*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). Multiple modelling techniques were used so that the results could be combined into an ensemble model, which 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 that predicts the potential geographic distribution of suitable habitat for species using species occurrence (presence-only) data and various potential combinations of environmental data layers as input. This method was chosen because it is one of the most widely used SDM algorithms due to 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 et al., 2019). This was complemented with modelling using the biomod2 package, another well-known and widely used SDM tool (Hao et al., 2019) selected because it allows multiple models to be run on the same training and testing datasets to better ensure the consistency and accuracy of resultant ensemble models. A suite of four models were run with biomod2: Generalized Linear Model (GLM), Random Forest (RF), Artificial Neural Network (ANN), and BIOCLIM because each represents a fundamentally different modelling technique. More detailed information about each model may be found in Thuiller et al. (2009).

MaxEnt generates background data to compare with known presence points. This study used the default option, which generates 10,000 random background points. In contrast, 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 models, 10,000 pseudo-absence points were selected randomly using the biomod2 default settings.

Model predictive power, for MaxEnt and biomod2, 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 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 training and validation iterations (Hijmans, 2012). For biomod2 the data were partitioned randomly. Both native and invaded ranges were used for training and evaluating all models (Verbruggen et al., 2013) since invaded areas provide valuable information on species’ tolerances 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 in MaxEnt was used as it produces complex yet smoothed and ecologically meaningful response curves and has been shown to improve model performance (Merow et al., 2013; Phillips & Dudík, 2008).

To select 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 that have been used in other modelling studies (Table 1) (Barnes, 1999; Belanger et al., 2012; Chust et al., 2016; Cusson et al., 2007; Gallardo et al., 2015; Jensen et al., 2017; Leidenberger, et al., 2015; Wagner, 1977), or that were identified as being important based on consultation 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 given to layers that were highly correlated (correlation coefficient ≥ 0.7) (Dormann et al., 2013), as correlated predictors may lead the model to produce erroneous response curves to layers that do not reflect species physiological tolerances (Marcelino & Verbruggen, 2015). Correlations were tested using the SDMtoolbox (Brown, 2014) (Table S2) and a single 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 analyze importance of variables; and c) the estimates of the contribution of each variable to model prediction. A minimum of three environmental layers was 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 and …. 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 et al., 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). … Models with a TSS score of < 0.7 were excluded from the final ensemble.

MaxEnt and the biomod2 models were each run five times. The continuous values produced by 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 model were averaged to create a single binary result per model. Averages from the five models were then averaged to produce the final ensemble binary projection for each species. Binary transformation was done using the maximum training sensitivity plus specificity threshold, which maximizes TSS values to create binary maps and has been shown to produce the most accurate predictions (Andrade et al., 2020; Jiménez-Valverde & Lobo, 2007; Liu et al., 2013) in other studies with similar objectives (e.g. Bellard et al., 2013; Duffy et al., 2017; Wisz et al., 2015). Following transformation, all binary models were 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.