*The following supplement accompanies the article*

# Explicitly integrating a third dimension in marine species distribution modelling

**G. A. Duffy*, S. L. Chown**

*Corresponding author: grant.duffy@monash.edu

*Marine Ecology Progress Series 564: 1–8 (2017)*

---

## Methods for 2.5D-SDMs and figures

*Model code and simulated species example data are available at https://dx.doi.org/10.4225/03/58507ddae3022*

### Environmental data

Environmental data used were from NOAA's 2013 World Ocean Atlas (WOA13; Locarnini et al. 2013; Zweng et al. 2013 Garcia et al. 2014a; 2014b) and the Bio-ORACLE dataset (Tyberghein et al. 2012). Raster stacks for the region of interest were created (0.25 ° resolution) for each depth layer (0 m – 4900 m, 96 depth layers total) and contained mean at-depth temperature, salinity, and oxygen data from WOA13 and surface net primary productivity data from Bio-ORACLE, resampled to 0.25 ° resolution. Environmental data for all depth layers were tested for collinearity using variance inflation factor testing. No significant collinearity problems were identified (VIF values for all layers: temperature, 6.80; salinity, 7.27; oxygen, 1.29; NPP, 1.28).

### Occurrence records

All occurrence records for all species of Class Actinopterygii were downloaded from the Global Biodiversity Information Facility (GBIF; accessed 20/05/2015; doi:10.15468/dl.glqfjs) and geographically plotted. The North-East Atlantic (45 – 55 °N, 0 – 30 °W) was identified as the area with the highest concentration of points and therefore chosen as the region of interest for all subsequent analyses. Data were cleaned to remove conspicuous spatial errors and records not identified to species level were deleted. All occurrence points without depth data or with a depth value of 0 m (a default value that is likely erroneous as some species known to only have bathypelagic distributions had occurrence points with depth 0 m) were removed from the dataset. Species with ≥ 10 valid presence points were used for subsequent analyses.

Environmental data were extracted from raster stack data layers and associated with occurrence points based on the spatial coordinates of each point (latitude, longitude, depth). The depth field of each occurrence point was used to identify the nearest depth layers from which environmental data were estimated using linear interpolation. This created matrices of presence records of each species with associated environmental data.

### 2.5D-SDMs

Ensemble SDMs for four representative fish species, inhabiting various depths, were generated using occurrence matrices. Habitat suitability of each species was modelled through the application of ensemble SDMs (Thuiller, 2004). Ensemble modelling uses a suite of different SDMs to estimate habitat suitability by consensus. This provides a more robust methodology than using singular SDMs and reduces the impact of individual model biases. Ten different SDM methodologies contributed to the ensemble models created in this study (Generalized Linear Model, Generalized Additive Model, Boosted Regression Trees, Classification Tree Analysis, Artificial Neural Network, BIOCLIM, Flexible Discriminant Analysis, Multiple Adaptive Regression Splines, Random Forest, and Maximum Entropy). Each SDM was built using default settings using the 'biomod2' package (Thuiller et al. 2009). All models and outputs are for demonstrative purposes only.

Following recommendations by Barbet-Massin et al. (2012) 1000 pseudo-absence points, randomly selected from non-occupied 3D space were generated for every species and environmental data were also associated with these points using identical methods. Other methods for selection of pseudo-absence points are available (e.g. see discussions of Stokland et al. 2011; Barbet-Massin et al. 2012), though which method is most appropriate in 3D marine systems, where imperfect detection is problematic, is as yet undetermined and likely varies on a case-by-case basis depending on the species studied, the extent of the study area, and the confidence in occurrence data quality. The influence of pseudo-absence selection on 2.5D/3D-SDMs warrants further investigation. Presence and pseudo-absence matrices for each species were stacked with an 'occ' field identifying presence (1) or pseudo-absence (NA) data. These matrices were used to build 2.5D-SDMs.

Occurrence and pseudo-absence data were randomly divided with 70 % of data used for model calibration with the remaining 30 % reserved for model evaluation. Calibration and evaluation were repeated three times for each species with data randomly assigned to calibration or evaluation for each repetition. The true skills statistic (TSS), identified as the best metric of SDM performance (Allouche et al. 2006), was used to evaluate models. TSS values can range from -1 to +1 where a value of +1 indicates perfect model performance and a value $\leq 0$ indicates performance no better than random. To exclude poorly performing models, only models with a TSS score higher than 0.6 were used for subsequent ensemble model construction and projection to create maps of habitat suitability.

Ensemble models were projected onto raster stacks of all original depth layers (0 m – 4900 m, 96 depth layers total) to provide habitat suitability maps for all modelled species at all modelled depths across the region of interest.

**Figure 2**

A review of marine SDM literature was made using Google Scholar and all combinations of the search terms: 'marine', 'SDM', 'species distribution model', 'distribution model' 'maxent' 'maxent model', linked with 'OR' and 'AND' statements. All literature cited by or which cited any search result was also retrieved. Results were restricted to those studies published between 2005 and 2015 and only publications listed on the first 100 pages of each search were considered. A total of 564 unique publications were identified and downloaded. The first author reviewed all publications and those that applied SDM methods to organisms living at depths greater than 200 m were kept and the remaining publications were discarded. Fifty-five publications, which met the above criteria, were identified. These 55 publications were subsequently categorised based on the number of genera they modelled, the environmental data they used, and methods that the authors used (Figure 2; Table S1).

**Figure 3**

Mean (± 1 sd) water temperature of occurrence for each fish species. Temperature data were extracted from either a single surface layer (grey points) or through linear interpolation of the multiple depth layers of the WOA13 (black points) using geographic coordinates and depth of each occurrence point. Species ordered by mean depth of occurrence. See supporting table (S2) for full dataset and species ID numbers.

**Figure 4**

2D-SDM modelling was performed using the same occurrence records described above but with ensemble models trained and projected using only surface environmental data (temperature, salinity, oxygen concentration, and net primary productivity; i.e. assuming all species occurrence points were at 0 m). Representative species were selected based on mean depth occurrence and total number of occurrence records (i.e. species with many records and representing different depth distribution (shallow – abyssal). Distribution of occurrence points shown in Figure S1.

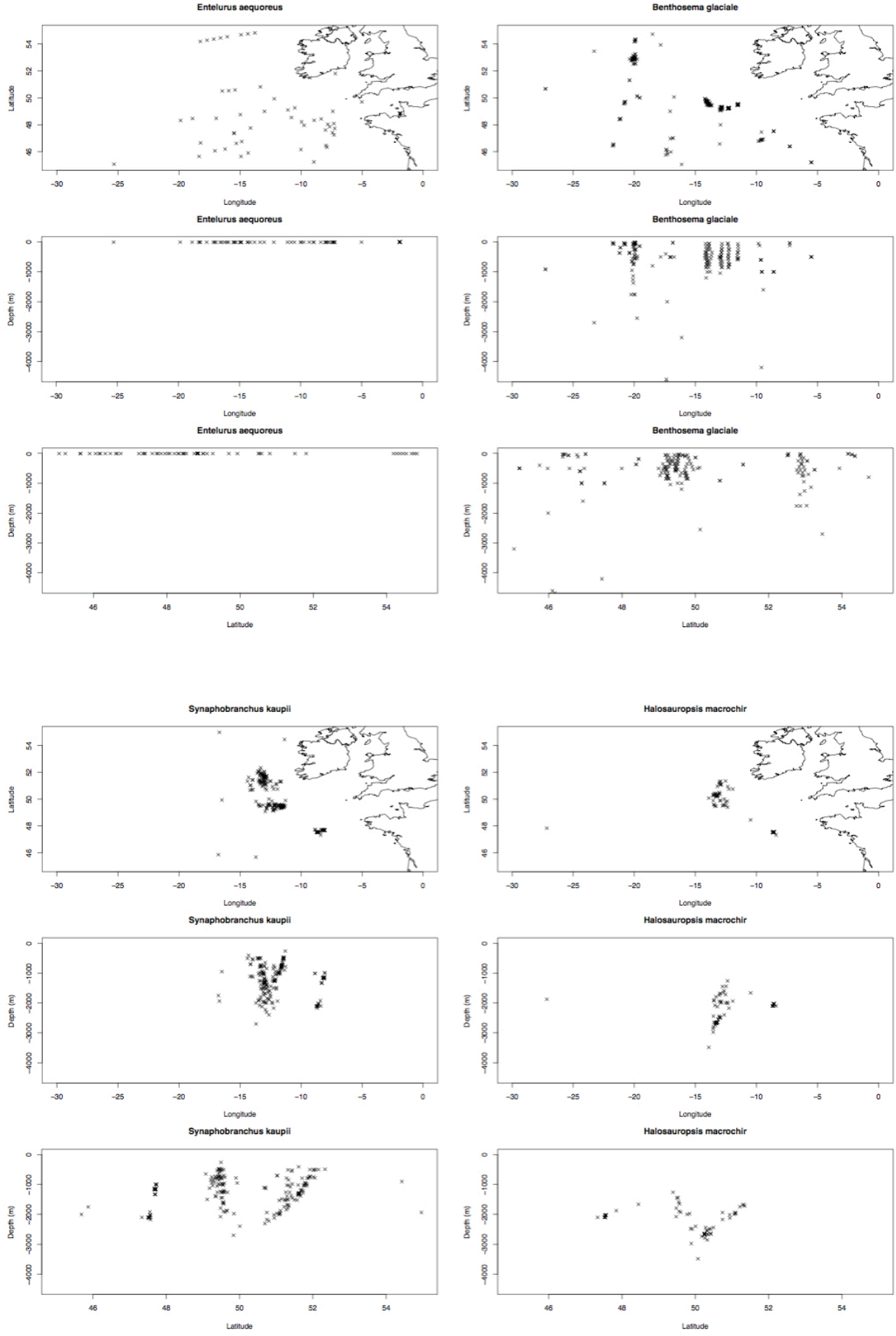


**Figure S1.** Distribution of occurrence points used for modelling for each of the four representative species. For an interactive rendering of points in 3D space see the attached html file at www.int-res.com/articles/suppl/m564p001_supp3.html.

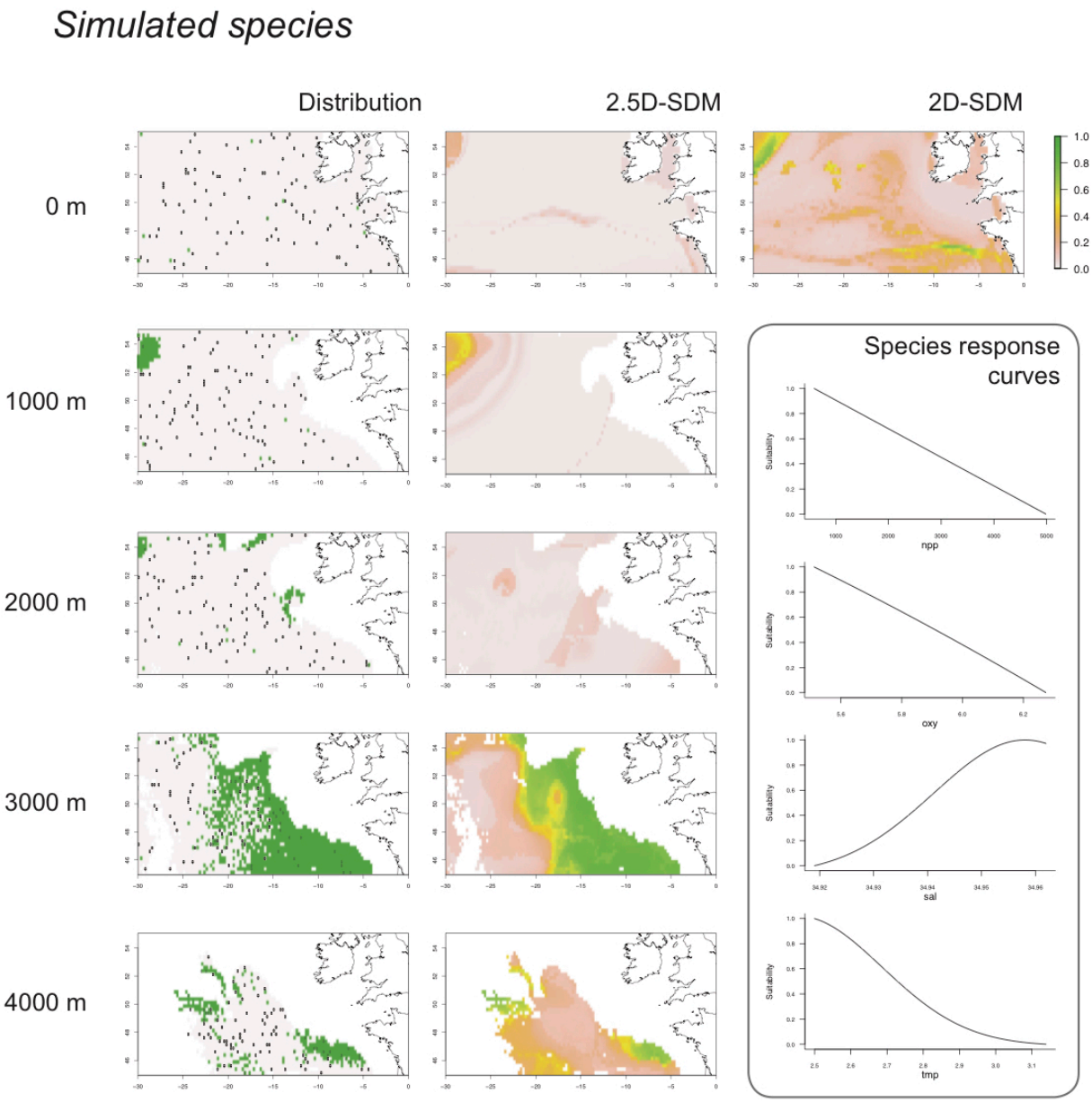


**Figure S2**

2.5D and 2D species distribution model probability of occurrence predictions (second and third columns respectively) for a simulated species compared to the 'true' distribution of the species (first column, green areas). A deep-water species was randomly generated using the 3000m depth layer (package virtualspecies:: generateRandomSp) and the species response curves (inset) for this species were then used to map the species distribution across all depth layers. From each of the depth layers, 50 random presence-absence points were sampled (indicated on the distribution plots) and these were used for all subsequent 2D and 2.5D species distribution modelling. The 2.5D-SDM was run using the methods described above (full code is included in SI) and incorporated the depth (z) from which an occurrence point was taken. The traditional 2D-SDM was run without considering depth and used only the xy-coordinates of each point and the uppermost depth-layer (0 m). A reduced TSS threshold (0.3 for 2D-SDM vs 0.6 for 2.5D-SDM) was also required so that at least two models passed TSS filtering to allow for projection.

All analyses were performed in R Statistical Software (R Core Team 2015) using the 'raster' (Hijmans & van Etten, 2015), 'biomod2' (Thuiller et al. 2009), 'ggplot2' (Wickham 2009), and 'virtualspecies' (Leroy et al. 2015) packages.

## LITERATURE CITED


Allouche O, Tsoar A, Kadmon R (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol 43:1223–1232

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012). Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol and Evol 3:327–338

Garcia, HE, Locarnini RA, Boyer TP, Antonov JI, Baranova OK, Zweng MM, Reagan JR, Johnson DR (2014) World Ocean Atlas 2013, volume 3: Dissolved oxygen, apparent oxygen utilization, and oxygen saturation. In: Levitus S, Mishonov A (eds) NOAA Atlas NESDIS 75. US Government Printing Office, Washington, DC

Garcia HE, Locarnini RA, Boyer TP, Antonov JI, Baranova OK, Zweng MM, Reagan JR, Johnson DR (2014) World Ocean Atlas 2013, volume 4: Dissolved inorganic nutrients (phosphate, nitrate, silicate). In: Levitus S, Mishonov A (eds) NOAA Atlas NESDIS 75. US Government Printing Office, Washington, DC

Hijmans RJ, van Etten J (2015) *raster: Geographic analysis and modeling with raster data*. http://CRAN.R-project.org/package=raster

Leroy B, Meynard CN, Bellard C, Courchamp F (2016) virtualspecies, an R package to generate virtual species distributions. Ecography 39: 599–607

Locarnini RA, Mishonov AV, Antonov JI, Boyer TP, Garcia HE, Baranova OK, Zweng MM, Paver CR, Reagan JR, Johnson DR, Hamilton M, Seidov D (2013) World Ocean Atlas 2013, volume 1: Temperature. In: Levitus S, Mishonov A (eds) NOAA Atlas NESDIS 75. US Government Printing Office, Washington, DC

R Core Team (2015) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Stokland JR, Halvorsen R, Støa B (2011) Species distribution modelling—Effect of design and sample size of pseudo-absence observations. Ecol Model 222:1800–1809

Thuiller W (2004) Patterns and uncertainties of species' range shifts under climate change. Glob Change Biol 10:2020–2027

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD - a platform for ensemble forecasting of species distributions. Ecography 32:369–373

Tyberghein L, Verbruggen H, Pauly K, Troupin C, Mineur F, De Clerck O (2011) Bio-ORACLE: a global environmental dataset for marine species distribution modelling. Glob Ecol Biogeogr 21:272–281

Wickham H (2009) *ggplot2: Elegant graphics for data analysis*. Springer New York

Zweng MM, Reagan JR, Antonov JI, Locarnini RA, Mishonov AV, Boyer TP, Garcia HE, Baranova OK, Johnson DR, Seidov D, Biddle MM (2013) World Ocean Atlas 2013, volume 2: Salinity. In: Levitus S, Mishonov A (eds) NOAA Atlas NESDIS 75. US Government Printing Office, Washington, DC