Review and Application of Statistical Techniques to Model the At-sea Distributions of Breeding Sandwich Terns

Authors

Chris B. Thaxter, Aonghais S.C.P. Cook, Alison Johnston & Niall H.K. Burton

March 2011

Review and Application of Statistical
Techniques to Model the At-sea
Distributions of Breeding Sandwich Terns

BTO Research Report No. 582

Chris B. Thaxter, Aonghais S.C.P. Cook, Alison Johnston & Niall H.K. Burton

CONTENTS

LIST OF TABLES

Page No.

LIST OF FIGURES

Page No.

EXECUTIVE SUMMARY

1.   A large number of seabird colonies have been designated as Special Protection Areas (SPAs) in the UK due to the international or national importance of their breeding seabird populations. However, to date, the associated marine foraging grounds of these seabirds have not been designated. While this requirement has long been recognised, standard methodologies for assessing the precise boundaries of foraging grounds for designation are still being developed.

2.   The aim of this work was to identify and implement the most suitable approach for modelling the at-sea foraging distribution of Sandwich Terns *Sterna sandvicensis* breeding in the North Norfolk Coast SPA, using data collected on the species' foraging locations from that SPA and associated environmental datasets. This will aid Natural England in the identification of a possible extension to the North Norfolk Coast SPA, and also inform on the Joint Nature Conservation Committee's anticipated UK-wide approach for identifying marine extensions, or additional areas, to breeding seabird colony SPAs.

3.   Here, we initially reviewed the suitability of different modelling approaches to be applied to two breeding Sandwich Tern colonies in North Norfolk: Blakeney Point and Scolt Head, for two types of tern data collected on this species at sea, namely (1) 'presence-only' visual tracking data of individual breeding birds and (2) boat-transect data which can be used to estimate abundances or densities.

4.   We proposed two methodologies that should be used for both tracking and transect datasets to produce probability of occurrence surfaces representing the foraging distributions of Sandwich Terns off the North Norfolk Coast: mixed-effect General Additive Models (GAMMs) and Maximum Entropy (MaxEnt). These methods were selected on the basis of suitability of the techniques for the data in question, and successful documented application of these techniques in similar circumstances. For both approaches, the breeding colony at Blakeney Point was initially modelled. These results were then applied to the data from the Scolt Head colony, for model validation and to assess predictive accuracy. For further cross-validation, the reverse approach was then also undertaken.

5.   For the boat-transect data, foraging locations of Sandwich Terns were recorded, which can provide information on both the presence-absence of foraging terns, and tern abundance or density estimates. Hence an abundance surface from GAMMs and a probability of occurrence surface from MaxEnt were produced. These were compared to the equivalent foraging probability surface produced from the tracking data where birds were recorded foraging. We also produced probability surfaces from GAMMs using all the transect data. Transect data encompassed a portion of potential non-breeders/failed birds whereas tracking data was for breeding birds only, thus caution is needed when interpreting the results.

6.   A third approach, Multi-Adaptive Regression Splines (MARS), was initially recommended for trial analyses; however, eventual time constraint limitations prevented this technique being tested.

7.   Both GAMMs and MaxEnt approaches were applicable for the tracking datasets. However, comparison of AUC showed that models varied in their predictive ability; this was most evident for the GAMMs. Using all the data, the most important predictor variables of occurrence were: distance to colony and depth for MaxEnt and distance to colony, salinity and depth for binomial GAMMs (where birds present = 1 and absent = 0).

8.   GAMMs worked less well than MaxEnt for the transect dataset, proving a poor fit for the data and severely under-predicting the distribution of foraging areas. MaxEnt provided a better set of predictions for the transect data, but overall MaxEnt performed better for the tracking than the transect data.

9.   Despite the poorer model fit of GAMMs compared to MaxEnt, comparisons between the two approaches using foraging locations from both datasets revealed that both predicted similar areas of occurrence.

10.   These results suggest that modelling habitat association of Sandwich Terns can be used to identify potential areas for inclusion within an SPA colony-extension. The same models could be applied and tested for Sandwich Tern data elsewhere in the UK.

1. INTRODUCTION

A large number of seabird colonies have been designated as Special Protection Areas (SPAs) in the United Kingdom due to the international or national importance of their breeding seabird populations. However, to date, the associated marine foraging grounds of these seabirds have not been designated. While this requirement has long been recognised, standard methodologies for accurately assessing the boundaries of foraging grounds for designation are still being developed.

The Sandwich Tern *Sterna sandvicensis* is a feature of 19 SPAs in the UK (Stroud *et al.* 2001). Several years worth of data have been collected on the at-sea foraging locations of birds breeding at colonies at Scolt Head and Blakeney Point, within the North Norfolk Coast SPA. These data have been collected by Econ Ecology in order to assess the impact on the terns of proposed offshore wind farms. Similar data have also been collected by Econ Ecology for Common Terns *Sterna hirundo* also breeding within the North Norfolk Coast SPA and Little Terns *Sternula albifrons* breeding at the Great Yarmouth North Denes SPA.

The marine SPA team of the Joint Nature Conservation Committee (JNCC) have further undertaken boat-based visual tracking of breeding terns of various species from various colonies around the UK. In addition to these data, further information on the at-sea locations of terns from several colonies around the UK has been collected through transect surveys of seabirds conducted for studies addressing the environmental impacts of proposed offshore wind farms.

Radio-tracking data have previously been used to suggest seaward extensions to SPAs for rafting Manx Shearwaters *Puffinus puffinus* in the UK (Wilson *et al.* 2009). While the data available for terns have also provided great detail of the foraging locations used by terns from the particular locations studied and could similarly be used to suggest seaward extensions to those particular SPAs, the foraging distributions of terns around the rest of the UK coast are much less well understood.

There is thus a need for a standard approach that can accurately describe tern foraging distributions where these are known and which can therefore, with some degree of confidence, be used to predict likely tern foraging distributions where these are not known and so be used to describe areas around the UK worthy of statutory designation by Natural England and the other country agencies.

To meet this need, JNCC have an ongoing programme (currently up to the end of 2012) of fieldwork, data collation and analysis that will inform the development of such an approach. As part of this programme, JNCC have recently started investigating the potential use of habitat-association modelling approaches for predicting the foraging distributions of breeding terns. This approach will ultimately lead to the identification of those at-sea areas around the coasts of the UK that constitute the most suitable territories for designation, either as extensions or additional areas to existing SPAs with tern interest features.

The aim of this work is to identify and implement the most suitable approach for modelling the at-sea foraging distribution of Sandwich Terns breeding in the North Norfolk Coast SPA, using the data collected on the species' foraging locations from that SPA and associated environmental datasets. This will not only help Natural England in the identification of a possible extension to the North Norfolk Coast SPA, but also inform on JNCC's anticipated UK-wide approach for identifying marine extensions to breeding seabird colony SPAs.

Here, we initially review the suitability of different modelling approaches for the two types of tern distribution data collected to date, i.e. a) 'presence-only' visual tracking data of individual breeding birds and b) boat-transect data (which can be assessed as presence only, presence/absence, or as densities) (Chapter 2). We then applied the most suitable modelling approaches decided upon in this review to produce and compare probability of occurrence and abundance surfaces describing the foraging distribution of Sandwich Terns off the North Norfolk Coast. This analysis was based on a) presence-only tracking data and b) boat-transect data (Chapter 3). A summary of the findings and associated conclusions is provided in Chapter 4.

2.   CHAPTER 2: REVIEW OF MODELLING TECHNIQUES

2.1   Introduction: Datasets Available

An increasing number of studies have examined the environmental and oceanographic variables that determine the at-sea distributions of seabirds, with increasing use of spatial statistical methods (Tremblay *et al.* 2009). A number of statistical approaches exist that are potentially suitable for the modelling of the tracking data and the boat-transect data available for Sandwich Terns from North Norfolk. This literature review aims to assess the suitability of these approaches so that the most suitable techniques can be identified and subsequently used to model the datasets. The resultant models will be used to produce probability or abundance surfaces describing the foraging locations of Sandwich Terns off the North Norfolk coast using presence-only tracking data and boat-based transect data, which can be treated as abundance or presence-absence data.

2.1.1   Boat-transect Data

Between 2004 and 2008 the area surrounding the North Norfolk Coast was the focus of a number of boat based surveys undertaken to inform Environmental Impact Assessments (EIAs) for proposed offshore wind farms. Surveys were carried out between May and June in the Lincs and Docking Shoal wind farms and June to August in the LID and Race Bank wind farms. During this time, a number of systematic transects were surveyed to provide information on species abundance and distributions.

Boat-based methods survey a predefined transect, recording species into distance bands out to 300m, following standard protocols (Camphuysen *et al.* 2004). Thus, observations of birds are recorded as the boat progresses. One assumption of all the modelling approaches is that data points are independent. However, it is likely that these data will not be independent, e.g. because several foraging locations are recorded for the same bird and birds may have individual foraging preferences. If this structure of the data is ignored, then this would directly violate the assumption of independence. Often this lack of independence is tested by examining spatial autocorrelation (exogenous and endogenous) in the data. Exogenous factors include extrinsic factors (climate, soil type etc) that if left unaccounted for, could lead to a more similar occurrence probability in neighbouring sites, due to specific autocorrelation patterns in these extrinsic factors (see Dorman 2006). Their inclusion within models can reduce residual spatial autocorrelation (e.g. Warren *et al.*, 2005). Endogenous spatial autocorrelation arises due to the biology of the species under consideration (e.g. dispersal, interspecific interactions, disturbance) (Dormann 2006), and can be expected to operate at smaller spatial scales (Guisan & Thuiller, 2005). Ignoring these sources of spatial autocorrelation can lead to either biased parameter estimates, or overly optimistic standard errors (Keitt *et al.*, 2002), the latter of which being important for species distribution models when model results are then used for predicting species distributions, e.g. for environmental change scenarios (Dormann 2006). In this study, endogenous and exogenous spatial autocorrelation may exist, and since the models could be used to predict distribution of terns elsewhere, its consideration is of prime importance. Furthermore, since a range of processes may contribute to spatial autocorrelation once also cannot assume that spatial autocorrelation is a small-scale problem, that can simply be overcome by resampling data at a coarser spatial scale (Fortin & Dale, 2005).

Typical treatment of at-sea boat survey data can include adjustments for detectability for birds on the sea surface (Camphuysen *et al.* 2004), accounting for declining detectability with increasing distance (Buckland *et al.* 2001). However, distance correction is only recommended for birds recorded on the sea surface (Camphuysen *et al.* 2004), since oscillations in wave heights mean that visibility is more severely affected for sea observations than for flight observations. For these reasons, adjustments for

detectability have not previously been carried out for terns, almost all of which will have been flying as they forage from the air, only occasionally sitting on the surface (Allcorn *et al.* 2003).

The position of the boat was recorded every 500m, approximately equal to every two minutes. The time encountered and behaviour of all terns was recorded. Snapshot counts were completed every 500m and each tern observation was recorded as occurring within a snapshot or not. These snapshot counts are conducted to avoid an overestimate of bird numbers in flight, and are conducted by observers recording birds both over the transect and within a 300m distance ahead of the ship (Camphuysen *et al.* 2004). Birds recorded in the scan are typically used to calculate densities (when this is the objective of the survey). Density assessments make an assumption that there is no net inflow and outflow within the study area. In particular, data collected between snapshot counts may overestimate the density due to the flux of flying birds through the area with time as the boat sails through it (Allcorn *et al.* 2003). Therefore, these observations between the snapshots shouldn't strictly be used as count data, but can be used (with the snapshots) as presence-absence data. Therefore, we used all data for presence-absence models. However, for abundance we tested models excluding those data in-between snapshots for assessments. However, when snapshots were used in isolation the dataset was greatly reduced (e.g. Docking Shoal 301/1327), thus we had insufficient data available to allow this. Therefore, we chose to include all data for assessment of abundance, but with a caveat to acknowledge the issue of flux stated above. We believe this constituted a sound approach. Distance bands (corresponding to A: <50 m, B: 50-100 m, C: 100-200 m, D: 200-300 m and E: >300 m) were still recorded within the data but as stated, detectability was not incorporated here due to flying birds. Each bird was also recorded as being within one of four height bands (0 m, 0-20 m, 20-120 m and >120 m).

If, after accounting for exogenous spatial autocorrelation, the residuals from the model or original data are autocorrelated (i.e. perhaps related to endogenous autocorrelation), then this may indicate that data are spatially pseudo-replicated and p-values may be spuriously small. One approach commonly used is to check for such autocorrelation is Morans I, and can be conducted for a number of Monte Carlo permutations based on a weight matrix of k-nearest neighbours, using the size of the study area to give a radius. Variograms and correlograms are typically used to check for strong signals of autocorrelation, and can also provide information about an appropriate radial distance over which spatial autocorrelation may act. Following discovery of autocorrelation, it can then be incorporated into models, through appropriate correlation structures; for instance, in the statistical software package R using package 'spdep'), or alternatively, including an autocovariate in the model as an independent variable, reflecting the values of response variable at neighbouring sites within a radius (Dorman *et al.* 2007).

Analyses to obtain general area distribution can be conducted irrespective of habitat covariates. These methods can be useful to understand general distribution of the raw data and have been used previously to define extensions to breeding seabird colony SPAs through targeted boat-survey effort near to seabird colonies (McSorley *et al.* 2003). Identification of behaviours adjacent to breeding SPAs enabled simple area use to be defined for those birds likely to be from that SPA, i.e. bathing, preening and display (McSorley *et al.* 2003; Wilson *et al.* 2009). Semi-variograms can then be used to provide the basis for interpolation, and kriging is a geostatistical method that expresses the autocorrelation described by the variogram to generate a grid of interpolated values (e.g. Bellier *et al.* 2010). In turn, this can then be used to obtain a density surface. Simple kernel density or minimum convex polygons may also be useful initial assessments.

2.1.2 Tracking Data

During the 2006 and 2007 breeding seasons, 69 individual Sandwich Terns were visually tracked from the Sandwich Tern colonies at Scolt Head Island (20 birds in 2006, 16 birds in 2007) and Blakeney Point (33 birds in 2007) (Perrow *et al.* 2006, 2011). For each trip, individuals were followed by boat from their breeding colony out to sea. GPS locations were recorded at least once per minute and at every location at which the bird dived for food. Individual data regarding the sex, age and breeding status of tracked individuals were not available.

A number of assumptions are common to all statistical approaches of the analysis of tracking data. These are that: 1. The tracked individuals are representative of the population as a whole. This may be violated if boats are more likely to follow birds from a single part of the colony, or those flying at a certain height. 2. The tracking of birds does not influence their behaviour. This is not an unrealistic assumption as a number of studies have shown that terns are relatively unaffected by the disturbance associated with boats (Garthe & Hüppop 1999; Skov & Durinck 2001; Perrow *et al.*2006, 2011). 3. The detectability of birds is equal in all habitats. There is a chance that birds may be 'lost' when tracked further out to sea in rougher environments – but here, survey were conducted in sea states of Beaufort scale 3 or less and data were not biased by these fair-weather conditions because Beaufort Scale ≥4 only occurred on 17% days during the study period (Perrow *et al.* 2006, 2011). The assumption that visually-tracked birds are lost "randomly" is also reasonable given the consistency of the above water marine environment. Likewise, in very shallow areas, boats may not be able to follow birds, however, as boats were able to track birds as they left the breeding colony such issues were not encountered in this study. Assuming birds are lost randomly, and not just further from the colony, then there will also be no issue of bias towards any one habitat.

2.1.3 Covariates

A variety of covariates have been collated by JNCC to describe the physical habitat within which the Sandwich Terns forage. These include basic information such as the distance from shore of the colony, likely to be related to the energetic costs incurred by the birds during foraging. Data such as salinity, water temperature and sediment type, have all been shown to be useful proxies for sandeel abundance (Cyrus 1991, Becker *et al.* 1993, Bertolero *et al.* 2005, Paiva *et al.* 2008, Schwemmer 2009). Data regarding the topography of the environment are also available, including the depth and slope of the seabed and the extent to which individual areas have easterly or northerly aspects. These variables are likely to relate to seabed sediment characteristics and the level of suspended sediments within the water column. In Chapter 3, we present output of the modelling procedures chosen, including the selection of variables based on ecological importance.

2.1.4 Autocorrelation

Autocorrelation here refers to the proximity of observations in space which are not independent, and therefore not a result of the process of interest, and it can lead to spurious results, due to pseudoreplication. There are a number of ways that autocorrelation can arise in these data. Firstly, in the tracking data, an individual bird is followed and foraging locations noted (Perrow *et al.* 2006, 2011). Successive foraging locations for the same individual are not independent, because they are likely to be spatially close to each other. In addition, the foraging preferences of an individual in relation to environmental covariates may represent a smaller range than the foraging preferences of the population, and all the foraging locations of an individual will therefore not be independent. In the line transect data, individuals are not uniquely identifiable, and therefore the same individual may be recorded twice in two locations in close proximity on a transect, again violating the assumptions of

independence. A variety of options are available to accommodate autocorrelation explicitly in the modelling and some of these are discussed below in sections 2.2.1 and 2.2.2.

For tracking data, there is a slim possibility that the same individual could have been tracked more than once, which would further reduce independence in the data. Here, we consider each individual tracked to be a random sample from all the individuals in the colony, thus the potential of tracking the same individual twice is not an issue, since this would ordinarily be expected occasionally in random sampling. Additionally, the probability of double-tracking occurring in this study is fairly low, given the large number of birds at the colony.

2.1.5   Selecting the Correct Model Type

The stated objective here is to implement the most suitable modelling approach for at-sea foraging distributions of Sandwich Terns in relation to environmental datasets. To meet this objective, the appropriate output could take a form of either a presence/absence or density surface model predicting where terns are more likely to be recorded foraging in relation to the habitat variables. Thus, in choosing an appropriate model, consideration needs to be given not only to the available data, but also the output desired.

The data here can be broadly delineated into the three main categories: (1) Presence-only data; (2) Presence-absence data; (3) Count data (abundance and density). As such, response variables require particular treatment under the most appropriate statistical method. The use of boat-based surveys means that the resultant data can be treated as all three, allowing a broader range of modelling techniques to be considered. Tracking data are essentially presence-only data and therefore can be treated to create presence-absence data through insertion of pseudo-absences (see section 3). The response variables available for each dataset are given in Table 1.

Table 2.1   Response variables per type of data available for Sandwich Terns

Response variable	Boat-transect	Visual tracking
Presence-only	yes	yes
Presence-absence	yes	requires pseudo-absences
Count(density and abundance)	yes	no

To represent overall distribution, models could be conducted on all of the data, but to indicate likely areas of importance for feeding, foraging locations could be modelled. The latter may therefore be best to prioritise.

2.2   Methods: Potential Statistical Approaches

Several modelling approaches are common to the analysis of both tracking and transect data, because as seen from Table 2.1, both can be modelled as presence-only and presence/absence data. However, the two datasets are not equivalent, even when reduced to the same data type. Technically the same methods may be applied to the two datasets, although the results may be indicating different things, if the data are not treated carefully. For instance, a model of the foraging locations of the tracking data (which is presence only) will elucidate the environmental covariates which affect where terns forage. Care would need to be taken when applying the same model to the transect data (when reduced to presence only), since this includes foraging and non-foraging (i.e. travelling/searching) locations. Thus, both transect and tracking data will need to be filtered in the same manner before comparisons under the same statistical technique can be made. Ultimately, if there are a lack of data in the transect

data, then environmental covariates will be highlighted that affect where terns forage *and* those in flight/searching behaviours, which would then require different interpretation.

Many traditional modelling approaches rely on knowledge not only of where a species is found, but also where it is absent. However, a common feature of some ecological datasets is that only presence data are available. Consequently, a variety of techniques have been developed to model this type of data, for instance, MaxEnt and regression tree approaches.

Here, we list available statistical methods in approximately increasing order of complexity. Since both types of data can be used in most techniques, where applicable we discuss the merit of the techniques to the application of both tracking and transect data.

2.2.1 Simple Hypothesis Testing

Conventional hypothesis testing techniques such as ANOVA (Analysis of Variance), Chi-squared tests and Wilcoxon tests can be used to test whether the use of certain habitats is more or less than would be expected by chance, given the proportion of available habitats (e.g. Skov & Durinck 2000; Roycroft *et al.* 2007; Alldredge & Ratti 1992). These analyses require separating habitats into discrete classes, which can be difficult to produce for continuous environmental variables. Furthermore, the number of categories can impact the results (Aarts *et al.* 2008) and may not represent important biological habitat distinctions for the species under study (Aarts *et al.* 2008; Wakefield *et al.* 2009). Proportions of available habitats are also normally defined in an anthropocentric way, being derived from arbitrary limits of study areas or other non-biological boundaries, although they can be defined from estimates of home range (Aebischer *et al.* 1993) or in other ways (e.g. Aarts *et al.* 2008).

The proportion of data points within each habitat, are assumed to represent the proportion of time that the species spend foraging there. One problem which arises when dealing with proportions of time, is that preference for one habitat, automatically leads to apparent avoidance of others, i.e. there is non-independence in the data, because the proportions sum to one. Such problems are noticeable for tracking datasets in particular, but may also apply to transect data if one considers that to be a snapshot of the distribution which continually moves around the study area. Compositional Analysis deals with the problem of the unit sum and re-configures the data (Aitchison 1986). Statistical analyses (e.g. MANOVA (Multivariate Analysis of Variance)) can then be conducted on the transformed data (Aebischer *et al.* 1993).
When using basic hypothesis testing there are two ways of dealing with spatial and temporal autocorrelation for both transect and tracking data. Data can be removed from the analysis such that the remaining data points are not serially correlated (Swihart & Slade 1985), or p-values can be adjusted to take account of the degree of autocorrelation (Legendre 1993).

2.2.2 Spatial Methods

Spatial methods, as we are referring to them here, are spatially explicit analyses of tracking locations. Several methods can highlight hotspots in distributions of locations, e.g. GAMs (Wood 2006) or kernel methods, but do not fit relationships with environmental variables. Shaffer *et al.* (2009) used Utilization Distribution kernels to characterise spatiotemporal patterns of habitat use. These methods do not in themselves quantify habitat preference, but variation in habitat use. However, this approach could be used to calculate habitat preference (e.g. Wood *et al.* 2000), and as a preliminary analysis to help inform relevant covariates for the modelling procedure (e.g. Olivier & Wotherspoon 2006a; Olivier & Wotherspoon 2006b).

Poisson point process modelling assumes the locations are generated by a point process and therefore, unlike logistic regression, this approach to modelling does not require pseudo-absences. As the number of pseudo-absences increases, however, the results from a logistic regression converge to those from a point process model (Warton & Shepherd 2010). It is possible to conduct point process modelling in standard software, but there is an assumption that the points are independent.

2.2.3   Regression Modelling Approaches

General Additive Models (GAMs) are widely used in habitat modelling (Clarke *et al.* 2003; Moisen & Frescino 2002; MacLeod *et al.* 2008), including for seabirds at sea data, both for raw counts and densities (Huettmann & Diamond 2006). GAMs are extensions of General Linear Models (GLMs) that facilitate inclusion of semi-parametric smoothing functions, allowing for greater flexibility in the shape of the response curve to the predictor variable (Hastie & Tibshirani 1990). However, GLMs too have been used in various instances (Yen *et al.* 2004; Canadas *et al.* 2005; MacLeod *et al.* 2008; Valavanis *et al.* 2008). GAMs in particular are useful because they avoid making untested assumptions on relationships between predictors and the response (Wood 2006), but may produce wide confidence limits on predictions if using the models to predict beyond the range of values included in the model. These methods are available in a wide variety of packages such as SAS and R.

Leading on from the above GAM and GLM approaches, another option is to adopt a two-stage approach (two-step delta model), defining both a presence absence part of the model, with a binomial error distribution, and a density part, with the desired distribution, which can be then multiplied together to give predictions at new predictor values. These are known as hurdle models and are useful when data are zero inflated (Potts & Elith 2006). Such approaches have been used for seabird abundance in colony based studies, in which the principles (i.e. spatial distribution of data) are the same as those at-sea studies (Scott *et al.* 2009). Likewise, such an approach was taken by Heinänen *et al.* (2008) in modelling occurrence and abundance of Arctic Terns *Sterna paradisaea* in southwest Finland at breeding colonies. Here, GAMs were used to model both presence-absence (binomial) and abundance. Although several options exist for dealing with overdispersion (above, e.g. Welsh *et al.* 1996, Martin *et al.* 2005), hurdle models can also be fitted to allow presence-absence and positive parts of the model defined separately, given different processes that may drive each (Potts & Elith 2006), and hurdle models could be better when the data is both zero inflated and highly overdispersed (Heinanen & von Numers 2009). However, such extreme overdispersion was not encountered in this current study.

2.2.3.1 Application to Transect Data

Both GAMs and GLMs have been used to model boat-based transect data in the past. Yen *et al.* (2004) used GLMs to produce a spatially explicit large scale distribution model of the Marbled Murrelet *Brachyramphus marmoratus*. Similarly, Huettmann & Diamond (2006) modelled the distribution of a variety of seabird species in relation to depth, sea-surface temperature and salinity using GAMs.

A variety of model families would be possible for modelling densities under GAMs and GLMs, including corrections for overdispersion using quasi-distributions, negative binomial distributions, zero-inflated models, or modelling more extreme overdispersion by adjusting error-variance relationships (Welsh *et al.* 1996, Martin *et al.* 2005; Potts & Elith 2006; Zuur *et al.* 2007; Heinänen *et al.* 2008). However, another family, the Tweedie family is also available; this family is a form of exponential distribution that allows the user greater flexibility in defining this error variance relationship (Tweedie 1984; Smyth 1996; Smyth & Verbyla 1999), potentially allowing for a better-fitting model.

Boat-based transect data can be treated as both presence/absence and abundance (or as a density) (Table 1), and thus GLMs and GAMs can be produced using both model types, using a binomial family for presence/absence, and Poisson, negative binomial, Tweedie, or zero inflated models for densities. Given the likely output could be a density surface within the defined study area, and thus requires a within-study variation in density assessment, snapshot counts were investigated. However, too few data were available for this assessment, and thus all data are included with a caveat that outputs may be slightly inflated due to flux within transects as the boat moves (see section 2.1). Furthermore, terns are recorded in transects in flight rather on the sea surface, therefore distance correction will not be applied (Allcorn *et al.* 2003; Camphuysen *et al.* 2004). Furthermore, boat-transect data can also be treated as presence only, thus making the data applicable to approaches such as MaxEnt (see below).

2.2.3.2 Application to Tracking Data

With continuous environmental covariates, and a grid of space within which the data fall, logistic regression can be carried out. The response is a '1' if there is a foraging location within each grid cell, and '0' otherwise. Such data treatment is also applicable to transect data for presence/absence. However, this violates assumptions of independence because individual points are not independent, and the data themselves are not independent in space or time.

The correlations within individuals can be accommodated using mixed effects models, where each individual bird is treated as a random effect; likewise, cross-validation as part of the model selection procedure can also help to limit spurious relationships that can otherwise arise due to serial autocorrelation (Aarts *et al*. 2008). Alternatively, serial autocorrelation can be dealt with by using repeated measures, and treating successive tracking locations as serially correlated repeated measures on the same individual (Fieberg *et al*. 2010). This approach has been carried out in a number of recent tracking studies for a number of species (e.g. Lewis *et al.* 2002; Hamer *et al.* 2009; Thaxter *et al.* 2010). More details of this approach are given in Chapter 3.

Tracking data provides information about the locations of individual birds, but no explicit information about where birds are not. It therefore only contains information about presences and not absences, and often data of this type is modelled with pseudo-absences which are used to fit relative proportions of presences. Modelling presence-only data with pseudo-absences can improve the predictive power of the model (Wisz & Guisan 2009), but may introduce very small bias into the results (Keating & Cherry 2004). There are several ways to select pseudo-absences, which then enable the use of logistic models with a presence/pseudo-absence response variable, and measurements of environmental variables at those locations as predictor variables. Pseudo-absences can be selected randomly from a delimited space, for example a home-range, regardless of how the 'availability' of different regions in space to species is increasingly considered (e.g. Aarts *et al*. 2008; Beyer *et al*. 2010). For central place foragers, such as terns during the breeding season (Gaston 2004), availability can simply be defined as a function of distance from the breeding colony or other factors (Aarts *et al*. 2008; Matthiopoulos 2003; Fieberg *et al*. 2010). Distance from the colony can also be included as a covariate, which may account for inaccurate specification of the availability surface (Aarts *et al*. 2008). The number of pseudo-absences can also impact on the results, and the model may be considered to have converged when an increase in the number of pseudo-absence points does not result in a change in the results (Aarts *et al*. 2008). Aarts *et al*. (2008) state that the number of pseudo-absences required to reach this point varies between datasets, but in their dataset twice as many pseudo-absence points as presence points were required.

Johnson *et al*. (2006) propose an alternative solution of using use-availability (Patil & Rao 1978) designs instead of using pseudo-absences. Calculating the likelihood and optimal model using this approach is not straightforward (Lele & Keim 2006; Lele 2009), although can be more powerful (Beyer *et al.* 2010).

2.2.4   Regression Tree Approaches

The use of tree structured methods is becoming increasingly popular in the analysis of wildlife-habitat relationships (e.g. Andersen *et al.* 2000; Karels *et al.* 2004; Schwartz *et al.* 2006; Davidson *et al.* 2010). These methods make no assumptions about either the distribution of the data or the relationships between the independent and dependent variables. Instead, the data are divided into a hierarchical sequence of groups based on the predictive power of the dependent variables. The results are then presented in the form of a tree diagram, which are typically easy to understand and interpret. Furthermore, these approaches are growing in popularity as they are often insensitive to outliers, irrelevant predictors are seldom selected and missing data can be modelled (Elith *et al.* 2008). Such approaches have been applied successfully to the analysis of seabirds at sea data in species including the Marbled Murrelet and Balearic Shearwater *Puffinus mauretanicus* (Yen *et al.* 2004; Oppel *et al.* in prep.). A variety of forms of regression tree are available, including Classification and Regression Trees (CART: De'ath & Fabricius 2000), Boosted Regression Trees (BRT: Schonlau 2005) and Random Forests (Brieman 2001).

Classification and Regression Trees (CART) model a single response variable against one or more explanatory variables by repeatedly splitting the data. At each split the data are partitioned into two mutually exclusive groups, which are as homogenous as possible. The splitting procedure is then applied to each group separately. The objective of this analysis is to partition the response variable into homogenous groups, whilst minimizing the size of the resultant tree. However, a common criticism of CART is that it is highly dependent on the sample of training data used, and often it has difficulty in modelling smooth functions.

Boosted regression trees (BRT) are an extension of CART that aims to improve model accuracy based on the idea that it is easier to find and average through many rough rules of thumb than to find a single, highly accurate prediction rule (Schapire 2003). The first step, as with CART, is to fit a tree that best explains the data. In subsequent steps, the model is updated to contain a tree fitted to the residuals of the preceding tree. This process is stagewise, rather than stepwise, so that the existing trees are left unchanged and only the fitted value for each observation is re-estimated to reflect the contribution of the newly added tree (Elith *et al.* 2008). The final model is a linear combination of all trees. However, this methodology is prone to mis-classification error, and the results can be difficult to interpret for larger trees. Despite these potential problems, it has been shown to outperform CART algorithms on a variety of datasets (Friedman *et al.* 2000). It has also been applied to the modelling of presence-absence data in the marine copepod *Oithona similas* and both the presence-absence and abundance of the Balearic Shearwater (Oppel *et al.* in prep.).

Random Forest models use bootstrapped samples to construct a large number of trees (typically 500-2000) (Breiman 2001; Prasad *et al.* 2006). Each tree is 'grown' from a randomized sample of the independent variables, in the manner outlined above. The results from all trees are then aggregated. Variable importance is evaluated based on how much worse predictions would be were the variable data to be permuted randomly. A key advantage of this approach is that the large number of trees makes generalisation errors extremely limited, thus making over-fitting impossible (Prasad *et al.* 2006).

Regression Tree approaches have been shown to perform well in comparison to more traditional approaches such as GAMs and GLMs, and have successfully been applied to the analysis of seabird at sea data (De'ath & Fabricius 2000; Moisen & Frescino 2002; Munoz & Felicisimo 2004; Yen *et al.* 2004; Prasad *et al.* 2006; Oppel *et al.* in prep.). Approaches such as Random Forests and BRT which are based on the construction of multiple trees are more effective than CART, which is based on a single tree. However, concerns remain about the ability of these approaches to fit smooth terms.

2.2.5   Multi-Adaptive Regression Splines (MARS)

Multi-Adaptive Regression Splines (MARS) is a technique that has been widely used in vegetation science (i.e. Moisen & Frescino 2002; Munoz & Felicisimo 2004; Leathwick *et al.* 2006; Prasad *et al.* 2006, Elith & Leathwick 2007) and has recently been applied to animal ecology (i.e. Yen *et al.* 2004; Leathwick *et al.* 2006; Heinänen & von Numers 2009). MARS builds flexible regression models by fitting separate splines to distinct intervals of the predictor variables. The variables and the interval end points used are selected following an extensive search procedure utilising a special class of basis functions (Prasad *et al.* 2006). As with other methods, resulting outputs from MARS, as with other techniques, can be easily transferred to other computational environments, such as GIS (Leathwick *et al.* 2006; Elith & Leathwick 2007).

Comparisons between MARS and other methodologies, such as CART, Logistic Multiple Regression and GAMS, show that MARS performs as well as, and more consistently than these techniques (Moisen & Frescino 2002; Munoz & Felicisimo 2004; Leathwick *et al.* 2006). Furthermore, MARS is capable of identifying a parsimonious set of environmental correlates of community composition, and robustly modelling species distributions in relation to these variables.

2.2.6   Genetic Algorithm for Rule-set Prediction (GARP)

Genetic Algorithm for Rule-set Prediction (GARP) models are based on occurrence only data with pseudo-absence data generated by randomly sampling points at which the species has not been detected (Stockwell & Noble 1992). Occurrence data are divided evenly into training data (for model development) and test data (for model evaluation). An iterative process of rule selection is used. Initially, a method is selected from a range of possibilities and applied to the training data to develop a rule. This rule is then applied to the test data and pseudo-absence points and the change in predictive accuracy from one iteration to the next is used to decide whether the rule should be adopted.

However, comparison between GARP and other models, such as BIOCLIM, GLMs and Maximum Entropy, suggests that GARP has a relatively low predictive ability (Stockman *et al.* 2006; Peterson *et al.* 2007), although this view has been challenged by others (McNyset & Blackburn 2006). Direct comparisons between GARP and Maximum Entropy found that whilst AUC statistics are broadly similar for both approaches, GARP had a tendency to substantially over-predict species' distributions (Phillips *et al.* 2006; Peterson *et al.* 2007).

2.2.7   Maximum Entropy (MaxEnt)

Maximum Entropy (MaxEnt) models were developed in the field of Statistical Mechanics, but their use has been growing in popularity in ecology. MaxEnt is a general purpose method for making predictions or influences from incomplete information (Jaynes 1957). MaxEnt aims to estimate a target probability distribution by finding the probability distribution that is most spread out or uniform (i.e. has maximum entropy). It does this subject to a set of restraints representing our incomplete knowledge. The available information about the target distribution is presented as real-

valued variables called 'features', representing environmental variables. However, as with other models, it can run into problems predicting values in areas where the range of the environmental variables is outside that observed within the original study area (Phillips *et al.* 2006), but compounded by the fact it is an exponential model. Despite this, MaxEnt approaches have proven very effective at predicting species' distributions (Phillips *et al.* 2006; Shipley *et al.* 2006 Kumar & Stohlgren 2009), and are capable of performing well in spite of small sample sizes (Wisz *et al.* 2008).

2.2.8   Ecological Niche Factor Analysis (ENFA*)*

Ecological Niche Factor Analysis (ENFA) was proposed by Hirzel *et al.* (2002) as a way to circumvent problems associated with 'false absences'. False absences can occur in areas where a species is present but not recorded, or in areas with suitable habitat which have yet to be colonised. . ENFA has been used successfully to model the at-sea distribution of Northern Gannet (Skov *et al.* 2008). However, Skov *et al.* (2008) highlight concerns over using ENFA within a predictive framework as the methodology does not provide any information on causal relationships between species distributions and individual parameters. Furthermore, ENFA is highly sensitive both to the computational algorithms chosen and to the input order of variables (Brotons *et al.* 2004).

2.2.9   Multivariate Analysis

Eigen value-type methods such as Principal Component Analysis (PCAs) don't involve model selection, and are thus a more objective description of habitat preferences (Calenge *et al.* 2005). However, as they are hard to interpret, they are not as suitable for forecasting or predicting.

2.2.10   Hierarchical Bayesian Methods

Bayesian models are a growing field in all areas of ecology, including habitat preference modelling (e.g. Jonsen *et al.* 2005; Patterson *et al.* 2008). The hierarchical model structure can easily incorporate point location observation error and deal with the correlated data structure, which are both inherent in tracking studies (Fieberg *et al*. 2010). Currently Bayesian methods are computationally expensive and require substantial specialist statistical knowledge.

2.2.11   Movement and Habitat Models

Many studies now incorporate the modelling of habitat preference into a model of animal movements (e.g. Christ *et al*. 2008; Smouse *et al*. 2010). The model of habitat preference can thus become dynamic (e.g. Dalziel *et al.* 2008) and depend on the spatially immediately available habitats. State-space models are very useful for modelling interactions between animals and their habitat, including movement (e.g. Jonsen *et al*. 2005; Eckert *et al*. 2008), and machine-learning techniques such as Artificial Neural Networks also enable the fitting of complicated models (Dalziel *et al*. 2008). However, as is the case with hierarchical Bayesian methods, these models can be computationally intensive, and require substantial statistical knowledge.

2.3   Conclusions and Recommendations

2.3.1   Boat-based Data

A variety of techniques are available for the analysis of the boat-based data, which can be treated as presence only, presence-absence or abundance data. However, where absence data are available, approaches, such as GAMs or GLMs, which account for both presence and absence can perform better than those that account for presence only, for example Regression Trees, GARP, MARS,

ENFA and MaxEnt (Brotons *et al.* 2004; Meynard & Quinn 2007), although there are concerns over the transferability of such models (Randin *et al.* 2006). Data from boat-based transects are likely to be auto-correlated, overdispersed and zero-inflated. Within GAMs in particular, there are well established protocols capable of accounting for these issues. Consequently, the distribution of the data obtained from the boat-based transects may, potentially, drive the type of model used.

2.3.2 Tracking Data

In contrast with the boat-based transect data, the tracking data must be treated as presence only. A wide variety of methodologies are available for analyzing presence-only data, and the optimal approach is less clear than that for dealing with presence-absence data. Regression type approaches have been used, with pseudo-absences generated within the data, constrained, for example, to be within an animal's home range to improve the predictive power of the model (i.e. Wisz & Guisan 2009). Alternatives to regression type approaches include regression trees, GARP, MARS, ENFA, MaxEnt and a number of spatial, multivariate and Bayesian approaches. Of these techniques, some can be computationally expensive or produce results that are hard to interpret. Regression trees are widely used, but have difficulty in fitting smooth terms to the data. They may also produce complicated trees which are hard to interpret. Methodologies such as ENFA and GARP have been shown to have relatively low predictive ability in comparison to GLMs. In contrast, both MARS and MaxEnt have been shown to work well in comparison with other methodologies (Moisen & Frescino 2002; Leathwick *et al.* 2006; Phillips *et al.* 2006; Kumar & Stohlgren 2009; Munoz & Felicisimo 2004).

2.3.3 Considerations for Analysis

A number of issues must be borne in mind in potential analysis of both datasets. First and foremost, the observed distribution, both from the tracking data and also from the transect data, may not necessarily be an accurate representation of the habitat requirements for Sandwich Terns. Occupancy of an area does not necessarily represent selection of that area (Beyer *et al.* 2010). This issue is particularly pertinent for the transect data and care needs to be taken to distinguish between areas that are actively selected for foraging, and those that are used incidentally, for example whilst individuals are in transit between the breeding colony and foraging areas. Furthermore, with regards to the tracking data, it is not possible to distinguish between the species' realised niche, the optimal area in which a species is found, and the species' fundamental niche, the total area in which the habitat matches the species requirements some of which is not used as it is sub-optimal or subject to pressures such as predation (Guisan & Thuiller 2005). Consequently, models are likely to represent the habitats optimally selected by Sandwich Terns, rather than the full range of habitats at their disposal. This may not necessarily be transferable between sites. Furthermore, models assume static habitat selection through time. This may not necessarily be the case with habitat preferences varying with weather and throughout the breeding season.

When comparing the distributions obtained from the boat-transect and tracking studies, it is important to consider the scale at which the studies were conducted (Bellier *et al.* 2010). The boat-transect data will refer more generally to overall habitat use, whilst the tracking data refers specifically to finer-scale foraging locations. Furthermore, the transect data are likely to include a number of birds which are either failed or non-breeders and birds that are on passage to or from other colonies. Consequently, direct comparisons between these data must be treated with caution.

For both sets of models, colinearity amongst the explanatory variables is likely to be an issue. To counteract this, pair-wise plots and variance inflation factors will be used to help decide which variables to include. Ultimately, those with the best *a priori* reason for affecting distribution should

be included in models. It is possible to calculate orthogonal linear combinations of variables, for example through the use of PCA, however, this can lead to difficulties in interpretation.

2.3.4   Recommended Modelling Approach

Table 2 provides an overview of the various methods reviewed here and their applicability to the datasets available. Ultimately there is no one correct modelling approach. Therefore, final chosen methods are selected based on their established application in the literature, their applicability to the datasets in question and their use of algorithms that are not computationally intensive.

2.3.4.1 Tracking Data

Two methodologies are taken forward in Chapter 3 to produce probability of occurrence surfaces representing the foraging distributions of Sandwich Terns off the North Norfolk Coast, based on the presence-only tracking data: mixed-effect General Additive Models (GAMMs) and Maximum Entropy (MaxEnt). Mixed effects models were chosen because data were essentially repeated measurements either on individual birds or proposed wind farm areas (see Chapter 3 for more details). For MaxEnt, background locations from the covariate grid are randomly sampled and the conditional density of covariates at presence locations is compared to the conditional density of covariates at these background locations (Elith *et al.* 2011). For GAMMs, similar pseudo-absences will be generated, but not as an integrated part of the modelling. Covariates are modelled in relation to the presence and pseudo-absence data. For each approach, data from the colony at Blakeney Point will initially be modelled.

A third approach, Multi-Adaptive Regression Splines (MARS), was initially recommended for trial analyses; however, eventual time constraint limitations prevented this technique being tested.

2.3.4.2 Transect Data

The boat-transect data are likewise also modelled in Chapter 3 using both MaxEnt and GAMM frameworks. These data can be treated as presence-only, and thus for the MaxEnt approach, data are treated in a similar manner to tracking data. For GAMMs, abundance information are modelled directly to produce a density probability surface, while counts are also converted to provide presence/absence data so that a binomial logistic-regression GAMM could be applied. Where absence data are available, GAMMs which account for both presence and absence data perform better than approaches that account for presence only. Our analyses thus use a subset of transect data for foraging birds only, to produce probability of occurrence surfaces that may be comparable to the probability surface from the tracking data using foraging-only locations.

Table 2.2. Summary of the approaches reviewed, together with advantages and disadvantages of each approach.

Approach	Summary	Advantages	Disadvantages
1. General Linear Models (GLMs)	Fits linear slopes to a response variable, with link function to define specific error-variance relationships	Not computationally intensive; simple output	Parametric constraints, lack of flexibility extracting more complex relationships between predictor and response
2. General Additive Models (GAMs)	Extensions of GLM, that include semi-parametric smoothing functions	Not computationally intensive; simple output; can fit smoothed terms unlike GLM	Can over-fit model; sensitive to over-smoothing, thus producing unrealistic ecological patterns
3. Classification and Regression Trees (CART)	Models response against predictor(s) by repeatedly splitting a starting "tree", and partitioning into homogenous groups whilst minimizing the size of the resultant tree	Copes with any data structure; has simple classification form; does not vary when variables are transformed; robust to outliers	Dependent on training data; overly sensitive to tree structure; has difficulty in modelling smoothed functions; variable combinations not used; optimal tree splits may not be globally optimal.
4. Boosted regression trees (BRT)	Initial tree to explain response~predictor(s) relationship; further stagewise steps update to the residuals of previous tree - new parameters reflect contribution of the newly-added tree	Can handle different types of predictors; no need for data transformation; fits complex non-linear relationships; outperforms CART on a variety of datasets	Prone to mis-classification error and thus predictive performance for individual trees; results difficult to interpret for larger trees
5. Random Forest	Bootstrapped samples to construct multiple trees - each 'grown' from a randomised sample of independent variables; results of all trees are aggregated	Many trees makes generalisation errors limited, and over-fitting less likely; trains rapidly; can identify outliers/anomalies easily; handling of highly dimensional data; good for clustering/classifying	Computationally intensive; cannot handle large numbers of irrelevant features alongside groups of entropy-reducing decision trees
6. Multi-Adaptive Regression Splines (MARS)	Builds flexible regression models by fitting separate splines to distinct intervals of the predictor variables	Ideal where relationships are non-linear; suitable where many variables interact; is non-parametric; out-performs CART and logistic methods; handles large datasets well	Simpler parametric techniques may be preferred if data are linear with lack of multi-colinearity and interactions; parameter confidence limits not calculated directly; cross-validation needed for model evaluation
7. Genetic Algorithm for Rule-set Prediction (GARP)	Generates pseudo-absences, and uses an iterative process of rule selection applied to a training dataset, using a change in predictive accuracy over iterations to decide to adopt the rule	Easily comprehendible; use of multiple models adds to robustness; can deal with a range of relationships in data; can handle correlated variables	As a stochastic method GARP needs averaging of many lengthy runs of algorithm; GARP vs GLMs and Maximum Entropy, suggests low predictive ability, or over-predicts species distributions;
8. Maximum Entropy (MaxEnt)	Estimates target probability distribution by finding the one most spread out or uniform, subject to restraints representing incomplete knowledge	Proven very effective at predicting species' distributions; clean, effective model fits; easy to interpret; can be extended to handle sample biases; can handle small sample sizes	As an exponential model, it can give particularly large predicted values where the range of the environmental variables is outside that of the observed
9. Ecological Niche Factor Analysis (ENFA)	Circumvents problems of 'false absences' that can occur where a species is present but not recorded, or in areas yet to be colonised	Species niches and potential distribution identified within a landscape - useful for tracking data of marine species that have non-random distribution	Predictive limitations due to lack of causal relationships between distributions and parameters. Sensitive to algorithms chosen, and order of input variables
10. Spatial methods	Explicit analyses of tracking locations - e.g. utilisation distribution kernels to characterise spatio-temporal variation in habitat use and habitat preference	Poisson point process modelling does not require pseudo-absences	Point processes have assumption that the points are independent
11. Multivariate Analysis	Eigen value-type methods such as Principal Component Analysis (PCAs), looking for single components that explain general patterns of multiple variables	Don't involve model selection, thus a more objective description of habitat preferences	Hard to interpret, not as suitable for forecasting or predicting
12. Hierarchical Bayesian Methods	Hierarchical Bayesian model structures	Can easily incorporate point location observation error, and deal with the correlated data structure	Computationally intensive
13. Movement and Habitat Models	Modelling of habitat preference into a dynamic model of animal movements (e.g. State-space models, Artificial Neural Networks)	State-space models useful for modelling interactions between animals and their habitat; Machine-learning techniques enable fitting of complicated models	Computationally intensive

3.   MODELLING THE HABITAT USE OF SANDWICH TERNS

3.1   Introduction

In the UK, breeding seabird species are protected at their colonies in Special Protection Areas (SPAs), designated under the EU Birds Directive (79/409/EEC, updated as 2009/147/EC). However, this directive also places an obligation on the UK government to identify marine SPAs for rare, vulnerable and migratory birds.

Sandwich Terns are a designated feature of the North Norfolk Coast SPA, the colonies at Blakeney Point and Scolt Head hosting an estimated 3457 pairs on an annual basis (Stroud *et al.* 2001). Sandwich Terns have a limited capacity to switch their diet to make use of alternative prey items (Furness & Tasker 2000; Stienen *et al.* 2000) and demonstrate distinct preferences in their use of foraging areas (Becker *et al.* 1993). By modelling habitat use by terns in the offshore environment surrounding the North Norfolk Coast SPA we seek to make recommendations about how best to define the marine foraging grounds of Sandwich Terns and thus inform Natural England's identification of a possible extension to the North Norfolk Coast SPA, and JNCC's anticipated UK-wide approach for identifying marine extensions to breeding seabird colony SPAs.

3.2   Methods

3.2.1   Sandwich Tern Data

Sandwich Tern distributional data from the North Norfolk Coast come from two sources.

Tracking data were obtained by following terns on foraging flights from their breeding colony and recording foraging activity using GPS, following the methodology of Perrow *et al.* (2006, 2011). In total 69 individuals were tracked from the Scolt Head and Blakeney Point breeding colonies in 2006 and 2007. Here, we use those foraging locations of birds for further modelling.

To investigate the predictive power of each methodology, initially, models were developed for the foraging locations identified from each colony individually. These models used environmental data extracted from the surrounding area using minimum area polygons developed using ArcGIS. Each model was cross-validated by applying it to the other colony and comparing the resultant values at each point both to the observed data and to the values predicted by the model for that colony. Finally, a "global" model was developed to incorporate foraging locations from both colonies in relation to the study area as a whole. It is these global models that would be put forward for use in making predictions for foraging locations from additional colonies.

In addition to this tracking dataset, a series of boat-based transect surveys were undertaken as part of the Environmental Impact Assessments for the Race Back, Docking Shoal, LID and Lincs wind farms. Surveys were carried out using standard methodology (Camphuysen *et al.* 2004). In total, data were available from 147 surveys across the four windfarms between 2004 and 2007. Data were insufficient to consider wind farm areas individually, consequently data for all four wind farm were combined in a "global" model. For transect data, we tried modelling both data for foraging birds only and all birds together, but these approaches gave very similar results. Therefore, we only present results relating to foraging birds. Resultant models were applied to making predictions for the study area as a whole.

3.2.2 Environmental Data

A range of environmental variables were available for inclusion in the analysis (Table 3.1.). These variables were pre-processed by JNCC to a 1 km² level and covered aspects of the topography, chemistry and temperature of the study area.

3.2.3 Generalised Additive Mixed Models

Both transect and tracking datasets were modelled using Generalised Additive Mixed Models (GAMMs). The finest scale for which habitat data were available was 1 km², consequently bird observations were summed to the 1 km² level prior to analysis. Habitat data were extracted for each 1 km² within the study area using GIS and then matched with observation data using the R 2.11.0 statistical package (R Development Core Team 2010).

For both datasets, data were modelled using GAMMs; for boat transects, we modelled count data with a Poisson distribution (consistent with plots of the data), summing the number of observations per 1 km square, and as presence-absence data with a binomial distribution. Whilst when modelling with count data from transects it is preferable to use snapshot counts, in this instance, insufficient data were available to allow this.

For tracking data, we modelled data as presence-only (and presence-absence). As data were repeated measurements either on individual birds or proposed wind farm areas, models included random effects for these variables. For transect data, wind farm nested within date was fitted as a random effect to account for the fact that data included repeated counts on multiple dates. For the tracking data, 'individual' was fitted as a random effect as data were repeated observations of individuals. As only presence data were available for the tracking data, pseudo-absences were generated by sampling 1 km grid squares from the study area at a ratio of 2 pseudo-absences to every presence, following the methodology of Aarts *et al.* (2008). Initially, pseudo-absences were selected based on availability as function of distance from colony. However, the high concentration of foraging locations within 10 km² of the breeding colony meant an insufficient area was available from which to generate pseudo-absences. Consequently, pseudo- absences were drawn randomly from the appropriate area surrounding each colony.

Forward selection, which adds predictor variables sequentially until no more additions have a significant effect on model fit, was used in final model selection. Cross - correlation plots were used to determine which environmental variables to include in models with a value of 0.7 taken as a cut-off point. Model residuals were assessed for autocorrelation. All models were fitted using the mgcv library in R (Wood 2006). So that model fit could be comparable between MaxEnt and GAMMS, Area Under Curve of Receiver-Operator Characteristic Plot (AUC) values were calculated for the GAMMs using the ROCR package in R (Sing *et al.* 2009). However, this package does not enable the calculation of AUC for non-binary data, consequently no AUC values were calculated for the models of transect abundance data. The AUC value relates the true-positive rate from the model to the false positive rate, the higher the value, the better performing the model.

Table 3.1.   Environmental variables available for modelling.

Group	Variable	Source
Topography	Depth (m)	SeaZone Digital Elevation Model
	Aspect Eastness[1]	SeaZone Digital Elevation Model
	Aspect Northness[2]	SeaZone Digital Elevation Model
	Slope (^{o})[3]	SeaZone Digital Elevation Model
Chemistry	Salinity (%)[4]	Holt (2006)
Temperature	Seabed Temperature (^{o}C)[5]	Met. Office
	Temperature Stratification (^{o}C)[5]	Holt (2006)
	Summer Front Frequency[6]	Miller *et al.* (2010)
Energy	Wave Shear Bed Stress (N/m^2)[7]	JNCC SeaMap
	Current Shear Bed Stress (N/m^2)[7]	JNCC SeaMap
	Sediment Type (coarse sediment, sand and muddy sand, mixed sediment, mud and sandy mud, rock or reef)	British Geological Survey (DigSBS250)
Features	Distance to Shore (km)	Nearest coastline identified from an Ordinance Survey high water polygon
	Distance to Colony (km)	Nearest colony identified from JNCC tern colony maps

[1] Aspect from -1 south to +1 north; [2] Aspect from -1 west to +1 east; [3] incline between adjacent grid cells; [4] sea surface salinity in summer (%); [5]seabed temperature, and surface to seabed temperature difference in summer; [6]summer thermal front probability based on satellite observations of SST; [7]Tidal force stress at seabed from waves and currents; Variables that varied temporally are averaged specific to the periods investigated.

3.2.4   Maximum Entropy Models

Maximum Entropy models for both transect and tracking datasets were constructed using MaxEnt 3.3.3e (Phillips *et al.* 2004). For each model, twenty-five replicates were used for cross-validation. A jack-knife approach was used to measure variable importance. This looks at the predictive ability of models fit for each predictor variable alone, and for models fit for all predictor variables but missing out each one in turn. Initially full models, containing all variables, were fitted. Environmental variables with values outside of the range present within the training data can have a strong impact on subsequent predictions. Consequently, variables in which this was the case were not included within these models. Following this, variables which had the lowest permutation importance – the drop in the AUC value of the model when values of a variable are randomly permuted - and made the lowest overall percent contribution – the sum of the regularized gain from the addition of the relevant variable to each iteration of the training algorithm - to the model were dropped and the associated change in the Area-Under-Curve (AUC) of the receiver operating characteristic (ROC) curve was noted. The final models presented are those from which no more variables can be dropped without affecting the AUC value.

For the tracking datasets, individual models were initially constructed separately for the Scolt Head and Blakeney point colonies. For each colony 50 % of the data were set aside and used as test data whilst the other 50 % were used as training data. Models for each colony were then identified and used to predict the distribution of foraging locations surrounding the remaining colony. To determine the accuracy of these predictions, the strength of the correlation between the modelled and predicted values for each colony was measured. Following this, a model was constructed combining data from both colonies. In the first instance, the data from Blakeney Point were used as training data and the data from Scolt Head were used as test data. For further cross-validation, the reverse approach was then also undertaken.

A MaxEnt model was also constructed for the transect data. In this instance, data from Race Bank and LID were used as training data and data from Lincs and Docking Shoal were used as test data.

3.3   Results

As environmental data were only available at the level of 1 km², all bird observations were summed to this level. For the tracking data 33 birds were recorded feeding within 93 1 km² grid squares from Blakeney Point in 2007, 20 birds were recorded feeding within 71 1 km² grid square from Scolt Head in 2006 and 16 birds were recorded within 134 1 km² grid squares from Scolt Head in 2007 (Figure 3.1a). During transect surveys, 520 birds were recorded feeding within 156 1 km² grid squares (Figure 3.b).

Prior to analyses, environmental datasets were assessed for correlation and spatial autocorrelation using cross-correlation plots and semi-variograms. Correlations were observed between salinity and temperature (r = 0.75), salinity and distance to coast (r = 0.76) and distance to coast and distance to colony (r = 0.70). Correlations between distance to colony and salinity and distance to colony and mean summer temperature were not significant. Environmental variables showed significant spatial autocorrelation over approximately 15 km and this was accounted for by using an exponential spatial autocorrelation term covering the appropriate area within subsequent models (Wood 2006; Pinheiro *et al.* 2009).

3.3.1   Tracking Data

3.3.1.1 Generalised Additive Mixed Models

GAMMs based on tracking data performed reasonably well, with R² values ranging from 0.376 for a model covering both colonies in 2007, excluding distance to colony, to 0.648 for the model for the colony at Blakeney Point in 2007. Predictions arising from models tended to accurately differentiate between observed presences and generated absences (Figure 3.2). Despite this, there was a strong tendency to under-predict the probability of foraging birds being observed within the study area, particularly for birds from Scolt Head in 2006 (Figures 3.1 & 3.3).

(a)   (b)
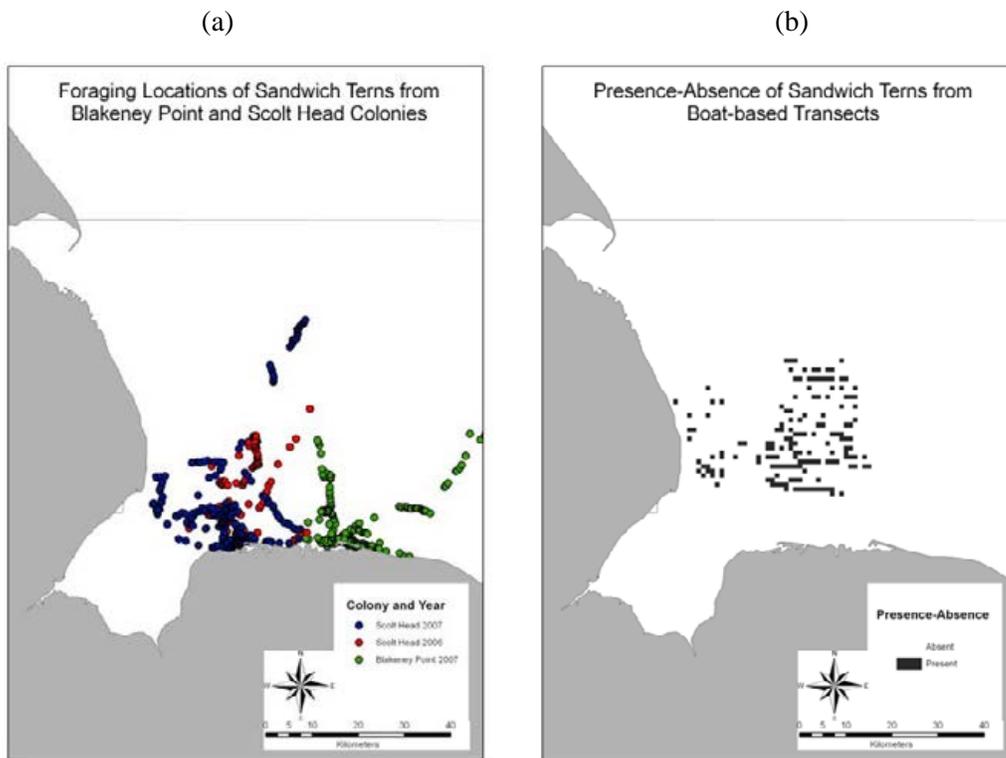
Figure 3.1   (a) Foraging locations of tracked birds from the Scolt Head and Blakeney Point
Sandwich Tern colonies in 2006 and 2007, bounded by study area and (b) Locations of
Sandwich Terns recorded foraging during boat-based transect surveys of the Lincs,
Race Bank, LID and Docking Shoal wind farm areas, bounded by study area.

(a)
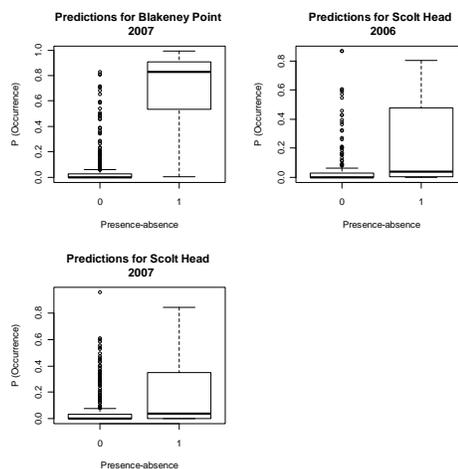
(b)
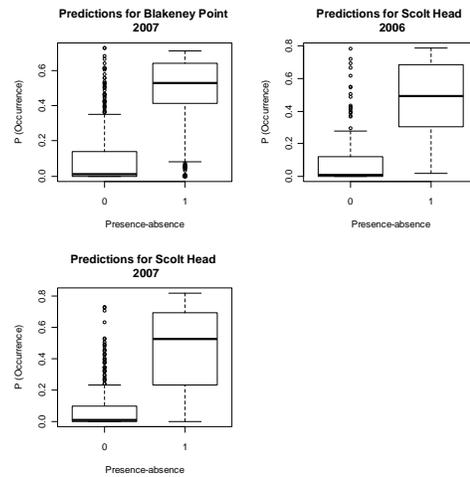
(c)
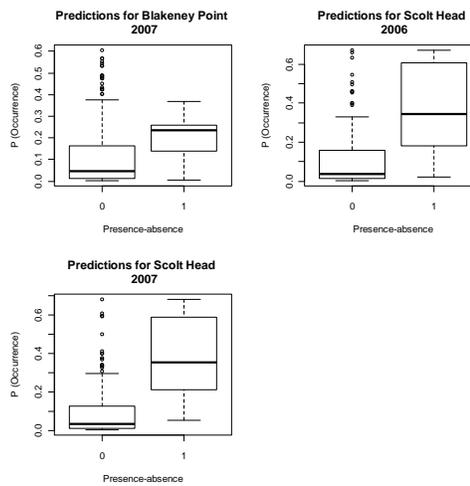
Figure 3.2   Predicted values for observed foraging (1) and pseudo-absence (0) locations for the Sandwich Tern breeding colonies at Blakeney Point in 2007 and Scolt Head in 2006 and 2007 plotted against predicted probability of occurrence of foraging birds from GAMMs built using the foraging locations for (a) the Blakeney Point colony in 2007 (b) the Scolt Head colony in 2006 and (c) the Scolt Head breeding colony in 2007

(a)

(b)

Predicted Probability of Occurrence of Foraging Terns
Using Blakeney Point 2007 GAMM

Predicted Probability of Occurrence of Foraging Terns
Using Scolt Head 2006 GAMM

(c)

Predicted Probability of Occurrence of Foraging Terns
Using Scolt Head 2007 GAMM
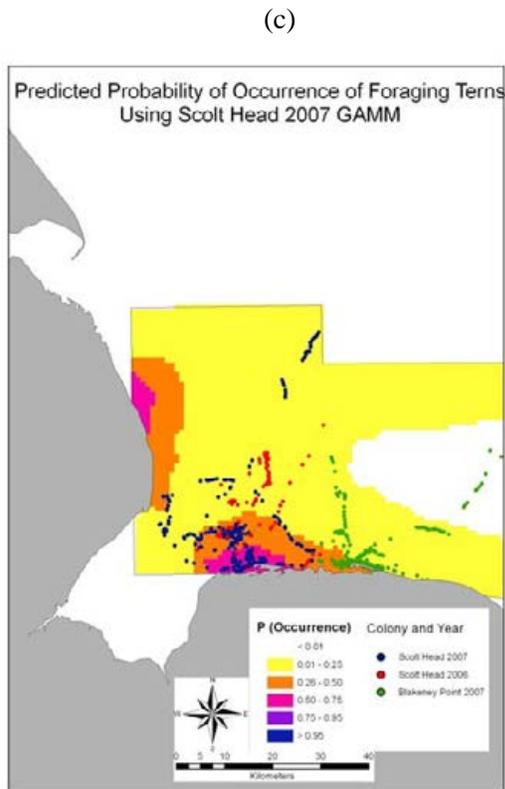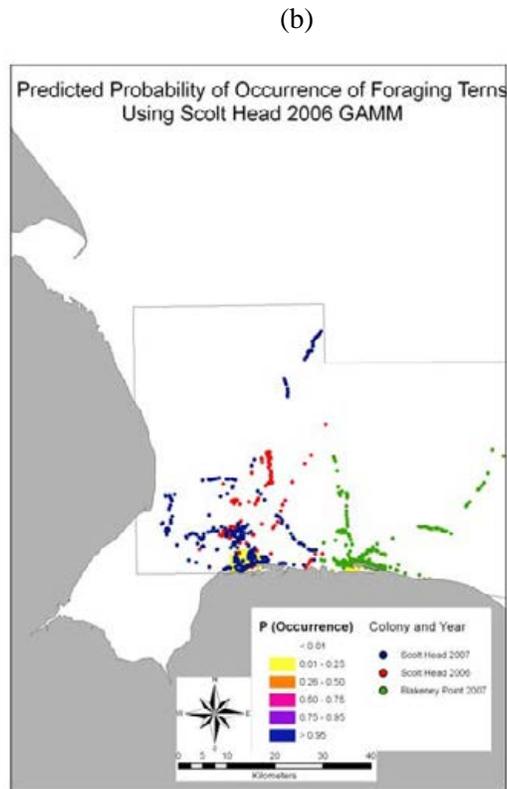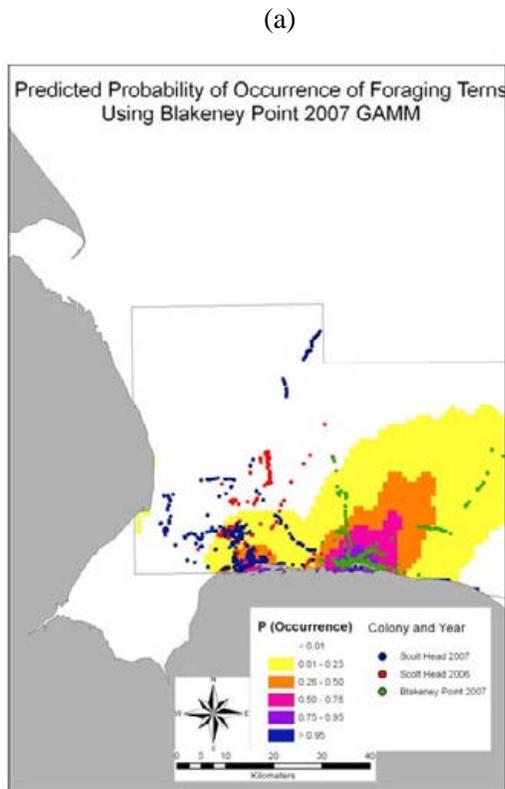
Figure 3.3   Probability of Sandwich Tern foraging activity as predicted using a GAMM and data for terns from (a) Blakeney Point in 2007 (b) Scolt Head in 2006 and (c) Scolt Head in 2007

Three initial models were constructed, one for Blakeney Point in 2007, one for Scolt Head in 2006 and one for Scolt Head in 2007. Of these, the best performing model was that for Blakeney Point in 2007 with an R² of 0.648 and an AUC of 0.949. This model showed that the probability of foraging birds being observed declined as distance from the breeding colony increased (coeff. $-0.47 \times 10^{-4} \pm 6.93 \times 10^{-5}$, P < 0.0001) and increased in response to increased salinity (coeff. 1.27 ± 0.28, P < 0.0001). Neither of these variables showed evidence of smoothing and were consequently fitted as linear terms. These models proved reasonably accurate at predicting observed foraging locations for terns from the Blakeney Point in 2007, with a mean probability of occurrence of 0.68 ± 0.31 for observed foraging locations (Figures 3.2a & 3.3a). In contrast, mean probabilities of 0.19 ± 26 and 0.17 ± 0.23 for observed foraging locations of birds from the Scolt Head breeding colony were calculated for 2006 and 2007 respectively.

The final model for Scolt Head in 2006 had an R² value of 0.412 and an AUC of 0.882. Again, in this case, the probability of foraging birds being observed showed a significant decline as distance from the breeding colony increased (coef. $-1.48 \times 10^{-4} \pm 3.34 \times 10^{-5}$, P < 0.0001) while the model also included a non-significant association with higher temperatures (coef. 1.92 ± 1.3, P = 0.14). Neither of these variables showed evidence of smoothing and were consequently fitted as linear terms. The predictive ability of this model was consistent across both colonies in both years, although probability of foraging birds being observed was under-predicted (Figures 3.2b & 3.3b) with mean probabilities of occurrence of 0.45 ± 0.23, 0.45 ± 0.25 and 0.46 ± 0.20 for observed foraging locations of birds from the Scolt Head breeding colony in 2006 and 2007 and from the Blakeney Point colony in 2007 respectively.
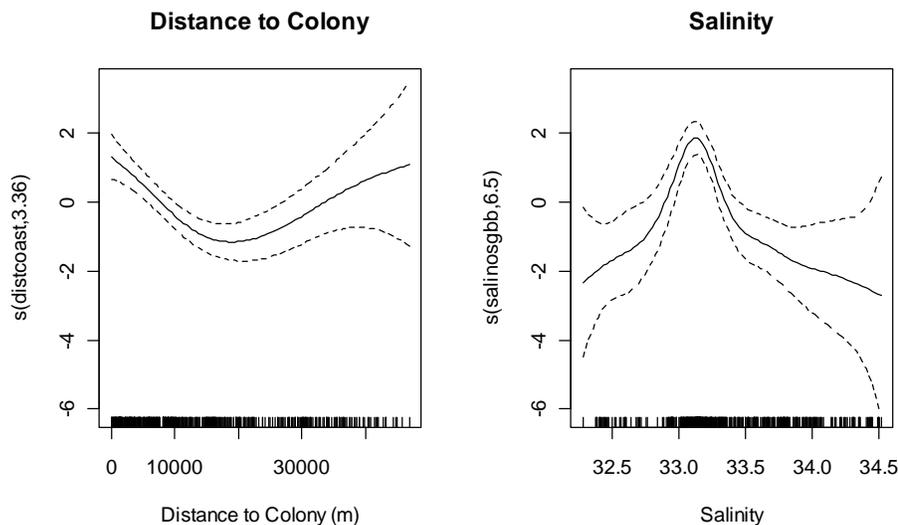
Figure 3.4   Smoothed terms from Generalised Additive Mixed Model for Sandwich Tern Foraging locations from Scolt Head breeding colony in 2007.

The final model for Scolt Head in 2007 had an R² of 0.397 and an AUC of 0.885. Again, the probability of foraging birds being observed showed a significant decline as distance from the breeding colony increased (edf 6.50, P <0.0001) and an association with salinity values of 33 - 33.5 (edf 3.36, P < 0.0001) (Figure 3.4); a non-significant association was also observed with deeper water (coef. -0.0015 ± 0.017, P = 0.92). The predictive ability of this model was consistent between 2006 and 2007 for the Scolt Head breeding colony with means of 0.37 ± 0.20 and 0.38 ± 0.20 of observed foraging locations of birds being correctly predicted for 2006 and 2007 respectively. However, a mean of just 0.19 ± 0.10 of observed foraging locations of birds from the Blakeney Point breeding colony in 2007 were correctly predicted (Figures 3.2c & 3.3c).

In the analysis of foraging locations using 2007 data from both the Scolt Head and Blakeney Point breeding colonies, two models were identified. The first model, which had the highest R^2 and AUC values, 0.466 and 0.887 respectively, showed a significant decline in the probability of foraging birds being observed as distance from the colony increased (edf 2.46, $P < 0.0001$) (Figure 3.5a). However, a mean probability of occurrence of only 0.47 ± 0.28 at observed foraging locations (Figures 3.6, 3.7).

(a)   (b)
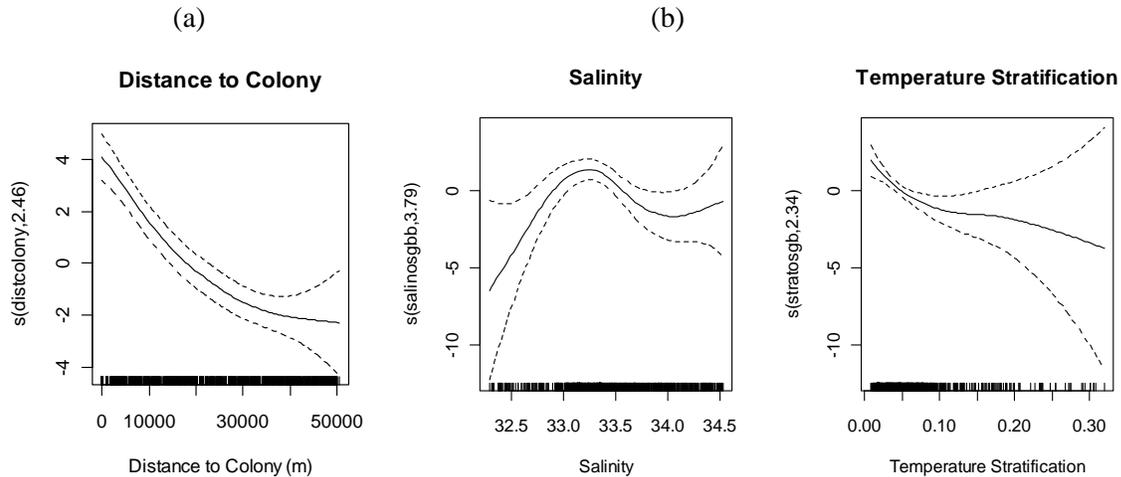
Figure 3.5   Smoothed terms from Generalised Additive Mixed Models for Sandwich Tern Foraging locations from Scolt Head and Blakeney Point breeding colonies in 2007 with (a) Distance to colony as the only explanatory variable and (b) Distance to colony excluded from the model.
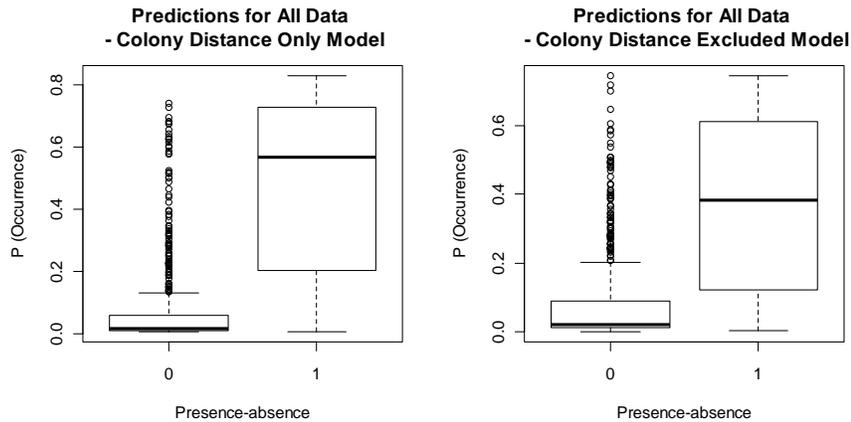
Figure 3.6   Predicted values for modelled presence (1) and pseudo-absence (0) data for Sandwich Tern foraging locations from the Blakeney Point and Scolt Head breeding colonies in 2007 from models combining both datasets. The first model contained distance to colony as its only explanatory variable, whilst the second excluded distance to colony.

Inclusion of additional explanatory variables failed to improve the predictive power of this model, or resulted in a singular convergence. Consequently, a second model, excluding distance from colony as an explanatory variable was fitted to the data. The resultant model had an R^2 of 0.376 and showed a significant association between the probability of foraging birds being observed and salinity levels of

around 33.0 – 33.5 (edf. 3.79, P = 0.0023) and with water exhibiting a lower level of temperature stratification (edf. 2.33, P = 0.0066) (Figure 3.5b). There was also a significant linear relationship indicating a preference for foraging within shallower water (coef. 0.06 ± 0.03, P = 0.0300). However, a mean of only 0.38 ± 0.24 of observed foraging locations of birds were correctly predicted.

(a) (b)
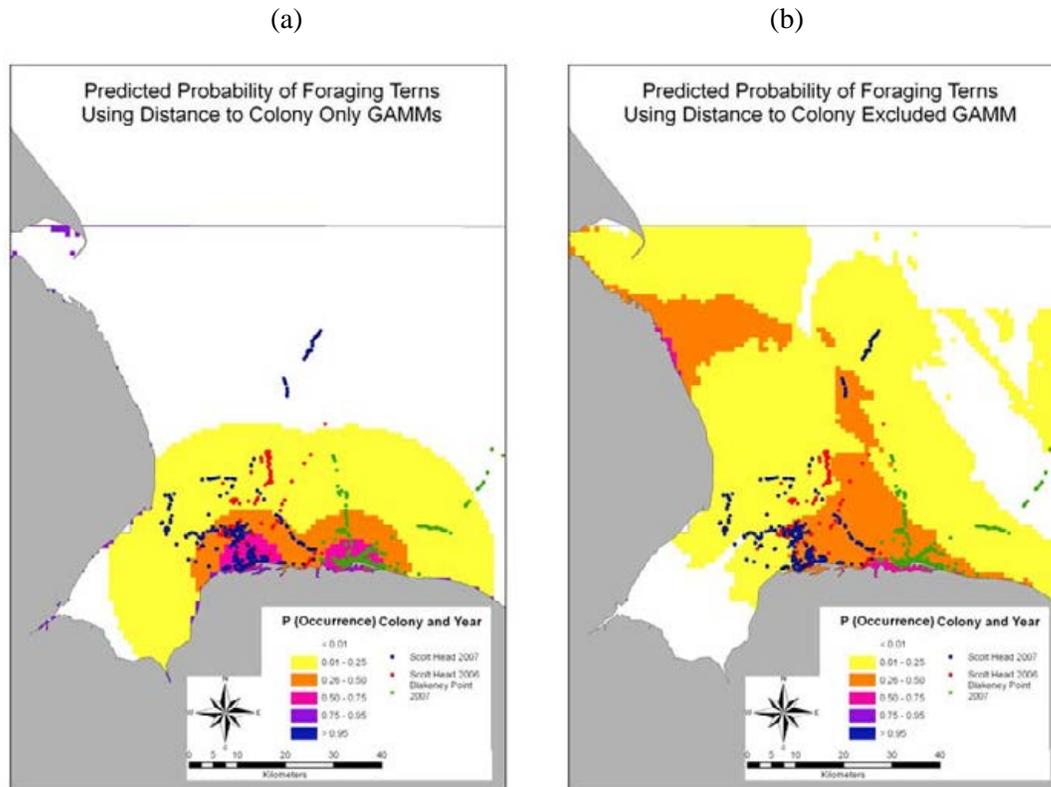
Figure 3.7   Probability of Sandwich Tern foraging activity as predicted using a GAMM for all foraging locations recorded during 2007 (a) with distance to colony as the only variable and (b) excluding distance to colony.

3.3.1.2 Maximum Entropy

MaxEnt models performed well with mean AUC values from 25 cross-validation replicates ranging from 0.925 (± 0.035) in the model using data from Scolt Head in 2007 to 0.979 (± 0.006) for a model using data from both colonies in both years. When models were applied across colonies, the resultant predicated values were fairly consistent (Figures 3.8). However, there was a slight tendency to over-predict the distribution of foraging areas in comparison to the observed data (Figure 3.9). Temperature within the area used by the birds from Scolt Head was outside the range of the area used by birds from Blakeney Point. As such differences can unduly influence predictions, temperature was not included in the models for individual colonies; however, it was considered for inclusion in the model using data from both colonies. This was not the case for any other environmental variable.

The final MaxEnt model using data from Blakeney Point in 2007 had a mean AUC value of 0.976 (± 0.005). Distance to colony was highlighted as the most important variable (Table 3.2, Figure 3.10a). The probability of occurrence of foraging birds was highest within the first 1 km from the colony and declined steeply over the subsequent 10 km. At a distance of 30 km from the colony, the probability of encountering a foraging bird is close to zero. Depth and temperature stratification also had a strong influence on the distribution of foraging areas in this model, with probability of occurrence of foraging birds highest in areas with low levels of temperature stratification and in water less than 10

m deep (Table 3.2, Figure 3.10a). Predictions from this model applied to the Scolt Head area provided a high level of correlation with the predictions from the models for Scolt Head in both 2006 (r = 0.69) and 2007 (r = 0.51). There was however, a slight tendency to over-predict the distribution of foraging locations.
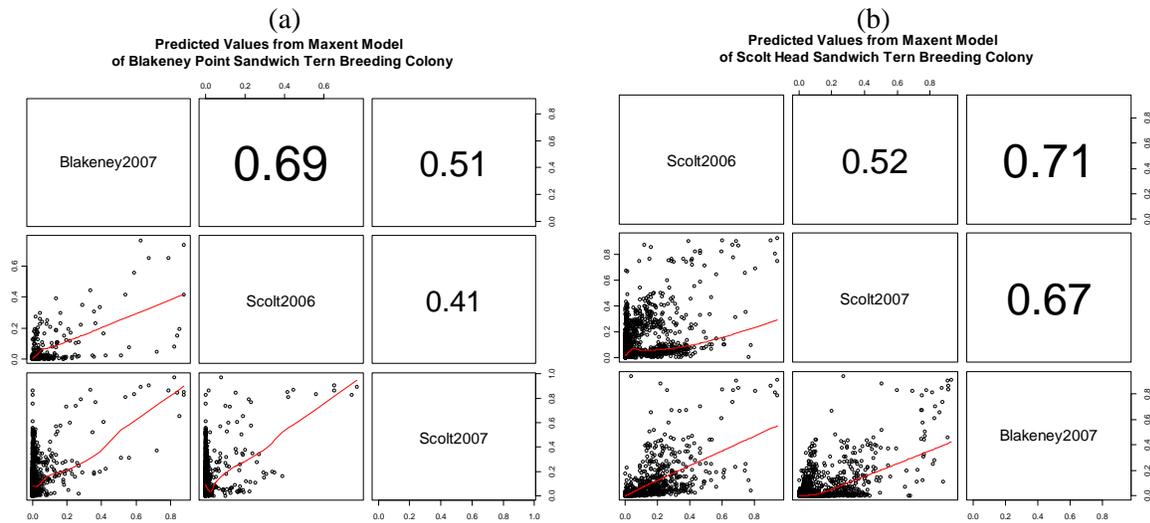

Figure 3.8   The correlation between the foraging probability of (a) Sandwich Terns at Scolt Head breeding colony in 2006 and 2007 predicted using the model for Blakeney Point in 2007 and the foraging probability predicted using the models for Scolt Head breeding colony in 2006 and 2007 and (b) Sandwich Terns at Blakeney Point breeding colony in 2007 predicted using the model for Scolt Head in 2006 and 2007 and the foraging probability predicted using the model for Blakeney Point breeding colony in 2007.

The final MaxEnt models for foraging areas surrounding Scolt Head in 2006 and 2007 had mean AUC values of 0.952 (± 0.027) and 0.925 (± 0.035) respectively. Distance to colony was highlighted as the most important variable in both years (Table 3.2, Figures 3.10 b & c). Again, the probability of foraging birds being observed was highest within the first 1 km from the colony, with a sharp decline up to 10 km. The probability of foraging birds being observed remained between 0.1 and 0.2 between 10 and 20 km from the breeding colony, before approaching 0 at distances in excess of 30 km. As with the model for Blakeney Point in 2007, depth was also important in this model, while for both years the model highlights a tendency for Sandwich Terns to forage over water that is less than 10 m deep. Current shear bed stress was identified as an important variable in the model for 2006 with a tendency for terns to forage in water where current shear bed stress was weak. A similar relationship was observed in the model for 2007, though this variable made a less significant contribution to the overall model. Predictions from these models applied to the Blakeney Point area provided a high level of correlation with the predictions from the model for Blakeney Point (Figure 3.8) – 0.71 and 0.67 for the 2006 and 2007 models respectively. Again however, there was a slight tendency to over-predict the distribution of foraging locations.

The final MaxEnt model considering both colonies in both years had a mean AUC value of 0.979 (± 0.006). Distance to colony was again highlighted as the most important variable with a steep decline in the probability of foraging birds being observed over the first 10 km (Table 3.2, Figure 3.11). Between 10 and 20 km the probability of foraging birds being observed was between 0.1 and 0.2, before falling to close to 0 at distance of more than 30 km. Other variables made a less significant contribution to the overall model (Table 3.2). However, of these variables, salinity and depth were the

most important, with the probability of foraging birds being observed highest in water with a depth of less than 10 m and a salinity value of around 33 (Figure 3.11).
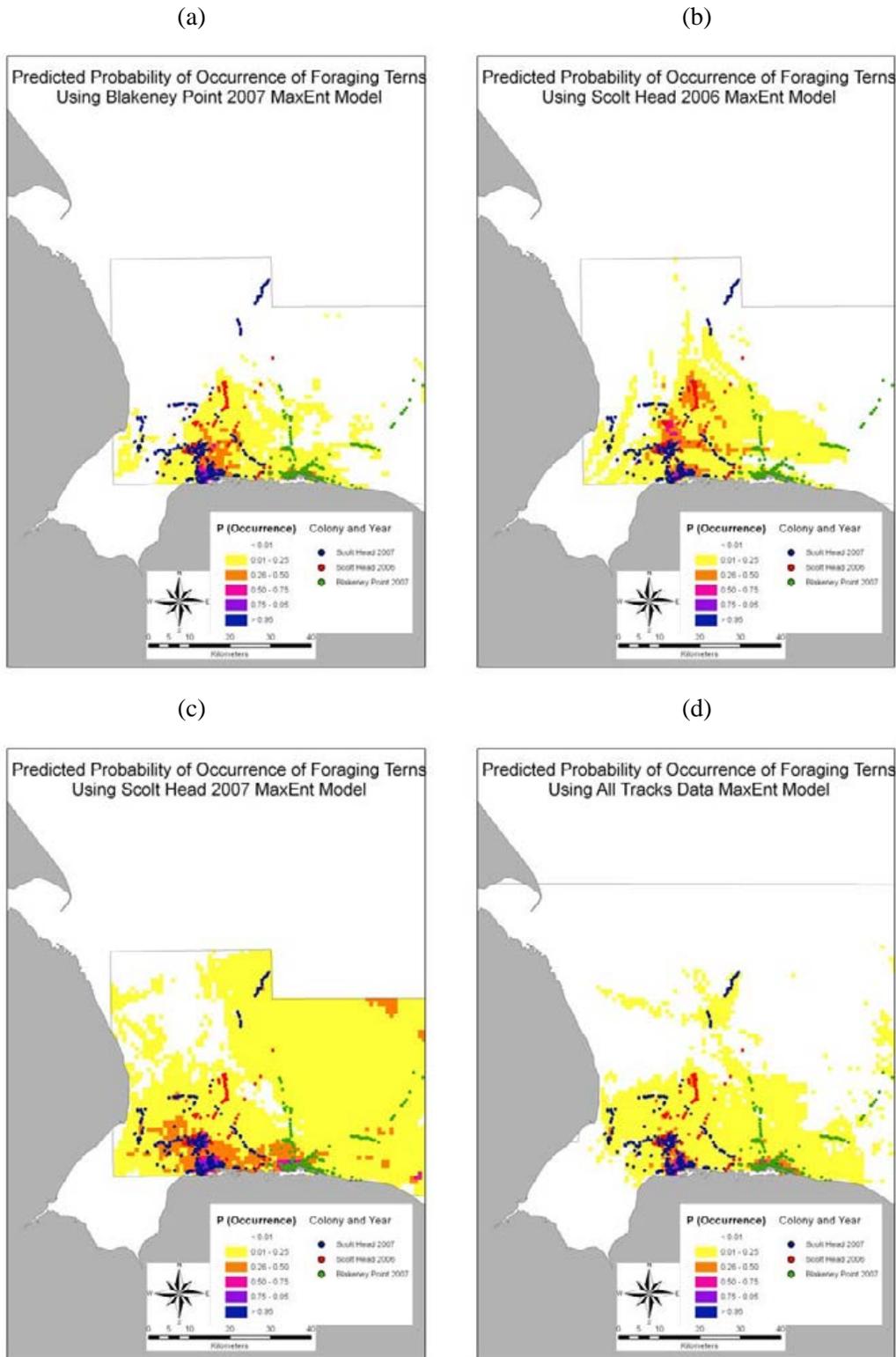
Figure 3.9   Probability of occurrence of Sandwich Tern foraging area based on predictions from the MaxEnt model of foraging locations from (a) Blakeney Point in 2007 (b) Scolt Head in 2006 (c) Scolt Head in 2007 and (d) Both colonies and both years
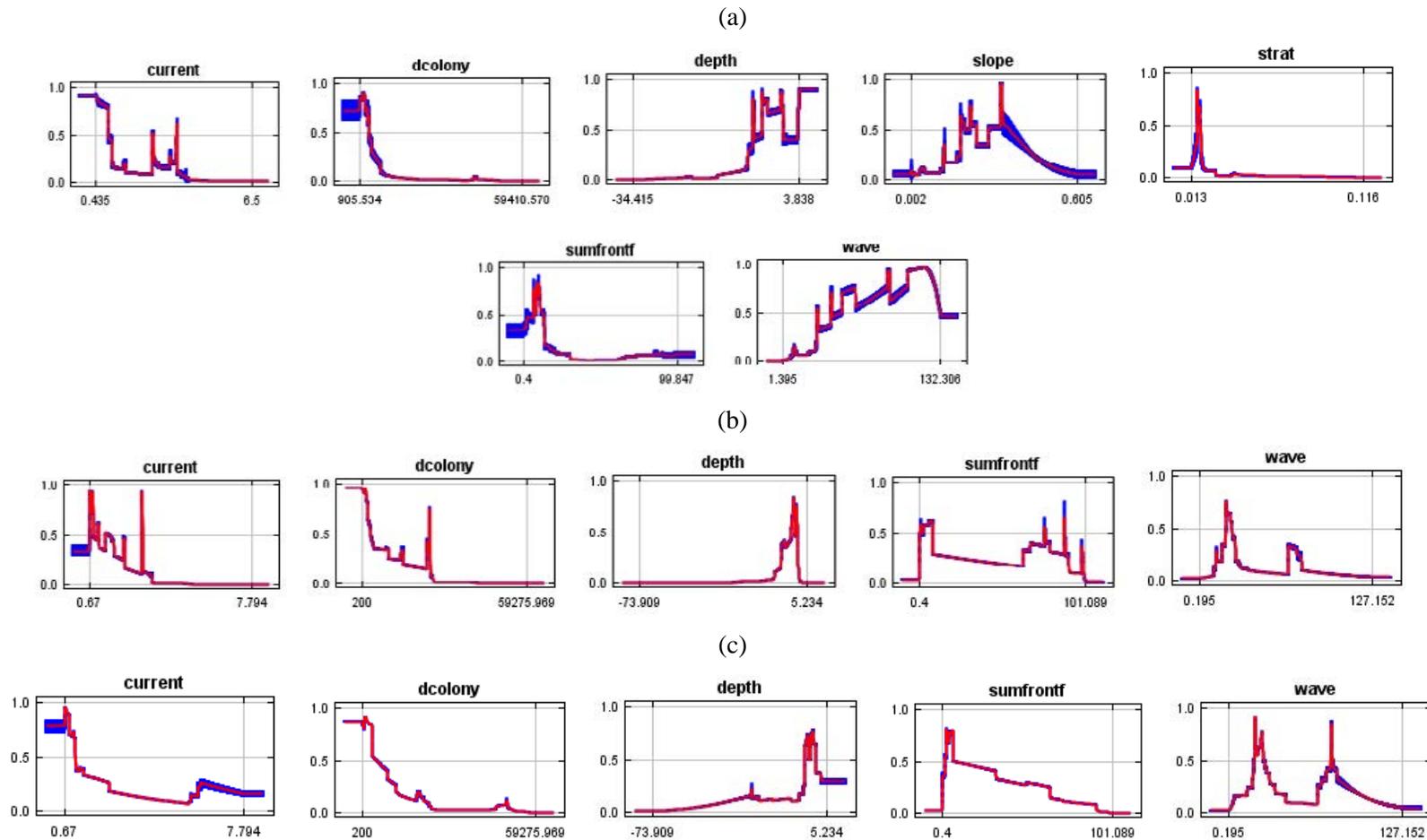
Figure 3.10 Mean response curves for each of the variables included in the final MaxEnt model for foraging locations of Sandwich Terns from (a) Blakeney Point in 2007 (b) Scolt Head in 2006 and (c) Scolt Head in 2007: Current Shear Bed Stress (current), Distance to Colony (dcolony), Depth, Frequency of Summer Thermal Fronts (sumfrontf), Wave Shear Bed Stress (wave), Temperature Stratification (strat), Slope. Red lines indicate mean value from 25 model replicates, blue lines indicate the standard deviation of these replicates.

Table 3.2 Relative contribution of explanatory variables to each MaxEnt model describing Sandwich Tern foraging locations from tracking data.

	Blakeney Point 2007		Scolt Head 2006		Scolt Head 2007		All Track Data	
	Percent Contribution[1]	Permutation Importance[2]	Percent Contribution[1]	Permutation Importance[2]	Percent Contribution[1]	Permutation Importance[2]	Percent Contribution[1]	Permutation Importance[2]
Current Shear Bed Stress	2.7	0.9	15.5	4.7	3.9	5.2	1.6	0.4
Depth	17.1	2.1	10.1	23.5	3.5	2.9	6.4	1.1
Distance to Colony	45.7	90.7	74.1	50.2	74.1	50.2	77.4	87.7
Frequency of Summer Thermal Fronts	2.4	0.4	12	5.7	8.1	23.3	1	0.5
Salinity	n/a	n/a	n/a	n/a	n/a	n/a	4.4	2.4
Slope	4.1	0.7	n/a	n/a	n/a	n/a	2	0.5
Temperature Stratification	n/a	n/a	n/a	n/a	n/a	n/a	3.8	1.6
Wave Shear Bed Stress	7.3	4	4.3	3.4	4.4	18.5	3.3	5.8

[1] The sum of the regularized gain from the addition of the relevant variable to each iteration of the training algorithm
[2] The drop in the AUC value of the model when values of a variable are randomly permuted
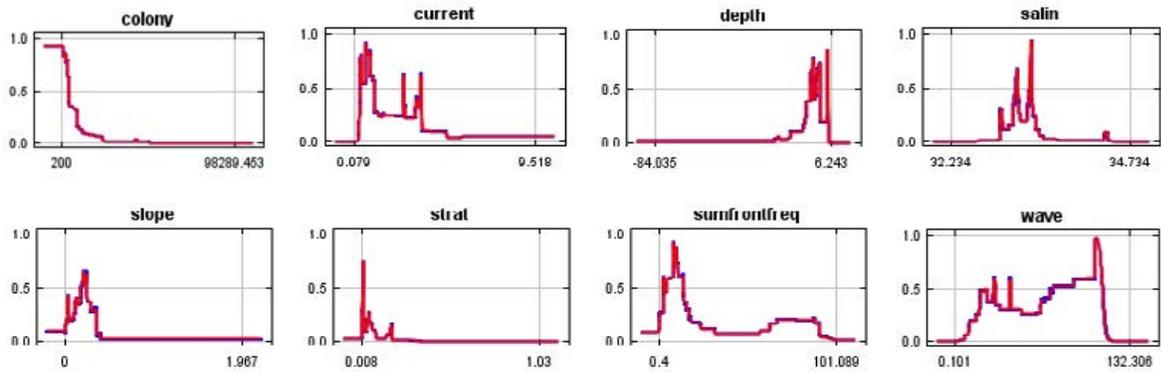
Figure 3.11 Mean response curves for each of the variables included in the final MaxEnt model for foraging locations of Sandwich Terns from the Blakeney Point and Scolt Head breeding colony in 2007: Current Shear Bed Stress (current), Distance to Colony (colony), Depth, Salinity (salin), Slope, Temperature Stratification (strat), Frequency of Summer Thermal Fronts (sumfrontf), Wave Shear Bed Stress (wave). Red lines indicate mean value from 25 model replicates, blue lines indicate the standard deviation of these replicates.

3.3.2 Transect Data

3.3.2.1 Generalised Additive Mixed Models

Using GAMMs, transect data were modelled both as presence-absence data with a binomial distribution and as count data with a Poisson distribution. When used for predictions, both models performed poorly (Figures 3.12 & 3.13). The binomial presence-absence model had an R^2 value of 0.003 and an AUC value of 0.543 and the Poisson count model had an R^2 of 0.002.

(a)   (b)
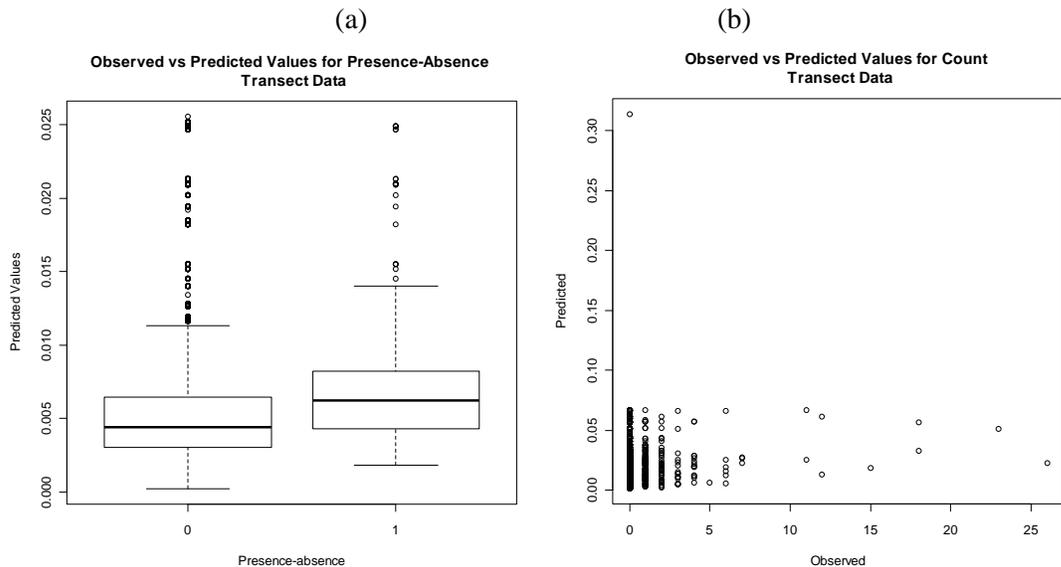
Figure 3.12 Predicted values for (a) observed foraging (1) and non-foraging (0) and (b) observed n birds foraging / km² at locations within the Race Bank, Docking Shoal, LID and Lincs wind farm area survey transects.

When modelled as presence-absence data, the best model included distance to colony (edf. 7.73, P < 0.0001) and seabed slope (edf. 3.57, P = 0.0002) as smoothed terms (Figure 3.14). The model indicated that foraging activity was most likely within areas that were no more than 10 km from the breeding colony, which had a slight slope (Figures 3.13a & 3.14). The mean probability of occurrence for observed presences and absences were 0.022 ± 0.015 and 0.014 ± 0.011 respectively (Figure 3.12a) indicating that models were poor at differentiating areas in which Sandwich Terns were observed foraging from those in which they were not observed foraging

(a) (b)
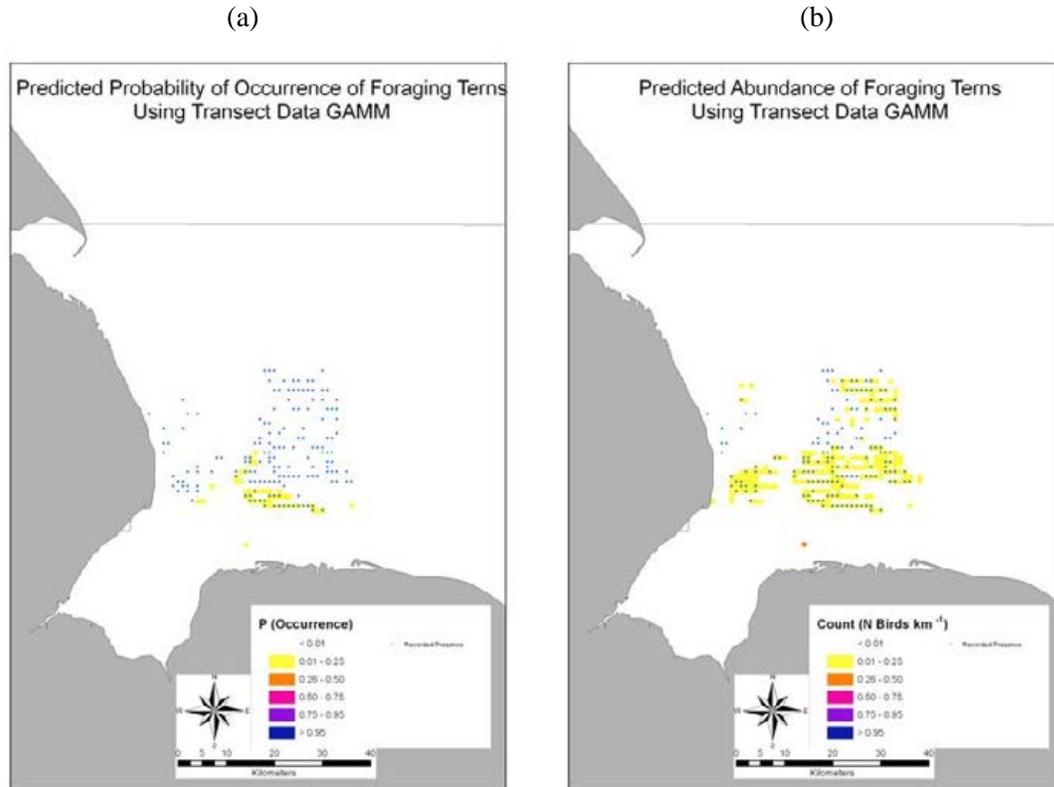
Figure 3.13 Predicted (a) probability and (b) abundance of foraging Sandwich Terns predicted using a GAMM, for boat-based transect surveys of the Race Bank, Docking Shoal, LID and Lincs wind farm areas.

The best model for abundance data included distance to colony (edf. 7.77, P < 0.0001), "Northness" (edf. 5.96, P <0.0001) and wave shear bed stress (edf. 1.00, P 0.0092) as smoothed terms and summer front frequency (coef. -0.014 ± 0.004, P = 0.0031) as a linear term. Foraging was concentrated in areas within 10 km of the breeding colony in areas with a low summer front frequency and in relatively flat areas with low levels of wave stress (Figure 3.15). The model severely under-predicted the abundance of foraging Sandwich Terns within the transect areas (Figures 3.12b & 3.13b).

3.3.2.2 Maximum Entropy

The final MaxEnt model for the transect data had a mean AUC value of 0.958 (± 0.016) Salinity and distance to colony were highlighted as the most important variables. The probability of foraging birds being observed was greatest in areas that were within 10 km of the breeding colony and in which salinity values were between 33 and 33.5 (Figures 3.16 & 3.17). Other variables indicated that the probability of foraging birds being observed was limited to water with a depth of no more than 10 m

and was greatest in warmer water with had weak currents and a low level of temperature stratification (Table 3.3). The model proved a reasonable fit for the data although it had a tendency to over-predict distributions.
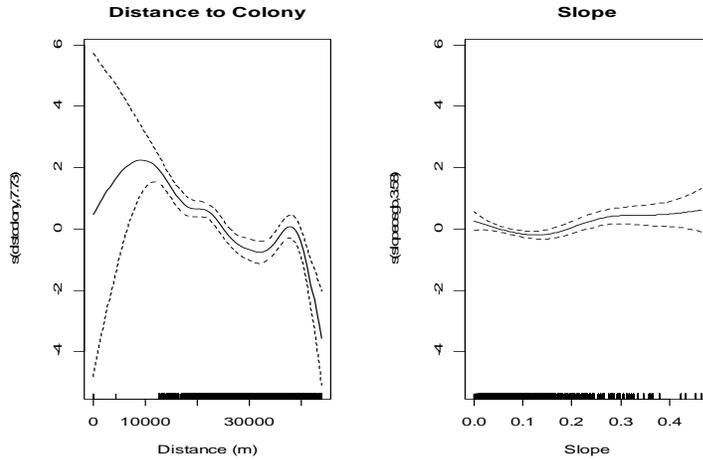
Figure 3.14 Smoothed terms from presence-absence Generalised Additive Mixed Model for Sandwich Tern Foraging locations from boat-based transect surveys of the Race Bank, Docking Shoal, LID and Lincs wind farm areas.
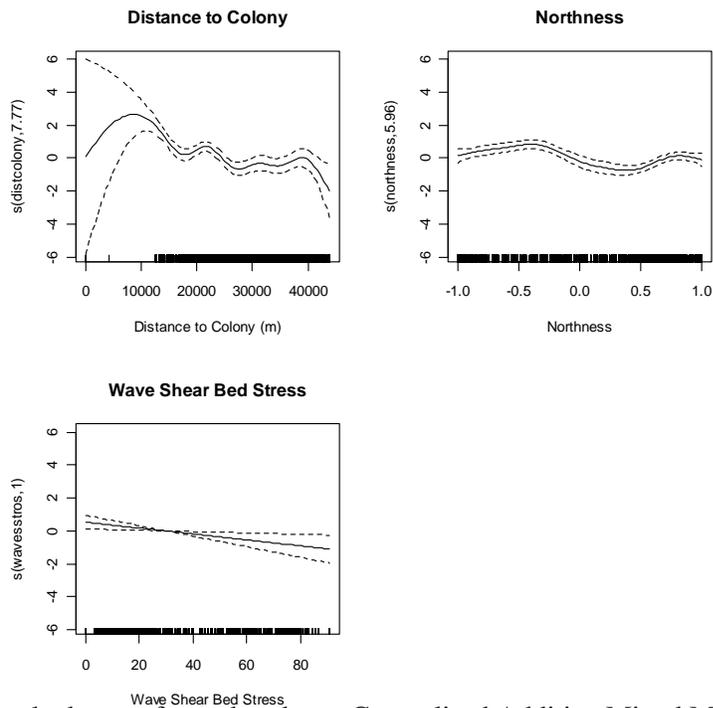
Figure 3.15 Smoothed terms from abundance Generalised Additive Mixed Model for Sandwich Tern Foraging locations from boat-based transect surveys of the Race Bank, Docking Shoal, LID and Lincs wind farm areas.
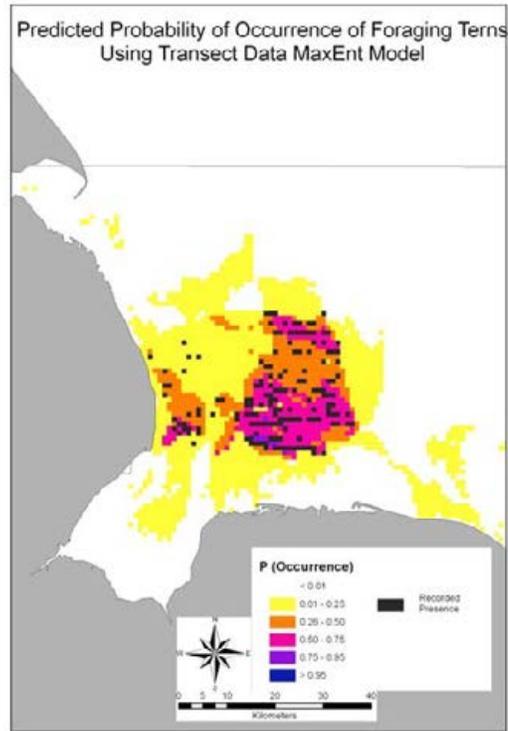
Figure 3.16 Probability of Sandwich Tern foraging activity as predicted using a MaxEnt model for boat-based transect surveys of the Race Bank, Docking Shoal, LID and Lincs wind farm areas.
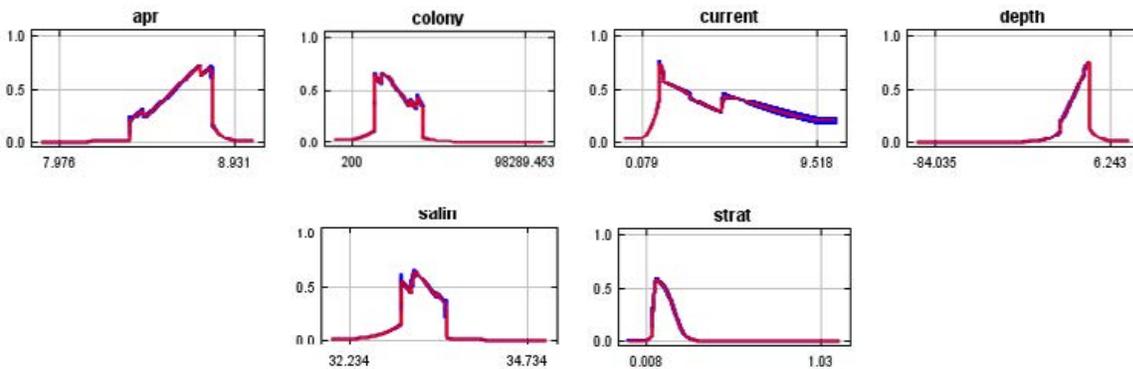
Figure 3.17 Mean response curves for each of the variables included in the final MaxEnt model for foraging locations of Sandwich Terns within the Race Bank, Docking Shoal, LID and Lincs wind farm areas: April Temperature (apr), Current Shear Bed Stress (current), Distance to Colony (colony), Depth, Salinity (salin), Temperature Stratification (strat). Red lines indicate mean value from 25 model replicates, blue lines indicate the standard deviation of these replicates.

3.4. Discussion and Conclusions

For the tracking dataset, there was a reasonable degree of consistency between the results from the GAMMs and the MaxEnt. Both methodologies highlighted a core area in which the probability of encountering a foraging bird was highest close to the colonies. The major differences in the predicted probability of encountering foraging birds between GAMMs and MaxEnt models were observed at greater distances from the colony within areas in which the probability of foraging was low.

For both the tracking data and the transect data, MaxEnt models provided more accurate predictions of foraging locations used by Sandwich Terns than GAMMs. In particular, GAMMs performed poorly for the transect data. In general, GAMMs tended to under-predict the distribution of foraging areas whereas MaxEnt models tended to slightly over-predict the distribution of foraging areas.

Table 3.3   Relative contribution of explanatory variables in the MaxEnt model of all Sandwich Tern foraging locations recorded during boat-based transects of the Race Bank, Docking Shoal, LID and Lincs wind farm areas.

Variable	Percent contribution	Permutation importance
Salinity	33.5	19.5
Distance to Colony	32.2	58.5
Temperature Stratification	12.9	9.1
Depth	7.9	3.8
April Temperature	7.8	4.5
Current Shear Bed Stress	5.7	4.5

Models for tracking data performed reasonably in predicting likely foraging areas. Distance to colony was consistently identified as an important explanatory variable by both the GAMMs and the MaxEnt models. Results from GAMMs were consistent between colonies and years with all three models showing a significant relationship with distance to colony and with salinity, or with temperature which was significantly correlated with salinity.

MaxEnt models were found to have a greater predictive ability than the GAMMs in the analysis of the tracking data. Despite this, there were a number of similarities in the variables identified as important. Distance to colony was again highlighted as a key variable, with few birds feeding in excess of 10 km from the colony. Furthermore, the MaxEnt models highlighted the same salinity values of around 33-33.5 as were identified by the GAMMs using data for birds from Scolt Head in 2007 and both colonies in 2007. These salinity values are similar to those identified as preferential for the sandeel *Ammodytes marinus*, a key prey species for Sandwich Terns (van der Kooij *et al.* 2008). The MaxEnt model appeared to identify a shift in foraging areas used by birds from the Scolt Head colony between 2006 and 2007. However, this may be the result of individual differences in the birds being followed, and potentially an issue related to the sample size of the tracked birds.

Models for the transect data did not perform as well as those for tracking data. In part, this may be because the areas covered by the transects are not as widely used by individuals breeding at the Scolt

Head and Blakeney Point colonies as nearer shore areas. Models for the tracking data consistently identified distance to breeding colony as a key explanatory variable, with most individuals foraging within areas that were within 10 km of the breeding colony. The area covered by the transects is mostly well in excess of 10 km from either Scolt Head or Blakeney Point breeding colony.

It has previously been suggested that habitat heterogeneity may decrease as distance offshore increases (Reise & Bartsch 1990). It may be that there is a lower level of habitat heterogeneity within the transect areas as they are at a greater distance offshore than the areas covered during the tracking studies. A lower level of habitat heterogeneity would make it harder to differentiate between areas that were used for foraging and those that were not. Certainly, initial exploratory analysis at a 1 km², 5 km² and 10 km² resolution failed to reveal significant differences between variables within areas which were used for foraging and those which were not. No further exploratory analyses were conducted.

Whilst the foraging areas identified using the tracking data were concentrated around the coast, those identified using the transect data were concentrated in a more central location offshore. There are likely to be a number of reasons for this. Firstly, unlike the tracking data the months covered by the boat-based surveys extend outside the nesting period. Post-fledging, birds may change their foraging behaviour as they are no longer constrained by the need to return to their nest site. Furthermore, the distance between the study colonies and the transect areas is at the limit of the typical foraging range for the Sandwich Tern (Thaxter *et al.* submitted.). As such, it is likely that a number of the birds observed foraging within the transects may be non-breeders, or from different breeding colonies. Finally, it is possible that since foraging has only been observed in a relatively small number of 1 km squares over a number of years, the model coefficients may be unduly influenced by the values from these squares such that when applied to predictions over a wider area they fail to accurately identify areas outwith the initial extent of the survey.

MaxEnt models show a greater predictive ability, and may therefore provide a better framework for predicting the distribution of Sandwich Tern foraging areas than GAMMs. MaxEnt models were remarkably consistent between colonies and years, with models for individual colonies all identifying foraging areas as being likely to be within 10 km of the colony, over water that was no more than 10 m deep and had salinity values of around 33-33.5. These characteristics were also highlighted by some of the GAMMs applied to tracking data. It should be noted that the models describe the realised niche of foraging sandwich terns as opposed to the fundamental niche of foraging sandwich terns. In practical terms this means that the birds forage over an optimal proportion of the total available habitat. In particular, the distance over which Sandwich Terns forage identified in these models is likely to be an under-estimate of the maximum distance over which they could potentially forage. The model utilising data from both colonies and both years is likely to be more valuable in applying predictions to other areas than any of the models using data from individual colonies and years as it makes use of more information and is consequently likely to prove more robust.

4.   GENERAL DISCUSSION AND CONCLUSIONS

This study has identified and implemented suitable approaches for modelling the at-sea foraging distribution of Sandwich Terns breeding in the North Norfolk Coast SPA, specifically for two colonies within the SPA: Blakeney Point and Scolt Head. After a thorough literature review, we applied two approaches (GAMMs, and Maximum Entropy) using data collected on the species' foraging locations from that SPA and associated environmental datasets. These methods have previously been demonstrated as highly applicable to these types of data.

As summarised in section 3.4, MaxEnt performed better than GAMMs providing more accurate predictions of foraging locations of Sandwich Terns, albeit with slight over-predictions. In contrast, tended to under-predict the distribution of foraging areas. The under-prediction of foraging areas by GAMMs is of greater concern than the over-prediction of foraging areas by MaxEnt. Primarily, this is due to the severity of this under-prediction, but also because: (1) these models may be used to identify potential marine extensions to existing SPAs, and such a conservative approach which results in all foraging habitat used being protected is preferable to an approach that results in an area that is too small being protected, and (2) because only a subset of individuals from each colony were observed it is likely that a wider area is used than was recorded during this study.

An advantage of MaxEnt is that it provides clean, effective model fits, and can also handle small sample sizes (Phillips *et al.* 2006, Shipley *et al.* 2006, Wisz *et al.* 2008, Kumar & Stohlgren 2009). MaxEnt also seems particularly well suited to species that are not stationary, temporarily absent, and seems to work well with fewer observations in a large area of suitable habitat. These points may also offer explanations for better model fit for MaxEnt. Furthermore, MaxEnt may have picked up more complex relationships between the distributions and the environmental variables than GAMMs. Whilst for MaxEnt it is not always as intuitive as GAMMs, it may provide a better framework for modelling the distribution of Sandwich Tern foraging areas, and produce predictions that are more readily transferable between areas.

In comparing predictions of foraging locations from models using tracking and transect data some similarities were evident for MaxEnt models. However, models based on transect data missed areas that were closer to the colonies that were identified by MaxEnt. Nevertheless we suggest caution when comparing these methods, not least because transect data may include a proportion of birds that may not be breeders – indeed most survey transects were located further away from the colonies than the more extreme edges of the visual-tracking ranges.

A number of limitations were identified with the datasets used. In particular, the observed distribution, both from the tracking data and also from the transect data, may not necessarily be an accurate representation of the habitat requirements for Sandwich Terns because occupancy does not necessarily translate into selection (Beyer *et al.* 2010). This issue is particularly pertinent for the transect data, thus we distinguished between areas actively selected for foraging, and those that were used incidentally (e.g. during transit). Furthermore, for the tracking data, it was not possible to distinguish between the species' realised niche, the optimal area in which the species was found, and the species' fundamental niche. Consequently, the models are likely to represent habitats optimally selected, rather than the full range at their disposal. As such, this approach should still be used with caution if the models are to be used to provide predictions at other sites, and would need to be validated with further observed data. Finally, the models here assume static habitat selection through time, which may not necessarily be the case for variables such as temperature, salinity gradients that can very over time throughout the season. Dynamic oceanographic modelling is another option (e.g. Skov *et al.* 2008; Schwemmer *et al.* 2009), which incorporates the changing nature of oceanographic input variables to predict species distributions across ranges in these variables (for instance across

tidal states, eddies, velocity and flow components). Such an approach should also be considered in future modelling.

In this study, we also suggested trialling a third technique: Multi-adaptive Regression Splines (MARS) is a technique which has also been successful applied elsewhere, including modelling of data of foraging terns from their colonies (e.g. Heinänen *et al.* 2008). However, eventual time constraints prevented investigation of the method. We would therefore recommend further study of this technique to these data. Other techniques such as regression trees were not tested here, mainly due to likely intensive computer time and time limitations. However, these techniques could also be fully applied to these data, and are currently being trialled at the BTO for other distributional data. A full comparison of all methods is rare in the literature, but would be highly valuable in this instance given the importance of selecting the most appropriate areas for SPA extensions.

MaxEnt was the preferred modelling framework in this study. However, we would also recommend that in any future modelling study, multiple statistical approaches should be tested. The results presented here are encouraging and will not only help Natural England in the identification of a possible extension to the North Norfolk Coast SPA, but also inform on JNCC's anticipated UK-wide approach for identifying marine extensions to breeding seabird colony SPAs.

Acknowledgements

References

Aarts, G., MacKenzie, M., McConnell, B., Fedak, M. & Matthiopoulos, J. 2008. Estimating space-use and habitat preference from wildlife telemetry data. *Ecography* **31**: 140-160.

Aebischer, N.J., Robertson, P.A. & Kenward, R.E. 1993. Compositional analysis of habitat use from animal radio-tracking data. *Ecology* **74**: 1313-1325.

Aitchison, J. 1986. *The statistical analysis of compositional data*. Chapman & Hall, London.

Allcorn, R., Eaton, M.A., Cranswick, P.A., Perrow, M., Hall, C., Smith, L., Reid, J., Webb, A., Smith, K.W., Langston, R.H.W., Ratcliffe, N., 2003. A Pilot Study of Breeding Tern Foraging Ranges in NW England and East Anglia in Relation to Potential Development Areas for Offshore Windfarms. Sandy: RSPB/ WWT/JNCC

Alldredge, J.R. & Ratti, J.T. 1992. Further comparison of some statistical techniques for analysis of resource selection. *Journal of Wildlife Management* **56**: 1-9.

Andersen, M.C., Watts, J.M., Freilich, J.E., Yool, S.R., Wakefield, G.I., McCauley, J.F. & Fahnestock, P.B. 2000. Regression-tree modeling of desert tortoise habitat in the central Mojave Desert. *Ecological Applications* **10**: 890-900.

Becker, P.H., Frank, D. & Sudmann, S.R. 1993. Temporal and spatial pattern of Common Tern (*Sterna hirundo*) foraging in the Wadden Sea. *Oecologia* **93**: 389-393.

Bellier, E., Certain, G., Planque, B., Monestiez, P. & Bretagnolle, V. 2010. Modelling habitat selection at multiple scales with multivariate geostatistics: an application to seabird in open sea. *Oikos* **119**: 988-999.

Bertolero, A., Oro, D., Martinez Vilalta, A. & Angel Lopez, M. 2005. Selection of foraging habitats by Little Terns *Sterna albifrons* at the Ebro Delta (NE Spain). *Revista Catalana d'Ornitologia* **21**: 37-42.

Beyer, H.L., Haydon, D.T., Morales, J.M., Frair, J.L., Hebblewhite, M., Mitchell, M. & Matthiopoulos, J. 2010. The interpretation of habitat preference metrics under use-availability designs. *Philosophical Transactions of the Royal Society B* **365**: 2245-2254.

Brieman, L. 2001. Random forests. *Machine Learning* **45**: 5-32.

Brotons, L., Thuiller, W., Araujo, M. B. & Hirzel, A. H. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* **27**: 437-448.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D. & Thomas, L. 2001. *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*. Oxford University Press, Oxford.

Calenge, C., Dufour, A.B. & Maillard, D. 2005. K-select analysis: a new method to analyse habitat selection in radio-tracking studies. *Ecological Modelling* **186**: 143-153.

Camphuysen, C.J., Fox, A.D., Leopold, M.F. & Petersen, I.K. 2004. *Towards standardised seabirds at sea census techniques in connection with environmental impact assessments for offshore wind farms in the U.K. A comparison of ship and aerial sampling methods for marine birds, and their applicability to offshore wind farm assessments*. Koninklijk Nederlands Instituut voor Onderzoek der Zee Report commissioned by COWRIE.

Canadas, A.R., Sagarminaga, R., De Stephanis, E., Urquiola, & Hammond, P.S. 2005. Habitat preference modelling as a conservation tool proposals for marine protected areas for cetaceans in southern Spanish waters. *Aquatic Conservation: Marine and Freshwater Ecosystems* **15**: 495-521.

Christ, A., Ver Hoef, J. & Zimmerman, D.L. 2008. An animal movement model incorporating home range and habitat selection. *Environmental and Ecological Statistics* **15:** 27-38.

Clarke, E.D., Spear, L.B., McCraken, M.L., Marques, F.F.C., Borchers, D.L., Buckland, S.T. & Ainley, D.G. 2003. Validating the use of generalized additive models and at-sea surveys to estimate size and temporal trends of seabird populations. *Journal of Applied Ecology* **40**: 278-292.

Cyrus, D.P. 1991. The influence of turbidity on the foraging behaviour of Little Terns *Sterna albifrons* off the St Lucia Mouth, Zululand, South Africa. *Marine Ornithology* **19**: 103-108.

Dalziel, B.D., Morales, J.M. & Fryxell, J.M. 2008. Fitting probability distributions to animal movement trajectories: using artificial neural networks to link distance, resources and memory. *The American Naturalist* **172**: 248-258.

Davidson, T.A., Sayer, C.D., Perrow, M., Bramm, M. & Jeppesen, E. 2010. The simultaneous inference of zooplanktivorous fish and macrophyte density from sub-fossil cladocern assemblages: a multivariate regression tree approach. *Freshwater Biology* **55**: 546-564.

De'ath, G. & Fabricius, K.E. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* **81**: 3178-3192.

Dorman, C.F. (2006) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecology and Biogeography **16**: 129-138.

Dorman, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davids, R.G., Hirzel, A., Jetz, W., Kissling, D., Kühnm I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M. & Wilson, R. 2007. Methods to account for spatial autocorrelation in the analysis of species distribution data: a review. *Ecography* **30**: 609-628.

Eckert, S.A., Moore, J.E., Dunn, D.C., van Buiten, R.S., Eckert, K.L. & Halpin, P.N. 2008. Modeling loggerhead turtle movement in the Mediterranean: importance of body size and oceanography. *Ecological Applications* **18**: 290-308.

Elith, J. & Leathwick, J.R. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* **13**: 265-275.

Elith, J., Leathwick, J.R. & Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* **77**: 802-813.

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

Fieberg, J., Matthiopoulos, J., Hebblewhite, M., Boyce, M.S. & Frair, J.L. 2010. Correlation and studies of habitat selection: problem, red herring or opportunity? *Philosophical Transactions of the Royal Society B* **365**: 2233-2244.

Fortin, M.J. & Dale, M.R.T. 2005. *Spatial analysis – a guide for ecologists*. Cambridge University Press, Cambridge.

Friedman, J.H., Hastie, T. & Tibshirani, R. 2000. Additive logistic regression: a statistical view of boosting. *Annals of Statistics* **28**: 337-407.

Furness, R. W. & Tasker, M. L. 2000. Seabird-fishery interactions: quantifying the sensitivity of seabirds to reductions in sandeel abundance, and identification of key areas for sensitive seabirds in the North Sea *Marine Ecology Progress Series* **202**: 253-264

Garthe, S. & Hüppop, O. 1999. Effect of ship speed on seabird counts in areas supporting commercial fisheries. *Journal of Field Ornithology* **70**: 28-32.

Gaston, A.J. 2004. *Seabirds: A Natural History*. Christopher Helm, A & C Black, London.

Guisan, A. & Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters* **8**: 993-1009.

Hamer, K.C. Humphreys, E.M., Magalhães, M.C.. Garthe, S., Hennicke, J., Peters, G., Grémillet, D., Skov, H. & Wanless, S. 2009. Fine-scale foraging behaviour of a medium-ranging marine predator. *Journal of Animal Ecology* **78**: 880-889.

Hastie, T. & Tibshirani, R.J. 1990. *Generalized Additive Models*. Chapman & Hall, London.

Heinänen, S., Rönkä, M., & von Numers, M. 2008. modelling the occurrence and abundance of a colonial species, the Arctic Tern *Sterna paradisaea* in the archipelago of SW Finland. *Ecography* **31**: 601-611.

Heinänen, S. & von Numers, M. 2009. Modelling species distribution in complex environments: an evaluation of predictive ability and reliability in five shorebird species. *Diversity and Distributions* **15**: 266-279.

Hirzel, A. H., Hausser, J., Chessel, D. & Perrin, N. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data. *Ecology* **83**: 2027-2036.

Holt, J. (2006) Geophysical characteristics of UK waters for habitat classification. Document to accompany datasets supplied by POL to JNCC. Proudman Oceanographic Laboratory.

Huettmann, F. & Diamond, A.W. 2006. Large-scale effects on the spatial distribution of seabirds in the Northwest Atlantic. *Landscape Ecology* **21**: 1089-1108.

Jaynes, E.T. 1957. Information theory and statistical mechanics I. *Physics Review* **106**: 620-630

Johnson, C.J., Nielsen, S.E., Merrill, E.H., McDonald, T.L. & Boyce, M.S. 2006. Resource selection functions based on use-availability data: theoretical motivation and evaluation methods. *Journal of Wildlife Management* **70**: 347-357.

Jonsen, I.D., Flemming, J.M. & Myers, R.A. 2005. Robust state-space modeling of animal movement data. *Ecology* **86**: 2874-2880.

Karels, T.J., Bryant, A.A. & Hik, D.S. 2004. Comparison of discriminant function and classification tree analyses for age classification of marmots. *Oikos* **105**: 575-587.

Keating, K.A. & Cherry, S. 2004. Use and interpretation of logistic regression in habitat selection studies. *Journal of Wildlife Management* **68**: 774-789.

Keitt, T.H., Bjørnstad, O.N., Dixon, P.M. & Citron-Pousty, S. 2002. Accounting for spatial pattern when modeling organism–environment interactions. Ecography **25**: 616-625.

Kumar, S. & Stohlgren, T.J. 2009. MaxEnt modeling for predicting suitable habitat for threatened and endangered tree *Canacomyrica monticola* in New Caledonia. *Journal of Ecology and Natural Environment* **1**: 94-98.

Leathwick, J.R., Elith, J. & Hastie, T. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modeling of species distributions. *Ecological Modelling* **199**: 188-196

Legendre, P. 1993. Spatial autocorrelation – trouble or new paradigm? *Ecology* **74**: 1659-1673.

Lele, S.R. 2009. A new method for estimation of resource selection probability function. *Journal of Wildlife Management* **73**: 122-127.

Lele, S.R. & Keim, J.L. 2006. Weighted distributions and estimation of resource selection probability functions. *Ecology* **87**: 3021-3028.

Lewis, S., Benvenuti, S., Dall'Antonia, L., Griffiths, R., Money, L., Sherratt, T.N., Wanless, S. & Hamer, K.C. 2002. Sex-specific foraging behaviour in a monomorphic seabird. *Proceedings of the Royal Society of London B* **269**: 1687-1693.

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.

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J. & Possingham, H.P. 2005. Zero tolerance ecology: improving ecological interference by modelling the source of zero observations. *Ecology Letters* **11**: 1235-1246.

Matthiopoulos, J. 2003. The use of space by animals as a function of accessibility and preference. *Ecological Modeling* **159**: 239-268.

McNyset, K.M & Blackburn, J.K. 2006. Does GARP really fail miserably? A response to Stockman *et al.* . (2006). *Diversity and Distributions* **12**: 782-786.

McSorley, C.A., Dean, B.J. Webb, A. & Reid, J.B. 2003. *Seabird Use of Waters Adjacent to Colonies: Implications for Seaward Extensions to Existing Breeding Seabird Colony Special Protection Areas*. Joint Nature Conservation Committee, JNCC Report, No, 329, Aberdeen, 102pp.

Meynard, C.N. & Quinn, J. F. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, **34**, 1455-1469

Moisen, G.G. & Frescino, T.S. 2002. Comparing five modeling techniques for predicting forest characteristics. *Ecological Modelling* **157**: 209-225.

Munoz, J. & Felicisimo, A.M. 2004. Comparison of statistical methods commonly used in predictive modeling. *Journal of Vegetation Science* **15**: 285-292.

Olivier, F. & Wotherspoon, S.J. 2006a. Modelling habitat selection using presence-only data: case study of a colonial hollow nesting bird, the Snow Petrel. *Ecological Modelling* **195**: 187-204.

Olivier, F. & Wotherspoon, S.J. 2006b. Distribution and abundance of Wilson's Storm Petrels *Oceanites oceanicus* at two locations in East Antarctica: testing habitat selection models. *Polar Biology* **29**: 878-892.

Oppel, S., Meirinho, A. & Ramirez, I. (in prep) Comparing four different modelling techniques to predict the spatial distribution of Balearic Shearwaters off the coast of Portugal. Poster presented at the World Seabird Conference 2010, Victoria, Canada.

Paiva, V.H., Ramos, J.A., Martins, J., Almeida, A., Carvalho, A. 2008. Foraging habitat selection by Little Terns *Sterna albifrons* in an estuarine lagoon system of southern Portugal. *Ibis* **150**: 18-31.

Patil, G. P. & Rao, C. R. 1978. Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics* **34**: 179-189

Perrow,M.R., Skeate, E.R., Lines, P., Brown, D. & Tomlinson, M.L. (2006) Radio telemetry as a tool for impact assessment of wind farms: the case of Little Terns *Sterna albifrons* at Scroby Sands, Norfolk. *Ibis*, **148**, 57-75

Perrow, M.R., Skeate, E.R. & Gilroy, J.J. (2011) Visual tracking from a rigid-hulled inflatable boat to determine foraging movements of breeding terns. *Journal of Field Ornithology*, 82, 68-79.

Patterson, T.A., Thomas, L., Wilcox, C., Ovaskainen, O. & Matthiopoulos, J. 2008. State-space models of individual animal movements. *Trends in Ecology and Evolution* **23**: 87-94.

Peterson, A. T., Papes, M. & Eaton, M. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and MaxEnt. *Ecography* **30**: 550-560

Phillips, S. J., Dudik, M. & Schapire, R. E. 2004. A maximum entropy approach to species distribution modelling. *Proceedings of the Twenty-First International Conference on Machine Learning* 665-662

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

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. 2009. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-96.

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

Prasad, A.M., Iverson, L.R. & Liaw, A. 2006. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **9**: 181-199.

Randin, C. F., Dirnbock, T., Dullinger, S., Zimmermann, N. E., Zappa, M. & Guisan, A. 2006. Are niche-based species distribution models transferable in space. *Journal of Biogeography*, **33**, 1689-1703

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reise, K. & Bartsch, I. 1990. Inshore and offshore diversity of epibenthos dredged in the North Sea. *Netherlands Journal of Sea Research* **25**: 175-179

Roycroft, D., Kelly, T.C. & Lewis, L.J. 2007. Behavioural interactions of seabirds with suspended mussel longlines. *Aquaculture International* **15**: 25-36.

Schapire, R. 2003. The boosting approach to machine learning – an overview. *MSRI Workshop on Nonlinear Estimation and Classification, 2002*, Springer, New York

Schonlau, M. 2005. Boosted Regression (Boosting): An introductory tutorial and a Stata plugin. *The Stata Journal*, **5,** 330-354.

Schwartz, M.W., Iverson, L.R., Prasad, A.M., Matthews, S.N. & O'Conner, R.J. 2006. Predicting extinctions as a result of climate change. *Ecology* **87**: 1611-1615.

Schwemmer, P. 2009. Influence of water flow velocity, water depth and colony distance on distribution and foraging patterns of terns in the Wadden Sea. *Fisheries Oceanography* **18**: 161-172.

Scott, D., Moller, H., Fletcher, D., Newman, J., Aryal, J., Bragg, C. & Charleton, K. 2009. Predictive habitat modeling to estimate petrel breeding colony sizes: Sooty Shearwaters (*Puffinus griseus*) and Mottled Petrels (*Pterodroma inexpectata*) on Whenua Hou Island. *New Zealand Journal of Zoology* **36**: 291-306.

Shaffer, S.A., Weimerskirch, H., Scott, D., Pinaud, D., Thompson, D.R., Sagar, P.M., Moller, H., Taylor, G.A., Foley, D.G., Tremblay, Y. & Costa, D.P. 2009. Spatiotemporal habitat use by breeding Sooty Shearwaters *Puffinus griseus*. *Marine Ecology Progress Series* **391**: 209-220.

Shipley, B., Vile, D. & Garnier, E. 2006. From plant traits to plant communities: a statistical mechanistic approach to biodiversity. *Science* **314**: 812-814.

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. 2009. ROCR: Visualizing the performance of scoring classifiers.. R package version 1.0-4. http://CRAN.R-project.org/package=ROCR

Skov, H. & Durinck, J. 2000. Seabird distribution in relation to hydrography in the Skagerrak. *Continental Shelf Research* **20**: 169-187.

Skov, H. & Durinck, J. 2001. Seabird attraction to fishing vessels is a local process. *Marine Ecology Progress Series* **214**: 289-298.

Skov, H., Humphreys, E., Garthe, S., Geitner, K., Hamer, K., Hennicke, J., Pamer, H., Gremillet, D. & Wanless, S. (2008) Application of habitat suitability modelling to tracking data of marine animals as a means of analyzing their feeding habitats. *Ecological Modelling*, **212**, 504-512.

Smouse, P.E., Forcardi, S., Moorcroft, P.R., Kie, J.G., Forester, J.D. & Morales, J.M. 2010. Stochastic modelling of animal movement. *Philosophical Transactions of the Royal Society B* **365**: 2201-2211.

Smyth, G.K. 1996. *Regression modelling of quantity data with exact zeroes*. Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management. Technology Management Centre, University of Queensland, pp. 572-580.

Smyth, G.K. & Verbyla, A.P. 1999. Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics* **10**: 695-709.

Stienen, E. W. M., Van Beers, P. W. M., Brenninkmeijer, A., Habraken, J. M. P. M., Raaijmakers, M. H. J. E. & Van Tienen, P. G. M. 2000. Reflections of a specialist: patterns in food provisioning and foraging conditions in Sandwich Terns *Sterna sandvicensis Ardea* **88**: 33-49

Stockman, A.K., Beamer, D.A. & Bond, D.A. 2006. An evaluation of a GARP model as an approach to predicting the spatial distribution of non-vagile invertebrate species. *Diversity and Distributions* **12**: 81-89.

Stockwell, D.R.B. & Noble, I.R. 1992. Induction of sets of rules from animal distribution data – a robust and informative method of data-analysis. *Mathematics and Computers in Simulation* **33**: 385-390.

Stroud, D.A., Chambers, D., Cook, S., Buxton, N., Fraser, B., Clement, P., Lewis, P., McLean, I., Baker, H. & Whitehead, S. 2001 *The UK SPA Network: its Scope and Content.* JNCC, Peterborough.

Swihart, R.K. & Slade, N.A.1985. Testing for independence of observations in animal movements. *Ecology* **66**: 1176-1184.

Thaxter, C.B., Wanless, S., Daunt, F., Harris, M.P., Benvenuti, S., Watanuki, Y., Grémillet, D. & Hamer, K.C. 2010. Influence of wing loading on the trade-off between pursuit-diving and flight in Common Guillemots and Razorbills. *Journal of Experimental Biology* **213**: 1018-1025.

Thaxter, C.B., Lascelles, B., Sugar, K., Cook, A.S.P.C., Roos, S., Bolton, M., Langston, R.H.W. & Burton, N.H.K. Submitted. Seabird Foraging Ranges as a Tool for Identifying Candidate Marine Protected Areas. Biological Conservation.

Tremblay, Y., Bertrand, S., Henry, R.W., Kappes, M.A., Costa, D.P. & Shaffer, S.A. 2009. Analytical approaches to investigating seabird-environment interactions: a review. *Marine Ecology Progress Series* **391**: 153-163.

Tweedie, M.C.K. 1984. An index which distinguishes between some important exponential families. In *Statistics: Applications and New Directions*. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. (Eds. J. K. Ghosh and J. Roy), pp. 579-604. Indian Statistical Institute, Calcutta.

Valavanis, V.D., Pierce, G.J., Zuur, A.F., Palialexis, A. Saveliev, A., Katara, I., Wang, J. 2008. Modelling essential fish habitat based on remote sensing, spatial analysis and GIS. *Hydrobiologia* **612**: 5-20.

van der Kooij, J., Scott, B. E. & Mackinson, S. 2008. The effects of environmental factors on daytime sandeel distribution and abundance on the Dogger Bank. *Journal of Sea Research* **60**: 201-209

Wakefield, E.D., Phillips, R.A. & Mattiopoulos, J. 2009. Quantifying habitat use and preferences of pelagic seabirds using individual movement data: a review. *Marine Ecology Progress Series* **391**: 165-182.

Warren, T.L., Betts, M.G., Diamond, A.W. & Forbes, G.J. 2005. The influence of local habitat and landscape composition on cavity-nesting birds in a forested mosaic. Forest Ecology and Management **214**: 331-343.

Warton, D.I. & Shepherd, L.C. 2010. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics* **4**:1383-1402.

Welsh, A.H., Cunningham, R.B., Donelly, C.F. & Lindenmayer, D.B. 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling* **88**: 297-308.

Wilson, L.J., McSorley, C.A., Gray, C.M., Dean, B.J., Dunn, T.E., Webb, A. & Reid, J.B. 2009. Radio-telemetry as a tool to define protected areas for seabirds in the marine environment. *Biological Conservation* **142**: 1808-1817.

Wisz, M.S. & Guisan, A. 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology* **9**: 8 doi:10.1186/1472-6785-9-8.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A. & NCEAS Predicting Species Distributions Working Group. 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions*. **14**, 763-773

Wood, A.G., Naef-Daenzer, B., Prince, P.A. & Croxall, J.P. 2000. Quantifying habitat use in satellite-tracked pelagic seabirds: application of kernel estimation to albatross locations. *Journal of Avian Biology* **31**: 278-286.

Wood, S.N. 2006. *Generalized Additive Models: an Introduction with R*. Chapman & Hall, London.

Yen, P.P.W, Huettmann, F. & Cooke, F. 2004. A large-scale model for at-sea distribution and abundance of Marbled Murrelets (*Brachyramphus marmoratus*) during the breeding season in coastal British Columbia, Canada. *Ecological Modelling* **171**: 395-413.

Zuur, A.F., Ieno, E.N. & Smith, G.M. 2007. *Analysing Ecological Data. Statistics for Biology and Health*. Springer, New York.