Journal of Applied Ecology
February 2008, Volume 45, Issue 1 : Page 204-217
http://dx.doi.org/10.1111/j.1365-2664.2007.01392.x

Modelling species distributions using regression quantiles

Sandrine Vaz[1, *], Corinne S. Martin[2], Paul D. Eastwood[3], Bruno Ernande[4, 5], Andre Carpentier[1], Geoff J. Meaden[2], Frank Coppin[1]

[1] Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER), Laboratoire Ressources Halieutiques, 150 quai Gambetta, BP699, F-62321, Boulogne/mer, France
[2] Department of Geographical and Life Sciences, Canterbury Christ Church University, Canterbury CT1 1QU, Kent, UK
[3] Centre for Environment, Fisheries, and Aquaculture Science, Lowestoft Laboratory, Lowestoft, Suffolk NR33 0HT, UK
[4] Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER), Laboratoire Ressources Halieutiques, avenue du Général de Gaulle, F-14520 Port-en-Bessin, France
[5] Evolution and Ecology Program, International Institute for Applied Systems Analysis, Schlossplatz 1, A-2361 Laxenburg, Austria

*: Corresponding author : S. Vaz, email address : Sandrine.Vaz@ifremer.fr

Abstract:

1. Species distribution modelling is an important and well-established tool for conservation planning and resource management. Modelling techniques based on central estimates of species responses to environmental factors do not always provide ecologically meaningful estimates of species-environment relationships and are increasingly being questioned.
2. Regression quantiles (RQ) can be used to model the upper bounds of species-environment relationships and thus estimate how the environment is limiting the distribution of a species. The resulting models tend to describe potential rather than actual patterns of species distributions.
3. Model selection based on null hypothesis testing and backward elimination, followed by validation procedures, are proposed here as a general approach for constructing RQ limiting effect models for multiple species.
4. This approach was successfully applied to 16 of the most abundant marine fish and cephalopods in the Eastern English Channel. Most models were successfully validated and null hypothesis testing for model selection proved effective for RQ modelling.
5. Synthesis and applications. Modelling the upper bounds of species-habitat relationships enables the detection of the effects of limiting factors on species' responses. Maps showing potential species distributions are also less likely to underestimate species responses' to the environment, and therefore have subsequent benefits for precautionary management.

Keywords: habitat, marine fish, distribution models, limiting factors, Geographical Information Systems

Introduction

Species distribution modelling is becoming an important tool for conservation planning, resource management, and understanding the effects of changing environmental conditions on biogeographical patterns (Guisan & Thuiller 2005; Austin 2007). Models are constructed from estimates of species' responses to one or more environmental attributes (Austin 2002; Oksanen & Minchin 2002). These typically comprise of habitat factors that affect the species either directly (e.g. temperature, dissolved oxygen), or indirectly (e.g. topography, latitude) (Austin 2002).

Ecologists are faced with a growing number of species distribution modelling techniques: for recent reviews see Guisan & Zimmerman (2000), Boyce *et al*. (2002), Guisan, Edwards & Hastie (2002), Guisan & Thuiller (2005), Redfern *et al*. (2006) and Austin (2007). In parallel to discussions over the suitability of different modelling techniques, there is ongoing debate surrounding approaches to model selection, as the past few years have seen a gradual shift from the more traditional use of null-hypothesis testing to information-theoretic approaches (Pearce & Ferrier 2000; Rushton, Ormerod & Kerby 2004; Stephens *et al*. 2005). Model validation is also an important issue, with strong pleas for the use of robust methods for model validation to ensure outputs maps are attributed with a measure of confidence (Guisan & Zimmerman 2000; Olden, Jackson & Peres-Neto 2002; Vaughan & Ormerod 2005).

Besides these issues, attention is also needed on the concepts underlying model design and the ecological interpretation of the different modelling approaches (Austin 2002; Austin 2007). The majority of species distribution modelling approaches in current use (e.g. GLM, GAM) are based on estimation of mean or median (central tendency) species responses to environmental factors (Oksanen & Minchin 2002). Although they provided very valuable insights, these widely used techniques, do not address some ecological aspects of species-habitat relationships (Huston 2002; Cade & Noon 2003; Eastwood & Meaden 2004; Austin 2007). One important limitation of central tendency modelling is that it does not properly estimate the limiting effects of the environment. Estimates near the upper bounds of species-habitat relationships relate to one of the central tenets of ecology theory, the law of limiting factors, which predicts that the growth rate of a species is determined by the most limiting resource (Hiddink & Kaiser 2005). When plotted, the relationship between the abundance of an organism to an environmental factor often takes the form of a bounding polygon. The upper boundary describes how abundance is limited by this factor, while the variation below the upper boundary reflects the limiting effect on the abundance of environmental attributes other than the factor of interest (Cade *et al.* 1999). In the context of habitat conservation required by ecosystem-based management, a precautionary approach would consist in considering the maximum abundance of a species that environmental factors can bring about, thus modelling potential (maximum) species abundance distribution instead of realised (mean) abundance.

While a number of techniques have been used to estimate the effects of limiting factors on species' responses (e.g. expert knowledge of species environmental tolerance limits, species-environment response curves, habitat suitability indices), only one, quantile regression (or regression quantiles (RQ)), is based on well-established statistical theory (Koenker & Bassett 1978; Koenker 2005). The statistical concepts behind RQ have been well described (for a recent review see Yu, Lu & Stander 2003), as have their general utility for estimating limiting effects (Cade, Terrell & Schroeder 1999; Cade & Noon 2003). Predictions from upper RQ models overestimate species density and distribution to illustrate the species maximum abundance given ideal environmental conditions. As such, they tend to describe potential patterns of species distribution. In recent years, RQ have increased in popularity among ecological modellers as a way of estimating a more complete range of species' responses to environmental gradients (Terrell *et al.* 1996; Cade, Terrell & Schroeder 1999; Eastwood, Meaden & Grioche 2001; Dunham, Cade & Terrell 2002; Eastwood *et al.* 2003). Modelling species distributions with RQ is more complex compared to central response modelling as a much greater range of quantile models can be estimated (1-99th). Therefore, model selection needs to be based on a range of quantiles. However, for this technique to be transferred to the field of habitat modelling and species distribution prediction in an operational way, after the selection phase, some objective criteria have to be proposed to choose a single quantile for the model's application as opposed to having to explore several RQ models on an interval of quantiles for the same species. In this context, a generic methodology for RQ-based species distribution model selection and application is needed, similar to those developed for other modelling techniques (e.g. Lehmann, Overton & Leathwick 2003).

We constructed distribution models for 16 of the most abundant fish and cephalopods in the Eastern English Channel, an area of increasing human activity and resource exploitation (Carpentier *et al.* 2005). We propose a methodology that could form part of an operational approach to RQ modelling, when the aim is to produce distribution models for several species. The aim of our model selection procedure was to select an upper quantile model able to best define limiting factors and delineate potential habitat given the environmental data available for model construction. Our intention was to

construct models that could be used to provide reliable and precautionary estimates of species' responses to their environment. We did not aim to develop or describe a generic RQ analysis of species response rates permitting the detailed study of the strength and direction of species relationship to the environment at different quantiles of the data distribution.

1. Materials and methods

1.1. fish distribution and environmental data

Marine fish and cephalopods in the Eastern English Channel are sampled annually for the purposes of developing indices of abundance for the principal commercial stocks. The UK Centre for Environment, Fisheries and Aquaculture Science (Cefas) undertakes an annual beam trawl survey (BTS) in August at a series of depth-stratified, fixed survey stations (Figure 1a). Samples are collected using 30 minute tows of a 4 m beam trawl fitted with an 80 mm diamond mesh cod-end net and a 20 mm square mesh liner. Water column depth, temperature, and salinity are recorded using sensors attached to the beam trawl, although data are not comprehensively available for all stations and years.
The Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) undertakes an annual bottom trawl survey (Channel GroundFish Survey - CGFS) in October. One or two randomly placed 30 minute hauls are taken within rectangles measuring 15' latitude and 15' longitude (Figure 1b). Sampling is with a high opening bottom trawl fitted with a 10 mm mesh size. Water column depth is recorded using sensors onboard the vessel. Since 1997, temperature and salinity (surface and bottom) have also been measured using a sensor attached to the head rope of the trawl.
For both surveys, fish counts at each trawl station were converted to catch densities based on the area swept by the gear, and expressed as numbers of fish per km^2. Data were available from 1989 (BTS) and 1988 (CGFS) to 2004. Of the total number of trawls available over this time period, only a limited number could be used for model development due to the large number of stations where environmental data were missing (Table 1).
Both surveys are designed to target different components of the fish fauna and do not catch all species with equal efficiency. Also, they operate in different seasons, thus catch densities for some species will vary as a result of ontogenic shifts in geographic distribution patterns. For species found to be well represented in both surveys and marked by different seasonal patterns, two separate models were constructed, one using BTS data and one using CGFS data (Table 2), hence a total of 25 models.
Three environmental variables - temperature, salinity, and depth - were collected during the surveys at each trawl station and were subsequently available for model construction. To increase the chance of detecting a relationship between species distributions and benthic environments, two further environmental predictors were attributed to the catch data: seabed sediment type and bed shear stress, an estimate of the pressure exerted across the seabed from tidal forcing (M_2 constituent) known for its strong relationship with patterns of species distribution (Freeman and Rogers 2003). Estimates of shear stress (in N/m^2) came from an 8 km resolution hydrodynamic model originally developed for the Irish Sea (Aldridge & Davies 1993) but extended to cover the north-west European shelf. Seabed sediment types were extracted from a digital version of the sediment map of the English Channel originally developed by Larsonneur *et al.* (1982). The original 29 sediment classes were aggregated into 5 broader classes considered to have ecological relevance to the 16 selected species, namely mud, fine sand, coarse sand, gravels, and pebbles. Sediment type was coded as dummy variables where 'mud' was the default category, i.e. the constant in the regression model, and the remaining sediment categories were 4 regression variables, coded 0 or 1 to indicate the absence or presence of the associated sediment type

1.2. regression quantiles

Regression quantiles are the linear model equivalent of one sample quantiles in that they allow a data distribution to be split into quantile classes, such as the 25th, 50th, 75th, 90th etc. One sample quantiles are extended to regression modelling through the use of an optimisation function that minimises the sum of weighted absolute deviations, where the weights are given by the specified quantile, τ, on a scale of 0 to 1 (e.g. 0.5 = 50th regression quantile). To estimate regression quantile

parameters, we used the freely available BLOSSOM statistical software programme[1] (Cade & Richards 2005).

1.3. model selection and quantile choice

Species catch densities were first $\log_{10}(y_i + 1)$) transformed to reduce heteroscedasticity in the data and limit the effect of heterogeneous error distribution models on the rank-score test statistic (Cade & Richard 2005), which was used to select significant variables. An initial exploratory analysis was performed to assess the form of the relationship between transformed species catch densities and environmental variables. This revealed that second order polynomials of continuous variables were at times necessary to describe the form of the relationships. Polynomials of higher order were not considered as they were more likely to estimate extreme catch densities at the upper and lower limits of the environmental ranges. All possible regression quantiles were estimated for a series of single variable models (e.g. catch density vs. water column depth). This helped to develop an understanding of the form of the estimated response before running more complex models containing a larger number of environmental variables.
Based on these preliminary results, we proceeded with the actual model selection by initially fitting a full, response surface model, i.e. a model including second order polynomials (main effects and their quadratics) and first order interactions between continuous environmental parameters. Sediment type was present in the full model as a categorical factor, both as a main effect and in first order interactions with the continuous environmental variables. Starting from the initial full model, we removed terms by backward elimination based on average P-values across a range of quantiles, until arriving at a model where all terms remained significant ($P < 0.05$) for at least one quantile:

1. Regression quantiles were estimated at 5 quantile intervals from the 75th to 95th. Significance tests of all polynomials and interactions were performed and the variable associated with the largest average *P*-value across the 5 quantiles, contingent on being greater than 0.05, was removed from the model.
2. Having removed one variable, reduced models were re-run across all 5 quantiles and significance tests again performed to eliminate additional variables according to the same rule. Main effects were tested only when associated interactions and polynomials had been eliminated.
3. Backwards elimination stopped when all remaining variables were significant ($P < 0.05$) at least for one quantiles. In case the resulting model was found to have all variables significant over more than one quantile, the highest of these quantiles was chosen to best represent the upper bounds of species catch density imposed by the environmental variables.

Levels of significance were evaluated within BLOSSOM using a rank-score test statistic which is appropriate for models associated with a heterogeneous error distribution (Cade, Terrell & Schroeder 1999). Weighted quantile regression may be more appropriate for non-homogenous data (Cade *et al*., 2005, Cade *et al*. 2006) but is computationally difficult to implement. Instead we used data transformation to reduce the heteroscedasticity in our data. Moreover, in doing so, we assumed that the effect of the variables are multiplicative instead of additive. In the conceptual frame of limiting factors, a multiplicative model is more relevant than an additive model. Indeed, if one factor is truly limiting a species' abundance, a multiplicative model insures the predicted abundance is low whatever the other factors' values, whereas in additive model, the predicted abundance might still be high depending on the other factors.

1.4. selected model evaluation

Stepwise variable selection has been challenged by several authors because of potential drawbacks regarding spurious variable selection (e.g. Whittingham et al. 2006). Quantile regression models are of course not immune to these potential problems. Moreover, the selection procedure, stopping as soon as a fully significant model is found for at least one quantile, may result in overfitting the model.
To assess whether the selection procedure resulted in an appropriate model, we compared the Akaike's Information Criterion (AIC) (Akaike 1974) values of our final models to a number of alternative models of varying complexity levels at the same quantile. The AIC for RQ models was calculated as

[1] www.fort.usgs.gov/products/software/Blossom/Blossom.asp

$$AIC = n \times \ln(SAF(\tau)/n) + 2p$$

where n is the number of observations and $SAF(\tau)$ is the weighted sum of absolute deviations minimised when estimating the τ th regression quantile with p parameters. AIC balances the degree of fit of a model with the number of parameters, so as to find the most parsimonious model based on these two properties. Absolute differences (dAIC) between AIC values of the least informative model (constant model used as base comparison) and the selected models were computed. Positive dAIC values indicated that the tested model was associated with a better (i.e. lower) AIC (Table 3). Largest dAIC values indicated which models achieved the best compromise between fit and complexity.

1.5. predicting species spatial distributions

Digital (raster) maps of the five environmental parameters, created in ArcGIS 8 (® ESRI), were used to predict species' spatial distributions by recoding the environmental maps using the predicted species catch densities obtained from the final RQ models (Figure 2).

1.6. model validation

Model validation was based on direct comparisons of observed vs. predicted catch densities. The first validation dataset, VALL, comprised of groundfish survey data from stations where environmental data were missing and which could therefore not be used for model development (Table 1). For CGFS' data, there was a marked temporal difference between the data used for model estimation and validation, but by using all available data we increased the chances of developing representative models. By comparison, BTS data used for model estimation and validation had a greater degree of temporal overlap. As no environmental data were available for VALL, predicted catch densities could not be generated by the models. These were instead extracted directly from the predicted species distribution maps at the VALL dataset trawl stations. The second validation dataset, V2004 (catch densities and complete set of environmental data), comprised of CGFS and BTS data from 2004. Predicted densities for V2004 were obtained by using the environmental data collected in 2004 as input to the relevant RQ models.

The bootstrap procedure allows obtaining standard errors and confidence intervals for a wide variety of statistics and this approach was used to produce a more robust validation of the models. For each validation dataset of observed and predicted densities, bootstrap datasets were generated each comprising n values (equal to n in the original dataset) by resampling with replacement within the range of observed and predicted densities. A preliminary study showed that 600 bootstrap datasets were necessary to obtain stable values for the tests means and confidence intervals. Two separate validation tests were carried out: correct classification test and rank correlation test (Eastwood et al. 2003).

Correct classification for a regression quantile model aimed at estimating limiting effects is defined by the proportion of observed values in the validation dataset that fall below those predicted. For example, if a species distribution model was developed from a 90th regression quantile, correct classification would require at least 90% of all observed values to fall below and at most 10% above those predicted. The bootstrap samples were used to provide estimates of the mean and confidence limits for the correct classification statistic for each final model. We considered a model to be successfully validated if the quantile was less than the upper confidence limit of the bootstrapped mean correct classification statistic. To assess the degree of correct classification, the difference (dCC) between the upper confidence limit of the bootstrapped correct classification statistic and the selected model quantile was calculated: increasing positive values of dCC relate to an improvement in validation success.

Because the values predicted by an upper quantile regression model are, by construction, higher than most of the observed values, Spearman's Rank Correlation Coefficient (r_s) was preferred as it does not assume a linear relationship between the variables. This correlation test was computed for all bootstrap datasets along with mean and 95% confidence intervals for r_s and associated P-values. For the test to be successful, a positive and significant correlation between observed and predicted catch densities would be expected.

In summary, Correct Classification tests were meant to assess the exactness of quantitative predictions of species abundance and Spearman Rank Correlations to assess the correctness of relative changes in predicted abundance, thus testing the spatial component of predictions.

2. Results

2.1. initial exploratory analysis

Results of the initial exploratory analysis are presented for two species with contrasting life histories: lesser spotted dogfish *Scyliorhinus canicula* L. and flounder *Platichthys flesus* L. (CGFS survey). These species were chosen because their contrasting life histories resulted in markedly different models. Lesser spotted dogfish is a bentho-demersal species known to occupy both shallow and deep waters and a range of seabed types, whereas flounder is a benthic flatfish living on sandy and muddy bottoms and inhabiting coastal waters and estuaries. Figures 3 and 4 illustrate the environmental preferences observed for both species respectively as well as the typically zero inflated and polygonal-shaped form of the relationship between species abundance and the environmental variables. To illustrate linear quantile regression relationships, three regression lines were fitted at the 75^{th}, 85^{th} and 95^{th} quantiles respectively. It is clear that central response modelling would have missed the effect of these environmental limiting factors. Triangle-like shape relationships (Figures. 3c, 4a, and 4d) suggested that a linear quantile model could be used to estimate responses near the upper bounds of the data distribution, whereas more complex relationships (Figures 3a and 4b) may require the use of second order polynomial regression. The slopes of the regressions lines at the three given quantiles (Figures 3 and 4) reflect the value of the regression coefficients at the same quantiles (Figures 5 and 6). For lesser spotted dogfish, regression coefficients were generally found to be non-zero at quantiles > 50th (Figure 5), whereas for flounder quantiles in the range 50 - 80th were in a number of cases found to be zero (Figure 6). Univariate RQ models for lesser spotted dogfish estimated catch densities to increase with depth (Figures 3a and 5a), temperature (Figures 3b and 5b), salinity (Figures 3c and 5c) and bed shear stress (Figures 3d and 5d) and over coarse sand and gravelly sediments (Figures 3e and 5e), while flounder catch densities showed the opposite pattern (Figures 4 and 6). This initial analysis of univariate responses highlighted how the value of the regression estimates may vary over the range of upper quantiles (Figures 5a-d and 6a-d) and may also switch between positive to negative throughout the range of possible quantiles (Figures 5e and 6e).

2.2. Model selection

Most final models (17 out of 25) included all 5 environmental predictors (Table 3). Seabed sediment type was found to be a significant predictor for all the species considered. Depth was significant in all but two models, water temperature in all but three models, whilst bed shear stress and salinity were significant in all but four models. Of the 25 models, 21 included at least one significant quadratic term, depth being the most common environmental factor represented as a polynomial. Of the 10 possible first order interactions tested, the number of significant interactions ranged from 0 to 7.

2.3. selected model evaluation

When comparing the difference in AIC values, the selected model (i) often constituted a good compromise between model fit and complexity (Table 3). When comparing selected models with initial full models (ii), the later yielded higher dAIC only four times out of 25. When comparing selected models with less complex alternatives (without interactions and/or polynomials, (iii to v)), selected models seemed to have as high and often much higher dAIC values. These results suggest that the chosen backward selection procedure often produces appropriate parameter selection and model selection.

2.4. predicted species response and spatial distributions

To better interpret regression coefficients in the presence of interactions and polynomial terms, we plotted predicted catch densities against each of the environmental variables, while holding all other environmental variables constant at their mean values (Figures 7 and 8). The relationships described by the model largely mirrored those estimated by the univariate models during the exploratory phase (Figure 3), namely higher catch densities in deeper waters and over coarser sediments for lesser spotted dogfish. As a result of interactions between sediment type and the three other explanatory

variables, the form of the relationship between these three and the predicted catch density differed for each sediment category. The map of predicted catch densities for lesser spotted dogfish described a relatively broad distribution across the central region of the central Eastern English Channel, corresponding to the increased depth and coarser sediment types that are characteristic of this area (Figure 9a).
The model for flounder predicted catch densities in October (CGFS data) as a function of all 5 environmental variables (Table 3). The model predicted linear relationships with the 4 continuous environmental factors and zero catch densities over gravel and pebble sediments (Figure 8). A strongly negative affinity for depth was estimated over muds and fine sands, which switched to weakly positive over coarse sands. A similar pattern was observed for the relationship with bed shear stress, temperature and also salinity, although for the latter two the response switched direction for different sediment types. The model predictions emphasize the preference shown by flounder for finer sediment types (muds and sands) in shallow waters with little tidal currents and low salinity, which generally corresponds to coastal areas and estuarine conditions. Highest catch densities for flounder (Figure 9b) were predicted in inshore areas in close proximity to the bays and estuaries found along both the English and French coasts. Both maps agreed with known distribution patterns at this time of the year (Carpentier *et al.* 2005).

2.5. model validation

Based on the VALL dataset, 15 of the 25 models successfully passed the correct classification test and all models passed the Spearman correlation test (Table 4). Species whose models performed the least well in the correct classification test, i.e. largely negative dCC, were herring *Clupea harengus* in October and flounder, both in August and October.
When using the V2004 dataset, 17 models successfully passed the correct classification test and all but one model passed the Spearman correlation test (Table 4).
Out of the 25 models, 8 passed all four tests (i.e. 2 methods x 2 validation datasets), 14 passed three tests, and 3 passed two tests, with no model passing fewer than two tests.

3. Discussion

The model selection procedure successfully arrived at models that estimated the limiting effect of the environment on fish catch densities as shown by the predicted species distribution maps. Aside from RQ modelling, Pearce & Ferrier (2000) found that the use of strict 5% significance level criteria for parameter selection arrived at models with the highest predictive power when based on generalized linear (GLM) and generalized additive modelling (GAM). We successfully extended this approach to RQ modelling and used it to select a model from a range of candidates across a number of quantiles using backward selection procedure and then choose a quantile for its application. Although stepwise selection may have some potential disadvantages (Whittingham *et al*. 2006), our 25 habitat models suggest that it could safely be used to reduce model complexity.
AIC approaches to model selection can potentially offer more flexibility than null-hypothesis testing (Stephens *et al.* 2005). In contrast to Cade *et al.* (2005) who successfully used dAICc to select appropriate variables for models at the same quantile, we computed AIC post hoc to compare our final models with alternative ones at the same quantile. Values of dAIC for the selected models were generally higher or very close to that of the equivalent RQ model containing all possible predictors . This suggests that while our selection procedure may have resulted in a slight loss of fit for some models, the degree of loss was relatively small and allowed the construction of parsimonious models. The selected model often returned the best fit compared to the alternative models we tested, with further improvements made only by increasing model complexity. Such improvements were, however, never sufficient to justify increasing the number of parameters.
For certain species, the selected models were found to differ between the two surveys, which broadly reflect conditions in summer (BTS) and autumn (CGFS). The case of lemon sole *Microstomus kitt* L. is striking since environment imposed limits were defined by eleven parameters in summer, whilst only three parameters in autumn. Spatially, the predicted distributions were also very different, highlighting how biogeographical patterns in the upper limits of catch density may vary seasonally (Carpentier *et al.* 2005).
The models performed relatively well under the two validation tests, with 22 models passing at least three validation tests. The correct classification test was the most conservative as it is based on threshold criteria, whilst Spearman's rank correlation only provided an assessment of general correlative trends. Some species models were not fully validated, possibly due to low catchability of

certain fish species by the sampling gear. The effects of gear efficiency on the predictive power of the models is expected to vary with the species considered if the abundance-environment relationship is not adequately sampled. For example, the CGFS bottom trawl is not ideally suited to catch common sole *Solea solea*, lemon sole, or flounder, and similarly for the BTS beam trawl with respect to cuttlefish.

The RQ models were developed from data collected over several years and thus represented average environment imposed limiting effects for a particular season. The validation datasets, however, represented either a single year (i.e. 2004), equating to a snapshot of the environmental conditions at that time, or an earlier time range (CGFS data only, 1988-1996) than that used for model construction (1997-2003). In the case of herring, the overall population abundance was lower during the period used for model development than during the years covered by the VALL validation dataset (notably a large abundance peak observed in 1990). This would explain why the RQ model for herring failed the classification test using VALL, in which many observed catch densities exceeded those predicted. In contrast, herring population abundance in 2004 was similar to the years used for model construction, hence an unsurprising successful validation. Discrepancies in population abundance between the periods used for model development (1997-2003) and for VALL (1988-1996) were also noted for East Atlantic red gurnard *Aspitrigla cuculus* L. and veined squid *Loligo forbesi* L. (both CGFS survey), both of which failed the classification test.

RQ has several unique advantages for species distribution modelling, which have been largely overlooked. Modelling the upper bounds of species-environment relationships enables to detect the effects of limiting factors on species' responses (Terrell *et al,* 1996, Cade, Terrell & Schroeder 1999). RQ models also yield statistical advantages as they can accommodate a relatively wide range of data distributions. Also, in contrast to GLM and GAM requiring a two step modelling procedure (Barry & Welsh 2002), RQ models prove effective in dealing with zero-inflated count data, which are common in species distribution models of abundance data.

Despite these unique advantages, RQ modelling, like most species distribution modelling techniques (e.g. GLM, GAM), does not account for spatial autocorrelation between the different environmental predictors (Legendre *et al.* 2002). Spatial autocorrelation can be caused by aggregation behaviour, competitive exclusion, density dependence (Keitt *et al.* 2002; Legendre *et al.* 2002), and species distribution patterns may differ when using spatially implicit vs. explicit approaches (Hui *et al.* 2006). However, species distribution maps constructed using geostatistical analyses (spatially explicit modelling) (Carpentier *et al.* 2005) were found to show similar spatial patterns to those constructed using our RQ models, suggesting that not accounting for spatial autocorrelation still lead to acceptable results

Our RQ models successfully estimated the limiting effects of the environment on catch densities and highlighted the importance of five environmental descriptors. These results agreed with those already reported by Vaz *et al.* (2007), where fish community structure in the Eastern English Channel was strongly shaped by its environment. Seabed sediment type, in particular, was included in all models and present in most significant interactions, illustrating its strong effect as a structuring and limiting factor.

Human pressures, such as from fishing or eutrophication, might also be used as predictors of fish abundance as these will have varied over time and will likely impact on patterns of species distributions. However, Hiddink & Kaiser (2005) pointed out that upper bound modelling is less suited to monitoring temporal variation because limiting factors, in general, remain relatively stable over time. Moreover, modelling upper bounds with RQ focuses attention on estimating the limiting effects of variables and in doing so partially accounts for variation caused by uncontrollable or unknown factors that may have affected distributions patterns. The impact and variability of human pressures on species distributions is thus implicitly taken into account in RQ upper models.

Predictions from RQ models, based on maximum instead of average species' response, especially when constructed from long time-series of occurrence data, also tend to describe potential rather than actual patterns of species distribution (Eastwood *et al.* 2003, Carpentier *et al.* 2005). Potential habitat describes areas where the environmental conditions are suitable, as opposed to realised habitat which is the region of the potential habitat where the species actually occurs and which depends on biotic parameters such as resource availability, intra and inter-specific competition or predation. These descriptions of habitat may be generalised to the fundamental and realised niche concepts being, respectively, the range of condition where a species could exist in the absence of other species, and that part of the fundamental niche to which the species is restricted due to interspecific interactions (Schoener 1989, Chase & Leibold 2003). The methodology we proposed here could be applied to biotic descriptors, such as primary production and food availability, in order to better predict species' realised niches. Final maps of species distributions are less likely to underestimate species responses to the environment, and therefore have subsequent benefits for precautionary approach of site-based and regional species habitat management principles, the mainstay of many national and international sustainable development initiatives (FAO 1996; WSSD 2002). Thus, in addition to closing the gap

between ecological theory and statistical modelling of species distributions, RQ upper bound models have unique, practical, and relevant benefits for species and habitat conservation and management.

Acknowledgements

This work was part funded by the EU under the INTERREG IIIA scheme and ERDF. P. Eastwood received additional financial support by the UK Department for Environment, Food, and Rural Affairs through contract AE0916. B. Ernande acknowledges financial support by the Conseil Régional de Haute Normandie through contract 04/1215378/MF. Thanks to Philippe Koubbi for constructive comments.

References

Akaike, H. (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control, **19**, 716–723.

Aldridge, J.N. & Davies, A.M. (1993) A high-resolution three-dimensional hydrodynamic tidal model of the Eastern Irish Sea. *Journal of Physical Oceanography*, **23**, 207-224.

Austin, M.P. (2007) Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1-19.

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

Barry, S.C. & Welsh, A.H. (2002) Generalized additive modelling and zero inflated count data. *Ecological Modelling*, **157**, 179-188.

Boyce, M.S., Vernier, P.R., Nielsen, S.E., & Schmiegelow, F.K.A. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281-300.

Cade, B.S., Richards, J.D. & Mielke, P.W. (2006). Rank score and permutation testing alternatives regression quantile estimates. *Journal of Statistical Computation and Simulation*, **76**, 331-355.

Cade, B.S., Noon, B.R. & Flather, C.H. (2005). Quantile regression reveals hidden bias and uncertainty in habitat model. *Ecology*, **86**, 786-800.

Cade, B.S. & Noon, B.R. (2003) A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, **1**, 412-420.

Cade, B.S. & Richards, J.D. (2005). User manual for BLOSSOM statistical software. Midcontinent Ecological Science Center, U. S. Geological Survey, Biological resources Discipline, Open-File Report 2005-1353, 124pp..

Cade, B.S., Terrell, J.W., & Schroeder, R.L. (1999) Estimating effects of limiting factors with regression quantiles. *Ecology*, **80**, 311-323.

Carpentier, A., Vaz, S., Martin, C.S., Coppin, F., Dauvin, J.-C., Desroy, N., Dewarumez, J.-M., Eastwood, P.D., Ernande, B., Harrop, S., Kemp, Z., Koubbi, P., Leader-Williams, N., Lefèbvre, A., Lemoine, M., Meaden, G.J., Ryan, N., & Walkey, M. (2005). Eastern Channel Habitat Atlas for Marine Resource Management (CHARM). INTERREG IIIa.

Chase, J.M. & Leibold, M.A. (2003). Ecological niches : linking classical and contemporary approaches. The University of Chicago Press.

Dunham, J.B., Cade, B.S., & Terrell, J.W. (2002) Influence of spatial and temporal variation on fish-habitat relationships defined by regression quantiles. *Transactions of the American Fisheries Society*, **131**, 86-98.

Eastwood, P.D. & Meaden, G.J. (2004). Introducing greater ecological realism to fish habitat models. *GIS/Spatial Analyses in Fishery and Aquatic Sciences (Vol. 2)* (eds T. Nishida, P.J. Kailola & C.E. Hollingworth), pp. 181-198. Fishery-Aquatic GIS Research Group, Saitama, Japan.

Eastwood, P.D., Meaden, G.J., Carpentier, A., & Rogers, S.I. (2003) Estimating limits to the spatial extent and suitability of sole (*Solea solea*) nursery grounds in the Dover Strait. *Journal of Sea Research*, **50**, 151-165.

Eastwood, P.D., Meaden, G.J., & Grioche, A. (2001) Modelling spatial variations in spawning habitat suitability for the sole *Solea solea* using regression quantiles and GIS procedures. *Marine Ecology Progress Series*, **224**, 251-266.

FAO (1996). Precautionary approach to capture fisheries and species introductions., Rep. No. No. 2. Food and Agricultural Organisation of the United Nations, Rome.

Freeman S.M. & Rogers, S. I. (2003) A new analytical approach to the characterisation of macro-epibenthic habitats: linking species to the environment. Estuarine, Coastal and Shelf Science, 56, 749-764.)

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. & Zimmerman, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186.

Hiddink, J. G. & Kaiser, M. J. (2005) Implications of Liebig's law of the minimum for the use of ecological indicators based on abundance. *Ecography*, **28**, 264-271.

Huston, M.A. (2002). Critical issues for improving predictions. *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall & F.B. Samson), pp. 7-21. Island Press, Washington.

Hui, C., McGeoch, M. A., Warren, M. (2006) A spatially explicit approach to estimating species occupancy and spatial correlation. *Journal of Animal Ecology*, **75**, 140-147.

Keitt, T.H., Bjornstad, O.N., Dixon, P.M., & Citron-Pousty, S. (2002) Accounting for spatial pattern when modelling organism-environment interactions. *Ecography*, **25**, 616-625.

Koenker, R. & Bassett, G. (1978) Regression quantiles. *Econometrica*, **46**, 33-50.

Koenker, R. (2005) Quantile regression. Economic Society Monographs. Cambridge University Press.
Larsonneur, C., Bouysse, P., & Auffret, J.-P. (1982) The superficial sediments of the English Channel and its western approaches. *Sedimentology*, **29**, 851-864.

Legendre, P., Dale, M.R.T., Fortin, M.-J., Gurevitch, J., Hohn, M., & Myers, D. (2002) The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, **25**, 601-615.

Lehmann, A., J. M. Overton, Leathwick, J. R. (2003). GRASP: generalized regression analysis for spatial prediction. *Ecological Modelling*, **160**, 165-183

Oksanen, J. & Minchin, P.R. (2002) Continuum theory revisited: what shape are species responses along ecological gradients? *Ecological Modelling*, **157**, 119-129.

Olden, J.D., Jackson, D.A., & Peres-Neto, P.R. (2002) Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329-336.

Pearce, J. & Ferrier, S. (2000) An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, **128**, 127-147.

Redfern, J.V., Ferguson, M.C., Becker, E.A., Hyrenbach, K.D., Good, C., Barlow, J., Kaschner, K., Baumgartner, M.F., Forney, K.A., Ballance, L.T., Fauchald, P., Halpin, P., Hamazaki, T., Pershing, A.J., Qian, S.S., Read, A., Reilly, S.B., Torres, L., Werner, F. (2006) Techniques for cetacean–habitat modelling, Marine Ecology Progress Series, **310**, 271–295.

Rushton, S.P., Ormerod, S.J., & Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41**, 193-200.

Schoener, T.W. (1989). The ecological niche. In Ecological concepts, ed. Cherrett, J.M., 79-113. Oxford: Blackwell.

Stephens P.A., Buskirk S.W., Hayward G.D., Martinez Del Rio C. (2005) Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology*, **42** , 4-12.

Terrell, J.W., Cade, B.S., Carpenter, J., & Thompson, J.M. (1996) Modelling stream fish habitat limitations from wedge-shaped patterns of variation in standing stock. *Transactions of the American Fisheries Society*, **125**, 104-117.

Vaughan, I.P. & Ormerod, S.J. (2005) The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, **42**, 720-730.

Vaz, S., Carpentier, A., & Coppin, F. (2007). Eastern English Channel fish assemblages: measuring the structuring effect of habitats on distinct sub-communities. – ICES Journal of Marine Science, **64**, 271–287.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B., Freckleton, R.P.(2006). Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology, **75**, 1182-1189.

WSSD (2002). Plan of Implementation of the World Summit on Sustainable Development. Division of Sustainable Development, UN Department of Economic and Social Affairs., New York.

Yu, K., Lu, Z., & Stander, J. (2003) Quantile regression: applications and current research areas. *The Statistician*, **52**, 331-350.

Tables

Table 1. Number of observations available for model estimation and validation. BTS = Beam Trawl Survey, CGFS = Channel Groundfish Survey

	n for model estimation		*n* for model validation	
YEAR	BTS	CGFS	BTS	CGFS
1988				68
1989	36		28	61
1990	57		11	69
1991			86	74
1992			79	54
1993	66		7	58
1994	71		3	80
1995			77	81
1996			78	64
1997		100	70	
1998	63	76	11	
1999	66	95	7	
2000	75	93		
2001	82	102		
2002	71	67	3	23
2003		90		
2004			71	86
TOTAL	587	623	531	718

Table 2. Marine fish species selected for model development. See Table 1 for details.

Code	Latin name	Common name	Survey
CHELCUC	*Aspitrigla (Chelidonichthys) cuculus*	East Atlantic red gurnard	BTS, CGFS
CLUPHAR	*Clupea harengus*	Atlantic herring	CGFS
GADUMOR	*Gadus morhua*	Atlantic cod	CGFS
LIMDLIM	*Limanda limanda*	dab	BTS, CGFS
LOLIFOR	*Loligo forbesi*	veined squid	CGFS
LOLIVUL	*Loligo vulgaris*	European squid	CGFS
MERNMER	*Merlangius merlangus*	whiting	CGFS
MICTKIT	*Microstomus kitt*	lemon sole	BTS, CGFS
MULLSUR	*Mullus surmuletus*	red mullet	CGFS
PLATFLE	*Platichthys flesus*	flounder	BTS, CGFS
PLEUPLA	*Pleuronectes platessa*	plaice	BTS, CGFS
RAJACLA	*Raja clavata*	thornback ray	BTS, CGFS
SCYOCAN	*Scyliorhinus canicula*	lesser-spotted dogfish	BTS, CGFS
SEPIOFF	*Sepia officinalis*	common cuttlefish	BTS, CGFS
SOLESOL	*Solea solea*	common sole	BTS, CGFS
SPONCAN	*Spondyliosoma cantharus*	black seabream	CGFS

Table 3. Selected models and AIC compared to alternative models. # : model estimated from BTS data, otherwise CGFS data. In 'Fitted parameters', Dep = depth, Str = seabed stress, Temp = temperature, Sal = salinity, Sed = sediment type, 2 alongside the variable indicates the use of 2nd order polynomial. dAIC is the absolute difference between the null model and selected or alternative model AIC. The models tested were: (i) selected model; (ii) full models containing all terms; (iii) selected models minus any quadratic terms; (iv) selected models minus any interactions; and (v) selected models with significant main effects only. Highest positive values are indicated in bold.

Species	*Quantile*	*Fitted parameters*	*No. of interactions*	*dAIC*				
				(i)	*(ii)*	*(iii)*	*(iv)*	*(v)*
CHELCUC#	*95*	*Dep2+Str+Temp+Sal+Sed*	*2*	***163***	*140*	*142*	*145*	*119*
CHELCUC	*85*	*Dep+Str+Temp+Sal2+Sed*	*7*	***241***	*234*	*225*	*126*	*121*
CLUPHAR	*75*	*Dep2+Str+Temp+Sal+Sed*	*4*	***228***	*210*	***230***	*128*	*80*
GADUMOR	*80*	*Dep+Str+Temp+Sal+Sed*	*3*	***59***	*41*	***59***	*41*	*41*
LIMDLIM#	*75*	*Dep+Str2+Temp+Sal+Sed*	*4*	*333*	***343***	*277*	*270*	*223*
LIMDLIM	*85*	*Dep2+Str2+Temp2+Sal+Sed*	*5*	***230***	*205*	***232***	*228*	*225*
LOLIFOR	*85*	*Dep+Str+Temp2+Sal+Sed*	*2*	***97***	*84*	*82*	*70*	*54*
LOLIVUL	*90*	*Dep+Temp+Sed*	*2*	***40***	*23*	***40***	*25*	*25*
MERNMER	*90*	*Dep2+Str2+Temp2+Sal+Sed*	*4*	***247***	*242*	*187*	*203*	*167*
MICTKIT#	*80*	*Dep+Str+Temp+Sal2+Sed*	*5*	*187*	***214***	*96*	*187*	*96*
MICTKIT	*85*	*Temp2+Sed*	*0*	***196***	***197***	***195***	*134*	*133*
MULLSUR	*90*	*Dep+Str+Temp+Sal2+Sed*	*3*	***93***	*89*	*86*	*66*	*67*
PLATFLE#	*80*	*Str2+Temp+Sal+Sed*	*3*	*236*	***245***	*238*	*172*	*172*
PLATFLE	*90*	*Dep+Str+Temp+Sal+Sed*	*5*	***421***	*405*	***421***	*345*	*345*
PLEUPLA#	*80*	*Dep2+Str2+Temp+Sal+Sed*	*4*	***181***	*169*	*167*	*121*	*114*
PLEUPLA	*90*	*Dep2+Str+Temp2+Sed*	*3*	***282***	***281***	*212*	*248*	*193*
RAJACLA#	*90*	*Dep2+Sed*	*1*	***73***	*48*	*69*	*49*	*50*
RAJACLA	*90*	*Dep+Str+Sal+Sed*	*2*	***41***	*26*	***41***	*16*	*16*
SCYOCAN#	*85*	*Dep+Str+Temp2+Sal+Sed*	*3*	***124***	*109*	*110*	*85*	*80*
SCYOCAN	*80*	*Dep2+Temp+Sal2+Sed*	*3*	***174***	*164*	*169*	*128*	*130*
SEPIOFF#	*80*	*Dep2+Str+Temp+Sal+Sed*	*2*	***43***	*32*	*27*	*29*	*18*
SEPIOFF	*90*	*Dep+Str+Temp2+Sal2+Sed*	*7*	***69***	*46*	*64*	*65*	*61*
SOLESOL#	*75*	*Dep+Str2+Temp+Sal+Sed*	*3*	*233*	***246***	*195*	*200*	*186*
SOLESOL	*85*	*Dep2+Str2+Sal+Sed*	*2*	***188***	*172*	***186***	*158*	*159*
SPONCAN	*90*	*Dep+Str+Temp2+Sal+Sed*	*3*	***67***	*53*	***66***	*15*	*16*

Table 4. Spatial distribution model validation results. # : model built using BTS data, otherwise CGFS data. dCC is the difference between the upper confidence limit of the bootstrapped correct classification rate and the selected model quantile. r_s is the Spearman's rank correlation coefficient and associated significance where ns = not significant (P >= 0.05), ** P <0.01, *** P <0.001.

		VALL			V2004			Total no. of
SPECIES	Quantile	dCC	r_s	P	dCC	r_s	P	tests passed
CHELCUC#	95	1.1	0.72	***	3.6	0.70	***	4
CHELCUC	85	-9.4	0.65	***	4.5	0.61	***	3
CLUPHAR	75	-53.5	0.46	***	20.4	0.25	**	3
GADUMOR	80	12.0	0.25	***	15.4	0.10	ns	3
LIMDLIM#	75	5.9	0.64	***	-3.2	0.49	***	3
LIMDLIM	85	0.7	0.75	***	-0.1	0.75	***	3
LOLIFOR	85	-6.1	0.43	***	10.2	0.60	***	3
LOLIVUL	90	8.7	0.28	***	7.7	0.43	***	4
MERNMER	90	2.3	0.66	***	5.4	0.58	***	4
MICTKIT#	80	-0.5	0.37	***	11.6	0.39	***	3
MICTKIT	85	-3.8	0.52	***	-0.1	0.31	**	2
MULLSUR	90	4.8	0.36	***	-9.8	0.47	***	3
PLATFLE#	80	-40.7	0.41	***	11.6	0.42	***	3
PLATFLE	90	-23.3	0.38	***	-0.5	0.44	***	2
PLEUPLA#	80	0.1	0.52	**	-3.9	0.55	***	3
PLEUPLA	90	1.4	0.70	***	-2.8	0.74	***	3
RAJACLA#	90	0.0	0.30	***	4.4	0.43	***	3
RAJACLA	90	-0.6	0.26	***	6.5	0.36	***	3
SCYOCAN#	85	4.7	0.59	***	0.2	0.46	***	4
SCYOCAN	80	7.7	0.59	***	9.5	0.71	***	4
SEPIOFF#	80	-9.4	0.56	***	-18.0	0.61	***	2
SEPIOFF	90	6.9	0.33	***	6.5	0.32	***	4
SOLESOL#	75	-0.2	0.63	***	53	0.70	***	3
SOLESOL	85	9.9	0.36	***	12.7	0.31	**	4
SPONCAN	90	2.7	0.38	***	1.9	0.35	***	4

Figures

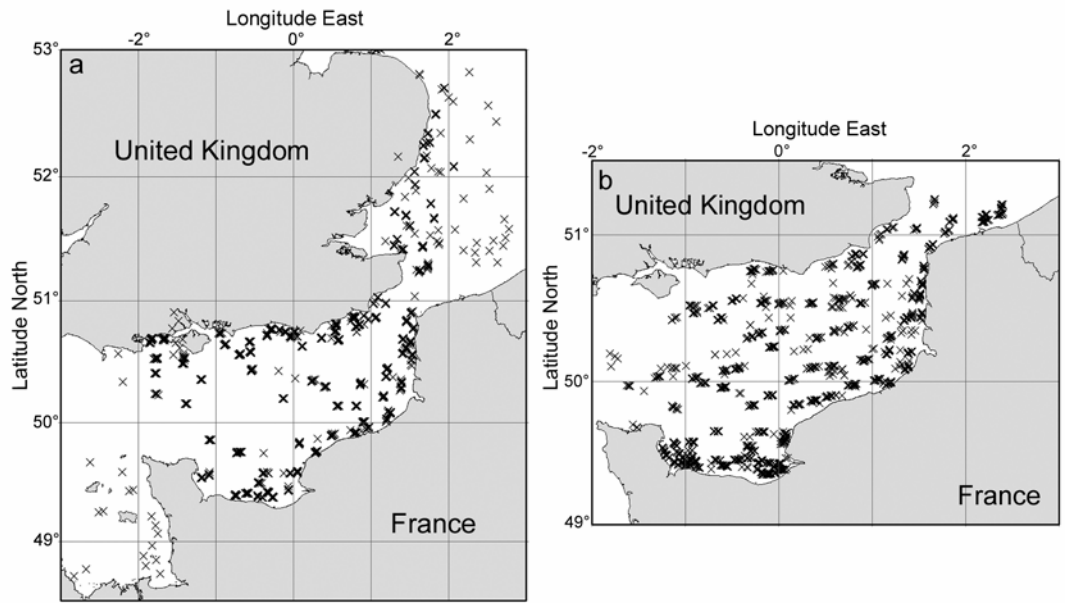

Figure 1. Station positions for the (a) BTS (1989-2004) and (b) CGFS (1988-2004) surveys.

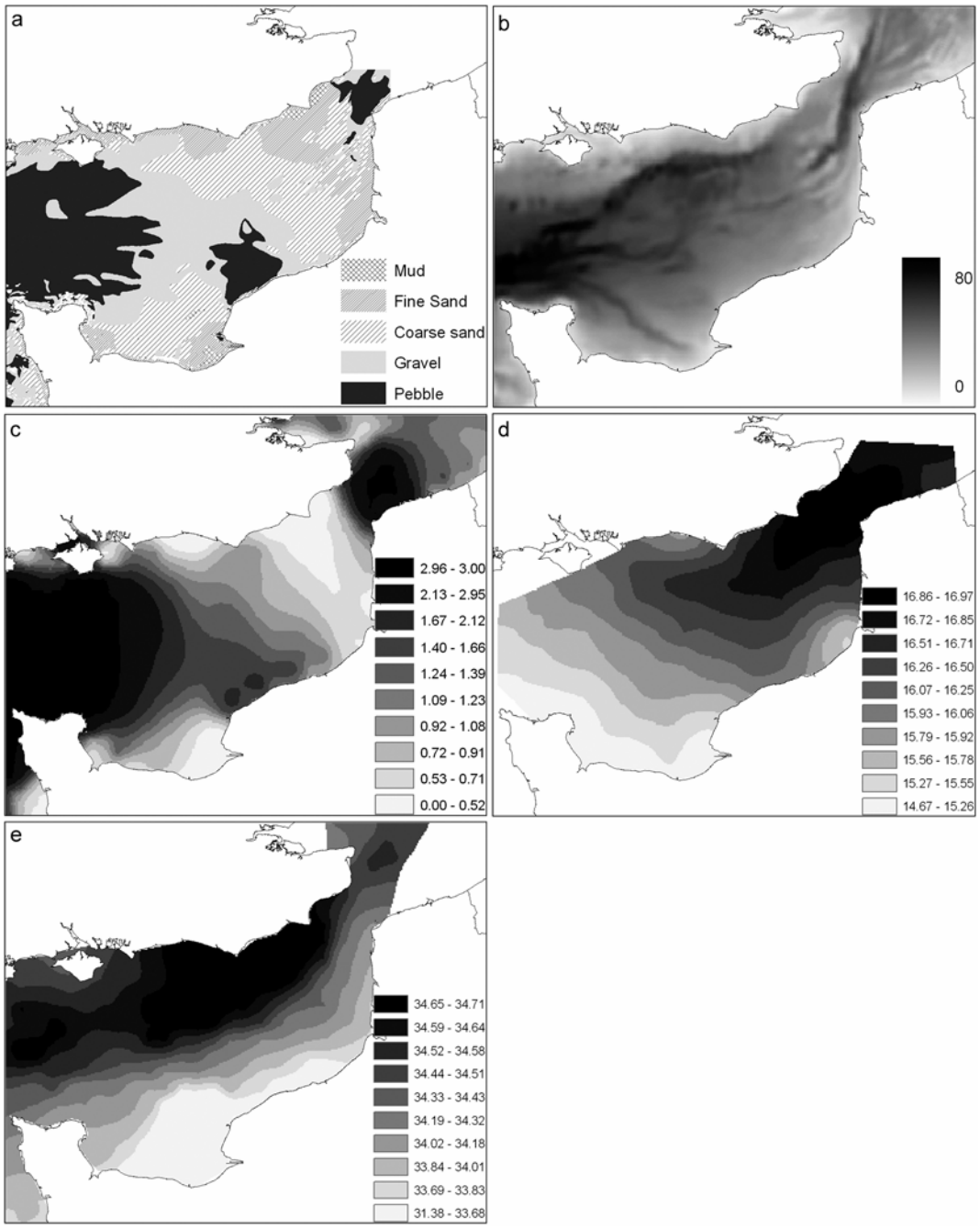

Figure 2. Digital environmental layers used to generate spatial predictions of fish catch densities from the selected RQ models: (a) seabed sediment types; (b) depth plus mean sea level (m); (c) bed shear stress in ($N.m^{-2}$); (d) mean sea surface temperature (°C, CGFS); (e) mean sea surface salinity (BTS).

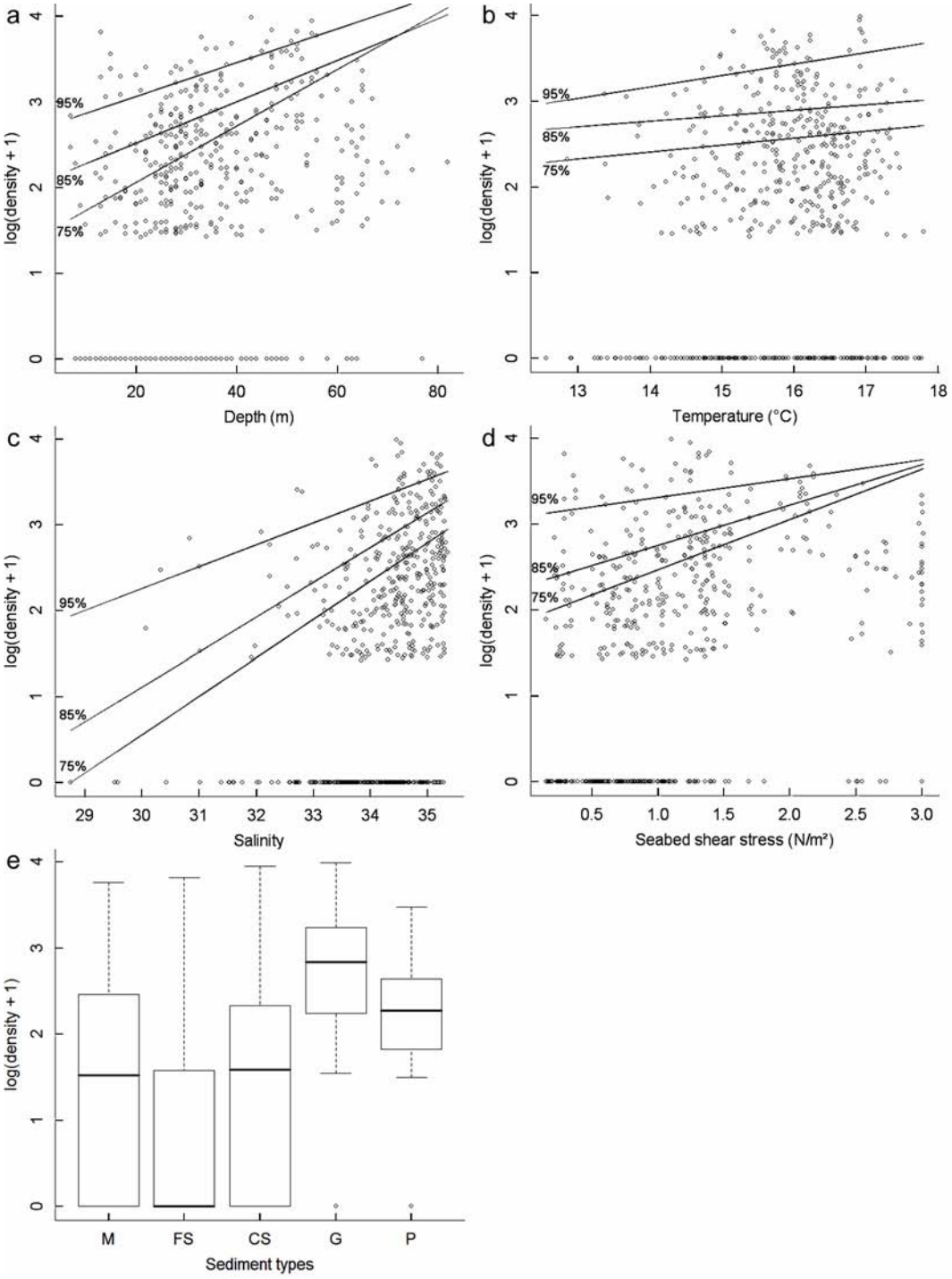

Figure 3. Observed catch densities (CGFS) for lesser spotted dogfish as a function of environmental variables: (a) depth, (b) temperature, (c) salinity and (d) seabed shear stress. Each plot illustrates the species response along one given environmental gradient along with regression lines for quantiles 75, 85, 95th. (e) The species' response associated to each sediment type: mud (M), fine sand (FS), coarse sand (CS), gravel (G), pebbles (P).

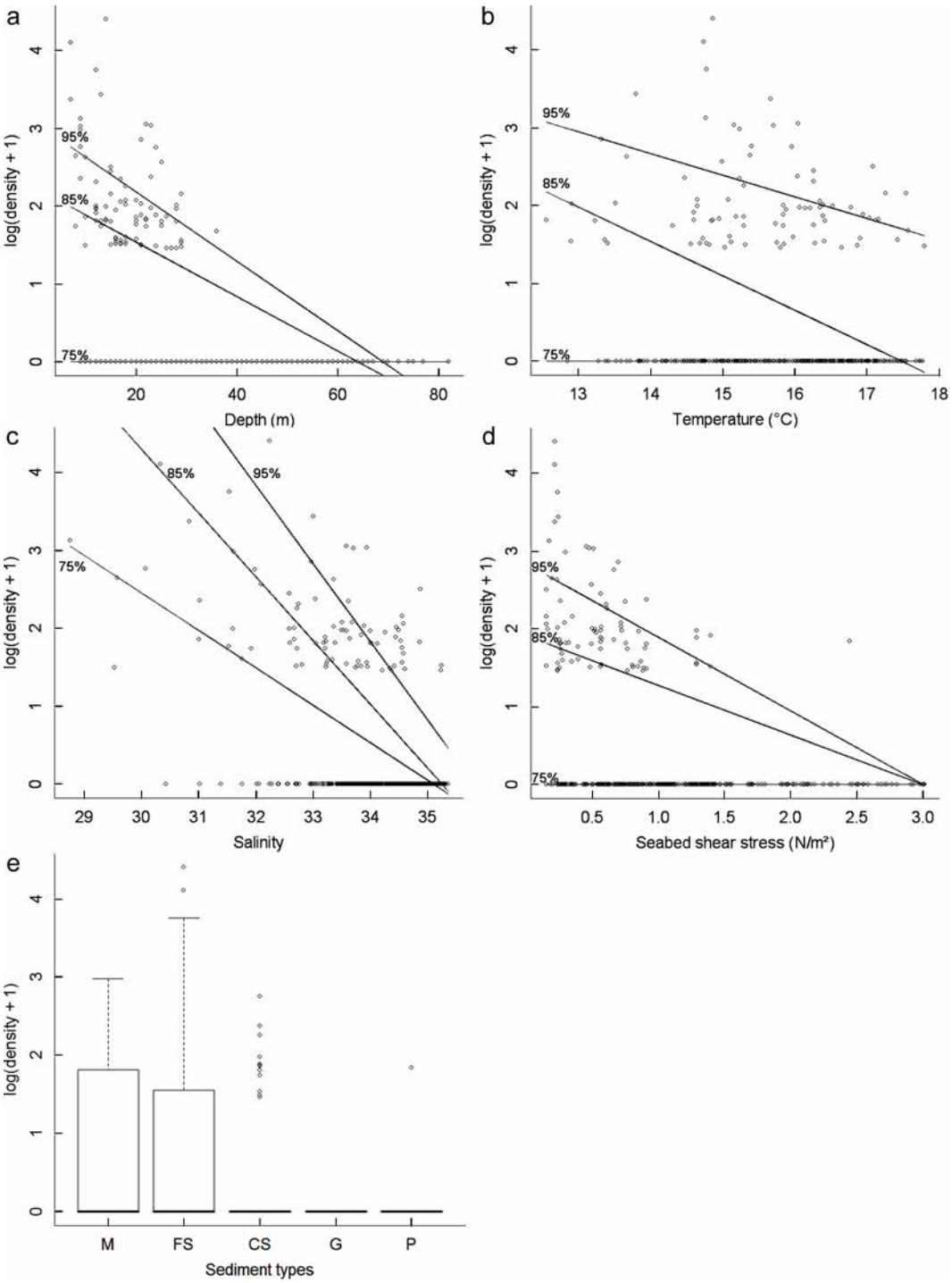

Figure 4. Observed catch densities (CGFS) for flounder as a function of environmental variables. See Figure 3 for details.

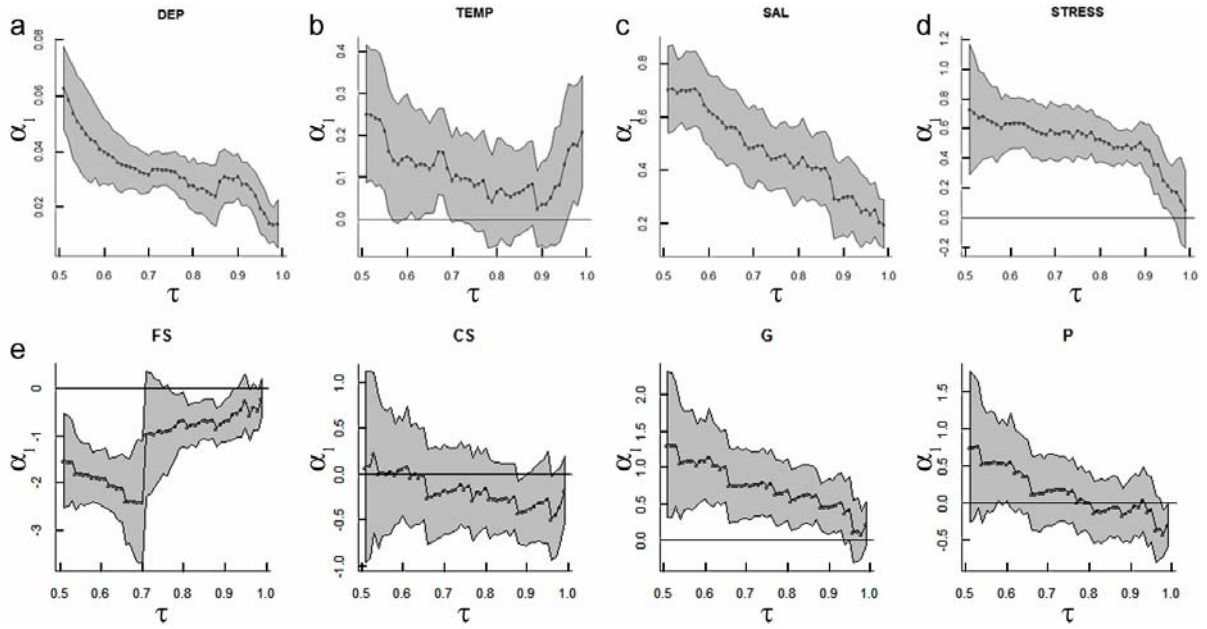

Figure 5. Regression coefficients (α_1) for quantiles (τ) over the range 0.50 - 0.99 for linear univariate models ($Log_{10}(y_i +1) = \alpha_o + \alpha_1 x + \varepsilon$) of lesser spotted dogfish catch densities (CGFS) according to (a) depth, (b) temperature, (c) salinity, (d) seabed shear stress and (e) sediment types: fine sand (FS), coarse sand (CS), gravel (G) and pebbles (P).

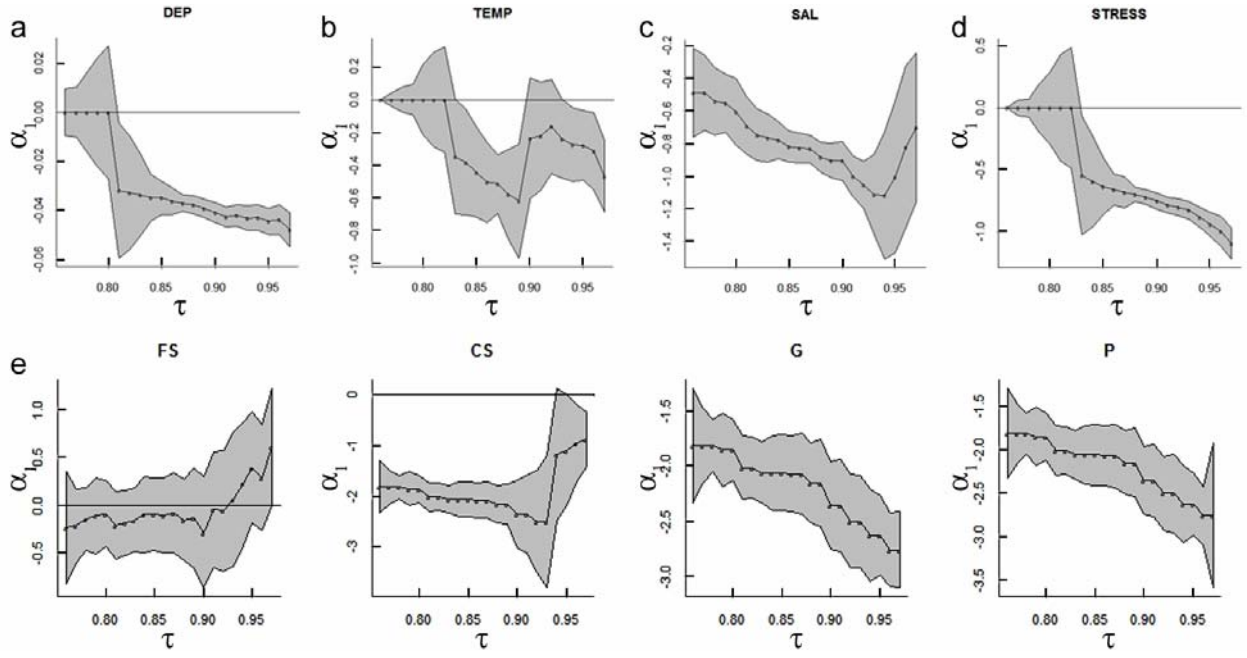

Figure 6. Regression coefficients (α_1) for quantiles (τ) over the range 0.75 - 0.99 for linear univariate models ($Log_{10}(y_i +1) = \alpha_o + \alpha_1 x + \varepsilon$) of flounder catch densities (CGFS). See Figure 5 for details.

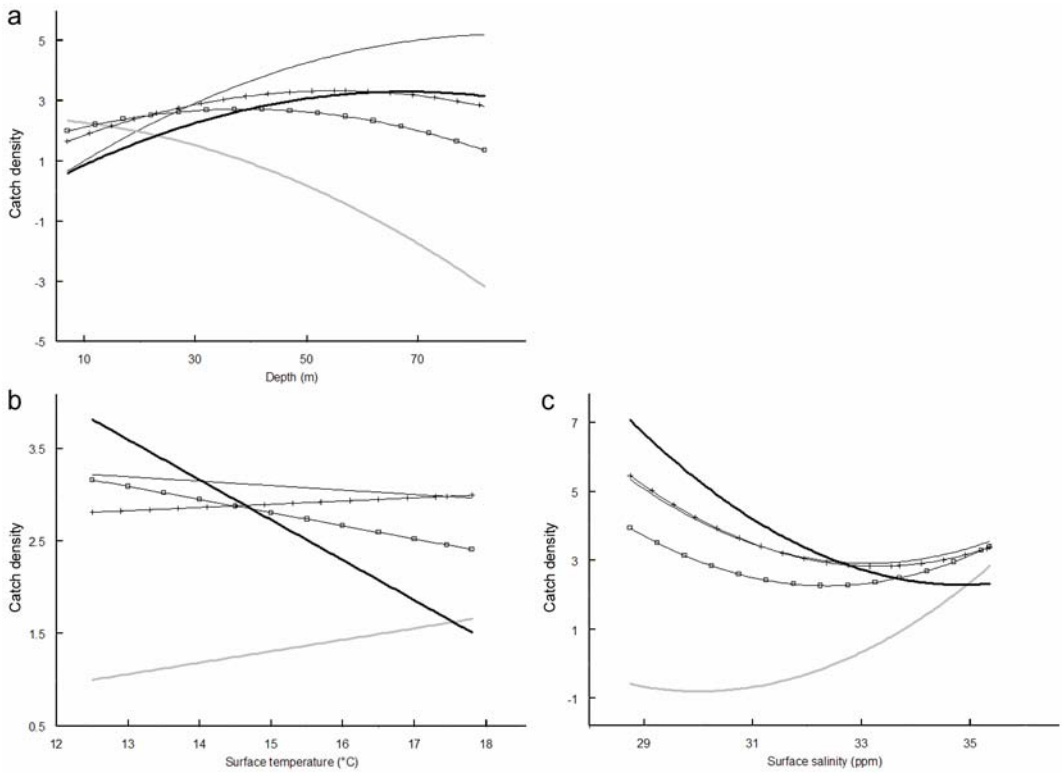

Figure 7. Predicted catch densities for lesser spotted dogfish (CGFS) as a function of the significant explanatory variables: (a) depth, (b) temperature, and (c) salinity. Each plot illustrates the species' response along one given environmental gradient, all other variables remaining constant at their mean value. The effect of each sediment type on the species' response is given by five lines or curves : —— mud, ——— fine sand, ▬▬ coarse sand, —ı— gravel, —□— pebbles.

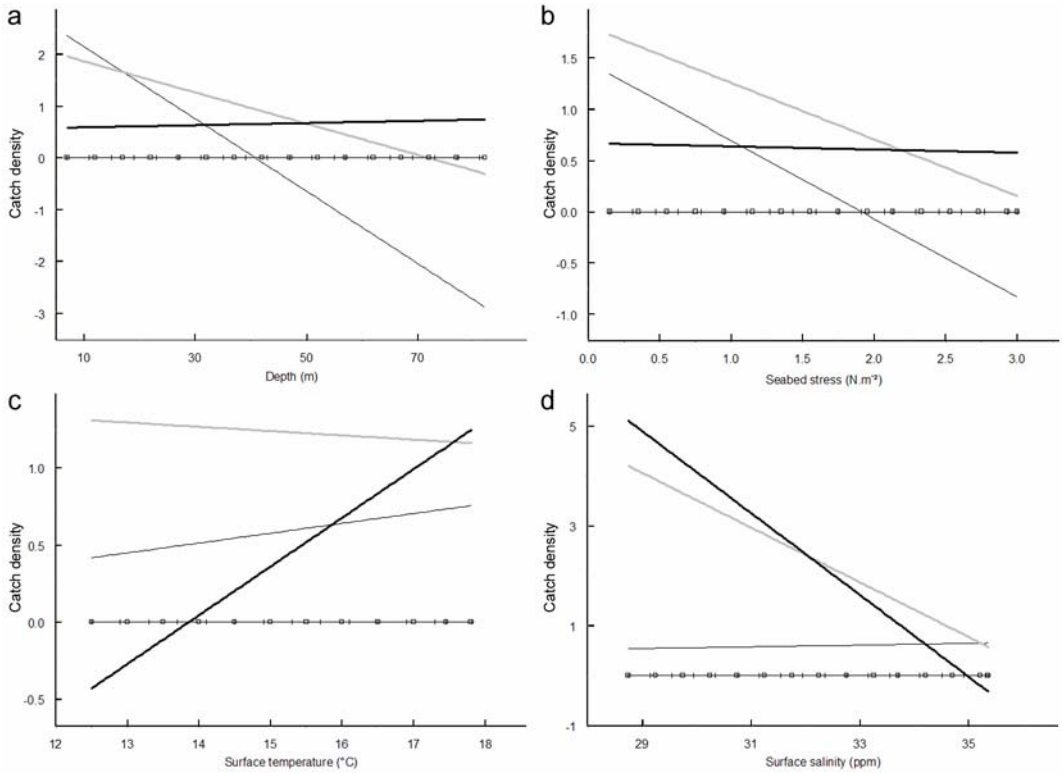

Figure 8. Predicted catch densities for flounder (CGFS) as a function of the significant explanatory variables. See Figure 7 for details.

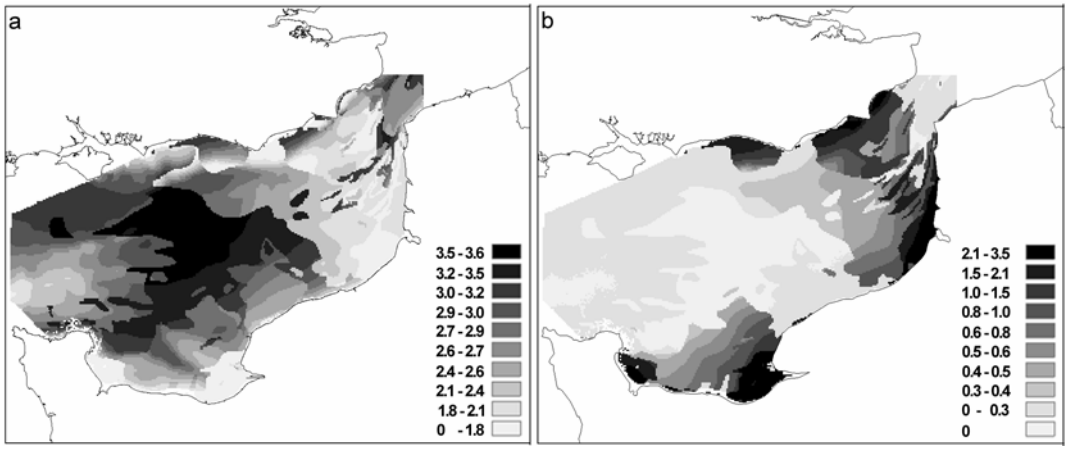

Figure 9. Predicted catch densities (number of fish per km^2) for (a) lesser spotted dogfish and (b) flounder in October (CGFS).