	**GREEN MATRIX UPLOADED: A NEW ECOSYSTEM VARIABLE FOR MARINE RESOURCES SECTOR**
GREENUP use case:	**Distribution of Mackerel**
Deliverable:	Final report
Date:	20 December2017
GREENUP partner:	DTU Aqua

The objective of GREENUP is to extend the CMEMS product catalogue by developing a new product covering a key ecosystem component at the mid-trophic level (MTL), i.e., the micronekton, to better address the Marine Resources sector. Two applications benefiting from the CMEMS products and the new proposed MTL variables will illustrate the interest of these new developments for the research, management and monitoring of marine resources. This report describes the use case developed by DTU-AQUA for the spatial distribution of Mackerel.

Abstract

Northeast Atlantic Mackerel is a highly migratory species that supports a commercially important fishery. Changes in the spawning distribution of this species create problems for the scientific monitoring of this stock and also for the performance of the fishery: however, critical data about factors driving these changes, such as the distribution and abundance of their food, is currently lacking. This work aims to assess the ability of the GREENUP Mid-Trophic Level (MTL) products to fill this gap in understanding. ***We developed species distribution models describing the relationship between the environment and the distribution of Mackerel egg production, based on scientific egg surveys in this region, and then examined the ability of the GREENUP MTL products to improve these models.*** We applied two species distribution modelling (SDM) approaches, a Generalised Additive Model (GAM) and a Random Forest (RF). Whilst it was possible to characterise the spatial-temporal correlation structure of the observations using the GAM model, technical difficulties with model stability and convergence prevented this model being taken through to a full SDM. The RF SDM approach, however, proved fruitful and produced models of good predictive skill. ***Those models that incorporated the GREENUP MTL products were shown to have appreciably better skill than baseline models.*** We therefore conclude that the GREENUP MTL products show important potential for characterising and predicting the spatial distribution and spawning intensity of this fish stock.

Background

The fishery for northeast Atlantic Mackerel (*Scomber scombrus* L.) is one of the largest and most valuable in Europe – annual landings of this species in recent years have exceeded 1.2 million tonnes with a market value of close to € 400 million. The species is widely distributed throughout European waters of the NE Atlantic, ranging from Gibraltar to the Norwegian Sea, and in recent years, individuals have been reported as far north as Svalbard (> 75°N) (Berge *et al.*, 2015). However, the distribution is highly dynamic and large interannual variations in distribution have been seen, giving rise to international conflicts (Hannesson, 2012), and making monitoring and managing the stock challenging. Understanding the processes that drive this distribution and its variability is key to maximising the economic potential of the resource, avoiding conflicts and ensuring its future sustainability.

The distribution of mackerel is characterised by large, long distance migrations of several thousand kilometres. Spawning typically takes place in the southern part of the range during spring, along the continental shelf edge from Gibraltar and Portugal, through the Bay of Biscay and to the waters west of the British Isles, with limited amounts of spawning in the North Sea and to the south of Iceland. The main centres of spawning however, are on the Cantabrian coast (northern Spain) and to the west of Ireland: an unresolved question in the biology of this species is whether these centres are biologically independent of each other. After spawning, mackerel migrate northwards towards their feeding grounds in the productive Norwegian seas, ranging from the Norwegian coast in the east to Iceland and Greenland in the west. Overwintering takes places in the southern part of the Norwegian Sea and northern part of the North Sea, before returning to the south to spawn again.

Shifts in this migration pattern, particularly during the feeding season, have received much focus in recent years. In 2007, mackerel appeared unexpectedly during summer in waters to the east of Iceland and in subsequent years expanded westwards to reach the east coast of Greenland (Astthorsson *et al.*, 2012). This substantial expansion in distribution and ingress into Icelandic waters lead to a breakdown in the arrangements for the management of this stock (Hannesson, 2012) and subsequent international conflicts over access rights. While the mechanisms underpinning these distributional shifts remain unclear, various hypotheses have been proposed, including the effect of climate change / variability (ICES, 2013), declines of nutrient availability (Pacariz *et al.*, 2016), and density dependent processes (van der Kooij *et al.*, 2015).

In the southern part of the range, important shifts in the spawning distribution have also been noted. Mackerel in this region are surveyed triennially by a scientific egg-survey focused on monitoring the biomass of spawning adults: these abundance estimates are incorporated directly into the stock assessment and management of the stock. However, substantial interannual differences between years in the spatial distribution and timing of spawning have been observed by this survey (ICES, 2013). Analysis of these shifts (Bruge *et al.*, 2016) suggests a relationship to temperature variations, although the mechanisms remain unclear.

Central to both of these distributional shifts is the role of food. Mackerel lack a swim bladder (the organ that most fish possess to regulate their buoyancy) and therefore need to maintain a high activity level to maintain their position in the water column. These high activity levels have an energetic cost associated with them, and mackerel therefore are known as voracious feeders. Furthermore, their high activity rate and strong swimming ability means that they can both seek out new feeding grounds and respond to variations in local productivity with relative ease. This combination of a strong need for food together with the ability to respond readily to the local environment means that their distribution is thought to follow local productivity conditions closely.

However, appropriate characterisation of the ideal habitat for mackerel, as for any fish species, is challenging. Traditional approaches have been hampered by the availability of appropriate environmental data: while temperature fields are readily available there are few other alternatives. Primary production derived from satellite-based chlorophyll measurements can be used as a proxy for food availability, but this metric is still several steps removed from the availability of food for mackerel. However, the estimates of micronekton abundance produced in GREENUP are potentially of direct relevance, as they are representative of the primary food source for this species. Here we investigate the appropriateness of the GREENUP MTL product for understanding, modelling and potentially forecasting the distribution, and distributional shifts, of mackerel in the North East Atlantic. We address this question by applying powerful state-of-the-art statistical modelling and machine learning techniques to model observations of the mackerel spawning activity: we then assess the changes in the quality of these models resulting from incorporating the GREENUP MTL products. This approach allows us to assess both the relevance of the products and their importance relative to other explanatory variables, and thereby the value of such products for describing changes in the distribution of Mackerel.

Description of R&D activities

Modelling Approach

Preparation of Observation Data

The distribution and abundance of spawning Mackerel along the European continental shelf edge is monitored by the Mackerel Egg Survey (MEGS). This survey has been performed every third year since 1977, and the methodology and coverage have been largely constant since 1992. The survey is a large undertaking that has involved vessels from 11 different nations, and typically involves around 2000 hauls per survey, covering the time period from January to as late as August, and a spatial domain from Gibraltar to Iceland. The sampling gear involved varies by nation and vessel, and has also changed over time: the majority of the hauls have been performed by horizontally-towed Gulf VII samplers (particularly in the most recent years), with 40cm and 60cm Bongo nets also common. Sampling covers the upper 200m of the water column (or to within 5m of the bottom) and volumes filtered are typically around 100-200m^3 of water. Icthyoplankton collected by these samples is then sorted by species and mackerel eggs are assigned to one of four developmental stages.

Data reported by the survey was first quality controlled and a number of cross checks performed to ensure the appropriateness of the data and to prepare it for subsequent spatial modelling. Missing values in the database were identified and checked with the survey coordinators. The temporal and spatial location of the hauls were checked for consistency and reasonableness, and to ensure that no points are present on land, out of the survey area or period. Reported sampling depths were cross-checked against bathymetric datasets to ensure reasonableness. The volume of water filtered, as reported by flowmeters, was cross-checked against the characteristics of the sampling gear and haul to ensure both proper functioning of the equipment and that the haul was performed in agreement with survey protocols (particularly with respect to trawl speed): erroneous hauls were removed from the analysis. An example of the spatial and temporal distribution of samples is shown from the 2013 survey (Figure 1).

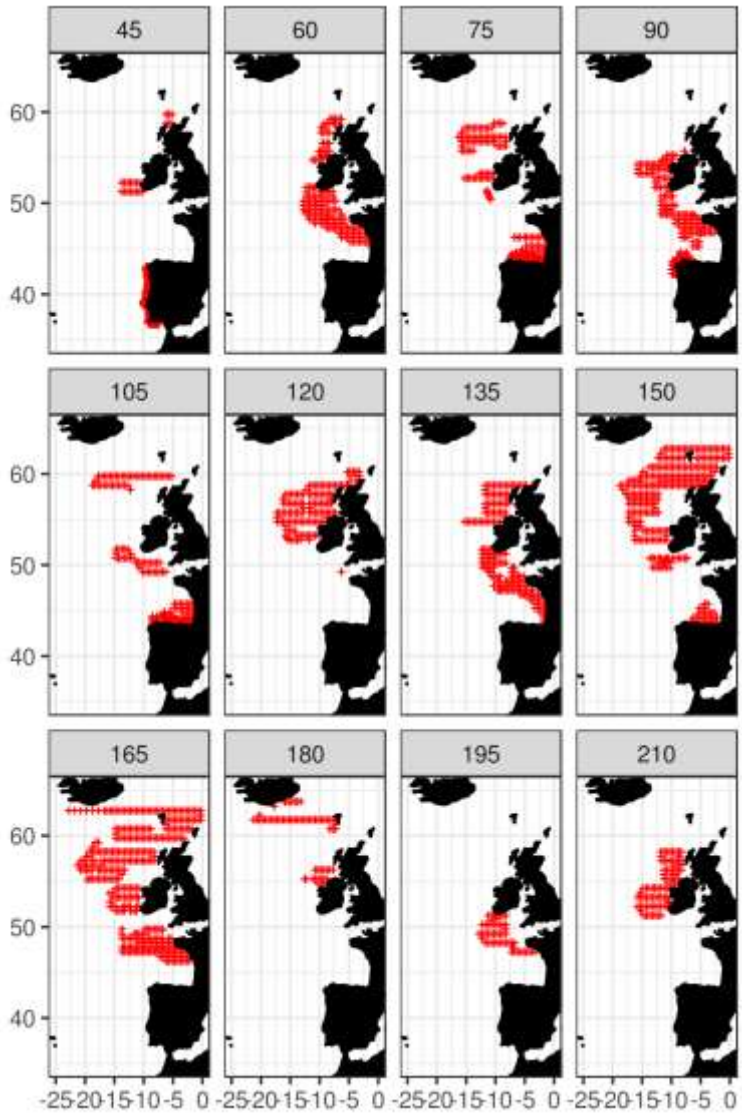

Figure 1 Example of the distribution of haul data from the 2013 ICES Mackerel Egg Survey (MEGS). Red points correspond to the spatial location of a haul. Individual panels correspond to a 15 day time-window, centered on the day of year indicated at the top of the panel (e.g. the "45" panel shows hauls taken between day-of-year 37.5 and 52.5)

Models Employed

The spawning intensity, and thereby distribution, of mackerel was modelled as a function of the environmental conditions using empirical species distribution models (SDMs), also known as environmental niche models (Elith and Leathwick, 2009; Elith *et al.*, 2010). While a range of such tools exist, each with varying strengths and weaknesses, we focus here on two main SDM model types: the Generalised Additive Model (GAM) and the RandomForest.

The GAM-based SDM is the more rigorous of the two approaches. GAMs can be best thought of as extensions of classical multi-variate regressions, with modifications to allow observation error structures more complex than the Gaussian (normal) distributions normally employed. Furthermore, GAMs can also incorporate curved response forms (e.g. splines) and the smoothness of these splines can be estimated directly within the model (Wood, 2006): in the form employed here, correlations between observations in space and time can also be

incorporated (Cameletti *et al.*, 2011; Lindgren and Rue, 2015). Finally, and most valuably, being set in a rigorous statistical framework, inference and hypothesis-testing approaches are well developed for GAM models.

The GAM models developed here employ an egg-production approach based on observations of stage 1 eggs, in line with the standard approach taken to this data. We fit the following model structure to the observations

$$n_i \sim Poisson(\lambda_i E_i) \quad (1)$$

$$log(\lambda_i) = f_i + \sum_j s_j(X_{ij}) \quad (2)$$

$$E_i = \frac{V_i \tau_i}{d_i r_i} \quad (3)$$

where n_i is the number of mackerel eggs of stage 1 observed in haul *i*, λ_i is the modelled local egg production (eggs per m^2 per day) corresponding to that haul, and E_i is a scalar that accounts for variations in the sampling process. The egg production is modelled using a log-link as a linear sum of terms associated with each environmental variable, *j*, where $s_j()$ is a spline-smoother for environmental variable *j* and X_{ij} is the value of that variable at the point in space and time corresponding to haul *i*. Spatial and temporal autocorrelation in these observations are accounted for by incorporating a space-time using structure, f_i (Cameletti *et al.*, 2013). The relationship between the egg production rate and the number of eggs counted varies from haul-to-haul and is accounted for by the variable E_i, where V_i is the volume of water filtered, d_i is the sample depth, and r_i accounts for any subsampling (i.e. between the eggs caught in the haul and those actually counted). The egg-development time, τ_i, (in days), is modelled as a function of temperature using the relationship developed by Mendiola et al. (2006).

$$\tau_i = exp(-1.31 \log(T_i) + 6.90$$

where T_i is the temperature experienced by the eggs (deg C). The model is fitted using the Integrated Nested Laplace Approximation (INLA) framework (Cameletti *et al.*, 2013; Lindgren and Rue, 2015).

The second SDM modelling technique applied is that of the Random Forest. In contrast to the statistical rigour of GAMs, the RandomForest approach is a machine learning algorithm that has become popular in the "Big-Data" community: amongst some of the more well-known applications are the suggestion of movies in Netflix and it's use by the US military to identify targets in Afghanistan and Pakistan (Robbins, 2016). A key advantage of Random Forests is that, unlike most linear models, including GAMs, it does not require matrix inversion, meaning that it scales well and can handle particularly large problems including thousands of predictors and millions of observations. Random Forests also can readily account for non-linear effects and interactions between variables in a way that cannot easily be handled in GAMs. Finally, the approach is focused on predictive power, rather than explanatory power, and therefore generalises well to out-of-sample data.

We used Random Forest (RF) models in two different modes. Firstly, we employed a classification approach, where the models were used with presence-absence (PA) data to estimate the probability of observing Mackerel eggs at a given point for a given set of

environmental predictors. Secondly, we used the models in a regression mode to estimate the egg production (λ_l in Equation 1). RF models were fitted using the "randomForest" package (Liaw and Wiener, 2002) in R.

Environmental Predictors

Both the GAM and RF species distribution models use environmental variables as predictors of their response (i.e. of egg production or egg-presence-absence). Haul metadata was therefore complemented by a catalogue of relevant environmental variables. The choice of environmental variables depends on both the availability of data and it's appropriateness for the task of modelling the distribution of Mackerel. The following variables were available from outside the GREENUP catalogue:

- Bathymetric depth based on the ETOPO1 database (Amante and Eakins, 2009) at the latitude and longitude where the haul was taken. Depth was log10 transformed for all analyses.
- Day length, based calculations performed by the maptools package in R, at the latitude and longitude and day where the haul was taken
- Temperature at 5m depth observed by CTD measurements performed in conjunction with sampling for Mackerel eggs
- Net primary production estimates derived from satellite-based observations of ocean-colour and temperature using the VGPM model (Behrenfeld and Falkowski, 1997). This variable was log-transformed for all analyses.

In addition the following variables were employed or generated based on those available from the GREENUP catalogue

- Sea-surface temperature
- Near-surface velocity, calculated from the norm of u and v velocity components provided by the model near the surface.
- Primary productivity, as modelled by the biogeochemical component of the model. This variable was log-transformed for all analyses.
- Potential biomass of MTL groups near the surface during the day, defined here as the concentration of the epipelagic functional group. This variable was log-transformed for all analyses.
- Potential production of MTL groups near the surface during the day, defined here as the production of the epipelagic functional group. This variable was log-transformed for all analyses
- Potential biomass of MTL groups near the surface during the night, defined here as the sum of the concentration of the epipelagic, migratory mesopelagic and highly-migratory bathypelagic functional groups. This variable was log-transformed for all analyses
- Potential production of MTL groups near the surface during the night, defined here as the sum of production by the previously mentioned groups. This variable was log-transformed for all analyses

Each of these model-based products was available based on three different physical reanalysis models:

- GLORYS2V4, on a weekly temporal resolution and 1/4 degree spatial from 1998 to 2015 (inclusive)
- ARMOR3D, on a weekly temporal resolution and 1/4 degree spatial resolution, from 1998 to 2016 (inclusive)
- PSY4, on a daily resolution and 1/12 degree spatial resolution, from 2013 to 2016 (inclusive).

For each of these variables and models, the appropriate quantity at the point in time and space where the hauls were made was extracted from the corresponding database by bilinear spatial interpolation at the nearest temporal point.

An important point for the development of SDMs is that the difference in temporal coverages varies between these models e.g. between the PSY4 (2013-2016) and GLORYS2V4 / ARMOR3D models (1998-2015/16). Such differences in coverage could potentially create problems with respect to comparison of model predictive skill i.e. it would be difficult to attribute differences in SDM model skill to differences in the observational dataset or due to the GREENUP MTL products. The analysis was therefore formed around two datasets – the first based on all data points that were common to GLORYS2V4, ARMOR3D and the MEGS dataset (approximately 6000 hauls, covering 1998-2013), and the second based on data points covered by all three models and the MEGS dataset (approximately 1400 points from 2013).

As both a check on the extraction process and to understand the differences and similarities between the GREENUP MTL products, the correlation between the versions of the individual variables extracted from each model was calculated (Figure 2). There are clear differences in the consistency of the various variables: sea surface temperature (SST), for example, is highly consistent between models, with correlation coefficients exceeding 97%, whilst the surface velocity (vel) shows poor agreement between models (correlations of 20-40%). The values of the GREENUP MTL products used here lie between these two extremes, with agreement ranging between 40 and 80%. These results suggest that we should expect important differences in the ability of these models to predict the distribution of mackerel eggs, and that it is therefore important to consider all model-variable combinations when developing and comparing the SDM models.

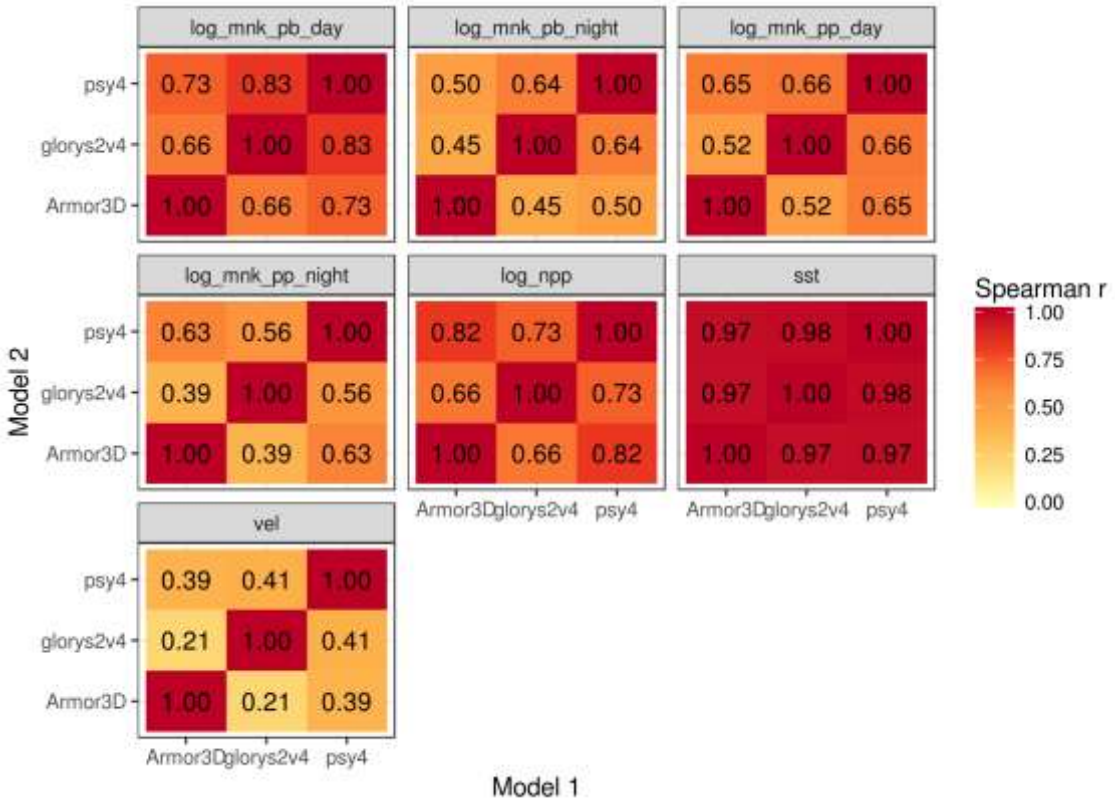

Figure 2 Correlations between different versions of the variables (panels) considered as environmental predictors derived from each of the three physical models (ARMOR3D, GLORS2V4 and PSY4). Correlations are expressed as spearman rank correlation coefficients, r.

Development of GAM Model & Correlation Structure

Correctly accounting for the spatial and temporal correlation between observations is widely recognised as being critical for the development of appropriate and reliable species distribution models, and is particularly important in cases where we wish to make statistical inference (Dormann, 2007; Cameletti *et al.*, 2013; Brun *et al.*, 2016). However, developing models with spatial and temporal correlation structures is technically and computationally challenging, and the steps taken towards developing the model here described below.

Before incorporating environmental explanatory terms into the species distribution model, it is first necessary to ensure that the structure of the rest of the modelling framework is appropriate. In particular, it is necessary to consider the appropriate choice of observational model, and the temporal and spatial correlation structures employed. We do this by first developing a spatial-temporal only model (i.e. without environmental covariates included) that can be used to both characterise the distribution of spawning in both space and time, and to function as a baseline against which models incorporating environmental predictors can be compared. Environmental predictors are then incorporated into this model at a later stage to create a species-distribution model.

Firstly, the spatial domain over which the analysis is to be performed was defined. To facilitate analysis on a spatially isotropic coordinate system, all hauls were transformed to a UTM grid (zone 29): a general spatial domain was then defined based on a non-convex hull around all of these points with a buffer region of 200km. This domain was then intersected with coastlines

derived from the Global Self-consistent, Hierarchical, High-resolution Shoreline Database (GSHHS) to define the oceanic domain of interest. Finally, this domain was discretised into an unstructured mesh based on constrained Delaunay triangulation to reflect the distribution of samples, with a minimum grid spacing of 50km and a maximum of 200km. The use of such a mesh is highly desirable when modelling the ocean, as it maintains the inherent spatial structure of the wet regions, allowing for correlations for example, around Ireland via water, but not across land (Figure 3). The spatial field on this grid was then represented using a Matern correlation structure and solved using a stochastic partial differential equation approach (Cameletti *et al.*, 2013).

Figure 3 Mesh used in the development of the spatial model. Red points represent the distribution of hauls, while the black lines link nodes on the mesh and the blue line is the boundary conditions.

Temporal variability in spawning distribution was incorporated into the model by allowing for correlation in time using a first-order autoregressive (AR1) process. The time at which an individual haul was taken in a particular survey was represented as the day of year (doy), and then binned into temporal blocks.

Choice of Observation Model

The appropriate choice of observation model to characterise the nature of the sampling noise in these models is not immediately clear. Given the count nature of the egg observations, a Poisson structure (Equation 1) would be the obvious first choice. However, this might not necessarily be the most appropriate: the presence of small-scale patchiness in the distribution of eggs, for example, seems highly likely, and can introduce both zero-inflation and over dispersion (Zuur *et al.*, 2009). In addition to the Poisson observation error, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial versions of this model were therefore also considered, and their appropriateness assessed.

Models were fitted using the structure described above with a range of available observation models. Models based on the Poisson observation structure (and its zero-inflated variants) were generally poorly suited to the modelling task, did not converge well, and often incorporated unrealistically-high local estimates of egg production: these models were therefore removed from consideration. The remaining set of observation models, based on the negative binomial error structure with and without zero-inflation, were then compared based on common model selection criteria (i.e. the deviance information criterion, DIC, and the Watanabe-Akaike information criterion, WAIC). These model selection criteria show (Figure 4) that the negative-binomial parameterisation is consistently the best model (has the lowest DIC and WAIC) and is therefore used as the observation model for the rest of the GAM SDM development.

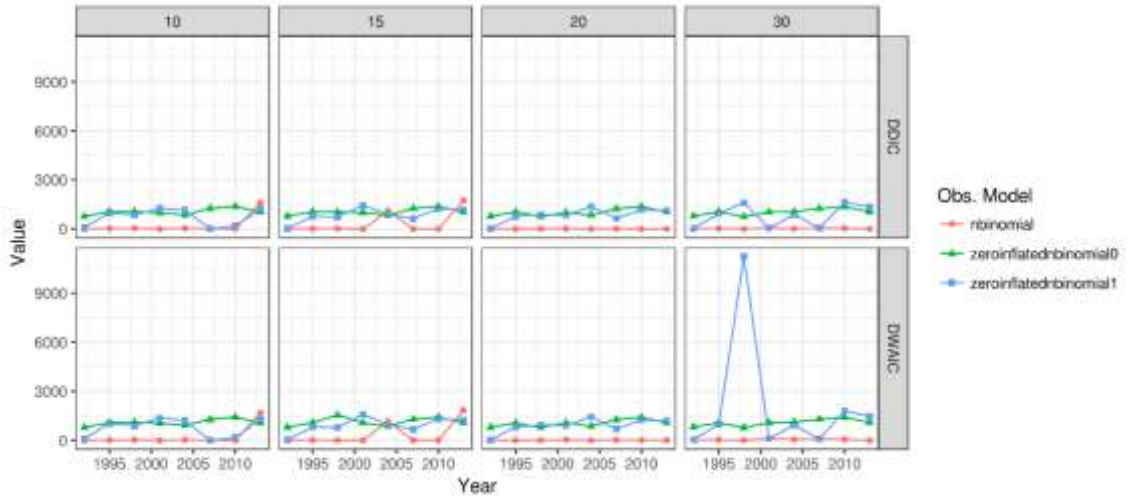

Figure 4 Model selection for observation models. Models are fitted for each survey year (horizontal axis) and the goodness of fit assessed using the deviance information criterion, DIC, and Watanabe-Akaike information criterion (WAIC)(horizontal rows of panels): in each case, these are expressed as the difference from the model with the lowest value I.e. DDIC and DWAIC). Observation models (coloured lines and symbols) considered are the negative binomial (nbinomial) and two different parameterisations of the zero-inflated model. The effect of different temporal binning schemes (columns of panels) corresponding to the width of the temporal bin in days is also considered.

Choice of Time Binning

The appropriate size of the time bins for use in this analysis is not clear *a priori* but involves a clear trade-off. On one hand, using a fine temporal resolution gives a high degree of model complexity, while a coarse resolution limits the ability to represent the relevant processes. The choice of bin size was therefore also considered as a key parameter in the development of the model.

The effect of time binning on the model selection criteria was inconsistent, and tended to be dependent on the particular year in question. In some cases, a fine time resolution was supported, whilst in other cases a coarse resolution was preferred (Figure 5). In the absence of a clear signal favouring one time resolution, a 30 day binning was employed, due to the greatly reduced model run-times associated with this resolution

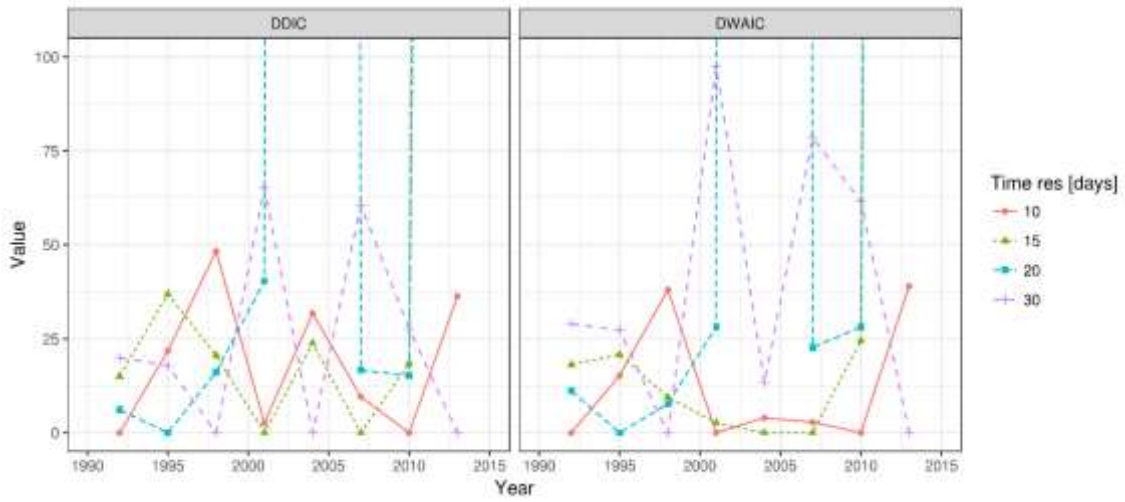

Figure 5 Model selection for time binning. Models are fitted for each survey year (horizontal axis) and the goodness of fit assessed using the deviance information criterion, DIC, and Watanabe-Akaike information criterion (WAIC)(panels): in each case, these are expressed as the difference from the model with the lowest value I.e. DDIC and DWAIC). The effect of different temporal binning schemes (coloured lines and symbols, corresponding to the width of the temporal bin in days) is considered. For all models, the negative binomial observation model was employed.

Consistency of parameters

The GAM model structure employed was fitted to each survey year individually. As a result, the key parameters were estimated based on a single survey year, and not, as would be ideal, based across the entire set of surveys. Checking consistency of these parameters between surveys is therefore a key aspect of ensuring the stability and credibility of the models.

There is generally good agreement between the parameters estimated in individual years (Figure 6). The negative binomial over-dispersion parameter is perhaps the least consistent: however, this parameter is also difficult to estimate reliability, and the variation is within the confidence intervals associated with this variable. The effect of temporal binning on the parameters is generally less than the interannual variability, although it is often systematic in nature: a notable case is the temporal correlation coefficient (GroupRho), which increases as the temporal resolution becomes finer, a result consistent with expectations. Finally, the parameters associated with the 1992 survey are often appreciably different from the rest of the years, a feature that may relate to the development and refinement of the survey methodology in its early years.

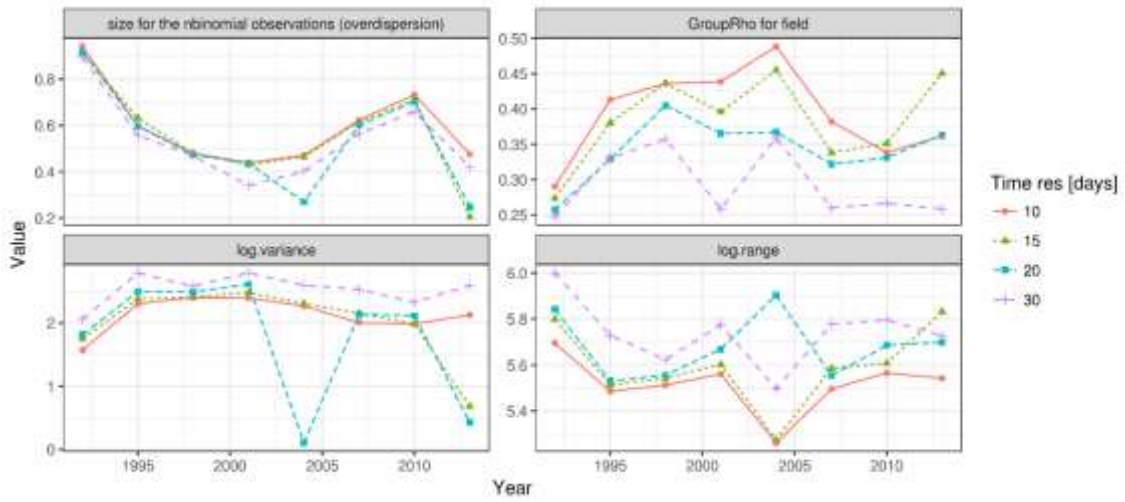

Figure 6 Consistency of parameters between model fits. Four key parameters were extracted from the model ensemble and plotted as a function of survey year: Top left, negative binomial over-dispersion parameter. Top right, temporal correlation parameter. Bottom left: variance (amplitude) of the spatial field. Bottom right: spatial correlation parameter. Coloured lines correspond to the various time resolutions.

Visualisation of Model Fit

The fitted spatial-temporal GAM model agrees well with expectations from previous work (e.g. survey and stock assessment reports). Egg production generally follows the continental shelf edge (Figure 7), progressing from the Bay of Biscay early in the year northwards towards Iceland by the middle of the year (days 150-180). The spawning distribution also becomes more diffuse later in the spawning period, spreading westwards away from the continental shelf and into the deeper waters south of Iceland. However, the majority of the spawning takes place during spring (days 60-120), with contributions outside this period being relatively minor (Figure 8).

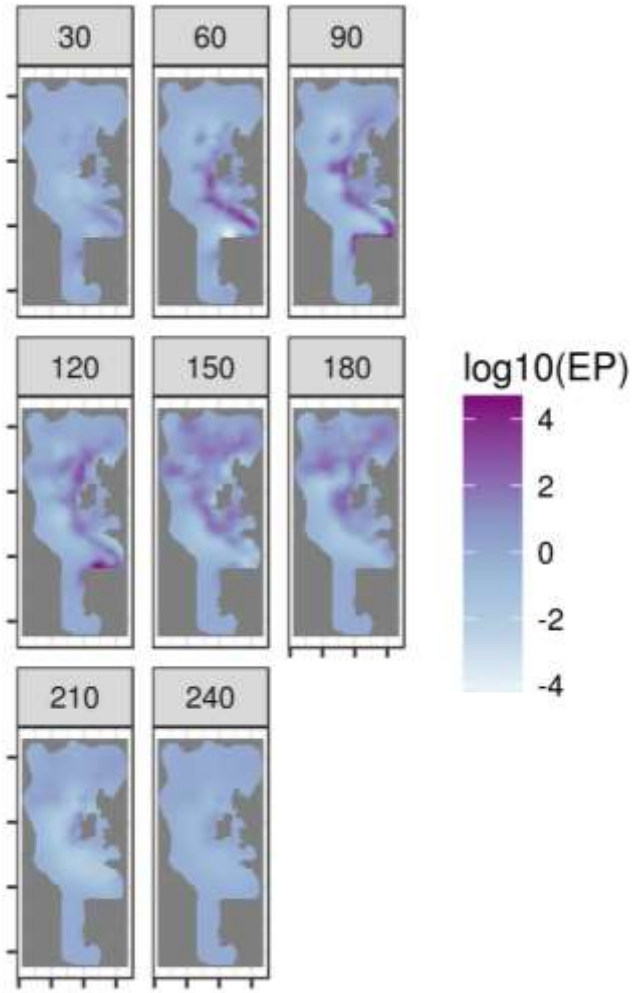

Figure 7 Modelled egg production as a function of time for the 2013 survey. Egg production (EP) is shown on a log10 scale. Individual panels correspond to a 30 day time-window, centered on the day of year indicated at the top of the panel (e.g. the "60" panel shows hauls taken between day-of-year 45 and 75).

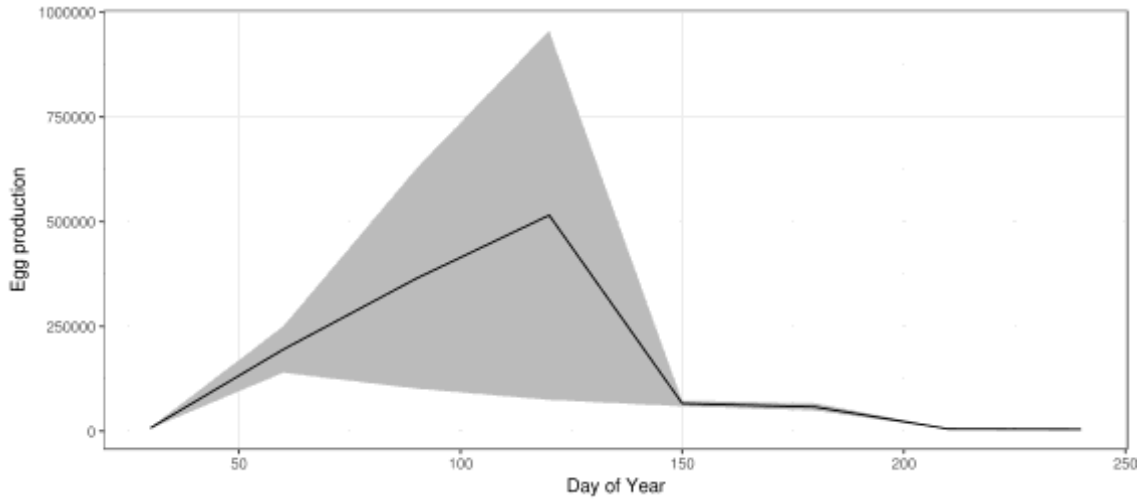

Figure 8 Temporal distribution of spawning from the 2013 egg survey. Total egg production within a period is plotted as a function of day of year (black lines) with 95% confidence intervals (grey area)

Development of Species Distribution Models

Selection of Environmental Variables

An important first-step in developing species distribution models is the choice of environmental predictors. As noted above, the set of predictors to be incorporated is often constrained by the availability of relevant data, particularly in the ocean. It is also important to ensure that the predictors selected give biological meaning, and those employed here are thought to be the most relevant for the task at hand. However, a third criteria that is often over-looked is the question of co-variance and co-linearity between predictors (Zuur *et al.*, 2010; Dormann *et al.*, 2013). Including explanatory variables that are correlated can inflate the variance of estimated regression parameters, potentially lead to the incorrect identification of relevant predictors and degrade the out-of-sample predictive power of a model. A widely recognised rule-of-thumb is that action should be taken to avoid issues of collinearity when a pair of predictors have a coefficient of determination (R^2) greater than 0.5 (Dormann *et al.*, 2013).

Our analysis highlight several sets of variables where care should be taken to avoid collinearity issues (Figure 9). In particular, the most striking is the correlation between the two different SST variables, one (ObsTemp) derived from measurements made in conjunction with the MEGS survey, and the other from the physical component of each GREENUP MTL model. However, given that the GREENUP MTL products are based on reanalysis models that assimilate observational data, this result is not surprising and is indeed a reassuring indicator that these models are performing well: indeed, a similar, although weaker result can be seen between the modelled and satellite-derived estimates of Net Primary Productivity, which are also assimilated into the physical models. Strong correlations are also seen between the MTL products, particularly between day-night values and to a lesser extent between potential product and potential biomass variables: again, this is not an entirely surprising result when considering that all of these variables reflect the underlying biological environment for growth. To address these issues of potential collinearity in the SDM model, only one of the variables from each of the groups of correlations (i.e. SST, NPP and MTL products) should be considered in the SDM models at a time.

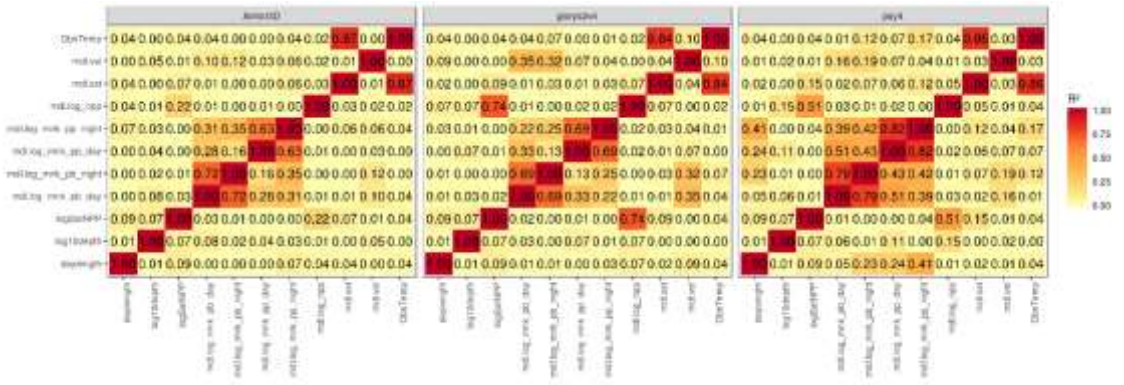

Figure 9 Collinearity analysis. The coefficient of determination (R^2) between each combination of variables is plotted as a colour on the grid, and also as the text value. The three panels correspond to the three MTL models considered in this work. Correlations are calculated on a pairwise-complete basis across the entire MEGS dataset where it was possible to match-up the environmental variable with the corresponding spatial and temporal coordinates.

Developing the SDMs

The modelling framework developed above was used as the basis to further develop a GAM-based species distribution model. Environmental predictors were incorporated into the space-time correlation structure model via the second term on the right-hand side of Equation 2. Second-order random walk terms were added for each environmental variable to estimate the effect of the variable on local egg production, and suites of models fitted accordingly.

Unfortunately, the fitting of these models proved to be wholly unsuccessful. Problems with model convergence became immediately apparent when incorporating these additional variables, and in many cases models simply failed to converge, even after several days of run time on a large multi-core cluster. In cases where the models did converge, the uncertainty estimates associated with parameters were unrealistically wide. Such model behaviours are characteristic of over- and/or poorly-parameterised models that lack clear optima: it may be that the addition of environmental variables, possibly in conjunction with complex (although necessary) observational error structures introduced too much complexity into the model in comparison to the size and information content of the observational data set. Such issues are not entirely unprecedented with this type of model (e.g. Brun *et al.*, 2016) and can be extremely challenging to debug: numerous attempts to get to the bottom of this problem proved to be unsuccessful.

The decision was there made to focus instead on the second type of species distribution model being considered, the Random Forest. As noted above, these models perform well with very large datasets and this type of work was therefore expected to be well within their capabilities. Whilst they lack the statistical rigor of the GAM SDM that we have mainly focused on here (and in particular the ability to account for spatial and temporal correlation structures), the desired end result of this work (an assessment of the predictive ability of the GREENUP MTL products for modelling Mackerel distribution) can also be obtained from these models. Within the limited resources available for this work, and the problems with fitting the GAM SDMs, it was therefore viewed as more productive to focus on the RF approach.

In contrast to the GAM-SDM approaches, the RF SDM models were well behaved, including both Presence-Absence (PA) Categorical models and Egg-Production (EP) based regression models. Standard model diagnostics associated with technique suggested that the models had converged appropriately and were giving a good fit. The remainder of this work is therefore based on using these models to assess the performance of the GREENUP MTL products in this context.

Assessing performance of the GREENUP MTL products

Predictive Performance

The performance of the GREENUP MTL products and their ability to predict the distribution of spawning mackerel were examined by comparing the predictive ability of models with and without the product as an explanatory variable. The following metrics were used to assess the predictive skill of the models:

- The positive predictive value (PPV) i.e. the probability that model-based predictions of egg presences are correct (PA models).
- The negative predictive value (NPV) i.e. the probability that model-based prediction of egg absences are correct (PA models)
- The true-skill score (TSS), which combines model sensitivity (fraction of correctly predicted presences) and specificity (fraction of correctly predicted absences) into a single metric (PA models).
- Mean-squared error (MSE), defined as the mean of the squared difference between the log Egg-production observed in MEGS and that predicted by the model (EP Models)
- Coefficient of determination (R^2), the proportion of the variation in observed log-Egg-Production that is explained by the models.

For both the PA and EP RF models, three broad classes of models were created and their skill scores derived:

- "Baseline" model, which serves as a reference against which other models are compared. Incorporates log10depth, day length, SST and surface velocity as predictors, but information about the biological environment (lower- or mid- trophic levels)
- "Baseline + NPP" as above, but also incorporating model-based estimates of Net Primary Production as a predictor
- "MNK" models, incorporating the "Baseline + NPP" model, plus one of the four micronekton (MTL) products available.

The PA RF models showed a high degree of skill, regardless of the model configuration (Figure 10). NPV values typically exceeded 80% and PPV values 75%, indicating the model has some skill in discriminating between presence and absence: for contrast, a coin-toss model would have a skill of around 50% for this data set. The TSS skill scores, which range between -1 and 1, and take a value of 0 for a coin-toss model, are also good, typically sitting above 0.5.

Incorporating biological variables, such as the GREENUP MTL products, improves the skill of these models further, relative to the baseline models. There is a clear increase in skill moving from the "baseline" to "baseline+NPP" to models incorporating MTL variables, indicating that these models are being improved by the addition of these variables. Furthermore, the potential biomass during the day ("pb_day" variable) is consistently the best performing of these MTL variables. Finally, the difference between the GREENUP models used as sources of this data appears to be relatively minor compared to the differences due to changes in the RF model configuration.

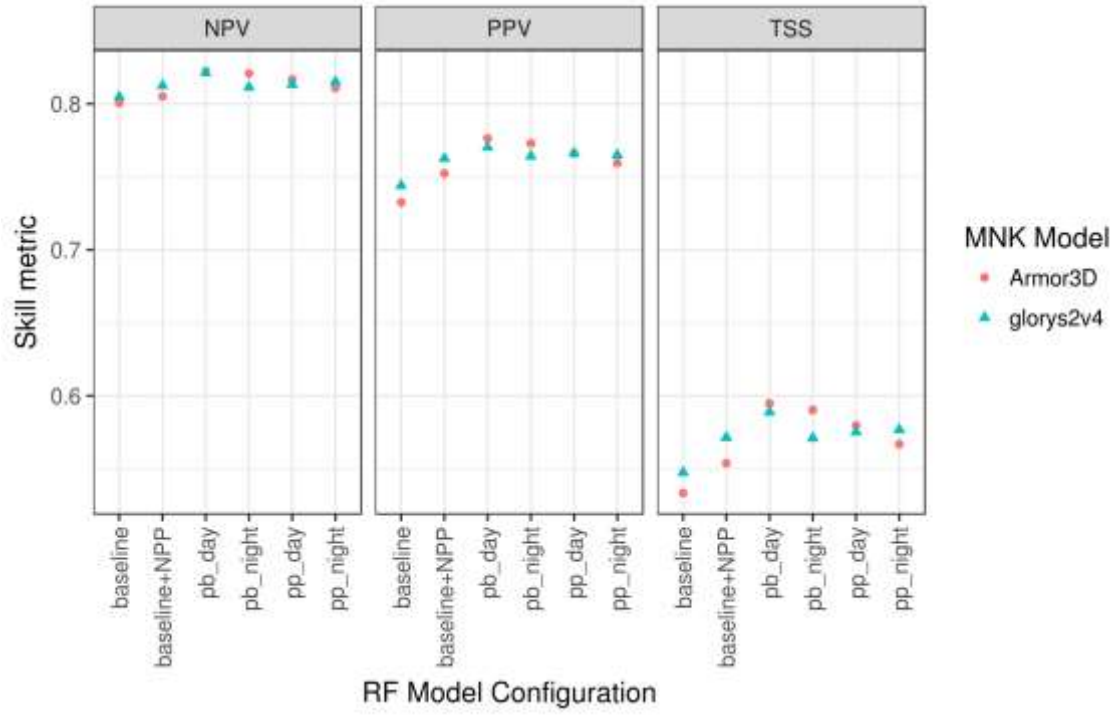

Figure 10 Skill of Random Forest Presence-Absence models. The various model configurations are plotting on the horizontal axis, with the two baseline models at the left of the panel. "pb" indicates potential biomass whilst "pp" indicates potential production of MTL for both day and night time. Coloured points correspond to the two main GREENUP MTL models considered, ARMOR3D and GLORYS2V4. The individual panels correspond to the skill metrics considered: NPV (negative predictive value), PPV (Positive predictive value) and TSS (True-skill score). Higher values of each skill metric indicate greater predictive skill.

Comparable results can be seen for the skill of the RF Egg-production regression models (Figure 11). Model performance is generally good, with a coefficients of determination up to 50%: importantly, adding MTL products leads to an increase in this metric from around 38% to 46%. The mean-squared error of prediction is relatively high, however, (corresponding to a mean prediction error of 2 EP units on a logarithmic-scale), but is tolerable compared to the range of log-EP values in the dataset (-6 to +6 log-EP units), indicating that the model is clearly capable of distinguishing between areas of high and low spawning activity.

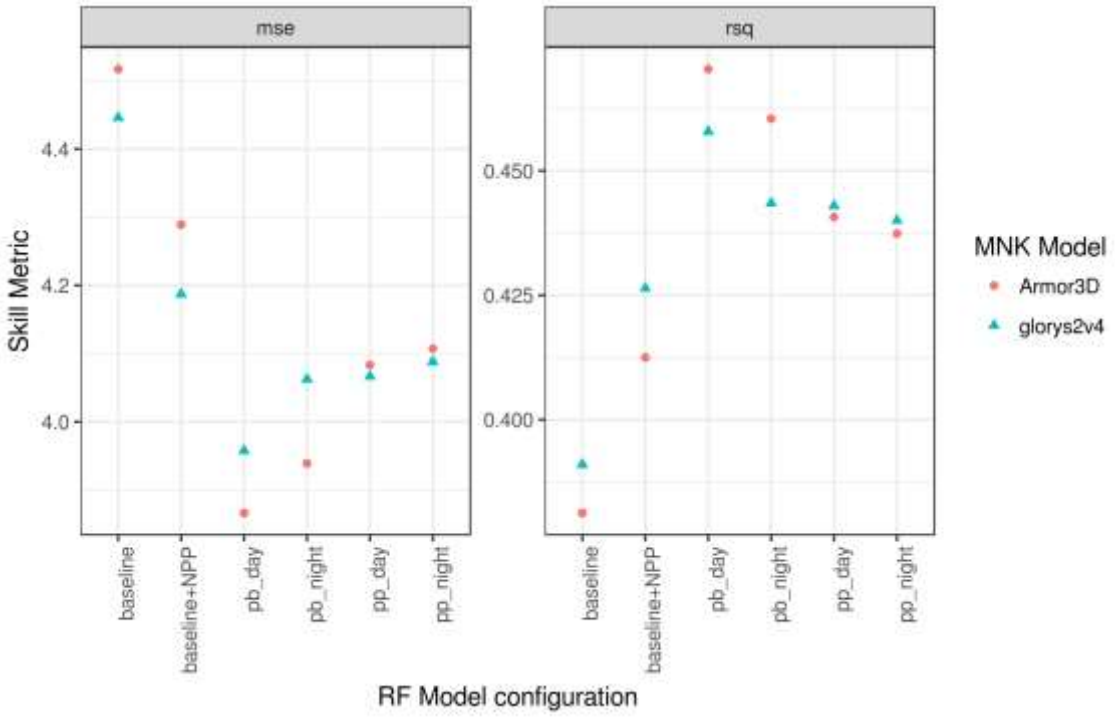

Figure 11 Skill of Random Forest Egg-production regression models. The various model configurations are plotting on the horizontal axis, with the two baseline models at the left of the panel. "pb" indicates potential biomass whilst "pp" indicates potential production of MTL for both day and night time. Coloured points correspond to the two main GREENUP MTL models considered, ARMOR3D and GLORYS2V4. The individual panels correspond to the skill metrics considered: mse (Mean squared error), rsq (coefficient of determination). More skilful models have higher values of rsq and lower values of mse.

Variable Importance

A second way to understand how individual variables are used in a species distribution model is via the concept of variable importance. Although individual implementations vary, the basis of this concept is to examine how the performance of a model worsens when a model is either removed or scrambled (thereby breaking the relationship between the explanatory and response variables).

The results of the variable importance analysis show consistency between both the presence-absence (Figure 12) and egg-production (Figure 13) models. The most important variables are clearly day length, log10depth and the sea surface temperature. Variations between RF model configurations and GREENUP models are generally less compared to differences in importance between variables. There are subtle differences between the importance of the MTL variable, depending on the particular form of the variable used: as above, the "potential biomass during the day" is consistently ranked as the most important of these variables.

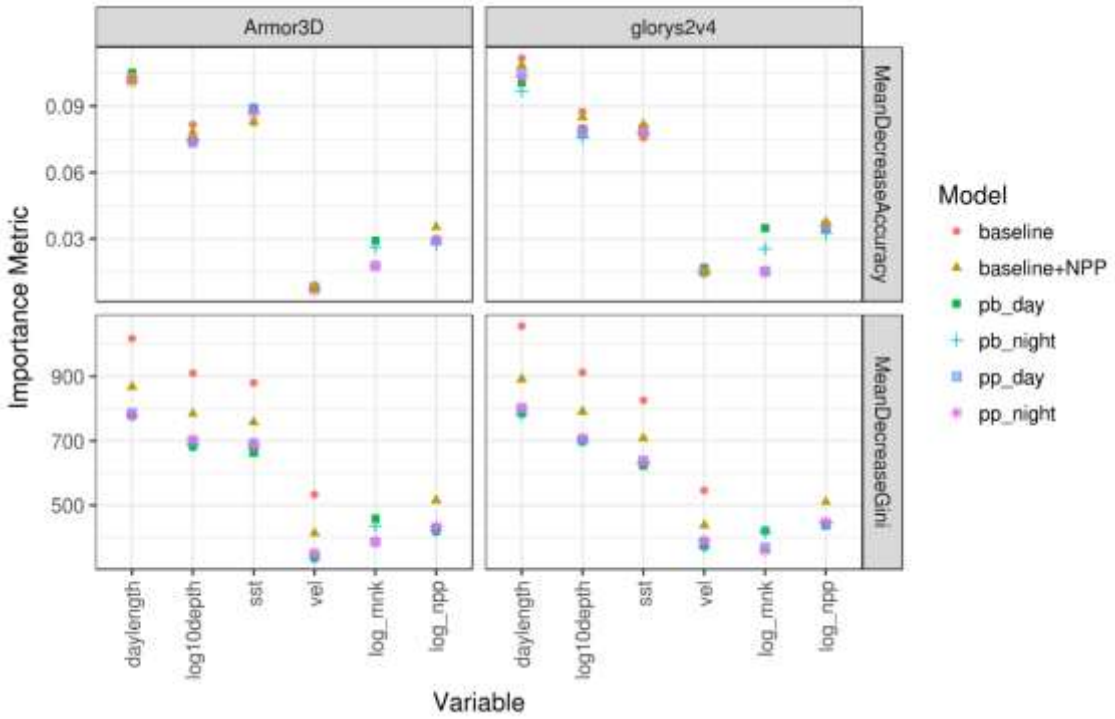

Figure 12 Importance of individual variables in each of the RF Presence-Absence models. The variable in question is plotted on the horizontal axis and the value of the importance metric on the vertical axes: coloured symbols correspond to the different RF model configurations. Vertical columns of panels correspond to the different GREENUP MTL models used as explanatory variables, whilst the horizontal rows of panels correspond to different importance metrics. In all cases, a higher numeric value of the importance metric corresponds to greater importance to the model.

i

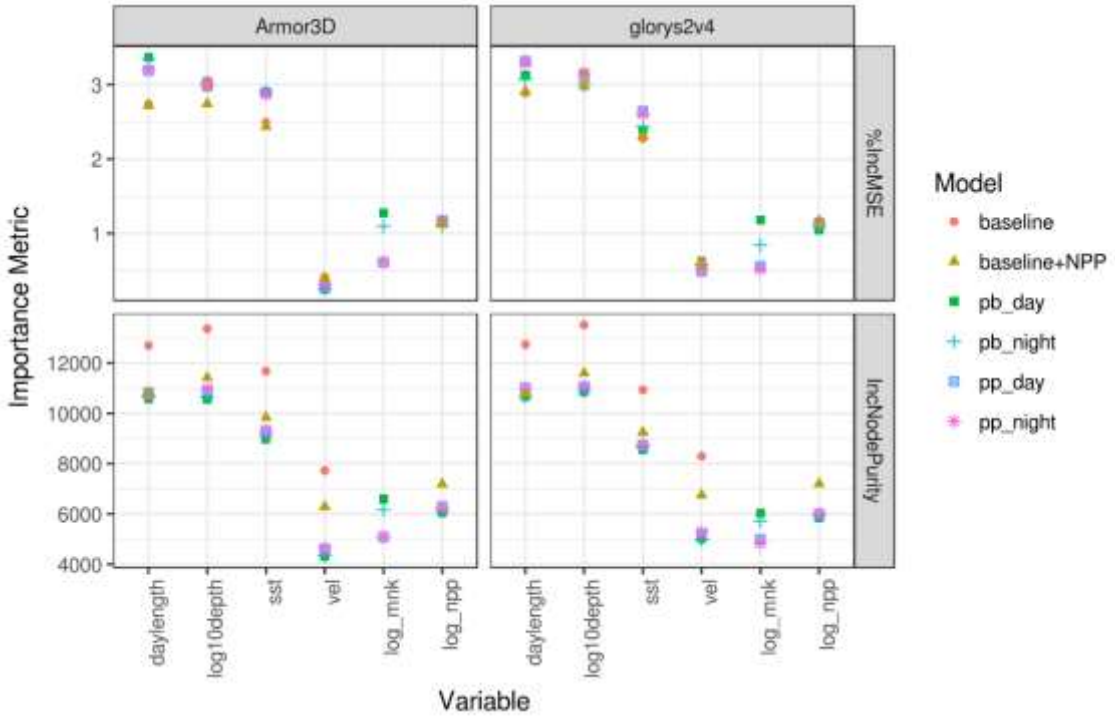

Figure 13 Importance of individual variables in each of the RF Egg-Production models. The variable in question is plotted on the horizontal axis and the value of the importance metric on the vertical axes: coloured symbols correspond to the different RF model configurations. Vertical columns of panels correspond to the different

GREENUP MTL models used as explanatory variables, whilst the horizontal rows of panels correspond to different importance metrics. In all cases, a higher numeric value of the importance metric corresponds to greater importance to the model.

Comparison with PSY4 Model

As noted above, the temporal coverage of the PSY4 reanalysis model used to produce the GREENUP MTL products is significantly different from the other two reanalysis products (GLORYS2V4 and ARMOR3D), and only covers the 2013 MEGS survey. On the other hand, this model has a substantially higher spatial (1/12 degree vs 1/4 degree) and temporal (daily vs weekly) resolution, and could therefore be expected to much better at resolving the small-scale processes of relevance to fish species such as Mackerel. We have therefore made a comparison between the MTL products based on the three reanalysis products for the one year (2013) where this was possible, and examined the skill of this model.

The performance of RF models based on PSY4 are broadly comparable to those seen for the other reanalysis products, and more generally throughout this work. The ability of the RF SDM models to predict egg-production (Figure 14) is comparable in this year to other years, with the coefficient of determination reaching 0.45 and the mean-squared error around 3.3. Again, SDMs using the pb_day variable represent MTL processes have the best performance. SDM models based on the PSY4 reanalysis appears to line between the ARMOR3D and GLORYS2V4 reanalyses: however, the results are within the general range of variability observed. Similar results are also seen for the RF Presence-Absence SDM. There is therefore little reason to believe that GREENUP MTL products based on the PSY4 reanalysis are appreciably better (or worse) than those from the other reanalysis products.

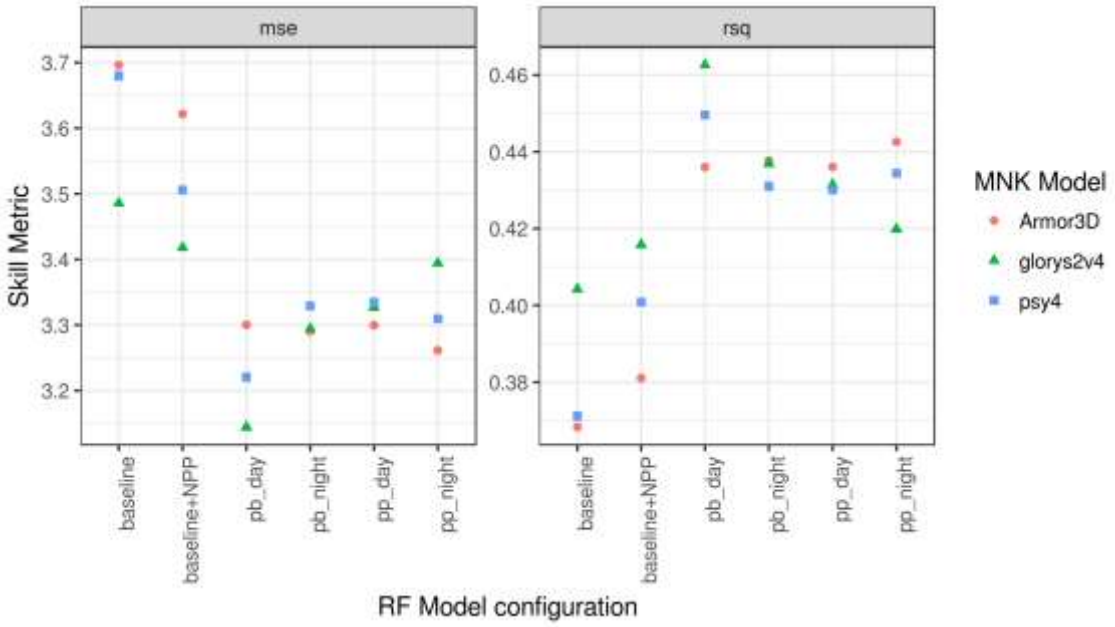

Figure 14 Skill of Random Forest Egg-production regression models for all three reanalysis products, including the PSY4 reanalysis. The various model configurations are plotting on the horizontal axis, with the two baseline models at the left of the panel. "pb" indicates potential biomass whilst "pp"indicates potential production of MTL for both day and night time. Coloured points correspond to the two main GREENUP MTL models considered, ARMOR3D and GLORYS2V4. The individual panels correspond to the skill metrics considered: mse (Mean squared error), rsq (coefficient of determination). More skilful models have higher values of rsq and lower values of mse. These results

are based on MEGS data and RF models calibrated and validated against the 2013 survey only, where all reanalysis models are available.

Discussion and Conclusions

The results obtained here show that the GREENUP MTL products improve our ability to both model and predict the distribution of Mackerel spawning activity along the European continental shelf margin. In both the case of dealing with presence-absence data and with the full set of egg-production, the addition of MTL products to the models uniformly improved their performance over a baseline model.

In our analysis we considered MTL variables describing both the potential biomass and the potential productivity during both the day and night. It was not possible, *a priori*, to see which of these would be the most appropriate for characterising the distribution of Mackerel – good arguments can be made for both - and all four combinations were therefore considered in our modelling framework. The results consistently supported the potential biomass during the day (pb_day) variable as the best of these. In retrospect, this is perhaps an unsurprising result: Mackerel are visual predators and therefore it seems reasonable that this variable gives the best predictive skill. However, this does not necessarily mean that better variables cannot be found: a weighted averaged of the daytime and night-time biomasses, for example, could also be considered and should be investigated in future work.

The models produced here have achieved pleasingly high predictive skills that may well lend themselves to futures applications. We have been able to predict the presence or absence of eggs with around 80% accuracy, and the level of egg production with an R^2 approaching 50%. These models could certainly therefore be used to both understand and improve the Mackerel survey. For example, these models can potentially be used to fill gaps in the survey coverage, and to improve the survey design. It is also possible to envisage these models being used in a near-real time context to inform the execution of the survey as it progresses or even in a forecast context, should MTL-based forecast products become available. Future work will therefore involve a direct collaboration with the survey coordination group to establish how such models can be applied to improve the monitoring of this stock.

The approach applied here employed two approaches: GAM SDM and Random Forest SDMs. Unfortunately while we were able to develop an appropriate modelling structure using the GAM approach, it was not possible to take this further into a full-blown SDM model. This experience highlights important differences in the applicability of the different modelling frameworks: whilst the GAM approach is more statistically satisfying, it is also substantially more complex. Nevertheless, the work invested in developing this framework has not been wasted: the spatial-temporal model of the distribution of egg-production is still extremely useful for survey practitioners as a way to characterise the distributions of egg production and their variation between years, and to produce survey indices for further use in the assessment of this stock. Furthermore, if it proves possible to resolve the technical difficulties encountered here at some point in the future, making a comparison between the results from the two different approaches would both bolster our confidence in, and the rigour of, the results obtained here.

The usefulness of the GREENUP MTL products demonstrated here also opens the door to other applications. For example, the summer feeding distribution of Mackerel has also been shown to vary greatly between years, and has led to conflicts between nations over access to fishing rights (Hannesson, 2012). Food distribution is also thought to be a key factor in this process and the GREENUP MTL products can therefore potentially shed insight into these changes and the processes driving them. Similarly, the recruitment (productivity) Blue Whiting (*Micromesistius poutassou* L), a large and commercially important fish stock found in similar regions to the Mackerel stock examined here, has varied tremendously between years (Payne *et al.*, 2012), and the productivity and abundance of food in this region is also thought to be important for the survival of their juveniles. Many other similar queries exist within fisheries research, and it can be hoped that the development of models such as the GREENUP MTL products will shine fresh light on these problems in the future.

References

Amante, C., and Eakins, B. W. 2009. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24.

Astthorsson, O. S., Valdimarsson, H., Gudmundsdottir, A., and Oskarsson, G. J. 2012. Climate-related variations in the occurrence and distribution of mackerel (Scomber scombrus) in Icelandic waters. ICES Journal of Marine Science, 69: 1289–1297. http://icesjms.oxfordjournals.org/content/69/7/1289.short (Accessed 7 December 2014).

Behrenfeld, M., and Falkowski, P. G. 1997. A consumer's guide to phytoplankton primary productivity models. Limnology and Oceanography, 42: 1479–1491.

Berge, J., Heggland, K., Lønne, O. J., Cottier, F., Hop, H., Gabrielsen, G. W., Nøttestad, L., *et al.* 2015. First records of Atlantic mackerel (Scomber scombrus) from the Svalvard Archipelago, Norway, with possible explanations for the extension of its distribution. Arctic, 68: 54–61. http://dx.doi.org/10.14430/arctic4455.

Bruge, A., Alvarez, P., Fontán, A., Cotano, U., and Chust, G. 2016. Thermal Niche Tracking and Future Distribution of Atlantic Mackerel Spawning in Response to Ocean Warming. Frontiers in Marine Science, 3: 1–13. http://journal.frontiersin.org/article/10.3389/fmars.2016.00086.

Brun, P., Payne, M. R., and Kiørboe, T. 2016. Trait biogeography of marine copepods - an analysis across scales. Ecology Letters, 19: 1403–1413. http://doi.wiley.com/10.1111/ele.12688.

Cameletti, M., Ignaccolo, R., and Bande, S. 2011. Comparing spatio-temporal models for particulate matter in Piemonte. Environmetrics, 22: 985–996.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. 2013. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. AStA Advances in Statistical Analysis, 97: 109–131.

Dormann, C. F. 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data: 129–138.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., *et al.* 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36: 27–46. http://doi.wiley.com/10.1111/j.1600-

0587.2012.07348.x (Accessed 6 October 2012).

Elith, J., and Leathwick, J. R. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics, 40: 677–697. http://www.annualreviews.org/eprint/HWR4cusJrXYCSPZ9sUDj/full/10.1146/annurev.ecolsys.110308.120159 (Accessed 5 October 2012).

Elith, J., Kearney, M., and Phillips, S. 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution, 1: 330–342. http://dx.doi.org/10.1111/j.2041-210X.2010.00036.x (Accessed 15 July 2014).

Hannesson, R. 2012. Sharing the Northeast Atlantic mackerel. ICES Journal of Marine Science, 70: 259–269. http://icesjms.oxfordjournals.org/cgi/doi/10.1093/icesjms/fss134 (Accessed 21 February 2013).

ICES. 2013. Report of the Ad hoc Group on the Distribution and Migration of Northeast Atlantic Mackerel (AGDMM) 30-31 August 2011 and 29-31 May 2012 ICES Headquarters , Copenhagen. 215 pp.

Liaw, A., and Wiener, M. 2002. Classification and Regression by randomForest. R News, 2: 18–22. http://cran.r-project.org/doc/Rnews/.

Lindgren, F., and Rue, H. avard. 2015. Bayesian Spatial Modelling with \pkgR-INLA. Journal of Statistical Software, 63: ??–?? http://www.jstatsoft.org/v63/i19.

Mendiola, D., Alvarez, P., Cotano, U., Etxebeste, E., and de Murguia, a. M. 2006. Effects of temperature on development and mortality of Atlantic mackerel fish eggs. Fisheries Research, 80: 158–168. http://linkinghub.elsevier.com/retrieve/pii/S0165783606001755 (Accessed 25 March 2014).

Pacariz, S. V, Hátún, H., Jacobsen, J. A., Johnson, C., Eliasen, S., and Rey, F. 2016. Nutrient-driven poleward expansion of the Northeast Atlantic mackerel (Scomber scombrus) stock: A new hypothesis. Elementa: Science of the Anthropocene, 4: 105. http://elementascience.org/article/info:doi/10.12952/journal.elementa.000105.

Payne, M. R., Egan, A., Fässler, S. M. M., Hátún, H., Holst, J. C., Jacobsen, J. A., Slotte, A., *et al.* 2012. The rise and fall of the NE Atlantic blue whiting (*Micromesistus poutassou*). Marine Biology Research, 8: 475–487. http://www.tandfonline.com/doi/abs/10.1080/17451000.2011.639778 (Accessed 3 June 2014).

Robbins, M. 2016. Has a rampaging AI algorithm really killed thousands in Pakistan? The Guardian. https://www.theguardian.com/science/the-lay-scientist/2016/feb/18/has-a-rampaging-ai-algorithm-really-killed-thousands-in-pakistan.

van der Kooij, J., Fässler, S. M. M., Stephens, D., Readdy, L., Scott, B. E., and Roel, B. A. 2015. Opportunistically recorded acoustic data support Northeast Atlantic mackerel expansion theory. ICES Journal of Marine Science: Journal du Conseil. http://icesjms.oxfordjournals.org/content/early/2015/12/18/icesjms.fsv243.abstract.

Wood, S. N. 2006. Generalized additive models: an introduction with R. Chapman Hall / CRC Press, Boca Raton, FL.

Zuur, A. F., Ieno, E., Walker, N., and Saveliev, A. 2009. Mixed effects models and extensions in

ecology with R. Springer, New York. http://books.google.com/books?hl=en&lr=&id=vQUNprFZKHsC&oi=fnd&pg=PA1&dq=Mixed+effects+models+and+extensions+in+ecology+with+R&ots=katKvQZF0t&sig=ngYfGeT-yybvzJww7K9ystKk40M (Accessed 4 May 2012).

Zuur, A. F., Ieno, E. N., and Elphick, C. S. 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution, 1: 3–14. http://doi.wiley.com/10.1111/j.2041-210X.2009.00001.x (Accessed 8 March 2012).