EUROPEAN JOURNAL OF ECOLOGY

EJE 2017, 3(1): 55-75, doi: 10.1515/eje-2017-0006

Biologically informed ecological niche models for an example pelagic, highly mobile species

Kate Ingenloff*

Biodiversity Institute, University of Kansas, Lawrence, KS, USA, Mailing Address: 1345 Jayhawk Blvd, Lawrence, KS, 66045, Phone: 785.864.4540
*Corresponding author, E-mail: kate.ingenloff@ku.edu

ABSTRACT

Background: Although pelagic seabirds are broadly recognised as indicators of the health of marine systems, numerous gaps exist in knowledge of their at-sea distributions at the species level. These gaps have profound negative impacts on the robustness of marine conservation policies. Correlative modelling techniques have provided some information, but few studies have explored model development for non-breeding pelagic seabirds. Here, I present a first phase in developing robust niche models for highly mobile species as a baseline for further development.
Methodology: Using observational data from a 12-year time period, 217 unique model parameterisations across three correlative modelling algorithms (boosted regression trees, Maxent and minimum volume ellipsoids) were tested in a time-averaged approach for their ability to recreate the at-sea distribution of non-breeding Wandering Albatrosses (Diomedea exulans) to provide a baseline for further development.
Principle Findings/Results: Overall, minimum volume ellipsoids outperformed both boosted regression trees and Maxent. However, whilst the latter two algorithms generally overfit the data, minimum volume ellipsoids tended to underfit the data. Conclusions: The results of this exercise suggest a necessary evolution in how correlative modelling for highly mobile species such as pelagic seabirds should be approached. These insights are crucial for understanding seabird–environment interactions at macroscales, which can facilitate the ability to address population declines and inform effective marine conservation policy in the wake of rapid global change.

KEYWORDS

Boosted regression trees; digital accessible knowledge; distribution modelling; Maxent; minimum volume ellipsoids; pelagic seabird distribution; *Diomedea exulans*

INTRODUCTION

Impacts of global change are increasingly evident, and long-term changes in marine systems are likely to be quite profound (Doney et al. 2012). In spite of these changes, spatial planning and conservation implementation in marine systems are lagging compared to terrestrial regimes (Game et al. 2009; Croxall et al. 2012; Lewison et al. 2012). Of particular concern are pelagic zones, which currently lack adequate protection compared to other marine ecoregions. Seabirds and other marine predators can serve as proxies to help identify potential marine conservation sites (Piatt et al. 2007; Lascelles et al. 2012).

The strong spatio-temporal heterogeneity inherent in marine systems (Hyrenbach et al. 2000; Weimerskirch et al. 2005) is mirrored in the movements and behaviour of pelagic seabird species tracking marine resources. Many pelagic seabird species appear to behave as generalists overall, whilst maintaining individual- or group-level specialisations (Ceia et al. 2012). In view of their complex behavioural biology (Grecian et al. 2012; Catry et al. 2013), such as high individual variability (Phillips et al. 2005), complex habitat partitioning and movement strategies (Phillips et al. 2005; Weimerskirch et al. 2006), and colonial nesting habits, available data for many of these species are highly biased towards breeding individuals. This information bias has left gaps in knowledge about at-sea distributions of non-breeding individuals (Weimerskirch et al. 2003; Taylor et al. 2011; Grecian et al. 2012; Lascelles et al. 2012).

Correlative niche modelling approaches, termed species distribution modelling (SDM) or ecological niche modelling (ENM), have the potential to fill knowledge gaps regarding species' distributions (Rodríguez et al. 2007; Lewison et al. 2012; Mateo et al. 2013), aid in conservation planning (Peterson 2006; Rodríguez et al. 2007), assess conservation-human conflicts (Rodríguez et al. 2007), and provide insight into impacts of climate change on species' distributions (Peterson 2006). Ongoing conservation concerns regarding pelagic seabirds make them an important focus group in such studies. To date, however, nearly all applications of these approaches to pelagic seabirds have focused on individual populations rather than species as a whole; and, few have explored algorithm function for seabirds outside the breeding season (Thiebot et al. 2011; Wakefield et al. 2011; Oppel et al. 2012; Ramos et al. 2015).

Wandering Albatrosses (*Diomedea exulans* Linnaeus, 1758) are biennial breeders (Prince et al. 1992; Milot et al. 2008) with multiple life stages (juvenile, immature, non-breeding adult, breeding adult) marked by distinct behaviours (Phillips et al. 2005; Ceia et al. 2012). Classified as Vulnerable (IUCN 2016), they are protected under the Agreement on the Conservation of Albatrosses and Petrels (ACAP) and are amongst the best-studied pelagic seabirds (ACAP 2009). Because occurrence data for this species are relatively rich, gaps in knowledge of their natural history are less drastic than in other pelagic species. This is a critical consideration for application to developing and testing methodological improvements. Studies have already noted impacts of global climate change on *D. exulans* and other pelagic marine species (Weimerskirch et al. 2003, 2012).

The aim of this study is to identify hemisphere-scale environmental associations of the geographic distribution of non-breeding *D. exulans* and work towards addressing the challenge of modelling these associations in highly mobile species. I present results of a first phase of correlational ecological niche modelling using traditional modelling techniques based on three algorithms and multiple parameterisations; I assessed each models' ability to anticipate seasonal environmental preferences of non-breeding *D. exulans*. This initial exploration focused on issues of algorithm selection and parameterisation in time-averaged correlative modelling.

1. MATERIALS AND METHODS

1.1. Input Data

Models were calibrated using digitally accessible knowledge (DAK; Sousa-Baena et al. 2014) in the form of *D. exulans* primary occurrence data and remotely sensed environmental data for December 2000 to November 2012. As temporal averaging of models may generalise spatial distributions and environmental associations (Peterson et al. 2005), the year was divided into three 'seasons' for this study (Table 1) based loosely on breeding biology and designed to respond to possible shifts in foraging behaviour by breeding adults: I = December–March (egg laying/incubation), G = April–July (brood guard/chick rearing) and P = August–November (fledging). The study area was restricted to −20˚S to −60˚S latitude, as this latitudinal range comfortably encompasses the generalised distributional extent of *D. exulans* (BirdLife International and NatureServe 2015a), reduces concern for significant gaps in environmental data coverage – particularly in seasonal, high-latitude regions – and constrains the extent to which implemented modelling algorithms must extrapolate.

To characterise the sampling process that produced the data (Anderson 2003), observation- and specimen-based occurrence data for all members of the order Procellariiformes were obtained from the Global Biodiversity Information Facility (GBIF; accessed on 26 May 2015, doi:10.15468/dl.fquf8g; Appendix S1). *D. exulans* observation data were separated from the greater dataset and divided by season (see above), cleaned of duplicates, gridded to the spatial resolution and extent of the environmental data and rarefied to one point per pixel to reduce spatial bias (Kramer-Schadt et al. 2013; Phillips et al. 2009). As no information about sex or breeding status was associated with the occurrence data, distinguishing non-breeding from breeding individuals was impossible. However, because these analyses aimed to assess the capacity for estimating non-breeding distributions, occurrence data south of 50˚S latitude were excluded from analyses (Weimerskirch et al. 1985; Weimerskirch et al. 2006). Of an initial 7,903 *D. exulans* records, 1,982 were available for use in modelling after cleaning; 136,947 records of the Order were used to characterise spatial sampling bias (see below).

Seven environmental layers were used to summarise the complex environmental landscape of the high-latitude marine systems under analysis. *Dynamic data* included four monthly variables of global MODIS Terra L3 SMI data at 4.6 km spatial resolution downloaded from the NASA OceanColor Web (Table S1.2; NASA 2014). Daytime and nighttime sea surface temperatures (SST) were used to average uneven heating/cooling of the ocean surface. Chlorophyll-*a* (Hyrenbach et al. 2007; Wakefield et al. 2009) and chromophoric dissolved organic matter (Coble 2007; Nelson & Siegel 2013; Urtizberea et

Table 1. Delineation of seasons for time-averaged correlative models, associated breeding stage and the number of Diomedea exulans *occurrence data used in model calibration (Calibration), model calibration testing (Calibration testing) and testing after model transfer (Projection testing).*

			D. exulans observation data			
Seasons	Period	Breeding stage	Calibration	Calibration testing	Projection testing	Total
I	December–March	Egg laying/incubation	553	239	269	1,061
G	April–July	Brood guard/chick rearing	281	121	185	587
P	August–November	Fledging	140	60	130	330

al. 2013) were incorporated as proxies for ocean productivity. Imagery were converted from native HDF to ASCII grids, projected to WGS 84 using the Marine Geospatial Ecology Tools (MGET) ArcGIS toolbox (Roberts et al. 2010), and 'No Data' values in raster layers were filled using a temporal filter followed by a spatial filter in R v 3.2.2 (R Development Core Team 2009). Next, environmental data layers were stacked by season, and the mean, maximum, minimum and range of values were calculated for each variable. The resulting 16 time-averaged rasters were subjected to principle component analyses (PCA) to reduce collinearity. The first five principle components (PCs) from each PCA per season were used in the analyses; in all three seasons, the first PC explained ≥95% of variation (Table S1.3). *Geophysical (static) data* included bathymetry – ETOPO1 global relief data (Amante 2009) – and a derivative bathymetric slope layer. All seven environmental layers were standardised to 0.2083° resolution and projected in geographic coordinates (WGS 84). Additional information regarding input data is available in Appendix S1.

1.2. Model Calibration

The biotic–abiotic–mobility (BAM) framework is a useful heuristic for developing strategies for model calibration (Soberón & Peterson 2005). The calibration region should match the mobility area (= the area that has been accessible to the species over relevant periods of time; Barve et al. 2011). Mobility is not a major distributional constraint for *D. exulans* (Soberón & Peterson 2005; Milot et al. 2008; Saupe et al. 2012). As such, calibration regions were delineated as marine areas within a 500 km buffer around known occurrences in a particular season (Fig. 1; Barve et al. 2011; Saupe et al. 2012). To permit rigorous model evaluation, 30% of occurrence records were selected randomly and set aside for model evaluation. Models were calibrated using the remaining 70% of occurrence records.

A total of 217 model calibrations were tested for each of the three time-averaged seasons across three correlative niche modelling algorithms, yielding 651 models following the 'no silver bullet' ideas of Qiao et al. (2015), in which many candidate approaches and algorithms are tested to identify the best-performing method for a particular situation. Two presence-only algorithms – Maxent (Phillips et al. 2004; Phillips et al. 2006) and minimum volume ellipsoids (MVEs) – and one presence–absence algorithm – boosted regression trees (BRT; Elith et al. 2008) – were selected for testing.

1.2.1 Presence-only algorithms

Maxent version 3.3.3k (Phillips et al. 2004; Phillips et al. 2006) was calibrated under different settings for three parameters: prevalence, regularisation multiplier and bias layer. Initial sensitivity analyses using the jackknife procedure within Maxent identified an ideal combination of environmental variables for model calibration (bathymetry, PCs 1–4). All models were run using 100 bootstrapped replicates, 30% random test percentage and 1,000 maximum iterations; all other settings remained at 'default'.

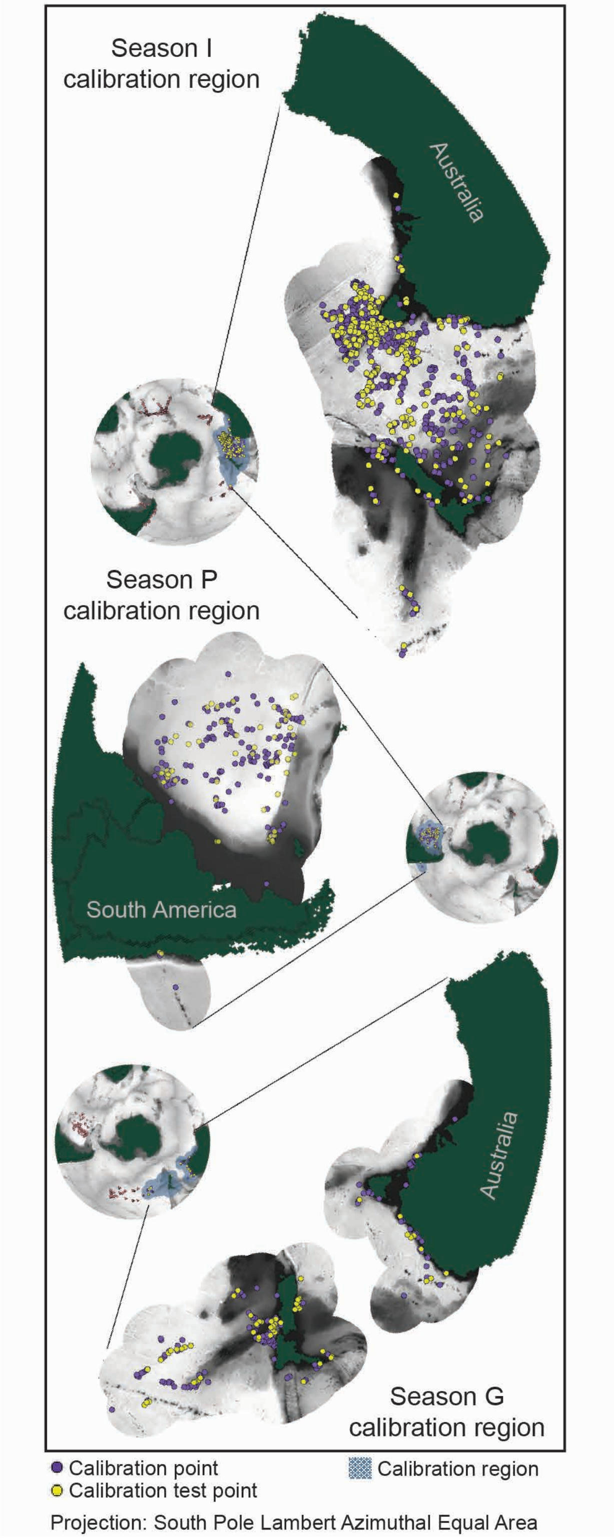


Figure 1. Model calibration regions for seasons I (December–March), P (April–July) and G (August–November). Base layers: ETOPO1 global relief data (Amante 2009) and Global Administrative Areas global shapefile (http://www.gadm.org).

Prevalence was tested over a range of 0.3–0.9 at intervals of 0.1. Prevalence has no impact on raw output scores in Maxent but does affect the 'logistic' output (Elith et al. 2011;

Merow et al. 2013); Elith et al. (2011) and Merow et al. (2013) provide an in-depth discussion on the impact of prevalence on model performance. The regularisation multiplier (RM) impacts model fit by loosening or tightening the constraints of a model around the training data (Elith et al. 2011; Shcheglovitova & Anderson 2013). RM was tested at three levels: 1 (default), 1.5 and 2. Bias layers are incorporated into Maxent to account for sampling bias in the data and reflect relative sampling effort (Phillips et al. 2009; Kramer-Schadt et al. 2013). Two sets of bias layers derived from the procellarid occurrence data set aside during data cleaning and matching the grid system of the environmental grids were tested. Procellariiform observations per pixel were summed to produce the 'raw' bias layer. To develop a more refined layer for comparison, the raw bias layer was subjected to a Log_2 transformation and kernel smoother to scale the value distribution more evenly (Table S1.4).

MVEs were calibrated for two sets of parameters: variable inclusion and threshold. MVEs calculate environmental distance using Mahalanobis distances based on an MVE drawn around the training (calibration) data. The simplicity of MVEs means few parameters. Six levels of variable inclusion (Runs; Table S1.5) and three thresholds (T; 0.9, 0.95, 0.99) were analysed using R v 3.2.2 (R Development Core Team 2009). The threshold designates the central percentage of training data to be used in calculating the MVE, such that a higher threshold (e.g. 0.99) indicates greater confidence in the input data used for training compared to a lower threshold value (0.95 or 0.9). Scripts were modified from code provided by J. Soberón (*pers. comm*.) and are available in Supplementary Materials (Appendix S3). As MVEs calculate relative environmental distance, model predictions were inverted and re-scaled (0–1) to render them comparable to the other algorithms.

1.2.2 Presence–absence algorithm

Boosted regression trees (BRTs) were calibrated under various settings of four parameters: pseudo-absence (PA), tree complexity (TC), learning rate (LR) and bag fraction (BF). Two levels of PAs were tested after Barbet-Massin et al. (2012) (Table S1.6): the first (PA-1) was set at 1,500 randomly selected absences; the second (PA-2) was two times the total number of model calibration training points. TC, which controls the maximum level of interactions permitted in model fitting, was tested over a range of 1–5, wherein TC = 1 indicates no variable interactions and TC = 5 permits interactions between ≤5 variables. Four LRs were tested: 0.01, 0.005, 0.0025 and 0.001. LR determines the relative contribution of each tree as the model grows, such that a slower LR tends to smooth effects of stochastic processes and reduces between-model variance. BF controls stochasticity by designating the random subset used for model calibration and testing; a smaller BF is likely to lead to an increase in the chance of fitting of unusual variables. Four levels of BF were tested: 0.5, 0.6, 0.7 and 0.75. More in-depth explanations of BRTs are provided by Elith et al. (2008). Models were run to a minimum of 1,000 trees using all seven environmental variables in R v 3.2.2 (R Development Core Team 2009) following scripts from Elith et al. (2008).

1.3. Model evaluation

Significance was evaluated for all model calibration and model transfer scenarios. In light of issues highlighted by Peterson et al. (2008) and Lobo et al. (2008), typical receiver operating characteristic (ROC) routines implemented within Maxent were not used. Instead, models were evaluated external to the Maxent package using the partial receiver operating characteristic (pROC) metric, for which the critical value is $AUC_{ratio} = 1.0$. pROC scores were calculated using the randomly selected 30% test data set aside prior to model calibration (see above) and occurrence data from the broader projection region, to provide two levels of testing. pROC scores were calculated in R v 3.2.2 (R Development Core Team 2009) at an omission threshold of E = 5% over 2,000 iterations using code provided by L. Osorio (*pers. comm.*). Significance was determined by direct count of numbers of replicate analyses in which $AUC_{ratio} \leq 1$. Although AUC ratios are difficult to compare amongst very different calibration areas or modelling contexts, they can be used to assess within-algorithm, within-season performance (e.g. to evaluate the best performing model calibration for an individual algorithm).

Final model performance was assessed using two metrics to permit comparison of models across algorithms. The first metric was omission rate. Omission rates (percent of test data predicted as 'absent') were calculated using the *D. exulans* observation data set aside and an 80% threshold (e.g. E = 20%). As a second measure of performance, BirdLife International's important bird area (IBA) polygons for the Southern Hemisphere (BirdLife International and NatureServe 2016) were overlaid on the best model for each algorithm and season to evaluate visually the ability of each model to anticipate areas of known importance to *D. exulans*. A query of BirdLife International's marine e-atlas (BirdLife International and NatureServe 2015b) facilitated generation of a subset of 130 of the original 1,275 polygons identified specifically as valuable to non-breeding *D. exulans* and classified as proposed or confirmed IBA areas. It is critical to note that these designations are based on limited data (e.g. a handful of tracking datasets) and do not necessarily encompass all areas of importance to non-breeding *D. exulans*. They do, however, provide a simple, qualitative view of model performance, thus their use in model evaluation here is considered secondary and supplemental to calculated omission rates.

2. RESULTS

2.1. Significance Testing

All 651 model calibration scenarios were significant ($p < 0.05$). In model transfer, only 52.7% (343 of 651) of models (across all combinations of algorithms, parameter settings and environmental datasets) performed statistically significantly better than random. All 54 MVE models (18 per season) were significant. Of the 480 BRT models (160 per season), 36.3% (58) were significant in season I, 93.8% (150) in season G and 11.9% (19) models in season P. And, of the 117 Maxent models (39 per season), 94.9% (37) were significant in season G and 64.1% (25) in season P. None of the Maxent models transferred in season I were significant. Results from the top-performing model for season I are presented for comparison.

2.2. Overall Model Performance

MVEs outperformed both Maxent and BRTs in all three seasons for both model evaluation metrics (Figure 2). MVEs thresholded at 0.9 and run = 3 (all variables included except bathymetry and bathymetry slope) yielded the best models in seasons I (December–March) and G (August–November), and MVEs thresholded at 0.9 and run = 1 (all variables included) yielded the best models in season P (April–July). Model projections following an 80% threshold for the top models produced by each algorithm are presented in Figures 3 (season G) and 4 (seasons I and P). To provide a more detailed view of model predictions relative to occurrence data and IBAs, three particularly well-sampled focus regions are shown: the waters surrounding New Zealand and Australia (Figs S2.1–S2.3), the vicinity of the sub-Antarctic Islands off South America (Fig. S2.4) and the vicinity of the sub-Antarctic Islands near southern Africa (Fig. S2.5). The top five model projections for each algorithm are summarised by season in Table S2.1

2.2.1 Model Calibration

In model calibration, MVE had the lowest omission rates in seasons G (16.5%) and P (0%). Maxent had the lowest omission rate in season I (24.7%).

2.2.2 Model Transfer

Though overall model performance declined in model transfer, MVE models maintained the lowest omission rates across all three seasons, with omission rate never exceeding 35% (Figure 2); its greatest drop in performance was in season P in which the omission rate jumped to 24.6% during model transfer. BRT and Maxent suffered the most drastic increases in omission. Omission rates for BRT and Maxent peaked at 95.4% and 93.8%, respectively, in season P. The greatest loss in performance occurred in season G for BRT, where omission rose by 64.1% (from 28.9% in model calibration to 93.0% in model transfer), and in season P for Maxent, where omission rates rose by 67.1% (from 26.7% in model calibration to 93.8% in model transfer). MVEs successfully predicted no less than 69.5% of IBA in all three seasons (Figure 2). Maxent and BRT models, on the other hand, never predicted greater than 32.9% of IBA.

2.3. Parameterising Algorithms

2.3.1 Boosted Regression Trees

In all, 46.7% (224) of the 480 BRT models (160/season) were significant in model transfer. PA, TC and LR parameter selections all impacted evaluation statistics. BRT tended to overfit in model transfer (Figs 3a and 4a,b). Models parameterised for PA-2 and TC ≤ 2 performed best in seasons I and P, whereas models parameterised at PA-1 and TC ≥ 3 excelled in G (Table S2.1). Higher performance was also associated with a faster LR (0.01) in season I, a moderate LR (0.005–0.0025) in season G and a slower LR (≤0.0025) in season P.

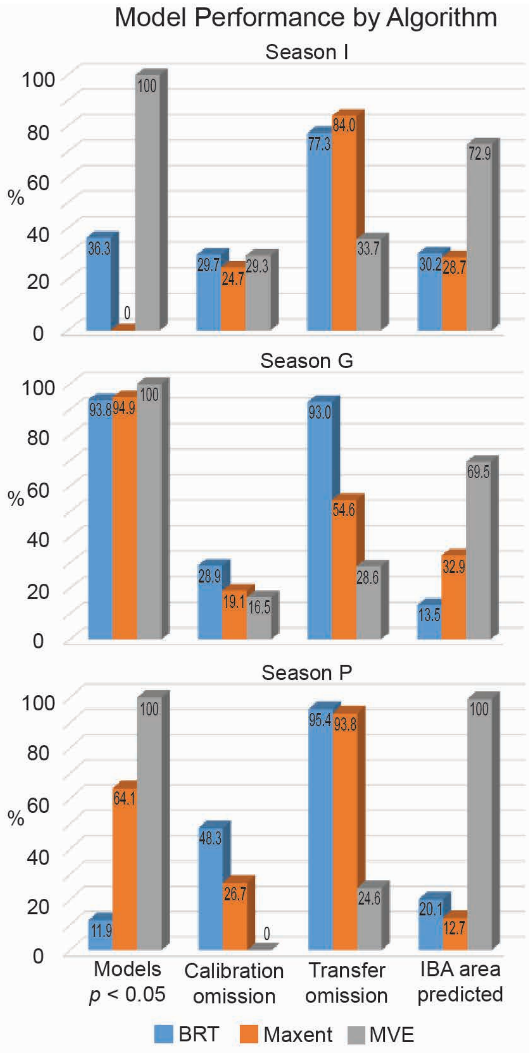


Figure 2. Model performance for best-performing models by algorithm and season (I: December–March, P: April–July and G: August–November): omission rates for model calibration and transfer and percentage of total IBA area predicted.

2.3.2 Maxent

In all, 53.0% (62) of 117 models (39 per season) parameterised in Maxent were significant after model transfer. BRT tended to overfit in model transfer (Figs 3b and 4c,d). Bias and RM played the biggest role in model performance. The top models were calibrated with the Log_2 bias layer and RM ≥ 1.5 (season G) and the raw bias layer (season P) (Table S2.1). None of the season I model projections were significant.

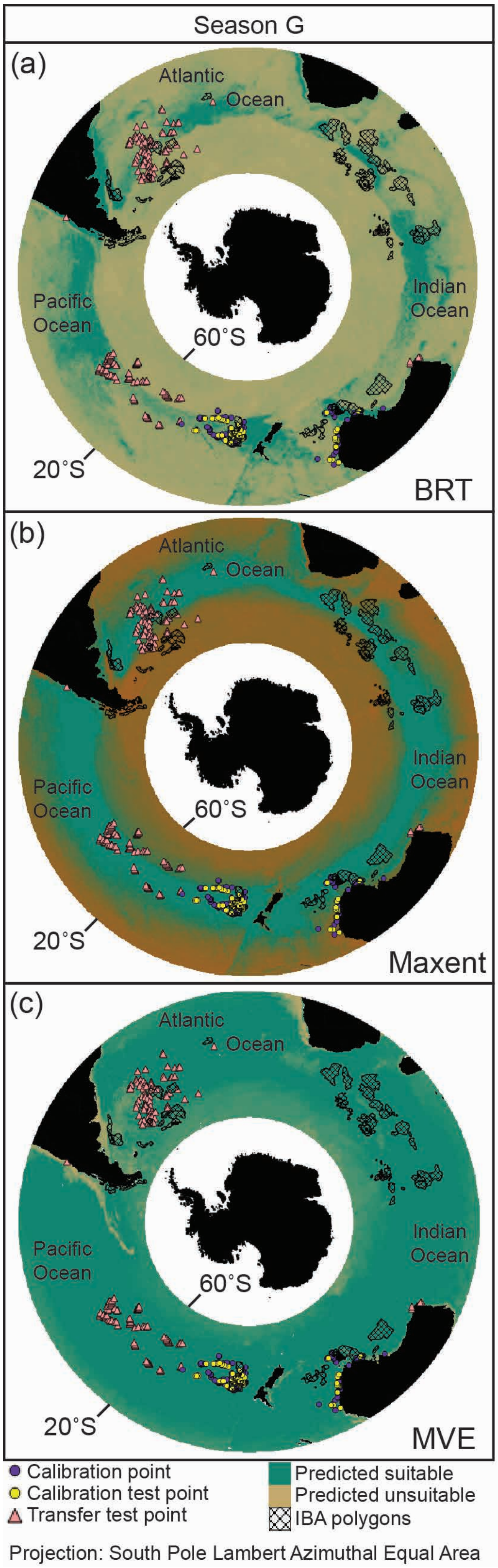


Figure 3. Season G (August–November) model projections following an 80% threshold for the top models produced by (a) BRT (PA-2, tree complexity = 5, learning rate = 0.0025, bag fraction = 0.5), (b) Maxent (Log2 bias layer, prevalence = 0.7, regularisation multiplier = 2) and (c) MVE (threshold = 0.9, run = 3). Base layer: Global Administrative Areas global shapefile (http://www.gadm.org).

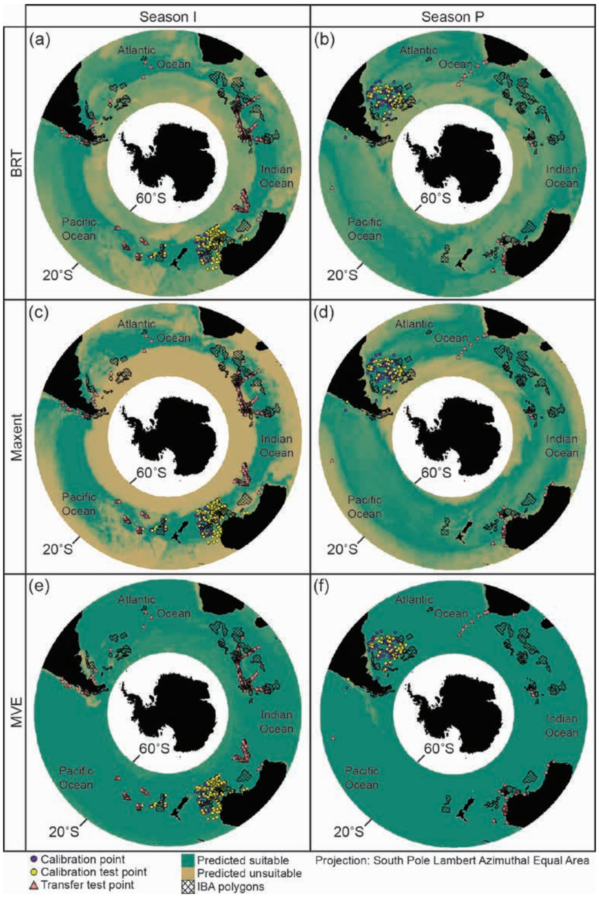


Figure 4. Season I (December–March) model projections following an 80% threshold for the top models produced by (a) BRT (PA-2, tree complexity = 1, learning rate = 0.01, bag fraction = 0.5), (c) Maxent (not significant; no bias layer, prevalence = 0.3, regularisation multiplier = 2) and (e) MVE (threshold = 0.9, run = 3); season P (April–July) model projections following an 80% threshold for the top models produced by (b) BRT (PA-2, tree complexity = 1, learning rate = 0.05, bag fraction = 0.6), (d) Maxent (no bias layer, prevalence = 0.9, regularisation multiplier = 1) and (f) MVE (threshold = 0.9, run = 1). Base layer: Global Administrative Areas global shapefile (http://www.gadm.org).

2.3.3 Minimum Volume Ellipsoids

All 54 models (18 per season) calibrated using MVEs were statistically significant in model projection. MVE predictions were generally underfit (i.e. overly general) (Figs 3c and 4e,f). Top models incorporated more moderate numbers (2–4) of environmental variables (2 ≤ Run ≤ 4) for season G and more variables (Run ≥ 2) in seasons I and P.

3. DISCUSSION

Correlative modelling offers a method by which the complexities of distributional dynamics of pelagic seabirds can be explored at the species level. Researchers have used these methods to address specific aspects of seabird distributional ecology such as habitat suitability (Ceia et al. 2012; Oppel et al. 2012; Catry et al. 2013; Louzao et al. 2013; McGowan et al. 2013; Scales et al. 2016), identification of hotspots in the present and past (Louzao et al. 2013), selection of potential conservation areas and potential climate change impacts (Krüger 2017). But, many of the more recent applications use ensemble modelling (Scales et al. 2016; Krüger 2017) or incorporate seabird movement data (Clay 2016; Quillfeldt 2017) that, whilst increasing in quantity and availability, is nowhere near as prevalent or accessible as point observation data.

The purpose of this exercise was to develop a baseline of model performance across a suite of parameterisations with an eye towards a step-wise approach to improving correlative niche modelling techniques for pelagic and other highly mobile species. Although just under half of all models tested performed significantly better than random, predictive performance was adequate only for MVEs (low omission rates, high percentage of IBA areas predicted). Indeed, MVE models consistently indicated the greatest potential for capturing the complexity of *D. exulans* distributional ecology with all calibrations significant in model calibration and model transfer. The best performing BRT and Maxent calibrations either had omission rates > 50% or predicted < 35% of the total area covered by BirdLife International's Marine IBAs (BirdLife International and NatureServe 2016) of known importance to *D. exulans*.

Though methods such as BRT and Maxent have a history of higher predictive performance (Elith et al. 2006; Phillips et al. 2009), the results presented here suggest that these more complex algorithms may not be ideal for summarizing the complexity of highly pelagic species. Parameterisations for both BRT and Maxent tended to overfit models: although Maxent exhibited a more moderate fit and higher predictive performance overall, Maxent models were still not necessarily 'good' at anticipating test occurrence data. Overall performance declined substantially during model transfer (i.e. extrapolation to the full study region) for both BRT and Maxent. Performance for the two complex algorithms was particularly poor in seasons I and P, though Maxent improved slightly in season G. This less-than-stellar overall performance likely results from the combination of the spatial exclusion of data (i.e. method of determining breeding vs non-breeding data), sampling bias within the observation data and inability to discern breeding from non-breeding individuals (i.e. lack of biological information in the observation data).

My results highlight one of the major roadblocks for correlational niche-modelling methodologies: the loss of detail in signals because of over-generalisation. Correlative modelling characterises a species' use of environmental space to create a model that can then be used to address questions regarding the species' distribution in geographic space (Barve et al. 2011). Recent studies have shown that no single 'best' algorithm or parameter setting for SDM or ENM applications exists or is likely to exist (Saupe et al. 2012; Shcheglovitova & Anderson 2013; Merow et al. 2013; Qiao et al. 2015), and the results of this study are in close agreement. Therefore, algorithm selection and parameterisation should be an iterative, hypothesis-driven or question-driven process. Myriad factors affect the performance of correlational models, including the limitations of the specific algorithms, input data quality, appropriateness of selected explanatory (environmental) variables, spatial resolution (Weimerskirch et al. 2005; Hyrenbach et al. 2007; Wakefield et al. 2009; Bellier et al. 2010) and study region extent (Barbet-Massin et al. 2012; Barve et al. 2011; Hyrenbach et al. 2007). As a result of this complexity, a key point is that averaging environmental data across each of the three seasons limits the detail available in the modelling outputs (Peterson et al. 2005; Scales et al. 2016).

The most obvious limitation encountered in this preliminary study of model assessment for pelagic bird distributions is the quality of the occurrence data, lack of absence data (Elith et al. 2011), sampling bias inherent in opportunistically collected data (Weimerskirch et al. 2006; Phillips et al. 2009; Elith et al. 2011; Grecian et al. 2012) and lack of relevant additional biological information (i.e. sex, age or breeding status; Grecian et al. 2012). These factors – lack of biological information and bias – necessarily influence calibration region designation, ultimately impacting overall model performance.

Bias is a significant concern in assessing biodiversity patterns at macroscales (Beck et al. 2014), and these biases are amplified when data are derived in bulk from portals such as GBIF (Graham et al. 2004; Yesson et al. 2007; Beck et al. 2013; Beck et al. 2014). *D. exulans* point occurrence data used here are strongly biased towards the Argentine Basin, the Tasman Sea, south Pacific Ocean south of Tasmania, the Campbell Plateau and Chatham Rise around Australia and New Zealand and areas directly adjacent to breeding colonies (Fig. 5); occurrence data are minimal for the high seas in the South Pacific Ocean east of the Pacific Rise, east of the Atlantic Ridge in the South Atlantic Ocean, south of South Africa around Agulhas Basin and Plateau and the Crozet Basin and the Southeast Indian Ridge in

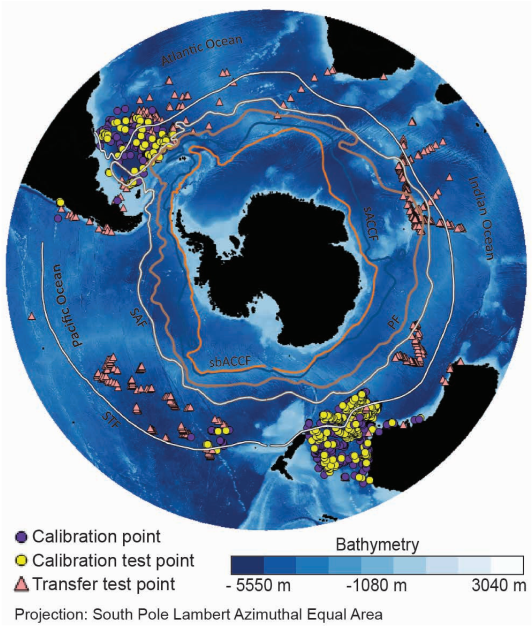


Figure 5. Diomedea exulans occurrence data overlaid with Southern Ocean front lines (STF, subtropical front; SAF, sub-Antarctic front; PF, polar front; sACCF, southern Antarctic Circumpolar Current; sbACCF, southern boundary Antarctic Circumpolar Current). Note the distinct spatial bias in observation data, particularly in the lack of data on the high seas.

the Indian Ocean. This uneven sampling leads to gaps in documentation of the species' response to some environmental conditions, limiting model generality (Owens et al. 2013).

A final concern associated with DAK is taxonomic uncertainty (Graham et al. 2004). Great albatross taxonomy has undergone multiple revisions, only recently 'stabilising' with four species in the *D. exulans* complex (Nunn et al. 1996; Burg & Croxall 2004; Chambers et al. 2009; Rains et al. 2011). Morphological similarities between species and significant overlap of non-breeding individuals within the complex make differentiation of species at-sea quite difficult (Onley & Scofield 2007). Lack of cohesive taxonomic resolution only further increases the potential for homogenisation of species ecology, an important factor often not discussed, which reduce the confidence and accuracy of correlative models.

Despite the intriguing result in which MVEs outperformed more complex algorithms, deriving ecological conclusions from low- performing or even moderate-performing models such as those that I have presented herein is premature. Rather, this study provides a baseline for the development of better and more predictive models that will eventually be capable of accounting for the complex behaviours of such species. Further, it serves as a reminder that correlative niche modelling approaches are sensitive to a large suite of factors and are impacted inherently by the study question itself. In light of the limitations of readily available seabird data (e.g. strong spatiotemporal biases, no information on sex and age of the individuals involved), development of such a baseline of algorithm behaviour is necessary for successfully evaluating the efficacy of more complex data treatment strategies.

Improved correlative modelling approaches, building on the baseline presented herein, can significantly enhance understanding of macroscale factors driving distributional dynamics of species, including pelagic seabirds and other highly mobile species, and provide crucial information to fill important information gaps necessary to project and explore the future distributional potential (Louzao et al. 2011; Catry et al. 2013). These insights, in combination with increasing knowledge of species' natural history and ecology, can inform conservation planners and offer information vital to the research priorities identified by Lewison et al. (2012) including identification and mapping of movement corridors and foraging areas to understand impacts of global change on the distributions of pelagic seabirds and other highly mobile species.

Acknowledgements: Thanks to Jorge Soberón for sharing the minimum volume ellipsoid R script, Luis Osorio for sharing his pROC script and delegates of the 2nd World Seabird Conference for insight into albatross behaviour and ecology. Funding and support were provided by the University of Kansas, Department of Ecology & Evolutionary Biology. I thank AT Peterson for insightful discussion and review of the manuscript.

Conflict of Interest: The author declares that there is no conflict of interest.

REFERENCES

Anderson, R.P. (2003) Real vs. artefactual absences in species distributions: tests for Oryzomys albigularis (Rodentia:Muridae) in Venezuela. Journal of Biogeography, 30, 591–605.

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.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., et al. (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecological Modelling, 222, 1810–1819.

Beck, J., Ballesteros-Mejia, L., Nagel, P. & Kitching, I.J. (2013) Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? Diversity and Distributions, 19, 1043–1050.

Beck, J., Boller, M., Erhardt, A. & Schwanghart, W. (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecological Informatics, 19, 10–15.

Bellier, E.G., Certain, G., Planque, B., Monestiez, P. & Bretagnolle, V. (2010) Modelling habitat selection at multiple scales with multivariate geostatistics: an application to seabirds in open sea. Oikos, 119, 988–999.

Birdlife International and Natureserve (2015b) Marine IBA e-Atlas: http://maps.birdlife.org/marineIBAs/default.html.

Burg, T.M. & Croxall, J.P. (2004) Global population structure and taxonomy of the Wandering Albatross species complex. Molecular Ecology, 13, 2345–2355.

Catry, P., Lemos, R.T., Brickle, P., Phillips, R.A., Matias, R. & Granadeiro, J.P. (2013) Predicting the distribution of a threatened albatross: the importance of competition, fisheries and annual variability. Progress in Oceanography, 110, 1–10.

Ceia, F.R., Phillips, R.A., Ramos, J.A., Cherel, Y., Vieira, R.P., Richard, P., et al. (2012) Short- and long-term consistency in the foraging niche of Wandering Albatrosses. Marine Biology, 159, 1581–1591.

Chambers, G.K., Moeke, C., Steel, R. & Trueman, J.W. (2009) Phylogenetic analysis of the 24 named albatross taxa based on full mitochondrial cytochrome b DNA sequences. Notornis, 56, 82–94.

Clay, T.A., Manica, A., Ryan, P. G., Silk, J.R.D., Croxall, J.P., Ireland, L. & Phillips, R.A. (2016) Proximate drivers of spatial segregation in non-breeding albatrosses. Scientific Reports, 6, 29932.

Coble, P.G. (2007) Marine optical biogeochemistry: The chemistry of ocean color. Chemical Reviews, 107, 402–418.

Croxall, J.P., Butchart, S.H.M., Lascelles, B., Stattersfield, A.J., Sullivan, B., Symes, A., et al. (2012) Seabird conservation status, threats and priority actions: a global assessment. Bird Conservation International, 22, 1–34.

Doney, S.C., Ruckelshaus, M., Duffy, J.E., Barry, J.P., Chan, F., English, C.A., et al. (2012) Climate change impacts on marine ecosystems. Annual Review of Marine Science, 4, 11–37.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., et al. (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29, 129–151.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. Journal of Animal Ecology, 77, 802–813.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. & Yates, C.J. (2011) A statistical explanation of MaxEnt for ecologists. Diversity and Distributions, 17, 43–57.

Game, E.T., Grantham, H.S., Hobday, A.J., Pressey, R.L., Lombard, A.T., Beckley, L.E., et al. (2009) Pelagic protected areas: the missing dimension in ocean conservation. Trends in Ecology & Evolution, 24, 360–369.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution, 19, 497–503.

Grecian, W.J., Witt, M.J., Attrill, M.J., Bearhop, S., Godley, B.J., Grémillet, D., et al. (2012) A novel projection technique to identify important at-sea areas for seabird conservation: an example using Northern Gannets breeding in the North East Atlantic. Biological Conservation, 156, 43–52.

Hyrenbach, K.D., Forney, K.A. & Dayton, P.K. (2000) Marine protected areas and ocean basin management. Aquatic Conservation—Marine and Freshwater Ecosystems, 10, 437–458.

Hyrenbach, K.D., Veit, R.R., Weimerskirch, H., Metzl, N. & Hunt, G.L. (2007) Community structure across a large-scale ocean productivity gradient: marine bird assemblages of the southern Indian Ocean. Deep-Sea Research Part I-Oceanographic Research Papers, 54, 1129–1145.

Iucn (2016) IUCN Red List of Threatened Species v2015-4: http://www.iucnredlist.org.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schroder, B., Lindenborn, J., Reinfelder, V., et al. (2013) The importance of correcting for sampling bias in MaxEnt species distribution models. Diversity and Distributions, 19, 1366–1379.

Krüger, L., Ramos, J.A., Xavier, J.C., Grémillet, D., González-Solís, J., Petry, M.V., Phillips, R.A., Wanless, R.M. & Paiva, V.H. (2017) Projected distributions of Southern Ocean albatrosses, petrels and fisheries as a consequence of climatic change. Ecography,

Lascelles, B.G., Langham, G.M., Ronconi, R.A. & Reid, J.B. (2012) From hotspots to site protection: identifying Marine Protected Areas for seabirds around the globe. Biological Conservation, 156, 5–14.

Lewison, R., Oro, D., Godley, B., Underhill, L., Bearhop, S., Wilson, R.P., et al. (2012) Research priorities for seabirds: improving conservation and management in the 21st century. Endangered Species Research, 17, 93–121.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography, 17, 145–151.

Louzao, M., Pinaud, D., Péron, C., Delord, K., Wiegand, T. & Weimerskirch, H. (2011) Conserving pelagic habitats: seascape modelling of an oceanic top predator. Journal of Applied Ecology, 48, 121–132.

Louzao, M., Aumont, O., Hothorn, T., Wiegand, T. & Weimerskirch, H. (2013) Foraging in a changing environment: habitat shifts of an oceanic predator over the last half century. Ecography, 36, 57–67.

Mateo, R.G., De La Estrella, M., Felicísimo, Á.M., Munoz, J. & Guisan, A. (2013) A new spin on a compositionalist predictive modelling framework for conservation planning: a tropical case study in Ecuador. Biological Conservation, 160, 150–161.

Mcgowan, J., Hines, E., Elliott, M., Howar, J., Dransfield, A., Nur, N., et al. (2013) Using seabird habitat modeling to inform marine spatial planning in central California's National Marine Sanctuaries. PLoS One, 8, e71406.

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.

Milot, E., Weimerskirch, H. & Bernatchez, L. (2008) The seabird paradox: dispersal, genetic structure and population dynamics in a highly mobile, but philopatric albatross species. Molecular Ecology, 17, 1658–1673.

Nelson, N.B. & Siegel, D.A. (2013) The global distribution and dynamics of chromophoric dissolved organic matter. Annual Review of Marine Science, 5, 447–476.

Nunn, G.B., Cooper, J., Jouventin, P., Robertson, C.J.R. & Robertson, G.G. (1996) Evolutionary relationships among extant albatrosses (Procellariiformes: Diomedeidae) established from complete cytochrome-B gene sequences. Auk, 113, 784–801.

Onley, D. & Scofield, P. (2007) Albatrosses, petrels, & shearwaters of the world. Princeton University Press, Princeton, New Jersey.

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'connell, A.F., Miller, P.I., et al. (2012) Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biological Conservation, 156, 94–104.

Owens, H.L., Campbell, L.P., Dornak, L.L., Saupe, E.E., Barve, N., Soberon, J., et al. (2013) Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. Ecological Modelling, 263, 10–18.

Peterson, A.T., Martínez-Campos, C., Nakazawa, Y. & Martínez-Meyer, E. (2005) Time-specific ecological niche modeling predicts spatial dynamics of vector insects and human dengue cases. Transactions of the Royal Society of Tropical Medicine and Hygiene, 99, 647–655.

Peterson, A.T. (2006) Uses and requirements of ecological niche models and related distribution models. Biodiversity Informatics, 3, 59–72.

Peterson, A.T., Papeş, M. & Soberón, J. (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecological Modelling, 213, 63–72.

Phillips, R.A., Silk, J.R.D., Croxall, J.P., Afanasyev, V. & Bennett, V.J. (2005) Summer distribution and migration of nonbreeding albatrosses: individual consistencies and implications for conservation. Ecology, 86, 2386–2396.

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., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., et al. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications, 19, 181–197.

Piatt, J.F., Sydeman, W.J. & Wiese, F. (2007) Introduction: a modern role for seabirds as indicators. Marine Ecology Progress Series, 352, 199–204.

Prince, P.A., Wood, A.G., Barton, T. & Croxall, J.P. (1992) Satellite tracking of Wandering Albatrosses (Diomedea exulans) in the South Atlantic. Antarctic Science, 4, 31–36.

Qiao, H.J., Soberón, J. & Peterson, A.T. (2015) No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. Methods in Ecology and Evolution, 6, 1126–1136.

Quillfeldt, P., Engler, J.O., Silk, J.R., Phillips, R.A. (2017) Influence of device accuracy and choice of algorithm for species distribution modelling of seabirds: a case study using black-browed albatrosses. Journal of Avian Biology,

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: http://www.r-project.org.

Rains, D., Weimerskirch, H. & Burg, T.M. (2011) Piecing together the global population puzzle of Wandering Albatrosses: genetic analysis of the Amsterdam albatross Diomedea amsterdamensis. Journal of Avian Biology, 42, 69–79.

Ramos, R., Sanz, V., Militao, T., Bried, J., Neves, V.C., Biscoito, M., et al. (2015) Leapfrog migration and habitat preferences of a small oceanic seabird, Bulwer's petrel (Bulweria bulwerii). Journal of Biogeography, 42, 1651–1664.

Roberts, J.J., Best, B.D., Dunn, D.C. & Halpin, P.N. (2010) Marine Geospatial Ecology Tools: an integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. Environmental Modelling and Software, 25, 1197–1207.

Rodríguez, J.P., Brotons, L., Bustamante, J. & Seoane, J. (2007) The application of predictive modelling of species distribution to biodiversity conservation. Diversity and Distributions, 13, 243–251.

Saupe, E.E., Barve, V., Myers, C.E., Soberón, J., Barve, N., Hensz, C.M., et al. (2012) Variation in niche and distribution model performance: the need for a priori assessment of key causal factors. Ecological Modelling, 237, 11–22.

Scales, K.L., Miller, P.I., Ingram, S.N., Hazen, E.L., Bograd, S.J. & Phillips, R.A. (2016) Identifying predictable foraging habitats for a wide-ranging marine predator using ensemble ecological niche models. Diversity and Distributions, 22, 212–224.

Shcheglovitova, M. & 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.

Soberón, J. & Peterson, A.T. (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics, 2, 1–10.

Sousa-Baena, M.S., Garcia, L.C. & Peterson, A.T. (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. Diversity and Distributions, 20, 369–381.

Thiebot, J.B., Lescroel, A., Pinaud, D., Trathan, P.N. & Bost, C.A. (2011) Larger foraging range but similar habitat selection in non-breeding versus breeding sub-Antarctic penguins. Antarctic Science, 23, 117–126.

Urtizberea, A., Dupont, N., Rosland, R. & Aksnes, D.L. (2013) Sensitivity of euphotic zone properties to CDOM variations in marine ecosystem models. Ecological Modelling, 256, 16–22.

Wakefield, E.D., Phillips, R.A. & Matthiopoulos, J. (2009) Quantifying habitat use and preferences of pelagic seabirds using individual movement data: a review. Marine Ecology Progress Series, 391, 165–182.

Wakefield, E.D., Phillips, R.A., Trathan, P.N., Arata, J., Gales, R., Huin, N., et al. (2011) Habitat preference, accessibility, and competition limit the global distribution of breeding Black-browed Albatrosses. Ecological Monographs, 81, 141–167.

Weimerskirch, H., Inchausti, P., Guinet, C. & Barbraud, C. (2003) Trends in bird and seal populations as indicators of a system shift in the Southern Ocean. Antarctic Science, 15, 249–256.

Weimerskirch, H., Gault, A. & Cherel, Y. (2005) Prey distribution and patchiness: factors in foraging success and efficiency of Wandering Albatrosses. Ecology, 86, 2611–2622.

Weimerskirch, H., Åkesson, S. & Pinaud, D. (2006) Postnatal dispersal of Wandering Albatrosses Diomedea exulans: implications for the conservation of the species. Journal of Avian Biology, 37, 23–28.

Weimerskirch, H., Jouventin, P., Mougin, J.L., Stahl, J.C. & Vanbeveren, M. (1985) Banding recoveries and the dispersal of seabirds breeding in French Austral and Antarctic Territories. Emu, 85, 22–33.

Weimerskirch, H., Louzao, M., De Grissac, S. & Delord, K. (2012) Changes in wind pattern alter albatross distribution and life-history traits. Science, 335, 211–214.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., et al. (2007) How global is the Global Biodiversity Information Facility? PLoS One, 2, e1124.

APPENDIX S1. ADDITIONAL MODEL CALIBRATION AND PARAMETERISATION INFORMATION.

Biologically-informed ecological niche models for an example pelagic, mobile species

Ingenloff, K.
KU Biodiversity Institute, kate.ingenloff@ku.edu

Input Data – Occurrence Data

Data Acquisition. Occurrence data for all members of the order Procellariiformes were obtained from the Global Biodiversity Information Facility (GBIF; accessed 5/26/2015, doi:10.15468/dl.fquf8g). GBIF search was restricted to observation data for all error-free procellariiform records within the study period of December 2000–November 2011 between −20˚S and −70˚S latitude and was requested on 26 May 2015. The search returned 144,850 observations of 105 species from 28 genera and 4 families (Diomedeidae, Procellariidae, Hydrobatidae and Pelecanoididae) species from 31 collections/institutions. Occurrence data were derived from human and machine observations and preserved specimens.

Table S1.1. GBIF procellariiform occurrence data contributors.

GBIF Institution Code	Institution
AADC	Australia Antarctic Data Centre
ABBBS	Bird Banding Records, Australian Antarctic Territory & Heard Island
AM	Australian Museum
AMNH	American Museum of Natural History
ANWC	Australian National Wildlife Collection
Anymals.org	Anymals.org; Anymals+Plants Mobile Application
APN-AR	*Administración de Parques Nacionales, Argentina*
BAS	British Antarctic Survey
BGBM	Botanical Garden and Botanical Museum Berlin-Dahlem
Birds Australia, Birdata	BirdLife Australia, BirdLife International
CAML	Census of Antarctic marine Life
CLO	Cornell Laboratory of Ornithology
CTALA_LB	*Ministerio del Medio Ambiente de Chile*
CUML	Cornel University Macaulay Library
Eremaea Pty Ltd	Eremaea eBird
iNaturalist	iNaturalist.org
Individual Sightings	Individual sightings
IRSNB	*Institut Royal des Sciences Naturelles de Belgique*
naturgucker	*Natur Gucker*
NMR	*Natuurhistorisch Museum*
NMV	National Museum Victoria
OBIS-SEAMAP	Ocean Biogeographic Information System: Spatial Ecological Analysis of Megavertebrate Populations
SAMA	South Australia Museum
SA Fauna	South Australia Department of Environment & Natural Resources
TMAG	Tasmanian Museum & Art Gallery

GBIF Institution Code	Institution
UCT-ADU	University of Cape Town Animal Demography Unit
USNM	Smithsonian Institution Natural History Museum
UWBM	University of Washington Burke Museum
QM	Queensland Museum, Australia
QVMAG	Queen Victoria Museum & Art Gallery
ZMA	Zoological Museum Amsterdam, University of Amsterdam

Input Data – Environmental Data

Four MODIS Terra L3 standard mapped image (SMI) environmental datasets at 4.6 km spatial resolution were downloaded from the NASA OceanColor Web (Table S1.2; NASA 2014). Imagery were converted from native HDFs to ASCII grids and reprojected to WGS 84 using the Marine Geospatial Ecology Tools (MGET) ArcGIS toolbox extension (Roberts et al 2010). ‘No Data’ values were filled using a temporal filter followed by a spatial filter. The mean, maximum, minimum and range of values were calculated by season for each variable; the resulting time-averaged rasters were then incorporated into a series of principle component analyses (PCA).

Table S1.2 MODIS Terra raster data accessed from NASA’s OceanColor Web.

Variable		Unit	Date accessed
Sea surface temperature	(SST)	11 µm	18 Feb 2015
Nightly sea surface temperature	(NSST)	11 µm	14 Feb 2015
Chromophoric dissolved organic matter index	(CDOM)		8 Feb 2015
Chlorophyll-a concentration	(CHL)	mg/m3	16 Feb 2015

PCAs: PCAs were run to reduce dimensionality and collinearity. The first five principle components (PCs) per season were used in the analyses; in all three seasons, the first PC explained ≥95% of variation (Table S1.3). The final PCs selected for use in the analyses were resampled from 0.041667 to 0.20833.

Table S1.3 PCA Loadings for the first five components of each set used in the analyses by season.

Season	Variable	PC1	PC2	PC3	PC4	PC5
I	CDOM_max	0.034	−0.235	0.636	0.217	−0.005
	CDOM_range	0.041	−0.238	0.621	0.218	0.009
	CHL_max	0.012	−0.677	−0.246	−0.011	−0.008
	CHL_range	0.011	−0.648	−0.246	−0.010	−0.009
	NSST_max	−0.418	−0.039	0.096	−0.268	−0.201
	NSST_mean	−0.412	−0.015	0.020	0.014	−0.059
	NSST_min	−0.405	0.013	−0.085	0.309	0.340
	NSST_range	−0.013	−0.052	0.182	−0.576	−0.541
	SST_max	−0.410	−0.039	0.092	−0.245	0.285
	SST_mean	−0.406	−0.015	0.026	0.016	−0.061
	SST_min	−0.393	0.019	−0.067	0.274	−0.317
	SST_range	−0.017	−0.058	0.158	−0.520	0.602
	Cumulative Proportion	96.69	98.30	99.10	99.83	99.93

Season	Variable	PC1	PC2	PC3	PC4	PC5
G	CDOM_max	0.053	−0.602	−0.335	0.003	0.008
	CDOM_range	0.056	−0.638	−0.334	0.010	0.006
	CHL_max	0.002	−0.068	0.109	−0.715	−0.013
	CHL_range	0.001	−0.059	0.099	−0.677	−0.018
	NSST_max	−0.428	−0.168	0.225	0.051	−0.309
	NSST_mean	−0.411	−0.016	−0.061	−0.005	−0.054
	NSST_min	−0.394	0.112	−0.282	−0.063	0.250
	NSST_range	−0.034	−0.280	0.507	0.114	−0.559
	SST_max	−0.420	−0.144	0.222	0.039	0.368
	SST_mean	−0.402	0.002	−0.055	−0.006	−0.043
	SST_min	−0.383	0.117	−0.268	−0.044	−0.216
	SST_range	−0.036	−0.261	0.490	0.084	0.585
	Cumulative Proportion	96.15	98.32	99.23	99.81	99.92
P	CDOM_max	0.0680	−0.5201	0.4385	−0.1618	0.0130
	CDOM_range	0.0751	−0.5295	0.4426	−0.1546	0.0038
	CHL_max	0.0084	−0.4177	−0.5581	−0.1779	−0.0147
	CHL_range	0.0078	−0.3955	−0.5410	−0.1803	−0.0236
	NSST_max	−0.4173	−0.1304	0.0061	0.2663	−0.2843
	NSST_mean	−0.4081	−0.0322	0.0374	−0.0378	−0.0660
	NSST_min	−0.4033	0.0682	0.0268	−0.3055	0.2590
	NSST_range	−0.0140	−0.1986	−0.0208	0.5718	−0.5432
	SST_max	−0.4108	−0.1214	−0.0081	0.2468	0.3756
	SST_mean	−0.4020	−0.0313	0.0360	−0.0287	−0.0676
	SST_min	−0.3942	0.0677	0.0392	−0.2645	−0.2218
	SST_range	−0.0166	−0.1891	−0.0472	0.5113	0.5974
	Cumulative Proportion	95.84	97.79	99.04	99.81	99.92

Model Calibration

Table S1.4 Cell value ranges for raw and Log2 kernel smoothed seasonal bias layers tested in Maxent model calibrations.

Season	Bias layer	
	Raw	Log_2 Smoothed
I	0:1604	0:114.0237
G	0:469	0:39.2755
P	0:1070	0:84.2523

Table S1.5 MVE model calibration parameterisations.

Parameter	Parameter Range	
Threshold (T)	0.90	
	0.95	
	0.99	
Variables included (Run)	Bathymetry, bathymetry slope, PC 1–5	Run 1
	Bathymetry, PC 1–5	Run 2
	Bathymetry, bathymetry slope, PC 1–4	Run 3
	Bathymetry, PC 1–4	Run 4
	Bathymetry, PC 1–3	Run 5
	Bathymetry, PC 1–2	Run 6

Table S1.6 *Pseudo-absence (PA) levels used in boosted regression tree calibrations. The first level, PA-1, was standardised at 1,500 randomly selected points in the calibration region; PA-2 values were calculated at double the total Diomedea exulans observation data available for use in model calibration and testing.*

Season	PA-1	PA-2
I	1,500	1,106
G	1,500	562
P	1,500	280

References

Global Biodiversity Informatics Facility (2015) www.gbif.org. Accessed 5/26/2015. DOI:10.15468/dl.fquf8g

NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group (2014) MODIS-Terra Ocean Color Data; NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group. http://dx.doi.org/10.5067/TERRA/MODIS_OC.2014.0. Accessed on 2/8-18/2015

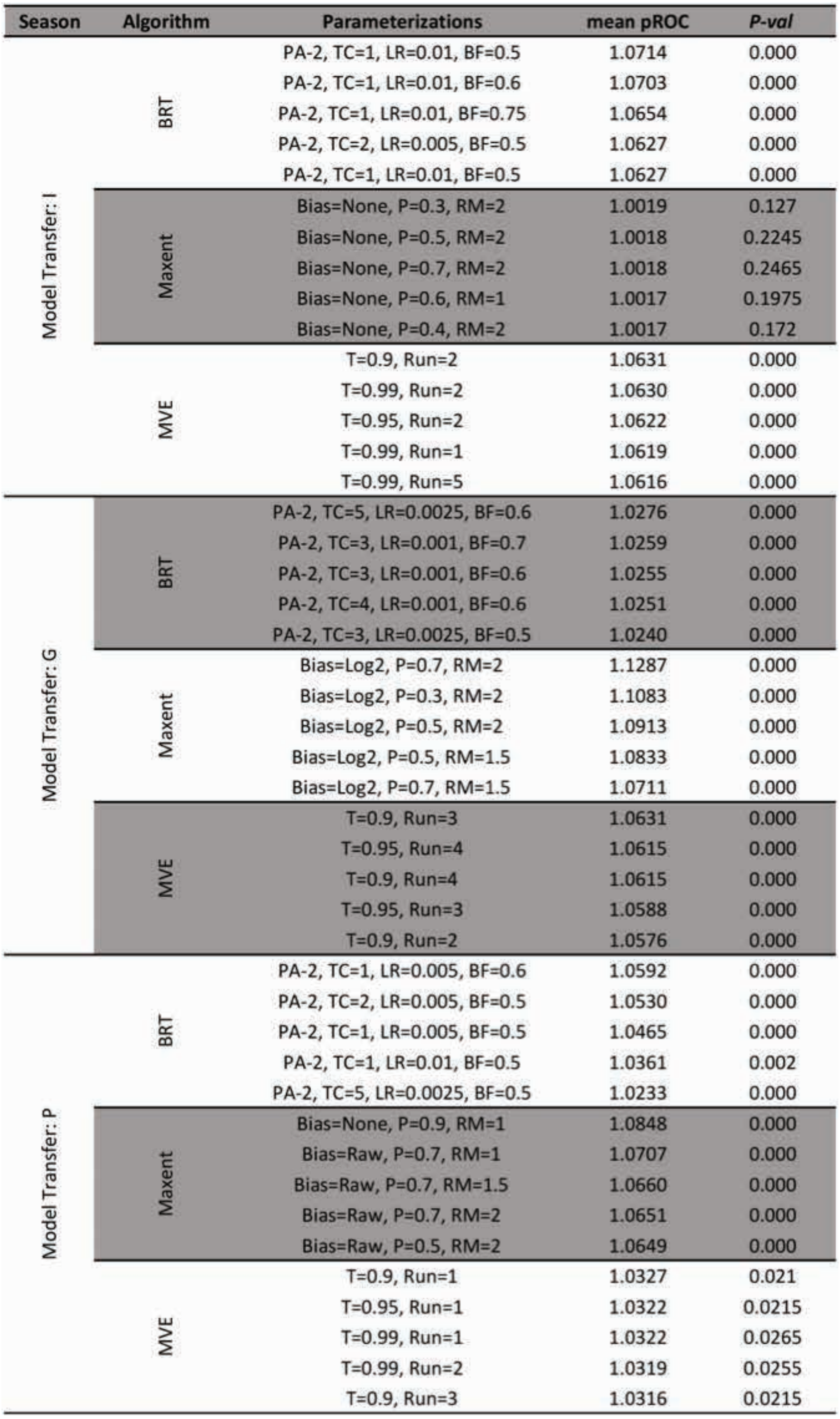
Roberts JJ, Best BD, Dunn DC, Treml EA, Halpin PN (2010) Marine Geospatial Ecology Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. Environmental Modelling & Software, 25: 1197-1207. doi: 10.1016/j.envsoft.2010.03.029.

APPENDIX S2 ADDITIONAL TABLES AND FIGURES.

Biologically-informed ecological niche models for an example pelagic, mobile species

Ingenloff, K.
KU Biodiversity Institute, kate.ingenloff@ku.edu

Table S2.1. Model transfer summary statistics – mean pROC score and the overall significance – for the top five model parameterizations from each algorithm by season



Season	Algorithm	Parameterizations	mean pROC	*P-val*
Model Transfer: I	BRT	PA-2, TC=1, LR=0.01, BF=0.5	1.0714	0.000
		PA-2, TC=1, LR=0.01, BF=0.6	1.0703	0.000
		PA-2, TC=1, LR=0.01, BF=0.75	1.0654	0.000
		PA-2, TC=2, LR=0.005, BF=0.5	1.0627	0.000
		PA-2, TC=1, LR=0.01, BF=0.5	1.0627	0.000
	Maxent	Bias=None, P=0.3, RM=2	1.0019	0.127
		Bias=None, P=0.5, RM=2	1.0018	0.2245
		Bias=None, P=0.7, RM=2	1.0018	0.2465
		Bias=None, P=0.6, RM=1	1.0017	0.1975
		Bias=None, P=0.4, RM=2	1.0017	0.172
	MVE	T=0.9, Run=2	1.0631	0.000
		T=0.99, Run=2	1.0630	0.000
		T=0.95, Run=2	1.0622	0.000
		T=0.99, Run=1	1.0619	0.000
		T=0.99, Run=5	1.0616	0.000
Model Transfer: G	BRT	PA-2, TC=5, LR=0.0025, BF=0.6	1.0276	0.000
		PA-2, TC=3, LR=0.001, BF=0.7	1.0259	0.000
		PA-2, TC=3, LR=0.001, BF=0.6	1.0255	0.000
		PA-2, TC=4, LR=0.001, BF=0.6	1.0251	0.000
		PA-2, TC=3, LR=0.0025, BF=0.5	1.0240	0.000
	Maxent	Bias=Log2, P=0.7, RM=2	1.1287	0.000
		Bias=Log2, P=0.3, RM=2	1.1083	0.000
		Bias=Log2, P=0.5, RM=2	1.0913	0.000
		Bias=Log2, P=0.5, RM=1.5	1.0833	0.000
		Bias=Log2, P=0.7, RM=1.5	1.0711	0.000
	MVE	T=0.9, Run=3	1.0631	0.000
		T=0.95, Run=4	1.0615	0.000
		T=0.9, Run=4	1.0615	0.000
		T=0.95, Run=3	1.0588	0.000
		T=0.9, Run=2	1.0576	0.000
Model Transfer: P	BRT	PA-2, TC=1, LR=0.005, BF=0.6	1.0592	0.000
		PA-2, TC=2, LR=0.005, BF=0.5	1.0530	0.000
		PA-2, TC=1, LR=0.005, BF=0.5	1.0465	0.000
		PA-2, TC=1, LR=0.01, BF=0.5	1.0361	0.002
		PA-2, TC=5, LR=0.0025, BF=0.5	1.0233	0.000
	Maxent	Bias=None, P=0.9, RM=1	1.0848	0.000
		Bias=Raw, P=0.7, RM=1	1.0707	0.000
		Bias=Raw, P=0.7, RM=1.5	1.0660	0.000
		Bias=Raw, P=0.7, RM=2	1.0651	0.000
		Bias=Raw, P=0.5, RM=2	1.0649	0.000
	MVE	T=0.9, Run=1	1.0327	0.021
		T=0.95, Run=1	1.0322	0.0215
		T=0.99, Run=1	1.0322	0.0265
		T=0.99, Run=2	1.0319	0.0255
		T=0.9, Run=3	1.0316	0.0215

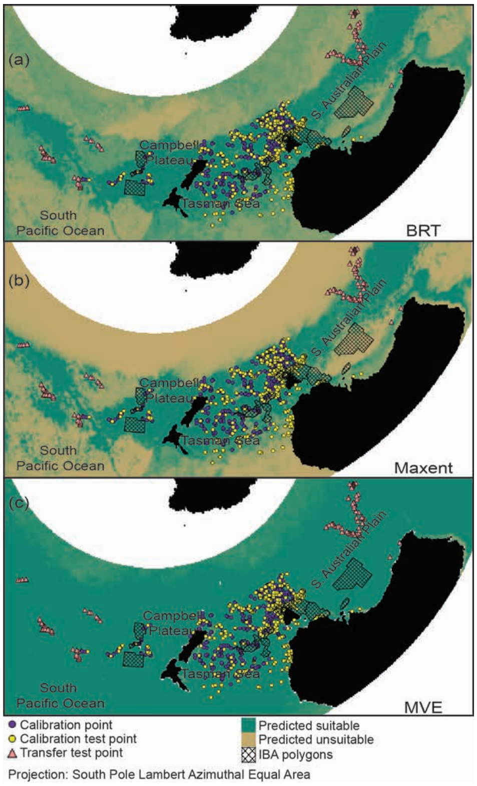


Figure S2.1 Season I projections for each algorithm: (a) BRT, (b) Maxent and (c) MVE overlaid with Diomedea exulans IBAs in waters around Australia and New Zealand.

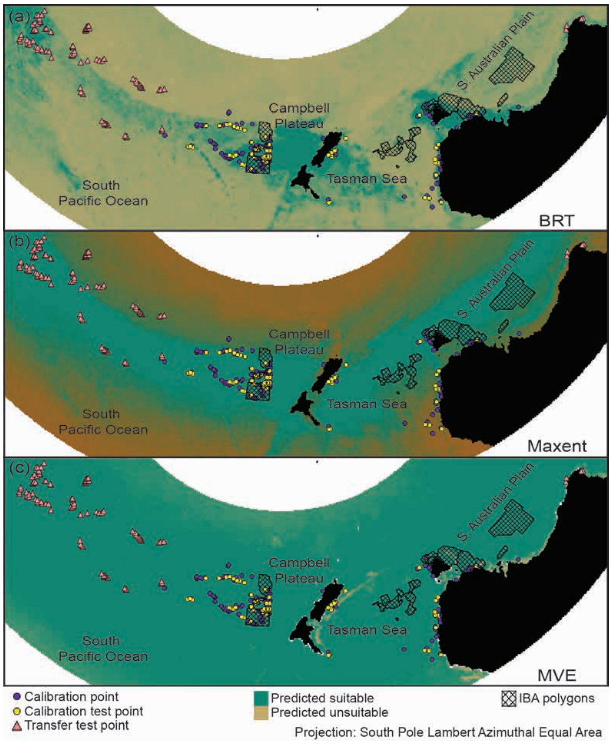


Figure S2.2 Season G projections for each algorithm: (a) BRT, (b) Maxent and (c) MVE overlaid with Diomedea exulans IBAs in waters around Australia and New Zealand.

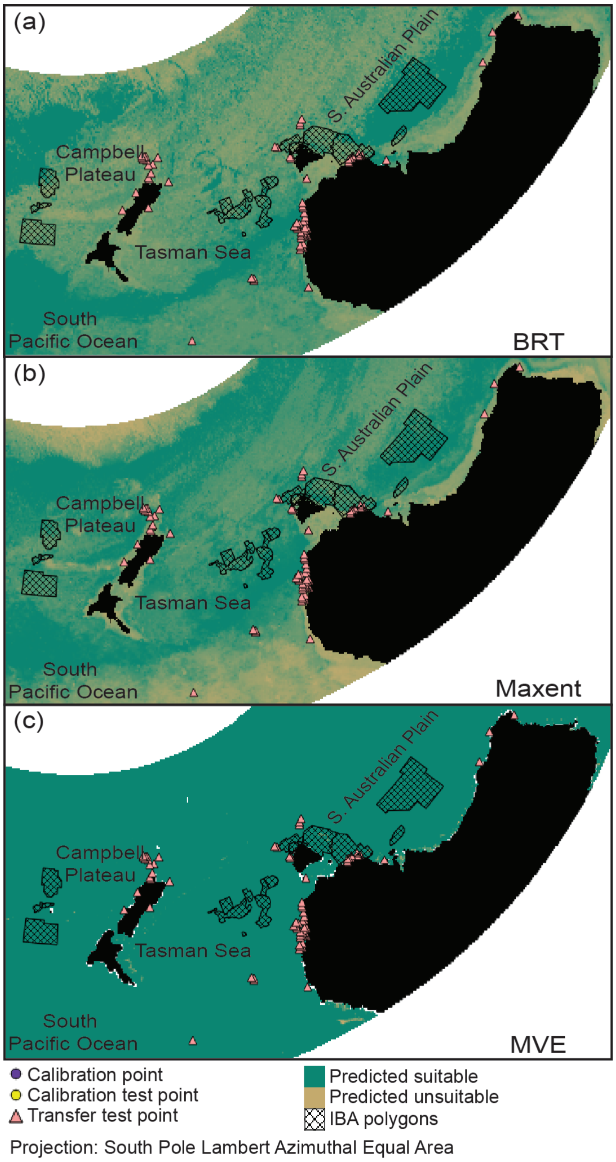


Figure S2.3 Season P projections for each algorithm: (a) BRT, (b) Maxent and (c) MVE overlaid with Diomedea exulans IBAs in waters around Australia and New Zealand.

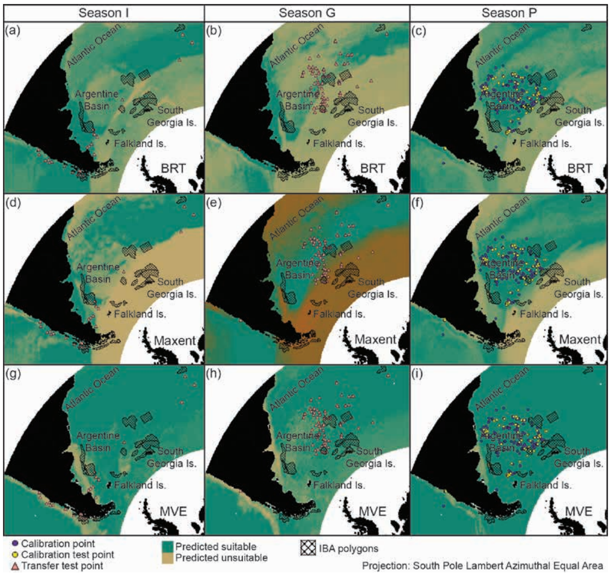


Figure S2.4 Binary model predictions for Diomedea exulans in the waters east of southern South America for (a,d,g) season I, (b,e,h) season G and (c,f,i) season P for (a–c) BRT, (d–f) Maxent and (g–i) MVE.

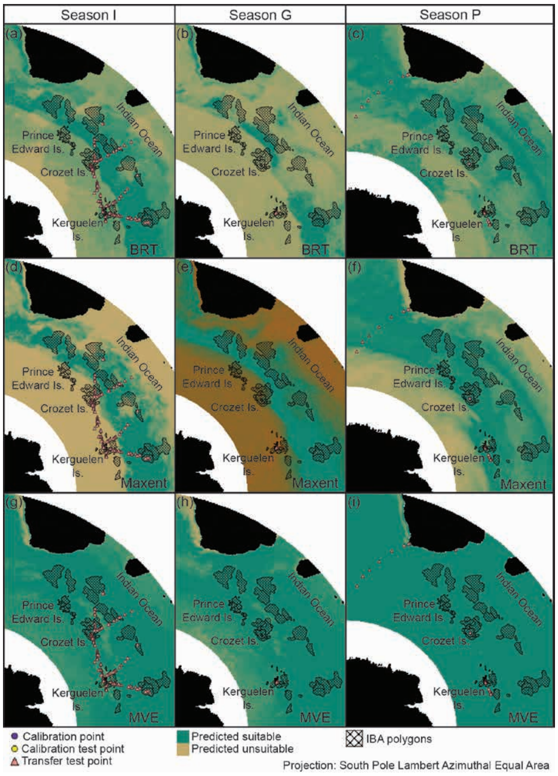


Figure S2.5 Diomedea exulans IBAs in marine regions southeast of southern Africa for (a,d,g) season I, (b,e,h) season G and (c,f,i) season P by algorithm: (a–c) BRT, (d–f) Maxent and (g–i) MVE.

Citations

BirdLife International and NatureServe. (2015) Marine IBA e-Atlas: http://maps.birdlife.org/marineIBAs/default.html.

BirdLife International and NatureServe. (2016) Marine IBAs. BirdLife International, Cambridge, UK and NatureServe, Arlington, USA.

APPENDIX S3 R CODE FOR RUNNING MINIMUM VOLUME ELLIPSOIDS AS NICHE MODELS.

Biologically-informed ecological niche models for an example pelagic, mobile species

Ingenloff, K.
KU Biodiversity Institute, kate.ingenloff@ku.edu

--------------------Fitting Minimum Volume Ellipsoids as Niche Models----------------------

# Original code provided by Jorge Soberón, August 2015.

## Minimum volume ellipsoids (MVEs) can be used as niche models, mostly when one is interested in fitting a niche not too constrained by the details of the observed data. To do this, we must (1) calculate ellipsoids and (2) calculate, for all pixels in a region of interest, the environmental distance of each pixel to a centroid of the ellipse.

## Ellipsoids can be calculated in many dimensions and are characterised by a centroid and by a matrix (symmetric) that describes the directions of the axes and their lengths.

# load required libraries

```
library(raster)
library(sp)
library(rgdal)
library(maptools)
library(MASS)
library(foreign)
```

# Define the Mahalanobis function that calculates the distance from a point (‘p’) to an ellipse of centroid (‘m’) and matrix (‘s’). Then the parameters are *p*, the test point; *m*, the centroid of the ellipse (of a distribution); and *s,* which is the INVERSE of the covariance matrix of the ellipse.

```
maja = function(p, m, s)((p - m)%*%s%*%t(p - m))^0.5;
```

--------------- DATA PREPARATION ---------------

# Set working directory

```
setwd(“<path to chosen working directory>”);
```

# Load environmental rasters (ASCII format)

```
EnvArchives <- list.files(path = “<path to environmental variables>”, pattern
= “*.asc$”, full.names = F);
EnvArchives;
```

# Rasterise and name each environmental variable to be used in analyses

```
V1 = raster(EnvArchives[1]);
V2 = raster(EnvArchives[2]);
V3 = raster(EnvArchives[3]);
V4 = raster(EnvArchives[4]);
V5 = raster(EnvArchives[5]);
```

…<and so forth>…

# Stack the environmental layers

```
layers = stack(V1, V2, V3, V4, V5, …);
layers;
```

# Read in the .csv file containing the ‘training points’ (species occurrence data to be used in model calibration) and check formatting. The .csv should contain three columns: species ID, longitude and latitude.

```
refined = read.csv(“<path to occurrence data file>.csv”, header = T);
head(refined);
```

# Index by species ID. This is *only* necessary if there are point observation data for multiple species in the .csv.

```
i1 = which(refined[, 1] == “<speciesID>”);
i2 = which(refined[, 1] == “<speciesID>”);
…
```

# Convert to matrix

```
refined = as.matrix(refined[, 2:3]);
```

# Extract the values of the environmental variables (using the raster stack) to the observation points. NOTE: specifying ‘i#’ here is not necessary if the occurrence data file only includes one species.

```
vars = extract(layers, refined[i1, 2:3]);
crds_vrs = cbind(refined, vars);
```

# Check that the new matrix contains SpeciesID, longitude, latitude and extracted environmental data for each point

```
head(crds_vrs);
```

--------------- CALCULATING MVEs ---------------

# Define the function to calculate the number of points to be included in MVE calculation. NOTE: ‘nD’ designates the species (i.e. ‘i1’) and ‘level’ designates the model threshold.

```
NDquantil = function(nD, level) return(round(nD * level/1))
```

# Specify the species, assign a threshold and calculate the number of points to be included in the analyses. In the following code, the threshold is 0.95, or E = 5%. What you’re doing here calculating the number of occurrence points for species ‘il’ excluding the most extreme 5%, which will then be used to generate the MVEs to be used in model calibration.

```
# only one species in occurrence dataset
n1 = NDquantil(refined, 0.95);
n1;


# for occurrence datasets with multiple species, run a count for each spe-
cies
n1 = NDquantil(length(i1), 0.95);
n1;
…
```

# Generate ellipsoids. Ellipsoids are represented by a (1) centroid and (2) matrix of covariance. NOTE: The values of the highlighted column range below will depend on the number of environmental variables to be extracted. The range below (4:8) indicate that ellipsoids are being generated based on five variables.

```
### only one species in occurrence dataset
mve1 = cov.mve(crds_vrs[i1, 4:8], quantile.used = n1);


#### for occurrence datasets with multiple species, run a count for each
species
mve1 = cov.mve(crds_vrs[i1, 4:8], quantile.used = n1);
…
```

# Create a matrix of the covariances

```
mu1 = matrix(mve1$center, nrow=1);
```

# Take the inverse of the covariances

```
invs1 = solve(s1);
```

--------------- MODEL CALIBRATION ---------------

# To proceed with model calibration, you must first generate a regular grid (also known as ‘fishnet’) of the training/calibration region. QGIS is highly recommended for this process because it is (a) non-proprietary (read, open-source) and (b) a lot more efficient in this process than the competing ESRI product.

## NOTE: The grid must be set to match the spatial resolution of the environmental data; be sure to add XY coordinates to labels. The resulting .dbf will be used to then apply the defined ellipsoids to every point in raster.

# ---- Creating the regular grid in QGIS (v 2.8.2 Wien):

```
# [1] Load one of the environmental rasters that will be used in analyses
# [2] Navigate to: Vector → Research Tools → Vector Grid
# [3] Set the “Grid extent” to match the environmental raster
# [4] Check “Align extents and resolution to selected raster layer”
# [5] Select “Update extents from layer”
# [6] Check “Output grid as polygons”
# [7] Assign a name to and pathway to the output shapefile
# [8] Press “OK” … processing does take a few minutes with processing time
increasing as resolution and geographic area increase.
```

# Read in regular grid .dbf file for the calibration region

```
randT = read.dbf(“<path to regular grid of calibration region>.dbf”);
head(randT);  # check that the grid read in properly (e.g. the longitude and latitude are there)
```

# Extract environmental data from the raster stack to the calibration region grid. NOTE: there will be A LOT will be NAs.

```
vrsT = extract(layers, randT[, 2:3]);
head(vrsT);  # check that everything read in and extracted properly
```

# For shits and giggles, you can calculate the percentage of NAs…

```
vrsTsna = na.omit(vrsT);
pNA = dim(vrsTsna) / TotalNumberPixelsInGrid;
pNA;
```

# Create the matrix that will contain the distance of environment to centroid. The matrix size will be [Total number of pixels in grid x 1].

```
dT1 = matrix(0, ncol = 1, nrow = TotalNumberPixelsInGrid);
```

# Calculate environmental distance of each ellipsoid from the centroid

```
for(i in 1:TotalNumberPixelsInGrid)dT1[i, 1] = maja(vrsT[i, ], mu1, invs1);
```

# Check that it worked. The resulting table should have the following: longitude, latitude, one column for each environmental variable and a column for dT1.

```
Mcalib = cbind(randT, vrsT, dT1);
head(Mcalib);
```

# You’re more than halfway through your application of MVEs to ENM approaches! Save model calibration in .csv to the path or your choosing and then continue on to the final step of the process – model projection.

```
write.csv(Mcalib, “<YourAwesomeMVEmodelCalibrationFilenameHere.csv>”);
```

--------------- MODEL PROEJCTION ---------------

# Again, a regularised grid is necessary to apply the defined ellipsoids to every point in the projection region.

### If the model projection region is geographically different from the model calibration region, create a regularised grid of the full projection region at the spatial resolution of the environmental data (remember to add XY coordinates to labels).

### If the model projection region is geographically the same as the model calibration region (e.g. if model projection is to different time periods only), you can use the same grid generated for model calibration.

# Read in the regular grid

```
randT_fullproj = read.dbf("<path to regular grid of projection area>.dbf");
```

# Extract environmental data from the raster stack to the calibration region grid. NOTE: there will be A LOT will be NAs.

```
vrsT_fullreg = extract(layers, randT_fullproj[, 2:3]);
head(vrsT_fullreg);  # check that everything read in and extracted properly
```

# For shits and giggles, let's calculate the percentage of NAs again.

```
vrsTsna_full = na.omit(vrsT_fullreg);
pNA_full = dim(vrsTsna_full) / TotalNumberPixelsInGrid;
pNA_full;
```

# Create the matrix that will contain the distance of environment to centroid. The matrix size will be equivalent to the dimension of 'randT_fullproj' (e.g. TotalNumberPixelsInGrid x 1).

```
dT_full = matrix(0, ncol = 1, nrow = TotalNumberPixelsInGrid);
```

# Calculate environmental distance of each ellipsoid from the centroid

```
for(i in 1:TotalNumberPixelsInGrid)dT_full[i, 1] = maja(vrsT_fullreg[i, ],
mu1, invs1);
```

# Check that it worked. The resulting table should have the following: longitude, latitude, one column for each environmental variable and a column that is dT_full.

```
modProj = cbind(randT_fullproj, vrsT_fullreg, dT_full);
head(modProj);
```

# Congratulations! You've now completed your application of MVEs to ENM approaches. Save full projection as .csv to the pathway of your choosing!

```
write.csv(modProj, "<YourAwesomeMVEmodelProjectionFilenameHere>.csv");
```