Supplementary material: Acoustic detectability of whales amidst underwater noise off the west coast of South Africa

Fannie W. Shabangu, Dawit Yemane, George Best, Bobbi J. Estabrook

Equations for calculating the absolute wind speed and direction

Equations used to calculate the absolute wind speed and wind direction from u and v vectors:

$$ws = \sqrt{u^2 + v^2} \quad (1)$$

$$wdt = atan2(\frac{u}{ws}, \frac{v}{ws}) \quad (2)$$

$$wdd = wdt * 180/\pi \quad (3)$$

$$wdn = wdd + 180 \quad (4)$$

where *ws* in equations (1) and (2) represents the absolute wind speed, *u* in equations (1) and (2) represents zonal wind speed, *v* in equations (1) and (2) represents meridional wind speed, *wdt* in equations (2) and (3) represents wind direction trigonometry, *wdd* in equations (3) and (4) represents wind direction trigonometry to degrees, *wdn* in equation (4) represents meteorological wind direction angle (degrees North).

Noise level relative to current speed

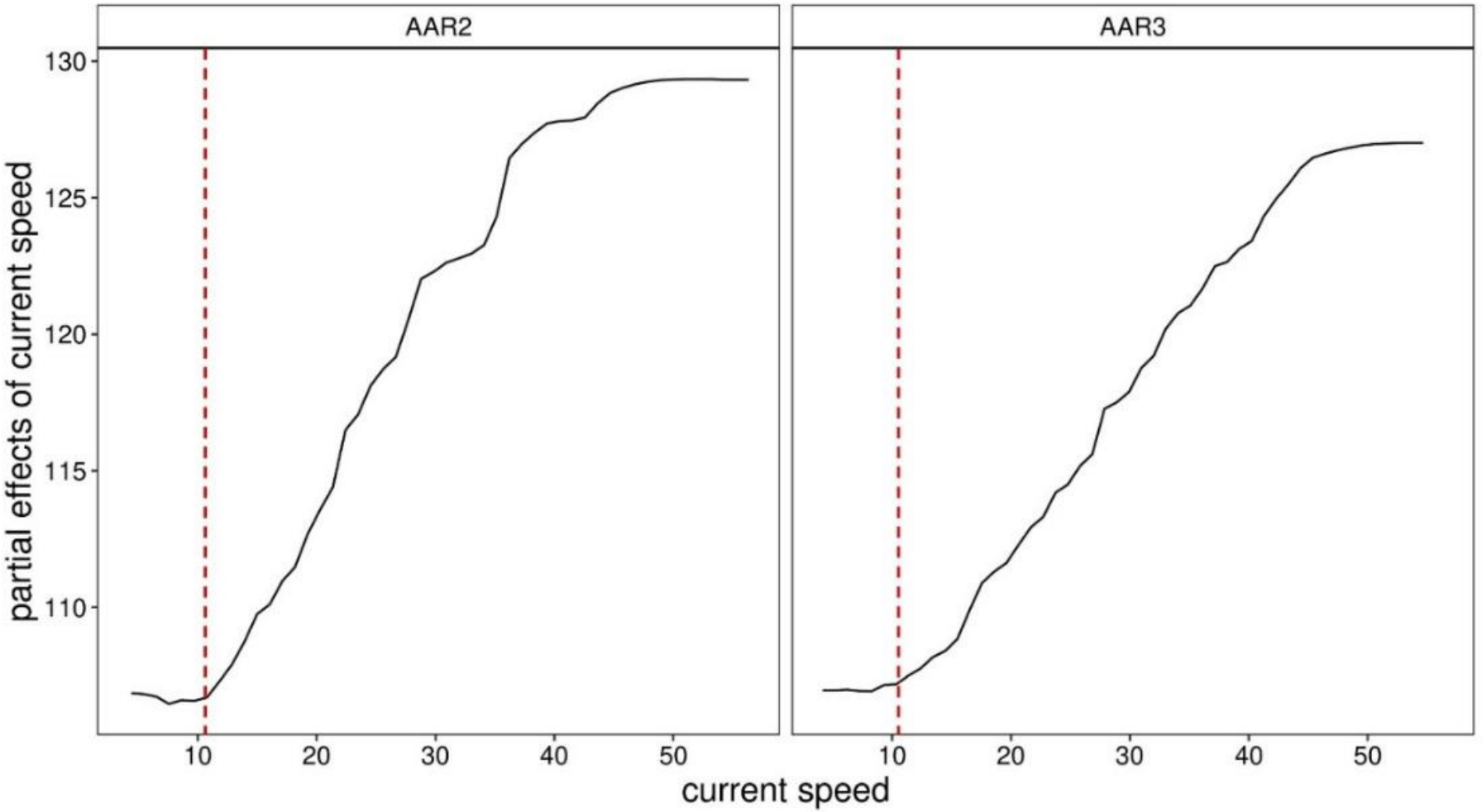

Fig. S1. Partial effects of ocean current speed on the recorded noise (including pseudo-noise) level at 10-500 Hz frequency band from AARs 2 and 3 based on random forest model regression. The red dashed vertical lines represent the 11 cm s^{-1} ocean current speed point used for filtering out pseudo-noise.

Long term spectrograms

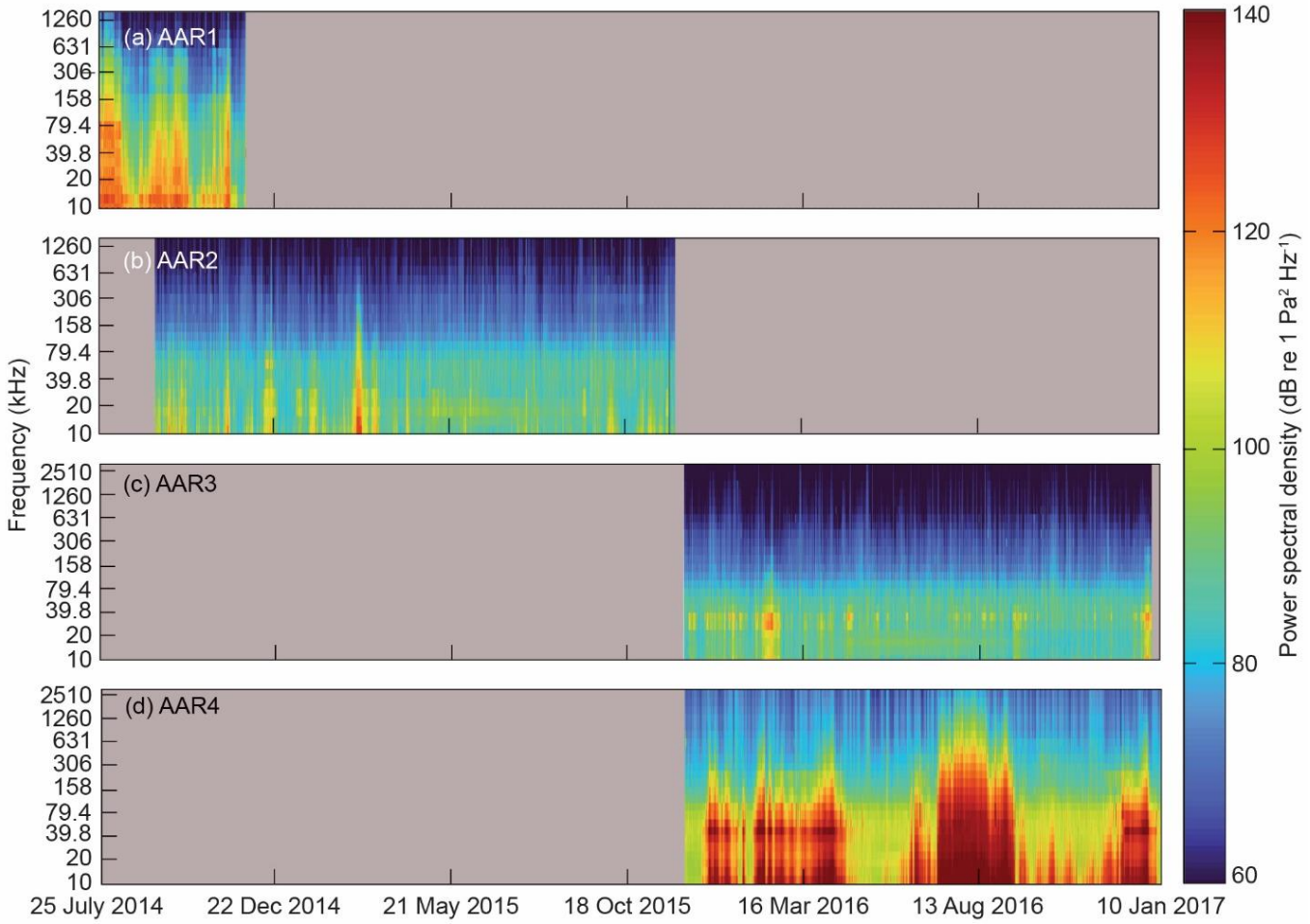

Fig. S2. 1/3-octave long-term spectrograms of the recorded noise level including pseudo-noise expressed as power spectral density (dB re 1 μPa^2 Hz^{-1}), averaged over 1 hour time intervals for each recording site: (a) AAR1, (b) AAR2, (c) AAR3 and (d) AAR4. Grey shaded areas indicate time periods without acoustic recordings. Frequency scales (y-axes) for panels (a) and (b) differ from panels (c) and (d).

Spatio-temporal variation of recorded noise including pseudo-noise

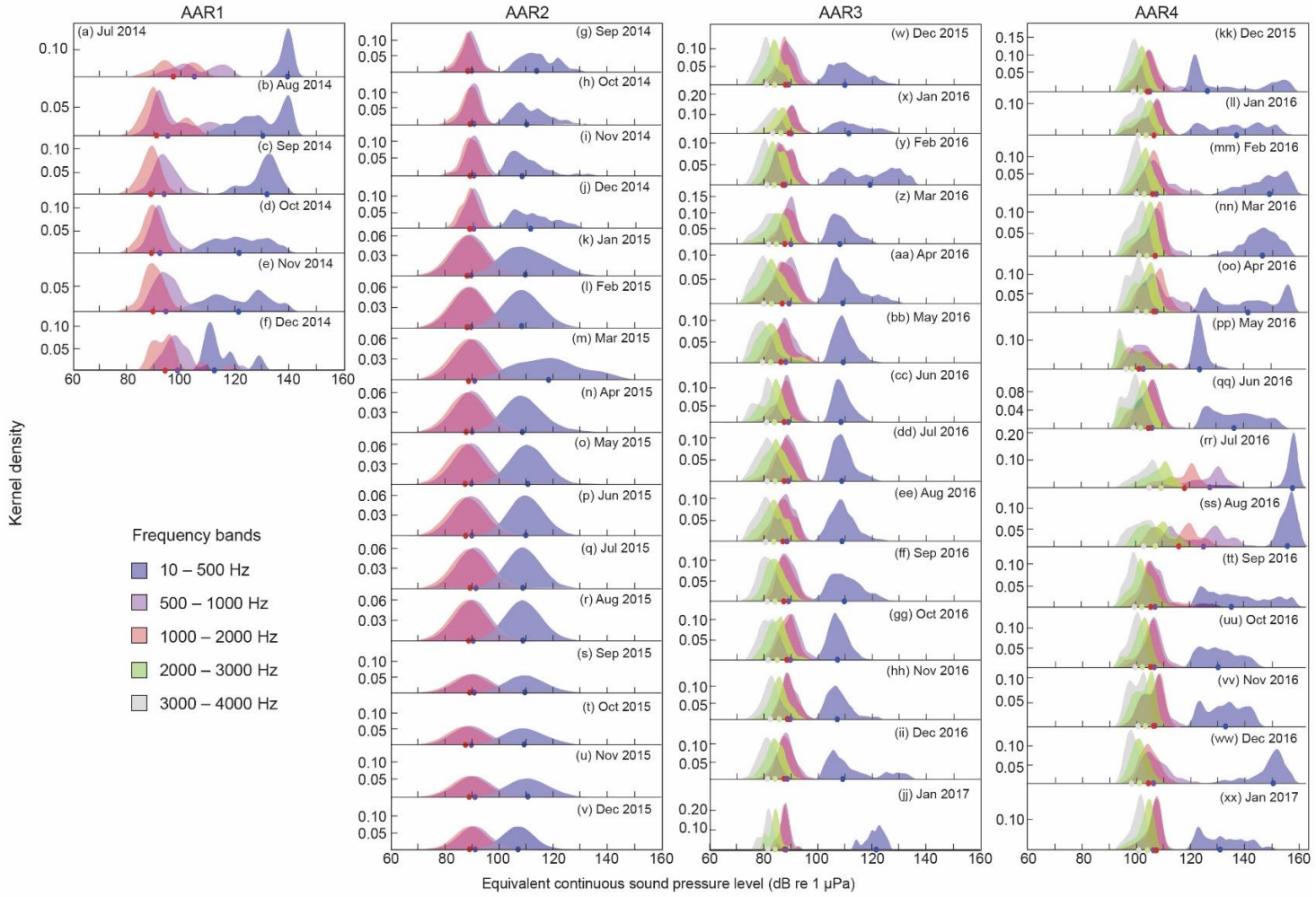

Fig. S3. Full timeseries of kernel density plots showing distributions and trends of the recorded equivalent continuous sound pressure level (L_{eq}) including pseudo-noise for each autonomous acoustic recorder (AAR) across different frequency bands. Closed circles at the bottom of each density plot represent the average L_{eq} for each frequency band. Y-axes scales are different between plots**.**

Exploring environmental data

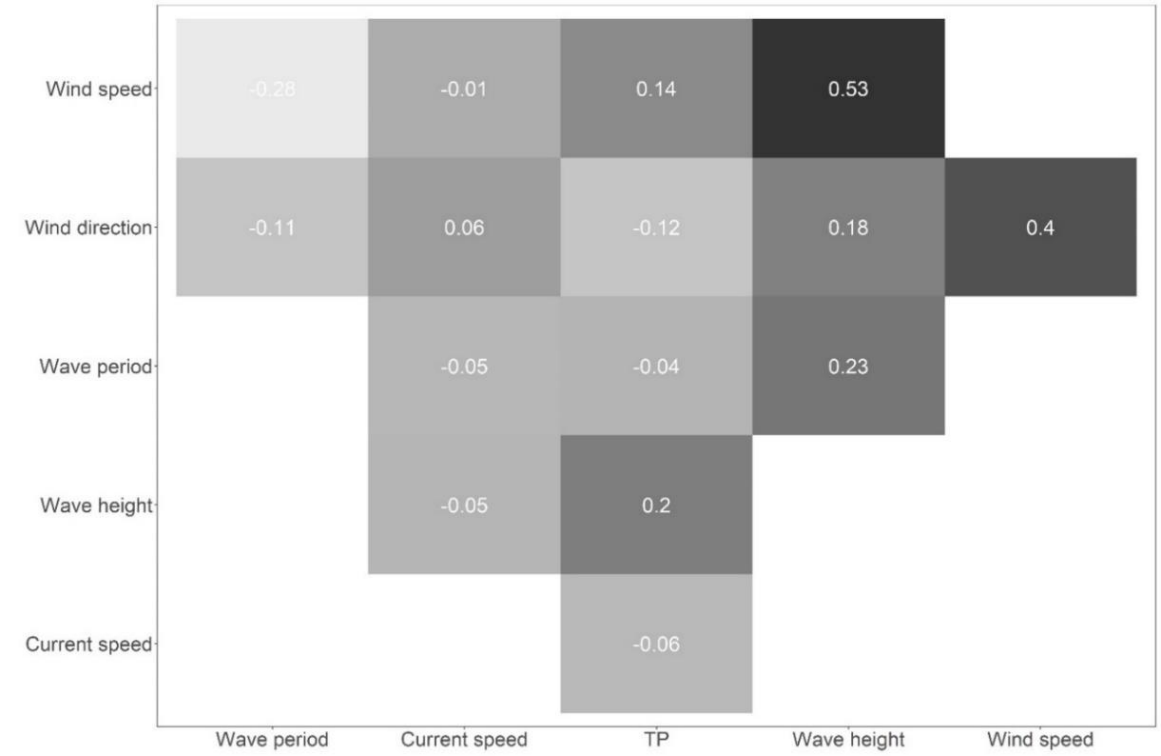

Fig. S4. Pair-wise correlation matrix among environmental condition associated predictors of AAR2 ambient noise.

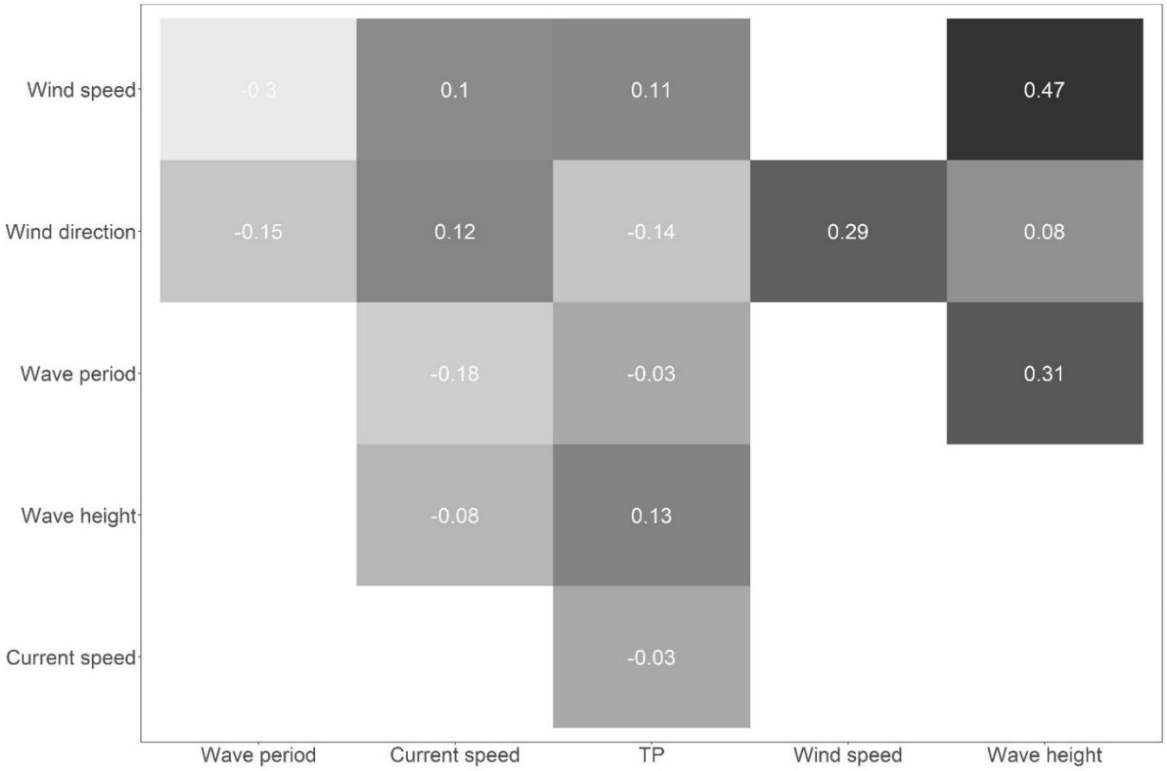

Fig. S5. Pair-wise correlation matrix among environmental conditions associated predictors of AAR3 ambient noise.

Exploring noise data

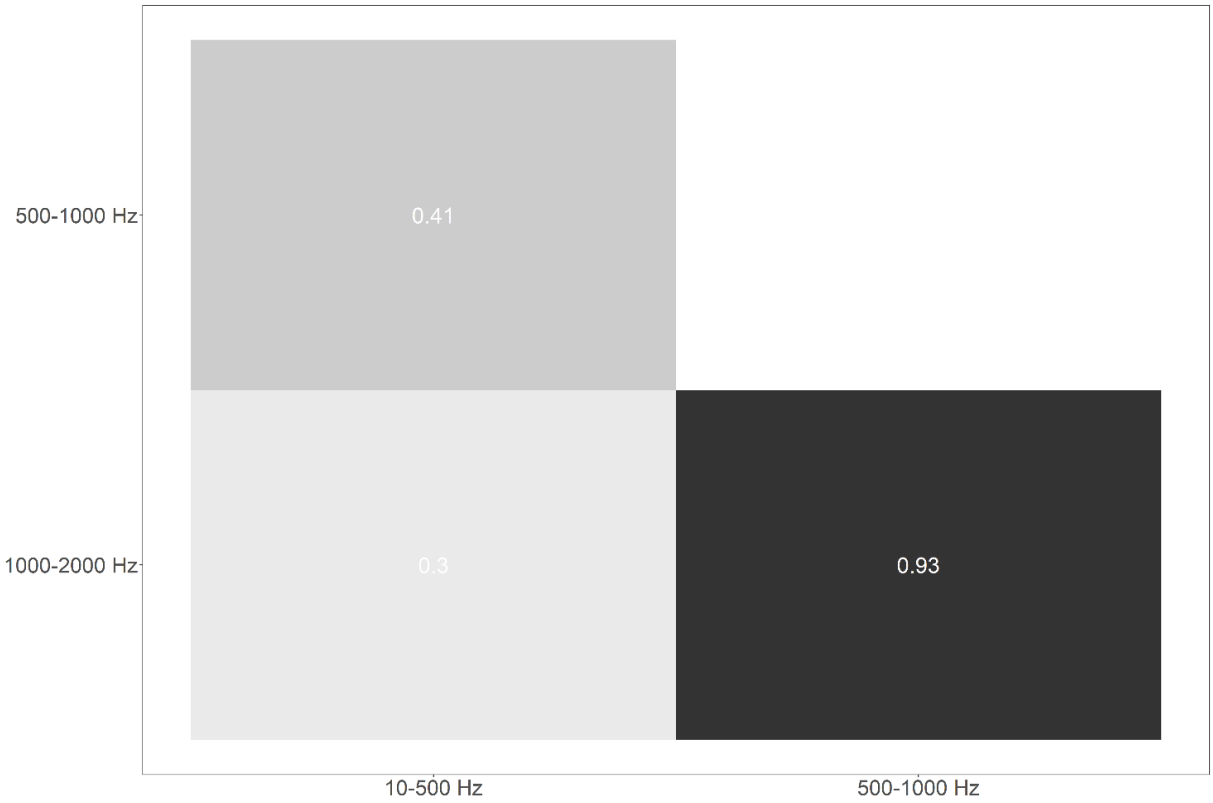

Fig. S6. Pair-wise correlation matrix of ambient noise level among frequency bands for AAR2 station.

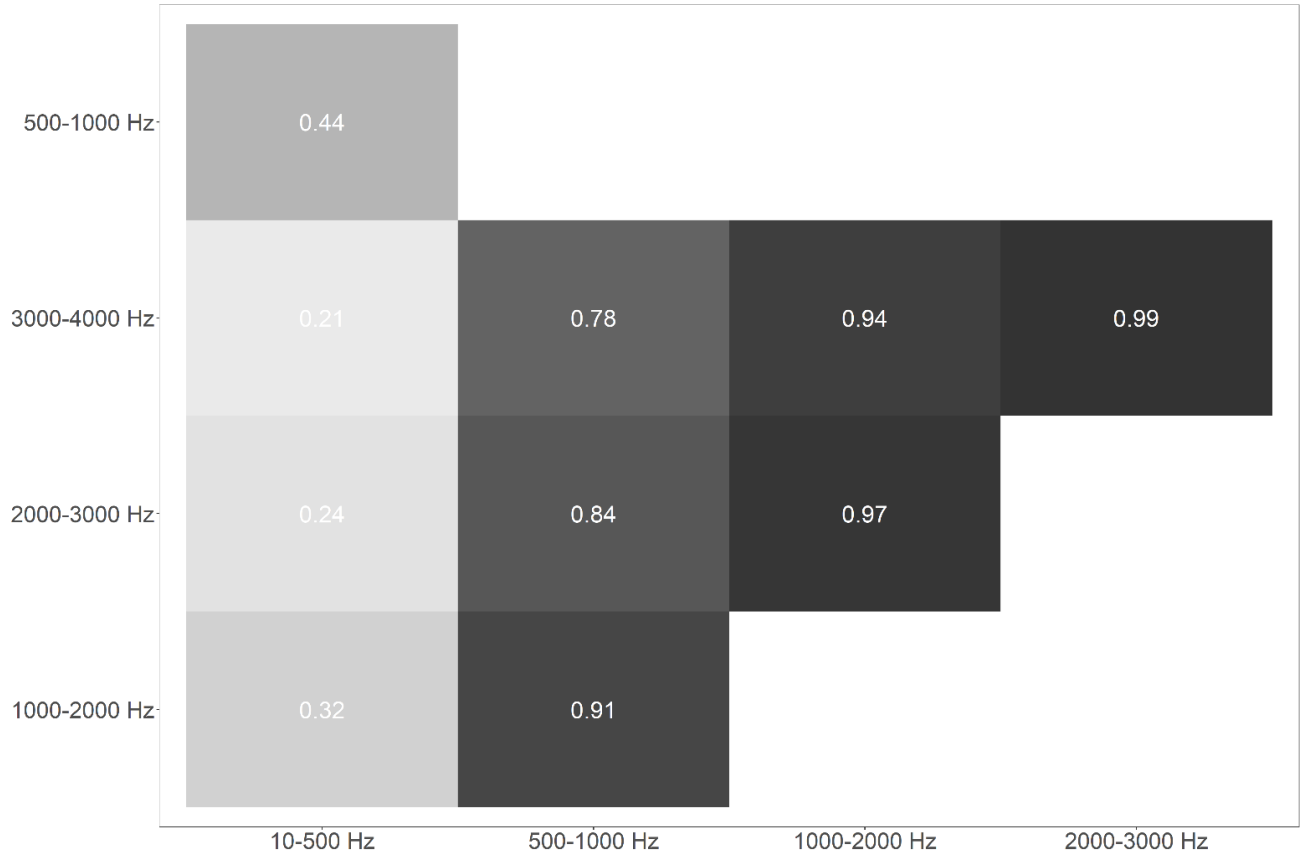

Fig. S7. Pair-wise correlation matrix of ambient noise level among frequency bands for AAR3 station.

GLM and RF formulas

Table S1. Formula for generalized linear model (GLM) fitted to evaluate effects of ambient noise level on acoustic detectability of the five whale species, and for random forest (RF) fitted to evaluate predictors of the ambient equivalent continuous sound pressure level (L_{eq}) from AARs 2 and 3. Lowest frequency band was 10-500 Hz for all AARs; highest frequency band was 1000-2000 and 3000-4000 Hz for AARs 2 and 3 respectively.

Model	Formula
Noise level RF model	L_{eq} ~ *Closest vessel distance + Month + Hour + Wind speed + Wind direction + Total precipitation + Wave height + Wave period + Number of vessels + Speed over ground + Current speed*
Whale detectability GLM model	*Species detection ~ Lowest frequency band + Highest frequency band*

R packages used

Table S2. List of R packages used for data manipulation, visualization, and analyses.

Package name	Use	Reference
broom	Data manipulation	Robinson, D., Hayes, A., Couch, S., 2021. broom: Convert statistical objects into tidy tibbles. https://CRAN.R-project.org/package=broom.
dplyr	Data manipulation	Wickham, H., François, R., Henry, L., Müller, K., 2021. dplyr: A grammar of data manipulation. https://CRAN.R-project.org/package=dplyr.
ggplot2	Visualization	Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.
knitr	Report writing	Xie, Y., 2021. knitr: A general-purpose package for dynamic report generation in R. https://yihui.org/knitr/.
lubridate	Data manipulation	Grolemund, G., Wickham, H., 2011. Dates and times made easy with lubridate. J. Stat. Softw. 40(3), 1–25.
purrr	Data manipulation and analyses	Henry, L., Wickham, H., 2020. purrr: Functional programming tools. https://CRAN.R-project.org/package=purrr.
rmarkdown	Report writing	Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., Iannone, R., 2021. rmarkdown: Dynamic documents for R. https://CRAN.R-project.org/package=rmarkdown.
tidymodels	Data manipulation and analyses	Kuhn, M., Wickham, H., 2020. tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org.
vip	Data analyses	Greenwell B.M., Boehmke B.C., 2020. Variable Importance Plots-An Introduction to the vip Package. The R Journal 12, 343–366.

Temporal noise trends excluding pseudo-noise

Fig. S8. Full timeseries of kernel density plots showing monthly (right scale) trends of ambient equivalent continuous sound pressure level (L_{eq}) for AAR2 across different frequency bands. Closed circles at the bottom of each density plot represent the average L_{eq} for each frequency band. Y-axes scales are different between plots.

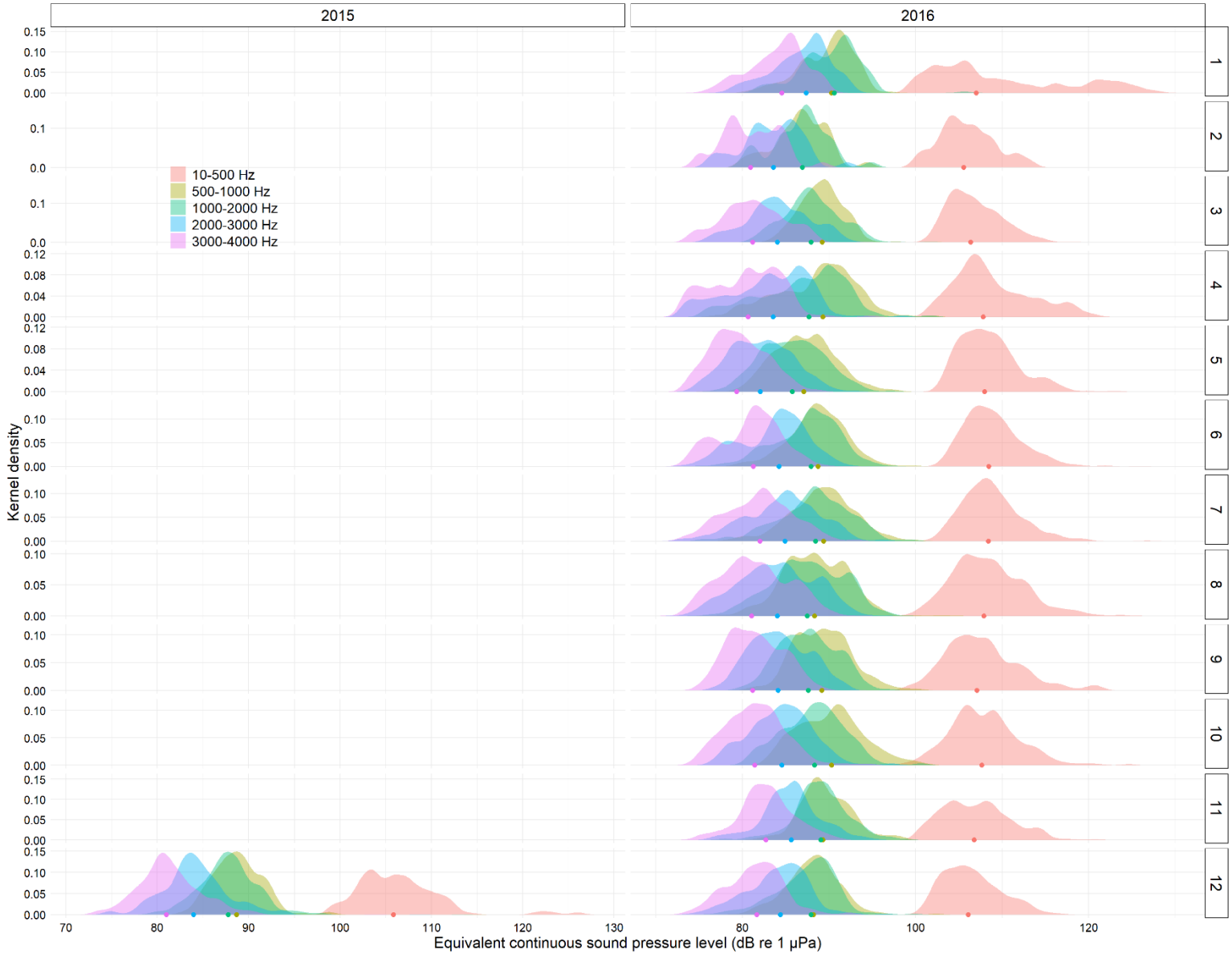

Fig. S9. Full timeseries of kernel density plots showing monthly (right scale) trends of ambinet equivalent continuous sound pressure level (L_{eq}) for AAR3 across different frequency bands. Closed circles at the bottom of each density plot represent the average L_{eq} for each frequency band. Y-axes scales are different between plots.

Model performance: ambient noise model

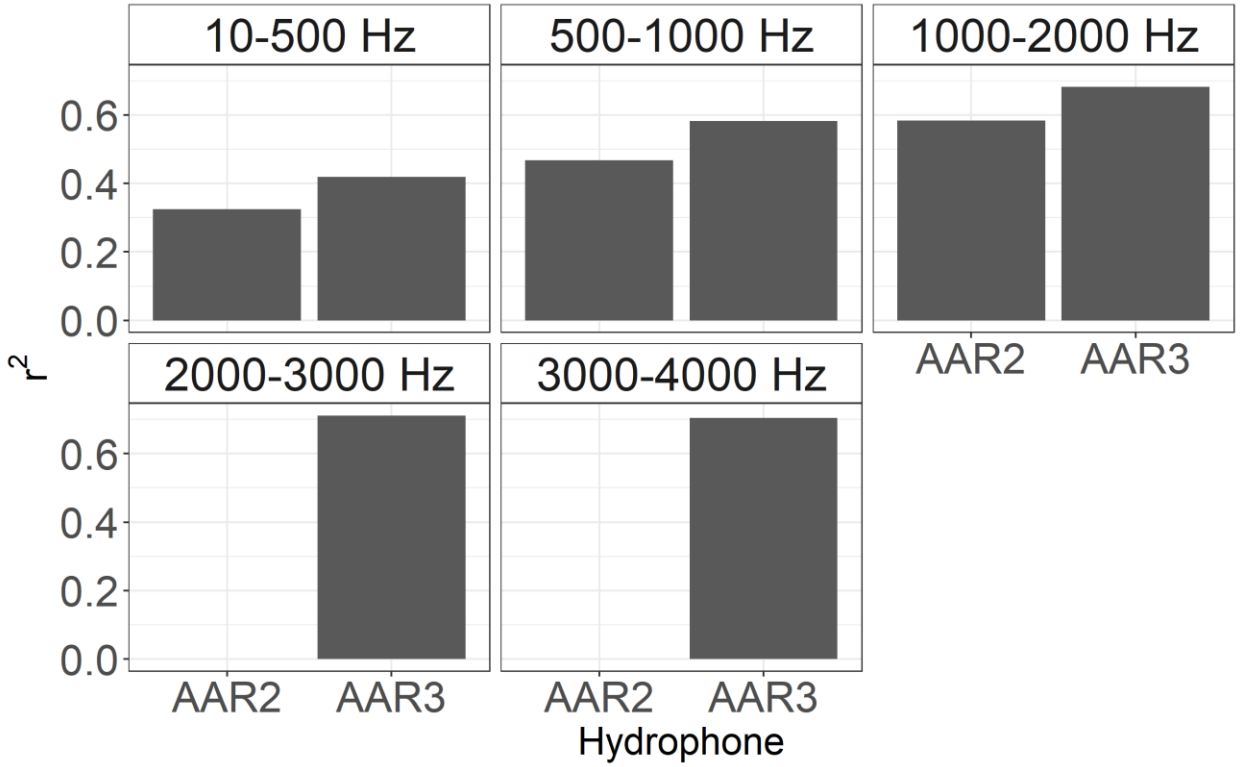

Fig. S10. Random forest model predictive performance, as measured by r-squared, of the ambient noise model for AAR2 and AAR3 at different frequency bands.

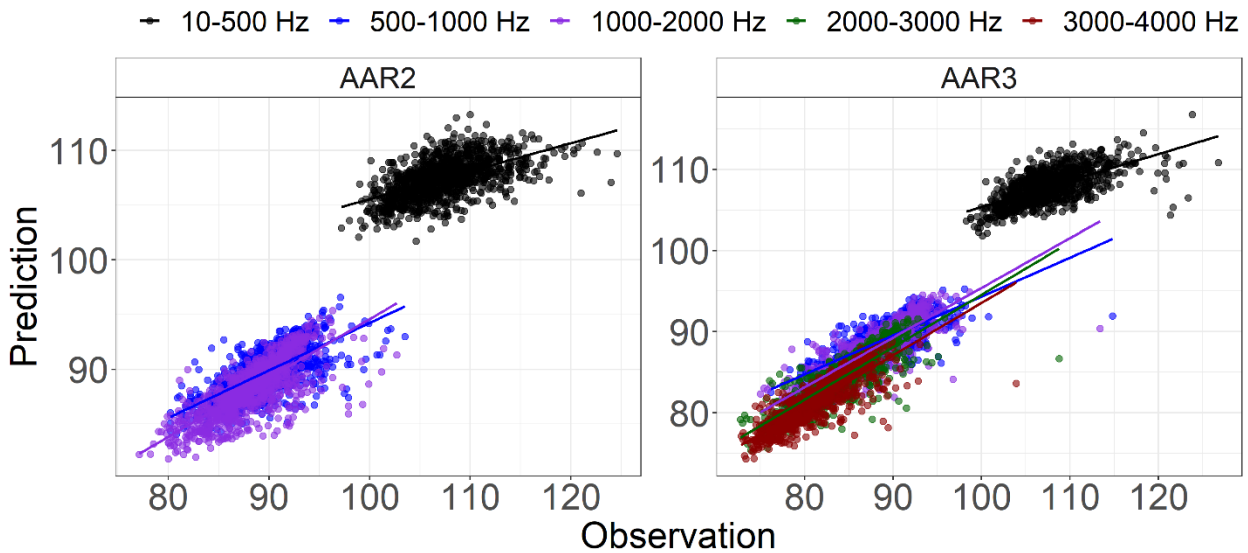

Fig. S11. Performance of random forest models on test data set for AAR2 and AAR3 at different frequency bands.

Model performance: acoustic detectability model

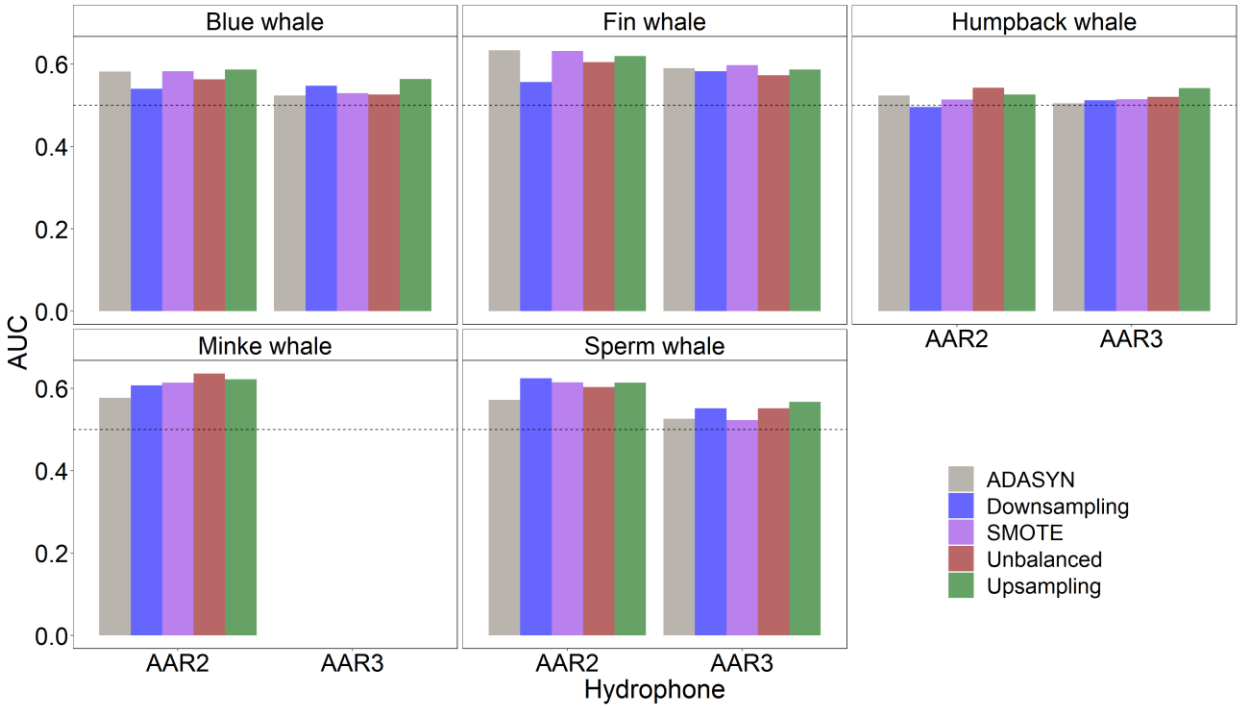

Fig. S12. Predictive generalized linear model performance, as measured by AUC, on the acoustic detection model for the six whale species around the four AAR stations. Horizontal dashed lines represent the 0.5 AUC threshold.

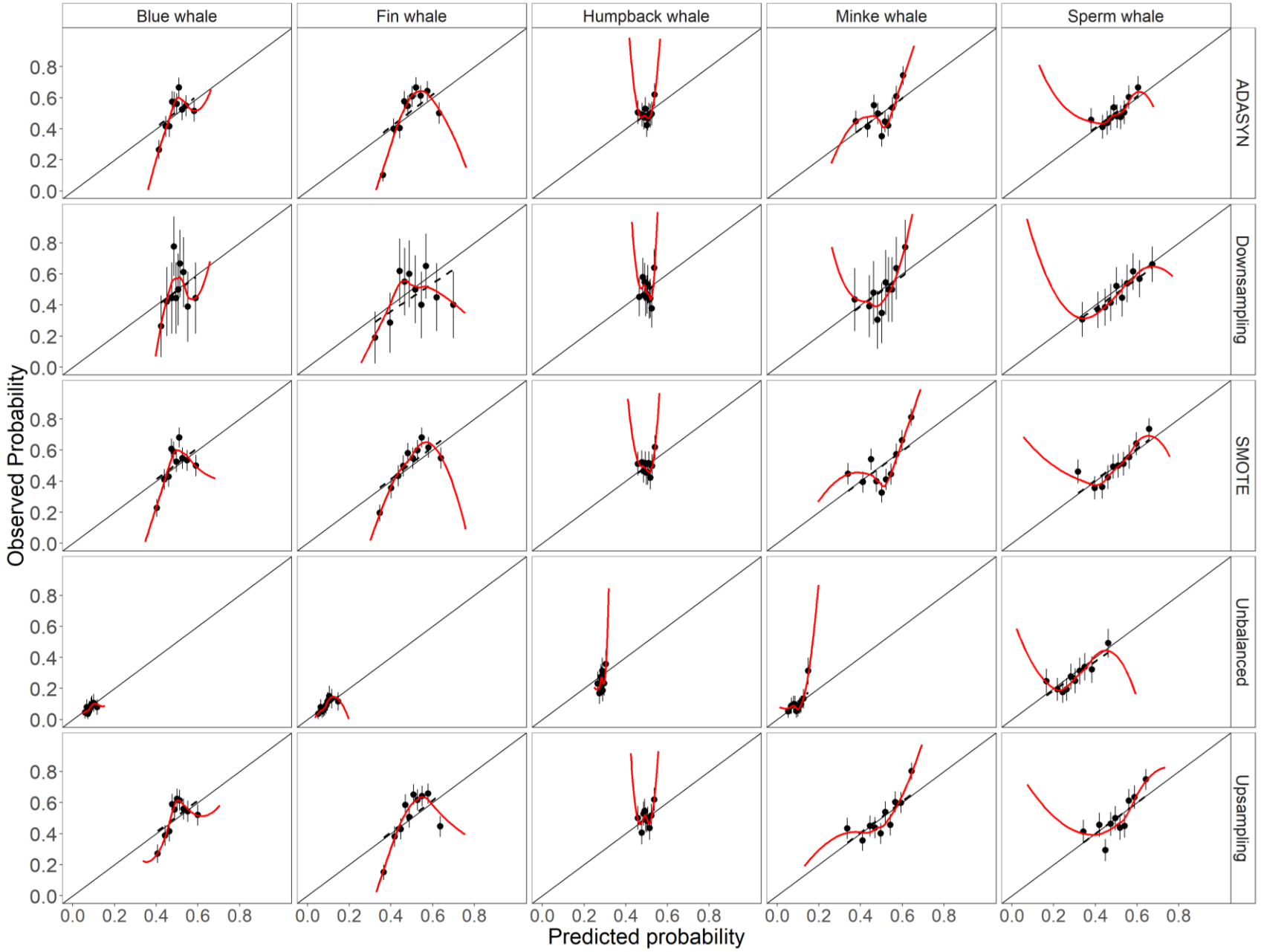

Fig. S13. Visual evaluation of generalized linear model predictive performance through calibration plots for whale species around AAR2 location.

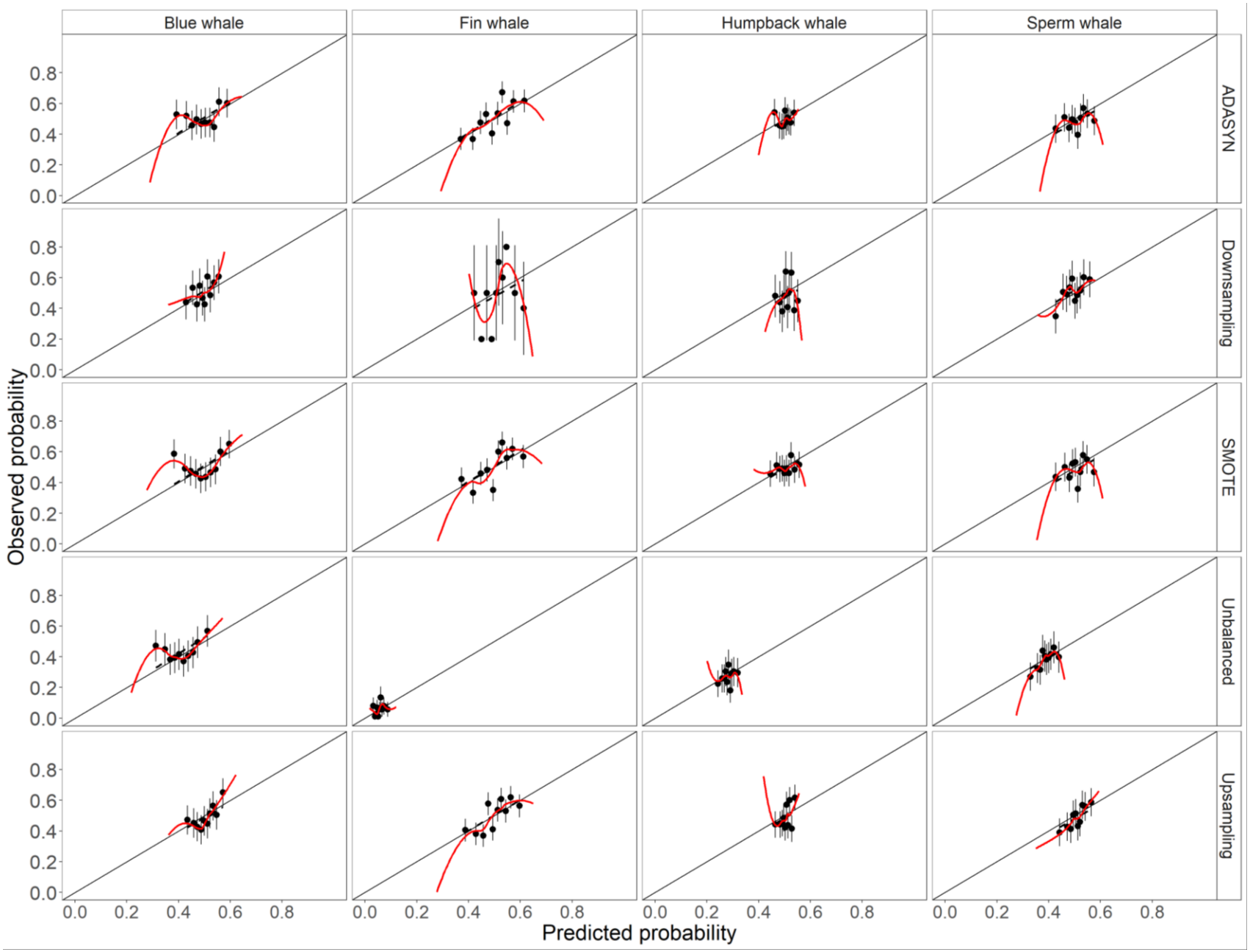

Fig. S14. Visual evaluation of generalized linear model predictive performance through calibration plots for whale species around AAR3 location.

Acoustic detection of whales

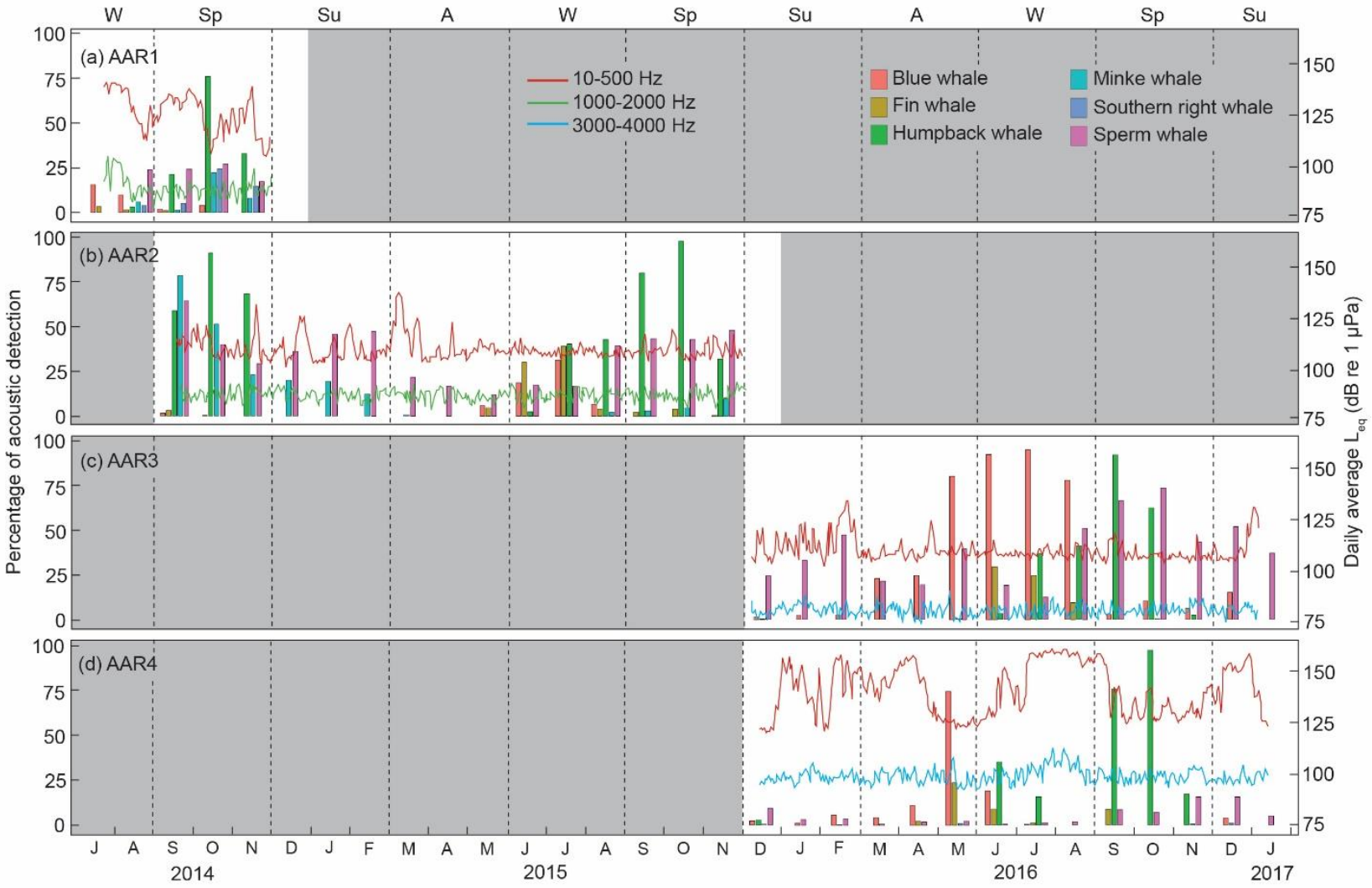

Fig. S15. Monthly percentage of whale acoustic detection (bar plots) relative to daily average recorded equivalent continuous sound pressure level (L_{eq}) at the lowest and highest frequency band (line plots) per season for all recording sites. Gray shaded areas represent periods without passive acoustic monitoring effort. W is winter, Sp is spring, Su is summer, and A is autumn.

GLM curves of ambient noise effect on whale detectability

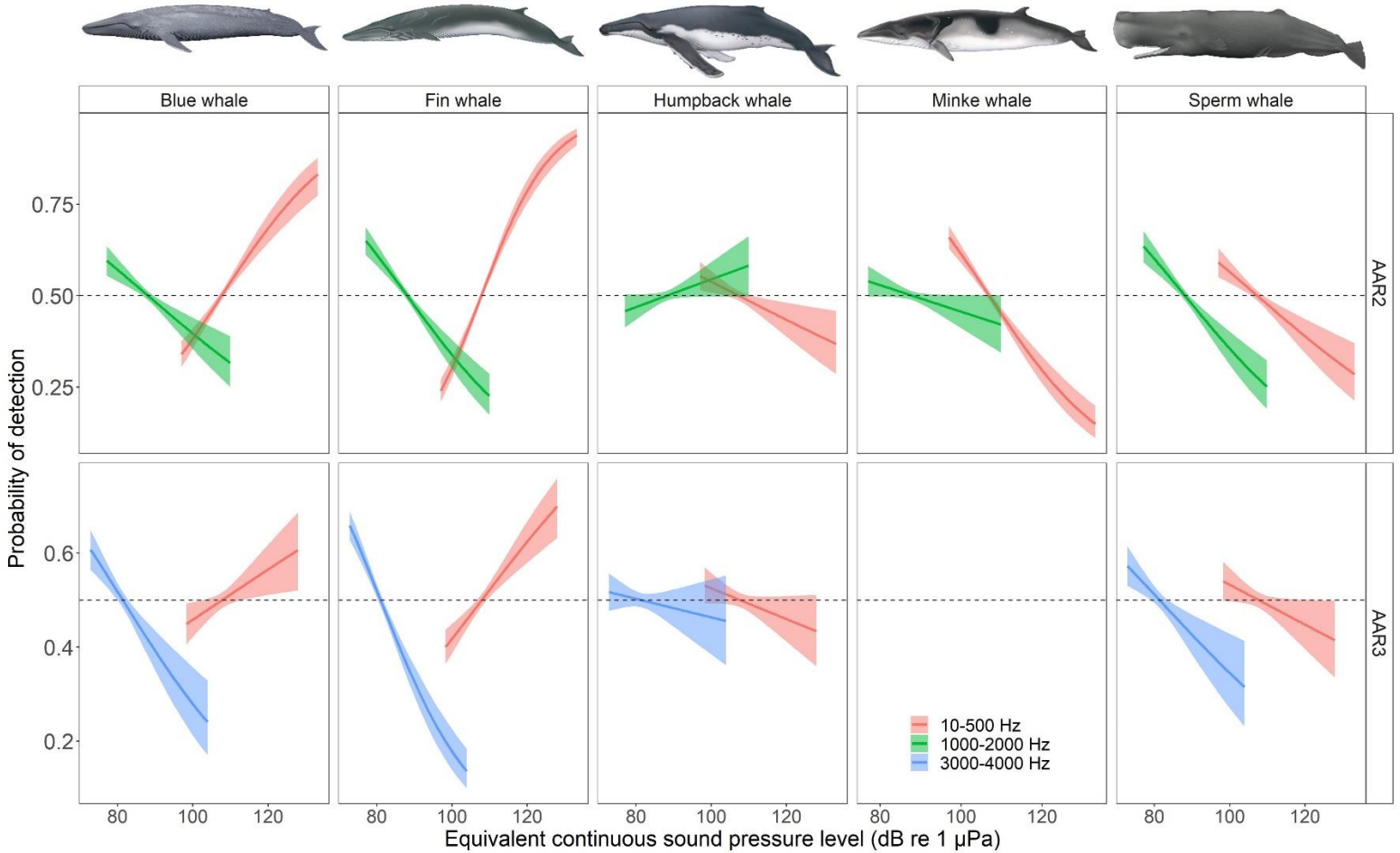

Fig. S16. Generalized linear model partial effect of the ambient equivalent continuous sound pressure level on probability of acoustic detecting whale species at the lowest and highest frequency bands for each autonomous acoustic recorder (AAR) stations based on ADASYN sample balancing method. Shadings around line plots represent 95% confidence intervals. Empty boxes represent cases where whale acoustic detection was <20 occurrences for that species. The dashed horizontal line at 0.5 represents the 50% point for whale acoustic detection at a given noise level. Y- and x-axes scales are different among plots.

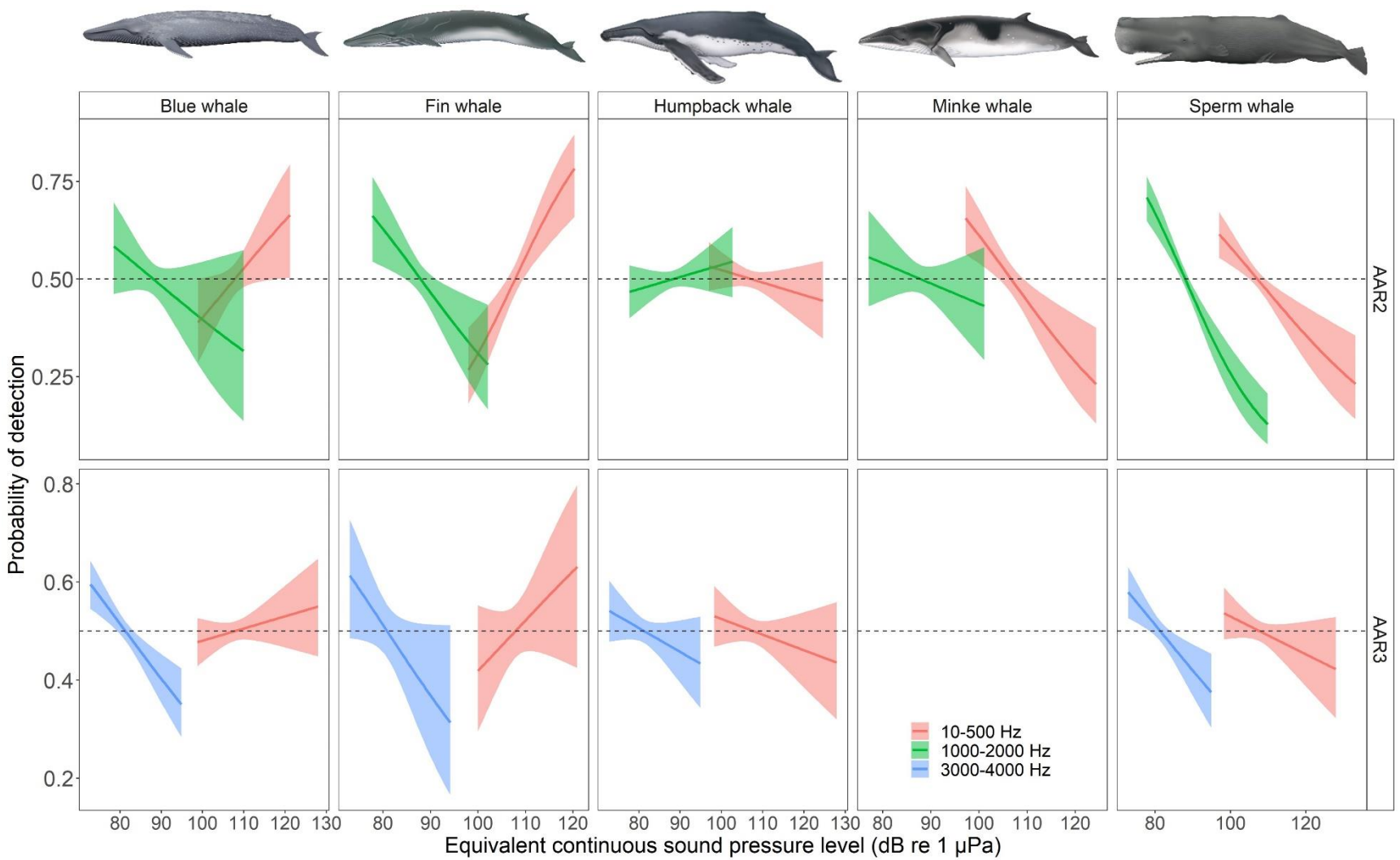

Fig. S17. Generalized linear model partial effect of the ambient equivalent continuous sound pressure level on probability of acoustic detecting whale species at the lowest and highest frequency bands for each autonomous acoustic recorder (AAR) stations based on downsampling method. Shadings around line plots represent 95% confidence intervals. Empty boxes represent cases where whale acoustic detection was <20 occurrences for that species. The dashed horizontal line at 0.5 represents the 50% point for whale acoustic detection at a given noise level. Y- and x-axes scales are different among plots.

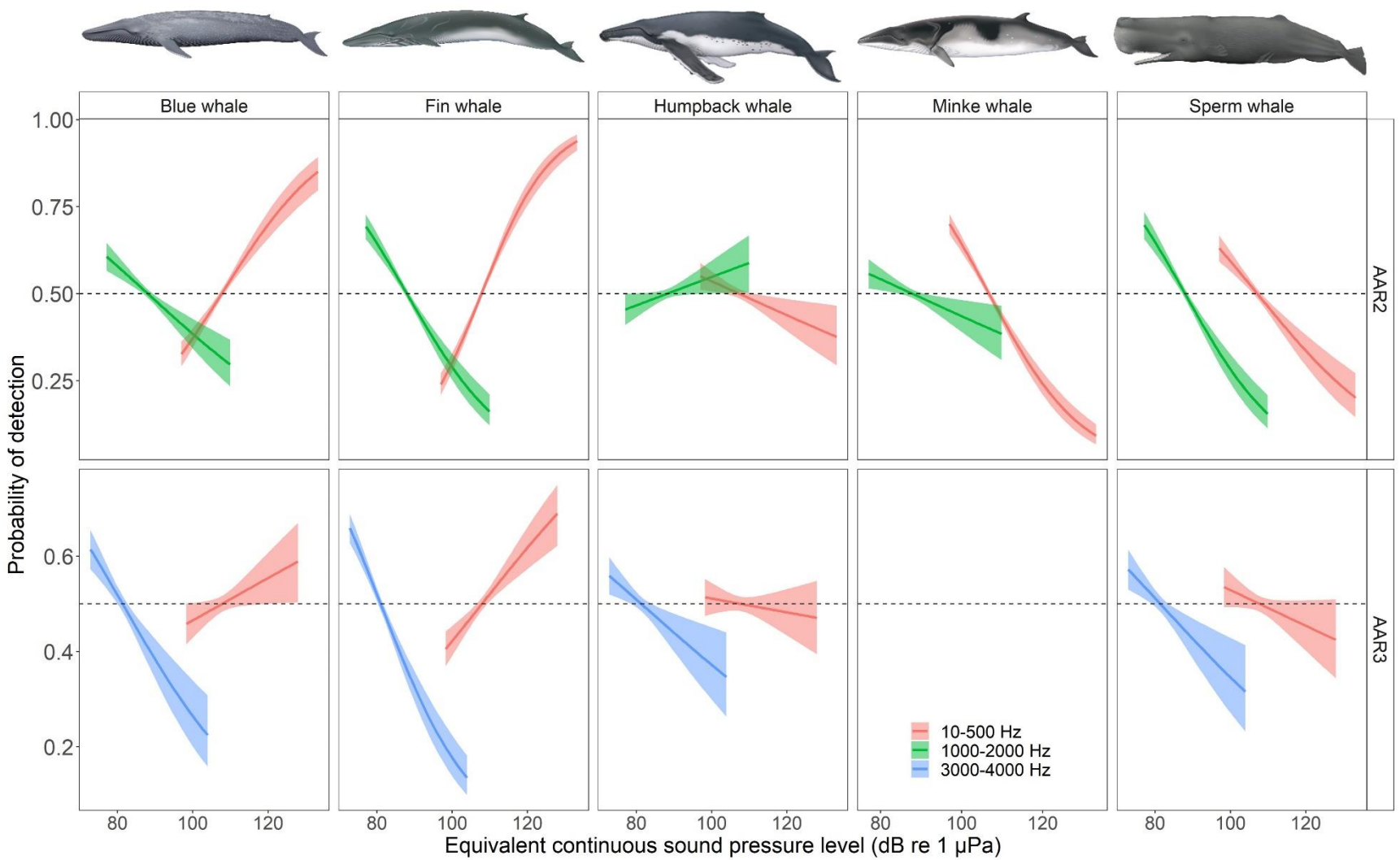

Fig. S18. Generalized linear model partial effect of the ambient equivalent continuous sound pressure level on probability of acoustic detecting whale species at the lowest and highest frequency bands for each autonomous acoustic recorder (AAR) stations based on SMOTE sample balancing method. Shadings around line plots represent 95% confidence intervals. Empty boxes represent cases where whale acoustic detection was <20 occurrences for that species. The dashed horizontal line at 0.5 represents the 50% point for whale acoustic detection at a given noise level. Y- and x-axes scales are different among plots.

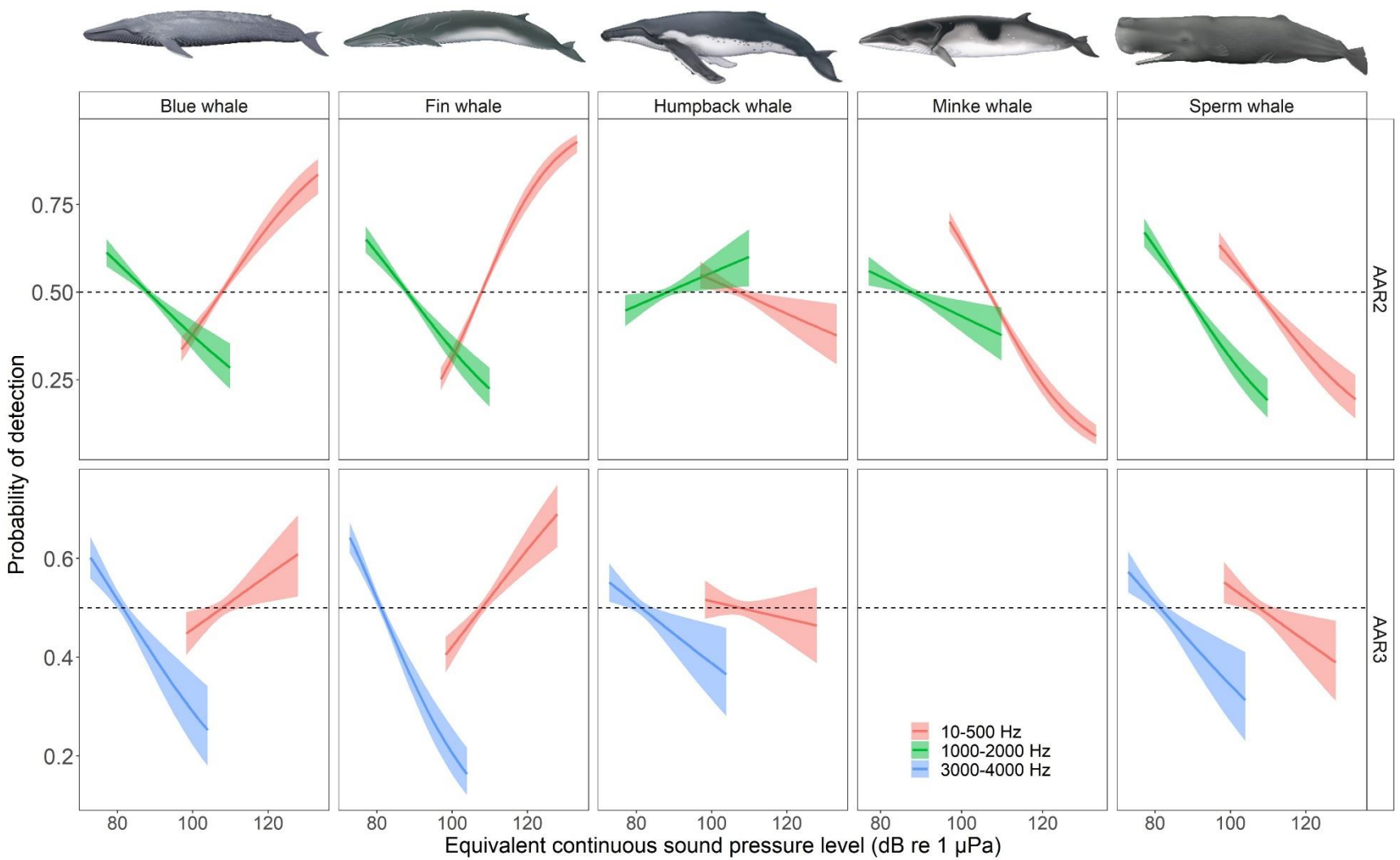

Fig. S19. Generalized linear model partial effect of the ambient equivalent continuous sound pressure level on probability of acoustic detecting whale species at the lowest and highest frequency bands for each autonomous acoustic recorder (AAR) stations based on upsampling method. Shadings around line plots represent 95% confidence intervals. Empty boxes represent cases where whale acoustic detection was <20 occurrences for that species. The dashed horizontal line at 0.5 represents the 50% point for whale acoustic detection at a given noise level. Y- and x-axes scales are different among plots.

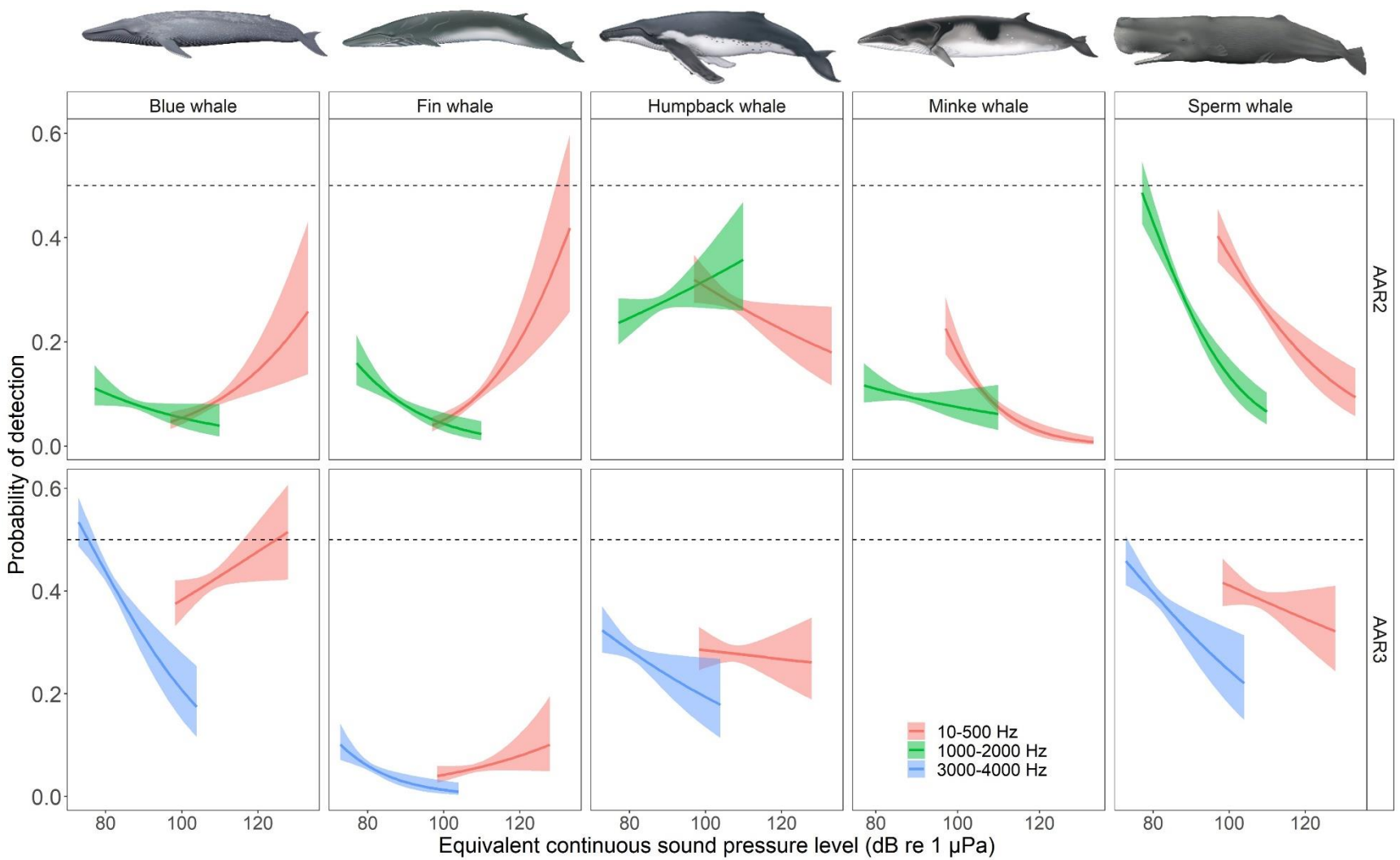

Fig. S20. Generalized linear model partial effect of the ambient equivalent continuous sound pressure level on probability of acoustic detecting whale species at the lowest and highest frequency bands for each autonomous acoustic recorder (AAR) stations based on unbalanced data. Shadings around line plots represent 95% confidence intervals. Empty boxes represent cases where whale acoustic detection was <20 occurrences for that species. The dashed horizontal line at 0.5 represents the 50% point for whale acoustic detection at a given noise level. Y- and x-axes scales are different among plots.

GLM results of recorded noise (including pseudo-noise) effect on whale detectability

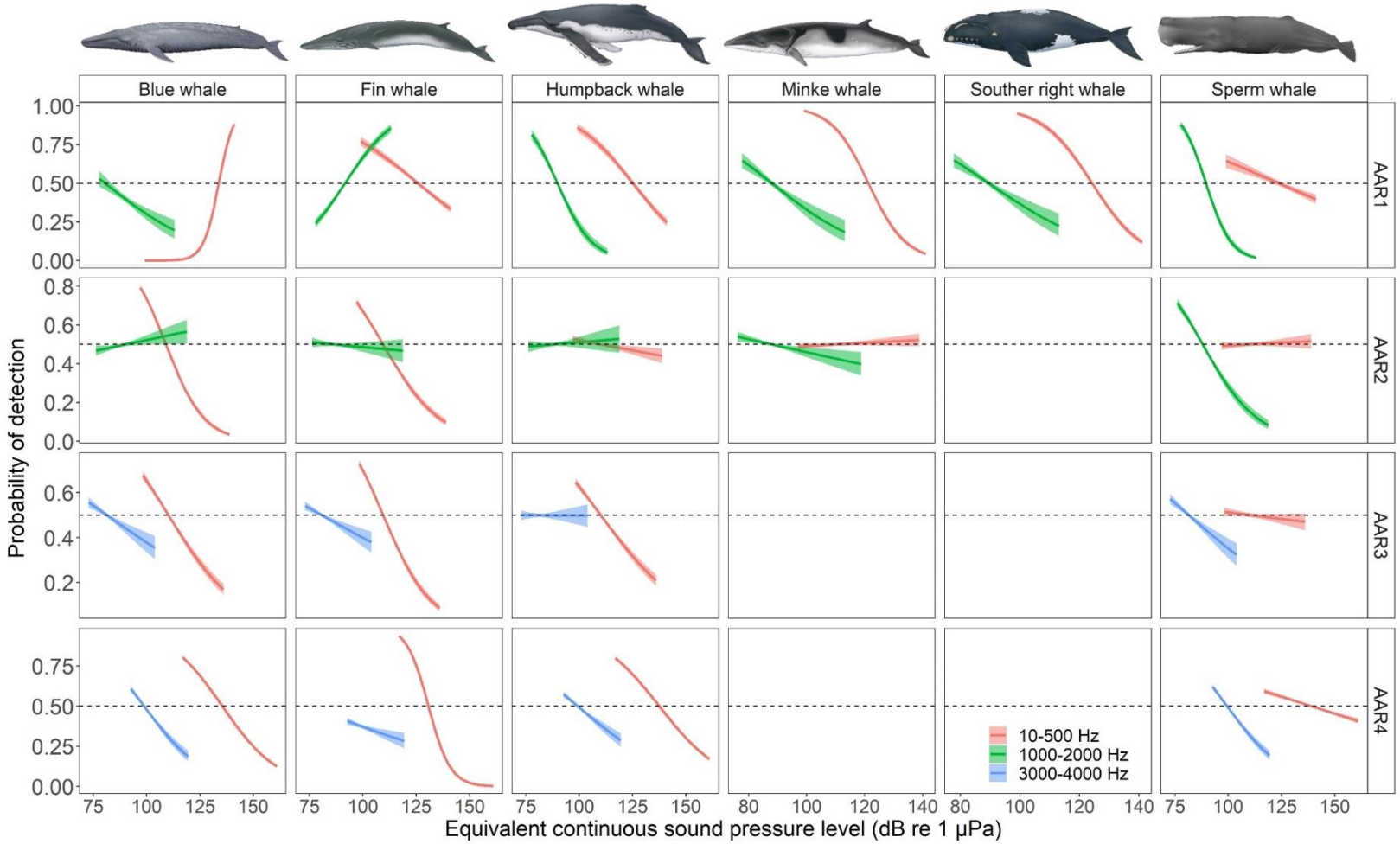

Fig. S21. Generalized linear model partial effect of the recorded equivalent continuous sound pressure level on probability of acoustic detecting whale species at the lowest and highest frequency bands for each autonomous acoustic recorder (AAR) stations based on ADASYN sample balancing method. Shadings around line plots represent 95% confidence intervals. Empty boxes represent cases where whale acoustic detection was <20 occurrences for that species. The dashed horizontal line at 0.5 represents the 50% point for whale acoustic detection at a given noise level. Y- and x-axes scales are different among plots.

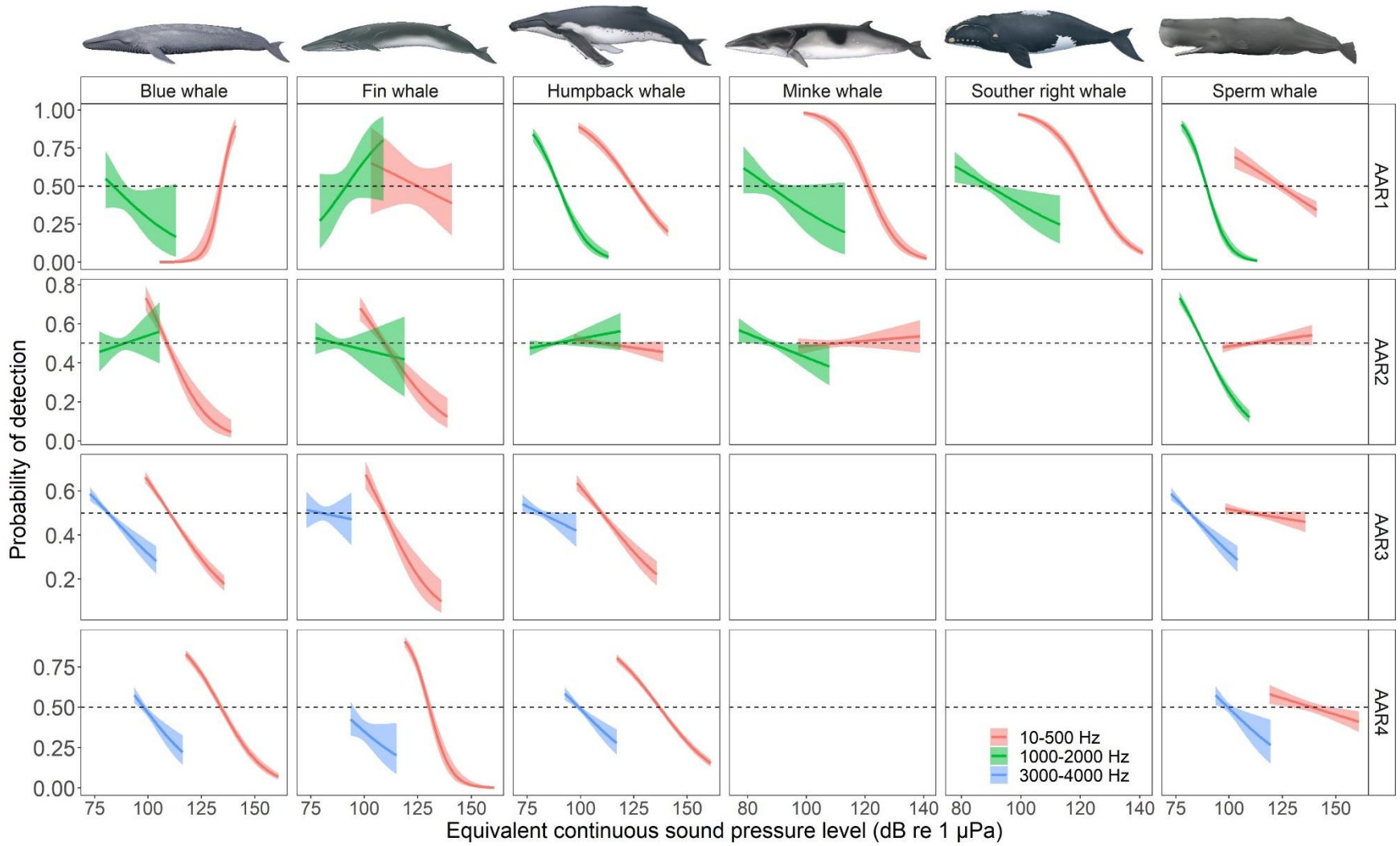

Fig. S22. Generalized linear model partial effect of the recorded equivalent continuous sound pressure level on probability of acoustic detecting whale species at the lowest and highest frequency bands for each autonomous acoustic recorder (AAR) stations based on downsampling method. Shadings around line plots represent 95% confidence intervals. Empty boxes represent cases where whale acoustic detection was <20 occurrences for that species. The dashed horizontal line at 0.5 represents the 50% point for whale acoustic detection at a given noise level. Y- and x-axes scales are different among plots.

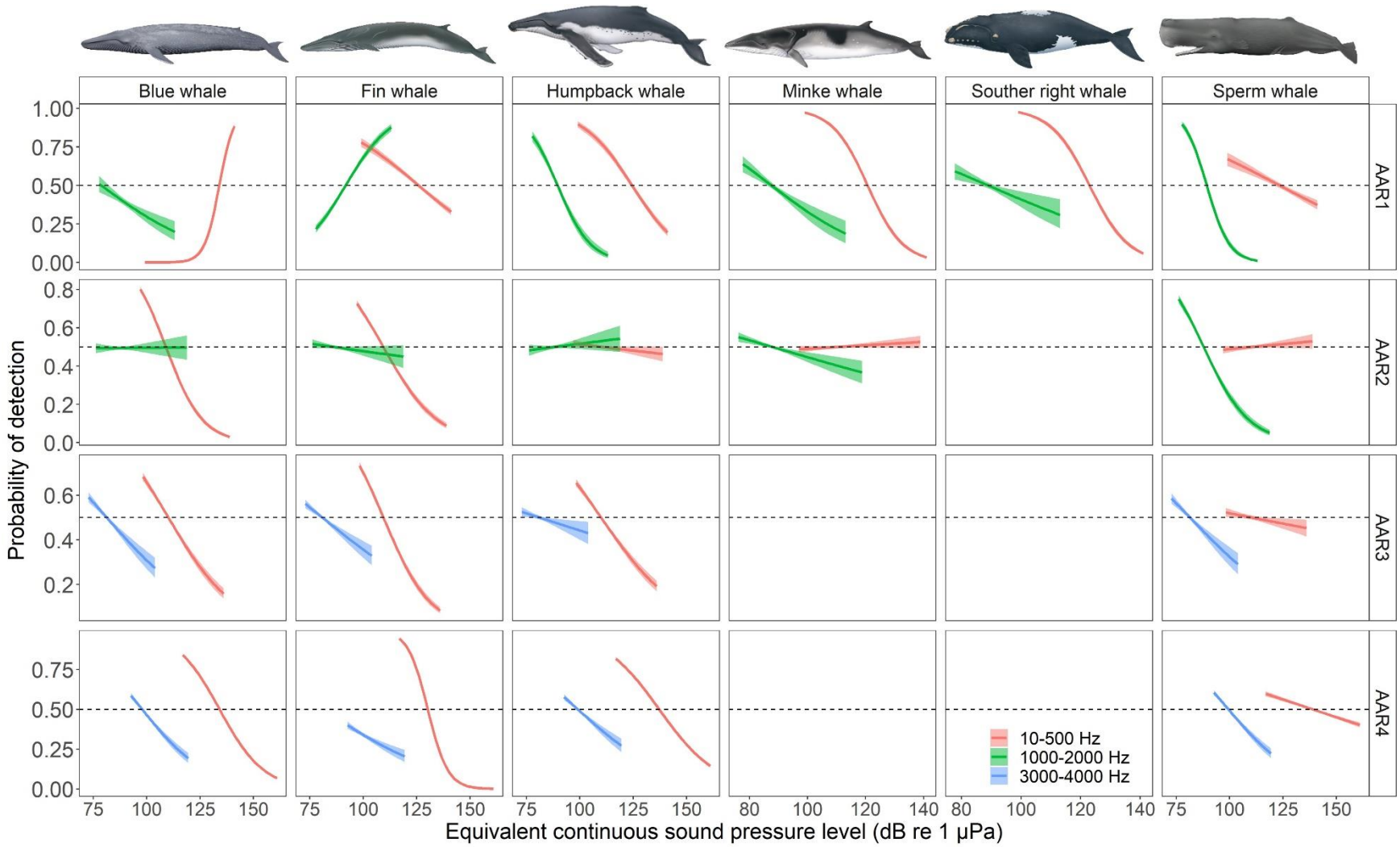

Fig. S23. Generalized linear model partial effect of the recorded equivalent continuous sound pressure level on probability of acoustic detecting whale species at the lowest and highest frequency bands for each autonomous acoustic recorder (AAR) stations based on SMOTE sample balancing method. Shadings around line plots represent 95% confidence intervals. Empty boxes represent cases where whale acoustic detection was <20 occurrences for that species. The dashed horizontal line at 0.5 represents the 50% point for whale acoustic detection at a given noise level. Y- and x-axes scales are different among plots.

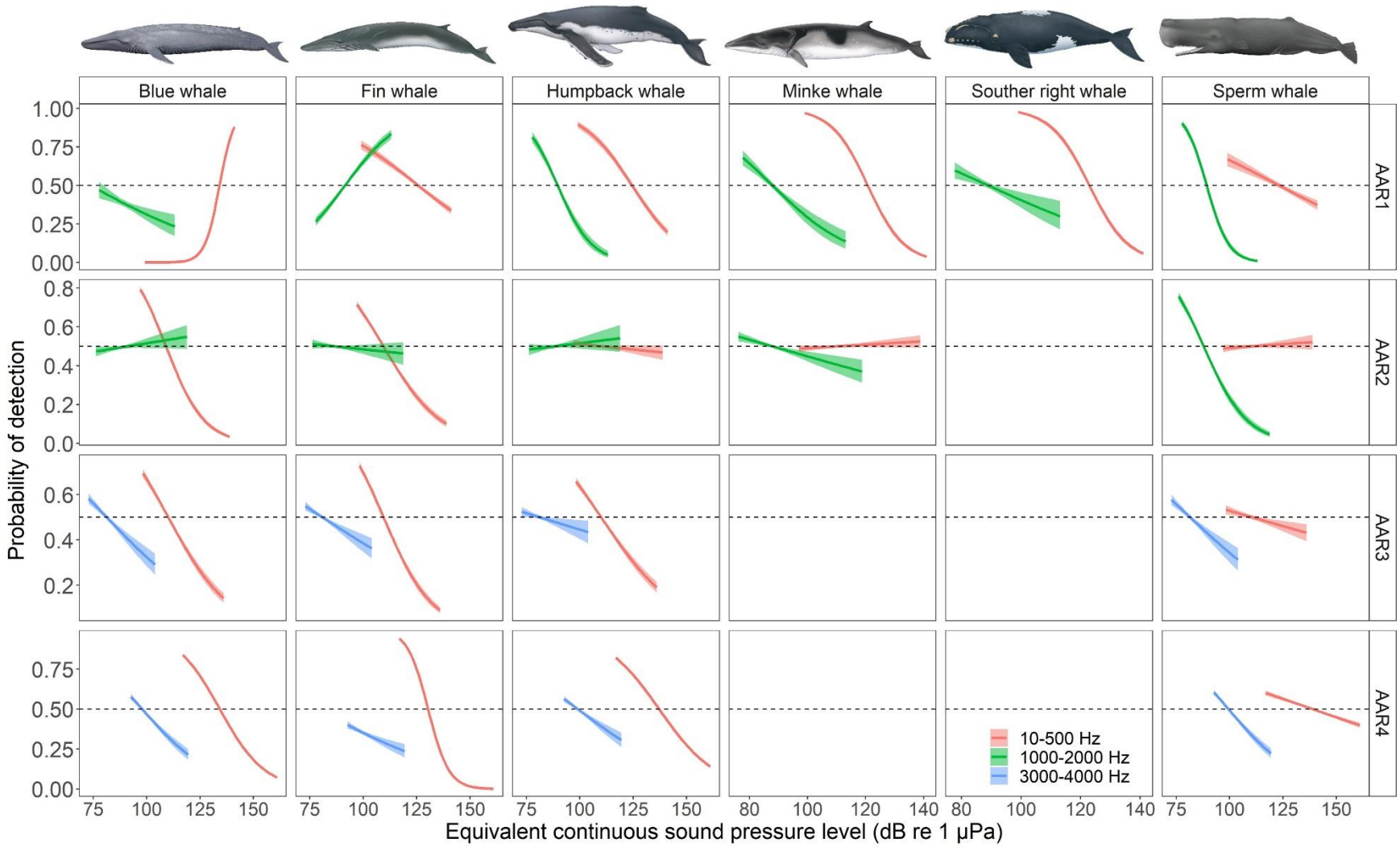

Fig. S24. Generalized linear model partial effect of the recorded equivalent continuous sound pressure level on probability of acoustic detecting whale species at the lowest and highest frequency bands for each autonomous acoustic recorder (AAR) stations based on upsampling method. Shadings around line plots represent 95% confidence intervals. Empty boxes represent cases where whale acoustic detection was <20 occurrences for that species. The dashed horizontal line at 0.5 represents the 50% point for whale acoustic detection at a given noise level. Y- and x-axes scales are different among plots.

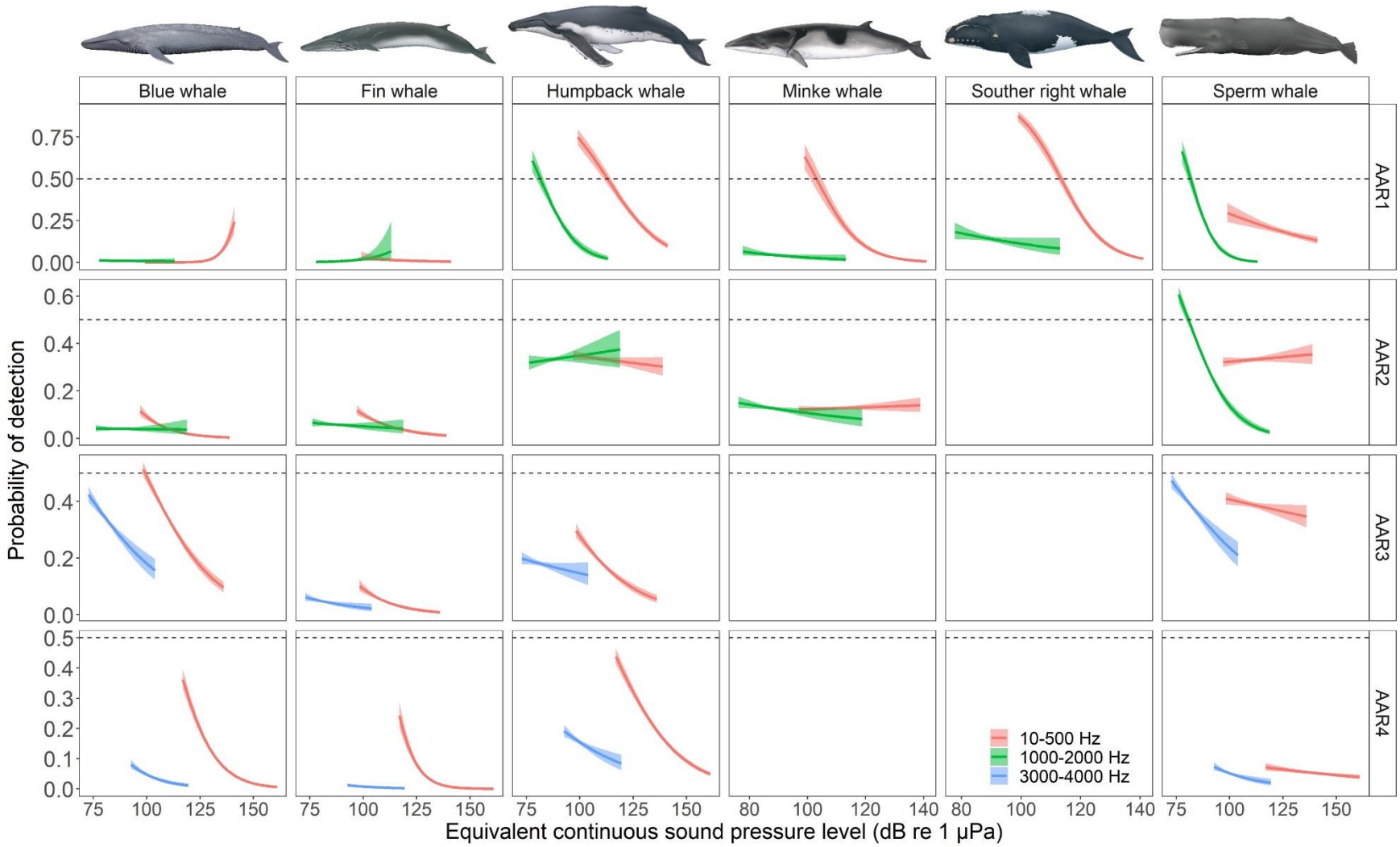

Fig. S25. Generalized linear model partial effect of the recorded equivalent continuous sound pressure level on probability of acoustic detecting whale species at the lowest and highest frequency bands for each autonomous acoustic recorder (AAR) stations based on unbalanced data. Shadings around line plots represent 95% confidence intervals. Empty boxes represent cases where whale acoustic detection was <20 occurrences for that species. The dashed horizontal line at 0.5 represents the 50% point for whale acoustic detection at a given noise level. Y- and x-axes scales are different among plots.

Seismic survey map

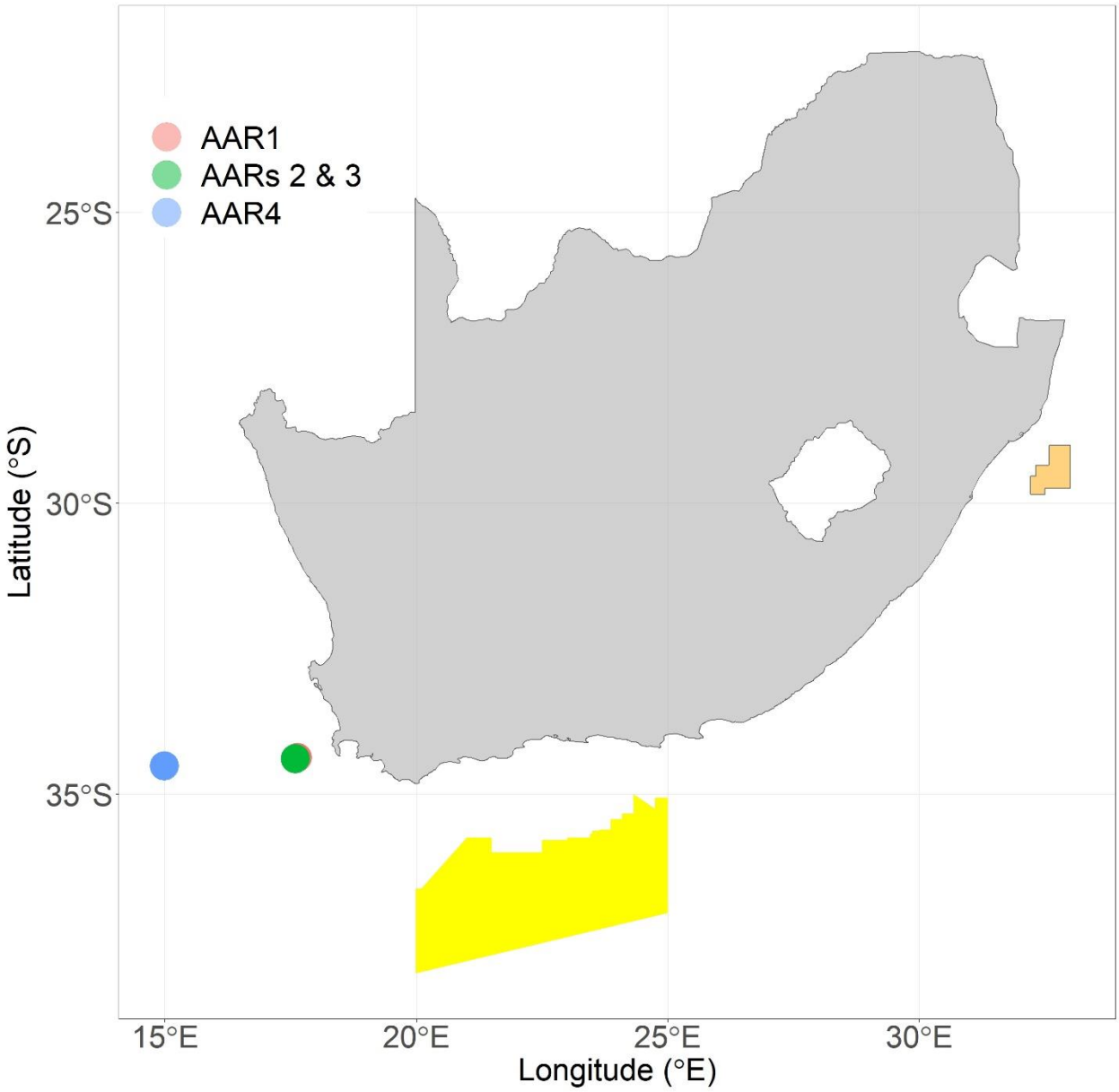

Fig. S26. Locations of 2D (yellow polygon) and 3D (orange polygon) seismic surveys conducted between July 2014 and January 2017 relative to the locations of autonomous acoustic recorders (AARs). 2D seismic survey was conducted in summer 2015/2016 and 3D seismic survey was conducted in summer through autumn 2016. Data on seismic survey locations was provided by Petroleum Agency South Africa (2021, Shape files and map gallery of seismic surveys in South Africa. Accessed on 19 February 2022. http://www.petroleumagencysa.com/index.php/data/viewing-our-technical-data).