JNCC Report
No: 482

A comparison of different techniques for mapping cetacean habitats

Booth, C.G. & Hammond, P.S.

July 2014

ISSN 0963 8091

For further information please contact:

Joint Nature Conservation Committee
Monkstone House
City Road
Peterborough PE1 1JY
www.jncc.defra.gov.uk

This report should be cited as:

Booth, C.G. & Hammond, P.S. 2014. A comparison of different techniques for mapping cetacean habitats. *JNCC Report No 482.*

Summary

An extensive visual and acoustic dataset on harbour porpoises west of Scotland, collected between 2003 and 2008, was used as a test bed for investigating the effectiveness of various analytical methods to provide information on habitat use. Methods used were geospatial analysis (kriging), presence-only analysis (MAXENT), Generalised Linear Modelling (GLM), Generalised Additive Modelling (GAM) and Generalised Estimating Equations (GEEs). The full dataset (visual and acoustic) was subsetted to generate data that were more sparse and had an uneven spatial distribution of effort.

Results showed that GLM, GAM and GEE generated broadly similar results but with much more variability in the predictions from the smaller subsetted datasets. Geospatial analysis generated much more patchy distribution maps and was sensitive to the high percentage of zeroes in the data; it was unable to generate results for the subsetted datasets. Presence-only analysis generated much coarser distribution maps and generally predicted larger areas of high-use habitat.

Overall, the geospatial method worked least well. The presence-only, GLM, GAM and GEE methods all generated useable maps. Presence-only methods are subject to bias if effort is not distributed representatively across the area of interest and may generate misleading results in such cases. In addition, the results were much coarser and may not show sufficient detail for conservation and management purposes. The GLM, GAM and GEE methods incorporate the effects of effort and gave results at a finer scale but also with considerable variability, especially for the smaller subsetted data. The results from GEE were similar to those from GLM and GAM.

Overall, effort-based methods of analysis (particularly GLM or GAM) should be preferred to reduce bias and provide a greater level of detail. However, it is important that measures of precision are also used to inform what inferences can be made from the results.

Contents

1 Introduction 1

2 Methods 2

2.1 Data collection 2

2.2 Data Processing and Analysis 2

2.2.1 Geospatial modelling (kriging) 4

2.2.2 Presence-only modelling (MAXENT) 5

2.2.3 Statistical inference modelling (GLM, GAM, GEE) 5

3 Results 7

3.1 Survey data 7

3.2 Full datasets 8

3.2.1 Acoustic 8

3.2.2 Visual 8

3.2.3 Subset 1 9

3.2.4 Subset 2 10

3.2.5 Areas covered 10

4 Discussion 10

4.1 Geospatial methods 10

4.2 Presence-only (MAXENT) models 11

4.3 Statistical Inference Models 12

4.4 Analysis of other datasets 12

4.5 Further Analysis 12

5 References 14

1 Introduction

Article 11 of the Habitats Directive places an obligation for Member States to undertake surveillance on the conservation status of all cetacean species occurring in their waters and to report on this every six years. The aim of the Habitats Directive is that listed species and habitats achieve and maintain a favourable conservation status (FCS). The FCS, as defined by the Habitats Directive, is measured primarily by assessing changes in three elements: natural range, population size and habitat.

Understanding how cetaceans utilize the biotic and abiotic resources that are potentially available to them (their habitat) is complex, partly because of the complexity of the usage itself. An individual cetacean must find suitable resources for various aspects of reproduction, feeding and avoiding predators/competition; the habitat requirements for these functions may be quite different. The distribution and habitat use of species with complex ecology are likely to be influenced by characteristics of the animals themselves, such as reproductive status and foraging strategies, as well as by abiotic and biotic features of the marine environment.

Describing species-habitat relationships is also complex because the study of cetacean ecology poses particular problems of access and scale. Individuals spend a substantial majority of their time underwater but are typically available to the researcher in only two dimensions at the surface, although the development of towed hydrophone arrays is proving very useful for the study of some species of odontocete (Gillespie *et al* 2008). Populations of many species (e.g. harbour porpoise, large whales, oceanic delphinids) may be large and range over wide areas making data collection logistically and financially challenging (SCANS-II 2008; CODA 2009). Conversely, some populations are small (e.g. coastal bottlenose dolphins; vaquita) and this presents problems of obtaining sufficient data for analysis (Wilson *et al* 1999; Jaramillo-Legorreta *et al* 2007). It is these small populations that are often the most threatened and thus in need of good information to guide conservation efforts.

In the 2007 FCS assessments, the habitat element was incompletely reported on because of the acknowledged difficulties associated with defining habitat for cetaceans. Where assessments were made, the judgement of Favourable was based on the relatively high spatial and temporal variability in the behaviour and ecology of all cetaceans. Additionally, where range and/or population is considered to be in a Favourable condition, it has been assumed that habitat must also be considered to be Favourable. Relatively recent and ongoing developments in analytical techniques (e.g. Redfern *et al* 2006; Matthiopoulos & Aarts 2010) as well as the availability of additional datasets (e.g. SCANS-II 2008; CODA 2009; Marubini *et al* 2009; Embling *et al* 2010) mean that at the next FCS reporting round in 2013, improved assessments of the habitat element of FCS should be possible.

Member States are required to report on the FCS of all cetacean species. As a test case, this project uses harbour porpoise datasets to compare different statistical methods for modelling distribution patterns: geospatial, presence only, Generalised Linear Models (GLMs), Generalised Additive Models (GAMs) and Generalised Estimating Equations (GEEs). The aim of this work was to investigate the spatial predictions generated by each statistical method. The species-habitat relationships underlying those distributions, while interesting, were not the primary focus of this work. How each statistical method performed under a range of effort scenarios (extensive vs relatively sparse effort; evenly spread vs uneven effort) was investigated.

The aim here was that these effort scenarios would provide some indication of how the models coped with relatively small, large, even and uneven-spread effort. However, some datasets for other species are much smaller and further work is required to explore the extent to which statistical modelling can help describe the habitat use of such species. The intention here is that the findings of this pilot study will provide the first stage in informing the distribution/habitat modelling of cetacean species in UK waters, to build towards providing the best information for FCS reporting in 2013.

2 Methods

2.1 Data collection

The data used in analysis were visual and acoustic detections of the harbour porpoise (*Phocoena phocoena*) recorded during systematic line transect surveys carried out from the Hebridean Whale and Dolphin Trust's (HWDT) 18m motor-sailer vessel *Silurian*. Surveys were conducted off the west coast of Scotland (55° 10' - 58° 40' N, 5° 0' - 8° 35' W; Figure 1), between April and September (inclusive) during daylight hours. Visual surveys were carried out from 2003 to 2008 and towed-hydrophone acoustic surveys were conducted simultaneously during the 2004 to 2008 seasons.

Visual observations were carried out by teams of two trained observers, from the front deck of the vessel, (2m above water level) surveying one side each. When cetaceans were sighted, species, group size and other data were entered into the data recording software *Logger 2000* (Gillespie *et al* in press) which ran continuously, logging GPS positional and NMEA feed data, and stored in a Microsoft Access database in real-time. Passive acoustic monitoring (PAM) was conducted using a towed hydrophone array from 2004 to 2008 in all sea conditions during daylight hours in waters >10m depth. The hydrophone array consisted of two high frequency elements with highest sensitivity at 150kHz and a near flat frequency response between 2-140kHz. Elements were housed in a streamlined sensor section consisting of 10m of 35 mm diameter polyurethane tubes filled with ISOPAR-M oil, which was towed 100 m behind the boat attached by Kevlar-strengthened cable.

The signal was fed into porpoise detection software Porpoise Detector (2004-2005) or Rainbow Click (2006-2008) (Gillespie & Chappell 2002; Gillespie *et al* in prep.). These programs automatically classified sounds detected on the hydrophone array, identified detections of harbour porpoises and calculated a bearing (no left/right discrimination) to the source based on the difference in arrival time at each element allowing for estimation of the number of animals echolocating in each detection event. Encounters were identified and the porpoise clicks in each click train were marked and porpoise clicks were arranged into 'groups' with an estimated number of porpoises in each event. Each porpoise group was linked to a GPS position from *Logger 2000* using a pre-written macro (Gillespie, pers. comm).

2.2 Data Processing and Analysis

The visual and acoustic datasets collected were investigated using five analytical methods: geospatial models (Kriging), presence-only (MAXENT) models, Generalised Linear Models (GLMs), Generalised Additive Models (GAMs) and Generalised Estimating Equations (GEEs).

In the absence of explanatory covariate data, a geospatial density-estimation based method can be used to estimate spatial distribution (Matthiopoulos & Aarts 2010). The most rigorous of these interpolation methods is Kriging (Isaaks & Srivastava 1990).

Presence-only models exist as an alternative to presence-absence methods in situations where reliable sampling (searching) effort data do not exist and it is not possible to determine absences. There are a wide range of presence-only models (see Pearson 2007 for a summary) and one commonly used presence-only approach is Maximum Entropy (MAXENT) (Elith *et al* 2011). Another presence-only method that has been used to describe species distributions is Ecological Niche Factor Analysis (ENFA) (Fernandez *et al* 2009; MacLeod *et al* 2008; Praca & Gannier 2008; Skov *et al* 2008). MAXENT can be considered a more robust approach and we use it here.

GLMs and GAMs have been used in a number of cetacean modelling studies (e.g. Bailey & Thompson 2009; Cañadas & Hammond 2008; Cañadas *et al* 2005; Embling *et al* 2010). One issue in the modelling of survey data, such as those used here, is that data collected close together in time and space may lead to autocorrelation. Not accounting for autocorrelation in modelling can result in standard errors of fitted coefficients being underestimated and more variables than necessary being retained in models. This is known as over-fitting and can lead to potentially misleading models (Lennon 2000).

One way to deal with spatial and temporal autocorrelation is to include autocorrelation structure in the model itself, such as in Generalised Linear Mixed Models (GLMMs), Generalised Additive Mixed Models (GAMMs) and Generalised Estimating Equations (GEEs) (e.g. Ballinger 2004; Hardin & Hilbe 2002). One study of cetacean distribution has shown an improvement in model selection over a GAM unadjusted for autocorrelation (Panigada *et al* 2008). These models have also been used to study harbour porpoise habitat preferences and distribution (Booth 2010). In this study we chose to use GLMs and GAMs with no autocorrelation structure and to investigate the effect of incorporating autocorrelation using GEEs.

In the construction of these models, a range of potential explanatory covariates (Table 1) was used to try to explain harbour porpoise distribution. When constructing the presence-only (MAXENT) models, only the spatially-varying covariates (i.e. seabed depth, seabed slope, distance from land, maximum spring tidal range and the percentage of sand, mud and gravel in the sediment) were included.

Survey effort track lines were divided into 2km segments, equivalent to the coarsest resolution of the available oceanographic covariates in the models and the total number of porpoise detections in each segment was calculated. Prior to segmenting, values for covariates were calculated for each GPS data point of track line. Visual and acoustic data were analysed separately to allow for differences in data collection methods to be incorporated in models and to investigate differences in results. Survey effort was limited to data collected in sea conditions Beaufort ≤ 2 for the visual data models. For the acoustic survey effort, data collected in all sea conditions were used.

Spatially uneven data collection can result in some areas being surveyed more often leading to unrepresentative spatial coverage of the data. This in common in cetacean studies and it may therefore be instructive to investigate the effects of small and uneven datasets on predicted distributions.

To investigate the impact of different effort scenarios on predicted distributions, the full visual and acoustic datasets were both subdivided. Two reduced effort scenarios were constructed incorporating different levels of effort and spatial unevenness. To generate the subsets, effort and detection data were partitioned using the survey day and location as the sampling unit. Surveys days were pseudo-randomly selected so that, in subset 1, effort collected on days with a low mean latitude (i.e. more southerly) were more likely to be retained than on days with a high mean latitude (i.e. more northerly). In subset 2, the opposite was the case, with a bias forcing almost all survey effort from northerly areas such as the Minch to be included, but very little effort being retained from some of the southern regions, e.g. the Sound of Jura.

Co-linearity between explanatory covariates, if unaccounted for, can cause inflated or underestimated standard errors and *p*-values leading to poor model selection and affecting the resultant predicted distributions. To avoid this, co-linearity between covariates was investigated prior to modelling using 'generalised variance inflation factors' (GVIF) (Cox & Snell 1989; Fox & Monette 1992) implemented through the *vif* function in the *car* package in R (R Core Development Team 2009). Only low co-linearity was found to exist between covariates (all GVIF scores were <5) indicating it was reasonable to include them all in the model selection phase of modelling.

2.2.1 Geospatial modelling (kriging)

Kriging generates estimates of a variable that are distance-weighted combinations of other observations/measurements. A semi-variogram assesses and quantifies how similarity between observations/measurements changes as a function of distance between points.

A disadvantage of this method in investigating the spatial distribution of marine mammal species, is the assumption that the data are normally distributed, which presence/absence or count data usually are not. There are methods of kriging that do not require this assumption but these can be computationally expensive (Matthiopoulos & Aarts 2010) and we did not implement them.

Kriging was used to model harbour porpoise distribution by interpolating relative densities (animals detected per km of effort). Survey effort and detection data were partitioned using a 4 x 4km grid in Manifold (Version 8.00. 32-bit, Manifold® Systems) and the total survey effort and number of animals detected in each grid cell was determined. Detections per unit effort (DPUE, animals per km of survey effort) were then calculated for each grid cell. Only grid cells with at least 2km of survey effort were included in the analysis.

DPUE values were interpolated in two-dimensions by kriging within Manifold. The model used to perform the kriging in each scenario was determined using the 'Automatic' setting, allowing Manifold to choose the best fitting model in each case.

2.2.2 Presence-only modelling (MAXENT)

Maximum Entropy (MAXENT) involves using a machine-learning algorithm to compare the covariate distribution where a species has been identified (presences) with the overall covariate distribution in the study area, thus identifying the types of environment where a species is known to occur (Phillips *et al* 2004, 2006; Elith *et al* 2011). The advantage of this method is that it allows the use of data where it is not possible to record absences. However, not incorporating effort data into models can lead to misleading conclusions about model results if effort is not representative of the study area. If some areas are sampled more regularly than others, this may lead to an over-exaggeration of the importance of the habitat that has been repeatedly sampled.

To generate presence-only data, effort data were discarded and the study area was gridded as with the geospatial model construction. Whether or not a presence was recorded in each grid cell was determined. A grid for each of the spatially-varying environmental variables was also created and these were fed into the MAXENT program. As part of the modelling process, a distribution map based on those presence data and the environment was generated.

2.2.3 Statistical inference modelling (GLM, GAM, GEE)

Generalized Linear Modelling (GLM), Generalized Additive Modelling (GAM) and Generalized Estimating Equations (GEE) are regression methods that predict a response variable as a function of explanatory covariates.

Here, the number of harbour porpoises detected per 2 km segment of survey effort was modelled as a function of survey and environmental explanatory covariates.

The statistical package R (64-bit Mac version 2.9.0, R Core Development Team 2009) was used for implementation, including the *mgcv* (GAM) (Wood 2006), *splines* and *geepack* (GEE) (Halekoh *et al* 2006) packages.

2.2.3.1. Generalized Linear Models (GLMs)

A GLM is a linear regression that allows for relaxation of the assumption of a normal distribution for the error structure of the response variable. GLMs have been used extensively to investigate species distribution patterns (e.g. Cañadas *et al* 2005). GLMs have the general form:

$$g(E(Y_i)) = \beta_0 + \beta_1 x_{1i} +,...,+ \beta_p x_{pi}$$

Where $E(Y_i)$ is the expected value of the response variable, $g(.)$ is the function linking the response to the fitted functions of the covariates x, β_0 is the intercept and β_p is the slope of the term x_p. Because of over-dispersion in the data, a quasi-Poisson error structure was chosen for the response variable (number of porpoises per 2km of survey effort) fitted with a log link function. Negative binomial and zero-inflated models were also considered but not implemented.

GLMs assume that, when transformed, there is a linear relationship between the response and the covariates. Here, polynomial (quadratic) transformations of the covariates were used to capture any additional non-linearity in the relationships between the response and the explanatory covariates. Model selection was conducted using a forwards-backwards stepwise selection method. The 'best' model is determined by moving both forwards (starting from an intercept-only model and adding variables that improve the model) and backwards (removing variables from the full model one by one) dropping and/or adding covariates through a series of steps until a final model (a model where no further changes improve the model) is reached.

2.2.3.2. Generalized Additive Models (GAMs)

GAMs are increasingly being used in modelling marine mammal distributions and investigating habitat preferences (Bailey & Thompson 2009; Cañadas & Hammond 2006: 2008; Embling *et al* 2010; Marubini *et al* 2009; SCANS-II 2008; CODA 2009). GAMs use combinations of non-linear smooth functions of explanatory covariates to predict response variables that are normally or non-normally distributed and have the general form:

$$g(\mathrm{E}(Y_i)) = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + \ldots$$

where $E(Y_i)$ is the expected value of the response variable, $g(.)$ is the function linking the response to the non-linear smooth functions s_j of the covariates x_k and β_0 is the intercept term. Because of over-dispersion in the data, a quasi-Poisson error structure was chosen for the response variable (number of porpoises per 2 km of survey effort) fitted with a log link function. Negative binomial and zero-inflated models were also considered but not implemented.

One advantage of using GAMs over GLMs is that it is not necessary to specify additional explanatory covariates to allow for non-linearity in modelled relationships.

In order to reduce over-fitting of the smooth functions to the data, when constructing the models α was set to 1.4 as recommended by Kim and Gu, (2004) (α is a constant multiplier which can be used to inflate the model degrees of freedom). By fixing it at 1.4, the GAM is penalized if it uses too many degrees of freedom (Wood 2006). Thin-plate regression shrinkage splines were used to govern model selection when constructing the GAMs. These splines allow smooth functions to be reduced to zero in the models, indicating there is no need for those covariates in the model (Wood 2006). Therefore, when constructing the models using these splines, any terms that should not be retained in the final model appear as a horizontal line in the residual plots. Any terms appearing in the model selection residual plots as a horizontal line were removed from the model and it was run again until no such terms remained in the model.

2.2.3.3. Generalized Estimating Equations (GEEs)

The final statistical method used here was to use Generalized Additive Models (GAMs) built within a Generalized Estimating Equations (GEEs) model construct. GEEs can perform better than GAMs when data are temporally or spatially autocorrelated (Panigada *et al* 2008; Booth 2010). Such autocorrelation violates the assumption that the model errors are independent and, if unaccounted for, can result in covariates being wrongly retained in the final model, potentially affecting predicted distributions. Here, GEEs were used to generate the standard errors and *p*-values which are used to govern model selection (Liang & Zeger 1986) to investigate whether accounting for autocorrelation significantly impacted model predicted distributions.

The over-dispersion in the data means that GEE model fitting must be based on quasi-likelihoods, so stepwise model selection was based on the QIC statistic (Ballinger 2004). Autocorrelation function plots (using *acf* in R) were used to determine the panel size to be used in the models and the models were fitted with an 'independent' correlation structure. For all models, each covariate was permitted to be present in the model as a curve (with a *B*-spline (deBoor 1978) fitted with knots placed at the mean for each covariate), as a linear term or removed from the model. Year and month were fitted as factor variables.

The full model was fitted using the *geeglm* function in the geepack package (Halekoh *et al* 2006) and GEE-based *p*-values were used to determine if covariates should stay in the model. The function *anova.geeglm* in the geepack package performs stepwise selection using QIC but, as this will only add terms sequentially and the model selection results depended on the order that covariates were inputted, it was necessary to identify a suitable input ordering for the covariates. Reduced models were therefore created, each one of which had one covariate omitted. Each of these models was then compared to the relevant full model (containing all the covariates) using a simple ANOVA (*anova.glm*) method, to determine if each covariate was important in explaining that dataset. The 'important' terms were then fitted in order of significance and investigated using the sequential *anova.geeglm* to determine the final 'best' model.

2.3. Prediction and Kriging

Once the final model was constructed for each method, predictions were made over the same 4 x 4km grid used for kriging. For the GLM, GAM and GEE models, this resolution was chosen as it is twice the segment length used in the models as recommended by Hedley (2000).

The predicted distributions from each statistical technique (kriging, MAXENT, GLM, GAM, GEE) for each dataset (visual and acoustic, full, subset 1 and subset 2) were then contoured to facilitate interpretation of the results. To highlight the most important regions (areas of highest predicted density), contours were selected to represent the 50th, 60th, 70th, 80th and 90th percentiles of densities (or habitat suitability in the case of MAXENT) from each model and scenario. The area covered by each percentile band was calculated in Manifold.

3 Results

3.1 Survey data

In total 38,708 segments were included in analysis (visual: 17,353; acoustic: 21,355) corresponding to 34,699km of visual survey effort from 2003-2008 (in Beaufort sea state ≤ 3) and 42,653km of acoustic survey effort 2004-2008 (in all sea states). Full details are shown below (Table 2). In total, 2,381 harbour porpoises were detected visually in Beaufort ≤ 2 (0.069 animals per km) and 4,927 acoustic detections were made in all sea conditions (0.12 detections per km). Porpoise detections were generally most common in regions close to shore (Figure 2 a & d). Data subset 1 consisted of 2,994km of visual effort and 5,637km of acoustic effort with 238 (0.079 animals per km) and 786 (0.14 detections per km) detections, respectively (Fig 2 b, e). Subset 2 was larger, with 4,487km of visual effort and 6,907km of acoustic effort with 276 visual sightings (0.062 animals per km) and 859 acoustic detections (0.12 detections per km), respectively (Fig 2 c, f).

3.2 Full datasets

The full visual and acoustic datasets were modelled using the five statistical techniques outlined above. The retained explanatory variables for the final models are shown in Table 3(a). The predicted distributions are shown in Figure 3.

3.2.1 Acoustic

The GLM, GAM and GEE models constructed using the full acoustic dataset predicted very similar distribution patterns. The GAM explained about 10% of the variability in the data. All of the models predicted a strongly inshore distribution for harbour porpoises throughout the west coast of Scotland. The highest relative densities (> the 80th percentile) were predicted in the northern Sound of Jura, northeast Firth of Lorn, within the Sound of Mull, around the Isle of Skye and throughout the Small Isles (see Figure 1 for location names). Additionally, there were high-predicted relative densities along the east coast of the Outer Hebrides, throughout the Little Minch (between Skye and the Outer Hebrides) and within the more coastal reaches of the Minch. Low relative densities were predicted in the southwest part of the study region and to the west of the Outer Hebrides. The areas covered by the most important regions were similar for each of the modelling techniques (Figure 4 and Table 1 in Appendix 1). Each of these techniques had areas including the highest 50% of relative densities covering 21,052 - 22,729km^2. The highest density areas covered 8,909 - 9,580km^2 for the three statistical inference techniques.

The predicted distribution generated using the geospatial method looked much more patchy than those generated using the statistical inference methods, but the highest relative densities of harbour porpoises were predicted in some of the same areas. High (> the 80th percentile) relative densities were predicted in the Small Isles, Sound of Jura and the Little Minch in particular. The area covered by each of the percentiles was smaller than all the other modelling techniques used. The 50th percentile area was 19,953km^2 and the highest density area was 2,676km^2, which is approximately one quarter the area predicted using statistical inference methods.

The presence-only model predicted that most of the west coast of Scotland was suitable habitat for harbour porpoises. The regions containing the highest habitat suitability were similar to those from the statistical inference methods, however, the size of the area was much larger (30,918km^2) and almost the entire study area was contained within the >50th percentile of values. The highest density >80th percentile area was also larger (12,364km^2) than predicted using any other method.

3.2.2 Visual

The predicted distributions for the visual models constructed using the GLM, GAM and GEE models were very similar to the distribution patterns from the acoustic models. The GAM explained about 20% of the variability in the data. These models predicted a strongly inshore distribution for harbour porpoises throughout the study region. The highest relative densities were predicted in the northern Sound of Jura, northeast Firth of Lorn, within the Sound of Mull, around the Isle of Skye and throughout the Small Isles. Additionally, there were high predicted relative densities along the east coast of the Outer Hebrides, between Skye and the Outer Hebrides and in the coastal mainland areas of the Minch. Low relative densities were predicted in the southwest part of the study region and to the west of the Outer Hebrides islands. There appear to be some boundary issues ("edge effects") in the predicted output, with some high predicted relative densities apparent at the edge of the study region where there was little effort. The areas covered by the highest relative density regions were similar for each of the modelling techniques (Figure 5 and Table 1 in Appendix 1). The highest 50% of relative densities covered 21,635 - 25,723km^2 and the highest density areas (>80th percentile) covered 9,016 - 10,010km^2.

The predicted distribution generated using the geospatial method looked very different to those patterns generated using the statistical inference methods/. Because of the large number of zeroes in the visual dataset it was not possible to show the >50th - >70th percentiles in the visual geospatial distribution because the 50th, 60th and 70th percentiles were equal to 0. Highest (> the 80th percentile) relative densities were predicted around the Minch and the Isle of Skye and in patches throughout the Small Isles, Sound of Jura and the Little Minch in particular. The area covered by the >80th percentile was similar to the statistical inference methods but the >90th percentile areas was much smaller (2,694km^2) (Figure 5).

As with the acoustic full model, the presence-only (MAXENT) model predicted that most of the west coast of Scotland was suitable habitat for harbour porpoises. The regions containing the highest habitat suitability were again similar to those from the statistical inference methods but again the size of the area was much higher (30,918km^2) and almost the entire study area was contained within the >50th percentile of values.

The subsets of the survey effort were designed to provide smaller datasets upon which to construct distribution models. Subset 1 had effort skewed to the south of the study area and subset 2 was skewed to have more effort in the north. Because of the small sample sizes in both subsets and the large number of segments with effort but no detections (zeroes), it was not possible to construct geospatial models for these datasets. Only presence-only (MAXENT), GLM, GAM and GEE models were constructed.

3.2.3 Subset 1

The retained explanatory variables for the final models are shown in Table 3(b). The predicted distributions are shown in Figure 6. For the visual and acoustic subset 1 models, the GAM explained 28% and 17% of the variability in the data, respectively, and there was little consistency between the predicted distributions generated using the different statistical methods. However, for each method, the models constructed on visual and acoustic subsetted data yielded similar distribution patterns. The presence-only model predicted a distribution similar to those using from the full dataset presence-only models. In these, high-suitability areas were predicted throughout the inshore regions of the west coast of Scotland.

The visual and acoustic GLMs predicted a very patchy distribution pattern, with the highest density areas found in the Little Minch, Small Isles and Sound of Jura. However, many regions that were high density in the full models, were low (<50th percentile) in the GLMs.

The visual and acoustic GAMs predicted high density areas to be scattered throughout the Minch, Small Isles and in the Sound of Jura. GEE models for both visual and acoustic data indicated larger high density areas around the Isle of Skye and in the very southern extent of the study region.

3.2.4 Subset 2

The retained explanatory variables for the final models are shown in Table 3(c). The predicted distributions are shown in Figure 7. Data subset 2 had a larger amount of effort (and visual and acoustic detections) than subset 1 and the model distribution patterns appeared more consistent across the different techniques used here. The GAMs explained 34% and 17% of the variability in the visual and acoustic data, respectively. The distribution patterns from the GLM, GAM and GEE acoustic models appear very similar. High density areas were predicted around the Minch, Little Minch, Small Isles, in the northeast of the Firth of Lorn and throughout the Sound of Jura.

Results from the presence only model constructed from the acoustic data were similar to those from the full datasets, but there was only a small area of 'high suitability' in the south of Jura, where there was little effort.

In the visual models, there was more variation between the predicted distributions produced using the different techniques. The GLM and GAM model produced similar distribution with high density regions around the Isle of Skye, in the Small Isles and in the north of the Sound of Jura. The GEE model produced a generally patchy distribution pattern with small areas of highest density throughout the study region.

The presence-only visual model predicted the highest density regions to the west of Mull, around the Small Isles and in the Little Minch and in the middle of the Minch. This differed from all the other models where a more coastal distribution was predicted. This model predicted low-medium suitability for the Sound of Jura, which other models predicted as higher density.

3.2.5 Areas covered

The area coverage predicted by the four modelling methods showed a consistent pattern across all the data subsets. In all cases, the presence-only model predicted a larger area than the GLM, GAM and GEE models, which all predicted similar, smaller high density areas (Figures 8-11 and Table 1 in Appendix 1). However, the distribution patterns for subset 1 appeared quite different.

4 Discussion

The aim of this project was to investigate modelling methods to predict habitat use for harbour porpoises as a first step towards considering how such methods could be used to inform the assessment of Favourable Conservation Status (FCS) for all cetacean species in the UK. A comprehensive harbour porpoise dataset was used because this species has a relatively high sighting/acoustic detection rate and consequently a relatively high rate of presences in the datasets making the construction of models easier than for species that are more rarely encountered. Subsets of the data were used to investigate the effect of using sparser datasets; the results from analyses of these data provide some information on the likely performance of the methods for modelling habitat use of less dense species.

4.1 Geospatial methods

In general, the geospatial methods struggled to generate consistent distribution patterns and it was very difficult to glean meaningful information from the maps produced. It was only possible to construct models using the full datasets, which had many grid cells with detections in them and so fewer zeroes. The subsetted data contained a large number grid cells with no detections or effort in them, making it difficult for the models to generate results.

These methods may be more suitable for data from more systematically designed surveys, in which the spatial spread of effort data would be more even.

4.2 Presence-only (MAXENT) models

The predicted distributions produced using MAXENT were reasonably consistent across the datasets, predicting an inshore distribution pattern across the west coast of Scotland. The visual and acoustic full datasets yielded very similar distribution patterns, with large, coarse areas of the west coast determined to be highly suitable habitat for harbour porpoises. However, while the presence-only models managed to identify general patterns satisfactorily, they lacked the detail required to identify smaller areas of high habitat use. Identifying such areas may be important for determining where to site SACs and how large they should be.

MAXENT (and presence-only models in general) are unable to take into account potential biases generated by data collected where search effort is not representative of the study area. Spatially uneven data caused by some areas being visited more often than others can lead to bias. This may lead to misleading conclusions being drawn regarding a species distribution and habitat use. However, the direction and extent of such bias will be unknown and will depend on each dataset.

The results from the presence-only models constructed using data subset 2 may provide a good illustration of this. Using the full datasets, the models were satisfactory in generating a general pattern of distribution. In subset 2, however, the effort data were skewed strongly so that there was reasonable effort in the Minch region (in the north), but only a single transect conducted in the Sound of Jura (in the southern part of the study area). For the visual data, the Sound of Jura was much less important than in other MAXENT models, most likely because there was little effort and therefore a small number of presences there.

In studies, especially smaller scale ones, where data may not be evenly spread across an area of interest, presence-only analysis would not usually be the method of choice because the results may show where effort was distributed as much as where animals are concentrated. Presence-only analysis of datasets in which search effort is focused in areas where it is believed likely that animals will be encountered may generate a self-fulfilling prophecy, whilst ignoring areas that were not visited.

The disadvantage of presence-only methods, therefore, is that they can be biased by where effort has or has not been distributed but the extent of that bias cannot be evaluated without the use of effort data. As a general rule, therefore, analyses should use effort data in analysis where possible. If effort data are not available, the strength of the inferences that can be made from the results will depend on how well it is believed that the effort was representatively or evenly distributed.

4.3 Statistical Inference Models

The GLM, GAM and GEE methods all incorporate effort data, which accounts for any unevenness in the distribution of survey coverage. The results are therefore unbiased in this respect. These methods generated more detailed distribution maps that may be useful in the context of assessing habitat use and the placement of SACs.

However, smaller sample sizes will result in results that are subject to higher sampling variability. For the full datasets, the models produced generally similar distribution patterns, but the results from the subsetted models showed more variability. Although not estimated or presented here for reasons of limited time in this project, it is therefore important to consider measures of uncertainty when interpreting the results from these methods. For example, these could be in the form of maps of the coefficient of variation of predicted density, which would show areas of relatively low or high precision and therefore confidence in the results, or in the form of maps of the lower and upper 95% confidence limits of predicted density, which would show the "extremes" of the predictions.

GLMs and GAMs produced very similar results and the more laborious GEE method of dealing with autocorrelation in the data did not markedly change the resulting predictions of distribution. In future analyses, it is therefore probably sufficient to limit analysis to the use of GLMs and GAMs. Sparse datasets are less likely to be affected by autocorrelation.

4.4 Analysis of other datasets

The results of this project illustrate that the method of analysis to provide information on habitat use is less important for a large dataset in which effort is well distributed over a study area that for smaller, less evenly distributed datasets. The harbour porpoise is by far the most abundant cetacean species in UK waters; datasets for other species will all be smaller, in some cases very much smaller. The implications for analysis are two-fold. First, it will be important to use methods that minimise bias, which means using methods that incorporate searching effort data. Second, results will be subject to considerable variability, which means that it will be important to show prediction errors. The best methods to use will likely be GLMs or GAMs.

For species for which effort-related data are unavailable, presence only methods could be used but the value of the results will depend on how well it is believed the data are representative of the area in question.

4.5 Further Analysis

There are a number of steps that were outside of the scope of this project that could be taken to build on the work presented here. Future analysis of these or other data could include:

- Investigations of the coefficients of variation (CVs) associated with the predicted relative density values from each of the models. Assessing the uncertainty around model predictions is a key step and could provide insight into which modelling methods are most efficient in capturing distribution patterns.

- A cross-validation technique could be used to assess how well the models perform in capturing the patterns that generate the data. This model evaluation could provide additional information on which statistical techniques perform best.

- Techniques exist for the quantitative comparison of statistical models (e.g. see Potts & Elith 2006). Future analysis could be expanded to include such an analysis.

- This study used a real dataset and it was not possible to 'ground-truth' results. Future analysis could include the development of more robust simulated datasets to act as a 'truth' against which model outputs could be compared.

- Datasets with a varying proportion of zeroes in them, as proxies for data-rich (low zeroes) and data-poor (high zeroes) species could be simulated to allow investigation of which statistical techniques were most appropriate for such species.

5 References

Bailey, H., Thompson, P.M., 2009. Using marine mammal habitat modelling to identify priority conservation zones within a marine protected area. Marine Ecology-Progress Series 378, 279-287.

Booth, C.G., 2010. Variation in habitat preference and distribution of harbour porpoises west of Scotland. Unpublished PhD thesis, University of St Andrews. 240pp.

Ballinger, G.A., 2004. Using Generalized Estimating Equations for Longitudinal Data Analysis. Organizational Research Methods 7, 127 - 149.

Cañadas, A., Sagarminaga, R., de Stephanis, R., Urquiola, E. & Hammond, P.S. 2005. Habitat selection modelling as a conservation tool: proposals for Marine Protected Areas for cetaceans in southern Spain. Aquatic Conservation: Marine and Freshwater Ecosystems 15: 495-521.

Cañadas, A . & Hammond, P.S. 2006. Model-based abundance estimates for bottlenose dolphins off southern Spain: implications for conservation and management. Journal of Cetacean Research and Management 8(1): 13-27.

Cañadas, A., & Hammond, P.S., 2008. Abundance and habitat preferences of the short-beaked common dolphin (*Delphinus delphis*) in the South-western Mediterranean: implications for conservation. Endangered Species Research 4, 309-331.

CODA, 2009. Cetacean Offshore Distribution and Abundance in the European Atlantic. Report available from SMRU, Gatty Marine Laboratory, University of St Andrews, St Andrews, Fife KY16 8LB, UK. http://biology.st-andrews.ac.uk/coda/.

Cox, D.R., & Snell, E.J., 1989. Analysis of Binary Data (Second Edition), 2nd Edition edn. Taylor & Francis Ltd Chapman & Hall/CRC.

deBoor, C., 1978. A Practical Guide to Splines. Springer-Verlag. pp372.

Elith, J., Phillips, S.J., Hastie, T., Dudik, M., Chee, Y.E. & Yates, C.J. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17: 43-57.

Embling, C.B., Gillibrand, P.A., Gordon, J, Shrimpton, J, Stevick, P.T. & Hammond PS. 2010. Using habitat models to identify suitable sites for marine protected areas for harbour porpoises (*Phocoena phocoena*). Biological Conservation 143: 267-279.

Fernandez, M., Oviedo, L., Hartman, K., Sousa, B., & Azevedo, J.M.N. 2009. Differences in spatial distribution of two small delphinids (*Delphinus delphis* and *Stenella frontalis*) in two islands of the Azores Archipelago. Reports of the International Whaling Commission: SC/61/SC8, May 2009.

Fox, J., & Monette, G., 1992. Generalized collinearity diagnostics. Journal of the Acoustical Society of America 87, 178-183.

Gillespie, D., & Chappell, O., 2002. An automatic system for detecting and classifying the vocalisations of harbour porpoises. Bioacoustics-the International Journal of Animal Sound and Its Recording 13, 37-61.

Gillespie, D., Leaper, R., Gordon, J. & Macleod, K.M., in press. An integrated data collection system for line transect surveys. Journal of Cetacean Research and Management.

Gillespie, D.M., Swift, R.J., Caillat, M., Gordon, J., & Hammond, P.S. in prep. Acoustic detection and tracking of harbour porpoise on line transect surveys.

Halekoh, U., Højsgaard, S., & Yan, J., 2006. The R Package geepack for Generalized Estimating Equations. Journal of Statistical Software 15, 12-22.

Hardin, J.W., & Hilbe, J.M., 2002. Generalized estimating equations, 2nd Edition edn. Chapman and Hall CRC.

Hedley, S.L., 2000. Modelling heterogeneity in cetacean surveys. Unpublished PhD thesis, University of St. Andrews. 139pp.

Isaaks, E. H. & Srivastava, R. M. 1990. *An Introduction to Applied Geostatistics.* xix + 561 pp. New York, Oxford: Oxford University Press.

Jaramillo-Legorreta, A., Rohas-Bracho, L., Brownell, R.L. Jr., Read, A.J., Reeves, R.R., Ralls, K. & Taylor, B.L. 2007. Saving the Vaquita: immediate action, not more data. Conservation Biology 21:1653-1655.

Kim, Y.J. & Gu, C. 2004. Smoothing spline gaussian regression: more scalable computation via efficient approximation. Journal of the Royal Statistical Society: Series B. 66, 337-356.

Lennon, J. 2000. Red-shifts and red herrings in geographical ecology. Ecography 23, 101 - 113.

Liang, K.-T., & Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13-22.

MacLeod, C.D., Mandleberg, L., Schweder, C., Bannon, S.M., & Pierce, G.J., 2008. A comparison of approaches for modelling the occurrence of marine animals. Hydrobiologia 612, 21-32.

Marubini, F., Gimona, A., Evans, P.G.H., Wright, P.J., & Pierce, G.J., 2009. Habitat preferences and interannual variability in occurrence of the harbour porpoise *Phocoena phocoena* off northwest Scotland. Marine Ecology-Progress Series 381, 297-310.

Matthiopoulos, J. & Aarts, G. 2010. The spatial analysis of marine mammal abundance. *In*: Boyd, IL, Bowen, WD & Iverson, S (Eds). *Marine mammal ecology and conservation: a handbook of techniques*. Oxford University Press, Oxford. pp 68-97.

Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J-P., & Guinet, C. 2006. Geostatistical modelling of spatial distribution of *Balenoptera physalus* in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. Ecological Modelling 193, 615--628.

Palka, D. 1996. Effects of Beaufort Sea State on the sightability of harbor porpoises in the Gulf of Maine. Reports of the International Whaling Commission 46, 575-582.

Panigada, S., Zanardelli, M., MacKenzie, M., Donovan, C., Melin, F., & Hammond, P.S. 2008. Modelling habitat preferences for fin whales and striped dolphins in the Pelagos Sanctuary (Western Mediterranean Sea) with physiographic and remote sensing variables. Remote Sensing of Environment 112, 3400-3412.

Pearson, R.G. 2007. Species' Distribution Modeling for Conservation Educators and Practitioners. Synthesis. American Museum of Natural History. Available at http://ncep.amnh.org.

Phillips, S.J., Dudik, M. & Schapire, R.E. 2004. A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 655-662.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. 2006. Maximum entropy modeling of species geographic distributions**.** *Ecological Modelling*, 190:231-259.

Potts, J. & Elith, J. 2006. Comparing species' abundance models. Ecological Modelling 199, 153-163.

Praca, E. & Gannier, A., 2008. Ecological niches of three teuthophageous odontocetes in the northwestern Mediterranean Sea. Ocean Science 4, 49-59.

SCANS-II, 2008. Small Cetaceans in the European Atlantic and North Sea. Final report to the European Commission LIFE Nature programme on project LIFE04NAT/GB/000245. Report available from SMRU, Gatty Marine Laboratory, University of St Andrews, St Andrews, Fife KY16 8LB, UK. http://biology.st-andrews.ac.uk/scans2/.

Skov, H., & Thomsen, F. 2008. Resolving fine-scale spatio-temporal dynamics in the harbour porpoise *Phocoena phocoena*. Marine Ecology-Progress Series 373, 173-186.

Skov, H., Gunnlaugsson, T., Budgell, W.P., Horne, J., Nottestad, L., Olsen, E., Soiland, H., Vikingsson, G., & Waring, G. 2008. Small-scale spatial variability of sperm and sei whales in relation to oceanographic and topographic features along the Mid-Atlantic Ridge. Deep-Sea Research Part II-Topical Studies in Oceanography 55, 254-268.

Wood, S.N. 2006. Generalized Additive Models - An Introduction with R. Clapham & Hall, Boca Raton, Florida.

Table 1. Covariates used in models showing details of temporal/spatial resolution, units and sources of data used. UKHO – United Kingdom Hydrographic Office; MESH – Mapping European Seabed Habitats; POL – Proudman Oceanographic Laboratory; SAMS – Scottish Association for Marine Science.

Covariate	Information	Resolution	Unit	Source
Year/Month	Recorded *in situ* from vessel GPS	every 10 seconds (≈ 30m)	-	
Boat Speed	Recorded *in situ* from vessel GPS	every 10 seconds (≈ 30m)	Knots	
Sea Conditions	Recorded by Observers	every 30 minutes (≈ 5.2km)	Beaufort Scale	
Time of Day	Ratio: Time from Sunrise/Total time between sunrise and sunset for day	at every GPS location	-	POLTIPS
Position Relative to Tidal Range	Ratio: (Tidal Range at location on day – The minimum tidal height at location on day)/ Maximum Spring Tidal Range for location	at every GPS location	-	POLTIPS
Position in Daily Tidal Cycle	Ratio: Time since Low water for nearest tidal port / Time between successive low waters for nearest tidal port	at every GPS location	-	POLTIPS
Max. Spring Tidal Range	Maximum Spring Tidal Range for nearest tidal port	at every GPS location	m	POLTIPS
Distance from Land	Calculated in Manifold	at every GPS location	m	-
Percentage Gravel	Calculated from RSDB codes	Variable	%	UKHO / MESH EUNIS
Percentage Sand	Calculated from RSDB codes	Variable	%	UKHO / MESH EUNIS
Percentage Mud	Calculated from RSDB codes	Variable	%	UKHO / MESH EUNIS
Depth	Depth of seabed	0.2km	M	EDINA
Slope	Slope of seabed	0.2km	°	EDINA
Current Speed	Maximum current speed	POL: 1.8km / SAMS: 0.1 or 0.2km	m / s	POLCOMS / SAMS

Table 2. Survey effort, detections and detection rates for the visual and acoustic datasets collected in favourable conditions (visual: sea states 0-2; acoustic: sea state 0-6) in 2003-2008.

	Visual			Acoustic		
	Full	**Subset 1**	**Subset 2**	**Full**	**Subset 1**	**Subset 2**
Total Survey Effort (km)	34,699	2,994	4,487	42,653	5,637	6,907
Total # of Segments	17,353	1,497	2,244	21,355	2,822	3,457
# of Segments w/ Detections	1,015	101	137	3,367	472	562
% Segments w/ Detections	6%	7%	6%	16%	17%	16%
Number of Detections	2,381	238	276	4,927	786	859
Detection Rate (det/km)	0.069	0.079	0.062	0.12	0.14	0.12

Table 3. Covariates retained in the GLM, GAM and GEE models of the visual and acoustic datasets: (a) full dataset, (b) subset 1, (c) subset 2.

(a) Full dataset

		Temporal		Survey			Sediment			Tidal				Topographic		
Model	**Data**	**Year**	**Month**	**Sea State**	**Boat Speed**	**Time of Day**	**Percentage Gravel**	**Percentage Sand**	**Percentage Mud**	**Position Relative To Tidal Range**	**Position in Daily Tidal Cycle**	**Current Speed**	**Maximum Spring Tidal Range**	**Depth**	**Slope**	**Distance to Land**
GLM	*Visual*	X	X	X	X		X	X	X	X	X	X	X	X	X	X
	Acoustic	X	X		X		X		X	X		X	X	X	X	X
GAM	*Visual*	X	X	X	X		X		X	X	X	X	X	X	X	X
	Acoustic	X	X		X		X	X	X	X		X		X	X	X
GEE	*Visual*	X	X	X	X				X				X	X	X	X
	Acoustic	X	X		X								X	X	X	X

(b) Subset 1

		Temporal		Survey			Sediment			Tidal				Topographic		
Model	**Data**	**Year**	**Month**	**Sea State**	**Boat Speed**	**Time of Day**	**Percentage Gravel**	**Percentage Sand**	**Percentage Mud**	**Position Relative To Tidal Range**	**Position in Daily Tidal Cycle**	**Current Speed**	**Maximum Spring Tidal Range**	**Depth**	**Slope**	**Distance to Land**
GLM	*Visual*	X	X	X	X		X	X	X	X	X	X	X	X	X	X
	Acoustic	X	X	X	X					X	X	X	X	X	X	
GAM	*Visual*	X	X	X	X		X			X	X	X	X	X	X	X
	Acoustic	X	X	X	X				X	X	X	X	X	X	X	X
GEE	*Visual*		X	X									X		X	
	Acoustic	X	X	X	X					X			X	X		

(c) Subset 2

		Temporal		Survey			Sediment			Tidal				Topographic		
Model	**Data**	**Year**	**Month**	**Sea State**	**Boat Speed**	**Time of Day**	**Percentage Gravel**	**Percentage Sand**	**Percentage Mud**	**Position Relative To Tidal Range**	**Position in Daily Tidal Cycle**	**Current Speed**	**Maximum Spring Tidal Range**	**Depth**	**Slope**	**Distance to Land**
GLM	*Visual*	X	X	X	X						X			X	X	
	Acoustic	X	X	X	X			X	X		X	X	X	X	X	
GAM	*Visual*	X	X	X	X							X	X	X	X	
	Acoustic	X	X	X	X		X		X			X	X	X	X	X
GEE	*Visual*	X	X	X	X							X		X		
	Acoustic	X			X				X				X	X	X	

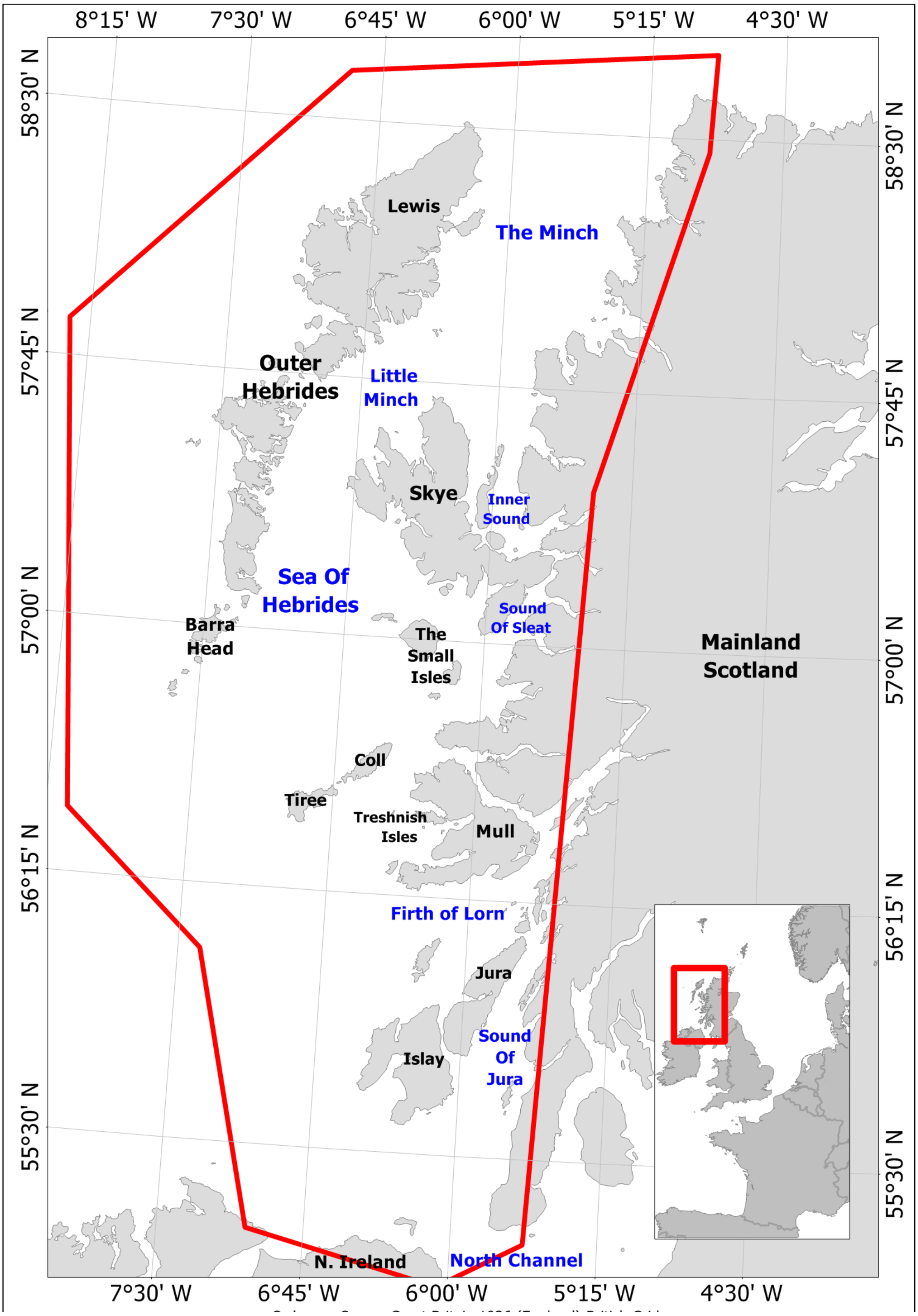

Figure 1. Study area and notable regions on the west coast of Scotland.

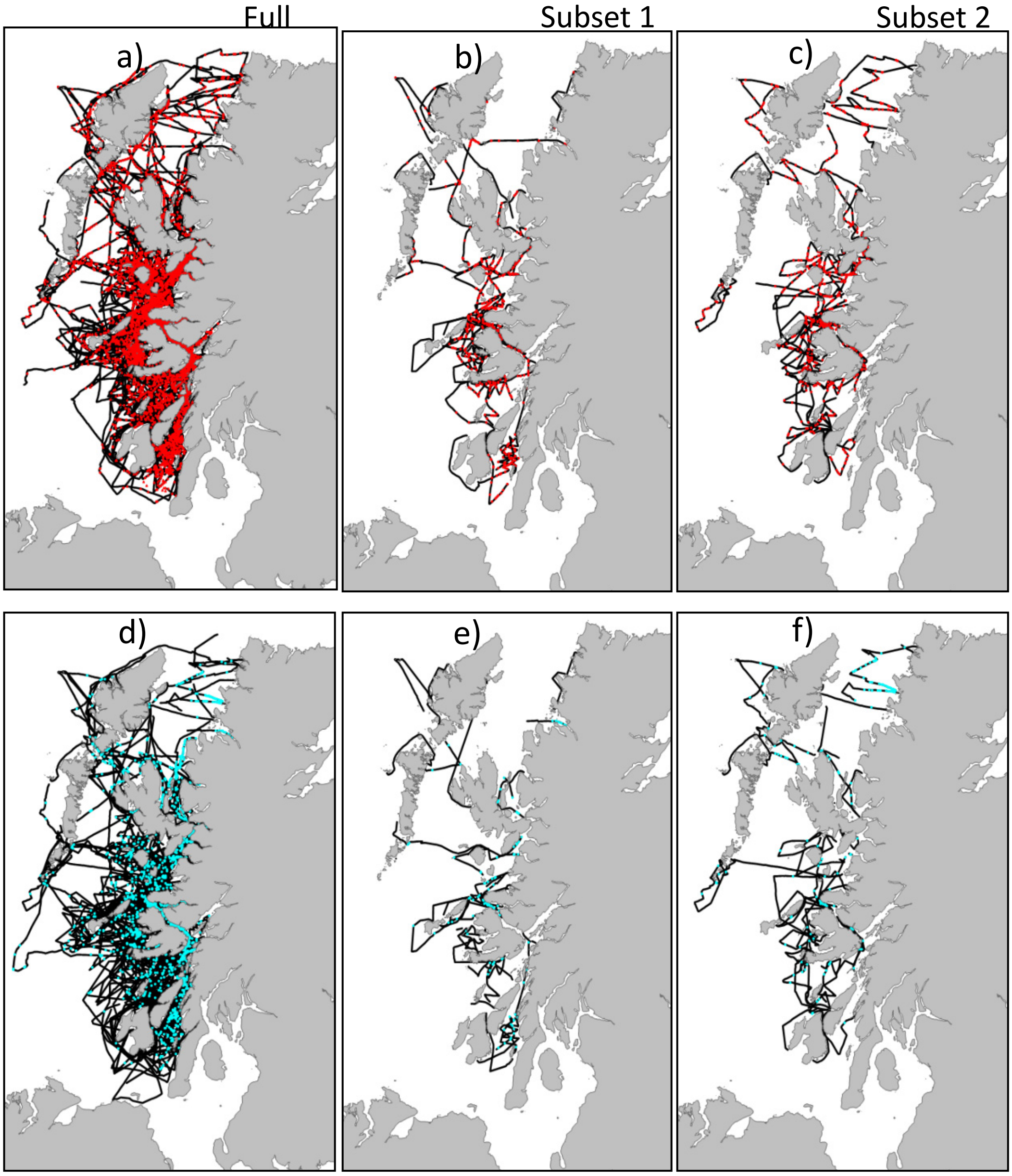

Figure 2. Survey effort tracklines and 2 km segments with detections from (a-c) acoustic surveys (2004-2008) and (d-f) visual surveys (2003-2008) and the subsetted datasets. Acoustic detections are shown in red and visual detections in light blue.

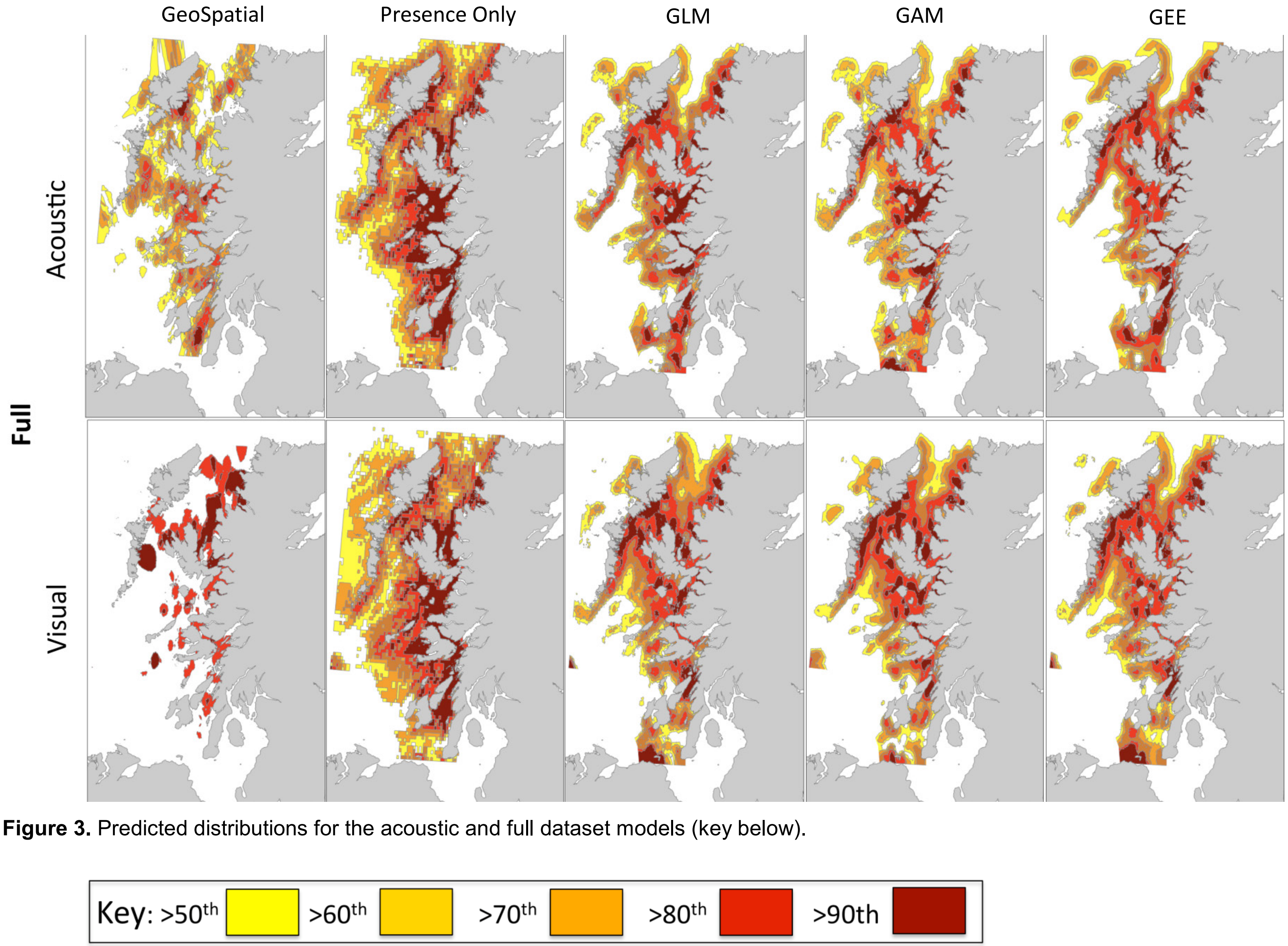

Figure 3. Predicted distributions for the acoustic and full dataset models (key below).

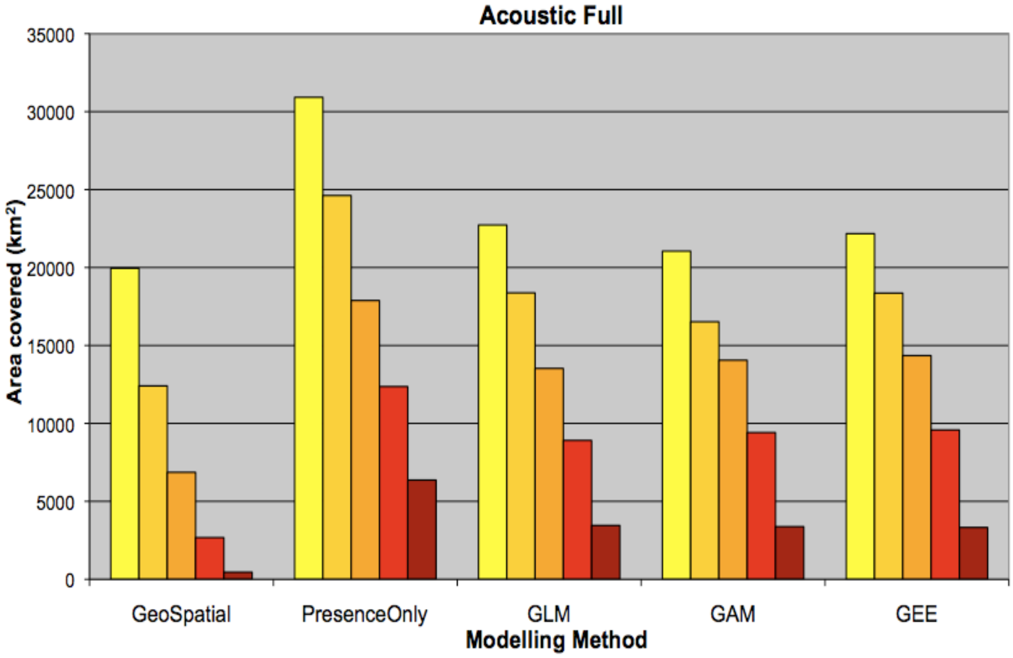

Figure 4. Area covered by the >50th, >60th, >70th, >80th, and >90th percentiles of relative density (or habitat suitability for presence-only) from the predicted distributions for the acoustic full dataset models.

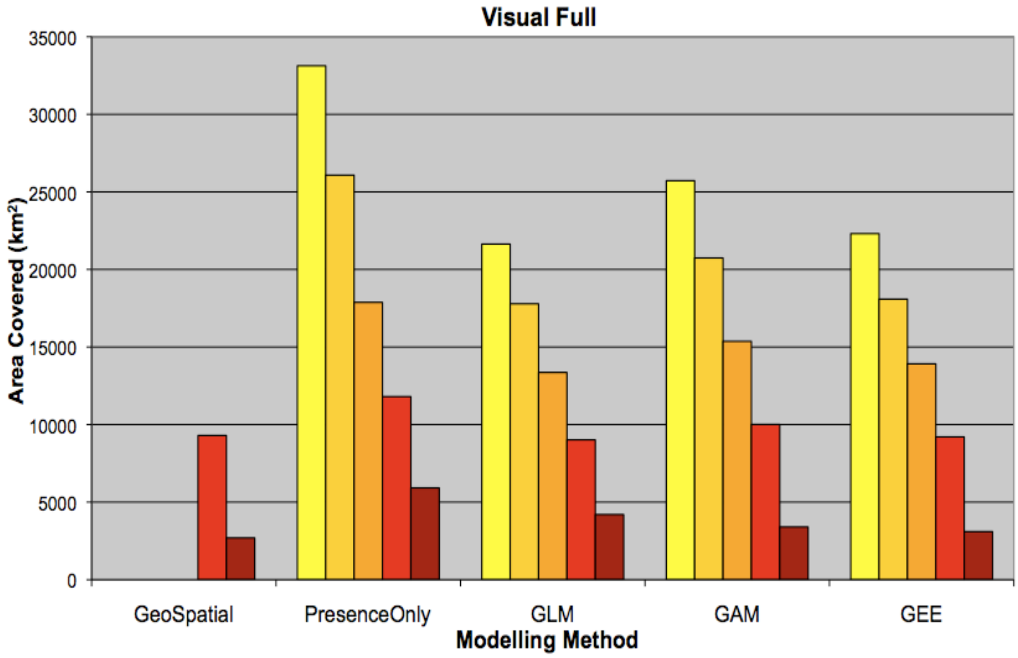

Figure 5. Area covered by the >50th, >60th, >70th, >80th, and >90th percentiles of relative density (or habitat suitability for presence-only) from the predicted distributions for the visual full dataset models.

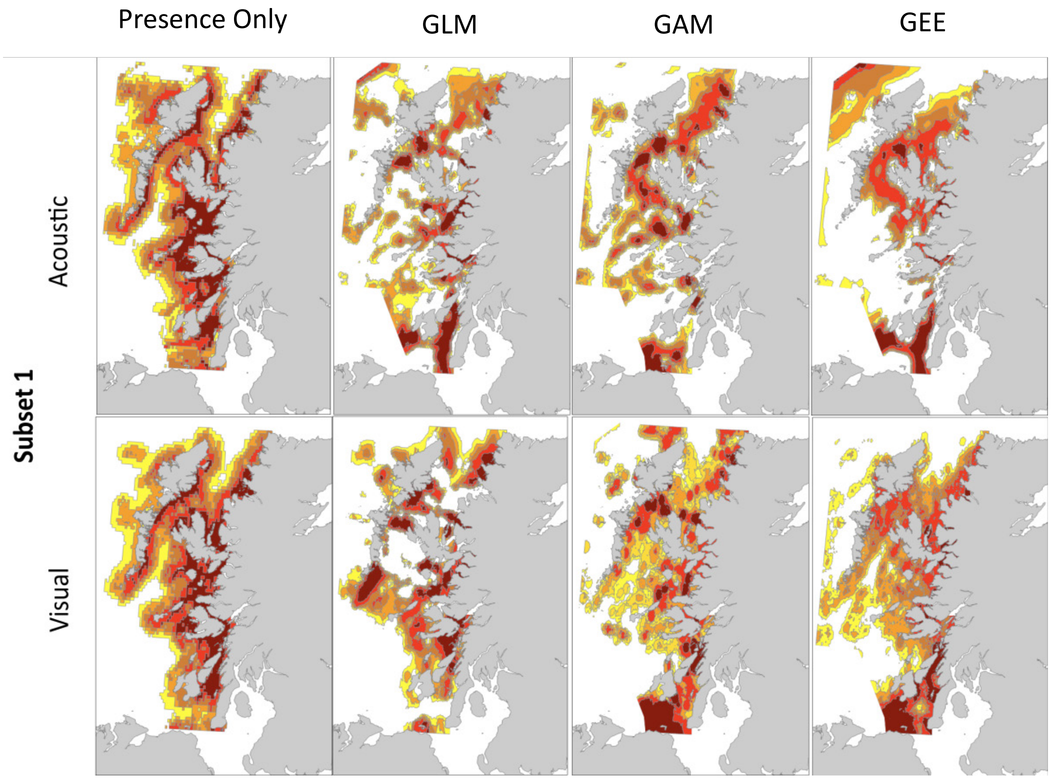

Figure 6. Predicted distributions for the acoustic and visual subset 1 models.

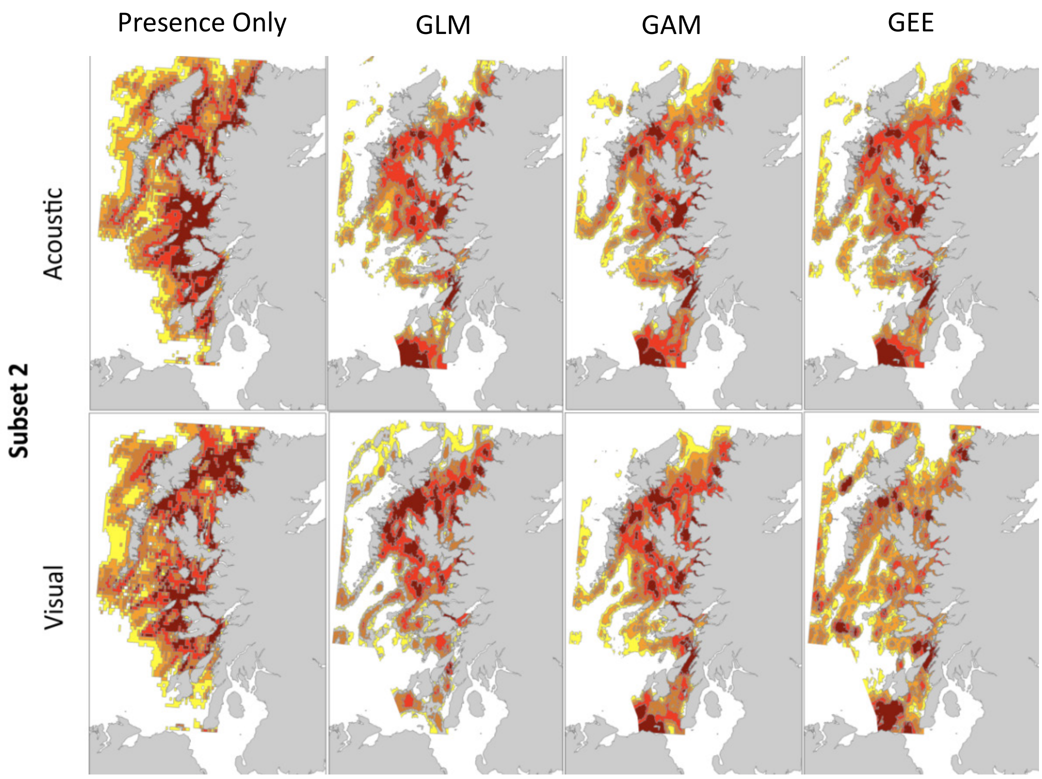

Figure 7. Predicted distributions for the acoustic and visual subset 2 models.

Key: >50th >60th >70th >80th >90th

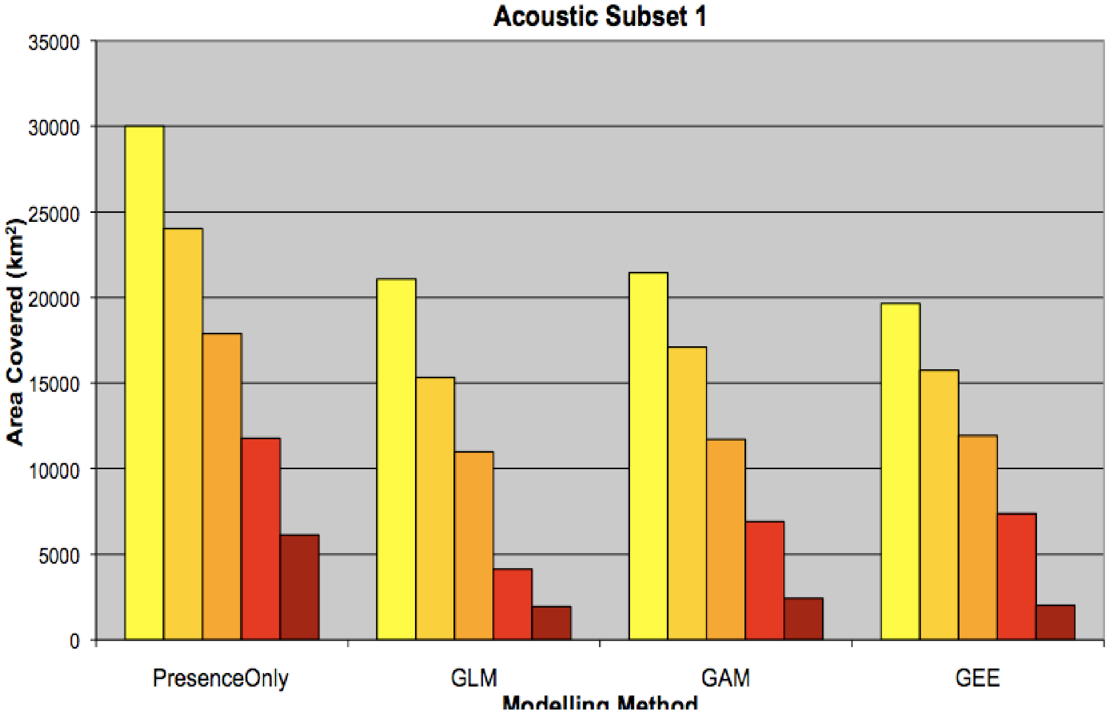

Figure 8. Area covered by the predicted distributions from the acoustic subset 1 models (key at base of the page).

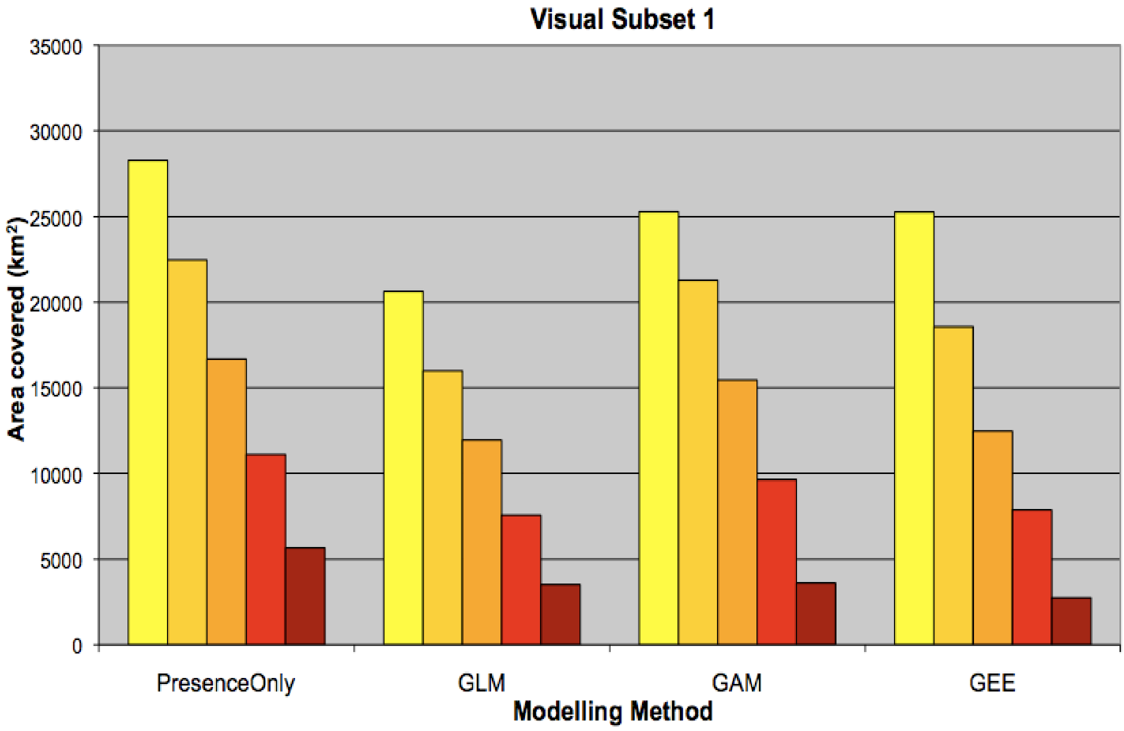

Figure 9. Area covered by the predicted distributions from the visual subset 1 models

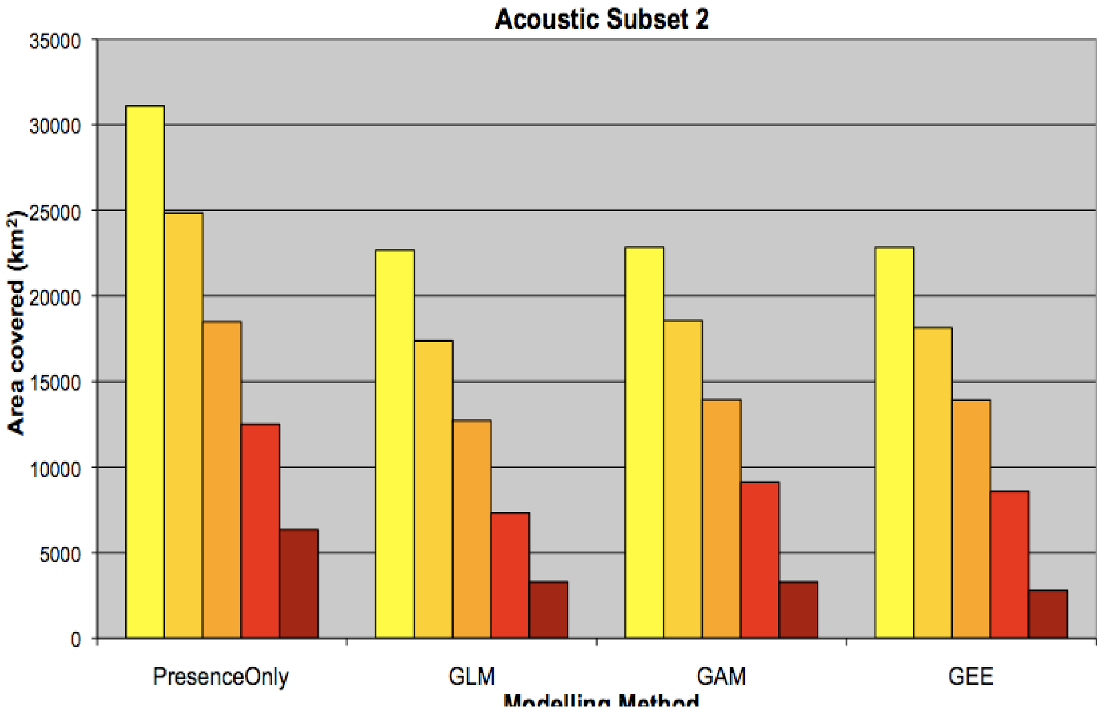

Figure 10. Area covered by the predicted distributions from the acoustic subset 2 models.

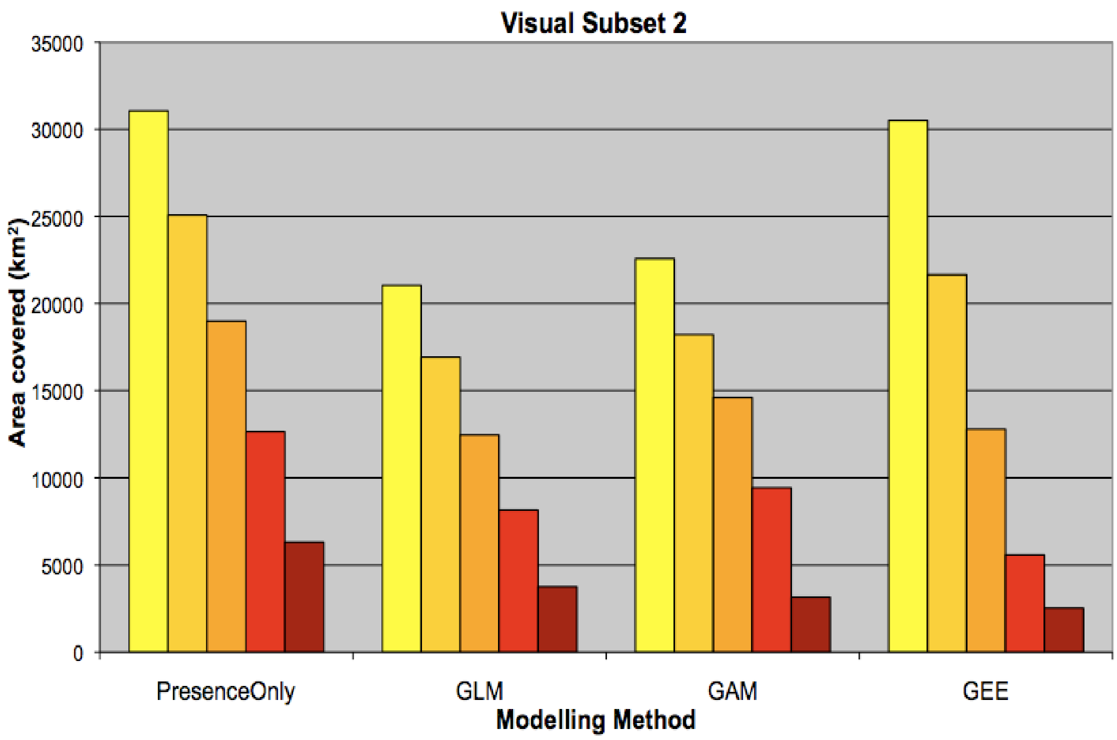

Figure 11. Area covered by the predicted distributions from the visual subset 2 models