*Habitat suitability models*

Habitat suitability for selected AIS was modelled with a combination of MaxEnt v3.3.3k (Phillips et al 2006) and ’biomod2’ v3.4.6 (Thuiller et al., 2020) within R v3.6.3 (R Core Team, 2019). Different modelling techniques were used so that the results could be combined into an ensemble model as this method has been shown to reduce the biases that single models may have (Araújo & New, 2007).

MaxEnt is a machine learning method based on maximum entropy. It predicts the potential geographic distribution of suitable habitat for species using species occurrence data (presence-only) and various potential combinations of environmental data layers as input. MaxEnt is one of the most widely used SDM algorithms because of its high predictive accuracy and efficiency in modelling range shifts under future climate change scenarios (Bucklin et al., 2015; Elith et al., 2006; Elith et al., 2011; Hijmans & Graham, 2006; Pearson, 2007). In addition, it has recently been shown to outperform other modelling techniques to accurately predict the distributions of invasive species (Battini, Farías, Giachetti, Schwindt, & Bortolus, 2019).

The ‘biomod2’ package was used to run the following four models: Generalized Linear Model (GLM), Random Forest (RF), Artificial Neural Network (ANN), and BIOCLIM. These four models were chosen as they represent fundamentally different modelling techniques. More detailed information about each model may be found in Thuiller et al., (2009). We chose to use the ‘biomod2’ package here as it is one of the most well-known and widely used SDM tools (Hao et al. 2019). It is designed to allow researchers to run multiple different models on the same training and testing datasets in order to better ensure the accuracy of the resultant ensemble models it produces.

MaxEnt generates background data to compare with known presence points. This study used the default option, which generates 10,000 random background points. On the contrary, models from ‘biomod2’ require presence *and* absence data, hence it was necessary to create pseudo-absence values before running any of the models (Thuiller et al., 2020). For consistency with MaxEnt, 10,000 pseudo-absence points were selected randomly using the ‘biomod2’ default settings.

Model predictive power was evaluated using cross-validation with 70% of the occurrence points chosen randomly and used to train the model and the other 30% to test it for the complete set of models (Araújo, Pearson, Thuiller, & Erhard, 2005). In this validation procedure for MaxEnt, data were partitioned by a random process of *k*=500 iterations of training and validation (Hijmans, 2012). Both native and invaded ranges were used for training and evaluating all of the models (Verbruggen et al., 2013), since invaded areas provide valuable information on species’ tolerance to climatic conditions that may not be present in their native range (Marcelino & Verbruggen, 2015). Training points for MaxEnt were selected by random seeding with the convergence threshold set at 0.00001. The hinge feature while running MaxEnt was used as it produces complex yet smoothed and ecologically meaningful response curves and has been shown to improve model performance (Merow, Smith, & Silander, 2013; Phillips & Dudík, 2008).

For making the selection of predictors, initial models were run only with MaxEnt. This was done for each species individually with pre-selected environmental layers. The selected layers were those known to typically be the most important limiting factors for each type of taxa to have been used in other modelling studies (Table 1) (Barnes, 1999; Belanger et al., 2012; Chust et al., 2016; Cusson, Archambault, & Aitken, 2007; Gallardo et al., 2015; Jensen, Mousing, & Richardson, 2017; Leidenberger, Obst, et al., 2015; Wagner, 1977), or that were identified as being important by personal communications with taxonomic experts (G. Winkler pers. comm. 2017, and A. Rochon pers. comm. 2018). After the first run, layers with a relative contribution score of <4% were excluded (Jueterbock et al., 2016). Special attention was made to layers that were highly correlated (correlation coefficient ≥ 0.7) (Dormann et al., 2013), as correlated predictors can lead the model to produce erroneous response curves to layers that do not reflect species physiological tolerances (Marcelino & Verbruggen, 2015). Correlation was calculated using the SDMtoolbox (J. L. Brown, 2014) (Table S2) and only one such correlated predictor – that with the highest contribution to the modelling exercise – was retained for model construction. As per Goldsmit et al. (2018), predictors were identified by evaluating the combination of: a) the response curves for each species – to evaluate if the predictor behaves in a biologically meaningful way in the model (Marcelino & Verbruggen, 2015); b) a species-specific Jackknife test – to evaluate the contributions of the various parameters and analyse importance of variables; and c) the estimates of the contribution of each variable to model prediction. A minimum of three environmental layers were included in each model. Each species was then modelled again with the selected layers (Table S3). For MaxEnt, model performance was evaluated as the area under the curve (AUC ) and true skill statistic (TSS). TSS was used to evaluate the performance of the ‘biomod2’ models. In presence-only models, AUC is the probability that the model correctly ranks a random presence site versus a random site from the study area (Phillips et al., 2009). Values close to 1 indicate good prediction in site discrimination, while a value of 0.5 indicates a prediction equal to random and values lower than 0.5 indicate a performance that is worse than random predictions. TSS assesses the accuracy of predictions using sensitivity (proportion of correctly predicted presences) and specificity (proportion of correctly predicted absences) in its equation (TSS= sensitivity + specificity – 1). TSS is an appropriate evaluation alternative for model predictions converted to binary (presence–absence) maps using a threshold (Allouche, Tsoar, & Kadmon, 2006). It ranges from −1 to +1, where values between 0 and -1 indicate performance no better than random, while a statistically reliable model performance is indicated by values > 0.4, excellent models by a minimum of 0.7, and 1 indicates perfect agreement with the model (Allouche et al., 2006). No models with a TSS score under 0.7 were included in the final ensemble.

MaxEnt and the ‘biomod2’ models were run five times each.The continuous values produced in the models were transformed to binary values to identify predicted suitable and unsuitable habitat since continuous model projections may present some difficulties for interpretation. In addition, binary presence/absence maps are more useful for risk assessment exercises. For each species, the binary results for the five runs per modelwere averaged to create one binary result per model, with all five different model averagesthen averaged together to produce the final ensemble binary projection.This transformation was done using the maximum training sensitivity plus specificity threshold, which maximizes TSS values to create binary maps and it has been shown to produce the most accurate predictions (Jiménez-Valverde & Lobo, 2007; Liu, White, & Newell, 2013; Andrade et al 2020), and as it has been used in similar objectives (e.g. (Duffy et al 2017 ; Bellard et al 2013 ; Wisz et al 2015). All binary models were then delimited using a threshold for the maximum depth each species could inhabit according to their ecological requirements (Goldsmit et al., 2018) (Table S4). An exception was made for phytoplankton as they were all surface dinoflagellates; thus, there was no need to consider maximum depth for their distribution (A. Rochon pers. comm. 2018). Heat maps showing the total number of modelled AIS that may find suitable habitat in a region – hereafter AIS richness – were then created using combined maps representing the cumulative number of species (of the 23 modelled) predicted to find suitable habitat in a given grid cell at global and pan-Arctic scales. It should be kept in mind that, at a global scale, richness includes native and invaded ranges of all species modelled, while at pan-Arctic scale, richness includes mainly predicted invaded ranges.

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species dis-

tributions. Trends in Ecology & Evolution, 22, 42–47. https://doi.

org/10.1016/j.tree.2006.09.010.

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting

pseudo-absences for species distribution models: How, where and

how many? Methods in Ecology and Evolution, 3, 327–338. https://doi.

org/10.1111/j.2041-210X.2011.00172.x.

Bellard, C., Thuiller, W., Leroy, B., Genovesi, P., Bakkenes, M., & Courchamp,

F. (2013). Will climate change promote future invasions? Global Change

Biology, 19, 3740–3748. https://doi.org/10.1111/gcb.12344.

Duffy, G. A., Coetzee, B. W., Latombe, G., Akerman, A. H., McGeoch, M. A., & Chown, S. L. (2017). Barriers to globally invasive species are weakening across the Antarctic. *Diversity and Distributions*, *23*(9), 982-996.

Hao, T., Elith, J., Guillera‐Arroita, G., & Lahoz‐Monfort, J. J. (2019). A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Diversity and Distributions*, *25*(5), 839-852.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD - a platform for

ensemble forecasting of species distributions. Ecography, 32, 369–373.

Wilfried Thuiller, Damien Georges, Robin Engler and Frank Breiner (2020). biomod2: Ensemble Platform for Species Distribution Modeling. R package version 3.4.6. https://CRAN.R-project.org/package=biomod2