Appendix S1: Explanation of the effect of learning rate on predictive stability in boosted regression trees

As explained in Elith, Leathwick & Hastie (2008), the learning rate (*lr*) sets the shrinkage applied to each tree in the boosted regression tree (BRT) model. For example, a model with 1000 trees fitted with *lr* = 0.005 will produce predictions that are the sum of predictions from each of the 1000 trees multiplied by 0.005. However, it is important to recognize that while the same shrinkage is applied to all the trees in a BRT model, not all the trees have equal contributions to error reduction. This is because those trees fitted initially will describe the most general and strongest patterns in the data, and will therefore explain large amounts of deviance. This is clearly shown by the rapid initial reduction in predictive deviance in Figure 5 of Elith *et al.*(2008). By contrast, those trees fitted later in a BRT model will explain more particular features of the data, and will therefore result in lower incremental reductions in predictive deviance than the initial trees, as reflected in the flattening out of the deviance reduction curve in Fig. 5 of Elith *et al.* (2008).

This strong influence of the initial trees in a BRT model has important practical implications related to model variability, which become particularly apparent if a "fast" learning rate (i.e. large value of *lr*) is used in combination with stochastic boosting. While repeated fits of a model with large *lr* may produce models with similar predictive performance (as measured for example using AUC) and similar partial plots, marked differences (instability or variance) can occur in the fitted and predicted values for individual observations. This variation is caused by the stochastic component of model building, in which random subsets of data are selected for fitting each tree. Effects of stochasticity are more apparent with (i) larger learning rates, because individual trees have increased influence, and (ii) as the bag fraction is decreased, because smaller random subsets lead to increased chances of "unusual" trees being fit.

In effect, the variance in predictions or fitted values represents uncertainty around a mean estimate. Note that such uncertainty is also present in more deterministic models (e.g., GLM), despite the apparent stability of their fitted or predicted values, which remain constant if they are fitted repeatedly. In many of these models, the magnitude of the variation can be assessed by calculating estimates of the standard errors associated with the fitted values or predictions. In ecological and conservation applications a stable and accurate estimate of the mean will often be required. This can be achieved by building a BRT model with no stochasticity (bag fraction of 1), repeat runs of which produces identical values (i.e., no run to run variability). However, this will be achieved at the cost of some loss in predictive performance, because stochasticity often improves predictive performance (Friedman 2002). A more effective way to control this instability is to retain the stochastic component of model fitting but using a slower learning rate, reducing the amount by which the individual initial trees contribute to the final model, and smoothing out the effects of the stochastic selection process. Alternatively, repeat BRT models can be fitted and the predictions averaged, as demonstrated by De'ath (2007) with his "aggregated boosted trees" (ABT). A disadvantage of the ABT approach is that the models used to contribute to the final average are all built on subsets of the training data used within cross-validation, leading to some information loss. We prefer using one BRT model built with a slow learning rate on the entire dataset.

The magnitude of the variation can be surprising, as demonstrated here on the case study data of Elith *et al.* (2008). Using the cross-validation *gbm.step* function (see online tutorial and code), the 1000-site training dataset, a tree complexity of 5 and bag fraction of 0.5, we determined the optimal number of trees for each of five learning rates (Table A1). These were then used as the fixed number of trees in a BRT model, fitted 10 times to the same data; for each repetition the fitted model was then used to predict to 1000 independent sites. The mean and variance for the fitted and predicted probabilities were then calculated over the 10 repeats (Fig. A1). Results clearly show how the between-model variances in the predicted values are reduced when a slower *lr* is used (Fig. A1, left column). A similar effect is also seen in the fitted values for the training data (Fig. A1, right column), but as expected, variances for these are smaller than those for predictions made to new sites.

Table S1: Learning rates and numbers of trees used to fit BRT models to the training data (1000 sites). These were estimated using cross-validation and *gbm.step* (see online tutorial)

Learning rate	Number of trees
0.025	250
0.01	600
0.005	1200
0.0025	2400
0.001	5000

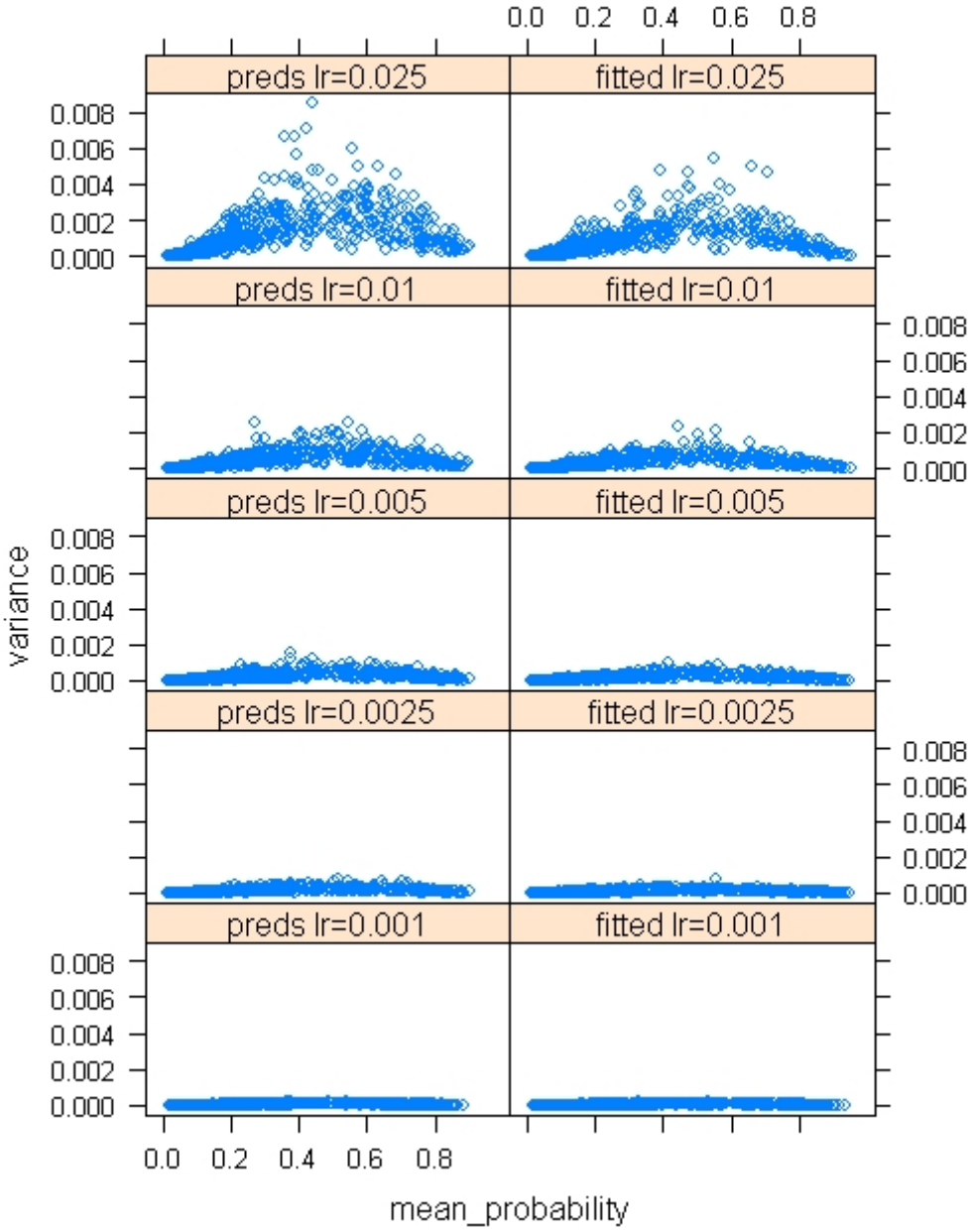

Figure S1: Variances in predicted (left column) and fitted values (right column) as a function of probability (x axis), calculated across 10 repeats of BRT models fitted to data from 1000 sites with learning rates from 0.025 (top panel) to 0.001 (bottom panel). For numbers of trees in each model see Table A1.

Appendix S2: Simplifying the predictor set.

Our strategy for elimination of non-informative variables involves simplifying the model by dropping the least important predictor, then re-fitting the model and sequentially repeating the process until some stopping criterion is reached (e.g., the reduction in predictive performance exceeds some threshold). This simplification process is run within a 10-fold cross-validation (CV) procedure, progressively simplifying the model fitted to each fold, and using the average CV error to decide how many variables can be removed from the original model without affecting predictive performance (Fig. A2, and see tutorial and code online)

In the case study of Elith, Leathwick & Hastie (2008) the two least important variables in the original model fitted to the training set of 1000 sites had a total influence of only 3% in the full model. They were removed through simplification, resulting in a minor reordering of contributions for the retained predictors (Table A2).This had virtually no effect on predictive performance, estimated to be 28.2% deviance explained and AUC = 0.86 on independent data (Table A3). This is a good result – the simplification resulted in a more parsimonious model, without degradation of model fit. In smaller data sets, we have frequently observed an improvement in estimated predictive performance with model simplification.

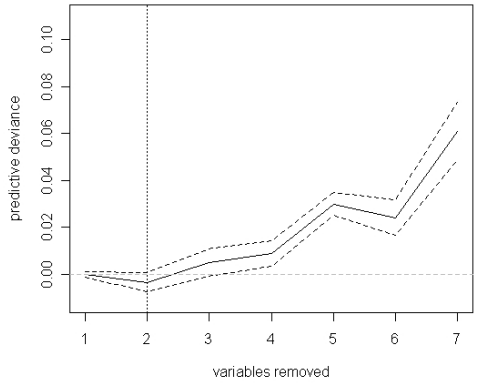

Figure S2 - Simplification of the model presented in Table 2 and Figure 6 (Elith *et al.* 2008), showing that removal of 2 predictors (vertical line) does not adversely affect model predictive performance. The solid line indicates the mean change in predictive deviance, and the dotted line one standard error, calculated over the 10 folds of the cross-validation.

Table S2: Summary of the relative contributions (%) of predictor variables for boosted regression tree models developed with cross-validation on data from 1000 sites and a tree complexity of 5. The base model (Elith *et al.* 2008) was fitted with all 12 predictors, and two were removed with simplification.

Predictor	Base model	Simplified model
SegSumT	24.7	25.3
USNative	11.3	12.1
Method	11.1	11.1
DSDist	9.7	10.2
LocSed	8.0	8.1
DSMaxSlope	7.3	7.9
USSlope	6.9	6.9
USRainDays	6.5	7.1
USAvgT	5.7	5.6
SegTSeas	5.7	5.6
SegLowFlow	2.9	-
DSDam	0.1	-

Table S3: Characteristics of models and their predictive performance, as evaluated on the independent 12,369 records, within a cross-validation, or on training data. All models developed with cross-validation on data from 1000 sites , learning rate of 0.005, tree complexity of 5, using variables detailed in Table 1, Elith, Leathwick & Hastie (2008)

		Base model	Simplified model
Number of sites		1000	1000
No. trees		1050	1000
No. predictors		12	10
% deviance explained	independent	28.3	28.2
	CV[1]	31.3 (0.96)	30.4 (0.97)
	train	52.6	51.6
AUC[2]	independent	0.858	0.858
	CV[1]	0.869 (0.015)	0.868 (0.014)
	train	0.958	0.955

[1] Mean, with standard errors in brackets
[2] AUC = area under the Receiver Operating Characteristic curve

References

De'ath, G. (2007) Boosted trees for ecological modeling and prediction. Ecology, 88, 243-251.

Elith, J., Leathwick, J.R., & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*.

Friedman, J.H. (2002) Stochastic gradient boosting. *Computational Statistics and Data Analysis*, **38**, 367-378.