The following supplement accompanies the article

Riding the tide: use of a moving tidal-stream habitat by harbour porpoises

Steven Benjamins*, Andrew Dale, Nienke van Geel, Ben Wilson

*Corresponding author: steven.benjamins@sams.ac.uk

Marine Ecology Progress Series 549: 275–288 (2016)

SUPPLEMENTARY MATERIAL: GAM-GEE MODELLING OF MOORED AND DRIFTING C-POD DATA

Porpoise presence was modelled using binomial-based GAM-GEEs with an independent correlation structure and a logit link function to describe the relationship between covariates and porpoise click train detection presence (the response variable, described in a binary presence/absence format) based on the methods described by Pirotta et al. (2011). Models are only intended to describe available records and should not be extrapolated to other datasets. The independent correlation structure was used because of uncertainty in the actual underlying structure within the datasets, and because GEEs were considered robust against correlation structure misspecification (Liang and Zeger 1986; Pan 2001). The logit link function was chosen because it allowed the probability of porpoise detections to be modelled as a linear function of covariates, one of the core assumptions of GEEs (Zuur et al. 2009; Garson 2013). Moored C-POD data from different locations (Nearfield, Farfield, and Scarba) and drifter data (2011, 2012, and 2013) were all modelled separately to assess the relative significance of covariates for each deployment. Data exploration protocols described by Zuur et al. (2010) and Zuur (2012) were used to identify outliers, data variability, relationships between covariates and response variable, and collinearity between covariates. Modelling was initiated using a basic GLM as a means to assess collinearity of covariates, following Zuur (2012). Collinear and non-significant covariates were removed during subsequent analyses.

GAMs offer the ability to incorporate nonlinear responses to variables and therefore provide a more flexible and powerful tool than Generalised Linear Models (GLMs) to clarify the interactions between marine mammals and their environment (e.g. Hastie et al. 2005). GAMs assume independence

between model residuals, which is likely to be violated in the case of both stationary and drifting recorders where conditions at time *t* may closely resemble those at *t*-1 and *t*+1. This temporal autocorrelation could cause the uncertainty surrounding model estimates to be underestimated. To address this problem, autocorrelation in the data was investigated using the R autocorrelation function *acf* (Venables and Ripley 2002). These results were used to define blocks of data within which autocorrelation was present, using Generalised Estimation Equations (GEEs; Liang and Zeger 1986). Using this approach, uniform autocorrelation was expected within the blocks but not between them (Garson 2013). This is appropriate when studying population-level effects (in contrast to animal-specific response patterns, e.g. GAMMs; Fieberg et al. 2009, 2010) and particularly suitable for binomial distributions. GEEs are considered to be relatively robust even if block sizes are misspecified (Hardin & Hilbe 2003). Block sizes were individually estimated for each moored C-POD and drifter. For moored C-PODs, block sizes were 6, 9 and 13 tide-degrees for Nearfield, Farfield and Scarba, respectively. For drifters, block sizes varied between years, ranging from 16-42 tide-degrees in 2011, from 1-8 tide-degrees in 2012, and from 6-38 tide-degrees in 2013. The number of porpoise click train detections per tide-degree for each tidal cycle were reworked to binary records (1 = presence, 0 = absence; subsequently referred to as the response variable) to avoid overdispersion (McCullagh and Nelder 1989).

Covariates used for analysis of moored C-POD data included Tidal Cycle, Tidal Phase Angle, Date, Diel Hour and Period of Day (Table S1). Several additional covariates were included in the drifter models to incorporate movement effects: Latitude, Longitude, C-POD identifier code, average drift speed (m s^{-1}), water depth (m), distance from shore (m), and distance from the approximate centre of the Gulf of Corryvreckan (m; Table S1).

As the standard maximum limit for C-PODs of 4096 clicks per minute was relaxed to maximise coverage during noisy periods in these deployments, the standard % of Minute Lost parameter generated by POD.exe software on the basis of these data (which is often used to indicate levels of noise) could not be compared to other such datasets. It was therefore decided to use Average N of raw clicks received (which includes trains from harbour porpoises as well as from other sound sources such as sediment movement and boat sonars) as a similar proxy, providing a crude metric of ambient noise variability.

Tidal Cycle and Tidal Phase Angle were defined on the basis of Oban tidal data generated by POLTIPS-3™ tidal prediction software. Days since Deployment and Diel Hour were generated on the basis of deployment times and the 24-hour clock. Diel Hour was not used for analyses of drifter data as individual drifts were too short for any differentiation between Tidal Phase Angle and Diel Hour.

To generate Period of Day scores, the diel cycle was divided into morning, day, evening, and night. Morning and evening were defined on the basis of beginning and end of civil twilight, which was obtained for Oban from the U.S. Naval Observatory's Astronomical Applications Department (http://aa.usno.navy.mil/data/index.php). Following Carlström (2005), morning duration was defined as twice the time between the beginning of civil twilight and sunrise, while evening was similarly defined as twice the time between sunset and end of civil twilight. As the experiment took place at ~56°N, there was considerable variability in duration of day and night to which moored C-PODs were exposed; in August, average durations of days was ~14 hours vs. ~6.7 hours of nights, whilst mornings and evenings lasted ~1.5 hours on average. This covariate was not used when analysing drifter data due to relatively brief deployment durations.

For the Farfield mooring, Flow Speed (m s^{-1}, as measured by ADCP) was initially considered as a covariate, but ultimately rejected on the grounds that it was too closely linked to Tidal Phase Angle (in that current speeds are driven by tidal phase, with the Spring-Neap tidal cycle imposing additional variability in terms of strength, duration and timing). For drifters, a Drift Speed parameter was used to describe drifter movement in response to surface currents. For similar reasons, C-POD Angle was not used for moored C-PODs given the close causal link between tidal phase, flow speeds and C-POD deflection. C-POD Angle for drifting C-POD data could only be compared within deployment years due to different drifter designs. C-POD ID was only used for drifters to address cases of multiple drifters sampling the same areas. For drifter data, Latitude and Longitude were used to uniquely identify each drifter's position over time and thus account for survey effort. Because of this, these covariates were included in all models and retained irrespective of their significance (Pirotta et al. 2011).

Table S1. Covariates considered in the analysis of moored and drifting C-POD data. All covariates were averaged to tide-degrees.

Covariate	Unit	Scale	Description	Treatment in model	Moored PODs	Drifters
Tidal Cycle	Number	1 - 71	Identifier of consecutive ebb-ebb tidal cycles	Cubic B-spline	Used	Not used
Tidal Phase Angle	Degree (°)	0 - 360° (cyclic)	Circular scale describing each tidal cycle, where 0° = 360° = Low Tide at Oban	Cyclic spline	Used	Used
Days since Deployment	Number	1 – 37 (moored); 1 - 3 (Drifters)	Number of days since deployment, where 1 = day of deployment	Factor	Used	Used
Diel Hour	Hour	0 – 23 (cyclic)	Hour of day	Cyclic spline	Used	Not used
Period of Day	Number	1 – 4	1 = Morning, 2 = Day, 3 = Evening and 4 = Night	Factor	Used	Used
Drift Speed	m s^{-1}	0 – 5 m s^{-1}	Average drifter speed as determined by successive GPS positions	Cubic B-spline	Not used	Used
C-POD ID	Number	Varies	Individual identifying number	Factor	Not used	Used
C-POD Angle	Degree (°)	0 - 180°	Average deflection from vertical, where 0° = C-POD pointing straight up	Cubic B-spline	Not used	Used
Latitude	Degree (°)	56 - 57° N	Component of DPD coordinates	Cubic B-spline	Not used	Used
Longitude	Degree (°)	5 - 7° W	Component of DPD coordinates	Cubic B-spline	Not used	Used
Depth	Metres	0 – 275 m	Derived from INISHydro bathymetry data	Cubic B-spline	Not used	Used
Distance from shore	Metres	0 – 12 km	Calculated from SeaZone coastline data	Cubic B-spline	Not used	Used
Distance from central Gulf of Corryvreckan	Metres	0 – 12 km	Calculated from SeaZone coastline data	Cubic B-spline	Not used	Used
Average N of raw clicks received	Number	0 - 65536	Number of raw clicks received each minute	Linear/Cubic B-spline	Used	Used

For both moored and DPD data, all covariates were considered as either 1) linear terms, 2) factors, or 3) 1-dimensional smooth terms with 4 degrees of freedom. The latter were modelled as either cubic B-splines with one internal knot positioned at the average value of each variable, or as cyclic penalized cubic regression splines (specifically those covariates identified as 'cyclic' in Table S1).

The Quasi-likelihood under Independence model Criterion (QICu; Pan 2001), a modification of Akaike's Information Criterion (Akaike 1974) appropriate for GEE models, was used to identify which covariates should be retained in the final model, using the R library *yags* (Carey 2004). Covariates were removed one at a time in a backwards stepwise model selection process, and models with the lowest QICu values were taken forward up to the point where removal of further covariates no longer resulted in lower QICu values. At this point, the final GAM model was fitted using the R function *geeglm* (contained within R package *geepack*; Halekoh et al. 2006) to assess the statistical significance of the remaining covariates within the correlation structure specified within the GEE. The Wald's Test (Hardin & Hilbe 2003) was used to determine each covariate's significance; non-significant covariates were removed from the model using backwards stepwise model selection.

Model performance was evaluated through confusion matrices which assessed how well the binary model predictions matched observed values (e.g. how often an observed detection was predicted by the model), thereby summarising the goodness of fit of the model (Fielding & Bell 1997; Pirotta et al. 2011). Plots were generated describing the probabilistic relationship between each contributing explanatory covariate and the model response variable (click train presence/absence). Confidence intervals around these plots were based on the standard errors of the GAM-GEE model.

MODELLING RESULTS – MOORED C-PODS

Data exploration following Zuur et al. (2010) confirmed temporal autocorrelation in the data, providing justification for the GAM-GEE approach. The covariates Date and Hour after Low Tide at Oban were excluded from the modelling process at all three sites due to collinearity. Two covariates (Tidal phase angle and Diel Hour) were modelled in all subsequent steps using cyclic splines based on

variance-covariance matrices. Step-wise model selection based on the QICu scores and Wald's tests resulted in the following final model structures (most important covariates first):

Nearfield: Tidal Phase Angle, Diel Hour, Tidal Cycle

Farfield: Tidal Phase Angle, Tidal Cycle, Diel Hour

Scarba: Tidal Cycle, Diel Hour, Tidal Phase Angle

Based on the Wald's test outcomes, Tidal Phase Angle, Diel Hour and Tidal Cycle were significant covariates for all three sites, but their relative significance varied between sites (Table S2).

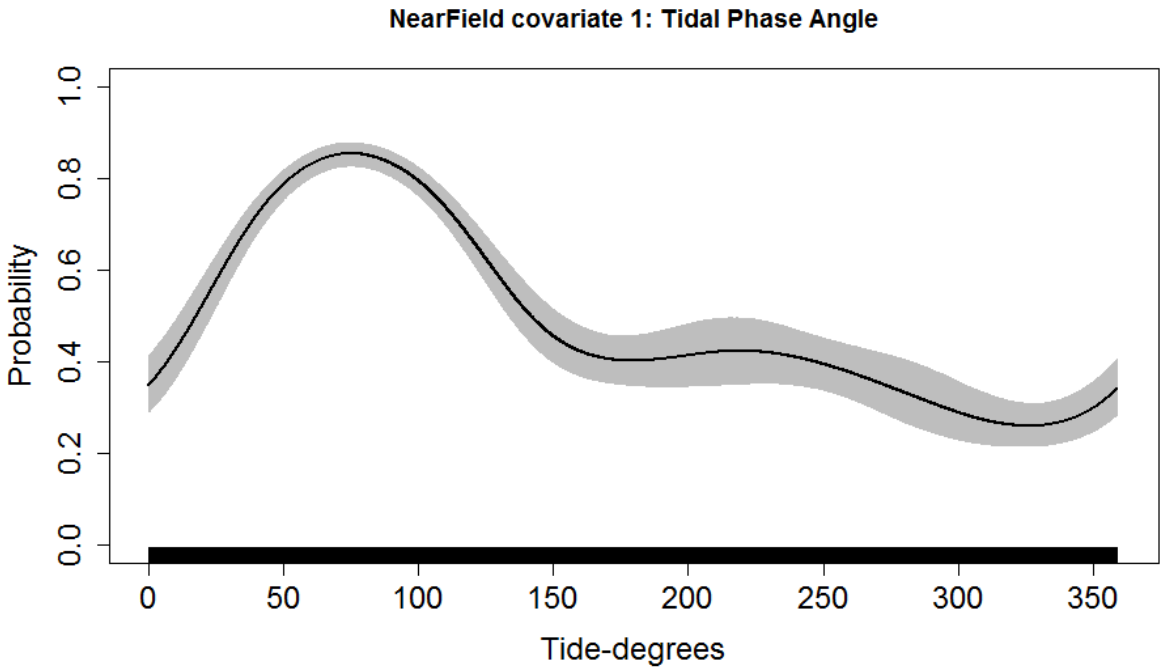
Table S2. Results of Wald's tests for all significant covariates for the final model for each site.

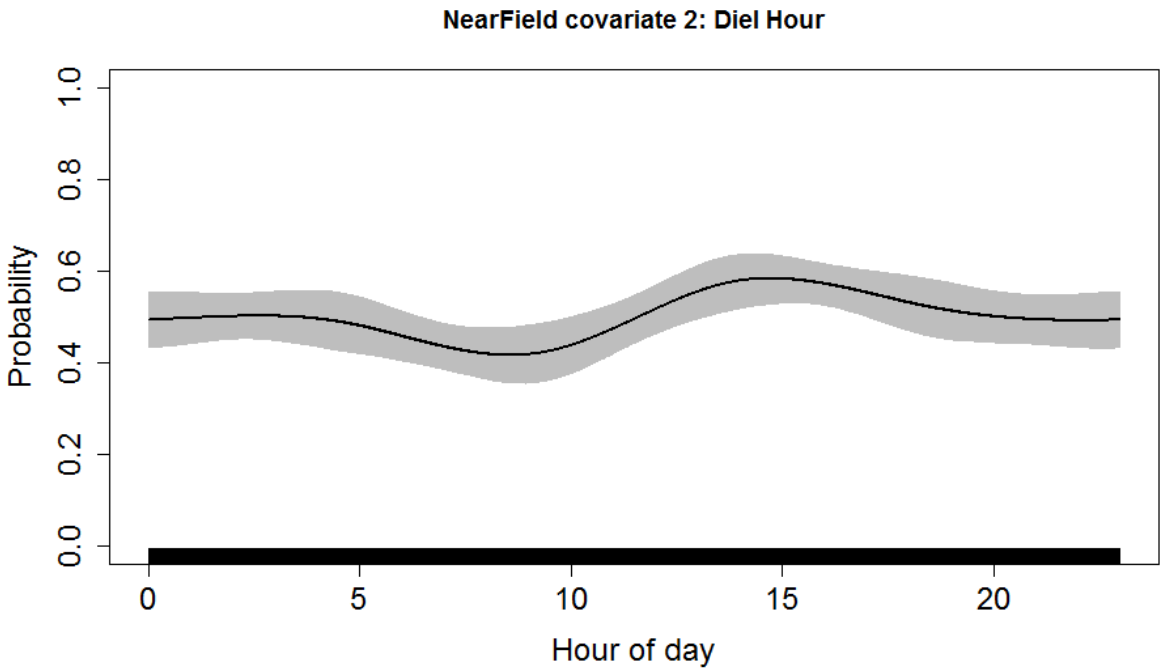
	Nearfield			**Farfield**			**Scarba**		
Covariate	**DF**	**χ^2 score**	**P-value**	**DF**	**χ^2 score**	**P-value**	**DF**	**χ^2 score**	**P-value**
Tidal cycle	4	12.94	0.0116	4	37.56	$1.4 \cdot 10^{-7}$	4	47.79	$1.0 \cdot 10^{-9}$
Tidal Phase Angle	4	318.76	$<2.2 \cdot 10^{-16}$	4	49.92	$3.7 \cdot 10^{-10}$	4	10.50	0.03282
Diel Hour	4	15.03	0.00464	4	10.42	0.0339	4	15.87	0.00320

Covariates were plotted independently to visualise the probabilistic relationship between each covariate and the binary response variable (porpoise detection) at each site (Figure S1-S3). Covariates were plotted in declining order of significance in terms of their explanatory power. Less significant covariates' relationships to the response variable were dependent upon the inclusion of more significant covariates in the model, and should therefore be interpreted as explaining residual amounts

of variation in the presence of more significant covariates, rather than seen in isolation. The significance of covariates varied between sites:

- At the Nearfield site, porpoise detection probability was very high early in the tidal cycle when the Great Race is known to be most active (Figure S1). At the Farfield site a comparable increase in detection probability occurred, but somewhat later in the tidal cycle. It was followed by a second, albeit smaller increase during early ebb tide, suggesting porpoises were also present during ebb tides (Figure S2). The Scarba site showed a bimodal pattern in response to Tidal Phase Angle similar to the Farfield site (Figure S3) although the covariate was less significant here.
- Tidal Cycle influenced detections at all three sites in a similar pattern, but the significance of that influence varied between sites (Figure S1-S3), with the covariate being most important at Scarba. Increasing detection probability over time at all sites suggested small-scale variability in porpoise distribution across the area during the deployment period.
- Diel Hour was never the most important covariate, but varied between sites suggesting short-term variability in porpoise detection probability. (Figure S1-S3).


NearField covariate 1: Tidal Phase Angle
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Tide-degrees


NearField covariate 2: Diel Hour
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Hour of day

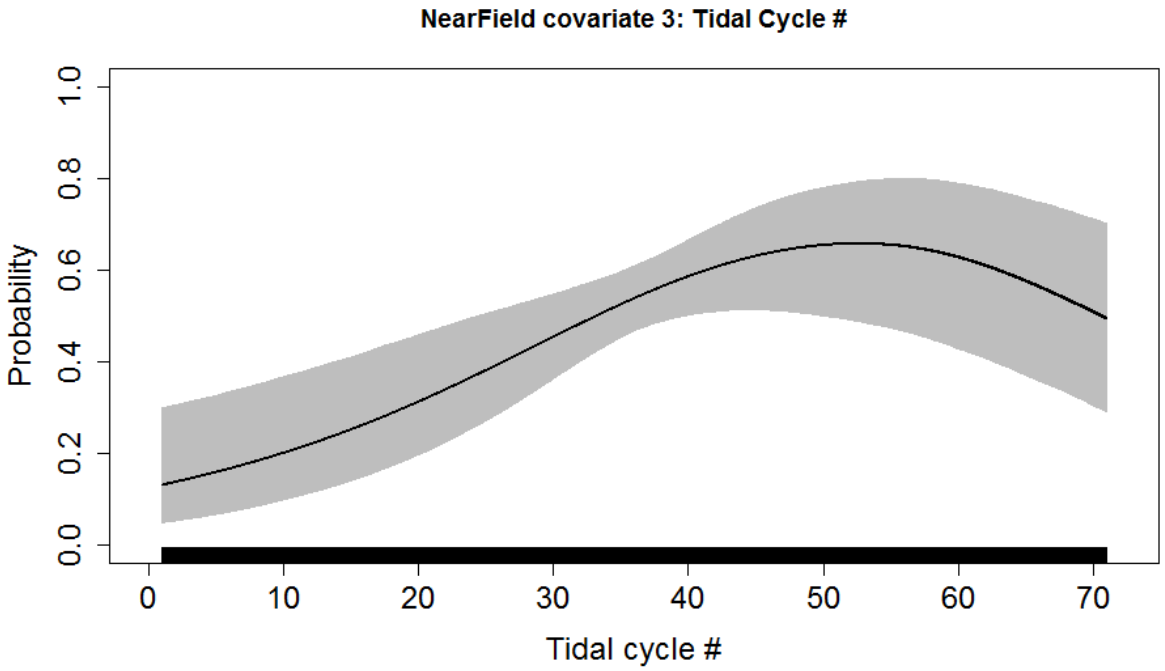
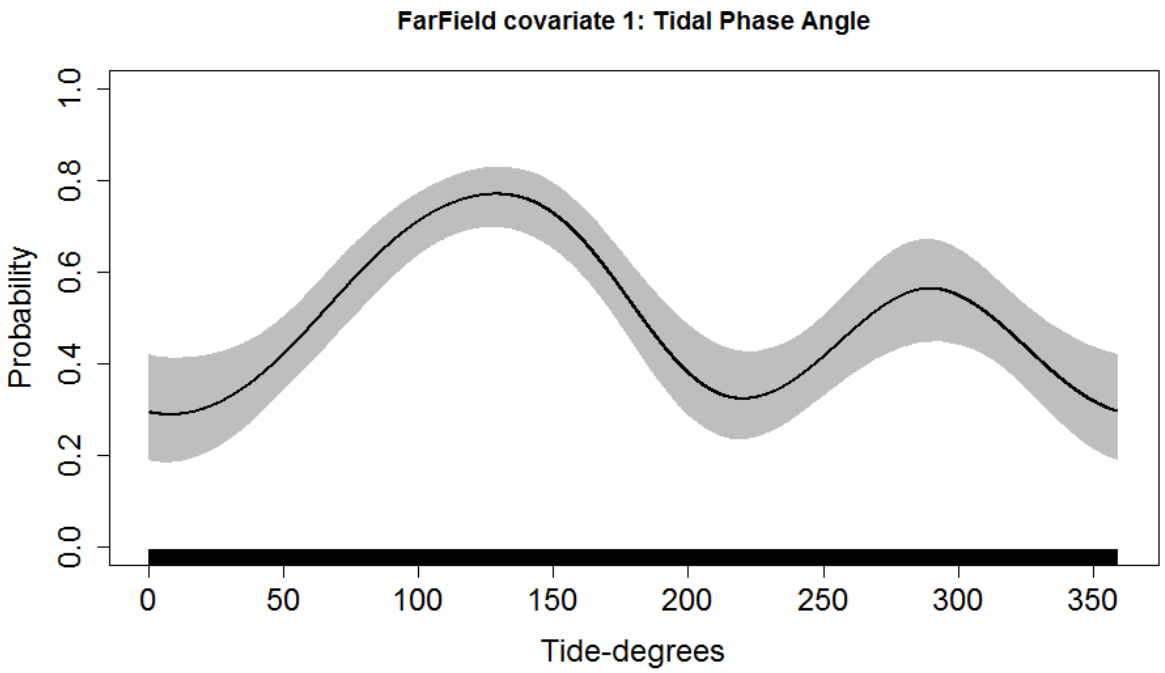
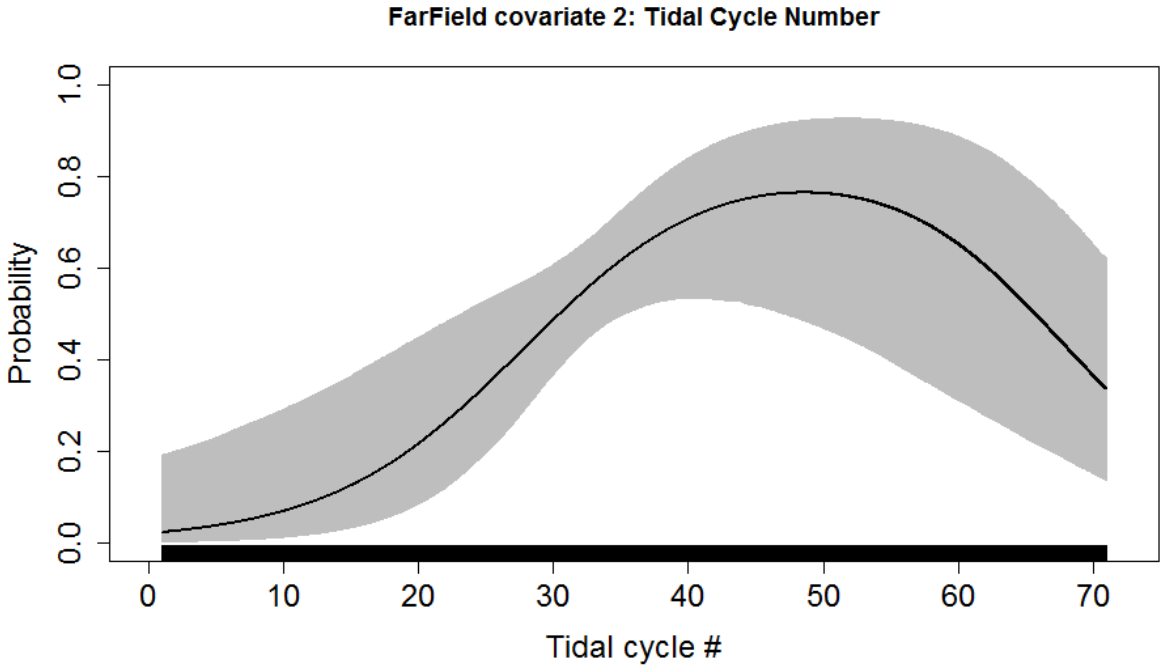


Figure S1. Results of the final Nearfield model for the presence/absence of porpoise click train detections. Vertical axes depict the probability of porpoise detections across the covariate ranges. Covariates are numbered from 1 (most significant) to 3 (least significant), each explaining progressively smaller amounts of residual variability. These graphs should therefore not be interpreted independently.


FarField covariate 1: Tidal Phase Angle
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Tide-degrees


FarField covariate 2: Tidal Cycle Number
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Tidal cycle #

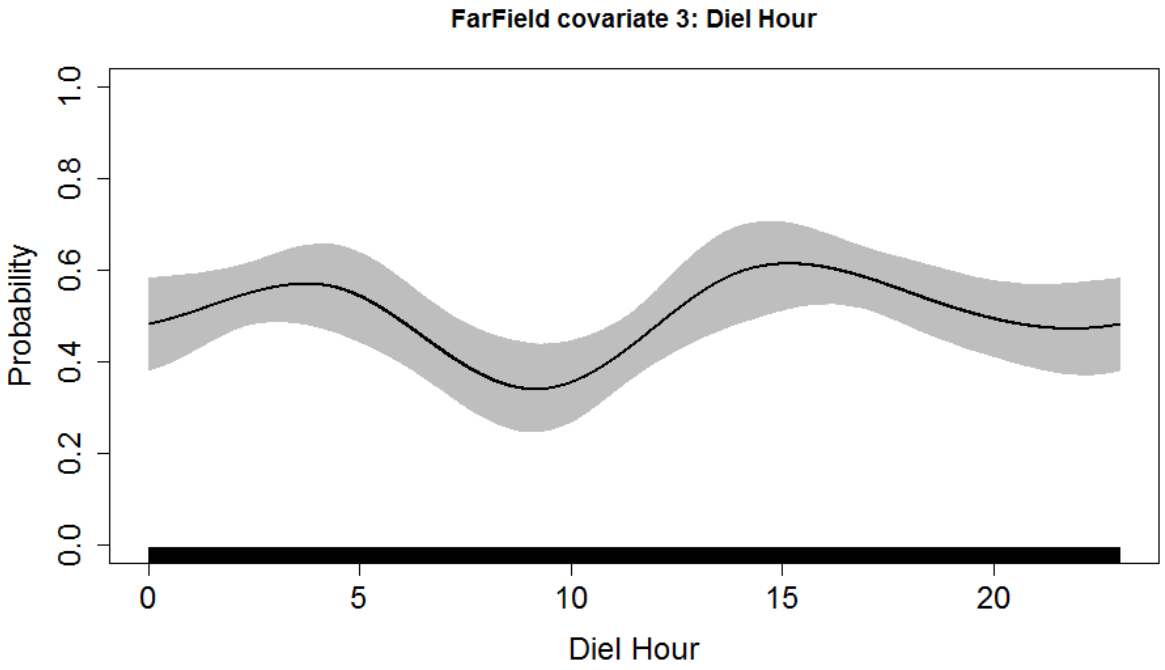
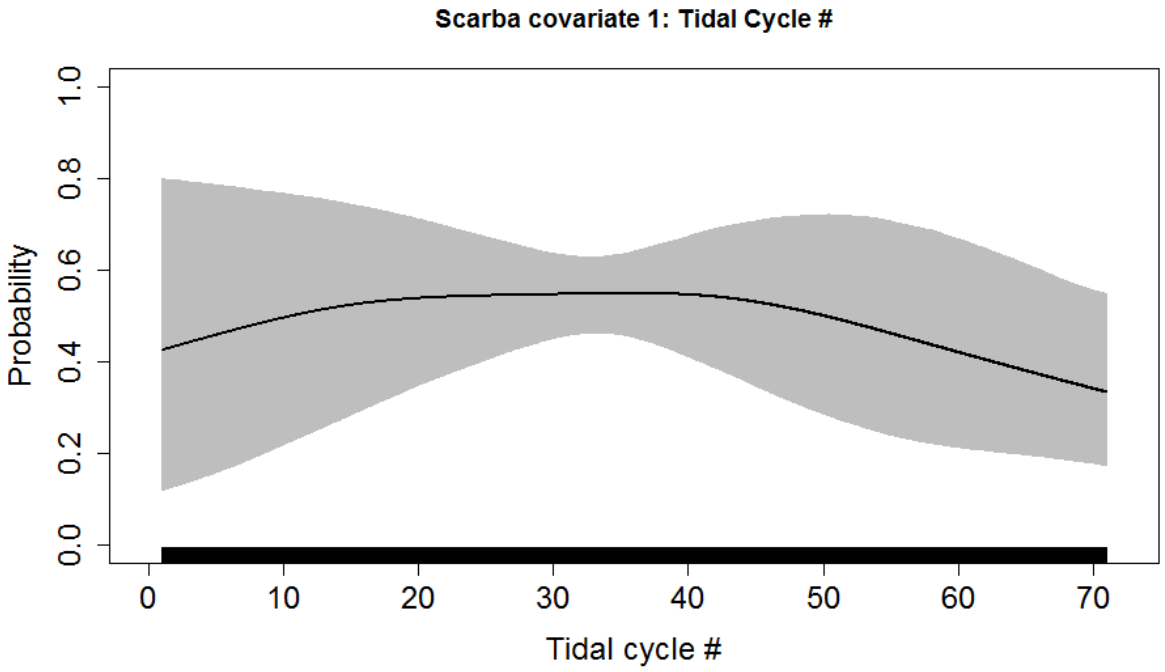
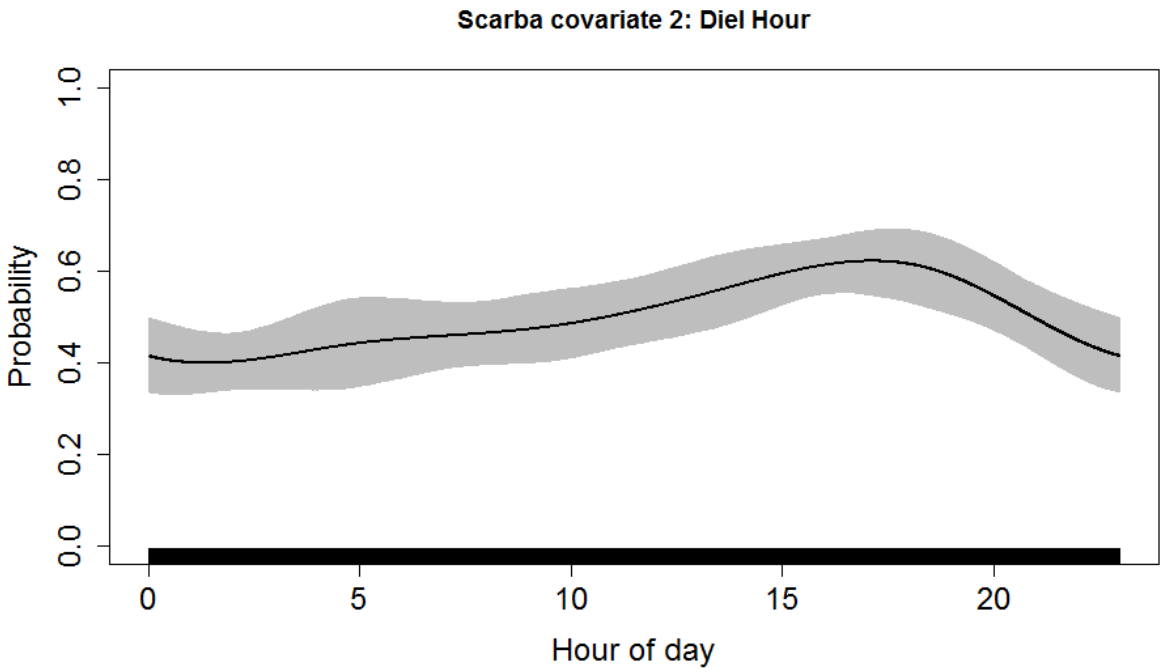


Figure S2. Partial residual plots for significant covariates for the final model at the Farfield site. Vertical axes depict the probability of porpoise detections across the covariate ranges. Covariates are numbered from 1 (most significant) to 3 (least significant), each explaining progressively smaller amounts of residual variability. These graphs should therefore not be interpreted independently.


Scarba covariate 1: Tidal Cycle #
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Tidal cycle #


Scarba covariate 2: Diel Hour
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Hour of day

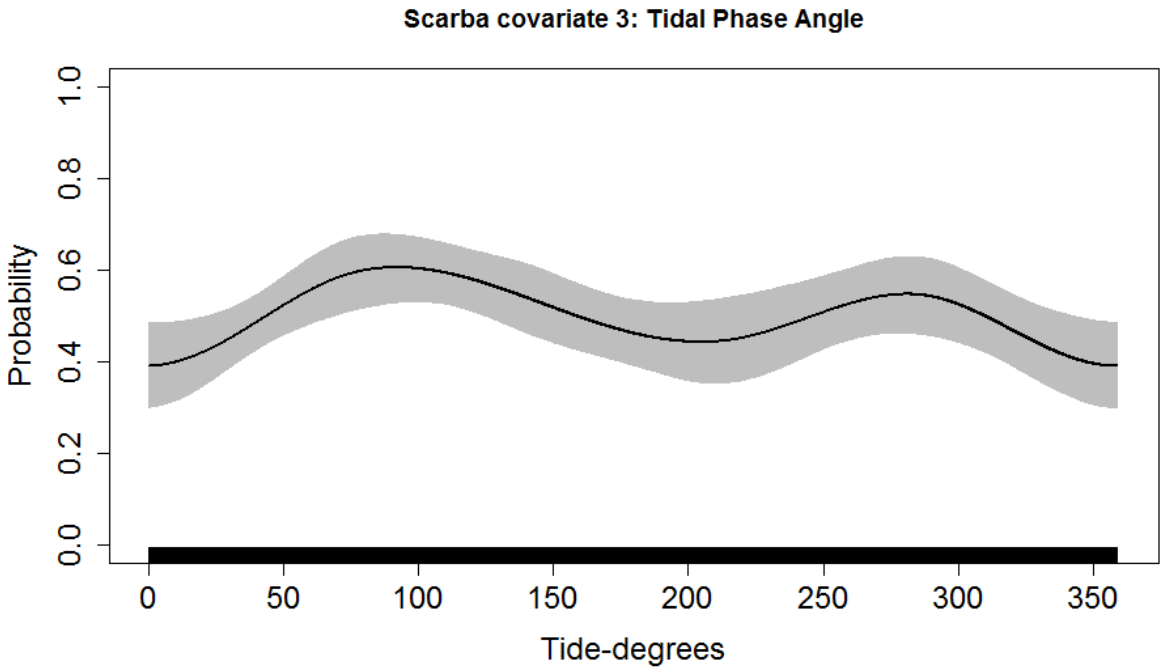


Figure S3. Partial residual plots for significant covariates for the final model at the Scarba site. Vertical axes depict the probability of porpoise detections across the covariate ranges. Covariates are numbered from 1 (most significant) to 3 (least significant), each explaining progressively smaller amounts of residual variability. These graphs should therefore not be interpreted independently.

Results from confusion matrices indicate that the Farfield final model worked best in terms of predicting both presence and absence of porpoises, closely followed by the Nearfield final model. The Scarba final model did less well in correctly predicting porpoise absence than the others (Table S3).

Table S3. Summary of confusion matrices (transformed into percentages) to assess performance of the final model for each location. Percentages indicate for each model what fraction of predicted porpoise detection events (Porpoise vs. No Porpoise) corresponded to factual observations at each site. Green cells = correctly predicted fractions, pink cells = incorrectly predicted fractions. Higher values in Green cells indicate a better working model.

Nearfield

	Observed	
Predicted	Porpoise	No Porpoise
Porpoise	68%	23%
No Porpoise	32%	77%

Farfield

	Observed	
Predicted	Porpoise	No Porpoise
Porpoise	73%	36%
No Porpoise	27%	64%

Scarba

	Observed	
Predicted	Porpoise	No Porpoise
Porpoise	70%	43%
No Porpoise	30%	57%

MODELLING RESULTS - DRIFTERS

As for the moored C-PODs, data exploration following Zuur et al. (2010) confirmed temporal autocorrelation in the data, providing justification for the GAM-GEE approach. All data were aggregated by tide-degrees, as described above for moored C-POD data. All three drifter datasets were modelled separately to assess which covariates might be most important for each deployment. Two covariates (Tidal Phase Angle and Diel Hour) were modelled using cyclic splines based on

variance-covariance matrices. Overall final model structures were as follows (most important covariates first, beyond Latitude/Longitude):

June 2011: Tidal Phase Angle, Days since Deployment, C-POD ID, Average N of raw clicks received

September 2012: Days since Deployment, Average N of raw clicks received

October 2013: Drift Speed, Period of Day, Tidal Phase Angle, Depth (m)

Models differed between years, reflecting the highly stochastic nature of drifter data aggregated across large areas in a spatiotemporally variable environment. Model results, using the Wald's test, indicated that Drift Speed and Tidal Phase Angle were important during flood tides (2011 and 2013 data). None of the environmental covariates proved significant in 2012, apart from Average N of raw clicks received. Days since Deployment was important in 2011 and 2012, indicating changes in detection rates as drifters moved away from the Great Race, but variability in 2012 was considerable. There was evidence for a diurnal cycle in the 2013 data (Table S4).

Table S4. Results of Wald's tests for all significant covariates for the final model for each drifter deployment bout.

	2011			**2012**			**2013**		
Covariate	**DF**	**χ^2 score**	**P-value**	**DF**	**χ^2 score**	**P-value**	**DF**	**χ^2 score**	**P-value**
Latitude	4	15.23	0.00425	4	21.50	0.00025	4	13.61	0.00866
Longitude	4	8.89	0.06386	4	13.39	0.00950	4	16.11	0.00287
Days since	1	22.67	$1.9 \cdot 10^{-6}$	2	9.90	0.00710	Not significant		

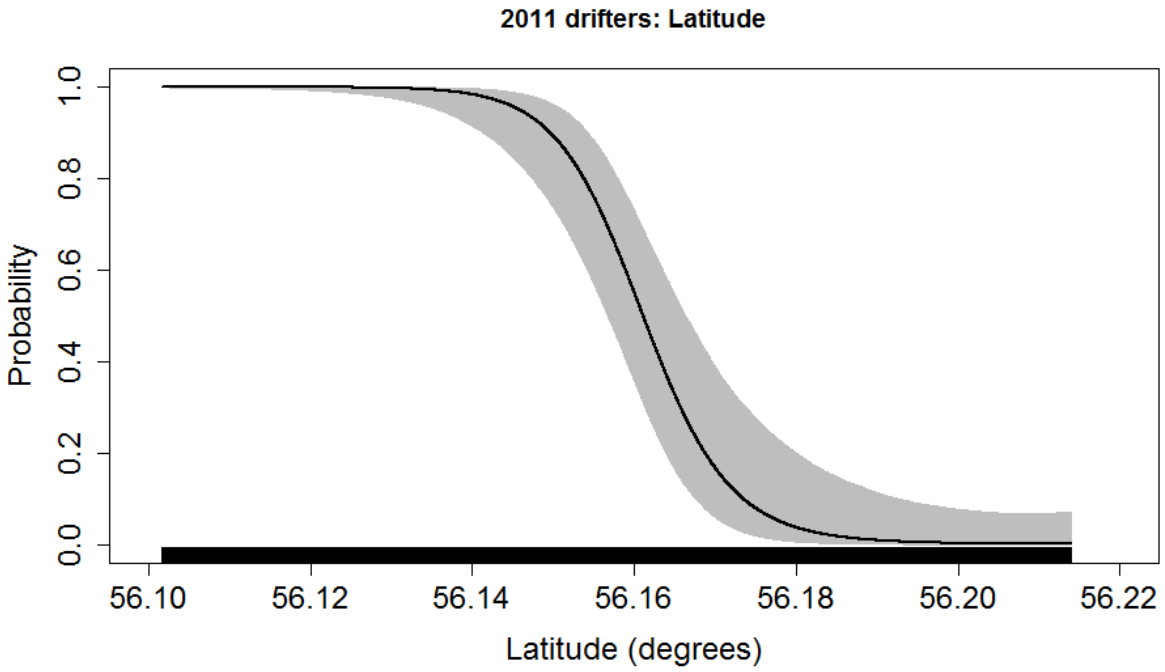
Deployment								
C-POD ID	2	12.46	0.00198	Not significant			Not significant	
Average N of raw clicks received	1	5.37	0.02047	4	17.70	0.00141	Not significant	
Drift Speed	Not significant			Not significant			4	74.55	$2.4 \cdot 10^{-15}$
Period of Day	Not significant			Not significant			3	8.99	0.02949
Tidal Phase Angle	4	59.59	$1.9 \cdot 10^{-12}$	Not significant			4	17.32	0.00168
Average Depth	Not significant			Not significant			4	10.76	0.02940

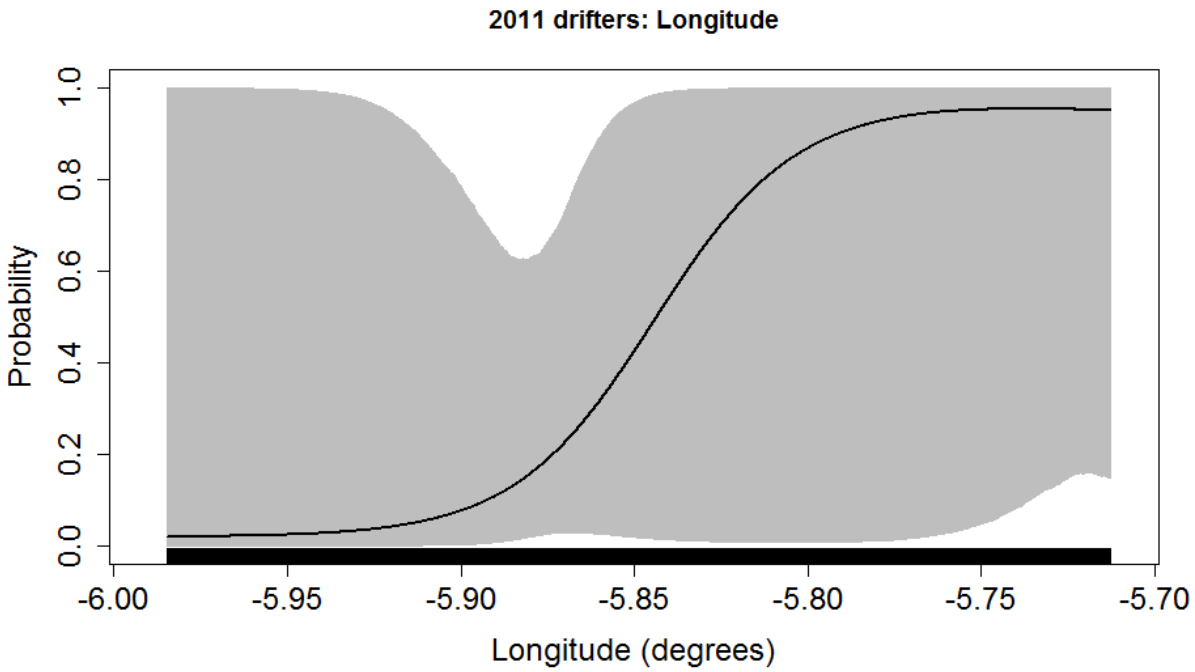
As for the moored C-PODs, covariates were plotted independently to visualise the relationship between each covariate and the response variable (porpoise detection) at each site (Figure S4-S6). The 2012 dataset suffered from small numbers of porpoise detections, making it difficult to illustrate clear relationships between covariates and detections. The significance of particular covariates also varied between years:

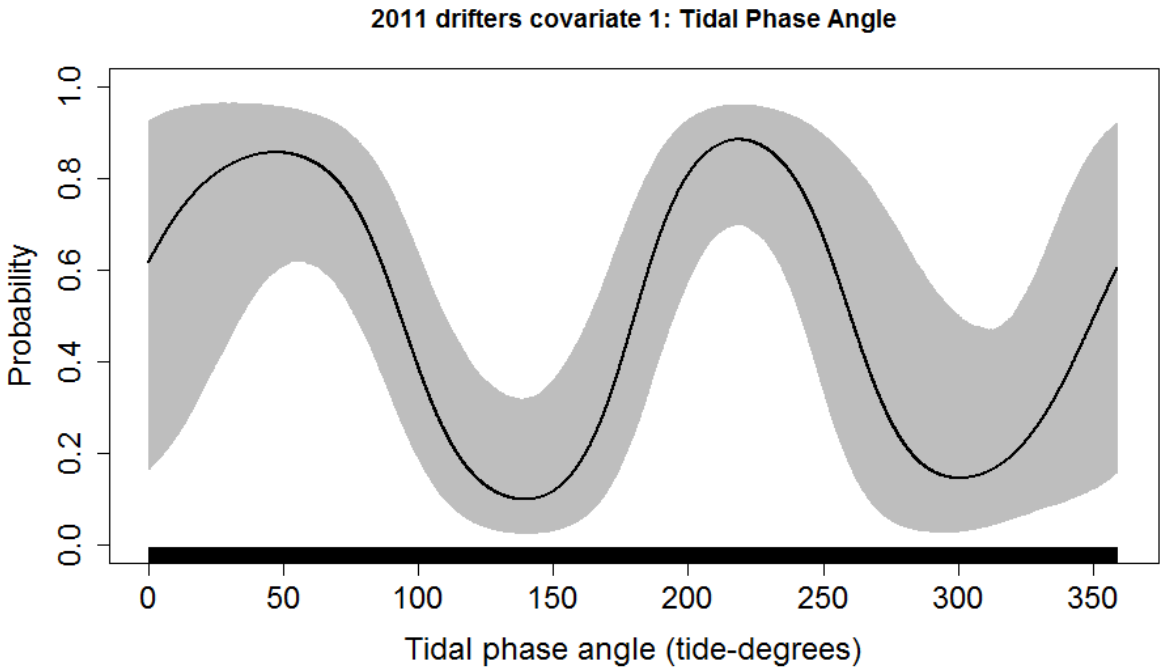
- In 2011, Tidal Phase Angle was the most significant covariate, with a cyclic pattern indicating increased porpoise detection probability during early stages of both rising and falling tide. Results from the 2013 model were quite different – here Tidal Phase Angle was less significant than Drift Speed and Period of Day, and indicated higher porpoise detection probability around High Tide at Oban than during Low Tide at Oban (Figure S4).
- Days since Deployment was significant in 2011 and 2012; in both cases more detections occurred in the 1st day of deployment than on subsequent days, although large confidence intervals obscured this outcome in 2012 (Figure S5).
- C-POD ID was only important in 2011, suggesting that one device (#1653) detected fewer click trains than the others.
- The Average N of raw clicks received was a significant covariate in 2011 and 2012. In the 2011 model, the covariate was included as a linear term, whilst for 2012 data it was incorporated as a smoothed variable. In 2011, detection probability declined under noisier

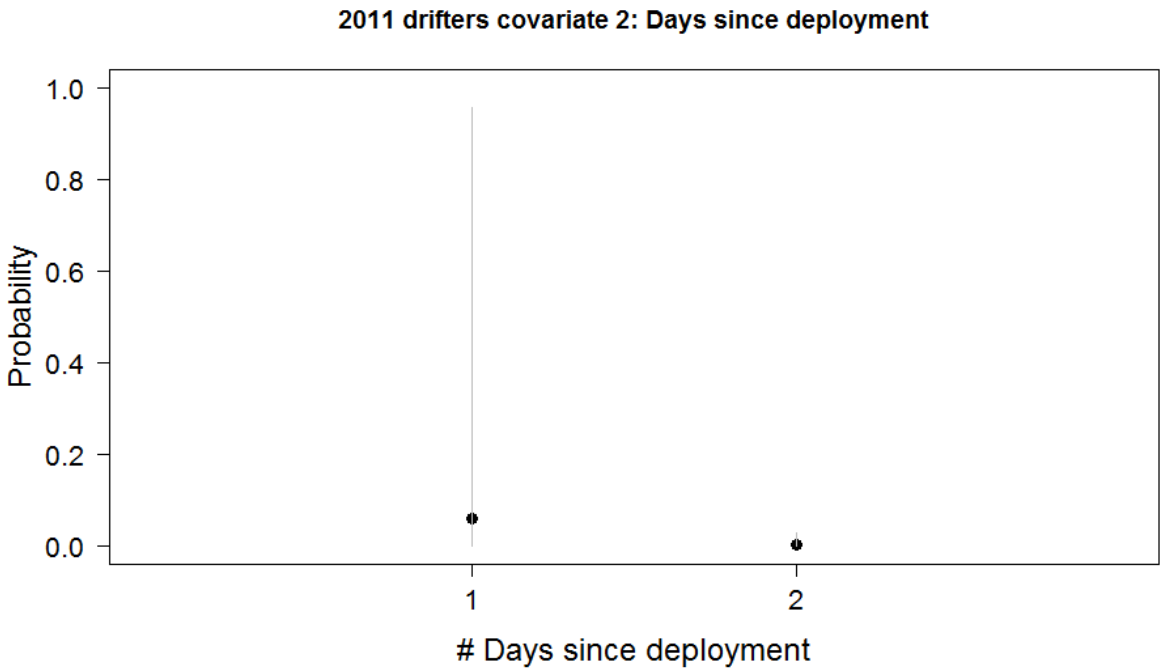
conditions, whereas the pattern was more complex in 2012. However, variability was very large (Figure S4, S5).

- Drift Speed was by far the most significant covariate in 2013, with increased detection probability during faster flows (>1.5 m s^{-1}; Figure S6). This may have been caused by deployments in 2013 occurring further east in the Gulf of Corryvreckan than in 2011, resulting in a wider range of drift speeds.
- Period of Day was important in 2013, with increased detection probability during either day or night (Figure S6).
- In 2013, Depth was a significant covariate, implying increased detection probability in depths <150 m (Figure S6).
- Finally, although not technically part of the covariate selection process, Latitude and Longitude had a significant effect in all models, suggesting that detection distribution was geographically nonuniform, with most detections occurring in the Great Race rather than the Gulf of Corryvreckan.


2011 drifters: Latitude
Probability
0.0
0.2
0.4
0.6
0.8
1.0
56.10
56.12
56.14
56.16
56.18
56.20
56.22
Latitude (degrees)


2011 drifters: Longitude
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-6.00
-5.95
-5.90
-5.85
-5.80
-5.75
-5.70
Longitude (degrees)


2011 drifters covariate 1: Tidal Phase Angle
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Tidal phase angle (tide-degrees)


2011 drifters covariate 2: Days since deployment
Probability
0.0
0.2
0.4
0.6
0.8
1.0
1
2
Days since deployment

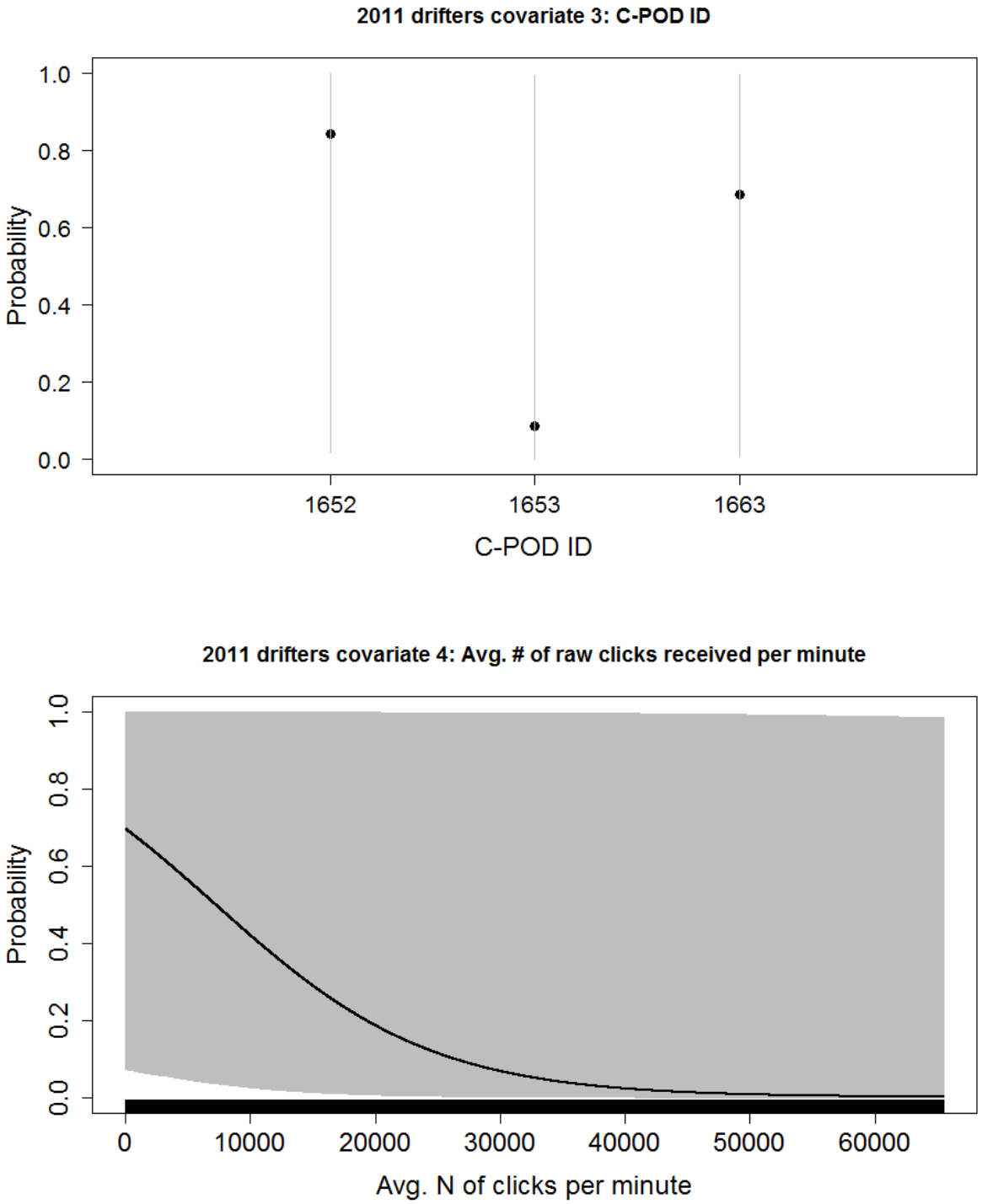
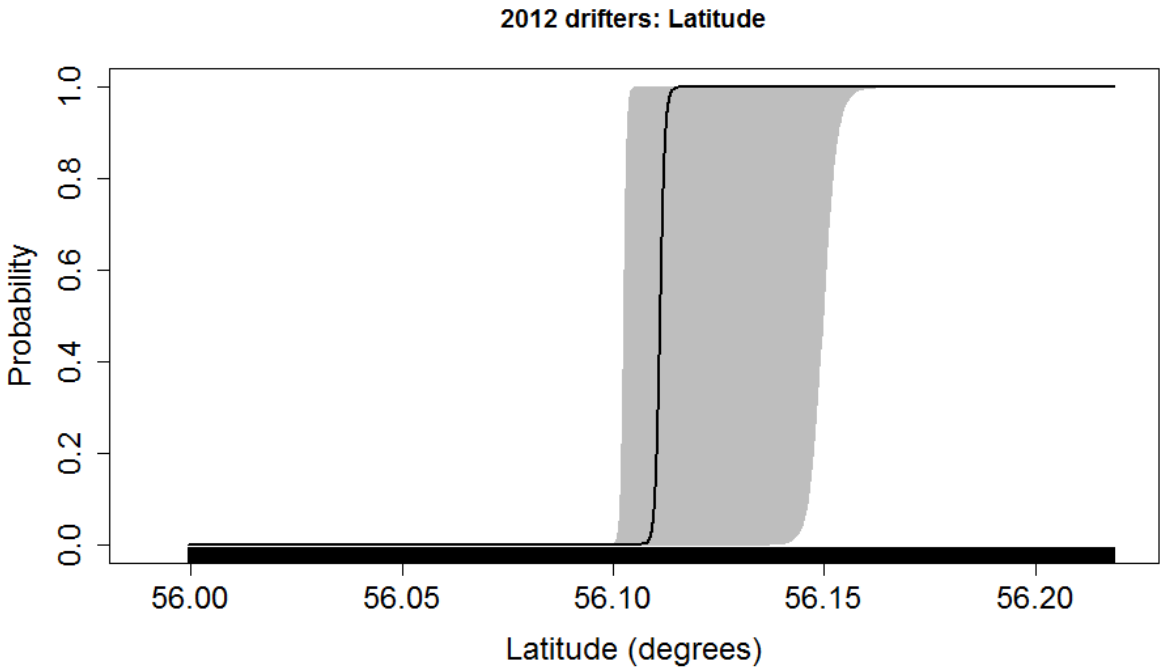
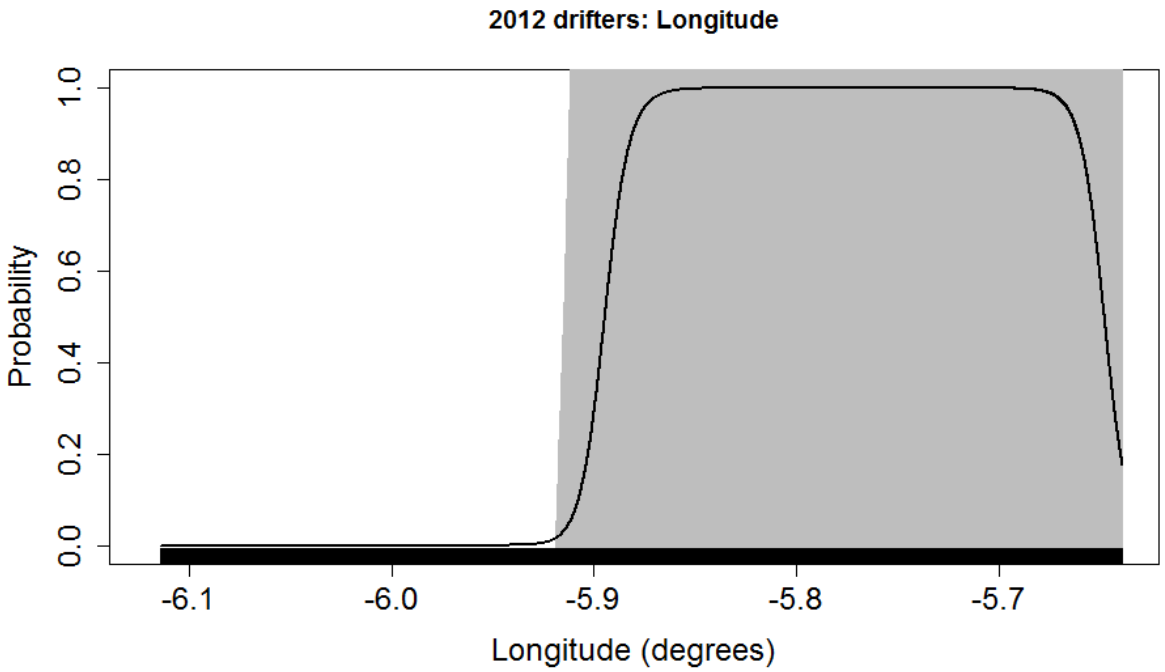
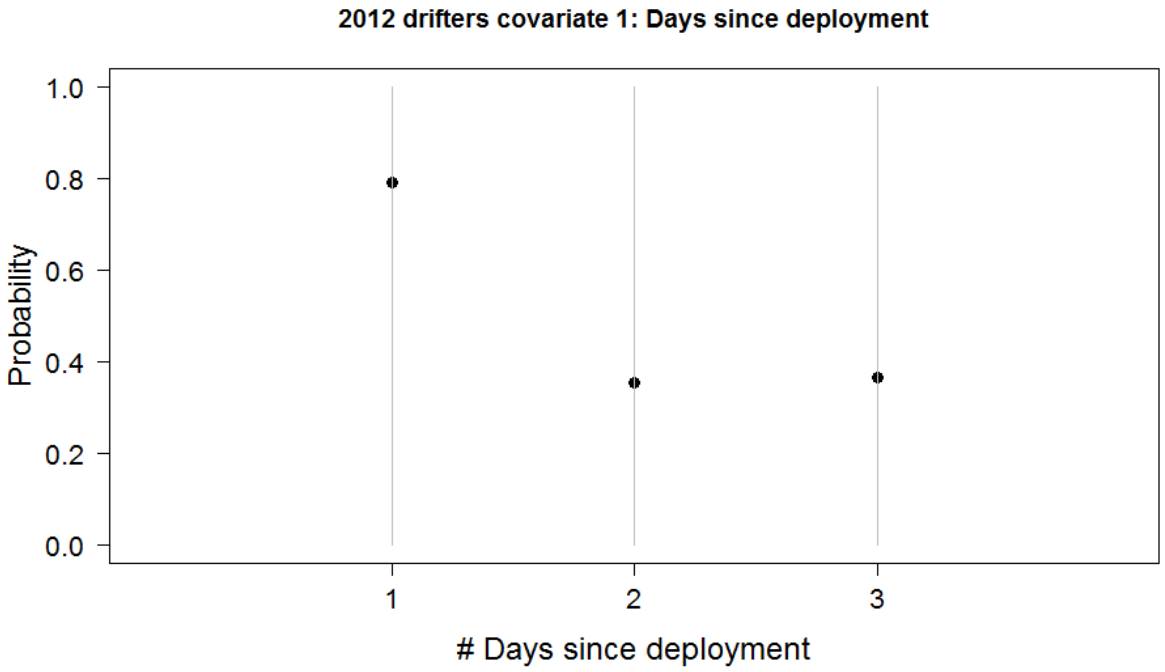


Figure S4. Partial residual plots for significant covariates for the final model for 2011 drifter data. Vertical axes depict the probability of porpoise detections across the covariate ranges. Latitude and Longitude were included in all models without being subject to model selection. Covariates are numbered from 1 (most significant) to 4 (least significant), each explaining progressively smaller amounts of residual variability. These graphs should therefore not be interpreted independently.


2012 drifters: Latitude
Probability
0.0
0.2
0.4
0.6
0.8
1.0
56.00
56.05
56.10
56.15
56.20
Latitude (degrees)


2012 drifters: Longitude
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-6.1
-6.0
-5.9
-5.8
-5.7
Longitude (degrees)



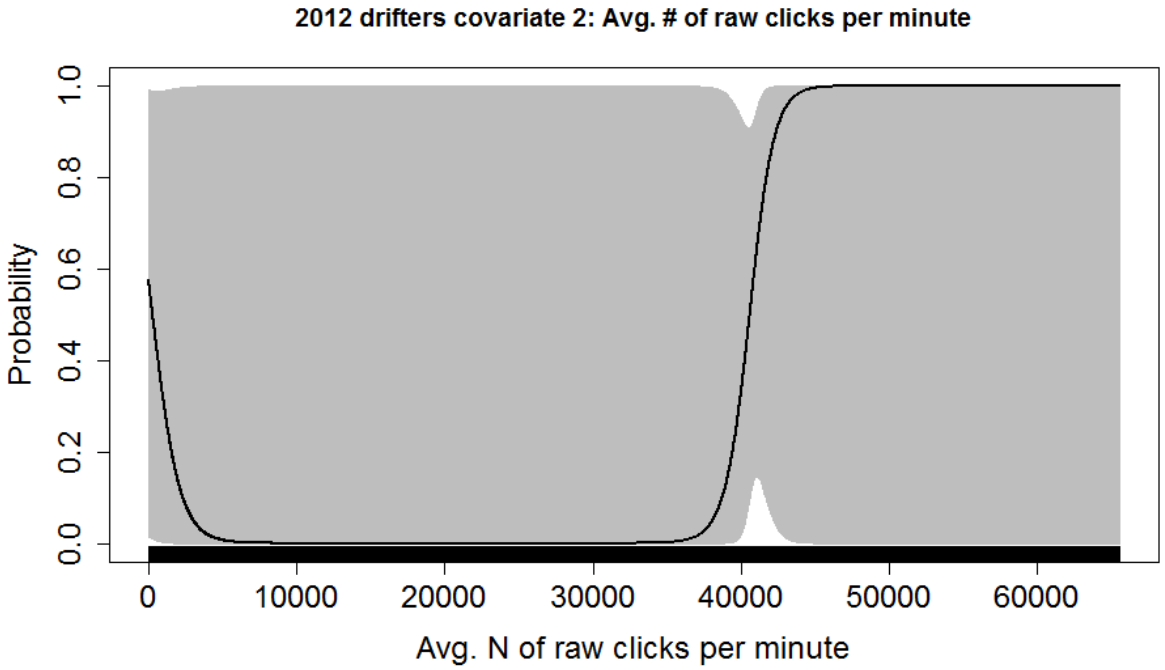
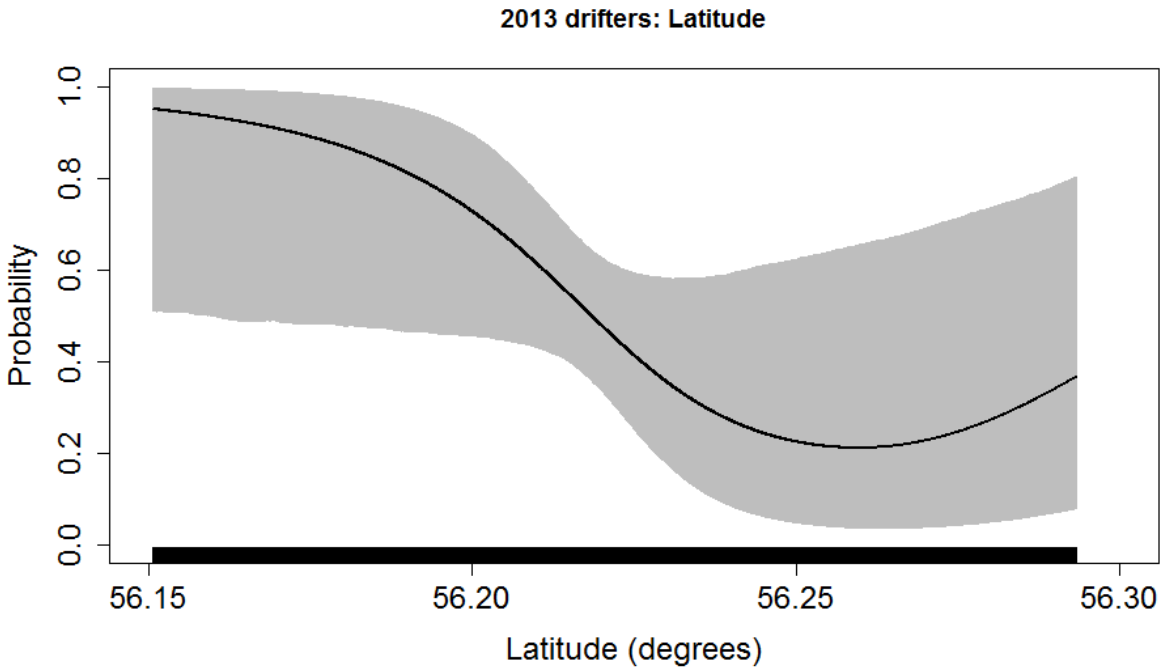
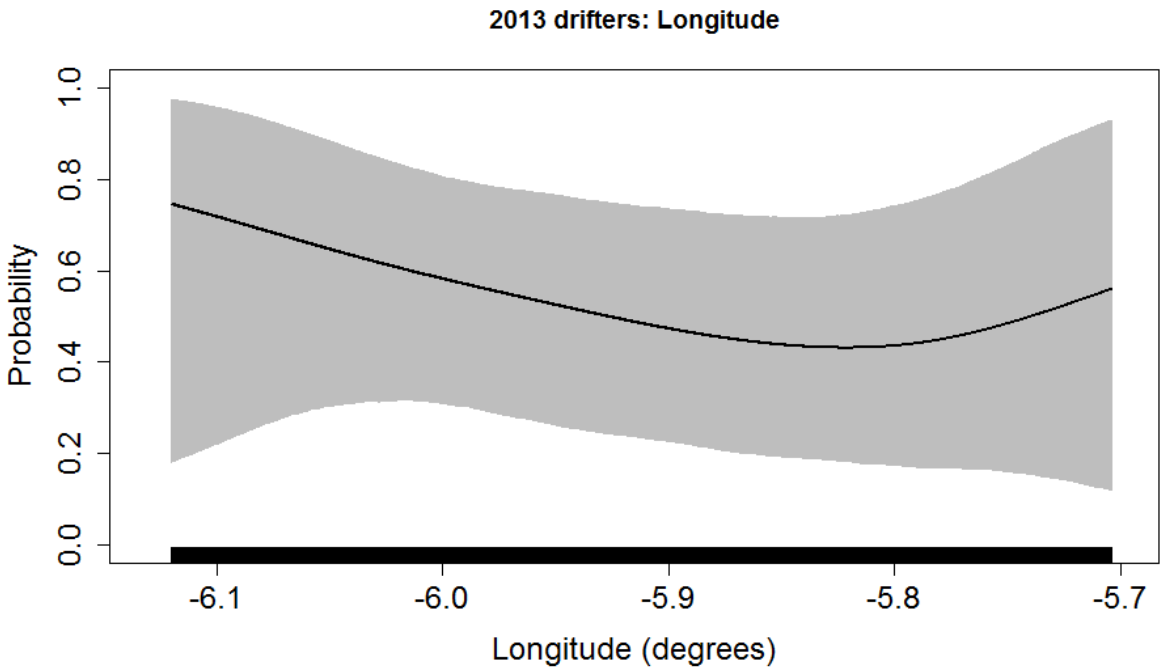
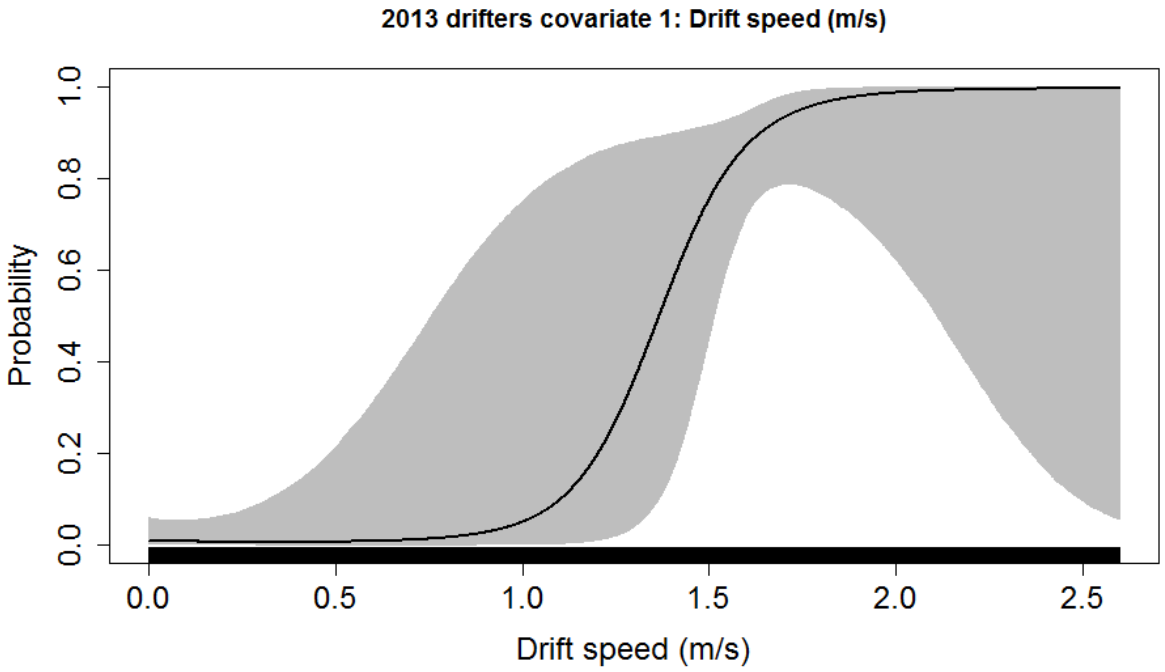
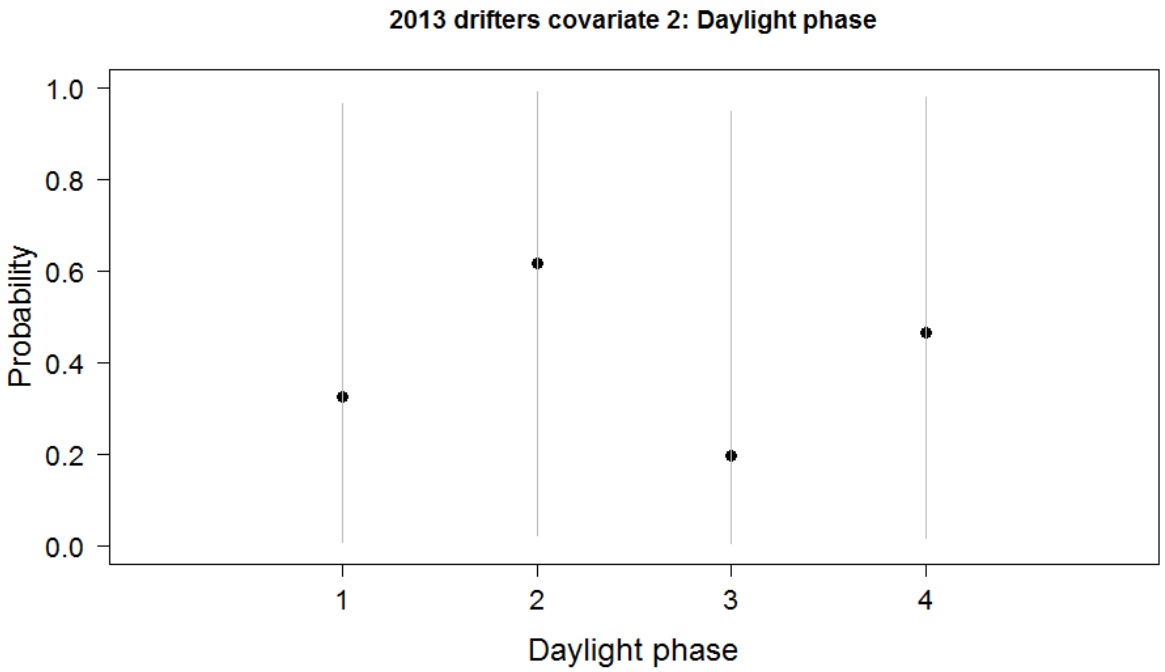


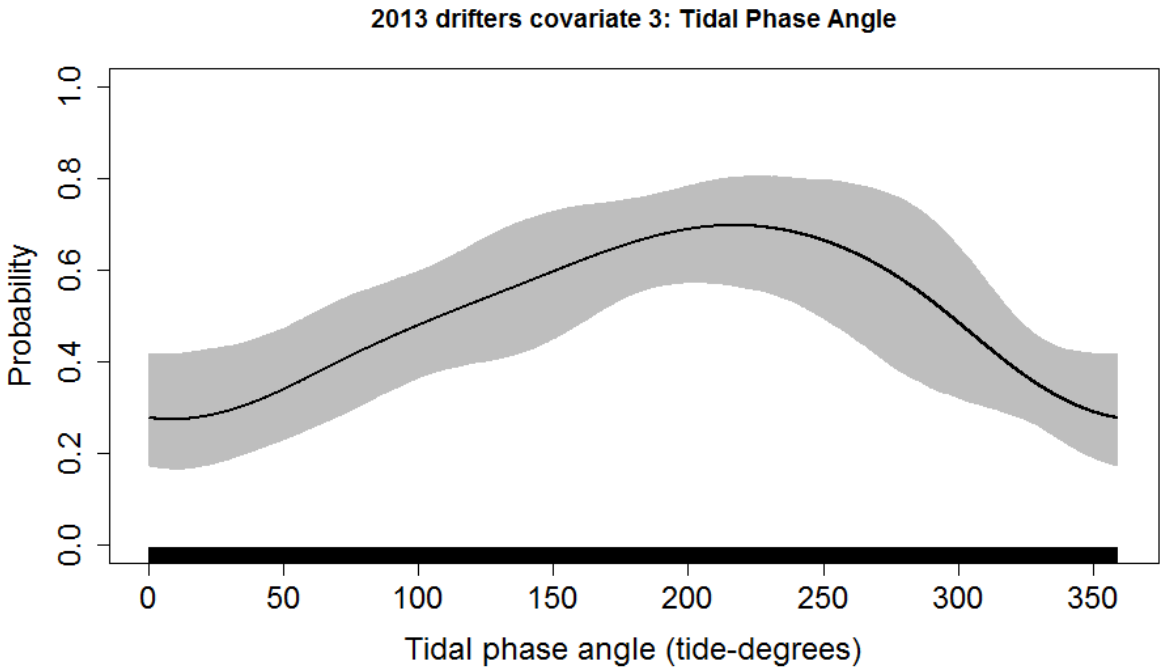
Figure S5. Partial residual plots for significant covariates for the final model for 2012 drifter data. Vertical axes depict the probability of porpoise detections across the covariate ranges. Latitude and Longitude were included in all models without being subject to model selection. Covariates are numbered from 1 (most significant) to 2 (least significant), each explaining progressively smaller amounts of residual variability. These graphs should therefore not be interpreted independently.


2013 drifters: Latitude
Probability
0.0
0.2
0.4
0.6
0.8
1.0
56.15
56.20
56.25
56.30
Latitude (degrees)


2013 drifters: Longitude
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-6.1
-6.0
-5.9
-5.8
-5.7
Longitude (degrees)


2013 drifters covariate 1: Drift speed (m/s)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
Drift speed (m/s)


2013 drifters covariate 2: Daylight phase
Probability
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Daylight phase



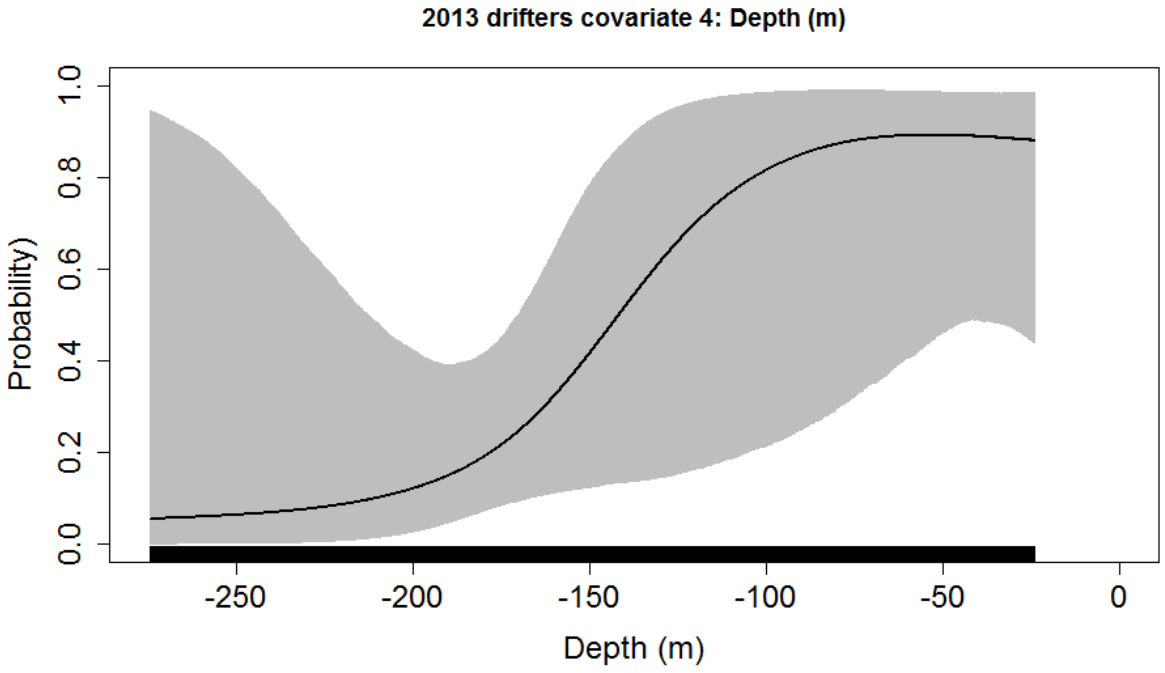


Figure S6. Partial residual plots for significant covariates for the final model for 2013 drifter data. Vertical axes depict the probability of porpoise detections across the covariate ranges. Latitude and Longitude were included in all models without being subject to model selection. Covariates are numbered from 1 (most significant) to 4 (least significant), each explaining progressively smaller amounts of residual variability. These graphs should therefore not be interpreted independently.

Results from confusion matrices indicated that all models performed relatively well in terms of predicting both presence and absence of porpoises, although the 2013 model performed least well (Table S5).

Table S5. Summary of confusion matrices (transformed into percentages) to assess performance of the final model for each year. Percentages indicate for each model what fraction of predicted porpoise detection events (Porpoise vs. No Porpoise) corresponded to factual observations at each site. Green cells = correctly predicted fractions, pink cells = incorrectly predicted fractions. Higher values in Green cells indicate a better working model.

2011

	Observed	
Predicted	Porpoise	No Porpoise
Porpoise	86%	15%
No Porpoise	14%	85%

2012

	Observed	
Predicted	Porpoise	No Porpoise
Porpoise	89%	20%
No Porpoise	11%	80%

2013

	Observed	
Predicted	Porpoise	No Porpoise
Porpoise	70%	26%
No Porpoise	30%	74%

BIBLIOGRAPHY

Akaike H (1974) A new look at the statistical model identification. IEEE T Automat Contr 19(6): 716-723

Carey VJ (2004) *yags*: yet another GEE solver. *R package version*, 4-0

Carlström J (2005) Diel variation in echolocation behavior of wild harbor porpoises. Mar Mam Sci 21(1): 1-12

Fielding AH, Bell JF (1997) A review of methods for assessment of prediction errors in conservation presence/absence models. Environ Conserv 24(1): 38-49

Garson GD (2013) Generalized Linear Models & Generalized Estimating Equations. Statistical Associates Publishers: Asheboro, NC, USA

Halekoh U, Højsgaard S, Yan J (2006) The R package *geepack* for generalized estimating equations. J Stat Softw 15(2): 1-11

Hardin JW, Hilbe JM (2003) Generalized estimating equations. Chapman & Hall/CRC Press, London, UK

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer, New York, NY, USA

Hilbe JM (2011) Negative Binomial Regression. Second Edition. Cambridge University Press, Cambridge, UK

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

McCullagh P. Nelder JA (1989) Generalized Linear Models, Second Edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, London, UK

Pan W (2001) Akaike's information criterion in generalized estimating equations. Biometrics 57(1): 120-125

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer, New York, NY, USA

Zar JH (1999) Biostatistical Analysis. Fourth Edition. Prentice-Hall International, Upper Saddle River, New Jersey, USA

Zuur AF (2012) A Beginner's Guide to Generalised Additive Models with R. Highland Statistics Ltd., Newburgh, UK

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. Springer, New York, NY, USA

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Method Ecol Evol 1: 3–14