# S1 | UNCERTAINTY ESTIMATION DETAILS

We used posterior simulation to propagate uncertainty through , , and . Each model was fitted via restricted maximum likelihood (REML), so the resulting estimates were empirical Bayes estimates. In this case we generated 5,000 samples from the (approximately multivariate normal) posterior of the model parameters. We generated a sample of the model parameters, , where is the estimate of the model coefficients and is the posterior covariance matrix. Here the for each model included the coefficients for the smooth terms in the model and fixed effects (e.g., intercept) if present. We then used the matrix that maps the model parameters to the predictions on the linear predictor scale (; Wood et al. 2017; section 7.2.6), along with the inverse link function, to generate predictions for each posterior sample. Denoting the vector of predictions , we calculate as follows:

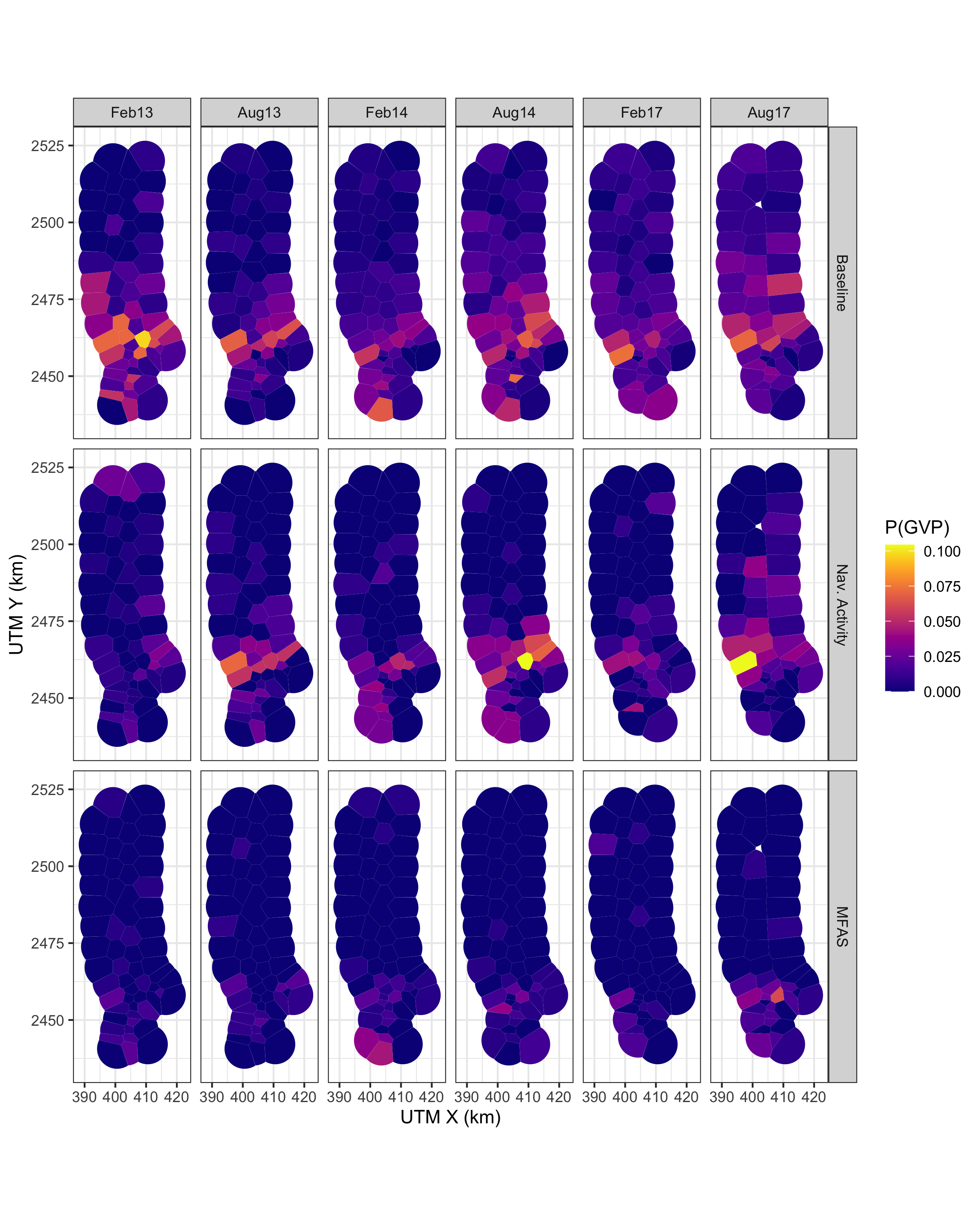
where was the link function, was the linear predictor and was any offset used by this prediction. Variance estimates can be obtained by taking the empirical variance of the resulting predictions (Wood et al. 2017; section 7.2.6). The prediction grid contained all possible combinations of covariates within the realized covariate space; i.e., each hydrophone for each SCC with associated location, hydrophone depth, and area of the tessellation tile, presence/absence of naval activity, and, if naval activity was present, then either sonar absence or sonar received level between 35 and 190 dB in intervals of 5 dB. This procedure was repeated for each model, with refitting to updated offsets from the previous model.

An algorithm for calculating the variance from our multi-stage approach is as follows. First define as the number of samples to take (= 5,000 here), let for be the matrix that maps coefficients to the predictions for model . For times:

1. Draw a sample from the posterior of M1:
2. Calculate a new offset for M2,
3. Refit M2 with as the offset to obtain M2′
4. Draw a sample from the posterior of M2′:
5. Calculate a new offset for M3, (predictions for the sonar data locations for M2′ when no sonar was present)
6. Refit M3 with offset to obtain M3′
7. Predict , , and over prediction grid and store them

We then calculated summary statistics (means and variances) of the values of , , and we generated. The empirical variance of the values of gave the uncertainty, incorporating components from all three models. We took appropriate pointwise quantiles (e.g., 2.5 and 97.5 for a 95% interval) to form confidence bands for the functional relationships between sonar received level and estimated probability of detecting GVPs.

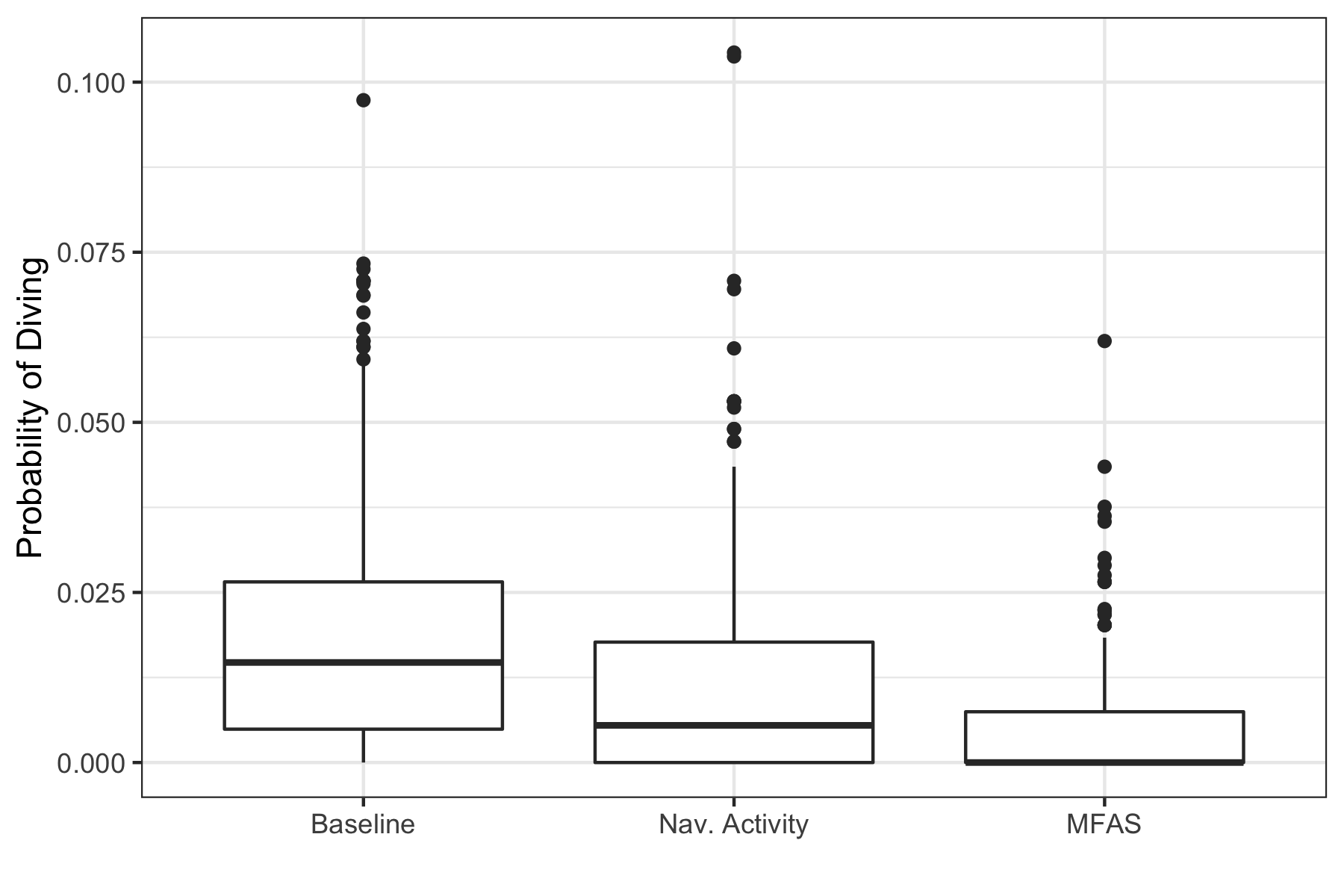
# S2 | SUPPLEMENTARY FIGURES



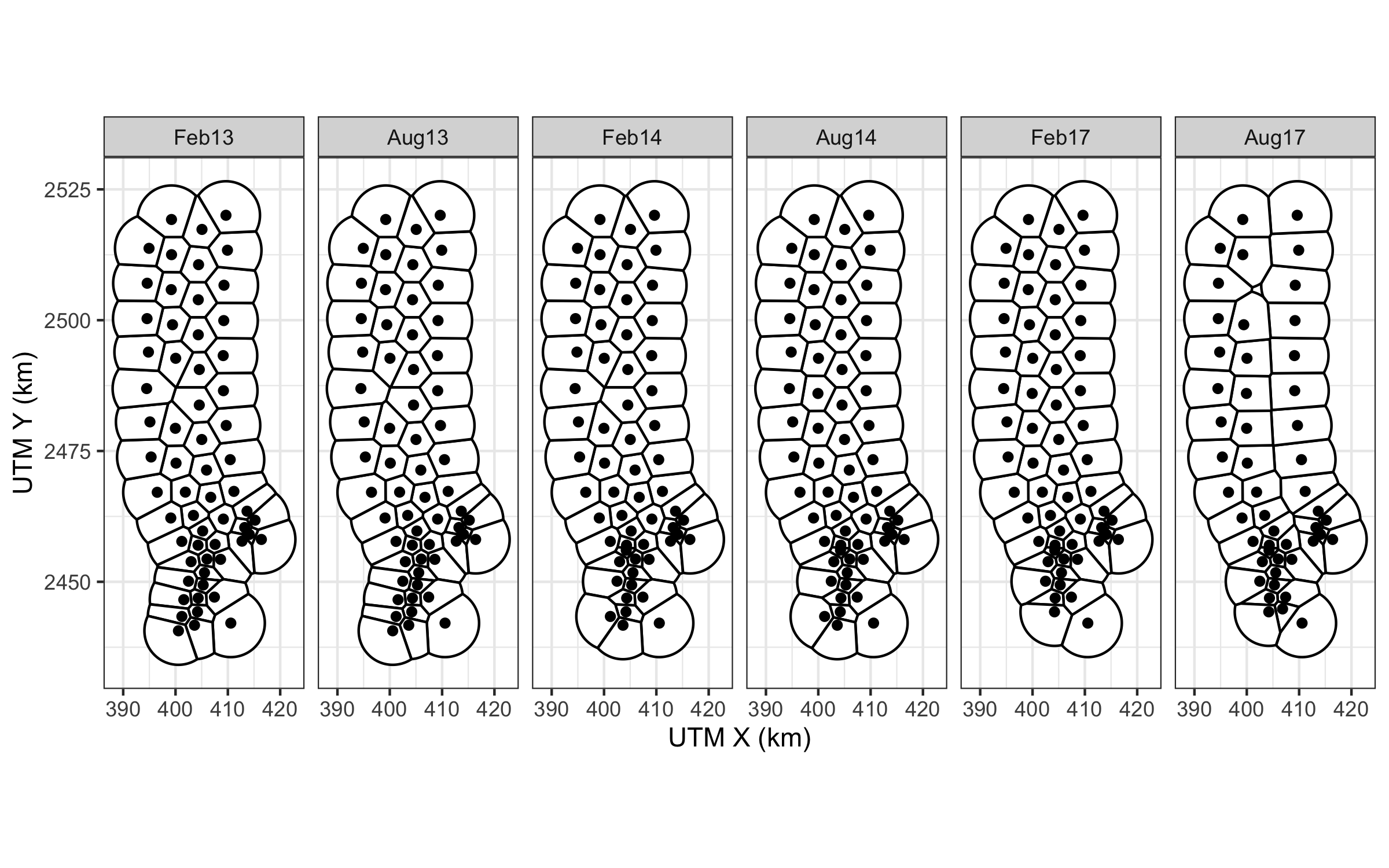
**FIGURE S2.1**: Map of observed probability of detecting a group vocal period (GVP) at each hydrophone (color scale) during the baseline period, when naval activity was present, and when mid-frequency active sonar (MFAS) was present (rows) for each submarine commander course (columns). Note that values of the probability of detecting a GVP are not corrected for effort (size of the hydrophone tile).

Median received level (dB re. 1 \muPa) when mid-frequency active sonar was present (color scale) for all hydrophones and submarine commander courses.

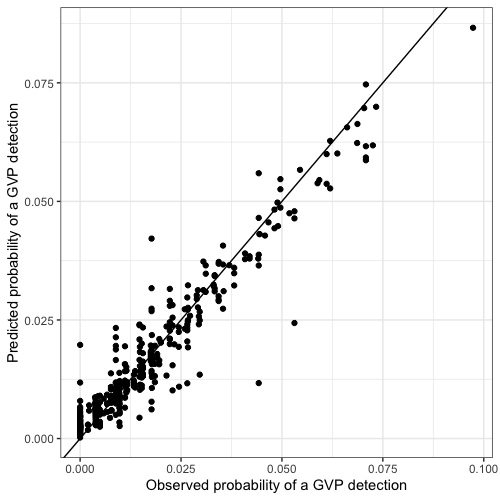
**FIGURE S2.2** Median received level (dB re. 1 Pa) when mid-frequency active sonar was present (color scale) for all hydrophones and submarine commander courses.



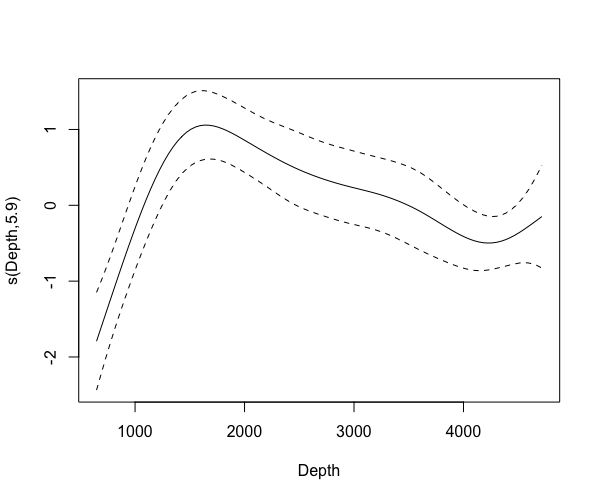
**FIGURE S2.3** Boxplot of observed probability of a group vocal period (GVP) for all hydrophones and submarine commander courses (SCCs; vertical axis) during baseline period, when naval activity was present, and when mid-frequency active sonar (MFAS) was present (horizontal axis). Each data point represents one hydrophone during one SCC and one phase of the training exercise.



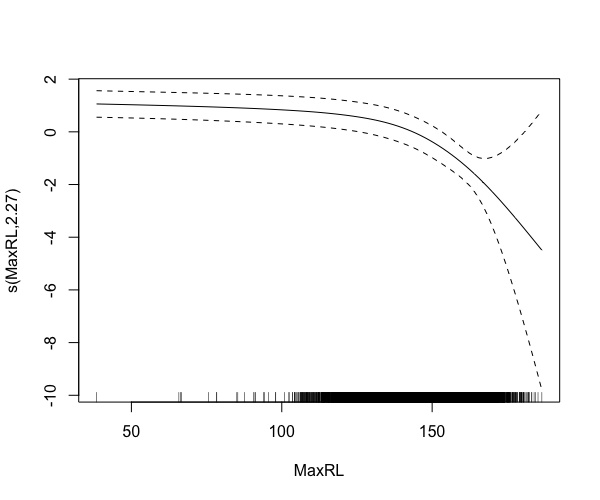
**FIGURE S2.4** Pacific Missile Range Facility range tessellations for each of six recorded submarine commander courses. Black lines indicate boundaries of hydrophone tiles. Black dots indicate approximate hydrophone locations.



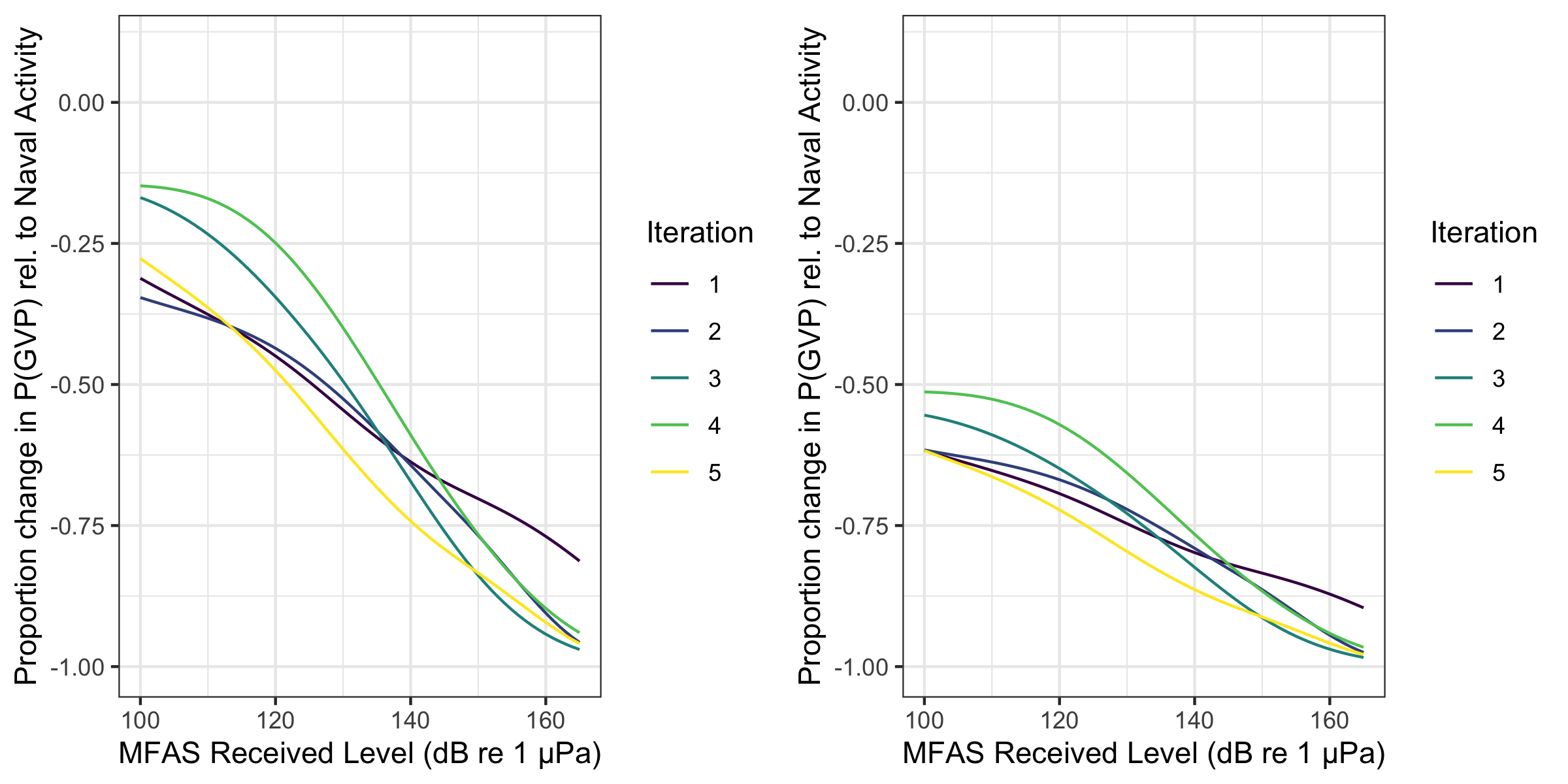
**FIGURE S2.5** Observed (horizontal axis) versus M1 predicted (vertical axis) probability of detecting a group vocal period (GVP) at each hydrophone during the baseline period.



**FIGURE S2.6** Spline for the relationship between the probability of detecting a group vocal period (GVP) and depth from M1 on the logit-link scale. Solid line: best fit; dashed lines: 95% CIs.



**FIGURE S2.7** Spline for the relationship between the probability of detecting a group vocal period (GVP) and maximum received level from M3 on the logit-link scale. Solid line: best fit; dashed lines: 95% CIs.



**FIGURE S2.8** Example of five iterations (colored lines) of the 5,000 posterior samples of the expected change in the probability of detecting a group vocal period (vertical axis) with increasing mid-frequency active sonar (MFAS) received level (horizontal axis) relative to when naval training activity but no MFAS was present on the range (left panel) and to when neither naval training activity nor MFAS were present on the range.

# S3 | SINGLE GENERALIZED ADDITIVE MODEL

A single generalized additive model (GAM) could be used to quantify the effect of naval sonar on Blainville’s beaked whales. Here, we present such a model and compare the results to the results obtained using the multi-stage model presented in the main text of the manuscript.

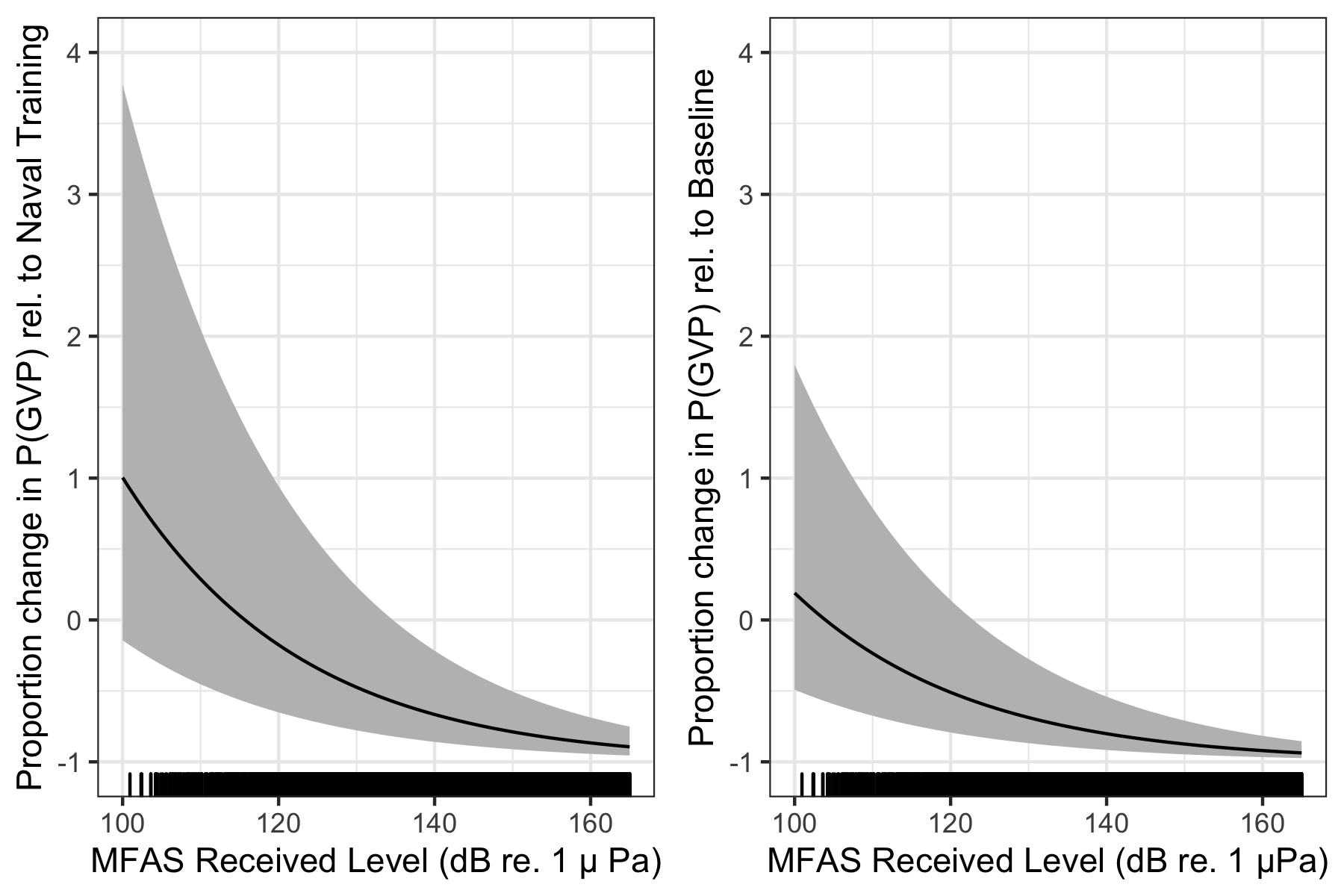
We modelled the probability of a group vocal period (GVP) at tile in submarine commander course (SCC) at time as a Bernoulli trial: . The linear predictor on the logit scale was given as:

where is an intercept, is the effect of naval training times an indicator variable for whether naval training was present or absent at time , denotes the Markov random field used to smooth space, is a smooth of depth (using a thin plate spline; Wood et al. 2003), is a smooth of sonar received level (using a thin plate spline) times an indicator variable for whether sonar was present or absent at time , and is an offset for the area (in ) of each tile, .

We fit the model to the same data used in M1, M2, and M3 (see Methods section of main manuscript for details) using mgcv (Wood, 2017).

This single GAM (Fig. S3.1) predicts a 41% (95% CI 34%-46%) decrease in when naval training is present compared to the baseline period, whereas the multi-stage GAM (Fig. 4) predicts a decrease of 44% (95% CI 38%-49%). The single GAM predicts that at a mid-frequency active sonar (MFAS) received level of 150 dB re 1 Pa, decreases by 87% (95% CI 71%-95%) relative to when only naval training is present, whereas the multi-stage model predicts the same decrease of 87% with a narrower credible interval (95% CI 81%-92%). Relative to when only naval training is present, the single GAM predicts a 50% reduction in at a MFAS received level of 120 dB re 1 Pa, whereas the multi-stage model predicts a 50% reduction at a MFAS received level of 132 dB re 1 Pa.

The major difference between this single GAM and the multi-stage model presented in the main text of the manuscript is that here, the spatial smooth is constructed using data from the baseline, naval training, and MFAS periods of each SCC. Therefore, the spatial distribution of MFAS may influence the predicted distribution of Blainville’s beaked whales. Using a single GAM leads to similar point estimates of the impact of sonar with greater uncertainty than the multi-stage model.



**FIGURE S3.1** Results from a single generalized additive model: Median (black line) and 95% CIs (gray shading) expected change in the probability of detecting a group vocal period (vertical axis) with increasing mid-frequency active sonar (MFAS) received level (horizontal axis) relative to when naval training activity but no MFAS was present on the range (left panel) and to when neither naval training activity nor MFAS were present on the range.