Supplementary Information

Use of openly available occurrence data to generate biodiversity maps within the South African EEZ

D Yemane, SP Kirkman and T Samaai

African Journal of Marine Science 2020, 42(1): 109–121

https://doi.org/10.2989/1814232X.2020.1737573

Appendix S1

To show both spatiotemporal pattern in sampling intensity and pattern in biodiversity (species richness), raw occurrence data for the three functional groups were compiled into 20-year periods, except for the early and present period. For the earliest period, all occurrence data prior to 1900 were lumped to construct the biodiversity maps. Similarly, for the present period all the data after 2000 were combined to construct biodiversity maps. The occurrences of the species in each functional group were aggregated in a grid with 10º × 10º resolution to generate the species richness by grid cell. The biodiversity maps for zooplankton, fishes, and benthos are presented in Figures S1, S2, and S3, respectively.

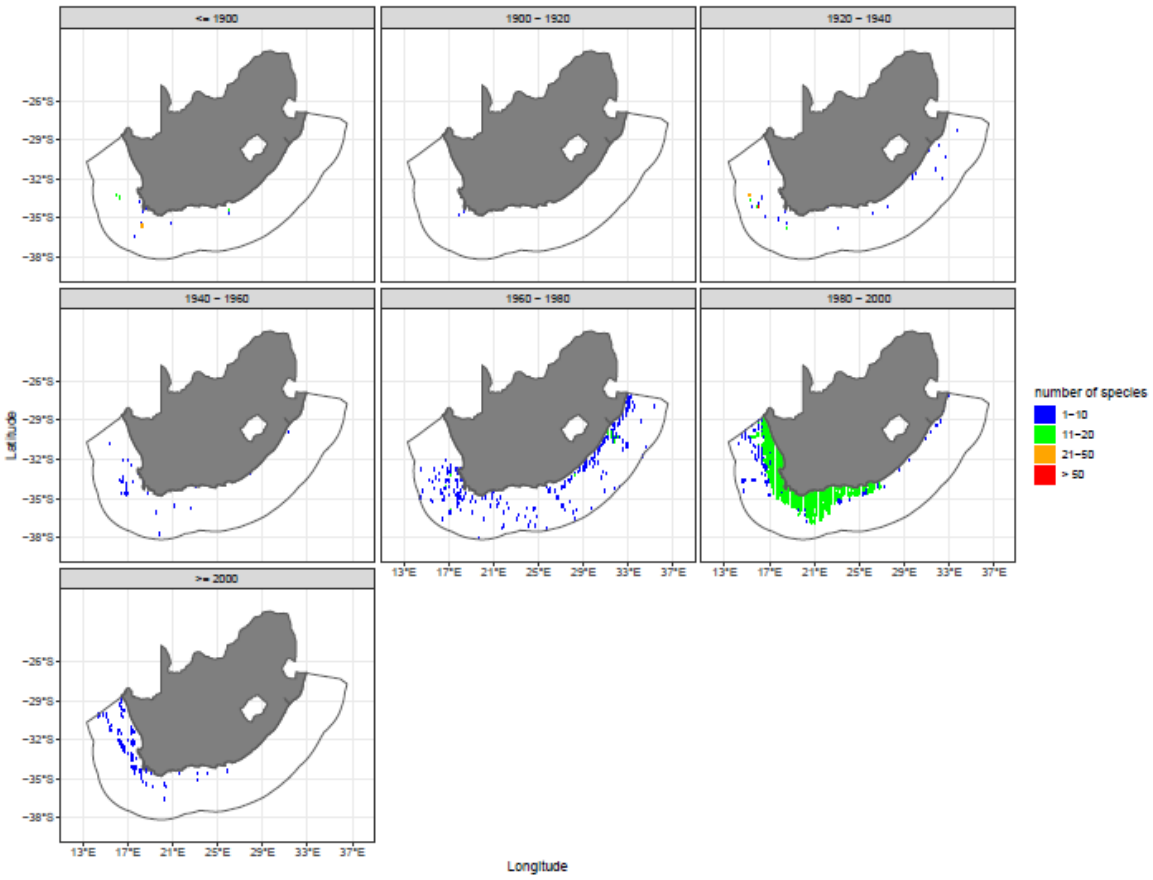

Figure S1: Zooplankton – summary of the distribution of sampling activity, and pattern in species richness, in the different periods

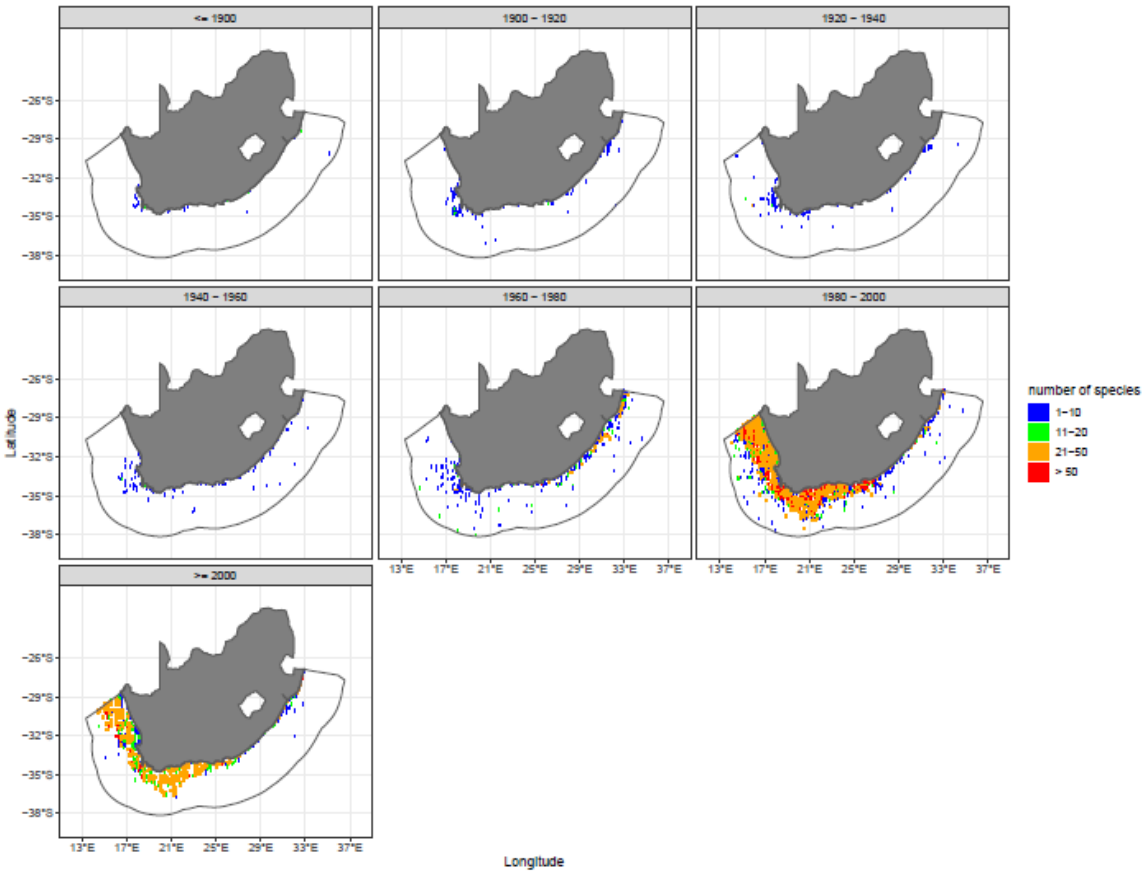

Figure S2: Fishes – summary of the distribution of sampling activity, and pattern in species richness, in the different periods

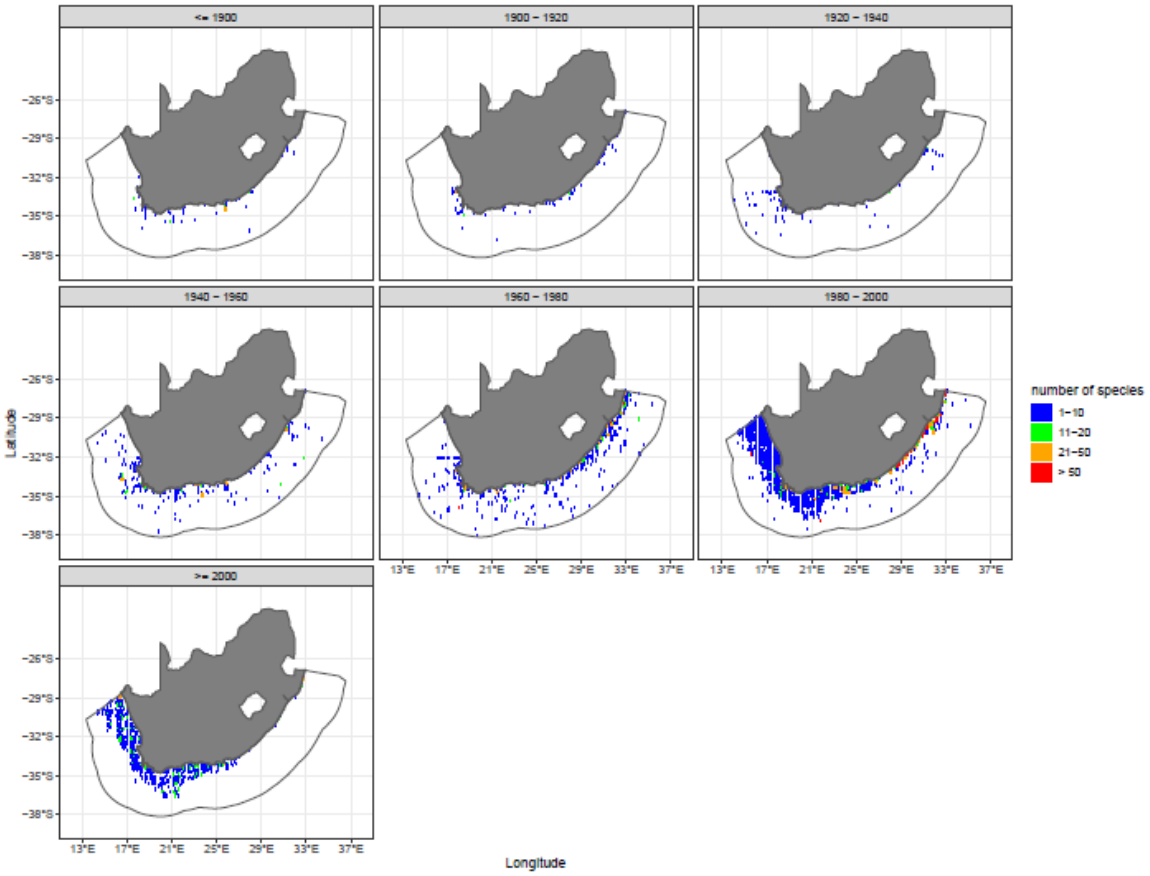

Figure S3: Benthos – summary of the distribution of sampling activity, and pattern in species richness, in the different periods

Appendix S2

Statistical models used for species distribution modelling (SDM)

There is a wide range of correlative statistical and machine-learning methods that are commonly used in SDM, but only five model types were considered for this study. These were: generalised linear models (GLM), multivariate adaptive regression spline (MARS), artificial neural network (ANN), random forest (RF), and classification tree analysis (CTA). The choice of these models was based on their minimum computational demands, given the need to run repeated model-fitting for more than 500 species, and because they are commonly used for SDM (e.g. Leathwick et al. 2006; Elith and Leathwick 2007; Shabani et al. 2016; Barbet-Massin et al. 2018). The fundamentals of the different models have been well-documented, and hence only a brief summary is provided below for each. A review by Norberg et al. (2019) of the performance of commonly used SDMs, including those considered in this study and other, emerging SDMs, found substantial differences in the performance of the models, especially in the modelling of community data with large numbers of rare species.

Generalised linear model (GLM)

GLMs are generalisations of ordinary linear regression that allow for response variables that have error-distribution models other than a normal distribution, such as the Poisson, binomial, Gamma, and quasi-Poisson distributions. The GLM generalises linear regression by allowing the response variable to be related to the linear model via a link function, and by allowing the magnitude of the variance of each measurement to be a function of its predicted value (James et al. 2013). The three main components of a GLMs are its error distribution, the linear predictor and the link function (Gerrard and Johnson 2015).

Classification tree analysis (CTA)

Classification tree analysis is used to model categorical/discrete response variables where either a prediction of observations in classes or the probability of belonging to classes can be obtained. In the context of species-distribution modelling, categories are presence/absence (Ramasubramanian and Singh 2017). Classification trees are grown by recursive binary splitting where splitting decision are made based on misclassification error rate. But two other measures that are recommended are the Gini index or entropy measure, as they are more sensitive to node impurity than is miss-classification error rate (James et al. 2013; Ramasubramanian and Singh 2017). CTA, as compared with standard linear regression-type models, allows one to implicitly model interaction among predictors and to model nonlinearity naturally.

Artificial neural network (ANN)

Artificial neural networks are modelling techniques inspired by how the brain is believed to process information and generate insight. Typical ANNs are composed of input, hidden, and output layers. Associations/connections between the different layers are determined by sets of mathematical models that pass information from one layer to the next. ANN can be considered a non-linear, or partially nonlinear, two-stage regression model (Azzalini and Scarpa 2012). Some of the strengths of ANN include: (i) its flexibility to approximate any regression functions; (ii)

its ability to estimate regression functions identified by limited numbers of components; and (iii) the fact that its estimated parameters can be updated with the arrival of new data. A limitation of ANN is that there is arbitrariness in the choice of numbers of hidden layers, instability in the estimation stage, and the lack of standard error estimates, which affects inference (Azzalini and Scarpa 2012).

Multivariate adaptive regression spline (MARS)

MARS is a nonparametric regression procedure that makes no assumption about the underlying functional relationship between the dependent and independent variables. Instead, MARS constructs this relation from a set of coefficients and basis functions that are entirely 'driven' from the regression data. MARS does not assume or impose any particular type or class of relationship (e.g. linear, logistic, etc.) between the predictor variables and the dependent (outcome) variable of interest. MARS is useful when there are many predictors and is able to model the effect of predictor(s) on multiple response variables simultaneously (e.g. in the context of this study, modelling the distribution of all species as the function of all the environmental variables, jointly) (Leathwick et al. 2006).

Random forest (RF)

The basis of RF is the standard decision tree but it improves upon the predictive accuracy of standard decision trees by building a large number of decision trees (from in the hundreds to a few thousands), based on training data. In addition, to remove correlation among trees, random samples of predictors are used when considering a split in a tree, and only one of these random subsets of predictors is used in a split. In comparison with other statistical methods, RF models are not prone to the problem of multicollinearity, implicitly allow for interactive effects, and can handle nonlinear effects (James et al. 2013).

References

Azzalini A, Scarpa B. 2012. *Data analysis and data mining: an introduction*. New York: OUP USA.

Barbet-Massin M, Rome Q, Villemant C, Courchamp F. 2018. Can species distribution models really predict the expansion of invasive species? *PLoS ONE* 13: e0193085.

Elith J, Leathwick J. 2007. Predicting species distributions from museum and herbarium records using multi-response models fitted with multivariate adaptive regression splines. *Diversity and Distributions* 13: 265–275.

Gerrard P, Johnson RM. 2015. *Mastering scientific computing with R*. Mumbai, India: Packt Publishing Ltd.

James G, Witten D, Hastie T, Tibshirani R. 2013. *An introduction to statistical learning*, vol. 112. New York: Springer.

Leathwick JR, Elith J, Hastie T. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling* 199: 188–196.

Norberg A, Abrego N, Blanchet FG, Adler FR, Anderson BJ, Anttila J et al. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs* 89: e01370.

Ramasubramanian K, Singh A. 2017. *Machine learning using R* (No. 1). New Delhi, India: Apress.

Appendix S3

For the sake of completeness and to serve as a comparison to the biodiversity maps presented in the Results, biodiversity maps constructed by ‘stacking binaries’ and ‘stacking probabilities’ are presented. Biodiversity maps for the three functional groups (zooplankton, benthos, and fishes) obtained using the two alternative methods are respectively shown in Figures S4 and S5.

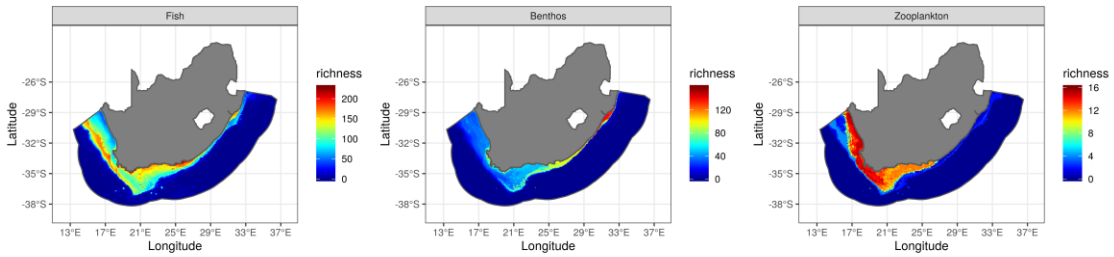

Figure S4: Overall pattern in biodiversity of zooplankton, benthos, and fishes based on the stacking-binaries method

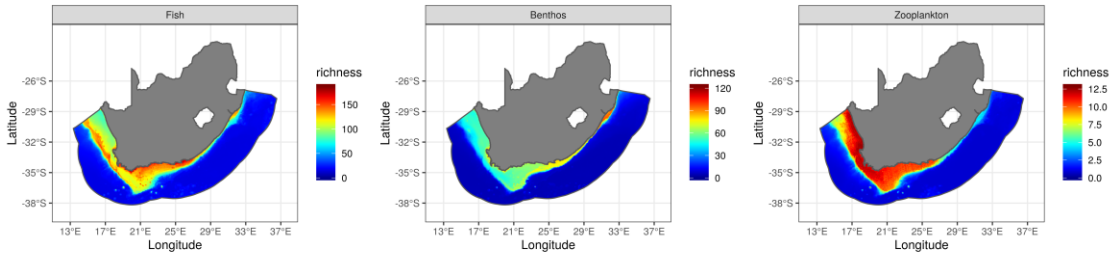

Figure S5: Overall pattern in biodiversity of zooplankton, benthos, and fishes based on the stacking-probabilities method

Appendix S4

Distribution maps of selected species from the three functional groups with observed occurrence data. For selected sets of fishes, benthos and zooplankton, distribution maps overlayed with the known occurrence are shown in Figures S6 and S7 for fishes, in Figures S8 and S9 for benthos, and in Figures S10 and S11 for zooplankton.

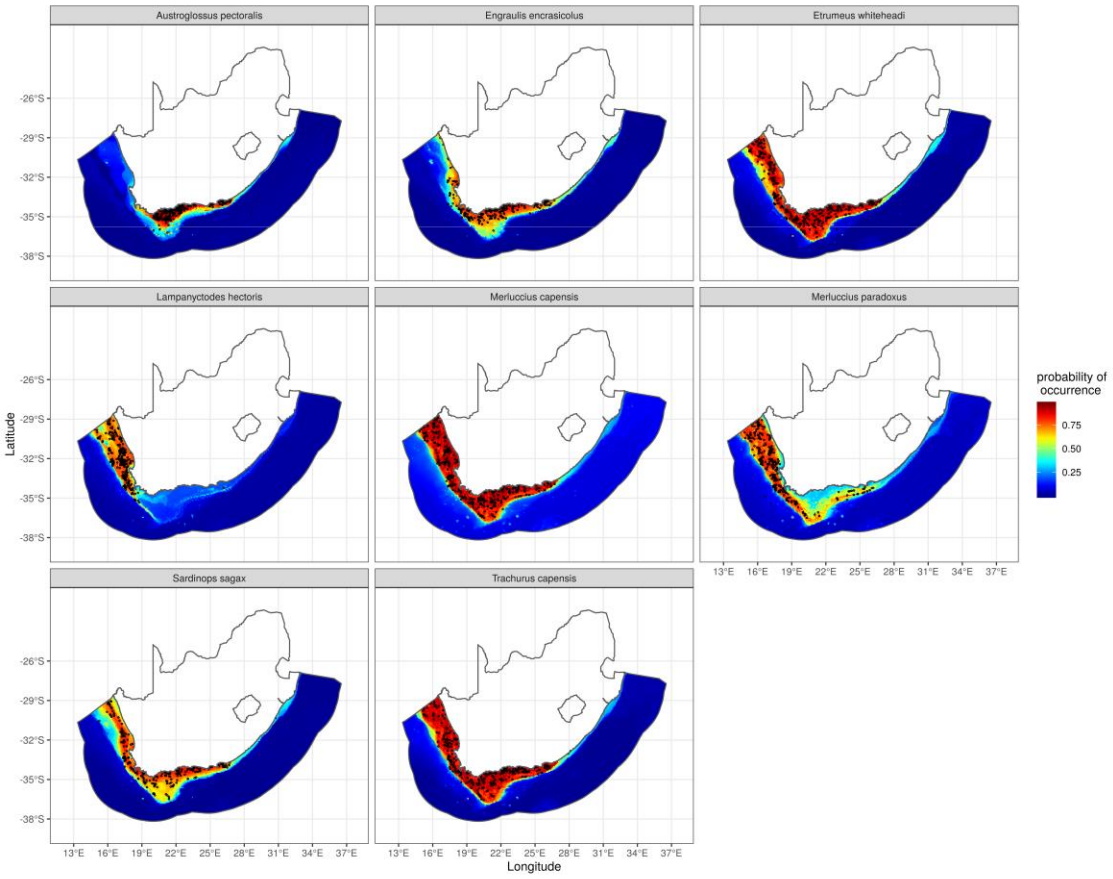

Figure S6: Fishes (Group1) – sample distribution maps for selected species. Open circles denote location of known presence

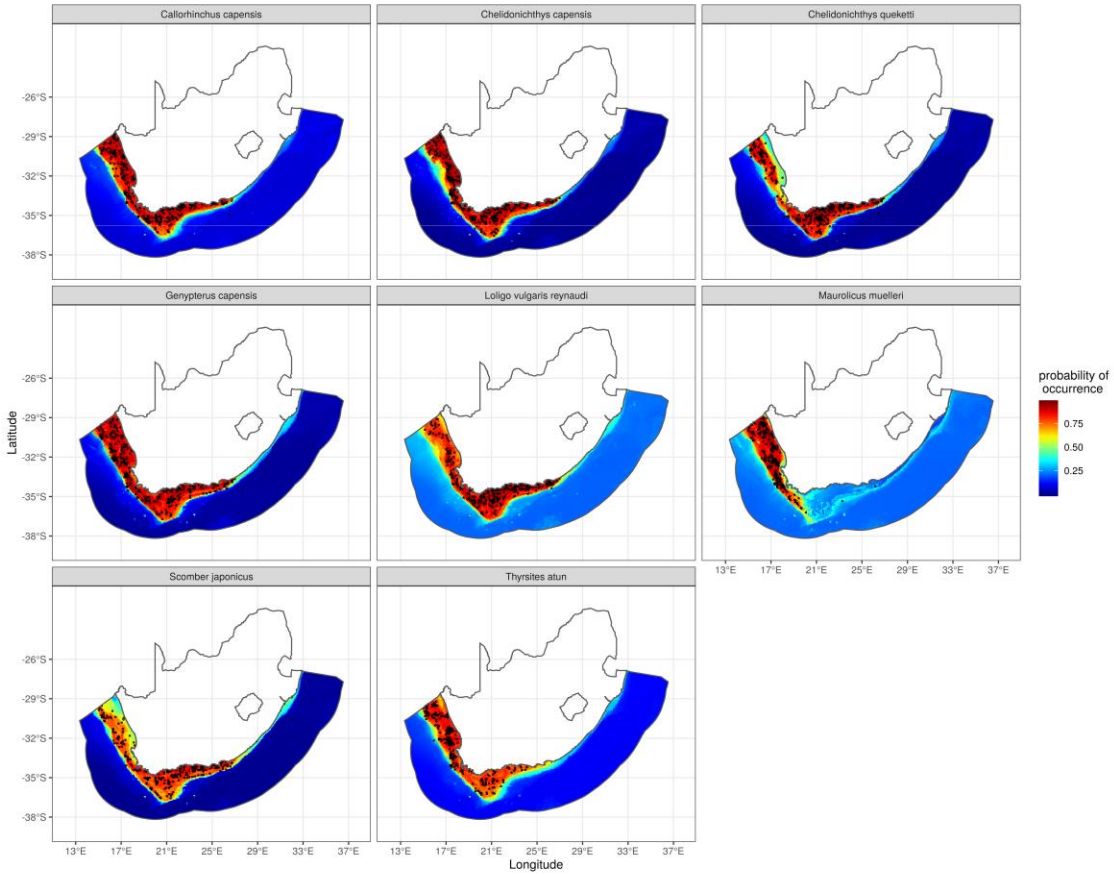

Figure S7: Fishes (Group 2) – Sample distribution maps for selected species. Open circles denote location of known presence

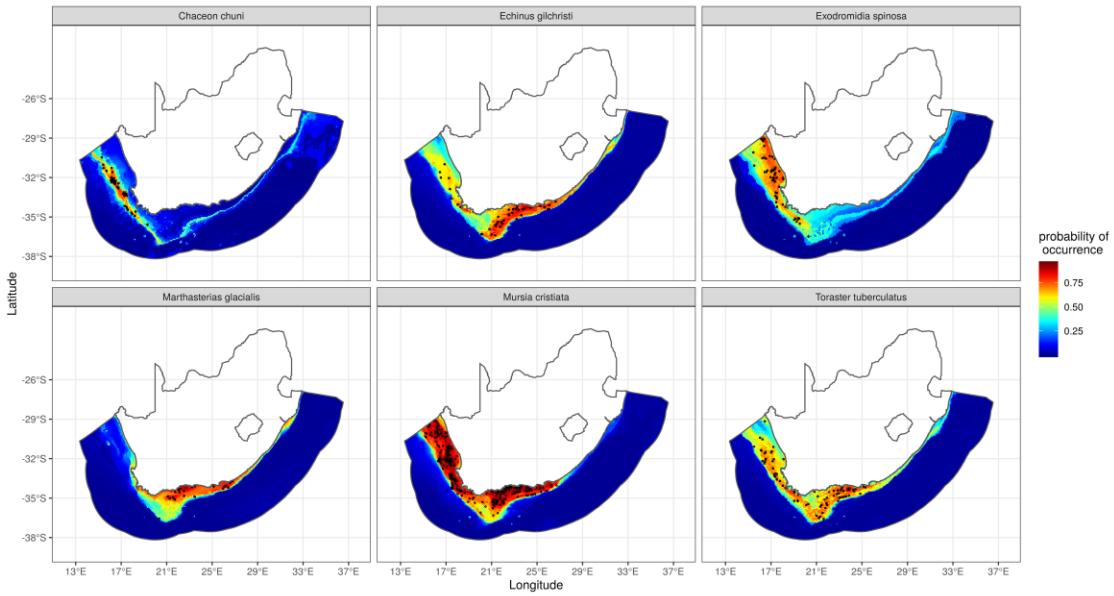

Figure S8: Benthos (Group 1) – sample distribution maps for selected species. Open circles denote location of known presence

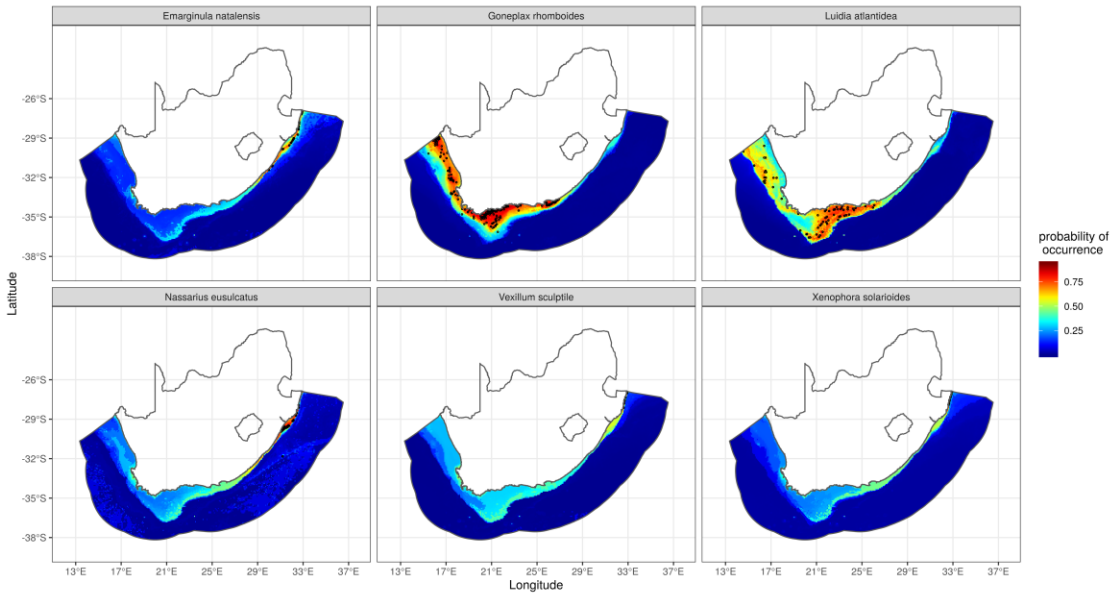

Figure S9: Benthos (Group 2) – sample distribution maps for selected species. Open circles denote location of known presence

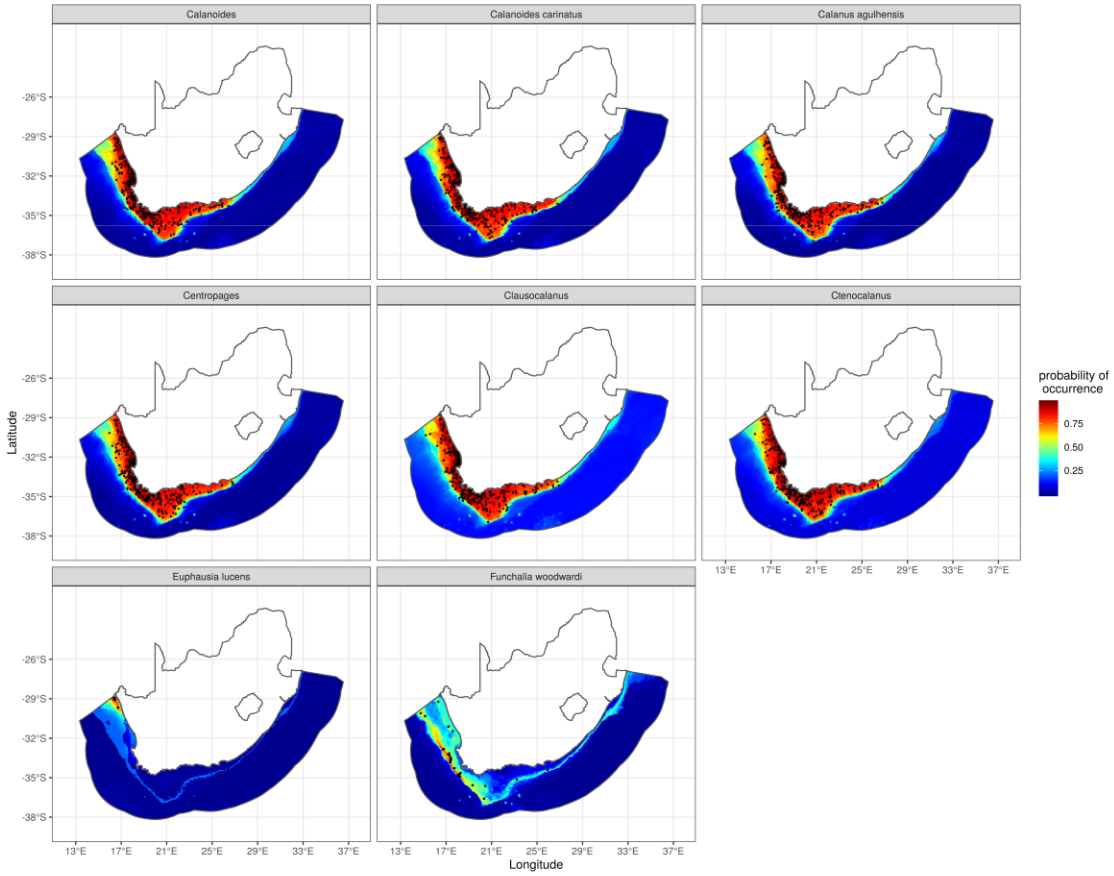

Figure S10: Zooplankton (Group 1) – distribution maps. Open circles denote location of known presence

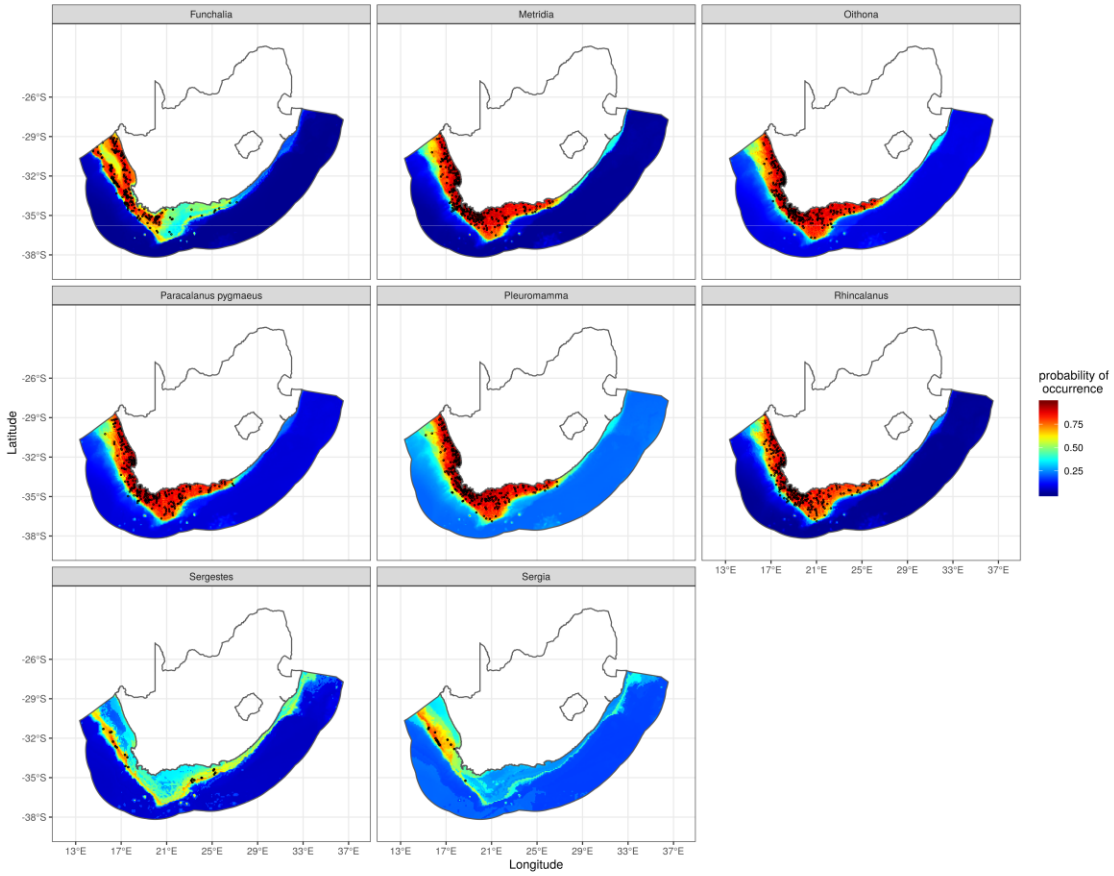

Figure S11: Zooplankton (Group 2) – distribution maps. Open circles denote location of known presence

Appendix S5

Species richness by selected classes: Actinopterygii, Cephalopoda and Elasmobranchii. These classes generally made up 98% of the species modelled.

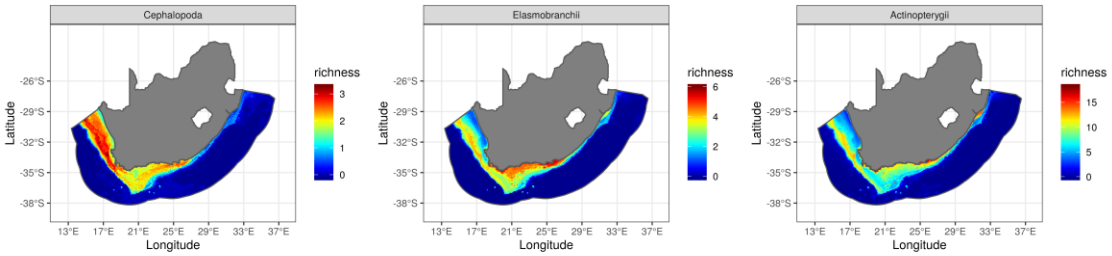

Figure S12: Pattern in biodiversity for the three major classes: Actinoptergyii, Cephalopoda, Elasmobranchii