Australian Centre of Excellence for Risk Analysis

Report Cover Page

ACERA Project
ARC Project Spatial Methods
Title
Reasons for differing performances of species distribution models
Author(s) / Address (es)
Jane Elith, Botany, ACERA, University of Melbourne and Catherine Graham, State University of New York at Stony Brook
Material Type and Status (Internal draft, Final Technical or Project report, Manuscript, Manual, Software)
Project final report #1
Summary

The aim of this project was to develop and evaluate methods to improve the prediction of species in geographic space. For some applications such as climate change and invasive species, these methods must be used to make predictions outside the ranges of the data used to build the models, making validation even more important. New methods have arisen in statistics and machine learning to deal with these issues.

This study uses simulated data to compare the performance of modeling methods. It concludes that some of new machine learning methods, in particular, boosted regression trees and maximum entropy methods, outperform statistical and climate matching methods under a number of performance measures, for a wide range of data types, data qualities and prediction circumstances. More importantly, the paper provides advice on how a user might evaluate the performance of predictive models, to make best use of the alternatives available to them for specific purposes.

This paper provides details of how to access software and statistical code to implement several of the better behaved and most promising methods.

ACERA Use only	Received By:	Date:
ACERA Use only	ACERA / AMSI SAC Approval:	Date:
ACERA Use only	ACERA / AMSI SAC Approval:	Date:

Title: Do they / how do they / WHY do they differ? -- on finding reasons for differing performances of species distribution models.

Authors: Jane Elith[1] and Catherine Graham[2]

[1] School of Botany
The University of Melbourne
Parkville, Victoria. 3010
Australia
j.elith@unimelb.edu.au
(corresponding author)

[2]Department of Ecology and Evolution
650 Life Sciences Building
Stony Brook University
Stony Brook, NY 11794, USA

Introduction

Species distribution models (SDMs) are increasingly being used to address a diverse range of applied and theoretical questions (Graham *et al.* 2004, Guisan and Thuiller 2005). Their expanding use means that models are now being fitted to new forms of data, including occurrence records from museums or herbaria. For some applications, such as climate change or invasive species research, model predictions are extended beyond the geographic or environmental region from which training samples were drawn (e.g., Araújo *et al.* 2005). Newer applications also move beyond the traditional ecological focus of SDM into new fields such as evolutionary biology, where they are used to study topics such as speciation or hybrid zones (e.g., Kozak *et al.* 2008). As a result of these ongoing changes in the field, and spurred on by advances in data analysis within statistics and computing sciences (Breiman 2001a), new modelling methods continue to be implemented. Model complexity has generally increased over time from simple environmental matching (e.g. BIOCLIM, Busby 1991; DOMAIN, Carpenter *et al.* 1993) to fitting more complex non-linear relationships between species presence and the environment (e.g., generalised additive models, GAM, Hastie and Tibshirani 1990, Yee and Mitchell 1993; and maximum entropy modelling, Maxent, Phillips *et al.* 2006). Recent emphases on machine-learning and Bayesian methods indicate that new methods will continue to be developed (Prasad *et al.* 2006, Latimer *et al.* 2006).

This wide array of methods, data types and diverse research questions imply different requirements of modelling methods, creating the need for choice of method to be matched to application. It has spurred a growing literature aimed at evaluating different modelling methods in varied contexts. Numerous studies compare how well different methods perform (e.g., Elith *et al.* 2006, Seguardo and Araújo 2004, Moisen and Frescino 2002). Generally, in these types of studies a set of distributions are predicted for one or more species using a range of methods or different ways of applying one method, and then predictive performance is tested, usually against withheld or independently collected data (i.e., independent presence/absence data, additional field surveys, fossils etc.). This type of approach has also been used to explore the relationship between model performance and data characteristics such as: the number of species localities for modelling (eg.

Hernandez *et al.* 2006, McPherson and Jetz 2007a, Reese *et al.* 2005), species-level attributes (e.g., Guisan *et al.* 2007, McPherson and Jetz 2007b), and quality of locality data (e.g., Loiselle *et al.* 2008, Graham *et al.* 2008). While this extensive effort has yielded new insights and advice for modellers (a review beyond the scope of this paper), there is still considerable debate as to what methods are useful for which applications, as evidenced in diverse views and conflicting conclusions in recent papers comparing methods (see recent opinions in Ecography by Peterson *et al.* 2007, Phillips 2008).

Additional method comparison studies using the same basic approach of evaluating how well a series of methods can recover the geographic distribution of a species will likely provide diminishing returns because what method performs best, or is most appropriate, can vary by study system, data quality, the way the method is used, or the question addressed. This makes it hard in most cases to learn general principles about the models and to garner the right sort of knowledge to inform choice of method. The problem is that, unless there are large amounts of appropriate data available, it is difficult to separate the effects of the particular data from more general questions about the methods and their performance. In our opinion, much more progress would be made in understanding the keys links between model and outcome if model comparison studies focussed more on determining why a given method performs as it does.

This demands a wider array of evaluation criteria. There are several possibilities, and here we mention a few. First, one could approach the problem by learning how a method works. This is usually the domain of those who develop the methods, commonly statisticians and computer science professionals, but practitioners would certainly benefit from such understanding. Given knowledge of how methods work, one could decide *a priori* which method should be most appropriate given the question being addressed and the biology of the species. Alternatively, a researcher could use a large number of settings for a given method and conduct a very thorough evaluation of model performance using an comprehensive range of performance measures (for useful measures see, for example, Pearce and Ferrier 2000, and Caruana and Niculescu-Mizil 2006). This can shed light on the model's behaviour and limitations under various settings, but will not lead to understanding of why the differences in settings lead to differing performance in the absence of knowledge of how the method

works. However it is an important part of using a method well. Effects of varying settings on model performance are rarely reported in ecological applications of SDM (but see Phillips and Dudik 2008) yet the information is crucial to identifying the sensitivities and tendencies of methods.

Another approach is to use data with known characteristics to test model fit and prediction. These can be sets of data with known properties (e.g. drawn from a normal distribution with mean x and standard deviation y; see for example Moisen and Frescino 2002) or simulated (artificial) species (Austin *et al.* 2006). Simulated species are particularly useful if relevant features such as population processes, competition, or typical landscape properties can be included (e.g. Kearney *et al.* in press, Tyre *et al.* 2001). In examples of more static correlation-based approaches, researchers define the true distribution of the species or the relationship between the species and its environment (e.g. Meynard and Quinn 2007; Reineking and Schröder 2006). The features of a simulated species can be matched to the aspects of model performance that are important to the researcher.

The simulated data approach is particularly important for methods where the structure of the model itself is unclear. We envisage this will be an increasingly relevant issue as more machine learning methods are appropriated for ecological prediction tasks. This is because, unlike statistical models that have underlying and clear assumptions about the data and the processes that generated it, machine learning and other data mining methods approach the data-generating mechanism (nature) as unknown, but learn patterns from the observed data to make predictions (Breiman 2001a). These models are not always easy to interpret. In Breiman's words: " the models that best emulate nature in terms of predictive accuracy are also the most complex and inscrutable". Ecologists are unlikely to want inscrutable models; we want and sometimes need to understand what is behind the predictions. In the machine learning community there is some focus on how to interpret and visualise model output, and statisticians are reinterpreting some methods from a statistical viewpoint. This is particularly useful for ecologists (e.g. Cutler *et al.* 2007, Elith *et al.* 2008). Modelling simulated data provides another opportunity to explore how new methods behave.

In summary, we believe that deeper insights into the causes of varying model performance require an expansion of model evaluation approaches (Araújo and Guisan 2005). Comparative studies

on one or two regions and a few species do not tend to give insight into why methods differ, and without that knowledge it is difficult to make well informed selection of methods. Perhaps we should de-couple the method comparison from the interesting questions about species and biodiversity. Method comparisons need to focus on approaches that give insight to why methods differ. A range of tests is required to determine whether the model does what the application requires.

To demonstrate these ideas, here we use a simulated species approach to compare methods. We pose three applications: (1) understand the relationships between the species and its environment; (2) predict which parts of the landscape are more or less suitable for the species by creating a map of relative suitabilities; (3) extrapolate to environmental conditions outside those in the sample space. In doing so we aim to develop a straightforward example of the type of research that helps to move beyond standard model comparisons to those where the emphasis is on determining why methods differ. Our underlying premise is that it is useful to think more about the reasons for variation in model performance, and to match the evaluation criteria (whether a simulation, particular subsets of data, certain statistics, or map comparisons) to the questions that the models are intending to answer.

1. Simulating the species and sampling it.

We created a simulated species and mapped it onto a landscape. The species responds to three variables: *wetness*, aspect ("*southness*") and *geology*. It represents a southern hemisphere plant preferring wet and south-facing (shaded) sites and fertile substrates, with an interaction that results in *wetness* dominating the response (equation 1 and Figure 1, top panel). Details of how each index (SI.*wetness* etc) was specified is presented in the online supplement, Appendix S1.

Suitability of a cell = SI.*wetness* * 0.5*(SI.*southness* + SI.*geology*) - eq'n 1

where SI = suitability index (see Appendix)

The predictor variables (*wetness*, *southness* and *geology*) existed as real mapped data; the species' response to them was invented. The mapped distribution based on these relationships shows the suitability (scaled 0 to 1) of each grid cell for the plant (Figure 2, top left). We chose to use models that use presence-only or presence-absence data, so needed to make a presence-absence

realisation of the species. We used the suitability value in each grid cell of the map as the success rate for one sample of the binomial distribution, i.e., a cell with a suitability of 0.6 has a 60% chance of being occupied. This uses the "rbinom" function in R (R Development Core Team 2006) and is preferable to choosing a threshold. A threshold would alter the response of the species to one that has vertical cut-offs of suitability, not gradual changes as in our "truth". The idea of the binomial realisation is consistent with an interpretation of species relationships that says: "if the environment is only partially suitable for the species, in some cases the species will occur there and in some it will not". Our presence-absence realisation of the simulated species occupied 12% of the cells in the study region (Figure S2a, online supplement).

To provide the data set for modelling we sampled from the simulated presence-absence distribution by taking 1000 sites at random; these contained 115 presences and 885 known absences and will from here on be called the PA sample. This number of sites is somewhat arbitrary, but is enough to fit models well, and is typical of the amount of data often used in modelling. Other protocols such as stratified sampling across environmental gradients could be used instead of random sampling, but our aim here is only to demonstrate the insights from this type of model comparison, not to test sensitivities to data samples. The PA sample is visualised in Figure 3a; consistent with the full data set (Figure 3b), samples at high *wetness* are relatively rare. For methods only requiring presence data, the presence records in the PA sample were used as presence-only (PO) data. In the online supplement (Appendix S5) we describe additional data samples, in which 1000 or 3000 pseudo-absences were selected and combined with the PO data, to test the effect of using presences with pseudo-absences instead of true absence data.

2. Modelling and evaluation methods

Given that our intent is to demonstrate an approach, not to conduct an exhaustive study, we included only five algorithms. For the PA data we selected a generalised linear model (GLM; McCullagh and Nelder 1989) as a standard regression model and two more recent methods, boosted regression trees (BRT; Friedman *et al.* 2000) and random forests (RF; Breiman 2001b). For the PO data, we used Maxent and GARP as featured in recent comparisons (Peterson *et al.* 2007, Phillips

2008). Maxent, GARP and RF are machine learning (ML) methods, BRT is a ML method reinterpreted into a statistical paradigm, and GLM is a statistical method. More details and references are given in the online supplement. These methods were selected because they allow a number of contrasts, including comparison of models developed on PA compared with PO data, comparison of Maxent and GARP, and comparison of two methods using ensembles of trees, BRT and RF. Each of the methods are assumed capable for at least some of our three posed applications, given that they have all been used for related tasks.

For details of model building methods see Table 1 and Appendices S2, S3 and S4 in the online supplement. In those appendices the settings for some methods are tested in detail; this reflects our need to explore the effect of a range of settings for some methods, either because we were relatively inexperienced with that method or because the recommended settings did not produce good results.

Because one of our applications required output showing modelled relationships and not all of the methods provide visualization of fitted functions, we created a second set of environmental grids with an evaluation strip inserted, following Elith *et al.* (2005). The strip is a simple data arrangement that holds two of the three environmental variables at a constant value (here, a value achieving maximum suitability for the species) whilst varying the value of the third environmental variable across its numerical range. To visualise interactions (and to check that results do not depend on the value at which variables are held constant), pairs of variables can be covaried. For this, we included a comprehensive set of combinations of *wetness* and *southness* for each of the four classes of *geology*. Predictions were made to the evaluation strip in its tabled form, for those methods that could predict to that, or by making a "projection" (Maxent and GARP) to the second set of grids. This process allowed us to determine the partial response to one or two variables while holding other variables constant. It is analogous to the partial plots from standard regression methods (see Elith *et al.* 2005 for details). To determine how the models extrapolate we included values for variables in the evaluation strip that were outside the variable range in the study area for both *wetness* and *southness*. This is only one measure of ability to extrapolate, and other possibilities exist - for example, creating landscapes

or evaluation data with new combinations of variables. The simple test of extrapolation outside limits will suffice here.

Finally, to evaluate ability to predict habitat, mapped predictions were visually and quantitatively assessed. Summary statistics were calculated across predictions in all 80000 cells, in comparison to the presence-absence realisation and the true suitabilities. Analyses of predicted probabilities in relation to the true, sampled presence-absence data were: (i) the area under the receiver operating characteristic curve (AUC; Hanley and McNeil 1982); (ii) the remaining per-observation deviance (i.e. the variation left unexplained, as measured by the mean binomial deviance across sites; Elith and Leathwick 2007); (iii) the point biserial correlation coefficient (a Pearson correlation, Elith *et al.* 2006, COR.pa); (iv) elements of the confusion matrix for predictions converted to binary values, using a threshold that gave prevalence as close to possible to truth but that gave the same prevalence from each method. Predictions were compared with true suitabilities with a Pearson correlation coefficient (COR.si).

The important point about using more than one measure is that they quantify different aspects of predictive performance (see Murphy and Winkler 1987 and Pearce and Ferrier 2001 for interesting discussions of this topic). Amongst the measures tested against PA observations, AUC measures the ability of predictions to discriminate between observed presence and absence, regardless of the absolute value of the predictions. COR.pa also measures discrimination, but includes consideration of the actual value of the prediction, and how it compares to the observation. Deviance puts more emphasis on the model calibration i.e. to whether predictions reliably predict frequencies of occurrence. In other words, for deviance the actual value of the prediction is important. These different metrics measure different attributes of the predictions and are a useful and complementary set (Murphy and Winkler 1987).

3. Results

Relationship of species to environment: Relationships were tested by predicting to the evaluation strip. Figure 1 shows the modelled responses to the 3 variables when others are held at their optima, and in the right column, the response to co-varying *wetness* and *southness* for *geology* class one. Here

we analyse the results for the ranges of the variables in the data – i.e., within the blue vertical lines of Figure 1; in the later section "Extrapolation" we deal with predictions outside the range of the data.

We would not expect any of the methods to fully retrieve the species environment relationship of the simulated species because a relatively small sample of a binary realisation of the data was used to build the models (Figure 3a). Nonetheless, several of the methods were able to fit reasonably accurate functions. This was an easier task for the methods using smoothed functions (GLM and Maxent), given that the true relationships were smooth. All methods except GARP modelled *geology* correctly (Figure 1, left column). GARP's modelled response to *geology* varied across different data samples, different runs, and different summaries of the data; see online Supplement Appendix S2). GARP might be expected to model categorical data properly because it uses logistic rules, but the implementation is not a true logistic regression and does not properly deal with unordered categorical variables (Elith 2002 and Peterson pers.comm.). However, the atomic rules provide some opportunity to model *geology* correctly, and in some cases (online appendix Fig. S3 and S6) the result was better than the one presented.

The low suitability of dry areas (*wetness* < 15) was correctly captured by all methods (Figure 1, second and fourth columns). At higher levels of *wetness*, where the data are more sparse, BRT modelled the response most accurately, followed by Maxent (slightly reduced amplitude), RF (smaller amplitude) and GLM (unnecessary complexity around *wetness* of 50) and last, GARP (small amplitude and wrongly predicted that high *wetness* was unsuitable).

Southness (Figure 1, third and fourth columns) was difficult to model well, probably partly because the response included an interaction between *southness* and *wetness*, and also because it was less dominant than *wetness* in the suitability equation. RF, BRT and Maxent did best, GLM was good but gave an increasing response at low values of *southness* and, without an interaction, could not capture the response to *southness* at high *wetness* (see right side of 3D plot). GARP completely failed to model the true response and this result was consistent across all tests (Appendix S2).

Mapped predictions; visual assessments: Differences between the methods were also evident based on the mapped predictions. We present the results as maps and plots so that any arbitrariness

introduced by choice of legend in the map (Figures 2 and 4) can be checked against the plotted results (Figure 5). Note that the COR.si in Table 2 acts as a summary measure of the plotted data in Figure 5.

Because all methods predicted the response to low *wetness* correctly, they all correctly predicted absence at dry sites (the white areas on the maps, Figure 2). Modelling this part of the response correctly meant that all methods produced a broadly correct mapped pattern. GARP tended to predict high values across any areas that had at least some suitability for the species, whereas the other methods predicted gradations in suitability more accurately (Fig. 2 and, a closeup in Fig. 4). Given that the true niche of the species is known and consisted of varying suitability depending on the combination of environmental parameters, failure to predict gradations is an error. The overprediction can be traced to the errors in retrieving the true underlying species-environment relationships. Simulated data like these could be used to explore whether another example or different settings for GARP improve model performance. For these data and other realisations of it, we spent some time attempting this but could not improve performance substantially beyond that shown (Appendix S2). The results of the different trials shown in the Appendix are interesting in that they suggest that means of all runs of GARP are slightly better for these data than the subsets that are generally advised. They also demonstrate considerable run-to-run variation (where one run is 500 models or summaries thereof).

As expected, none of the methods perfectly retrieved the true mapped suitabilities – sample size, use of binary data rather than suitabilities, and algorithmic limitations all contribute to this result. Maxent recreated the general mapped pattern of the simulated species well and only failed, and only moderately so, with respect to calibration (Figures 2 and 4). Perfect calibration would have resulted in all records in Figure 5 sitting on the diagonal, but a presence-only method cannot be well calibrated unless information on the species prevalence in the region is available. In contrast, the three methods trained on the PA data should be properly calibrated. The results are all reasonable, with BRT predictions slightly better constrained than RF and GLM (Figures 2, 4 and 5). We tested various settings for RF (seven combinations are presented in Appendix S4); the ones we present here are

those that would have been selected from the out-of-bag error estimates, and are consistent with those used in other published studies (Appendix S4).

Mapped predictions; quantitative analyses: The summary statistics presented in Table 2 focus on predictions as continuous values, and these demonstrate some differences between the methods, with the extent of the difference varying with the statistic. Note that the first row in Table 2 ("Truth") indicates the best possible performance for AUC, Deviance and COR.pa, because it measures the relationship between the true presence-absence realisation (sampled for modelling) and the true suitability. The AUC indicated that all models discriminated the broad patterns of presence and absence reasonably well, but with some variation between methods evident. The broad success in discriminating between presence and absence locations probably results from the correct modelling of dry conditions, because all the pairwise comparisons between predictions in these areas and ones in the wetter areas would have been correctly ranked. Measures that include consideration of the actual value of the predictions (all others) emphasise more clearly that Maxent, BRT and GLM all do well, followed by RF then GARP (Table 2). AUC and elements of the confusion matrix (Table 3) are the only truly fair basis here for comparing all methods, because the 3 models fitted to the PA data had more information than the PO methods, and this information allows the models to estimate probabilities and hence to be better calibrated. Nevertheless, for this exercise Maxent does well without this information. The fact that the order of model performance is consistent across metrics shows that errors in prediction are related to both discrimination and calibration. Given that the general shapes of the fitted functions for RF are reasonable, it is likely that its reduced performance in these summary metrics results from the slightly noisier fit compared with BRT (Figure 1, right column) and the slightly poorer calibration.

It is possible that using predictions as continuous values unfairly discriminates against GARP, because GARP works by producing a presence-absence prediction, and non-binary predictions result from summaries across multiple runs. Because of this, the predictions for all methods were thresholded and a test of binary predictions applied. We did this without biasing the results towards methods with good calibration, by using thresholds that gave the same prevalence across the

landscape for all methods. This meant that we used a threshold of 1 for GARP (to get prevalence as close to truth as possible and to restrict the overprediction of GARP), then set others in relation to that. The results (Table 3) are consistent with other results, though the differences are less extreme. If we had set separate thresholds to give the best binary realisation for each method (using the known prevalence), the result for GARP would remain constant and the others would all improve.

Information on BRT and GLM models fitted to the presence - pseudo-absence data are presented in the online supplement, Appendix S5. These show that in some (but not all) cases the fitted functions are reasonably accurate, and that there are differential effects on the discrimination and calibration of the models. We note that alternative implementations of BRT that are better suited to the presence / pseudo-absence data structure will be available soon (Ward *et al.* in press).

***Extrapolation**:* The extrapolation behaviours of the models is presented in Figure 1, in which the responses to the highest and lowest values for the *wetness* and *southness*, outside the blue dashed lines, are where the models are extrapolating to unsampled conditions outside the range of these variables in the mapped region. The nonsensical negative values of *wetness* and *southness* don't matter here – the important question is how the models would extrapolate beyond the sampled values of data. We had no prior knowledge of GARP behaviour, and in this example extrapolation was either a stepped decline (at extremes of *southness*) or a constant value. The selected rules in the best-subsets models would explain this behaviour, but they are not accessible in the desktop program so could not be analysed. Extrapolation patterns varied in GARP according to the dataset, the particular run, and the selected subsets (online Appendix Figure S3). By contrast, Maxent acts consistently and is "clamped" so it extrapolates in a horizontal line from the fit at the most extreme environmental value in the training data, both presence and background. As expected from the way polynomials behave, a GLM fitted with cubic and quadratic functions extrapolates by continuing the fitted trend beyond the last observation, sometimes with unwanted results (Austin *et al.* 1990). For example, here the projected increase in suitability at the lower values of *southness* is not sensible (north-facing (low *southness*) sites should be least suitable for the species). Classification and regression trees always extrapolate at a constant value from the last "known" site, as seen for BRT and RF.

4. Reflections on the simulation

Before considering the wider implications of this simulation, we want to first emphasise what it does not do. Whilst our simulated species had some ecological realism (it is affected by more than one variable, and reacts to predictors in non-linear and non-additive way), we were not trying to emulate nature, but to interrogate the behaviour of different algorithms. We also only tested some of the many ways of fitting these methods - for example, a GLM can include interactions and splines can be used to control behaviours at the extremes of the sample ranges. We also postulated our application as mapping suitabilities, whereas mapping binary presence-absence predictions are an alternative output. Finally, we chose scenarios that give some insight but did not attempt a comprehensive study.

What then does this simulation demonstrate? First, knowing what an algorithm is doing can give insights into various features that are apparent in its predictions; it helps to answer why particular behaviour is observed. Testing an algorithm's performance in this way reveals problems and may help developers and modellers to understand and improve model performance. The systematic studies in Appendices S2, S3 and S5 demonstrate that looking carefully at model performance across settings helps to expose what is happening.

Second, our example clearly illustrates the value of evaluating models from several viewpoints. The summary statistics indicated which algorithm gave the best mapped predictions, and gave some hints about why through the different metrics. AUC only measures rank so did not reveal the more extreme differences between the models that were more related to calibration of the model. We do not believe that this means that any of these statistics are misleading (see the title of Lobo *et al.* 2008), but simply that different statistics measure different aspects of performance, and that appropriate statistics relevant to the application of the model need to be selected. Being able to visualise fitted functions not only satisfied our application of exploring modelled relationships, but also allowed us to understand what caused differences among the methods and how different fitted functions influenced mapped predictions.

Third, comparing the results for the partial responses with the quantitative assessments gives some useful insights. For example, even though the GLM modelled unncessary complexity in the

wetness response (giving the wave-like forms in the 3D plots), the evaluation statistics implied it did nearly as well as BRT. This is in spite of the fact that BRT had a better controlled fit overall. The distribution of environments (Figure 3b) gives the key: there are relatively few sites in the places where the GLM has failed to model the true response. Similarly, the environmental distribution of the training data (Figure 3a) explains why several methods (GLM, BRT and RF) tended to model a declining response (even if only a small trough) to *wetness* at values early on the plateau of the true response - the samples are sparse in some parts of this environment, and several have been realised as "absence" in this particular sample. These demonstrate that understanding the environmental distribution of the data in the region of interest, in the sample, and in regions that might be used for projection, is a critical part of understanding the implications for modelling and prediction.

Finally, the demonstration prompts a range of questions about what characteristics we want from models in certain situations. For example, when using species distribution models to predict how species ranges will shift with climate change or how they will extrapolate to new regions it is critical that we understand how the algorithm performs when projected into new environmental combinations combinations not sampled by the training data. In other words, is the way that the algorithm extrapolates appropriate from an ecological perspective? Different behaviours are apparent in the methods that we tested, but the choice as to which (if any) is correct is as much an ecological and/or a physiological question, as a statistical one. For a full investigation of extrapolation or forecasting behaviour a much larger range of tests is required including prediction to new combinations of environments, and the species should to be simulated with known responses to these. The important point is that we need to first recognise what different modelling applications require of SDMs and then research the best means for achieving what they require. Understanding how the models work and devising evaluation criteria that are closely matched to the questions being asked can inform decisions about the best modelling approach.

Conclusion

Given the multitude of applications of SDMs, many of them related to pressing conservation issues, we hope this forum paper stimulates research on how to effectively continue to develop and

test these methods. Additional model comparison studies may not be fruitful unless they start to ask why certain methods perform better than others. This is particularly important for applications that do not satisfy the underlying assumptions of species equilibrium with environment (Dormann 2007), such as range shifts with climate, or species invasions into a new area. By providing an example that we believe gives useful insight into model performance, we hope that both developers and users will increasingly question what their models are doing and whether that is most appropriate for the intended applications and outcomes.

Acknowledgements

Many thanks to Mark Burgman, John Leathwick, Yung En Chee, Stephen Baines, Michael Kearney, Steven Phillips and Town Peterson for comments on various drafts of the manuscript, and to Mike Austin, Simon Ferrier and Simon Barry for asking challenging questions that were the starting points for several of the ideas. The reviewers and editors comments gave important direction and we appreciate their efforts. Jane Elith was funded by ARC grant DP0772671 and the Australian Centre of Excellence for Risk Analysis. The colours used in the figures are from a palette developed for colour-blind people; thanks to its author:http://jfly.iam.u-tokyo.ac.jp/html/color_blind/#stain

References

Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. - J. Biogeogr. 33: 1677-1688.

Araújo, M. B. and Rahbek, C. 2006. How does climate change affect biodiversity? Science 313:1396 - 1397.

Araújo, M. B. *et al.* 2005. Validation of species-climate impact models under climate change. - Global Change Biology 11: 1504-1513.

Austin, M. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. - Ecol. Model. 200: 1-19.

Austin, M. P. *et al.* 2006. Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. - Ecol. Model. 199: 197-216.

Breiman, L. 2001a. Statistical modeling: the two cultures. - Statistical Science 16: 199-215.

405 Breiman, L. 2001b. Random Forests Technical Report.
406 http://oz.berkeley.edu/users/breiman/randomforest2001.pdf

407 Busby, J. R. 1991. BIOCLIM - a bioclimate analysis and prediction system. - In: Margules, C. R. and
408 Austin, M. P. (eds.), Nature Conservation: Cost Effective Biological Surveys and Data Analysis.
409 CSIRO, pp. 64-68.

410 Carpenter, G., Gillison, A. N. and Winter, J. 1993. DOMAIN: a flexible modelling procedure for
411 mapping potential distributions of plants and animals. - Biodivers. Conserv. 2: 667-680.

412 Caruana, R. and Niculescu-Mizil, A. 2006. An Empirical Comparison of Supervised Learning
413 Algorithms. - In, Proceedings of the 23 rd International Conference on Machine Learning,
414 Pittsburgh, PA.

415 doi:10.1111/j.1365-2699.2007.01779.x. - J. Biogeogr. 35: 105-116.

416 Dormann, C. F. 2007. Promising the future? Global change projections of species distributions. -
417 Basic and Applied Ecology 8: 387-397.

418 Elith, J. 2002. Predicting the Distribution of Plants. PhD thesis. School of Botany. The University of
419 Melbourne.

420 Elith, J. *et al.* 2005. The evaluation strip: a new and robust method for plotting predicted responses
421 from species distribution models. - Ecol. Model. 186: 280-289.

422 Elith, J. *et al.* 2006. Novel methods improve prediction of species' distributions from occurrence data.
423 - Ecography 29: 129-151.

424 Friedman, J. H., Hastie, T. and Tibshirani, R. 2000. Additive logistic regression: a statistical view of
425 boosting. - Ann. Statist. 28: 337-407.

426 Graham, C. H. *et al.* 2004. New developments in museum-based informatics and applications in
427 biodiversity analysis. - Trends in Ecology & Evolution 19: 497-503.

428 Graham, C. H. *et al.* 2008. The influence of spatial errors in species occurrence data used in
429 distribution models. - J. Appl. Ecol. 45: 239-247.

430 Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat
431 models. - Ecology Letters 8: 993-1009.

432 Hanley, J. A. and McNeil, B. J. 1982. The meaning and use of the area under a Receiver Operating
433 Characteristic (ROC) curve. - Radiology 143: 29-36.

434 Hastie, T. and Tibshirani, R. 1990. Generalized Additive Models. - Chapman and Hall.

435 Hernandez, P. A. *et al.* 2006. The effect of sample size and species characteristics on performance of
436 different species distribution modeling methods. - Ecography 29: 773-785.

437 Kearney, M. *et al.* in press. Modelling species distributions without using species distributions: the
438 cane toad in Australia under current and future climates. - Ecography.

439 Kozak, K., Graham, C.H. and Wiens, J.J. In press. Species distribution modeling in evolutionary
440 biology. TREE.

441 Latimer, A. M. *et al.* 2006. Building statistical models to analyze species distributions. - Ecol. Appl.
442 16: 33-50.

443 Latimer, A. M. *et al.* 2006. Building statistical models to analyze species distributions. - Ecol. Appl.
444 16: 33-50.

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

447 Loiselle, B. A. *et al.* 2008. Predicting species distributions from herbarium collections: does climate
448 bias in collection sampling influence model outcomes?

449 McCullagh, P. and Nelder, J. A. 1989. Generalized Linear Models. - Chapman and Hall.

450 McPherson, J. M. and Jetz, W. 2007a. Type and spatial structure of distribution data and the perceived
451 determinants of geographical gradients in ecology: the species richness of African birds. - Global
452 Ecology and Biogeography 16: 657-667.

453 McPherson, J. M. and Jetz, W. 2007b. Effects of species' ecology on the accuracy of distribution
454 models. - Ecography 30: 135-151.

455 Meynard, C. N. and Quinn, J. F. 2007. Predicting species distributions: a critical comparison of the
456 most common statistical models using artificial species. - J. Biogeogr. 34: 1455-1469.

457 Moisen, G. G. and Frescino, T. S. 2002. Comparing five modeling techniques for predicting forest
458 characteristics. - Ecol. Model. 157: 209-225.

459 Murphy, A. H. and Winkler, R. L. 1987. A general framework for forecast verification. - Monthly
460 Weather Review 115: 1330-1338.

461 Orr, D. W. 2000. Ideasclerosis: Part One. - Conserv. Biol. 14: 926-928.

462 Pearce, J. and Ferrier, S. 2000. An evaluation of alternative algorithms for fitting species distribution
463 models using logistic regression. - Ecol. Model. 128: 127-147.

464 Peterson, A. T., Papes, M. and Eaton, M. 2007. Transferability and model evaluation in ecological
465 niche modeling: a comparison of GARP and Maxent. - Ecography 30: 550-560

466 Phillips, S. 2008. Response to "Transferability and model evaluation in ecological niche modelling". -
467 Ecography 31: 272-278.

468 Phillips, S. in press. Response to "Transferability and model evaluation in ecological niche
469 modelling". - Ecography.

470 Phillips, S. J. and Dudik, M. 2008. Modeling of species distributions with Maxent: new extensions
471 and a comprehensive evaluation. - Ecography 31: 161-175.

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

474 R Development Core Team 2006. R: A Language and Environment for Statistical Computing. - In, R
475 Foundation for Statistical Computing.

476 Redford, K.H. and Taber, A. 2000. Writing the wrongs: developing a safe-fail culture in conservation.
477 - Conserv. Biol. 14: 1567-1568.

478 Reese, G. C. *et al.* 2005. Factors affecting species distribution predictions: A simulation modeling
479 experiment. - Ecological Applications 15: 554-564.

480 Reineking, B. and Schröder, B. 2006. Constrain to perform: regularization of habitat models. - Ecol.
481 Model. 193: 675–690.

482 Segurado, P. and Araujo, M. B. 2004. An evaluation of methods for modelling species distributions. -
483 J. Biogeogr. 31: 1555-1568.

484 Tyre, A. J., Possingham, H. P. and Lindenmayer, D. B. 2001. Matching observed pattern with
485 ecological process: can territory occupancy provide information about life history parameters? -
486 Ecol. Appl. 11: 1722-1738.

487 Ward, G. *et al.* in press. Presence-only data and the EM algorithm. - Biometrics.

488 Wu, J. and Hobbs, R. 2002. Key issues and research priorities in landscape ecology: An idiosyncratic
489 synthesis. - Landscape Ecol. 17: 355-365.

490 Yee, T. W. and Mitchell, N. D. 1991. Generalized additive models in plant ecology. - J. Veg. Sci. 2:
491 587-602.

492

493

494 Table 1 – Details of model fitting procedures and settings

Method	Name for model	Data	Settings and notes on further tests
Genetic Algorithm for Ruleset Prediction	GARP	PO	Used v 1.1.6. Details in online supplement Appendix S2 on tests to compare effects of different settings. Results presented here from: species data = 115 PO samples plus pseudo-absences selected by GARP. 50% data used for training, 50% for extrinsic evaluation. Created 500 models each with a convergence limit of 0.01 and 1000 maximum iterations. Allowed all rule types. From the 500models chose 20% with mid extrinsic omission error and from those 20 with mid commission error. Final prediction is mean of these. For predicting to evaluation strip projected to grids with strip inserted.
Maximum entropy	Maxent	PO	Used version 3.2.1 from the command line. Modelled the 115 PO samples and allowed Maxent to select a random 10000 background samples (the default). All other settings were the defaults except: flagging *geology* as a categorical variable, providing a separate set of grids to project to that contained the evaluation strip, and using the "-d" flag (see help file for Maxent). The -d flag forces Maxent to calculate the probability distribution over the background samples alone (rather than the default, which calculates it over the joint background and presence data), and providing it with the best chance to be well calibrated. For predicting to evaluation strip projected to grids with strip inserted.
Generalised linear models	GLM.pa	PA	Used R[1] and function *glm.* Created all possible subsets of models with the options for each variable being: exclude, or (if continuous): linear, quadratic or cubic fits. Used AIC to select the best model.
Boosted regression trees	BRT.pa	PA	Used R[1] and function *gbm* with custom scripts of Elith *et al,* 2008 to build an ensemble of regression trees. Selected tree complexity of 3, learning rate of 0.001, using prevalence-stratified cross-validation to determine optimal number of trees (4250). See Appendix S4 for details.
Random forests	RF.pa	PA	Used R[1] and function *randomForest* to build an ensemble of classification trees. Tests of a range of settings are presented in Appendix S3. Model presented here had 500trees with one variable randomly selected from the 3 candidates at each split. No class weights.

495

496
497

Table 2: Comparison of model results with truth, as realised by the presence-absence map (columns 2, 3 and 4) and the suitability values (column 5). For all statistics except deviance, higher is better.

Model	**AUC**	**Remaining deviance**	**COR.pa**	**COR.si**
Truth (suitabilities)	0.872	0.514	0.508	1.000
GARP	0.822	3.391	0.401	0.793
Maxent	0.861	0.612	0.467	0.922
GLM.pa	0.863	0.546	0.480	0.941
BRT.pa	0.862	0.537	0.485	0.954
RF.pa	0.834	0.736	0.448	0.875

Table 2: Method comparisons when predictions are reduced to binary results. Because GARp overpredicted, the highest possible GARP threshold (1) was used to convert GARP predictions to binary form, then thresholds were selected for all other methods that gave identical prevalence in the landscape (17.3%, compared with truth 11.9%). The values in the tables are the proportions of predictions that fell into each category (true negative etc) when cross-tabulated with truth.

Prediction:	true negative	true positive	false negative	false positive
GARP	0.771	0.064	0.056	0.109
Maxent	0.778	0.071	0.049	0.102
GLM	0.780	0.073	0.047	0.100
BRT	0.780	0.073	0.047	0.100
RF	0.777	0.070	0.050	0.103

Figure legends:

Figure 1: Partial responses to the 3 variables (left) and over co-varying wetness and southness (3D plots, right). The true responses (top panel) were generated by using the equations that define the simulated species to predict to the evaluation strip, then plotting the results (see Elith *et al.*2005 for details). The blue vertical lines show the extent of the variable values in the mapped region; outside these the models are extrapolating. The range on the *wetness* and *southness* axes of the 3D plots is that within the blue lines of the 2D ones, and predictions range from 0 to 1.

Figure 2: Mapped distributions of the virtual species (top left) and predictions of relative suitabilities from the methods detailed in the text, Legend: white < 0.1, cream 0.1 to 0.5, blue-lightblue-green-orange-vermillon at steps of 0.1 from blue (0.5 to 0.6) to vermillon (0.9 to 1) (note to editor - this legend could be included):

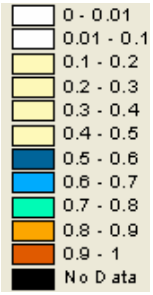
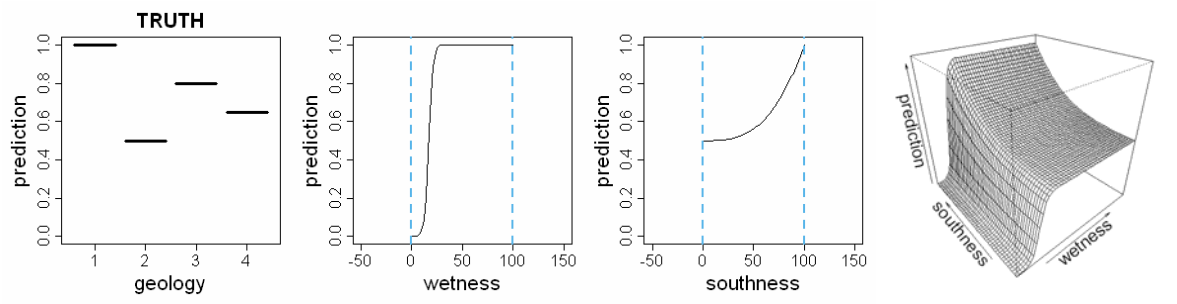


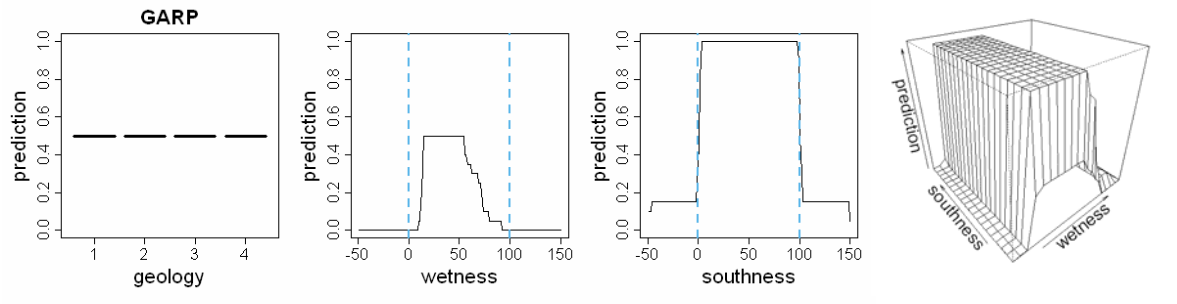
Figure 3: (a): The data samples for presence-absence models. Samples are shown at their original true suitability value (vertical axis), but were converted to presence (blue) or absence (orange) as described in the text. **(b):** The location of all 80000 grid cells in environmental space. The pale yellow mesh shows the full suitability surface from the simulated species, for geology class 1. The points of varying colours show sites in the four geology classes. Note the few sites with high wetness values.

Figure 4: Close-up of predictions from Figure 2. Choice of location was via random number selection for centre grid position. Predictions in greyscale, from white (zero) to black (one); fine grid lines are in the same position on each map.

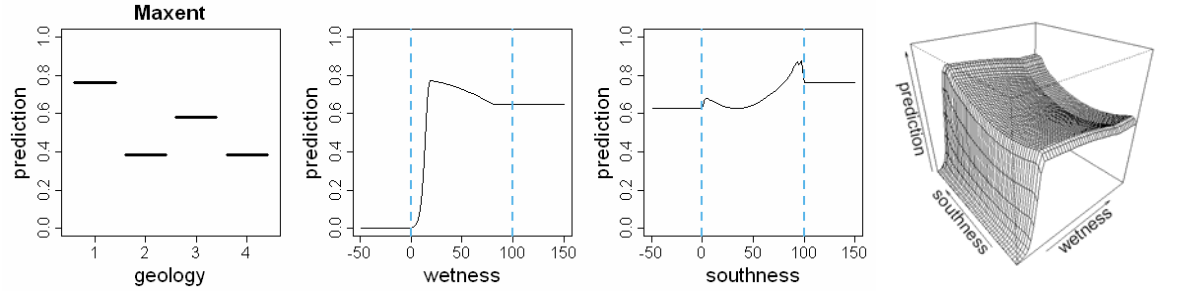
Figure 5: Predictions (y axis) versus the true suitability for all 80000 grid cells in the maps in Figure 2, covering five modelling methods described in Table 1. The blue diagonal line shows the 1:1 relationship.



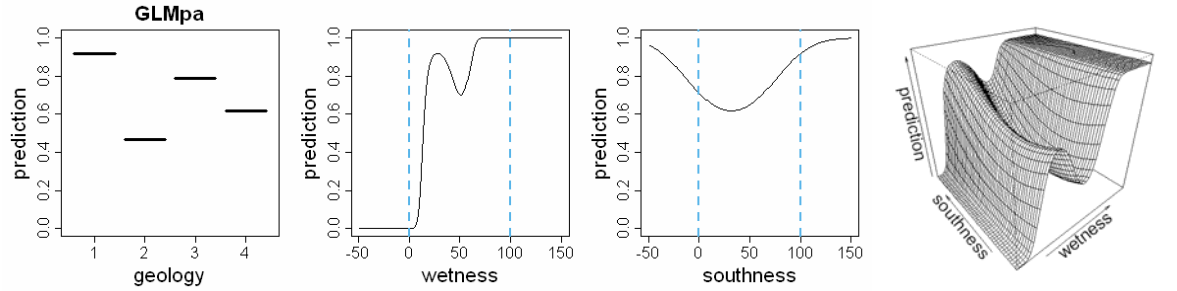
535



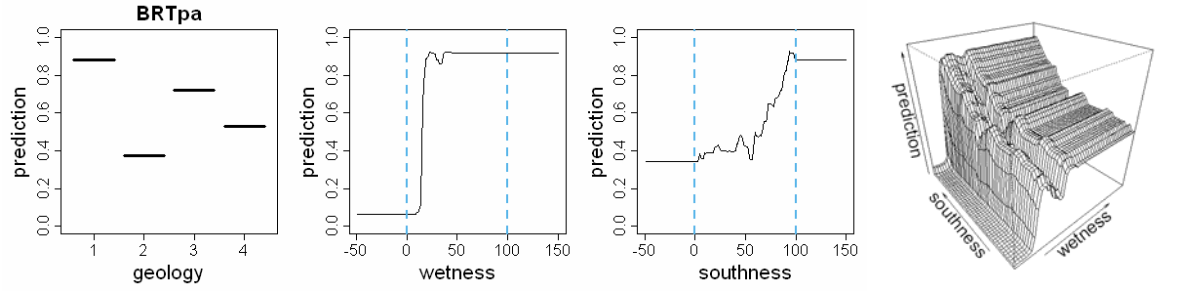
536



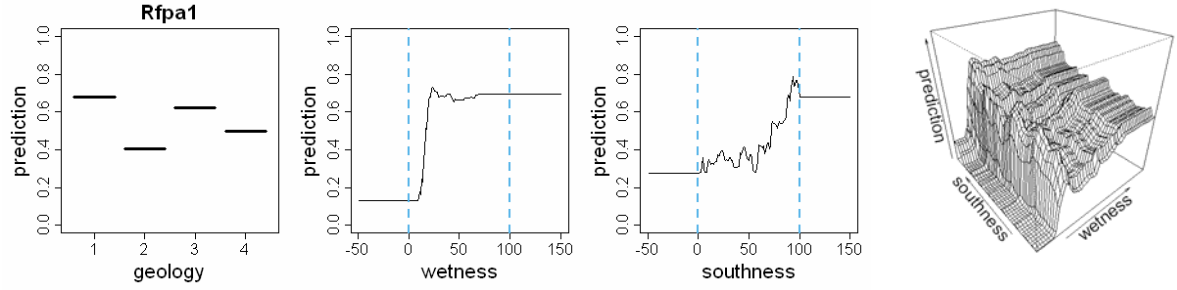
538



540



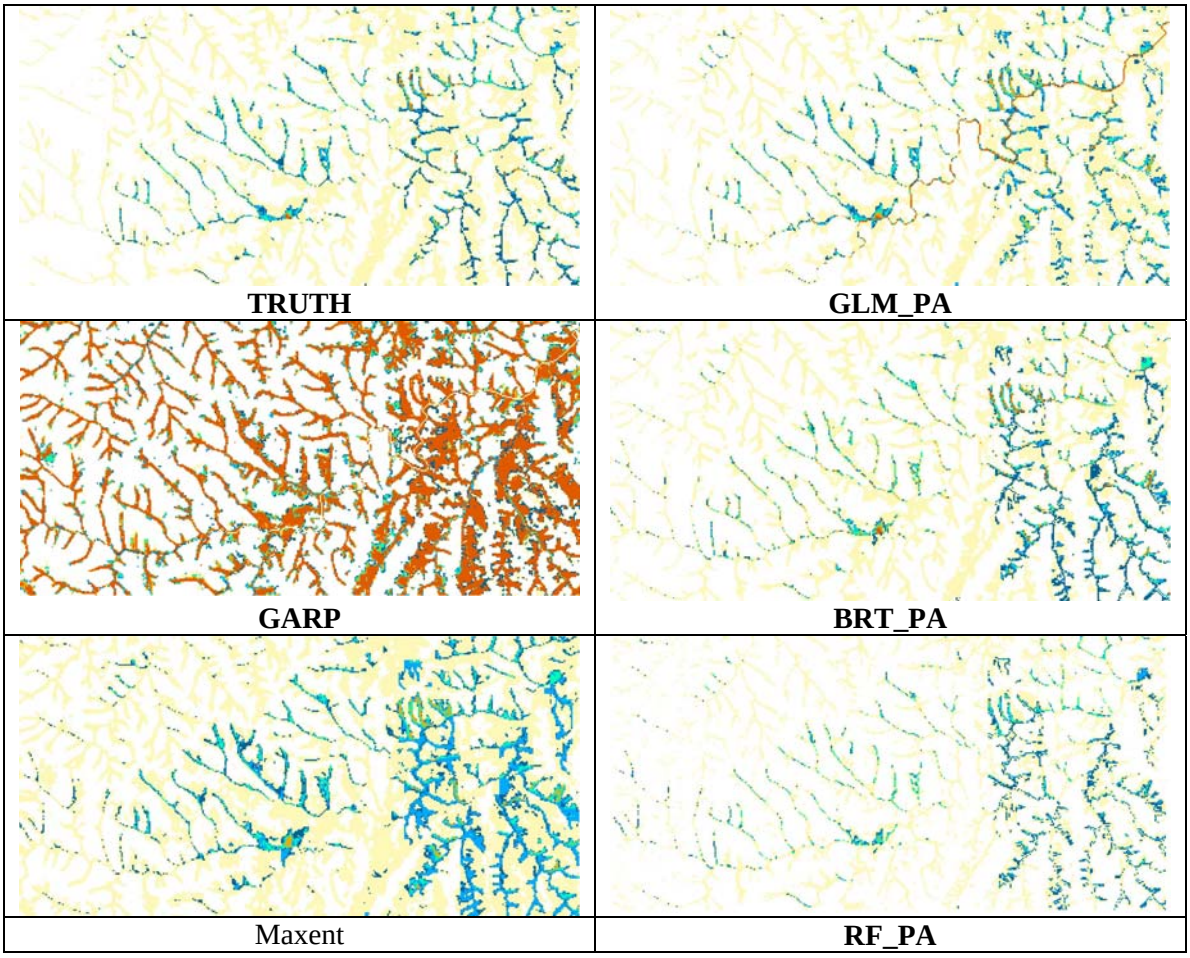
542



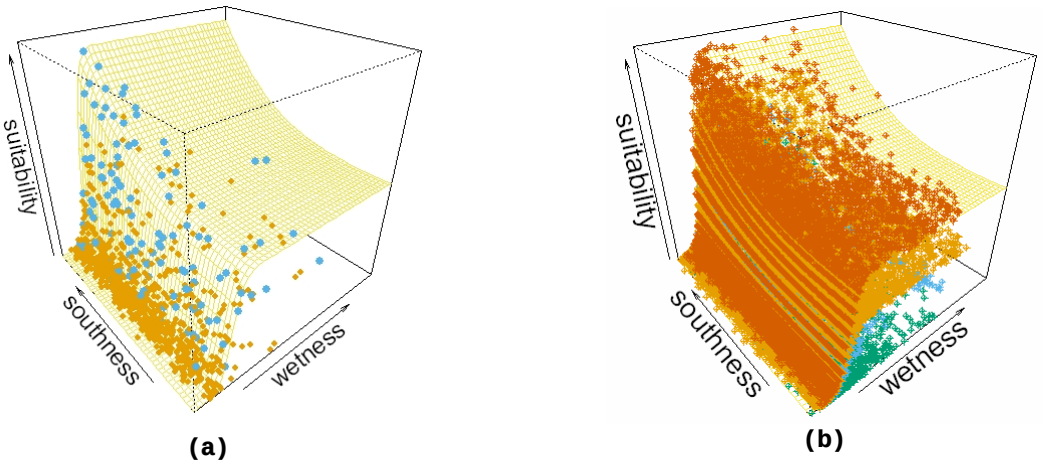
543

544 **Figure 1**

545



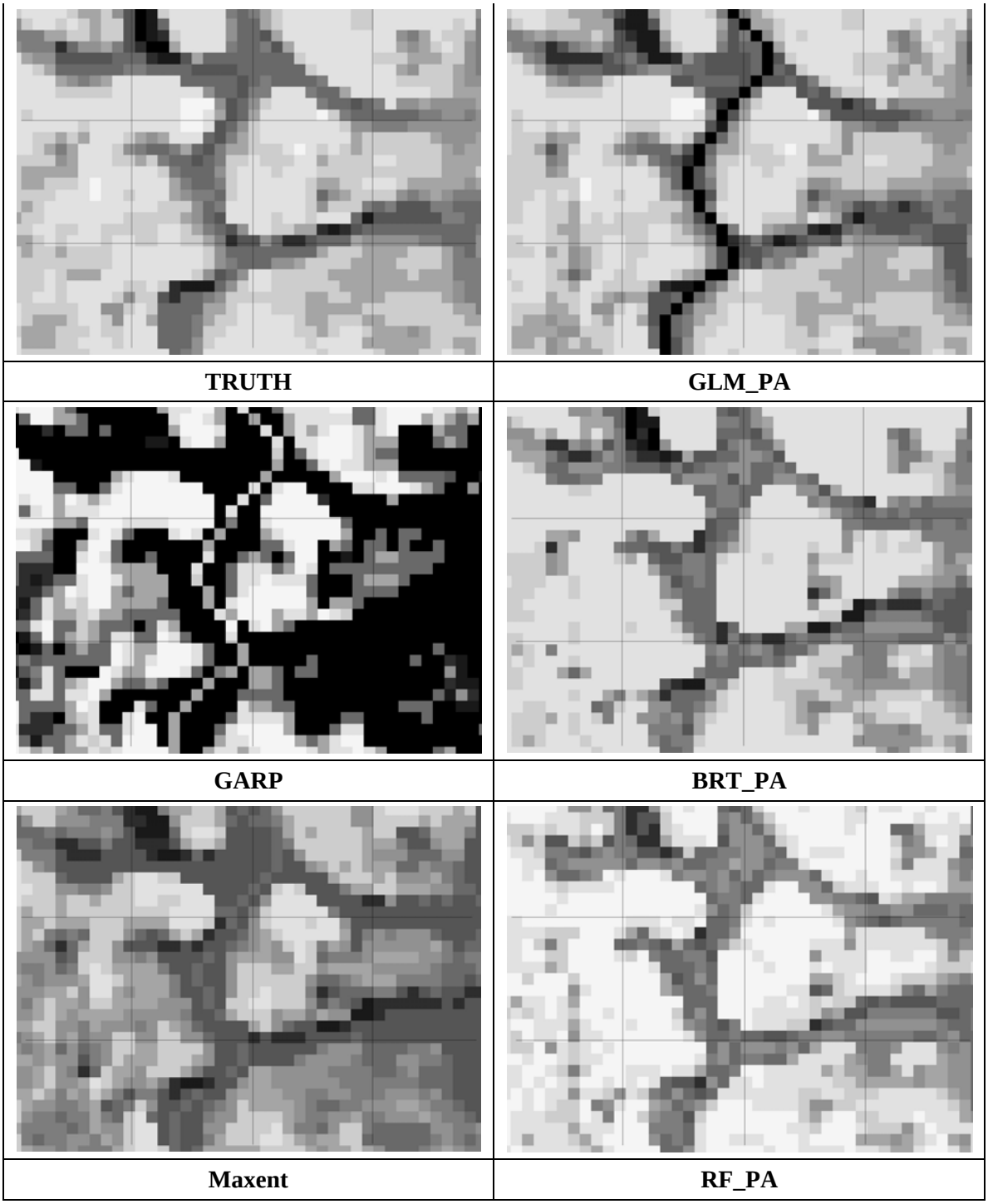
546
547 **Figure 3**
548
549



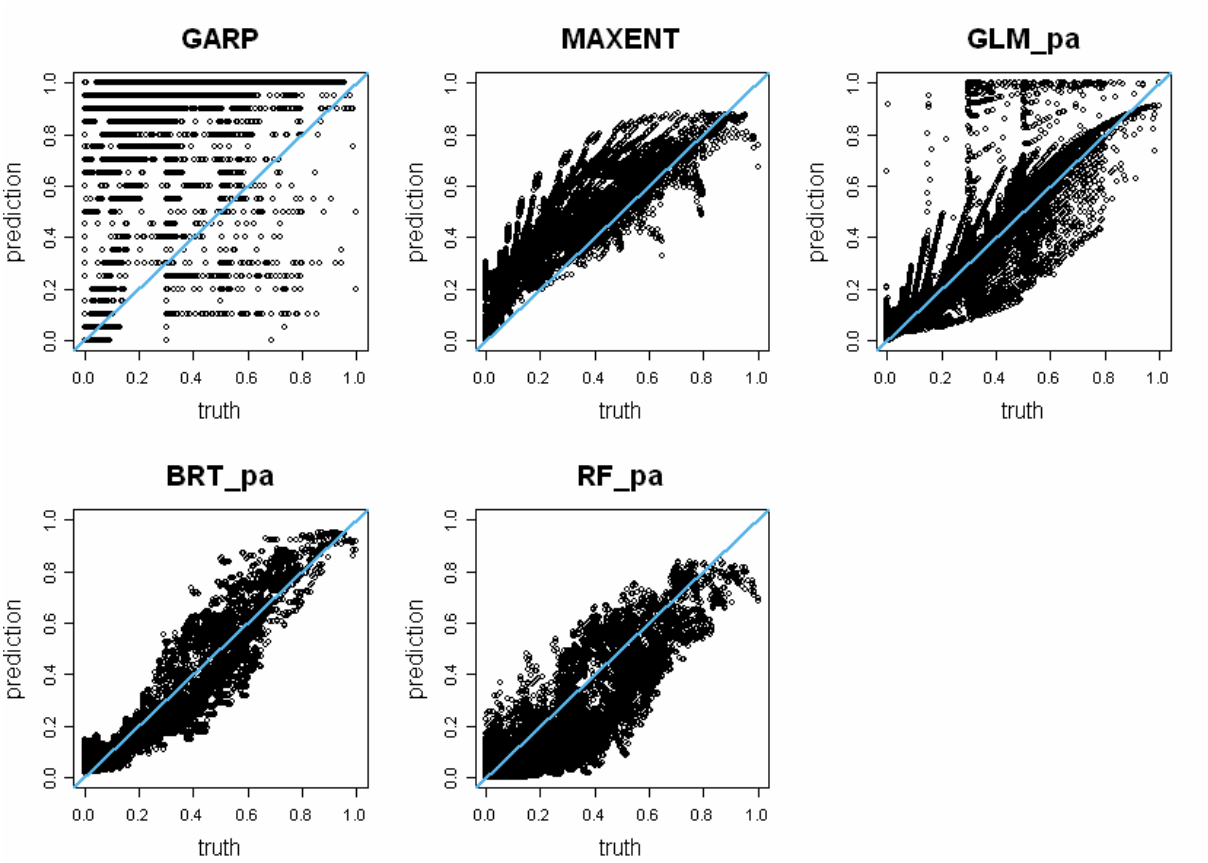
550
551 **Figure 3**

552

553



554
555 **Figure 4**
556



557
558

559 **Figure 5**

560

Online supplementary material: **Appendix S1: Details of the data generation**

The species, described in Elith and Graham (2008: equation 1 and Figure 1), responds to three environmental gradients:

Suitability (SI) = SI.wetness * 0.5*(SI.southness + SI.geology) - equation S1
where SI = suitability index

SI.wetness is the individual response to wetness, varying non-parametrically between 0 and 1 (Figure S1a), and SI.southness is the response to how south-facing a site is, varying parametrically between 0 and 1 (Figure S1b): SI.southness = 0.000001*(southness^3). The response to geology (SI.geology) is simply set at four levels: response to class 1= 1, class2 = 0, class3 = 0.6, class4 = 0.3.

The overall suitability is not a simple addition of these terms but involves an interaction between wetness and the sum of the responses to southness and geology. This implies that southness and geology substitute for one another but wetness overrides both.

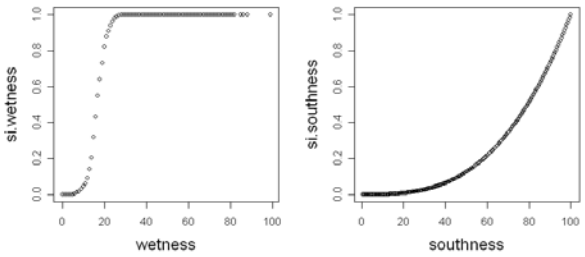


Figure S1. Individual suitability indices for (a) wetness and (b) southness

Using mapped grids of wetness, southness and geology (400 columns by 200 rows) from a real region in south-east Australia, we created the suitability indices for each, and a composite SI for the simulated species, from equation S1. The 80000 SI values were mapped, and also converted to binary values (using the function *rbinom* in R; R Development Core Team 2006; Figure S2a), which were then sampled (Elith and Graham 2008). In addition to the presence-absence (PA) and presence-only (PO) samples described in Elith and Graham (2008), we also created pseudo-absence samples. For these we randomly sampled from the 80000 grid cells in the region, with sample sizes of 1000 and 3000 sites. These will be called P0.1000 and P0.3000. In each of these pseudo-absence samples, some sites will, in reality, be inhabited by the species (i.e. as expected for pseudo-absences like this, we will create contaminated absences). In our samples there were 116 true presence sites used as pseudo-absences in P0.1000 and 355 in P0.3000. However in the model we treat all as absences to be consistent with the case where true absence is not known (Figure S3b).

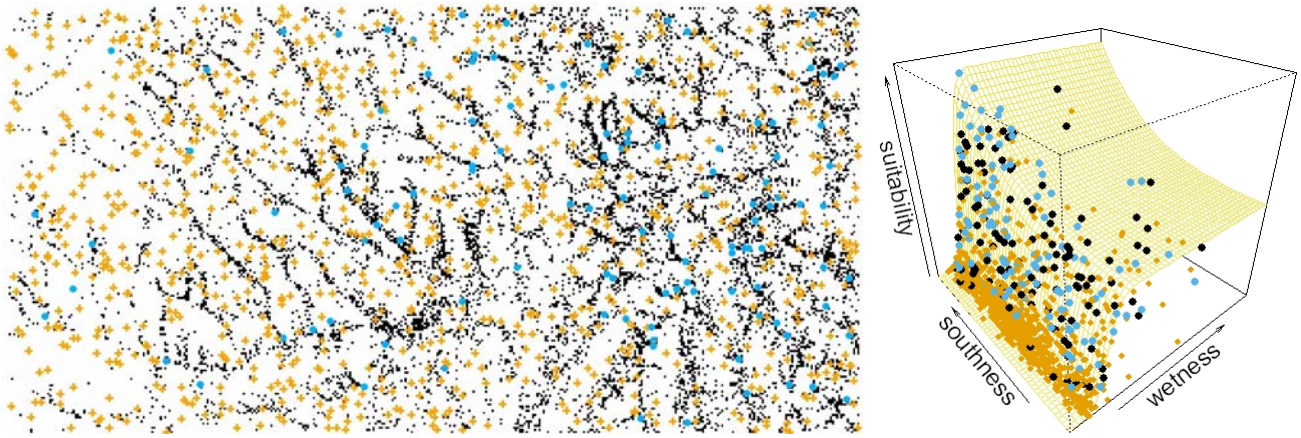


Figure S2: (a) Left: Map showing realised presence-absence data (black presence white absence, with the PA sample: blue (presence) and tan (absence); (b) Right: The data sample for presence plus pseudo-absence 1000. Samples are shown at their original true suitability value (vertical axis), but were converted to presence (blue and black) or absence (orange) as described in the text. The black points are those in the pseudo-absence sample that were subsequently changed to "absence" for modelling.

Online supplementary material: Appendix S2: Running and testing GARP

We used the current version of Desktop GARP (Genetic Algorithm for Rule-set Production; Stockwell and Noble 1992; version 1.1.6 from late 2007) and followed Peterson *et al.* (2007) for parameter settings. This meant that the sampled presence data (115 records) were supplied for model fitting, and we allowed GARP to select the pseudo-absences. We used 50% data for training, and 50% for extrinsic evaluation, and created 500 models each with a convergence limit of 0.01 and 1000 maximum iterations (500 models took about 8 hours on desktop PC). We offered all three environmental predictors and allowed all rule types, then for predictions evaluated the mapped predictions for performance measures and also projected to grids with the evaluation strip included. We chose two subsets from the 500 models: first, following Peterson *et al* (2007), selected the 20% of models[1] with the lowest extrinsic omission error, and then selected from that subset of approximately 100 models the twenty models with commission errors in the middle of the range of commission indices. Second, selecting models based on low omission is considered best for predicting potential distributions might not be optimal for this application where we are attempting to accurately model records of the true niche of the species. Therefore, we took a second subset like the first, but the first 20% were selected as those in the middle of the range of extrinsic omission errors. We processed both these subsets of 20 (using the mean prediction as the prediction per grid cell), and found that the results were similar (Figure S3, rows 2 and 3). In the paper we present the second variant – i.e. mid external omission and mid commission error - because conceptually it seemed more consistent with the task required. This meant that in the paper geology was not modelled well, but the test results (Table S1) were better - compare rows 4 and 5, Table S1. We also tested various other combinations of the 500 models, in an attempt to reduce commission error (Figure S3). None of our attempts improved GARP performance so that it was comparable to the other methods, though we note that the mean of 500 (fourth row, Figure S3)provided the best results. We did not include it in Elith and Graham (2008) because it is not the method recommended by those most experienced at running GARP.

The fitted responses across a range of pairwise values of wetness and southness (Figure S3, right column) revealed that, except for the minimum set, the responses wetness and southness are consistent across a range of values of the other and so are not dependent on the precise value at which wetness or southness were kept constant. The relevant predictive maps are shown in Figure S5.

To explore the consistency of these results we repeated the analysis (about 500 runs, same settings) on 2 repeats of the same data and one new sample of the simulated species, using 125 new presence records from the sample of 1000 (see earlier). Results were summarised for mid-omission and mid-commission errors as above (Figure S6). There is some variation amongst runs but again no results do as well as the other methods tested in the paper.

Table S1: Comparison of model results with truth, as realised by the presence-absence map (columns 2, 3 and 4) and the suitability values (column 5). For all statistics except deviance, higher is better. Models 2 to 5 are those described in paragraph 1, above, and 6 to 8, in paragraph 2. Models 5 to 7 (highlighted) are run with the same presence records with the same settings, so any differences are due to stochasticity in the model.

Model	AUC	Remaining deviance	COR.pa	COR.si
1.Truth (suitabilities)	0.872	0.514	0.508	1.000
2.GARPmin	0.564	3.233	0.224	0.441
3.GARPmean	0.842	1.856	0.407	0.805
4.GARPlow omission	0.812	4.812	0.385	0.752
5.GARPmid omission (paper)	0.822	3.391	0.401	0.793
6.GARPrepeat1	0.807	4.470	0.388	0.767
7.GARPrepeat2	0.814	3.944	0.391	0.772
8.GARPnew125pres	0.819	4.034	0.386	0.757

[1] If at the bounds of the subset there were multiple sites with the same extrinsic omission error, we expanded the sets to include all with that error rate. However when selecting the second set based on commission error, strictly selected 20 sites.

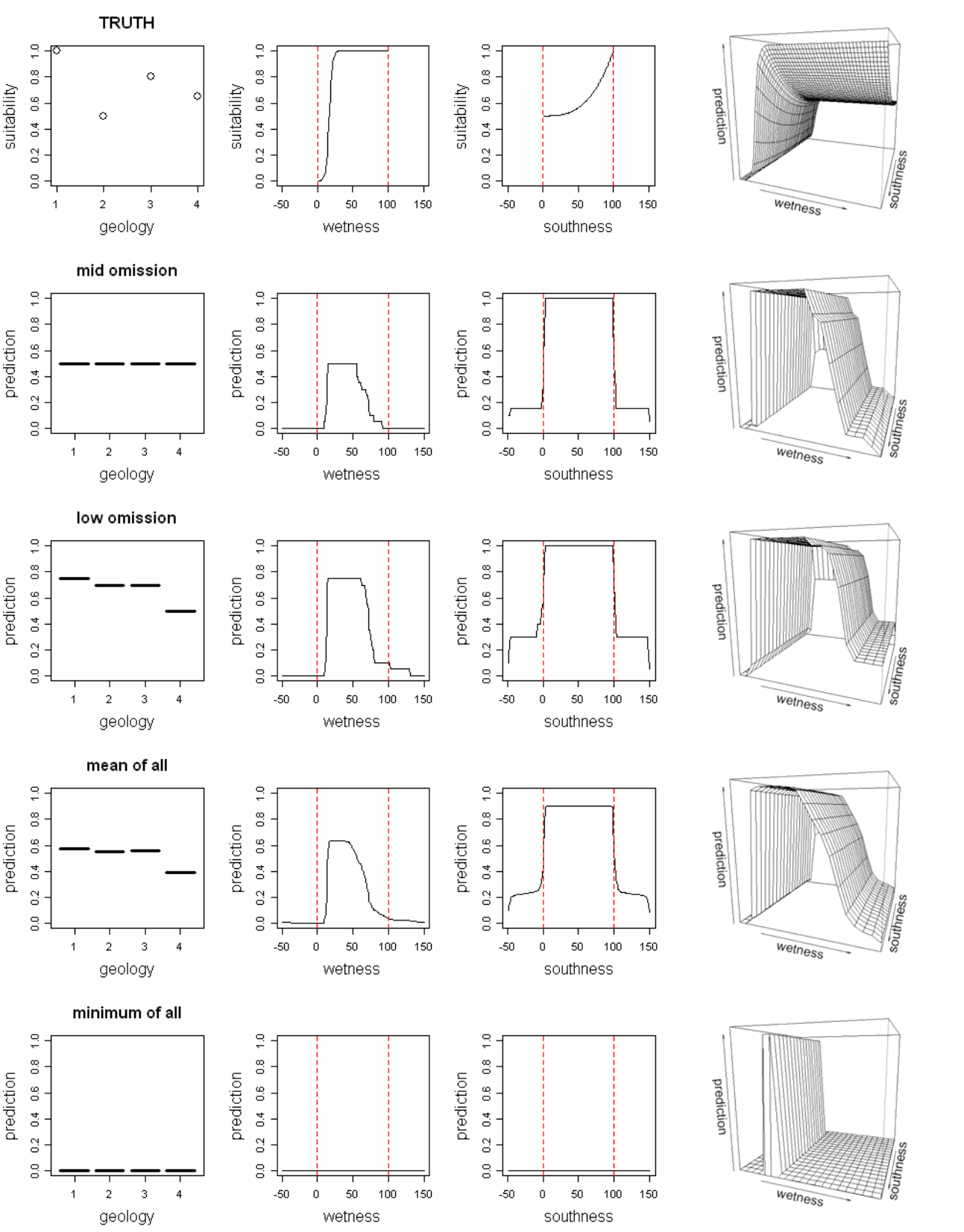


Figure S3: Fitted functions for truth plus four different summaries of the GARP run of 500 models. The second top one ("mid omission") is the one presented in Elith and Graham (2008). Related pairwise plots and predicted maps are in Figures S4 and S5 (not that these are rotated to a different perspective compared with those in Elith and Graham 2008).

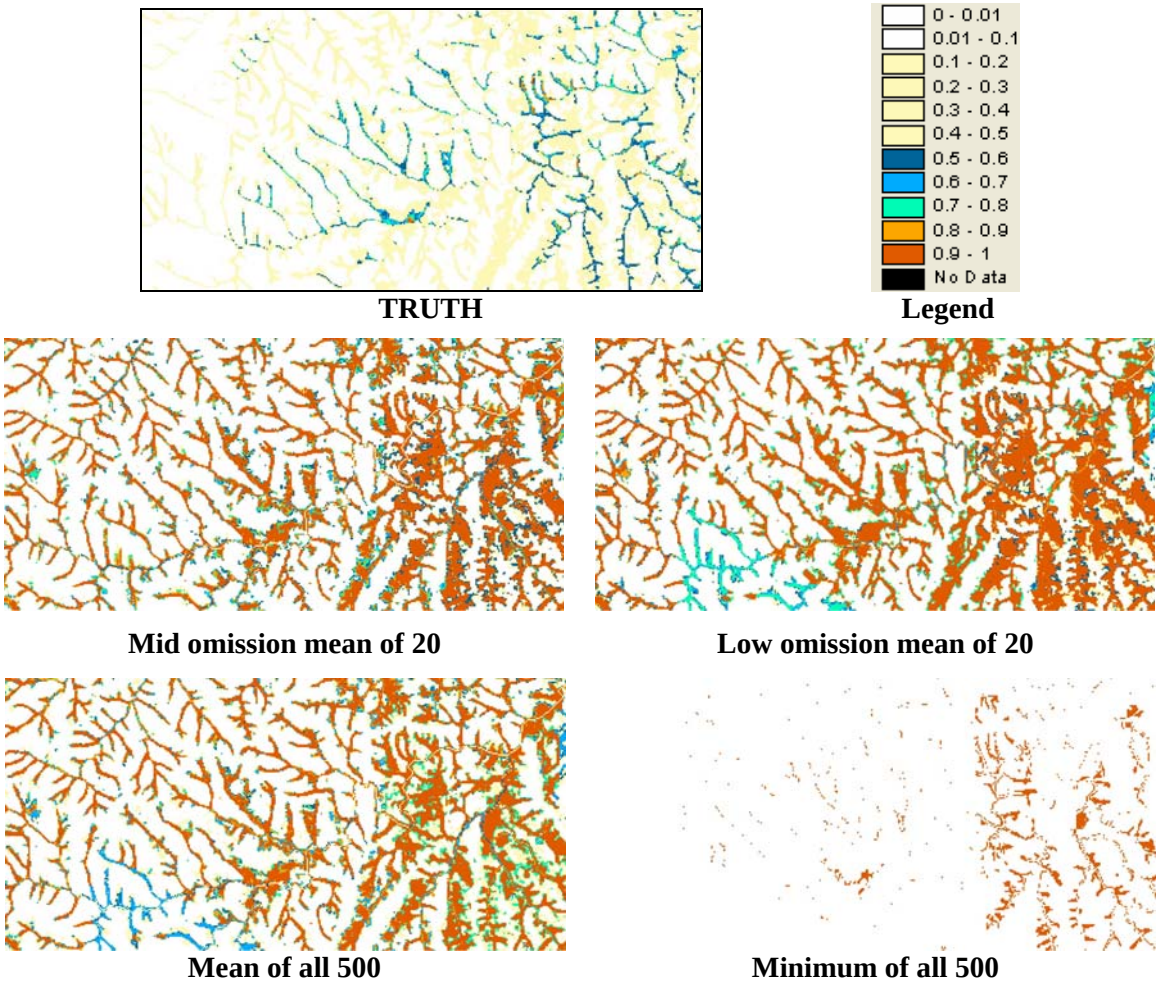


Figure S4 – Maps of predicted distributions, from the models illustrated in Fig. S3 and summarised in Table S1. Legend same as for main paper, and shown in top row

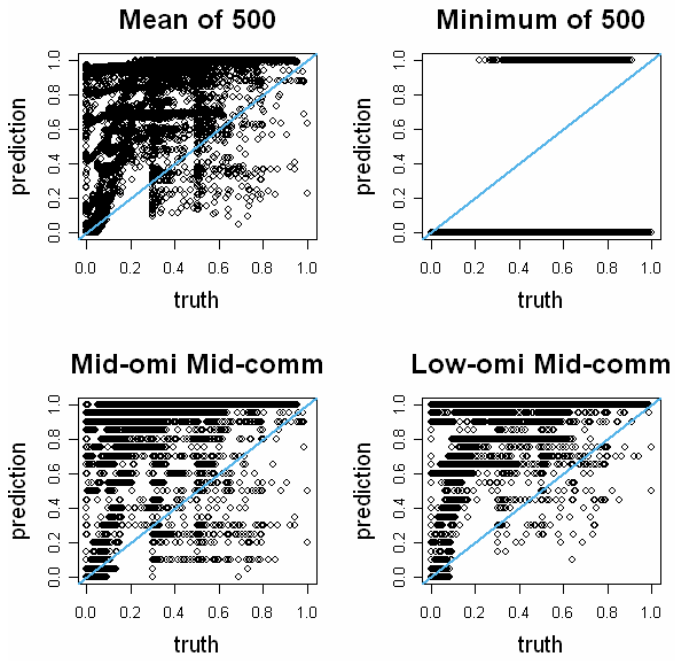


Figure S5: Predictions (y axis) versus the true suitability for all 80000 grid cells in the maps in Figure S4, for the models from Table S1 and Figures S3 and S4. The blue diagonal line shows the 1:1 relationship.

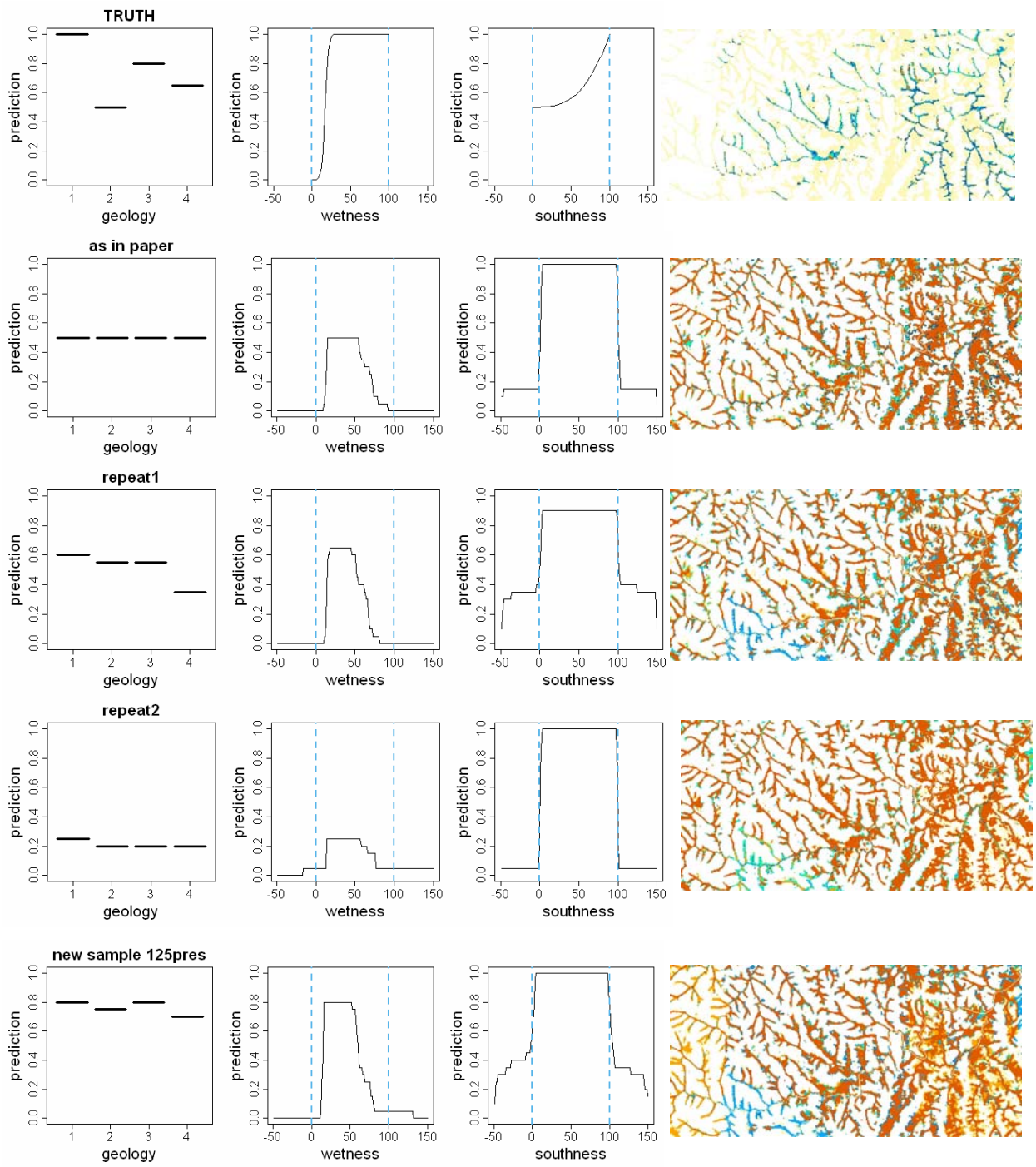


Figure S6: Results from 3 independent runs of GARP on the sample of 115 presence records(rows 2 to 4), and a new sample of 125 records. In each 500 models were produced then subsetted as described in the text. The models are based on the mid external omission / mid commission scenarios. Fitted functions (left panels) and mapped distributions (right). The legend for the distribution is the same as that in Figure S4.

Online supplementary material: Appendix S3: Running and testing RandomForests

Random Forests (RF) is a machine learning method that builds an ensemble of classification or regression trees. It uses *bagging* (bootstrap aggregation) to form the ensemble, taking a new bootstrap sample of the training data for each new tree. The reason for making many trees (a "forest") is that the variance of single trees, a known problem, is reduced by bagging. A useful side-effect of using bagging is that, at each step, there is an "out-of'bag" sample (i.e. those records not selected) that can be used for testing the model. RF are called "random" forests because at each split only a random subset of the candidate predictors are considered. This de-correlates the trees and improves the variance reduction. Trees are fully grown and not pruned For regression trees the results are averaged, for classification, each tree casts a vote for the predicted class. For binary data such as ours, classification trees are used but the final votes can give a probability rather than a binary output. Further details on the theory of RF can be found in the publications mentioned below.

RF were run using the R library *randomForest*. JE ran the models and understood the theory of RF but had little experience. Recent publications were read (Prasad *et al.* 2006, Benito Garzon *et al.* 2006, Breiman 2001, Cutler *et al.* 2007), and experts consulted in person or via web pages.

Random forests are generally considered easy to tune. The most important choices are the "*mtry*" and "*ntree*" settings in the R version, representing how many variables are randomly selected at each split of the tree as it is grown (*mtry*), and how many trees are allowed in the ensemble (*ntree*). Rules of thumb are used to estimate a good value for *mtry*, for classification often either sqrt(n) (Cutler *et al.* 2007), where n = number of candidate variables, or log(n) (D. Marginateau, pers.comm.); *mtry* can be as low as 1. Cutler *et al.* (2007) suggest that *ntree* as low as 50 can be suitable; in R the default is 500, and Prasad *et al.* (2006) used 1000 because it stabilised their results. There is also an R function called tuneRF that can be used to set mtry in relation to error rates. Cutler *et al.* (2007) comment on it in their appendix: "We have not used this function, in part because the performance of RF is insensitive to the chosen value of *mtry*, and in part because there is no research as yet to assess the effects of choosing RF parameters such as *mtry* to optimize out-of-bag error rates on the generalization error rates for RF".

For classification trees, the error rates on the classes in the out-of-bag estimates can be balanced, if this is appropriate for the application, by putting priors on the class weights. In other words, for a binary outcome the model can try to predict "0" as well as it predicts "1".

In our modelling we explored the effect of changing *mtry* and *ntree*, in various combinations. We also looked at the effect of balancing error rates (by use of class weights in R) and tested *tuneRF*. Our results were sensitive to *mtry*, with the best results using *mtry* = 1. This is consistent with *mtry* = log(3) and with the results from tuneRF, but not so clearly with sqrt(3) = 1.7, which perhaps might have been rounded to 2. Class weights and *ntree*s also affected the outcome. The best results were obtained with the settings used in the paper (*ntree*s = 500, *mtry* = 1 and no class weights) or the comparable model with 1000 trees. These are compared below with examples of some of the other settings tested. The results show that it is important to test settings. Models with *mtry* > 1 were more chaotic than those with mtry=1. With 50 trees the response to geology was not modelled properly. Out of bag estimates or cross-validations could be used to systematically test a range of settings to get best predictive performance for the given application.

Table S2: Details of the models described in the text above, showing the effect of varying parameter settings (first 3 rows) on the error estimates (rows 4 to 6) and evaluation statistics (rows 7 to 10). The meaning of the statistics is described in Elith and Graham (2008).

Model: / Measure:	rfpa1 (paper)	rfpa2	rfpa3	rfpa4	rfpa5	rfpa6	rfpa7
1. *mtry*	1	1	1	1	1	2	3
2. *ntree*s	500	500	500	50	1000	500	500
3. classwt (0,1)	null	1,1	2,3	null	null	null	null
4. oob[1] error overall	0.098	0.184	0.221	0.103	0.105	0.107	0.121
5. oob[1] error class1	0.011	0.168	0.220	0.018	0.008	0.038	0.054
6. oob[1] error class2	0.765	0.304	0.226	0.757	0.852	0.635	0.635
7. auc	0.834	0.843	0.838	0.814	0.835	0.828	0.823
8. remaining deviance	0.736	0.785	0.902	0.948	0.712	0.748	0.769
9. cor.pa	0.448	0.434	0.421	0.438	0.447	0.425	0.412
10. cor.si	0.875	0.853	0.827	0.855	0.874	0.822	0.793

[1] oob = out-of-bag estimate from R

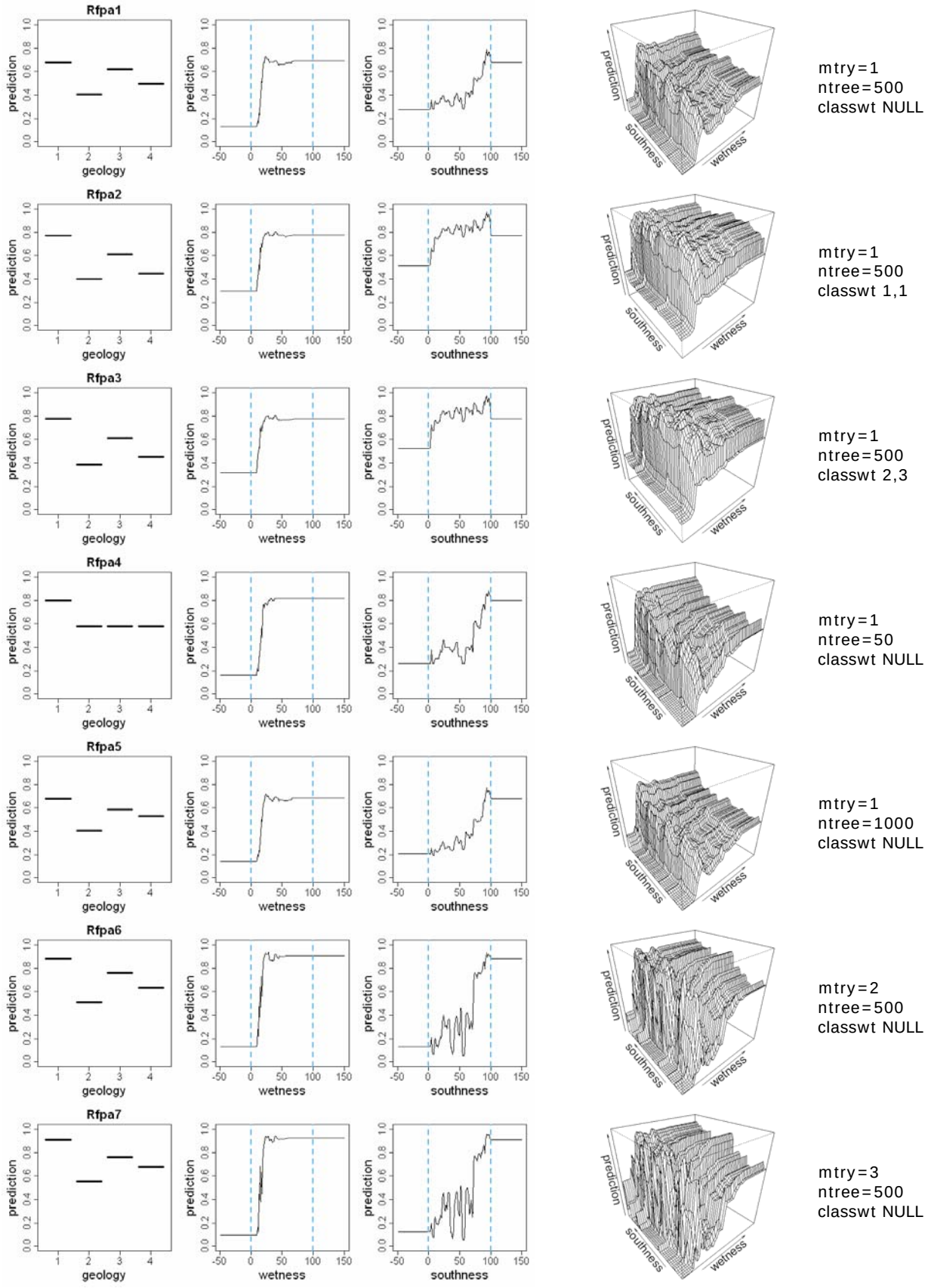


Figure S7: Fiited functions for the seven random forest models described in the text and Table 3. Note the effects of changing mtry (rows 1,6 & 7), ntree (rows 1, 4 & 5) and class weights. The first model was presented in Elith and Graham 2008.

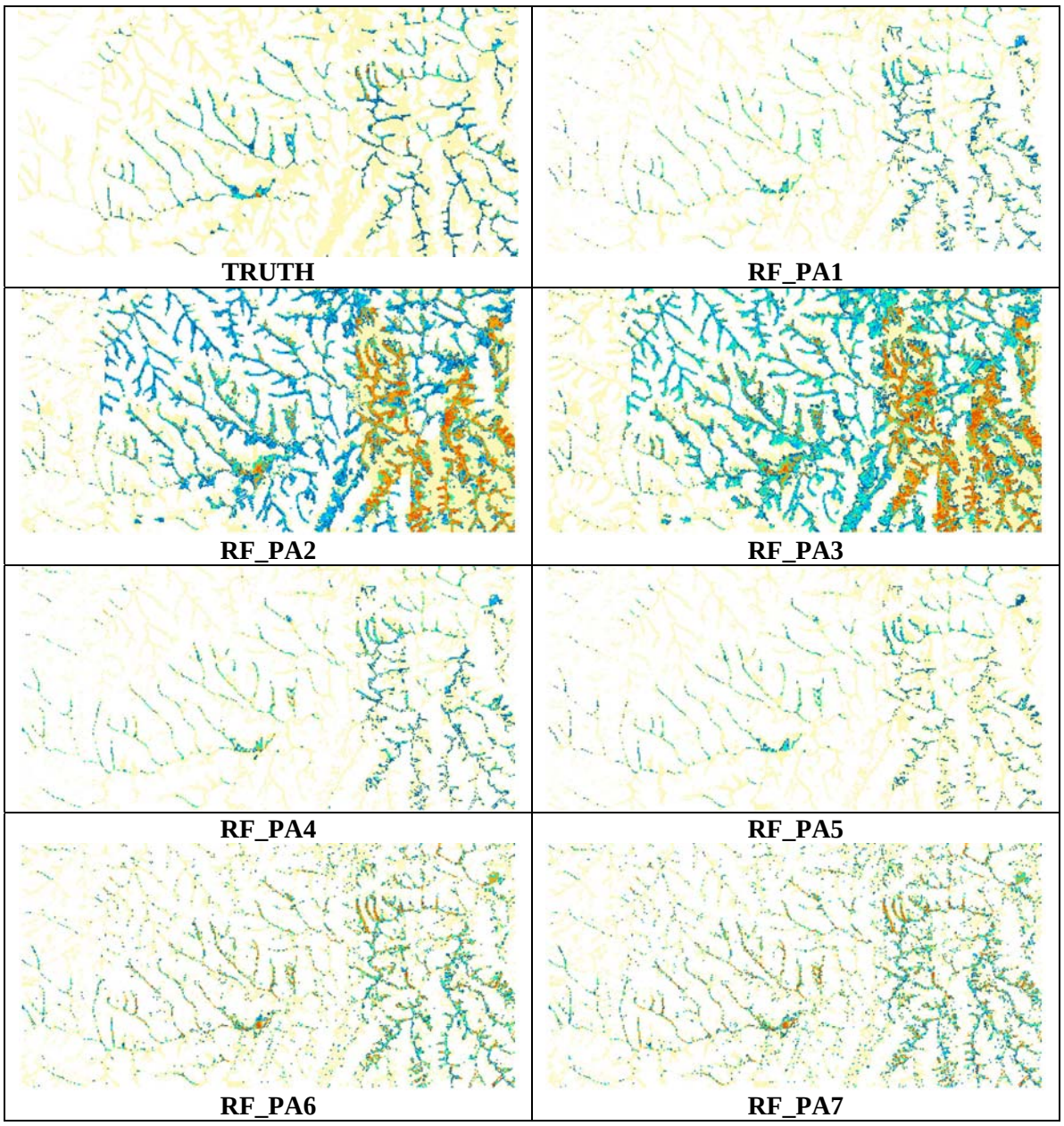


Figure S8: Mapped distributions of the virtual species (top left) and predictions of relative suitabilities from the methods detailed in the text, Legend: white < 0.1, cream 0.1 to 0.5, blue-lightblue-green-orange-vermillon at steps of 0.1 from blue (0.5 to 0.6) to vermillon (0.9 to 1), as in Figure S3

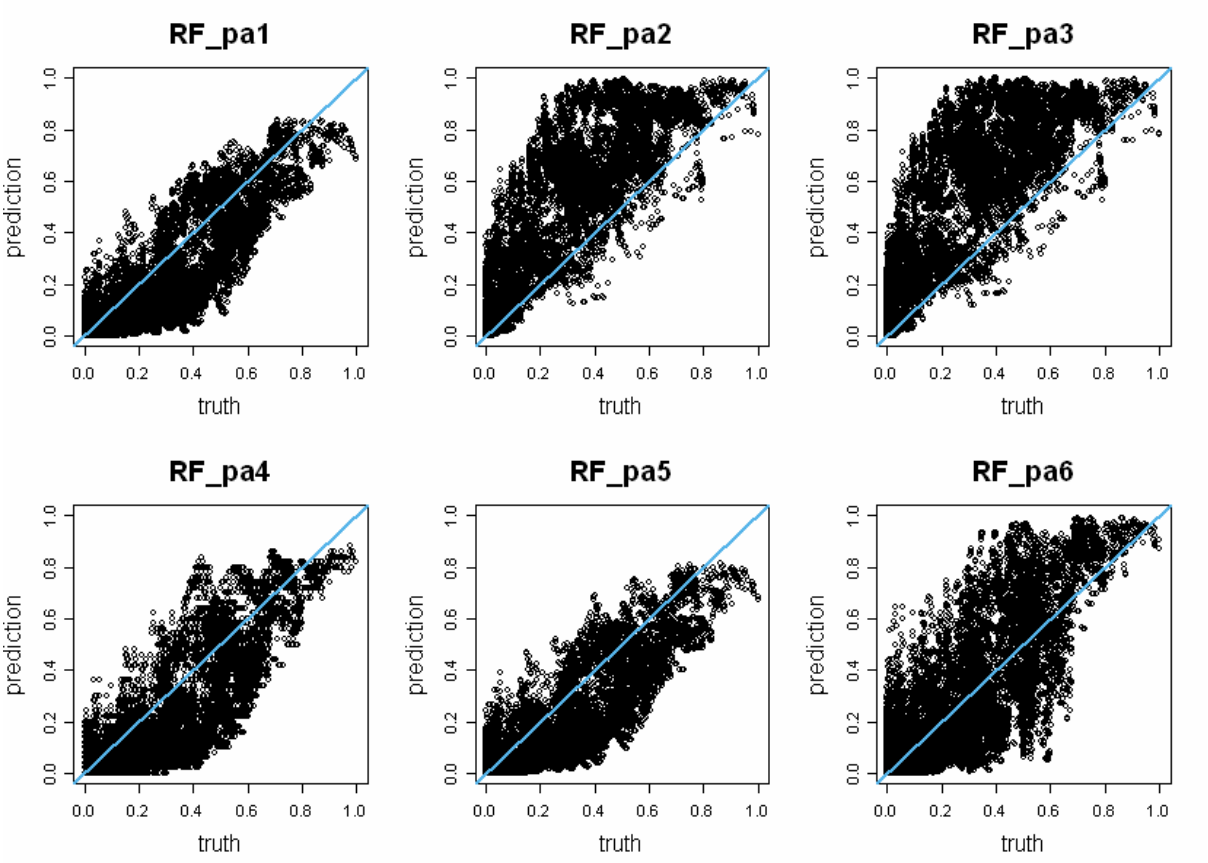


Figure S9: Predictions versus truth for all 80000 grid cells in the maps in Figure S8, for the models from Table S2 and Figures S7. The blue diagonal line shows the 1:1 relationship.

Online supplementary material: **Appendix S4: Running Maxent, BRT and GLM**

Maxent, boosted regression trees (BRT) and generalised linear models (GLMs) were straightforward to run and did not need extended testing, because the modeller was familiar with the methods. Whilst it is possible that performance might have improved with some changes in parameterisation, we were confident that the approaches used were well representative of the capacity of the methods.

Maxent (Phillips *et al.* 2006, Phillips and Dudik 2008) is a machine learning method, using the principles of maximum entropy models to model the species distribution. The idea is to set some constraints that enable the prediction to reflect patterns in the sample, and then select a model that maximises entropy (a uniform or spread out distribution) given that those constraints are met (either exactly or approximately). It considers the region (in this case, the full grid) then models the distribution of the species across that with a density estimation approach. The approach can be thought of as modelling the probability of the covariates (the predictor variables) conditional on species presence. Further information can be found in the papers cited above. Maxent version 3.2.1 was used, and run from the command line (specifically: -e env -s po.csv -t ge -o res -j proj -r -a -d -P). Settings for Maxent are presented in the paper, and are mostly the recommended defaults. Because Maxent is set up to model species distributions and the settings have been tested on large data sets, the defaults tend to perform well. The program sets feature selection and regularisation parameters (Phillips *et al.* 2006) in relation to the number of presence records supplied. In this case, with 115 records, it allowed all feature types with fairly strong regularisation (control over) the threshold features. This allows it to model flexible relationships without overfitting. The only exception to the usual defaults is that we used the "-d" flag (see help file provided with the program), which does not add the samples to the background data. This gives Maxent the best chance to have a well calibrated output.

The boosted regression trees were run in R (v. 2.6.1) with the *gbm* library and custom code written by John Leathwick and JE (Elith *et al.* 2008). That paper and others by the authors (e.g. Leathwick *et al.* in press) give detailed descriptions of the method. Briefly, an ensemble of regression trees are formed in a forward stagewise procedure ("stochastic gradient boosting"), where at each step the tree that is added is the one that best explains the residuals from the previous tree(s). The method models binary data accurately by using a logit link function, just as in a GLM. BRT's need careful choice of settings, but once the principles are understood, this is not difficult. We chose settings that would give at least 1000 trees, and that would grow trees deep enough to model interactions. The algorithm uses cross-validation to choose how many trees to add, stopping before it is too overfit (Elith *et al.* 2008). The final model comprised 4250 trees with a learning rate (shrinkage) of 0.001 and a tree complexity of 3.

The generalised linear model (GLM) was run in R version 2.6.1 using function *glm*. For all models we created all possible subsets of models where the allowable fit for each variable was: (1) geology: in as a 4-level factor, or out; (2) wetness and southness: out, in as linear, quadratic or cubic function. From all these models, we selected the model with the lowest Akaike Information Criterion (AIC).

Online supplementary material: **Appendix S5: Using pseudo-absences**

The test with pseudo-absences is a demonstration of the effect on model structure and performance of using presence records and pseudo-absences (appendix S1) in what has been described as a "naïve" model (Ward *et al*. in press, Phillips *et al.* in press). In this, a logistic model is fit to the presence (PO) and background data. If a species is rare, the background data will resemble true absences and the naive model will be close to the true model. But with higher levels of "contamination" (presences in the background sample) the naïve model can be biased. Both GLM and BRT are used here as logistic models. The models were fit with the same settings as described in Appendix S4, with the pseudo-absences replacing true absences. For each set of pseudo-absence data (PO.1000 and PO.3000, see Appendix S1), we made two models from each method, in one applying weights on the data so that the sum of the weights on the presence records is the same as the sum of the weights on the absence records (PO.1000.wt etc). This has often been done (e.g Ferrier *et al.* 2002) and produces fitted values and predictions that are distributed across the possible range of the response (here, 0 to 1), rather than predicting many very low values, as occurs if using many more pseudo-absences than presences.

The results are briefly summarised here and we suggest that they are worth pondering in some detail (Table S3; Figures S19 to S15). The results for the unweighted models with the sample of 1000 pseudo-absences were almost as good as those with pa data, mainly because the number of pseudo-absence samples compared with the 115 presences is relatively close to the true prevalence of the species (i.e. there were 885 true absences in the pa data), and because the species in not common in the landscape. With a more common species, contamination of the pseudo-absence sample would have a larger effect. Weighting the data in the models had no effect on the discrimination of the models as long as variable selection wasn't affected. So, for BRT the AUC for the unweighted / weighted pairs are very close, whereas the GLM tended to identify the best model as one without geology when the data were weighted. Dropping geology as a predictor negatively affected all evaluation statistics for the GLMs (Table S3). BRT models the data reasonably in all cases but models the response to southness as much too muted in the weighted models. The GLM never models the response to southness properly at high wetness values (see right side of 3-dimensional plots) but this doesn't affect the evaluation statistics much because there are few data in that part of the environmental space (see Figure 2, Elith and Graham 2008)

We note that there are statistical solutions to modelling data such as these with more statistical rigour. They are described in Ward *et al.* (in press), and software for running presence-only BRTs with these will be released soon (Gill Ward, pers.comm.).

Table S3: Comparison of model results with truth, as realised by the presence-absence map (columns 2, 3 and 4) and the suitability values (column 5). For all statistics except deviance, higher is better.

Model	AUC	Remaining deviance	COR.pa	COR.si
Truth (suitabilities)	0.872	0.514	0.508	1.000
Maxent	0.861	0.612	0.467	0.922
BRT.pa	0.862	0.537	0.485	0.954
BRT.po.1000	0.856	0.568	0.464	0.915
BRT.po.1000.wt	0.858	0.842	0.442	0.872
BRT.po.3000	0.855	0.703	0.435	0.851
BRT.po.3000.wt	0.857	0.848	0.441	0.871
GLM.pa	0.863	0.546	0.480	0.941
GLM.po.1000	0.853	0.560	0.468	0.922
GLM.po.1000.wt	0.843	0.924	0.419	0.825
GLM.po.3000	0.855	0.691	0.455	0.896
GLM.po.3000.wt	0.841	0.932	0.417	0.821

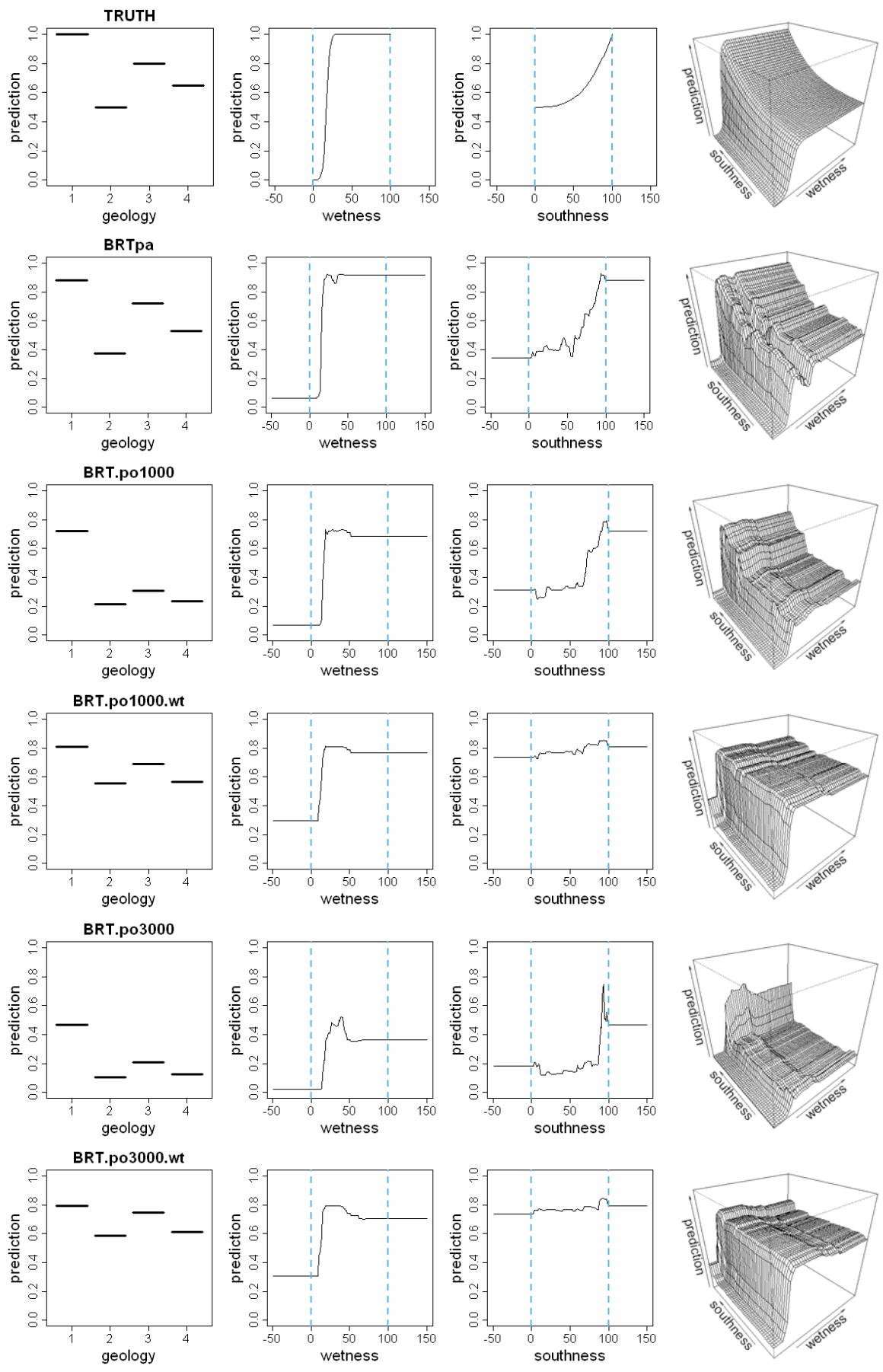


Figure S10: Fiited functions for the boosted regression tree models described in the text and presented in Table S3

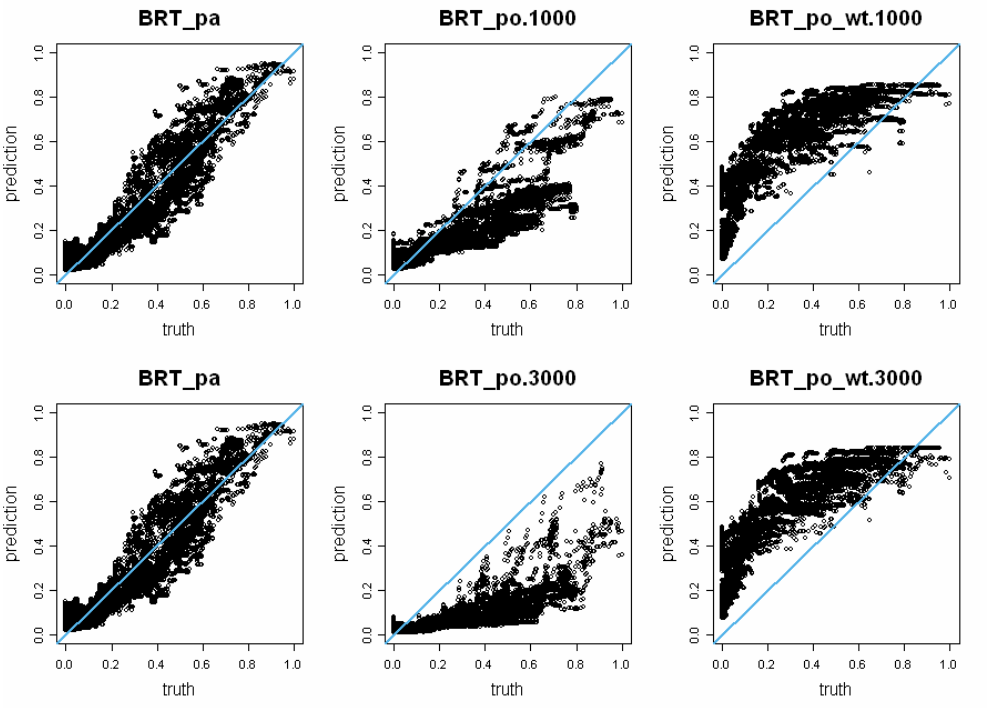


Figure S11: Predictions versus truth for all 80000 grid cells in the maps in Figure S12, for the models from Table S3 and Figure S10. The blue diagonal line shows the 1:1 relationship.

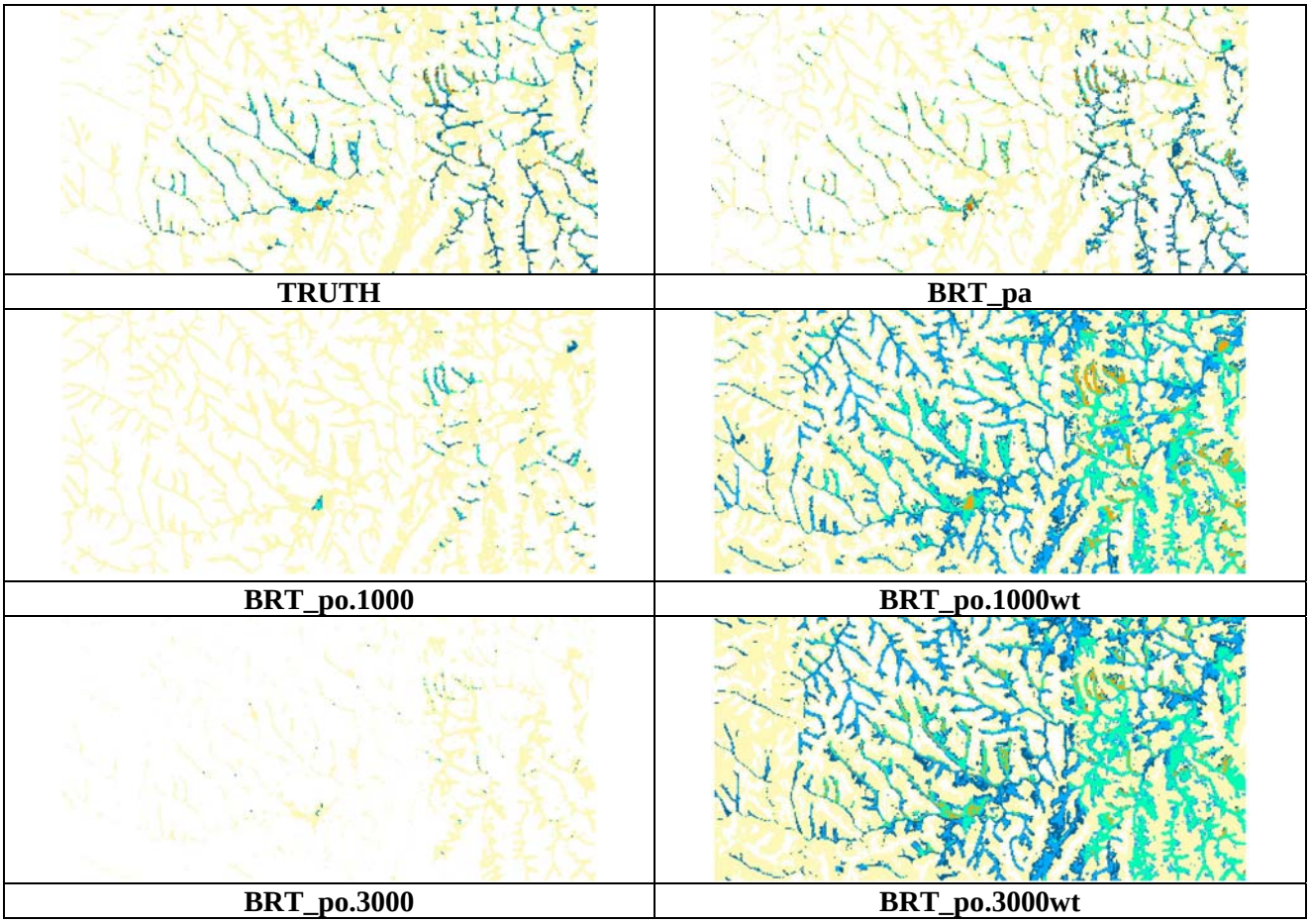


Figure S12: Mapped distributions of the virtual species (top left) and predictions of relative suitabilities from the methods detailed in the text, Legend: white < 0.1, cream 0.1 to 0.5, blue-lightblue-green-orange-vermillon at steps of 0.1 from blue (0.5 to 0.6) to vermillon (0.9 to 1), as in Figure S3

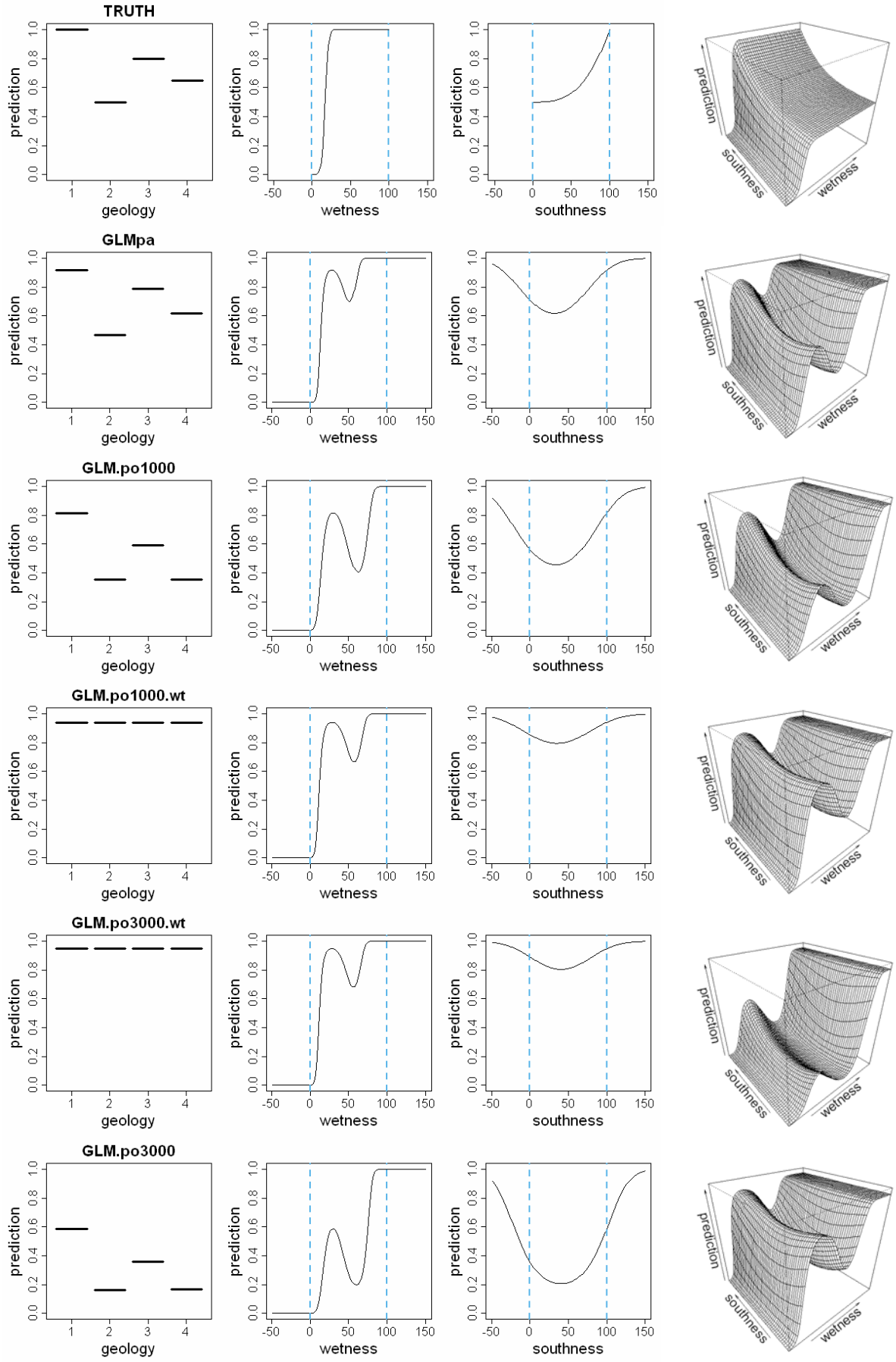


Figure S13: Fiited functions for the generalised linear models described in the text and presented in Table S3

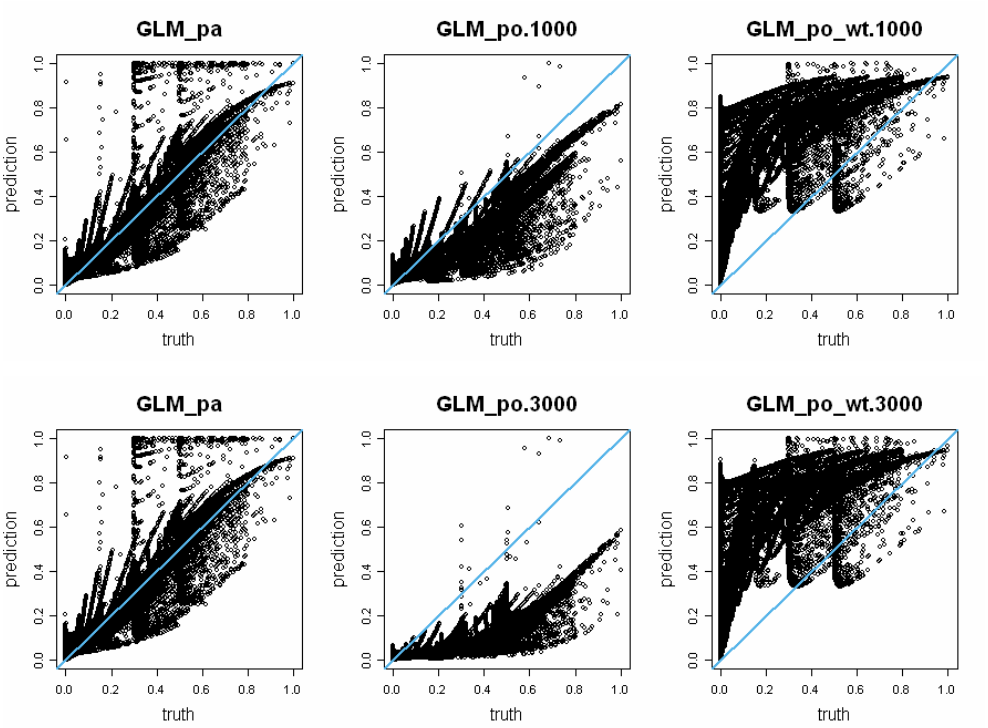


Figure S14: Predictions versus truth for all 80000 grid cells in the maps in Figure S14, for the models from Table S3 and Figure S13. The blue diagonal line shows the 1:1 relationship.

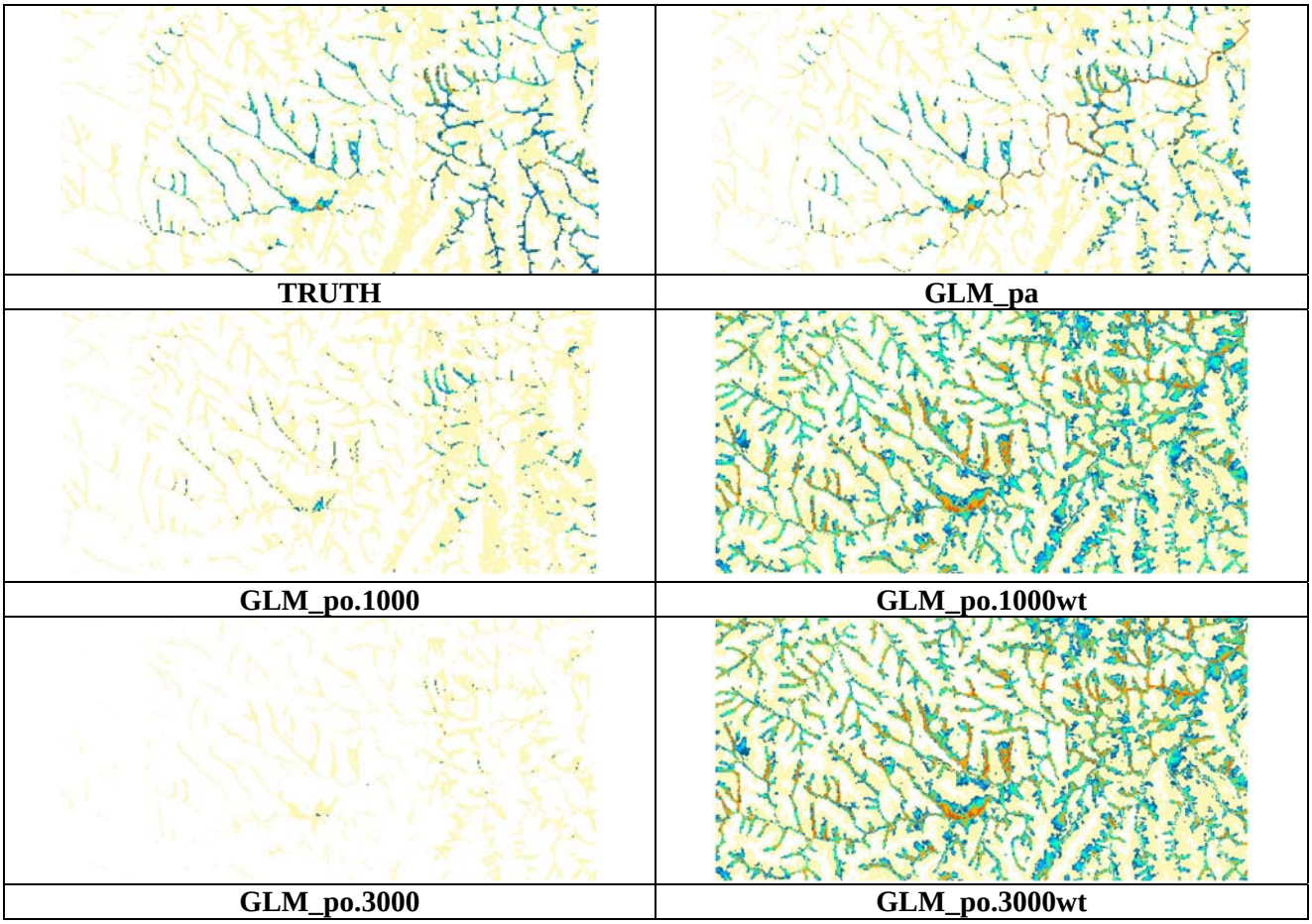


Figure S15: Mapped distributions of the virtual species (top left) and predictions of relative suitabilities from the methods detailed in the text, Legend: white < 0.1, cream 0.1 to 0.5, blue-lightblue-green-orange-vermillon at steps of 0.1 from blue (0.5 to 0.6) to vermillon (0.9 to 1), as in Figure S3. Note that the effect of dropping geology as a predictor is to lose the definition of poorer habitat towards the west (left)

References

Benito Garzon, M. *et al.* 2006 in press. Predicting habitat suitability with machine learning models: The potential area of Pinus sylvestris in the Iberian Peninsula. - Ecol. Model. 197: 383-393

Breiman, L. 2001. Random Forests Technical Report. http://oz.berkeley.edu/users/breiman/randomforest2001.pdf.

Cutler, D. R. *et al.* 2007. Random forests for classification in ecology. - Ecology 88: 2783-2792.

Elith, J. and Graham, C. 2008. On the need for a buyers guide to species distribution modelling. - Ecography.

Elith, J., Leathwick, J. R. and Hastie, T. in press. A working guide to boosted regression trees. - J. Anim. Ecol.

Ferrier, S. et al. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. I. Species-level modelling. - Biodivers. Conserv. 11: 2275-2307.

McCullagh, P. and Nelder, J. A. 1989. Generalized Linear Models. - Chapman and Hall.

Peterson, A. T., Papes, M. and Eaton, M. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. - Ecography 30: 550-560.

Phillips, S. J. et al. in press. Sample Selection Bias and Presence-Only Models Of Species Distributions. - Ecol. Appl.

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

Prasad, A. M., Iverson, L. R. and Liaw, A. 2006. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. - Ecosystems 9: 181-199.

R Development Core Team 2006. R: A Language and Environment for Statistical Computing. - In, R Foundation for Statistical Computing

Ridgeway, G. 2006. Generalized boosted regression models. Documentation on the R package "gbm", version 1.5-7. http://www.i-pensieri.com/gregr/gbm.shtml.

Stockwell, D. R. B. and Noble, I. R. 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. - Math. Comput. Simulat. 33: 385-390.

Ward, G. et al. in press. Presence-only data and the EM algorithm. - Biometrics.