology* 249: 403–410.

Roch, M.A., M.S. Soldevilla, J.C. Burtenshaw, E.E. Henderson, and J.A. Hildebrand. 2007. Gaussian mixture model classification of odontocetes in the southern California Bight and the Gulf of California. *Journal of the Acoustical Society of America* 121: 1737–1748.

Roch M.A., H. Klinck, S. Baumann-Pickering, D.K. Mellinger, S. Qui, M.S. Soldevilla, and J.A. Hildebrand. 2011. Classification of echolocation clicks from odontocetes in the Southern California Bight. *Journal of the Acoustical Society of America* 129: 467–475.

Soldevilla, M.S., E.E. Henderson, G.S. Campbell, S.M. Wiggins, J.A. Hildebrand, and M.A. Roch. 2008. Classification of Risso's and Pacific white-sided dolphins using spectral properties of echolocation clicks. *Journal of the Acoustical Society of America* 124: 609–624.

Steiner, W.W. 1981. Species-specific differences in pure tonal whistle vocalization of five western North Atlantic dolphin species. *Behavioral Ecology and Sociobiology* 9: 241–246.

Wang, D., B. Würsig, and W. Evans. 1995. Comparisons of whistles among seven odontocete species. Pages 299–323 in: R.A. Kastelein, J.A. Thomas, and P.E. Nachtigall (eds). *Sensory Systems of Aquatic Mammals*. De Spil Publishers, Woerden, Netherlands.

This page intentionally left blank.

A

Boxplots comparing original Atlantic classifier training data and AMAPPS 2013 data

This page intentionally left blank.

Appendix A. Boxplots Comparing Original Atlantic Classifier Training Data and AMAPPS 2013 Data

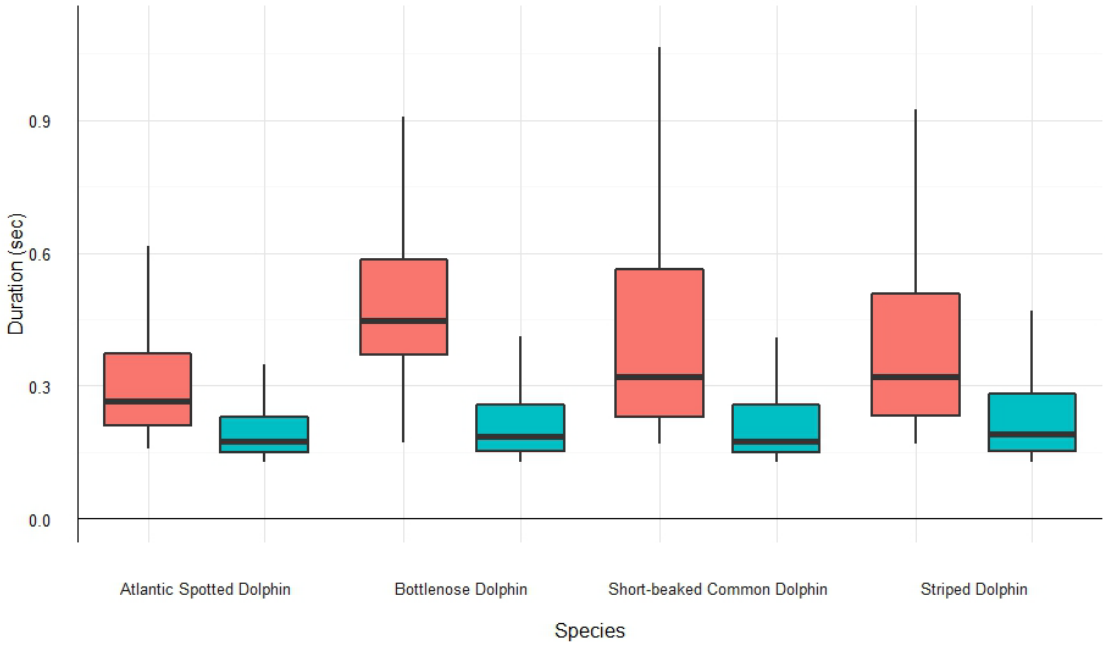

Figure 1. Boxplot comparing duration for whistles measured from the original Atlantic classifier training dataset (pink) and the AMAPPS 2013 dataset (blue). Lower and upper hinges correspond to the first and third quartiles. The upper whisker extends to the highest value that is within 1.5*IQR (IQR = Inter-quartile Range, the distance between the first and third quartiles). The lower whisker extends to the lowest value that is within 1.5*IQR. Outliers not shown.

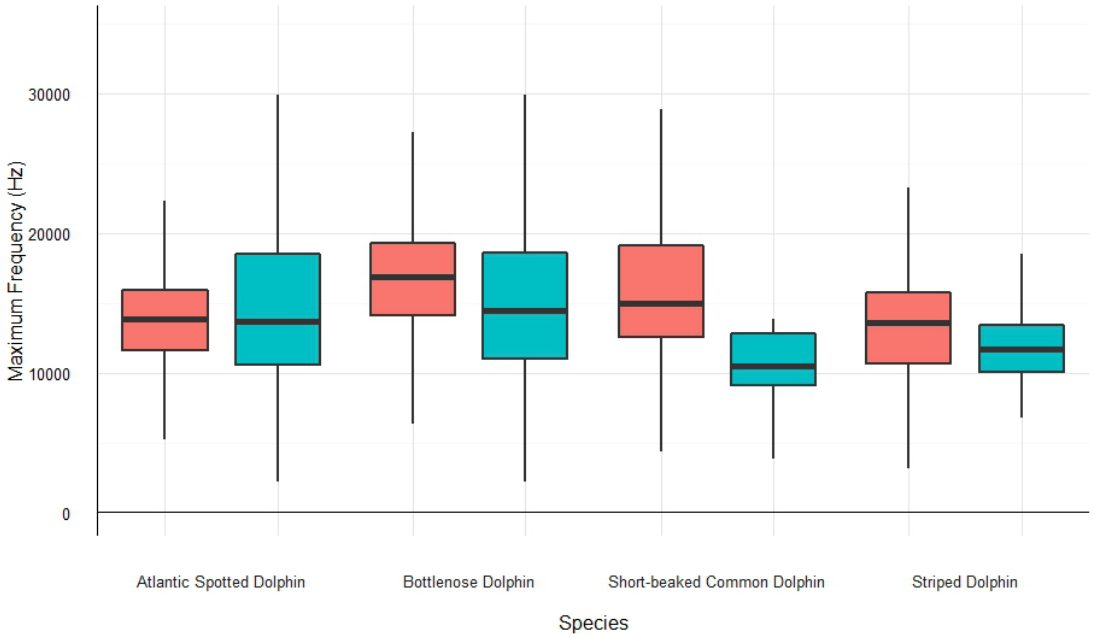

Figure 2. Boxplot comparing maximum frequency for whistles measured from the original Atlantic classifier training dataset (pink) and the AMAPPS 2013 dataset (blue). Lower and upper hinges correspond to the first and third quartiles. The upper whisker extends to the highest value that is within 1.5*IQR (IQR = Inter-quartile Range, the distance between the first and third quartiles). The lower whisker extends to the lowest value that is within 1.5*IQR. Outliers not shown.

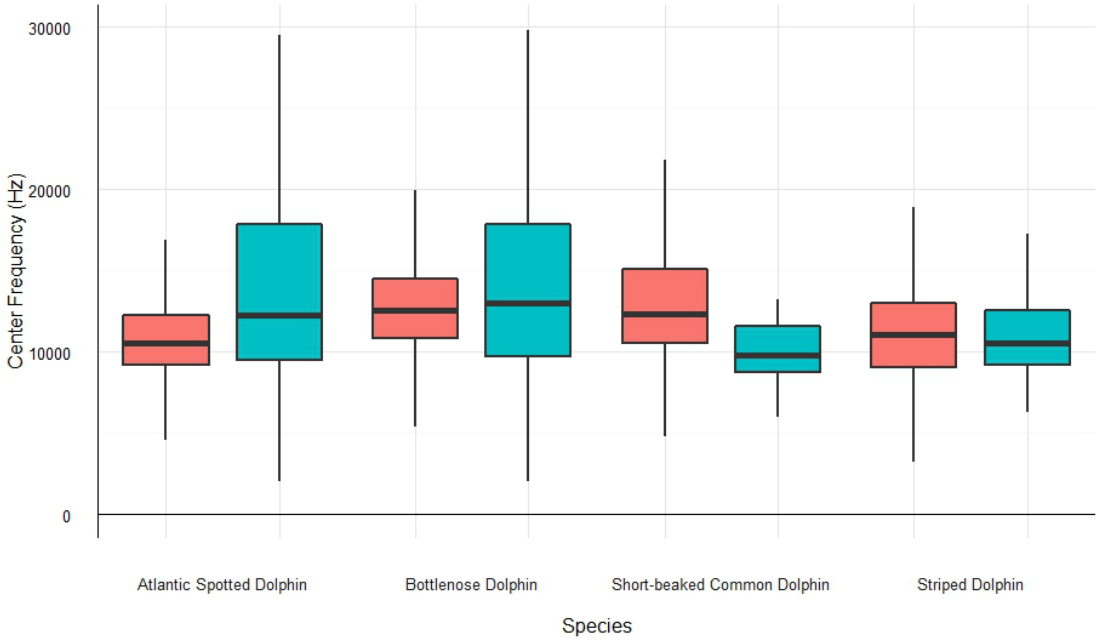

Figure 3. Boxplot comparing center frequency for whistles measured from the original Atlantic classifier training dataset (pink) and the AMAPPS 2013 dataset (blue). Lower and upper hinges correspond to the first and third quartiles. The upper whisker extends to the highest value that is within 1.5*IQR (IQR = Inter-quartile Range, the distance between the first and third quartiles). The lower whisker extends to the lowest value that is within 1.5*IQR. Outliers not shown.

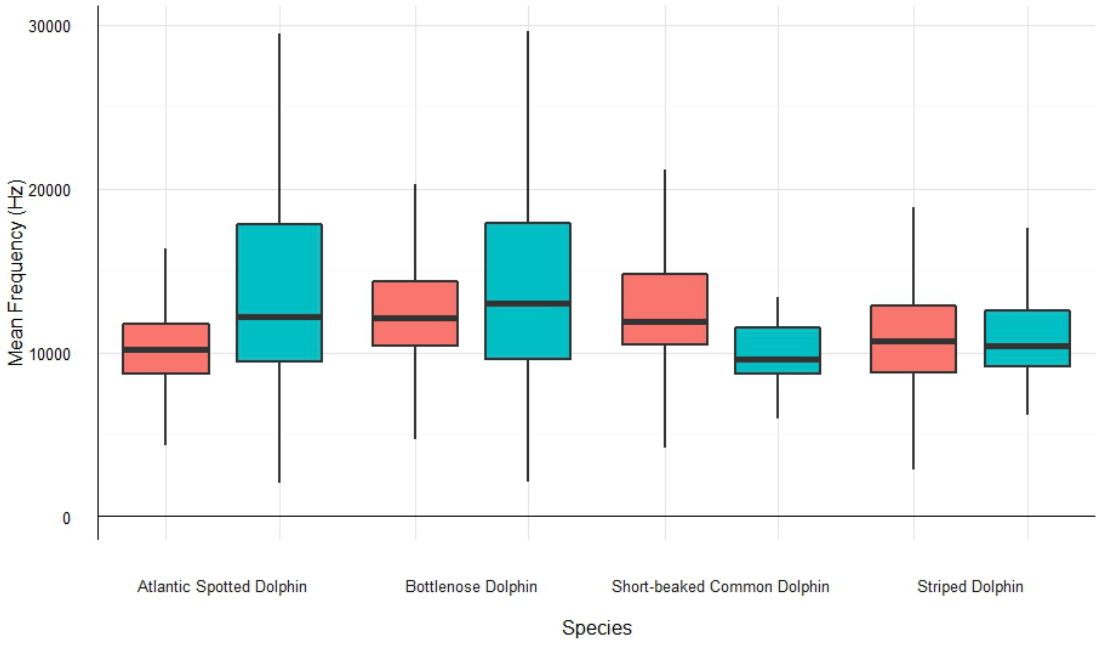

Figure 4. Boxplot comparing mean frequency for whistles measured from the original Atlantic classifier training dataset (pink) and the AMAPPS 2013 dataset (blue). Lower and upper hinges correspond to the first and third quartiles. The upper whisker extends to the highest value that is within 1.5*IQR (IQR = Inter-quartile Range, the distance between the first and third quartiles). The lower whisker extends to the lowest value that is within 1.5*IQR. Outliers not shown.

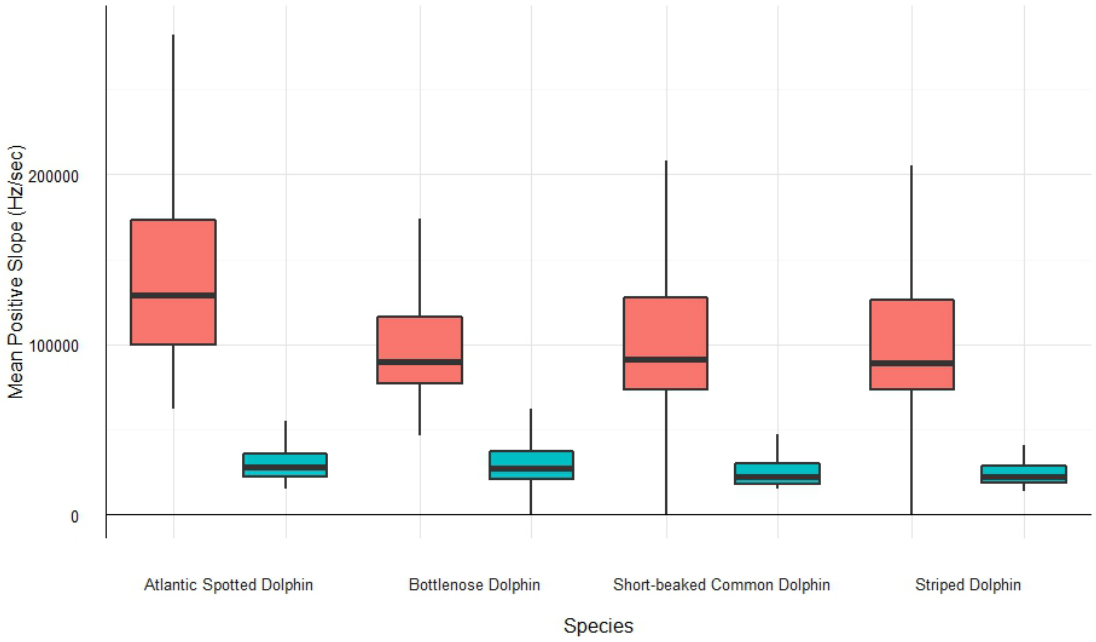

Figure 5. Boxplot comparing mean positive slope for whistles measured from the original Atlantic classifier training dataset (pink) and the AMAPPS 2013 dataset (blue). Lower and upper hinges correspond to the first and third quartiles. The upper whisker extends to the highest value that is within 1.5*IQR (IQR = Inter-quartile Range, the distance between the first and third quartiles). The lower whisker extends to the lowest value that is within 1.5*IQR. Outliers not shown.

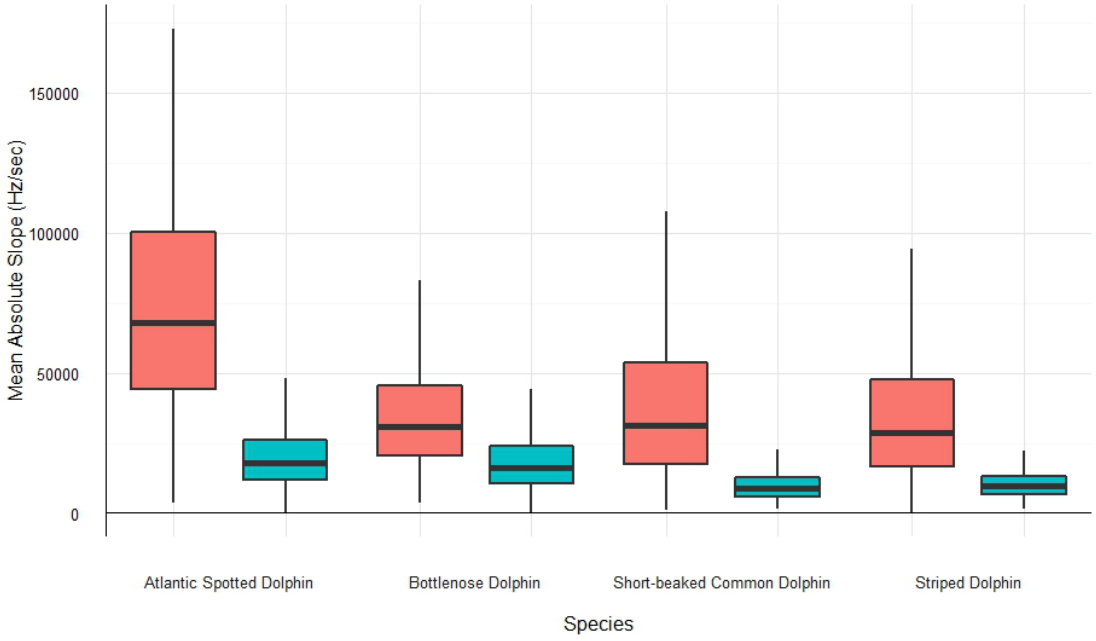

Figure 6. Boxplot comparing mean absolute slope for whistles measured from the original Atlantic classifier training dataset (pink) and the AMAPPS 2013 dataset (blue). Lower and upper hinges correspond to the first and third quartiles. The upper whisker extends to the highest value that is within 1.5*IQR (IQR = Inter-quartile Range, the distance between the first and third quartiles). The lower whisker extends to the lowest value that is within 1.5*IQR. Outliers not shown.

This page intentionally left blank.