Final Report

Project Name:	EMEC Fall of Warness
Reference:	010607EMEC
Project Manager:	Beth Mackey

Drafted by:	DMP Statistical Solutions	
Checked by:	Dr Gordon Hastie	
Approved by:	Dr Beth Mackey	
Date:	30th July 2009	

VAT reg. No. GB 607 6064 48

SMRU LIMITED is a limited company registered in Scotland, Registered Number: 296937.
Registered Office: 5 Atholl Crescent, Edinburgh EH3 8EJ

Analysis of Bird and Marine Mammal Data for the Fall of Warness Area: A report prepared for SMRU Ltd

DMP Statistical Solutions UK Ltd.

30 June 2009

Table of Contents

1 Introduction

This report gives a review of the statistical analyses of the Fall of Warness bird and marine mammal data. For clarity, technical details of the statistical models are separated from the main results in each section.

The main objectives of this analysis were to:

- assess the distribution of birds and marine mammals across the study site and across time;
- provide outputs that summarise their distribution;
- ascertain relationships between collected environmental variables and the observed relative bird and marine mammal abundances;
- determine the ability of the observation and analysis methods to detect a variety of tidal stream energy device impacts on bird and marine mammal abundances.

The data collection and analysis is thought to ultimately allow the detection of changes in the bird and marine mammal usage of the Fall of Warness region through time. This might be in terms of their use of the space, and/or their response to measured environmental conditions e.g. they may shift preferred locations of activity or become more active or visible at different times of day. This would form the basis of an environmental impact assessment for the placement of the proposed tidal stream energy devices.

Initial Caveats:
The terms *abundance/counts/numbers* will be used throughout this report – this requires initial clarification. The nature of the study (a fixed observation point) means that there is likely to be a lower probability of observing animals the greater the distance from the observation point. For example comparing 2 sites, one far and one near – even with equal numbers of animals, the further site will have an apparently lower abundance due to lower probability of detection.

Due to this, any numbers derived cannot be interpreted as *true* count estimates without proper account of this detection probability (inestimable for the current data). The terms used here will be synonymous with detected numbers. However, the intended use is for measuring relative change through *time*, which can be ascertained if the study design remains unchanged, detectability remains constant and the power to detect an impact is sufficiently high. Relative spatial sightings are similarly affected by detection probability.

It is assumed throughout that inference is restricted to times similar to those sampled i.e. general daylight hours.

The term *bird* and *marine mammal* is synonymous with the species represented in the dataset provided by SMRU Limited.

2 Data Details

This section briefly outlines the data manipulation, exploratory analysis and final analyses applied. Greater detail is given in the relevant sections.

2.1 Software

All data analysis was performed in the statistical data package R, with associated add-on packages as required[1].

2.2 Data Preprocessing

The following are notes relating to the data manipulation performed on the environmental and species count data contained within the files provided.

Summary:

Data pre-processing comprised of the actions previously detailed in the 2006 analysis (DMP, 2007). In general data pre-processing consisted of the following:

- Alteration of all mis-spelt zone codes, species codes, and species names (refer to Section 7.1 and Section 7.2 for more details).
- Reduction of the numbers of categories in variables (e.g. precipiation) to a smaller more tractable set.
- Inference of missing values where possible. E.g. tide heights inferred from time and date; missing values for wind-direction assumed to be zero when sea-state was zero.
- Actual tide heights were interpolated based on the known high/low tide times and heights, coupled with a sinusoidal curve.
- Inconsistencies in character field entries fixed e.g. “Hi. 0302” versus “Hi. 302”.

[1] Version 2.8, R Development Core Team (2008) *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

2.3 Data Available

All environmental data available were initially considered as candidates for prediction of bird and marine mammal abundance.

The variables considered in the statistical models for the animal counts were:

- *Wind strength*: a score measured from 0 to 6
- *Sea State*: a score measured from 0 to 6
- *Wind Direction*: 9 categories
    - East, None, North, North-East, North-West, South, South-East, South-West, West
- *Cloud cover*: a percentage score
- *Precipitation*: 4 categories
    - None, Rain, Showers, Snow
- *State of tide*: 3 categories
    - Ebb, Flood, Slack
- *Water flow speed*: 4 categories
    - Fast, Moderate, Slack, Slow
- *Water flow direction*: 5 categories:
    - North, North-West, Slack, South, South-East
- *Survey Month*: 37 months numbered in order from 0 to 36 (July 2005 – July 2008).
- *Grid Code*: 30 categories as combinations of A, B, C, D, E with 0, 1, 2, 3, 4, 5 which represent the spatial position of the birds sighted (Figure 20, page 20).
- *Time of day:* Observation time
- *Tide height:* Height of tide (m)

Time information was also trialled in the model using 'Month' and 'Year' in addition to the combination of the two ('Survey month') listed above.

3 Modelling relative bird and marine mammal abundances 2005 to 2008.

3.1 Overview

The modelling methods used here are flexible and naturally accommodate relative abundance data collected over time. This helps ensure sound model predictions and realistic confidence about these predictions.

Models for data collected over time deserve special consideration. Specifically, observations collected close together in time are likely to be more similar than observations collected from different days and when ignored this can mean random fluctuations in animal counts over time are confused with real underlying increases or decreases. As a result, time-based statistical methods which naturally accommodate data of this sort were used for this analysis.

Model flexibility was considered important for this analysis, since bird and marine mammal numbers are unlikely to increase (or decrease) at a constant average rate for each variable (e.g. across the year) and so flexible curves were permitted for all variables in the model, where appropriate e.g. Cloud cover, time of day and tide height.

The confidence attributed to model predictions was also considered important in this analysis. Without identifying a realistic range of plausible animal counts before any potential impact, it is virtually impossible to determine if there has been a real change in animal counts after any impact.

3.2 Technical Details

Models for the average number of birds sighted were fitted using Generalized Additive Models (GAMs, Hastie & Tibshirani, 1990) with log link and Quasi-Poisson errors. Splines were used to model the continuous explanatory variables while categorical variables were fitted as factors.

Robust standard errors estimated using Generalized Estimating Equations (GEEs, Liang & Zegar, 1986) were used to adjust for temporal auto correlation in the errors. Specifically, observations within days were permitted to be correlated. Since the data set was of considerable size and GEE standard errors are robust to the nature of the correlation specified, a working independence structure was used.

Variance inflation factors were used to detect collinearity in the model covariates and model selection was carried out using the QIC statistic (Pan, 2001) and GEE-based *p*-values. Specifically, an automated stepwise selection procedure based on the QIC statistic was followed by a backwards selection procedure based on GEE-based Wald tests. This ensured any autocorrelation was accounted for during the model selection process.

The full model contained continuous variables as splines, while categorical variables (such as Grid Code) entered the model as factors. Time information was trialled in the model using 'Month' and 'Year' and the combination of the two ('Survey month'). 'Survey month' was permitted to be modelled as a spline or factor variable during the model selection process.

Two pairs of candidate variables were too similar to be fitted together successfully in a model (i.e. the collinearity was prohibitive). Therefore, for each covariate pair, the term that returned greater predictive power was retained as a candidate for model selection (Table 1).

Variable pair	Variable chosen
Tide State/Flow Direction	Flow Direction
Flow Speed/Flow Direction	Flow Direction

Table 1: Pairs of collinear variables and variable chosen to be considered for selection in the fitted model.

3.3 Results for the Bird Data

This section gives the results of analyses for all bird species combined. Birds were more likely seen in calm conditions (low sea state, low winds), with a moderate water flow speed and in good light conditions.

The following variables were deemed predictors of relative bird abundance in this area:

- Wind Strength
- Sea State
- Cloud Cover
- Flow Speed
- Flow Direction
- Time of day
- Grid Code
- Survey Month

Figure 1 to Figure 5 represents the model relationships using coefficients on the vertical axis[2]. In all cases the estimate is given by a small central point, with the 95% confidence bounds represented by vertical lines.

Plot interpretation for Figure 1 to Figure 5. Higher coefficient values indicate greater predicted numbers of birds. Categories that have confidence bounds that are distinct from the horizontal line can be considered statistically different from the baseline level at the 5% level; baseline level information is included in figure captions. Additionally, all interpretations are made assuming all other terms in the model are held constant; e.g. all else being equal, significantly more birds are predicted to be observed in group A than in group B.

[2] Coefficient values are given on the scale of the link function – refer Technical Section §4.2

Wind Strength: Less birds were seen as conditions became progressively more windy (wind strengths 0 to 3) while bird numbers appeared to be similar for wind strengths 4to 6.

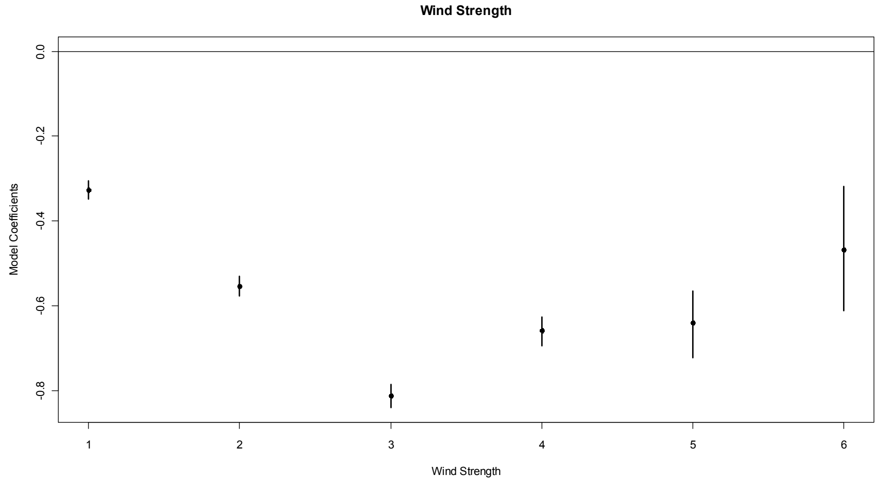

Figure 1: Relative effects of *Wind Strength* on estimated numbers of birds. Dots indicate estimates, vertical lines the 95% uncertainty bounds. A wind strength of 0 is the baseline category (horizontal line at 0). The *p*-value for the GEE-based Wald test for this term is 0.00000 (5 d.p.).

Sea State: In general, more birds were seen in calmer waters for smaller values of sea state, and bird numbers appeared to decline as sea state increased in value. The exception to this occurs for a sea state of 6, where larger numbers of birds were apparently seen in very rough seas. This was due to just two (of the 60) observations when sea state was recorded as 6, where bird counts were as high as 50 when 90% of these 60 observations were zero.

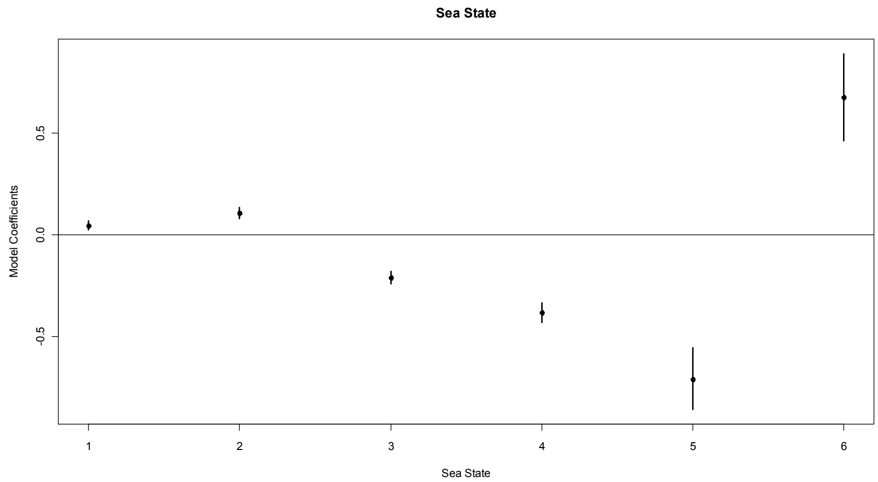

Figure 2: Relative effects of *Sea State* on estimated numbers of birds. Dots indicate estimates, vertical lines the 95% uncertainty bounds. A Sea State of 0 is the baseline category (horizontal line at 0). The *p*-value for the GEE-based Wald test for this term is 0.00000 (5 d.p.).

Flow Speed: Significantly more birds were seen in moderate flow speeds than fast flow speeds, and significantly fewer birds were seen in Slack and Slow water speeds compared with both the moderate and fast flow speeds.

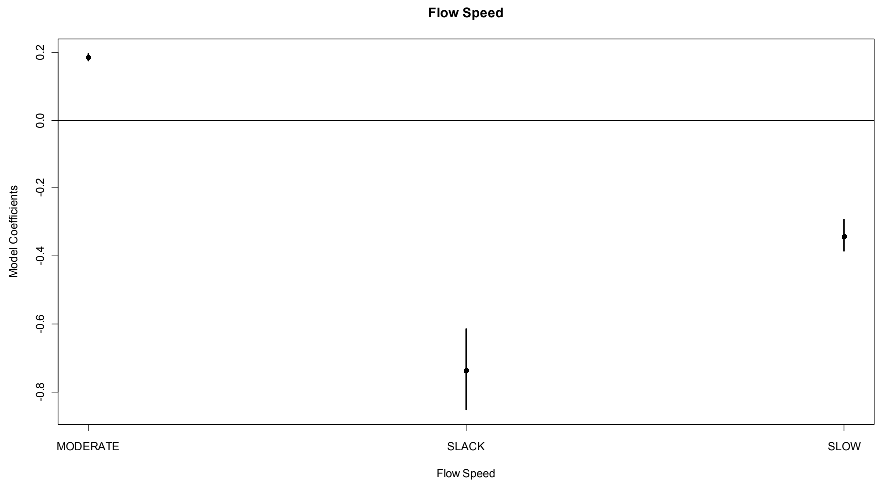

Figure 3: Relative effects of *Flow Speed* on estimated numbers of birds. Dots indicate estimates, vertical lines the 95% uncertainty bounds. The 'Fast' Flow Speed is the baseline category (horizontal line at 0). The *p*-value for the GEE-based Wald test for this term is 0.00005 (5 d.p.).

Flow Direction: Fewer birds were seen when the flow direction was from the North-West and South-East and significantly more when the flow was coming from the North, Slack and South directions.

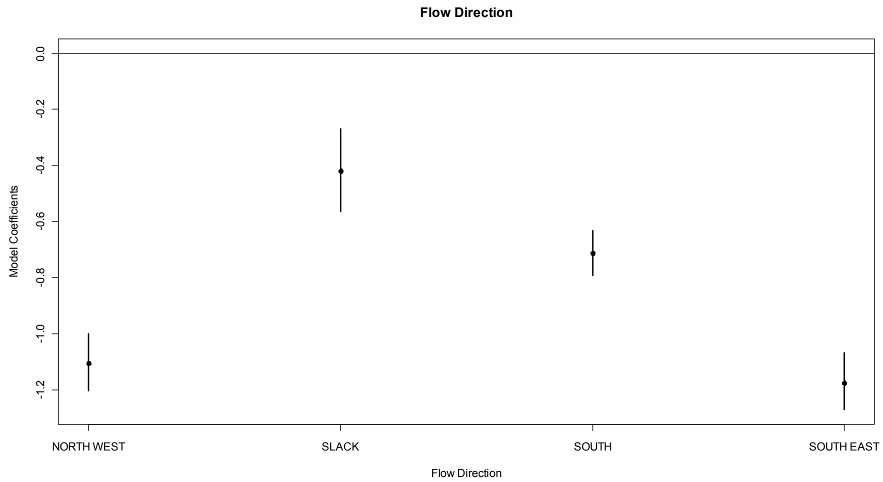

Figure 4: Relative effects of *Flow Direction* on estimated numbers of birds. Dots indicate estimates, vertical lines the 95% uncertainty bounds. The 'North' direction is the baseline category (horizontal line at 0). The *p*-value for the GEE-based Wald test for this term is 0.00000 (5 d.p.).

Cloud Cover: More birds were seen when cloud cover was moderate to high, however there is a great deal of uncertainty about this relationship and so caution must be used when interpreting this curve.

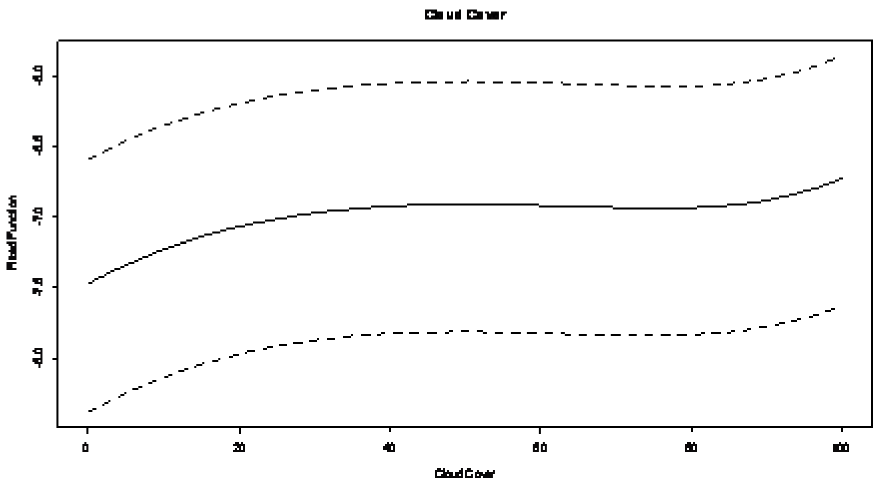

Figure 5: Effects of *Cloud Cover* on estimated numbers of birds. Solid curve indicates estimated function, dotted curve above and below indicate the 95% confidence envelope. The *p*-value for the GEE-based Wald test for this term is 0.01143 (5 d.p.).

Time of day: Under the model, the greatest numbers of birds are found early in the morning, falling away after this point (Figure 6). Naturally, low light levels at the end of the day may have a confounding effect on these results.

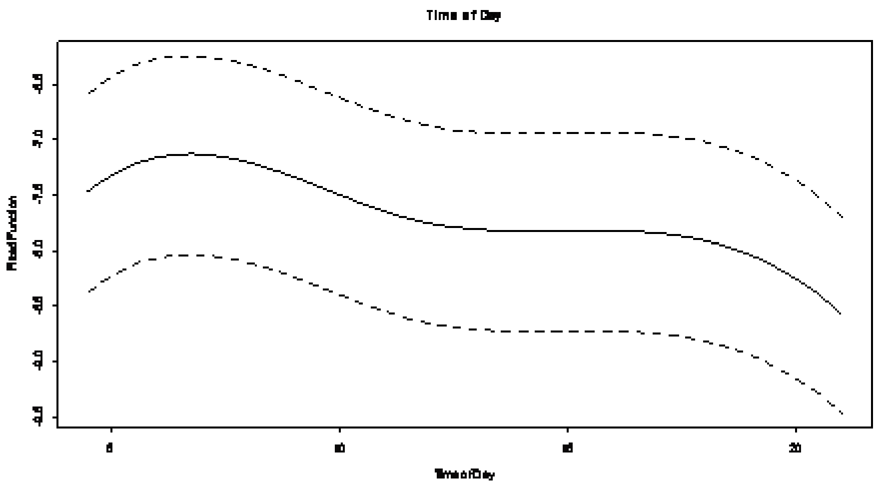

Figure 6: Effects of *Time of Day* on estimated numbers of birds. Solid curve indicates estimated function, dotted curve above and below indicate the 95% confidence envelope. The *p*-value for the GEE-based Wald test for this term is 0.00005 (5 d.p.).

Survey Month: Bird numbers appeared to cycle seasonally but appear to be relatively stable over time.

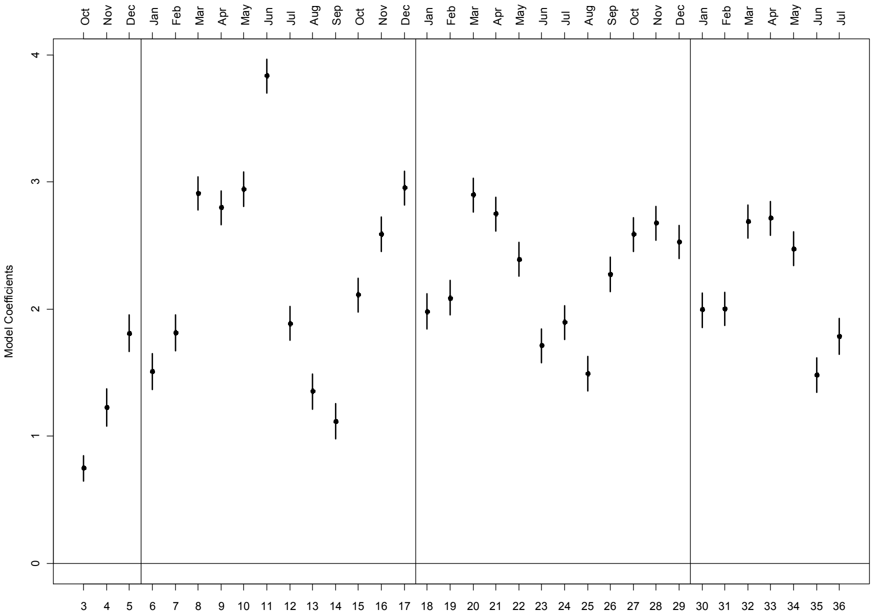

Figure 7: Relative effects of *Survey Month* on estimated numbers of birds. The x-range represents September 2005 (the baseline) to July 2008[3]. Dots indicate estimates, vertical lines the 95% uncertainty bounds. The *p*-value for the GEE-based Wald test for this term is 0.00000 (5 d.p.).

Grid Code: Grid codes generally gave significantly higher predicted bird numbers than grid code *A0* (the baseline) which was statistically indistinct from areas D5 and E5. In keeping with previous analyses, areas close to land (eg. *E0* to *E4,* and *A2* to *A3*) are predicted to exhibit more birds than grid codes far from land. See Figure 20, page 20 for grid code positions.

The grid codes with the highest estimated counts are those very near the island opposite the survey observation post, and the land adjacent to the survey position. Considering the uncertainty in these estimates this spatial pattern is still apparent. The caveat relating to detection probability (§1) is reiterated here.

[3] Survey Months 0 and 1 do not feature in the analysis due to universally missing values for Sea State in these months and inclusion of Sea State in the selected

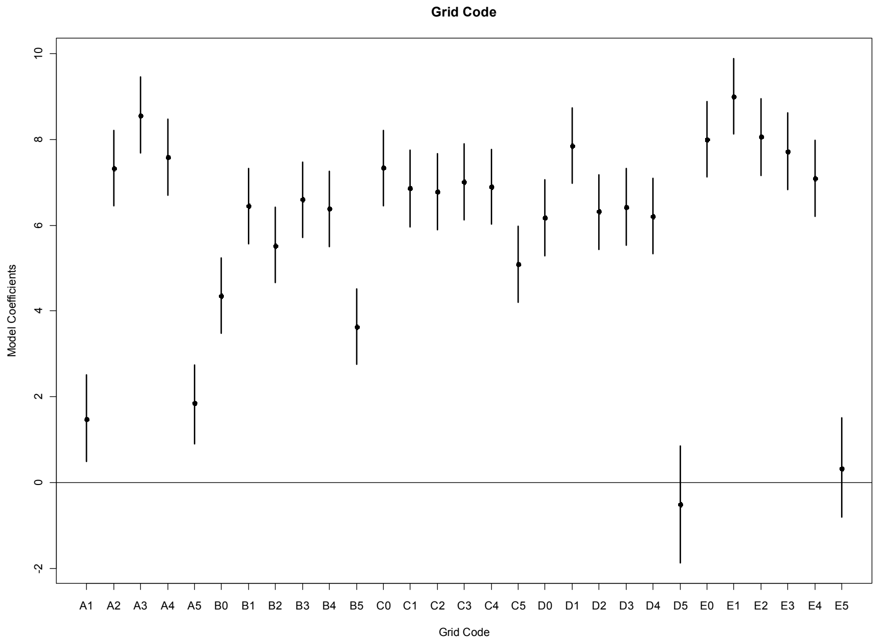

Figure 8: Relative effects of *Grid Code* on estimated numbers of birds. Dots indicate estimates, vertical lines the 95% uncertainty bounds. The *p*-value for the GEE-based Wald test for this term is 0.00000 (5 d.p.).

3.4 Results for the Marine Mammal data

This section gives the results of analyses for all marine mammal species combined. Marine mammals were more likely seen in low winds, calm seas, in good weather and in high tides.

The following variables were deemed predictors of relative marine mammal abundance in this area:

- Wind Strength
- Sea State
- Wind Direction
- Cloud Cover
- Precipitation
- Flow Speed
- Flow Direction
- Tide Height
- Time of day
- Grid Code
- Survey Month

Counts were universally zero in two of the grid codes located far from the observation point (Grid codes A0 and E5) and in rough conditions (a sea state of 6). Estimates of marine mammal numbers for these grid codes and in this sea state were therefore not possible.

Wind Strength: Less marine mammals were seen as conditions became progressively more windy.

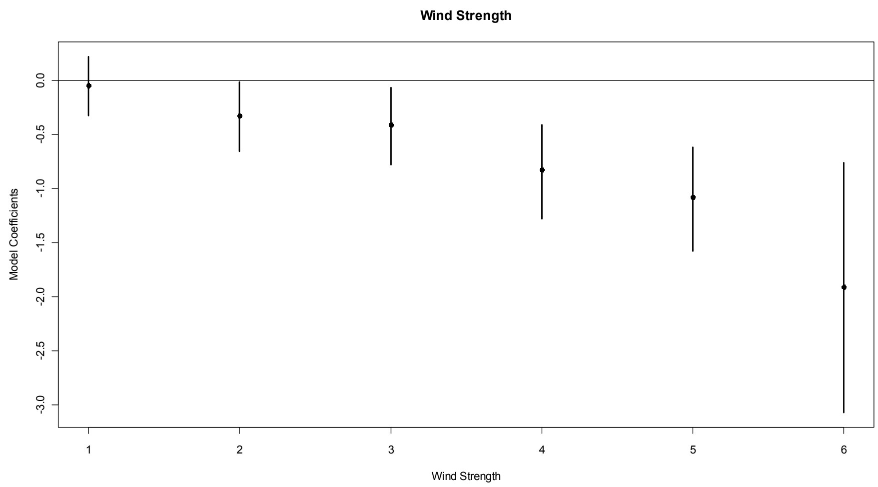

Figure 9: Relative effects of *Wind Strength* on estimated numbers of marine mammals. Dots indicate estimates, vertical lines the 95% uncertainty bounds. A wind strength of 0 is the baseline category (horizontal line at 0). The *p*-value for the GEE-based Wald test for this term is 0.00000 (5 d.p.).

Sea State: In general, more marine mammals were seen in calmer waters (when sea state 0 to 3). Mean estimates for sea states 4 & 5 appear to contradict this trend, but there is great uncertainty about these estimates and so are statistically indistinct from average animal numbers for any of the other sea states.

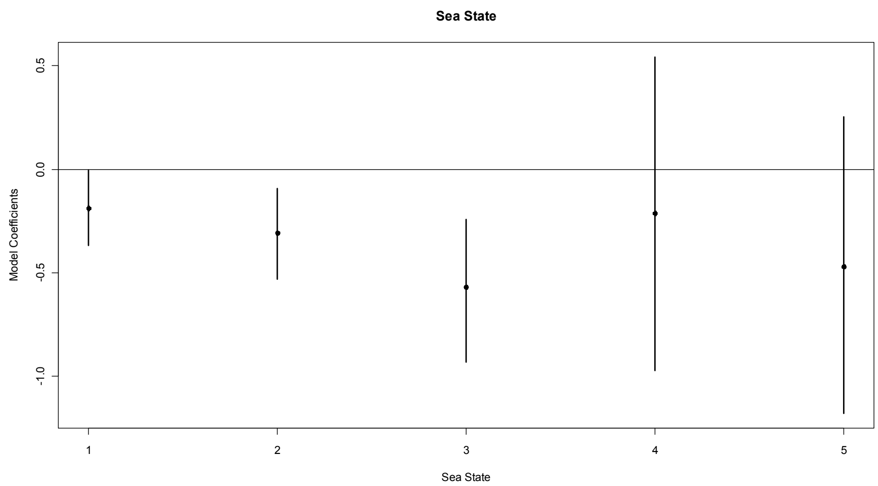

Figure 10: Relative effects of *Sea State* on estimated numbers of marine mammals. Dots indicate estimates, vertical lines the 95% uncertainty bounds. A Sea State of 0 is the baseline category (horizontal line at 0). The *p*-value for the GEE-based Wald test for this term is 0.01657 (5 d.p.).

Wind Direction: Very similar average numbers of marine mammals were seen in all wind categories however average marine mammal numbers in the North, North-East, South-East were all significantly higher than average numbers when the wind was coming from the East (baseline) direction.

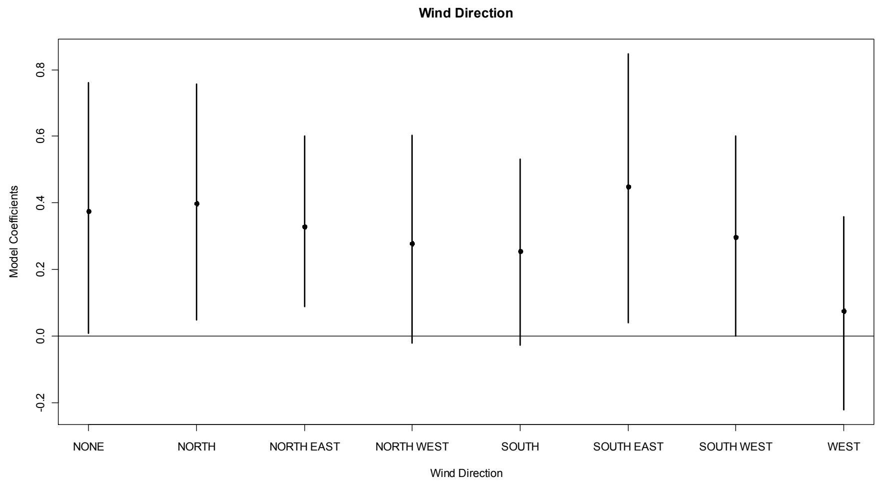

Figure 11: Relative effects of *Wind Direction* on estimated numbers of marine mammals. Dots indicate estimates, vertical lines the 95% uncertainty bounds. The 'East' wind direction is the baseline category (horizontal line at 0). The *p*-value for the GEE-based Wald test for this term is 0.00051 (5 d.p.).

Precipitation: Similar average numbers of marine mammals were seen in all precipitation categories however numbers were generally lower during rain and showers compared with when precipitation was recorded as 'None' (the baseline category).

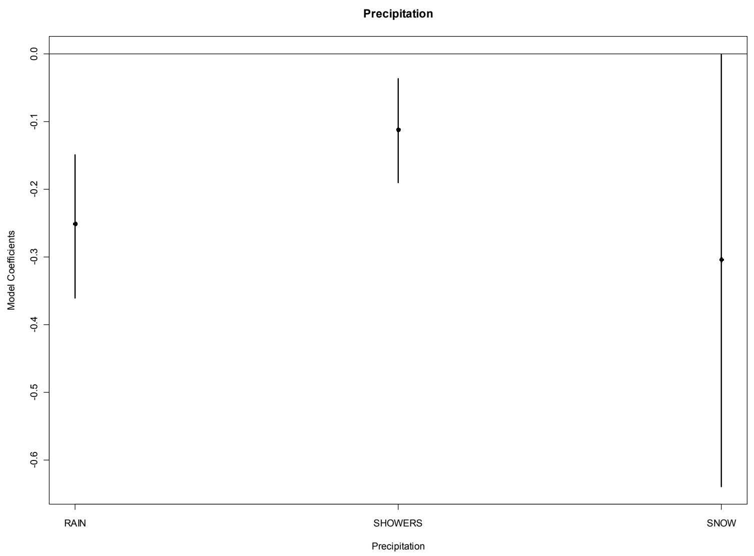

Figure 12: Relative effects of *Precipitation* on estimated numbers of marine mammals. Dots indicate estimates, vertical lines the 95% uncertainty bounds. 'None' is the baseline category (horizontal line at 0). The *p*-value for the GEE-based Wald test for this term is 0.00000 (5 d.p.).

Flow Speed: Average marine mammal numbers were generally lower when the water flow was slow compared to average numbers when flow speed was fast (the baseline category).

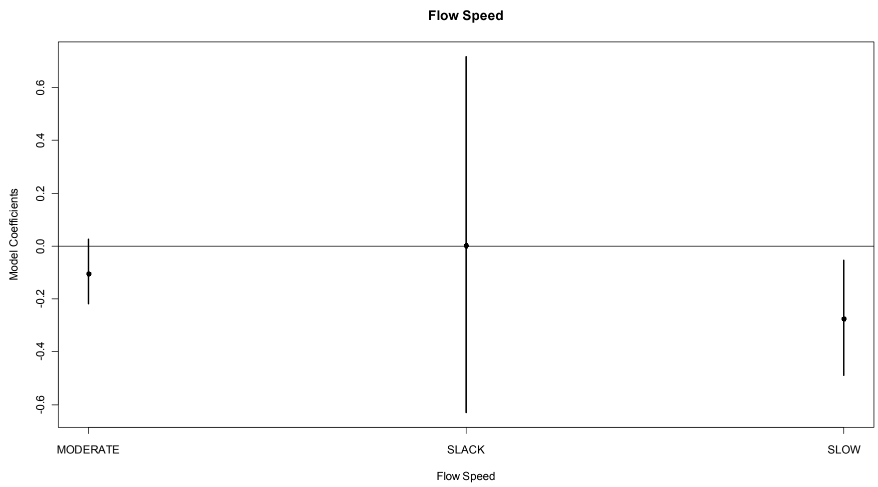

Figure 13: Relative effects of Water *Flow Speed* on estimated numbers of marine mammals. Dots indicate estimates, vertical lines the 95% uncertainty bounds. The 'Fast' flow speed is the baseline category (horizontal line at 0). The *p*-value for the GEE-based Wald test for this term is 0.00000 (5 d.p.).

Flow Direction: Average marine mammal numbers were similar across flow directions however average numbers were significantly lower when the flow direction was categorized as "South" compared to North-West (the baseline category).

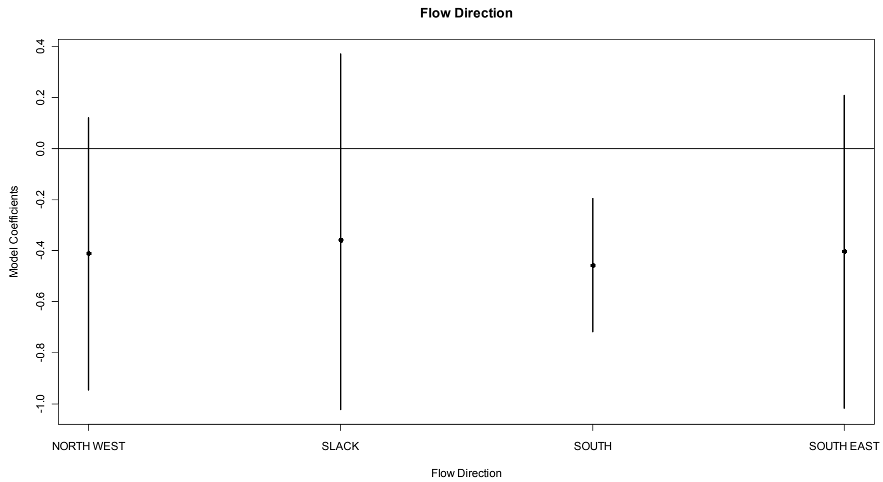

Figure 14: Relative effects of Water *Flow Direction* on estimated numbers of marine mammals. Dots indicate estimates, vertical lines the 95% uncertainty bounds. The 'North-West' flow direction is the baseline category (horizontal line at 0). The *p*-value for the GEE-based Wald test for this term is 0.00000 (5 d.p.).

Cloud Cover: Similar marine mammal numbers were seen during the full range of cloud cover. As for the bird data, there is a great deal of uncertainty about this relationship, so caution must be used when interpreting this curve.

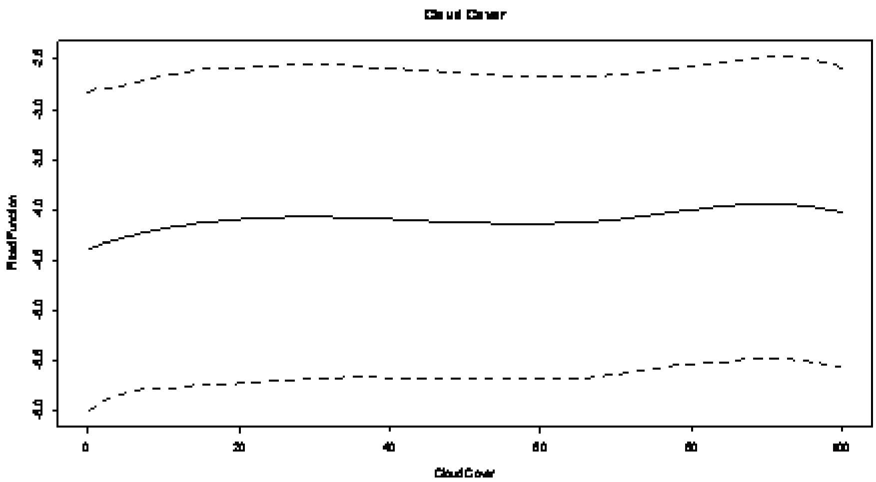

Figure 15: Effects of *Cloud Cover* on estimated numbers of marine mammals. Solid curve indicates estimated function, dotted curve above and below indicate the 95% confidence envelope. The *p*-value for the GEE-based Wald test for this term is 0.00022 (5 d.p.).

Tide Height: Higher average numbers of marine mammals were seen during higher tides, although there is great uncertainty about this relationship.

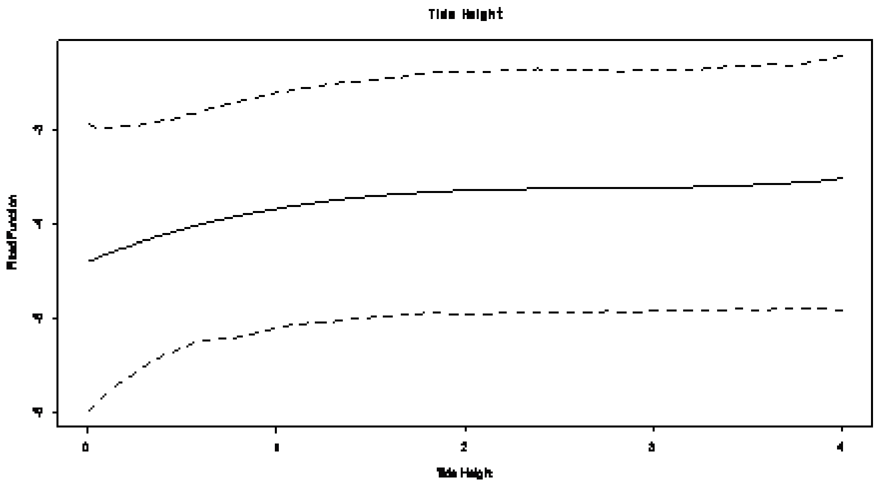

Figure 16: Effects of *Tide Height* on estimated numbers of marine mammals. Solid curve indicates estimated function, dotted curve above and below indicate the 95% confidence envelope. The *p*-value for the GEE-based Wald test for this term is 0.00438 (5 d.p.).

Time of day: Under the model, very similar numbers of marine mammals were observed throughout the day. There is also a great deal of uncertainty about this relationship.

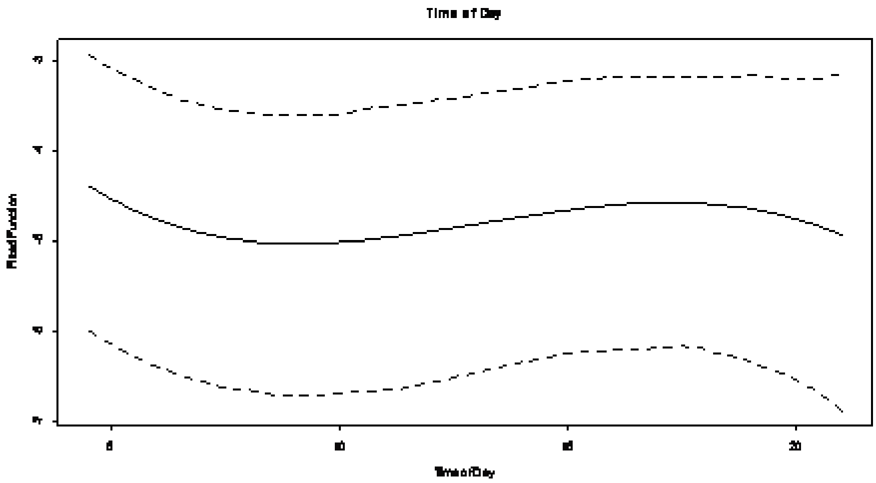

Figure 17: Effects of *Time of Day* on estimated numbers of marine mammals. Solid curve indicates estimated function, dotted curve above and below indicate the 95% confidence envelope. The *p*-value for the GEE-based Wald test for this term is 0. 00010 (5 d.p.).

Survey Month: Marine mammal numbers appeared to cycle seasonally with a consistent peak in numbers in September/October annually. Marine mammal numbers appear to be relatively stable across years.

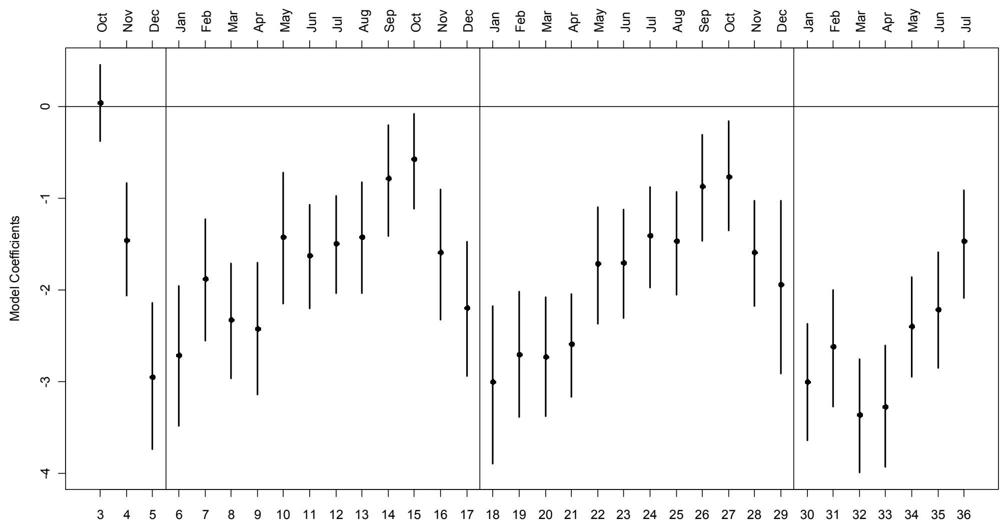

Figure 18: Relative effects of *Survey Month* on estimated numbers of marine mammals. The x-range represents September 2005 to July 2008. Dots indicate

estimates, vertical lines the 95% uncertainty bounds. Month 2 forms the baseline category for reasons stated for the bird data. The *p*-value for the GEE-based Wald test for this term is 0. 00000 (5 d.p.).

Grid Code: Grid codes generally gave significantly higher predicted marine mammal numbers than grid code *A1* (the baseline) which was statistically indistinct from areas A5, B5, C5 & D5. In keeping with previous analyses, the areas close to land (eg. *E0* to *E4,* and *A2* to *A4*) are predicted to exhibit more marine mammals than grid codes far from land.

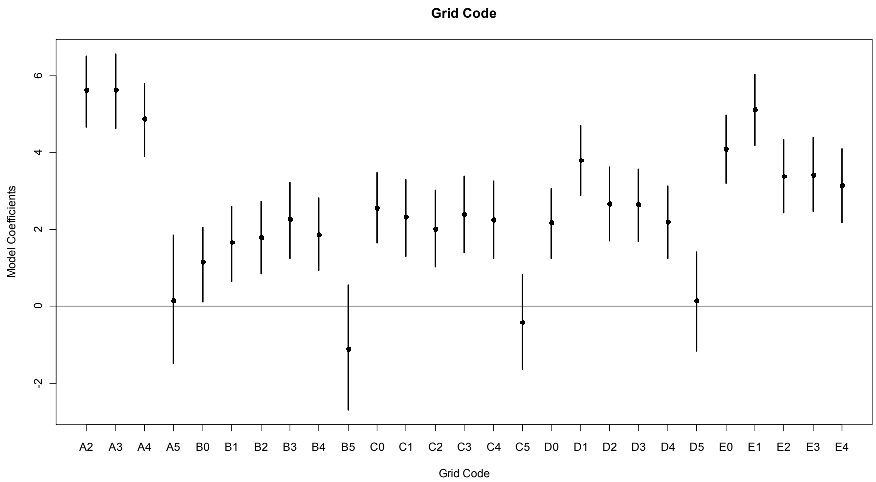

Figure 19: Relative effects of *Grid Code* on estimated numbers of marine mammals. Dots indicate estimates, vertical lines the 95% uncertainty bounds.

3.5 Reference Grid

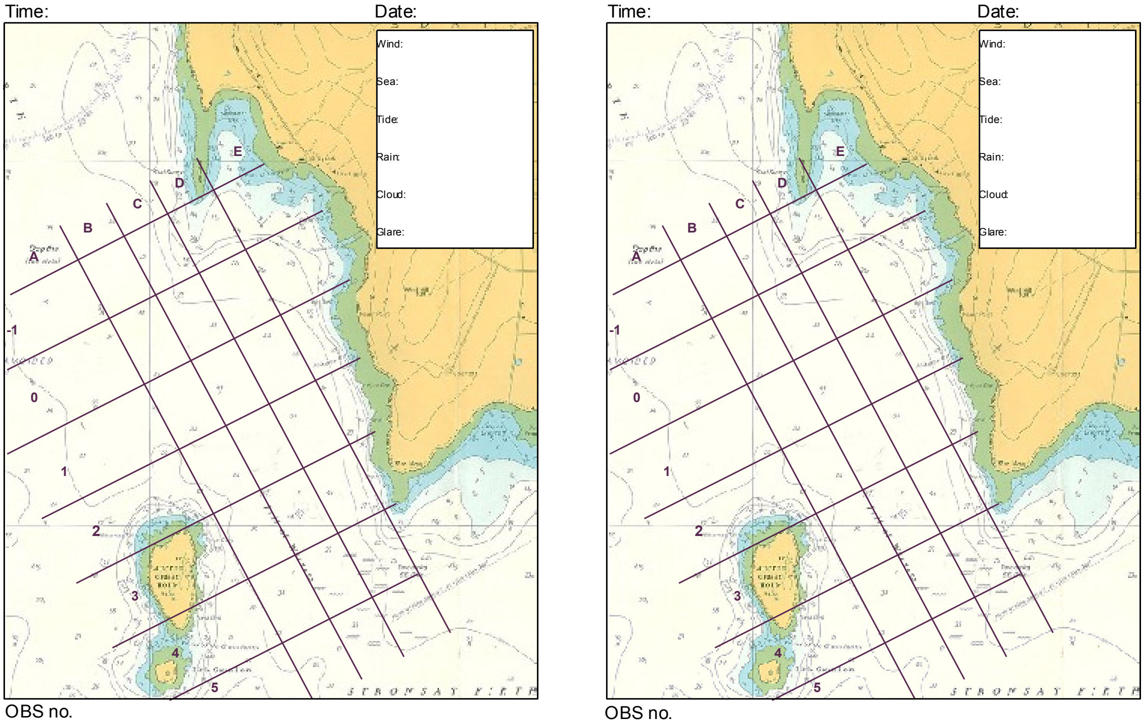

Figure 20: Eday survey grid as supplied by Aurora Environmental Limited.

4 Assessing the power to detect turbine effects on different time scales.

4.1 Overview

To assess the ability of the model to detect changes in bird numbers due to a 'turbine' effect over time a simulation exercise (based on current sampling effort) was used. Specifically, data were generated after different time periods using the existing observation process and the current model. The ability of these models for detecting a 'turbine' effect under these realities was measured.

The 'success rates' at detecting a turbine effect were determined for small, moderate and large turbine effects (causing a reduction in bird and marine mammal numbers by 5% to 30%) after a variety of post-impact monitoring periods: 1 to 6 and 12 months post (simulated) turbine installation/operation.

For rigour, the variability in the abundance data and the time dependent nature of the observation process was included in this simulation exercise; these factors can heavily dictate whether turbine effects are detected.

4.2 Technical details

Over-dispersed auto-correlated Poisson data was simulated using the parameter estimates (as parameter values) obtained for the model fitted to the pre-turbine data. Specifically, observations collected on the same day were assumed to have an AR(1) correlation structure (ρ=0.2 for both the bird and marine mammal data) and these errors were added on the scale of the link function. While inducing auto-correlation in this way is unlikely to be identical to the true auto-correlation in the data, the GEE approach used here is robust to misspecification of this type. The amount of overdispersion used to generate the data was chosen using the dispersion parameter estimate based on the fitted model (5.77 for the bird data and 1.93 for the marine mammal data).

A binary (intercept) term was used to simulate a turbine effect, and a range of turbine parameters were used to simulate a variety of turbine effects. Specifically turbine effects resulting in reductions of 5%, 10%, 20% and 30% of existing bird and marine mammal abundances were simulated, and the percentage of simulations which successfully detected a statistically significant turbine effect (at the 5% level) were determined for a range of monitoring periods post (simulated) turbine operation. The success of this modelling approach (given these data) at recovering the magnitude of the 'turbine effect' was also assessed using 95% confidence intervals based on GEE standard errors.

500 simulations were run for each combination of turbine effect and monitoring period post-installation/operation. For computational reasons, only survey month and turbine were used in the simulation approach. While in practice more covariates feature in each model (e.g. time of day), the estimates of the turbine effect and *p*-values were very similar under the small and larger models for the subset trialled.

When no turbine effect is present, this approach is expected to detect turbine effects approximately 5% of the time. This (Type I) error rate accompanies all statistical tests but can be reduced, if desired (1% is common). This reduction however increases the chance of a 'Type II' error which results in a genuine turbine effect being missed.

4.3 Results for the Bird Data

This approach was extremely successful (>90% success) at identifying a turbine effect after just one month when this event resulted in (at least) a 20% reduction in bird numbers. A monitoring period in excess of 3 months was required however to achieve this success for a 10% reduction in average bird numbers and approximately 12 months when the reduction in bird numbers was just 5%.

		Post-installation monitoring period (months)						
		1	2	3	4	5	6	12
Effect size	**0%**	7% (93%)	5% (95%)	6% (94%)	5% (95%)	6% (94%)	5% (95%)	8% (92%)
	5%	17% (93%)	30% (95%)	43% (92%)	50% (92%)	64% (93%)	71% (94%)	93% (94%)
	10%	54% (94%)	81% (94%)	94% (94%)	98% (94%)	100% (95%)	100% (94%)	100% (96%)
	20%	97% (91%)	100% (93%)	100% (95%)	100% (95%)	100% (95%)	100% (95%)	100% (94%)
	30%	100% (93%)	100% (94%)	100% (95%)	100% (95%)	100% (95%)	100% (95%)	100% (95%)

Table 2: Power to detect % abundance reductions (Effect size) for different post-installation monitoring periods. The cells shaded in grey indicate power for a simulated turbine effect is less than 90% for these combinations of monitoring period and effect size.

4.4 Results for the Marine Mammal Data

This approach was extremely successful (>90% success) at identifying a turbine effect after just one month when this event resulted in (at least) a 20% reduction in marine mammal numbers. However, a monitoring period of at least 12 months was required to achieve this success for a 5%-10% reduction in average marine mammal numbers.

		Post-installation monitoring period (months)						
		1	2	3	4	5	6	12
Effect size	**0%**	7% (93%)	7% (93%)	6% (94%)	5% (95%)	7% (93%)	5% (95%)	8% (92%)
	5%	17% (93%)	30% (95%)	43% (92%)	50% (92%)	64% (93%)	71% (94%)	93% (94%)
	10%	20% (94%)	49% (96%)	73% (94%)	82% (93%)	77% (93%)	84% (96%)	94% (95%)
	20%	97% (92%)	98% (95%)	100% (93%)	100% (94%)	100% (96%)	100% (95%)	100% (94%)
	30%	98% (94%)	100% (94%)	100% (95%)	100% (96%)	100% (96%)	100% (95%)	100% (96%)

Table 3: Power to detect % abundance reductions (Effect size) for different post-installation monitoring periods. The cells shaded in grey indicate power for a simulated turbine effect is less than 90% for these combinations of monitoring period and effect size.

5 Conclusions

This study provides information that may be used to determine bird and marine mammal abundances relatively *through time*. This can provide a basis for temporal monitoring by permitting comparisons across time.

Sound statistical models have been formulated to allow relative animal abundances to be predicted for each grid code under differing environmental conditions. Further, special attention has been paid when presenting best and worse case scenarios (using 95% confidence limits) to ensure comparisons made during this analysis (e.g. across time and grid code) and any future comparisons are reasonable.

The average number of birds and marine mammals sighted appears to differ across the grid codes sampled, with greater numbers detected in the sub-areas near land (both at the observation point and the off-shore island opposite). Bird and marine mammal abundances appear to be seasonal; however concrete confirmation of seasonal patterns can only be made by considering several years' worth of data.

There are clear relationships between monitoring times after turbine installation and the magnitude of the turbine effect. Longer monitoring times increase the probability of detecting a genuine effect and while large effects can be detected in short time frames, a monitoring programme of at least a year is required to detect smaller effect sizes (e.g. a reduction in average animal numbers of 5%).

However, it should also be noted that the ability to make such interpretations on the effects of tidal stream energy devices will be greatly enhanced by details of the accurate times and locations of the turbine operations within the Fall of Warness.

6 Limitations/Caveats

Absolute estimates of bird abundance are unable to be obtained from the data at hand. For instance, there may be the possibility of repeatedly counting the same individuals through time and this could seriously bias estimates of absolute abundance. However, relative abundance information can be extracted which is suitable for monitoring changes through time.

We continue to have concerns about the apparent relationship between the distance from the observer and observed bird abundance i.e. predicted bird abundance is highest near land. For instance it is well known that the probability of detecting an animal decreases as the distance from the observer increases and animals are more easily identified against a contrasting background. If detection primarily determines the number of birds recorded by the observer rather than the number of birds present, then this model will not adequately reflect underlying differences in grid-code to grid-code abundances[4]. That said, if the sampling and observer protocol and detection rates stay constant with time, valid comparisons can still be made across time using this approach, if the power to detect change is sufficiently high.

Cost-effective augmentation of the current survey design would go some ways to addressing these detection concerns. Specifically, additional observers placed at (randomly) chosen locations (on water or land) could provide animal counts concurrent with the observer already in place, and any bias in the current design could be objectively evaluated. This augmented design could be in place for a short period of time and used to correct for any biases which emerge; counts for affected grid codes could be inflated or deflated according to the extent and nature of the bias revealed.

[4] this is not an uncommon issue. If the detection function can be modelled then absolute abundance figures can be generated.

7 Appendices

7.1 Data cleaning phases

1. Tide height variable was very inconsistent in format. It contained tide height at high or low tide and an indicator of which tide state the height applies to. This is free-form text with an extremely large numbers of errors in recording. The variable was cleaned and the tide height (m) and tide state (Hi/Lo) components were extracted and, where necessary, interpolated.
2. On the basis of the data from step 1, functions were written that constructed sinusoidal predictions of actual tide height for every sighting.
3. Precipitation was entered as free-form text without a considered set of possible types. Consequently there are a number of errors in recording and numerous interpretations of precipitation state leading to 70+ precipitation categories. These were reduced in number by correcting entry errors and condensing similar classes.
4. Flow direction and speed were similarly affected by entry errors and excessive classes – these were corrected and reduced.
5. Zone codes were not consistently entered (e.g. numeric and character components were reversed) and subject to entry errors. These were corrected where the intended code was obvious.
6. Species names and codes were entered as free-form text and subject to entry errors and inconsistencies in format. These were corrected where the intended entry could be inferred.

7.2 Species ID data cleaning

The following gives the set of species names and codes prior to data cleaning (Table 4) and after (Table 5).

Species Name and Code	Frequency	Species Name and Code	Frequency
ARCTIC TERN:STA	64	LITTLE AUK :ALA	1
ARTIC TERN:STA	4	LITTLE AUK:ALA	5
BASKING SHARK:BAS	91	LONG-TAILED DUCK:CLH	26
BLACK GUILLEMOT:CEG	7910	LONG TAILED DUCK:CLH	525
BLACK GUILLEMOT:URA	1	MINKE WHALE:BAA	20
COMMOM SEAL:PHV	1	MINKIE WHALE:BAA	1
COMMON GUILLEMOT:URA	2268	NOTHING:	1
COMMON SCOTER:MEN	1	ORCA:ORO	1
COMMON SEAL:PHV	750	OTTER:LUL	9
CORMORANT:E4	1	PHALACROCORAX :PHS	20
CORMORANT:PHC	2890	PHALACROCORAX SPP:PHS	551
DIVER :GAV	2	PHALACROCORAX:PHS	2735
DIVER SPP:GAV	4	PUFFIN :FRA	142
DIVER:GA	4	PUFFIN:FRA	1732
DIVER:GAV	23	RAZORBILL:ALT	269
DIVER:GHV	1	RED-BREASTED MERGANSER:MES	1
EIDER DUCK:SOM	551	RED-THROATED DIVER:GAS	304
EIDER:S0M	4	RED THROATED DIVER:GAS	1
EIDER:SOM	3013	RED BREASTED MERGANSER:MES	241
GANNET:MOB	2651	RED THROATED DIVER :GAS	783
GOLDENEYE:BUC	46	RED THROATED DIVER:GAS	414
GREAT NORTHERN DIVER :GAI	210	RISSO'S DOLPHIN:GRG	1
GREAT NORTHERN DIVER:GAI	356	SEAL :SEA	2
GREAT NORTHERN DIVER:GAS	6	SEAL:SEA	873
GREAT NORTHERN DIVER:PHI	1	SHAG:PHA	5899
GREY SEAL:HAG	2237	SLAVONIAN GREBE:POA	6
GUILLEMOT:GA	1	UNIDENTIFIED CETACEAN:UNC	1
GUILLEMOT:URA	220	UNKNOWN BIRD:UNB	35
HARBOUR PORPOISE :PHP	39	UNKNOWN CETACEA:CET	1
HARBOUR PORPOISE:PHP	152	UNKNOWN CETACEAN:UNC	1
HARBOUR PORPOISE:PHP	3	WHITE-BEAKED DOLPHIN:LAA	6
KITTIWAKE:RIT	38	WHITE BEAKED DOLPHINS:LAA	1

Table 4: Species codes and names prior to data cleaning

Species Name and Code	Frequency	Species Name and Code	Frequency
ARCTIC TERN:STA	68	MINKE WHALE:BAA	21
BASKING SHARK:BAS	91	NOTHING:	1
BLACK GUILLEMOT:CEG	7911	ORCA:ORO	1
COMMON GUILLEMOT:URA	2489	OTTER:LUL	9
COMMON SCOTER:MEN	1	PHALACROCORAX:PHS	3306
COMMON SEAL:PHV	751	PUFFIN:FRA	1874
CORMORANT:PHC	2891	RAZORBILL:ALT	269
DIVER:GAV	34	RED-BREASTED MERGANSER:MES	1
EIDER:SOM	3568	RED BREASTED MERGANSER:MES	241
GANNET:MOB	2651	RED THROATED DIVER:GAS	1502
GOLDENEYE:BUC	46	RISSO'S DOLPHIN:GRG	1
GREAT NORTHERN DIVER:GAI	573	SEAL:SEA	875
GREY SEAL:HAG	2237	SHAG:PHA	5899
HARBOUR PORPOISE:PHP	194	SLAVONIAN GREBE:POA	6
KITTIWAKE:RIT	38	UNIDENTIFIED CETACEAN:UNC	3
LITTLE AUK:ALA	6	UNKNOWN BIRD:UNB	35
LONG TAILED DUCK:CLH	551	WHITE-BEAKED DOLPHIN:LAA	7

Table 5: Species codes and name after data cleaning

8 References

DMP Statistical Solutions UK Ltd. (2007). *Analysis of Bird Data for the Fall of Warness Area: A Report for prepared SMRU Limited*

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*. New York, Chapman & Hall.

Liang, K.-Y., and Zeger, S. L. (1986). "Longitudinal data analysis using Generalized Linear Models." *Biometrika* **73**(1): 13-22.

Pan, W. (2001). "Akaike's Information Criterion in Generalized Estimating Equations." *Biometrics* **57**(1): 120-125.

R Foundation for Statistical Computing (2008) *R: A Language and Environment for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0, www.R-project.org. Version 2.8