

SUPPLEMENTARY MATERIAL S1.**MODELLING THE POTENTIAL DISTRIBUTION AND RICHNESS OF CETACEANS IN THE AZORES FROM FISHERIES OBSERVER PROGRAM DATA**

Marta Tobeña, Rui Prieto, Miguel Machete, Mónica A. Silva

Tuning of MaxEnt models

Model tuning is essential for reaching the optimal level of complexity that balances the best possible predictions without sacrificing generality (the ability to predict to different times or geographic spaces). Generality (or transferability) is normally addressed by penalizing models for complexity, which is dealt internally in MaxEnt by using L1 (lasso) regularization, also applied to generalized linear and generalized additive models (Dudík et al., 2007; Guisan et al., 2002). Different default regularization values are included in MaxEnt, according to the type of fitting functions and number of presences used, which were based in the best results from tests using a large dataset of 226 species from 6 regions (Phillips and Dudík, 2008). Warren and Seifert (2011) suggested ways of refining regularization in MaxEnt, namely by using the Bayesian (BIC) and sample size corrected Akaike (AICc) information criteria to compare competing MaxEnt models with different regularization values. However the method of implementation of that approach is not suitable for “species with data” (SWD) datasets, such as our own dataset (Warren, *pers. comm.*). Instead, we directly addressed two of the major issues affecting overall model performance and generality: overfitting and variable selection (Araújo and Guisan, 2006).

Overfitting is an often overlooked problem with methods capable of fitting complex responses, such as MaxEnt (Radosavljevic and Anderson, 2014). Overfitting will occur when the model fits the training data too closely, hampering the generalization ability of the model (Dudík et al., 2007; Schaffer, 1993). MaxEnt can use six classes of functions to fit models: linear (L), quadratic (Q), product (P), threshold (T), hinge (H), and category indicator (C) features (Phillips et al., 2006; Phillips and Dudík, 2008); when relying on MaxEnt default settings, models can easily become overfit as more functions

are added to the model (Anderson and Gonzalez Jr., 2011; Radosavljevic and Anderson, 2014; Warren and Seifert, 2011). In order to address overfitting, we ran test trials using MaxEnt default settings and inspected variable response curves for biologically nonsensical responses (ie, highly jagged or multimodal). We acknowledge that complex response curves may reflect actual ecological responses in some instances; however they may also be caused by excessive complexity of the models and lead to overfitting (Merow et al., 2013; Peterson et al., 2011; Phillips and Dudík, 2008). Model complexity must be constrained based on the attributes of the data and study objectives, favoring simpler models when in the presence of small sample sizes and strong sampling bias (Merow et al., 2013).

To avoid overfitting, we explored a range of settings for each cetacean taxon, limiting the number and type of functions used to fit the models in order to obtain smoother fitted functions (Elith et al., 2010). We created models for each taxon using all feature types and then sequentially reducing the number of features from the most complex to the least complex. We also created models with a single feature for all feature types. We then inspected the response curves for ecological plausibility (for example, by investigating if sea surface temperature responses were consistent with the known thermal preferences of the species), and compared the model performance based on the test-AUC values returned by MaxEnt, which measure the model ability to discriminate between randomly chosen presence and absence sites (background sites in the case of MaxEnt) (Fielding and Bell, 1997; Phillips et al., 2006). Models were chosen based on the best combination of their ecological plausibility and predictive power based on test-AUC scores. For most models this procedure lead to the choice for simpler models based only on the simpler Linear and Quadratic features, but the more complex Product feature was kept in some models (Table 3 of the main text). In two cases the best combination of model plausibility and predictive power was obtained using only the Hinge feature (Table 3 of the main text). We acknowledge that this approach is highly subjective as it depends on expert opinion for model selection. Recently, Shchegovitova and Anderson (2013) proposed a jack-knife approach for tuning models based on small sample sizes that offers an alternative for subjective model tuning. Those authors alert to some issues that should be considered before widespread utilization of the method, but it is a promising way for improved creation of environmental niche models.

Central to ecological modeling is the choice of ecologically relevant variables with good explanatory power (Araújo and Guisan, 2006). Stripping (pruning) models of uninformative or redundant variables can improve predictive performance and generality (Araújo and Guisan, 2006; Parolo et al., 2008) and is especially relevant when sample sizes are small in relation to the number of predictors considered (Breiner et al., 2015). We addressed this issue by first evaluating collinearity between explanatory variables and then using an interactive process to prune variables with low contribution to the models.

Collinearity between environmental variables was investigated by calculating the Kendall's correlation coefficient (τ) in R (R Development Core Team 2012). Kendall's correlation coefficient $|\tau| > 0.7$ was interpreted as indicating high collinearity between pairs of environmental variables (Dormann et al., 2013; Warren et al., 2013). We found no indication for high collinearity among variables and thus pilot models using all 18 candidate variables were built for each taxon and relative variable importance was evaluated using the 'permutation importance' (PI) scores reported by MaxEnt.

The PI metric works by randomizing the values for each environmental variable between presences and environmental samples, to make that variable uninformative, and then measuring the drop in the area under the receiver operating characteristic curve metric values (AUC). Large drops in AUC value are expected from variables that are important for the overall model quality, and vice-versa (Searcy and Schaffer, 2016). The PI has been shown to reflect actual aspects of a species' ecology, with high scoring variables playing an important role in defining the environmental niche of the species and low scoring variables having marginal importance in driving the species' distribution (Searcy and Schaffer, 2016). Thus this metric is useful in sorting relevant variables from variables of marginal importance (Searcy, *pers. comm.*).

Here variables with a PI score < 5 in the pilot models were considered to have little predictive power and were excluded, creating a new pruned set of variables (Hastie et al., 2009; Merow et al., 2013). Since unknown interactions between variables may be at play, eliminating variables may result in a change regarding the nature of those interactions and requires re-estimation of the functions for the other variables. Thus after a first pruning round, models were ran again and this pruning process was repeated

iteratively until no more variables with a PI score < 5 were detected. In practice, the process was repeated no more than once, since in most cases the first iteration was sufficient to eliminate all uninformative variables. In the few cases when an uninformative variable was not readily excluded in the first iteration, no more uninformative variables were left after a second iteration.

Final SDMs (SDM_f) as well as the test-SDMs (see main text for explanation) were built using only the variables retained for each taxon after this pruning process (Table 3 of the main text).

REFERENCES

- Anderson, R.P., Gonzalez Jr., I., 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. *Ecological Modelling* 222, 2796-2811.
- Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography* 33, 1677-1688.
- Breiner, F.T., Guisan, A., Bergamini, A., and Nobis, M.P. (2015). Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution* 6, 1210-1218. doi: 10.1111/2041-210x.12403.
- Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 027-046.
- Dudík, M., Phillips, S.J., Schapire, R.E., 2007. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research* 8, 1217-1260.
- Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. *Methods in Ecology and Evolution* 1, 330-342.
- Fielding, A.H., and Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24, 38-49.
- Guisan, A., Edwards Jr, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* 157, 89-100.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. *The elements of statistical learning* (Second edition), 2nd ed. Springer-Verlag, New York.
- Merow, C., Smith, M.J., Silander, J.A., 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* 36, 1058-1069.
- Parolo, G., Rossi, G., Ferrarini, A., 2008. Toward improved species niche modelling: *Arnica montana* in the Alps as a case study. *Journal of Applied Ecology* 45, 1410-1418.

- Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190, 231-259.
- Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* 31, 161-175.
- Radosavljevic, A., Anderson, R.P., 2014. Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography* 41, 629-643.
- Schaffer, C., 1993. Overfitting avoidance as bias. *Machine Learning* 10, 153-178.
- Searcy, C.A., Schaffer, H.B., 2016. Do ecological niche models accurately identify climatic determinants of species ranges? *The American Naturalist* 187, DOI: 10.1086/685387.
- Shcheglovitova, M., and Anderson, R.P. (2013). Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecological Modelling* 269, 9-17. doi: <http://dx.doi.org/10.1016/j.ecolmodel.2013.08.011>.
- Warren, D.L., Seifert, S.N., 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications* 21, 335-342.
- Warren, D.L., Wright, A.N., Seifert, S.N., Shaffer, H.B., 2013. Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. *Diversity and Distributions* 20, 334-343.