Appendix A. Automated Detection and Classification of Marine Mammal Vocalizations[3]

A.1. Introduction

This appendix describes the methods developed by JASCO Applied Sciences for automated detection of beluga whistles, bowhead moans, bowhead songs, and walrus grunts within the data collected during the winter 2009–2010 and summer 2010 Chukchi Sea Joint Acoustic Monitoring Programs (AMPs). The algorithms and their performance are described.

Methods for automated detection and classification of marine mammal vocalizations in digital acoustic recordings have been developed over several decades. The variability of the target vocalizations influences the performance of detection algorithms. Some species such as fin and blue whales produce highly stereotyped vocalizations that are easier to detect automatically than are more variable sounds. For these stereotyped vocalizations, template-matching methods such as matched filter (Stafford 1995) and correlation of spectrograms (Mellinger and Clark 1997, 2000, Mouy *et al.* 2009) are generally effective (Mellinger *et al.* 2007). Other species produce more variable and complex tonal sounds that are more difficult to detect and classify. Such vocalizations generally require band-limited energy summation for detection, followed by statistical classification techniques for species identification (Fristrup and Watkins 1993, Oswald *et al.* 2003). Several classification methods have been investigated for belugas (Clemins and Johnson 2006, Mouy *et al.* 2008), dolphins (Oswald *et al.* 2007), humpback whales (Abbot *et al.* 2010), elephants (Clemins *et al.* 2005), and birds (Kogan and Margoliash 1998).

The performance of detection algorithms is also influenced by the acoustical surroundings. Noise generated by anthropogenic activities (shipping, seismic exploration) or weather (wind, rain, waves) may be mistaken as biological in origin. Increased ambient noise reduces the signal-to-noise ratio of vocalizations, making them harder to detect and classify. The sound propagation characteristics of the study area can alter the spectral and temporal structure of received vocalizations which can interfere with detection and classification algorithms that work well in a different propagation environment. Finally, the presence of other marine animals vocalizing in the frequency band of interest greatly increases the risk of misclassification. The influences of these factors generally also vary with time. Consequently, methods shown to be successful for a specific location, season, and species may not be successful under different circumstances.

The Chukchi Sea AMP recordings contain vocalizations produced by several species of marine mammals, including bowhead (*Balaena mysticetus*), beluga (*Delphinapterus leucas*), gray (*Eschrichtius robustus*), fin (*Balaenoptera physalus*), and killer (*Orcinus orca*) whales, walrus (*Odobenus rosmarus*), and various ice seals. Vocalizations produced by several of these species share frequency bands and can occur at the same period of the year. For instance, certain vocalizations produced by bowheads and walrus have similar durations and frequency ranges. While an experienced human analyst can usually distinguish between those vocalizations, training an automated machine to do the same is no simple task.

[3] Although many sounds made by marine mammals do not originate from vocal cords, the term “vocalization” is used as a generic term to cover all sounds discussed in this report that are produced by marine mammals. The term “call” will also be used in this sense for brevity.

Multiple sources contribute to ambient noise in the eastern Chukchi Sea. In winter, ice noise is highly problematic for automated detection algorithms. Ice cracking sounds can be emitted at surprisingly regular intervals and can resemble walrus knocks. Ice squeaking sounds are often in the frequency range of beluga vocalizations. Detection algorithms therefore must be well adapted to the variable and overlapping vocalizations of the species that frequent the eastern Chukchi Sea as well as robust against the surrounding noise background. Because many terabytes of data are collected during the Chukchi Sea AMPs, the automated analysis methods must also be computationally efficient, with computing times no less than 5 times real time (per processor).

A.2. Methods

A.2.1. Bowhead and Beluga Call Detection and Classification

The bowhead acoustic repertoire includes low-frequency moans (< 1000 Hz) produced in summer and higher-frequency, more complex songs produced in fall and early winter (Delarue *et al.* 2009). Belugas produce tonal whistles in the 1–8 kHz frequency band (Karlsen *et al.* 2002). Because these three sound-types are produced in different frequency bands, three unique detectors and classifiers were created for: (1) bowhead winter (and fall) songs, (2) bowhead summer moans, and (3) beluga whistles. Each detector had unique frequency, duration, and FFT settings to optimize performance on the call type of interest. The output of each detector was then run through its associated classifier.

The detection/classification process consists of the following steps (see Figure 150):

1. Creating the normalized spectrogram.
2. Extracting the time-frequency contours using the tonal detector developed by Mellinger *et al.* (2009).
3. Extracting 46 features from each contour to create binary random forest models.
4. Classifying the contours as either ‘target species’ (bowhead or beluga) or ‘other’ with the random forest models.
5. Post-processing of bowhead moans and songs to combine parts of single calls that were detected separately.

Once random forest models were created for bowhead moans, bowhead songs, and beluga whistles, they were tested on the test datasets described in Section A.2.4. The detection/classification process is described in detail in the following sections.

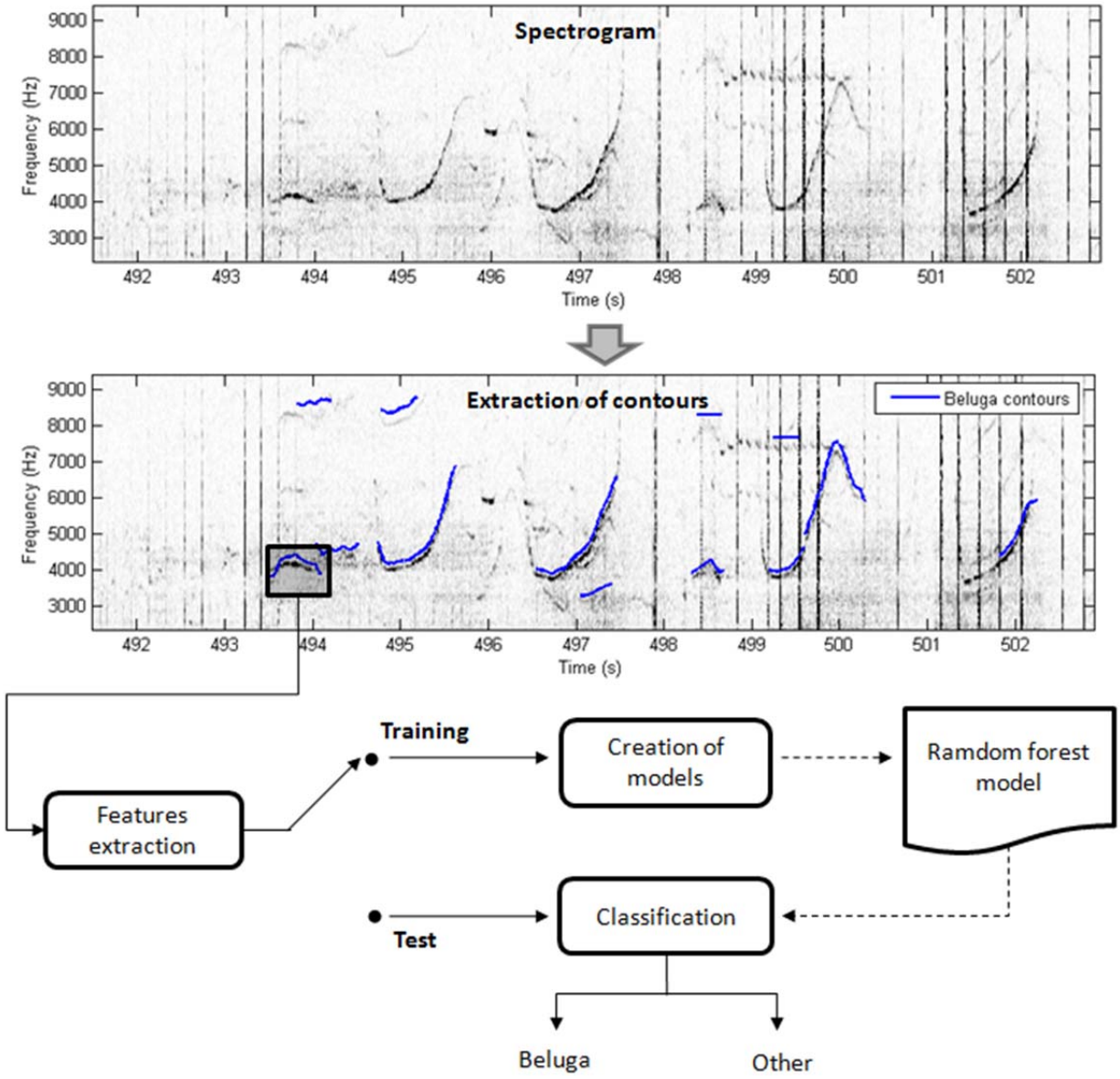

Figure 150. Steps in the detection/classification process.

Step 1: Spectrogram Processing

The first step of the detection process was the calculation of the spectrogram. Spectrogram resolutions differed for each species to ensure accurate time-frequency representation of the calls (Table 28). To attenuate long spectral rays in the spectrogram due to vessel noise and to enhance weaker transient biological sounds, the spectrogram was normalized in each frequency band (*i.e.*, each row of the spectrogram) with a split-window normalizer. The size of the window and the notch of the normalizer are indicated in Table 28. For the processing of beluga whistles the spectrogram was smoothed by convolving it with a 2-D Gaussian kernel (Gillespie 2004). Gaussian smoothing was not used for analyzing bowhead calls as it did not improve the performance of the contour extraction.

Table 28. Spectrogram parameters for each call type.

	Bowhead winter songs	Bowhead summer moans	Beluga whistles
Analysis frame size (samples)	4096	4096	1024
Overlap between frames (samples)	3500	3500	896
FFT size (sample)	16,384	16,384	1024
Window function	Hanning	Hanning	Blackman
Normalizer window size (s)	1.5	1.5	0.7
Normalizer notch size (s)	0.4	0.4	0.1
Gaussian kernel size (bins)	n/a	n/a	3×3

Step 2: Contour Extraction

Vectors representing the time-evolution of the fundamental frequency of marine mammal calls (referred to as *contours*) were extracted from the spectrograms with the MATLAB® version of a tonal detector developed by Mellinger *et al.* (2009). This tonal detector is implemented in the latest version of the widely-used Ishmael acoustic analysis software (Mellinger 2001). The algorithm works as follows based on user-defined parameters (chosen empirically, Table 29): First, candidate frequency peaks are identified for each time slice of the spectrogram in the frequency band $[f_0, f_1]$. Peaks of height h [dB] above the noise threshold (defined as the percentile P_{bg} of the spectrum values) that are the highest point in their neighborhood (n Hz wide) are selected. Second, successive peaks differing in frequency by less than f_d are connected together. Third, to accurately follow simultaneous calls, the location of the next candidate peak is estimated by fitting a line to the most recent k seconds of the contour and looking for spectral peaks where the line continues. Finally, candidate contours must persist for a minimum duration d. Figure 150 above shows an example of contours extracted from a recording containing beluga whistles.

Table 29. Contour extraction parameters for each call type.

Symbol	Description	Bowhead winter songs	Bowhead summer moans	Beluga whistles
P_{bg}	Percentile for estimating background noise	50	50	50
h	Height above that estimate (dB)	2	2	1.2
n	Neighborhood width (Hz)	50	50	250
f_d	Frequency difference from one step to the next (Hz)	25	25	300
d	Minimum duration (s)	0.5	0.5	0.3
k	Duration for estimating next spectral peak location (s)	0.2	0.2	0.2
f_0	Minimum frequency (Hz)	1000	50	50
f_1	Maximum frequency (Hz)	1000	50	8000

Step 3: Feature Extraction

Using custom MATLAB software, 46 features were measured from each extracted time-frequency contour. These features describe the frequency content, duration, and shape of the contour (slopes, number of inflection points, *etc*., Table 30).

Table 30. The 46 features measured from each time-frequency contour.

Feature	Definition
Beginning sweep	Slope at the beginning of the call (1=positive, -1=negative, 0=flat)
Beginning up	Binary variable: 1=beginning slope is positive, 0=beginning slope is negative
Beginning down	Binary variable: 1=beginning slope is negative, 0=beginning slope is positive
End sweep	Slope at the end of the call (1=positive, -1=negative, 0=flat)
End up	Binary variable: 1=ending slope is positive, 0=ending slope is negative
End down	Binary variable: 1=ending slope is negative, 0=ending slope is positive
Duration	Call duration (s)
Beginning frequency	Frequency at start of call (Hz)
End frequency	Frequency at end of call (Hz)
Minimum frequency, f_{min}	Minimum frequency (Hz)
Maximum frequency, f_{max}	Maximum frequency (Hz)
Frequency range	$f_{max} - f_{min}$ (Hz)
Mean frequency	Mean of frequency values (Hz)
Median frequency	Median of frequency values (Hz)
Standard deviation frequency	Standard deviation frequency values (Hz)
Frequency spread	Difference between the 75th and 25th percentiles of the frequency
Quarter frequency	Frequency at one-quarter of the duration (Hz)
Half frequency	Frequency at one-half of the duration (Hz)
Three-quarter frequency	Frequency at three-quarters of the duration (Hz)
Center frequency, f_c	$(f_{max} - f_{min})/2 + f_{min}$
Relative bandwidth	$(f_{max} - f_{min})/f_c$
Maxmin	f_{max} / f_{min}
Begend	Beginning frequency/end frequency
Steps	Number of steps (≥10% increase or decrease in frequency over two contour pts)
Inflection points	Number of inflection points (changes from positive to negative slope or *vice versa*)
Max delta	Maximum time between inflection points
Min delta	Minimum time between inflection points
Maxmin delta	Max delta/Min delta
Mean delta	Mean time between inflection points
Standard deviation delta	Standard deviation of the time between inflection points
Median delta	Median of the time between inflection points
Mean slope	Overall mean slope
Mean positive	Mean positive slope
Mean negative	Mean negative slope
Mean absolute	Mean absolute value of the slope
Ratio posneg	Mean positive slope/Mean negative slope
Percent up	Percentage of the call having positive slope
Percent down	Percentage of the call having negative slope
Percent flat	Percentage of the call having zero slope
Up-down	Number of inflection points going from positive to negative slope
Up-flat	Number of times the slope changes from positive to zero
Flat-down	Number of times the slope changes from zero to negative
Step-up	Number of steps with increasing frequency
Step-down	Number of steps with decreasing frequency
Step-duration	Number of steps/Duration
Inflection-duration	Number of inflection points/Duration

Step 4: Classification

A random forest classifier was created for each call type (bowhead winter songs, bowhead summer moans, and beluga whistles). Each of these random forests was a binary classifier, so contours were classified as 'target species' (*i.e.*, bowhead or beluga whale) or 'other'. A random forest is a collection of decision trees that are grown using binary partitioning of the data based on the value of one of the 46 features (see Table 30) at each branch, or node. Randomness is

injected into the tree-growing process by choosing the feature to use as the splitter based on a random subsample of the features at each node (Breiman 2001).

The number of decision trees to include in each random forest was determined by empirical trials on datasets of calls extracted from annotated recordings. Recordings made during the previous year's AMPs were used to train and optimize the random forests: winter 2008–2009 AMP data for the bowhead winter song and beluga whistle detectors, and summer 2009 AMP data for the bowhead summer moan detector. Contours were detected and extracted based on parameters specific to bowhead or beluga sounds (see Table 29). Sample sizes for each trial dataset are given in Table 31. These datasets were first randomly sampled so that each class ('target species' and 'other') had equal sample size. Sampling was performed such that the proportion of species and call-types within species in the 'other' class reflected those in the full dataset. Next, a random forest analysis was run on the sampled data. The sampling and random forest analysis was repeated 100 times. The output for each random forest analysis included out-of-bag error estimates for forests consisting of one to 800 trees. To calculate out-of-bag error, each tree was grown using about two-thirds of the trial data. The remaining third of the trial data was used as the 'out-of-bag' test data to evaluate the performance of the tree. The out-of-bag error estimates were averaged over the 100 runs (Figure 151). The point at which the out-of-bag error approaches its asymptote was considered the number of decision trees to include in the random forest because after this point, little gain was made in classification success with the addition of more trees. Based on these analyses, all three random forests consisted of 300 decision trees.

Table 31. Sample size of the trial datasets used to train and optimize the random forest classifiers for each call type.

Class	Winter 2008–2009 beluga whistles	Winter 2008–2009 bowhead songs	Summer 2009 bowhead moans
Beluga	1295	24	0
Bowhead	2837	3989	754
Bearded seal	20,331	17,887	269
Non-biological noise	9443	6491	536
Ribbon seal	530	0	0
Unknown	864	1148	1177
Walrus	483	199	625
Killer whale	0	0	13

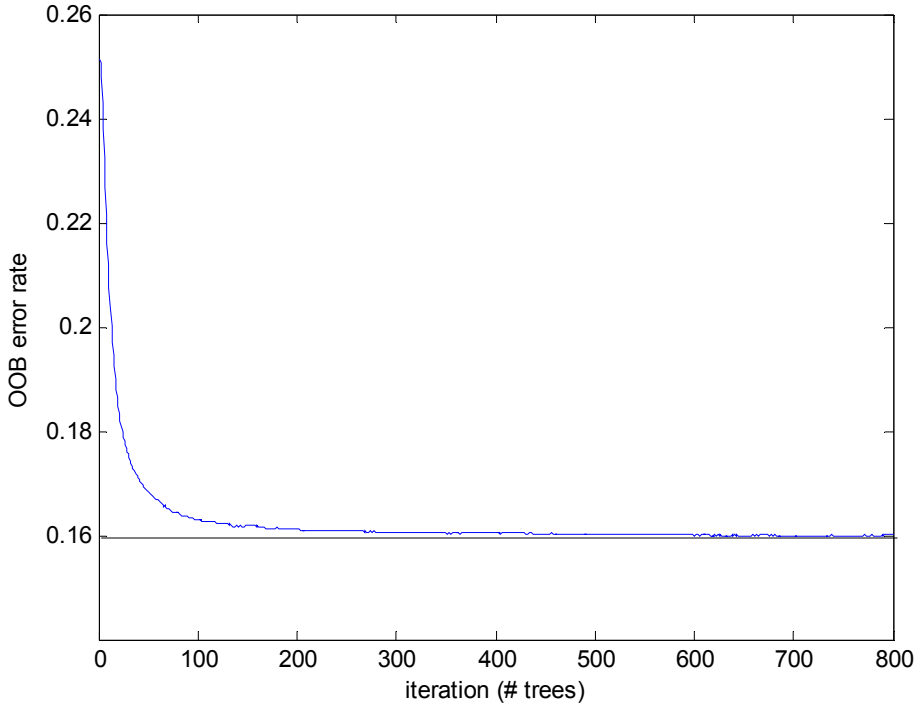

Figure 151. Out-of-bag (OOB) error rates averaged over 100 random forest runs (example of the beluga whistle classifier).

Another output of the random forest analysis is the Gini importance index (Breiman *et al.* 1984), which measures how strongly each feature contributes to the random forest model predictions. The optimal subset of features to include in each random forest was determined based on this importance index. Feature importance was averaged over all 100 runs described above (Figure 152). The features most important to the model predictions were chosen for inclusion in the three random forests (Table 32).

Table 32. Features included in bowhead moan, bowhead song, and beluga whistle random forests, listed in order of importance to the model.

Bowhead moan	Bowhead song	Beluga whistle
Minimum frequency	Maximum frequency	Mean frequency
Median frequency	Center frequency	End frequency
Mean frequency	Beginning frequency	Median frequency
Three-quarter frequency	Mean frequency	Three-quarter frequency
End frequency	End frequency	Center frequency
Half frequency	Mean slope	Half frequency
Quarter frequency	Median frequency	Maximum frequency
Beginning frequency	Quarter frequency	Quarter frequency
Duration	Three-quarter frequency	Minimum frequency
Center frequency	Half frequency	Beginning frequency
Mean negative slope	Mean absolute slope	

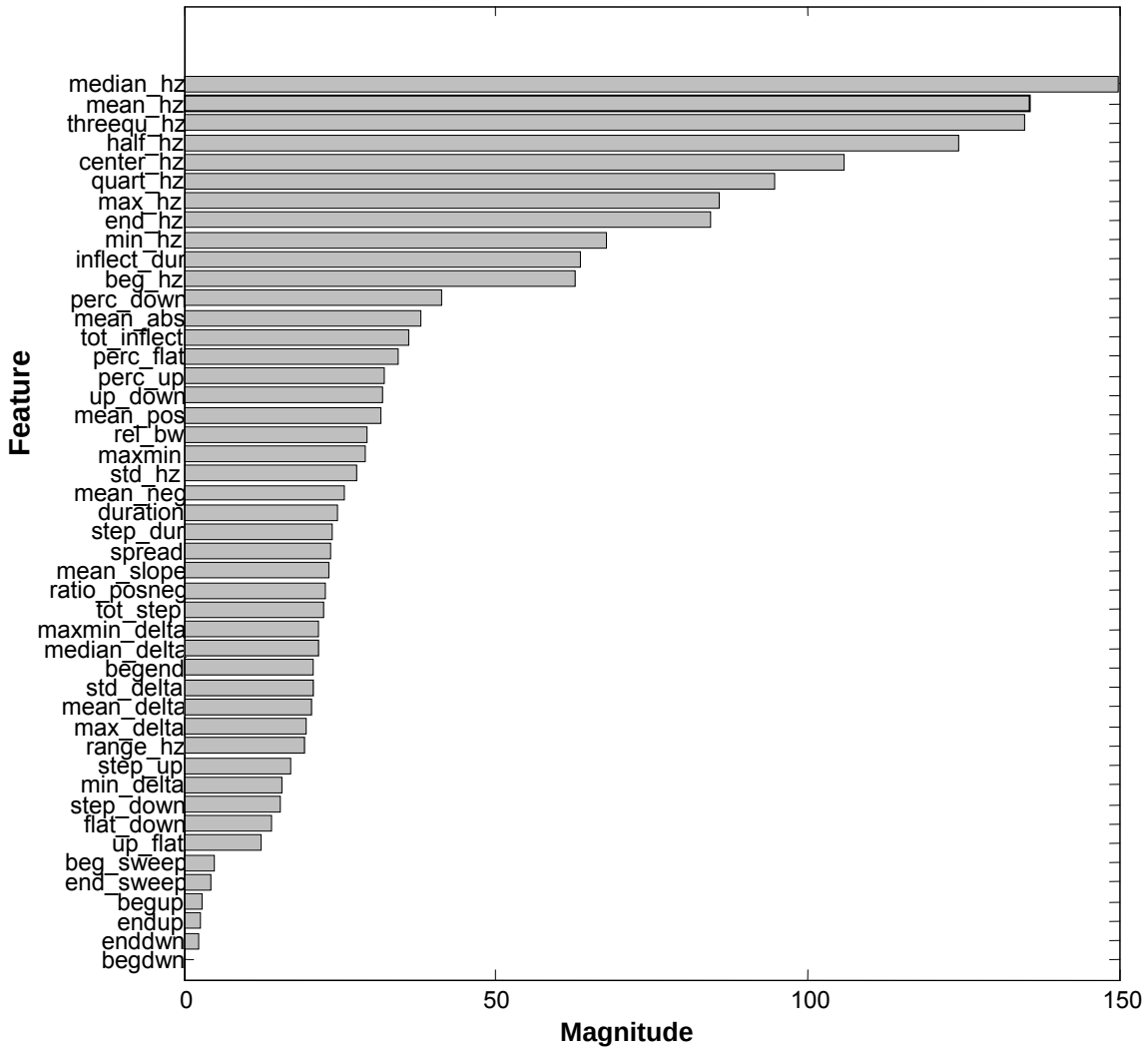

Figure 152. Gini feature importance indices, averaged over 100 random forest runs.

Step 5: Post-Processing

Bowhead calls recorded in the winter 2009–2010 AMP generally consisted of several harmonics that the automated detector considered as separate calls. This tended to overestimate the number of bowhead calls in the recordings. To avoid this, all bowhead detections overlapping in time were merged together to form a single detection. Also, only detections occurring below 300 Hz were considered. No post-processing was performed on beluga detections.

A.2.2. Walrus Grunt Detection and Classification

The algorithm first calculated the spectrogram and normalized it for each frequency band. The spectrogram was analyzed in consecutive 0.7 s frames overlapped by 50%. For each frame, a set of features representing salient characteristics of the spectrogram were extracted in the frequency band 50–800 Hz. Extracted features were presented to a two-class random forest classifier to determine the class of the sound in the analyzed frame (*i.e.*, 'walrus grunt' or 'other'). During the training phase, features of known sounds (*i.e.*, manual annotations) were extracted to create the random forest model. The detection process is illustrated in Figure 153.

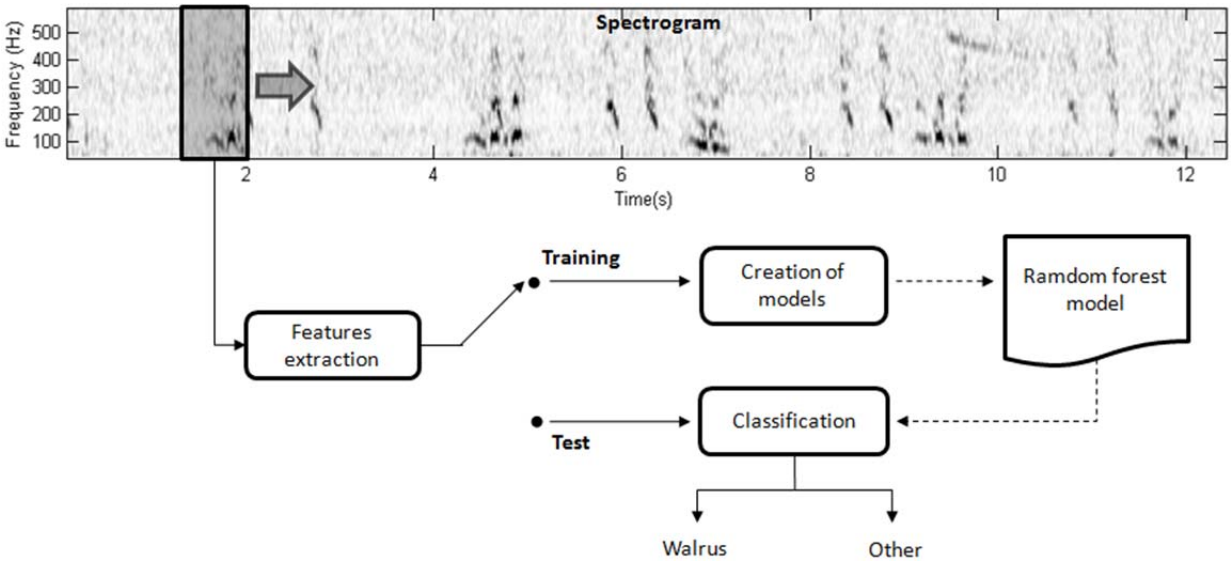

Figure 153. Steps of the walrus grunt detector.

Step 1: Spectrogram Processing

The spectrogram resolution was chosen to ensure accurate time-frequency representation of the walrus grunts (Table 33). The spectrogram was normalized by the averaged spectrum calculated over every 80 s of the recording.

Table 33. Spectrogram parameters used in the walrus grunt detector.

Spectrogram parameters	Walrus Grunts
Analysis frame size (samples)	1024
Overlap between frames (samples)	896
FFT size (sample)	2048
Window function	Blackman

Step 2: Feature Extraction

The spectrogram was analyzed in consecutive 0.7 s frames overlapped by 50%. Each 0.7 s frame was represented by 20 features. Several features were calculated following Fristrup and Watkins (1993) and Mellinger and Bradbury (2007). Features were calculated using the spectrogram (Figure 154a), frequency envelope (Figure 154b), and amplitude envelope (Figure 154c) of the signal. The frequency envelope is the sum of the spectrogram amplitude for each frequency. The maximum of the frequency envelope was normalized to 1. The amplitude envelope is the sum of the spectrogram amplitude values for each time step. The measured features are as follows:

- *Median frequency,* f_{med} (*F1*): Based on the frequency envelope. The cumulative sum of the spectrum was calculated by moving from low to high frequencies. The median frequency is the frequency at which the cumulative energy reaches 50% of the total energy (green dashed line in Figure 154b).
- *Spectral inter-quartile range* (*F2*): Calculated by defining the 25^{th} percentile of the energy on each side of the median frequency (dashed blue lines in Figure 154b). Each quartile was

defined as frequency for which the cumulative energy calculated from the median frequency equals 25% of the total energy. The spectral inter-quartile range is the difference between the higher quartile (f_{Q3}) and the lower quartile (f_{Q1}).

- *Spectral asymmetry* (*F3*): Skewness of the spectral envelope calculated as $(f_{Q1} + f_{Q3} - 2f_{med})/(f_{Q1} + f_{Q3})$.
- *Spectral concentration* (*F4*): Calculated by ranking amplitude values of the spectral envelope from largest to smallest. The cumulative sum of ranked amplitude values was computed beginning with larger values until 50% of the total energy was reached. The lowest frequency index included in the additive set was considered the minimum and the highest index was considered the maximum. Their difference provides the spectral concentration (red box in Figure 154b).
- *Maximum frequency peak* (*F5*): Frequency of the highest amplitude peak in the spectral envelope (red dot in Figure 154b).
- *Maximum frequency peak width* (*F6*): Width (Hz) of the maximum frequency peak measured at the point where amplitude values on each side of the peak reached the 75th percentile of all the spectral envelope amplitude values (red vertical line in Figure 154b).
- *Second frequency peak* (*F7*): Frequency of the second highest peak in the spectral envelope.
- *Comparison of the maximum and second frequency peaks* (*F8, F9*): Amplitude ratio and frequency difference between the maximum and second frequency peaks.
- *Variance and kurtosis of frequency envelope* (*F10, F11*): These describe the distribution of the amplitude in the spectral envelope (Balanda and MacGillivray 1988).
- *Frequency modulation index* (*F12*): Calculated as follows: First, the maximum frequency of the maximum amplitude peak was extracted for each time slice of the spectrogram. Frequency values of the selected peaks were stored in the vector F_{max} and their associated energy values in the vector E_{max}. Only peaks whose amplitude value exceeded the median amplitude of the spectrogram were considered (white dots in Figure 154a). Second, the weighted maximum frequency offset vector O was defined as $O = (F_{max} - X_{med}) \cdot E_{max}/max(E_{max})$, where X_{med} is a scalar representing the median frequency of the vector F_{max}. The frequency modulation index was defined as the standard deviation of the vector O.
- *Asymmetry of the maximum frequencies* (*F13*): The skewness of the vector O defined above.
- *Duration* (*F14*): Number of spectrogram frames with a maximum amplitude value above the 90th percentile of the amplitude values of the spectrogram. The resultant number of frames was then multiplied by the spectrogram time resolution to give the duration in seconds.
- *Amplitude modulation index* (*F15*): The 90th percentile of the first derivative of the amplitude envelope. An example of the derivative of the amplitude envelope is shown in Figure 154d.
- *Signal-to-noise ratio* (*F16*): Ratio of the 100th percentile and 25th percentile of the amplitude values of the spectrogram.

- *Overall spectral entropy* (*F17*): The Shannon entropy (Erbe and King 2008) calculated for each time slice of the spectrogram in the frequency band 50–600 Hz (Figure 154e). The overall spectral entropy is the 10^{th} percentile of these values.
- *Kurtosis of the spectral entropy* (*F18*): Kurtosis of the Shannon entropy values calculated on each time slice of the spectrogram.
- *Minimum of the spectral entropy* (*F19*): Minimum of the Shannon entropy values calculated on each time slice of the spectrogram.
- *Overall harmonicity* (*F20*): Harmonicity was calculated for each time slice of the spectrogram by calculating the Shannon entropy of the Harmonic Product Spectrum (*e.g.*, Figure 154f; see Ding *et al.* 2006). Low harmonicity means the frequency content of the analyzed signal is harmonic. The overall harmonicity is the 10^{th} percentile of all the harmonicity values.

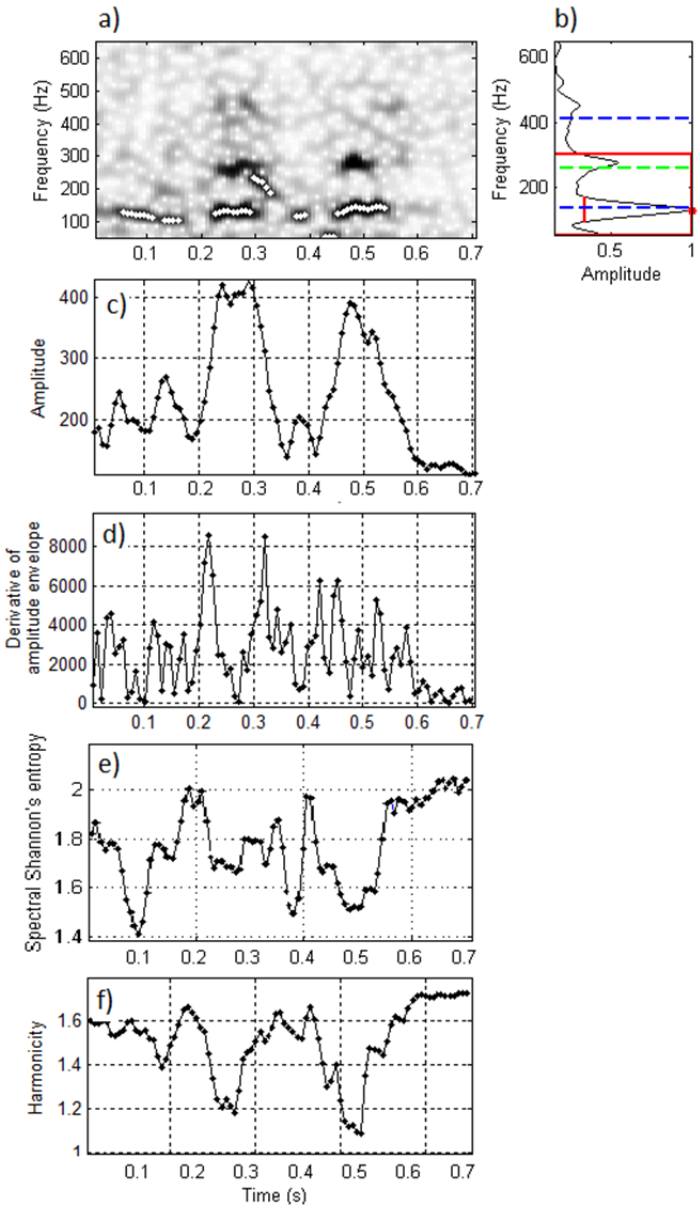

Figure 154. Extraction of features used in the walrus grunt classifier: (a) spectrogram of the analyzed frame; (b) Frequency envelope (black line), with the median frequency (green line), the upper and lower quartiles (blue lines), the maximum frequency peak (red dot), and the spectral concentration (red box); (c) Amplitude envelope; (d) first derivative of the amplitude envelope; (e) spectral entropy; (f) harmonicity index.

Step 3: Classification

Classification was performed using a random forest classifier (Breiman 2001). The random forest classifier was trained using all manual annotations in recordings from the summer 2009 AMP. The random forest was defined with two classes, 'walrus grunt' and 'other'. Training of the classifier, optimization of the number of decision trees in the forest and the selection of the most relevant features based on the Gini index were performed using the same process as described for bowhead and beluga call detection (Section A.2.1). The optimal number of

decision trees was 600. The importance of the features is illustrated in Figure 155. Because feature importance did not decrease abruptly, all 20 features were used for classification.

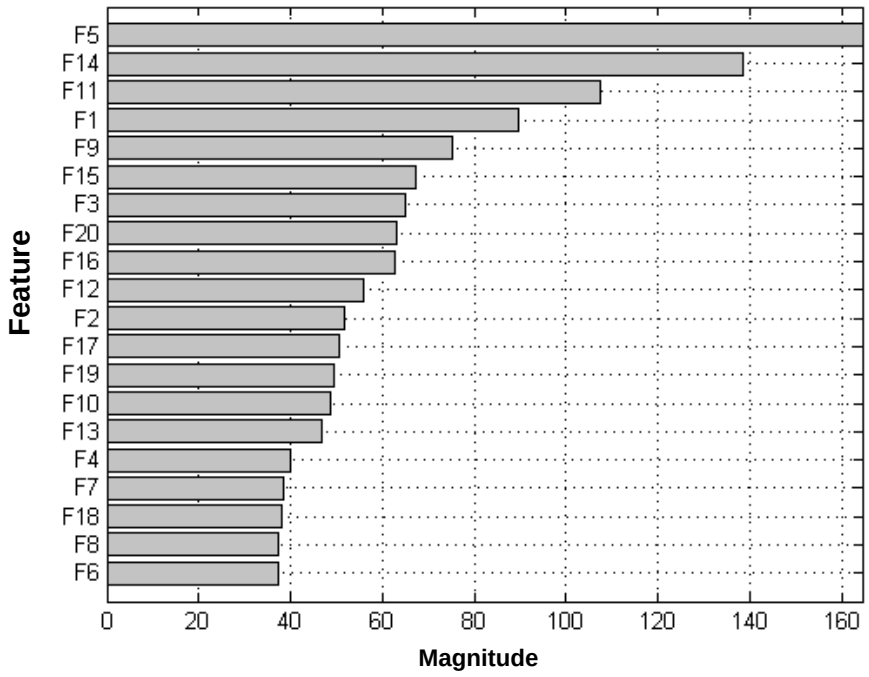

Figure 155. Gini feature importance indices averaged over 100 random forest runs.

Step 4: Post-Processing

Seismic airgun pulses can be distorted by propagation effects and appear similar to walrus grunts. To minimize false alarms due to seismic pulses, walrus grunt detections concurrent with airgun shots detected by the seismic detector (see Section 2.2.2.2) were removed.

A.2.3. Bearded Seal Call Detection

The automated detection and classification of bearded seal calls is performed in four steps: 1) the calculation and binarization of the spectrogram, 2) the definition of time-frequency objects, 3) the feature extraction, and 4) the classification.

Step 1: Spectrogram Processing

The first step of the detection process was the calculation of the spectrogram. The spectrogram parameters used are in the Table 34. To attenuate long spectral rays in the spectrogram due to vessel noise and to enhance weaker transient biological sounds, the spectrogram was normalized in each frequency band (*i.e.*, each row of the spectrogram) with a median normalizer (see Section 6.1). The size of the window used by the normalizer is indicated in Table 34. The normalized spectrogram was binarized by setting all the time-frequency bins exceeding a normalized amplitude of 4 (no unit) to 1 and the other bins to 0.

Table 34. Spectrogram parameters

	Bearded seal calls
Analysis frame size (samples)	4096
Overlap between frames (samples)	3072
FFT size (sample)	4096
Window function	Reisz
Normalizer window size (s)	120
Binarization threshold (no unit)	4

Step 2: Definition of Time-Frequency Objects

The second step of the detection process consisted in defining time-frequency objects (or events) by associating together contiguous bins in the binary spectrogram. The algorithm implemented is a variation of the *flood-fill* algorithm (Nosal 2008). Every spectrogram bins that equals 1 and separated by less than 3 bins in both time and frequency are connected together. Figure 156 illustrates the search area used to connect a spectrogram bin to another one. The bin connection process moves from oldest data to newest and from lowest frequency to highest. Also, a spectrogram bin can only belong to a single time-frequency object. Each group of connected bins is referred to as a *time-frequency object*.

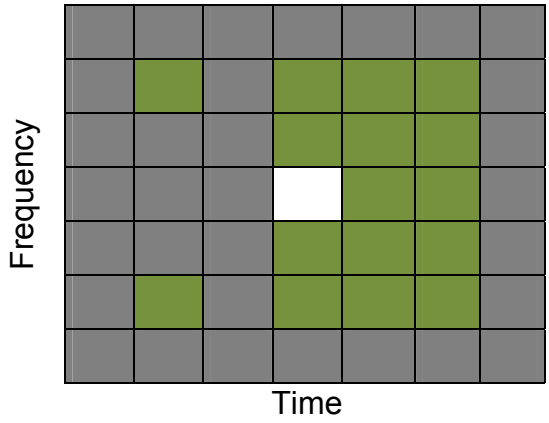

Figure 156. Illustration of the search area used to connect spectrogram bins together. The white bin represents a bin of the binary spectrogram equaling 1 and the green bins represent the potential bins that it could be connected to. The algorithm advances from left to right, therefore gray cells left of the test cell need not be checked. However, checking the far left cells may join broken contours.

The definition of time-frequency objects is sensitive to noise generated by small pleasure craft or fishing vessels near a recorder, which can generate many time-frequency objects that may be mistaken for marine life calls. Therefore, a vessel detector is incorporated into the time-frequency event definition process to reduce false detections. Vessel noise is considered detected when at least five frequencies have detected contours for 5 s. Files with at least two vessel detections are omitted from further processing.

Step 3: Feature extraction

The third step consists of representing each of the time-frequency objects extracted in the previous step by a set of features. Features were defined as the start time (date), the duration (s), the minimum and maximum frequency (Hz), and the bandwidth (Hz) of the time-frequency objects.

Step 4: Classification

The final step consists of classifying the time-frequency objects by comparing their features against a dictionary defining the features of the vocalizations present in the Chukchi sea based on the literature and analysts observations. In the present study, only bearded seal calls were represented in the dictionary (Table 35). Notice that the classification process has the ability to handle vocalizations that are made of several time-frequency objects such as vocalizations with harmonics (referred to as *MultiFrequencyComponents*) and vocalizations made of a succession of time-frequency objects such as seal trills and groups of beluga, dolphin, or beaked whale whistles (referred to as *MultiTimeComponents*). Vocalizations in the dictionary are defined by the following features:

1. Minimum frequency
2. Maximum frequency—either the maximum frequency expected for the call type, or the maximum frequency in the data, whichever is lower.
3. Minimum duration—at least one spectrogram time slice.
4. Maximum duration.
5. Minimum bandwidth.
6. Maximum bandwidth—not often used.
7. MultiFrequencyComponent (Boolean): for call types where contours should be grouped in frequency with some time overlap before applying the frequency, duration, and bandwidth constraints. Each contour that is added to the multi-component contour has the following constraints applied:
    a. minComponentDuration—minimum duration for a contour to be added to the multi-component contour.
    b. minComponentBW—minimum bandwidth for a contour to be added to the multi-component contour.
    c. Minimum and maximum frequencies as per the global definition.
8. MultiTimeComponent (Boolean): for call types where contours should be grouped in time before applying the frequency, duration, and bandwidth constraints. Each contour that is added to the multi-time-component contour has the following constraints applied:
    a. minTimeComponentDuration—minimum duration for a contour to be added to the multi-time-component contour.
    b. minTimeComponentBW—minimum bandwidth for a contour to be added to the multi-time-component contour.
    c. Minimum and maximum frequencies as per the global definition.

Table 35. Dictionary defining the time-frequency features of bearded seal calls in the Chukchi sea in the summer and in the winter

Species	Call Type	Min / Max frequency (Hz)	Min / Max duration (s)	Min / Max bandwidth (Hz)	Min / Max sweep rate	Multi-Frequency-component settings	Multi-time-component settings
Bearded seal – winter calls	Full Trill	250 / 5000	5 / 60	500 / -	-100 / -10	Min BW = 30 Max BW = 200 Min Dur = 0.5 Max Dur =5 MaxFreqShift = 100	0
	Trill end	250 / 1200	10 / 60	100 / -	-50 / -5	Min BW = 20 Max BW = 100 Min Dur = 0.5 Max Dur = 8 MaxFreqShift = 100	0
Bearded seal – summer calls	Downsweep	200 / 1500	0.6 / 10	38 / -	-200 / -20	N/A	0
	Upsweep	200 / 1500	0.6 / 4.5	100 / -	50 / 250	N/A	0

Figure 157 shows a block diagram of the several stages of the classification algorithm.

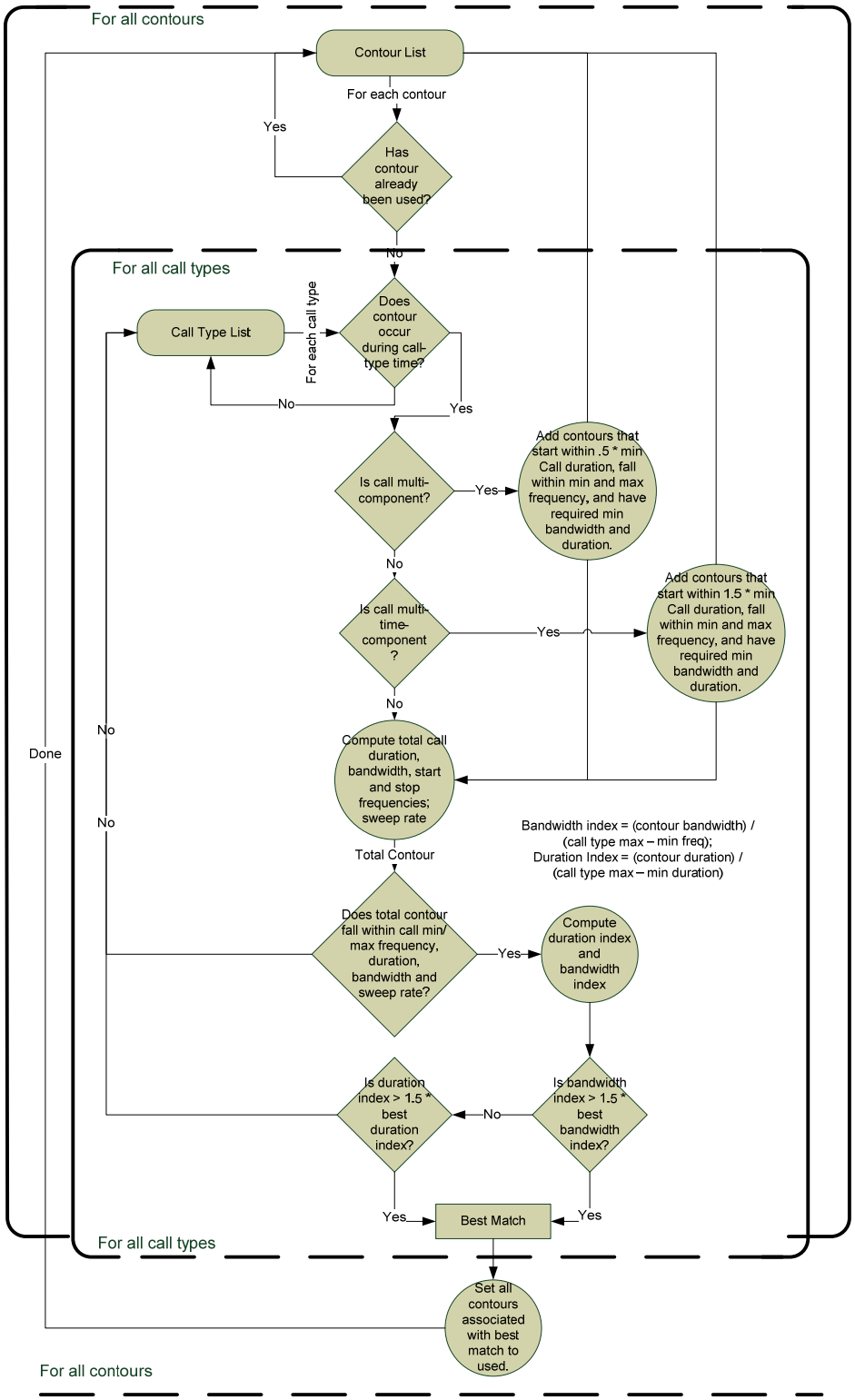

Figure 157. Block diagram of the classification algorithm.

The algorithm consists of two loops. The outer loop iterates through all the time-frequency objects. For each time-frequency object that has not yet been classified, the object's features are compared to each call in the dictionary. If the call is a multi-frequency-component or multi-time-component type, the list of time-frequency objects is searched for unsorted objects that meet the multi-components settings (see Table 35). The total time-frequency object duration, minimum and maximum frequencies, and frequency bandwidth are compared to the calls definitions in the dictionary. If the object features fall within the call type's bounds, then the bandwidth (BW_i) and duration (T_i) indices are computed:

$$BW_i = \frac{BW_{object}}{BW_{dictionary}} \qquad T_i = \frac{T_{object}}{T_{dictionary}}$$

If either of these indices exceed an empirically chosen threshold of 1.5 times the current best index, then the current best-match call type is updated. The 1.5 threshold for updating the best-match call type means that the algorithm prefers call types that are defined earlier. Therefore if for a particular recording, killer whales are more likely to occur than singing humpbacks, the killer whale call definitions should occur first in the *mammalContours.xml* definition file. **Error! Reference source not found.** is an example of all three types of contours applied to dolphin calls.

The classification algorithm also implements a time-based filter. Since the classification algorithm is intended to count calls of species that are expected in an area, it is reasonable to limit the algorithm with a priori knowledge. For instance, we will not detect any bowhead calls before 1 Sep or after 1 Jan in the Chukchi sea. The detection of extra-limital species and unusual detections as a function of time is left to the manual analysis.

Figure 158 shows an example of detection and classification of bearded seal calls.

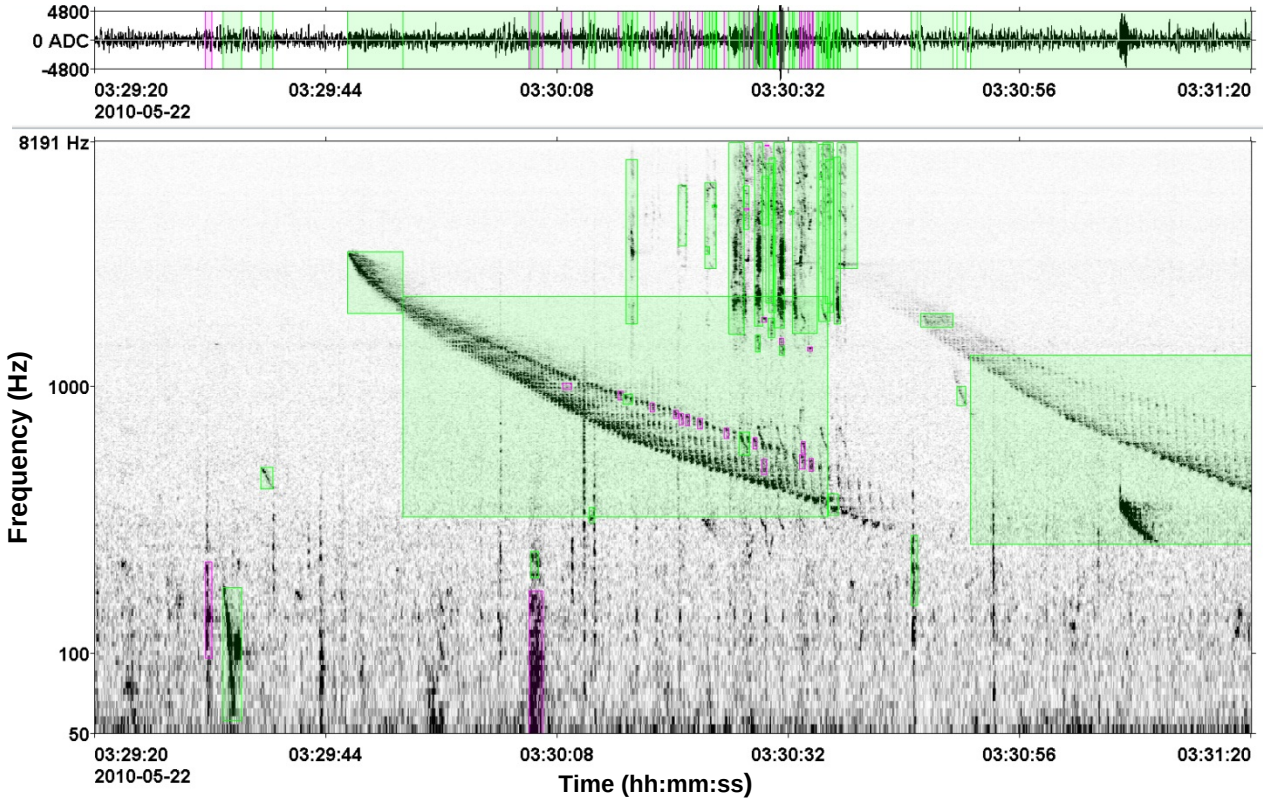

Figure 158. Pressure in digital units (top) and spectrogram (bottom) of bearded seal trills (500-200 Hz; downsweeps in center) detected using the multi-time component contour type. Beluga and bowhead calls are visible in this figure as well. (16 kHz sample rate, 4096-pt STFT, 1024-pt advance).

A.2.4. Performance Evaluation

Test Datasets

The automated detectors/classifiers must be verified with a test dataset that represents the spatio-temporal variations of the marine mammal calls and background noise in the entire dataset. Since the acoustic environment in the eastern Chukchi Sea differs between winter and summer, a unique test dataset was used to test the detection/classification algorithms for each season. For the winter 2009–2010 AMP data, marine mammal calls were fully manually annotated in the first 2 min of each day for recordings from Stations B05, CL50, PL50, W35, and WN40. This yielded a test dataset of 1376 2 min fully-annotated samples. For the summer 2010 AMP data, marine mammal calls were fully manually annotated in the first 1.5 min after midnight and the first 1.5 min after noon of each day for Stations B05, B30 (until 23 Oct 2010), W05 (until 25 Aug 2010), W35, WN20A (until 6 Oct 2010), CL20, CLN90, KL01, PL20, PL50, and SO01. This yielded a test dataset of 1779 1.5 min fully-annotated samples.

Performance Metrics

The decisions made by detectors/classifiers can be represented as a confusion matrix. The confusion matrix consists of four categories: true positives (*TP*), false positives (*FP*), true negatives (*TN*) and false negatives (*FN*). Table 36 depicts the confusion matrix, where *E* is the

signal event we want to detect/classify and $\overline{E}$ is a non-event that we want to ignore (*i.e.*, noise). The definition of $\overline{E}$ varies depending on the detector or classifier.

Table 36. Confusion matrix.

		True result	
		E	$\overline{E}$
Detection/ classification result	E	*TP*	*FP*
	$\overline{E}$	*FN*	*TN*

A true positive (*TP*) corresponds to a signal of interest being correctly classified as such. A false negative (*FN*) is a signal of interest being classified as noise (*i.e.*, missed). A false positive (*FP*) is a noise classified as a signal of interest (*i.e.*, a false alarm). A true negative (*TN*) is a noise correctly classified as such.

The numbers of *TP*s, *FP*s, and *FN*s were calculated for each detector and test dataset by comparing the manual annotations of marine mammal calls (considered true results, *i.e.*, ground truth) with the automated detections/classifications. Numbers of *FP*s, *TP*s and *FN*s were calculated on all dataset samples containing annotations of the target call type. If a manually-annotated call was automatically detected/classified, then the detection was considered a *TP*, if undetected, it was a *FN*. Each automated detection occurring in the sample that did not correspond to a manually-annotated call was considered a *FP*.

Precision and Recall

To assess the performance of the detectors/classifiers, precision (*P*) and recall (*R*) metrics were calculated based on the numbers (*N*) of *TP*s, *FP*s, and *FN*s:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \qquad R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{1}$$

P can be seen as a measure of exactness, and *R* is a measure of completeness. For instance, a *P* for beluga of 0.9 means that 90% of the detections classified as beluga were in fact beluga calls, but says nothing about whether all beluga vocalizations in the dataset were identified. An *R* for beluga of 0.8 means that 80% of all beluga calls in the dataset were classified, but says nothing about how many beluga classifications were wrong. Thus, a perfect detector/classifier would have $P = R = 1$. Neither *P* nor *R* alone can describe the performance of a detector/classifier on a given dataset; both metrics are required.

The *P*-*R* metric presents advantages over the True-Positive Rate (TPR) and False-Positive Rate (FPR) generally used in Receiver Operating Characteristic (ROC) curves. Firstly, the *P*-*R* metric is more adapted to skewed datasets. Secondly, it has been demonstrated that an algorithm dominates in ROC space if and only if it dominates in *P*-*R* space (Davis and Goadrich 2006). Finally, a significant advantage of *P*-*R* values over ROC values comes in defining N_{TN} in

continuous data. A subjective criterion is necessary to define the length of time that counts as one *TN* value over a continuous recording that contains no targeted vocalizations, whereas N_{TN} need not be calculated for the *P-R* metric. Therefore, using *P-R* values is better suited to the analysis of these time-continuous data.

Signal-to-Noise Ratio

The signal-to-noise ratio (SNR) is the ratio of signal power (P_s) to noise power corrupting the signal (P_n). It compares the level of the desired signal to the level of the background noise. The greater this ratio, the less obtrusive the background noise. SNR is defined in decibels as:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{P_s}{P_n}\right) \tag{2}$$

The signal power of a call in a spectrogram is the average power of the call within the frequency range of the vocalization, and the noise power is the average power before and after the call within the same frequency band (Mellinger 2004; Mellinger and Clark 2006). The duration of the noise signal measured before and after the call equals the duration of the call (Figure 159). This calculation was performed on the original spectrogram without noise reduction. To quantify detector performance for various SNRs, N_{FN} and N_{TP} were calculated for SNR intervals of < 0 dB, 0–5 dB, 5–10 dB, and > 10 dB. Values of *P* are influenced by the background noise and not by the SNR of the calls. Therefore *P* values per SNR intervals were not calculated since these values are less relevant.

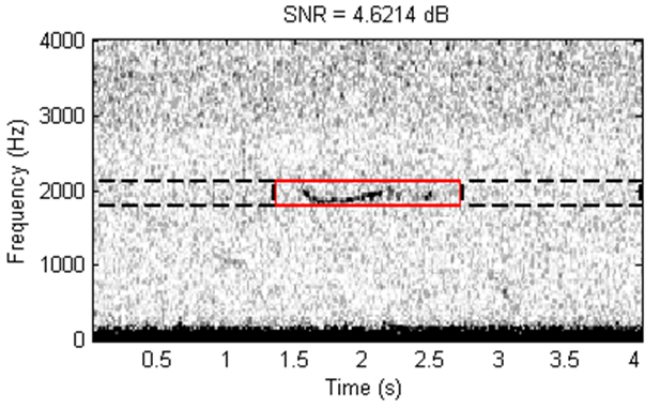

Figure 159. Calculation of the signal-to-noise ratio (SNR). The power of the call (P_s) is calculated in the red box and the power of the noise (P_n) is calculated in the black boxes on either side of the call.

A.2.5. Call Count Estimation

Because the detectors/classifiers are imperfect (having false alarms and missed calls), the number of automated detections is not exactly equal to the actual number of calls present in the recordings. A better estimate can be achieved using *P* and *R*. These values characterize the relationship between the detector/classifier and the dataset. Therefore, these values are specific to, and depend on, both the detector/classifier and the dataset. Provided that the subset of data used to characterize *P* and *R* are representative of the entire dataset, *P* and *R* can be used to extrapolate the total number of vocalizations from the number of detected vocalizations. The

total number of detections (N_{det}) found by the detector/classifier is the sum of the number of true and false positives:

$$N_{\text{det}} = N_{TP} + N_{FP} \tag{3}$$

and from the definition of P (Equation 1), N_{TP} can be defined as:

$$N_{TP} = P \cdot (N_{TP} + N_{FP}) = P \cdot N_{\text{det}} \tag{4}$$

The total number of vocalizations in the data (N_{voc}) is the sum of those correctly identified (*TP*) and those that were missed (*FN*):

$$N_{voc} = N_{TP} + N_{FN} \tag{5}$$

Therefore R (Equation 1) becomes:

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} = \frac{N_{TP}}{N_{voc}} \tag{6}$$

Combining Equations 4 and 6 yields the total number of vocalizations in terms of P, R, and the number of detections:

$$N_{voc} = \frac{N_{TP}}{R} = \frac{P \cdot N_{\text{det}}}{R} \tag{7}$$

All call-count estimation plots in the main report (bubble-plots) were produced using Equation 7.

A.3. Detector/Classifier Performance Results

The performance of each automated detector/classifier is provided for test datasets of both the winter 2009–2010 and summer 2010 AMPs. The test datasets consist of all fully manually-annotated data samples for each AMP. For each detector/classifier and AMP season dataset, the precision (*P*) and recall (*R*) of the detector/classifier on the entire test dataset are given. The SNR distribution of the test dataset over four SNR intervals and the *R* value calculated for each SNR interval are shown in Figure panels (a) and (b), respectively.

A.3.1. Bowhead Winter Songs

The bowhead winter song detector/classifier was tested against the fully manually-annotated recordings of the winter 2009–2010 AMP. The test dataset had a total of 1006 manually-annotated bowhead songs. The performance of the bowhead song detector/classifier on the test dataset yielded $P = 0.5$ and $R = 0.44$. As expected, the detector/classifier was able detect more calls at higher SNRs. The highest *R* value was 0.7, obtained for calls with SNR > 10 dB.

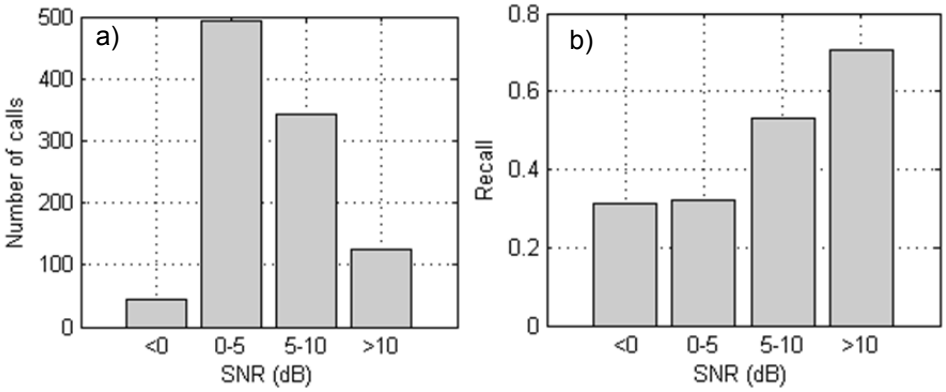

Figure 160. Performance of the bowhead winter song detector/classifier on the winter 2009–2010 test dataset: (a) signal-to-noise ratio (SNR) distribution of calls in the test dataset; (b) Recall of the detector/classifier per call SNR interval.

A.3.2. Bowhead Summer Moans

The bowhead summer moan detector/classifier was tested against fully-annotated recordings collected during the summer 2010 AMP. The test dataset had a total of 406 manually-annotated bowhead moans. The performance of the bowhead moan detector/classifier on the test dataset yielded $P = 0.84$ and $R = 0.22$. As expected, *R* increased with increasing SNR.

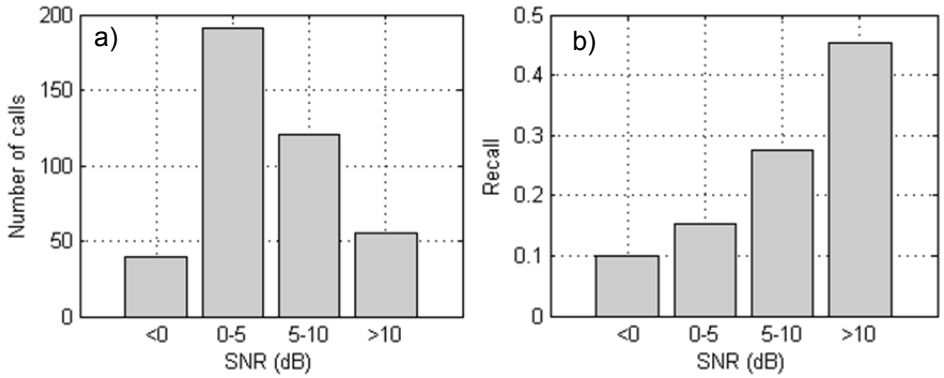

Figure 161. Performance of the bowhead summer moan detector/classifier on the summer 2010 test dataset: (a) signal-to-noise ratio (SNR) distribution of calls in the test dataset; (b) Recall of the detector/classifier per call SNR interval.

A.3.3. Beluga Whistles

The beluga whistle detector/classifier was used for analysis of only the winter 2009–2010 AMP data because no beluga whistles occurred in the summer 2010 AMP data. The test dataset had a total of 2191 manually-annotated beluga whistles. Most annotated whistles had a SNR between 0 and 5 dB. The beluga whistle detector/classifier had P = 0.66 and R = 0.26. R for calls with a SNR < 0 dB is higher than that for calls with a SNR of 0–5 dB due to bias in the estimation of SNR for concurrent beluga whistles. The highest R was 0.75, obtained for whistles with SNR > 10 dB.

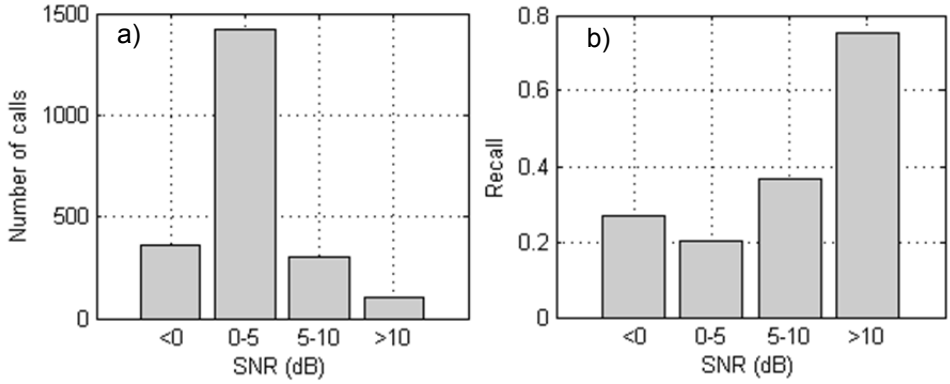

Figure 162. Performance of the beluga whistle detector/classifier on the winter 2009–2010 test dataset. (a) signal-to-noise ratio (SNR) distribution of calls in the test dataset. (b) Recall of the detector/classifier per call SNR interval.

A.3.4. Walrus Grunts

Walrus grunts were recorded only in summer, which included the last few days of the winter 2009–2010 AMP (*i.e.*, late June) and the entire summer 2010 AMP. Consequently, the performance of the walrus grunt detector/classifier was calculated using the summer 2010 and winter 2009–2010 test datasets combined (*i.e.*, one set of P and R values for both datasets). The combined test dataset had a total of 2228 manually-annotated walrus grunts. Most annotated calls had low SNR (1500 annotations with SNR = 0–5 dB). The detector/classifier had P = 0.52 and R = 0.26. R increased gradually with increasing SNR.

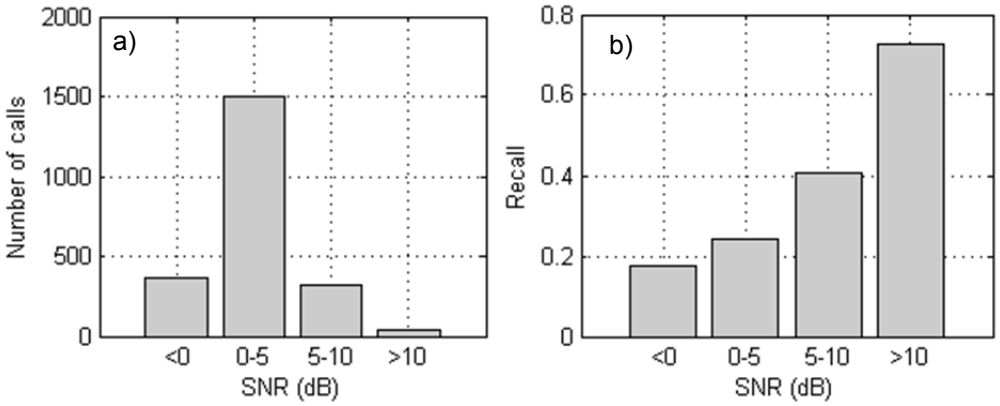

Figure 163. Performance of the walrus grunt detector/classifier on the winter 2009–2010 and summer 2010 test datasets. (a) signal-to-noise ratio (SNR) distribution of calls in the combined test datasets. (b) Recall of the detector per call SNR interval.

A.3.5. Bearded Seal Calls

Bearded seal calls were detected and classified in both winter 2009–2010 and summer 2010, with a greater vocal presence in the winter. The performance of the bearded seal call detector/classifier was evaluated separately for each AMP season.

Winter 2009–2010 AMP

The winter 2009–2010 AMP test dataset had a total of 6344 manually-annotated bearded seal calls. *P* and *R* were calculated on many more calls for the winter test dataset than for the summer (6344 *vs*. 86, respectively) due to high vocal presence of bearded seals in winter. The bearded seal call detector/classifier had $P = 0.68$ and $R = 0.5$ for the winter 2009–2010 AMP test dataset.

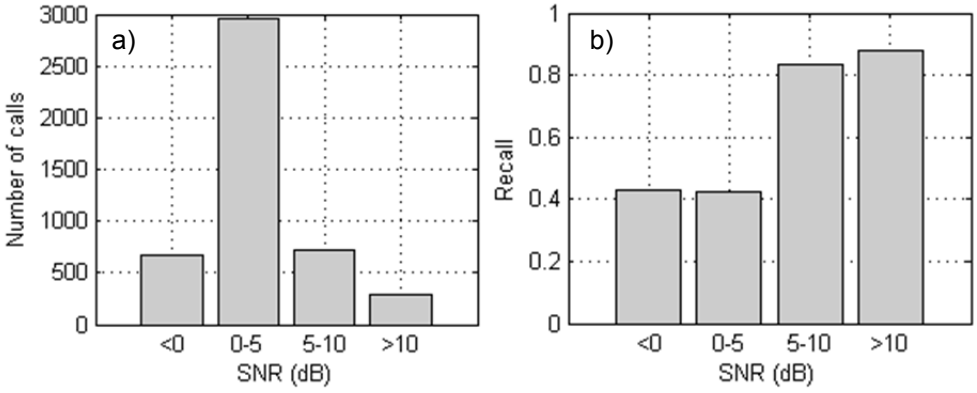

Figure 164. Performance of the bearded seal detector/classifier on the winter 2009–2010 test dataset. (a) signal-to-noise ratio (SNR) distribution of calls in the test dataset. (b) Recall of the detector/classifier per call SNR interval.

Summer 2010 AMP

The summer 2010 AMP test dataset had a total of 86 manually-annotated bearded seal calls. The detector/classifier has $P = 0.65$ and $R = 0.17$. *R* for calls with SNR greater than 10 dB is null because few manually-annotated bearded seal calls had SNR greater than 10 dB.

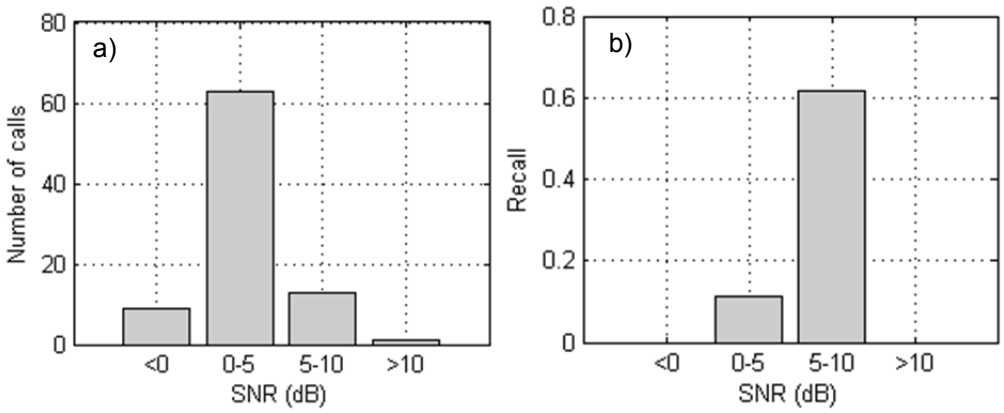

Figure 165. Performance of the bearded seal detector/classifier on the summer 2010 test dataset. (a) signal-to-noise (SNR) distribution of calls in the test dataset. (b) Recall of the detector/classifier per call SNR interval.

A.3.6. Summary

Table 37. Precision (*P*) and recall (*R*, for all SNRs) of each detector/classifier.

Detector	*R*	*P*
Bowhead winter songs	0.44	0.5
Bowhead summer moans	0.22	0.84
Beluga whistles	0.26	0.66
Walrus grunts	0.26	0.52
Bearded seal, summer	0.17	0.65
Bearded seal, winter	0.5	0.68

A.4. Discussion

Performance calculations are essential in developing automated acoustic detectors and classifiers. It quantifies how well the detector/classifiers work and allows estimation of the total number of calls present in recordings (both detected and undetected). Detector/classifier performance can also be evaluated by comparing the number of automated detections per day with the daily acoustic presence/absence of the target species (based on manual annotations) over a long period of time. Although this approach is less quantitative than the precision (*P*) and recall (*R*) metrics, it can provide context for the calculated performance metrics.

A.4.1. Bowhead Winter Songs and Summer Moans

Figure 166 shows the manual and automated detection/classifications of bowhead moans for the summer 2010 Station S01. Although the bowhead winter song and summer moan detectors/classifiers both had $R < 0.5$, they allowed the acoustic presence-absence of bowheads to be captured for most days during which they were detected manually. Figure 166 shows few false alarms occured. Most false alarms in the summer 2010 AMP data were caused by noise from the mooring. Figure 167 shows an example of mooring line noise detected and classified as bowhead moans. Such mooring noise was less common in the 2009 data than in the 2008 data. Therefore, because the bowhead summer moan random forest was trained with the 2009 data, mooring noise was under-represented during creation of the classification model. Further work will include accounting for mooring noise during training of the classification algorithm to reduce the number of false alarms.

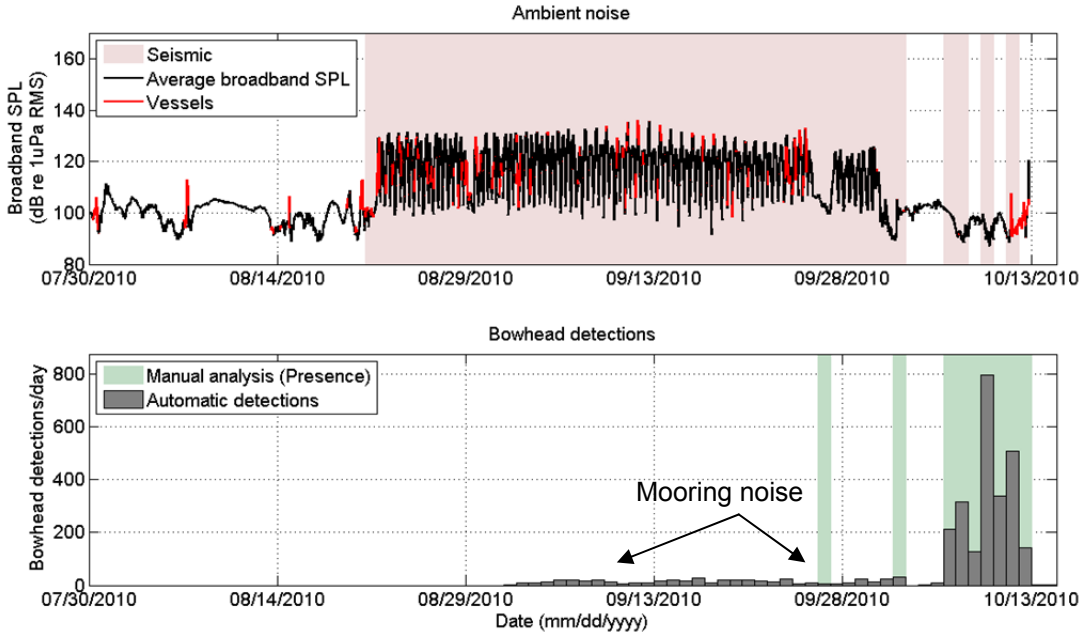

Figure 166. Detection/classification of bowhead summer moans at Station S01, 30 Jul to 13 Oct 2010: (top) Average broadband sound pressure level (SPL) of ambient noise, presence of seismic survey activity (from the automated seismic detector), and presence of vessels with time, and (bottom) number of automated detection/classifications compared to presence/absence of manual detections of bowhead summer moans.

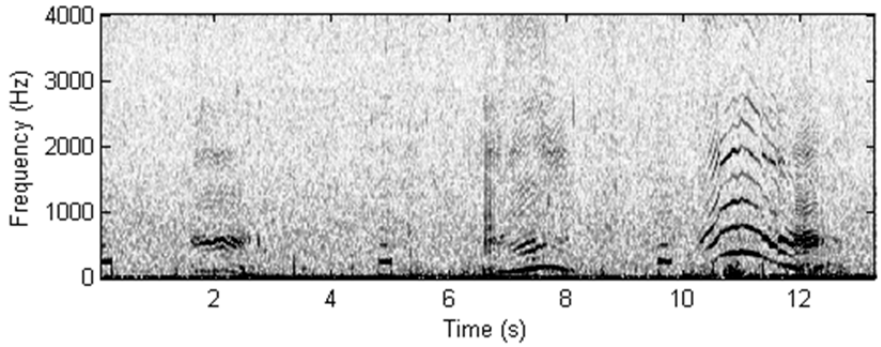

Figure 167. Spectrogram of mooring line noise falsely detected and classified as bowhead moans.

A.4.2. Beluga Whistles

Figure 168 compares the manual and automated detections/classifications of beluga whistles for winter 2009–2010 Station B05. No seismic activity was detected during this period. Most false beluga detections were due to ice noise. Even though ice noise was considered during training of the classification algorithm, some ice recordings that were very similar in duration and frequency to beluga calls (Figure 169) were falsely detected.

A possible solution to avoid false alarms by the beluga detector is to add an 'ice' class to the random forest classifier. This would better represent ice sounds in the classification model and avoid them being swamped by calls of other species in the 'other' class. Alternatively, the random forest model could be created such that the proportions of contours for each class in the 'other' category are equal, rather than representing that found in the performance test dataset (see 6.2.A.2.1). Both possibilities will be investigated in future.

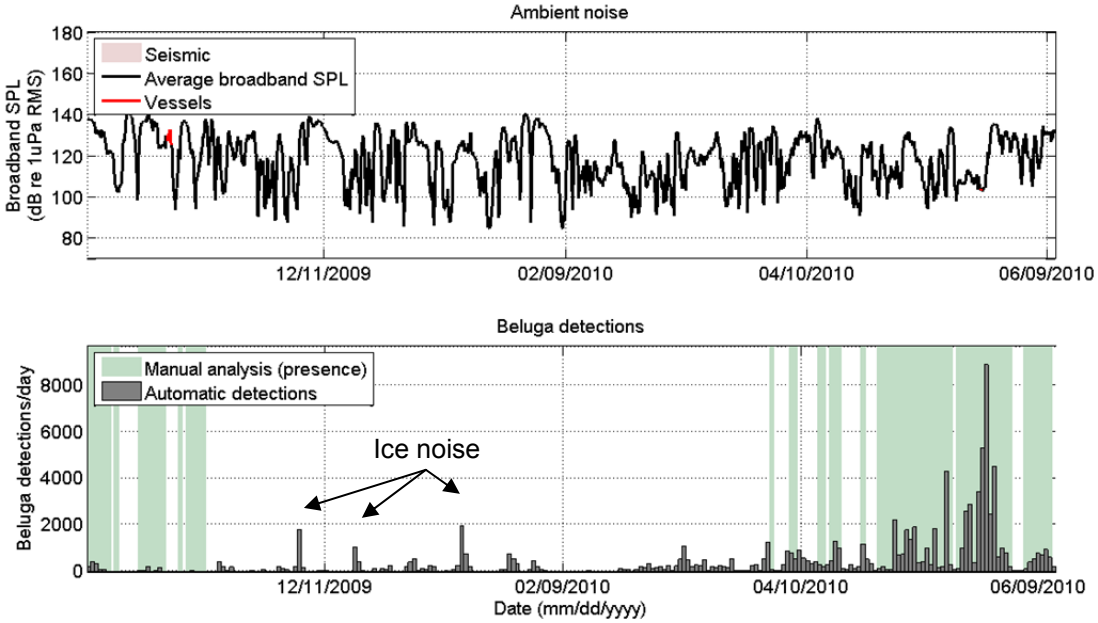

Figure 168. Detection/classification of beluga whistles at winter 2009–2010 AMP Station B05, 12 Nov 2009 to 9 Jun 2010: (top) Average broadband SPL of ambient noise, and the presence of vessels with time (no seismic activity occurred); and (bottom) number of automated detections/classifications compared to presence-absence of manual detections of beluga whistles.

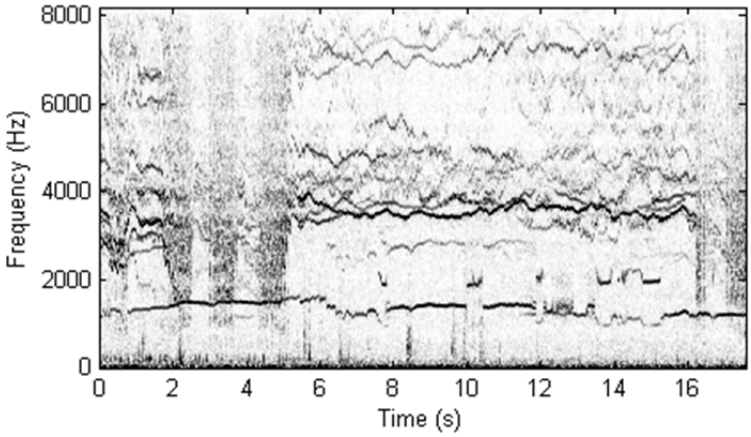

Figure 169. Spectrogram of ice squeaking noise falsely detected and classified as beluga calls.

A.4.3. Walrus Grunts

Figure 170 compares the manual and automated detections/classifications of walrus grunts for summer 2010 AMP Station W35. Most walrus false alarms were caused by seismic pulses. To minimize these false alarms, detection/classifications concurrent with automated seismic detections were removed in post-processing.

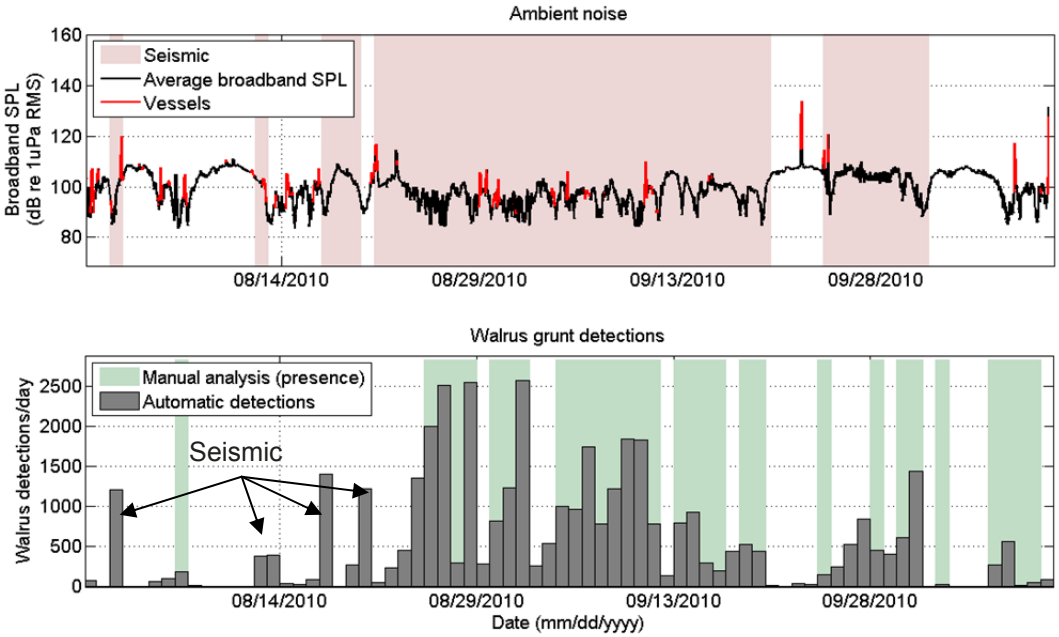

Figure 170. Detection/classification of walrus grunts at summer 2010 AMP Station W35, 30 Jul 2010 to 11 Oct 2010: (top) Average broadband sound pressure level (SPL) of ambient noise, presence of seismic survey activity (from the automated seismic detector), and presence of vessels with time, and (bottom) number of automated detection/classifications compared to presence/absence of manual detections of walrus grunts.

A.4.4. Summary

The detector/classifier performances depend greatly on the choice of parameters, such as spectrogram resolution (*i.e.*, FFT size, overlap, *etc*.). Most parameters were chosen empirically by testing a set of parameters on a small trial dataset of the previous year's AMP detections and choosing the set of parameters that provided the best detection results. The large number of detection parameters precludes testing all possible combinations of said parameters. Therefore, future work should include optimization of algorithms to determine the best set of parameters. The detector/classifier performances also depend on the choice of features used to characterize the calls. The bowhead and beluga classifiers used very different features than the walrus classifier, and these are just two examples of feature sets that could be chosen. Many additional or alternate features are possible, which will be investigated in future.

A.5. Literature Cited

Abbot, T.A., V.E. Premus, and P.A. Abbot. 2010. A real-time method for autonomous passive acoustic detection classification of humpback whales. *J. Acoust. Soc. Am.* 12:2894–2903.

Balanda, K. and H. MacGillivray. 1988. Kurtosis: a critical review. *The American Statistician* 42:111–119.

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.

Clemins, P.J. and M.T. Johnson. 2006. Generalized perceptual linear prediction features for animal vocalization analysis. *J. Acoust. Soc. Am.* 120:527–534.

Clemins, P.J., M.T. Johnson, K.M. Leong, and A. Savage. 2005. Automatic classification and speaker identification of African elephant *Loxodonta africana* vocalizations. *J. Acoust. Soc. Am.* 117:956–963.

Davis, J. and M. Goadrich. 2006. The relationship between precision-recall and ROC curves. *Proc. 23rd Intl. Conf. Machine Learning (ICML)*, Pittsburgh, PA.

Delarue J., M. Laurinolli, and B. Martin. 2009. Bowhead whale (*Balaena mysticetus*) songs in the Chukchi Sea between October 2007 and May 2008. *J. Acoust. Soc. Am.* 126:3319–28.

Ding Hui, Bo Qian, Yanping Li, Zhenmin Tang. 2006. A method combining lpc-based cepstrum and harmonic product spectrum for pitch detection. IIH-MSP, pp.537–540. *Intl. Conf. Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP'06)*.

Erbe, C. and J. King. 2008. Automatic detection of marine mammals using information entropy. *J. Acoust. Soc. Am.* 124:2833, DOI:10.1121/1.2982368.

Fristrup, K.M. and W.A. Watkins. 1993. Marine animal sound classification. *Technical Report WHOI-94-13*, Woods Hole Oceanographic Institution, Woods Hole, MA. 32 p.

Gillespie, D. 2004. Detection and classification of right whale calls using an "edge" detector operating on a smoothed spectrogram. *Can. Acoust.* 32:39–47.

Karlsen J.D., A. Bisther, C. Lydersen, T. Haug, and K.M. Kovacs. 2002. Summer vocalizations of adult male white whales (*Delphinapterus leucas*) in Svalbard, Norway. *Polar Biol.* 25:808–817.

Kogan, J.A., and D. Margoliash. 1998. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study. *J. Acoust. Soc. Am.* 103:2185–2196.

Mellinger, D.K. 2001. *Ishmael 1.0 User's Guide*. NOAA. Technical Memorandum OAR PMEL-120, available from NOAA/PMEL/OERD, 2115 SE OSU Drive, Newport, OR, 97365-5258. http://www.pmel.noaa.gov/pubs/PDF/mell2434/mell2434.pdf

Mellinger, D. 2004. A comparison of methods for detecting right whale calls. *Canad. Acoust.* 32(2):55-65.

Mellinger, D.K., and J.W. Bradbury. 2007. Acoustic measurement of marine mammal sounds in noisy environments. *Proc. Intl. Conf. Underwater Acoustic Measurements: Technologies and Results*, Heraklion, Greece, pp. 273–280.

Mellinger, D.K., and C.W. Clark. 1997. Methods for automatic detection of mysticete sounds. *Mar. Freshwater Behav. Physiol.* 29:163–181.

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

Mellinger, D.K. and C.W. Clark. 2006. MobySound: A reference archive for studying automatic recognition of marine mammal sounds. *Applied Acoustics* 67:1226–1242.

Mellinger, D.K., K.M. Stafford, S.E. Moore, R.P. Dziak, and H. Matsumoto. 2007. An overview of fixed passive acoustic observation methods for cetaceans. *Oceanography* 20(4):36–45.

Mellinger, D.K., S. Martin, R. Morrissey, N. DiMarzio, D. Moretti, and L. Thomas. 2009. An algorithm for detection of whistles, moans, and other tonal sounds. *4th Intl. Workshop on Detection, Classification, and Localization of Marine Mammals Using Passive Acoustics*, Pavia, 10–13 Sep.

Mouy, X., D. Leary, B. Martin, and M. Laurinolli. 2008. A comparison of methods for the automatic classification of marine mammal vocalizations in the Arctic. *New Trends for Environmental Monitoring Using Passive Systems, 2008 (Conf. Proc.)*. Institute of Electrical and Electronics Engineers, DOI:10.1109/PASSIVE.2008.4786984.

Mouy, X., M. Bahoura, and Y. Simard. 2009. Automatic recognition of fin and blue whale calls for real-time monitoring in the St. Lawrence. *J. Acoust. Soc. Am*. 126:2918–2928.

Nosal, E.M. 2008. Flood-fill algorithms used for passive acoustic detection and tracking. Proc. IEEE Workshop and Exhibition on New Trends for Environmental Monitoring using Passive Systems, Hyeres, France.

Oswald, J., J. Barlow, and T. Norris. 2003. Acoustic identification of nine delphinid species in the eastern tropical Pacific Ocean. *Marine Mammal Science* 19:20–037.

Oswald, J.N., S. Rankin, J. Barlow, and M.O. Lammers. 2007. A tool for real-time acoustic species identification of delphinid whistles. *J. Acoust. Soc. Am*. 122:587–595.

Stafford, K.M. 1995. *Characterization of Blue Whale Calls from the Northeast Pacific and Development of a Matched Filter to Locate Blue Whales on U.S. Navy SOSUS (SOund SUrveillance System) arrays*. M.Sc. Thesis, Oregon State University, Corvallis, OR.

Appendix B. Ambient Noise Results

B.1 Introduction

Bowhead whale (*Balaena mysticetus*) sounds recorded on the Burger and Klondike cluster recorder arrays were located using a location process developed by JASCO Applied Sciences. This appendix discusses JASCO's approach to performing the localization process.

A localization engine, that models the problem with synthetic data, was implemented to evaluate the performance of the localization algorithm. The localization engine consists of two parts: a data simulator and a localization processor (Figure B-171).

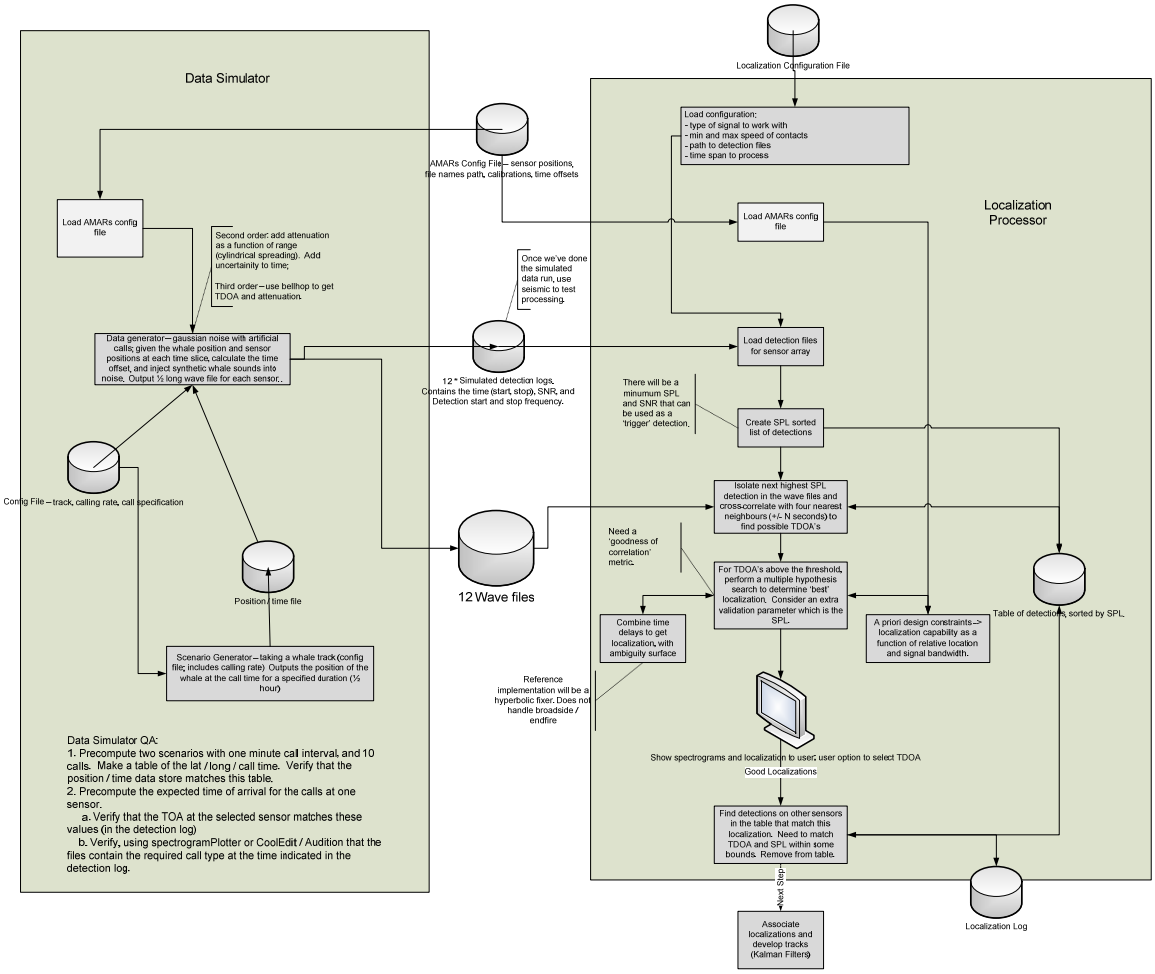

FIGURE B-171. (Left) Data simulator and (right) localization processor of the localization engine.

B.2 Source Localization

The localization processor constitutes the second block in the localization engine. The data simulator (first block) will be described in detail in the following section. The localization

processor was designed and developed in MATLAB version 7.11.0.584 (R2010b). The methodology consists of five main stages:

(1) Time-Alignment

(2) Data extraction

(3) Calls Associated with Multiple Recorders

(4) Time Difference Of Arrival (TDOA) synchronization

(5) Source localization.

A general overview of the localization processor is given in Figure B-172.

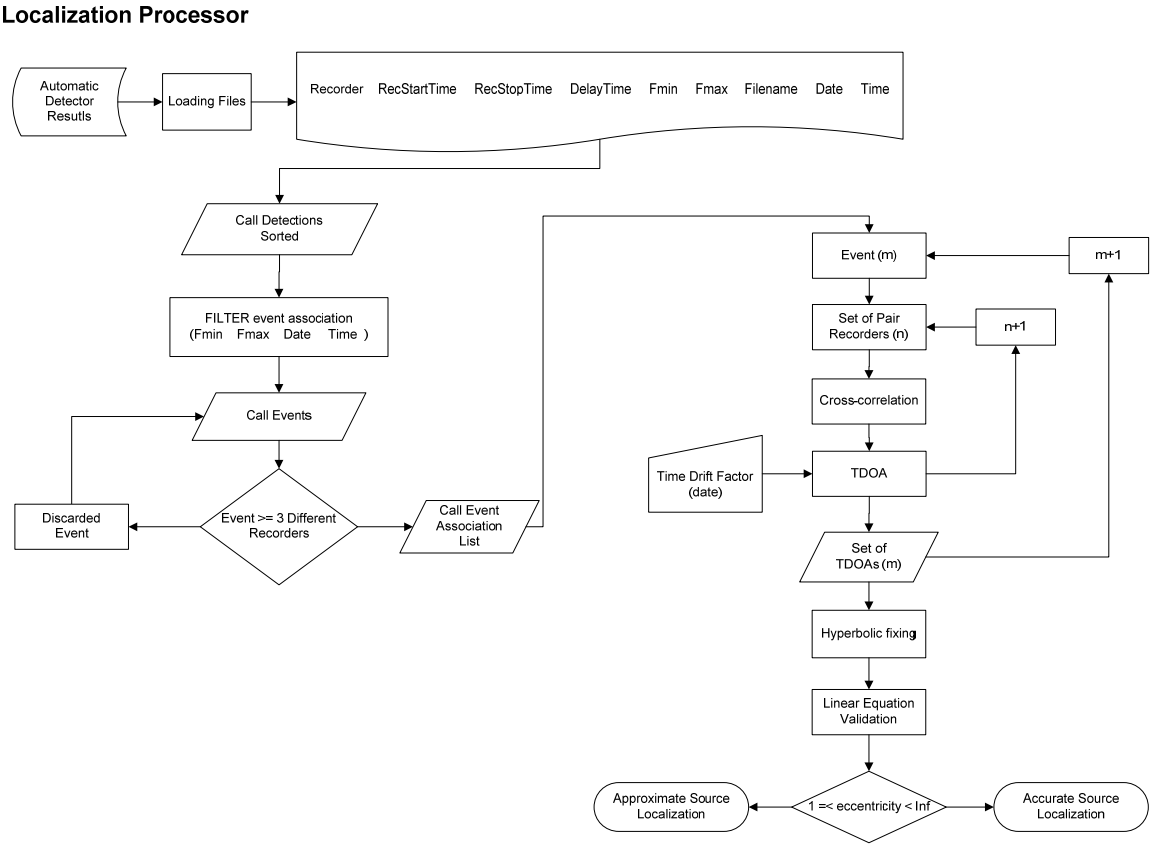

FIGURE B-172. Block diagram of the localization processor.

B.2.1 Time-alignment

Since the recorders have a drift in sampling time over the months of deployment it is essential to align the time on each recorder to perform localization. The time drift of each recorder, relative to an arbitrary reference recorder R_{ref} is therefore:

$$\text{Time Drift} = \frac{R_{\text{ref Samples}}}{\text{Eff}_{\text{Sampling rate}}} - R_{\text{ref Time}} + \Delta_{\text{Sync}}$$

where:

Δ_{Sync} is the delta sync time over the period of the deployment on each recorder,
$R_{\text{ref Time}}$ is the calculated time of reference recorder,
$R_{\text{ref Samples}}$ is the number of time on the reference recorder, and
$\text{Eff}_{\text{Sampling rate}}$ is the effective sampling rate calculated on each recorder.

B.2.2 Data Extraction

The automatic localization technique applied on this project constitutes a dependent algorithm of the automatic detector output. Bowhead calls identified by the automatic detector are saved into MATLAB `.MAT` files. A MATLAB function loads all the files and extracts the necessary information into a new `.MAT` file: recorder number, start and stop time of the file when the event was detected, delay-time, minimal and maximum frequency, file-number, and date and time of the event. Figure B-173 shows a block diagram summary of the data extraction.

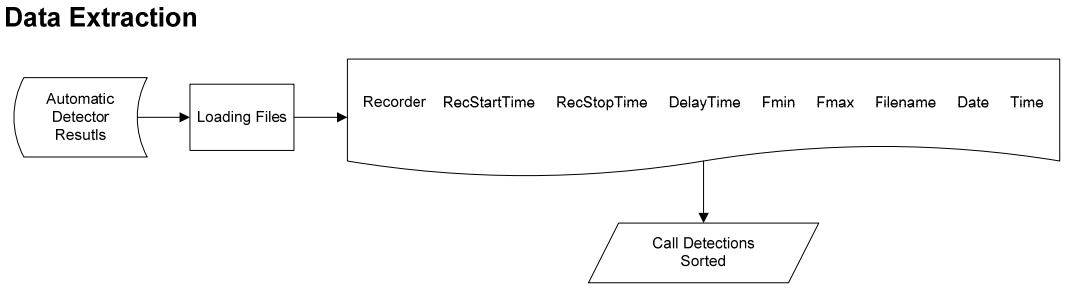

FIGURE B-173. Block diagram of the data extraction from the automatic detector.

B.2.3 Calls Associated with Multiple Recorders

To avoid misleading information, JASCO has developed an association method that eliminates redundant information from the rest of the receivers and secondly, discriminates false TDOAs that would generate false source locations. In addition, this procedure lessens the algorithm computing time. The aim is to find call detections of the same vocalization event as recorded by various receivers with different delay times. All call detection events are sorted and classified by frequency band, date and an elapsed delay time. A call associated with multiple recorders is a candidate for potential localization if the call was detected by at least three recorders. Figure B-174 shows a flow diagram of the event association.

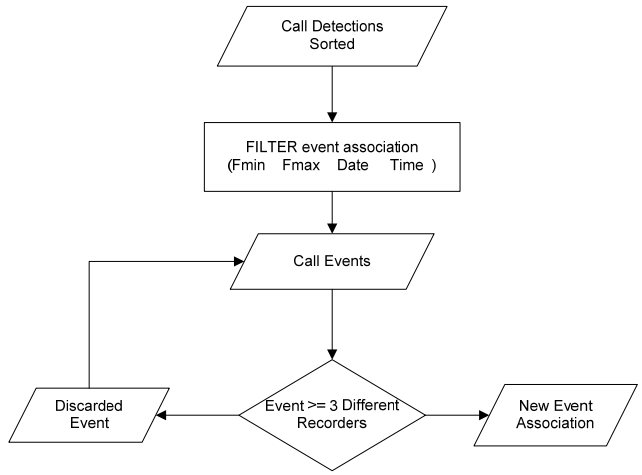

FIGURE B-174. Block diagram of the automatic calls associated with multiple recorders.

B.2.4 TDOA Synchronization

The association stage identifies events across multiple recorders that are in the same frequency bands, and occur at approximately the same time. The synchronization stage determines the exact time delay of arrival between a reference recorder, and each of the other recorders that detected the event. The TDOAs are computed via cross-correlation of the other recorder's data with the reference recorder. It then adds the *time drift factor* according to the date of the call detection and synchronizes the times of the different recorders to obtain accurate sets of TDOAs. Figure B-175 illustrates the synchronization in a block diagram.

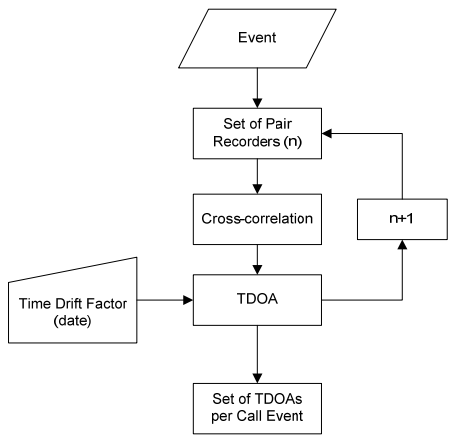

FIGURE B-175. Flow diagram of the TDOA synchronization stage.

B.2.5 Source Localization

Figure B-176 shows a block diagram of the event localization process, which computes the source location for each set of event TDOAs. The accuracy obtained in whale localization depends on the ambient noise, the acoustic characteristics of the calls, the instrumentation, and

the localization technique. Of those, we have control of the last two. JASCO's event localization approach produces a candidate source location based on hyperbolic fixing of a set of TDOAs from three separate recorders. The location is validated by a linear equation algorithm, and we check that both results are similar. The following sections explain both techniques in detail.

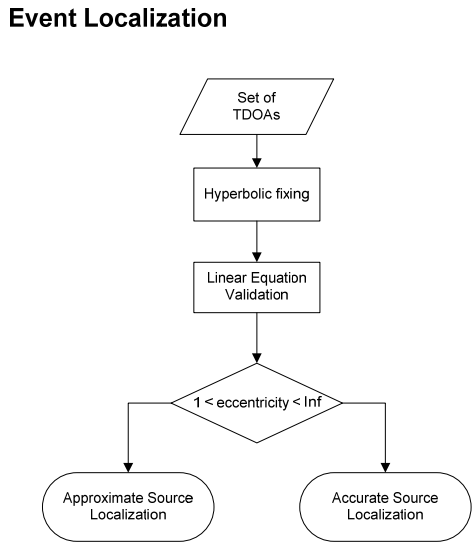

FIGURE B-176. Flow diagram of the event localization.

B.2.5.1 Hyperbolic Fixing Technique

Hyperbolic fixing continues to be the most widely used technique for localization research (Watkins and Schevill 1972, Spiesberger and Fristrup 1990) because of its simplicity and strong dependence on three main variables: TDOA, receiver position and sound velocity of the medium. TDOAs are easily obtained by cross-correlation methods. Sound velocity is measured *in situ* with Conductivity, Temperature and Depth (CTD) sensors. Hyperbolic fixing can be adapted to various scenarios. By using detection-classification methods and adequate receiver array geometries, TDOA measurements may cover cetacean vocalizations of a wide frequency range from 10 Hz to 200 kHz, including calls that vary from clicks to groans, buzzes, chirps and whistles. Hyperbolic fixing provides accurate two-dimensional localization for shallow underwater scenarios (Vallarta 2009).

In the source localization problem, the active source represents a point along the hyperbola solution. Each hyperbola focus point represents a passive receiver element. The difference in arrival times of a source signal between a pair of receivers is the TDOA. When multiplying the TDOA by the sound velocity of the medium, the constant value that defines a hyperbola is obtained. By using several receiver pairs, the intersecting hyperbolas will give an indication of source location.

The accuracy of hyperbolic fixing depends greatly on the number of receiver pairs and their relative locations. The absolute position of each receiver and TDOA must be accurately known for this localization to be effective. The intersection of several hyperbolas provides sufficient information of the source location including range and bearing (azimuth) (Vallarta 2009).

B.2.5.2 Linear Equation Approach

The linear equation approach describes the algebraic relation between the TDOA and the locations of the source and the receivers. It yields the same mathematical form for 2D and 3D recorder arrays. Defining Receiver 1 as the origin, the source location (*s*) is obtained from a three-receiver array as:

$$s = \frac{1}{2} R^{-1} b - c^2 \delta \tau_1 R^{-1} \tag{C.1}$$

where *c* is sound velocity, δ is the TDOA vector $\delta = [\delta_{12}, \delta_{13}]^T$, *b* is given by $b_i = r_{x(i)}^2 + r_{y(i)}^2 - c^2 \delta_{1(i)}^2$, τ_1 is time of arrival from source to receiver reference, and *R* represents the receiver matrix:

$$R = \begin{bmatrix} r_{x(2)} & r_{y(2)} \\ r_{x(3)} & r_{y(3)} \end{bmatrix} \tag{C.2}$$

Solving for τ_1:

$$\tau_1 = \frac{c a_2 \pm \sqrt{c^2 a_2^2 - \left(c^2 a_3 - 1\right) a_1}}{2c\left(c^2 a_3 - 1\right)} \tag{C.3}$$

where

$$a_1 = \left(R^{-1} b\right)^T \left(R^{-1} b\right) \qquad a_2 = \left(R^{-1} \tau\right)^T \left(R^{-1} b\right) \qquad a_3 = \left(R^{-1} \tau\right)^T \left(R^{-1} \tau\right) \tag{C.4}$$

Substituting Equation B.3 into Equation B.1, the source location *s* is obtained. Two positive solutions correspond to two possible source positions. Negative and complex solutions are discarded as they have no physical solution or meaning (Wahlberg *et al.* 2001, Vallarta 2009).

B.2.5.3 Source Locations

The intersections of the hyperbolas may delineate a region of uncertainty rather than an intersection point, due to the ambiguity of source positions within the discrete pair of intersection points of two hyperbolas. Hence, those source locations are considered *approximate locations* with higher uncertainty. JASCO has employed the eccentricity as a measure of localization uncertainty:

$$e(\mathrm{TDOA}) = \frac{2d}{c\ \mathrm{TDOA}} \tag{C.5}$$

where *e* is eccentricity of the localization hyperbola, *d* is the separation distance of the receivers, *c* is the sound velocity of the medium, and TDOA is the time difference of arrival of the signal. Eccentricity is a measure of the curvature (or wideness) of the hyperbola. An eccentricity greater than one ensures hyperbolic intersection, an eccentricity equal to one generates ambiguous end-fire locations, and an eccentricity smaller than one produces elliptical areas that do not generally intersect. Hyperbolas with eccentricity tending to infinity are also discarded if they do not intersect with another curve. Therefore, the hyperbolic eccentricity is used as a function of the TDOA to minimize the ambiguity of source localizations. We required that the eccentricity be greater than one.

B.3 Data Simulation

The data simulator was designed in MATLAB version 7.11.0.584 (R2010b). The simulator is synthetic with respect to generating whale calls within a random interval; however, receiver locations and all other aspects of the simulator are based on real data. The simulator provides controlled test data for validating the localization algorithm and implementation. Its main advantages are complete control in the development of the localization processor algorithm, a systematic approach to understanding the role of internal processes in the localization algorithm chain, and a modular/segmented tool allowing efficient tracking of the localization processor including identifying errors in parameter estimates relevant to localization.

Inputs to the data simulator are:

- GPS locations of recorder deployments
- Whale call rate, whale call duration, whale speed and track time
- Choice of tracking option: linear, directional, or omni-directional movement

Outputs of the data simulator are:

- A synthetic whale path (track-line)
- Position of synthetic calls along the path
- Set of TDOAs of the synthetic whale calls at receivers

B.3.1 Bowhead Call Synthesis

The data simulator synthesizes a bowhead whale call from a chirp signal—a group of samples of a linear swept-frequency cosine signal. Figure B-177 shows a sample synthetic bowhead call between 150 and 350 Hz of 2 s duration (Ljungblad *et al.* 1982), with white Gaussian noise added.

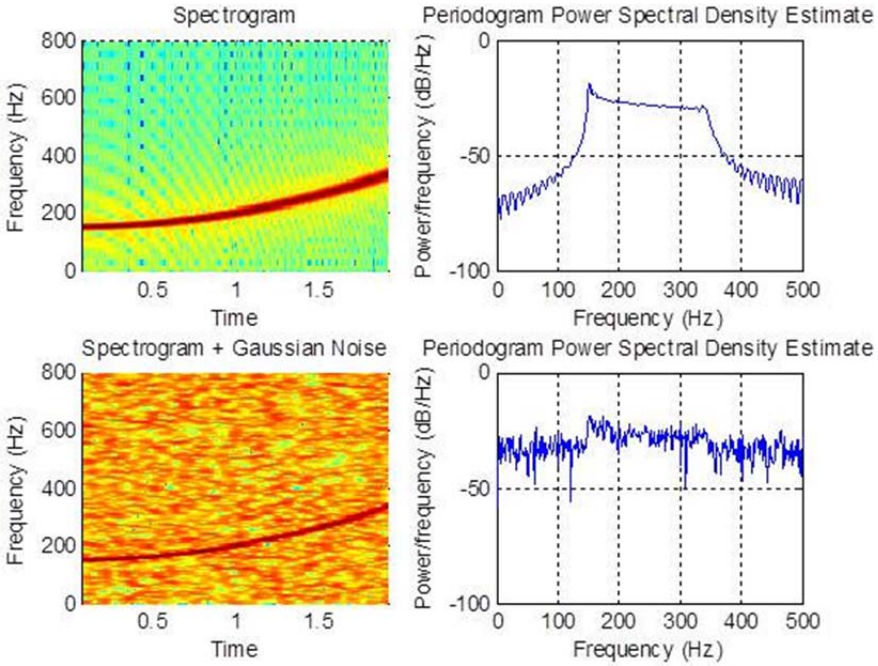

FIGURE B-177. (Left) Spectrogram and (right) power spectrum of a synthetic bowhead call generated from a chirp signal of 2 s duration and 200 Hz bandwidth.

B.3.2 Track Line Synthesis

The data simulator generates a synthetic track line with a random number of calls. Three tracking patterns (linear, directional, and omni-directional movement) are generated based on the types of movement observed via satellite monitoring of radio-tagged bowhead whales in the Beaufort and Chukchi Seas in 1992 (Mate *et al.* 2000). An average speed of 4 knots was assumed.

A *linear track* consists of an artificial bowhead whale following a simple linear path, for example, movement with no perceptible randomness in direction (Figure B-178, top). A *directional track* consists of a bowhead whale following a random path but in a specific direction, for example, a random migratory motion along a directional path (Figure B-178, middle). An *omni-directional track* consists of a bowhead whale following a non-migratory random path, typically localized within a given area. For example, feeding or resting/socializing activities (Figure B-178, bottom). Tracks incorporating mixed directional and omni-directional movements can also be generated, consisting of a combination of random movements in small areas interspersed with directional tracks.

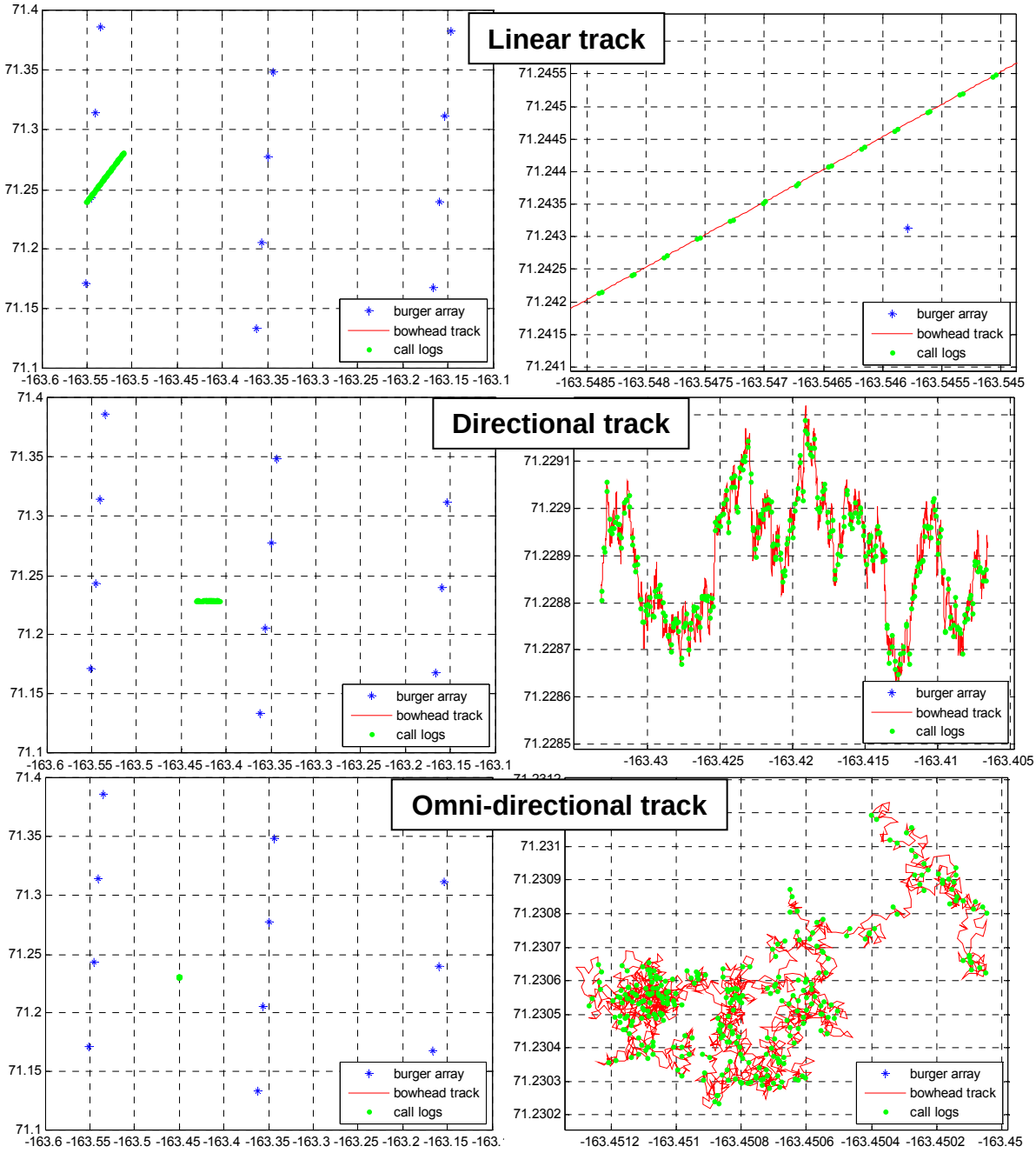

FIGURE B-178. (Left) Synthetic tracks of a bowhead whale within the Burger array and (right) the associated calls generated by the data simulator for a (top) linear track, (middle) directional track, and (bottom) omni-directional track.

B.3.3 Sound File and TDOA Synthesis

The data simulator creates WAV files and detector logs that are structurally identical to the real acoustic recordings. Time of arrivals (TOAs) for each call log is measured and recorded in a single file for each receiver. Next, a set of TDOAs for all the call logs is grouped in one master file. Figure B-179 shows three synthetic WAV files, the top panel represents the source in real

time, and the middle and bottom panels represent receivers at different positions. The difference of the TOAs to the start of the call gives the appropriate TDOA.

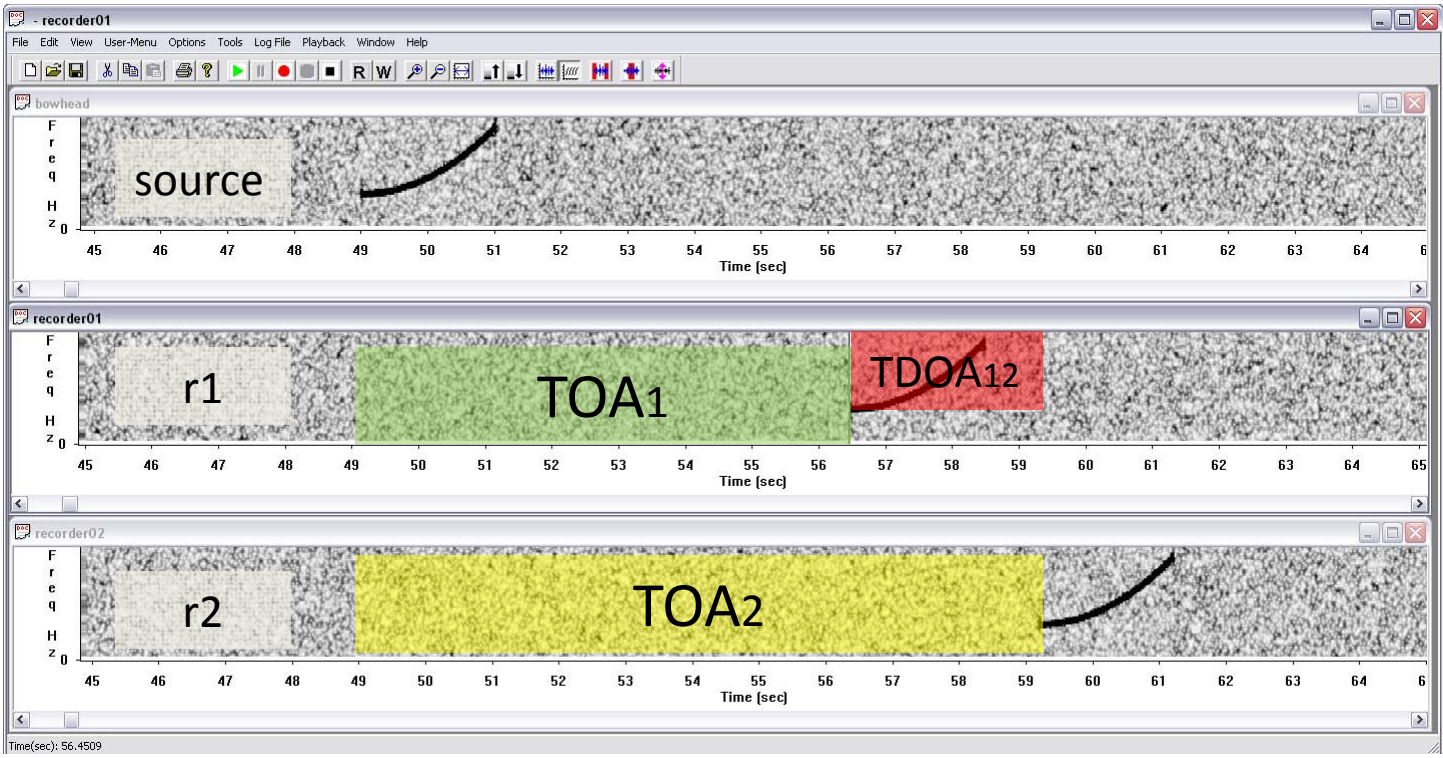

FIGURE B-179. Synthetic bowhead whale calls (top) at the source and (middle and bottom) as recorded by receivers r1 and r2. The TOA and TDOA for each set of receivers are recorded in one master TDOA file.

B.3.4 Localization of Synthesized Calls

To evaluate and validate the accuracy of the localization processor, a series of synthetic calls (chirp signals) was generated with the data simulator, as described previously. `WAV` and `MAT` files containing the detection call data were produced and used as input to the localization engine. Figure B-180 shows an individual source localized on a Cartesian plane within the Burger recorder array using hyperbolic fixing. A sequence of source localizations for a synthetic bowhead track line is shown in Figure B-181.

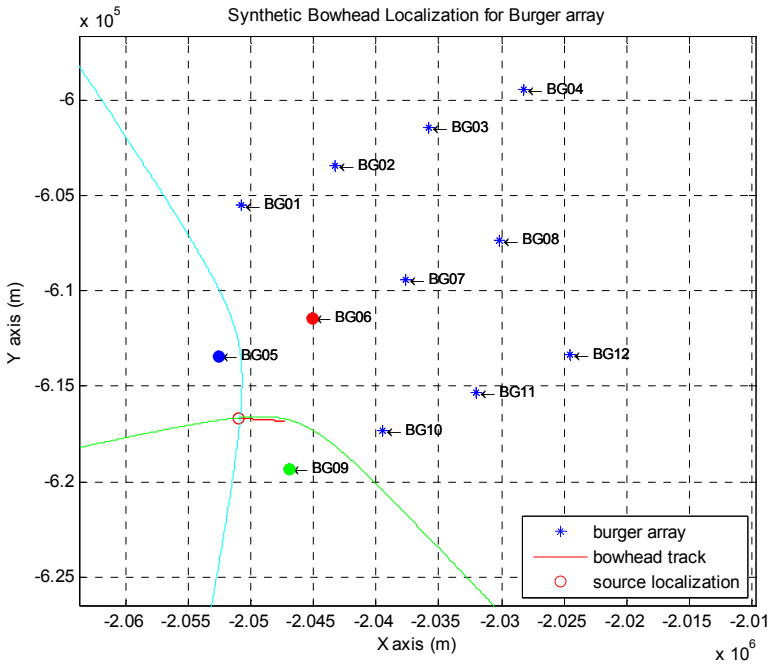

FIGURE B-180. Localization by hyperbolic fixing of a synthetic bowhead call within the Burger recorder array.

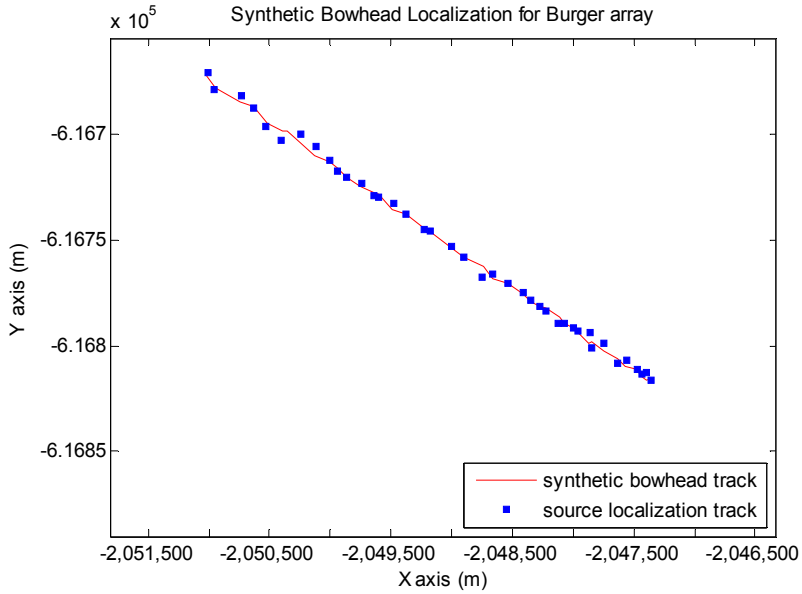

FIGURE B-181. Source localizations by hyperbolic fixing for a synthetic bowhead whale track line.

B.4 Literature Cited

Ljungblad, D.K., P.D. Scoggins, and W.G. Gilmartin. 1982. Auditory thresholds of a captive eastern Pacific bottle-nosed dolphin, *Tursiops spp*. *J. Acoust. Soc. Am.* 72(6):1726-1729.

Mate, B.R., G.K. Krutzikowsky, and M.H. Winsor. 2000. Satellite-monitored movements of radio-tagged bowhead whales in the Beaufort and Chukchi seas during the late summer. *Can. J. Zool.* 78:1168-1181.

Spiesberger, J.L. and K.M. Fristrup. 1990. Passive localization of calling animals and sensing of their acoustic environment using acoustic tomography. *The American Naturalist* 135:107-153.

Vallarta, J. 2009. *The Significance of Passive Acoustic Array-Configuration on Sperm Whale Range Estimation When Using the Hyperbolic Algorithm*. Ph.D. Thesis. Heriot-Watt University, Edinburgh, UK.

Vallarta, J. 2010. *Eccentricity discrimination of hyperbolic localizations to minimize positioning uncertainty*. J. Acoust. Soc. Am. 128, 2328. DOI:10.1121/1.3508222

Wahlbergh, M., B. Mohl, and P.T. Madsen. 2001. Estimating source position accuracy of a large-aperture hydrophone array for bioacoustics. *J. Acoust. Soc. Am.* 109(1):397-406.

Watkins, W.A. and W.E. Schevill. 1972. Sound source location by arrival-times on a non-rigid three-dimensional hydrophone array. *Deep-Sea Research* 19:691-70

Appendix C. Ambient Noise Results

C.1. Winter 2009–2010

C.1.1. Power Spectral Density Levels

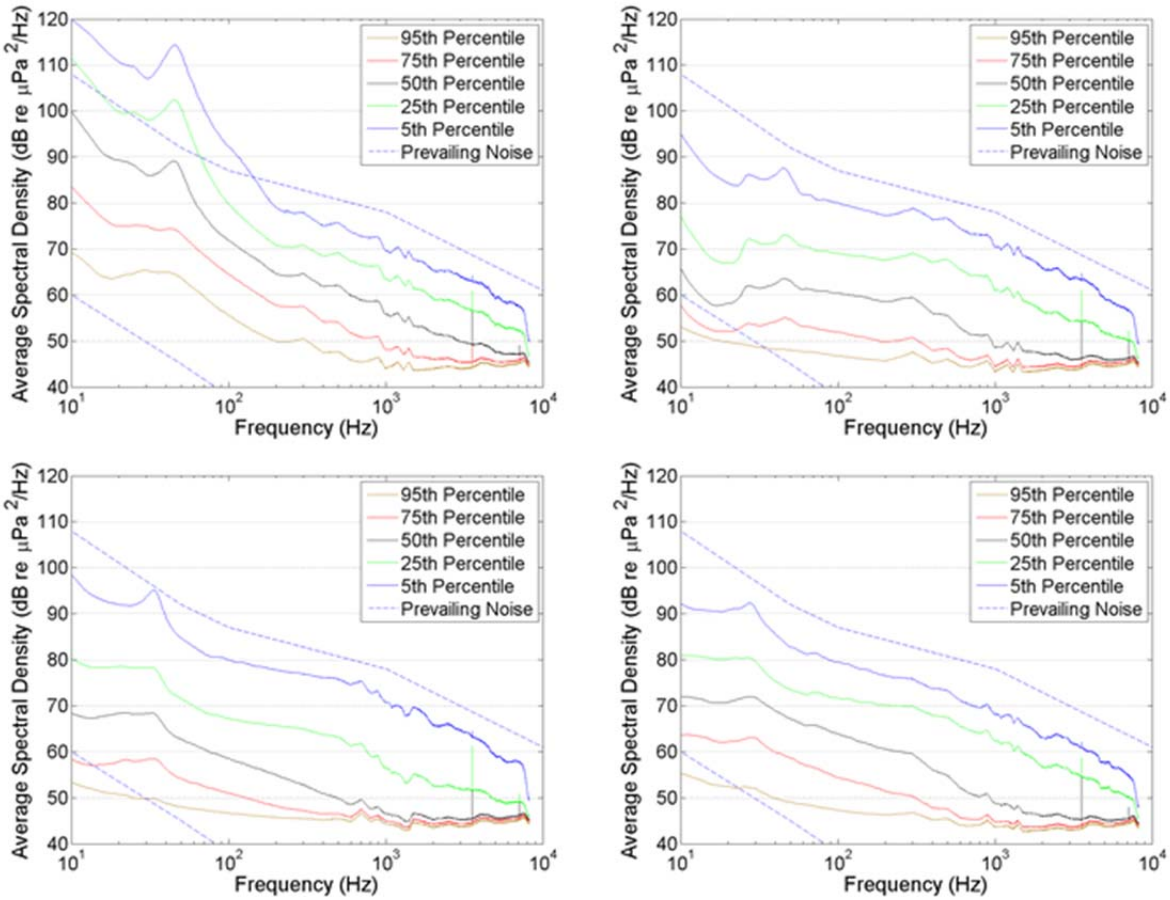

Figure B-1. Percentile 1 min power spectral density levels for winter 2009–2010 Stations (top left) B05, (top right) CL50, (bottom left) PL50 and (bottom right) PLN40, October 2009 to August 2010.

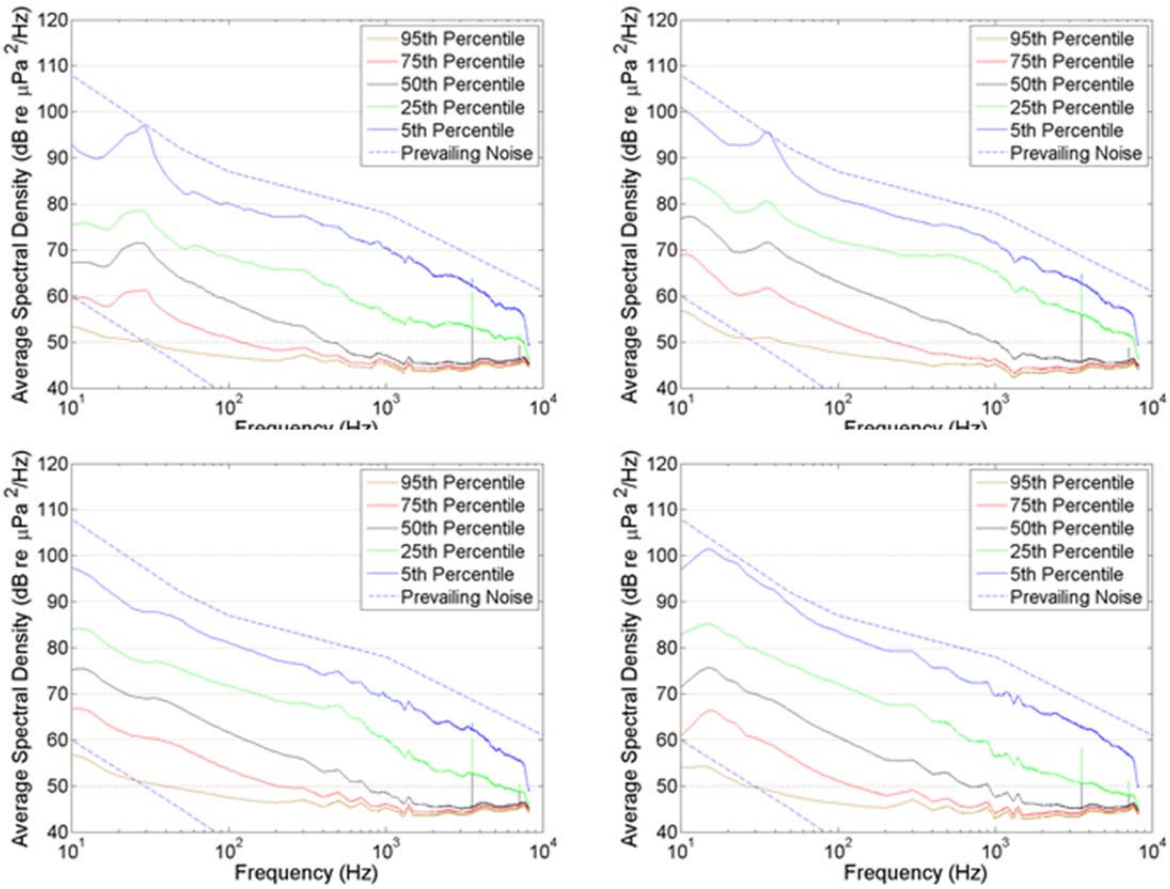

Figure B-2. Percentile 1 min power spectral density levels for winter 2009–2010 Stations (top left) PLN80, (top right) W35, (bottom left) W50 and (bottom right) WN40, October 2009 to August 2010.

C.1.2. Sound Pressure Levels

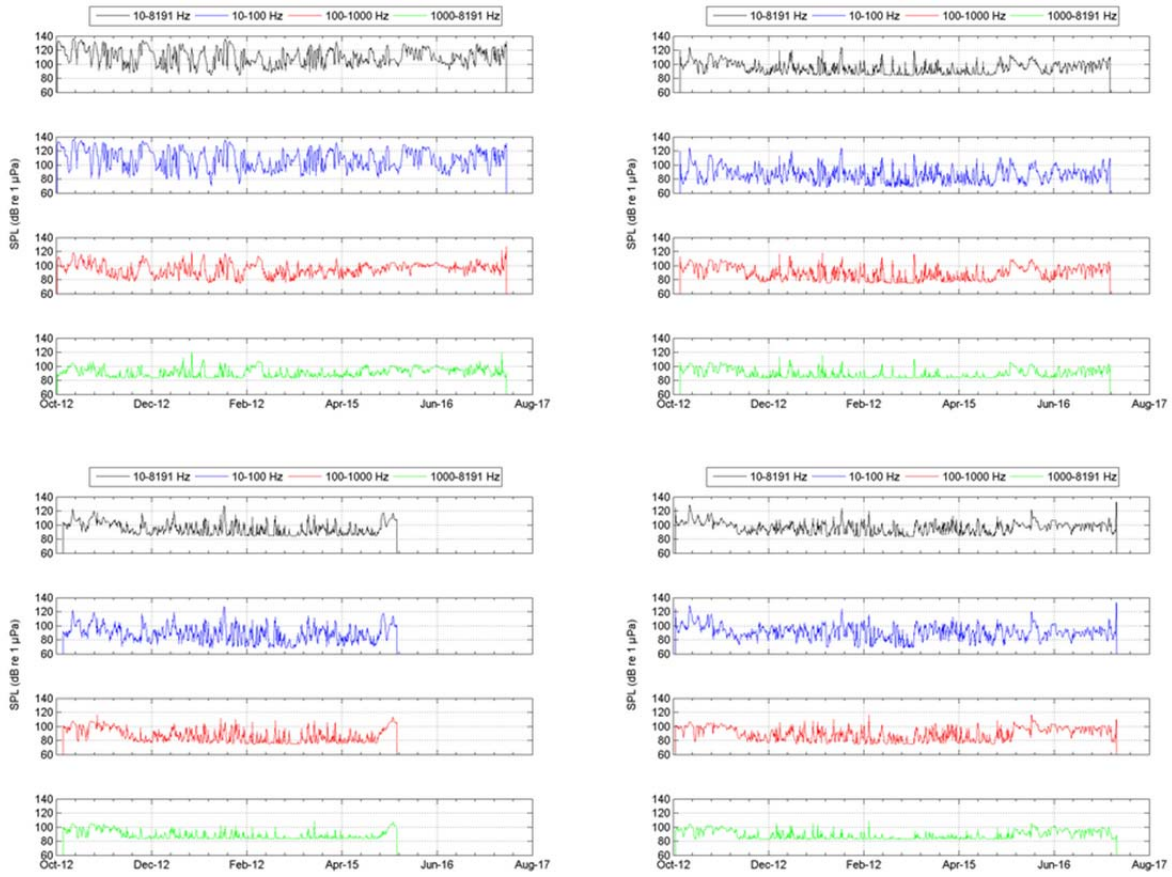

Figure B-3. Broadband and decade band sound pressure levels (SPL) for winter 2009–2010 Stations (top left) B05, (top right) CL50, (bottom left) PL50 and (bottom right) PLN40, October 2009 to August 2010.

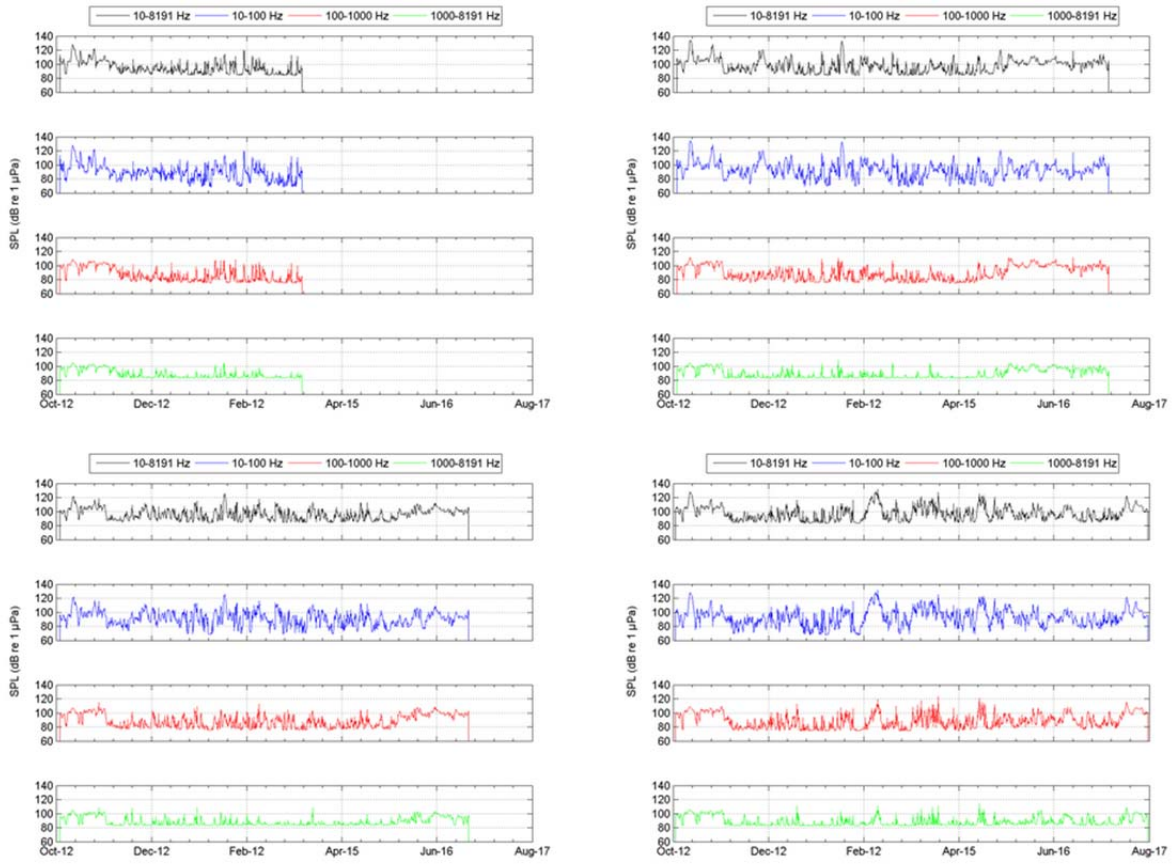

Figure B-4. Broadband and decade band sound pressure levels (SPL) for winter 2009–2010 Stations (top left) PLN80, (top right) W35, (bottom left) W50 and (bottom right) WN40, October 2009 to August 2010.

C.1.3. Spectrograms

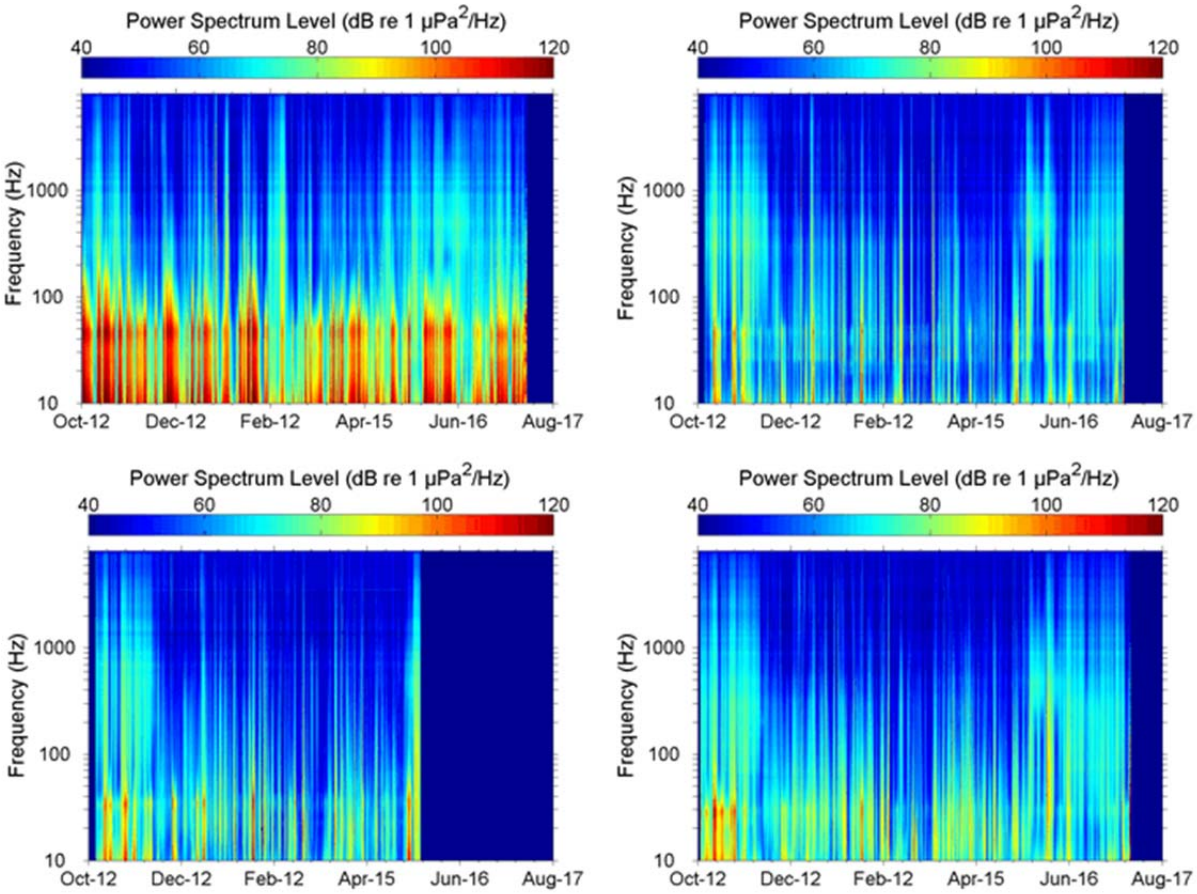

Figure B-5. Spectrogram of underwater sound at winter 2009–2010 Stations (top left) B05, (top right) CL50, (bottom left) PL50 and (bottom right) PLN40, October 2009 to August 2010.

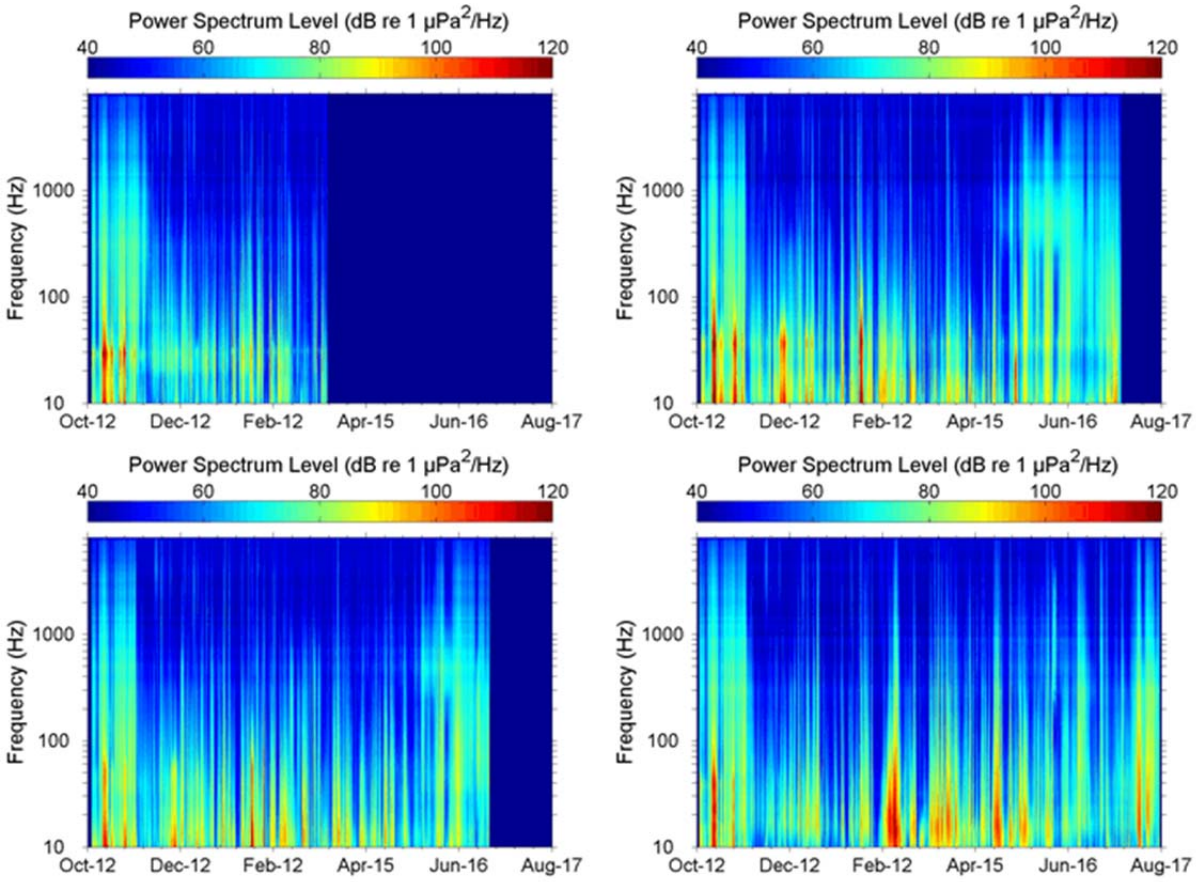

Figure B-6. Spectrogram of underwater sound at winter 2009–2010 Stations (top left) PLN80, (top right) W35, (bottom left) W50 and (bottom right) WN40, October 2009 to August 2010.

C.2. Summer 2010

C.2.1. Power Spectral Density Levels

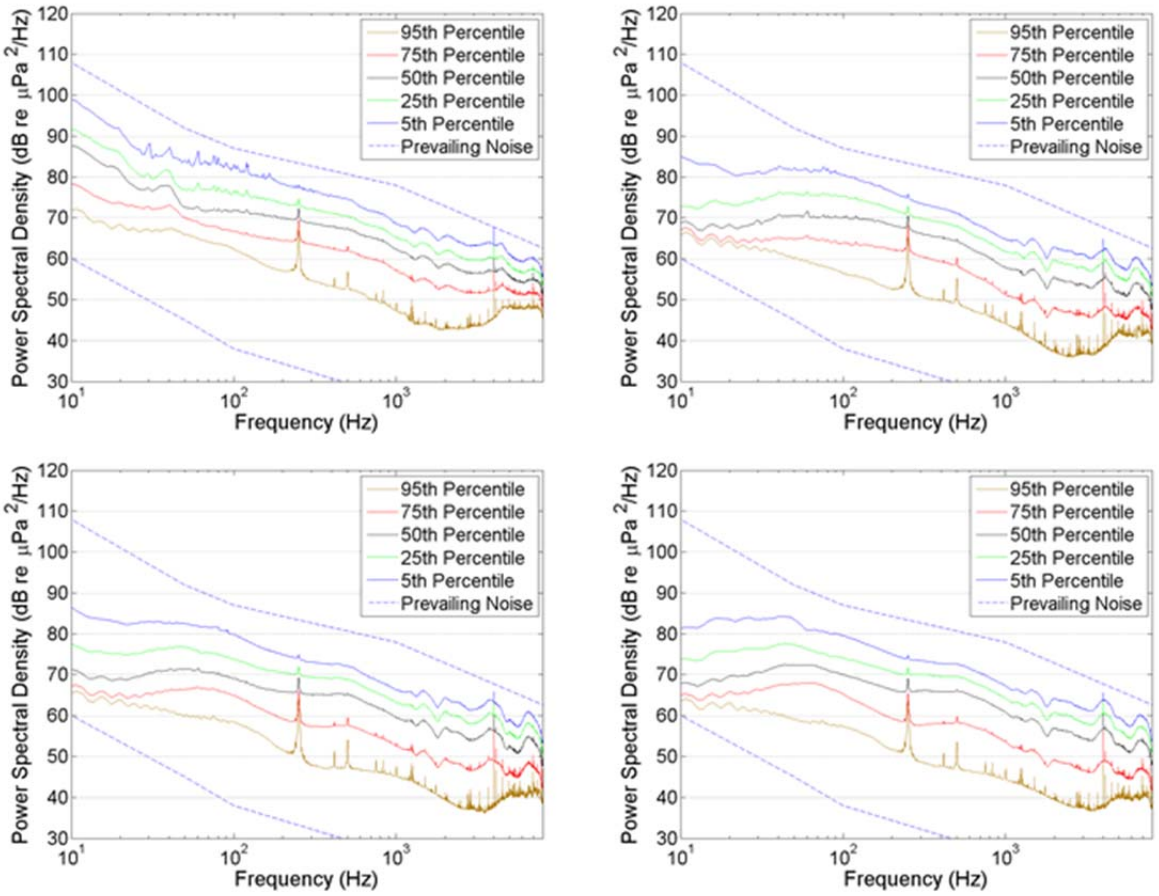

Figure B-7. Percentile 1 min power spectral density levels for summer 2010 Stations (top left) B05, (top right) B15, (bottom left) B30, and (bottom right) B50, July 2010 to October 2010.

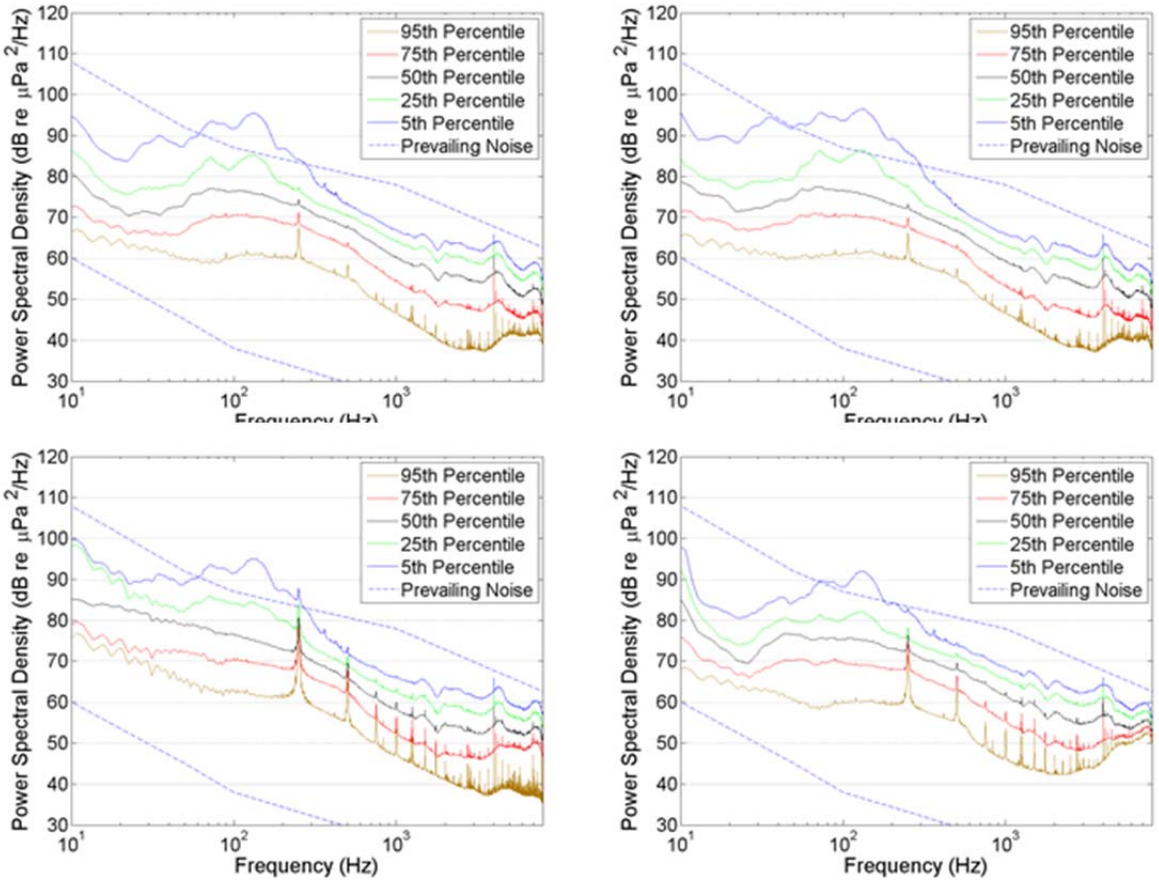

Figure B-8. Percentile 1 min power spectral density levels for summer 2010 Stations (top left) BG01, (top right) BG02, (bottom left) BG03, and (bottom right) BG04, July 2010 to October 2010.

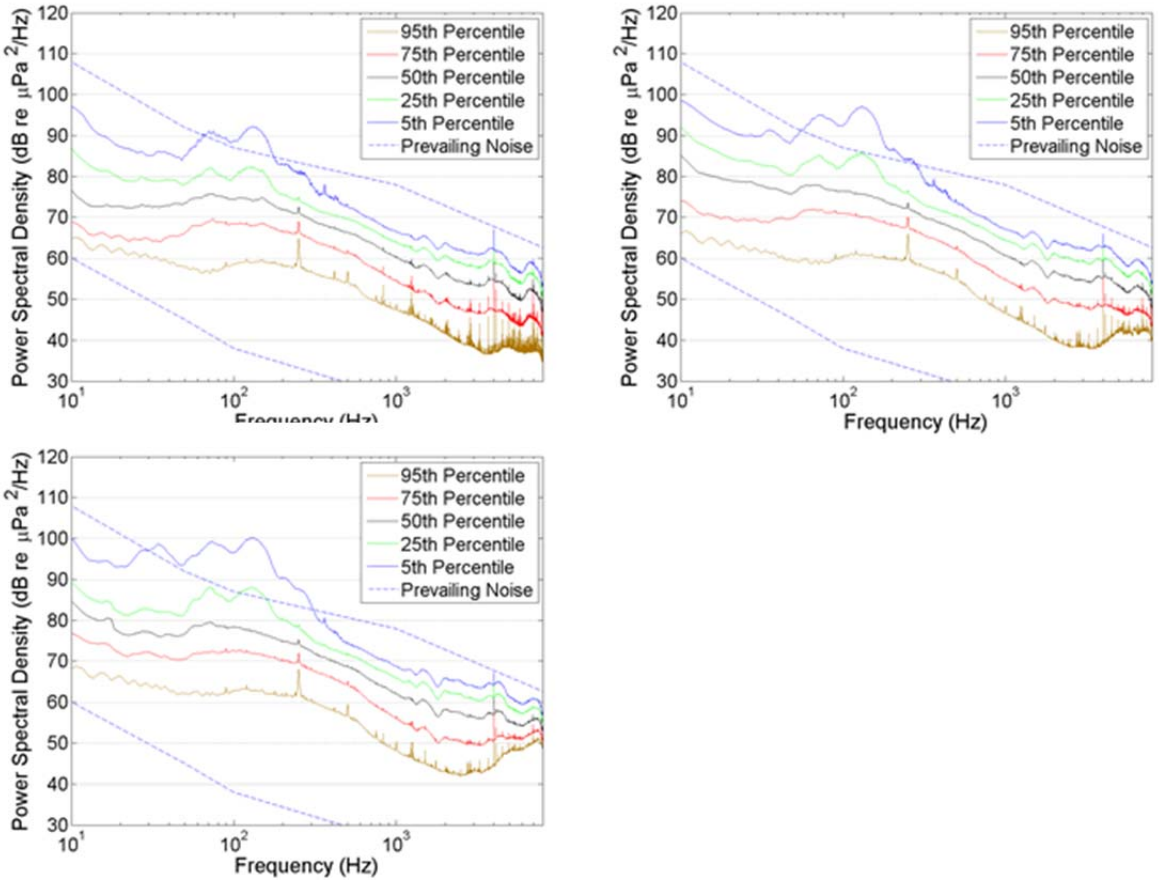

Figure B-9. Percentile 1 min power spectral density levels for summer 2010 Stations (top left) BG05, (top right) BG06, and (bottom) BG07, July 2010 to October 2010.

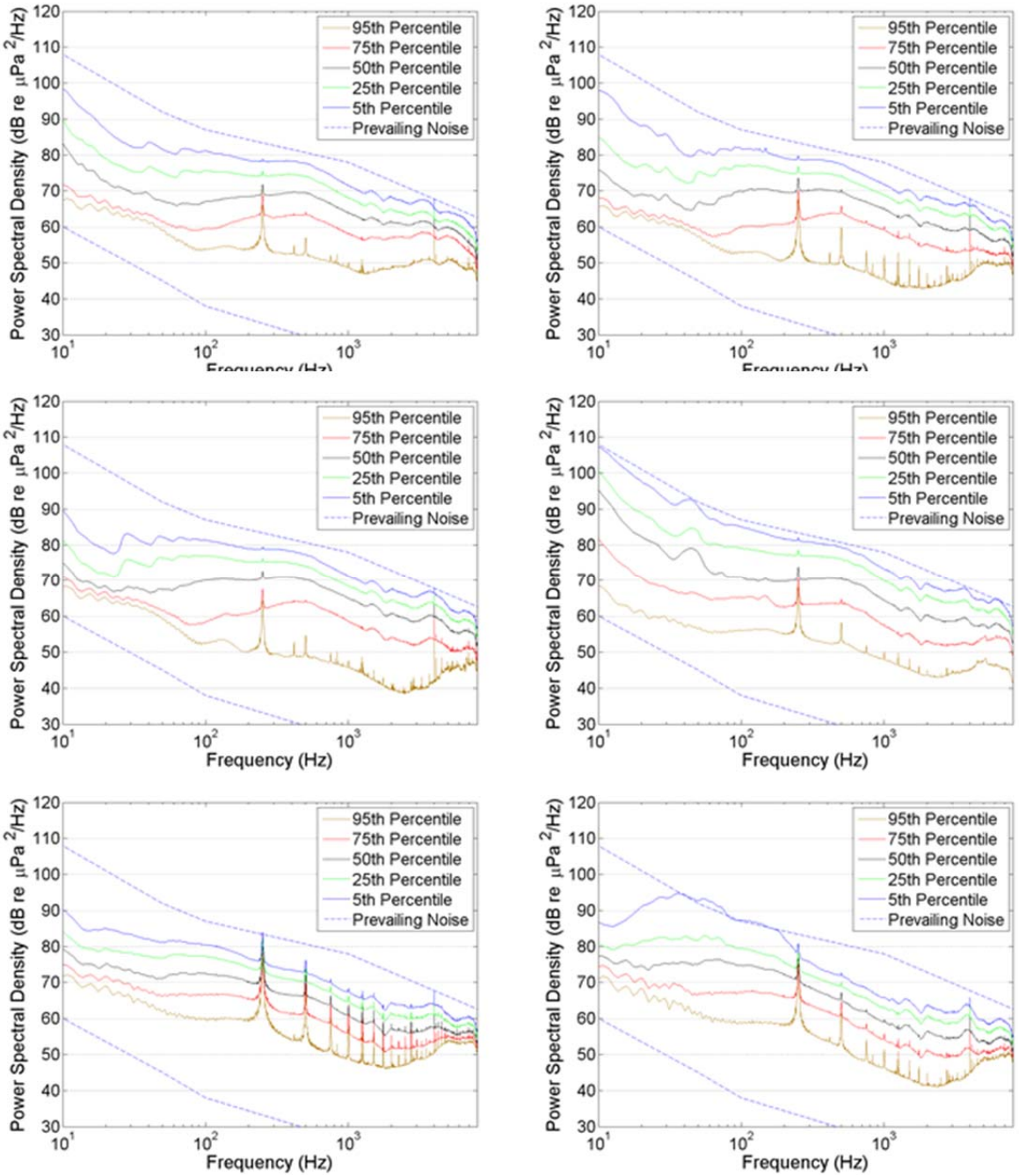

Figure B-10. Percentile 1 min power spectral density levels for summer 2010 Stations (top left) CL05, (top right) CL20, (middle left) CL50, (middle right) CLN40, (bottom left) CLN90, and (bottom right) CLN120, July 2010 to October 2010.

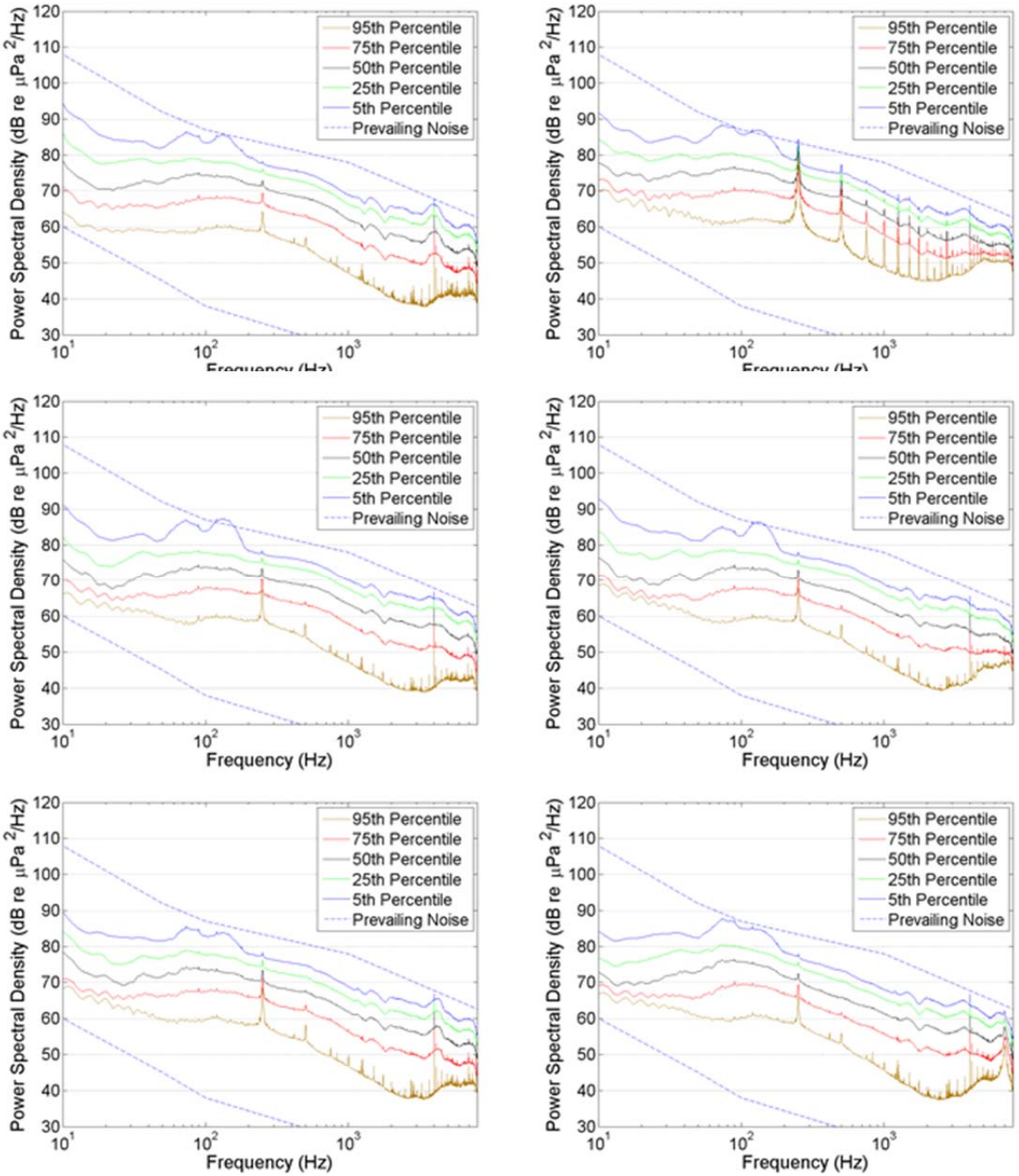

Figure B-11. Percentile 1 min power spectral density levels for summer 2010 Stations (top left) KL01, (top right) KL02, (middle left) KL03, (middle right) KL04, (bottom left) KL06, and (bottom right) KL07, July 2010 to October 2010.

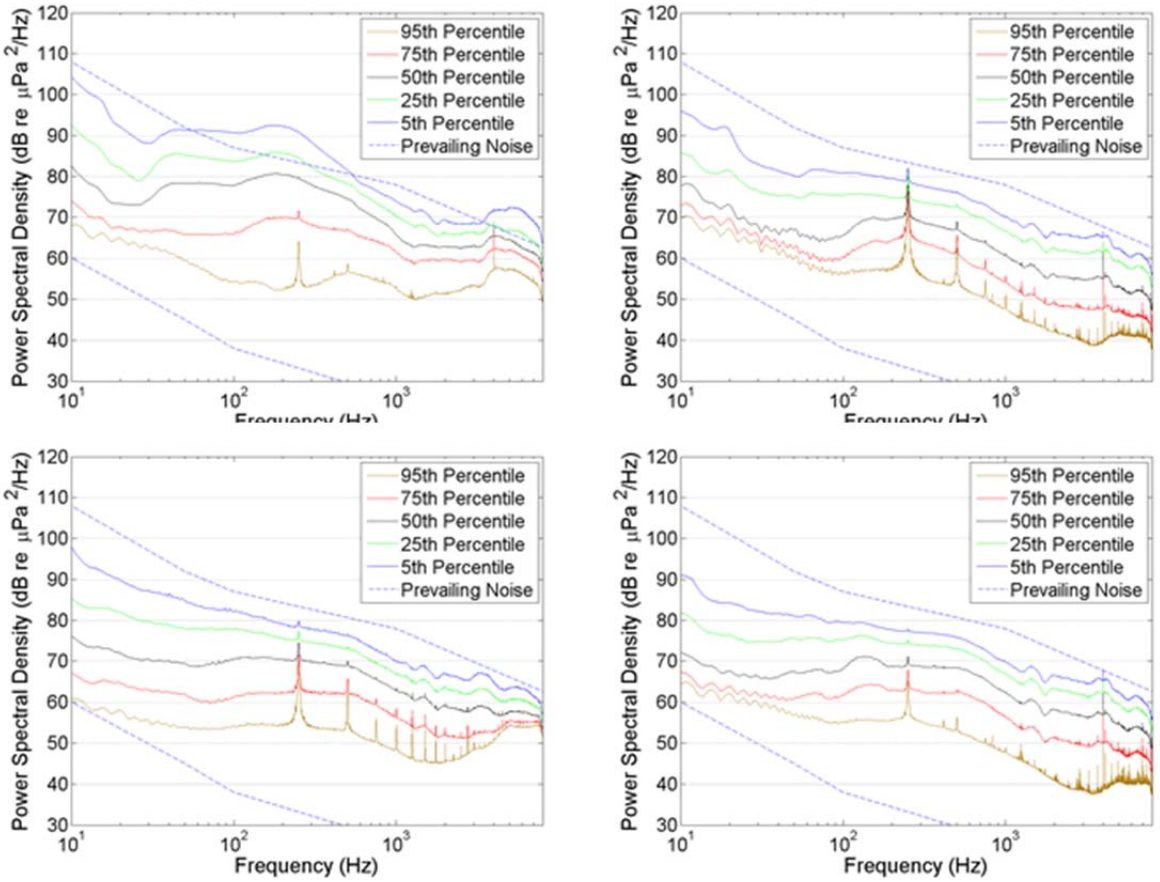

Figure B-12. Percentile 1 min power spectral density levels for summer 2010 Stations (top left) PL05, (top right) PL20, (bottom left) PL35, and (bottom right) PL50, July 2010 to October 2010.

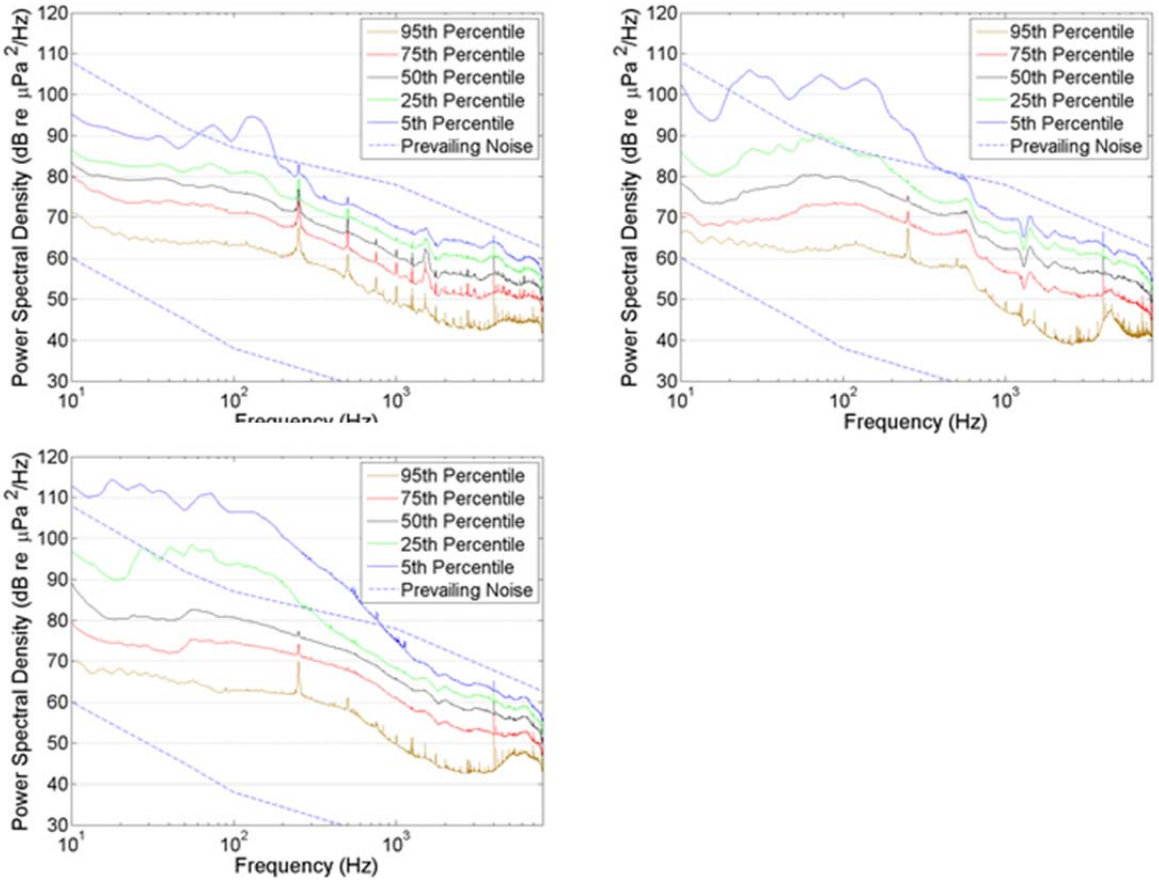

Figure B-13. Percentile 1 min power spectral density levels for summer 2010 Stations (top left) PLN40, (top right) PLN60, and (bottom) PLN80, July 2010 to October 2010.

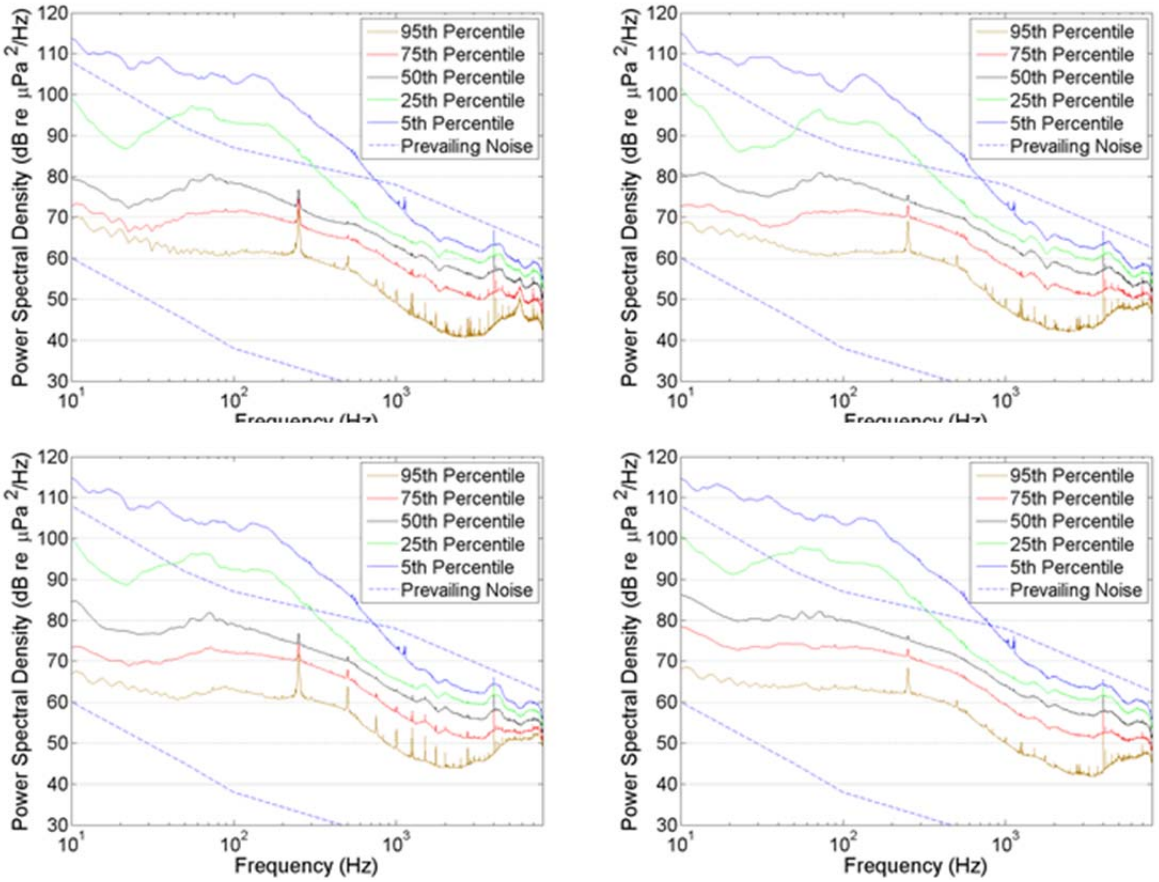

Figure B-14. Percentile 1 min power spectral density levels for summer 2010 Stations (top left) S01, (top right) S02, (bottom left) S03, and (bottom right) S04, July 2010 to October 2010.

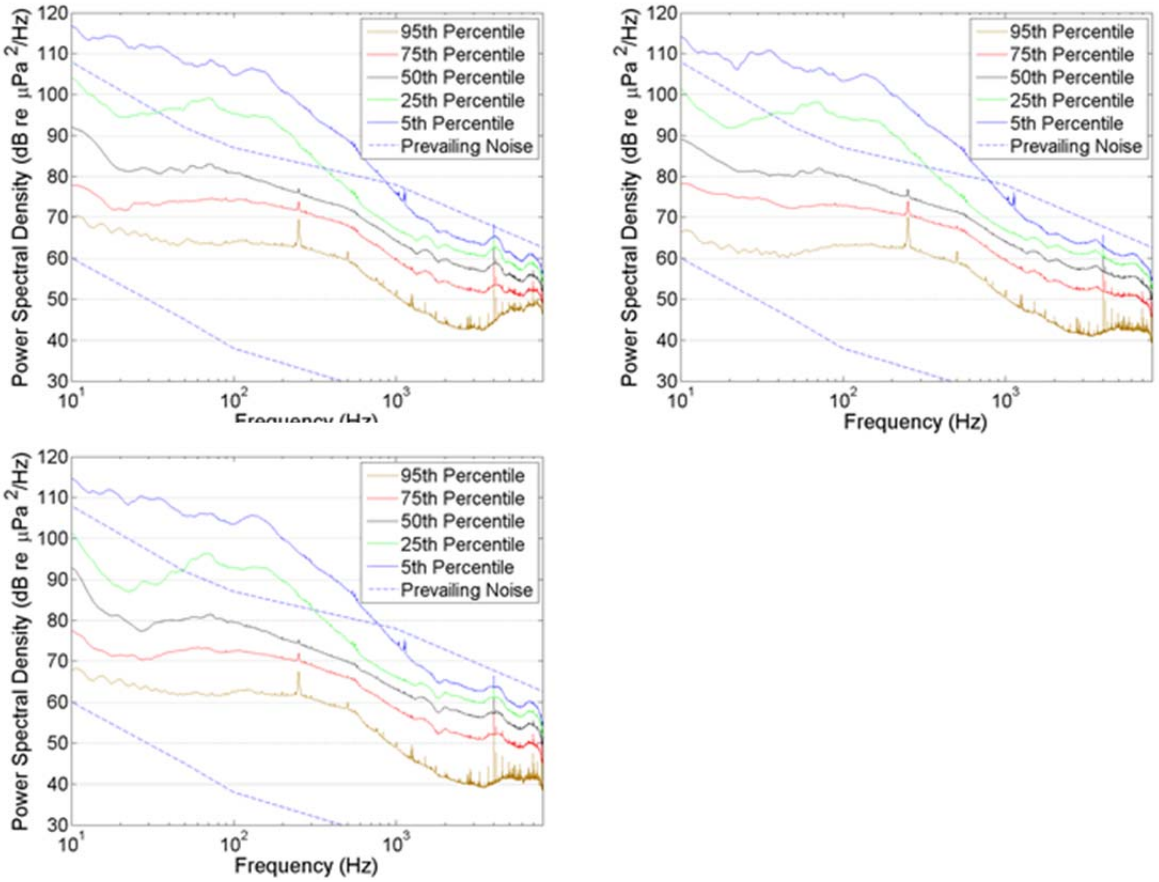

Figure B-15. Percentile 1 min power spectral density levels for summer 2010 Stations (top left) S05, (top right) S06, and (bottom) S07, July 2010 to October 2010.

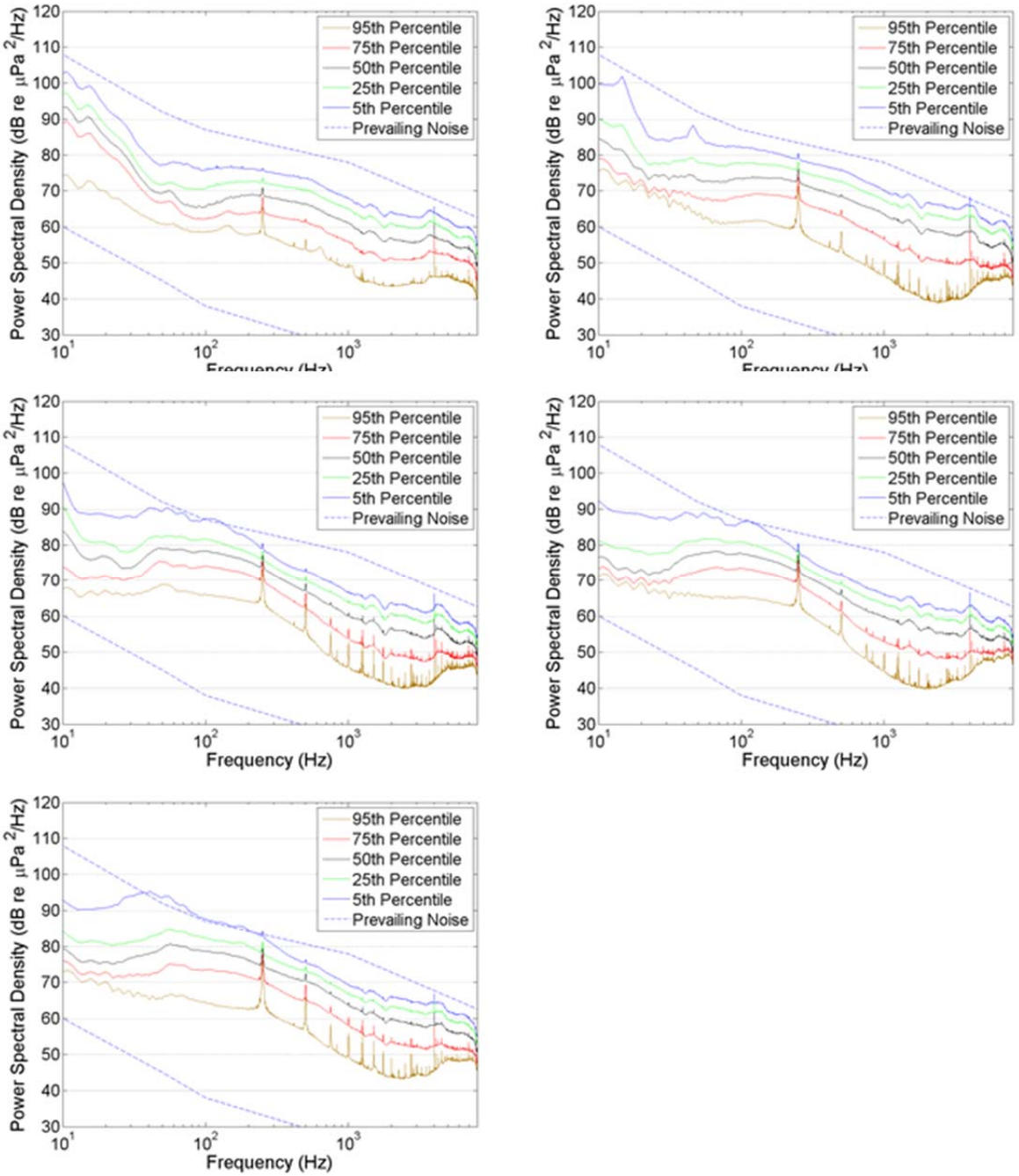

Figure B-16. Percentile 1 min power spectral density levels for summer 2010 Stations (top left) W05, (top right) W35, (middle left) WN20A, (middle right) WN20B, and (bottom) WN40, July 2010 to October 2010.

C.2.2. Sound Pressure Levels

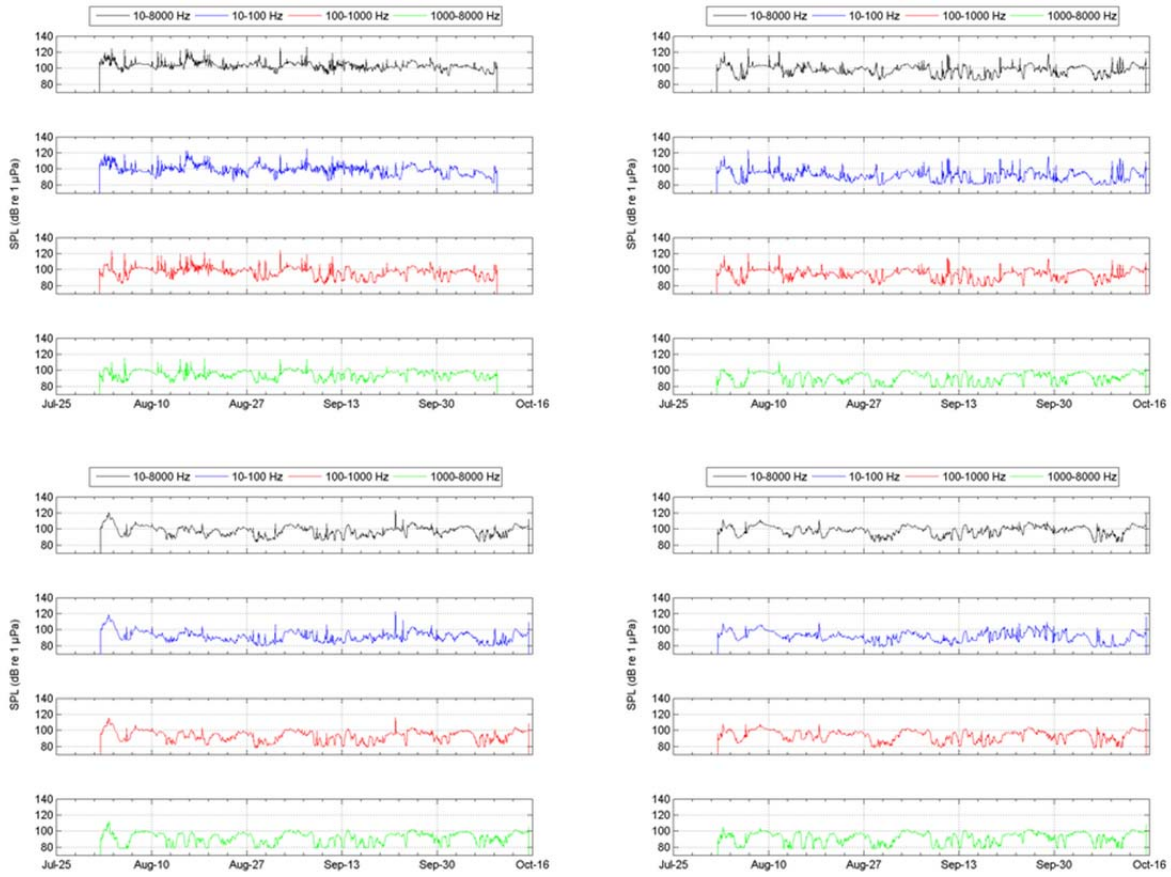

Figure B-17. Broadband and decade band sound pressure levels (SPL) for summer 2010 Stations (top left) B05, (top right) B15, (bottom left) B30, and (bottom right) B50, July 2010 to October 2010.

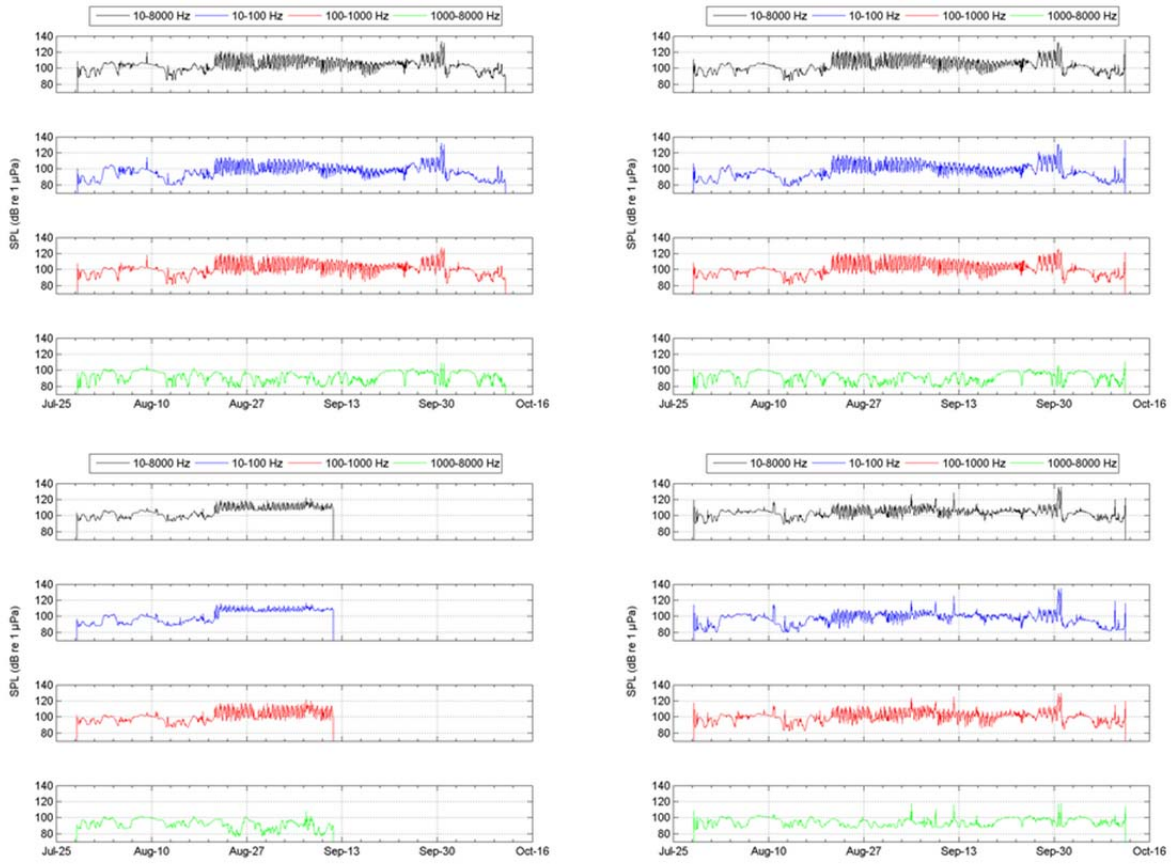

Figure B-18. Broadband and decade band sound pressure levels (SPL) for summer 2010 Stations (top left) BG01, (top right) BG02, (bottom left) BG03, and (bottom right) BG04, July 2010 to October 2010.

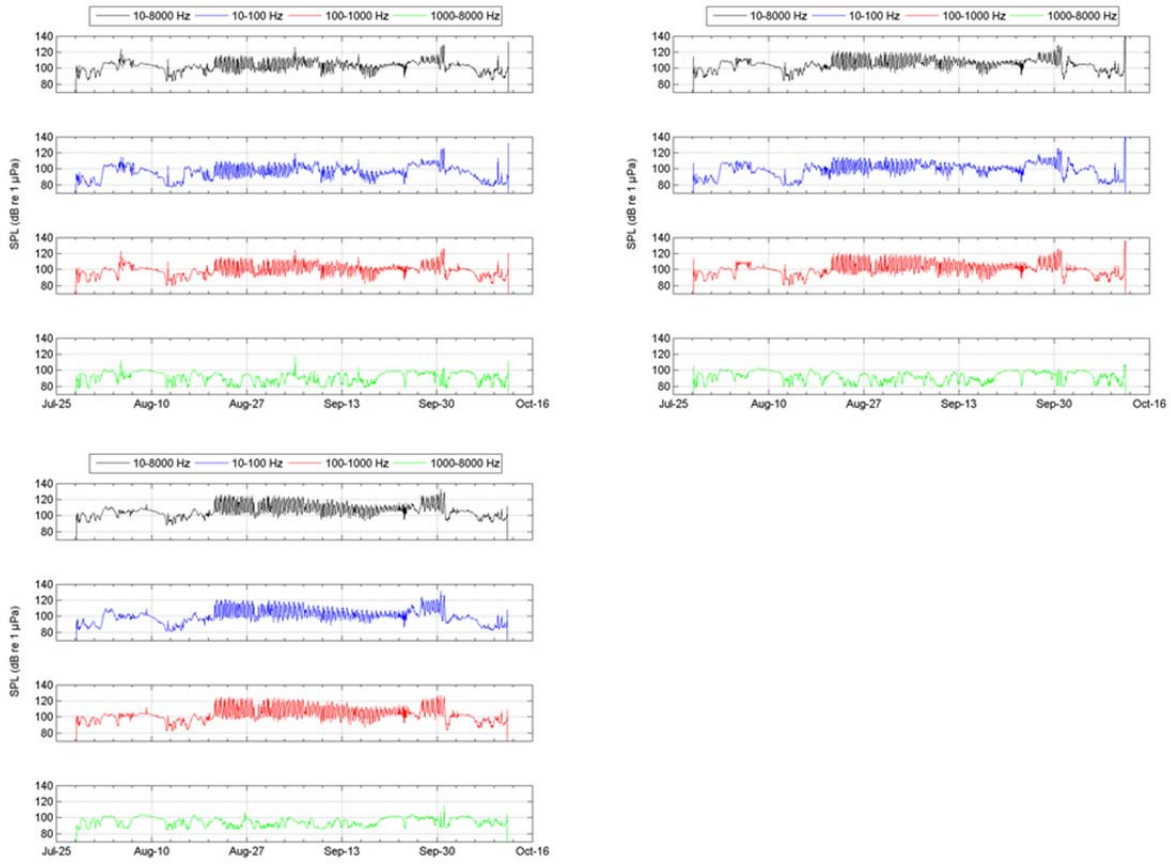

Figure B-19. Broadband and decade band sound pressure levels (SPL) for summer 2010 Stations (top left) BG05, (top right) BG06, and (bottom) BG07, July 2010 to October 2010.

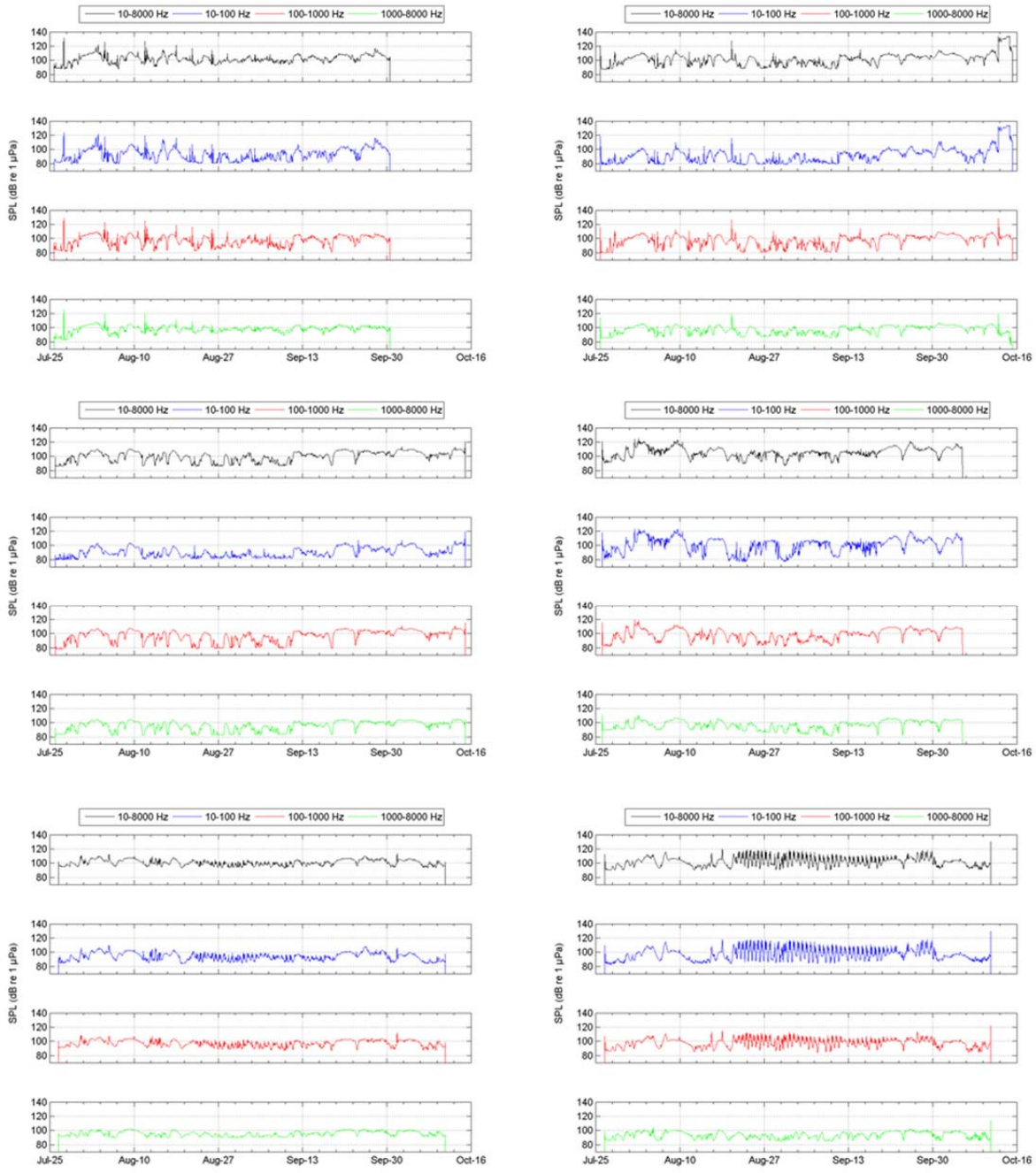

Figure B-20. Broadband and decade band sound pressure levels (SPL) for summer 2010 Stations (top left) CL05, (top right) CL20, (middle left) CL50, (middle right) CLN40, (bottom left) CLN90, and (bottom right) CLN120, July 2010 to October 2010.

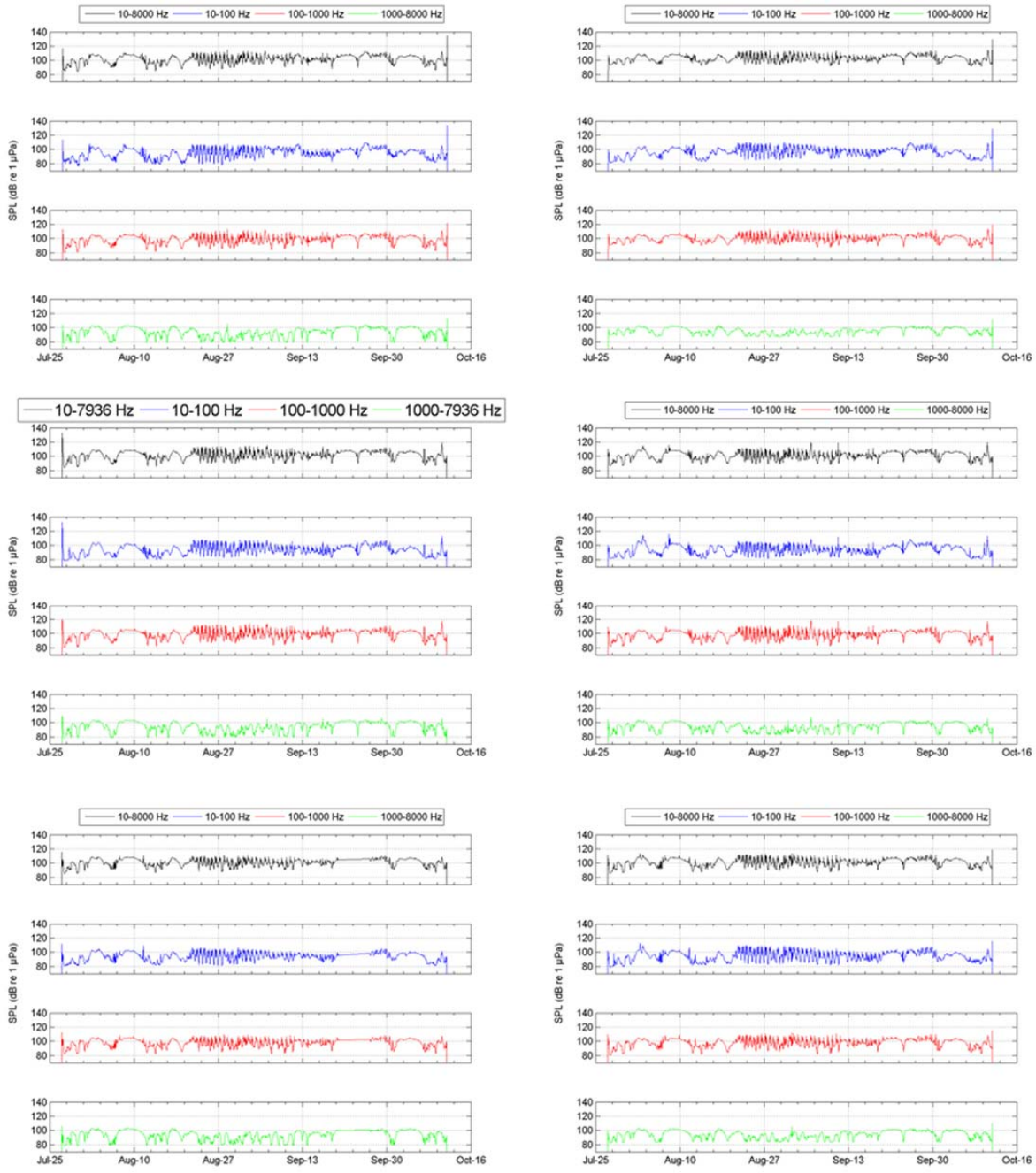

Figure B-21. Broadband and decade band sound pressure levels (SPL) for summer 2010 Stations (top left) KL01, (top right) KL02, (middle left) KL03, (middle right) KL04, (bottom left) KL06, and (bottom right) KL07, July 2010 to October 2010.

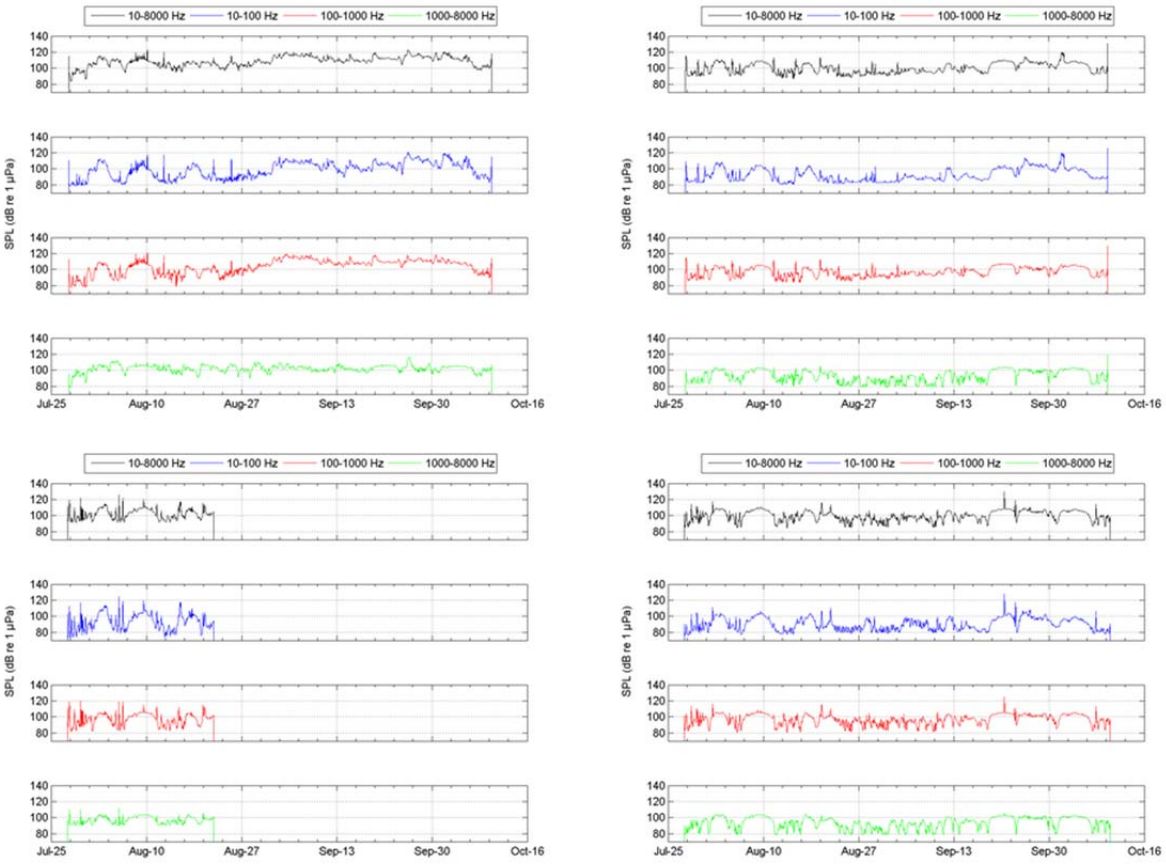

Figure B-22. Broadband and decade band sound pressure levels (SPL) for summer 2010 Stations (top left) PL05, (top right) PL20, (bottom left) PL35, and (bottom right) PL50, July 2010 to October 2010.

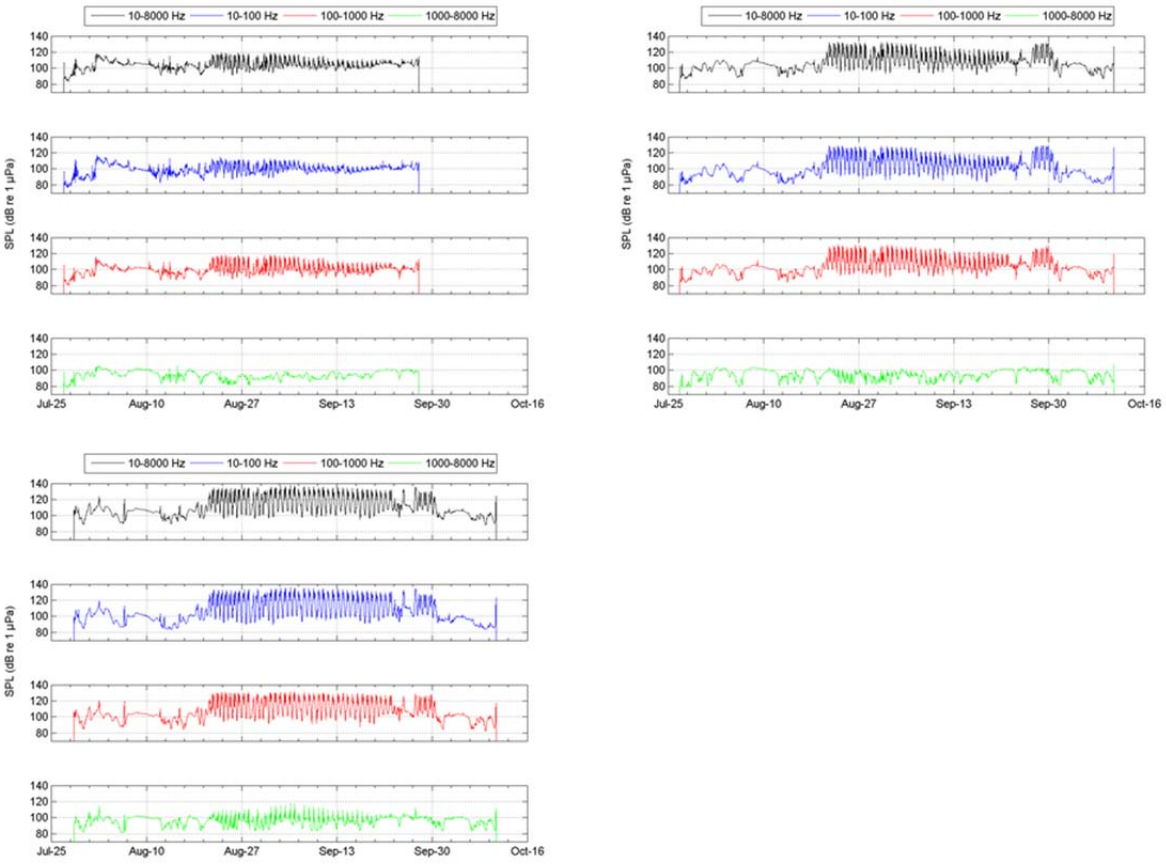

Figure B-23. Broadband and decade band sound pressure levels (SPL) for summer 2010 Stations (top left) PLN40, (top right) PLN60, and (bottom) PLN80, July 2010 to October 2010.

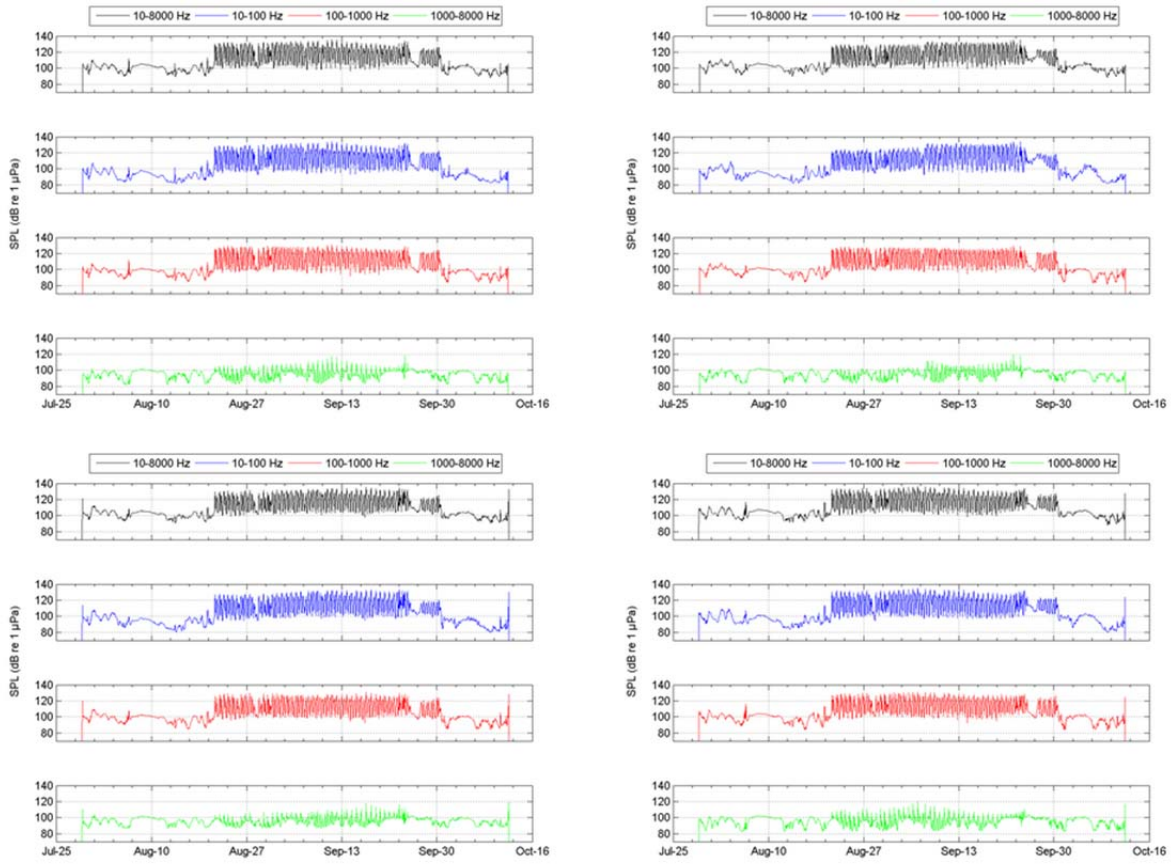

Figure B-24. Broadband and decade band sound pressure levels (SPL) for summer 2010 Stations (top left) S01, (top right) S02, (bottom left) S03, and (bottom right) S04, July 2010 to October 2010.

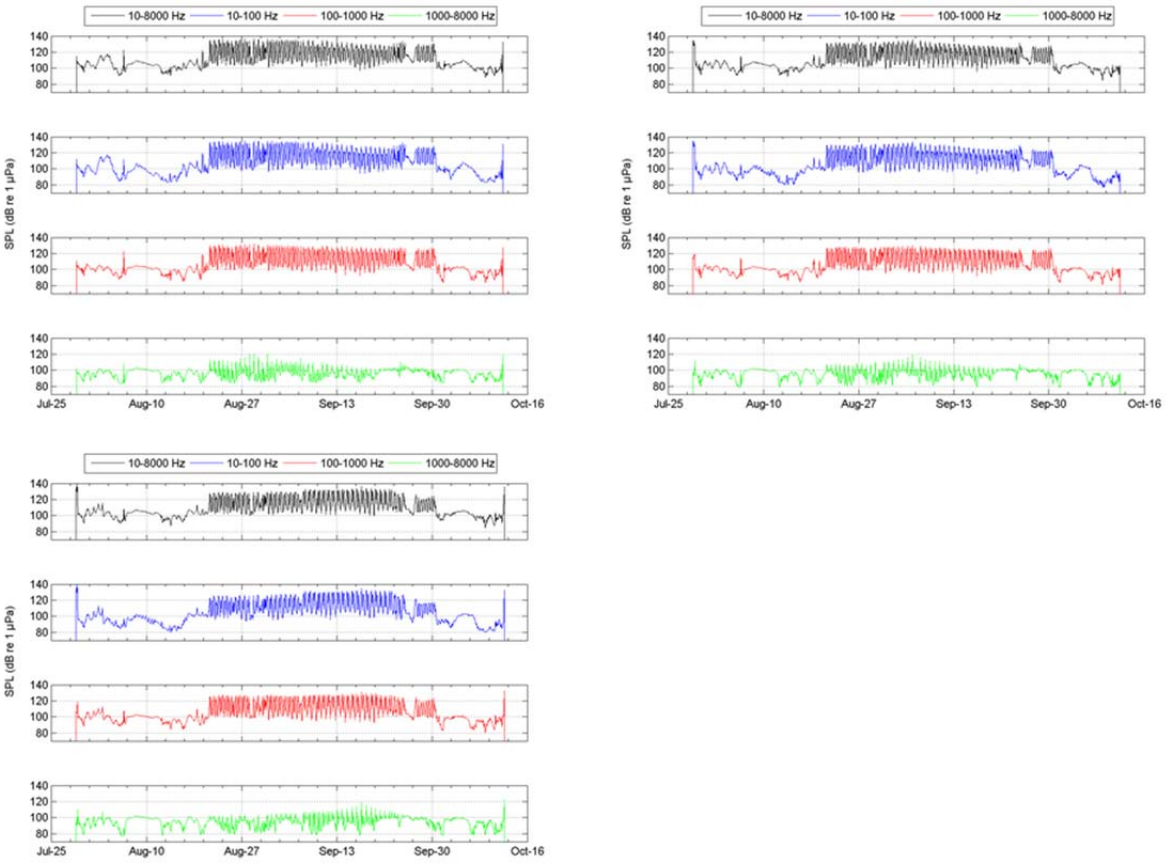

Figure B-25. Broadband and decade band sound pressure levels (SPL) for summer 2010 Stations (top left) S05, (top right) S06, and (bottom) S07, July 2010 to October 2010.

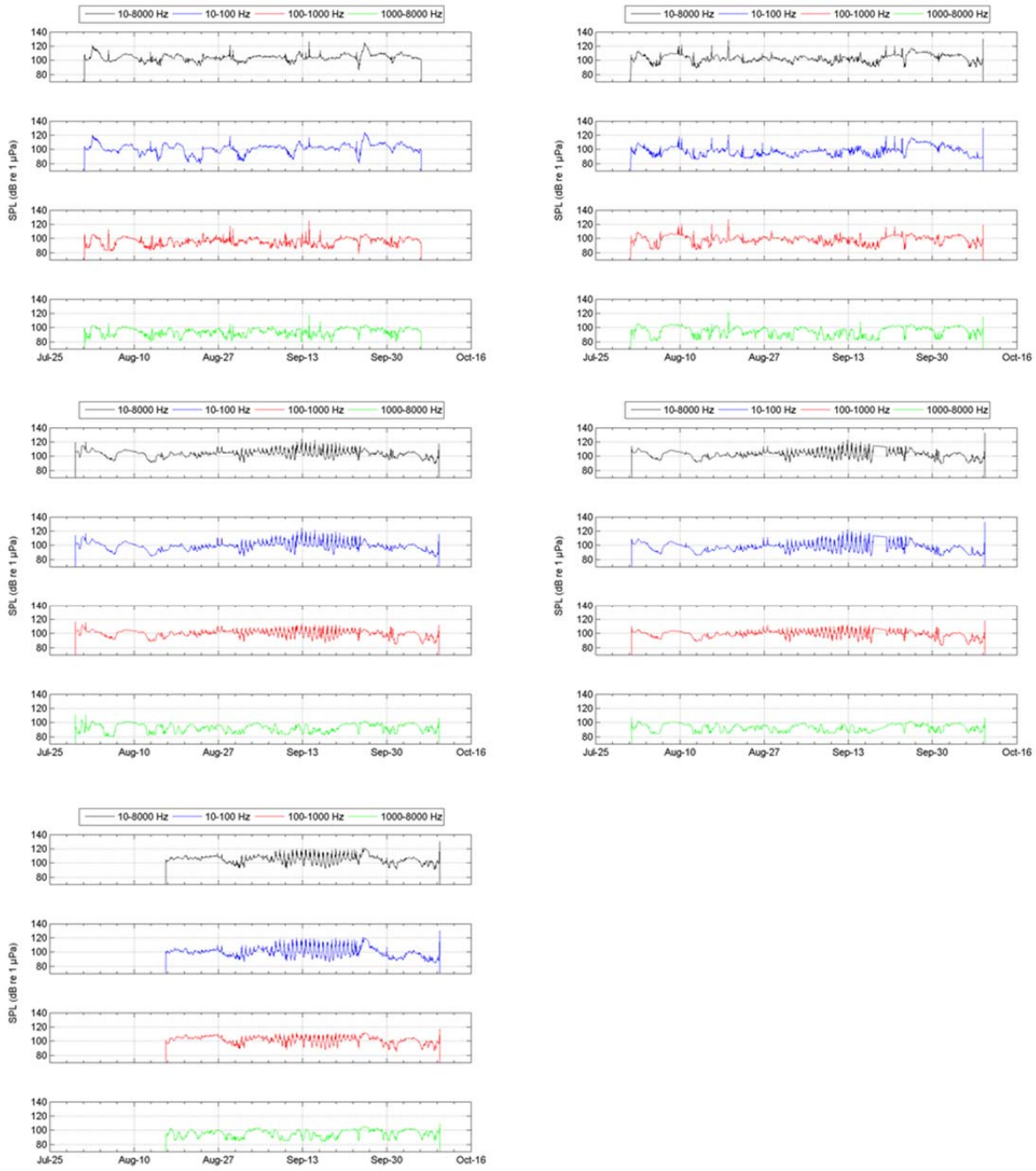

Figure B-26. Broadband and decade band sound pressure levels (SPL) for summer 2010 Stations (top left) W05, (top right) W35, (middle left) WN20A, (middle right) WN20B, and (bottom) WN40, July 2010 to October 2010.

C.2.3. Spectrograms

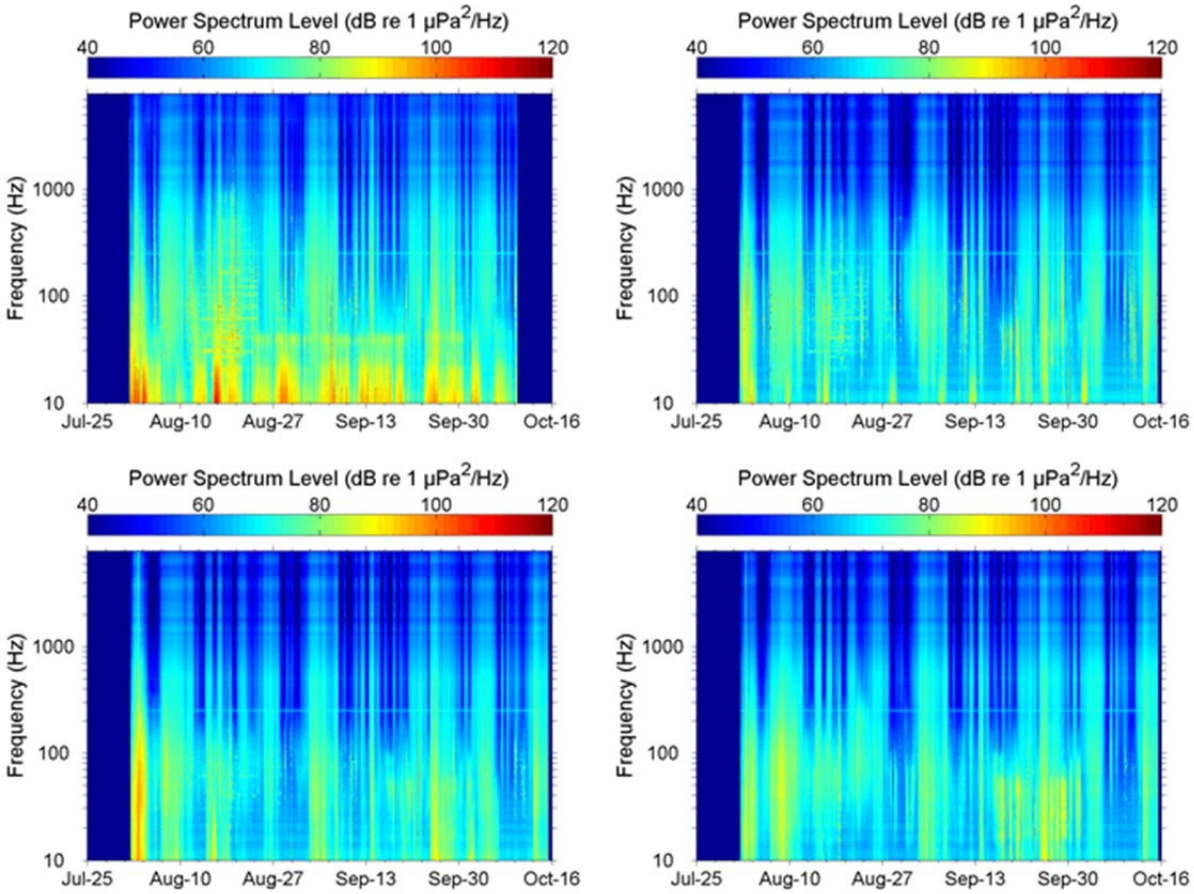

Figure B-27. Spectrogram of underwater sound at summer 2010 Stations (top left) B05, (top right) B15, (bottom left) B30, and (bottom right) B50, July 2010 to October 2010.

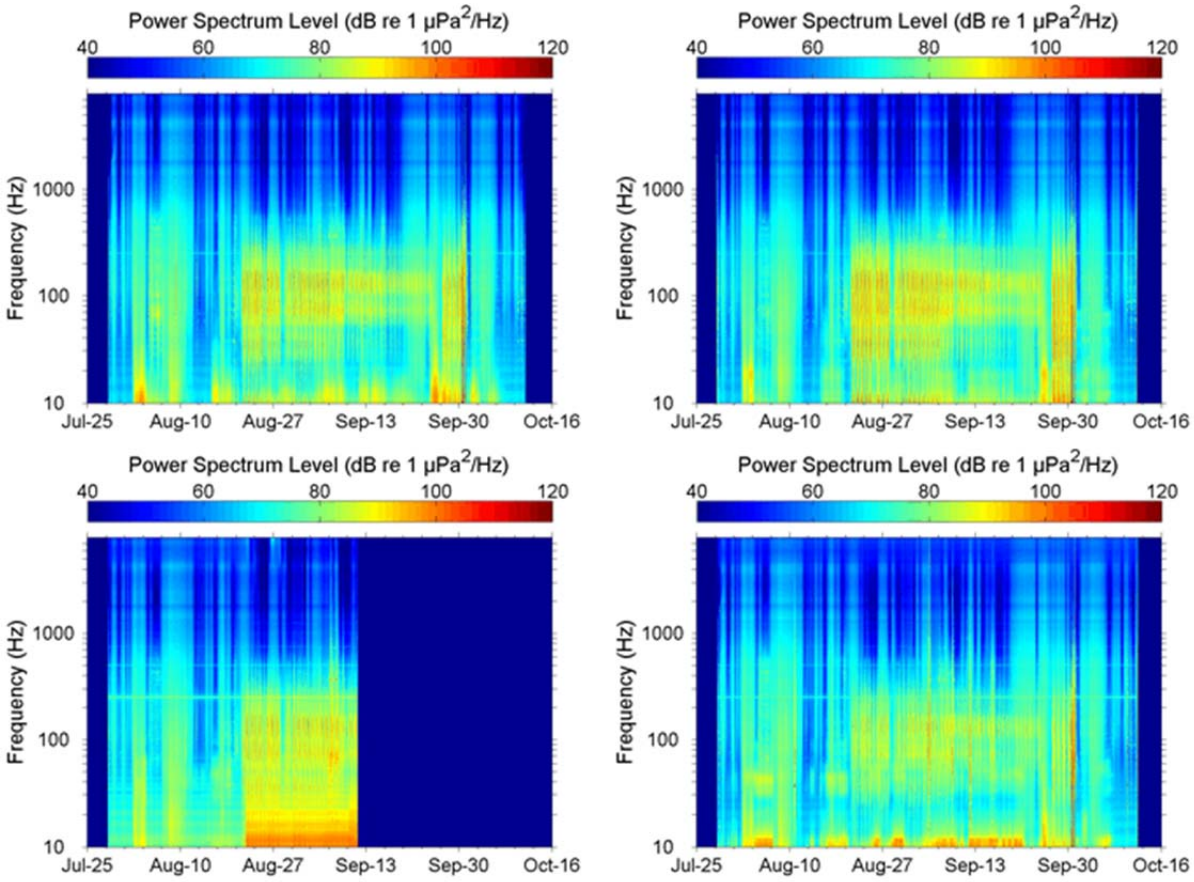

Figure B-28. Spectrogram of underwater sound at summer 2010 Stations (top left) BG01, (top right) BG02, (bottom left) BG03, and (bottom right) BG04, July 2010 to October 2010.

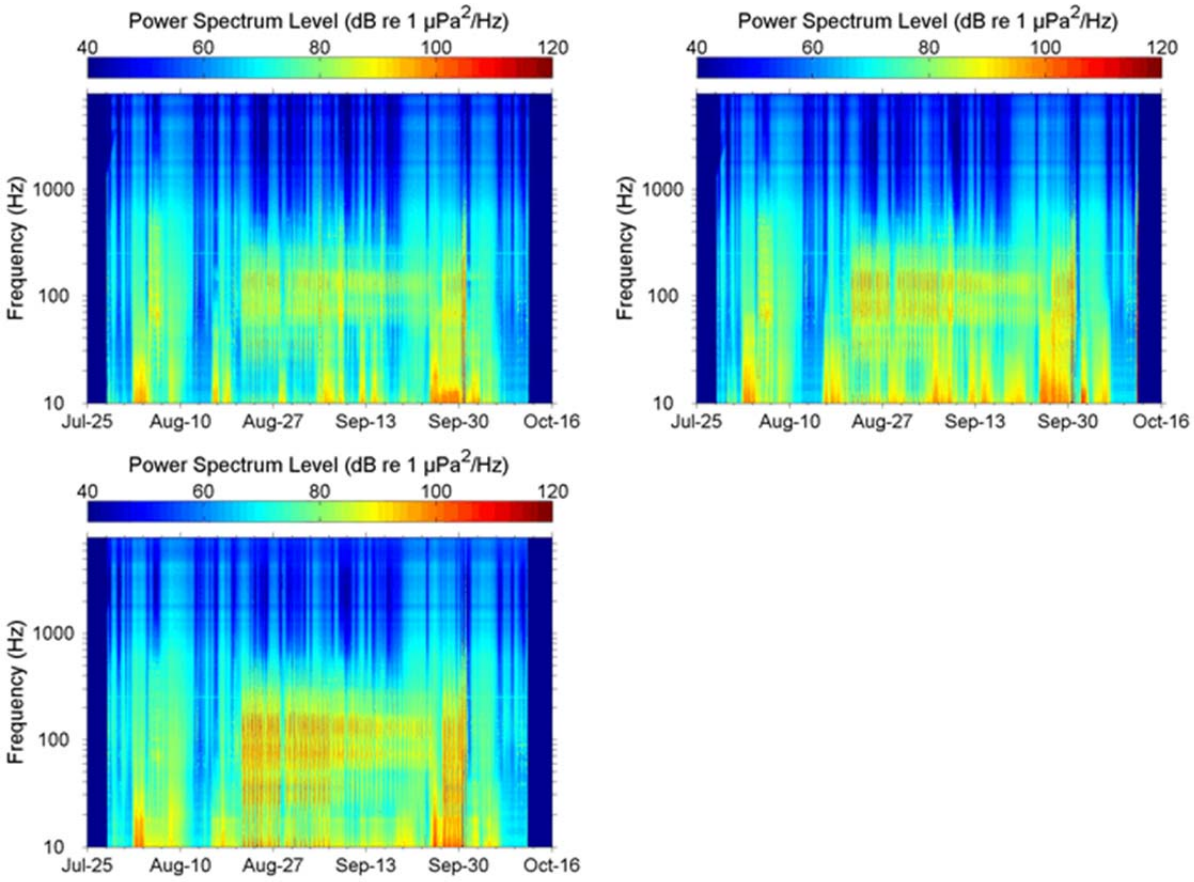

Figure B-29. Spectrogram of underwater sound at summer 2010 Stations (top left) BG05, (top right) BG06, and (bottom) BG07, July 2010 to October 2010.

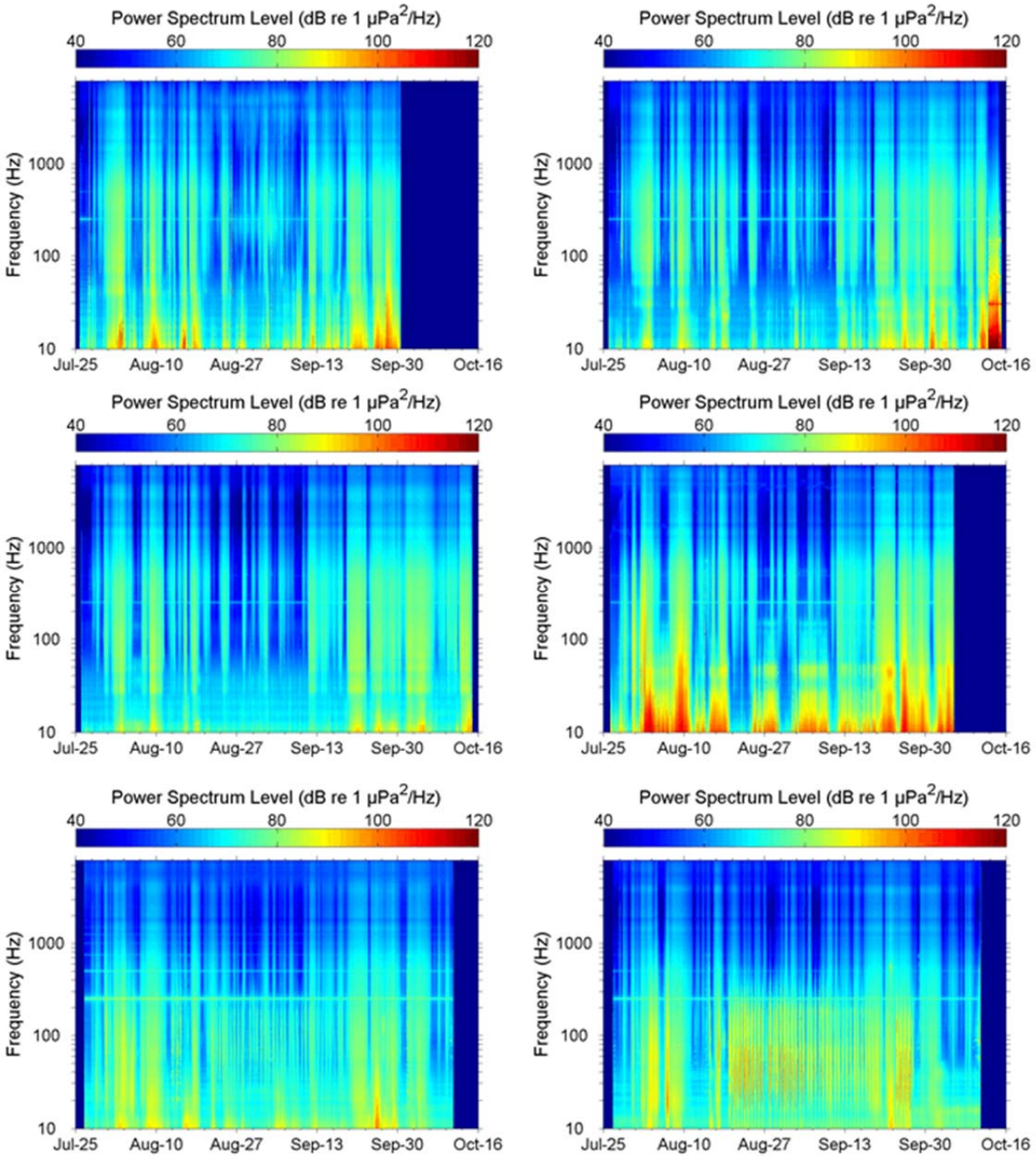

Figure B-30. Spectrogram of underwater sound at summer 2010 Stations (top left) CL05, (top right) CL20, (middle left) CL50, (middle right) CLN40, (bottom left) CLN90, and (bottom right) CLN120, July 2010 to October 2010.

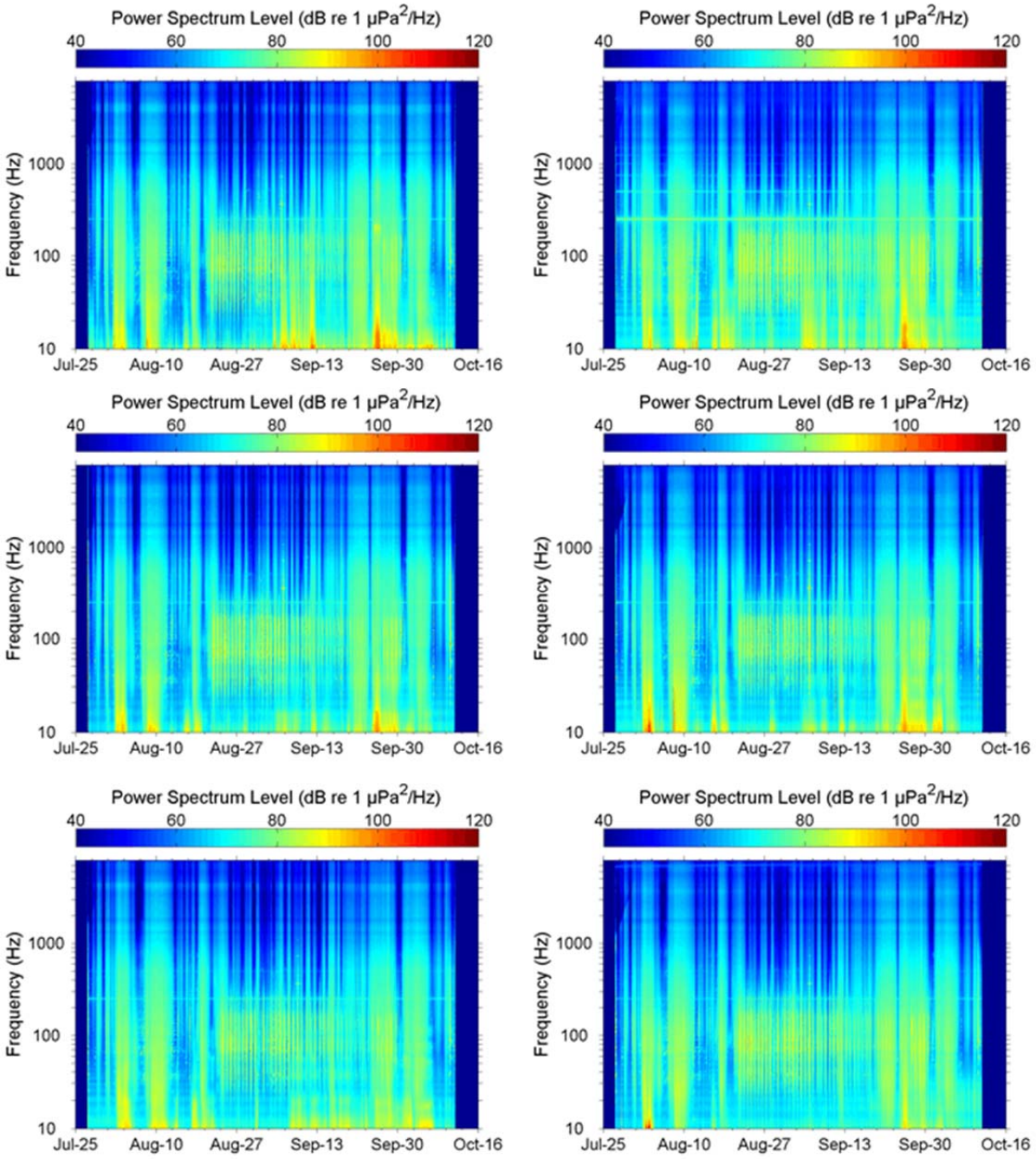

Figure B-31. Spectrogram of underwater sound at summer 2010 Stations (top left) KL01, (top right) KL02, (middle left) KL03, (middle right) KL04, (bottom left) KL06, and (bottom right) KL07, July 2010 to October 2010.

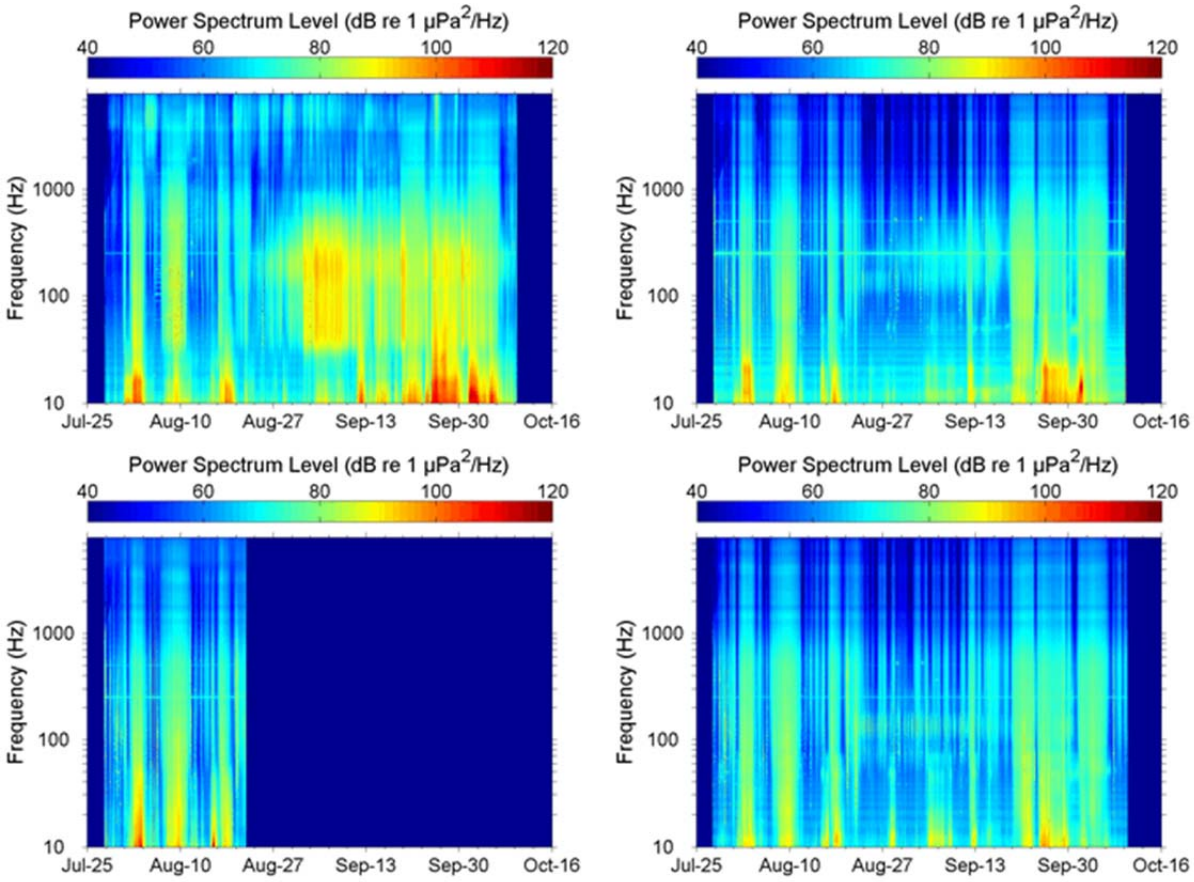

Figure B-32. Spectrogram of underwater sound at summer 2010 Stations (top left) PL05, (top right) PL20, (bottom left) PL35, and (bottom right) PL50, July 2010 to October 2010.

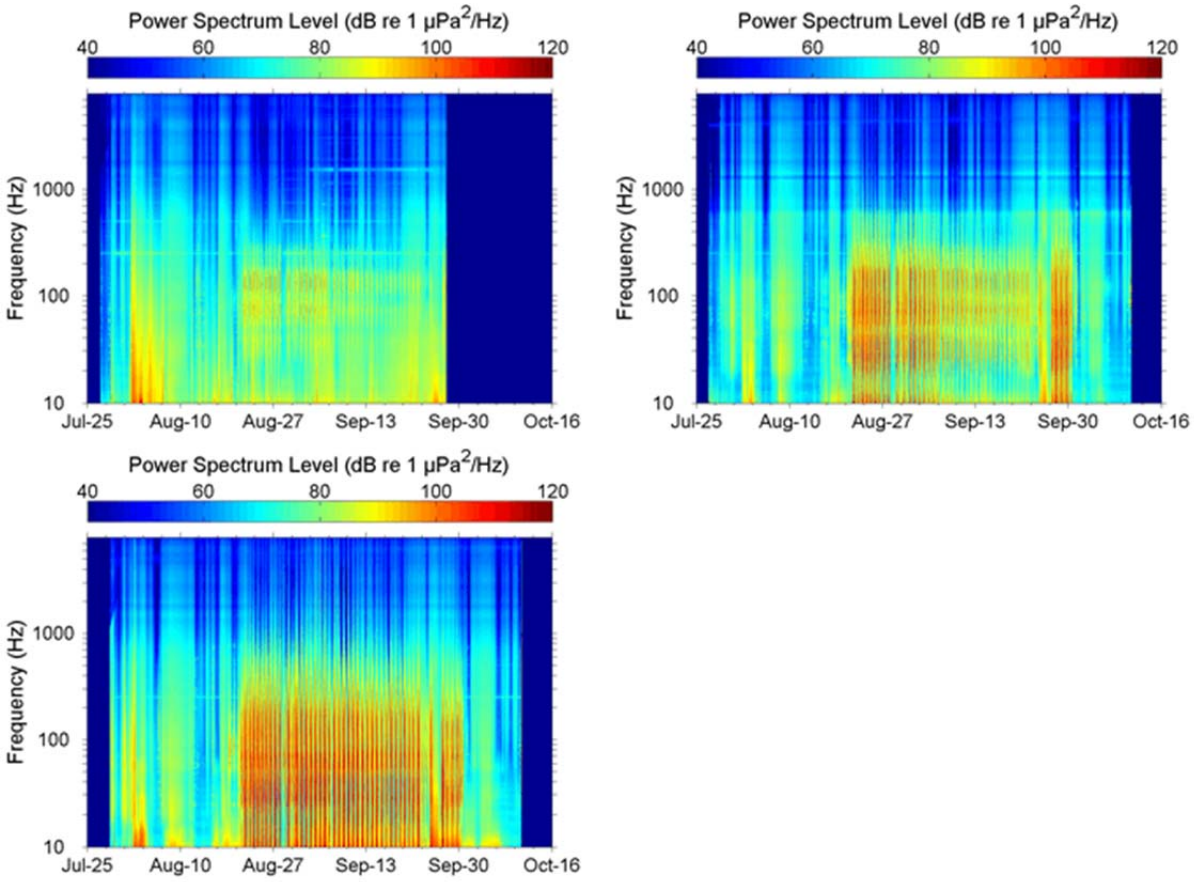

Figure B-33. Spectrogram of underwater sound at summer 2010 Stations (top left) PLN40, (top right) PLN60, and (bottom) PLN80, July 2010 to October 2010.

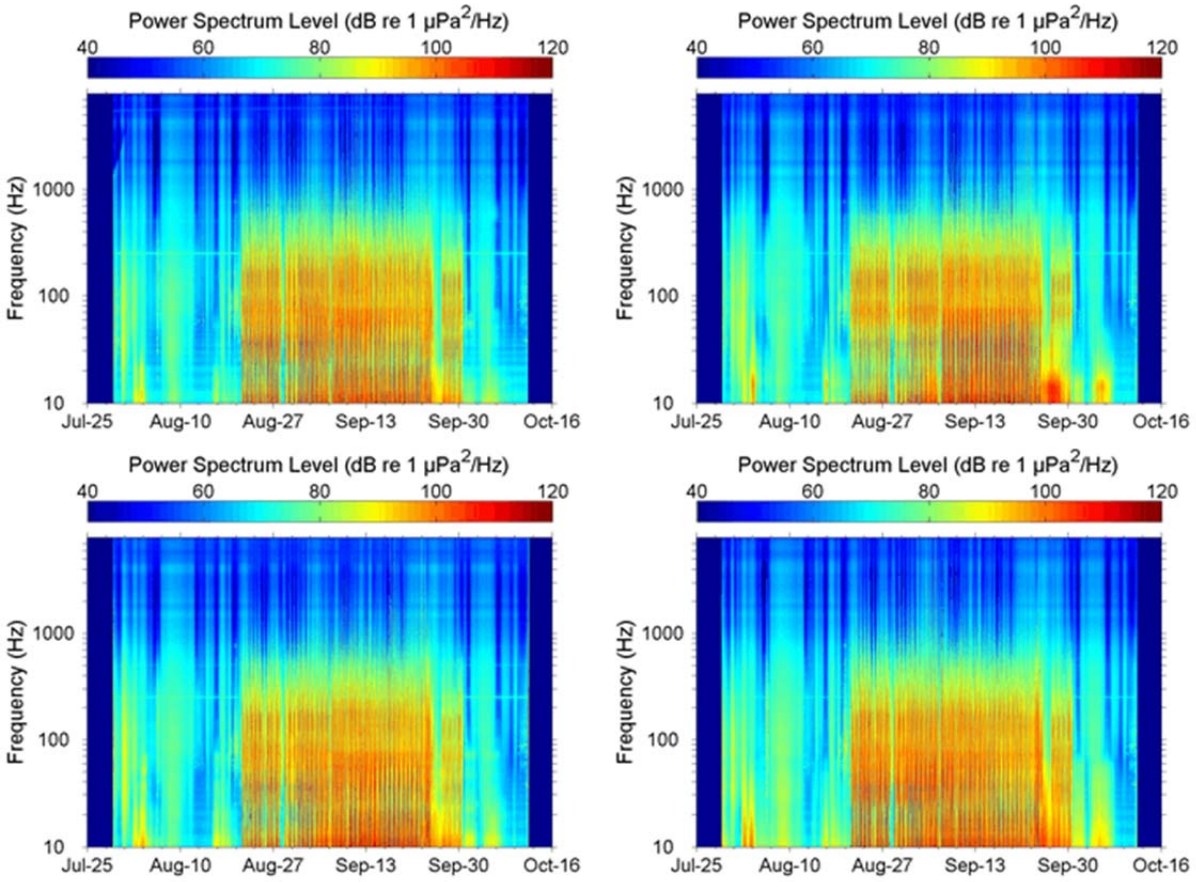

Figure B-34. Spectrogram of underwater sound at (top left) S01, (top right) S02, (bottom left) S03, and (bottom right) S04, July 2010 to October 2010.

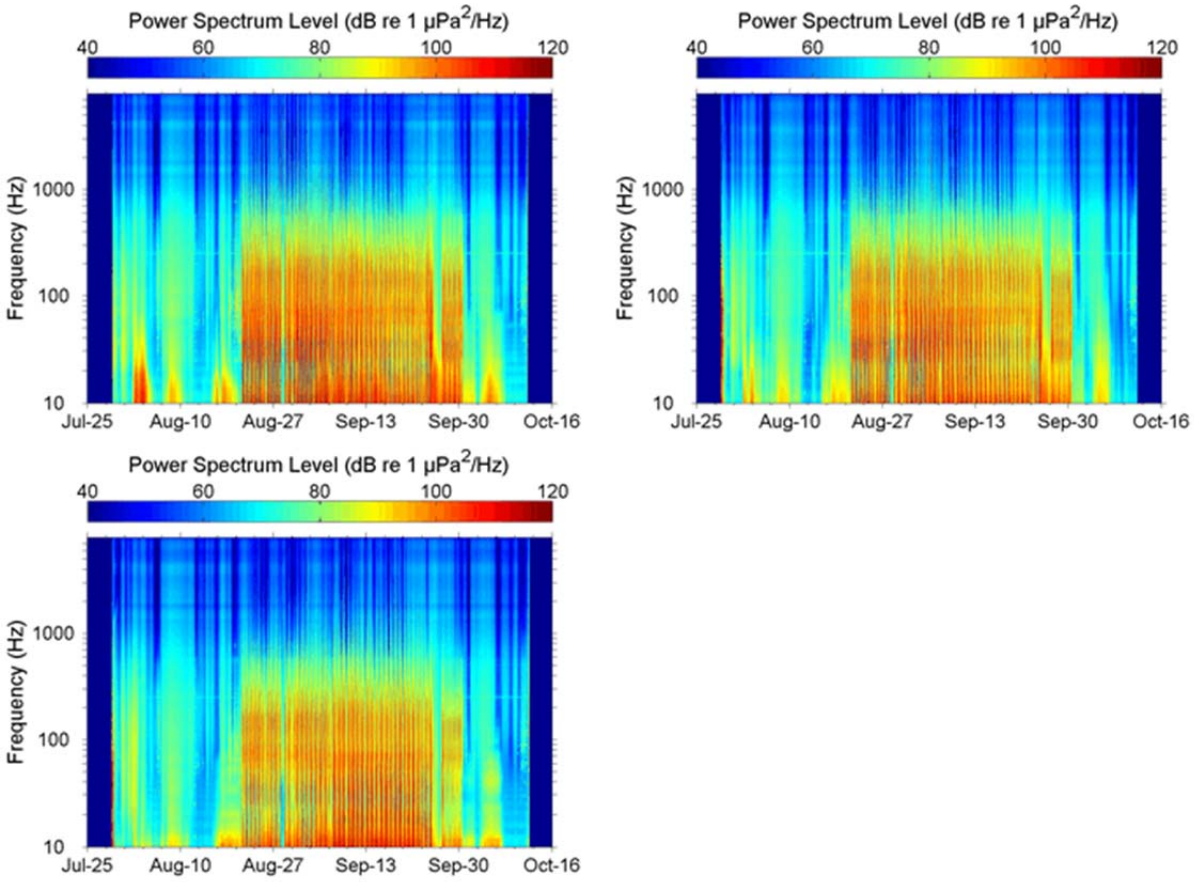

Figure B-35. Spectrogram of underwater sound at (top left) S05, (top right) S06, and (bottom) S07, July 2010 to October 2010.

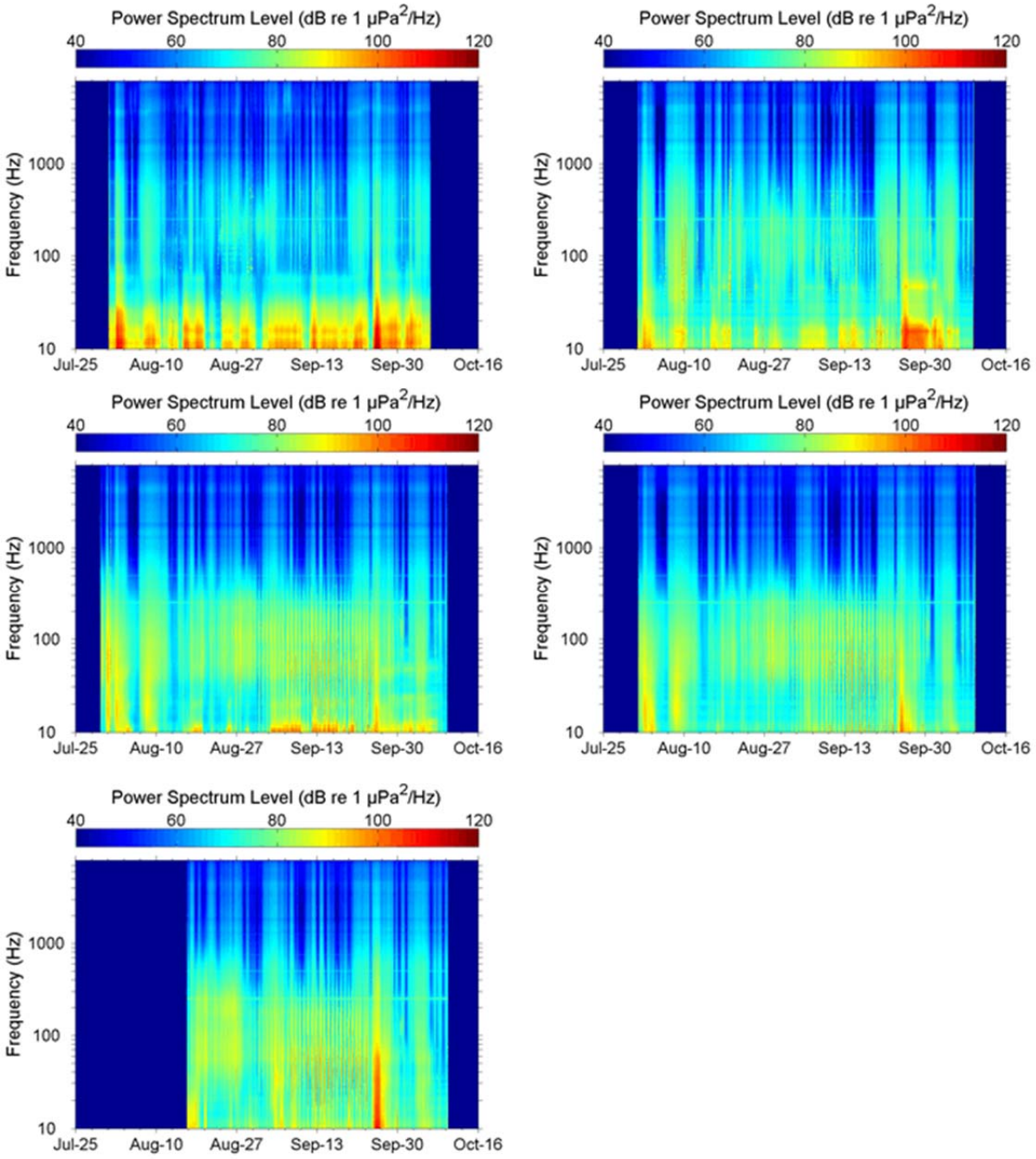

Figure B-36. Spectrogram of underwater sound at (top left) W05, (top right) W35, (middle left) WN20A, (middle right) WN20B, and (bottom) WN40, July 2010 to October 2010.