TITLE: Evaluating 318 continental-scale species distribution models over a 60-year prediction horizon: what factors influence reliability of predictions?

SHORT TITLE: Temporal transferability of species distribution model predictions

Authors: Alejandra Morán-Ordóñez [1,2], José J. Lahoz-Monfort[1], Jane Elith[1], Brendan A. Wintle[1]

[1] School of BioSciences, The University of Melbourne, VIC 3010, Australia

[2] Centre Tecnològic Forestal de Catalunya (CTFC), Ctra. Antiga St. Llorenç km 2, 25280 Solsona, Spain.

Email addresses: alejandra.moran@ctfc.es; jose.lahoz@unimelb.edu.au; j.elith@unimelb.edu.au; brendanw@unimelb.edu.au

Current details corresponding author: Alejandra Morán-Ordóñez, CTFC Centre Tecnològic Forestal de Catalunya, Ctra. Antiga St. Llorenç km 2, 25280 Solsona, Spain. Telephone: 973481752 - Ext. 330, email: alejandra.moran@ctfc.es

Keywords: AUC, geographic range, MaxEnt, predictions into the future, presence-only data, rarity, sample size, species distribution models, species traits, temporal transferability.

ABSTRACT

Aim. Species Distribution Models (SDMs) are currently the most widely used tools in ecology to evaluate suitability of environments for biodiversity in the face of future environmental change. In this study, we seek to provide an assessment of the predictive performance of SDMs over time. How well do SDMs predict to future time periods and what factors influence predictive performance?

Innovation. We used a historical spatially-explicit database of 1.8M occurrence records for 318 tetrapod species from across continental Australia over the period 1950-2013. We fit distribution models for each species to data from four multi-decadal time-slices and used these to predict the species distributions up to 60 years after the data collection period for the fitted models. We evaluated predictions against observed data from the relevant time period. Predictions were made assuming either complete knowledge of changes in climatic and environmental conditions or assuming the environment and climate remained unchanged between the fitting and evaluation time periods. We used generalized linear mixed models to model variation in the predictive performance of SDMs over time in relation to a variety of factors, including the length of time between fitting and evaluation, species traits, taxonomic group, and attributes of the dataset used to fit models.

Main conclusions. We found that most models provided useful predictions even when the period between model fitting and evaluation was 60 years (AUC > 0.7 in 80% of the species evaluated). Variation in predictive performance over time were strongly related to the species range breadth (models for species with broad geographic ranges tended to perform worse than models for locally restricted species) and to the environmental coverage of occupancy data. Conversely, taxonomic group, habitat preferences and body size were not highly influential in describing the variation in predictive performance over time.

INTRODUCTION (A)

Species distribution models (SDMs) are widely used in ecology to explain variation in a species' occupancy in response to variation in environmental conditions (Franklin, 2010). SDMs can be used to make spatial predictions about the probability (or relative likelihood) of a species' occupancy at locations under present or future anticipated environmental conditions (Elith & Leathwick, 2009; Franklin, 2010). Making good predictions about the future availability and spatial distribution of suitable environments for species is of fundamental importance for the prioritization of biodiversity conservation efforts (Guisan *et al.*, 2013; Guillera-Arroita *et al.*, 2015).

However, there are many reasons to be cautious about the ability of SDMs to accurately project the future spatial distribution of a species. SDMs can perform well when modelling a stationary process, when the environmental conditions onto which projections will be made are covered by the data used to train the models (i.e. the projections do not extrapolate to un-sampled environments) and when the fitted response of the dependent variable to environmental gradients is not confounded by un-modelled gradients influencing occupancy (Elith & Leathwick, 2009). There is understandable concern about whether these conditions can realistically be met when projecting SDMs into distant future conditions, including non-analogue climates (Dormann, 2007; Fitzpatrick & Hargrove, 2009).

Given such concerns, and considering the number of highly influential studies that utilize SDM projections, there are relatively few studies that assess the predictive accuracy of SDM projections over long time periods. Studies attempting to assess the temporal transferability of SDM predictions either utilize relatively short time lags between model development and evaluation, and/or include a restricted number of species and functional groups such that generalizations about the factors leading to accurate or inaccurate projections are limited (Araújo *et al.*, 2005; Pearman *et al.*, 2008; Kharouba *et al.*, 2009; Dobrowski *et al.*, 2011; Rubidge *et al.*, 2011; Rapacciuolo *et al.*, 2012; Eskildsen *et al.*, 2013; Watling *et al.*, 2013; MCfarland *et al.*, 2015). This is due in part to the difficulty of finding sufficiently large empirical datasets collected over long enough time periods for a wide enough range of species such that generalizations can be drawn. In studies that have attempted to assess the

temporal transferability of SDM predictions, species traits including niche specialization and dispersal ability have been identified as likely drivers of temporal and geographic transferability (Menéndez *et al.*, 2006; Dobrowski *et al.*, 2011; Eskildsen *et al.*, 2013), though the magnitude of those effects remains largely unquantified and highly uncertain.

Drawing strong statistical inference about a complex phenomenon such as predictive performance is difficult. Many small, additive and interacting effects are likely to determine predictive performance. Therefore, large sample sizes are required to allow comparison of predictive performances within and between time periods and to analyse the factors that drive variation in predictive performance over time. We have not been able to identify any study that compares within (cross-validated) and between time period model predictive performances for many species over long time periods at a continental scale, drawing inference about the drivers of performance. Yet many unresolved questions remain about the causes of variation in the spatial and temporal predictive accuracy of SDMs. Understanding what drives degradation of predictive performance over time, or conversely what predisposes a model to high predictive performance, would help to determine whether any given model prediction should be trusted or not in any given application.

In this study we seek to evaluate whether the predictive accuracy of SDMs changes as the time horizon over which models are projected increases, and to investigate factors influencing variation in predictive accuracy. To date, published evidence points to effects on predictive performance into the future from factors including the elapsed time over which a prediction is made (Pearman *et al.*, 2008), the degree of spatial and temporal autocorrelation of data (Segurado & Araujo, 2004; Araújo *et al.*, 2005; Swanson *et al.*, 2013) and the degree of extrapolation of predictions into novel environments (Araújo *et al.*, 2005; Dobrowski *et al.*, 2011). Further studies suggest impacts on the quality of current SDM predictions from sample size (Hernandez *et al.*, 2006; Wisz *et al.*, 2008), habitat distinctiveness (Brotons *et al.*, 2004; McPherson & Jetz, 2007; Pöyry *et al.*, 2008; Kharouba *et al.*, 2009; Morán-Ordóñez *et al.*, 2012), range size (Segurado & Araujo, 2004; Guisan *et al.*, 2007; McPherson & Jetz, 2007) and species body size (Seoane *et al.*, 2005; McPherson & Jetz, 2007; Pöyry *et al.*, 2008) among other factors. Presumably, through their impact on the quality of the fitted model, some of these

factors will also impact the accuracy of predictions into the future. Here we focus on the following questions about predictive accuracy: (1) Does the predictive accuracy of SDMs decrease as the time lag between prediction and evaluation increases, and if so, by how much? (2) Which measurable features of the species observation data used to fit and evaluate models most influence the predictive accuracy of SDMs? (3) Do quantifiable ecological features of the modelled species influence SDM predictive accuracy?

METHODS (A)

Occurrence data (B)

We based our analyses on an extensive dataset of presence-only records gathered for all terrestrial native tetrapods (mammals, reptiles, amphibians and birds) across Australia. Species occurrence data were accessed between May and August 2013 from the Atlas of Living Australia (http://spatial.ala.org.au/) and individual agencies in the Australian states and territories (see acknowledgments). We filtered the dataset to retain only those spatially-valid records collected from 1950 onwards, with maximum point location error of less than 1 km. We followed an iterative approach to maximize the number of time periods (replicates) and species available to test for temporal transferability of model predictions, whilst requiring a minimum of 30 records for each species in each time period (Hernandez *et al.*, 2006; Wisz *et al.*, 2008): we defined iteratively a range of decadal and multi-decadal time divisions and evaluated the number of presence observations available for each species within each of the defined time periods after removal of duplicate records (each cell contained only one record of each species within a time period). The reduced dataset that satisfied our criteria consisted of 1,887,653 records belonging to 255 bird, 48 mammal, and 33 reptile species available for modelling across four time periods: *t1* (1950-1980), *t2* (1980-1990), *t3* (1990-2000) and *t4* (2000-2013) (Fig. 1). Due to data scarcity for amphibians in the first time period (*t1*), modelling analyses for this taxonomic group were restricted to 31 species, and only for the last three time periods (*t2*, *t3* and *t4*).

Environmental predictors (B)

We compiled an initial set of environmental variables that could be of ecological relevance for all species and across all trait groups. We generated 19 mean annual climatic variables for each time period considered in the analyses at 1km grid cell resolution across the whole Australian continent. Note that these are based on observations (i.e. interpolated from data of meteorological stations distributed across the continent; Appendix S1) and are not predictions from general circulation models. Because climatic variables were not topographically downscaled (McVicar *et al.*, 2007), we included predictors representing elevation diversity at the same resolution (to account for combined effects of topography and climate that could be of relevance for the modelled species), along with land cover and a variable accounting for water availability (one of the main limiting environmental factors for species distributions in Australia where 70% of the land is either arid or semi-arid; Young, 2000) (Appendix S1). We calculated Pearson's pairwise correlations and Variance Inflation Factors between all environmental predictors on a random sample of 100,000 points across the continent, retaining a subset of variables with maximum pairwise correlation of 0.7 at all time periods (Tabachnick & Fidell, 1996; Dormann *et al.*, 2013) (Appendix S1). Only four climatic variables were retained, representing climate variability and extremes (Appendix S1 & S2): temperature seasonality, mean temperature of the coldest quarter, precipitation seasonality and precipitation of the driest quarter.

Land clearing is one of the main drivers of habitat loss in Australia over the last 60 years (Young, 2000; Bradshaw, 2012). Currently cleared areas (e.g. agriculture, mining, urban areas) were masked out of all models in all time periods in order to minimize the effect of habitat loss on both the species records and the apparent predictive performance of distribution models (Fig. 1). Two land cover variables (forest and grassland cover) were included in SDMs. However, these were assumed constant over the 60 year period covered by this study since there is not sufficiently accurate mapping of land-cover change in Australia over the full period 1950-1990.

Modelling species distributions (B)

We modelled the distribution of all tetrapod species using MaxEnt (version 3.3.3k; Phillips *et al.*, 2006; Phillips & Dudík, 2008), a machine learning method designed for dealing with presence-only data (Elith *et al.*, 2006, 2011) while taking into account the distribution of environmental predictors in the background area of analysis. For each species we built four sets of MaxEnt models using observation data collected within each of the four nominated time periods (*t1*, *t2*, *t3*, *t4*: Fig. 1). Exploratory analyses showed that species records were biased towards urban areas, roads and in general, areas of high accessibility. Biased survey data can lead to environmentally and geographically biased predictions that might reflect sampling effort rather than true distributions across the study area (Phillips *et al.*, 2009; Kramer-Schadt *et al.*, 2013; Syfert *et al.*, 2013). To reduce the effect of sampling bias on SDM predictive performance, we provided the background points to MaxEnt in such a way as to mimic the sampling bias of the occurrence records (Phillips *et al.*, 2009; Syfert *et al.*, 2013) by using as background for analysis all available records (including those for the species of interest) for the same taxonomic group at the same time period (e.g. all birds at *t1*). This approach has been coined the "target-group background" approach (Phillips *et al.*, 2009), and has been shown to perform relatively well in dealing with sampling bias (Syfert *et al.*, 2013; Fithian *et al.*, 2015). We controlled the complexity of the response shapes by allowing only linear, quadratic and product features in the models. These are similar to linear, quadratic and interaction terms in regression models. Models with these restricted feature types will be smoother than those fitted with Maxent's default settings, less prone to fitting idiosyncrasies of the data, and potentially better at predicting to new times and places (Merow *et al.*, 2014). We also toggled off the default setting that adds samples to background points, because the presence points are already added to the background samples we created. Default values were used for the other MaxEnt settings.

Model evaluation (C)

Models fitted in each time period were used to generate species distribution predictions ("*projections*") into future time periods (Fig. 1). Predictive performance was assessed in terms of discrimination ability. This metric is suited to presence-background data, since calibration cannot be assessed and thresholding predictions loses information (Guillera-Arroita *et al.*, 2015). Discrimination

was measured using the area under the receiver-operator characteristic curve (AUC; Hanley & McNeil, 1982), adapted for its use with presence-background data (Phillips *et al.*, 2006). AUC values for models that were evaluated within the time period in which they were fitted were calculated using the ten-fold cross-validation procedure provided in MaxEnt (using as background the target-group background data set of the corresponding time period). The cross-validation was only used for evaluation; the final fitted model for projection was based on all training data. AUC values for projections into 'future' time periods were calculated by comparing the projections in those periods (e.g. projections made from *t1* to *t4*) with the data –both presence and target-group background data - collected in those 'future' time periods (e.g. *t4*), providing an independent test of predictive performance (Fig. 1). We note that we used presence-"target-group background" data for evaluation because presence-absence data were not available. This brings with it challenges that we will discuss later.

We used boxplots and conducted pairwise comparisons of AUC using Wilcoxon's rank sum and signed test to evaluate differences in models' predictive performance between taxonomic groups (birds, mammals, reptiles, amphibians) within the same time period and across time, respectively.

Bounding prediction scenarios (C)

We tested two scenarios addressing the impact of uncertainty about 'future' environmental conditions on model predictive performance. For convenience, we focused on uncertainty about medium-term (30–60 year) temporal variations in climate, though our approach could be used to assess the impact of uncertainty about the future state of any independent environmental variable. At one extreme (of optimism), we assume that the modeller acting in any time period had perfect knowledge of how the future climate would unfold. In our study, we implemented this assumption by projecting models built in one time period onto the *observed* environmental conditions in the latter time periods (Appendix S2). At the other end of the uncertainty spectrum, we assumed that nothing was known about future conditions or likely change and, as such, climatic conditions contemporary to the model fitting data were assumed to persist into the time periods for which projections were made. Those projections

were then evaluated with observation data from the periods into which projections were made. While it is possible that a worse scenario could emerge if modellers attempted to project future environmental conditions and got it more wrong than had they assumed no change, we expect that our scenarios bound the bulk of circumstances under which our projections are made.

Assessing temporal transferability (B)

Species were omitted from the study of the factors driving SDM predictive performance if their predictive performance within the first time period (evaluated using cross-validation) was below our chosen threshold of AUC < 0.7. This was based on the notion that models exhibiting relatively low 'within-period' predictive performance almost certainly would perform poorly when extrapolated to new observations in space *and* time and therefore, these would be unlikely to be used for predicting the future distribution of a species (see species retained and data summaries in Appendix S3). We note that whilst there are guidelines for interpretation of AUC values with presence-absence data (Pearce & Ferrier, 2000), the situation is less clear for presence-background data. For convenience, from here on we refer to AUC < 0.7 as 'poor performance' and AUC > 0.7 as 'useful' performance (sensu Swets, 1988).

Generalized Linear Mixed Models (GLMM; Bolker *et al.*, 2009) were used to describe variation in the predictive performance of models (i.e. discrimination ability measured using AUC) as a function of: (i) intrinsic aspects of distribution data (e.g. sample size, the geographic spread of presence data); (ii) time lag (years) between collection of model fitting data and observation data used to evaluate models; and (iii) ecological traits of the species (full description of predictors in Table 1). Models describing variation in AUC were developed for each taxonomic group individually as well as for all species combined. In all cases, species was treated as a random effect; this controls for species-specific idiosyncrasies that remain constant over time. We used the Akaike's Information Criterion (AIC; Burnham & Anderson, 2002) to select the most parsimonious model explaining variation in AUC (the model explaining the largest variance using the minimum set of predictor variables; detailed

methods in Appendix S4). GLMM were fitted using the *lme* function in the "nlme" package in R (Pinheiro *et al.*, 2013). We provide the data used to fit GLMM in Table S10.

RESULTS (A)

Model performance varies over time and between species groups (B)

After reducing the number of species used to test model transferability by removing those that performed poorly in *t1* and *t2* cross-validation testing, 217 birds, 40 mammals, 31 reptiles and 30 amphibians remained in the dataset (Appendix S3). The average AUC value obtained from MaxEnt models across all these species and over all time periods (including cross-validated testing within the same time period) was 0.834 ± 0.002 (mean ± sd). Overall, models trained in *t1* (or *t2* for amphibians) and evaluated using a temporally independent dataset from the later periods, showed lower predictive performance than expected based on 'within-period' cross-validation results (Fig. 2, Appendix S6). These differences were not observed between cross-validation estimates at *t3* and projection performance evaluations of *t3* models predicted to *t4* for any of the species groups except amphibians (Appendix S6). In general, predictive performance decreased as the time span between model training and model evaluation increased (Fig. 2). For all species groups, the lowest predictive performances were obtained for models built on *t1* data that were then projected to *t4* and evaluated using *t4* observation data. However, only 20% of the species modelled in *t1* and 10 % of the species modelled in *t2* had weak transferability to *t4* (AUC values for projection from *t1* or *t2* to *t4* < 0.7). Reptiles were the only group for which there were no significant difference between the cross-validated predictive performance (within a period) and the between-period predictive performance measured using AUC (Fig. 2).

On the whole, most species in all the species groups produced models that made useful predictions to future time periods (approx. 80 % of models showed AUC > 0.7 when transferred from *t1* to *t4*; Fig. 2). The largest AUC reduction was observed in birds (e.g. $\Delta AUC_{t1\text{-}t4} = 0.41$ for *Cincloramphus mathewsi*). The greatest variability in predictive performance between cross-validated models was

observed in birds, which in certain time periods had significantly lower predictive performance than both mammals and reptiles (Wilcoxon's rank sum tests, $P < 0.01$ in all cases, $\alpha = 0.05$; Appendix S6).

Lack of knowledge about future climatic conditions was approximated by naively assuming climate was constant between fitting and projection (rather than using the known climate at each step). This strongly impacted predictive performance for reptiles, but not for other groups (Fig. 2 and Appendix S7). Based on this observation, and given that we were primarily interested in exploring the impact of other factors such as data quality and species traits on predictive performance, we chose to conduct the remaining analysis using models and projections based on the scenario in which climates were known for all time periods. This allowed us to control for the effect of climate variations on models' predictive performance.

Factors affecting transferability of predictions (B)

Regression analysis provided insights into the factors related to variation in model predictive performance (AUC). These factors explained between 27 and 52% of the variation in AUCs depending on the taxonomic group modelled (Table 2).

Of the full set of covariates considered, only the number of biogeographic sub-regions in which a species occurred (BR) was consistently selected as an important predictor for each taxonomic group, showing a strong negative effect on performance in all instances that generally flattened off at high BRs (Table 2, Figure 3a & Figure S8.1). This effect was moderated by the time lag between training and evaluation data (time since training; TST) and the ratio between the training and background data of the training data set (TR), as evidenced by the significant negative interaction between the terms (Fig. 3f, 3g, Fig. S8.1). A combination of larger ranges (higher BR values), longer time-lags and higher training ratios tended to be associated with the lowest AUC values.

The set of explanatory variables selected for the most parsimonious model varied between species groups, though covariates showed consistent effects on the response variable when selected (Figure S8.1). For example, the larger the time interval between model training and model evaluation (TST) or the larger the ratio between the number of training points vs. number of background points

(training ratio, TR) the smaller the AUC values tend to be (Figures 3b, 3c and Fig. S8.1). The proportion of total environmental range sampled by the training dataset (PES) had a positive effect on AUC values, with the effect tending to level off at larger values (i.e. when approximately more than 60 % of the environmental range in which a species has been observed in was covered by the training data set) (Figures 3d, Fig S8.1). This variable showed a significant interaction with the number of biogeographic sub-regions in the reptiles' models. The predictive performance of models for reptiles occurring in a few biogeographic regions increased significantly as the proportion of their total environmental range sampled increased; this effect was not significant at large numbers of biogeographic sub-regions (Fig S8.4d). Body size had a positive effect on predictive accuracy for mammals but it did not contribute significantly to explain the variation in AUC in any other taxonomic group model (Figure S8.1). Forest-dependent species had lower AUC values than those preferring woodlands or shrublands (Fig. 3e, Appendix S8). Species classified as 'grassland' and 'wetland' dwellers were less common in the dataset and produced models with more variable performance (Fig. 3e, Appendix S8). Not surprisingly, the model fitted to the AUCs for all species together showed qualitatively similar structure and response shapes to those fitted separately to the AUC data from each of the taxonomic groups (Fig. 3, Figures S8.2 to 8.5). Taxonomic group was, in itself, not an important predictor of AUC (Table 2). The combined taxon model explained 42% of the variation in AUC across all species.

Datasets used from earlier time periods had less complete environmental coverage than those from more recent time periods (Fig. 4, Appendix S5). At *t1* and *t2*, most of the occurrence records were concentrated towards the eastern, south-eastern and south-western coasts of Australia, whereas the rest of the continent appeared to be largely un-sampled (or records were unavailable). The training datasets used to fit the models at early time periods covered an average of 10-25% of the total suitable environmental ranges currently known for the species. These proportions increased significantly at later time periods (60 % and 90 % at *t3* and *t4* respectively) linked to an increase in the geographic extent of the available samples towards the drier interior of the country. This, together with the results of the GLMM, indicates that incomplete sampling of species ranges could have been an important

factor in the apparent decline of predictive performance as the time lag between fitting and evaluation datasets increased.

DISCUSSION (A)

We found that SDMs for the bulk of the 318 tetrapod species modelled here performed fairly well – as evidenced by AUCs > 0.7 in around 80% of the models tested – in forward predicting species distributions up to 60 years into the future. This a promising result given that, in many cases, the occupancy data used to fit the models did not sample a large proportion of the geographic and environmental space that the species are now known to occupy. To the extent that we can test predictions with available data, the evidence suggests that the models were fitted well enough (with relevant predictors and fitted relationships) to predict reliably and with fairly stable predictive performance over the time-span studied.

The effect of species geographic extent and habitat specificity (C)

SDM predictive performance was most strongly (negatively) influenced by the geographic extent of the species (embodied here by the number of biogeographic sub-regions, estimated on all records for the species over all time periods). Models for species with broad geographic ranges tended to perform worse than models for locally restricted species (as previously observed in other studies: Guisan & Hofer, 2003; Brotons *et al.*, 2004; Guisan *et al.*, 2007; Morán-Ordóñez *et al.*, 2012). This factor remained consistently important across major taxonomic groups and itself explained the most variation in AUCs between species. Predictions for narrow-range species (records restricted to a small number of biogeographic sub-regions) were significantly better than those of widespread species (Fig.3a; Fig S5.3). This predictor is, most likely, acting as a proxy for the rarity, range and habitat specificity of the species (McPherson & Jetz, 2007). The impact of broad geographic range on predictive performance may arise from the fact that, for a given landscape, a widespread species is likely to have less restrictive, and more difficult to map environmental requirements than a specialist or narrow-range species. However if a species was once widespread but now has a narrow range due to over-harvesting, persecution or other dynamic factors that cannot easily be mapped, it will not

necessarily be well modelled or its distribution easy to predict. SDMs for widespread taxa are less likely to identify sharp environmental thresholds, that clearly delineate the most suitable environmental conditions; this naturally limits the model's ability to discriminate between suitable and unsuitable locations, leading to lower AUC values (Brotons *et al.*, 2004; McPherson *et al.*, 2004; Morán-Ordóñez *et al.*, 2012). This is not a fault with AUC as a measure of performance, but simply a reflection of the challenges associated with modelling and predicting distributions of widespread species. Previous findings about the impact of rarity on predictive performance were based on more robust presence-absence evaluation data (e.g. Elith *et al.*, 2006), but are consistent with this result. A contribution of this study is to confirm the result using presence- target-group background data for evaluation, a larger sample of species, and testing over a longer time lag between model fitting and evaluation. We do note, however, the result may partly be confounded by characteristics of our data, addressed next.

The effect of training and testing sample characteristics (C)

The number of training samples available in each time period (covariate 'training ratio') impacted AUCs, with high values of training ratio leading to lower discrimination. Given that we restricted modelling to species with more than 30 records (to avoid having to deal with the impacts of extremely few samples) this result may reflect the difficulty in modelling widespread (often generalist) species, for which we naturally tended to have a larger number of records. The interaction estimated between the training ratio and the number of biogeographic sub-regions in which a species occurs indicates that widespread and common species with large numbers of presence records in the training sample are particularly difficult to model (Fig 3g). We cannot be sure of the explanation for this due to our evaluation method. Because training ratios are moderately correlated with evaluation ratios ($r \sim 0.5$), characteristics of the evaluation data could be impacting this result. It is known that the maximum possible AUC varies across presence-background datasets because presence and background records overlap (Phillips *et al.*, 2006; Smith, 2013). This is a major disadvantage when evaluating a model with presence-background data (Jiménez-Valverde, 2012). However it is clearly not the main driver of our results given that, in our analyses, the covariate 'evaluation ratio' was not selected as the most

influential predictor in explaining variation in AUC. Additionally, because our analysis is based on the GLMM which treats species as a random effect, this impact will be somewhat ameliorated. Moreover, we cannot account for potential changes in species prevalence (or in evaluation ratio, if different) over time. Our inference would be impacted by a hypothetical systematic change in prevalence across species that was correlated with one of our explanatory variables, but this is unlikely to occur over the large pool of species included in our analysis. It is more likely that any real change in prevalence (or evaluation ratio) over time is random with respect to species, and therefore adds noise rather than bias to our GLMM analysis. Nevertheless, we see this as an ongoing challenge with presence-background evaluation that is worthy of further attention.

Cross-validated AUCs may be artificially inflated because they are only tested on the extent of data available at the time. A more rigorous representation of true predictive performance arises when predictions are evaluated using a more complete data set (collected at a later time). Such evaluations are expected to reveal poorer predictive performance (Fig. 3g, Fig. 4, Appendix S5 & S6). The relationship between AUC and proportion of environmental sampled by the data training set (PES; Fig 3d) demonstrates the expected positive link between predictive performance and environmental coverage (how well observations in a given period sample the environmental space of all observations from all periods), though the relationship does level off at higher values of PES, perhaps suggesting some level of saturation in the coverage required. In any case, the results emphasize that good environmental coverage of sampling effort can be as important as the overall amount of observation data used (Hortal *et al.*, 2008; Syfert *et al.*, 2013).

Positive temporal autocorrelation in the environmental conditions sampled by model training records could inflate the apparent predictive performance of SDMs built and tested within adjacent periods (Araújo *et al.*, 2005). In our observation datasets, temporal autocorrelation in raw observations (as evaluated with Euclidean distances between environments sampled by training and testing data) decays as the time gap between model building and model testing increases (Appendix S9). This could be driven by changes in environmental conditions that are themselves driving the distributions of species, or changes in observation effort over time. Changes in distributions due to changing

environmental conditions should not impact the predictive performance if the variation in environmental conditions is adequately modelled within the SDMs. Climatic changes were well modelled within our SDMs because we were in the unusual position of 'knowing' the climate when the model-training observation data were collected *and* when the model-testing observation data were collected, supporting the notion that observation processes, rather than changing environmental conditions are driving variation in AUC over time.

The impact of environmental change in the form of vegetation (habitat) clearing on SDM predictive accuracy was effectively masked out of the analysis undertaken here via our choice to include only records that intersect extant vegetation at the most recent year of our analysis (2013). By doing this, we largely avoided the direct impact of habitat loss on SDMs apparent predictive accuracy, though we acknowledge that landscape level effects of fragmentation *could* have impacted on species in areas of extant vegetation close to cleared areas and, therefore on the predictive accuracy of our models. As habitat loss and climate change were largely controlled in our analysis, we believe that reduced predictive performance with increased temporal lag between fitting and evaluation of SDMs is largely a result of the more comprehensive sampling of environmental space achieved in later years (Figure 4), as observed in other studies (Tuanmu *et al.*, 2011).

The role of unmodelled change in environmental conditions (C)

Despite the fact that the overall range of climate conditions did not change much at a continental scale over the 60 year period evaluated in this study (extreme values of the climatic predictors remained mostly unchanged), there were some minor regional and local climatic geographic shifts in climate (Appendix S2). However, for all taxonomic groups except reptiles, having perfect knowledge of how the climate changed afforded no appreciable advantage in terms of predictive performance compared with assuming (incorrectly) that no climate changes took place. However, that does not indicate that the larger climatic changes expected during this century will have a similarly innocuous effect on our ability to predict species' future distributions. For reptiles, accounting for changing climates significantly improved ($p<0.05$) the predictive performance of SDMs compared with naively

assuming no change. The variable that shifted the most over the 60 years of the study was Bio15 (rainfall seasonality), and this was selected as the most important predictor in all the reptile models, which may explain why it was important to utilize contemporary (to the model testing observation data) climate maps in order to achieve good predictive performance for reptiles. Further, the sensitivity of ectotherms to small climatic changes, and the capacity of many reptiles to track changing climatic conditions may partly explain why they were the only group for which SDM performance was sensitive to climatic changes between time periods. However, understanding the processes behind the sensitivity of this group to changing climates would require a more detailed, fine scale analysis than we present here.

The role of species ecological traits and habitat preferences (C)

Of the three species traits modelled, body size was the only trait to have any consistent influence on predictive performance. For mammals, this might be due to the fact that larger species are more conspicuous, increasing data availability and quality, which in turn can increase model performance (Seoane *et al.*, 2005). The positive effect of being 'large' may be counteracted to some extent by the positive correlation between body size and the geographic range size of a species, which appears to be associated with relatively poor predictive performance in this study. For mammals, the number of biogeographic sub-regions with observation data was correlated with body size ($r = 0.41$); correlation was weaker for other taxonomic groups ($r \leq 0.2$). The best predictive mammal models are likely to arise from large-bodied species with small geographic ranges. However, compared with the influence of data quality, the importance of species traits appears to be, at best, marginal.

Several possible explanations exist for the apparently lower performance of models for forest dependent species. It is possible that a greater number of wide-ranging generalist species were primarily recorded in forests in our dataset, or that forest species observation data are somehow less reliable due to larger numbers of records from inexperienced observers or variation in species detectability. It is also possible that the cleared areas that we excluded from our study were important former habitats of forest-dependent species. However, no particularly compelling hypotheses emerge.

Methodological challenges addressed and future opportunities (C)

A major limitation in generating inference about the drivers of variation in SDM predictive performance is a general lack of good quality biodiversity point data collected over large areas and long time periods. Other studies conducted at finer scales, usually utilizing far fewer species, and often over much shorter temporal scales, have managed to evaluate the performance of SDM projections with presence-absence data (Araújo *et al.*, 2005; Pearman *et al.*, 2008; Kharouba *et al.*, 2009; Dobrowski *et al.*, 2011; Rubidge *et al.*, 2011; Rapacciuolo *et al.*, 2012; Eskildsen *et al.*, 2013; Swanson *et al.*, 2013; Watling *et al.*, 2013), potentially providing a more precise picture of predictive performance in the specific context of their study, but possibly with lower generality due to the their reduced scope. Presence-absence data are less representative of the typical data used to fit SDMs for most applications. Despite being restricted to utilizing presence-only observation data, here we have the data volumes and time series that allow us to make generalizations about what drives variation in the performance of SDMs in predicting future distributions. Clearly the use of presence-only data to examine SDM predictive performance created technical challenges including: data compilation and cleaning, correction of spatial bias, selection of adequate model settings, selection of covariates with adequate temporal resolution and interpreting model evaluation results. Testing the drivers of predictive performance using a presence-absence (observed/not-observed) dataset with wide taxonomic, spatial, and temporal coverage is not currently possible and remains a challenge for the future. If this could be achieved, it would make a significant contribution to understanding of the performance and limitations of SDMs for prediction.

Finally, we have not dealt directly with the relationship between drivers of long-term predictive performance and the drivers of predictive performance when projecting to the same, or other parts of geographic or environmental space within a time period. Teasing apart those effects was beyond the scope of this study. However, several of the studies we cite assessed the predictive performance of models over geographic and environmental space (or over short time lags), and those studies identify predictors of performance that seem quite well matched to those we found to be relevant for long-term predictive accuracy such as specificity and generality of habitat requirement and the species range,

and geographic and environmental coverage of observation data (e.g. Thuiller *et al.*, 2004; Guisan *et al.*, 2007; Pöyry *et al.*, 2008; Wisz *et al.*, 2008; Morán-Ordóñez *et al.*, 2012). On the other hand, the question of whether the species can be considered to be in equilibrium with its environment over the period between model fitting and evaluation does seem to be an issue of particular relevance to long-term prediction (Nogués-Bravo, 2009). This influence may manifest in our result for the lizard species, where significantly declining predictive performance could not be easily explained by the GLMM analysis of static factors. However, lizard model predictive performance did appear sensitive to our assumptions about static or changing climates. A rigorous test of causation is not possible with the data available to this study, but appears to warrant further investigation.

While many questions remain unanswered, this study has provided significant insights about the circumstances under which strong or weak predictive performance might be expected based on empirical evidence drawn from a temporally and geographically extensive biodiversity data set. On the basis of our findings, it is clear that the robustness of predictions can be enhanced through efforts to ensure models are built on data that sample well the environmental and geographic space from which the species is known to exist, and increased efforts to identify and map the drivers of wide-spread and generalist species.

ACKNOWLEDGMENTS(A)

Data custodians: Department of Land Resource Management of Northern Territory; Department of Environment and Primary Industries of Victoria; New South Wales office of Environment and Heritage; Department of Environment and Land Conservation of Western Australia, and the Department of Environment and Heritage protection of Queensland. G. Guillera-Arroita, M. McCarthy, M. Kearney, M. Bode, R. Fuller, G. Luck, G. Heard, A.Whitehead, R. Tingley, contributed discussion, ideas and data. This work was supported by the ARC Centre of Excellence for Environmental Decisions (CEED) and the National Environment Research Program Environmental Decisions Hub (NERP ED). BW and JE were supported by ARC Future Fellowships (FT100100819, FT0991640).

REFERERENCES (A)

Anstis, M. (2013) *Tadpoles and Frogs of Australia*, New Holland.

Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. (2005) Validation of species–climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.-S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, **24**, 127–135.

Bradshaw, C.J.A. (2012) Little left to lose: Deforestation and forest degradation in Australia since European colonization. *Journal of Plant Ecology*, **5**, 109–120.

Brotons, L., Thuiller, W., Araújo, M.B. & Hirzel, A.H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*, Springer, New York, USA.

Cogger, H.G. (2000) *Reptiles & amphibians of Australia*, 6th ed. Reed New Holland, Australia.

Dobrowski, S.Z., Thorne, J.H., Greenberg, J.A., Safford, H.D., Mynsberge, A.R., Crimmins, S.M. & Swanson, A.K. (2011) Modeling plant ranges over 75 years of climate change in California, USA: temporal transferability and species traits. *Ecological Monographs*, **81**, 241–257.

Dormann, C.F. (2007) Promising the future? Global change projections of species distributions. *Basic and Applied Ecology*, **8**, 387–397.

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.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R. & Lehmann, A. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

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.

Eskildsen, A., Roux, P.C., Heikkinen, R.K., Høye, T.T., Kissling, W.D., Pöyry, J., Wisz, M.S. & Luoto, M. (2013) Testing species distribution models across space and time: high latitude butterflies and recent warming. *Global Ecology and Biogeography*, **22**, 1293–1303.

Fithian, W., Elith, J., Hastie, T. & Keith, D. a. (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, **6**, 424–438.

Fitzpatrick, M.C. & Hargrove, W.W. (2009) The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation*, **18**, 2255–2261.

Franklin, J. (2010) *Mapping species distributions: spatial inference and prediction*, Cambridge University Press.

Greer, A.E. (1997) *The biology and evolution of Australian snakes*, S. Beatty.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R. & Wintle, B.A. (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, **24**, 276–292.

Guisan, A. & Hofer, U. (2003) Predicting reptile distributions at the mesoscale: Relation to climate and topography. *Journal of Biogeography*, **30**, 1233–1243.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P. & Buckley, Y.M. (2013) Predicting species distributions for conservation decisions. *Ecology Letters*, **16**, 1424–1435.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S. & Peterson, A.T. (2007) What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, **77**, 615–630.

Hanley, J.A. & McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hernandez, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M. & Baselga, A. (2008) Historical bias in

biodiversity inventories affects the observed environmental niche of the species. *Oikos*, **117**, 847–858.

Jiménez-Valverde, A. (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, **21**, 498–507.

Jones, K.E., Bielby, J., Cardillo, M., Fritz, S.A., O'Dell, J., Orme, C.D.L., Safi, K., Sechrest, W., Boakes, E.H. & Carbone, C. (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184. *Ecology*, **90**, 2648.

Kharouba, H.M., Algar, A.C. & Kerr, J.T. (2009) Historically calibrated predictions of butterfly species' range shift using global change as a pseudo-experiment. *Ecology*, **90**, 2213–2222.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H. & Wilting, A. (2013) The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, **19**, 1366–1379.

Luck, G.W., Carter, A. & Smallbone, L. (2013) Changes in Bird Functional Diversity across Multiple Land Uses: Interpretations of Functional Redundancy Depend on Functional Group Identity. *PLOS ONE*, **8**, e63671.

Luck, G.W., Lavorel, S., McIntyre, S. & Lumb, K. (2012) Improving the application of vertebrate trait-based frameworks to the study of ecosystem services. *Journal of Animal Ecology*, **81**, 1065–1076.

MCfarland, T.M., Grzybowski, J.A., Mathewson, H.A. & Morrison, M.L. (2015) Presence-only species distribution models to predict suitability over a long-term study for a species with a growing population. *Wildlife Society Bulletin*, **39**, 218–224.

McPherson, J., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811–823.

McPherson, J.M. & Jetz, W. (2007) Effects of species' ecology on the accuracy of distribution models. *Ecography*, **30**, 135–151.

McVicar, T.R., Van Niel, T.G., Li, L., Hutchinson, M.F., Mu, X. & Liu, Z. (2007) Spatially distributing monthly reference evapotranspiration and pan evaporation considering topographic

influences. *Journal of Hydrology*, **338**, 196–220.

Menéndez, R., Megías, A.G., Hill, J.K., Braschler, B., Willis, S.G., Collingham, Y., Fox, R., Roy, D.B. & Thomas, C.D. (2006) Species richness changes lag behind climate change. *Proceedings of the Royal Society B: Biological Sciences*, **273**, 1465–1470.

Merow, C., Smith, M.J., Edwards, T.C., Guisan, A., McMahon, S.M., Normand, S., Thuiller, W., Wüest, R.O., Zimmermann, N.E. & Elith, J. (2014) What do we gain from simplicity versus complexity in species distribution models? *Ecography*, **37**, 1267–1281.

Morán-Ordóñez, A., Suárez-Seoane, S., Elith, J., Calvo, L. & de Luis, E. (2012) Satellite surface reflectance improves habitat distribution mapping: a case study on heath and shrub formations in the Cantabrian Mountains (NW Spain). *Diversity and Distributions*, **18**, 588–602.

Nogués-Bravo, D. (2009) Predicting the past distribution of species climatic niches. *Global Ecology and Biogeography*, **18**, 521–531.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.

Pearman, P.B., Randin, C.F., Broennimann, O., Vittoz, P., Knaap, W.O., Engler, R., Lay, G. Le, Zimmermann, N.E. & Guisan, A. (2008) Prediction of plant species distributions across six millennia. *Ecology Letters*, **11**, 357–369.

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.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & the R Development Core Team (2013) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-113.

Pöyry, J., Luoto, M., Heikkinen, R.K. & Saarinen, K. (2008) Species traits are associated with the quality of bioclimatic models. *Global Ecology and Biogeography*, **17**, 403–414.

Rapacciuolo, G., Roy, D.B., Gillings, S., Fox, R., Walker, K. & Purvis, A. (2012) Climatic Associations of British Species Distributions Show Good Transferability in Time but Low Predictive Accuracy for Range Change. *PLoS ONE*, **7**, e40212.

Rubidge, E.M., Monahan, W.B., Parra, J.L., Cameron, S.E. & Brashares, J.S. (2011) The role of

climate, habitat, and species co-occurrence as drivers of change in small mammal distributions over the past century. *Global Change Biology*, **17**, 696–708.

Segurado, P. & Araujo, M.B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555–1568.

Seoane, J., Carrascal, L.M., Alonso, C.L. & Palomino, D. (2005) Species-specific traits associated to prediction errors in bird habitat suitability modelling. *Ecological Modelling*, **185**, 299–308.

Smith, A.B. (2013) On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions*, **19**, 867–872.

Swanson, A.K., Dobrowski, S.Z., Finley, A.O., Thorne, J.H. & Schwartz, M.K. (2013) Spatial regression methods capture prediction uncertainty in species distribution model projections through time. *Global Ecology and Biogeography*, **22**, 242–251.

Swets, J.A. (1988) Measuring the Accuracy of Diagnostic Systems. *Science (New York, N.Y.)*, **240**, 1285–1293.

Syfert, M.M., Smith, M.J. & Coomes, D.A. (2013) The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLOS ONE*, **8**, e55158.

Tabachnick, B.G. & Fidell, L.S. (1996) *Using multivariate statistics*, HarperCollins College Publishers, New York, USA.

Thuiller, W., Brotons, L., Araújo, M.B. & Lavorel, S. (2004) Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, **27**, 165–172.

Tuanmu, M.-N., Viña, A., Roloff, G.J., Liu, W., Ouyang, Z., Zhang, H. & Liu, J. (2011) Temporal transferability of wildlife habitat models: implications for habitat monitoring. *Journal of Biogeography*, **38**, 1510–1523.

Watling, J.I., Bucklin, D.N., Speroterra, C., Brandt, L.A., Mazzotti, F.J. & Romañach, S.S. (2013) Validating Predictions from Climate Envelope Models. *PLOS ONE*, **8**, e63600.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H. & Guisan, A. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

Young, A.R.M. (2000) *Environmental change in Australia since 1788*, Oxford University Press Melbourne.

1

BIOSKETCH(A)

The authors belong to the ARC Centre of Excellence for Environmental Decisions (CEED; http://ceed.edu.au/) and the National Environment Research Program Decisions Hub (NERP ED; http://www.nerpdecisions.edu.au/). They have a common interest in the methodological development, use and application of species distribution models to better inform environmental decisions and policies.

SUPPORTING INFORMATION (A)

Appendix S1. List of climatic and environmental predictors used to fit the SDMs & pairwise correlations.

Appendix S2. Climate variation during the study period (1950- 2013).

Appendix S3. List of taxa and number of records used to fit the models at each time period.

Appendix S4. Factors pre-disposing SDMs to good predictive performance. Regression analyses.

Appendix S5. Effect of incomplete sampling on AUC changes over time. Results and R script.

Appendix S6. Wilcoxon signed-rank and sum-rank paired test results.

Appendix S7. Transferability results when environmental conditions are considered constant across all time periods.

Appendix S8. GLMMs response curves for each taxonomic group, validation plots and effect sizes.

Appendix S9. Analysis of temporal autocorrelation in the sampled environmental space.

Table S10. Data set used to fit GLMMs.

Table 1. List of covariates used in the regression of AUC values (measure of predictive performance).

Variable name (acronym)	Description
Intrinsic factors of the model	
Training Ratio (TR)	Ratio Number of presence records on the training dataset / Number of records on the target-group background dataset used for model fitting, log transformed.
Evaluation Ratio (ER)	Ratio Number of presence records on the evaluation dataset / Number of records on the target-group background dataset used for model evaluation, log transformed.
MESS	Based on MaxEnt MESS maps (multivariate environmental similarity surface; 53), this predictor accounts for the proportion of novel environmental space in each model prediction, estimated as: Number of grid cells with values outside the environmental ranges covered by the target-group background dataset used for model fitting / Total number of grid cells of the study area (Australia)
PES	Proportion of the total environmental range sampled by the training dataset. We assessed the environmental coverage of the presence data used to fit the models in each time period by calculating for each species the proportion of the environmental space (presence records for the species across all years) that was sampled in that time period (Appendix S5).
Factors accounting for temporal transferability	
Year (YEAR)	Training year (central year of each of the training time periods; e.g. 1965 for *t1*).
Time since training (TST)	Temporal gap (in years) between model training and evaluation, calculated as the difference between the central years of the training and evaluation periods (e.g. for models trained in *t1*-1965- and evaluated in *t2* -1985-, this parameter will have a value of 20).
Species ecological traits(and proxies)	
Number of biogeographic sub-regions (BR)	Number of biogeographic sub-regions in which a species was recorded, calculated using all available records for each species over all time periods. The 419 biogeographic sub-regions described in Australia are geographically distinct bioregions based on common climate, geology, landform, native vegetation and species assemblages. Used as proxy for species range (the larger BR, the more widespread a species is in the study area).
Type of preferred habitat (TPH)	Preferred species habitat based on six major land cover types derived from the NVIS data base (National Vegetation Information System, NVIS v 4.1, Australian Government), and calculated on a basis of where the majority of all records of each species were observed over all time periods. We considered six types of habitats: (1) forest, (2) woodlands, (3) shrublands, (4) grasslands, (5) waterbodies and (6) naturally bare, rock, claypan, etc.
Body size (BS)[1]	Body mass (g) or body length (mm) of each species, log transformed

[1] Body mass used for birds and mammals models. Bird trait data from a database constructed by Gary Luck for Australian birds (Luck *et al.*, 2012, 2013). Mammals trait data from a global data base (Jones *et al.*, 2009). Body length used for amphibians (as total length) and reptiles (snout to vent length), derived from published literature (Greer, 1997; Cogger, 2000; Anstis, 2013).

Table 2. Results of generalized linear mixed effect models (GLMM) used to assess the relationship between AUC values, intrinsic factors of the modeling process and species traits. We report (1) D^2: deviance explained (Guisan & Zimmermann, 2000); (2) ΔAIC: difference in Akaike Information Criterion of each model and that of a null model with no informative predictors; (3) ΔAIC_{random}: difference in AIC between each GLMM model and a model with the same structure of fixed effects but lacking random factors; (4) variables selected for the most parsimonious model –the model explaining the largest variance with as few predictor variables as possible - among the pool of nested models considered (see notes below).

D2	ΔAIC	ΔAIC_{random}	Selected model structure
Birds (n=2170)			
0.52	-1059.8	-1620.5	poly(BR,2) + TR + poly(PES, 2)+ TST+ TPH
Mammals (n=400)			
0.41	-132.2	-397.7	poly(BR,2) + BS+ TR + TST+ TPH
Reptiles (n=310)			
0.27	-107.5	-160.42	poly(BR,2) × poly(PES, 2) + TR
Amphibians (n=180)			
0.33	-75.8	-117.5	poly(BR,2) + TST + TR + TPH
All groups (n=3060)			
0.42	-1344.2	-2543.2	poly(BR,2) × TR + poly(PES,2) + poly(BR,2) ×TST+ TPH

Key to abbreviations: BR, number of biogeographic sub-regions; TST, time since training; TR, training ratio; PES proportion of the total environmental range sampled by the training dataset; TPH, type of preferred habitat; BS, body size. Full description in Table 1.

[1] "+": additive effect; "×": interaction; "poly (*var* , *n*)": polynomial of order *n* of predictor *var*.

FIGURE CAPTIONS

Figure 1. Experimental design. The timeline at the top shows the decades covered for each time period considered (*t1*, *t2*, *t3* and *t4*). For each species we made ten model predictions (except for amphibians, with only six predictions, from *t2* to *t4*): four of these corresponded to "*cross-validated* " models (within-period model training and evaluation - solid line rectangles) and six corresponded to "*projected*" predictions (extrapolations of *cross-validated* models into future time periods - dashed rectangles). The width of each rectangle corresponds to the time period –read from the top timeline - from which background (Bg) and species data were used to either train (Tr) or evaluate (E) the models (note that all models were fit using records from at least a 20-year period). The map on the left of the figure shows in white the areas of Australia that were masked out from our analyses due to land use change during the study period (1950-2013).

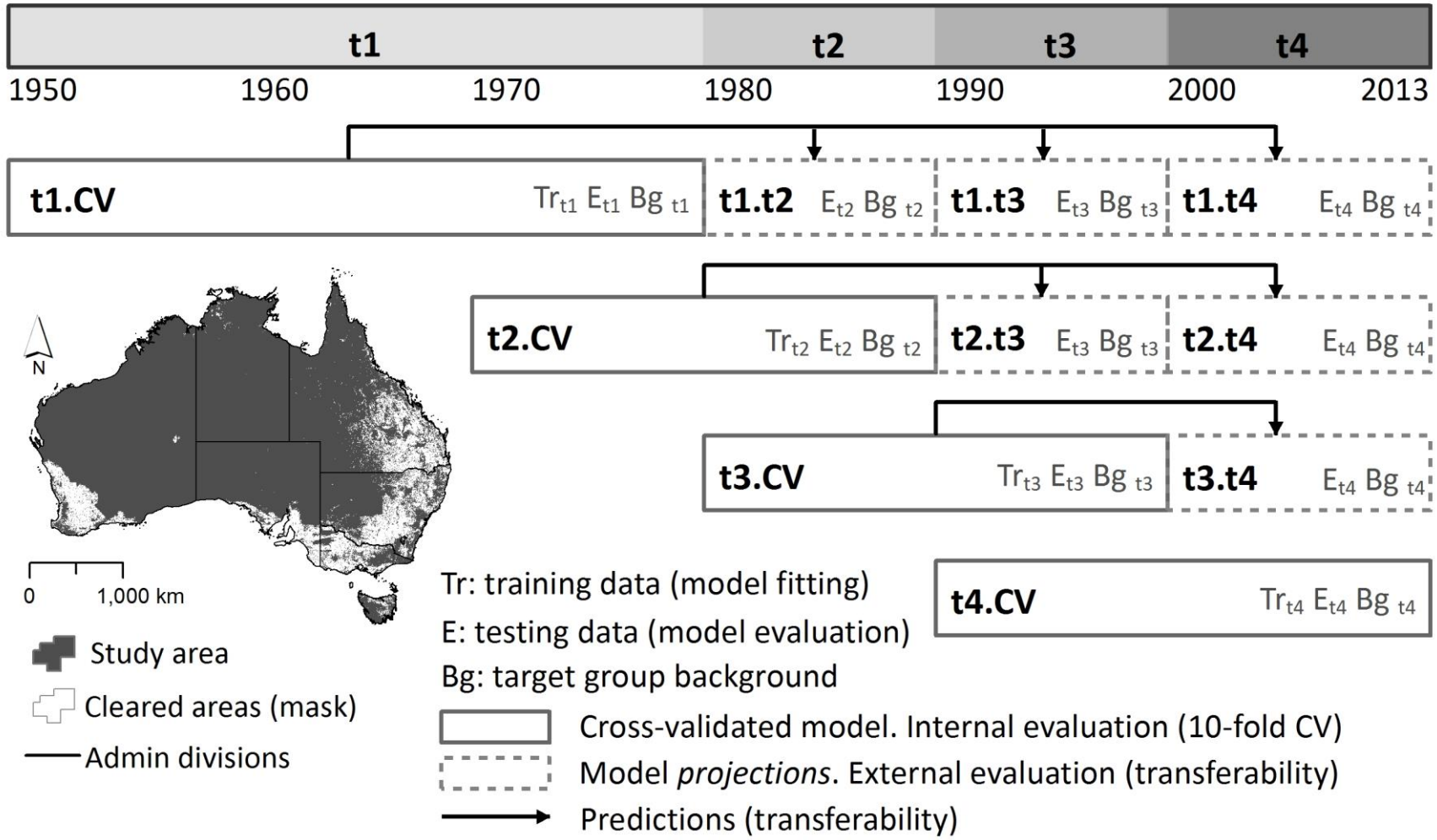
Figure 2. Notched boxplots for AUC values for cross-validated models (t1.cv, t2.cv, t3.cv, t4.cv) and their predictions into different "future" time periods (t1.t2, t1.t3, t1.t4, t2.t3, t2.t4, t3.t4), for each species group and across all groups. For example, AUC values for *t1.t2* refer to models trained with *t1* data and evaluated with *t2* data. Results correspond to those of the 'optimistic scenario' (projections into future time periods under complete knowledge of future environmental conditions). Models based on the same training dataset are shown under the same grey-scale colour. In each boxplot, boxes delimit interquartile ranges (25th and 75th percentiles), whiskers delimit ±2 standard deviations, notches are centred around AUC median values (horizontal bolded line), mean values are indicated with a solid black dot and outliers with empty dots. Horizontal dashed lines denote mean AUC values across all time periods for each species group. Note that the Y-axis ranges between 0.3 – 1 (and not between 0 – 1). The lack of overlap between the notch –narrowing around the median- of two boxes offers evidence of a statistically significant difference (α=0.05) between the medians (see also Appendix S6 for paired comparisons of median AUC values between different time periods) .

Figure 3. Response curves showing the estimated relationship between model accuracy (AUC values) measured across all species groups and all time periods (*t1*, *t2*, *t3* and *t4*) and: a) BR: number of biogeographic sub-regions in which a species occurs; b) TST: length of the time period between

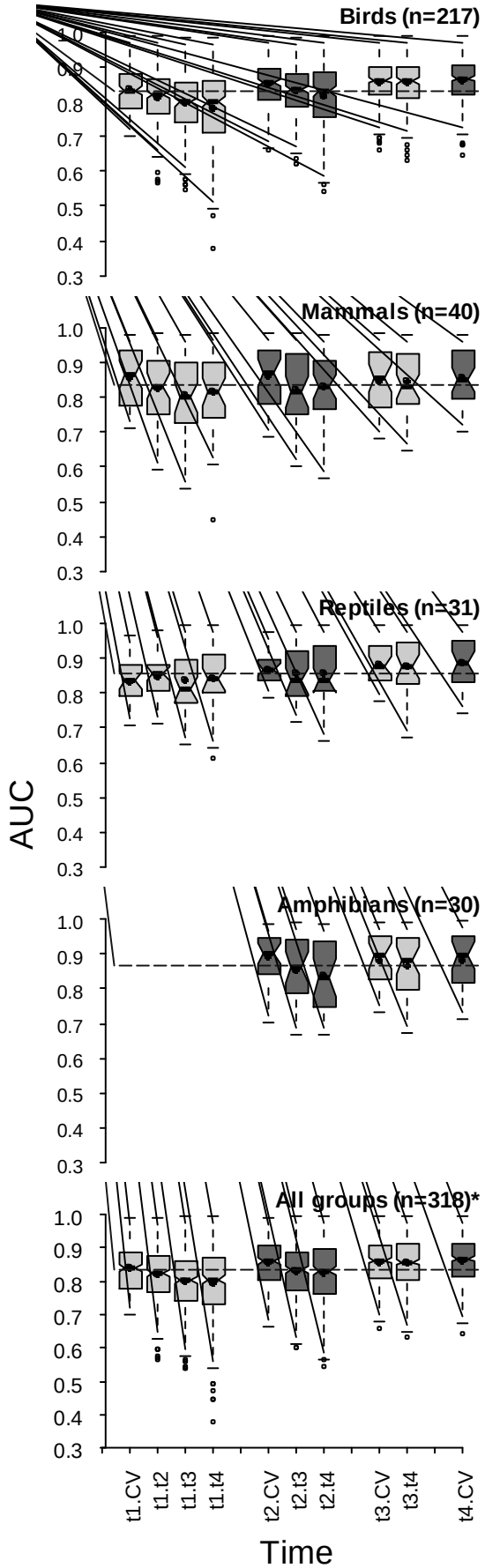
model training and evaluation, in years; c) TR: ratio between the number of presence and background points within the training dataset (log scale); d) PES: proportion of the total environmental range sampled by the training dataset; e) preferred habitat: *for*, forest; *wod*, woodland; *shd*, shubland, *grs*, grassland and *wat*, wetlands or water bodies; f) interaction between BR and TST; g) interaction between BR and TR (log scale). Plots f) and g) show the total effects of the variables displayed including their independent parts (that shown in a, b and c). Black lines show the effect of each variable on AUC value when keeping all other covariates at their average value. Tick marks and red lines along the x axis show the distribution of the original data. Shaded areas represent 95% confidence intervals. See Figs. S8.2-S8.5 for the AUC responses of the models fitted for each individual species group.

Figure 4. Distribution of the proportion of the total environmental range sampled by the training datasets at each time period (*t1* to *t4*) across all species and taxonomic groups. The larger the proportion of environmental range sampled, the more complete the information about the suitable environmental ranges of the species provided to the model. Density curves were calculated using the function "density" in R with default settings. The maps show the geographic distribution of the occurrence records used to fit the models at each time period (black points).

Figure 1

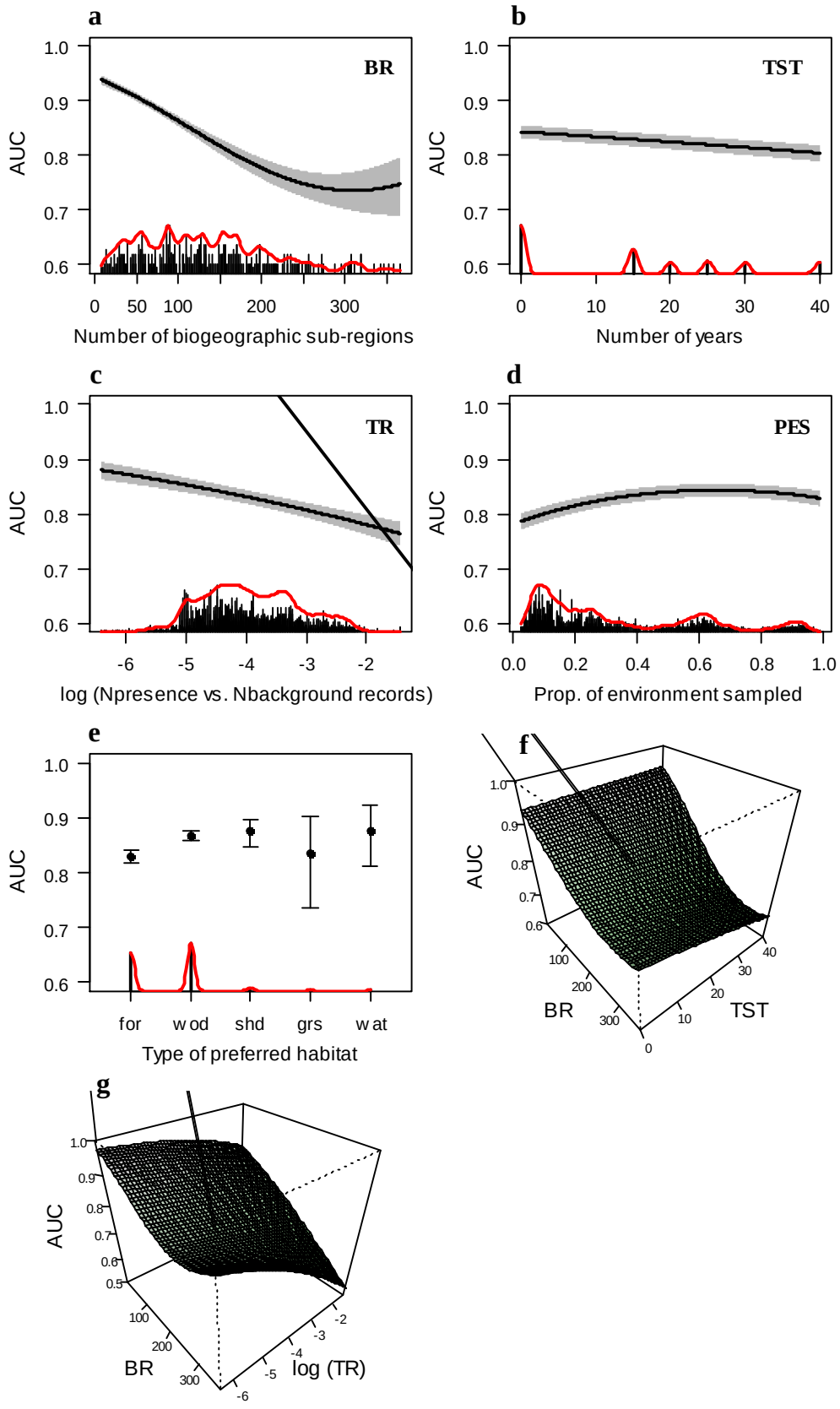

t1
t2
t3
t4
1950
1960
1970
1980
1990
2000
2013
t1.CV
Tr t1 E t1 Bg t1
t1.t2
E t2 Bg t2
t1.t3
E t3 Bg t3
t1.t4
E t4 Bg t4
t2.CV
Tr t2 E t2 Bg t2
t2.t3
E t3 Bg t3
t2.t4
E t4 Bg t4
t3.CV
Tr t3 E t3 Bg t3
t3.t4
E t4 Bg t4
t4.CV
Tr t4 E t4 Bg t4
N
0
1,000 km
Study area
Cleared areas (mask)
Admin divisions
Tr: training data (model fitting)
E: testing data (model evaluation)
Bg: target group background
Cross-validated model. Internal evaluation (10-fold CV)
Model *projections*. External evaluation (transferability)
Predictions (transferability)

86 **Figure 2**



87

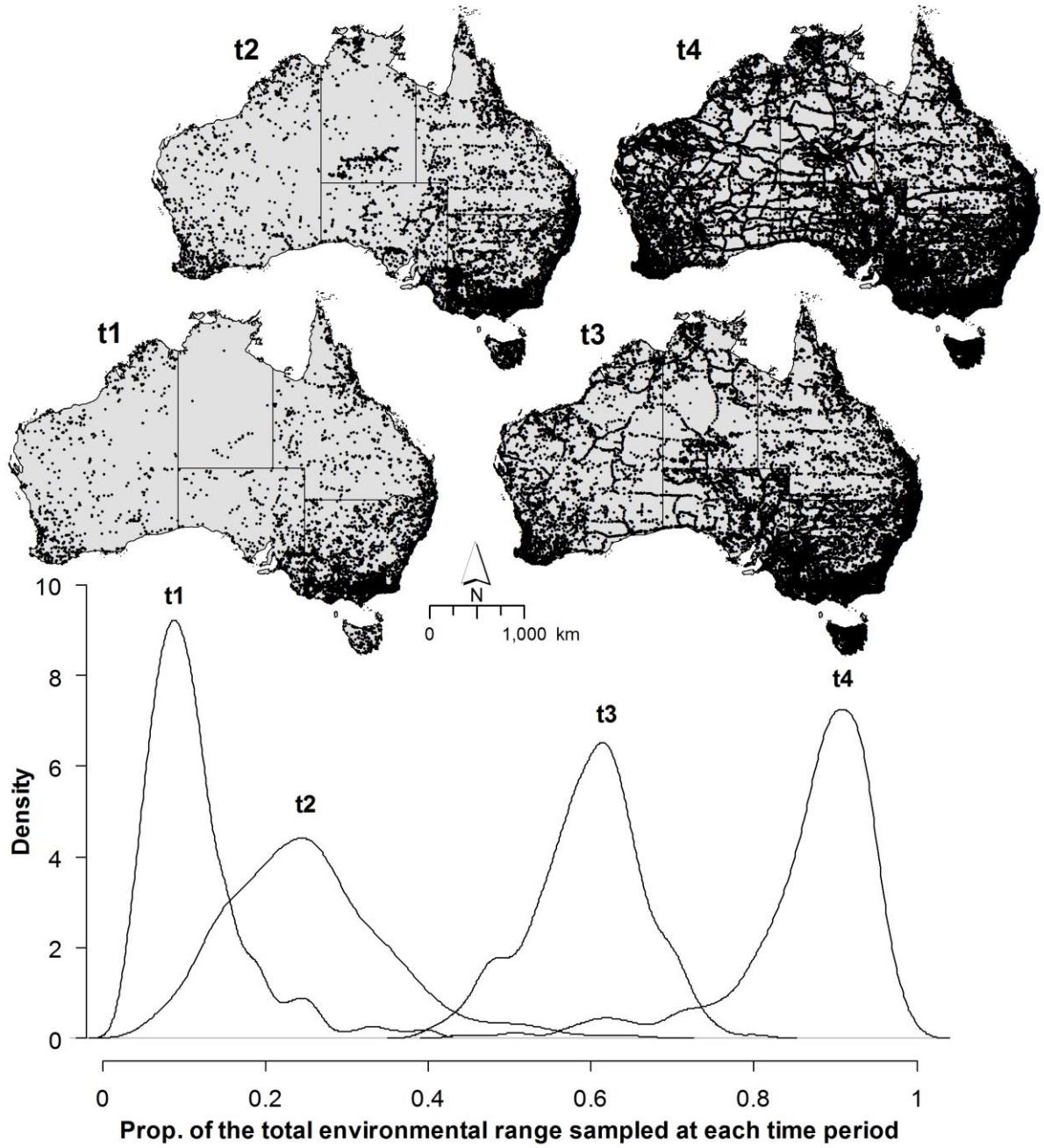
88 **Figure 3**



89

90

91 **Figure 4**



92

University Library

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Moran-Ordonez, A; Lahoz-Monfort, JJ; Elith, J; Wintle, BA

Title:
Evaluating 318 continental-scale species distribution models over a 60-year prediction horizon: what factors influence the reliability of predictions?

Date:
2017-03-01

Citation:
Moran-Ordonez, A., Lahoz-Monfort, J. J., Elith, J. & Wintle, B. A. (2017). Evaluating 318 continental-scale species distribution models over a 60-year prediction horizon: what factors influence the reliability of predictions?. Global Ecology and Biogeography, 26 (3), pp.371-384. https://doi.org/10.1111/geb.12545.

Persistent Link:
http://hdl.handle.net/11343/217217

File Description:
Accepted version