Goal-oriented evaluation of species distribution models' accuracy and precision: True Skill Statistic profile and uncertainty maps

Authors: Alejandro Ruete[1*], Gerardo C. Leynaud[2]

Affiliation:

[1] Swedish Species Information Centre, Swedish University of Agricultural Sciences (SLU). P.O. 7007, SE-750 07. Uppsala, Sweden.

*Corresponding author: alejandro.ruete@slu.se; +46-18-672453

[2] Centro de Zoología Aplicada, Facultad de Ciencias Exactas Físicas y Naturales, and Instituto de Diversidad y Ecología Animal (CONICET - UNC). Universidad Nacional de Córdoba. Córdoba (5000), Argentina.

Running title: Goal-oriented evaluation of SDM's performance

ABSTRACT

1. The use of species distribution models' (SDM) is limited by its performance in terms of accuracy, precision, or the spatial distribution of model errors. Despite the wide acceptance of some standard statistics used to evaluate SDM, there is currently a strong on-going debate as to their use. The "area under the curve" (AUC) is a popular measure used to evaluate SDMs; however, it does not provide complete information about model accuracy. The maximum True Skill Statistic (TSS) is another statistic that is gaining acceptance. However, evaluations of a model's accuracy solely based on this statistic may also be misleading. We investigate the use of alternative methods to evaluate the performance of SDMs, to objectively compare among different modelling approaches.
2. We evaluate the performance of SDMs fitted to simulated and real data by contrasting model predictions to additional validation datasets. We propose visualising TSS scores over the whole detection threshold range (TSS profile).
3. We show how models with similarly good performance according to AUC, present very different results and may serve to different purposes. Also, a high maximum TSS may not guarantee accurate predictions and should be accompanied by the threshold where the maximum is reached (t*). We observe that the higher t* the better predicted observations correlate with confirmed observations. Also, SDM predictions should be accompanied with the corresponding uncertainty map to avoid misleading conclusions. Too high or too widely spread uncertainty on such maps would question the overall accuracy of the model.
4. Whether the model is intended to detect all potential observation sites (sensitive model) or to accurately predict where confirmed observations could be found (specific model) sets a different performance targets to be achieved by the model.

The approach proposed helps to discern which SDM may best suit the intended goals. Furthermore, the TSS profile helps i) to evaluate the overall performance of SDMs and compare among them, ii) to identify the main source of error, and iii) to select a detection threshold depending on the maps intended use.

Keywords: AUC, Bayesian inference, *Chelonoidis chilensis*, MaxEnt, presence-only data, sensitivity, specificity, TSS.

INTRODUCTION

Species distribution models (SDMs) have been widely used to test biogeographic hypotheses (Mourrelle & Ezcurra 1996; Leathwick 1998), for species delimitation (Raxworthy *et al.* 2007), to assess the impact of global climatic changes on species (Araújo & New 2007; Coetzee *et al.* 2009), to establish conservation priorities (Margules & Pressey 2000; Nori *et al.* 2011), and to predict the impact and distribution of invasive species (Nori *et al.* 2011). According to its use we may choose a modelling approach that better suits the goal of the study as different modelling approaches involve different trade-offs between accuracy and generality (Guisan & Zimmermann 2000; Guisan & Thuiller 2005).

SDMs' performance, i.e. accuracy and precision of model predictions, depends on the quality of the observation data, the model formulation and assumptions, and the set of explanatory variables included. Based on the uncertainty arising from the many modelling approaches available, ensembles of models are used to improve prediction accuracy by reporting only the agreement among models (Araújo & New 2007). Although model ensembles are increasingly favoured in different disciplines (Collins 2007; Araújo & New 2007), there are suggestions that not every model is accurate enough to be included in an ensemble (Knutti 2010).

Evaluation of individual model accuracy is therefore still crucial, and objective statistics are required for comparison between alternative modelling approaches (Hirzel *et al.* 2006; Liu *et al.* 2011; Cheaib *et al.* 2012; Jiménez-Valverde 2012). Despite the wide acceptance of some standard statistics, there is currently a strong on-going debate as to their use (Allouche *et al.* 2006; Lobo *et al.* 2008; Jiménez-Valverde 2012).The "area under the receiver operating characteristic curve" (AUC; Hanley & McNeil 1982) is a statistic currently considered to be the standard method to assess the accuracy of predictive distribution models (Jiménez-Valverde 2012). It is used for evaluating both

binary and continuous probability maps. AUC scores range from 0 to 1, where models with scores higher than 0.5 predict better than random draws. However, the AUC statistic was designed to evaluate only the sensitivity of detection methods, e.g. radar signals or X-ray images (Green & Swets 1966; Hanley & McNeil 1982), not specificity (i.e. predicted absences). It does not give information about the spatial distribution of model errors (Lobo *et al.* 2008). The AUC provides information about the generalist or restricted distribution of a species along the range of predictor conditions in the study area, but it does not provide information about the performance of the model in terms of accuracy and precision (Lobo *et al.* 2008).

An alternative to AUC is the True Skill Statistic (TSS; Allouche *et al.* 2006): a simple and intuitive measure for the accuracy of species distribution models. Predictions are contrasted to a validation dataset to derive the model's sensitivity (i.e. proportion of presences accurately predicted) and specificity (i.e. proportion of absences accurately predicted). Sensitivity and specificity are independent of each other when compared between models, and are also independent of prevalence – i.e. the proportion of observed sites in which the species was recorded as present (Allouche *et al.* 2006). The TSS is defined as sensitivity + specificity – 1, and ranges from −1 to +1, where +1 indicates perfect agreement and values of zero or less indicate a performance no better than random. However, this statistic is restricted to binary (presence-absence) maps requiring an arbitrary detection threshold, and TSS varies significantly depending on that threshold. There is a detection threshold at which TSS is maximized (max(TSS)). This maximum value has been used as a threshold independent-accuracy statistic (Liu *et al.* 2011), and even as a criterion for including models in ensembles (Diniz-Filho *et al.* 2009). However, as we will discuss, model selection based solely on max(TSS) can be misleading.

Model precision is also critical when evaluating SDM predictions. Particularly with low quality data (e.g. presence-only data), SDM's precision will depend on how the model accounts for data uncertainty (Congdon 2003). A direct way to evaluate SDMs' precision is to observe the spatial distribution of the models' confidence (or credible) interval on 'uncertainty maps'. Uncertainty maps show the precision of predicted continuous probabilities, augmenting the information contained on prediction maps based on point estimates. However, even with the advances in modelling techniques that account for different sources of uncertainty (Congdon 2003; Argáez *et al.* 2005; Clark & Gelfand 2006; Soberon & Nakamura 2009) few studies report or explore uncertainty maps for single SDMs (but see e.g. Argáez *et al.* 2005).

The goals of any study will influence whether: (i) continuous probability maps or binary presence-absence maps assuming a detection threshold are used (Liu *et al.* 2005; Jiménez-Valverde & Lobo 2007), (ii) accurate point estimates of predictions are enough or high precision on predictions are also needed, and (iii) it is needed to compromise the models' ability to detect true absences (model specificity) by the models' ability to detect true presences (model sensitivity). Thus, our aim is to show how different modelling approaches may best suit different goals depending on their performance, and therefore should not necessarily be contrasted with each other in an ensemble. We offer to researchers and practitioners tools to discern which models may best suit these goals. We first compare AUC scores and explore the usefulness of visualising TSS scores over the whole detection threshold range (TSS profile), for five simulated SDMs with known accuracy and precision, based on simulated data. We show how the TSS profile allows an evaluation of general model accuracy and precision, and to perform a goal-oriented selection of a detection threshold. Then, we fit SDMs to real presence-only data using two different modelling approaches to assess the utility of each approach based on the

models TSS profile. We base the following study on presence-only data to highlight and overcome some of the problems associated with low quality data, and the consequential model evaluation; however, these evaluation methodologies are also applicable for presence-absence data.

METHODS

Simulated data exercise

To control for the response of the statistical measures of model performance (AUC and TSS scores) to different amount of accuracy and precision we created a set of five simulated SDM with known accuracy and prediction around the simulated observed data. In total we generated 1000 presence-absence data points consisting in 94 presences and 906 absences (Fig. 1a), over a grid of 100x100 pixels (10000 prediction values). Specifically, we generated an overlap between two 2D Gaussian kernels from normal distributions x1 = Normal(0,0.2), y1= Normal(0,0.5) and x2 = Normal(-0.5,0.2), y2 = Normal(1,0.5), normalized the distribution values and placed the 1000 simulated observations points randomly over the kernel. Points overlapping probabilities of observation ≥ 0.8 where set as presence. Model 1 was then set as a normalized Gaussian kernel such that observations overlap with observation probabilities $p \geq 0.8$ (Fig. 1). Model 2 is the same as Model 1 where we added to each pixel a random value (noise) drawn from a Normal distribution with mean = 0 and standard deviation = 0.05, and then normalized (0-1). Model 3 is as Model 1, but observation probabilities where homogeneously reduced by 40%. Model 3 then predicts consistently low observation probabilities, as if e.g. the data was not enough to properly inform the model. Model 4 is as Model 1, but adding to each pixel noise drawn from a Normal distribution with mean = 0 and standard deviation = 1, and then normalized (0-1). That is, Model 4 is close to a

totally imperfect prediction. Model 5 predicts a different core area for the species, as if the model predictions are inaccurate for certain areas. That is, Model 5 systematically predicts observations in regions where there were no presence data, and fails to predict observations where there were presence data.

We calculated the AUC score for each simulated SDM with the *SDMtools* package for R (VanDerWal *et al.* 2012) using the complete dataset (n = 1000). We also calculated for each simulated SDM the True Skill Statistic (TSS; Allouche *et al.* 2006) for every detection thresholds (i.e. $0 \leq t \leq 1$) describing the TSS profile with a resolution of 0.01 units. A model performs accurately at a certain detection threshold if it scores a TSS higher than 0.5 (Allouche *et al.* 2006; Liu *et al.* 2011). The TSS profile comparing observations with themselves (instead of with predictions) serve as a reference profile for a model with perfect fit to the data (perfect-fit TSS profile henceforth, Fig. 1b). The TSS profile for each simulated SDMs was calculated contrasting 50, 100 and 1000 prediction values (pixels) with observation data points (used as validation data).

Real data exercise

The study species

The common Chaco tortoise, *Chelonoidis chilensis* (Testudinidae, Gray 1870), is found mainly in the ecoregions of Monte and Chaco (Fig. 2) in Argentina, Bolivia and Paraguay (Cei 1993; Cabrera 1998). It is a burrow-nesting species, found on sandy soils in scrublands or dry forests (Cei 1993; Cabrera 1998) up to 1200 m.a.s.l. (Cerro Nevado, Mendoza; Richard 1988). In Argentina, the species is mainly threatened by habitat degradation and poaching (Chebez 2009); thus is categorized as Vulnerable by the IUCN (Tortoise & Freshwater Turtle Specialist Group 2010) and is CITES listed. In the current study the species is defined after Fritz et al. (2012), who concluded that *Chelonoidis*

178 *chilensis* (Gray, 1870), *C. donosobarrosi* (Freiberg, 1973) and *C. petersi* (Freiberg 1973)
179 are the same species (i.e. *C. chilensis*).
180
181 *Data collection*
182 We collected confirmed observations of the Chaco tortoise dated 1950-2012 from the
183 EMYSystem World Turtle Database (http://emys.geo.orst.edu/), and from scientific
184 literature (Waller 1986; Buskirk 1993; Ergueta & Morales 1996; Cabrera 1998; Ernst
185 1998; Richard 1999; Gonzales *et al.* 2006; Fritz *et al.* 2012). We merged in a GIS vector
186 layer all reported observations using QuantumGIS 1.8 (Quantum GIS Development Team
187 2012). In case of overlap within 5 km we kept only the latest observation. For a complete
188 list of the 244 observations and corresponding sources see Table S1 in Supporting
189 Information. We arbitrarily defined the study area (Fig. 2) larger than the observed
190 species distribution to include surrounding areas where the species is known to be absent.
191 We excluded Chile from the study area because the Andean Mountain Range is a
192 physical barrier the species cannot pass.
193 We obtained geographic and bioclimatic data from raster layers with 5 km
194 resolution from world databases (WorldClim, Hijmans *et al.* 2005; WorldMaps, Hengl
195 2009). The complete list of variables included in the study is presented in Table S2. We
196 did not included in the analysis land-use variables because the data collected covers a
197 wide temporal range (1950-2012), and the landscape has changed dramatically over this
198 time period.
199
200 *Modelling the species distribution*
201 We developed a Bayesian spatially expanded logistic (BSEL) model (Casetti 1997;
202 Congdon 2003) to obtain the probability of observation at non-visited locations. Non-

visited locations were randomly located with the same density as the observed locations (~0.0004/km^2). Given the nature of presence-only data, predicted probabilities combine the probability of the species being at the location, the probability of an observer being at the same location, and the probability of the observer finding the species (Lobo *et al.* 2010). The Bayesian approach allows us accounting for all three uncertainty sources on each observation, and displaying the model uncertainties on an uncertainty map. We assume that observations at every non-visited location *i* are distributed according to a Bernoulli distribution $Obs_i \sim \text{Bernoulli}(p^*_i)$, where p^*_i is an *a priori* probability distribution generated from confirmed observations (Fig. 2b). We generated the *a priori* probability distribution as a quadratic density kernel raster layer using the R package "*splancs*" (Rowlingson *et al.* 2013). By generating a prior distribution from the observations, we assume that the entire study region has been sampled with the same intensity.

We then modelled observations Obs_i according to a logistic model, $Obs_i \sim \text{Bernoulli}(p_i)$, The spatially expanded model (Casetti 1997; Congdon 2003) assumes that the effect of an explanatory variable on the response variable p_i can vary among the observed locations. This assumption is particularly convenient when fitting species distribution model along large ranges, where the species can be locally adapted to e.g. temperature ranges (Turchin & Hanski 1997; Nilsson-Örtman *et al.* 2013). For further details on the modelling approach see Appendix S1.

The final model presented (Table 1) is the result of a forward stepwise selection procedure based on the deviance information criterion (DIC), an information-theoretic criterion similar to Akaike's information criterion (AIC; Burnham & Anderson 2002), that is appropriate for Bayesian hierarchical modelling (Spiegelhalter et al. 2002). For further details on the selection procedure and all tested variables see Appendix S1 and Table S2.

Once the final model was obtained, we generated maps for the observation probability. We predicted observation probabilities for regularly distributed locations with the same resolution as the raster images for environmental variables (i.e. 5 km). We generated raster layers for the mode and for the length of the 95% credible interval (95% CI). The length of the 95% CI is a measure of precision ranging from 0 (precise) to 1 (imprecise).

For comparison, we generated a map with MaxEnt (Elith *et al.* 2011) using the same sets of variables as the final BSEL model. MaxEnt is a widely used free program for species distribution models based on machine learning algorithms and maximum entropy (Elith *et al.* 2011). We are aware that better performance may have been obtained with MaxEnt adding more variables, however, for the comparison purpose we used the same selection of variables than those chosen for BSEL.

Model evaluation

We calculated the AUC index for both models (i.e. BSEL and MaxEnt) with the *SDMtools* package for R (VanDerWal *et al.* 2012), contrasting predictions against data generated from the *a priori* observation probability distribution. Then, to calculate the TSS profiles, we contrasted model predictions with two independent data sets of observations of Chaco tortoises in Argentinean and in Bolivian protected areas (a Paraguay dataset was not available). The first data set is mainly based on park rangers reports, and includes 144 Argentinean protected areas in the study area (Sistema de Información de Biodiversidad, SIB; Administración de Parques Nacionales 2012). The second data set was put together in the framework of a doctoral thesis (Embert 2007), and includes museum and field systematic collections for 38 Bolivian protected areas in the study area. The species were reported in 12 Argentinean and 3 Bolivian protected areas

(Table S3). With these independent observations as a validation, we calculated the TSS profile for both the BSEL and MaxEnt predictions.

RESULTS

Simulated data exercise

Only a nearly imperfect prediction (Model 4) could be separated from accurate models using AUC scores. An inaccurate model (Model 5) scored an AUC of 0.92, but this is still a very high AUC score. Alternatively, the more accurate and precise a model is, the more similar a model's TSS profile is to the perfect-fit TSS profile. We observe that the detection threshold where the maximum TSS score is obtained (t^* henceforth) is the threshold at which the best compromise between sensitivity and specificity is reached (Fig. S1). Except for the TSS profile of a perfectly fitting model, in which t^* is infinitely close to 1 from below, t^* is always lower than 1. For accurate models (Models 1, 2 and 3), regardless of their precision, we observe an abrupt decrease in TSS scores at detection thresholds higher that t^*, indicating a drastic loss of sensitivity. TSS scores at detection thresholds lower than t^* decreases because of loss of model specificity. Models that only provide weak signals (low probabilities; e.g. Model 3) could score high max(TSS) at lower t^* that more precise models do. Inaccurate models (Model 5) show a general decrease in both max(TSS) and t^*, reflecting a serious compromise between sensitivity and specificity to acquire the best information from the model.

Regarding the sample size of validation data points, it is important to be aware about how small validation sample size could affect the estimate of t^*. However, this problem seems more relevant to inaccurate models than to imprecise ones.

279 **Real data exercise**

280 The species distribution predicted with the Bayesian spatially expanded logistic (BSEL)
281 model for *Chelonoidis chilensis* was mainly driven by temperature related variables, but
282 included water availability in the reproductive period (Table 1). From this we generated
283 probability and uncertainty maps for the species' distribution (Fig. 3). Both BSEL and
284 MaxEnt predictions suggest that the fundamental niche (i.e. potential suitable sites) of the
285 species is continuous across Argentina, West Paraguay and South Bolivia, in
286 consideration of the variables, scale and resolution used. In general terms, temperature
287 related variables constrain the latitudinal and altitudinal range of the species, while
288 precipitation related variables constrain it in longitude.

289 The uncertainty of the BSEL model (the opposite of its precision) was generally
290 low (i.e. 95% CI length < 0.5, Fig. 3b) and is lower in areas where the observation
291 probability is close to either 0 or 1 (Fig. S2). However, uncertainty is highest in poorly
292 sampled areas.

293 According to the AUC the performance of both BSEL and MaxEnt predictions is
294 high and equally good in terms of accuracy (AUC = 0.92). Despite this, there are major
295 differences between the predicted distribution maps of these two approaches at the north
296 and east of the species' distribution (Fig. 3). Both models perform better describing the
297 species distribution in Argentina than in Bolivia. In Argentina max(TSS) is higher for the
298 BSEL than for MaxEnt. In other words, with this particular set-up, BSEL is more
299 accurate predicting observations than MaxEnt. However both models performed
300 accurately, i.e. max(TSS) > 0.5 (Fig. 4a). Maximum TSS was 0.88 at $t^* = 0.45$ for BSEL
301 and 0.73 at $t^* = 0.25$ for MaxEnt. However, the TSS scores for BSEL are higher than 0.8
302 for thresholds of up to 0.6. That is the model's sensitivity is very high up to $t = 0.6$; Fig.
303 4a). Both models' predictions generally overlaps with published distribution maps for the

species (Waller 1986; Ernst 1998; Richard 1999; Administración de Parques Nacionales 2012; Fritz *et al.* 2012) and with the ecoregions where the species has been described (Fig. 2a). The main difference between the models' predictions is that MaxEnt predicted higher observation probabilities for protected areas in the Espinal and to the east (Fig. 3c) where the species has not been observed. On the other hand, comparing to confirmed observations in Bolivian protected areas both models' TSS profiles were very different than the perfect-fit TSS profile (Fig. 4b). Maximum TSS was 0.77 at $t^* = 0.02$ for BSEL and 0.97 at $t^* = 0.04$ for MaxEnt. Both models are very imprecise, but MaxEnt is more sensitive than BSEL at low detection thresholds.

DISCUSSION

We put forward alternative methods to evaluate and compare the performance of species distribution models (SDMs). We show how models with similarly good performance according to AUC and max(TSS) present very different results and may serve different purposes. Therefore, we suggest analysing the complete TSS profile to evaluate and compare the overall quality of SDM results. Uncertainty maps and TSS profiles can help to objectively evaluate and compare the performance of SDMs, to select a detection threshold depending on the intended use of the map, and to identify the main source of error of a continuous probability map. In general we now understand that model uncertainty, i.e. lack of precision to distinguish between presence and absences reduces the model specificity (high commission error). In other words, high max(TSS) scores reached at low detection thresholds (t^*) are a sign of low model precision. Lack of accuracy in predictions, however, reduces both max(TSS) and t^*.

329 *Model precision and accuracy*

330 An honest display of model uncertainties (i.e. as the opposite of model precision) is
331 crucial to evaluate and validate model predictions, no matter if continuous or binary maps
332 are used. In general, probabilities obtained for each pixel on the map have uncertainties
333 associated to the observation events (Lobo *et al.* 2010), as well as to the model that
334 generated those probabilities (Congdon 2003; Clark & Gelfand 2006). Model uncertainty
335 complements the information contained on point estimate predictions, and should be
336 displayed as yet another SDM result. However, uncertainties are generally lacking from
337 most SDM reports (Congdon 2003; Clark & Gelfand 2006), even if the approach used
338 can produce them. Species distribution maps generated with low quality data (e.g.
339 presence-only data) could be dangerously misleading if not accompanied with the
340 corresponding uncertainty map. Too high or too widely spread uncertainty would also
341 question the accuracy of the model, suggesting that more observations or alternative
342 explanatory variables should be considered in the study. On the real data exercise we
343 observe that higher uncertainty is expected on transition areas between high and low
344 estimated probabilities or on poorly sampled areas (Figs. 3b and S2). Uncertainty maps
345 can be a valuable tool for designing field work efficiently. The researcher can then decide
346 to focus future sampling effort either on areas with high uncertainty to validate the model
347 or on areas with high probabilities of observation and low error to sample more
348 efficiently.

349 Adding models into an ensemble could increase precision in SDM predictions
350 (Araújo & New 2007; Garcia *et al.* 2012). However, not all models should be included on
351 the same ensemble or average (Knutti 2010), especially when any of the models is
352 particularly inaccurate. It is therefore crucial to evaluate individual models' accuracy.
353 The AUC score is not a good measure of model accuracy (Lobo *et al.* 2008; Jiménez-

Valverde 2012). Different models (or even modelling approaches) with similarly high AUC can return significantly different results. As discussed before, high max(TSS) scores alone may not guarantee good performance either, as high TSS scores at low t^* are a sign for imprecise models (Fig. 4 comparing TSS profiles for Argentina and Bolivia). However, max(TSS) scores are sometimes reported without specifying the detection threshold at which it is reached (e.g. Soininen *et al.* 2012; Comte & Grenouillet 2013).

We argue that t^* is also necessary to evaluate the accuracy of a SDM. The TSS profile shows TSS scores as the detection threshold (t) change from 0 to 1. If we think of a hypothetical perfectly-fitting model that can separate presences from absences, one would expect a "flat" TSS profile (i.e. TSS = 1 for $0 \leq t < 1$; Fig. 1b). That is, the only predicted values would be 0 or 1, and the model would perform equally at any t lower than 1. For any other model, we observe that the higher t^* the better the correlation between predicted high probabilities of observations and confirmed observations are. That is, the higher t^* the better the model explains the variability along the species niche dimensions. In that case, any thresholds lower than t^* implies higher sensitivity (less omissions) but lower specificity (more commissions).

The TSS profile can also help to determine whether the model is suited to the intended goal of the study. As stated above, the higher t^* the more accurately our predictions can discard unsuitable sites for the species, without losing sensitivity for sites where the species can be. In general, algorithmic models like MaxEnt are expected to present low omission error (high sensitivity) but high commission error (low specificity) (Guisan & Zimmermann 2000). For example, we observe that MaxEnt predictions have a lower max(TSS) and t^* than BSEL predictions, because of the higher commission error in the Espinal ecorregion and the higher omission in the High Monte ecoregion (Figs. 2 and 3). The lower t^* the larger the assumed distribution area needs to be not to miss sites

where the species could be (i.e. high commission error). That is, as we observe for predictions in Bolivia, one should assume that any probability higher than 0.04 could be a confirmed observation. Such models are good for detecting low signals of species presences, identifying most of the potential suitable sites for the species, but do not help to understand the relationship of the species with the environment.

Selection of detection threshold

For many practical applications it is necessary to transform continuous maps to binary presence-absence maps assuming a (more or less) objective detection threshold (Liu *et al.* 2005; Jiménez-Valverde & Lobo 2007). Liu et al. (2005) and Jiménez-Valverde and Lobo (2007) previously discussed that a threshold of 0.5 is not always the best option, although it is often used. Theoretically, in a perfectly-fitting model, predicted probabilities could be interpreted as the expectancy of a Bernoulli probability distribution, where $p_i = 0.5$ describes a site on which an observation would be a purely random event. In such a case, a threshold of 0.5 separates sites where it is likely to find the species from those where it is not. However, the further away the model is from the perfect fit the less the model predictions reflect the true probability of observation. Alternatively, max(TSS) has been used as a criteria to select detection thresholds (Albouy *et al.* 2012; Cheaib *et al.* 2012). From our results, we conclude that it would be unfair to convert continuous predictions to binary at an arbitrary $t = 0.5$. Basically, when the detection threshold changes from 0 to 1, the rate of well-predicted presences decreases while the rate of well-predicted absences increases. The best compromise between sensitivity and specificity is reached at t^*. Therefore, we may select different t for each model, reducing comparability.

Here is how the complete TSS profile can help to determine one detection threshold for all models being compared. Despite BSEL max(TSS) is scored at t = 0.45, the model's sensitivity and specificity are still very high for t = 0.25 - 0.6 (Fig. 4a). That is, accurate niche description are also obtained without big losses in sensitivity at t = 0.6. Similarly for MaxEnt predictions, max(TSS) is scored at t = 0.25, while TSS is not much lower at t = 0.4. Therefore, we could compare both model's predictions using t = 0.4.

It is the researcher's task to decide (depending on the study's goal) whether it is needed not to miss potential observations, or if it is preferable to be conservative with predictions. If t is selected below t* predictions are less specific, but probably captures more observations. Inversely, if t is selected above t* predictions may be more specific but the loss of sensitivity may be much greater than any gain in specificity.

It is also the researcher's task to decide on which side of the detection threshold he/she wants the most of the model's uncertainty. We observed that a non-perfectly fitting model has the highest uncertainty (length of 95% CI) on regions where predicted observation probabilities are close to 50% (Fig. S2). As previously discussed, using the BSEL model in Argentina, either 0.4 or 0.6 may be good thresholds alternatives for respectively detecting the species or for predicting its presence (Fig. 4). Choosing t = 0.4 would leave higher uncertainties on values interpreted as presences (Fig. S2). The opposite is also true for t = 0.6.

Further practical applications and consideration

The sources of commission error (i.e. false positives) can be identified by contrasting different evaluation approaches, i.e. uncertainty maps, previous distribution maps and TSS profile. Commission errors could be caused by i) overestimation of probability of observation, ii) incomplete validation dataset (i.e. lack of complete surveys

or reports for some protected areas), or iii) local extinction of the species by the time of the validation data is collected. High probabilities of observation (beyond a set threshold) with high uncertainty on areas where the species has never been described before, is likely to be due to bad performance of the model (i). Alternatively, high probabilities of observation with low uncertainty on protected areas where the species was not reported, but that overlaps previous delimitations of the species distribution are likely to be due to lack of information on single protected areas (ii) or local extinction (iii). When using protected areas as the set up for independent data, it is important to consider the possible bias present on their distribution, and how it affects commission error. For example, because of heavily unbalanced distribution of protected areas, commission error on the east of the species distribution is underestimated (Espinal and Pampas ecoregions, <1% protected) if compared to the cover on the core distribution area (Monte and Chaco ecoregions, 3.7% protected)(Chebez 2009).

It is important to note that TSS is not sensitive to variations in prevalence in the validation dataset (Allouche *et al.* 2006), but it is to validation sample size. TSS profiles are rougher the smaller the validation dataset (Fig. 1). However, poor model performance in localized areas due to low number of samples cannot be detected with subsamples of the original dataset. Therefore independent validation datasets are needed.

ACKNOWLEDMENTS

Thanks to M. Low for valuable comments on the manuscript.

LITERATURE CITED

Administración de Parques Nacionales. (2012). Sistema de información de biodiversidad. *www.sib.gov.ar*. Retrieved from

http://www.sib.gov.ar/busqueda.php?qry=Chelonoidis&qrydo.x=-1067&qrydo.y=-145

Albouy, C., Guilhaumon, F., Araújo, M.B., Mouillot, D. & Leprieur, F. (2012). Combining projected changes in species richness and composition reveals climate change impacts on coastal Mediterranean fish assemblages. *Global Change Biology*, **18**, 2995–3003. Retrieved August 22, 2013,

Allouche, O., Tsoar, A. & Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232. Retrieved May 20, 2013,

Araújo, M.B. & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, **22**, 42–47. Retrieved December 2, 2010,

Argáez, J.A., Andrés Christen, J., Nakamura, M. & Soberón, J. (2005). Prediction of potential areas of species distributions based on presence-only data. *Environmental and Ecological Statistics*, **12**, 27–44. Retrieved August 9, 2012,

Buskirk, J.R. (1993). Distribution, status and biology of the tortoise, Geochelone chilensis, in Río Negro Province, Argentina. *Studies on Neotropical Fauna and Environment*, **28**, 233–249. Retrieved October 2, 2012,

Cabrera, M. (1998). *Las tortugas continentales de Sudamérica austral*. Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina.

Casetti, E. (1997). The expansion method, mathematical modeling, and spatial econometrics. *International Regional Science Review*, **20**, 9–33. Retrieved August 18, 2012,

Cei, J.M. (1993). *Reptiles del noroeste, nordeste y este de Argentina. Herpetofauna de las Selvas Subtropicales, Puna y Pampas*. Museo Regionale di Scienze Naturali, Torino.

Cheaib, A., Badeau, V., Boe, J., Chuine, I., Delire, C., Dufrêne, E., François, C., Gritti, E.S., Legay, M., Pagé, C., Thuiller, W., Viovy, N. & Leadley, P. (2012). Climate change impacts on tree ranges: model intercomparison facilitates understanding and quantification of uncertainty. *Ecology Letters*, **15**, 533–544. Retrieved August 22, 2013,

Chebez, J. (2009). *Los que se van: Fauna Argentina amenazada*. Albatros, Argentina.

Clark, J.S. & Gelfand, A. (2006). *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications*. Oxford University Press, USA.

Coetzee, B.W.T., Robertson, M.P., Erasmus, B.F.N., Van Rensburg, B.J. & Thuiller, W. (2009). Ensemble models predict Important Bird Areas in southern Africa will become less effective for conserving endemic birds under climate change. *Global Ecology and Biogeography*, **18**, 701–710. Retrieved August 22, 2013,

Collins, M. (2007). Ensembles and probabilities: a new era in the prediction of climate change. *Philosophical Transactions of the Royal Society A: Mathematical*,

Physical and Engineering Sciences, **365**, 1957 –1970. Retrieved September 18, 2010,

Comte, L. & Grenouillet, G. (2013). Species distribution modelling and imperfect detection: comparing occupancy versus consensus methods. *Diversity and Distributions*, **19**, 996–1007. Retrieved August 22, 2013,

Congdon, P. (2003). *Applied Bayesian modelling*. John Wiley and Sons, West Sussex, England.

Diniz-Filho, J.A.F., Mauricio Bini, L., Fernando Rangel, T., Loyola, R.D., Hof, C., Nogués-Bravo, D. & Araújo, M.B. (2009). Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, **32**, 897–906. Retrieved August 22, 2013,

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. Retrieved April 24, 2013,

Embert, D. (2007). *Distribution, diversity and conservation status of Bolivian reptiles*. Doctoral Thesis thesis, Boon University, Bonn. Retrieved from http://hss.ulb.uni-bonn.de/2008/1441/1441-engl.htm

Ergueta, P.S. & Morales, C.B. de. (1996). *Libro rojo de los vertebrados de Bolivia*. Asociación para la Biología de la Conservación - Bolivia, La Paz, Bolivia.

Ernst, C.H. (1998). Geochelone chilensis. *Catalogue of American Amphibians and Reptiles*, **668**, 1–4.

Fritz, U., Alcalde, L., Vargas-Ramírez, M., Goode, E.V., Fabius-Turoblin, D.U. & Praschag, P. (2012). Northern genetic richness and southern purity, but just one species in the Chelonoidis chilensis complex. *Zoologica Scripta*, **41**, 220–232. Retrieved April 23, 2012,

Garcia, R.A., Burgess, N.D., Cabeza, M., Rahbek, C. & Araújo, M.B. (2012). Exploring consensus in 21st century projections of climatically suitable areas for African vertebrates. *Global Change Biology*, **18**, 1253–1269. Retrieved February 15, 2012,

Gonzales, L., Muñoz, A. & Cortéz, E. (2006). Primer reporte sobre la herpetofauna de la reserva natural 'El Corbalán', Tarija, Bolivia. *Kempffiana*, **2**, 72–94.

Green, D.M. & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley and Sons.

Guisan, A. & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009. Retrieved March 13, 2012,

Guisan, A. & Zimmermann, N.E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186. Retrieved December 8, 2010,

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.

Hengl, T. (2009). *A Practical Guide to Geostatistical Mapping*. University of Amsterdam, Amterdam, The Netherlands. Retrieved from http://spatial-analyst.net/book/About

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978. Retrieved October 20, 2010,

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, **199**, 142–152. Retrieved August 18, 2011,

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. Retrieved August 22, 2013,

Jiménez-Valverde, A. & Lobo, J.M. (2007). Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica*, **31**, 361–369. Retrieved August 22, 2013,

Knutti, R. (2010). The end of model democracy? *Climatic Change*, **102**, 395–404. Retrieved February 21, 2012,

Leathwick, J.R. (1998). Are New Zealand's Nothofagus species in equilibrium with their environment? *Journal of Vegetation Science*, **9**, 719–732.

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**, 385–393. Retrieved June 13, 2013,

Liu, C., White, M. & Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence–absence data. *Ecography*, **34**, 232–243. Retrieved August 22, 2013,

Lobo, J.M., Jiménez-Valverde, A. & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, **33**, 103–114. Retrieved September 23, 2013,

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151. Retrieved June 13, 2013,

Margules, C.R. & Pressey, R.L. (2000). Systematic conservation planning. *Nature*, **405**, 243–253. Retrieved August 18, 2011,

Mourrelle, C. & Ezcurra, E. (1996). Species richness of Argentine cacti: A test of biogeographic hypotheses. *Journal of Vegetation Science*, **7**, 667–680.

Nilsson-Örtman, V., Stoks, R., De Block, M., Johansson, H. & Johansson, F. (2013). Latitudinally structured variation in the temperature dependence of damselfly growth rates. *Ecology Letters*, **16**, 64–71. Retrieved March 15, 2013,

570 Nori, J., Akmentins, M.S., Ghirardi, R., Frutos, N. & Leynaud, G.C. (2011). American
571 bullfrog invasion in Argentina: where should we take urgent measures?
572 *Biodiversity and Conservation*, **20**, 1125–1132. Retrieved October 4, 2011,

573 Quantum GIS Development Team. (2012). *Quantum GIS Geographic Information*
574 *System*. Open Source Geospatial Foundation Project. Retrieved from
575 http://qgis.osgeo.org

576 Raxworthy, C.J., Ingram, C.M., Rabibisoa, N. & Pearson, R.G. (2007). Applications of
577 ecological niche modeling for species delimitation: a review and empirical
578 evaluation using day geckos (Phelsuma) from Madagascar. *Systematic Biology*,
579 **56**, 907–923. Retrieved September 8, 2012,

580 Richard, E. (1988). Las Yataché (Chelonoidis donosobarrosis: Chelonii, Testudine) de la
581 región del Nevado (Mendoza, Argentina). Apuntes sobre la hitoria natural.
582 *Amphibia y Reptilia*, **1**, 79–92.

583 Richard, E. (1999). *Tortugas de las regiones aridas de Argentina*. L.O.L.A., Buenos
584 Aires Argentina.

585 Rowlingson, B., Diggle, P., Bivand, R., Petris, G. & Eglen, S. (2013). *splancs: Spatial*
586 *and space-time point pattern analysis*. Retrieved from http://CRAN.R-
587 project.org/package=splancs

588 Soberon, J. & Nakamura, M. (2009). Niches and distributional areas: Concepts, methods,
589 and assumptions. *Proceedings of the National Academy of Sciences*, **106**, 19644–
590 19650. Retrieved August 9, 2012,

591 Soininen, J., Korhonen, J.J. & Luoto, M. (2012). Stochastic species distributions are
592 driven by organism size. *Ecology*, **94**, 660–670. Retrieved August 22, 2013,

593 Tortoise & Freshwater Turtle Specialist Group. (2010). Chelonoidis chilensis. *IUCN*
594 *2010. IUCN Red List of Threatened Species*. Retrieved December 21, 2010, from
595 http://www.iucnredlist.org/apps/redlist/details/9007/0

596 Turchin, P. & Hanski, I. (1997). An empirically based model for latitudinal gradient in
597 vole population dynamics. *The American Naturalist*, **149**, 842–874. Retrieved
598 May 23, 2012,

599 VanDerWal, J., Falconi, L., Januchowski, S., Shoo, L. & Storlie, C. (2012). *SDMTools:*
600 *Tools for processing data associated with species distribution modelling*
601 *exercises*. Retrieved from http://CRAN.R-project.org/package=SDMTools

602 Waller, T. (1986). Distribucion, habitat y registro de localidades para Geochelone
603 chilensis (Gray , 1870) (Syn donosobarrosi, petersi) (Testudines, Testudinidae).
604 *Amphibia & Reptilia*, **1**, 10.

605
606
607
608
609

610 **SUPPORTING INFORMATION**
611 Additional Supporting Information may be found in the online version of this article:

612 **Fig. S1.** Sensitivity and specificity of simulated SDM

613 **Fig. S2.** BSEL model uncertainty

614 **Appendix S1**. Bayesian spatially expanded logistic (BSEL) model and model selection
615 procedure

616 **Table S1.** Complete list of observations and sources.

617 **Table S2.** Explanatory variables and model selection.

618 **Table S3.** Presence of *Chelonoidis chilensis* on protected areas in Argentina and Bolivia.

619

TABLES

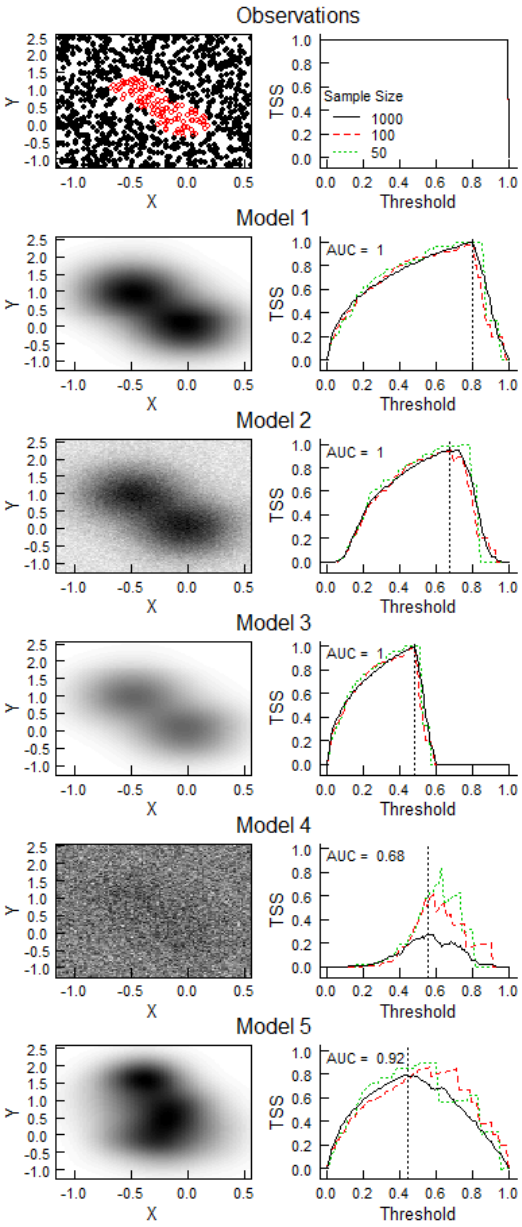
Table 1: Explanatory variables included in the final model.

	DIC[a]	$\bar{\delta}$[b]	$\bar{\delta}$ 95% CI	
Mean annual temperature	930.4	0.56	-4.68	5.14
Max. temperature of warmest month	869.4	1.61	-2.05	5.86
Temperature annual range	853.7	0.03	-2.25	2.12
Precipitation of warmest quarter	824.5	-1.57	-2.35	-0.80

[a]: Deviance Information Criterion (progressive)

[b]: mode of the effect parameter.

641 **FIGURES**



642

643 **Figure 1**: left panels: simulated observations (open circles = 1; filled circles = 0) and

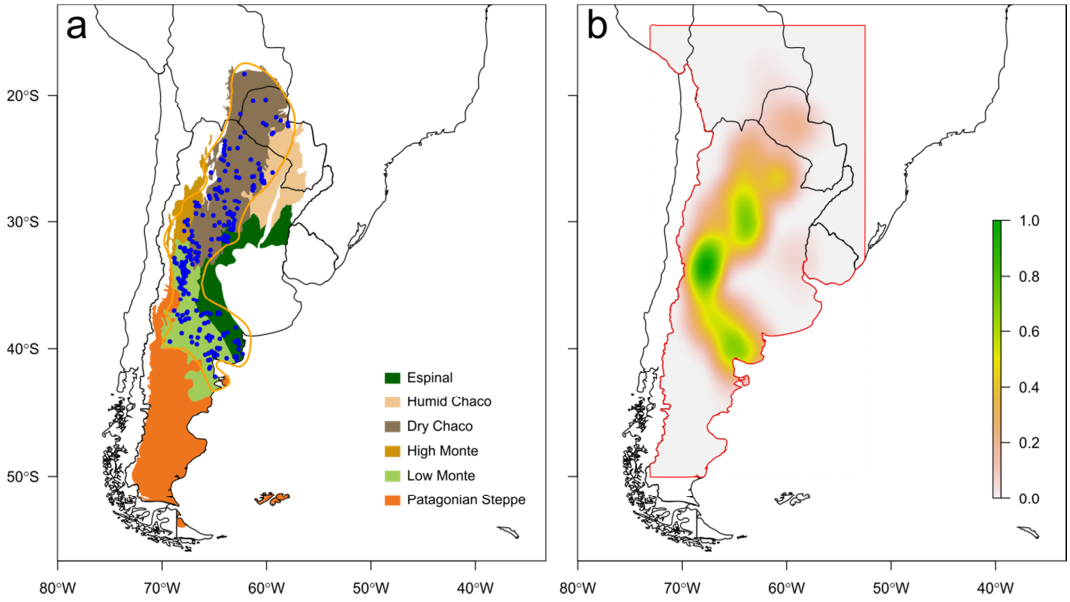
644 species distribution models 1 to 5 (probabilities in linear scale black = 0 to white = 1).

645 Simulated models assume varying accuracy and precision. Right panels: TSS profiles

646 calculated with different sizes of validation dataset. Vertical dashed lines indicate *t**.

647

648



649

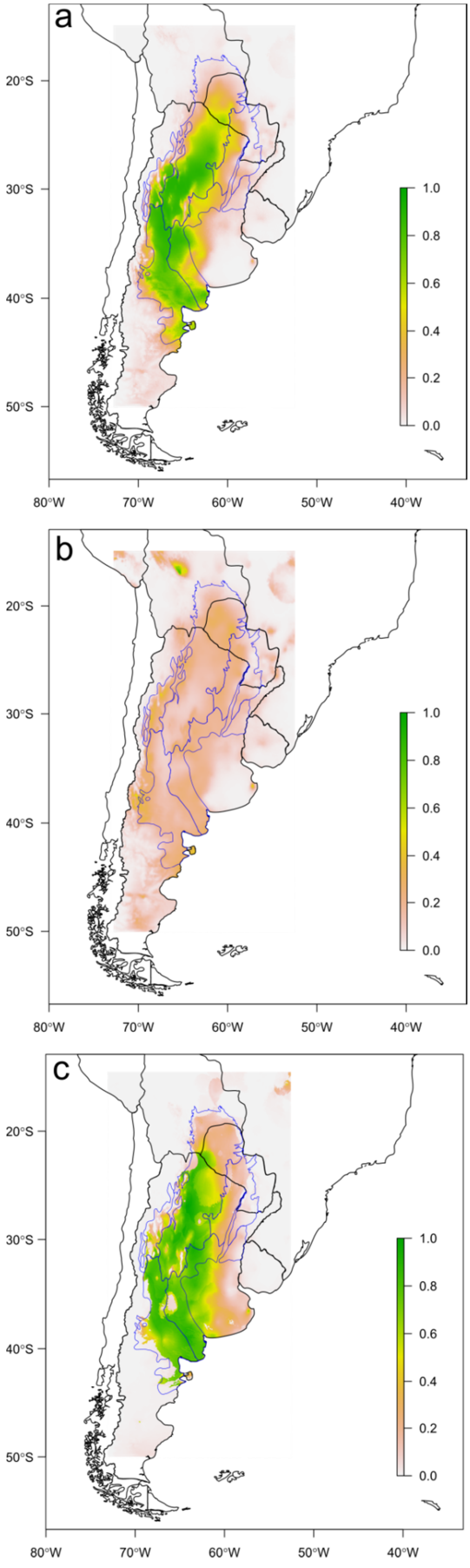
650 **Figure 2**: Map of austral South America, showing a) sites of confirmed observations of

651 *Chelonoidis chilensis* (blue dots) and ecoregions where the species has been observed

652 (coloured polygons); b) a *priori* probabilities of observation (colour scale) estimated from

653 observation densities.

654



655

Figure 3: Maps showing a) mode and b) length of the 95% Credible Interval (CI) of probabilities of observation generated with the Bayesian Spatially Expanded Logistic model (BSEL), and c) probabilities of observation generated with MaxEnt. Both models were fitted to the same set of variables detailed in Table 1. Blue lines show ecoregions delimitation for comparison with Figure 2a.

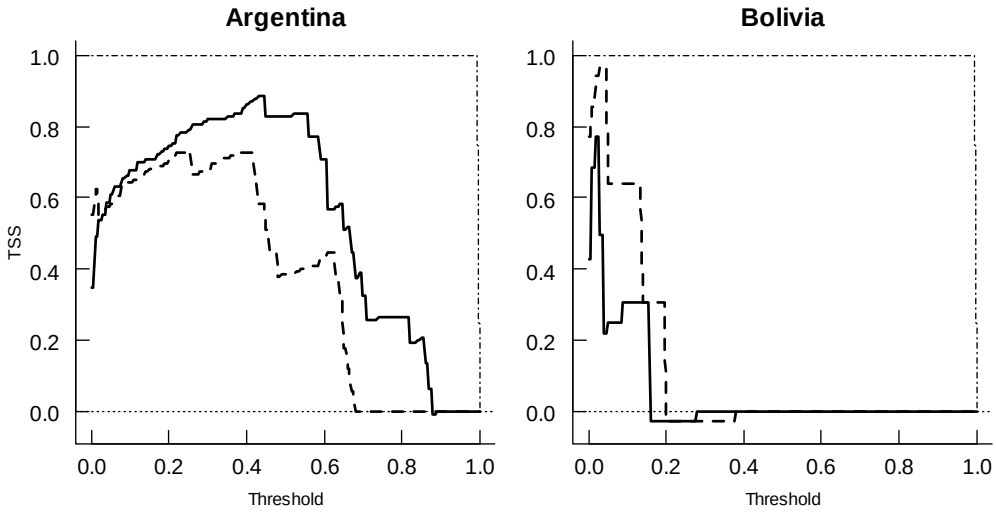


Figure 4: True Skill Statistic (TSS) profile over different detection thresholds, for the Bayesian Spatially Expanded Logistic model (BSEL; solid) and MaxEnt (dashed) predictions compared to independent data sets of confirmed observation on protected areas in Argentina and Bolivia. Perfect fit profile is shown with dot-dashed line.