

A comparison of approaches for modelling the occurrence of marine animals

Colin D. MacLeod · Laura Mandleberg ·
Caroline Schweder · Sarah M. Bannon ·
Graham J. Pierce

© Springer Science+Business Media B.V. 2008

Abstract Approaches for modelling the distribution of animals in relation to their environment can be divided into two basic types, those which use records of absence as well as records of presence and those which use only presence records. For terrestrial species, presence–absence approaches have been found to produce models with greater predictive ability than presence-only approaches. This study compared the predictive ability of both approaches for a marine animal, the harbour porpoise (*Phocoena phocoena*). Using data on the occurrence of harbour porpoises in the Sea of Hebrides, Scotland, the predictive abilities of one presence–absence approach (generalised linear modelling—GLM) and three presence-only approaches (Principal component analysis—PCA, ecological niche factor analysis—ENFA and genetic algorithm for rule-set prediction—GARP) were compared. When the predictive ability of the

models was assessed using receiver operating characteristic (ROC) plots, the presence–absence approach (GLM) was found to have the greatest predictive ability. However, all approaches were found to produce models that predicted occurrence significantly better than a random model and the GLM model did not perform significantly better than ENFA and GARP. The PCA had a significantly lower predictive ability than GLM but not the other approaches. In addition, all models predicted a similar spatial distribution. Therefore, while models constructed using presence–absence approaches are likely to provide the best understanding of species distribution within a surveyed area, presence-only models can perform almost as well. However, careful consideration of the potential limitations and biases in the data, especially with regards to representativeness, is needed if the results of presence-only models are to be used for conservation and/or management purposes.

Guest editor: V. D. Valavanis
Essential Habitat Mapping in the Mediterranean

C. D. MacLeod (✉) · L. Mandleberg ·
C. Schweder · S. M. Bannon · G. J. Pierce
School of Biological Sciences (Zoology),
University of Aberdeen, Tillydrone Avenue,
AB24 2TX Aberdeen, UK
e-mail: c.d.macleod@abdn.ac.uk

G. J. Pierce
Centro Oceanográfico de Vigo, Instituto Español de
Oceanografía, P.O. Box 1552, 36200 Vigo, Spain

Keywords Species distribution modelling ·
Marine species · Harbour porpoise ·
GLM · ENFA · GARP

Introduction

A detailed knowledge of species' distribution in relation to their environment is essential for understanding many aspects of their ecology, as well as for

effective conservation, management and assessment of possible impacts from anthropogenic activities (Lindenmayer et al., 1991; Beerling et al., 1995; Schulze & Kunz, 1995; Austin et al., 1996). However, knowledge on the true distribution of many marine animals remains limited, especially for species that are hard to detect. In the marine environment, poor detectability is primarily a function of the fact that humans can only directly observe surface waters close to the coast with any ease and usually require expensive and complex equipment to conduct studies on species that occur only in waters far from shore (e.g. large research vessels) or below the surface (e.g. underwater vehicles and deep-water camera sleds—see Robison (2004)).

One solution to this lack of knowledge is to use mathematical approaches to model species distribution relative to various quantifiable aspects of their physical environment known as eco-geographic variables (EGVs). These modelled relationships can then be used to predict where species are most likely to occur and investigate ecological relationships between a species and its environment (Lindenmayer et al., 1991; Zaniwski et al., 2002). Many traditional modelling approaches require presence–absence data (Guisan & Zimmerman, 2000; Hirzel et al., 2001). That is, they require data on locations where a species is known not to occur (absence data) as well as data on locations where a species does occur (presence data). It is essential that any absence data used for such modelling are accurate and that none of the data points represent ‘false’ absences—locations where a species occurs but for some reason was not detected during data collection (Hirzel et al., 2002). For hard-to-detect species, even in terrestrial environments, it can be difficult to obtain datasets that do not include a substantial number of false absences. In the marine environment, accurate absence data may be all but impossible to collect for many species, particularly those that occur at great depth, far from shore, are very mobile, avoid survey vessels or that are difficult to detect in other ways.

The problem of false absences has led to the development of modelling approaches that do not use absence data (e.g. Robertson et al., 2001; Hirzel et al., 2002; Ortega-Huerta & Peterson, 2004). Such presence-only approaches are generally based on constructing a model of a species’ niche from locational records. This modelled niche can then be

used to predict distribution within the available environment.

The validity of such modelled niches is contingent on having unbiased distribution data available to build the models. If survey effort data are available, it is possible to both determine whether all habitat types have been adequately sampled and to correct for bias by using effort as a weighting factor in the model. However, as presence-only models do not take survey effort into account such models may be affected by biases in the collection of presence data. While this is less likely to be a problem with large numbers of records, as can often be available for terrestrial species from sources such as museum collections (e.g. Robertson et al., 2001; Reutter et al., 2003), this may be an issue when a small number of records is used to generate the model.

When presence–absence and presence-only modelling approaches have been compared using the same datasets, presence–absence models have generally been found to perform better and have higher predictive abilities (Hirzel et al., 2001; Brotons et al., 2004), leading to most researchers to prefer the use of presence–absence models whenever possible. However, these comparative studies have been limited to terrestrial species (Brotons et al., 2004) and theoretical populations (Hirzel et al., 2001) and it is not known whether the same relationship will hold in the marine environment where detectability of many species is much lower than for terrestrial species. Here, the abilities of presence–absence and presence-only modelling approaches to predict the distribution of a marine species, the harbour porpoise (*Phocoena phocoena* Linnaeus 1758), in relation to EGVs are compared for the first time.

Harbour porpoises are one of the smallest members of the order Cetacea and are known to be hard to detect, particularly in rougher seas (Palka, 1996; Laake et al., 1997; Teilmann, 2003). This low detectability is primarily a function of small body size, small group sizes, boat avoidance and unobtrusive surface behaviours. Traditionally, problems with detectability have been dealt with by introducing a correction factor to estimate the number of animals missed, especially for abundance estimates (Teilmann, 2003). However, such correction factors can be difficult to calculate (Laake et al., 1997; Teilmann, 2003). In particular, visual detectability of harbour porpoises varies in relation to many factors,

such as changes in group size with season (Bannon Pers. Obs.), behaviour, time of day and sea state (Palka, 1996).

Four modelling approaches were compared in this study. These were Generalised Linear Modelling (GLM), a widely used presence–absence technique (Sparholt et al., 1991; Guisan & Zimmerman, 2000; Garcia-Charton & Perez-Ruzafa, 2001; Guisan & Hofer, 2003; MacLeod et al., 2004; Evans & Hammond, 2004) which has been compared to presence-only techniques in previous studies (Hirzel et al., 2001; Brotons et al., 2004), and three presence-only approaches: Ecological niche factor analysis (ENFA), Genetic algorithm for rule-set prediction (GARP) and a PCA-based approach. Presence-only techniques were selected based on their previous successful application in the terrestrial environment (Robertson et al., 2001; Hirzel et al., 2002; Stockwell & Peters, 1999; Ortega-Huerta & Peterson, 2004). Currently, there are no published applications of these presence-only approaches to model the distribution of marine animals. The aim of this study was to directly compare the ability of these approaches to predict the occurrence of harbour porpoises within a surveyed area using a single data set, and, in particular, to explore the potential application of presence-only models to the marine environment.

Materials and methods

Study area and eco-geographic variables (EGVs)

This study was conducted in the Sea of Hebrides, an area of shelf waters to the west of Scotland, UK (Fig. 1). A geographic information system (GIS) consisting of 15,520 1 km² grid cells was created using ESRI Map Info software to cover this study area. Each cell was assigned a value for water depth, seabed slope, standard deviation of seabed slope, aspect of seabed and distance from the nearest coast using ESRI ArcView 3.2 software. The EGVs used in this study were primarily related to topography and included a number that are commonly used when studying the distribution of cetacean species (e.g. MacLeod et al., 2004; MacLeod & Zuur, 2005; Ingram et al., 2007) and that are known to be important for porpoise habitat use in the west of Scotland (MacLeod et al., 2007). While other variables, not included in this analysis, may also

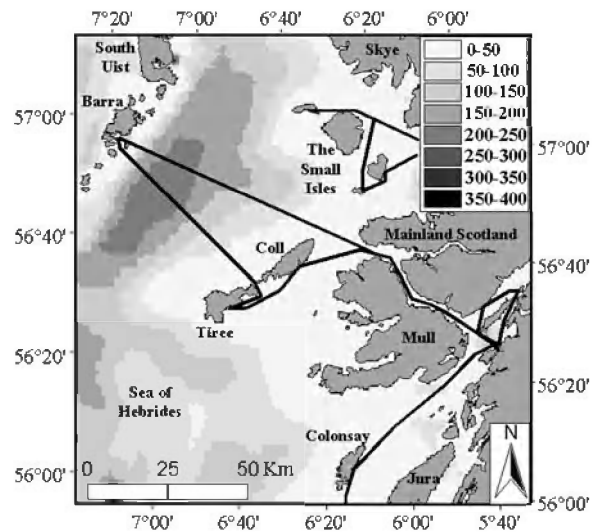


Fig. 1 The study area used to investigate the ability of different modelling approaches to predict the occurrence of harbour porpoises in the Sea of Hebrides. Black lines indicate route travelled by ferries used to survey for harbour porpoise. Shading indicates water depth. Latitudes are in degrees north and longitude in degrees west

relate to porpoise distribution, the aim of this study was not to identify all factors that relate to porpoise distribution but rather to compare modelling approaches using the same variables. Therefore, while this limitation should be borne in mind when considering the actual habitat preferences identified by the models presented here, it will not affect the results in relation to the comparisons of the predictive abilities of the different modelling approaches using this standardised data set.

Water depth was interpolated from the ETOP02 global 2' elevation dataset (National Geophysical Data Centre 2001) at a 1 km by 1 km resolution, and slope, standard deviation of slope and aspect for each cell were derived using ArcView functions. In order to make aspect a suitable parameter for inclusion in the analysis, it was converted into two linear components: aspect easting (the sine of the aspect value) and aspect northing (the cosine of the aspect value). For all modelling approaches, the modelling process started with all six variables. However, the EGVs included in the final model were identified through the modelling process independently for each modelling approach. Finally, each grid cell was assigned a random number using the random grid function in ArcView.

Data collection

Data on the occurrence of harbour porpoises were collected from repeated surveys along five fixed routes in the months of May to July 2003 and 2004 using passenger ferries as research platforms (Fig. 1). While these ferry routes may not cover a representative sample of habitat within the study area, the same data set was used for all four models and therefore allows a direct comparison of the predictive abilities of the different modelling approaches for the surveyed areas. In addition, the repeated coverage of these routes allowed a large number of grid cells to be surveyed on multiple occasions, a feature that was important for reducing the likelihood of false absences within the dataset, at low cost. The surveys were conducted by a single observer situated approximately 15 m above sea level to one side of the vessel. This gave a field of view that covered from 90 degrees on the observer's side of the bow to 20 degrees to the other side.

This field of view was continuously swept with 7×50 reticulated binoculars and with the naked eye. At the start and end of each survey, as well as every 15 min during the surveys, the position, direction of travel and speed of the ship were recorded using a GPS receiver, along with environmental variables such as sea state. Assuming a straight line course between the locations of the ship recorded every 15 min allowed the ship's track between these two points to be plotted. When any harbour porpoises were detected, the distance to the animals was estimated with the reticules in the binoculars (following the trigonometric methods of Lerczak & Hobbs, 1998) and a relative bearing to the animals was recorded using a compass rose, along with the group size, the ship's position, course and speed. This information allowed the actual position of each group to be estimated, in terms of latitude and longitude, and plotted in the GIS.

In order to identify those cells that were surveyed, all 15-min track segments surveyed in sea states of Beaufort 3 or less were identified. Around these survey segments, a cut off point of 750 metres from the vessel was defined as the point beyond which the observer could not accurately detect harbour porpoise at the surface (although even within this distance animals that were underwater would still be missed). This distance was based on previous experience with

surveys from these vessels, the binoculars used and the distances over which porpoises could be visually detected. Due to the restricted field of view, this resulted in a survey swath width of 1,000 m, 750 m on the side of the vessel where the observer was positioned and 250 m on the opposite side. Since results are not used to estimate absolute abundance of porpoises, no bias will result if the real swath width was not exactly 1,000 m. A cell was defined as surveyed if the survey swath covered a portion of the cell defined by a triangle with a hypotenuse of at least 500 m, or approximately one-eighth of the cell, although for the majority of surveyed cells the proportion of the cell within the survey swath was much greater than this. The total number of times each cell was surveyed throughout the study was then calculated. Finally, the sightings data associated with the 15-min segments conducted in sea states 3 or less were compared to the survey swathe and only those where the estimated position fell within it sightings for which used to identify which surveyed cells could be assigned as porpoise presence.

The surveyed cells within the study area were divided into a model construction dataset and a model testing dataset in a ratio of 2:1 using the random number assigned to each cell. Within each set, any cell where one or more groups of harbour porpoises were recorded were classified as 'presence', while all cells that were surveyed at least three times without recording any harbour porpoises were classified as 'absence'. This provided a relatively strict rule for classifying cells as 'absence' and reduced the likelihood of false absences (i.e. cells that are used by harbour porpoises but where they were not detected) within the datasets used for presence-absence modelling and intermodel comparisons.

Model construction

GLM

As the data were binary (presence/absence), a binomial regression was applied to the presence-absence data in the construction dataset. All linear and quadratic terms were included as potential predictors in the building of the model. Co-variance between each variable was assessed using pair plots and only variables with co-variance <0.8 were considered for the GLM. In order to select the model that explained

the most variation using the fewest number of variables, a ‘backwards stepwise’ procedure was used (BRODGAR software, Highland Statistics Ltd). The statistic used to select the final linear model was the Akaike Information Criterion (AIC—Chambers & Hastie, 1997). For the final model, the probabilities of harbour porpoise occurrence were calculated for all grid cells in the study area by substituting the intercept value and the coefficients for each of the variable into the following equation:

$$\text{Probability of occurrence} = e^{g(x)}$$

where $g(x)$ is the regression equation from the GLM.

PCA

PCA-based modelling followed the method provided by Robertson et al. (2001). For presence cells within the model construction dataset, a mean and standard deviation was calculated for each EGV. The values for each EGV for each presence cell were then standardised by subtracting this mean and dividing by the standard deviation. Standard PCA analyses were conducted using Minitab statistical software (Minitab Ltd) on these standardised values using all possible combinations of three or more variables. For each PCA, the predicted likelihood of occurrence in each cell was calculated by first standardising the values for each EGV of every cell in the study area by dividing it by the species mean and subtracting the species standard deviation for that variable. Then a total eigen score was calculated for each cell for each principal component by weighting each EGV used to construct the model with its principal component-specific eigen score. The total eigen score for each principal component was then divided by its eigen value. Finally, the resulting values for each principal component were squared and summed until the accumulated variation explained by the principal components was >90%. The Chi-squared distribution was then used to produce a likelihood of occurrence based on this value. The model-testing dataset was used to assess the predictive ability of all models using a receiver operating characteristic (ROC) plot.

ROC plots provide a threshold-independent method for assessing the predictive ability of ecological models and allow the predictive abilities of models constructed using different techniques to be directly compared (Fielding & Bell, 1997). For every

possible threshold value for separating model predictions into predicted presence and predicted absence, sensitivity and specificity values were calculated. Sensitivity values indicate the proportion of cells where the model correctly predicted presence in relation to all presence cells in the testing dataset. Specificity values indicate the proportion of cells where the model correctly predicted absence in relation to all absence cells in the testing dataset. When one minus the specificity value (on the X-axis) and the sensitivity value (on the Y-axis) at every possible threshold value are plotted on a scatter plot, the area under curve (AUC) provides a measure of predictive ability. A random model (i.e. does not predict occurrence better than randomly selecting cells from the testing dataset) would be expected to have an AUC of 0.5, while a model that was in perfect agreement with the testing dataset would have an AUC of 1.0 (Fielding & Bell, 1997). The higher the AUC, the greater the predictive ability of the model under consideration and the further it differs from a random model.

ROC analysis was conducted using the Analyse-It ‘Add-In’ to Microsoft Excel produced by Analyse-It, LTD. The PCA model with the highest AUC was defined as the best PCA model of harbour porpoise occurrence within the study area.

ENFA

ENFA was conducted using Biomapper 3 software (Hirzel et al., 2000). An EGV grid for each variable was imported into the Biomapper programme along with a grid identifying which cells were classified as ‘presence’ within the model construction dataset. The EGV grids were standardised using a Box–Cox transformation. The broken stick rule was used to suggest how many niche factors should be used to construct the final habitat suitability map. This habitat suitability map classified cells on a scale of 0–100 based on its combination of values for the EGVs, weighting each one in a similar way to the PCA analysis. A cell with a habitat suitability value of zero would have the least suitable combination of values for all variables, while a cell with a value of 100 would have the most suitable combination. This habitat suitability map was then assessed using jack-knife cross-validation and area-adjusted frequencies (Boyce et al., 2002).

GARP

GARP was conducted using GARP Desk Top software (University of Kansas Centre for Research, Inc.). This software was set to automatically conduct 20 runs of every possible combination of the EGVs consisting of at least three EGVs and using four-fifths of the presence cells in the construction dataset. The final fifth was used for an assessment of each model to identify the best combination of EGVs based on the lowest mean omission error across the 20 runs. For the best model, the output maps of all 20 runs were imported into the GIS and summed. This resulted in a map that gave each cell a value between 0 and 20. A zero value meant that presence was not predicted in a cell in any of the 20 runs, while a value of 20 meant that presence was predicted in all 20 runs.

Intermodel comparison

ROC plots were calculated for each model using the testing dataset, allowing a direct comparison to be made between the predictive abilities of each model within the surveyed area (Fielding & Bell, 1997). In addition, the spatial predictions of the models were compared by using the models to predict species occurrence for all cells (including those not surveyed) within the study area. The study area was then divided into 12 sub-areas based on coarse oceanographic similarities and differences (Fig. 5). The average predicted occurrence for cells within these 12 sub-areas for each model was then compared using Pearsons correlation to assess whether each model was predicting relatively high and relatively low occurrences in the same spatial areas.

Results

Harbour porpoises were recorded on 159 occasions in sea states of 3 or less, in 101 separate grid cells (Fig. 2). This surveyed area constitutes a substantial proportion of the Sea of Hebrides (around 10%), however all results presented below are only applied to the surveyed areas. Of these presence cells, 68 were partitioned into the model construction dataset and 33 into the testing dataset. Of the remaining cells in the study area, 965 were surveyed three times or

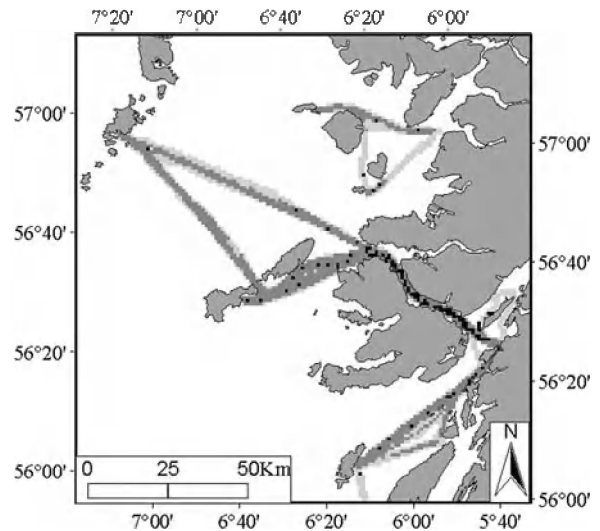


Fig. 2 Cells defined as surveyed during this study. Black—cells where harbour porpoises were recorded; dark grey—Cells surveyed three or more times without harbour porpoises being recorded; Light grey—Cells surveyed only once or twice times without harbour porpoises being recorded

more. Of these, 679 were classified as absence data for model construction and 286 for model testing.

For GLM, all six variables considered were found to have a sufficiently low co-variance to be included in the model as separate terms. The model with the best 'fit' used three variables: (i) distance from coast (ii) standard deviation of slope and (iii) aspect northing. The AIC value for this model was 363.6. Both distance from coast ($P = 0.004$) and standard deviation of slope ($P = 0.002$) had highly significant effects, with porpoise presence decreasing with increasing distance from the coast (co-efficient: -0.0002537) and increasing with greater standard deviation of slope (co-efficient: 0.8957). Aspect northing had a positive effect on porpoise presence (co-efficient: 0.3642), but this was not significant ($P > 0.05$). However, including it increased the fit of the model as measured by the AIC. For the PCA, the model with the highest AUC used four EGVs: distance from the coast, water depth, and aspect easting and aspect northing. Four principal components were used to construct this model accounting for 100% of the variation in the presence data (Table 1). In the ENFA, four niche factors were selected accounting for 88.4% of the variation (Table 1). For GARP, the best model (the one with the lowest omission error for the internal testing

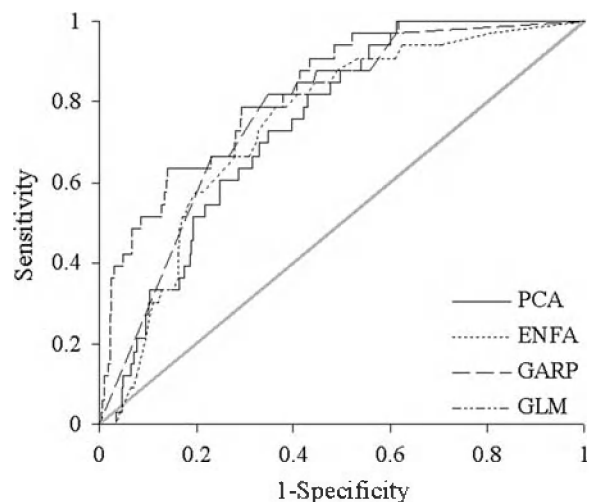
Table 1 Components and niche factors used for PCA and ENFA modelling of harbour porpoise distribution. The most important variable for each component/niche factor is highlighted in bold; the second most important is highlighted in italics

PCA	Variable	Component 1	Component 2	Component 3	Component 4
	Distance from coast	<i>−0.645</i>	0.153	−0.293	<i>0.689</i>
	Aspect easting	0.380	<i>0.340</i>	−0.856	−0.083
	Aspect northing	−0.139	0.916	<i>0.322</i>	−0.197
	Water depth	0.648	0.149	0.280	0.693
	Eigen Value	1.6089	1.0246	0.8808	0.4848
	Accumulated Variation Explained	0.402	0.659	0.879	1.000
ENFA	Variable	Niche Factor 1	Niche Factor 2	Niche Factor 3	Niche Factor 4
	Distance from coast	−0.599	<i>0.694</i>	−0.521	0.107
	Water depth	0.523	0.703	0.087	−0.279
	Aspect northing	0.107	−0.102	−0.357	−0.733
	Aspect easting	−0.016	−0.065	0.043	0.098
	Slope	0.157	0.035	0.395	−0.219
	Standard deviation slope	<i>0.576</i>	0.092	−0.661	<i>0.563</i>
	Eigen Value	6.805	2.943	1.116	1.046
	Accumulated Variation Explained	0.505	0.723	0.806	0.884

procedure) was produced using three EGVs, distance from coast, slope and standard deviation of slope.

The ROC plots revealed that all four models differ significantly from a random model ($AUC = 0.5$), indicating that all four approaches produced models that could predict harbour porpoise occurrence in relation to EGVs (Fig. 3). Of the four approaches, the GLM had the highest AUC (0.828) followed by the GARP model (0.773), PCA (0.746), and ENFA (0.745—Table 2).

While these comparisons showed that GLM had the greatest predictive ability, the only significant differences (at $P = 0.05$) were that the GLM had a significantly greater predictive ability than the PCA. However, multiple statistical comparisons were used to test the null hypothesis that there was no difference in the predictive ability between the modelling techniques. As a result, the Bonferroni correction (the usual threshold for significance divided by the number of statistical tests conducted) should probably be applied to reduce the chance of a type 1 error (but see Devlin et al., 2003; Garcia, 2004). This would shift the threshold P -value for a significant difference in predictive ability from 0.05 to 0.0083. At this corrected P -value, there were no significant differences in the predictive ability between any of the models (Table 3).

**Fig. 3** Receiver operating characteristic (ROC) plots used to assess and compare the predictive abilities of the different modelling approaches (as recommended by Fielding & Bell, 1997). Black lines—ROC plots for individual models; Light grey line—Random model with area under curve (AUC) of 0.5. See Table 2 for AUC values of each model

In terms of the predicted spatial occurrence, all models predicted similar areas of high and low occurrence. For example, all four models predicted the highest likelihood of occurrence within shallow coastal areas, such as the Sound of Mull, and the

Table 2 Area under curve (AUC) for ROC plots of each model type, including the repeat of GARP for the testing dataset

	GLM	PCA	ENFA	GARP
Area under curve (AUC)	0.828 (0.762–0.895)	0.746 (0.676–0.816)	0.745 (0.667–0.823)	0.773 (0.701–0.846)

Numbers in brackets are 95% confidence intervals. All models differed significantly from a random model at $P < 0.0001$

Table 3 Difference in predictive abilities of the models as measured by ROC plots

Probability that AUCs are different due to chance					
Difference in AUC	Model	GLM	PCA	ENFA	GARP
	GLM	–	$P = 0.0256$	$P = 0.0810$	$P = 0.1472$
	PCA	0.083	–	$P = 0.9847$	$P = 0.4877$
	ENFA	0.084	0.001	–	$P = 0.4925$
	GARP	0.055	0.028	0.029	–

Probabilities in bold indicate significant differences at $P = 0.05$. However, if the Bonferroni correction for multiple statistical tests of a null hypothesis is applied there is no significant difference for any pair-wise comparison at $P = 0.0083$

lowest likelihood of occurrence in deeper waters further from shore, such as the Sea of Hebrides (Fig. 4). This apparent similarity was confirmed by the correlation of the average predicted occurrence in the 12 sub-areas, as there was a strong and significant correlation between the spatial predictions of all four models (Table 4). Therefore, the relative spatial occurrence predicted by each model within the study area was very similar.

Discussion

Ecological modelling offers the opportunity to investigate species distribution and to increase the understanding of the biology of individual species. However, while mathematically sound, modelling approaches can often be difficult to implement due to the imperfections and limitations of biological data. This can reduce the usefulness of a specific approach to model the distribution of a specific species. In particular, problems associated with detecting species can lead to errors in assigning locations into presence/absence categories (Hirzel et al., 2002; Williams, 2003) and violate assumptions of accurate absence data required for modelling approaches such as GLM (although it may be possible to use the amount of survey effort at a specific location as a weighting factor to at least partially control for the risk of ‘false’ absences within the dataset). This is likely to be an issue for many marine species that are

inherently hard to detect due to problems associated with undertaking surveys for species presence in the marine environment. Therefore, modelling approaches that do not require accurate absence data would appear to offer a solution to these problems, provided that the survey coverage is adequate.

The results of this study suggest that presence–absence approaches provide the best predictive ability, and therefore presumably the best understanding of species distribution, in relation to ecogeographic variables. As a result, when it is possible to implement them, such presence–absence approaches should be used. However, this study also suggests that when no sufficiently accurate and/or suitable absence data are available, presence-only approaches, such as ENFA, can potentially produce models of the distribution of marine species which perform significantly better than random models and that do not necessarily have a significantly poorer performance than presence–absence modelling approaches for the same surveyed area. In addition, the predicted spatial distributions of the presence–absence model and the three presence-only models were similar, with all predicting the highest likelihoods of occurrence in similar areas. Therefore, while their application may be limited to specific data sets, these modelling approaches do appear to offer an opportunity to increase our understanding of the distribution of marine species.

The results of this study differ from previous studies, such as Brotons et al. (2004) that found a

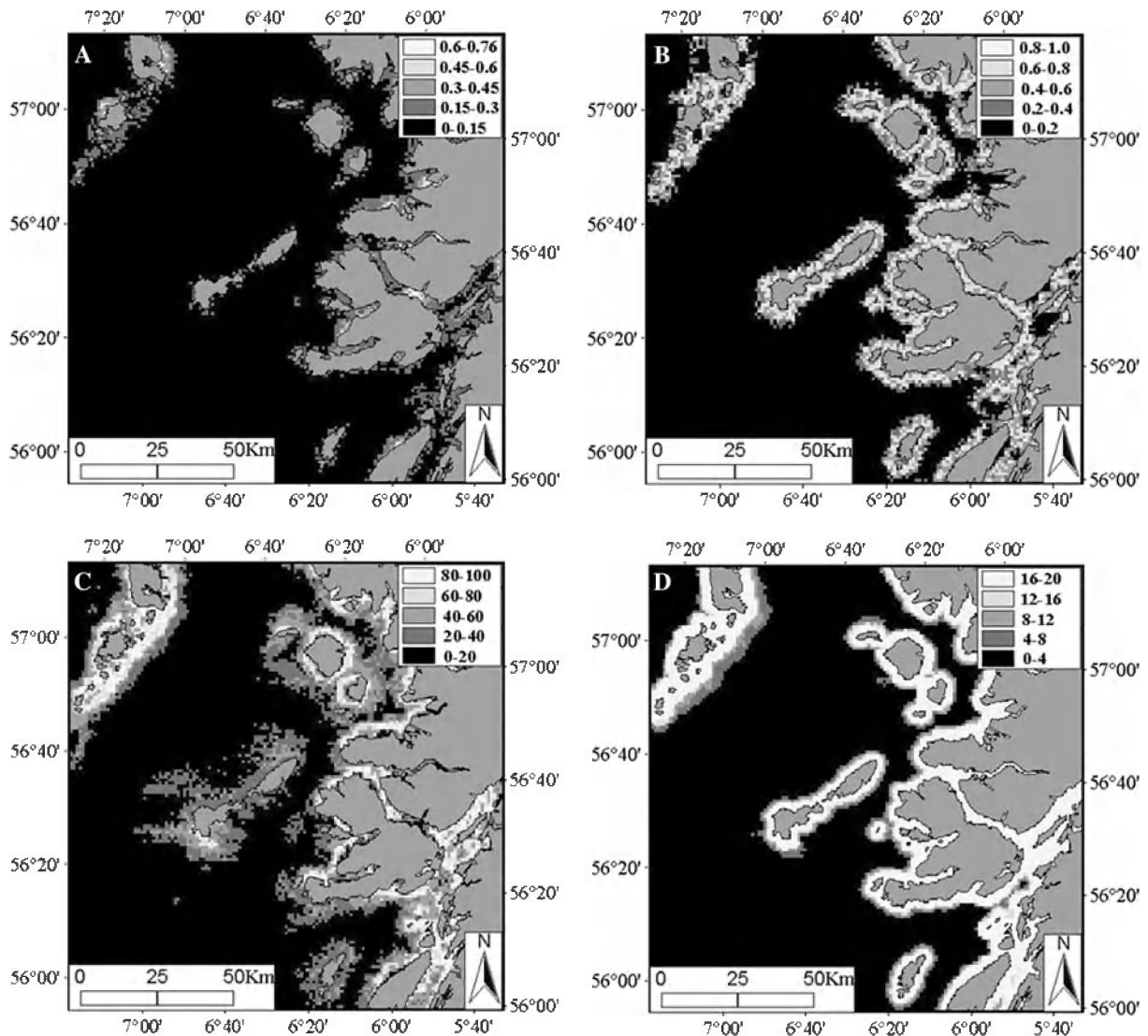


Fig. 4 Maps of predicted occurrence of harbour porpoises within the study area from each of the four modelling techniques. **(A)** GLM—Predicted probability of occurrence for individual cells ranging from 0 to a highest probability of 0.755; **(B)** PCA—Predicted likelihood of occurrence ranges from 0 for cells with habitat furthest from the centre of the calculated niche to 1.0 for cells with habitat closest to the

centre; **(C)** ENFA—Habitat suitability index ranges from 0 for least suitable habitat to 100 for most suitable habitat based on niche preferences calculated during analysis; **(D)** GARP—Values range from 0 to 20 with 20 indicating that occurrence was predicted in all 20 runs and 0 that it was not predicted on any runs

significant difference in the predictive ability of ENFA and GLM for forest-dwelling bird species. However, this significant difference was identified by comparing the combined outcomes of models for 30 different species rather than by directly comparing the models for individual species. In this study, only a single species was examined, so it may be that the differences between ENFA and GLM are only

significant when compared across a large number of species to take individual variation between species into account. Certainly, in over 20% of species modelled by Brotons et al. (2004) the AUCs of the GLM and ENFA models were similar (within 0.03) or the ENFA had the higher AUC, suggesting a degree of variation between species in the comparative predictive abilities of these approaches. The cause of

Table 4 Results of the correlation between mean predicted occurrences of cells within 12 sub-areas on the West Coast of Scotland (see Fig. 5)

	GLM	PCA	ENFA	GARP
GLM	–	0.93	0.841	0.953
PCA	$P < 0.001$	–	0.894	0.951
ENFA	$P < 0.001$	$P = 0.001$	–	0.957
GARP	$P < 0.001$	$P < 0.001$	$P < 0.001$	–

Statistical Test: Pearsons Correlation Coefficient (top right), P values (lower left)

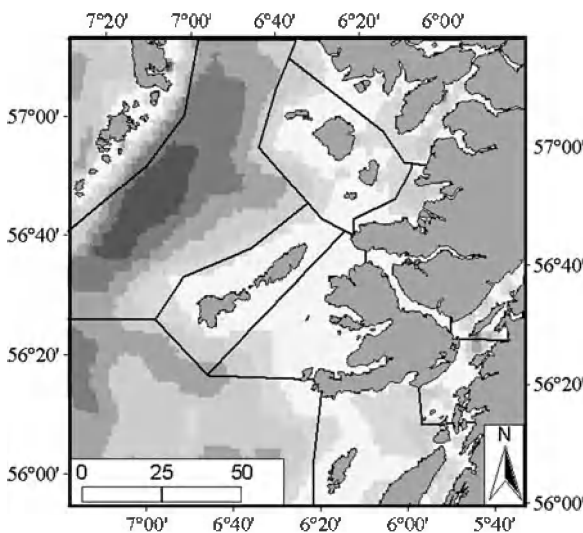


Fig. 5 The 12 sub-areas used to compare the spatial predicted occurrence from the four modelling approaches. These sub-areas were assigned based on coarse oceanographic similarities. Shading shows water depth (white: 0–20 m, black: >300 m)

this variation is unclear, although the majority of these species (six out of seven) had low prevalence (were recorded in a relatively small number of grid cells in comparison to the total number surveyed) and high marginality (how the habitat occupied differed from the average habitat in the study area). As a result, Brotons et al. (2004) suggest that presence-only approaches may be particularly useful for modelling the distribution of such species when absence data are not available. For this study, the ENFA found that the marginality of harbour porpoise was relatively high at 0.907 (see Hirzel et al., 2002 for how marginality is calculated), while the prevalence was relatively low (68 cells out of 679, or 0.10, within the model construction dataset).

However, there is another possible explanation for the difference between the results of this study and that of Brotons et al. (2004). Williams (2003) found that the predictive ability of some ecological modelling approaches varies with species detectability. While presence–absence approaches generally have higher predictive abilities for species with high detectability, they do not perform as well as presence-only approaches when detectability is low (Williams, 2003). Marine species, such as harbour porpoises, may have sufficiently low levels of detectability that the numbers of false absences within the model construction dataset are sufficient to violate the requirement of presence–absence approaches that all absence data are accurate. As a result, the predictive ability of any models generated using presence–absence approaches may be reduced in comparison to ones produced from datasets that do not contain such high numbers of false absences. If low detectability is the underlying reason for the difference between this study and previous comparative studies, this has important implications for modelling the distribution of other marine animals. While it is hard to detect in comparison to many terrestrial species, the harbour porpoise is relatively easy to detect when compared to many other marine species, including other cetaceans such as beaked whales (MacLeod, 2000; Barlow & Gisiner, 2006). However, further research is required to test if this is in fact the case.

Even though they may not perform as well as presence–absence approaches, all the presence-only models applied here provided models with significantly greater predictive ability than random models. In addition, the predicted spatial distribution of these models was very similar to that predicted from the presence–absence model. Therefore, these approaches could potentially allow presence data collected opportunistically, non-systematically or held in databases collated from surveys using incompatible methods to be used to investigate a species distribution. In particular, presence-only approaches may be useful when a species occurrence needs to be understood to allow potential environmental impacts to be assessed and conservation strategies developed in the short term rather than waiting for logistically complex, time-consuming and expensive systematic surveys to collect data of sufficient quality for presence–absence approaches

to be applied. However, clearly due caution is necessary since models based on unrepresentative (biased) surveys could generate misleading results. This can be avoided, even if the quality of the survey is unknown, by adequate testing of the model's predictive ability, although assessing the accuracy of presence-only models can be problematic. The PCA approach requires absence data to test the predictive ability of the model and to identify the best combination of variables to use to model species distribution. This can be a sub-sample of the total available data and, if they can be identified, the most accurate absence data can be assigned to the testing dataset. For example, for harbour porpoises, it would be possible to use data collected under the best conditions, such as sea state zero, when they are most detectable and when absence data may be most accurate (Palka, 1996) to test the models, while still allowing presence data collected under poorer sightings conditions when detectability is lower to be used for model construction.

Neither ENFA nor GARP necessarily require any absence data and both rely on internal verification procedures to test whether a model has a high predictive ability (jack-knife cross validation) and as a result, there is always the possibility that models produced using these approaches, while internal verification suggests a good fit to the data, may not be biologically sensible due to unidentified biases in the presence data associated with the way they were collected. Both approaches assume that the presence data are representative of the species' niche in terms of the EGVs used in the model. If this is not the case, the model may under-predict species occurrence in some locations. While this is unlikely to be a problem with very large datasets, such as those used by Hirzel et al. (2002), this is more likely to be a problem with small datasets. Therefore, when applying these modelling approaches, particularly to the small datasets that likely be available for hard-to-detect marine species, it is important to consider this possibility and try to ensure that the presence data are likely to be representative of the species niche in terms of the EGVs to be used for modelling. If, for some reason, it is suspected that a certain EGV is under-represented in the presence data, it may be prudent to exclude that EGV from any presence-only modelling.

One possible solution to this limitation of using the results of presence-only models for conservation and/

or management purposes is to conduct surveys to specifically test the models' predictive ability. This could involve intensively sampling a representative, but small, portion of an area of interest in order to use the data to assess how any model performs. This combination of presence-only modelling followed by the collection of a data set to specifically test the models' performance from a more limited, but representative, area would potentially allow much greater use to be made of currently available data sets which contain only locational records, rather than presence-absence records, while still retaining a strong assessment criterion for the model's predictive ability. With specific reference to cetaceans, such surveys could be conducted from platforms of opportunity, such as passenger ferries or research vessels conducting other activities, as long as they pass through representative areas, and this would keep costs to a minimum.

However, there may be circumstances where these limitations of presence-only models are not as important. For example, presence-only models may be particularly useful for comparing the relative distributions of a number of species. If these data come from a single data set, it can be assumed that the survey coverage for each species was similar. Therefore, any detected differences in the distributions of species are likely to relate to real differences between them. This may be particularly useful when assessing whether marine protected areas for one species are likely to also protect areas that are important for other species.

Acknowledgements This project would not have been possible without the co-operation of the staff and crew of the Caledonian MacBrayne passenger ferries throughout summer 2003 and 2004. Fieldwork was conducted by both L. Mandleberg and C. Schweder as part of M.Res./M.Sc. degrees at Aberdeen University. S. Bannon and C.D. MacLeod initiated the ferry survey programme used to collect the data, while G. J. Pierce supervised these projects. L. Mandleberg was funded for this M.Sc. by a grant from the NERC. Funding for fieldwork in 2004 was provided by DSTL. G. J. Pierce was supported by the EU under the EnviefH project (CEC FP6 Specific Support Action, 022466).

References

- Austin, G. E., C. J. Thomas, D. C. Houston & D. B. A. Thompson, 1996. Predicting the spatial distribution of buzzard *Buteo buteo* nesting areas using a Geographical

- Information System and Remote Sensing. *Journal of Applied Ecology* 33: 1541–1550.
- Barlow, J. & R. Gisiner, 2006. Mitigating, monitoring and assessing the effects of anthropogenic sound on beaked whales. *Journal of Cetacean Research and Management* 7: 239–250.
- Beerling, D. J., B. Huntley & J. P. Bailey, 1995. Climate and the distribution of *Fallopia japonica*: Use of an introduced species to test the predictive capacity of response surfaces. *Journal of Vegetation Science* 6: 269–282.
- Boyce, M. S., P. R. Vernier, S. E. Nielsen & F. K. A. Schmiegelow, 2002. Evaluating resource selection functions. *Ecological Modelling* 157: 281–300.
- Brotons, L., W. Thuiller, M. B. Araujo & A. H. Hirzel, 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* 27: 437–448.
- Chambers, J. M. & T. J. Hastie, 1997. *Statistical Models in Science*. Chapman and Hall, New York.
- Devlin, B., K. Roeder & L. Wasserman, 2003. False discovery or missed discovery? *Heredity* 91: 537–538.
- Evans, P. G. H. & P. S. Hammond, 2004. Monitoring cetaceans in European waters. *Mammal Review* 34: 131–156.
- Fielding, A. H. & J. F. Bell, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24: 38–49.
- Garcia, L. V., 2004. Escaping the Bonferroni iron claw in ecological studies. *Oikos* 105: 657–663.
- Garcia-Charton, J. A. & A. Perez-Ruzafa, 2001. Spatial pattern and the habitat structure of a Mediterranean rocky reef fish local assemblage. *Marine Biology* 138: 917–934.
- Guisan, A. & U. Hofer, 2003. Predicting reptile distributions at the mesoscale: Relation to climate and topography. *Journal of Biogeography* 30: 1233–1243.
- Guisan, A. & N. E. Zimmerman, 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* 135: 147–186.
- Hirzel, H. A., J. Hausser & N. Perrin, 2000. *Biomapper 2.0*. Laboratory for Conservation Biology, University of Lausanne.
- Hirzel, A. H., V. Helfer & F. Metral, 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling* 145: 111–121.
- Hirzel, A. H., J. Hausser, D. Chessel & N. Perrin, 2002. Ecological Niche-factor analysis: How to compute habitat suitability maps without absence data? *Ecology* 83: 2027–2036.
- Ingram, S. N., L. Walshe, D. Johnston & E. Rogan, 2007. Habitat partitioning and the influence of benthic topography and oceanography on the distribution of fin and minke whales in the Bay of Fundy, Canada. *Journal of the Marine Biological Association of the United Kingdom* 87: 149–156.
- Laake, J. L., J. Calambokidis, S. D. Osmek & D. J. Rugh, 1997. Probability of detecting harbor porpoise from aerial surveys: Estimating $g(0)$. *Journal of Wildlife Management* 61: 63–75.
- Lerczak, J. A. & R. C. Hobbs, 1998. Calculating sightings distances from angular readings during shipboard, aerial and shore-based marine mammal surveys. *Marine Mammal Science* 14: 590–599.
- Lindenmayer, D. B., H. A. Nix, J. P. McMahon, M. F. Hutchinson & M. T. Tanton, 1991. The conservation of Leadbeater's possum, *Gymnodelidius leadbeateri* (McCoy): A case study of the use of bioclimatic modelling. *Journal of Biogeography* 8: 371–383.
- MacLeod, C. D., 2000. Review of the distribution of *Mesoplodon* species (order Cetacea, family Ziphiidae) in the North Atlantic. *Mammal Review* 30: 1–8.
- MacLeod, K., R. Fairbairns, A. Gill, B. Fairbairns, J. Gordon, C. Blair-Myers & E. C. M. Parsons, 2004. Seasonal distribution of minke whales *Balaenoptera acutorostrata* in relation to physiography and prey off the Isle of Mull, Scotland. *Marine Ecology Progress Series* 277: 263–274.
- MacLeod, C. D., C. R. Weir, C. Pierpoint & E. J. Harland, 2007. The habitat preferences of marine mammals west of Scotland (UK). *Journal of the Marine Biological Association of the United Kingdom* 87: 157–164.
- MacLeod, C. D. & A. F. Zuur, 2005. Habitat utilisation by Blainville's beaked whales off Great Abaco, Northern Bahamas, in relation to seabed topography. *Marine Biology* 147: 1–11.
- Ortega-Huerta, M. & A. T. Peterson, 2004. Modelling spatial patterns of biodiversity for conservation prioritisation in north-eastern Mexico. *Diversity and Distributions* 10: 39–54.
- Palka, D., 1996. Effects of Beaufort Sea state on the sightability of harbour porpoises in the Gulf of Maine. Report of the International Whaling Commission 46: 575–582.
- Reutter, B. A., V. Helfer, A. H. Hirzel & P. Vogel, 2003. Modelling habitat-suitability on the base of museum collections: an example with three sympatric *Apodemus* species from the Alps. *Journal of Biogeography* 30: 581–590.
- Robertson, M. P., N. Caithness & M. H. Villet, 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions* 7: 15–27.
- Robison, B. H., 2004. Deep pelagic biology. *Journal of Experimental Marine Biology and Ecology* 300: 253–272.
- Schulze, R. E. & R. P. Kunz, 1995. Potential shifts in optimum growth areas of selected commercial tree species and subtropical crops in southern Africa due to global warming. *Journal of Biogeography* 22: 679–688.
- Sparholt, H., E. Aro & J. Modin, 1991. The spatial distribution of cod *Gadus morhua* L. in the Baltic Sea. *Dana* 9: 45–56.
- Stockwell, D. & D. Peters, 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* 13: 143–158.
- Teilmann, J., 2003. Influence of sea state on density estimates of harbour porpoises (*Phocoena phocoena*). *Journal of Cetacean Research and Management* 5: 85–92.
- Williams, A. K., 2003. The influence of probability of detection when modelling species occurrence using GIS and survey data. PhD thesis, Blacksburg University, Blacksburg, USA.
- Zaniewski, A. E., A. Lehman & J. M. Overton, 2002. Predicting species spatial distributions using presence-only data: a case study of the New Zealand ferns. *Ecological Modelling* 157: 261–280.