

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

2

3 Cecilia Pinto<sup>1,2</sup>, James A. Thorburn<sup>1</sup>, Francis Neat<sup>2</sup>, Peter J. Wright<sup>2</sup>, Serena Wright<sup>3</sup>, Beth E.  
4 Scott<sup>1</sup>, Thomas Cornulier<sup>1</sup> and Justin M.J. Travis<sup>1</sup>.

5

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

10

11 **ABSTRACT**

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

17 **Location** West coast of Scotland, NE Atlantic.

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

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

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

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

35

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

38

## 39 (A) INTRODUCTION

40 Describing how a species is distributed in space, defining its preferred habitat and  
41 establishing which environmental characteristics best support its populations, are key to  
42 understanding the ecology of threatened or declining species (Guisan & Zimmermann, 2000)  
43 and planning for their conservation (Pulliam, 2000). Commonly a species' distribution is  
44 obtained from coupling field data with corresponding environmental variables within a  
45 modelling framework (Austin, 2002; Aarts *et al.*, 2008). One of the main advantages of  
46 species distribution models (SDMs) is that they may be used to generate predictions for the  
47 species distribution beyond the area originally sampled, provided the prediction is performed

48 within the environmental range sampled (Elith *et al.*, 2010). SDMs thus have the potential to  
49 inform broader scale management which is especially important for marine species where it  
50 is often difficult and costly to sample the entire range of a species.

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

59 Statistical tools commonly used in SDMs include random forest regression trees, MaxEnt  
60 (Phillips *et al.*, 2006), generalized linear models (GLMs) and generalized additive models  
61 (GAMs). GAMs have been shown to perform as well (Opper *et al.*, 2012), if not better  
62 (Moisen & Frescino, 2002; Aertsen *et al.*, 2010), than other predictive models. GAMs are  
63 more flexible in fitting complex non-linear responses (Aarts *et al.*, 2008), and can  
64 compensate for over-fitting through the use of a penalized likelihood (Venables & Ripley,  
65 2004). They do, however, require a high number of degrees of freedom in order to perform  
66 well and give reliable predictions (Wood, 2006; Drexler & Ainsworth, 2013). Thus, for  
67 predictions on environmental suitability obtained by GAMs, as well as for those obtained  
68 using other statistical or machine learning approaches, their reliability should be tested  
69 through field validation or by finding alternative ways of testing model outputs.

70 Model validation is important when extrapolating model outputs to non-sampled areas (Elith  
71 & Leathwick, 2009) and specifically when the ecology of the species of interest is poorly  
72 known. As field validation requires entails significant economic and time investment, test

73 datasets (Drexler & Ainsworth, 2013), testing against a null model (Raes & ter Steege, 2007)  
74 or bootstrapping (Elith & Leathwick, 2009) have been used as alternative methods.  
75 Comparing model outputs and the spatial distribution of independent data obtained from  
76 different sampling sources has been used as an option to validate model outputs (Grubbs &  
77 Musick, 2007), to reduce estimate and prediction uncertainty (Petit & Lambin, 2002; Jetz *et*  
78 *al.*, 2012) and to help cross-validate outputs of models obtained from independent sets of data  
79 (Rogers *et al.*, 2014).

80 There has been a rapid increase in the volume of spatial data derived from electronic tagging  
81 devices, often referred to as ‘biologgers’, on individual movements of animals across a broad  
82 range of taxa. Following the definition in Ropert-Coudert *et al.*, (2009), biologgers comprise  
83 storage tags, archival tags and electronic data recorders. Many of the earliest applications of  
84 this approach were on seabirds, pinnipeds, cetaceans and sea turtles (Ropert-Coudert *et al.*,  
85 2009). The rapid uptake of electronic tagging for marine species was due to the benefits  
86 provided from observations of underwater behaviour and the gathering of positional  
87 information at sea. Tagging of fish species, however, has lagged behind because of the  
88 greater difficulty of acquiring reliable positional information on sub-surface species.  
89 However, the advent of geolocation models, either using tidal signatures or light intensity  
90 levels, is now resulting in increased knowledge on the spatial ecology of an increasing  
91 number of fish species of both economic and conservation concern, including pacific bluefin  
92 tuna (Whitlock *et al.*, 2012), cod (Neuenfeldt *et al.*, 2013), white sharks (Jorgensen *et al.*,  
93 2009) and tiger sharks (Werry *et al.*, 2014). Geolocation models use contemporaneous  
94 environmental information, such as light level and depth, to estimate the individual’s most  
95 likely geographical locations at a certain time-step. Assuming that an individual spends more  
96 time in its preferred habitat, estimated individual tracks are an ideal independent source of  
97 information to infer environmental suitability and to test predictions obtained from SDMs.

98 This study examines the distribution of flapper skate (*Dipturus cf. intermedia*) off the west  
99 coast of Scotland. The flapper skate is listed in the IUCN Red List of Threatened Species as  
100 “Critically Endangered” ([www.iucnredlist.org](http://www.iucnredlist.org)). As a slow growing, late maturing and low  
101 fecundity species, its population growth rate is highly sensitive to fishing mortality (Brander,  
102 1981). The species suffered a rapid decline in the last 40 years with landings falling by 90 per  
103 cent (Du Buit, 1977; Brander, 1981; Philippart, 1998). The flapper skate is now only  
104 occasionally found in the North Sea, and its former distribution contracted throughout the  
105 period, leaving a number of relict populations off the West coast of Scotland (Brander, 1980;  
106 Walker & Hislop, 1998; Daan *et al.*, 2005). In order to estimate the potential for this species  
107 to recolonize its former range, it is fundamental to understand the environmental influences  
108 determining its distribution. To this aim we used SDMs to define environmental suitability  
109 for the flapper skate off the west coast of Scotland from presence-absence data obtained from  
110 trawl survey data, and used individual geolocation estimates from electronic tagging devices  
111 to validate model predictions. The aim of this study was to demonstrate the potential of  
112 integrating information from individual tracking data (obtained from tidal geolocation  
113 modelling of electronic data storage tagging devices in this study), for the validation of SDM  
114 predictions.

## 115 (A) METHODS

### 116 (B) *Species distribution data*

117 Presence-absence data on the flapper skate were obtained from several trawl surveys  
118 including the Marine Scotland Science northern shelf monkfish survey and the International  
119 Bottom Trawl Survey (Fig.1). The *Dipturus batis* complex was identified as two species  
120 (*Dipturus intermedia* and *Dipturus flossada*) in 2010 (Griffiths *et al.*, 2010; Iglésias *et al.*,  
121 2010), and therefore data on the flapper skate were generally available only from 2010  
122 onwards (330 records), although a few records (65) were obtained from surveys conducted in

123 2003, 2004 and 2005 during which catch records were matched to individual photographs  
124 which allowed identification to the species level. The surveys each used a different trawl gear  
125 and this may have led to different catchabilities for flapper skate (Appendix S3). The areas  
126 covered by the different surveys did, however, broadly overlap, just that some had a more  
127 restricted areal coverage than others.

#### 128 (B) *Predictor variables*

129 The environmental variables used as predictors in the SDM model were all projected in UTM  
130 29N (WGS84) and all had the same resolution of 30" (790.2m). Environmental variables  
131 included: trawl shot latitude and longitude, depth, slope (reported as angle measures), mean  
132 salinity, mean temperature, distance from the coast, seabed composition (sediments) and gear  
133 type. Depth and slope were obtained from OCEANWISE 6" ([www.oceanwise.eu](http://www.oceanwise.eu)) and  
134 INFOMAR ([www.infomar.ie](http://www.infomar.ie)); depth was square root-transformed for modelling purposes as  
135 the raw data had a skewed distribution. Salinity and temperature were obtained from the  
136 freely available oceanographic model EUROPEAN NORTH WEST SHELF – OCEAN  
137 PHYSICS REANALYSYS FROM METOFFICE (1985-2012) ([www.myocean.eu](http://www.myocean.eu)).  
138 Euclidean distance to the nearest coast was calculated in ArcGIS 10. The sediments layer was  
139 extracted from the British Geological Survey database (European Marine Observation and  
140 Data Network, EMODNET, [www.emodnet-geology.eu](http://www.emodnet-geology.eu)) and is represented by seven classes  
141 of sediments: coarse sands, mixed sediment, mud to sandy mud, rock, sand to muddy sand,  
142 seabed (unknown sediment) and till (mixed sediments). Preparation of the final dataset to be  
143 used for predictions was performed in R 3.1.1 (R Core Team, 2014).

#### 144 (B) *GAM fitting*

145 A binomial GAM with a logit link function was fitted using the "mgcv" package in R (Wood,  
146 2006, 2011). Thin plate regression splines were used as smoothing functions for the

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

170 (B) *Geolocation modelling*

171 Independent data from six data storage tags (DSTs) attached to common skate and recovered  
172 between 2012 and 2014 were used. DSTs were deployed on 18 individuals in an area known  
173 as the Sound of Jura (4 recaptures) and on 29 individuals from an area known as the Stanton  
174 Banks (2 recaptures). DSTs were attached externally to the fish (Neat *et al.*, 2015) and  
175 recorded hydrostatic pressure and environmental temperature every 2 minutes. For the  
176 biological characteristics of the tagged individuals and the total length of the time series see  
177 Appendix S1 in Supporting Information (where S indicates Supporting). Notably, three  
178 (7968, 7967 and 7972) of the six individuals tagged with DSTs were also tagged with  
179 acoustic transmitters connected to a set of acoustic receiver stations which were active for  
180 one year in the northern section of the Sound of Jura (Neat *et al.*, 2015).

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



196 depth. The spatial extents of the model of release and recapture locations at a resolution of  
197 approximately 7km, were -32W, 35N, 11E, 70N.

198 A second output of the model is the average of all possible tracks an individual could have  
199 covered during the tagging period, producing a density map called a utilization distribution  
200 map (UDM). As the UDM is a distribution of all possible tracks predicted by the model, it  
201 directly includes a measure of the uncertainty about the true track. The UDM was used to  
202 compare the geolocated locations against the predicted probability of presence obtained from  
203 the GAM. The UDM is a probability map where each cell has a value between 0 and 1  
204 ( $\Sigma = 1$ ), the higher the value the greater the probability the individual spent time in that cell.  
205 Because the model calculates a probability for each cell being part of the total grid (the final  
206 size of the grid is optimized by the model based on the diffusivity value (Pedersen *et al.*,  
207 2008)) the final output needs to be rescaled after excluding probabilities that are too close to  
208 zero (therefore far from the actual individual track), which would shrink all probabilities  
209 towards zero (see Appendix S1). This output is produced at a 2km resolution.

## 210 (B) *Model validation*

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

$$213 \quad RHU(x) = \frac{\exp(g(x))}{\sum_D \exp(g(x))}$$

214 where  $g(x)$  is the linear predictor of the GAM at location  $x$  part of study domain  $D$ , so that  
215 RHU is on a scale proportional to time spent by an animal at that location and to density of  
216 observations (Aarts *et al.*, 2012). The RHU was compared to geolocation estimates in three  
217 separate steps. First, in order to facilitate an initial visual estimation of whether the tracks  
218 produced by the geolocation model covered either high or low predicted RHU obtained from  
219 the GAM, the individual tracks obtained from the UDM predictions were plotted on top of

220 the RHU map. Secondly, the distribution of RHU values for an area surrounding each  
221 individual track (i.e. the area is defined by an individual track plus a 2' buffer) was compared  
222 to the distribution of RHU values extracted at track locations, in order to compare which  
223 values of the predicted distribution were actually selected by the individual along its track.  
224 This process was implemented separately for each individual. Lastly, to verify that the  
225 selection of high probability of presence areas by individual tracks was effectively better than  
226 a random selection of areas, we assimilated the RHU for the whole area to a likelihood  
227 distribution (termed 'RHU-likelihood' for simplicity thereafter). This last process was  
228 considering all the individual tracks together at once. To account for the probability assigned  
229 by the geolocation model to each track cell belonging to the UDM, and to give each  
230 individual equal weight in the analysis, track cells were resampled with replacement for  
231 2,000 draws (as this is the size of the locations of all the tracks put together) proportionally to  
232 cell probabilities' value ("standardized tracks") of the UDM. We then performed 10,000  
233 simulations of random track locations to calculate their respective RHU-likelihood. The final  
234 output was than the difference between the sum of the  $\log(\text{RHU-likelihood})$  at the 2,000  
235 observed track locations and the sum of the  $\log(\text{RHU-likelihood})$  at 2,000 simulated locations  
236 for each of the 10,000 randomly generated sets of tracks. In order to preserve the internal  
237 spatial structure of the animal tracks, simulations were done by anchoring six points  
238 generated at random over the study area which would define the new centroid of each  
239 individual track (see Appendix S2). The random tracks were generated as a set of locations  
240 with the same shape of the original tracks but centred around the position defined by the  
241 randomly generated centroid. When parts of random tracks were generated on land these  
242 locations were excluded and the section of the track generated at sea was resampled with  
243 repetition until the original sample size of the track was reached. The 10,000 simulations  
244 were performed twice using two nested spatial domains to produce both a regional and a

245 more stringent local test of the model performance. The regional polygon was drawn around  
246 the area covered by the raw data of presence-absence (Fig.1), and the local polygon was  
247 drawn around the area covered by the geolocated tracks (see Appendix S2). The use of two  
248 spatial scales allowed us to assess both the reliability of our predictions and the  
249 representativeness of the geolocation estimates.

## 250 (A) RESULTS

### 251 (B) *GAM fitting*

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

269 high depths but at the same time are surrounding islands or are constrained within islands and  
270 the main land.

271 (B) *Geolocation modelling*

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

290 (B) *Cross validation*

291 The visual exploration of locations estimated by the geolocation model and the corresponding  
292 area extracted from the GAM predictions (Fig.5) showed a high overlap between the UDM

293 predictions and the cells with the highest RHU values predicted by the GAM. The values  
294 corresponding to the single track locations for all six individuals were always distributed  
295 among the highest RHU values (Fig.6). At the regional scale, the likelihood that the observed  
296 geolocated tracks coincided with areas of high values of RHU was always higher than the  
297 likelihood that the randomly generated tracks would fall over high RHU values (Fig.7). At  
298 the local scale, a very similar result was obtained (this second result is not shown).

## 299 (A) DISCUSSION

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

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

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

343 sets of data are available, between data sets cross-validation should be used, taking advantage  
344 of the independency of data sets which reduces bias and increases statistical power.

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

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

### 383 (A) CONCLUSIONS

384 We showed how integrating independent sets of data and different modelling procedures can  
385 help validate model predictions reducing the uncertainty surrounding such estimates. This  
386 approach combined static observations with individual tracking data, taking advantage of the  
387 strengths of both information sources: the higher sample sizes of distribution data and the real  
388 time habitat use from individual tracks. The integration process can help in the definition of  
389 effective conservation measures for endangered species and to assess the efficacy of those  
390 already existing. Considering the increasing volumes of data collected at the individual level  
391 (Block *et al.*, 2011), the development of methods to integrate independent sources of data is



392 of high value in the marine environment. Visual comparison of outputs can be useful for  
393 communicating findings to stakeholders when defining ecosystem based management  
394 frameworks, after it has been formally backed-up with quantitative evidence. Finally model  
395 validation improved the confidence in using data with relatively low power to inform  
396 conservation management and to direct future data collection to improve on-going adaptive  
397 conservation planning.

#### 398 **(A) ACKNOWLEDGEMENTS**

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

408

#### 409 **(A) REFERENCES**

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

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

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

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

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

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

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

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

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

432 Brander K. (1981) Disappearance of common skate *Raja batis* from Irish Sea. *Nature*, **290**,  
433 48–49.

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

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

439 Daan N., Heessen H., & Hofstede R. (2005) North Sea Elasmobranchs : distribution ,  
440 abundance and biodiversity. *CM-International Council for the Exploration of the Sea*, **CM 06**  
441 1–15.

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

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

446 Elith J. & Leathwick J.R. (2009) Species distribution models: ecological explanation and  
447 prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**,  
448 677–697.

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

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

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

456 Griffiths A.M., Sims D.W., Cotterell S.P., El Nagar A., Ellis J.R., Lynghammar A., McHugh  
457 M., Neat F.C., Pade N.G., Queiroz N., Serra-Pereira B., Rapp T., Wearmouth V.J., & Genner  
458 M.J. (2010) Molecular markers reveal spatially segregated cryptic species in a critically

459 endangered fish, the common skate (*Dipturus batis*). *Proceedings of the Royal Society of*  
460 *London B: Biological Sciences*, **277**, 1497–503.

461 Grubbs R.D. & Musick J.A. (2007) Spatial delineation of summer nursery areas for juvenile  
462 sandbar sharks in Cheasepeak Bay, Virginia. *American Fisheries Society Symposium*, **50**, 63-  
463 86.

464 Grüss A., Drexler M., & Ainsworth C.H. (2014) Using delta generalized additive models to  
465 produce distribution maps for spatially explicit ecosystem models. *Fisheries Research*, **159**,  
466 11–24.

467 Guisan A., Edwards T.C., & Hastie T. (2002) Generalized linear and generalized additive  
468 models in studies of species distributions : setting the scene. *Ecological Modelling*, **157**, 89–  
469 100.

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

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

474 Heikkinen R.K., Bocedi G., Kuussaari M., Heliola J., Leikola N., Poyry J. & Travis J.M.J.  
475 (2014) Impacts of land cover data selection and trait parameterisation on dynamic modelling  
476 of species' range expansion. *PLoS ONE* **9**: e108436.

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

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

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

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

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

492 Maggini R., Lehmann A., Zimmermann N.E., & Guisan A. (2006) Improving generalized  
493 regression analysis for the spatial prediction of forest communities. *Journal of Biogeography*,  
494 **33**, 1729–1749.

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

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

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

504 Oppel S., Meirinho A., Ramírez I., Gardner B., O’Connell A.F., Miller P.I., & Louzao M.  
505 (2012) Comparison of five modelling techniques to predict the spatial distribution and  
506 abundance of seabirds. *Biological Conservation*, **156**, 94–104.

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

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

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

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

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

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

524 Pulliam H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**,  
525 349–361.

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

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

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

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

534 Schumaker, N.H. (2013) HexSim (Version 2.4). U.S. Environmental Protection Agency,  
535 Environmental Research Laboratory, Corvallis, Oregon, USA. [www.hexsim.net](http://www.hexsim.net)

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

539 Synes N.W., Watts K., Palmer S.C.F., Bocedi G., Bartoń K.A., Osborne P.E. & Travis J.M.J.  
540 (2015). A multi-species modelling approach to examine the impact of alternative climate  
541 change adaptation strategies on range shifting ability in a fragmented landscape. *Ecological  
542 Informatics*, **30**, 222-229.

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

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

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

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

553 Wood S.N. (2006) Generalized additive models: an introduction with R. *CRC press*.  
554 Wood S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood  
555 estimation of semiparametric generalized linear models. *Journal of the Royal Statistical*  
556 *Society (B)*, **73**, 3-36.

557

## 558 **SUPPORTING INFORMATION**

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

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

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

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

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

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

## 572 **BIOSKETCH**



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

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

585

586

587

588

589

590

591

592

593

594

595

596

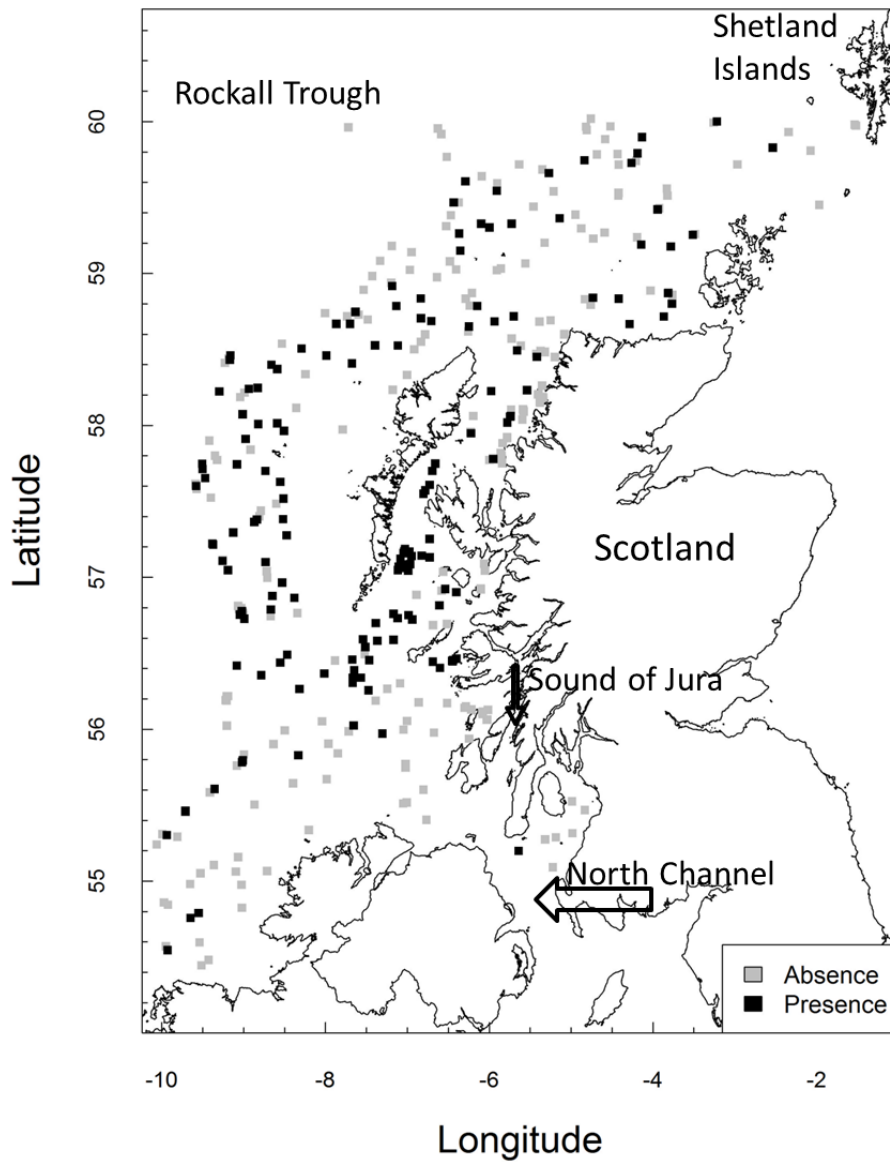
597

598 **TABLES**

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

	AIC	Log-lik
<b><math>g(\eta) \sim s(\text{latitude, longitude}) + s(\sqrt{\text{depth}}) + s(\text{distance from the coast}) + \text{factor}(\text{gear})</math></b>	<b>378.2613</b>	<b>-177.4526</b>
$g(\eta) \sim s(\text{latitude, longitude}) + s(\sqrt{\text{depth}}) + \text{factor}(\text{gear})$	383.2979	-179.5784
$g(\eta) \sim s(\sqrt{\text{depth}}) + s(\text{distance from the coast}) + \text{factor}(\text{gear})$	377.6462	-179.6911

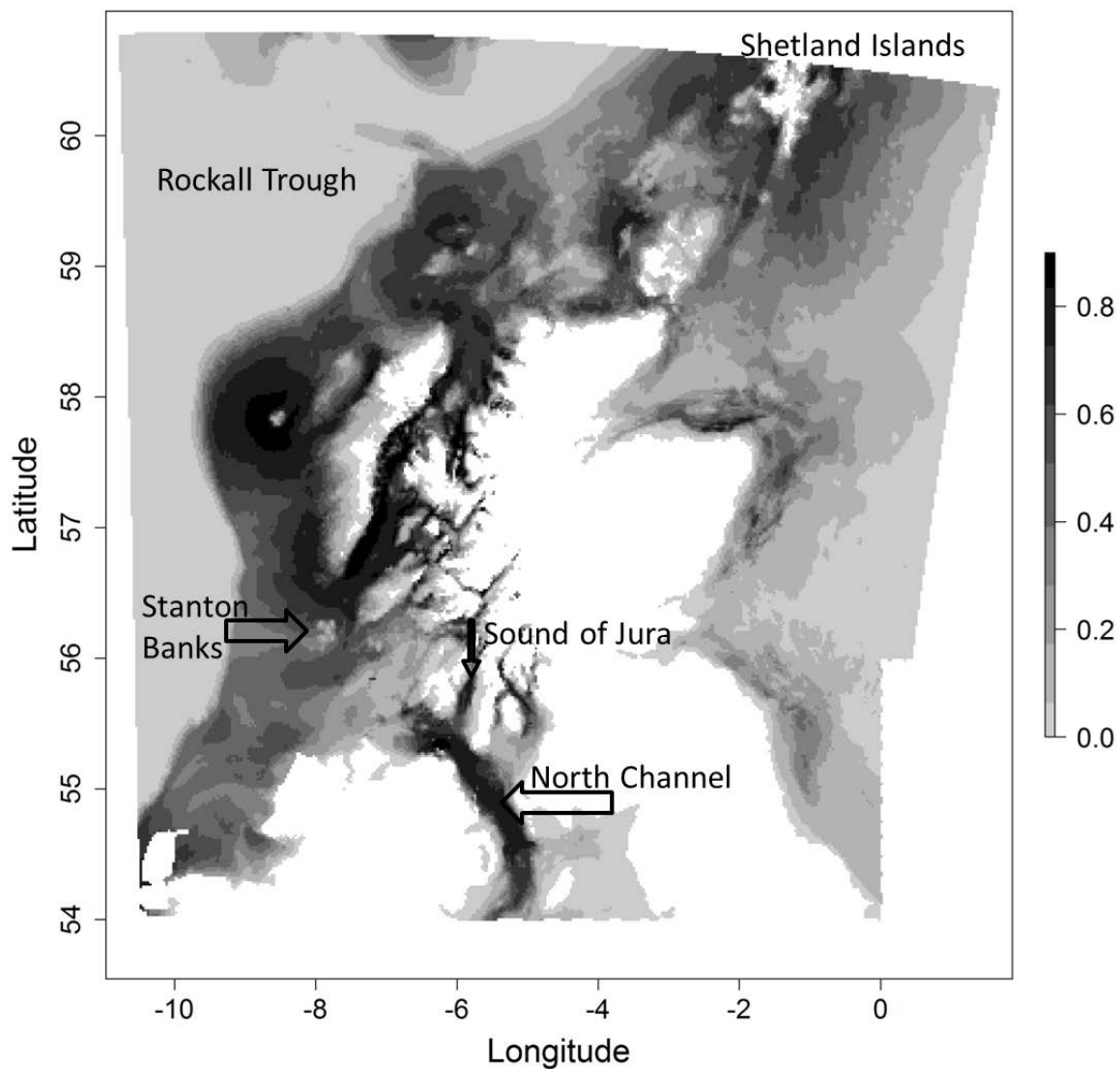
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617



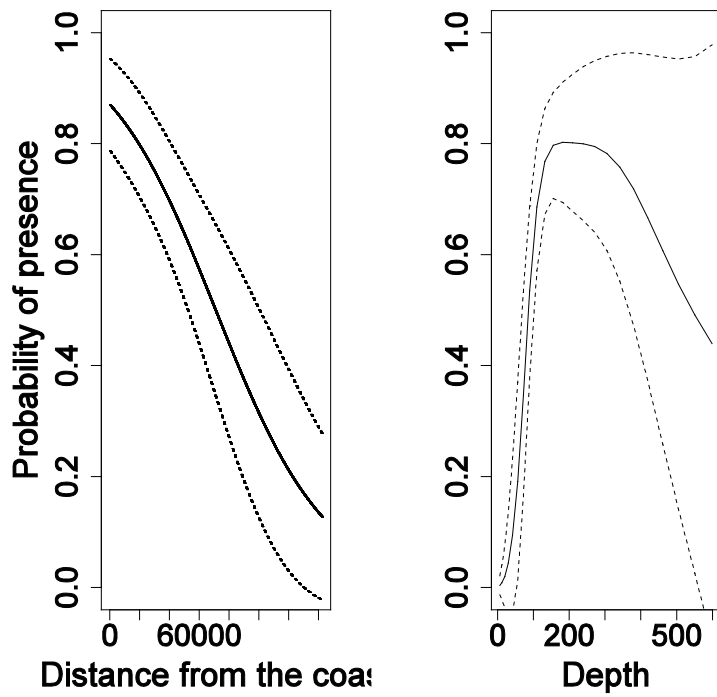
619  
620 Figure 1 Locations of all bottom trawl surveys around Scotland (UK) from which presence-  
621 absence records of flapper skate were extracted.

622

## Probability of presence



623  
624 Figure 2 Probability of presence of flapper skate around Scotland as predicted from the  
625 GAM. As no records from the east coast of Scotland were available, predictions in that area  
626 should not be considered reliable.



627

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

630

631

632

633

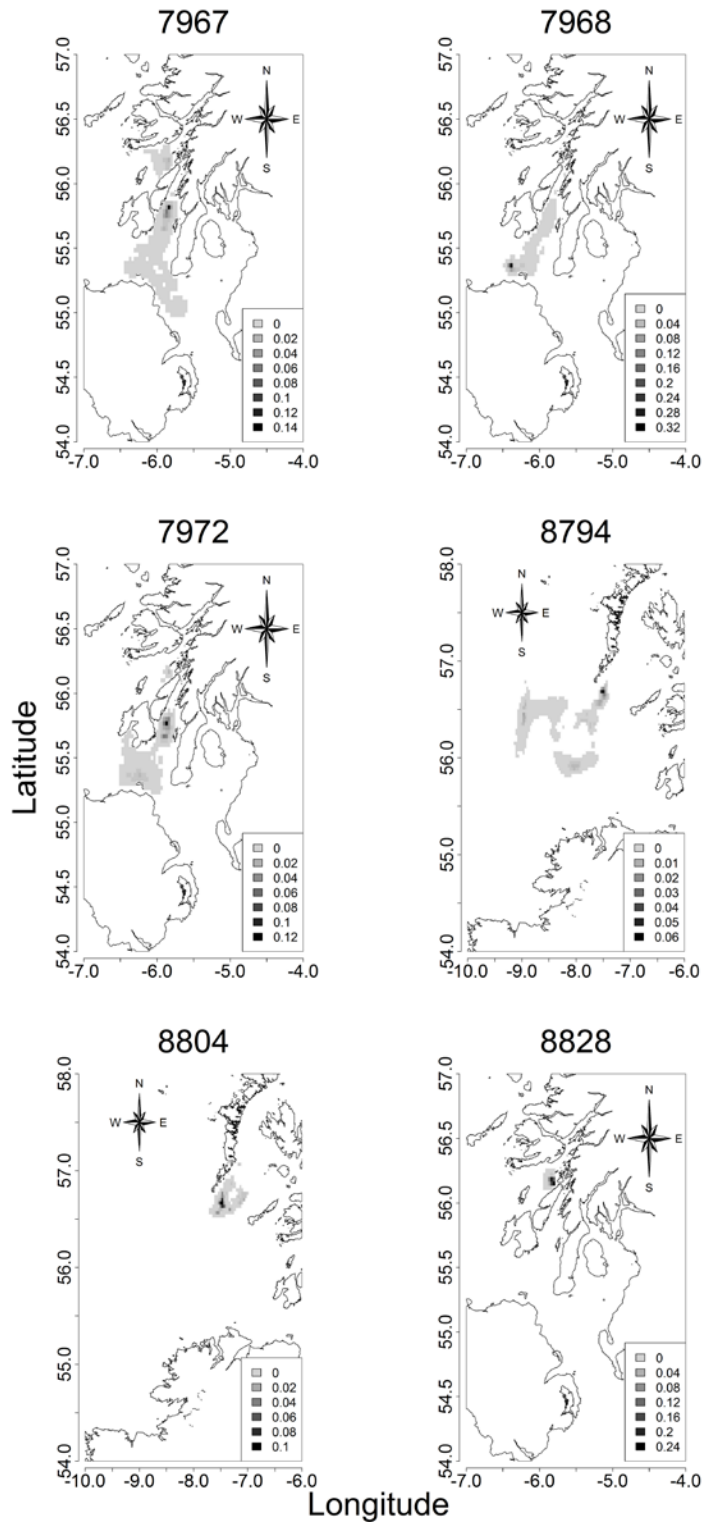
634

635

636

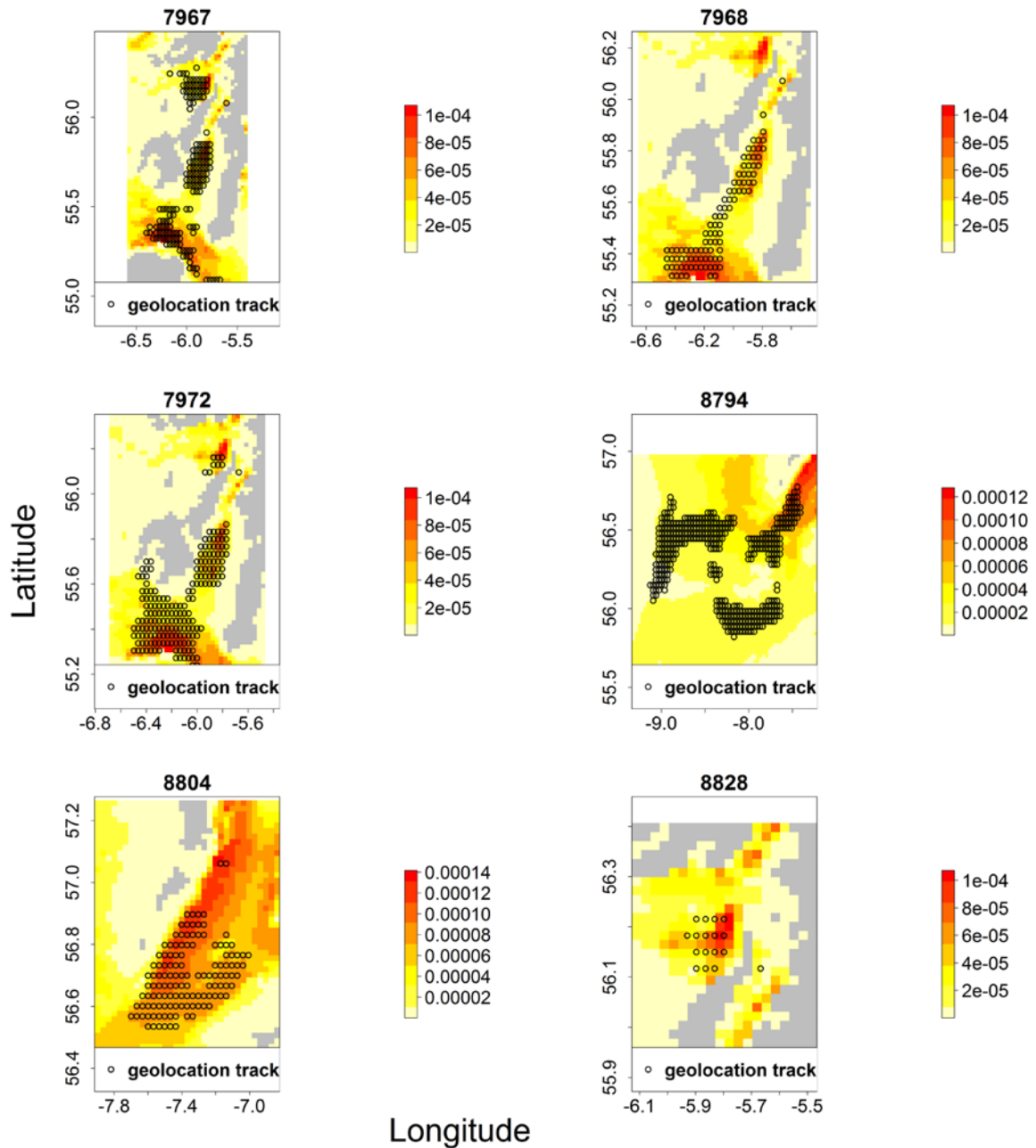
637

638

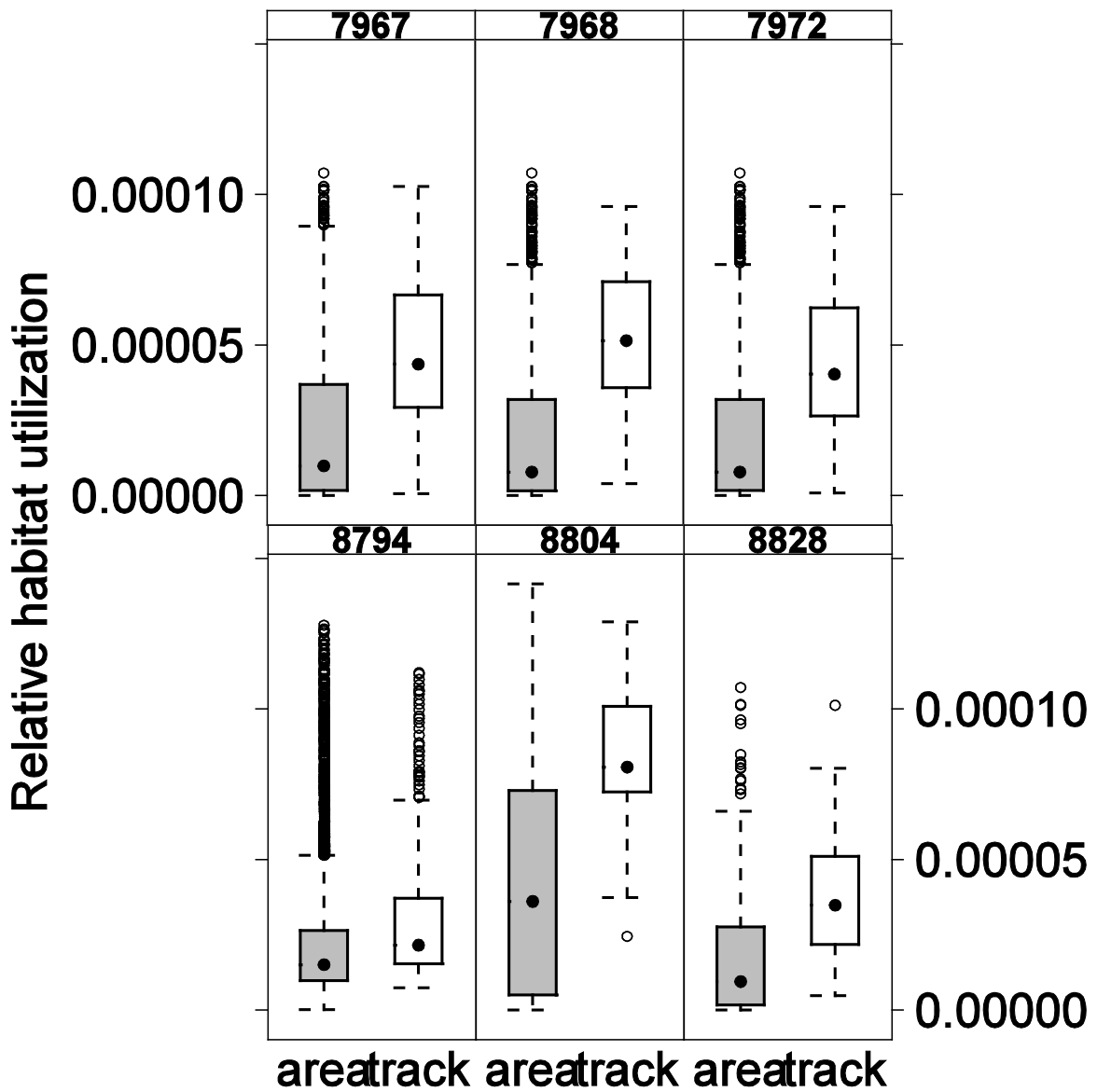


639

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

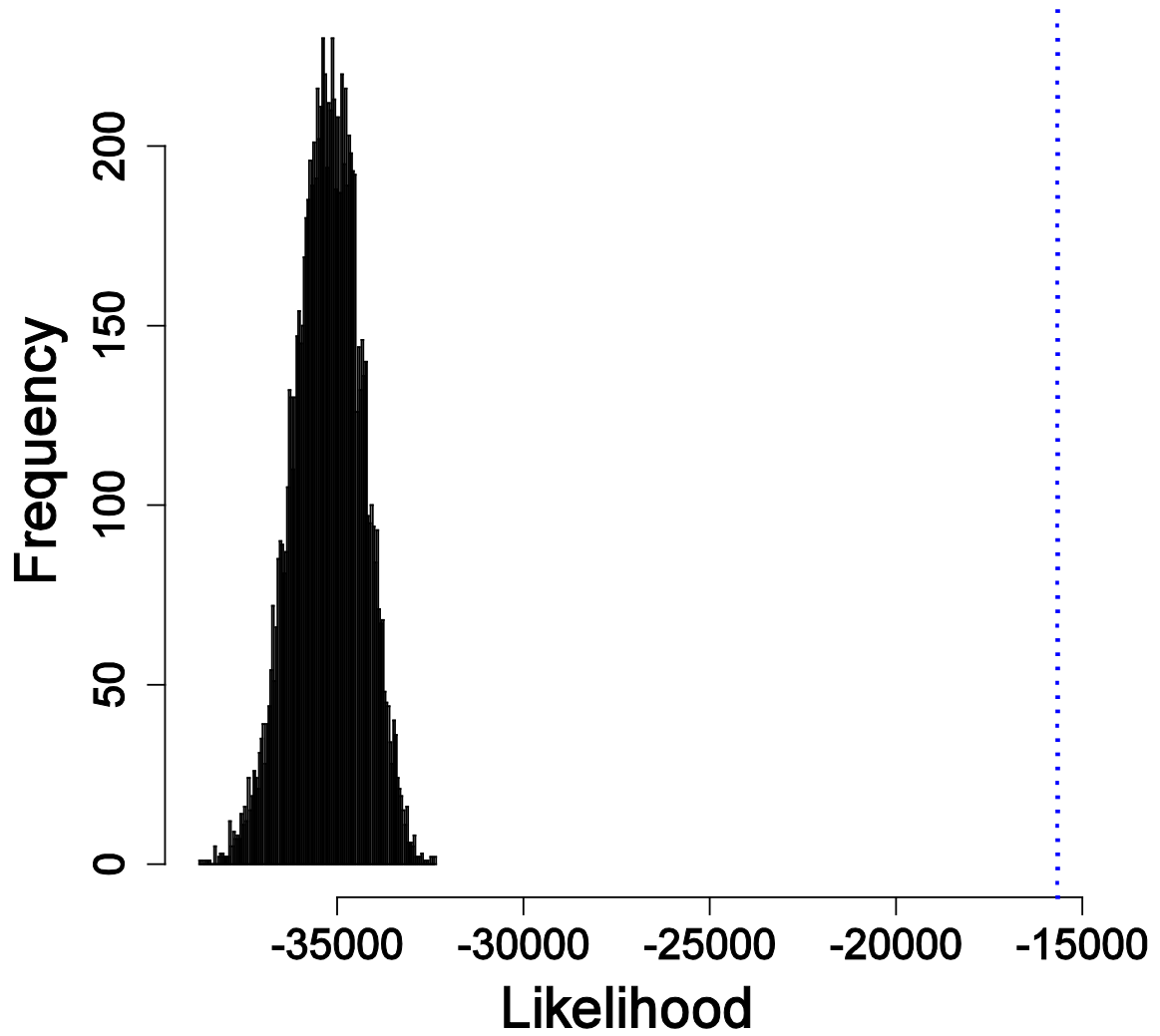


644  
 645 Figure 5 Estimated tracks of each individual (black circles) plotted over the relative habitat  
 646 utilization predicted from the GAM (see legend for values). Differently from Figure 4 here  
 647 the tracks' cells are plotted without representing the different probability values. The grey  
 648 areas correspond to land.



649  
 650 Figure 6 Distribution of the relative habitat utilisation predicted in the area covered by the  
 651 geolocated track plus a 2' buffer (grey boxplot) against the distribution of relative habitat  
 652 utilisation predicted at the track exact locations (white boxplot) for each tagged flapper skate.  
 653 Values of relative habitat utilisation at exact tracks' locations are always higher.





654  
655 Figure 7 The dashed vertical line represents the RHU-likelihood at the six observed tracks  
656 locations combined. The histogram represents the distribution of RHU-likelihoods at 10,000  
657 randomised tracks.