Using individual tracking data to validate the predictions of species distribution models.

Cecilia Pinto[1,2], James A. Thorburn[1], Francis Neat[2], Peter J. Wright[2], Serena Wright[3], Beth E. Scott[1], Thomas Cornulier[1] and Justin M.J. Travis[1].

[1]Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, AB24 2TZ, UK; [2]Marine Scotland Science, Marine Laboratory, 375 Victoria Road, Aberdeen, AB11 9DB, UK; [3]Centre for Environment Fisheries & Aquaculture Science, Pakefield road, Lowestoft, Suffolk NR33 0HT, UK.

ABSTRACT

Aim Estimating environmental suitability from species distribution data is crucial in defining spatial conservation measures. To this end, species distribution models (SDMs) are commonly applied, but seldom validated by completely independent data. Here we use data on individual tracks derived from electronic tags as an alternative means of validating SDM outputs.

Location West coast of Scotland, NE Atlantic.

Methods We used a binomial generalized additive model (GAM) to predict the environmental suitability for flapper skate (*Dipturus cf. intermedia*) in Scottish waters. The GAM modelled relative habitat usage as a function of environmental variables using presence/absence data obtained from scientific trawl surveys. Additional data obtained from electronic tags attached to six individual flapper skates were used to estimate individual tracks using a tidal based geolocation model. Concordance between individual tracks and

GAM-predicted maps of relative habitat usage (RHU) was tested by comparing predicted RHU between estimated tracks and randomly generated tracks.

Results Environmental suitability for the flapper skate was driven by depth and distance from the coast in the SDM. We found high spatial concordance between the estimated tracks of the six tagged individuals and regions of high RHU predicted by the SDM.

Main Conclusions Integrating outputs from an independent data source allowed us to validate predictions from a species distribution model (SDM). The integration of individual- and population-level data sources increases confidence in the outputs being used to define spatial conservation measures. The information on flapper skate distribution provided by this study provides a useful framework for considering spatial conservation measures for this species.

Keywords Data integration, *Dipturus cf. intermedia*, generalised additive model, individual movement, model validation, species distribution model, tidal geolocation model.

(A) INTRODUCTION

Describing how a species is distributed in space, defining its preferred habitat and establishing which environmental characteristics best support its populations, are key to understanding the ecology of threatened or declining species (Guisan & Zimmermann, 2000) and planning for their conservation (Pulliam, 2000). Commonly a species' distribution is obtained from coupling field data with corresponding environmental variables within a modelling framework (Austin, 2002; Aarts *et al.*, 2008). One of the main advantages of species distribution models (SDMs) is that they may be used to generate predictions for the species distribution beyond the area originally sampled, provided the prediction is performed within the environmental range sampled (Elith *et al.*, 2010). SDMs thus have the potential to inform broader scale management which is especially important for marine species where it is often difficult and costly to sample the entire range of a species.

In their most simple form SDMs couple data on a species distribution with environmental variables to quantify the ecological niche of the species, for example, Hutchinson's realized niche or Grinnell's fundamental niche (Guisan & Thuiller, 2005). In most cases, presence-only observations are available to define a species habitat preference, limiting the realism and precision of predictions and increasing their uncertainty (Elith & Leathwick, 2009). Precision and the reliability of predictions are affected by sample size as well, as small sample sizes can be a possible source of instability that will increase uncertainty of model outputs (Guisan & Thuiller, 2005; Barry & Elith, 2006).

Statistical tools commonly used in SDMs include random forest regression trees, MaxEnt (Phillips et al., 2006), generalized linear models (GLMs) and generalized additive models (GAMs). GAMs have been shown to perform as well (Oppel *et al*., 2012), if not better (Moisen & Frescino, 2002; Aertsen *et al.*, 2010), than other predictive models. GAMs are more flexible in fitting complex non-linear responses (Aarts *et al*., 2008), and can compensate for over-fitting through the use of a penalized likelihood (Venables & Dichmont, 2004). They do, however, require a high number of degrees of freedom in order to perform well and give reliable predictions (Wood, 2006; Drexler & Ainsworth, 2013). Thus, for predictions on environmental suitability obtained by GAMs, as well as for those obtained using other statistical or machine learning approaches, their reliability should be tested through field validation or by finding alternative ways of testing model outputs.

Model validation is important when extrapolating model outputs to non-sampled areas (Elith & Leathwick, 2009) and specifically when the ecology of the species of interest is poorly known. As field validation requires entails significant economic and time investment, test datasets (Drexler & Ainsworth, 2013), testing against a null model (Raes & ter Steege, 2007) or bootstrapping (Elith & Leathwick, 2009) have been used as alternative methods. Comparing model outputs and the spatial distribution of independent data obtained from different sampling sources has been used as an option to validate model outputs (Grubbs & Musick, 2007), to reduce estimate and prediction uncertainty (Petit & Lambin, 2002; Jetz *et al.*, 2012) and to help cross-validate outputs of models obtained from independent sets of data (Rogers *et al.*, 2014).

There has been a rapid increase in the volume of spatial data derived from electronic tagging devices, often referred to as 'biologgers', on individual movements of animals across a broad range of taxa. Following the definition in Ropert-Coudert *et al*., (2009), biologgers comprise storage tags, archival tags and electronic data recorders. Many of the earliest applications of this approach were on seabirds, pinnipeds, cetaceans and sea turtles (Ropert-Coudert *et al.*, 2009). The rapid uptake of electronic tagging for marine species was due to the benefits provided from observations of underwater behaviour and the gathering of positional information at sea. Tagging of fish species, however, has lagged behind because of the greater difficulty of acquiring reliable positional information on sub-surface species. However, the advent of geolocation models, either using tidal signatures or light intensity levels, is now resulting in increased knowledge on the spatial ecology of an increasing number of fish species of both economic and conservation concern, including pacific bluefin tuna (Whitlock *et al*., 2012), cod (Neuenfeldt *et al.*, 2013), white sharks (Jorgensen *et al.*, 2009) and tiger sharks (Werry *et al.*, 2014). Geolocation models use contemporaneous environmental information, such as light level and depth, to estimate the individual's most likely geographical locations at a certain time-step. Assuming that an individual spends more time in its preferred habitat, estimated individual tracks are an ideal independent source of information to infer environmental suitability and to test predictions obtained from SDMs.

This study examines the distribution of flapper skate (*Dipturus cf. intermedia*) off the west coast of Scotland. The flapper skate is listed in the IUCN Red List of Threatened Species as "Critically Endangered" (www.iucnredlist.org). As a slow growing, late maturing and low fecundity species, its population growth rate is highly sensitive to fishing mortality (Brander, 1981). The species suffered a rapid decline in the last 40 years with landings falling by 90 per cent (Du Buit, 1977; Brander, 1981; Philippart, 1998). The flapper skate is now only occasionally found in the North Sea, and its former distribution contracted throughout the period, leaving a number of relict populations off the West coast of Scotland (Brander, 1980; Walker & Hislop, 1998; Daan *et al*., 2005). In order to estimate the potential for this species to recolonize its former range, it is fundamental to understand the environmental influences determining its distribution. To this aim we used SDMs to define environmental suitability for the flapper skate off the west coast of Scotland from presence-absence data obtained from trawl survey data, and used individual geolocation estimates from electronic tagging devices to validate model predictions. The aim of this study was to demonstrate the potential of integrating information from individual tracking data (obtained from tidal geolocation modelling of electronic data storage tagging devices in this study), for the validation of SDM predictions.

(A) METHODS

(B) *Species distribution data*

Presence-absence data on the flapper skate were obtained from several trawl surveys including the Marine Scotland Science northern shelf monkfish survey and the International Bottom Trawl Survey (Fig.1). The *Dipturus batis* complex was identified as two species (*Dipturus intermedia* and *Dipturus flossada*) in 2010 (Griffiths *et al.*, 2010; Iglésias *et al.*, 2010), and therefore data on the flapper skate were generally available only from 2010 onwards (330 records), although a few records (65) were obtained from surveys conducted in

2003, 2004 and 2005 during which catch records were matched to individual photographs which allowed identification to the species level. The surveys each used a different trawl gear and this may have led to different catchabilities for flapper skate (Appendix S3). The areas covered by the different surveys did, however, broadly overlap, just that some had a more restricted areal coverage than others.

(B) *Predictor variables*

The environmental variables used as predictors in the SDM model were all projected in UTM 29N (WGS84) and all had the same resolution of 30” (790.2m). Environmental variables included: trawl shot latitude and longitude, depth, slope (reported as angle measures), mean salinity, mean temperature, distance from the coast, seabed composition (sediments) and gear type. Depth and slope were obtained from OCEANWISE 6” (www.oceanwise.eu) and INFOMAR (www.infomar.ie); depth was square root-transformed for modelling purposes as the raw data had a skewed distribution. Salinity and temperature were obtained from the freely available oceanographic model EUROPEAN NORTH WEST SHELF – OCEAN PHYSICS REANALYSYS FROM METOFFICE (1985-2012) (www.myocean.eu). Euclidean distance to the nearest coast was calculated in ArcGIS 10. The sediments layer was extracted from the British Geological Survey database (European Marine Observation and Data Network, EMODNET, www.emodnet-geology.eu) and is represented by seven classes of sediments: coarse sands, mixed sediment, mud to sandy mud, rock, sand to muddy sand, seabed (unknown sediment) and till (mixed sediments). Preparation of the final dataset to be used for predictions was performed in R 3.1.1 (R Core Team, 2014).

(B) *GAM fitting*

A binomial GAM with a logit link function was fitted using the “mgcv” package in R (Wood, 2006, 2011). Thin plate regression splines were used as smoothing functions for the continuous environmental predictors, while gear type was added as a factor variable and kept in all models in order to account for the different catchability of gears (Appendix S3). The effect of year was also included in the model to test for an effect of yearly variation in the presence of the species due to factors concerning the population dynamics of the species and not necessarily the variation of environmental covariates. In order to determine which covariates best predicted the distribution of the species model selection was done through minimizing the Akaike Information Criterion (AIC). We looked for potential spatial autocorrelation in the residuals by fitting generalized additive mixed models (GAMM) without a spatial correlation structure and with exponential and spherical correlation structures. We compared the models by checking the estimated range (the extent to which the correlation is detected across space or not) and the nugget effect (the level of correlation between two random points taken in close proximity). A confusion matrix was calculated in order to estimate accuracy, sensitivity (probability of true positives) and fall out (probability of false positives) proportions in model predictions through the library “PresenceAbsence” in R (Freeman & Moisen, 2008). Potential colinearity between covariates was examined with Pearson’s correlation coefficient and with the parameter correlation matrix from the model. To test its robustness, the best model was rerun after excluding extreme values of covariates (identified as outliers relatively to the central distribution of the covariate) from the dataset to test if the estimates would hold to the data reduction. The final model was run with the whole dataset as the exclusion of extreme values did not affect the results. Model predictions were first produced as probability of presence and then as relative habitat utilisation (RHU) as explained in the “Model validation” section. These were both produced at a 2km resolution to match the geolocation model resolution.

(B) *Geolocation modelling*

Independent data from six data storage tags (DSTs) attached to common skate and recovered between 2012 and 2014 were used. DSTs were deployed on 18 individuals in an area known as the Sound of Jura (4 recaptures) and on 29 individuals from an area known as the Stanton Banks (2 recaptures). DSTs were attached externally to the fish (Neat *et al.,* 2015) and recorded hydrostatic pressure and environmental temperature every 2 minutes. For the biological characteristics of the tagged individuals and the total length of the time series see Appendix S1 in Supporting Information (where S indicates Supporting). Notably, three (7968, 7967 and 7972) of the six individuals tagged with DSTs were also tagged with acoustic transmitters connected to a set of acoustic receiver stations which were active for one year in the northern section of the Sound of Jura (Neat *et al*., 2015).

Time series of pressure levels obtained from the DSTs were converted to time series of depth values, which were then matched to tidal time series for UK waters at 7km resolution using an adapted version of a hidden Markov model, as developed by Pedersen *et al.* (2008). This was used to geolocate flapper skate tagged with data storage tags from the point of release to the point of recapture. As outputs, for each day at liberty, a probability distribution (most probable track) was constructed using a model constrained by the maximum depth, tidal geolocation estimates (Hunter *et al*., 2003) and an automatically selected diffusivity value (Pedersen *et al*., 2008). Briefly, the model requires four parameters: variance, amplitude, mean square error (between the tidal signal recorded by the animal and the actual tidal cycle) and a tidal time window in which to search for the tidal signal. The optimized values of these parameters were estimated by Pedersen *et al*. (2008). The tidal grid was constructed using the Oregon State tidal inversion model with seven tidal constituents (M2, S2, N2, K2, O1, K1 and M4) as defined in Pedersen *et al*. (2008). On days where there was no tidal signal (i.e. the fish was away from the seabed and so no tidal signal could be selected), bathymetric depth (Gebco bathymetry) was used to exclude recorded positions shallower than the maximum depth. The spatial extents of the model of release and recapture locations at a resolution of approximately 7km, were -32W, 35N, 11E, 70N.

A second output of the model is the average of all possible tracks an individual could have covered during the tagging period, producing a density map called a utilization distribution map (UDM). As the UDM is a distribution of all possible tracks predicted by the model, it directly includes a measure of the uncertainty about the true track. The UDM was used to compare the geolocated locations against the predicted probability of presence obtained from the GAM. The UDM is a probability map where each cell has a value between 0 and 1 ($\Sigma = 1$), the higher the value the greater the probability the individual spent time in that cell. Because the model calculates a probability for each cell being part of the total grid (the final size of the grid is optimized by the model based on the diffusivity value (Pedersen *et al.*, 2008)) the final output needs to be rescaled after excluding probabilities that are too close to zero (therefore far from the actual individual track), which would shrink all probabilities towards zero (see Appendix S1). This output is produced at a 2km resolution.

(B) *Model validation*

To obtain predictions from the GAM model that would be comparable to the geolocation estimates we estimated RHU (relative habitat utilisation). RHU was calculated as:

$$RHU(x) = \frac{\exp(g(x))}{\sum_D \exp(g(x))}$$

where $g(x)$ is the linear predictor of the GAM at location x part of study domain D, so that RHU is on a scale proportional to time spent by an animal at that location and to density of observations (Aarts *et al.,* 2012). The RHU was compared to geolocation estimates in three separate steps. First, in order to facilitate an initial visual estimation of whether the tracks produced by the geolocation model covered either high or low predicted RHU obtained from the GAM, the individual tracks obtained from the UDM predictions were plotted on top of the RHU map. Secondly, the distribution of RHU values for an area surrounding each individual track (i.e. the area is defined by an individual track plus a 2' buffer) was compared to the distribution of RHU values extracted at track locations, in order to compare which values of the predicted distribution were actually selected by the individual along its track. This process was implemented separately for each individual. Lastly, to verify that the selection of high probability of presence areas by individual tracks was effectively better than a random selection of areas, we assimilated the RHU for the whole area to a likelihood distribution (termed 'RHU-likelihood' for simplicity thereafter). This last process was considering all the individual tracks together at once. To account for the probability assigned by the geolocation model to each track cell belonging to the UDM, and to give each individual equal weight in the analysis, track cells were resampled with replacement for 2,000 draws (as this is the size of the locations of all the tracks put together) proportionally to cell probabilities' value ("standardized tracks") of the UDM. We then performed 10,000 simulations of random track locations to calculate their respective RHU-likelihood. The final output was than the difference between the sum of the log(RHU-likelihood) at the 2,000 observed track locations and the sum of the log(RHU-likelihood) at 2,000 simulated locations for each of the 10,000 randomly generated sets of tracks. In order to preserve the internal spatial structure of the animal tracks, simulations were done by anchoring six points generated at random over the study area which would define the new centroid of each individual track (see Appendix S2). The random tracks were generated as a set of locations with the same shape of the original tracks but centred around the position defined by the randomly generated centroid. When parts of random tracks were generated on land these locations were excluded and the section of the track generated at sea was resampled with repetition until the original sample size of the track was reached. The 10,000 simulations were performed twice using two nested spatial domains to produce both a regional and a more stringent local test of the model performance. The regional polygon was drawn around the area covered by the raw data of presence-absence (Fig.1), and the local polygon was drawn around the area covered by the geolocated tracks (see Appendix S2). The use of two spatial scales allowed us to assess both the reliability of our predictions and the representativeness of the geolocation estimates.

(A) RESULTS

(B) *GAM fitting*

The best model defining the probability of presence of flapper skate included trawl latitude, trawl longitude, √depth, distance from the coast and gear type (Table 1) and explained 33% of the variance (n=395) (for details see Appendix S3). Trawl latitude and longitude were not significant in the model, but were kept in as their exclusion did not improve the AIC significantly (Table 1). There was no significant spatial autocorrelation in the residuals (see Appendix S3). Model accuracy calculated with the confusion matrix was 63%, from a fall-out value of 0.34 and a sensitivity of 0.78 (see Appendix S3), suggesting that model predictions are more accurate than at random. Model predictions suggest that flapper skate is a species that concentrates on inshore areas, showing the highest probability of presence in the sea lochs of the west coast of Scotland and in areas surrounding banks and islands (Fig.2). Probability of presence is limited by the extent of the continental shelf and the depths of the Rockall Trough, but continues to be high in the North Channel and around the Shetland Islands (Fig.2). As shown by the model outputs, flapper skate distribution is driven by depth, with low probability of presence at depth < 100m and decreasing again at depths > 400m, although the variance surrounding the estimates increases in the 300m-600m range where data points are more scattered (Fig.3). Probability of presence decreases strongly as distance from the coast increases (Fig.3). Therefore this species seems to prefer areas that can reach high depths but at the same time are surrounding islands or are constrained within islands and the main land.

(B) *Geolocation modelling*

The geolocation model obtained a number of tidal matches for each individual (Fig.4). The longer the time frame within which each individual was tracked the more information was available to describe the usage area of each individual. With the exception of individual 8828, which was recaptured after only two weeks at liberty, all other individuals were at liberty for between six months and one year. Four individuals spent most of their time where they were originally tagged, while individual 7968 moved south towards the top of the North Channel, and 8828 moved north. Individual 8794 was the only individual that had probabilities of area usage always lower than 0.1 and shows the largest area coverage across time (Fig.4). Thus our results across these six individuals suggest that the output probabilities produced by the geolocation model are highly affected by the time spent at liberty, the area covered by an individual and its level of activity during this time. Therefore, we suggest these outputs cannot be readily compared between individuals or with other measures of probabilities obtained from different modelling procedures. Data from the acoustic stations that detected individuals 7968, 7967 and 7972 confirm the reliability of the tracks shown by the geolocation model (see Appendix S4) as we can directly compare time steps at which each individual was recorded by the acoustic station and predicted in the same area by the geolocation model. This supports our suggestion that geolocation models' outputs have a high potential as a validation tool for predictions obtained from other modelling procedures.

(B) *Cross validation*

The visual exploration of locations estimated by the geolocation model and the corresponding area extracted from the GAM predictions (Fig.5) showed a high overlap between the UDM predictions and the cells with the highest RHU values predicted by the GAM. The values corresponding to the single track locations for all six individuals were always distributed among the highest RHU values (Fig.6). At the regional scale, the likelihood that the observed geolocated tracks coincided with areas of high values of RHU was always higher than the likelihood that the randomly generated tracks would fall over high RHU values (Fig.7). At the local scale, a very similar result was obtained (this second result is not shown).

(A) DISCUSSION

This study demonstrated the potential for integrating very different types of data to obtain and validate environmental suitability surfaces. These approaches typically deal with data collected across different spatial scales, involve very different sample sizes and provide different types of information and, as such, are generally used to address very different research questions. However, we demonstrate that the comparison of the different data and model approaches has considerable potential in validating reciprocal outputs, improving their reliability and strengthening inference. Predicting species distribution from model outputs carries varying levels of uncertainty depending on the quality and amount of data and the availability of covariates and movement parameters that could improve precision and accuracy (Elith & Leathwick, 2009). Uncertainty increases around SDMs outputs when information on dispersal characteristics is lacking in the modelling procedure (Pulliam, 2000) or the model is predicting far from the range of available data (Venables & Dichmont, 2004; Elith *et al*., 2010). Furthermore, modelling the habitat preference of an endangered species that has undergone range contraction is particularly problematic, i.e. absence from an area might not mean that the area is unsuitable, simply that the species has been extirpated from that area (Guisan & Thuiller, 2005). Therefore, as the estimation of environmental suitability is fundamental when defining conservation measures for an endangered species, predictions need to be carefully validated in order to provide increased confidence in their accuracy.

Here we demonstrate that by using estimated individual tracks, it is possible to observe habitat use of a single animal directly and verify if it preferentially moves within areas of high predicted RHU. Combining direct observation of habitat use from individual tracking data to validate predicted environmental suitability is particularly important when static distribution data are used to describe habitat utilisation of mobile species. An additional advantage of comparing model outputs from independent sets of data lies in increasing the confidence of predictions made from a small sample size. Individual tracking observations would be too few (only six individuals in this study) to make robust inference regarding population-level habitat use, but the combination of distributions model outputs with geolocation model outputs can be used to infer the potential drivers of the distribution of the flapper skate. Therefore combining independent datasets also increases the power of individual tracking and survey data which, taken separately, would be too sparse to be used in a management framework, specifically when dealing with an endangered species only occupying a severely contracted distribution.

There are other validation methods when field validation is not an available option. The most common practice is to split the data into a trial data set on which the model will be run, and the remainder to be used as a validation data set to see if model predictions correspond with these observations locations (Drexler & Ainsworth, 2013). The comparison between the predicted and observed values at the same location can be bootstrapped in order to create additional datasets and increase power and then fit correlation parameters to test for correspondence between the predicted and the observed value (Grüss *et al.*, 2014). These methods are an important development, specifically when data are available on a single area or a single population. However, despite these statistical advances, cross-validation has been found to be stronger than "split sample" methods already within a single dataset, specifically when the sample size is small (Drummond *et al.*, 2003; Maggini *et al.*, 2006). When different sets of data are available, between data sets cross-validation should be used, taking advantage of the independency of data sets which reduces bias and increases statistical power.

Understanding the environmental preferences of the flapper skate, an endangered species in urgent need of conservation, is a fundamental step towards its management. The spatial dynamics of a species are important in the context of conservation planning as they not only highlight areas of use but also their connectivity (Baguette *et al*., 2013). A significant portion of the study area was recently designated a marine protected area for flapper skate (www.scotland.gov.uk/Topics/marine/marine-environment/mpanetwork). Although this study suggests that the flapper skate is a species which concentrates close to the coast, its presence is also predicted to be high around offshore islands. Therefore, the environmental preference of the flapper skate seems to be defined by areas which are close to the coastline with deep areas in close proximity. The preference of areas defined by the combination of deep areas and limited by the distance from the coast is in agreement with findings from previous studies (Neat *et al*., 2015; Pinto & Spezia, 2015) showing that this species has a wide daily range of depths (from 20m to over 200m) potentially due to the following of its benthic preys daily migrations. The geolocation results suggest that individuals have a high probability to move out of the protected area. The protected area is currently only protecting individuals resident in the inner lochs, and these individuals (as 7967, 7968 and 7972) were observed to consistently use areas south of the protected area (towards the North Channel) (Fig.2 and Fig.4). This study therefore suggests further areas where additional protection might be beneficial and where more information needs to be collected. Connectivity between the inner lochs and offshore areas (Stanton Banks) (Fig.2 and Fig.4) should also be explored to investigate if these populations are connected or isolated, as this may influence conservation measures.

Layers of species' environmental preferences produced by suitability models are not the final step of spatial conservation modelling, but are a fundamental step towards it. An emerging approach is the application of spatially-realistic, individual-based simulation models, such as RangeShifter (Bocedi *et al*., 2014) and HexSim (Schumaker, 2013). These modelling platforms are already being used to address a range of conservation questions, related to improving landscape connectivity (Synes *et al*., 2015), reintroduction or assisted colonisation programmes (Huber *et al*., 2014) as well as for understanding and informing the management of spread of invasive species (Fraser *et al*., 2015). In all of these examples, the definition of landscape suitability is a vital step, and there is typically considerable uncertainty in model outputs when, as is often the case, the uncertainty in environmental preference is large. Notably, one recent study using RangeShifter highlighted that uncertainty in the environmental layer can be responsible for greater uncertainty in the outputs than that due to the uncertainty surrounding demographic estimates (Heikkinen *et al*., 2014). Thus the approach proposed here, using a combination of data sources to improve representations of environmental suitability, offers substantial promise for increasing the reliability of model outputs used to inform conservation management.

(A) CONCLUSIONS

We showed how integrating independent sets of data and different modelling procedures can help validate model predictions reducing the uncertainty surrounding such estimates. This approach combined static observations with individual tracking data, taking advantage of the strengths of both information sources: the higher sample sizes of distribution data and the real time habitat use from individual tracks. The integration process can help in the definition of effective conservation measures for endangered species and to assess the efficacy of those already existing. Considering the increasing volumes of data collected at the individual level (Block *et al*., 2011), the development of methods to integrate independent sources of data is of high value in the marine environment. Visual comparison of outputs can be useful for communicating findings to stakeholders when defining ecosystem based management frameworks, after it has been formally backed-up with quantitative evidence. Finally model validation improved the confidence in using data with relatively low power to inform conservation management and to direct future data collection to improve on-going adaptive conservation planning.

(A) ACKNOWLEDGEMENTS

The authors would like to thank the College of Life Sciences of Aberdeen University and Marine Scotland Science which funded CP's PhD project. Skate tagging experiments were undertaken as part of Scottish Government project SP004. We thank Ian Burrett for help in catching the fish and the other fishermen and anglers who returned tags. We thank José Manuel Gonzalez-Irusta for extracting and making available the environmental layers used as environmental covariates in the environmental suitability modelling procedure. We also thank Jason Matthiopoulos for insightful suggestions on habitat utilisation metrics as well as Stephen C.F. Palmer, and three anonymous reviewers for useful suggestions to improve the clarity and quality of the manuscript.

(A) 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.

Aarts G., Fieberg J. & Matthiopoulos J. (2012) Comparative interpretation of count, presence-absence and point methods for species distribution models. *Methods in Ecology and Evolution*, **3**, 177-187.

Aertsen W., Kint V., van Orshoven J., Özkan K., & Muys B. (2010) Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecological Modelling*, **221**, 1119–1130.

Austin M. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.

Baguette M., Blanchet S., Legrand D., Stevens V.M. & Turlure C. (2013) Individual dispersal, landscape connectivity and ecological networks. *Biological Reviews*, **88**, 310–326.

Barry S. & Elith J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.

Block B.A., Jonsen I.D., Jorgensen S.J., Winship A.J., Shaffer S.A., Bograd S.J., Hazen E.L., Foley D.G., Breed G.A., Harrison A.L., & Ganong J.E. (2011) Tracking apex predator movements in a dynamic ocean. *Nature*, **475**, 86-90.

Bocedi G., Palmer S.C.F., Pe'er G., Heikkinen R.K., Matsinos Y.G., Watts K., & Travis J.M.J. (2014) RangeShifter: A platform for modelling spatial eco-evolutionary dynamics and species' responses to environmental changes. *Methods in Ecology and Evolution*, **5**, 388–396.

Brander K. (1980) Fisheries management and conservation in the Irish Sea. *Helgoländer Meeresuntersuchungen*, **33**, 687–699.

Brander K. (1981) Disappearance of common skate *Raja batis* from Irish Sea. *Nature*, **290**, 48–49.

Du Buit M.H. (1977) Age et croissance de *Raja batis* et de *Raja naevus* en Mer Celtique. *J. Cons. int. Explor. Mer.*, **37**, 261–265.

Correia A.M., Tepsich P., Rosso M., Caldeira R., & Sousa-Pinto I. (2015) Cetacean occurrence and spatial distribution: Habitat modelling for offshore waters in the Portuguese EEZ (NE Atlantic). *Journal of Marine Systems*, **143**, 73–85.

Daan N., Heessen H., & Hofstede R. (2005) North Sea Elasmobranchs : distribution , abundance and biodiversity. *CM-International Council for the Exploration of the Sea*, **CM 06** 1–15.

Drexler M. & Ainsworth C.H. (2013) Generalized additive models used to predict species abundance in the Gulf of Mexico: an ecosystem modeling tool. *PLoS ONE*, **8**, e64458.

Drummond S.T., Sudduth K. a, Joshi a, Birrell S.J., & Kitchen N.R. (2003) Statistical and neural methods for site-specific yield prediction. *Transactions of the ASAE*, **46**, 1–10.

Elith J. & Leathwick J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith, J., Kearney M. & Phillips S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330-342.

Fraser E.J., Lambin X., Travis J.M.J., Harrington L.A., Palmer S.C.F., Bocedi G. & MacDonald D.W. (2015) Range expansion of an invasive species through a heterogeneous landscape - the case of American mink in Scotland. *Diversity and Distribution*, **21**, 888-900.

Freeman, E. A. & Moisen, G. (2008) PresenceAbsence: An R Package for Presence-Absence Model Analysis. *Journal of Statistical Software*, **23**,1-31.

Griffiths A.M., Sims D.W., Cotterell S.P., El Nagar A., Ellis J.R., Lynghammar A., McHugh M., Neat F.C., Pade N.G., Queiroz N., Serra-Pereira B., Rapp T., Wearmouth V.J., & Genner M.J. (2010) Molecular markers reveal spatially segregated cryptic species in a critically endangered fish, the common skate (*Dipturus batis*). *Proceedings of the Royal Society of London B: Biological Sciences*, **277**, 1497–503.

Grubbs R.D. & Musick J.A. (2007) Spatial delineation of summer nursery areas for juvenile sandbar sharks in Cheasepeak Bay, Virginia. *American Fisheries Society Symposium*, **50**, 63-86.

Grüss A., Drexler M., & Ainsworth C.H. (2014) Using delta generalized additive models to produce distribution maps for spatially explicit ecosystem models. *Fisheries Research*, **159**, 11–24.

Guisan A., Edwards T.C., & Hastie T. (2002) Generalized linear and generalized additive models in studies of species distributions : setting the scene. *Ecological Modelling*, **157**, 89–100.

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

Guisan A. & Zimmermann N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Heikkinen R.K., Bocedi G., Kuussaari M., Heliola J., Leikola N., Poyry J. & Travis J.M.J. (2014) Impacts of land cover data selection and trait parameterisation on dynamic modelling of species' range expansion. *PLoS ONE* **9**: e108436.

Huber P.R., Greco S.E., Schumaker N.H. & Hobbs J. (2014) A priori assessment of reintroduction strategies for a native ungulate: using HexSim to guide release site selection. *Landscape Ecology*, **29**, 689-701.

Hunter E., Aldridge J.N., Metcalfe J.D. & Arnold G.P. (2003) Geolocation of free ranging fish on the European continental shelf as determined from environmental variables. *Marine Biology*, **142**, 601-609.

Iglésias S.P., Toulhoat L. & Sellos D.Y. (2010) Taxonomic confusion and market mislabelling of threatened skates: important consequences for their conservation status. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **20**, 319–333.

Jetz W., McPherson J.M. & Guralnick R.P. (2012) Integrating biodiversity distribution knowledge: Towards a global map of life. *Trends in Ecology and Evolution*, **27**, 151–159.

Jorgensen S. J., Reeb C.A., Chapple T. K., Anderson S., Perle C., Van Sommeran S. R., Fritz-Cope C., Brown A.C., Klimley A. P.& Block B. A. (2009) Philopatry and migration of Pacific white sharks. *Proceedings of the Royal Society of London B: Biological Sciences*: rspb20091155

Maggini R., Lehmann A., Zimmermann N.E., & Guisan A. (2006) Improving generalized regression analysis for the spatial prediction of forest communities. *Journal of Biogeography*, **33**, 1729–1749.

Moisen G.G. & Frescino T.S. (2002) Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling*, ***157***, 209–225.

Neat F., Pinto C., Burrett I., Cowie L., Travis J., Thorburn J., Gibb F., & Wright P.J. (2015) Site fidelity, survival and conservation options for the threatened flapper skate (*Dipturus cf. intermedia*). *Aquatic Conservation: Marine and Freshwater Ecosystems*, **25**, 6-20.

Neuenfeldt S., Righton D., Neat F., Wright P.J., Svedäng H., Michalsen K., Subbey S., Steingrund P., Thorsteinsson V., Pampoulie C., Andersen K.H., Pedersen M.W., & Metcalfe J. (2013) Analysing migrations of Atlantic cod *Gadus morhua* in the North-East Atlantic Ocean: then, now and the future. *Journal of Fish Biology*, **82**, 741–763.

Oppel S., Meirinho A., Ramírez I., Gardner B., O'Connell A.F., Miller P.I., & Louzao M. (2012) Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. *Biological Conservation*, **156**, 94–104.

Pedersen M.W., Righton D., Thygesen U.H., Andersen K.H., & Madsen H. (2008) Geolocation of North Sea cod (*Gadus morhua*) using hidden Markov models and behavioural switching. *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 2367–2377.

Petit C.C. & Lambin E.F. (2002) Impact of data integration technique on historical land-use / landcover change : comparing historical maps with remote sensing data in the Belgian Ardennes. *Landscape Ecology*, **17**, 117-132.

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

Philippart C.J. (1998) Long-term impact of bottom fisheries on several by-catch species of demersal fish and benthic invertebrates in the south-eastern North Sea. *ICES Journal of Marine Science:Journal du Conseil*, **55**, 342-352.

Pinto C., & Spezia L. (2015) Markov switching autoregressive models for interpreting vertical movement data with application to an endangered marine apex predator. *Methods in Ecology and Evolution.* DOI: 10.1111/2041-210X.12494.

Porzig E.L., Seavy N.E., Gardali T., Geupel G.R., Holyoak M., & Eadie J.M. (2014) Habitat suitability through time: using time series and habitat models to understand changes in bird density. *Ecopshere*, **5**, 1–16.

Pulliam H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349–361.

R Core Team (2014) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. http://www.R-project.org/

Raes, N. & H. ter Steege (2007) A null-model for significance testing of presence-only species distribution models. *Ecography*, **30**, 727-736.

Rogers L., Olsen E., Knutsen H., & Stenseth N. (2014) Habitat effects on population connectivity in a coastal seascape. *Marine Ecology Progress Series*, **511**, 153–163.

Ropert-Coudert Y., Beaulieu M., Hanuise N., & Kato A. (2009) Diving into the world of biologging. *Endangered Species Research*, **10**, 21–27.

Schumaker, N.H. (2013) HexSim (Version 2.4). U.S. Environmental Protection Agency, Environmental Research Laboratory, Corvallis, Oregon, USA. www.hexsim.net

Shearer K. A, Hayes J.W., Jowett I.G., & Olsen D. A (2015) Habitat suitability curves for benthic macroinvertebrates from a small New Zealand river. *New Zealand Journal of Marine and Freshwater Research*, 37–41.

Synes N.W., Watts K., Palmer S.C.F., Bocedi G., Bartoń K.A., Osborne P.E. & Travis J.M.J. (2015). A multi-species modelling approach to examine the impact of alternative climate change adaptation strategies on range shifting ability in a fragmented landscape. *Ecological Informatics*, **30**, 222-229.

Venables W.N. & Dichmont C.M. (2004) GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. *Fisheries Research*, **70**, 319–337.

Walker P.A. & Hislop J.R.G. (1998) Sensitive skates or resilient rays? Spatial and temporal shifts in ray species composition in the central and north western North Sea between 1930 and the present day. *ICES Journal of Marine Science*, **55**, 392–402.

Werry J.M., Planes S., Berumen M.L., Lee K.A., Braun C.D., & Clua E. (2014) Reef-fidelity and migration of tiger sharks, *Galeocerdo cuvier*, across the coral sea. *PLoS ONE*, **9**, e83249.

Whitlock R.E., McAllister M.K. & Block B.A. (2012) Estimating fishing and natural mortality rates for Pacific bluefin tuna (*Thunnus orientalis*) using electronic tagging data. *Fisheries Research*, **119-120**, 115–127.

Wood S.N. (2006) Generalized additive models: an introduction with R. *CRC press*.

Wood S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, **73**, 3-36.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 {Details on the individual tracks and preparation for the analysis.}

Appendix S2 {Figures showing the areas were the tracks were randomly generated to test for consistency with results from the observed tracks}

Appendix S3 {Details on the GAM modelling and its parameters with tables and figures reporting additional results on variables' colinearity, mixed models selection and model accuracy.}

Appendix S4 {Individual depth profiles showing time steps when individuals were recorded also by acoustic stations}

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCH

Cecilia Pinto is interested in applying scientific research to conservation practices, in particular developing methods to assess the state of data poor species in need for conservation. This study was an aspect of her PhD which researched the potential of integrating multiple data sources in an individual based dynamic model to define conservation measures for the endangered species *Dipturus* cf. i*ntermedia.* The remaining authors have diverse interests in ecology and conservation and apply a combination of practical and theoretical approached to conservation and species management.

Author contributions: C.P., J.A.T. and F.N. collected the data. C.P. carried on the spatial distribution model analysis, interpreted the results and led the writing of the manuscript. J.A.T. carried on the geolocation analysis. S.W. developed the modified geolocation model. T.C. supervised the analysis and corrected and made suggestions to the text. J.M.J.T., F.N., P.W. and B.S. corrected and made suggestions to the text.

TABLES

Table 1_model selection was based on AIC. Log-likelihood values show model significance.

	AIC	Log-lik
g(η) ~ *s*(latitude, longitude) + *s*(√depth) + *s*(distance from the coast) + *factor*(gear)	**378.2613**	**-177.4526**
g(η) ~ *s*(latitude, longitude) + *s*(√depth) + *factor*(gear)	383.2979	-179.5784
g(η) ~ *s*(√depth) + *s*(distance from the coast) + *factor*(gear)	377.6462	-179.6911

FIGURES

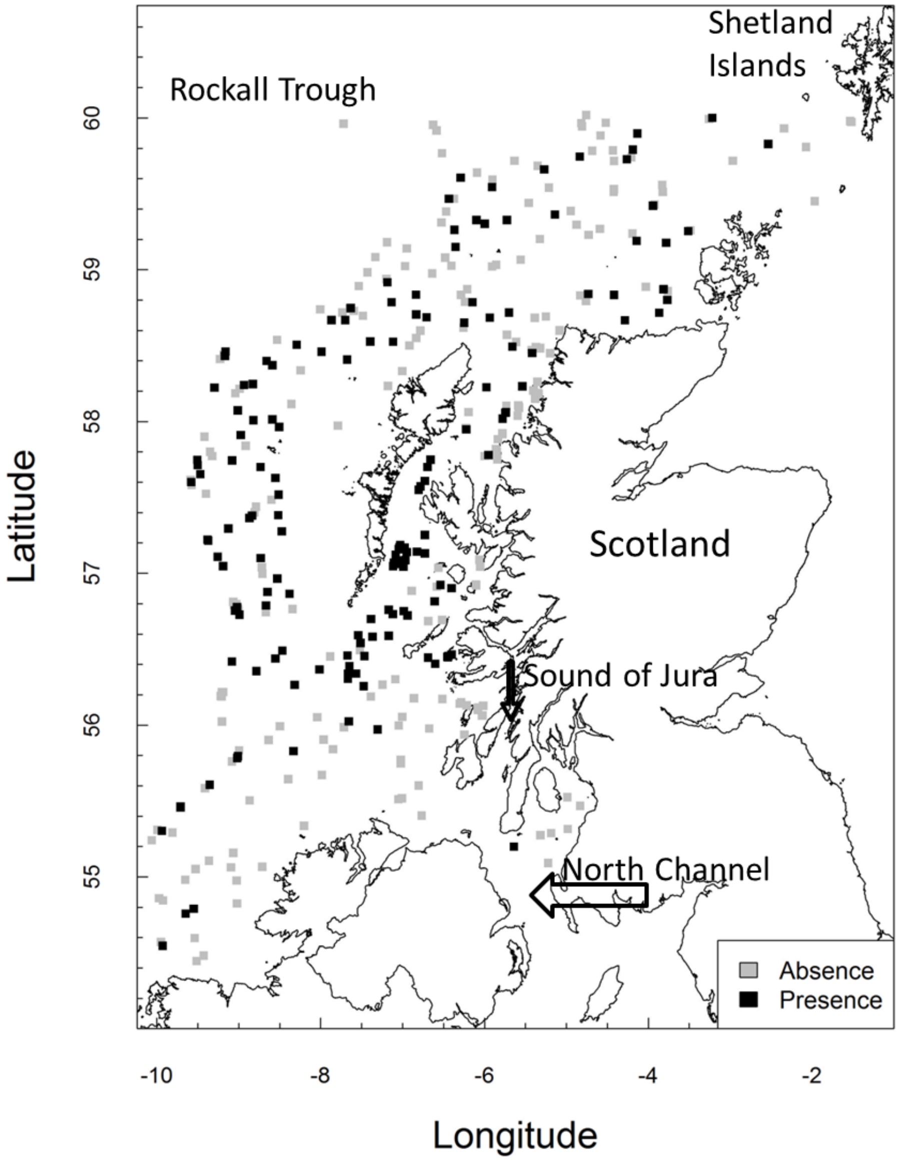

Figure 1 Locations of all bottom trawl surveys around Scotland (UK) from which presence-absence records of flapper skate were extracted.

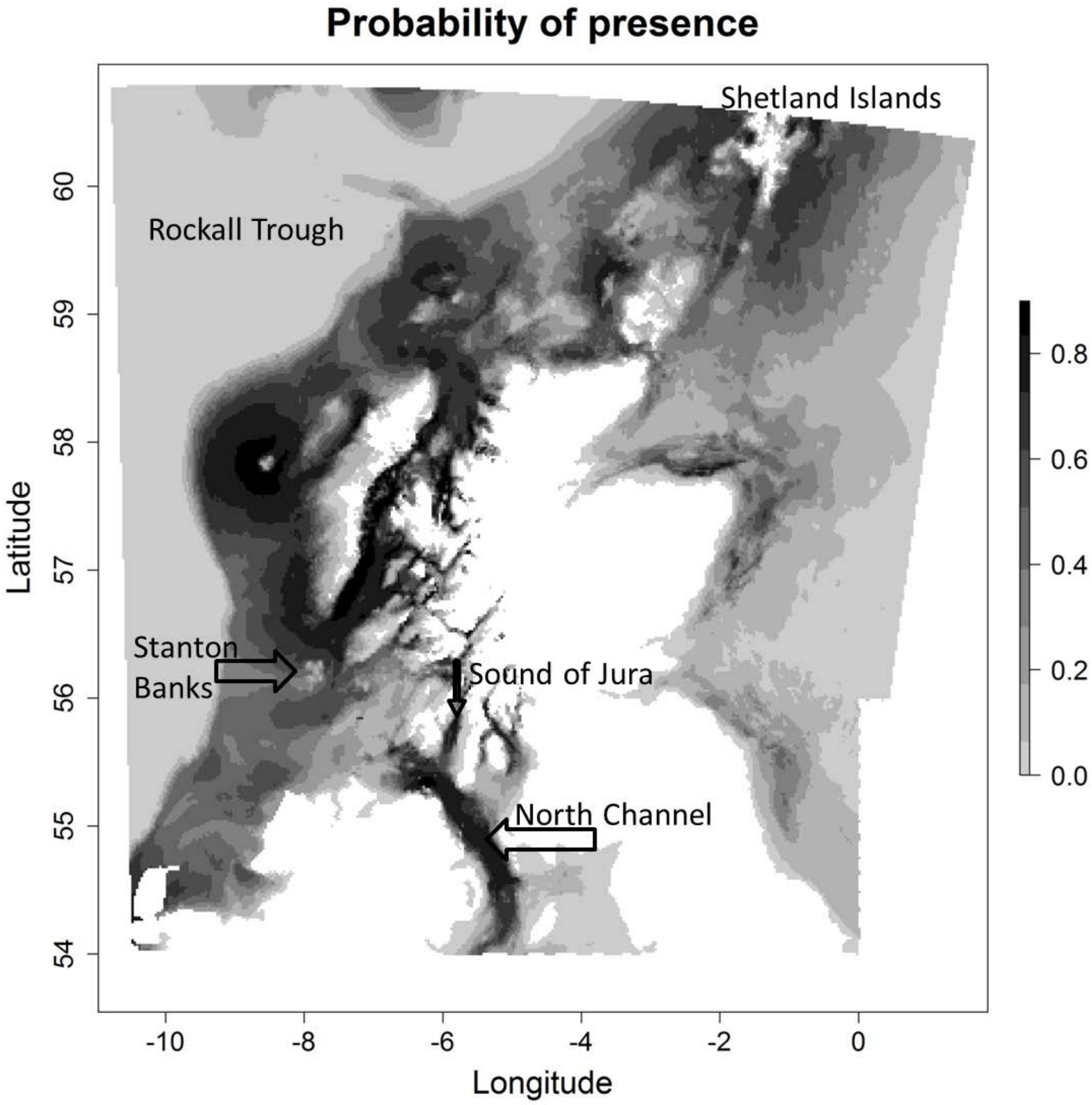

Figure 2 Probability of presence of flapper skate around Scotland as predicted from the GAM. As no records from the east coast of Scotland were available, predictions in that area should not be considered reliable.

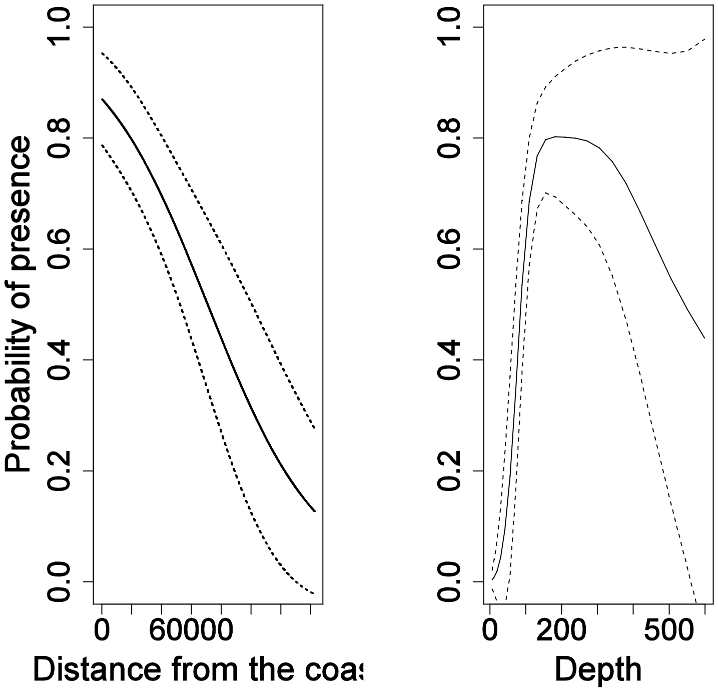

Figure 3 Predicted probability of presence of flapper skate from a GAM in relation to distance from the coast and depth. Dotted lines indicate 95% confidence intervals.

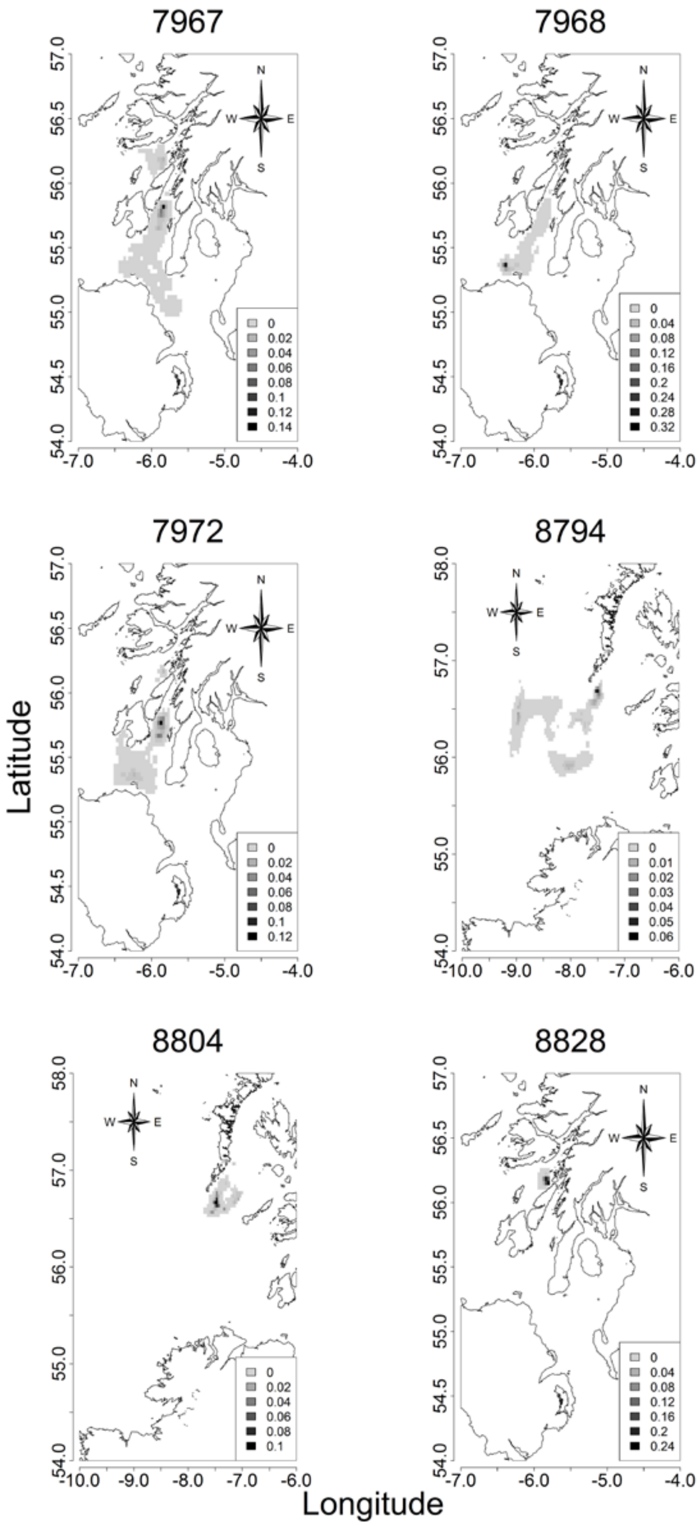

Figure 4 Utilization distribution map (UDM) estimated by the geolocation model for each tagged flapper skate off the west coast of Scotland. Each cell of a track has a different probability value as the UDM is an average of all possible tracks predicted by the model. This directly accounts for the model error in the UDM.

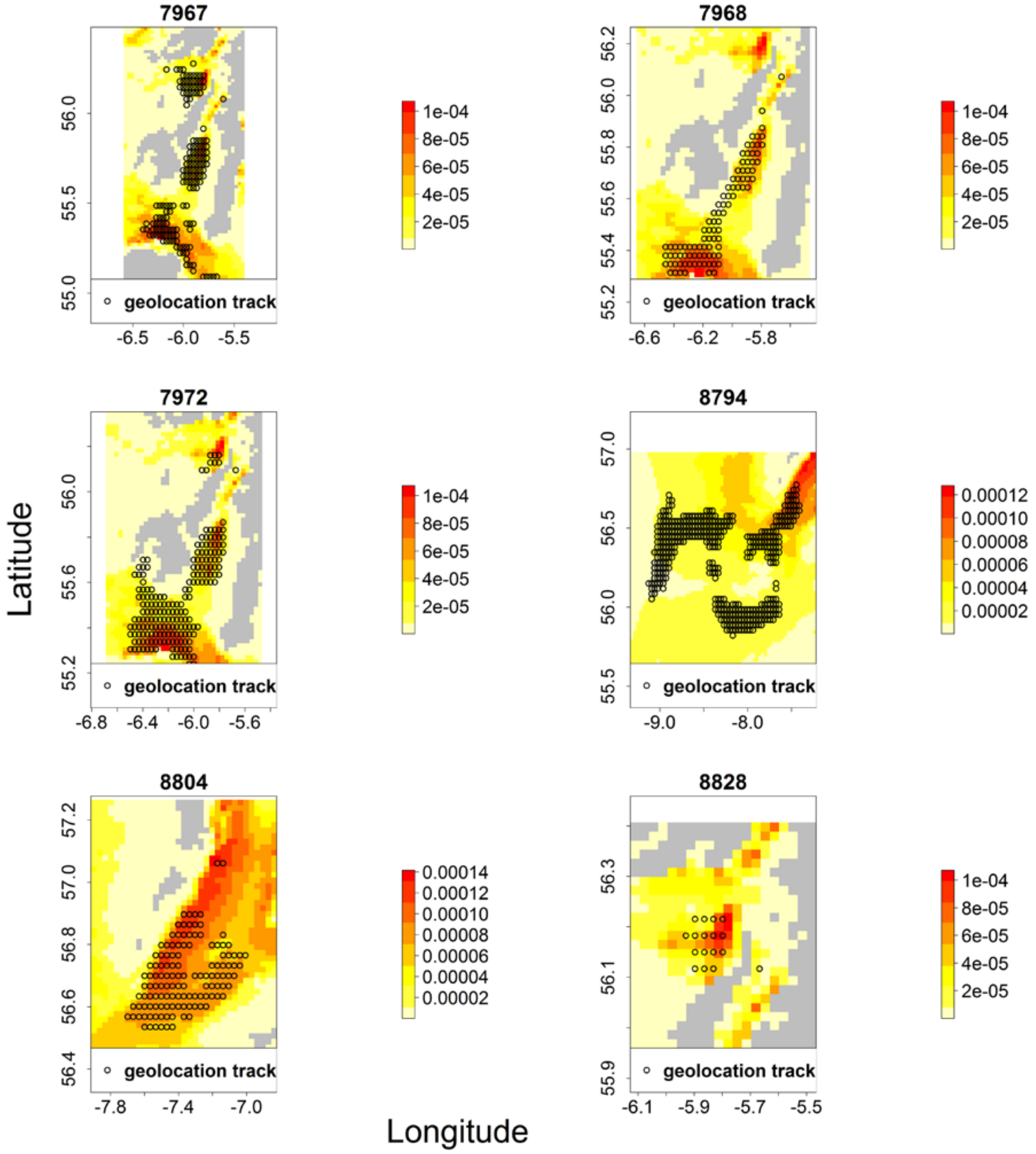

Figure 5 Estimated tracks of each individual (black circles) plotted over the relative habitat utilization predicted from the GAM (see legend for values). Differently from Figure 4 here the tracks' cells are plotted without representing the different probability values. The grey areas correspond to land.

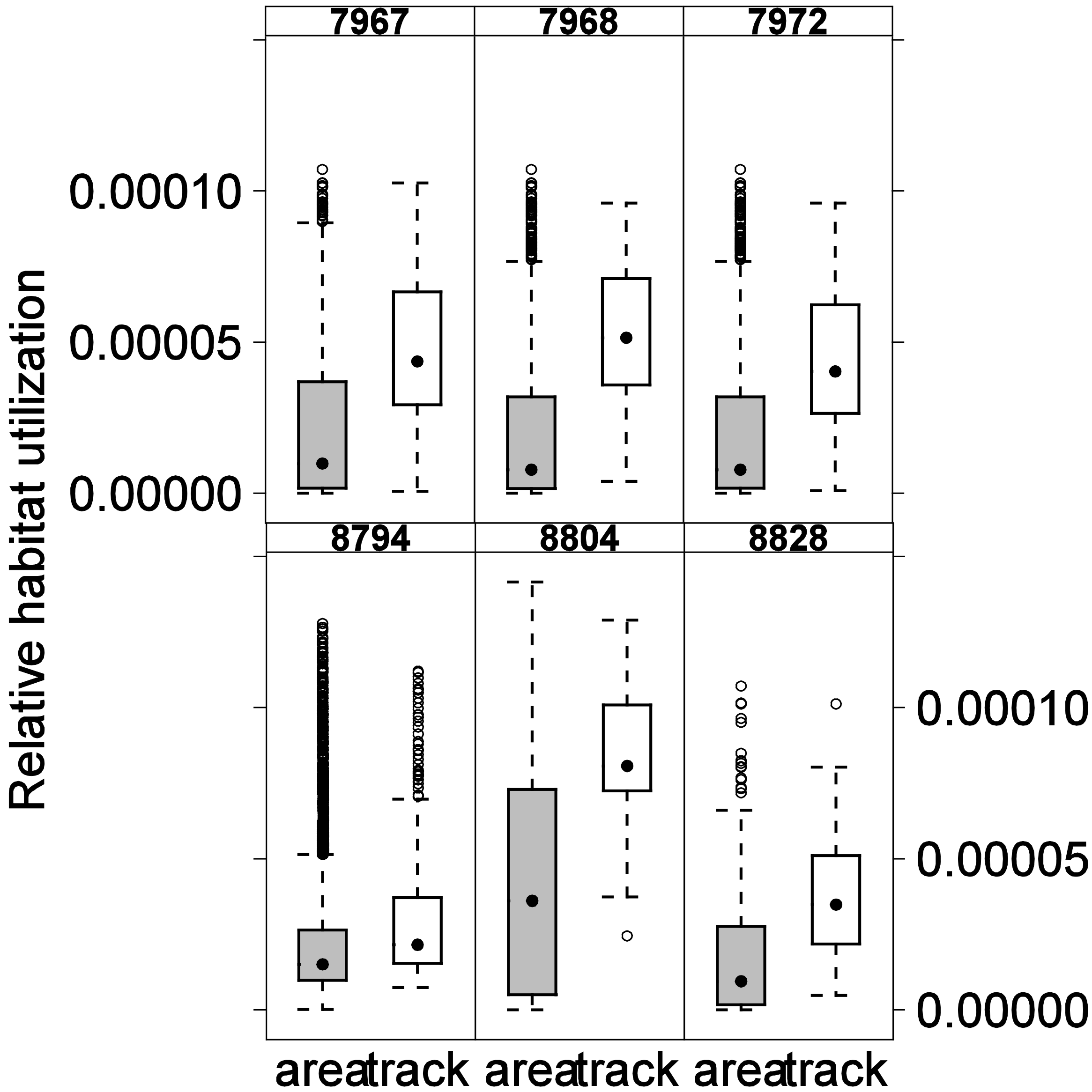

Figure 6 Distribution of the relative habitat utilisation predicted in the area covered by the geolocated track plus a 2’ buffer (grey boxplot) against the distribution of relative habitat utilisation predicted at the track exact locations (white boxplot) for each tagged flapper skate. Values of relative habitat utilisation at exact tracks’ locations are always higher.

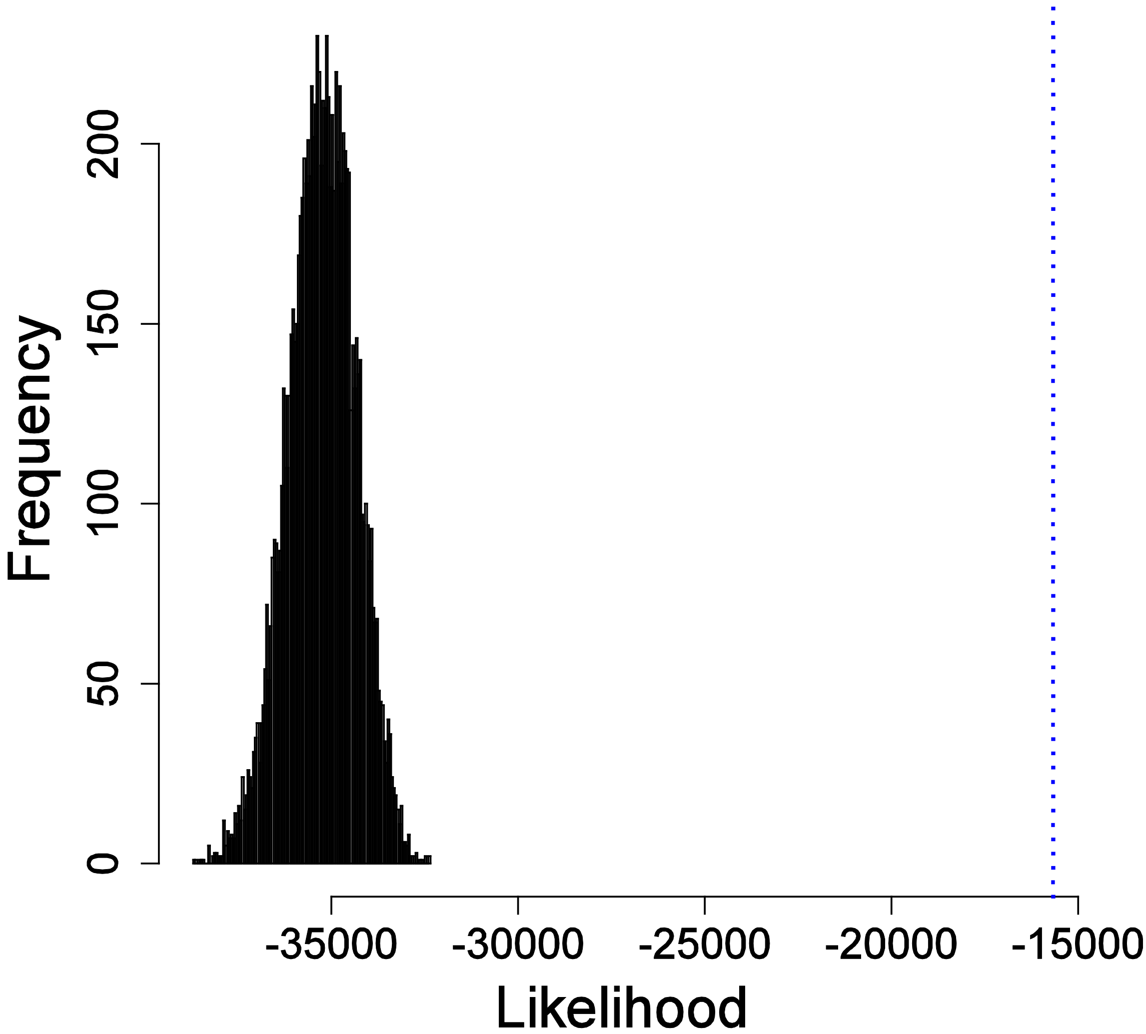

Figure 7 The dashed vertical line represents the RHU-likelihood at the six observed tracks locations combined. The histogram represents the distribution of RHU-likelihoods at 10,000 randomised tracks.