Running head:

Validating species distribution models

Title:

Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null-model.

Author:

Robert J. Hijmans

Department of Environmental Science and Policy

1023 Wickson Hall, University of California, Davis, CA 95616, USA

Email: rhijmans@ucdavis.edu

Telephone: 1 (530) 752-6555

Abstract

Species distribution models are usually evaluated with cross-validation. In this procedure evaluation statistics are computed from model predictions for sites of presence and absence that were not used to train (fit) the model. Using data for 226 species, from 6 regions, and two species distribution modeling algorithms (Bioclim and MaxEnt), I show that this procedure is highly sensitive to "spatial sorting bias": the difference between the geographic distance from testing-presence to training-presence sites and the geographic distance from testing-absence (or testing-background) to training-presence sites. I propose the use of "pair-wise distance sampling" to remove this bias, and the use of a null-model that only considers the geographic distance to training sites to calibrate cross-validation results for remaining bias. Model evaluation results (AUC) were strongly inflated: the null-model performed better than MaxEnt for 45% and better than Bioclim for 67% of the species. Spatial sorting bias and AUC values increased when using partitioned presence data and random-absence data instead of independently obtained presence-absence testing data from systematic surveys. Pair-wise distance sampling removed spatial sorting bias, yielding null-models with an AUC close to 0.5, such that AUC was the same as null-model calibrated AUC (*c*AUC). This adjustment strongly decreased AUC values and changed the ranking among species. Cross-validation results for different species are only comparable after removal of spatial sorting bias and/or calibration with an appropriate null-model.

Key words:

AUC, Bioclim, cross-validation, MaxEnt, model evaluation, niche model, pair-wise distance sampling, spatial sorting bias, spatial autocorrelation, species distribution model

Introduction

Species distribution models (SDMs), also known as climate envelope models, ecological niche models, and habitat suitability models, use environmental data for sites of occurrence (presence) of a species to predict all the sites where the environmental conditions are suitable for the species to persist, and may be expected to occur (see Elith and Leathwick 2009 and Zimmermann et al. 2010, for recent reviews). Here I only consider "presence-only" models, including models that use randomly sampled 'background' sites, but not models where both presence and absence sites are used to fit a SDM. Presence-only models are more frequently used than presence/absence models because of the wide-availability of occurrence data (e.g., from museum collections, Graham et al. 2004) compared to presence/absence data from systematic surveys. The application of SDMs is a very active field of research, but a number of methodological problems remain insufficiently addressed. A particularly pressing question is how to assess the quality of SDMs (Araújo and Guisan 2006, Lobo et al. 2007, Jiménez-Valverde 2011, Merckx et al. 2011).

The predictive power of SDMs is commonly evaluated through cross-validation. In this procedure, models are fitted with 'training' data and evaluated with a separate set of 'testing' data (Fielding and Bell 1997). That is, data that were not used to fit the model are used to evaluate its predictions for sites of known presence and for sites of known (or assumed) absence of a species. If independent data sets are not available, they can be created through partitioning

the entire data set into training and testing data, typically with "*k*-fold" sub-sampling. Testing data must have absence records, but as most species occurrence data used in SDM are not from systematic surveys, these are often not available. For that reason, most studies use sites that are randomly sampled from the study area ("pseudo-absence", "random-absence", or "background") instead of absence data. Cross-validation is considered the preferred approach to evaluate this type of models because (1) SDMs typically use climate variables that are strongly correlated with each other, such that more traditional (internal) goodness of fit statistics, that require that the variables used are statistically independent, are highly inflated and (2) because the objective of the use of SDM tends to be prediction, not explanation or hypothesis-testing, and hence an estimate of "predictive power" is often more relevant than "significance".

Several statistical measures can be computed from the predictions for the testing sites (the sites of known presence and (assumed) absence) (Fielding and Bell 1997, Pearce and Ferrier 2000, Liu et al. 2011). The most commonly used measure (Merckx et al. 2011) is the "area under the receiver-operator-curve" (AUC), which is a measure of discrimination that can be computed from the Wilcoxon (also known as Mann-Whitney) rank-sum test statistic (*W)* for the difference between two samples (Pearce and Ferrier 2000). If predictions for all sites of known presence are higher than predictions for all sites of known absence, AUC is 1 and an AUC value of 0.5 is considered to be equivalent to a random draw. However, if "background" data are used instead of absence data, these numbers are lower because the species will in fact be present in a fraction of the background sites (Phillips et al. 2006; Jiménez-Valverde 2011). The use of AUC in evaluating SDMs has been criticized (Lobo et al. 2008, Peterson et al. 2008, Jiménez-Valverde 2011), but here I do not discuss the merits of AUC relative to other evaluation statistics; I use it

as an example to discuss general problems with cross-validation, irrespective of the particular evaluation statistic used.

Because of spatial autocorrelation, the environment (climate in particular) of nearby sites is expected to be more similar than that of distant sites (Figure 1; Koenig 2002). It follows that the nearer testing-presence sites are to training-presence sites the more similar their environments, and the higher the predicted suitability of those sites will be (Segurado et al. 2006, Veloz 2009). Likewise, the more distant the testing-absence sites from the training-presence sites, the lower the model predictions for those sites will be (Elith et al. 2006, Bahn and McGill 2007, Lobo et al. 2007, Chefaoui and Lobo 2008, Lobo et al. 2010). Whilst these are not exact relationships because geographic patterns in environmental data can be quite complex, they will generally hold, and therefore model evaluation statistics should improve when the extent of a study area increases (absence sites tend to be further away from the presence sites), and when the range size of a species decreases (testing-presence sites tend to be nearer to training-presence sites).

The distance between training-presence and testing-presence sites will tend to be smaller than the distance between training-presence and testing-absence sites. I refer to this phenomenon of "attraction" of testing-presence sites and "repulsion" of testing-absence sites to presence-training sites as "spatial sorting bias". This bias is likely very common, such that cross-validation results should generally be inflated (Hampe 2004, Araújo et al. 2005, Segurado et al. 2006, Veloz 2009). The amount of inflation will vary from case to case, and, therefore, cross-validation statistics are not absolute numbers that can be compared across species or studies. For example, this type of inflation can explain why models for species with a narrow range tend to have very high AUC values (Elith et al. 2006, Lobo et al. 2007, Raes and Ter Steege 2007, Jiménez-Valverde et al. 2008), while, compared to widespread species, their distribution is more likely to

be constrained by non-climatic factors that are not variables in the model (*cf*. Schwarz et al. 2006). Note that in presence-only models there are no training-absence sites and hence training data and training-presence data are equivalent (some presence-only models use randomly sampled values in stead of absence data, but that is not relevant in the context of the present paper).

For unbiased cross-validation the issue is not that model testing data need to be *independent* of model training data, rather, the requirement should be that the attraction of testing-presence sites to training sites is the same as the attraction of testing-absence sites to training sites. It is possible to statistically test for differences in attraction/repulsion (Diggle and Cox 1983) but in the context of validating SDM the objective would be to remove such bias from the data. To achieve that, one could filter out test sites that are very close to training sites (Segurado et al. 2006, Pearson et al. 2007, Veloz 2009, Merckx et al. 2011) but such an approach is unlikely to eliminate spatial sorting bias entirely, because of the strong spatial autocorrelation, over long distances, in climate data (Figure 1). It follows that, whether or not an attempt is made to remove bias from the evaluation data, the effect of spatial sorting bias still needs to be quantified, such that evaluation results can be adjusted accordingly. This can be done with an appropriate null-model.

SDM null-models have been developed to test the strength of species/climate associations (Beale et al. 2009, Chapman 2010) and to improve testing for statistical significance (Raes and Ter Steege 2007, Merckx et al. 2011). As far as I know, null-models have not been used to address the problem of spatial sorting bias in cross-validation. Here I propose a new approach to cross-validation in which evaluation data is subsampled to minimize spatial sorting bias, and in which the evaluation results are adjusted with a geographic null-model that captures the effect of any

remaining spatial sorting bias. The proposed geographic null-model is solely based on the spatial pattern of the model training sites. It computes the inverse geographic distance to the nearest model training (presence) sites. The null-model can be evaluated with the same testing data, and the same evaluation statistic as used for the SDM. This procedure establishes how easy it is to predict presence/absence in the test data from the geographic position of the training data alone, very much like classical species range maps developed by drawing polygons around areas of known occurrence. The null-model does not use any environmental data and if there were no spatial autocorrelation, knowledge of the geographic location of training sites would be of no value to the model, and the expected AUC value for the null-model would be 0.5. Bahn and McGill (2007) showed, in a model comparison context with species abundance data, that geographic distance based models can outperform SDM approaches.

I illustrate the use of the null-model by computing AUC for 226 species showing that the predictions made with the null-model are highly correlated with those for the SDMs. This indicates that standard model evaluation statistics are inflated, and that there is a strong variation in the amount of bias between models for different species. I also show that using random-absence and data partitioning further increases AUC values; that sampling can be used to remove spatial sorting bias from evaluation data; and that removing bias leads to generally lower AUC valus, and to strong differences in the relative performance of models for different species.

Methods

I modeled the distribution of 226 species from 6 regions: the Australian Wet Tropics, Ontario (Canada), New South Wales (Australia), New Zealand, tropical South America, and Switzerland. These data are described in detail by Elith et al. (2006) and consist of a model training data

(presence-only), randomly sampled background data, and independently obtained model testing data (presence and absence data obtained from systematic surveys).

I used two species distribution modeling algorithms: Bioclim (Nix 1986) and MaxEnt (Phillips et al. 2006, Elith et al. 2011) via the 'dismo' package (Hijmans et al. 2011) in *R* (*R* Development Core Team 2010). Bioclim is a classical 'climate envelope model' that computes the suitability of a site by comparing the values of environmental variables at any site to the percentile distribution of the values at sites of known occurrence ('training sites'). MaxEnt, probably the most widely used SDM method, is a machine learning algorithm that uses presence and background data. I used a null-model that only considers the geographic distance to known occurrences, as implemented in the 'geoDist' function in the 'dismo' package. This function computes the inverse geographic distance to the nearest training site (sites of known occurrence). Distances smaller than 1 meter get the highest possible value of 1, all other values are lower.

I used two approaches to select testing-presence data (with or without pooling and subsampling of the training and testing data sets) and two approaches to select testing-absence data (survey or random absence). I evaluated these 4 treatments in two ways (with or without correcting the testing data for geographic bias), leading to a total of 8 treatments.

For all treatments I computed the mean distance of testing-presence sites to the nearest training-presence site (Dp) and the mean distance of testing-absence sites to the nearest training-presence site (Da), using the 'ssb' function in 'dismo'. SSB=Dp/Da is an indicator of spatial sorting bias, where SSB=1 suggests there is no spatial sorting bias and a SSB near 0 indicates extreme spatial sorting bias.

In the "baseline" treatment, models were fit with the training data and evaluated with the independently obtained presence/absence testing data, as in Elith et al. (2006). I also evaluated these models after sub-sampling the testing data using "pair-wise distance sampling" to attempt to remove spatial sorting bias from the testing data, using the "pwdSample" function in "dismo". The first step in this approach is to compute, for each testing-presence site, the distance to the nearest training-presence site. Each testing-presence site is paired with the testing-absence site that has the most similar distance to its nearest training-presence site. If the difference between the two distances is more than a specified threshold (I used 33%) the presence site is not used. Each testing-absence site was only used once. The baseline treatment, as described above, evaluated with these adjusted testing data, is referred to as the “baseline-adjusted" treatment.

In two additional treatments, I combined the presence sites from the test and training data and partitioned these into new training and testing sites sets by randomly taking 25% of the records for model testing, and 75% for model training. This mimics the common approach to data partitioning used in most SDM studies. With these new training data, I created a new set of 226 models and I evaluated these models in two ways: with the model testing data without attempting to correct for spatial sorting bias (the "combined" treatment) and after applying the pair-wise distance sampling method (the "combined-adjusted" treatment).

I also evaluated the models of the two main treatments described above (baseline or combined) using random sites instead of the absence sites from the presence/absence data; again to have treatments that are more comparable to common practice in species distribution modeling, as independently obtained absence data are seldom available for model testing. The number of absence sites selected was twice the number of presence records. In a final treatment, I adjusted

these evaluation data for spatial sorting bias. This yielded four additional treatments: "baseline-random", "baseline-random-adjusted", "combined-random", and "combined-random-adjusted".

For all eight treatments, I computed the AUC and the calibrated AUC (*c*AUC) for all species and for Bioclim and MaxEnt, according to Formula 1

$$c\text{AUC}_{s,m} = \text{AUC}_{s,m} + 0.5 - max\,(0.5,\ n\text{AUC}_s) \qquad \text{(Formula 1)}$$

Where $c\text{AUC}_{s,m}$ is the calibrated AUC, $\text{AUC}_{s,m}$ is the standard AUC, and $n\text{AUC}_s$ is the null-model AUC for species *s* and model *m*. In the probably rare case where *n*AUC < 0.5, *c*AUC is the same as AUC; although one could choose to adjust the value of *c*AUC as in the case where *n*AUC > 0.5, I see no reason to *increase* a cross-validation result when a SDM cannot benefit from spatial autocorrelation.

Results

The median AUC for the 226 modeled species was 0.64 for Bioclim and 0.73 for MaxEnt in the baseline treatment (Table 1, Figure 2A). These values are very close to those reported by Elith et al. (2006), who used a different implementation of Bioclim and an early version of the MaxEnt software. The median AUC for the geographic null-model was 0.69. This is higher than for Bioclim, and also slightly higher (not statistically significant) than the results that Elith et al. (2006) reported for the BRUTO, DOMAIN, GAM, GARP, GLM, and MARS methods. The median calibrated AUC, *c*AUC (Formula 1), was 0.46 for Bioclim and 0.52 for MaxEnt. The AUC of the null-model was a good predictor of the standard AUC values, particularly for MaxEnt (Figure 3).

Combining and then randomly partitioning training and testing presence data (without adjustment) resulted in higher AUC values than for the unadjusted baseline data, and in very low *c*AUC values (Table 1). Using random testing-absence data further increased AUC values: the median value was 0.78 for Bioclim, 0.92 for MaxEnt, and 0.93 for the null-model (Table 1). For MaxEnt, AUC increased with 0.13 when models were evaluated with "random-absence" rather than with "survey-absence".

In the unadjusted data, spatial sorting bias was strong, but with large variation between species and regions (Table 2). The median value for SSB=Dp/Da (in km) was 19/49=0.38 for the baseline treatment, 19/77=0.24 for the baseline-random treatment, 5/26=0.17 for the combined treatment and 5/43=0.11 for the combined-random treatment. All adjusted treatments had a SSB of 1. The ranking of SSB by region for unadjusted treatments closely resembled the ranking of AUC values (Table 2).

Pair-wise distance sampling removed, as expected, most spatial sorting bias, and the median AUC for the null-model was 0.5 in both the baseline- and combined-adjusted treatments (Table 1; Figure 2 C,D,G,H). The median *c*AUC values were also similar in the four treatments where the test data were adjusted for spatial sorting (0.56-0.58 for Bioclim and 0.59-0.68 for MaxEnt), and higher than in the unadjusted treatments (Table 1). In the baseline treatment, 37% of the Bioclim models and 58% of the MaxEnt models had an AUC greater than 0.7 (this threshold is often used to determine if a model is "good"). In the baseline-adjusted treatment, only 8% of the Bioclim and 17% of the MaxEnt models had a *c*AUC greater than 0.7; and in the combined-adjusted treatment this was 8% for Bioclim and MaxEnt. The highest *c*AUC values (26% of the models with a value greater than 0.7) were obtained with MaxEnt in the baseline-adjusted-random treatment (Table 1).

Despite the similarity in the distribution of *c*AUC values for the adjusted treatments, their relationship at the species level was noisy. For example, regression lines between *c*AUC for the baseline-adjusted and the combined-adjusted treatments had a slope of 0.38 (R^2=0.1) for Bioclim and 0.61 (R^2=0.24) for MaxEnt. The relation between the baseline AUC and the baseline-adjusted *c*AUC values was also weak (R^2 was 0.23 for Bioclim and 0.36 for MaxEnt) . This was also true for the relation between the combined treatment AUC and the combined-adjusted treatment *c*AUC (R^2 was 0.24 for Bioclim and 0.21 for MaxEnt). MaxEnt models had a higher AUC than Bioclim models for 167 species in the baseline treatment, and for 134 species in the combined-adjusted treatment, but only for 110 species (49% of the cases) was MaxEnt better in both treatments. These results show that the ranking of models can change when changing approaches to data selection for cross-validation and when considering *c*AUC rather than AUC. This was also illustrated by aggregating the results by region, which showed strong differences between the AUC and the *c*AUC results. For example, the models of the species from South America had the highest median AUC values for MaxEnt (0.81) in the baseline treatment, but the one but lowest median *c*AUC values in the baseline-adjusted treatment, and the lowest in the combined-adjusted treatment (Table 2). The models of the species from Switzerland were reasonably good overall: *c*AUC values were greater than 0.75 when using random background absence data and near 0.65 when using survey absence data (Table 2).

Discussion

The strong relation between cross-validation results for the geographic null-model and those for the two species distribution models clearly illustrates that uncalibrated cross-validation results

are strongly influenced by spatial sorting bias. This, and the strong variation in AUC values between treatments, shows that it is impossible to directly interpret uncalibrated cross-validation results, or to compare such results across species and data sets. Therefore, reporting and interpreting cross validation results for SDMs is only useful if spatial sorting bias is removed from the evaluation sites and/or if the results are calibrated with the results for a null-model. Common statements like "a model with an AUC that is higher than 0.7 is a good model" are inappropriate, unless it is shown that a null-model that accounts for spatial sorting bias has an AUC of about 0.5. Similarly, the notion that a SDM performs better than a random draw when AUC > 0.5, is in most cases incorrect, because that requires the generally unsupported assumption that there is no spatial sorting bias in the evaluation data.

Inverse-distance is an easy to understand and easy to compute null-model for evaluating SDMs. The null-model estimates the amount of spatial sorting bias in the testing data; but the exact effect of this bias on model evaluation depends on the amount of spatial autocorrelation in the environmental data, the sites used, and the particular SDM algorithm. It could well be that the null-model overestimates bias, as is suggested by the very low *c*AUC values in the unadjusted treatments, relative to those for the adjusted treatments. It could be of interest to develop a more refined null-model, perhaps by considering multi-site instead of nearest neighbor distance, and/or by using a different distance decay function than the hyperbolical (1/x) proposed here (which would matter for evaluation statistics like the correlation coefficient, but not for the AUC), perhaps informed by spatial autocorrelation in the predictor variables used by the model. However, as demonstrated in this paper, one can avoid "over-calibration", by first sub-sampling evaluation sites using pair-wise distance sampling or a comparable method. Sub-sampling is particularly easy to implement when using random absence data. Therefore, I recommend to first

use pair-wise distance sampling or a comparable method to remove spatial sorting bias, and then compute and report *c*AUC (*c*AUC values should be very close to AUC values after removing spatial sorting bias).

The AUC values obtained with random-absence data were higher than with observed-absence data. This difference was particularly large for the unadjusted treatment, which suggests that while using random-absence lowers the maximum AUC that can theoretically be obtained (Phillips et al. 2006) as well as the threshold that should be considered a random draw (Jiménez-Valverde 2011), using random absence may, in practice, lead to increased AUC values, even after removing spatial sorting bias from the evaluation data.

Adjustment with pair-wise distance sampling effectively removed spatial sorting bias. Another source of bias, which I have ignored in this paper, is the distribution of presence-testing sites relative to presence-training sites. Although absence-testing sites were selected to get matching patterns for testing-presence and testing-absence data, it is important to also assure that testing sites have a balanced distribution across the range of the species (Vaughan and Ormerod 2003), and perhaps data partitioning should use spatial stratification to assure this. In this paper, the null-model was applied to presence-only models, but a similar approach could be developed for presence/absence models using, for example, an inverse-distance weighted based prediction of presence (1) and absence (0) in the training data.

Following general rules of sampling design, it has been suggested that training and testing data should ideally be obtained from independent systematic surveys (Elith et al. 2006, Araújo and Guisan 2006) and that this can alleviate the problem of inflated cross-validation results (Veloz 2009, Gogol-Prokurat 2011). This might be true in many cases, as in the dataset analyzed this

paper, as AUC values were further inflated by combining and partitioning presence training and testing data. However, this may not always be the case because if two surveys are independent in the sense that they were carried out by different researchers, this does not imply that they produce spatially independent samples because, for a number of reasons, different surveys may tend to go to the same places (Hijmans et al. 2000). And even if the two samples are equivalent to two independent random samples, inflation of cross-validation results may still occur, because what affects cross-validation is the spatial sorting bias. Therefore, independent systematic survey data are not the "gold-standard" that they have been proclaimed to be (Jiménez-Velarde 2011). AUC and *c*AUC were highly sensitive to the evaluation approach used, even for the 'adjusted' treatments. It might therefore be best, even if independent survey data are available, to standardize SDM evaluation by using (*k*-fold) partitioned presence sites and random background sites that are adjusted for spatial sorting bias, because such data can be created for all species.

The problem of spatial sorting bias exists because of the presence of spatial autocorrelation. The ecological literature on spatial autocorrelation focuses on the problem of increased probability of Type-I error in hypotheses tests, and of biased variable selection and parameter estimation (Bahn et al. 2006; Dormann et al. 2007; although Hawkins et al. (2007) suggest that this bias is uncommon in the analysis of macro-ecological data). One can attempt to account for the effect of spatial autocorrelation through the use of, e.g., autoregressive models (see Dormann et al. 2007 for a discussion of different approaches), and this can improve the "internal" evaluation of model fit and statistical significance. However, when such models are evaluated with cross-validation, one still needs to remove spatial sorting bias and/or use a geographic null-model to obtain unbiased results.

The approach to cross-validation proposed in the present paper is not relevant when testing sites are very far away from the training data, for example, in another continent (but cross-validation may, in those cases, nevertheless, be strongly influenced by, e.g., the spatial extent from which the evaluation sites are drawn). The need for comparing with a null-model would seem unimportant in model comparison studies like Elith et al. (2006), as all models are evaluated with the same bias, and in model selection and model averaging (Thuiller et al. 2009). However, the results presented in this paper question that assumption because the difference in performance between MaxEnt and Bioclim was reduced after removing spatial sorting bias and calibrating AUC values; because a very good model before calibration can be a very poor model after calibration (as was the case for many of the South American species); and because the relations between cross-validation results for the different treatments were noisy.

Comparisons of model results between species, regions, or any other results obtained with a different combination of model training and testing data are questionable if the results are not calibrated. The results presented here therefore question the conclusions of previous work that was based on comparing uncalibrated AUC values between different models. For example, Pearman et al. (2010) found that models for groups below the species level (e.g., clades) are better than models for species. However, the smaller the range of the group studied, the smaller the distance between testing-presence and training-presence sites will get; and hence Pearman et al.'s results may only reflect this statistical artifact. Heikkinen et al. (2006) summarize the results of a number of studies as follows: "species at the margin of their range or with low prevalence were better predicted than widespread species, and species with clumped distributions better than widely scattered dispersed species". However, without a null-model that penalizes spatial sorting

bias, it is unclear to what extent the results of these studies reflect changes in model performance or mainly changes in the spatial configuration of model training relative to model testing data.

*c*AUC values were very low, supporting the notion that failure to account for spatial structure in the data may have led to inflated confidence in SDMs (Beale et al. 2008, Chapman 2011). Araújo et al. (2009) argued against this by showing that AUC values increased for the species modeled by Beale et al. (2008) when they used more occurrence records. However, the evidence presented in the present paper suggest that this increase in AUC might be an artifact of a larger sample size, as this will decrease the expected geographic distance between testing and training sites.

It is important to note, however, that a low *c*AUC means that there is no support for the statement that a SDM is good, but it does not prove that the model is bad. In some cases, the available data simply won't allow for establishing, through cross-validation, that a model has good predictive power. If the species occurs at a high density in a single small contiguous range, a SDM might be able to correctly model the distribution, but it would nearly impossible to outperform a geographic null-model. If, however, a species occurs in a number of clumps that are far apart, testing and training presence sites can be spatially separated, and *c*AUC values could be high. Thus, the *c*AUC removes spatial sorting bias, and allows for comparing results between studies, but it does not capture everything. While the use of AUC clearly leads to inflated model evaluation results, the use of *c*AUC will likely lead to a situation where cross-validation provides no support for some models even if these models provide a reasonably good description of the relation between the environment and the distribution of a species. In such cases, careful data collection and partitioning may allow for improved model evaluation (Vaughan and Ormerod 2003).

Because of the problematic nature of cross-validation, particularly when no attempt is made to correct for spatial sorting bias, it remains unclear if we can use SDMs to adequately extrapolate the distribution of species across space and time, particularly for species with a narrow range. More SDM validation work is needed with species for which there are independent data sets across large areas, or large time periods, especially if this can lead to insights about which species are more (un)suitable for use in SDM (Dobrowski et al. 2011). Cross-validating SDMs is inherently problematic, even when the results are deflated with a null-model, as it is affected by the approach use to creating training and testing data. Therefore, I believe that modelers should focus less on cross-validation, and more on the quality of the occurrence data used (Lobo 2008), on the biology of the species studied, and use a "comprehensive toolbox of evaluation measures" (Elith and Leathwick 2009). For example, they could build models that incorporate known mechanisms (Kearney and Porter 2009, Monahan 2009), use a-priori relevant predictor variables (Austin 2002), and use alternative model selection criteria such as AIC (Warren and Seifert 2011). The resulting models might have a lower cross-validation score, but, nevertheless, be a more relevant description of a species' environmental requirements.

Acknowledgements

This work builds on efforts by the working group on "Testing Alternative Methodologies for Modeling Species' Ecological Niches and Predicting Geographic Distributions", at the National Center for Ecological Analysis and Synthesis (Santa Barbara, California, USA). I thank all the members of the working group, as well as others who provided data used here: A. Ford, CSIRO Atherton; M. Peck and G. Peck, Royal Ontario Museum, and M. Cadman, Bird Studies Canada,

Canadian Wildlife Service of Environment Canada; the National Vegetation Survey Databank and the Allan Herbarium; Missouri Botanical Garden, especially R. Magill and T. Consiglio; and T. Wohlgemuth and U. Braendi from WSL Switzerland. I thank Jane Elith and Catherine Graham for compiling these data and I also thank them and Steven Phillips for helpful comments on this manuscript.

Literature cited

Araújo, M.B. and A. Guisan. 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33: 1677–1688.

Araújo, M.B., R.G. Pearson, W. Thuiller and M. Erhard. 2005. Validation of species-climate impact models under climate change. Global Change Biology 11: 1504–1513.

Araújo, M.B., W. Thuiller and N.G. Yoccoz. 2009. Reopening the climate envelope reveals macroscale associations with climate in European birds. Proceedings of the National Academy of Sciences 106: E45-46.

Austin, M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157: 101–118.

Bahn, V., R.J. O'Connor, and W.B. Krohn. 2006. Importance of spatial autocorrelation in modeling bird distributions at a continental scale. Ecography 29: 835–844

Bahn, V. and B.J. McGill. 2007. Can niche-based distribution models outperform spatial interpolation? Global Ecology and Biogeography 16: 733–742.

Beale, C.M., J.J. Lennon and A. Gimona. 2008. Opening the climate envelope reveals no macroscale associations with climate in European birds. Proceedings of the National Academy of Sciences 105: 14908–14912.

Beale, C.M., J.J. Lennon and A. Gimona. 2009. European bird distributions still show few climate associations. Proceedings of the National Academy of Sciences 106: E41-E43.

Chapman, D.S. 2010. Weak climatic associations among British plant distributions. Global Ecology and Biogeography 19: 831–841.

Chefaoui, R.M. and J.M. Lobo. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling 210: 478–486.

Diggle, P.J. and T.F. Cox. 1983. Some distance-based tests of independence for sparsely-sampled multivariate spatial point patterns. International Statistical Review 51: 11-23.

Dobrowski, S.Z., J.H. Thorne, J.A. Greenberg, H.D. Safford, A.R. Mynsberge, S.M. Crimmins and A.K. Swanson. 2011. Modeling plant distributions over 75 years of measured climate change in California, USA: Relating temporal transferability to species traits. Ecological Monographs doi:10.1890/10-1325.1.

Dormann, C.F., J.M. McPherson, M.B. Araújo, R. Bivand, J. Bolliger, G. Carl, R.G. Davies, A. Hirzel, W. Jetz, W.D. Kissling, I. Kühn, R. Ohlemüller, P.R. Peres-Neto, B. Reineking, B. Schröder, F.M. Schurr and R. Wilson. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30: 609-628.

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

Elith, J., C.H. Graham, R.P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J. Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. Overton, A.T. Peterson, S. Phillips, K. Richardson, R. Scachetti-Pereira, R. Schapire, J. Soberón, S. Williams, M. Wisz and N. Zimmerman. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151.

Elith, J., S.J. Phillips, T. Hastie, M. Dudik, Y.E. Chee and C.J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17: 43-57.

Fielding, A. H. and J.F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24: 38–49

Gogol-Prokurat, M. 2011. Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. Ecological Applications 21: 33–47.

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

Hampe, A. 2004. Bioclimatic envelope models: what they detect and what they hide. Global Ecology and Biogeography 13: 469–471.

Hawkins, B.A., J.A.F. Diniz-Filho, L.M. Bini, P. De Marco, and T.M. Blackburn. 2007. Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. Ecography 30: 375–384.

Heikkinen, R.K., M. Luoto, M.B. Araújo, R. Virkkala, W. Thuiller, and M.T. Sykes. 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. Progress in Physical Geography 30: 751–777.

Hijmans, R.J., K.A. Garrett, Z. Huamán, D.P. Zhang, M. Schreuder and M. Bonierbale. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. Conservation Biology 14: 1755-1765.

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965–1978.

Hijmans, R.J., S. Phillips, J. Leathwick and J. Elith. 2011. dismo: Species distribution modeling. *R package* version 0.6-3. http://cran.r-project.org/web/packages/dismo/

Jiménez-Valverde, A. 2011. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography (on-line early): DOI: 10.1111/j.1466-8238.2011.00683.

Jiménez-Valverde, A., J.M. Lobo and J. Hortal. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions 14: 885–890.

Kearney, M. and W. Porter. 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecology letters 12: 334-350.

Koenig, W.D. 2002. Global patterns of environmental synchrony and the Moran effect. Ecography 25: 283–288.

Liu, C., M. White and G. Newell. 2011. Measuring and comparing the accuracy of species distribution models with presence-absence data. Ecography 34: 232-243.

Lobo, J.M. 2008. More complex distribution models or more representative data? Biodiversity Informatics 5: 14-19.

Lobo, J.M., A. Jiménez-Valverde and R. Real. 2007. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17: 145-151.

Lobo, J.M., A. Jiménez-Valverde and J. Hortal. 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography 33: 103-114.

Merckx, B., M. Steyaert, A. Vanreusel, M. Vincx and J. Vanaverbeke. 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. Ecological Modelling 222: 588–597.

Monahan W.B. 2009. A Mechanistic Niche Model for Measuring Species' Distributional Responses to Seasonal Temperature Gradients. PLoS ONE 4: e7921.

Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes. Pages 4-15 in Longmore, R., editor. Atlas of Elapid Snakes of Australia. Australian Flora and Fauna Series 7. Australian Government Publishing Service, Canberra.

Pearman, P.B., M. D'Amen, C.H. Graham, W. Thuiller and N.E. Zimmermann. 2010. Within-taxon niche structure: Niche conservatism, divergence and predicted effects of climate change. Ecography 33: 990-100.

Pearce, J. and S. Ferrier. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133: 225-245.

Pearson, R.G., C.J. Raxworthy, M. Nakamura and A.T. Peterson. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 44: 102–117.

Peterson, A.T., M. Papeş and J. Soberón. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecological Modelling 213: 63-72.

Phillips, S.J., R.P. Anderson and R.E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190: 231-259.

R Development Core Team. 2010. *R*: A Language and Environment for Statistical Computing. *R* Foundation for Statistical Computing. http://www.R-project.org/

Raes, N. and H. ter Steege. 2007. A null-model for significance testing of presence-only species distribution models. Ecography 30: 727–736.

Schwartz, M.W., L.R. Iverson, A.M. Prasad, S.N. Matthews and R.J. O'Connor. 2006. Predicting Extinctions as a Result of Climate Change. Ecology 87: 1611-1615.

Segurado, P., M.B. Araújo and W.E. Kunin. 2006. Consequences of spatial autocorrelation for niche-based models. Journal of Applied Ecology 43: 433–444.

Thomas, C.D., A. Cameron, R.E. Green, M. Bakkenes, L.J. Beaumont, Y.C. Collingham, et al. 2004. Extinction risk from climate change. Nature 472: 145-148.

Thuiller, W. 2003. BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology 9: 1353-1362.

Thuiller, W. S. Lavergne, C. Roquet, I. Boulangeat, B. Lafourcade and M.B. Araújo. 2011. Consequences of climate change on the tree of life in Europe. Nature: doi:10.1038/nature09705.

Vaughan, I.P., and S.J. Ormerod. 2003. Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. Conservation Biology 17: 1601–1611.

Veloz, S.D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. Journal of Biogeography 36: 2290-2299.

Warren, D. and S. Seifert, S. 2011. Environmental niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. Ecological Applications: doi:10.1890/10-1171.1

Zimmermann N.E., T.C. Edwards, C.G. Graham, P.B. Pearman and J.-C Svenning. 2010. New trends in species distribution modelling. Ecography 33, 985-989.

Table 1. Median AUC and *c*AUC (distance null-model calibrated AUC) values for two species distribution models (Bioclim and MaxEnt) for 226 species from six regions and for 8 treatments. The numbers in parenthesis are the 10^{th} and 90^{th} percentile. The “baseline” treatments use independently obtained model training and testing data. For the “combined” treatment the original training and testing presence data were pooled and a random sample of 75% of the records was used for model training, and 25% for model testing. In the "random" treatments, random background data is used instead of absence data and in the “adjusted” treatments, evaluation data were corrected for "spatial sorting bias" with "pair-wise distance sampling" (see Methods).

	AUC			*c*AUC	
	Distance	Bioclim	MaxEnt	Bioclim	MaxEnt
Baseline	0.69 (0.50-0.92)	0.64 (0.50-0.83)	0.73 (0.53-0.91)	0.46 (0.30-0.59)	0.52 (0.42-0.62)
Baseline-random	0.80 (0.56-0.95)	0.70 (0.51-0.87)	0.86 (0.62-0.96)	0.42 (0.30-0.54)	0.52 (0.42-0.67)
Baseline-adjusted	0.50 (0.49-0.51)	0.56 (0.47-0.70)	0.60 (0.48-0.76)	0.56 (0.46-0.69)	0.60 (0.47-0.75)
Baseline-adjusted-random	0.50 (0.49-0.51)	0.57 (0.46-0.70)	0.68 (0.52-0.82)	0.56 (0.45-0.70)	0.67 (0.52-0.82)
Combined	0.83 (0.62-0.95)	0.71 (0.57-0.86)	0.80 (0.58-0.92)	0.41 (0.28-0.53)	0.48 (0.39-0.55)
Combined-random	0.93 (0.80-0.98)	0.78 (0.64-0.91)	0.92 (0.79-0.97)	0.38 (0.28-0.45)	0.48 (0.43-0.55)
Combined-adjusted	0.50 (0.49-0.51)	0.56 (0.47-0.70)	0.59 (0.48-0.70)	0.56 (0.47-0.69)	0.58 (0.47-0.69)
Combined-adjusted-random	0.50 (0.49-0.52)	0.58 (0.48-0.66)	0.66 (0.56-0.78)	0.58 (0.47-0.66)	0.66 (0.56-0.77)

Table 2. Median AUC or the geographic null-model ('D'), and for two species distribution models Bioclim (BC) and MaxEnt (ME) and *c*AUC (null-model calibrated AUC) for Bioclim and MaxEnt. And the median distance to the geographically nearest training-presence site for testing-presence (Dp) and testing-absence (Da) sites, and the number of testing presence (Np) and testing absence (Na) sites. For 226 species from six different regions and eight treatments. See Table 1 for an explanation of the treatments and Figure 3 for the region codes.

		AUC			*c*AUC					
Treatment	Region	D	BC	ME	BC	ME	Dp	Da	Np	Na
Baseline	AW	0.64	0.66	0.71	0.50	0.55	9	17	47	107
	CA	0.56	0.64	0.58	0.55	0.51	22	55	867	13705
	NS	0.71	0.63	0.69	0.42	0.48	15	43	76	844
	NZ	0.67	0.60	0.73	0.44	0.55	46	91	821	18299
	SA	0.85	0.77	0.81	0.42	0.48	106	645	9	143
	SW	0.76	0.70	0.78	0.46	0.55	4	11	255	9759
Baseline-random	AW	0.77	0.64	0.73	0.42	0.49	9	25	47	93
	CA	0.92	0.86	0.94	0.46	0.52	22	201	867	1733
	NS	0.82	0.67	0.87	0.41	0.51	15	50	76	151
	NZ	0.70	0.60	0.79	0.39	0.56	46	105	821	1642
	SA	0.87	0.75	0.86	0.37	0.50	106	464	9	50
	SW	0.78	0.78	0.90	0.49	0.62	4	13	255	509
Baseline-adjusted	AW	0.50	0.57	0.60	0.57	0.60	10	10	42	42
	CA	0.50	0.53	0.52	0.53	0.52	22	22	867	867
	NS	0.50	0.55	0.57	0.55	0.57	15	15	68	68

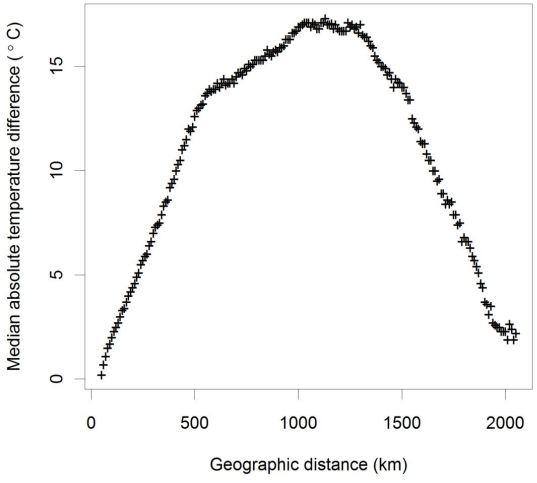
	NZ	0.50	0.55	0.65	0.55	0.64	46	46	820	820
	SA	0.50	0.58	0.56	0.58	0.56	110	111	9	9
	SW	0.50	0.60	0.67	0.60	0.67	4	4	252	252
Baseline-random-adjusted	AW	0.50	0.52	0.57	0.52	0.57	9	9	45	45
	CA	0.50	0.62	0.67	0.62	0.67	25	25	861	861
	NS	0.50	0.59	0.71	0.57	0.70	15	15	74	74
	NZ	0.50	0.52	0.69	0.52	0.69	46	46	817	817
	SA	0.50	0.61	0.63	0.58	0.62	108	108	9	9
	SW	0.50	0.68	0.81	0.68	0.81	4	4	254	254
Combined	AW	0.70	0.72	0.76	0.50	0.53	4	11	26	107
	CA	0.74	0.65	0.71	0.44	0.49	5	25	239	13705
	NS	0.81	0.69	0.79	0.39	0.46	5	27	37	844
	NZ	0.87	0.78	0.85	0.37	0.47	2	34	224	18299
	SA	0.81	0.72	0.76	0.41	0.47	111	631	17	143
	SW	0.88	0.70	0.83	0.34	0.47	2	10	210	9759
Combined-random	AW	0.90	0.72	0.84	0.36	0.46	4	19	26	93
	CA	0.97	0.92	0.96	0.43	0.48	5	162	239	1733
	NS	0.91	0.77	0.92	0.36	0.49	5	31	37	151
	NZ	0.97	0.86	0.94	0.38	0.47	2	54	224	1642
	SA	0.86	0.67	0.83	0.33	0.47	111	497	17	50
	SW	0.89	0.77	0.93	0.40	0.54	2	12	210	509
Combined-adjusted	AW	0.50	0.66	0.63	0.65	0.61	5	5	23	23
	CA	0.50	0.51	0.54	0.51	0.54	5	5	239	239
	NS	0.50	0.58	0.56	0.58	0.56	5	5	35	35
	NZ	0.50	0.56	0.59	0.56	0.59	2	2	224	224

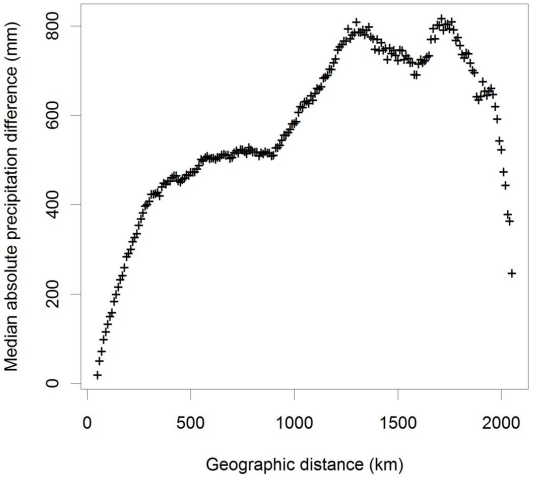
	SA	0.50	0.55	0.56	0.55	0.55	124	124	15	15
	SW	0.50	0.55	0.64	0.55	0.64	2	2	186	186
Combined-	AW	0.50	0.57	0.61	0.57	0.61	4	4	26	26
random-	CA	0.50	0.54	0.64	0.54	0.63	6	6	205	205
adjusted	NS	0.50	0.59	0.70	0.59	0.70	5	5	31	31
	NZ	0.51	0.61	0.65	0.60	0.64	4	4	151	151
	SA	0.50	0.54	0.60	0.54	0.60	113	113	17	17
	SW	0.50	0.62	0.77	0.61	0.77	2	2	174	174

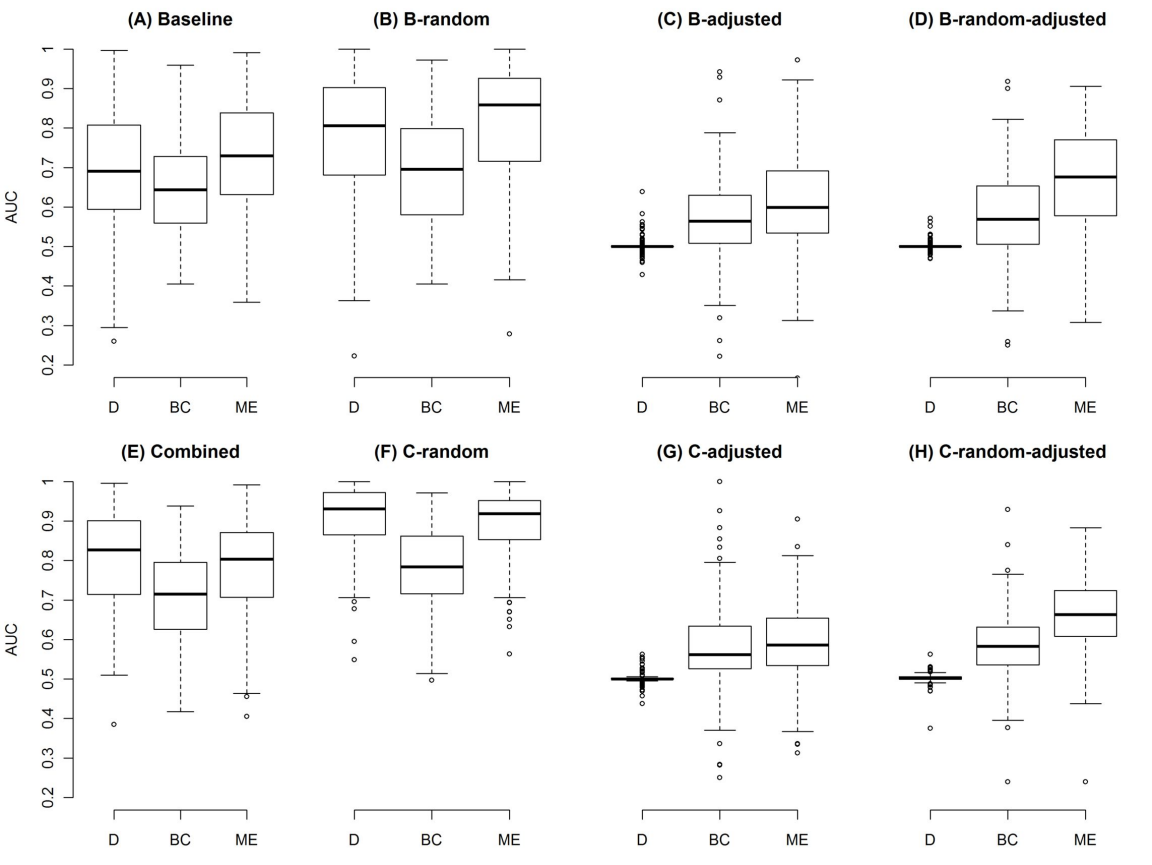
Figure 1. The relation between the geographic distance between sites and the similarity of their climate (average annual temperature and annual precipitation). Obtained by computing the geographic distance between 2500 random terrestrial sites (excluding Antarctica) and the absolute difference between their mean annual temperatures and total annual precipitation according to the WorldClim database (Hijmans et al. 2006) for 100 km wide bins. Climatic similarity is linearly dependent on geographic distance at geographic distances relevant for species distribution modeling (< 2500 km). Temperature values are also similar at very large distances because sites become near-antipodal (at the opposite side of the earth), and hence at the same latitude in another hemisphere.

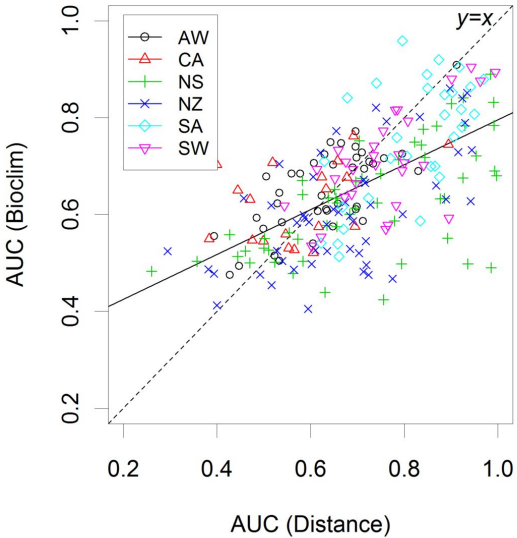
Figure 2. AUC for 226 species modeled with an inverse-distance geographic null-model (D) and two species distribution models: Bioclim (BC) and MaxEnt (ME) for eight treatments. See Table 1 for an explanation of the treatments. The median value is indicated by the black line and first and inter-quartile range by the box. Whiskers cover the full range of the data, except when there are outliers (indicated as dots).

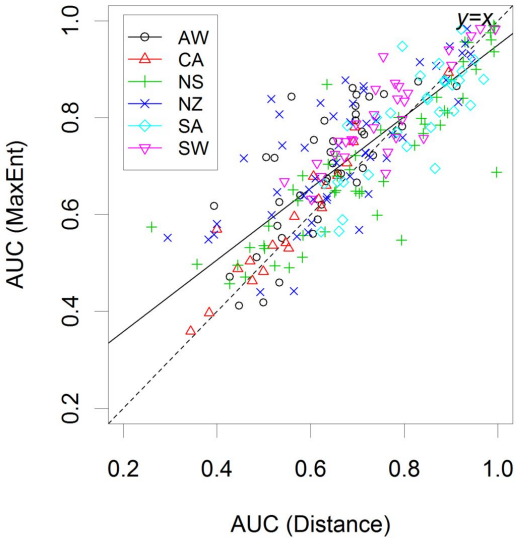
Figure 3. AUC for two species distribution models (Bioclim and MaxEnt) versus AUC for the geographic null-model, for 226 species from six regions. The thick line is the linear regression line (slope = 0.46, R^2=0.38 for Bioclim, and slope =0.74, R^2=0.68 for MaxEnt). The dashed line represents $y=x$. The symbols represent different regions (AW = Australian Wet Tropics, CA=Canada, NS=New South Wales, NZ=New Zealand, SA=South America, SW=Switzerland)


Median absolute temperature difference (° C)
0
5
10
15
0
500
1000
1500
2000
Geographic distance (km)


Median absolute precipitation difference (mm)
0
200
400
600
800
0
500
1000
1500
2000
Geographic distance (km)


(A) Baseline
(B) B-random
(C) B-adjusted
(D) B-random-adjusted
(E) Combined
(F) C-random
(G) C-adjusted
(H) C-random-adjusted
AUC
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
D
BC
ME


AW
CA
NS
NZ
SA
SW
y=x
AUC (Bioclim)
AUC (Distance)
0.2
0.4
0.6
0.8
1.0


AW
CA
NS
NZ
SA
SW
y=x
AUC (MaxEnt)
AUC (Distance)
0.2
0.4
0.6
0.8
1.0