00000100000
00001010000
00010001000
00100000100
01000000010
10000000001
Software notes

Ecography 32: 369–373, 2009
doi: 10.1111/j.1600-0587.2008.05742.x

Subject Editor: Carsten Rahbek. Accepted 30 December 2008

BIOMOD – a platform for ensemble forecasting of species distributions

Wilfried Thuiller, Bruno Lafourcade, Robin Engler and Miguel B. Araújo

W. Thuiller (wilfried.thuiller@ujf-grenoble.fr) and B. Lafourcade, Laboratoire d'Ecologie Alpine, UMR CNRS 5553, Univ. Joseph Fourier, BP 53, FR–38041 Grenoble Cedex 9, France. – R. Engler, Dépt d'Ecologie et d'Evolution (DEE), Univ. de Lausanne, Bâtiment de Biologie, CH–1015 Lausanne, Switzerland. – M. B. Araújo, Dept of Biodiversity and Evolutionary Biology, National Museum of Natural Science, CSIC, C/Gutierrez Abascal, 2, ES–28006, Madrid, Spain and Rui Nabeiro Biodiversity Chair, Univ. of Évora, Colégio Luís António, Rua Romão Ramalho no. 59, PT-7000-671, Évora, Portugal.

BIOMOD is a computer platform for ensemble forecasting of species distributions, enabling the treatment of a range of methodological uncertainties in models and the examination of species-environment relationships. BIOMOD includes the ability to model species distributions with several techniques, test models with a wide range of approaches, project species distributions into different environmental conditions (e.g. climate or land use change scenarios) and dispersal functions. It allows assessing species temporal turnover, plot species response curves, and test the strength of species interactions with predictor variables. BIOMOD is implemented in R and is a freeware, open source, package.

Species distribution models (SDM, Guisan and Thuiller 2005) are being used in nearly all branches of life and environmental sciences. A quick search in ISI Web of Science (18/02/08) using ''species distribution models'' OR ''niche models'' OR ''habitat models'' OR ''bioclimatic models'' highlights 21 973 papers, 74% of which published in the past 10 yr, in fields as varied as environmental sciences (53% of the records), zoology (15%), marine and freshwater biology (15%), life sciences and biomedicine (9%), biodiversity and conservation (8%), evolutionary biology (8%), fisheries (6%), forestry (6%), oceanography (5%), genetics and heredity (5%), amongst others. Advancement of knowledge in these fields is now intertwined with technical innovation in species distribution modelling and dependent on the existence of suitable software for fitting models and examining results. One difficulty with the use of species distribution models is that the number of techniques available is large and is increasing steadily, making it difficult for ''non-aficionados'' to select the most appropriate methodology for their needs (Elith et al. 2006, Heikkinen et al. 2006). Recent analyses have also demonstrated that discrepancies between different techniques can be very large, making the choice of the appropriate model even more difficult. This is particularly true when models are used to project distributions of species into independent situations, which is the example of projections of species distributions under future climate change scenarios (Thuiller 2004, Pearson et al. 2006). A possible solution to account for this inter-model variability is to fit ensembles of forecasts by simulating across more than one set of initial conditions, model classes, model parameters, and boundary conditions (for a review see Araújo and New 2007) and analyse the resulting range of uncertainties with bounding box, consensus and probabilistic methodologies rather than lining up with a single modelling outcome (Araújo and New 2007, Thuiller 2007). BIOMOD offers such a platform for ensemble forecasting (Fig. 1) using freeware and open-source R software (R Development Core Team 2008). It overcomes some of the limitations of existing software (e.g. being able to fit and compare different models) and incorporates several features for testing models (e.g. k-fold cross validation) and for examining species-environment relationships (e.g. using randomization tests) (Fig. 2).

Earlier implementations of BIOMOD (Thuiller 2003, 2004) provided limited ensemble simulations across model classes (i.e. four modelling techniques) and boundary conditions (i.e. up to five climate scenarios). Currently, BIOMOD enables larger simulations across initial conditions (i.e. by randomly re-sampling species distribution data and fitting different models for each sample), nine model classes (generalised linear models (GLM, McCullagh and Nelder 1989), generalised additive models (GAM, Hastie and Tibshirani 1990), multivariate adaptive regression splines (MARS, Friedman 1991), classification tree analysis (CTA, Breiman et al. 1984), mixture discriminant analysis (MDA, Hastie et al. 1994), artificial neural networks (ANN, Ripley 1996), generalised boosted models (GBM, Ridgeway 1999), random forests (Breiman 2001), and one rectilinear envelope similar to BIOCLIM (SRE, Busby

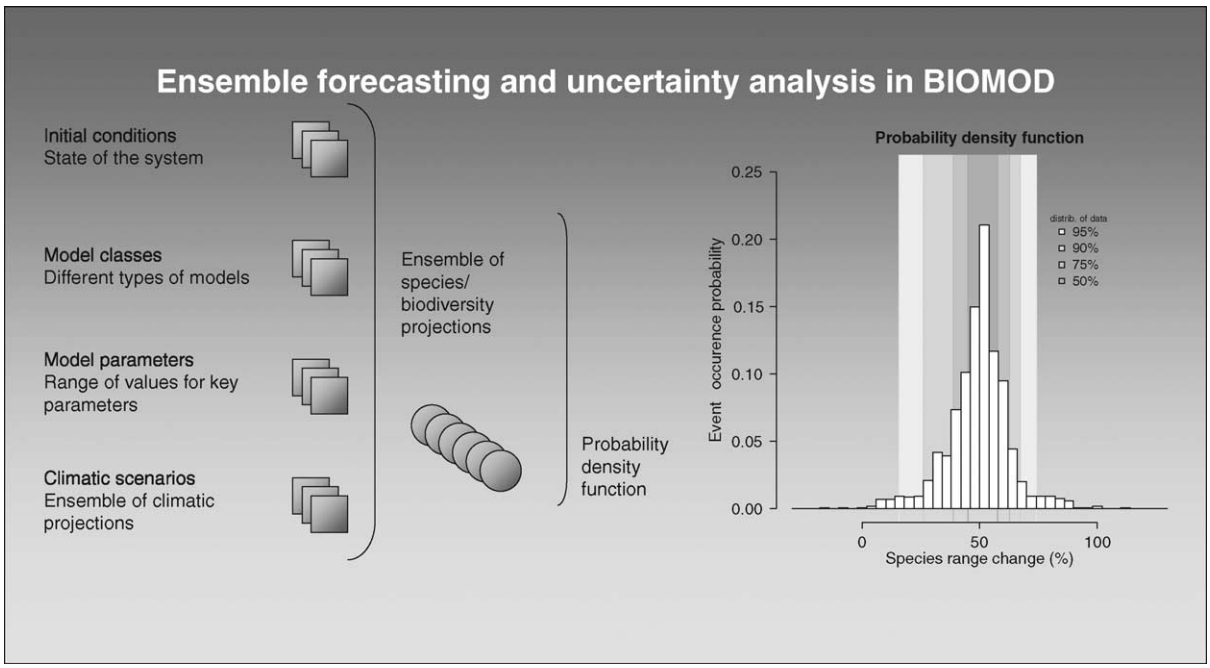

Figure 1. Probabilistic approach for forecasting species potential distributions (adapted from Thuiller 2007).

1991)), a variable number of model parameterizations (e.g. polynomials and smoothing splines of different orders in general linear or additive models, nodes in classification trees, hidden layers in neural nets), and a virtually unlimited number of boundary conditions. Most modelling techniques implemented in BIOMOD require that species distribution data are presence and absence. When data are presence-only, a simple solution is to generate random pseudo-absences. This can be done in BIOMOD using strategies of increasing complexity.

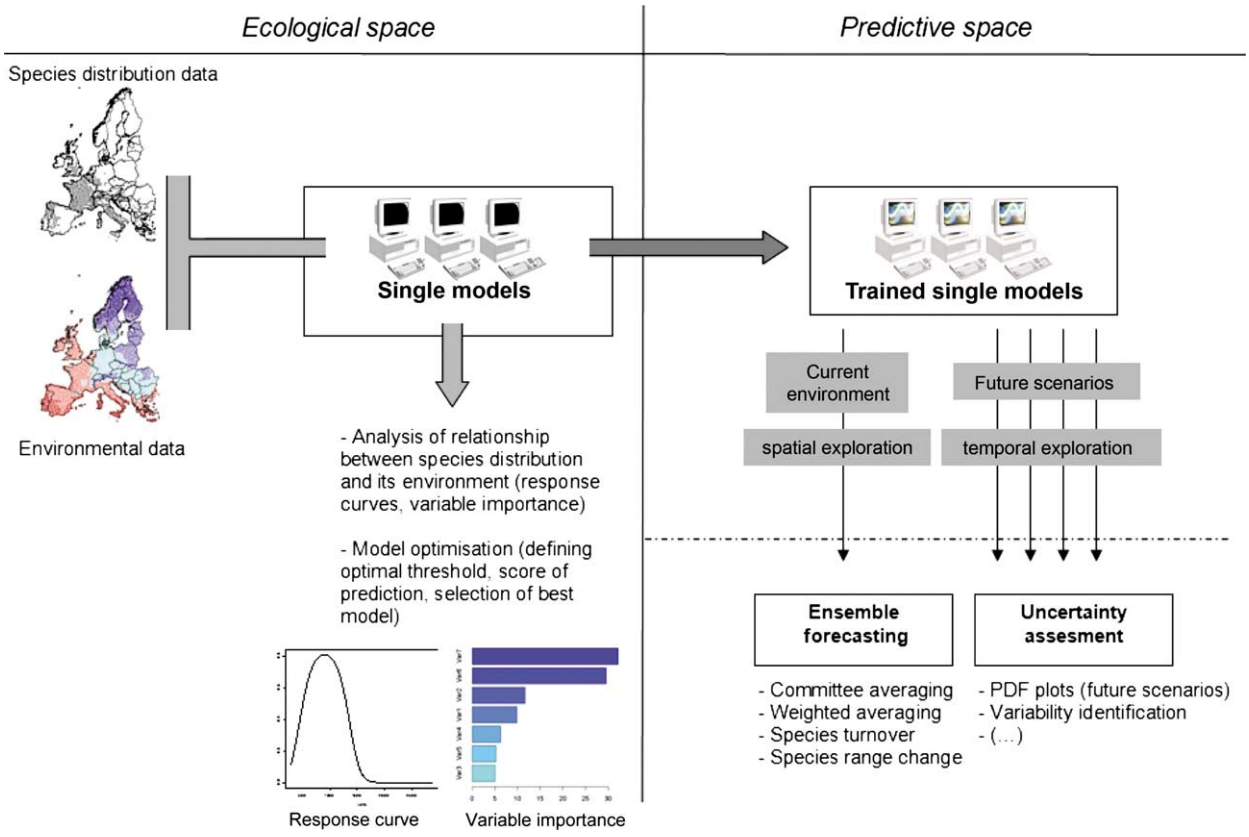

Figure 2. Schematic representation of the modelling procedure in BIOMOD.

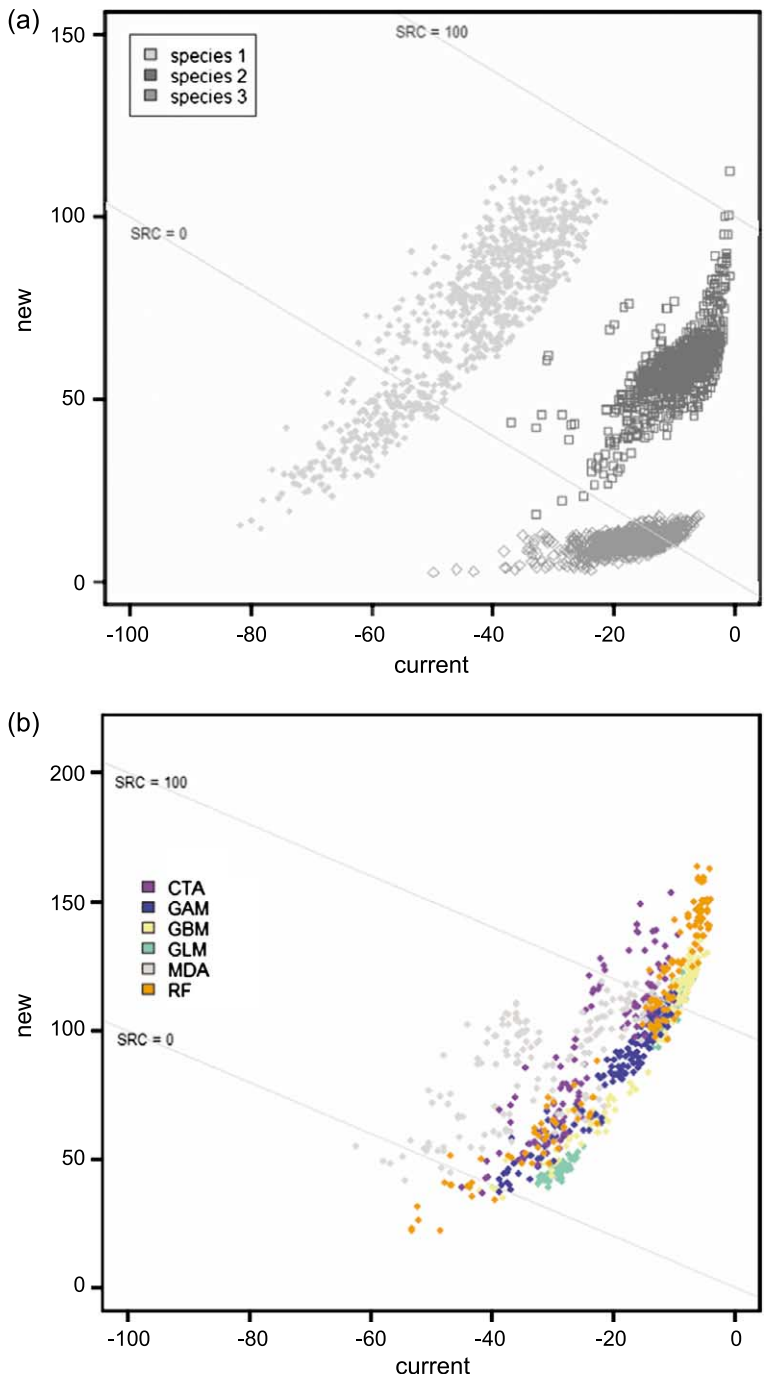

Figure 3. A representation of species potential range changes. Species range changes (SRC) are represented by a 2 axes plot: the first axis (x) represents the percentage of currently occupied sites projected to be lost; the second axis (y) represents the relative percentage – according to the current distribution's size – of currently unsuitable sites but projected to be suitable in future. For each dot, the sum of these two values gives the SRC. Every dot is a projection (e.g. different climate change scenarios, different model parameterisations). (a) Species range change projections for 3 species. Whilst species 2 keeps the majority of its current sites and gains around half of its potential habitat surface, species 3 is projected to keep as many of its sites, but will gain very few new suitable sites. Species 1 will lose suitability in many current sites but will also gain suitability in many new sites. (b) Species range change (SRC) according to different modellin techniques. This type of plot enables visual exploration of the sources of uncertainty accrued from different methodological sources of uncertainty, such as model algorithms, criteria to transform probabilities into presence and absence, climate change scenarios.

Evaluation of models in BIOMOD includes two sorts of analysis: assessments of the goodness-of-fit (=explanatory power) and of model accuracy (=predictive power). The former uses standard approaches associated with each technique; for example, ANOVA decomposition and AIC are available for both GLM and GAM, whereas rate of misclassification is used for CTA. The latter can be performed with three different procedures: the area under the relative operating characteristic curve (AUC, Hanley and McNeil 1982), Cohen's K (Monserud and Leemans

1992), and the true skill statistic (TSS, Allouche et al. 2006).

In an ideal world, model accuracy (e.g. AUC/Kappa/TSS) should always be evaluated with statistically independent data, i.e. training data that are not spatially autocorrelated with test data (Araújo et al. 2005a). When independent data are not available, an alternative is to use data-splitting procedures, whereby a proportion of the original data are used for training the models and the withheld data are used for model evaluation. A single random splitting of data was available in earlier implementations of BIOMOD (Thuiller 2003), but it proved to be a non-negligible source of variability when making predictions. Currently, BIOMOD allows much greater flexibility. Apart from the ability to define the size of the training and test datasets, BIOMOD also allows k number of data splitting runs to be computed. In each run, a model is fitted on one part of the data and evaluated on the left-out data. The evaluation values provided by each of the k splitting runs are then averaged, ensuring the final evaluation is quasi-independent of a particular realisation of random split. BIOMOD also provides a version of ''leave-one-out'' resampling. Users simply need to define the training sets as 100% of the data minus 1 record and then repeat the procedure a user-defined number of times (e.g. 1000 times). When non-independent data are used for model evaluation, variability in model accuracy should be interpreted as a measure of the sensitivity of model results to the initial conditions rather than as a measure of predictive accuracy (Araújo and Guisan 2006).

Model evaluation can be used to investigate the variability of predictions across modelling techniques. In BIOMOD a table displaying the AUC/K/TSS values is produced for each model and for each species. This table can be used for selecting the ''best'' model, i.e. the model providing greater accuracy on the test data for each species (Thuiller 2003). Assuming that no modelling procedure is always better, selecting the best model for each situation might be a useful option. The alternative ensemble forecasting paradigm draws on the assumption that model accuracy on non-independent test data is not representative of model accuracy on independent situations. In such cases, committee averaging of model predictions (giving the same weight to all predictions) can be implemented to derive a consensus prediction; an alternative is to combine models using some form of weighting (e.g. using PCA score value, Thuiller 2004, Araújo et al. 2006). There are a range of approaches to do this (for review see Araújo and New 2007), but in BIOMOD weights are currently calculated on the basis of models' predictive accuracy on test data (i.e. a form of ''stacking''). Empirical testing of consensus forecasting under climate change has shown that weighted approaches are promising (Araújo et al. 2005b, Marmion et al. 2008).

When using models to predict potential distributions in other regions, or times, it is often useful to visually examine species response curves (Austin and Gaywood 1994). To do so, BIOMOD uses an implementation of the ''evaluation strip'' procedure (Elith et al. 2005), making it possible to extract species' response curves independently of the model's algorithm.

There are some techniques available for characterising variable contribution in model predictions. However, these techniques are model-specific, so are the conclusions that one may extract from them. To overcome this limitation, BIOMOD uses a randomisation procedure to estimate the importance of each variable. The procedure is independent of the modelling technique, thus enabling direct comparison across models. This procedure uses Pearson correlation between the standard predictions (i.e. fitted values) and predictions where the variable under investigation has been randomly permutated. If the correlation is high, i.e. it is showing little difference between the two predictions, the variable permutated is considered not important for the model. This is repeated a user-defined number of times for each variable, and the mean correlation coefficient over the runs is kept. BIOMOD then gives a ranking of the variables for each of the model selected.

Finally, when projecting potential distributions of species into future environmental conditions, different dispersal assumptions can be made: no dispersal; unlimited dispersal; and user-defined species-specific dispersal. Measures of temporal turnover in potential species richness can then be calculated for each period (Thuiller et al. 2005), as well as species habitat change (Fig. 3a), and visualized according to the different models used (Fig. 3b) to emphasize the potential uncertainty coming from modelling technique or climate change scenarios' choice.

The BIOMOD R-package and a detailed user's guide of BIOMOD is available at the R-Forge website <biomod.r-forge.r-project.org>. To cite BIOMOD, or acknowledge its use, cite this Software Note as follows, substituting the version of the application that you used for ''Version 0'':

Thuiller, W., Lafourcade, B., Engler, R. and Araújo, M. B. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. – Ecography 32: 369–373 (Version 0).

Acknowledgements – BIOMOD is currently developed with funding from FP6 EU MACIS (No. 044399 SSPI) and ECOCHANGE (GOCE-CT-2003-506675) projects; MBA is also funded by FBBVA BIOIMPACTO project. We are indebted to the R community for their valuable contribution to the development and improvement of R code and libraries.

References

Allouche, O. et al. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.

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

Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.

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

Araújo, M. B. et al. 2005b. Reducing uncertainty in projections of extinction risk from climate change. – Global Ecol. Biogeogr. 14: 529–538.

Araújo, M. B. et al. 2006. Climate warming and the decline of amphibians and reptiles in Europe. – J. Biogeogr. 33: 1712–1728.

Austin, M. P. and Gaywood, M. J. 1994. Current problems of environmental gradients and species response curves in relation to continuum theory. – J. Veg. Sci. 5: 473–482.

Breiman, L. 2001. Random forests. – Mach. Learn. 45: 5–32.
Breiman, L. et al. 1984. Classification and regression trees. – Chapman and Hall.
Busby, J. R. 1991. BIOCLIM – a bioclimate analysis and prediction system. – In: Margules, C. R. and Austin, M. P. (eds), Nature conservation: cost effective biological surveys and data analysis. CSIRO, pp. 64–68.
Elith, J. et al. 2005. The evaluation strip: a new and robust method for plotting predicted responses from species distribution models. – Ecol. Model. 186: 280–289.
Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.
Friedman, J. 1991. Multivariate adaptive regression splines. – Ann. Stat. 19: 1–141.
Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.
Hanley, J. A. and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. – Radiology 143: 29–36.
Hastie, T. et al. 1994. Flexible discriminant analysis by optimal scoring. – J. Am. Stat. Assoc. 89: 1255–1270.
Hastie, T. J. and Tibshirani, R. 1990. Generalized additive models. – Chapman and Hall.
Heikkinen, R. K. et al. 2006. Methods and uncertainties in bioclimatic envelope modeling under climate change. – Prog. Phys. Geogr. 30: 751–777.
Marmion, M. et al. 2008. Evaluation of consensus methods in predictive species distribution modelling. – Divers. Distrib. 15: 59–69.
McCullagh, P. and Nelder, J. A. 1989. Generalized linear models. – Chapman and Hall.
Monserud, R. A. and Leemans, R. 1992. Comparing global vegetation maps with the Kappa statistic. – Ecol. Model. 62: 275–293.
Pearson, R. G. et al. 2006. Model-based uncertainty in species' range prediction. – J. Biogeogr. 33: 1704–1711.
R Development Core Team 2008. R: a language and environment for statistical computing. – R Foundation for Statistical Computing.
Ridgeway, G. 1999. The state of boosting. – Comput. Sci. Stat. 31: 172–181.
Ripley, B. D. 1996. Pattern recognition and neural networks. – Cambridge Univ. Press.
Thuiller, W. 2003. BIOMOD: optimising predictions of species distributions and projecting potential future shifts under global change. – Global Change Biol. 9: 1353–1362.
Thuiller, W. 2004. Patterns and uncertainties of species' range shifts under climate change. – Global Change Biol. 10: 2020–2027.
Thuiller, W. 2007. Biodiversity – climate change and the ecologist. – Nature 448: 550–552.
Thuiller, W. et al. 2005. Climate change threats to plant diversity in Europe. – Proc. Nat. Acad. Sci. USA 102: 8245–8250.